ENERGY AND QUALITY ANALYSIS OF WINE BOTTLE DATA

by

Cansu Dursun

Submitted to Graduate School of Natural and Applied Sciences
in Partial Fulfillment of the Requirements
for the Degree of Master of Science in
Computer Engineering

Yeditepe University
2022

ENERGY AND QUALITY ANALYSIS OF WINE BOTTLE DATA

APPROVED BY:

Assist. Prof. Dr. Onur Demir ....
(Thesis Supervisor)
(Yeditepe University)

Prof. Dr. Taner Akbay ....
(Yeditepe University)

Assoc. Prof. Dr. Mehmet Göktürk ....
(Gebze Technical University)

DATE OF APPROVAL: ..../..../20..

I hereby declare that this thesis is my own work and that all information in this thesis has been obtained and presented in accordance with academic rules and ethical conduct. I have fully cited and referenced all material and results as required by these rules and conduct, and this thesis study does not contain any plagiarism. If any material used in the thesis requires copyright, the necessary permissions have been obtained. No material from this thesis has been used for the award of another degree.

I accept all kinds of legal liability that may arise in case contrary to these situations.

Name, Last name ……………………………………

Signature …..………………….……………………..

ACKNOWLEDGEMENTS

It is with immense gratitude that I acknowledge the support and help of Professor Onur Demir. Pursuing my thesis under his supervision has been an experience that broadens my mind and presents an unlimited source of learning.

I would also like to thank Ahmet Doğan and Can Koloğlugil who assisted me with glass science in this study. Without their support, this thesis could not have been successfully conducted.

Also, I thank my family members Nazlı Yıldırım and Rafet Yıldırım for their support.

Finally, I would like to thank my husband Fatih Dursun and son Akif Barlas Dursun for their endless love and support, which makes everything more beautiful. This accomplishment would not have been possible without them. Thank you.

ABSTRACT

ENERGY AND QUALITY ANALYSIS OF WINE BOTTLE DATA

Glass materials have gone through many improvements from past to present. From the first day the glass was built to this day, a large number of glass materials can be produced thanks to the great developments in technology. All these glass materials types are heated at elevated temperatures to have glass melts and can be produced by different production techniques.

There are many prestigious glass manufacturers globally with a high degree of specialization in the glass industry. These companies have started to make profits for their companies by using the data and applying all the results from the data in their production. This thesis aims to analyze wine bottles, which are a branch of glass packaging production. In addition, it is to examine the parameters that may affect the energy and glass quality and to benefit the glass industry with the results obtained. The data examined in the research are anonymous datasets to determine various variables, including the numerical values of the glass production process, furnace structure, and energy consumption. To understand the detail of the data, support is obtained from production engineers specializing in glass production.

In this work, a big data visualization study on the glass industry has been carried out. Within the scope of the research, the glass furnace, quality, bubbles, energy, and production data in the big data structure of the 2019 and 2020 years are studied. With graphics produced using the Python programming language, infographics in the functional, and dynamic structures are obtained in which the interaction of glass properties with each other can be seen. In addition, correlations are created over the data sets, and parameters that are in a positive or negative relationship with each other are determined. The effect of furnace traction power on production quantities is presented with visuals in graphic form created. Moreover, the daily amount of natural gas consumption in a year is graphed.

Since the variables include different types of missing data in the studied dataset, the missing data removal process is an integral part of this study. Once the missing data removal processes are completed, descriptive statistics are obtained. The relationship structure

between the variables is examined using linear regression and multiple regression methods and correlation tests and the levels of significance among each other. Thus, it is aimed to evaluate the parameters that can be effective in determining the glass quality and glass color.

ÖZET

ŞARAP ŞİŞESİ VERİLERİNİN ENERJİ VE KALİTE ANALİZİ

Cam malzemeler geçmişten günümüze birçok iyileştirmeden geçmiştir. İlk cam yapıldığından beri, büyük gelişmelerin arkasındaki teknoloji sayesinde çok sayıda cam malzeme üretilebilmektedir. Tüm bu cam malzeme türleri, yüksek sıcaklıklarda ısıtılarak cam eriyik haline getirilmekte ve farklı üretim teknikleri ile üretilebilmektedir.

Dünyada, cam endüstrisinde yüksek uzmanlaşma derecesine sahip birçok prestijli cam üreticisi bulunmaktadır. Bu firmalar veriyi kullanarak veriden çıkan tüm sonuçları üretiminde uygulayarak firmalarına kar elde ettirmeye başlamıştır. Bu tezin amacı da cam ambalaj üretiminin bir kolu olan şarap şişelerinin analizini yapmaktır. Bununla birlikte, enerji ve cam kalitesini etkileyebilecek paramatereleri inceleyip elde edilen sonuçlarla cam sektörüne fayda sağlayabilmektir. Araştırmada incelenen veri anonim veri setleri olup içeriğinde cam üretim süreci, fırın yapısı ve enerji tüketimine ait numerik değerler bulunmaktadır. Verilerin detaylarını anlamak için cam alanında uzmanlaşmış üretim mühendislerinden destek alınmıştır.

Bu tezde, cam sektörü üzerine büyük veri görselleştirme çalışması yapılmıştır. Araştırma kapsamında, 2019 ve 2020 yılına ait büyük veri yapısındaki cam fırın, kalite, habbe, enerji ve üretim verileri ile çalışılmıştır. Python programlama dili kullanılarak üretilen grafikler ile; cam özelliklerin birbiri ile etkileşiminin görülebileceği fonksiyonel ve dinamik yapıda infografikler elde edilmiştir. Bununla birlikte veri setleri üzerinden korelasyonlar oluşturulmuş, birbiri ile pozitif yada negatif ilişkide olan parametreler belirlenmiştir. Power BI programı ile oluşturulan grafik biçimindeki görsellerle de üretim miktarlarında fırın çekiş gücünün etkisi sunulmuştur. Ayrıca Tableau Reader programı ile doğalgaz tüketiminin bir yıl içerisindeki günlük kullanım miktarı grafiklenmiştir.

Üzerinde çalışılan veri setinde, değişkenlerin farklı tiplerde eksik veriler içermesi nedeniyle, eksik veri giderme süreci bu çalışmanın önemli bir kısmını oluşturmuştur. Eksik veri giderme işlemleri tamamlandıktan sonra, tanımlayıcı istatistikler elde edilmiştir. Değişkenler arasındaki ilişki yapısı lineer regresyon ve çoklu regresyon yöntemi ile ilişki

testleri ve birbirleri arasındaki anlamlılık düzeyleri incelenmiştir. Böylece cam kalitesini ve cam rengini belirlemede etkili olabilecek parametrelerin değerlendirilmesi amaçlanmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

Na	Sodium
Mg	Magnesium
Al	Aluminum
Si	Silicon
K	Potassium
Ca	Calcium
Ba	Barium
Fe	Iron
SiO2	Silicon dioxide
Al2O3	Aluminum oxide
Fe2O3	Iron oxide
TiO2	Titanium dioxide
CaO	Oxocalsium
MgO	Magnesium oxide
Na2O	Sodium oxide
K2O	Potassium oxide
SO3	Sulfur trioxide
Cr2O3	Chromium oxide
CoO	Cobalt oxide
Pb	Lead
ASS	Lower spec boundary
ÜSS	Upper spec boundary
AUS	Lower warning limit
ÜUS	Upper warning limit
Müş_ASS	Customer lower spec boundary
Müş_ÜSS	Customer upper spec boundary
MAR	Missing at random
MNAR	Missing not at random
MCAR	Missing completely at random
CAM_K	Cullet
F_ak_dogalgaz1	Natural gas_1 flow

F_ak_dogalgaz2	Natural gas_2 flow
Cam Yüzey Optik	Glass surface optic
Riser Yüzey Optik	Riser surface optic
Renk Tanımı	Bottle color
Kabul	Accepted products
Ret	Rejected products
Taban	Furnace bottom
Fırın_çekiş	Pull
IS machine	Individual section machine

1. INTRODUCTION

Throughout history, data has been a tool that people have used to eternalize their existence. Even before people discovered the alphabet, they have created a method of conveying events, socio-cultural activities, tools, lifestyles, and clues about traditional clothing to us through visualization.

Nowadays, every minor data is a potential source of data and piles up to form big data such as operational, personal, historical, production and consumption. Many projects and academic studies are carried out to understand big data and draw meaningful results. Recently, it has become an indispensable phenomenon for companies. Since it is based on the analysis of real data, it allows us to make the right decisions in many different areas such as reducing costs, spending on the right channels, saving labor, and developing products that meet expectations.

At the same time, using big data and statistical methods contributes positively to the increase of sales of the companies, the more planned progress of its production, high-quality products, and the development of financial strategies if it is looked at the financial side. In this context, it also enables companies to be one step ahead of their competitors by predicting factors that may directly affect facts such as profitability and productivity. If a company has a regular data network, it can predict the daily-monthly-yearly sales rate and the actions that can be taken to increase its sales and develop its strategies in this direction. Even a small-scale manufacturing company directs its production methods within the statistical results presented by the data and increases the rate of production, sales, and profitability. It is also used in the banking sector, where customer behavior is of great importance and must be followed. In the field of health, it can also be used to make improvements in the education system on important issues such as early diagnosis of diseases by utilizing data visualization or drug development. Another benefit of big data, security units are used to prevent crimes. Hereby, most criminals are found thanks to the data.

However, understanding the analysis of big data can be difficult because of its maths and statistics so visualization plays an important role. It is possible with visualizations to convey the inferences about the data accurately and simply and to make a remarkable study. The techniques developed in data visualization have provided significant advantages and have become the preferred method of presenting information by increasing the importance of visualization.

While performing data analysis, visualizing the data rather than reading it makes it easier to perceive. Visualizations with the support of experts help shorten the completion time of projects by looking at the right points and seeing dynamic data with constantly updated graphics helps speed up the work.

Until today, a lot of research has been done about the glass structure in the glass industry and studies on big data have been accelerated. Data usage is increased with new projects such as more production, less energy consumption, and removal of environmental pollution.

In this research, the wine bottle data is analyzed by examining the important features in the datasets and making a comparison between the related data. To understand values, visualization has provided a great advantage in using time efficiently and getting quick results.

The study consists of three parts. In the introduction, information is given about the subject, purpose, scope, and contents of the study. In the first part of the study, information about glass is given and glass data is explained. In the second part, big data visualization applications are made with glass anonymous data. Data preparation processes such as the data set used in the application, the process of removing missing data, categorizing the variables, creating a new variable, and glass data visuals are presented. Missing Data Analysis and Regression Analysis topics used in the application are explained in detail. As a result of the analysis, graphs, correlations of different data sets, and regression analysis tables are added.

1.1. MATERIALS AND METHODOLOGY

With the automated production process, the melted glass is cut into drops at 1200 degrees at certain weights and passes through the production line phases such as the first forming, hot coating, bonding, stress reduction, cooling, and increasing the lubrication resistance.

The shaped glasses that are not found good enough in the automation system and are labeled as poor-quality products are automatically controlled and separated at each tab and left to melt. Faulty products that escape the notice of the machines are tested with samples taken by the quality laboratory. Low-quality products are separated according to their serial numbers and sent to be melted. Each pattern is checked hourly according to product shape and customer requests. The low-quality products in the automation system are detected making very strong determinations. It is observed that the separated products in the obtained dataset are not recorded. Therefore, when machine learning algorithms are applied to the data set, it is seen that the accuracy rate of high-quality products reaches 99% - 100%. It is said that the reason for such a high rate to be found by the professionals of glass science is the poor-quality glass bottle series that are not recorded. Since it is seen that the results do not reflect the reality in this research that I started to improve the glass quality, each data set is examined separately and the properties related to the glass are tried to be explained with visuals and it is aimed to have an idea about the values.

In the review, there are anonymous datasets for the years 2019-2020 regarding the furnace, bubbles, energy, quality laboratory, and production departments of the glass wine bottles. According to the information received from the experts, the data recorded in the 2019 data set is not entered in carefully as in 2020. For this reason, it is very important to consider the results obtained in 2020 more than the results of 2019 in terms of inferences. This information has been taken into account during the writing of the thesis.

Although there are many columns in the energy data, after the information is obtained from the experts, visualizations are applied by selecting the distinctly separated natural gas columns in the data set. In the energy data set, there are usually columns with the automatic entry within the automation system. Since these values are also seen and explained in the furnace set, it is said that only the natural gas columns can show the consumption clearly. In the general framework, the densities are determined by looking at min, max, average, and standard deviation values in which intervals these important columns and an idea about the values are obtained.

On the row where the quality standard value is determined on the bubble data, some deductions are made regarding the bubble numbers of the products by visualizing the wine bottle data.

Looking at the analysis from a different perspective, it is tried to explain how the production changes daily by using the "*date*" column on the production data and by dividing the data into daily and weekly with visual tools.

Quality tests are entered with the "*Kontrol_Türü*" (Control type) feature in the data determined to belong to the quality laboratory among the data sets, and it is observed that different processes are performed for the samples taken at certain times of the day. In addition, visual tools are used to obtain information about the specification limits and specification statuses according to the quality status, and a conclusion is drawn about the values. In addition, a hypothesis is established with the OLS regression test method, which is established to find the factors affecting the quality, and as a result, it is seen that the values of cullet rate, batch humidity, glass surface, air/fuel ratio, the bottom temperature of furnace columns are important. When the spec status is examined in the glass component data, the significance of the values is the same, although the glass colors in the 2019 and 2020 datasets are different. In the 2019 dataset, it is seen that all glass components are significantly as effective as the 2020 dataset with the same regression method.

Glass components have been particularly studied. In this section, visualizations are used to understand the interaction of the components with each other. In addition, the hypothesis and the components that give the color of the glass, as well as the spec, are compared in the 2019 and 2020 datasets. In addition, another data analysis work is to examine the relationship between glass color and cullet rate. It is clear from this observation that the rate of the cullet plays an important role in determining the color of the glass. In this section, a comparison is made with Anova one-way and OLS regression and it is observed that the results matched exactly.

Trying to understand the data with visualization and making inferences has been one of the most important goals. In this study, the creation of these tools, which facilitate the analysis by exemplifying the data and help us get to know the data more closely is explained.

1.2. GLASS AND GLASS INDUSTRY

Glass can be defined as a fragile substance that retains its form when it is shaped and cooled by melting silica soda, potash, and lime at high temperatures.[1]

Obsidian, a natural stone known to be formed by the sudden cooling of lava in the early ages, was introduced and it started to be used in the production of various materials such as jewelry. It is known that the glass was discovered by chance as a result of historical excavations. It was valuable to kings in ancient times and its history dates back to 5000 years ago. [2]

Thousands of years ago, ancient Egyptian pharaohs decorated their surroundings with glass while they were alive and when they died, leaving important specimens for archaeologists to uncover. For example, a decorative writing palette and two blue-colored headrests made of solid glass were found in Tutankhamun's tomb. Tutankhamun's funeral mask, on the other hand, was seen to consist of gold and blue glass inlays to frame the pharaoh's face.[3]

In the 1980s, blue glass was discovered thanks to the work of Polish chemist Alexander Kaczmarczyk, who discovered elements that give the glass a vast blue color, such as aluminum, manganese, nickel, zinc, and cobalt. By examining the relative amounts of these elements in ancient glass samples, Kaczmarczyk's team traced the cobalt ore used to give glass its blue color in certain Egyptian oases to the mineral source.[4]

After Kaczmarczyk, Shortland began another search to understand how the ancient Egyptians worked with the element cobalt during the first millennium B.C. A sulfate-containing compound called alum does not normally mix into glass. But Shortland and his colleagues reproduced a chemical reaction in vitro that Late Bronze Age artisans may have used to create a compatible pigment. Thus, they produced a dark blue glass that looks like Egyptian blue glass.[5]

Nowadays, the colors of the glass have stunning improvements. Also, the production stages of glass, which is used in many areas and is an important element for a healthy life, has advanced and made life easier with technological developments.

In the production of glass, the main material of glass is sand, which is one of the most abundant materials in nature. The material made of pure silica glass, obtained only by melting sand, is very fragile. Soda is added to the mixture to increase its strength and decrease the melting temperature. Minerals such as limestone, dolomite, and feldspar are added to eliminate the weakness against water and chemicals. Secondary raw materials in the form of a cullet are added to this mixture called blend and it becomes ready for melting in the furnace.

Glass can be made by a wide variety of methods in the glass industry but it is obvious that the tremendous majority are still produced by melting bulk components at a raised temperature. This method always involves the selection of raw materials, arrangement of the relative proportions of each to use in the bulk, and weighing and mixing these materials to provide a homogeneous starting material. The raw materials undergo a series of chemical and physical changes to produce the melt during the initial heating process. Conversion of this melt to a homogeneous liquid may require further processing, including the removal of any unmelted bulk residuals, impurities, and bubbles.

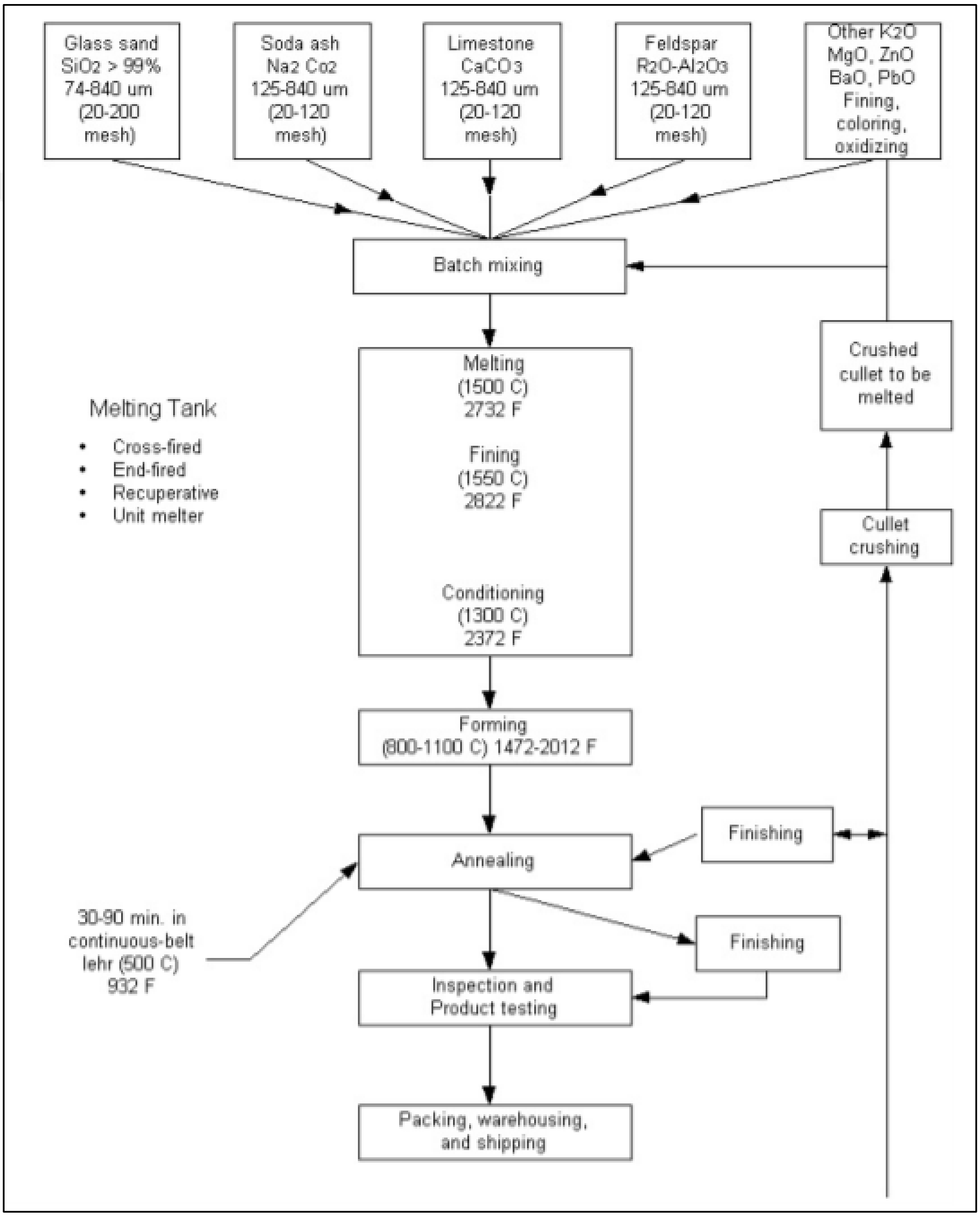
Production of bottle products needs forming of specific shapes, also heat treatments to remove stresses generated during the cooling process or to produce glasses strengthened by thermal tempering. However, the bulk components are melted at high temperatures between 1500°C and 1600°C in the glass furnace, it turns into soda-lime glass, which is a honey-like fluid of the desired color. The most important glass type is soda-lime which mainly gives the lower and upper percentage ratios of glass components which are silicon oxide, sodium oxide, and calcium-oxide as seen in Table 1.1. Glass components percentage ratios of soda-lime glass type [6].

Table 1.1. Glass components percentage ratios of soda-lime glass type

Component	Lower Percentage	Upper Percentage
Silicon dioxide	71%	75%
Sodium oxide	12%	16%
Calcium Oxide	10%	15%

In the next stage, the molten glass is cut into drops and automatically blown-blow or press-blow methods are applied according to the glass packaging type in the machines and shaped according to the desired shape. The glass products obtained are then cooled programmatically, subjected to precise quality inspections on the line and defective ones are separated. Thus, the production process is completed which is seen in Table 1.2. Production Process(Adapted from the Glass Industries of the Future Report, Pellegrino 2002). The glass can be poured, blown, pressed, and molded into a wide array of shapes. Therefore, the glass industry serves sectors covering different products, applications, and markets.

Table 1.2. Production process(Adapted from the Glass Industries of the Future Report, Pellegrino 2002)[7]



The glass industry has a goal of adequate maximum thermal efficiency, maximum glass quality, a maximum lifetime for furnace and refractory, minimum pollutants production and emission, and minimum energy consumption.[8]. It is the only way to achieve that furnace cleaning maintenance and repair must be routine. And also, furnaces must be suitable for the production amount and must be controlled at regular intervals. It is seen that the production of poor-quality glass in the furnace without regular maintenance increases, the consumption of electricity, energy, and natural gas increases and causes environmental pollution. [9]

The glass industry relies on electricity and natural gas to supply the bulk of its energy needs. Glass melting consumes the most energy of all the production processes and is accomplished using a combination of natural gas and electricity. Also about half of the electricity used in glassmaking is for process heating, mostly in electric boosting of furnaces.

The production of higher quality glass, which has the same importance as energy consumption by the glass industry, is also a very important issue. Parameters such as the number of bubbles in the glass, the formation of cracks in the glass, and, the durability of the glass should be carefully examined by the production experts. In addition, it is important for all glass manufacturers to produce by the quality standards determined in the production process and to transfer the quality glass to the customer. Because the company's bond with the customer is the quality products that are sold. Moreover, despite the financial benefits of the energy, electricity, and natural gas consumed, the production of quality glass is also at the top of the production pyramid for the companies to develop, grow and create a good image in the global market.

There are many glass manufacturing companies in the global market. These manufacturers are operating in the world's glass industry's main fields of activity in terms of flat glass, glassware, and glass packaging.

1.3. BIG DATA RESEARCH ON THE GLASS INDUSTRY

Turkey's leading companies in the glass industry, have carried out valuable academic studies as well as research with their R&D department and have guided valuable studies since awareness of big data. It is known that there are some big data projects conducted by branches of glass companies in line with the needs of plant technicians. The aim of these projects is to achieve different gains such as less energy consumption and keeping production at the maximum level. At the same time, the profit rates and investments of the companies increase thanks to these projects. Research in glass companies to produce quality products also ensures that the product reaches the customer with higher quality. These researches and analyses increase the worldwide awareness and prestige of glass companies in our country. Utilizing the power of data, glass companies appeal to more and more people day by day and can respond to the supply of the market.

At this point, Turkey has glass industrial enterprises that always strive to use technology at the highest level, has also increased its tendency towards this issue and, has increased its steps regarding digitalization.

It is known that there are values that must be focused on or paid attention to obtain better quality products owing to artificial intelligence in the automation system. Understanding and utilizing data is the basis of this workflow in the world. Therefore, most companies work on digitalization significantly. As long as the data can be understood, it can assist in the discovery of projects that involve understanding glass structures with artificial intelligence, machine learning and, deep learning methods, finding relationships, validating and, viewing data using statistical and numerical optimization disciplines.

In structural glass engineering, the use of artificial intelligence, machine learning, and, deep learning methods, which are used from the inspection of the cut-edge to the appearance of the glass, has increased. For example, crack contours were examined manually with the image processing program and then improvements were made in glass production, but thanks to methods of AI and especially the algorithms from the field of AI in computer vision, the step of detecting the glass cutting edge manually can now be automated. The software libraries of artificial intelligence such as Tensorflow and Keras are very beneficial in the development of projects and the production of higher-quality glass.[10]

In the global glass industry, big annual revenue companies that continue to exist in the market are trying to increase production efficiency and quality, reduce costs, support classify different wastes in recycling facilities for waste management, and solve logistics and supply chain problems by using data and artificial intelligence. It is much easier in today's technology to encounter the needs of manufacturers thanks to machine learning methods and artificial intelligence.

1.4. SUBJECT AND PURPOSE OF RESEARCH

The research aims to analyze and visualize energy, production, glass chemicals compositions, quality laboratory, furnace data, and also bubble values datasets by working on glass data. In addition, it is to ensure that the factors affecting quality and energy are understood with visualization methods and to create tools.

Moreover, the purpose of this study is to measure how visualizations can be enriched, to recognize the data with visualizations, and to obtain information about the ratios of the energy on the data, especially the factors that affect the quality. When it is intended to make an inference about quality, certainly the available datasets aren't enough to make a full determination in this situation. Working on visualization became more attractive after concluding that it is rather difficult to estimate quality. Since it is also willed to create visual tools on quality, care is taken to move forward and derive inferences in this direction by using the opinions and experiences of factory experts to understand the variables that directly affect the quality and other factors. Then, visual tools are plotted for energy, production, and bubble data as well as quality data.

The anonymous datasets from various departments used in this research are the wine bottle product family among the product types produced. By observing these data, it has been tried to obtain certain results and to get outputs with specific visualizations.

In this research, The wine bottle datasets are inferred and visualization maps are presented. The graphs and maps are interpreted and ideas are exchanged with experts in the research. Thanks to the feedback from experts, the course of the established hypotheses changes and the differences found are evaluated.

There are many visualization tools. They are explained and opinions about possible inferences are presented.

1.5. INTRODUCTION TO PROCESSING OF WINE BOTTLE DATA IN GLASS DATA

The data obtained from anonymous that has a daily capacity of 500 tons of glass melting is the main source of this research. In the automatic system, the melted glass is cut into drops at 1200 degrees at certain weights, and the IS(Individual Section) machine passes through the machines on the factory line such as initial shaping, hot coating, bonding, stress reduction, cooling, and lubricity resistance. In the automation system, poor-quality products are separated at each tab and left to melt. Faulty products that escape from the eyes of the machines are tested with samples taken by the quality laboratory. Products with low quality are separated according to their serial numbers and sent for melting. Also, samples are made checks for each mold per hour according to product shape, product card, and customer requests.

According to the information received from the experts, wine bottles are produced from soda-lime-silica glass, and the main raw materials are sand, feldspar, feldspathic sand, perlite, limestone, dolomite, soda, sodium sulfate, cullet. Apart from the main raw materials, colorants, decolorizers, and auxiliary raw materials that facilitate melting are also used as inputs. The rate of use of raw materials, whose input control is carried out, is determined as a result of periodic glass analysis and applied to the blend recipe. However; hot coating, cold coating, dimensional measurements, inhomogeneity, weight, volume, internal pressure resistance, proofing test, thermal shock test, vertical load resistance, impact resistance, thin wall, annealing, and corrosion specification values are also examined and these data are manual. The specification status for each feature is different and the production is made according to the quality procedure. The automation system, which is based on quality values, does not deliver products that are found out of specification at some points to the destination. A series of tests are applied with the quality criteria determined by the quality laboratory with the samples taken from the products that reach. The results of the tests are recorded and the results are analyzed and inferences are made.

By utilizing data to increase production efficiency and reduce errors, it is predicted that higher quality products can be obtained by saving energy in the production process. The datasets used in the research represent the production process, furnace values, bubble values, and quality data.

1.5.1. Dataset Used in Research

The dataset used in the study is the furnace, glass chemical components, production data, bubbles data, and quality laboratory data for the 2019 and 2020 years. It is expected to yield more significant results due to the size of the 2019 data. Wine bottles have been selected by me from the glass packaging product range to reduce product diversity within the facility and to deal with a particular product. Then, the data of wine bottle products are extracted with the product line from the whole data set through the Python Programming Language, analyze the quality and production and enrich with visualizations. In addition, visualization graphs are plotted by examining the energy, bubble, and glass components data. Noticing that there are some unregistered and unimportant features in the data set, it is cleared over the data set. Then, it is ready for visualization in this way, interpretation, and analysis.

There are datasets kept from the initial state of the production to the packaging part. Each department keeps the data and analyzes it within itself. A problem that product-based tracking was not possible arose during the analysis as all the data were not kept in a single set over a system. Therefore, evaluating the data separately helped me to get more accurate results.

1.5.2. Variables Used in Research

In this section, the datasets used during the study are introduced and the parameters that are considered important by the glass experts are explained. Detailed information about the data is presented in the Appendix section.

Natural gas columns are used for inference in energy dataset, related to production and quality data columns, glass components, valuable columns in furnace blend data by getting feedback from the specialist, especially features which are effective on glass quality in plants that are analyzed by using all in this research.

It is necessary to evaluate the missing values for quality, production, energy, bubbles, and furnace data separately. There are 21 columns and no missing values in the glass components dataset. There are 36 columns in the furnace data and the contents of Oxygen upper left, Oxygen upper right, Oxygen lower left, and Oxygen lower right data which keep the oxygen percentage are not available. There are 41 columns in the quality laboratory data and there is no data only in the content of the "*Mastör*" and "*Yorum*" columns. There are 13 columns in Bubble data and the energy data has 206 columns. Both have no missing values in their content. There are 28 columns in production data and no missing values.

Control machines features are used on a quality laboratory dataset that is assigned with its controlled samples. As is seen in Figure. 2.4., "*DSG_B*" is the large range between control machines that are used for dimensional measurement control. "*DSG_C*", and "*DSG_D*" which are the other most popular machines in the proportion of other machines are used also for dimensional measurement. "*RPT2_Lab1*" is related to getting information about pressure tests. *AGR_Kafa_Kaplama_1* and *2* are used for testing head and body hot coating and "*Terazi_1_Metter_Toledo*" is used for weight. ISIS is the control machine for the wall(cidar) and abnormal glass. However; "*Elektrophysik*" has not been used. "*AGR_Dik_Yük_2*" is a control mechanism that is made quality control manually and the system is not getting feedback that's why there isn't any value in the dataset. Control type also shows what type of controls the products have gone through in the control machines. This column provides the visibility of the test diversity of the products.

It is occurred a map by using values of cullet, furnace bottom temperatures, number of drops, glass components rates, control type, control machines on manual or automatic, specification status, natural gas consumption, bubble values, and some important variables.

Specification status feature is very important on quality data because of understanding quality status whether it passed or not. It gives an idea about whether the product is of good quality. It makes sense about glass quality through this column.

Ass value refers to product lower spec limit, Üss - upper spec limit, Üus - upper warning limit, Aus - lower warning limit, Müş_ass - customer lower spec limit, Müş_üss - customer upper spec limit, Müş_hedef - customer target value. These values should be a specific range to obtain quality glass so they are directly related to spec status.

There are 206 columns, 119.826 entries, and no missing data in the energy dataset. Natural gas and electric consumption values are important columns to understand the furnace energy issue. Figures in the Energy section are seen to show how the values can proceed on a scatter plot. At the same time, Energy consumption, Electricity consumption, Natural gas consumption, Cullet Rate, Batch Humidity, Air-Fuel Ratio, Throat 903, and Throat Riser values play an important role in making inferences about energy.

Bubble concept, which is a very important criterion for understanding glass quality, A and B furnace lines were taken to be examined on coding independently of wine bottle products. In the dataset, the values of bubbles, seeds, and blisters were separated as per 1 kg and 30

g. To understand its effect on quality, its distribution with Figure 2.72. Total30gr Distribution by Specification Status will be illustrated. As for the quality, the "*Toplam30gr*" column is used by experts to understand the Bubble values, so this column has to be focused on. According to the information given, it is considered that if the "Toplam30gr" column value is less than 64, it can be accepted as good as quality. Otherwise, when it is greater than 64, it is said that it is not good quality.

The correlation of the glass components with each other is examined in 3.3. Correlation between glass components data variables section. The closely related glass components are explained and the relationships are inferred as linear or not linear to have an idea about the glass proportions in the visualization chapter.

According to the feedback from the experts in the production data; Production Efficiency calculations (%) can be calculated with Production (kg) / furnace_pull (kg). An average efficiency value of 0.84 is obtained as the values can be illustrated graphically in Figure 2.82. The relationship between the *"net_çalışma_ süresi*", "*fiili_ağırlık", "üretim(kg)"* and "*Fırın_çekiş*" columns in the data set are also very important columns for understanding production capacity.

While examining the production data, a prediction about the future production is made by using the date column and the production capacity column. In this estimation, a one-year data set is used and it should be noted that the margin of error is high because it is insufficient to make inferences. In addition, the date column in the production data is separated as day, month, and year, and is examined the average production efficiency.

By the way, there is some complication between production and the furnace data set. The data recording rate of the bottle colors in the production data set and the furnace datasets vary. This change is clearly explained with visualizations in the relevant visualization sections.

After looking at the relationship between the glass components in the data set, it is applied the OLS regression model to examine p-values and which components are associated with quality are determined. Thus, in Table 4.5. 2019 OLS regression table for spec status and Table 4. 6. 2020 OLS regression table for spec status; all elements about specification status can be seen.

There are columns in furnace data that show the consumed electricity consumption. These values, in units of KWh, explain the amount of electricity consumed in the furnace blend chamber, in the hot zone, and, in total as can be seen in graphs in Section 2.1.2. After examining the furnace data with visuals, the OLS regression model in Table 4.2. 2019 OLS regression table for spec status and Table 4.3. 2020 OLS regression table for spec status is applied to some important features in the data and its relationship with quality is determined by looking at the p-values. In this way, It is understood that features in applied OLS Regression are closely related to the specification status.

It is graphed for each dataset and tries to make an inference to have an idea about these values and to understand the mean median and standard deviation values better.

When the furnace data are examined, there are 97 columns. 24 of these columns appear to be null and 28 of them do not have data in the column at the rate of 70%. After these data are cleaned and prepared for analysis, OLS Regression is applied to measure the effect of glass components on glass color and the results are added to tables.

1.5.3. Preparation of Research Data for Analysis

In data mining applications, the quality of the dataset directly affects the success of the applications. Therefore, in this process depending on the dataset, a lot of time and effort may be spent on processing to make sense.

In this research, it acquired information about datasets, created correlations for each independent data, and, dealt with visualizations between these values which have been studied to draw some conclusions using wine bottle data. It also touches upon the major phenomena in production, energy consumption, bottle color, and glass quality. Since the energy dataset is ready, it is made possible to be studied by clearing the missing values. Even so; if it is willing to examine the quality of wine bottles, it is really difficult to get definite results that the information conveyed by the experts, the reason is that defective products are parsed in the automation system and the data of these separated products have not been kept in the plant's data warehouse. Regardless, it is partially possible to examine the quality dataset among the datasets transmitted from the plant experts. It is combined all the data under mutual columns, which are the common property of the blend data formed in the combination of all items and the data obtained from the quality laboratory kept in the Vertech Database program. It is tried to occur the data organized by assigning 0 values to the missing values. To edit data is proceeded by clearing the features that are ineffective in correlation. Quality tests in the laboratory are entered into Vertech Database System which is used in the plant's internal system with the control type feature, and it is observed that different processes are performed for samples taken at certain times of the day.

To refer to the quality fact, the quality criteria for each glass bottle are different, and the observations and impressions for each different control type are said by experts. Therefore, to test its accuracy, it creates discrete data frames for control types, and, makes examinations, and visualizations.

When looking at the analysis from a different perspective, it is tried to understand how the production of wine bottles changed daily by dividing the data into daily and weekly production amounts, and how many are accepted among the products manufactured. At this point, another output is taken by utilizing data visualization.

1.6. VARIABLES TRANSFORMATION ADDING NEW VARIABLES

Analyzing the variables and adding new variables in line with the requirements of the data set and analyzing the data through those variables have been very useful during the research in terms of understanding and recognizing the data.

It is aimed to insert the new feature that is arranged the contents and transferred it to examine the columns that are difficult to read and not suitable for analyzing data. In this way, the errors are corrected that the program may give and facilitate data visualization and analysis.

Column names in the dataset that would complicate the analysis are arranged and transferred to a separate line. In addition, label encoding and one-hot encoding methods are used to convert categorical string values to numerical data, which will be useful for this study.

Table 1.3. Transformation of variables

Dataset	Column	Method	New column
Production	Date	Variable conversion	Day, Month, Year
Quality	Control Type	Remove spaces	_
Quality	Control Type	One_Hot Encoding	0,1

Quality	Specification Status	Label Encoding	0,1,2
Glass Component	Bottle Color	Label Encoding	0,1,2,3

1.6.1. Adding Date Variables Such as Day, Month, Year

To produce data on a daily and monthly basis, the data taken in a single column based on day-month-year and hour were taken into separate columns thanks to variable conversion, and the date data were separated by Python as a month, day, and year. In the allocated date variables, It managed to visualize the daily and monthly productions and see the upward and downward parts. Apart from that, it can select a specific day and see the most produced product daily with graphical methods in the last part of the research.

1.6.2. Transformation of Wine Bottle Control Types and Color

To categorize the wine bottle, it is used the One-Hot Encoding method which is very common in Data Science. This method assists categorical variables to be represented as binary. This process requires categorical values to be mapped to integer values first. Next, each integer value is represented as a binary vector with all values 0 (zero) except the integer index marked with 1 (one). It is especially used to prepare the data for machine learning algorithms. Because ML algorithm couldn't handle textual data and it always requires numeric values.[11]

It is utilized as a one-hot encoding method for the Control type feature in the Quality Lab Data that is categorical as seen in Figure 2.3. Control type, the column's data is converted to numerical values (1,0) instead of string values and processed into new columns. The purpose of this step is to manage to grasp the changes in the data for different types of control and analyze them from different perspectives. It is shown in Figure 2. 5. Control types for wine bottle types which product type has more percentage for the "Hot and Weight" control type, which is seen to be the most used among the samples taken.

To examine the control types of wine bottles on Python, the "_" symbol instead of the space between words so that the program could detect it. Thus, the analysis was facilitated and the risk of error was eliminated.

Another categorical encoding type is label encoding to convert each value in a column to a number. To make the data understandable or in human-readable form, the data is often labeled in words but it must be converted to numeric for machine learning algorithm-readable form.

An optimization problem is considered in the wine bottle color column, which is entered as text data when the glass components are considered to examine the data set. For the solution, it is predicted that the use of Label Encoding would be very useful in defining data and drawing conclusions. OLS Regression method is applied to the wine bottle by coding the colors as 0,1,2,3 in a separate column. This application can be seen in the Multiple Regression section.

1.7. MISSING VALUES ANALYSIS IN VARIABLES

Missing data is a very common problem in statistical studies. While data exploration, missing values are one of the pre-processes in which the data is prepared for analysis, and it can also be accepted as the first stage of statistical analysis. It is widely known that the existence of missing values in data causes serious problems for example; biassed parameter estimates and exaggeration of standard errors. Most of the missing data assignment methods are interested in datasets containing continuous variables. Before starting the analysis, the issue of which arguments and how to fill or delete missing data is important in terms of obtaining accurate results.

There are three types of missing data: [12]

Table 1.4. Types of missing data

Missing Completely at Random	MCAR
Missing at Random	MAR
Missing Not at Random	MNAR

In Type I, Missing completely at random (MCAR) assumes that missing data consist, randomly, and independently from the other observed or unobserved data. If there is no systematic difference between the observed values of the variables in the data set and the variables containing missing data, it is said that the missing data are completely random. It assumes that missing data consist randomly, and, independently of the other observed or unobserved data.

Type II is also referred to as "Negligible missing data". The random missing data is a very weak assumption. Often data are not MCAR, but they may be classifiable as Missing at random (MAR). In this case, missing and observed observations are no longer coming from the same distribution and this is a crucial distinction between the two methods.

Type III is the situation where the deficiency is due to both observed and missing values, and the situation where the values are non-random missing data.

If the data set is not MAR or MCAR, it is classified as Missing not at random (MNAR). It can be thought that this is the most difficult situation to deal with. There are times frankly when it is impossible to deal with it at all realistically. In summary, if data are missing for reasons that have nothing to do with the values of any of the data variables, the data are missing completely at random. If data are missing for reasons that may have to do with the other variables, but not with this variable, then the variable is missing at random. Missing data mechanism can be ignored if MAR and MCAR status are present but ignoring the MNAR condition will result in erroneous results.[13]

1.7.1. Missing Values Analysis

Missing values are one of the most common problems while preparing the analysis. The reason for missing values might be either interruption in the data flow or privacy concerns on the plant. It decreases the performance of analysis extremely. The simple solution to abolish missing values is to drop to rows or the whole column. Alternatively, a 70% threshold is used for columns and rows to drop the rows and columns which have missing values higher than the threshold percentage. By the way, the data size can be preserved. Replacing the missing values with mean values is a good option for handling columns.

The most important situation is deciding to choose which method is suitable for missing values. In this research, each type of missing value is observed in the datasets. Converting, deleting, or filling in missing values is one of the most important parts of the analysis. Type-I is Missing Completely and Random, which has a higher percentage of missing values. Data conversion, i.e. NaN values can be evaluated as " " to facilitate the operations during analysis. However, the mean value can be assigned in order not to affect the average values while analyzing.

Columns that are seen as insignificant and have no effect during correlation analysis can be completely deleted from the data set and removed from missing values. After progressing the analysis of missing values correctly, we will increase the reliability rate of the results by moving to the data analysis section. How to clear the missing values from the data set varies according to the structure and content of the data. I would like to reiterate that we can move forward by getting to know the data.

1.7.2. Filling Null Values

After combining excel sheets, which would be useful to examine together, null-valued columns were obtained. Since these null-valued columns cause an error while visualizing the data, spaces are assigned 0 to the values. As it can be seen in the visualization, some values are close to 0(zero) points which are missing data in some graphs due to the mean value being close to 0 dot. In this case, the values that reflect the clustered real values should be looked at.

At the same time, it has been deemed unnecessary by experts who have no value in the columns in the glass data, the unassigned columns are completely cleared from the dataset and the functionality of the data has been increased. Columns that are completely removed on the furnace data such as Oxygen regenerator temperatures, belt temperatures, and some furnace Bottom temperatures. In the quality laboratory data, there is no value in the “Master” column, and there is no value in the majority of the “Yorum”(Comment) column. There are no columns without values in Bubble, Energy, and glass components data thanks to the automation system.

1.8. FEATURE ENGINEERING

The feature selection method is intended to reduce either redundant or irrelevant features without incurring much loss of information. And feature selection is also related to dimensionality reduction which does not have to be a subset of the original set of features that aims to reduce the number of random variables under consideration by obtaining a set of principal variables. The difference is that feature selection selects features to keep or remove from the dataset, whereas dimensionality reduction creates a projection of the data resulting in entirely new input features. [14]

Firstly, the missing values are cleared on the datasets. One hot encoding to make improvements and fix the deficiencies and discover that some corrections are required because the programming language which is used in some columns gave a syntax error. Then, they transfer the misspelled columns by creating new columns. The mean, minimum, and maximum values are examined using Python to learn more about columns that have been considered important by experts, and the data are tried to be recognized by examining the distributions of the data through visualizations. Correlations and graphs are used in order to better understand the data related to natural gas consumption, which is important in the energy section. By trying to find the factors affecting natural gas, visualizations are presented to understand where it is consumed more. In quality laboratory data, on the other hand, the non-specification column is very important as it plays a determining role in quality. For this reason, non-specification is in our focus, charts are created for the factor values that affect it, and it is tried to have an idea about the values that should be. In another data set, production data, bubble data, and glass components data, a conclusion was tried to be drawn with graphics without any prior knowledge, you can see the explanations in the visualizations section.

2. REPRESENTING WINE BOTTLE DATA

2.1. DATA VISUALIZATION

A statistical plot or data graphic should balance functionality, interpretability, and complexity, all without sacrificing aesthetics. It can be easy to get an idea from vast amounts of data owing to data visualization. It also allows exploring the outliers in the data. However, there may be situations where the value difference between the data is very large and data can be handled using statistical normalization.

There are many popular programs such as Python, R, Power BI Desktop, Tableau Public, and Javascript to analyze and visualize the data which is providing a great deal of convenience to data scientists.

This part focuses on the visualization of wine bottle data using different tools with python which is mostly used as the programming language. It is aimed to draw meaningful results by using the data visualization libraries matplotlib, seaborn, plotly, and pygal. Tableau Public and Power BI Desktop are also used in a small part. Complex features could be easily graphed with Tableau and Power BI via bar graphs given as an example.

In this section, variables considered important by glass experts in the quality laboratory, energy, glass components, furnace, bubbles, and production data will be examined using visualization methods.

2.1.1. Visualizations of Quality Laboratory Data

By using the graphics in the quality laboratory data, it will be tried to obtain a lot of information such as seeing the value ranges of the variables, examining the number of wine bottle products produced, analyzing how much of the spec values are suitable for the quality in the control types applied within these numbers, and the distribution of the glass cullet rates.

2.1.1.1. Drop Count

The number of drop cuts is the name given to the glass before it is shaped with IS machine. The values in its content were formed by giving the type integer as "*1*" and "*2*" to represent the front and back values of the bottle. Values representing the value of *0* are the values designated as null. These values belong to weight, body hot coating, head hot coating, cold coating, and tension control types and have been entered as unknown which is 89370 quantity. Also, *"2"* represents the front of the bottle value and *"1"* represents the back of the bottle value. There are approximately 2500 bottles that are not controlled on the front side as can be seen in box plots.

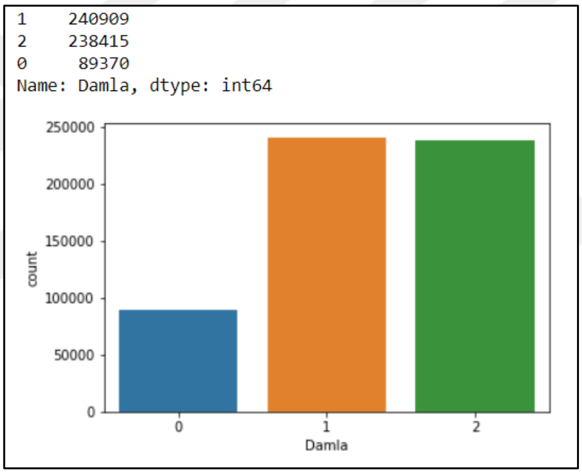


Figure 2.1. Count of "Damla"(Drop) column

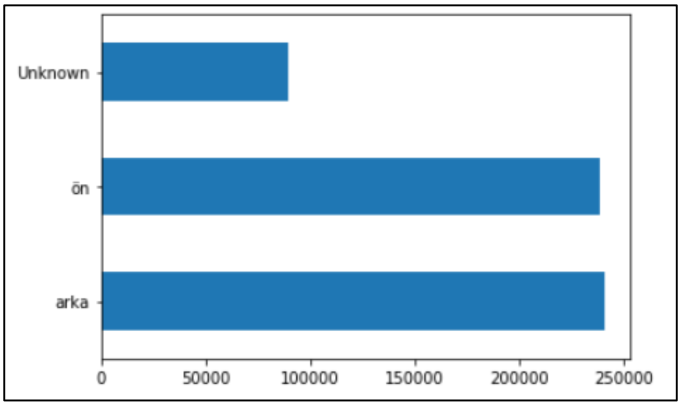


Figure 2.2. Front and back values

2.1.1.2. Control Type

The products sampled per hour in the quality laboratory are made manually and automatically. A small amount of data that was forced manually was entered as a different classification as shown below. In this feature, 92% of the part was realized automatically. The manual part has a very small percentage. Then, pressure control type data has an option "*Manuel Zorla*"(Manual force) on rows sorted by the operator in the Vertech screen. "*Manuel_zorlandı*"(Manual_forced) refers to pressure testing. The automation system is very effective in both the production process and testing of quality in all of the products as seen in Figure 2.3. Control type.

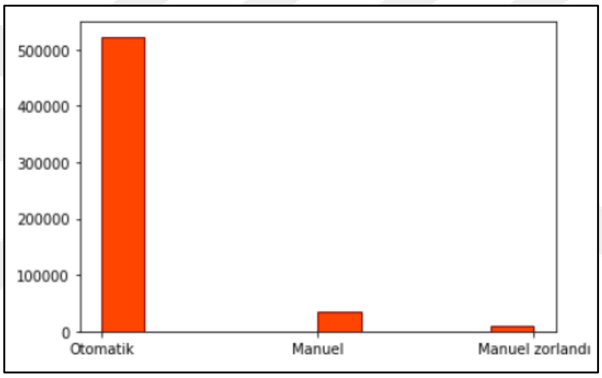


Figure 2.3. Control type

2.1.1.3. Quality Control Machines

Among the quality control machines, there are test machines listed below. It is the most used "*DSG_B"*(Dimensional measurement control_B production line) parameter. The entry is expressed as "*Yok*"(None) which features without access are the control types that are controlled manually. Detail of explanation regarding control machines can be found in part 1.5.2. Variables used in research.

Table 2.1. Quantity of quality machines

QUALITY MACHINES	QUANTITY OF COLUMNS
DSG_B	209907
DSG_C	157840

Yok	86540
DSG_D	57243
Terazi_1_Mettler_Toledo	12902
AGR_Sıcak_Kaplama_1	12561
AGR_Kafa_Kaplama_1	12549
Terazi_2_Mettler_Toledo	9818
RPT2_Lab1	5715
ISIS	2459
ElectroPhysik	676
AGR_Dik_Yük_2	478
AGR_Kafa_Kaplama_2	6

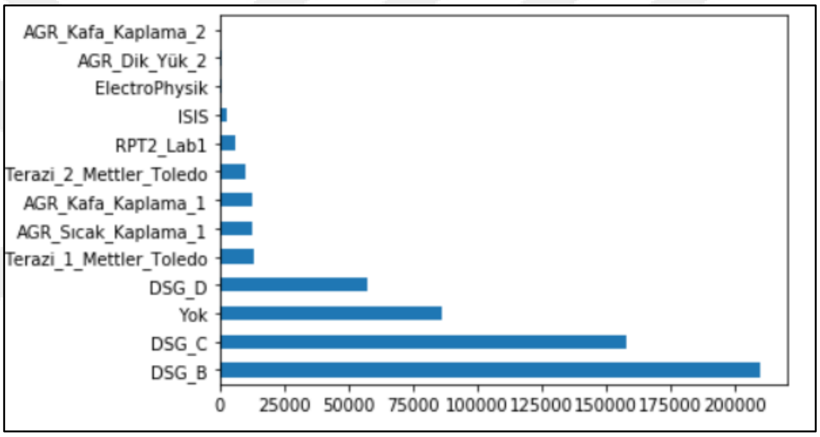


Figure 2.4. Quality control machines

2.1.1.4. Control Types

There are 117 control types in wine glass bottles and the most tested control type parameters are listed in the image below utilizing the countplot and value_counts() command in the python seaborn library. The most commonly used control types are "*Hot end Weight*" and then "*Ağırlık*"(Weight). Also, Other control types are seen with different numbers of distribution. The most used types in quality control types can be seen visually in Figure 2.5. Control types. In this way, it can be easily recognized as the dominant control type.

Table 2.2. Quantity of control types

QUALITY TYPES	QUANTITY OF COLUMNS

Hot_End_Weight_F1	31482
Ağırlık	16223
Soğuk_Kaplama	12603
Tansiyon	12594
Gövde_Sıcak_Kaplama	12582
……….	….
Ağız_İçi_Çapı_2_Ovallik	29
Gövde_Cidar_ACD	22
Gövde_Cidar_Maks	22
Gövde_Cidar_Min	22
Habbe_30gr	6

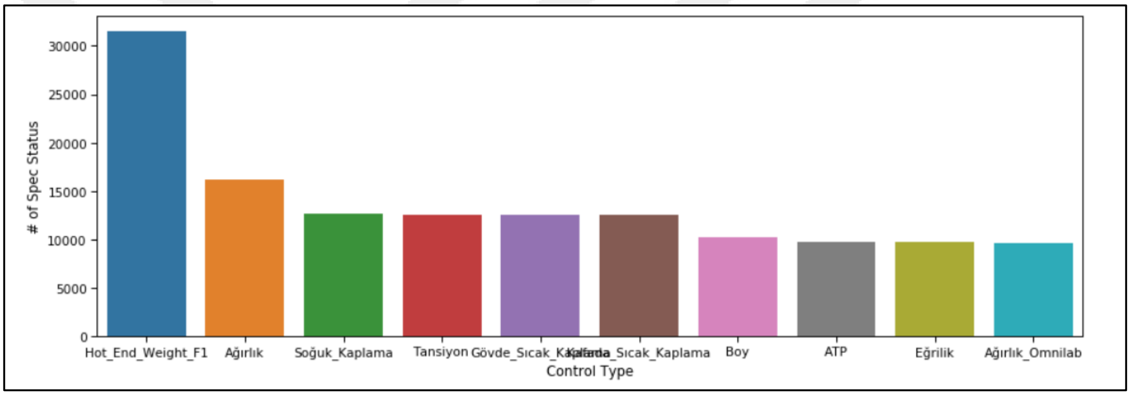


Figure 2.5. Control types

2.1.1.5. Wine Bottle Quantity

Have taken wine bottle data with product codes, names from the production department, and Table 2.3. Quantity of wine bottle types for 2019 illustrates quantities of bottle types. Significantly, it is clear that 221399 coded wine bottle product is high production according to the others and goes through more types of control, and as far as understood from the samples taken, it is more than other products. Then, the product with code 527575 is the lowest production. At the same time, it is shown visually in Figure 2.6. Types of Wine Bottles.

Table 2.3. Quantity of wine bottle types in 2019

PRODUCTS	QUANTITY OF COLUMNS

223199-1000 CC TANQUERAY BOTTLE	221385
833599-1.5 L WINE EMR SCR GREEN	59354
525675-75 CL WINE CHAMPAGNE	49753
529499-1 L WINE STD	43562
597675-75 CL WINE LGT LONG BORDO	40619
832799-1.5LT.WINE MANTAR	35942
521099 - 1 LT.VİD.BİRİCİK ŞARAP ŞİŞESİ	18635
521199 -1,5 LT.VİD.BİRİCİK ŞARAP ŞİŞESİ	15588
830275-75CL WINE NEW BORDO GREEN	12350
521299-2 LT WINE BOTTLE GR-FL	11065
250134-341 ML EMPTY STEAM WHISTLE	10786
590475-75 CL WINE SCR BORDO DUZ	8678
836899-1 L WINE CORK GREEN	8004
526799-1 L WINE KAVAKLIDERE	7502
599475-75 CL.OMUZLU BORDO SARAP SISESI	7197
580199-1 LT EFES MARMARA GOLD	4390
599975-75 CL.VIDALI BORDO SARAP SISESI	3928
508499-100CL Y.RAKI ESKI	3390
303375-75 CL SU DAMLA STD PALET	2866
303975-75 CL DAMLA MINERA	2593
527575-75 CL WINE BORDO	1107

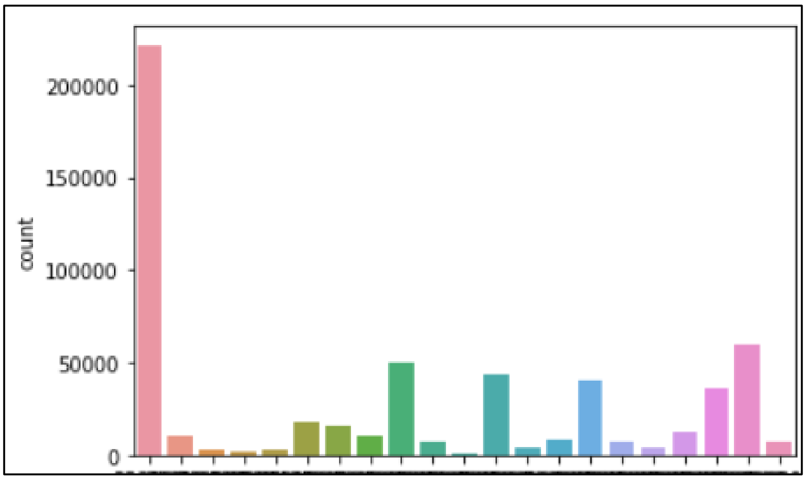


Figure 2.6. Control types

2.1.1.6. Specification States

The number of wine products taken from the quality laboratory can be examined according to their specification value. The distribution of quality status with the quantity of product can be seen on box plots in Figure 2.7. Specification States. The good quality bottle is called "*İyi*"(Good) or "*0*" numerical values that have the majority rate of data and others are very low quantities which are expressed with numerical values in the "*Spesifikasyon_dışı*"(Out of specification) or "2" and "Alarm" or "1" values.

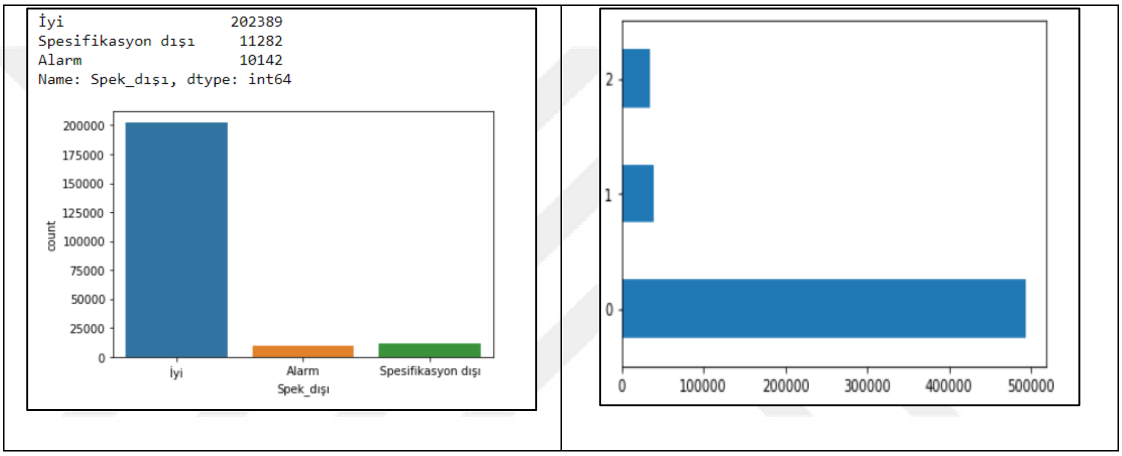


Figure 2.7. Specification States. (a) Box plot, (b) Count plot

The plotly Python package is an open-source library built on plotly.js which in turn is built on d3.js that is used in Figure 2.8. Specification status versus result value by products. It is used as a wrapper on plotly called cufflinks designed to work with Pandas dataframes.

In the graphic below, the specification status of the products according to the target values is examined. The distribution of products can be observed both collectively and product by product. The density of good quality products with a value of "0" is remarkable.

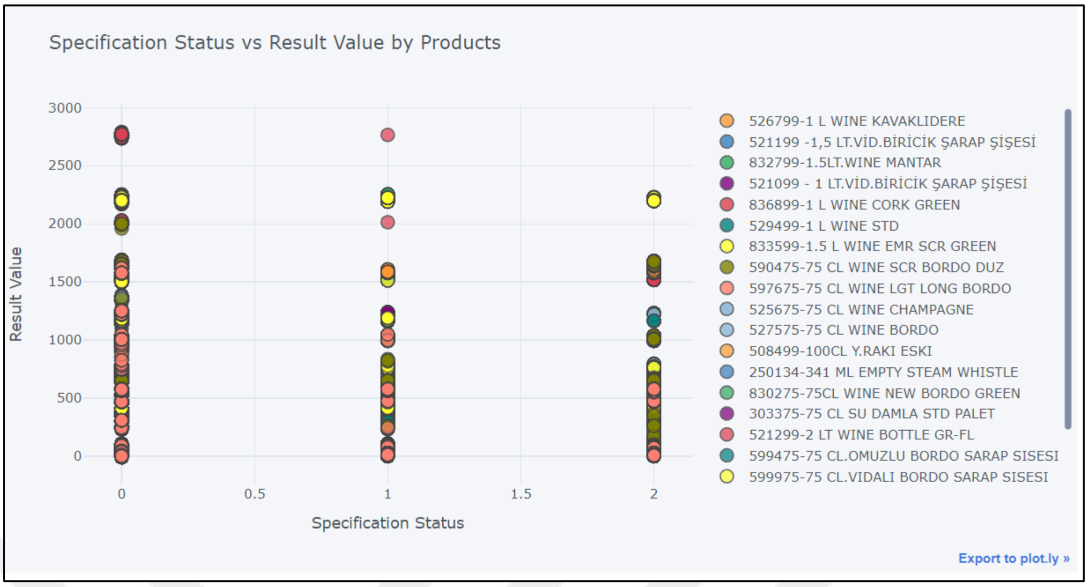
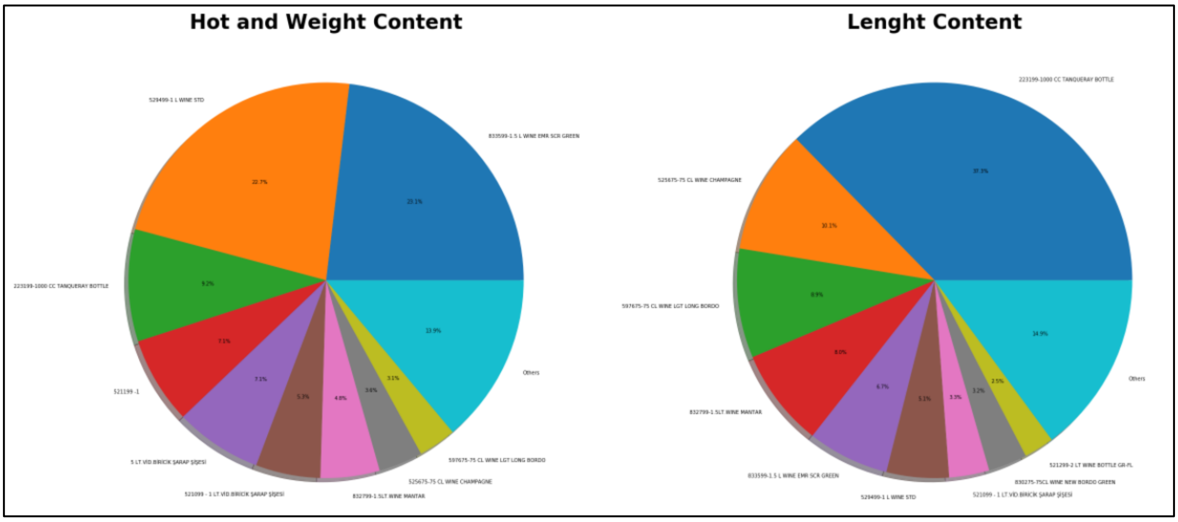


Figure 2.8. Specification status versus result value by products

In the 2.9. Figure Commonly used control types for wine bottle types, the percentages of the tests used on a product basis among the commonly used control types are given. *Hot and weight* control type has 833599-1.5 L wine emr scr green with 23.1%, following 529499-1 L wine std with 22.7%. 597675-75 cl wine lgt long bordo product is not used to be on it because of %3.1. *Length* content has 223199-1000 cc tanqueray bottles with 37.3% following 525675-75 cl wine champagne. *Pressure* content is used frequently in the *Hot and Weight* control type.



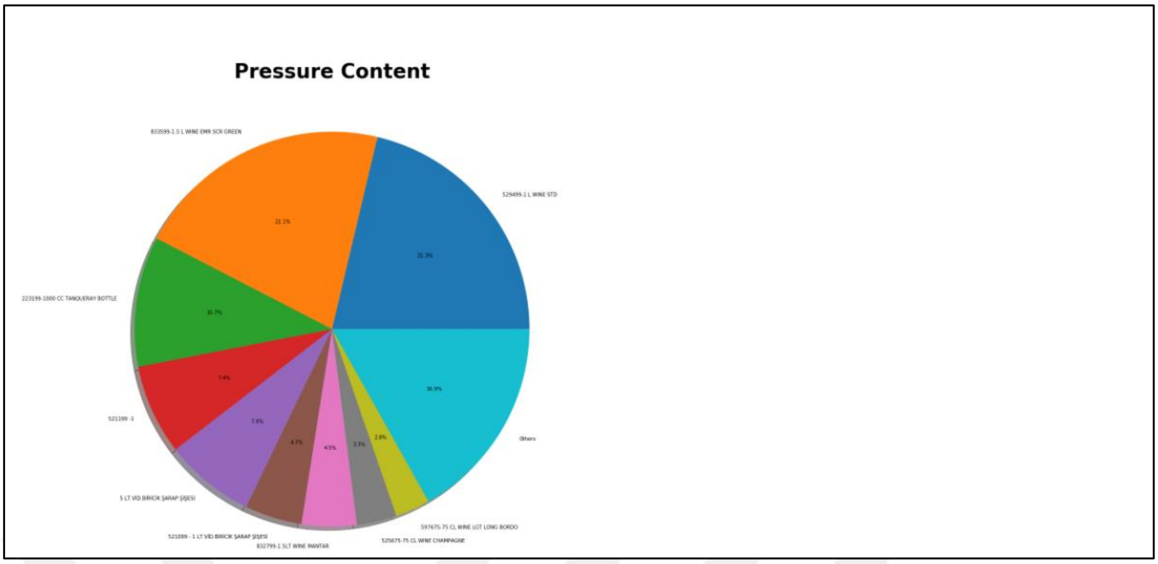


Figure 2.9. Commonly used control types for wine bottle types

2.1.1.7. Control Machines on Spec Status

In Figure 2.10. Control machines versus wine bottle types on specification state, illustrates the rates of distributions by drawing a scatter plot according to the specification situation in quality control types on a product basis. It is managed to demonstrate the products that passed all tests and which products failed in which for used mostly control type by scanning within control types according to product's quality status which is assigned "*0-iyi"(good), "1-alarm"(alarm)* and *"2-spek_dışı"(out of specification)* conditions.

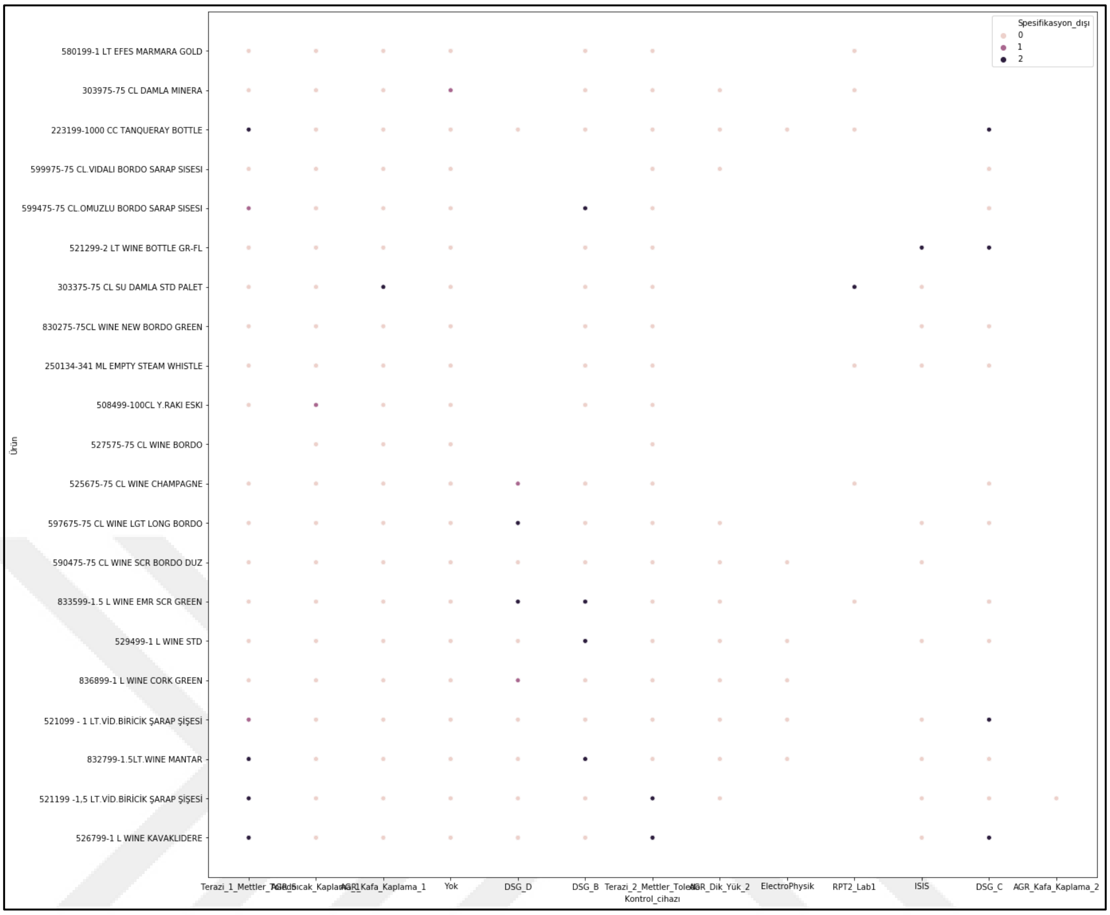


Figure 2.10. Control machines versus wine bottle types on spec. status

It can be clearly seen in the graphics below that wine products are quite successful in terms of overall quality and the distribution of defective product quality controls varies. 223199-1000 cc tanguery bottle is the most produced product that couldn’t pass at least control types. On closer inspection, the failure rate in control types is roughly proportional to the number of production of the product. As revealed by this figure, some control machines are not applied to specific products. It can be possible that they have not entered the data into the system and there is no explanation for inferences.

2.1.1.8. Lower Spec Status versus Customer Lower Spec Status

In the graphic below, the sub-specification status requested by the customer and the specification status obtained by the plant during production is compared. Generally, it has been observed that the values are on the same line. There are values that are very close to 0 on the graph because of not entering the data and filling with mean values. “0” dots and outliers can be ignored since there are very few numbers of them so it can be seen that the specification status in their production is progressing in line with the demands of the customer.

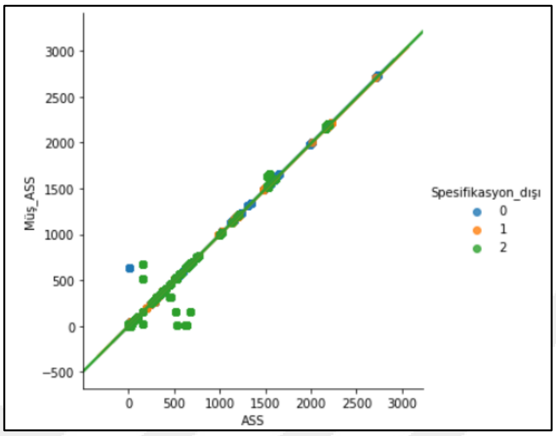


Figure 2.11. Ass versus müş_ass on spec. status

2.1.1.9. Cullet Distribution

The cullet parameter, which plays an active role in quality, the ranges of approximately 70.15% to 88.71% are used extensively. This parameter is the most important criterion as it is a factor that increases the glass quality. As it is known, if the cullet rate falls below or rises above a certain limit, it will cause big problems on pressure tests and not given quality products. At the same time, if the cullet rate increases too much and it is so high, the foreign materials from the cullet will be discarded by the mcal machine in the automation system. For, as the glass cullet increases, the number of foreign matter will also increase dependently.

Table 2.4. Statistical information about cullet(%)

count	119237
mean	70.154
std	14.205
min	34.715
25%	69.966
50%	75.002
75%	79.953

max	88.719

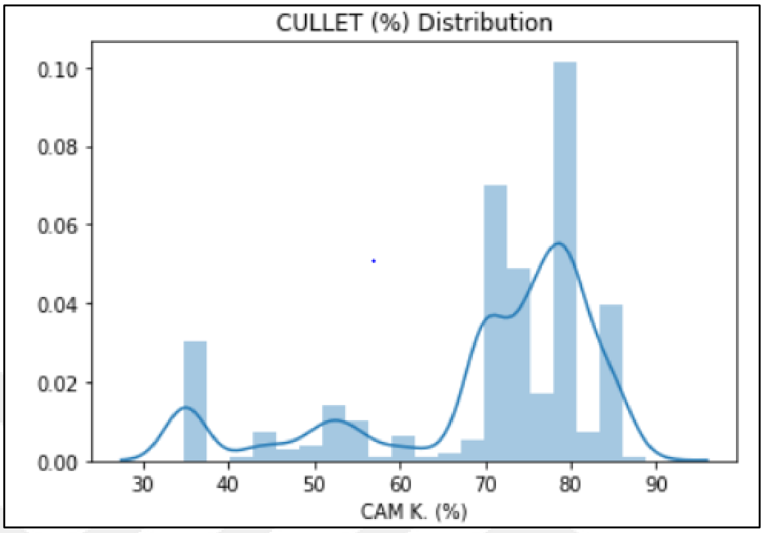


Figure 2.12. Cullet(%) distribution

2.1.1.10. Cullet Distribution on Spec Status

Below is a graph that helps us understand how the cullet values affect the specification status. It is frankly shown that the *Cullet(%)* range is intensely between 70 and 90, especially good quality(0 dots) on the "*Spesifikasyon_dışı*" column.

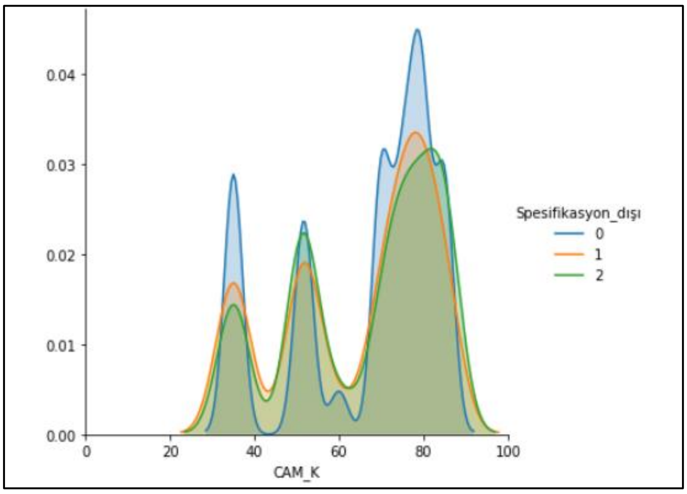


Figure 2.13. Cullet(%) distribution on spec. status

The most important factor to assign the value of the stone is the glass type. If it is exemplified from silica stone, its thermal expansion occurs in a much higher glass so, it is most likely surrounded by fractures. The same stone looks different in hard borosilicate glass with far fewer fractures. Because the borosilicate glass expansion coefficient is a low value compared to the bottle glass expansion coefficient. It is determined according to the production process. In Figure 2.14. Stone_button mean on spec. status, which indicates the average of "*Taş_düğme*"(Stone_button) values, the most preferred value is 3%, although the quality values are at the same rate for the values other than the average value of 3%.

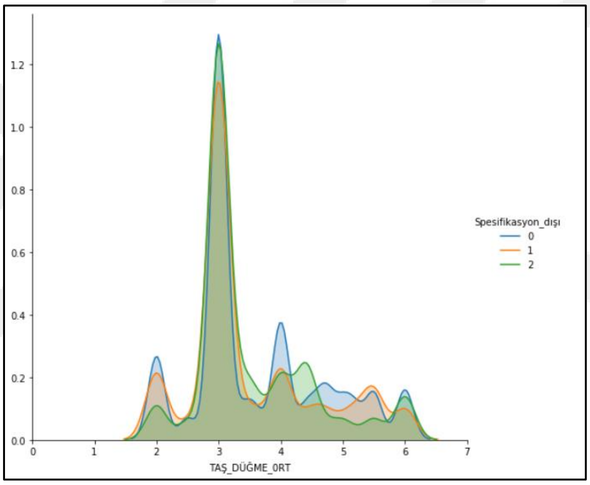


Figure 2.14. Stone_button mean on spec. status

2.1.1.11. Throat and Bottom_820 Temperatures

The Throat value taken from the furnace is approximately between 1275° C and 1285 ° C. In the throat region, the molten glass enters a process where it is cooled to a temperature of approximately between 1100°C and 1300°C. In this way, air bubbles are eliminated.

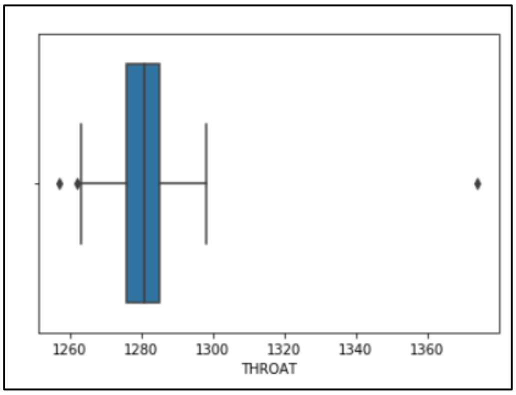


Figure 2.15. Throat temperature value

Since the glass experts are informed about the importance of the temperature of the Furnace bottom 820 region in the wine bottle, only this region is examined. It is generally between 1143° C and 1153° C values.

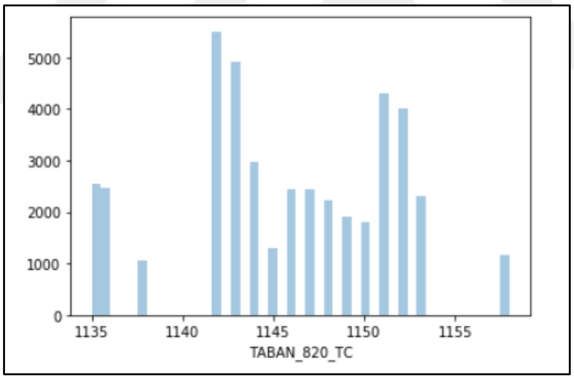


Figure 2.16. Bottom_820 region temperature value

Throat is a connection between the melting end and the riser that brings the molten glass into the refiner, working end or distributor. "*Taban_820*"(Bottom_820) column represents a region in the furnace. Then, when compared between the throat taken in the furnace and the temperature in the "*Taban_820*", there is a linear relationship as can be seen on the graph. The points are aligned homogeneously and spread on the line so it is realized there is a relationship between each other. Although these are linear, it is seen that some points intersect exactly and some points are far from the line.

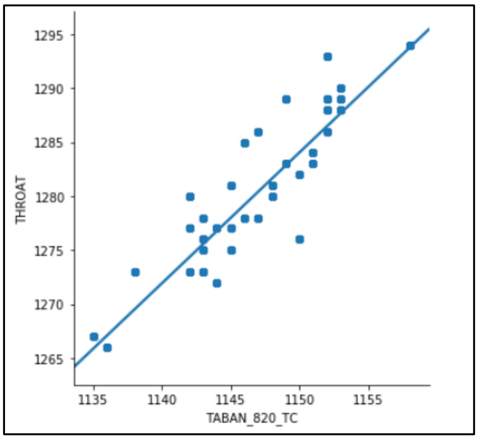


Figure 2.17. Bottom_820 region temp. versus throat temp.

Using the points spread linearly in the furnace, it is tried to understand the temperature zones of the product types within these points. 526799-1 L WINE KAVAKLIDERE wine bottle needs more heat than other products. It can be seen a list of other products in section 2.1.1.3. Wine bottle quantity.

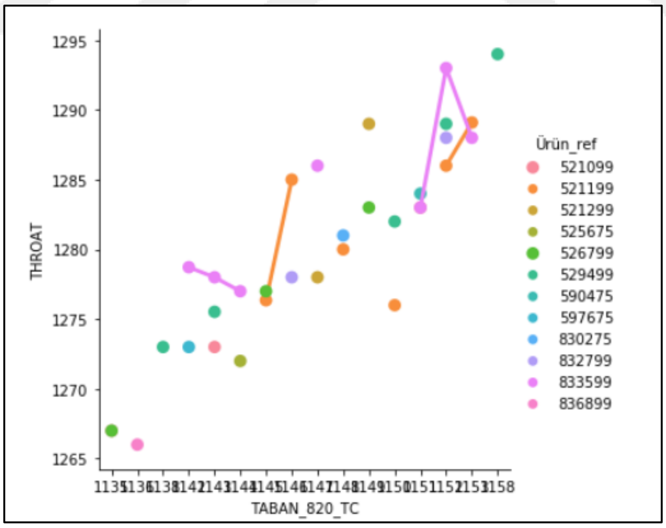


Figure 2.18. Bottom_820 region temp. versus throat temp. on wine bottle types

2.1.1.12. Spec Boundary

There is a linear relationship between the lower spec limit and the lower warning limit. In Figure 2.19. AUS versus ass, the density is between about 500 and 750 regions. On the other hand, you will see a graph comparing the lower spec boundary with the customer requested boundary in Figure 2.20. Customer target versus ass. If it is ignored the small differences and “0” values here, it is seen that it is generally linear and moving on a certain line. In this case, it can be assumed that the quality of the glass bottle satisfies the customer's request.

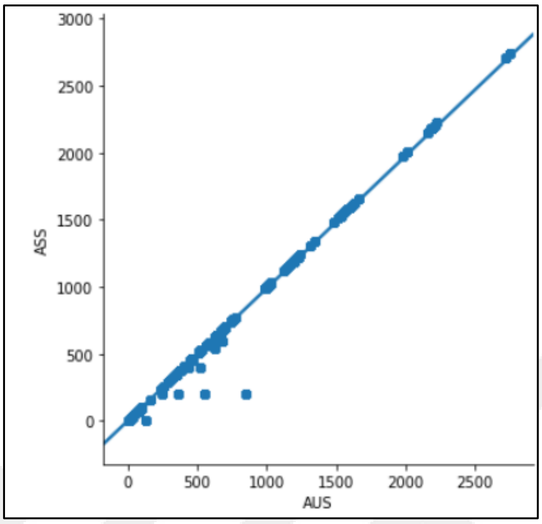


Figure 2.19. Aus versus ass

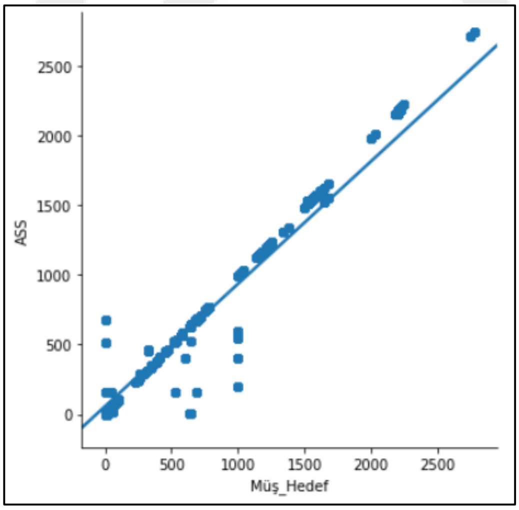


Figure 2.20. Customer target versus ass

A chart is drawn below showing the lower spec limit versus customer target values on specification status. It is clearly seen in Figure 2.21 that the points where the quality is good are on a certain line. “*İyi*”, “*Spesifikasyon_dışı*” and “*Alarm*” respectively refer to good quality, defective, and the quality value at a limit. In the graph, there are defective products determined as out of specification in a small part of the points. It can be thought that the products that are seen as defective in this section appear this way on the chart due to different factors.

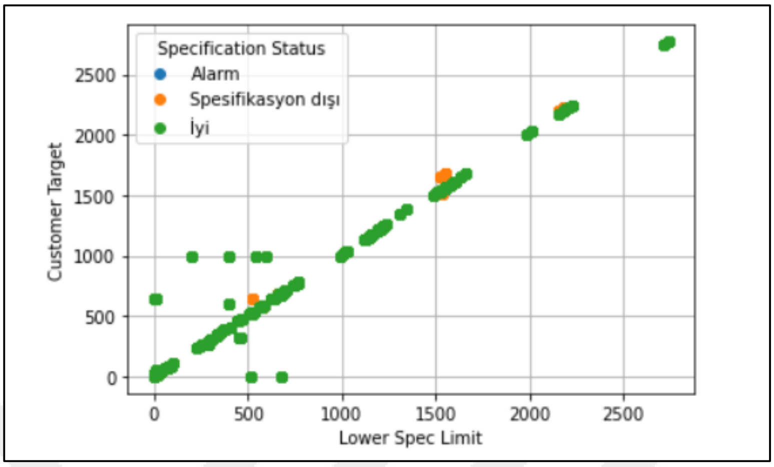


Figure 2.21. Lower spec limit versus customer target

In the quality laboratory data, the "*Harman_Altı_Elek_KWh*"(Under blend electricity) column value is roughly a minimum 7600 value, and also "*Sıcak_Nokta_Elek_KWh*"(Hot spot electricity) column value's minimum value is 20000. However, the value at which intensive observation is made starts from 25000 among these values.

The distribution is examined by drawing a 3D scatter plot in order to analyze the under-blend and hot spot electricity consumption according to the specification in Figure 2.22. Under-blend and hot spot electricity consumption on specification status, it is seen that these values are homogeneously distributed over the quality spectrum values. The purple dots seen in the graphic represent products with good specifications and it can be seen that these dots are of high density. The green color dots indicate the specification status alarm, while the yellow dots indicate the products that are out of specification value.

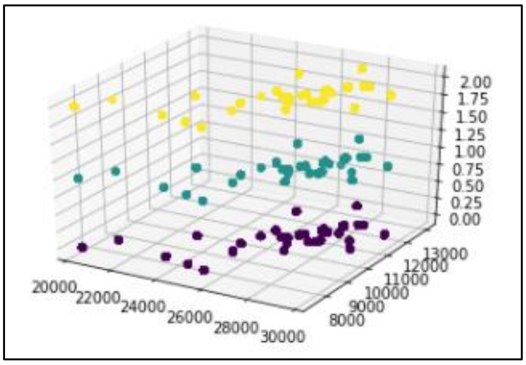


Figure 2.22. Under-blend and hot spot electricity consumption on spec. status

Likewise, the following results can be reached by looking at the Bottom_820 and cullet rates on the distribution of specification status in Figure 2.23. Bottom_820 region and cullet rates on spec. status. It can be said that when the bottom temperature where the cullet rate percentage is over 70% and the points are concentrated in different from each other and show homogeneous distribution. When examined deeply on a graph, it can be said that products with good specifications (purple color) are aligned more intensely. At his point, it is worth repeating that cullet rates and also furnace bottom temperature are significantly effective for the product quality.

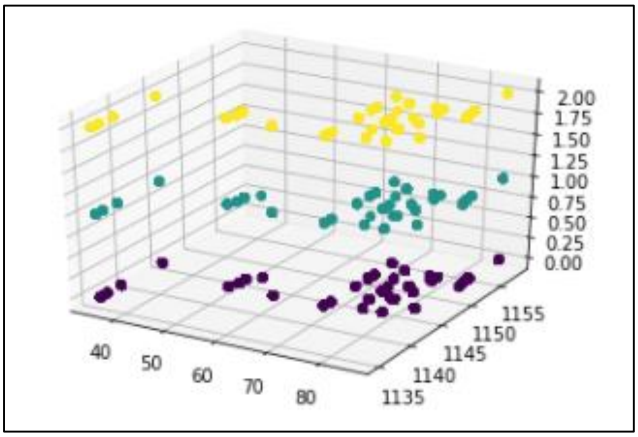


Figure 2.23. Bottom_820 region and cullet rates on spec. status

In the quality laboratory, the front and back surfaces of the wine bottles are examined. Values are entered as 1 and 2 (1-front, 2-back). Quality control is usually done automatically, and a small part is done manually. Conditions specified as manual forced refer to pressure control tests. There are 12 quality control machines and one manual control input is available. It is the most controlled dimensional measurement test. Control types are 117 and it is seen that the hot end weight type is used the most. When we look at the product range, if we consider the samples taken from the production, the most 1000 CC Tanqueray bottles are produced. In spec status, about 90% of the products are produced with good quality. It has been observed that only 5% of it is produced out of spec, that is, of poor-quality products. The distribution is homogeneous when looking at the spec states of the products. It is not possible for a particular product to be out of spec. When we look at the control types used in the products, generally 75cl wine bottles are products that are under pressure and weight too much. However, the 1000cc Tanqueray bottle is in the foreground in length measurement. When the spec conditions of the quality control machines in the products are examined, AGR_Sicak_Kaplama_1 works with weight, manual control, vertical load, and electrophysik zero error. It is seen that the sub-spec value of the product produced and the customer demand overlap with each other. It is seen that the rate of glass cullet is between 70% and 85% in terms of density and the spectra are homogeneous. This ratio is directly related to the glass color, as will be seen later. The stone button rate is 3%. Throat temperature is 1280 on average. The base 820 temperature is between 1143 and 1153. When comparing the bottom 820 with the throat temperature, there is a linear relationship between them. When the throat and bottom 820 temperatures on the products are compared, it is seen that each product does not have a certain temperature value. The lower spec value and the lower warning value are in a linear relationship with each other. Likewise, customer target value and sub-spec value are in a linear relationship and generally overlap with each other. When we chart for the subspec value and the customer target spec values, the differences are quite small, and overall a quality product was produced. By creating 3D graphics, under-blend electricity, hot spot electricity values, and spec states are visualized. At the same time, the bottom 820, cullet rate, and spec statuses are graphed.

2.1.2. Visualizations of Energy Data

In this section, important parameters such as energy, electricity, and natural gas will be examined on energy data. The distribution of energy consumption will be examined against the parameters of furnace traction, air/fuel ratio, batch humidity, throat temperatures, and cullet ratio. However, significant variables will be also visualized, indicating furnace temperatures.

2.1.2.1. Energy and Electric Consumption

When interpreted with a scatter plot for energy consumption kcal/kg glass and pull (kg/day) columns in Figure 2.24. Pull(kg/day) versus energy consumption kcal/kg glass, which is one of the important parameters in the energy data set used in the research, the glass pull means the furnace melting capacity value that usually is expressed in the number of (metric) tons of glass melted per day (24 hours) and it is concentrated around 360,000 kg/day and 390,000 kg/day, but the energy consumption varies.

As it can be seen in the graph, Pull(kg/day) value dots are spreading mostly on 360,000 kg/day, and significantly, energy consumption is less on 360,000 kg/day than 390,000 kg/day. It can be said that if the set pull value is 390,000 kg/day, it is more likely that energy consumption will increase.

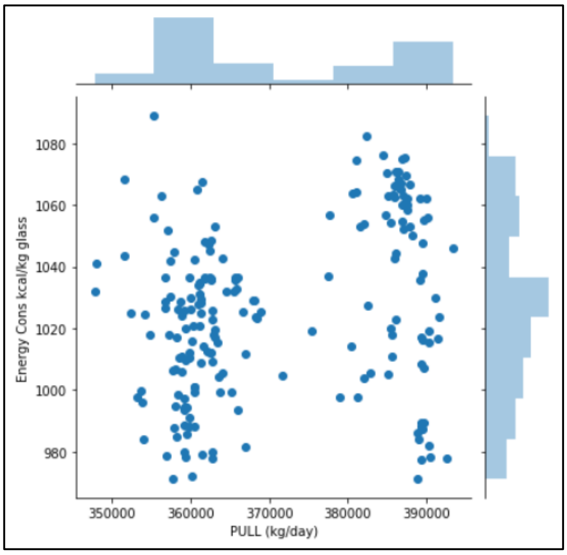


Figure 2.24. Pull(kg/day) versus energy consumption kcal/kg glass

The electric consumption value varies between 60 kcal/kg and 85 kcal/kg. The most intense point value is roughly 70 kcal/kg and above. The chart below describes the relationship between daily pull values and electric consumption, and it is not linear. Just like Figure 2.25.

Pull(kg/day) versus electric kcal/kg, Pull(kg/day) values are around 360,000 value and electric consumption is less than 390,000 value when it is compared with each other. At the same time, it can be seen the intensities of values in Figure 2.26. Pull(kg/day) versus electric kcal/kg distribution.

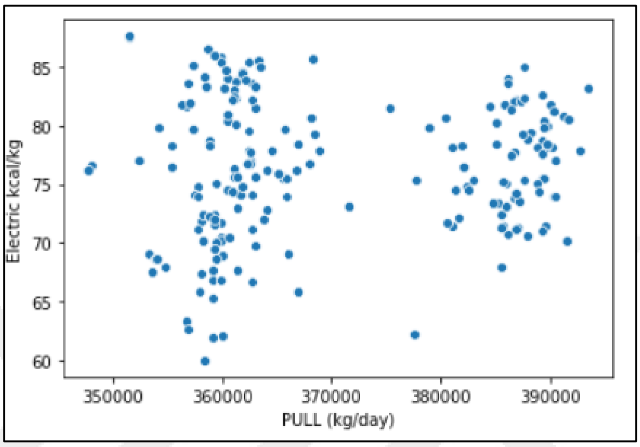


Figure 2.25. Pull(kg/day) versus electric kcal/kg

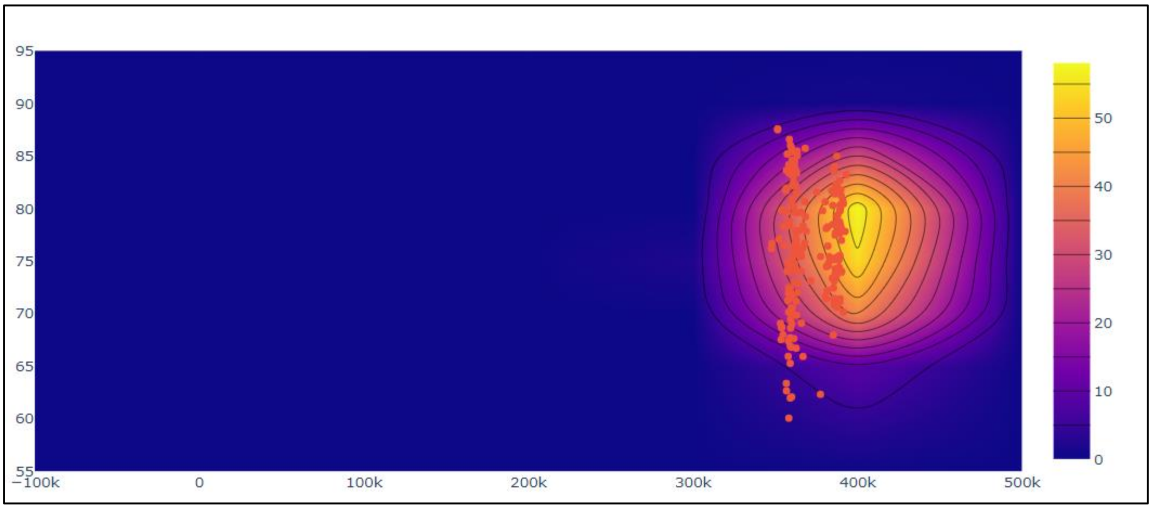


Figure 2.26. Pull(kg/day) versus electric kcal/kg distribution

The use of cullet and recycled glass will reduce energy requirements. When it is for each used 10% percentage of cullet in the manufacture of melt, It will be obtained at an estimated 2% savings. Thus air emissions will be up to 10% for 50% of cullet in the mix. [15]

When drawing a graph on the "energy consumption" kcal values and the "cullet" percentage rate data in the energy data, it is seen that the density in the percentage of the cullet is 70. On the other hand, it can be seen that the energy consumption increased more in the region where the cullet percentage rate is roughly 70% and its rate is high in the regions where the energy consumption is low.

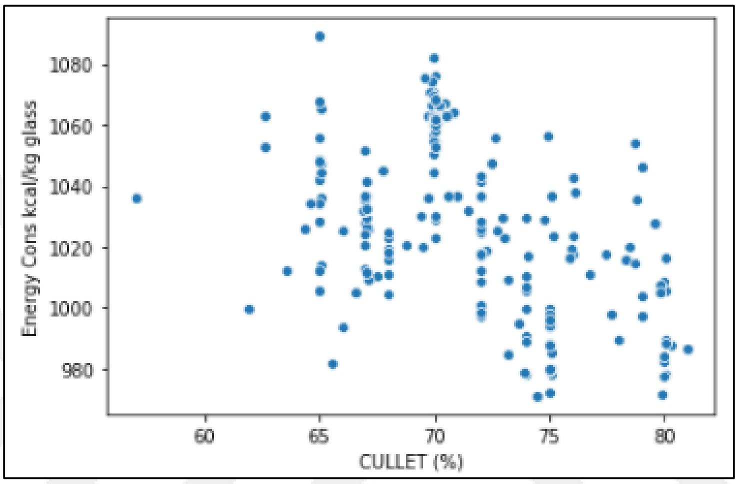


Figure 2.27. Cullet(%) versus energy cons kcal/kg glass

In the graphical view of the energy consumption and the throat 903 columns in the energy data representing a region at furnace temperature, it is understood that the throat temperature should be between 1250 ° C and 1280° C in order to consume less energy.

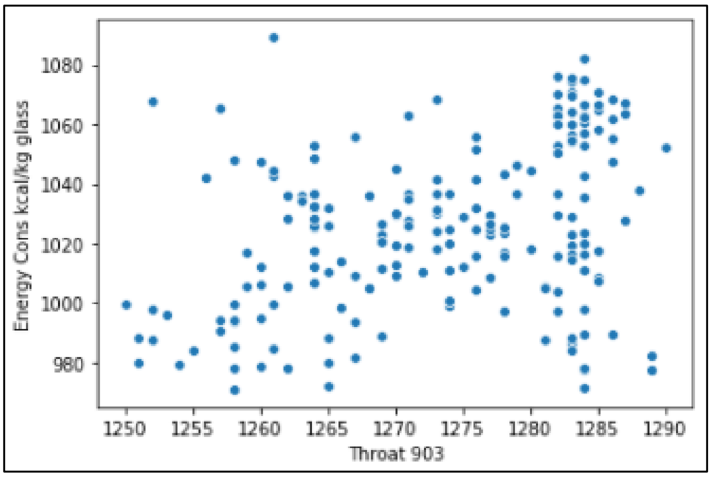


Figure 2.28. Throat 903 versus energy cons kcal/kg glass

In Figure 2.29. Throat riser versus energy cons kcal/kg glass, The energy consumption and Throat riser values in the energy data are examined. Then, it can be seen that the line progresses linearly when the line plot is drawn and the Throat riser value is concentrated around at 1200 kcal/kg and the energy consumption is around 1000 kcal/kg.

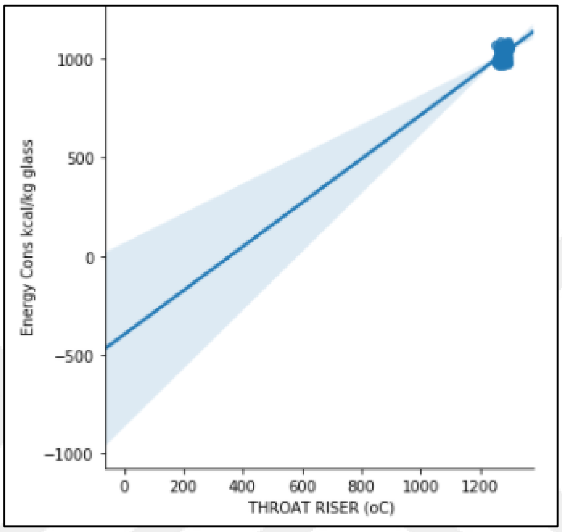


Figure 2.29. Throat riser versus energy cons kcal/kg glass

When it is examined energy consumption and Air/fuel ratio columns in the energy data, the Air/fuel ratio value with the highest energy consumption is approximately 9.3 and the minimum value is 9.2 in Figure 2.30. Air/fuel ratio versus energy cons kcal/kg glass.

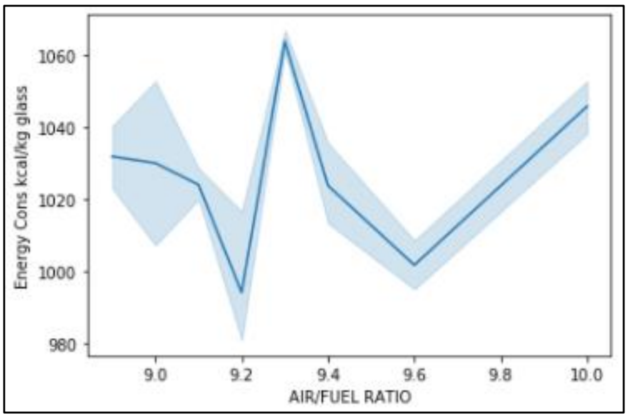


Figure 2.30.Air/fuel ratio versus energy cons kcal/kg glass

The batch humidity value found on the energy data is an important ratio in energy consumption. To understand its effects, it is compared with energy consumption. It is seen in Figure 2.31. Batch humidity(%) versus energy cons kcal/kg glass that the density is at the batch humidity rate 3.5% except extraordinary value and the energy consumption value is around 1000 kcal/kg glass.

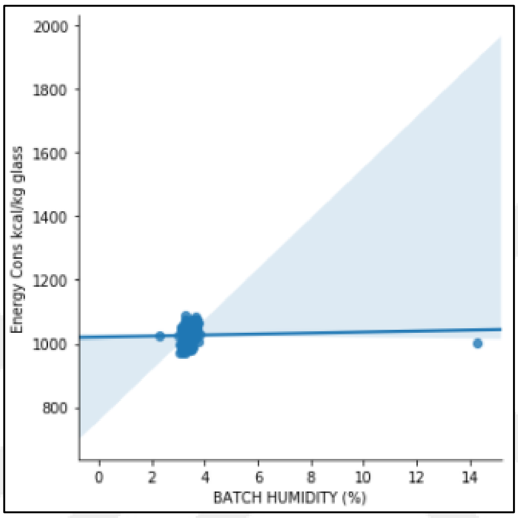


Figure 2.31. Batch humidity(%) versus energy cons kcal/kg glass

2.1.2.2. *Natural Gas Consumption*

As can be seen in Figure 2.32. Natural Gas Consumption Values and Figure 2.33. Natural gas consumption distributions, the minimum natural gas consumption value is around 722 kcal/kg. Its max value is higher than 1100 kcal/kg, median value is 952 kcal/kg.

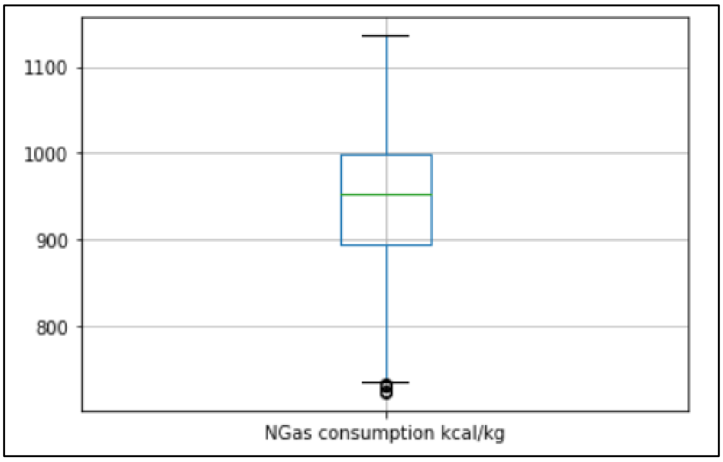


Figure 2.32. Natural gas consumption values

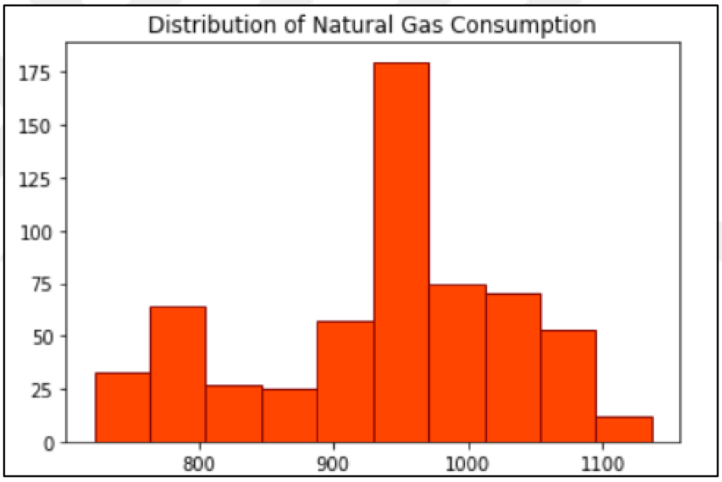


Figure 2.33. Natural gas consumption distribution

When looking at the energy data, histogram drawings related to furnace inputs-outputs values of natural gas flux. In Figure 2.34. Natural gas 1_2 distribution is between 0.1 and 0.8 for the doğal gaz_1(natural gas_1) column, while it is between 2000 and 2500 for the doğal gaz_2(natural gas_2).

It is clear that the maximum values fluctuate between 0.5 for doğal gaz_1 and between 2200 and 2300 for doğal gaz_2. These are important parameters for understanding furnace natural gas flux in energy data.

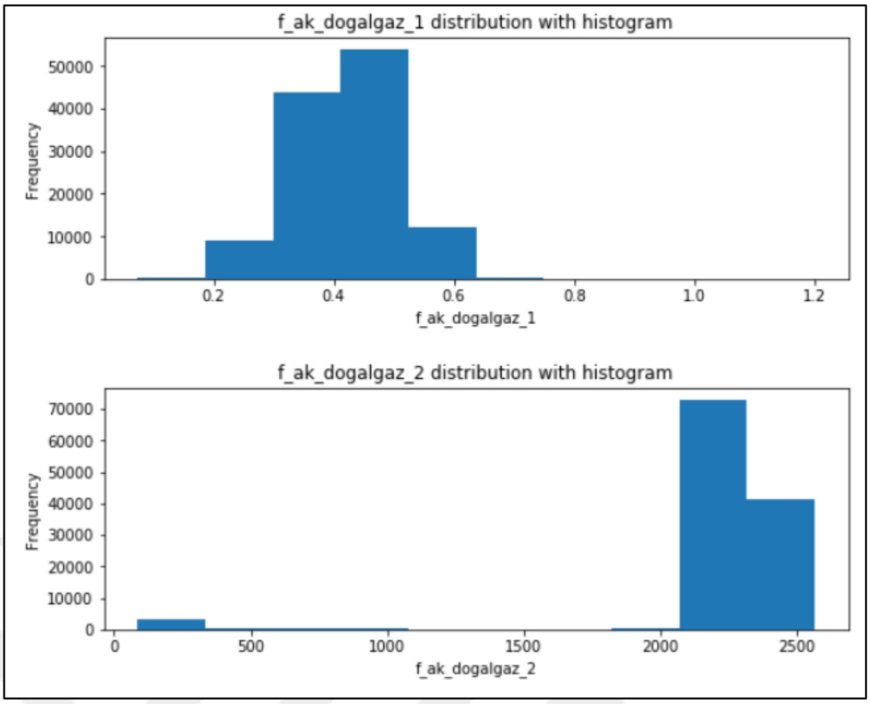


Figure 2.34. Natural gas 1_2 distribution

Looking at the natural gas consumption in terms of statistics, the average is around 952.51, the standard deviation is 952.51. It is seen that the maximum consumption is around 1136.37.

Table 2.5. Statistical information about n. gas consumption distribution

count	119237
mean	937.45
std	94.78
min	722.12
25%	892.93
50%	952.51
75%	998.93
max	1136.37

It is understood from the graph that the cullet rate in the energy data used is between 10% and 75%. As the cullet rate increases, natural gas consumption decreases. In case of less cullet rate, the amount of natural gas consumed by the furnace will increase due to the inverse relationship with each other.

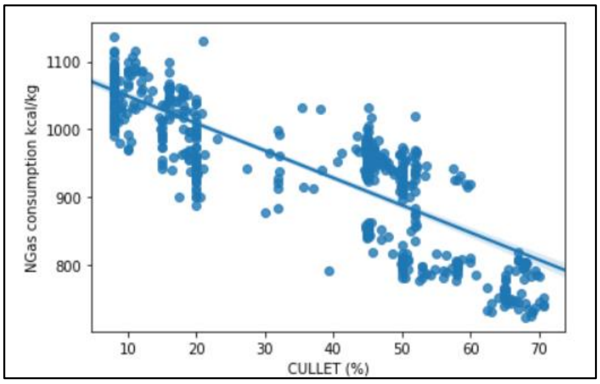


Figure 2.35. Natural gas consumption versus cullet(%)

Looking at the average Pull(kg/day) value here, it is observed that there is a density of 400 tons. While the pull value is around 500 tons, there is a decrease in the amount of natural gas consumption.

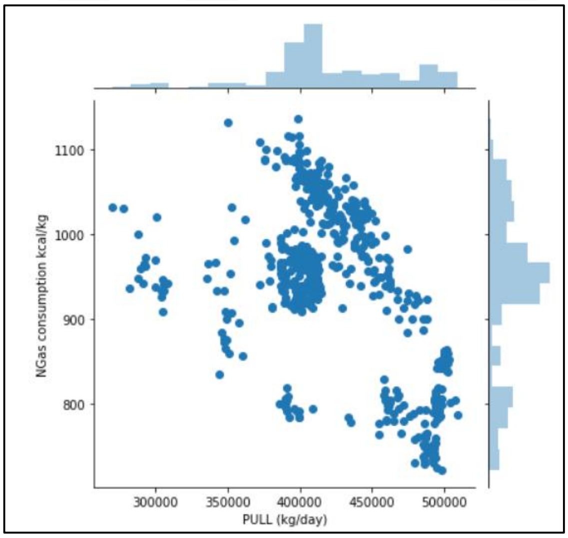


Figure 2.36. Natural gas consumption versus pull(kg/day)

When we look at the Air/fuel ratio, while it is 9.6, 9.8, and 10.1, the amount of natural gas consumption decreases, but the consumption increases at the values of 10, 10.2.

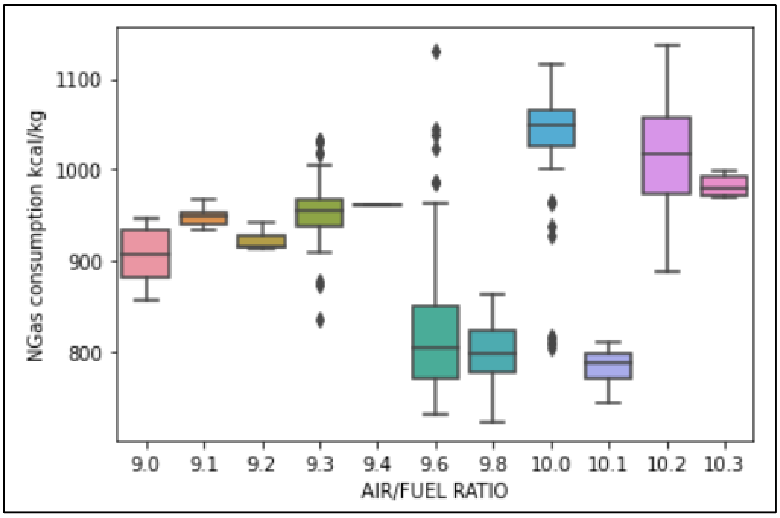


Figure 2.37. Natural gas consumption versus Air/fuel ratio

While the batch humidity value is between 3% and 3.5%, there is a decrease in natural gas consumption, just like in energy consumption.

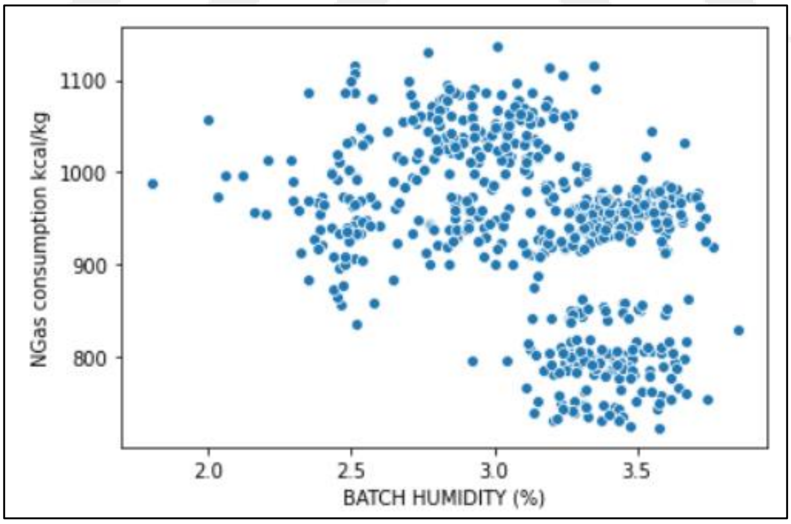


Figure 2.38.Natural gas consumption versus batch humidity(%)

Considering the daily energy data, the total natural gas consumption is presented in Figure 2.39. Consumption of natural gas daily for first quarter of 2020 which is an example of a tool drawn by utilizing the Tableau program. The highest consumption is seen on the 18th and 11th days, while the lowest consumption is on the 31st , 1st , 2nd, and 3rd days.

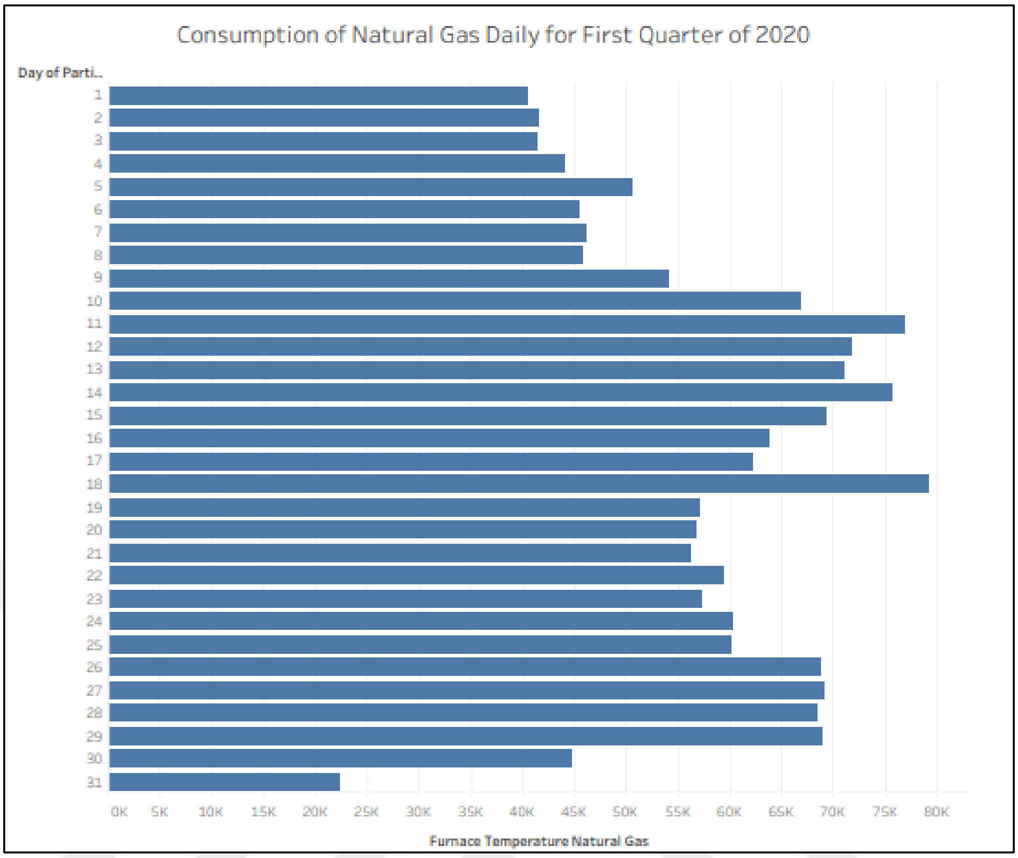


Figure 2.39. Consumption of natural gas daily for the first quarter of 2020

When looking at the following 3D graph consisting of natural gas and natural gas flow values to evaluate the consumption of natural gas; The x_axis represents the "*f_s_dogaz_gaz*" column, the y_axis the "*f_ak_dogalgaz_1*" column, and the z_axis the "*f_ak_dogalgaz_2*" column. The most obvious situation is that "*f_ak_dogalgaz_1*" shows the value of 0.4. The other features are spreading specific value ranges.

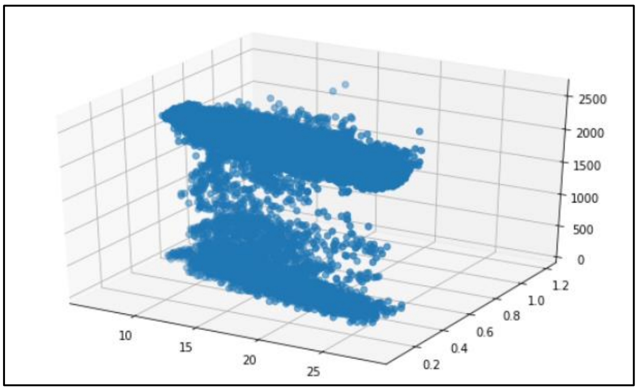


Figure 2.40. Natural gas and natural gas flow values

2.1.2.3. Furnace Temperatures

When the values in the chimney channels are transferred to the graph at the furnace temperature, a linear view emerges. The values that are closely related to each other differ in the middle and lower parts.

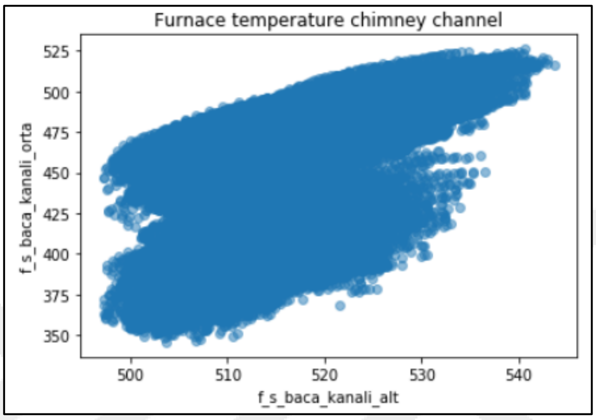


Figure 2.41. Furnace temperature chimney channel values

The Bottomline values at furnace temperatures are examined with box_plot in Figure 2.42. Furnace temp for 25 mm and Figure 2.43. Furnace temp for 245mm. It is 1385° C in the 25 mm region and approximately 1257° C in the 245 mm region. It is seen that there is a difference in the minimum and average values between these regions. And it can be said that the temperature is higher in the 25 mm region than in the 245 mm region.

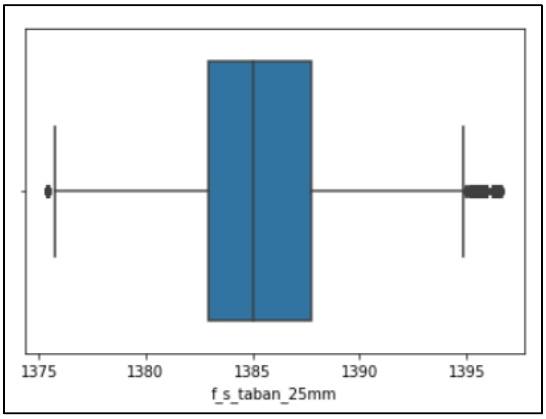


Figure 2.42. Furnace temp for 25 mm

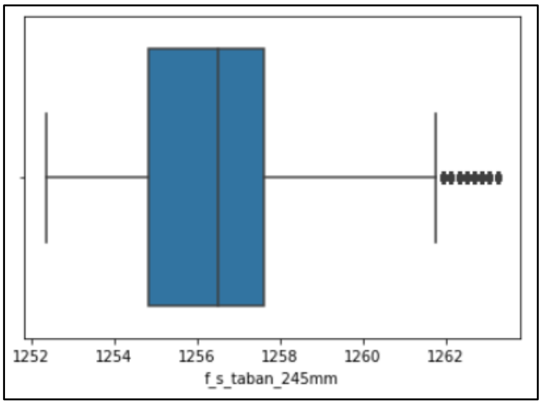


Figure 2.43. Furnace temp for 245mm

In the graphic below, visualization is made over the furnace belt temperatures. While the temperature is lower in the 801 region, it is higher in the 803 and 804 regions and has values close to each other. *Fırın_kemer_803*(Furnace_belt_803) region has a maximum approximately 1530 ° C and *Fırın_kemer_804*(Furnace_belt_804) region has maximum about 1565 ° C.

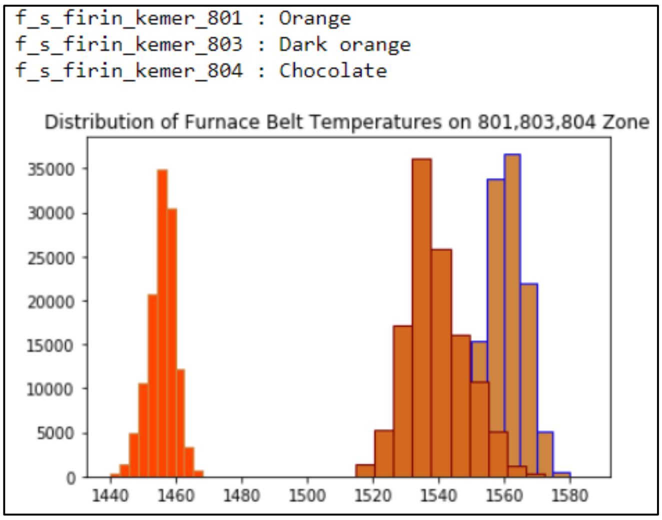


Figure 2.44. Distribution of furnace belt temperatures on 801,803,804 zone

When a small sample is taken from the furnace data and the Throat column is evaluated, the mean value is 1382.83° C, the max value is 1400.69° C and the standard deviation value is around 4.37.

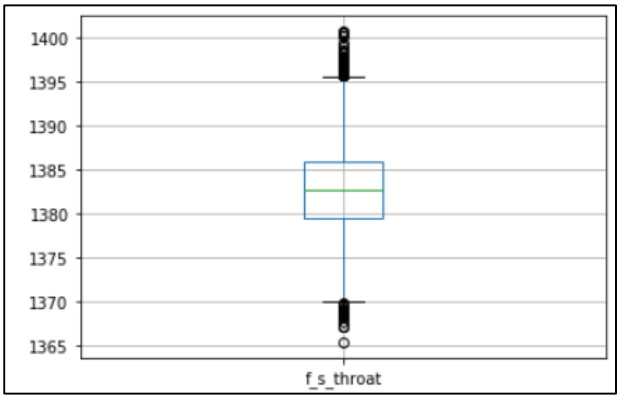


Figure 2.45. Furnace temperature throat

Looking at the left front wall values at furnace temperatures, it is observed that the mean value was around 1378° C and mostly 1384° C temperature.

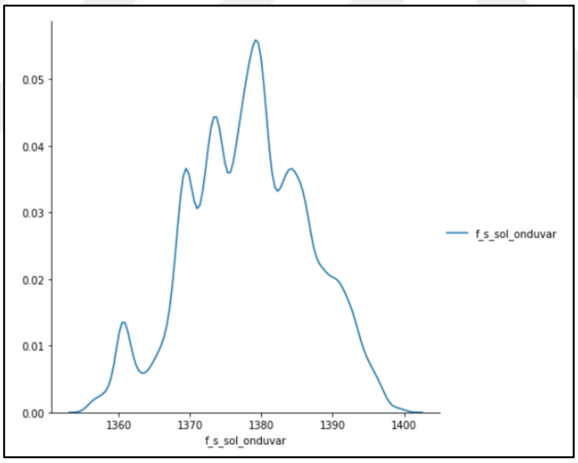


Figure 2.46. Furnace temperature left front wall

Looking at the temperature inside the glass at the bottom of the furnace, its mean value is 1309° C, and its maximum value is 1317° C. The standard deviation value of the data column is 2.5. It can be seen that between 1308 and 1309 data is more denser, between 1304° C and 1305° C and between 1313° C and 1317° C, data density decreases.

The mean of furnace temperature bottom glass_in in terms of statistics is around 1309.213, the standard deviation is 2.52. The maximum furnace temperature of bottom glass_in is around 1317.76 and the minimum temperature is 1304.02.

Table 2.6. Statistical information on furnace temp. bottom glass_in

count	119105
mean	1309.127
std	2.518
min	1304.016
25%	1307.138
50%	1308.972
75%	1310.800
max	1317.757

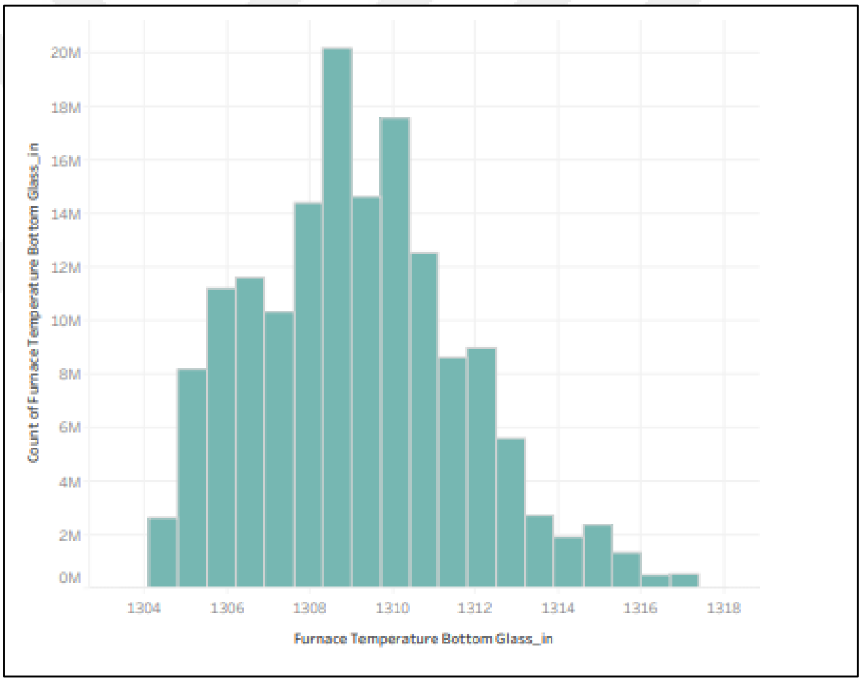


Figure 2.47. Furnace temperature bottom glass_in for first quarter of 2020

Looking at the energy data, it is seen that between 360 tons and 390 tons of glass is melted in the furnace. In the graphs, the dots showed a density of 360 tons. Considering the energy and electricity consumption, 360 tons of traction is lower than 390 tons. In another energy dataset showing natural gas, mainly pull values vary between 400 tons and 500 tons. When the natural gas consumption and furnace pull are compared, the density is seen as 400 tons. As the pull value increases, the amount of natural gas consumption decreases. As the cullet rate increases, energy and natural gas consumption decreases. As Throat and Throat

Riser(oC) temperature increases, energy consumption increases. The Air/Fuel Ratio value consumes the least at 9.2 and consumes the most at 9.3. At this point, the minimum and maximum values between the values close to each other suggest that the furnace humidity and other factors are effective. When the Air/Fuel Ratio values in natural gas consumption are examined, it is seen that there is a minimum natural gas consumption of 9.8 and a maximum consumption of 10.2. Considering the furnace humidity in energy consumption, it shows a density of 3.5%. Natural gas consumption varies between 3.0% and 3.5%. When the humidity is at 3.5%, natural gas consumption decreases. However, as the furnace humidity increases, the energy consumption also increases. Rate is 953 kcal/kg on average. It has been observed that the natural gas flow is from 2 channels. While small numbers are seen in the 1st channel, high flow values are observed in the 2nd channel. When the days are grouped and analyzed throughout the year, the most natural gas consumption is on the 18th day of the month. When the furnace temperature is compared to the middle and bottom with flue blood, it is observed that there is a linear relationship between them. Base 25mm and 245mm graphed. It is understood that the temperature of Base 25mm is higher than Base245mm. In belt temperatures, the 804 region has the highest temperature. The furnace throat region is generally seen to be between 1380° C and 1385° C. The temperature of the bottom floor inside the furnace is 1309° C degrees intensely.

2.1.3. Visualizations of Glass Components Data

In this section, In this section, the list of glass components used in glass making is defined. In addition, the relations of the variables that are considered important with each other were examined. SiO2, Al2O3, Na2O, Fe2O3, K2O, Cr2O3 components can be seen in the correlation table and visualizations are used to better see the relationship between them. The functions of the oxides, which are essential in the glass structure, can be explained as follows.

Table 2.7. Definitions of glass components[15]

Glass Components	**Definitions**

SiO_2	The component that makes up the structure of the cooled glass is SiO_2 which increases the viscosity of the glass. Therefore, it gives the glass a glassy feature.
Na_2O	Na_2O reacts at a temperature much lower than the melting point of silica. The soda allows the silica to flow at a lower temperature, creating a product that melts more easily. If the amount of soda in the soda-silica glass is increased; The chemical exposure of glass by water also increases. Therefore, in order to produce a usable glass that can be easily melted and shaped easily, it is necessary to make some additions and corrections in the glass composition.
CaO	To increase the resistance of the glass to chemical effects forms the glass skeleton by adding CaO.
MgO	MgO is also used to increase the resistance of the glass. But it is not as effective as CaO. While MgO slightly lowers the liquefaction temperature of the glass, it greatly slows down the crystal growth rate. At the same time, It increases the resistance of the glass against atmospheric effects.

K_2O	Potassium oxide (K_2O) is another alkaline oxide glass modifier. The effects resemble sodium oxide. Since the diameter of the potassium ion is larger than the diameter of the sodium ion, it also reduces the mobility in the glass. For this reason, the working range of potassium glasses is wider and electrical conductivity is low. When pure potassium oxide is used instead of sodium oxide, it becomes difficult to melt glass. The best results are obtained under the combined conditions of sodium oxide and potassium oxide.
PbO	Sodium oxide is used in crystal glasses with/without PbO.
Cr_2O_3	It is a dark green powdery substance known as chrome green, which is insoluble in water and acids, forms hexagonal crystals, is used as a pigment and catalyst in the production of colored glass and ceramics. It can be seen in Table 2.23 and Table 2.24 that the Cr_2O_3 component is also effective in determining the color of the glass bottle.

When graphing the values of SiO_2 and Al_2O_3 on the glass components data, it is seen that the data is clustered. Al_2O_3 value varies between 1.4% and 1.9%. The SiO_2 component fluctuates between 68.5% and 70.5%. And downward linearity is observed.

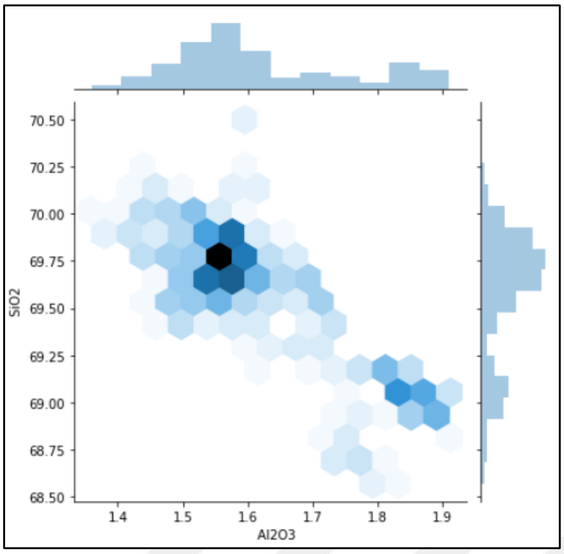


Figure 2.48. Al2O3 versus SiO2

On the graph drawn between *Al2O3* and *Na2O* glass components, it is observed that the *Al2O3* component showed a density between 1.5% and 1.7%, and could take values between 1.4% and 1.9%. It has been observed that the *Na2O* component shows an intensity between 14.3% and 14.8%, and can take values between 14.0% and 15.2%. There is upward linearity.

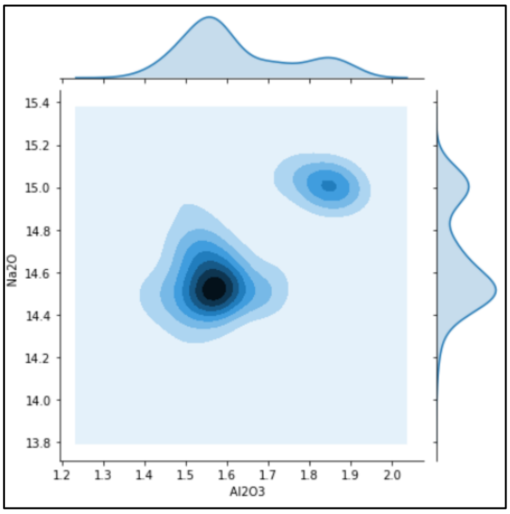


Figure 2.49. Al2O3 versus Na2O

When the values between *SiO2* and *Fe2O3* glass components are examined, it is seen that the values *of Fe2O3* columns are between 0.23% and 0.25% and that the data is stacked at a certain point and that the *SiO2* column values are between 68.75% and 69.25% in this region.

It can be said that *Fe2O3* column values are dense between 0.27% and 0.40% and SiO2 component has values between 69.25% and 70.5% at these points.

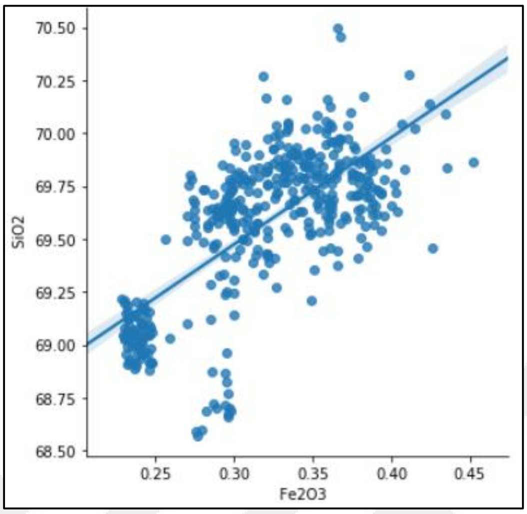


Figure 2.50. Fe2O3 versus SiO2

When *K2O* and *SiO2* glass component columns are examined, downward linearity is observed. The *K2O* component concentrated between 0.30% and 0.53%. *SiO2* takes values between 69% and 70.25%.

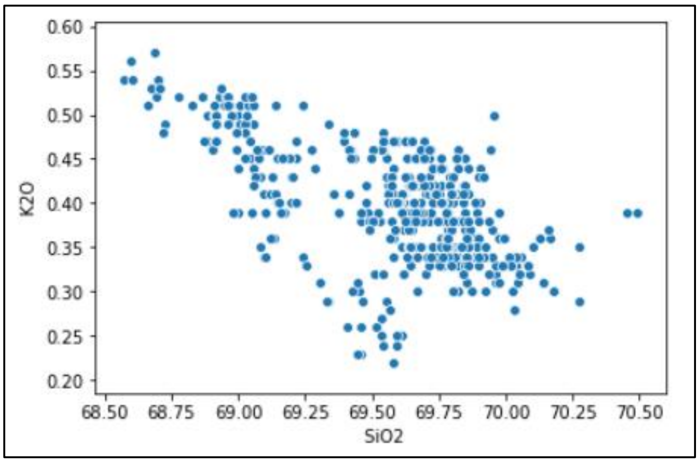


Figure 2.51. SiO2 versus K2O

There is a linearity with average correlation between *SO3* and *Cr2O3* glass component columns. The *Cr2O3* glass component is between 0.22% and 0.28%. *SO3* has a density between 0.05% and 0.14%.

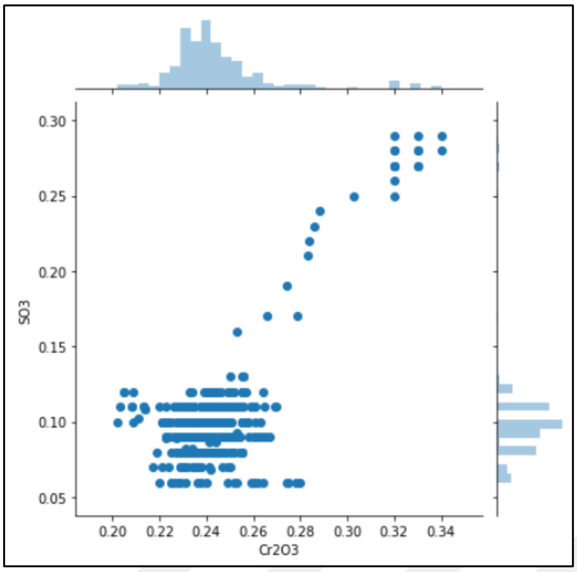


Figure 2.52. Cr2O3 versus SO3

In this section, glass components are examined. When SiO2 and Al2O3 are compared, Al2O3 decreases as SiO2 increases. A density of 69.75% in the SiO2 component and 1.55% in the Al2O3 component was observed. When SiO2 and Fe2O3 were graphed, it was seen that there was a linear relationship between them. While SiO2 was between 69.25% and 70%, Fe2O3 was between 0.25% and 0.40%. When the SiO2 and K2O components are evaluated, the K2O component shows a density between 0.30% and 0.50. When Al2O3 and Na2O were compared, density was determined at certain points. The Al2O3 component is 1.55% while the Na2O3 component is about 14.5%. When SO3 and Cr2O3 are graphed, it is seen that there is no correlation between them and the values are generally constant. SO3 is between 0.06% and 0.13%, while Cr2O3 is between 0.21% and 0.28%.

2.1.4. Visualizations of Furnace Data

In this section, it will be tried to understand the furnace temperatures, which are considered important in the production of wine bottles, within the furnace data. It will be presented with a figure to show the bottom 820 region and the throat 903 region in the furnace and a temperature display. In addition, it will be observed how the temperatures are related to each other. The distribution of wine bottle colors, culllet rate and the graphs of energy consumption will be discussed.

Figure 2.53. Furnace structure shows the structure of the glass furnace and to understand the zones before starting to visualize the furnace data. This table can be examined especially when it is desired to examine the bottom temperature and belt temperature regions.

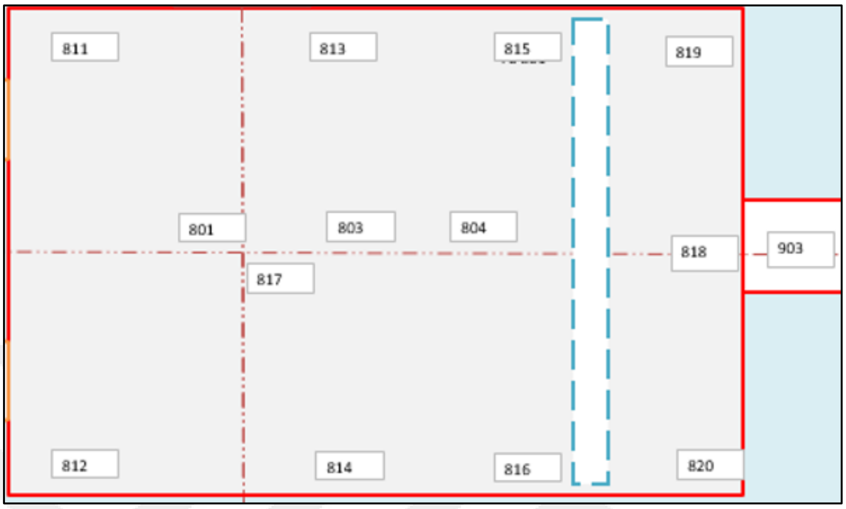


Figure 2.53. Furnace structure

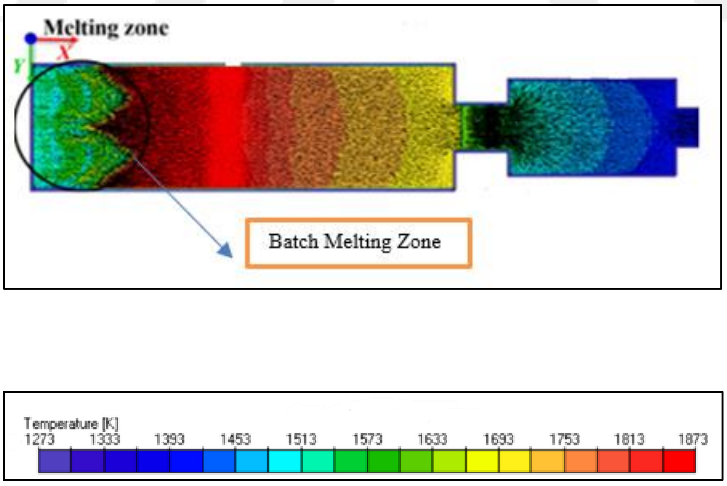


Figure 2.54. Furnace Temperature Display[16]

2.1.4.1. *Cullet Rate*

In the furnace data, when the cullet rate and natural gas data are examined, it can be said that the natural gas consumption decreases with the increase of cullet rate. Moreover if it is above

75%, it can be understood that there is low natural gas consumption.

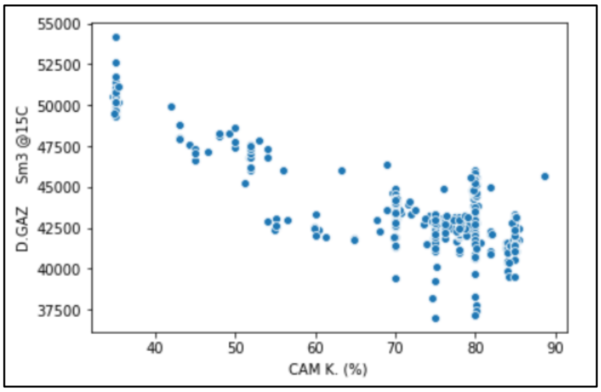


Figure 2.55. Cullet(%) versus natural gas

When the graph between the Taban_820 and the cullet rate is examined, it shows that the cullet value is between 70% and 85% and the Bottom temperature 820 region is between 1140° C and 1155° C. In addition, it is clearly seen that as the cullet rate increases, the Bottom temperature value increases.

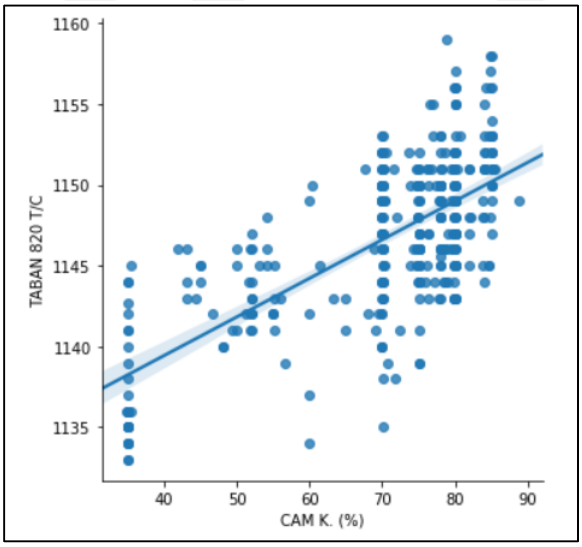


Figure 2.56. Cullet(%) versus bottom temperature 820 region

2.1.4.2. Wine Bottle Colors

In the dataset where there are four different wine bottle productions, it is seen that the most produced products are IR-Green and UV-Green.

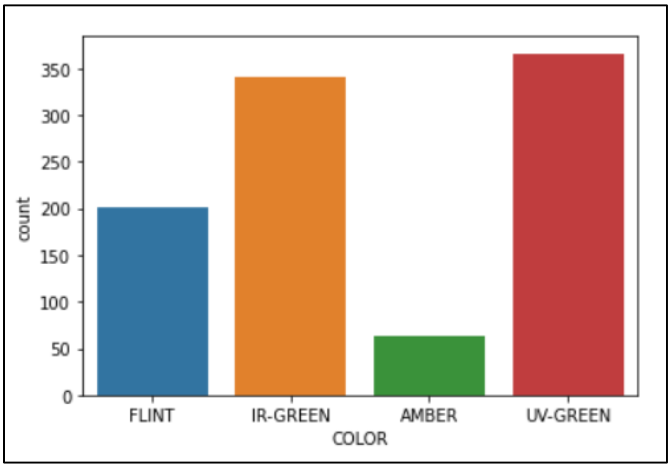


Figure 2.57. Glass bottle color distribution

In Figure 2.58, it is combined the power of pandas with plotly and cufflinks for a boxplot of the cullet percentage per story by bottle colors. However, it is used a pivot in python. As a result, it can be deduced that *UV-GREEN* needs much more cullet percentage than *IR-*GREEN. While IR-GREEN has a maximum 80% cullet, UV-GREEN has 69,98%.

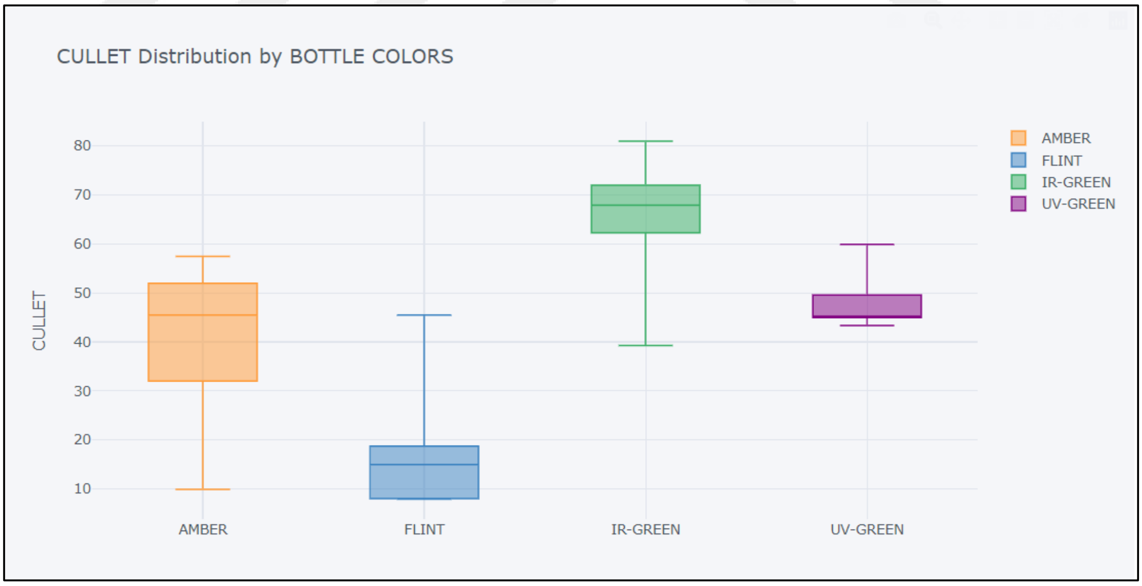


Figure 2.58. Cullet distribution by bottle colors

2.1.4.3. Furnace Temperatures

There is a positive trend between "*Taban_820*" and "*Taban 816*" temperatures. *Taban_820* temperature is between 1132° C and 1157° C, *Taban 816* temperature is between 1105° C and 1130° C.

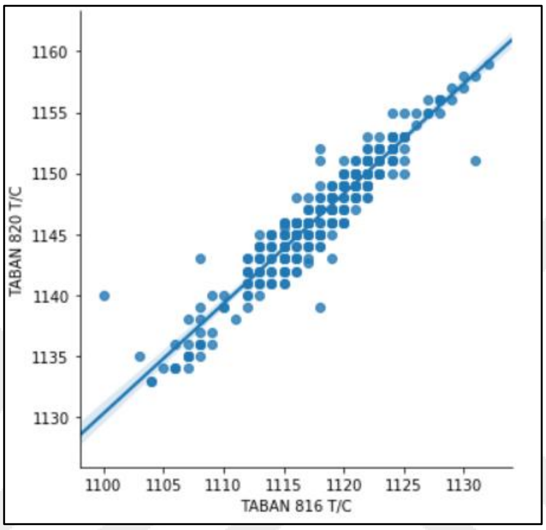


Figure 2.59. Bottom temperature 816 region versus bottom temperature 820 region

Another feature that is thought to be important is "*Riser_Optik*" values, which are densely located between 1275° C and 1290° C temperature values.

"*Cam Yüzey Optik*"(Glass surface optic) column's values have the highest ratio between 1570° C and 1575° C.

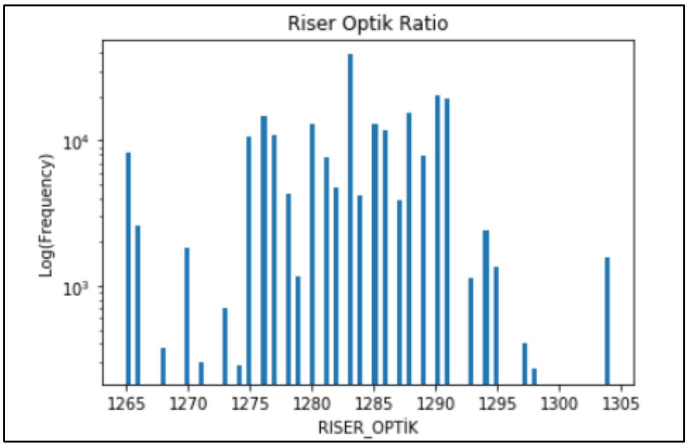


Figure 2.60. Riser optic

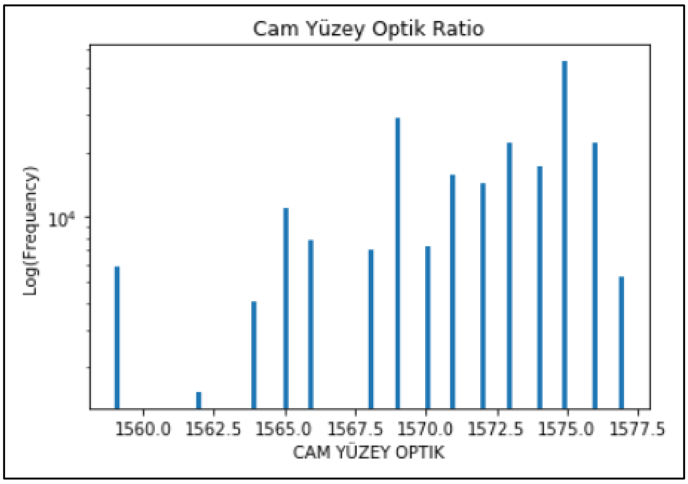
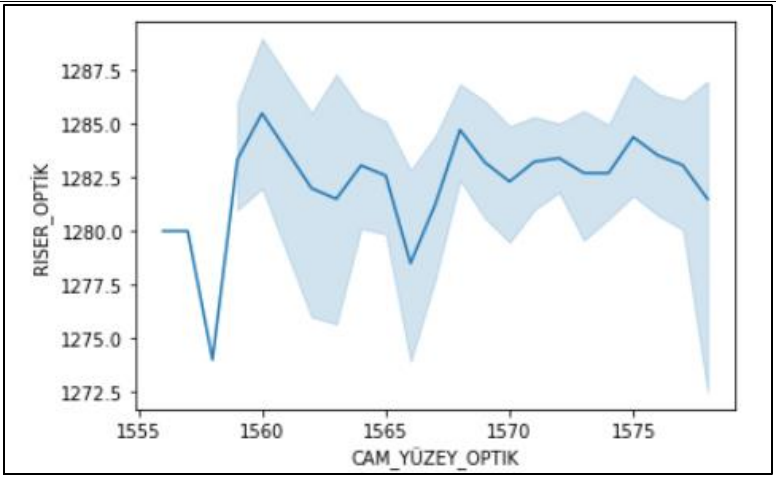
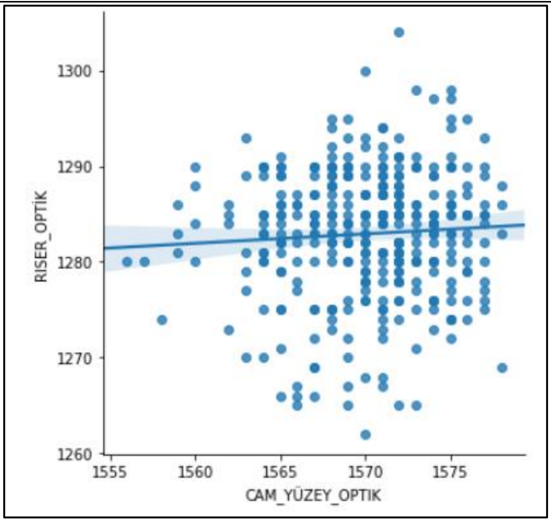


Figure 2.61. Glass surface optic

As can be seen from Figure 2.62. Glass surface optic versus riser optic, there is no linear relationship between the "*Cam yüzey optik"* and the"*riser optik"* column. "*Cam yüzey optik* " is between 1565° C and 1575° C, "*Riser optik*" is between 1270° C and 1295° C These are related to glass temperature and furnace throat zone temperature. It can be seen that the "*Cam Yüzey Optik*" column represents the glass temperature under the forehead wall which is around 1570° C.





(b)

Figure 2.62. Glass surface optic versus riser optic. (a) Line graph, (b) Scatter graph

When the column values of "*Taban 816*" and "*Riser Optik*" are examined, it is seen that the Riser Optic value is stacked between 1275° C and 1295° C, and at these points, Taban 816 temperature varies between 1105° C and 1125° C.

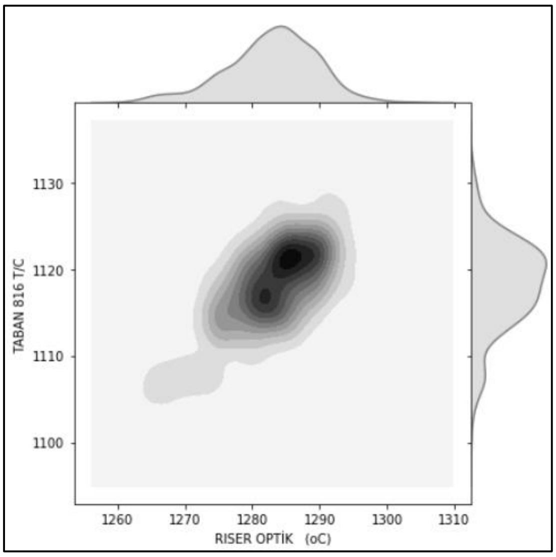


Figure 2.63. Riser optic versus bottom 816 region

As seen in correlation, there is a close relation between furnace temperature on the right forehead optic and glass surface optic. “*Sağ Alın optik*”(Right forehead optic) and “*Cam Yüzey Optik*”(Glass surface sptic) is on range of 1565° C and 1575° C intensely.

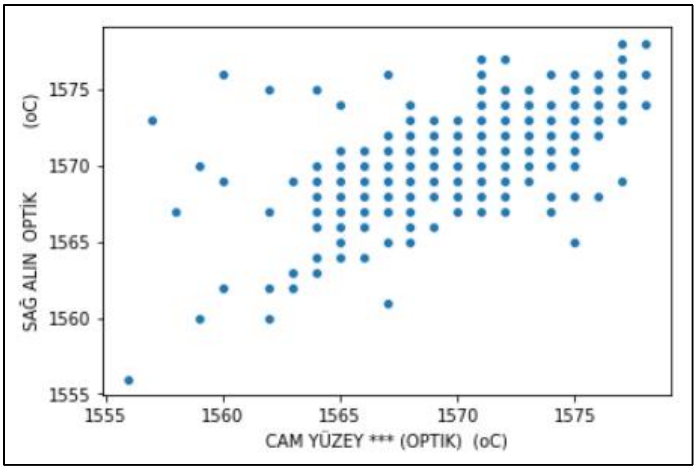


Figure 2.64. Glass surface optic versus right forehead optic

There is a negative and at the same time positive relationship between the regions of spout glass temperature. If it is necessary to see the positive relationship, C1 and C5 region temperatures can be taken. It can also be sampled with the C2 and C4 region temperatures in a negative relationship.

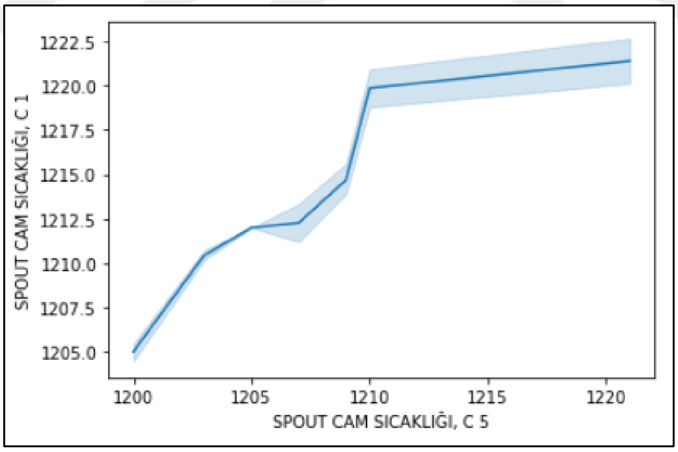


Figure 2.65. Spout glass temp. c5 versus spout glass temp. c1

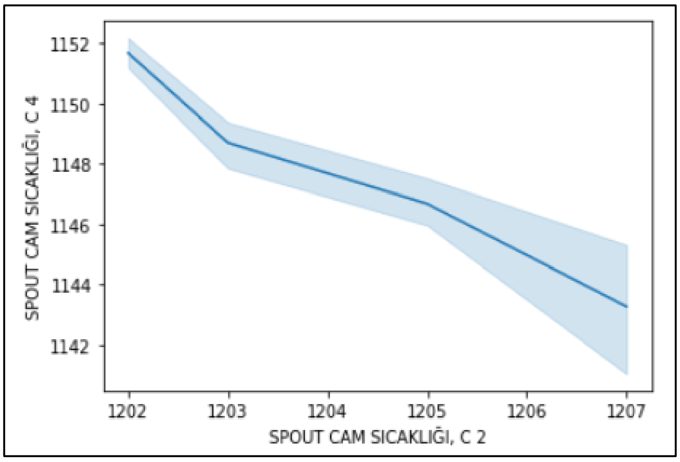


Figure 2.66. Spout glass temp. c2 versus spout glass temp. c4

The CH(Working pool) temperature zone has a highly positive correlation with each other. The representative graph of the highest correlation for çh(working pool) zone temperatures is on the graph with "Çh zone temperatures (˚C)" versus "Çh zone temperatures (˚C) 905".

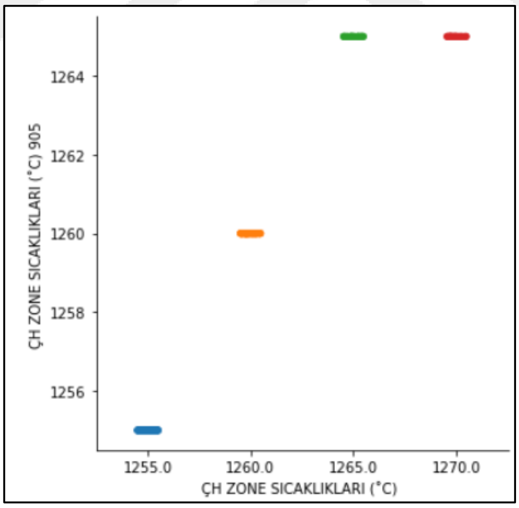


Figure 2.67. Ch zone temp. versus ch zone temp. 905

2.1.4.4. *Electric and Energy Consumption*

In the analysis of "*Toplam Elek_KWh*" amount and "*Harman_altı_elek_KWh*" consumption, *Harman_altı_elek_KWh* consumption varies between 8000 KWh and 12500 KWH values, *Toplam Elek_KWh* value is concentrated between 30000 KWh and 425000 Kwh. In addition, there is a linear relationship between them.

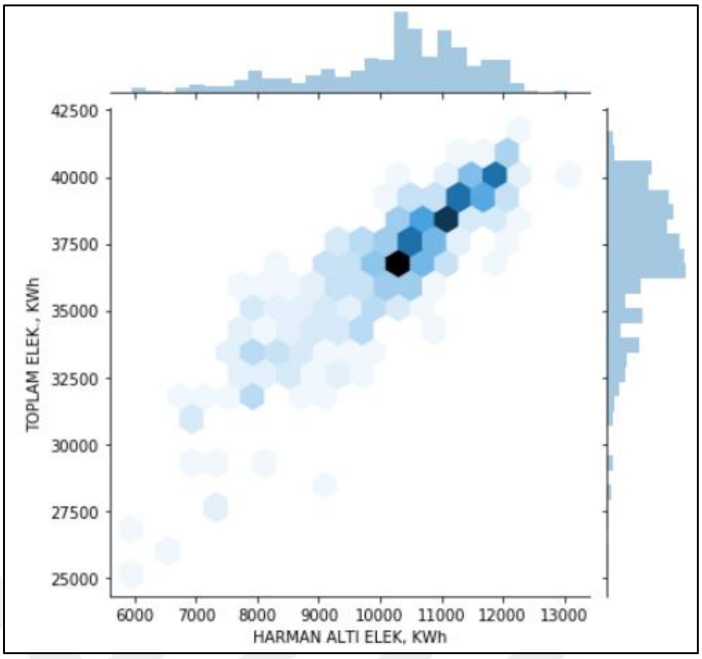


Figure 2.68. Underthreshing electric versus total electric

When the total electricity consumption and under threshing electric consumption are examined on the graph that includes the color types of wine bottles, the distribution of *IR_GREEN*, *UV_GREEN*, *AMBER*, and *FLINT* colors can be seen as follows. It can be understood that *UV_GREEN* color exists for 50K electricity consumption amount and it is also higher than approximately 1000 kcal/kg energy consumption. *IR_GREEN* has the lowest energy consumption and its electric consumption is between 40K and 85K. Amber has an average amount of both energy and electric consumption. On the other hand; Energy consumption is changing at the specific range but electric feature scatters randomly in *FLINT* bottle color.

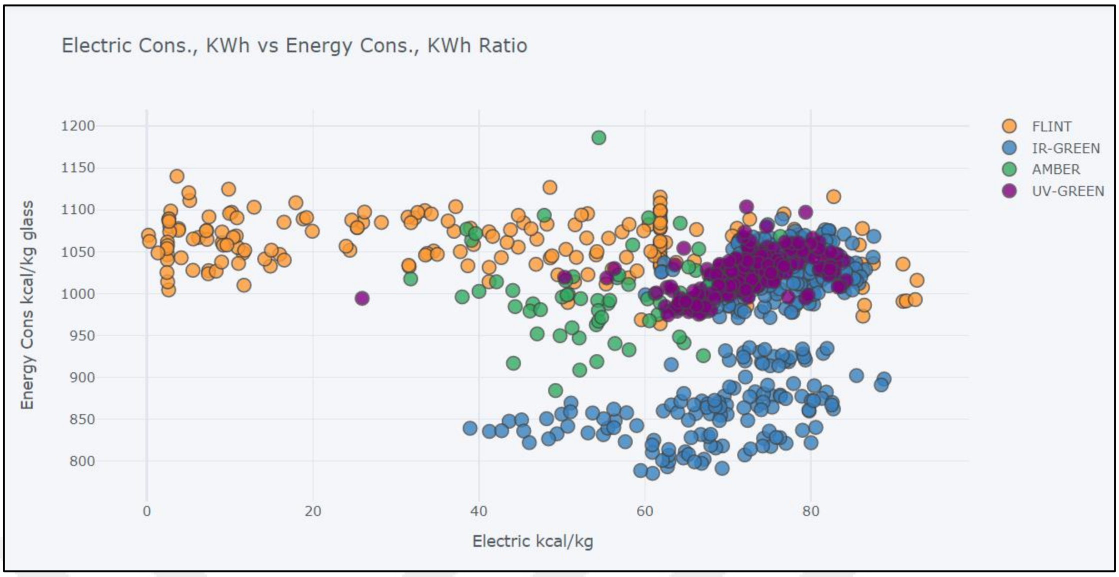


Figure 2.69. Underthreshing electric versus total electric on glass colors

When the cullet rate and natural gas consumption are examined in the furnace data, it is seen that the natural gas consumption is high when the cullet is low, as in the energy dataset. When the base 820 region and the cullet variables are examined, it is seen that the temperature increases as the cullet rate increases. Wine bottle colors are charted. Flint, IR-Green, Amber, and UV-Green colors are available and Flint has the least cullet rate. IR-Green has the highest rate of cullet. As it is known, the cullet rate varies according to the sensitivity of the color. Looking at the temperatures in the furnace data, there is a positive linear relationship between the Bottom 816 and the Bottom 820. The Riser Optic temperature is between 1275° C and 1290° C, while the glass surface optic temperature is between 1570° C and 1575° C. There is no relationship between them. When the bottom 816 temperature and Riser Optic are evaluated, there is a linear relationship between them. Although there is density at certain temperature points, the Bottom 816 is between 1110° C and 1130° C and the riser Optic is between 1265° C and 1300° C. Right forehead Again, there is a linear relationship between the optics and the glass surface. Glass surface optic with right forehead optic between 1555° C and 1580° C. There is a linear increase between Spout glass temperature-C1 and C5 up to C1 1220° C. After this temperature, C5 increases while C1 remains stable. There is also a linear relationship between Spout Glass Temperature C4 and C2. It is seen that the working pool temperatures are fixed at certain degrees. When the under threshing electricity consumption and total electric are evaluated, a linear relationship is observed. Total electricity is 37000 kWh, and the under threshing electric is 11000 kWh.

When we look at the glass colors in energy consumption, flint color generally consumes high energy and less electricity, while IR Green consumes low energy and high electricity.

2.1.5. Visualizations of Bubble Data

Bubbles are one of the most important quality problems caused by various reasons in the glass production process. It contains gases characteristic of the melting atmosphere, which may be air, combustion gases, or some gas deliberately introduced to control chemical reactions with the batch.

In terms of science, bubbles are formed two forms:[17]

Table 2.8. Information of bubbles

1	The physical trapping of atmospheric gases during the initial phase of batch melting.
2	The decomposition of batch components

The presence of bubbles in a glass sample is a subject for many scientific studies and bubbles are definitely undesirable in most glass bottles. Bubbles in products are almost always considered flaws and result in rejection of the product.

A small decrease in furnace floor temperature can convert a previously undersaturated melt into a highly supersaturated one, with the consequent formation of many bubbles.

Holding of the atmospheric gases is enhanced by use of very fine sand in the batch, or by the use of batch components of widely differing particle sizes. And, mechanical stirring of a melt can introduce bubbles by trapping air and forcing it into the melt.

On the other hand, dissociation of batch materials can produce extremely large quantities of gases. Reactions with metals in contact with the melt can generate oxygen, carbon dioxide, or hydrogen by electrolytic reactions[18]. Corrosion of refractories can open previously closed pores to the melt, releasing the gas contained in those pores into the melt. The products of all of these reactions can agglomerate to form bubbles.

In obtained data regarding of bubbles, it is handled to quantity of important features. "*Toplam30gr*"(Total30gr), which is the most valuable column in the bubble data, the bubble counts, analysis and evaluation of the operation data are carried out. Studies are carried out on the sources that may cause bubble faults in the production series that are seen as problematic. In the chart below, it can be seen that the values in the *Total30gr* column of the dataset consisting of 901 rows have increased at some points.

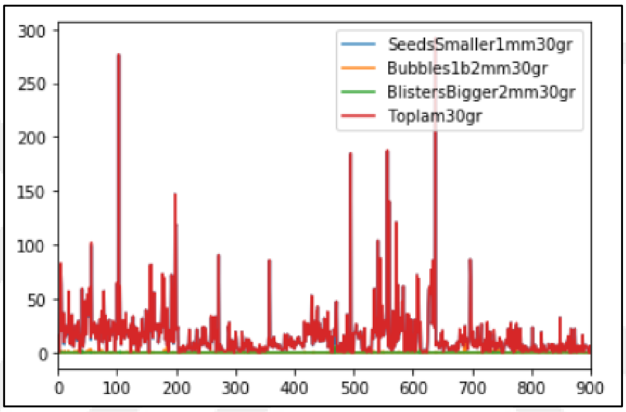


Figure 2.70. Bubble data features distribution

Let's consider important ratios such as min, max, mean. It is seen that the average of a "*Toplam30gr*" column is approximately 15.53. Standard deviation of the *Toplam30gr* is approximately 23.16. So it is not composed of big troubles in quality dataset because it can be given error graphs above values of 64.

Table 2.9. Statistical information of bubble data

	count	mean	std	min	25%	50%	75%	max
Ürün	901	-	-	-	-	-	-	-
SeedsSmaller1mm30gr	901	15.3	23.07	0	3.4	9.2	19.2	291.1

Bubbles1b2mm30gr	901	0.19	1.03	0	0	0	0	26
BlistersBigger2mm30gr	901	0.023	0.26	0	0	0	0	5.7
Toplam30gr	901	15.53	23.16	0	3.6	9.5	19.5	291.9

It is known that if the bubble dataset specified in the "*Toplam30gr"* column in the Bubble count is greater than 64, the product decomposes as poor-quality. In this case, it is clear that a very small amount of unquality products are produced.

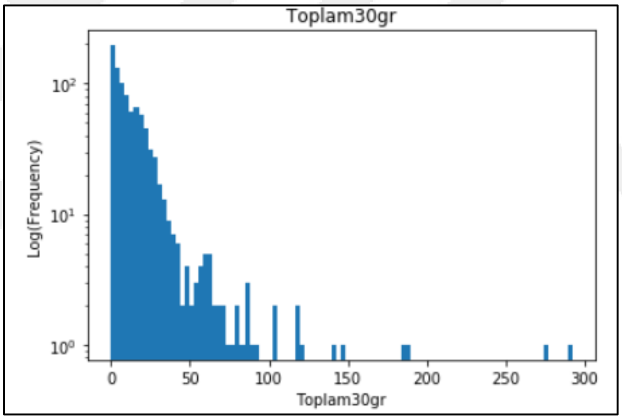


Figure 2.71. Total30gr frequency

It is informed that the product is of good quality when the number of bubbles is lower than 64 value. So, as can be seen in the chart below, the distribution of good quality and not good quality products are given in Figure 2.72. Total30gr distribution by specification status.

Figure 2.72. Total30gr distribution by Specification Status

In Figure 2.73. Bubbles1b2mm30gr versus toplam30gr by quality status, *Total30gr* column and *Bubbles1b2mm30gr* column relations on quality status is plotted . Since the values in the *Bubbles1b2mm30gr* column are very low, it remains between 0 and 64 on the x-axis and between 0 and 5 on the y-axis.

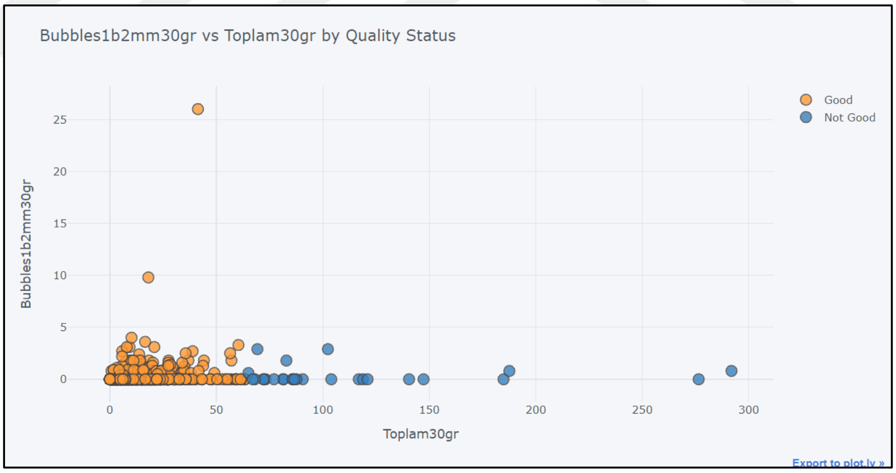


Figure 2.73. Bubbles1b2mm30gr versus toplam30gr by quality status

Figure 2.74. Toplam30gr over time by quality status is drawn using the pivot table in python within the given bubble data. As can be seen in this chart, the quality status in *Toplam30gr* column varies according to the dates. In addition, not good quality products are clearly shown with a blue icon and quality products are shown with yellow icon.

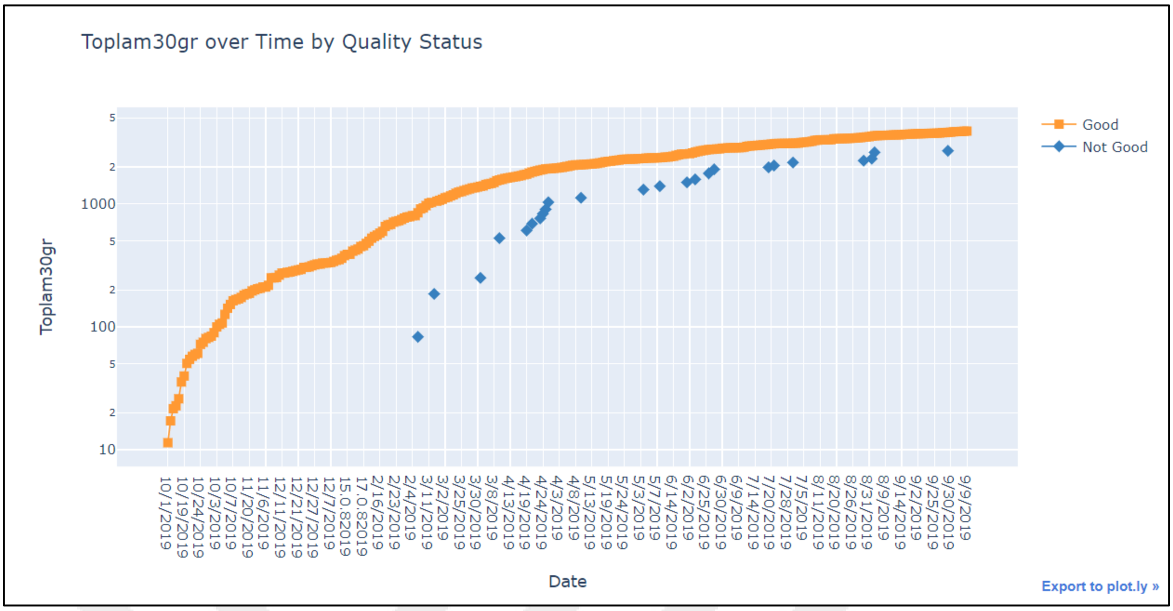


Figure 2.74. Toplam30gr over time by quality status

In summary, in Bubble dataset The 30gr column, which is an important feature for quality, was drawn on the value determined by the glass experts for the wine bottles. If the reference value is less than 64, the product is considered to be of good quality. The devices measure Bubbles1b2mm30gr column at 4cm² and the glass weight corresponds to 30gr. The values given in this column give the number of bubble made in 30 grams. A total of 30gr is the bubble count made in 1 kg glass. The values compared with the spec values are the Total 30gr column. A comparison was made in order to find the relationship between both columns. In addition, the measurements until the end of the year are determined with a graphic. It has been observed that quality products are produced in most of the production. The reference values for examining the bubble are given as follows.

Table 2.10. Bubble range

Bubble
(1-2mm. Max.3 ad/40 cm², >2-3 mm. Max.2 ad/ 40 cm²)
1mm> Bubbles>0.3 mm max. 80 ad/40 cm²

2.1.6. Visualizations of Production Data

This data set includes production data. Production quantities, working times, distribution of the products produced within the year and month will be seen. Preliminary information about production will be obtained with different graphics.

2.1.6.1. Wine Bottle Color

In the production data, two colors are seen for wine bottle production which is different from other dataset. It is shown that the color of Emerald Green is 196 pieces and there are 21 if it is colorless in Figure 2.75. Distribution of colors in production data.

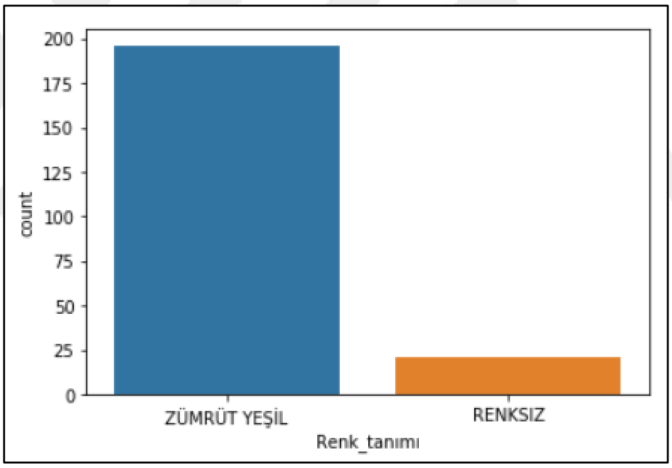
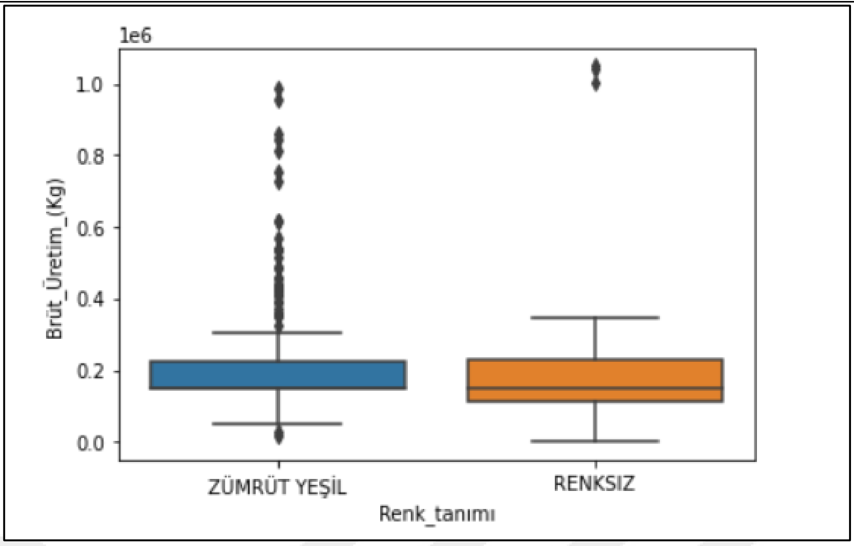
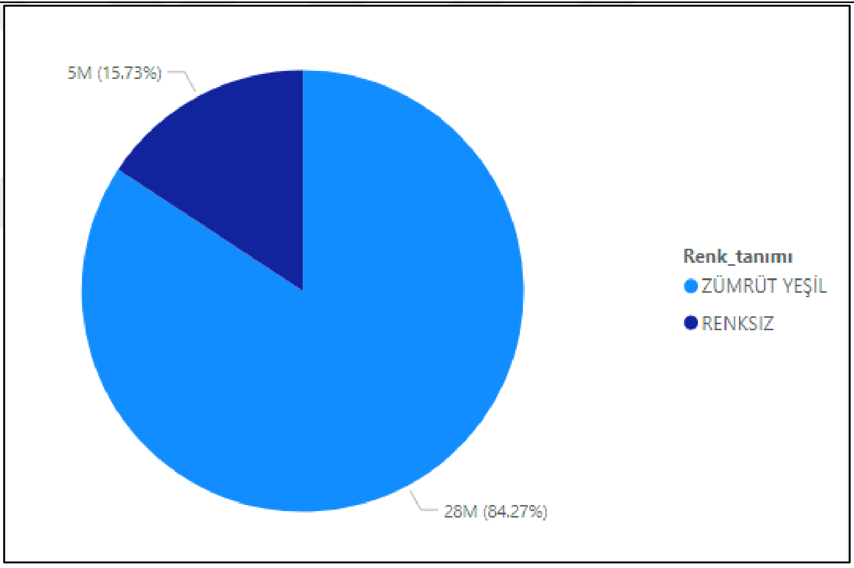


Figure 2.75. Distribution of colors in production data

When the gross production and color variety are compared, it is clear that the emerald green color is predominant. And, it can be seen that the circles are outlier, the maximum value of the emerald green color is 600000 and the colorless is around 500000.



(a)



(b)

Figure 2.76. Distributions of colors on gross production. (a) Box plot, (b) Pie plot

2.1.6.2. Wine Bottle Colors on Production Quantity

It analyzed wine bottle colors according to the mean of gross production quantity in handled production data in Figure 2.77. Wine bottle colors on mean of gross production quantity which is interactive visualization by utilizing Pygal library. Pygal allows the user to create a beautiful interactive plot with an optimal resolution for printing or being displayed on webpages using Flask or Django which needs to enable IPython display and HTML options. It is a really special library that can help us create eye-catching, stunning, interactive visualizations.

It can be seen that uncolourn wine bottle is much more than an emerald color as can be seen below.

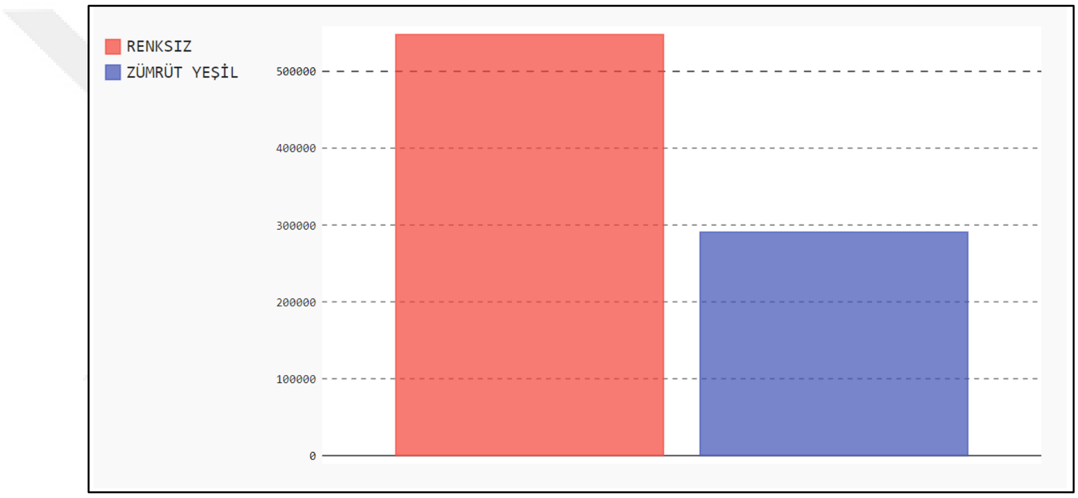


Figure 2.77. Wine bottle colors on mean of gross production quantity

On a monthly basis, it is noteworthy that the highest production is in December, July and November.

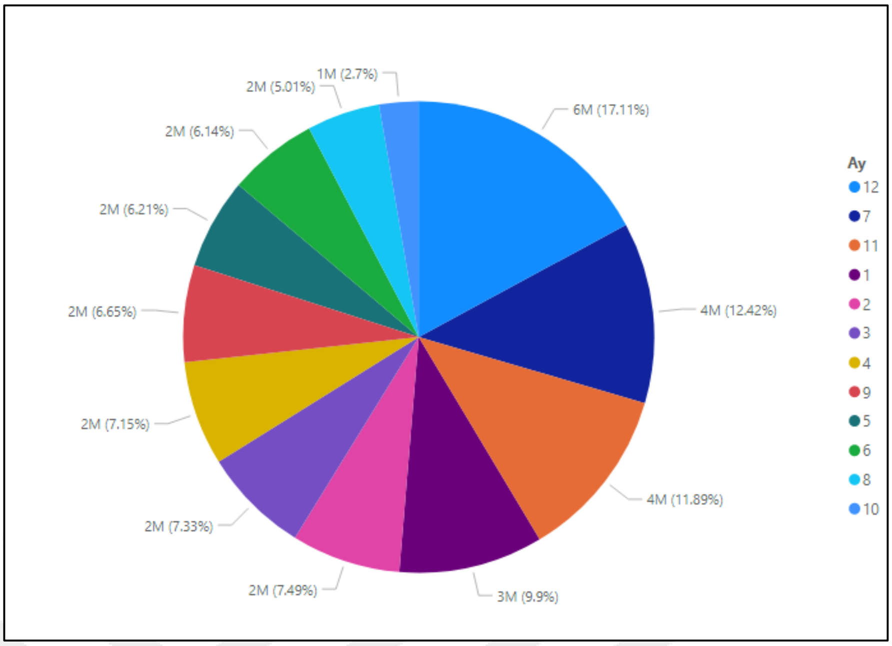


Figure 2.78. Gross production(kg) by month

It is intended to learn how month and *Pull(kg)* by *production(kg)*. As seen in the graph, balloon sizes are changing by production(kg) values in wine bottles.

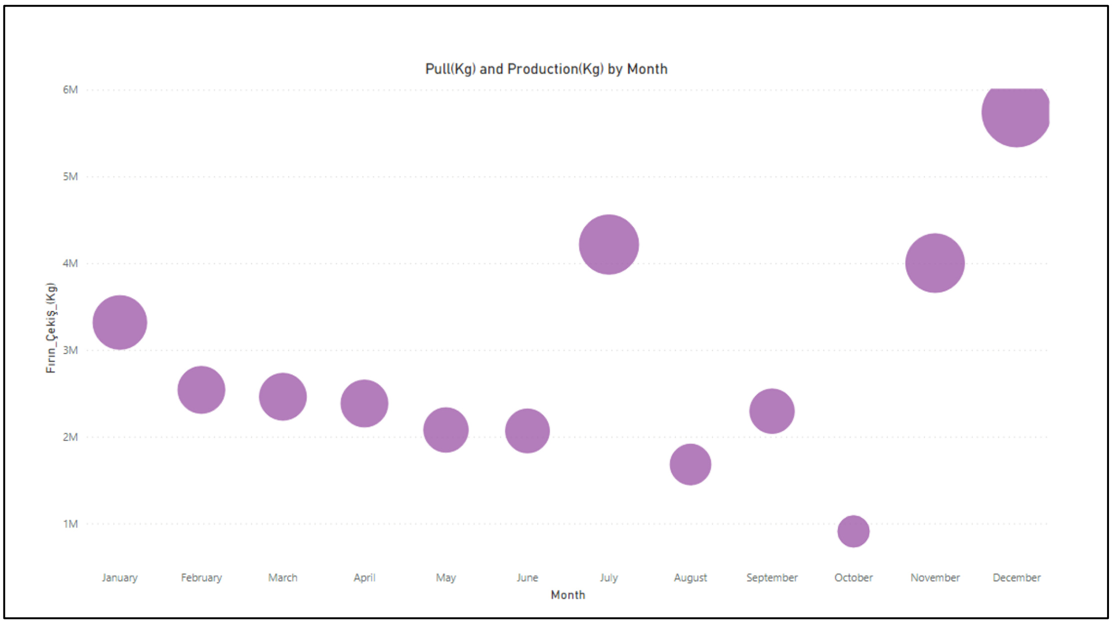


Figure 2.79. Pull(kg) and production(kg) by month

2.1.6.3. Furnace Traction

"*Fırın_Çekiş*" column values has a relation with "*Net_Çalışma_Süresi*"(Net working time) column linearly, which it can be seen very closely related in the production dataset correlation indicated in Figure 2.80. Furnace traction(kg) versus net working time. Then "Fırın_Çekiş_(kg)" (Furnace traction) values is around 200,000 kg.

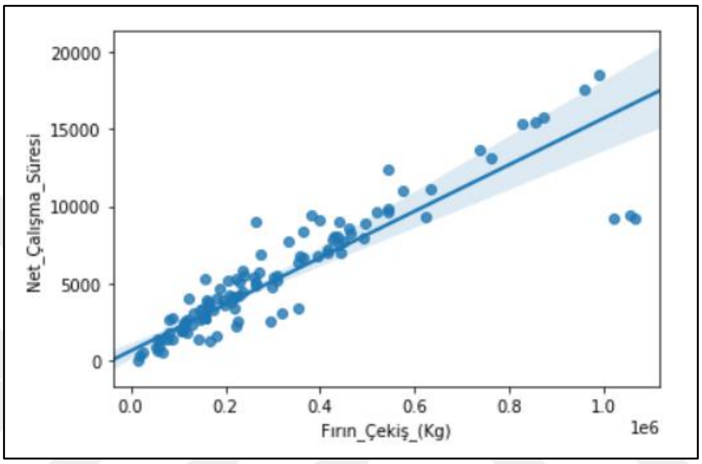


Figure 2.80. Furnace traction(kg) versus net working time

When the relationship between "*Fiili_Gramaj*"(Actual Weight) column and "*Fırın_Çekiş*" (Furnace Traction) column is compared, "*Fiili_Gramaj*" values are between 350 kg and 700 kg. "*Fırın_Çekiş_(Kg)*" cloumn has increased on a horizontal plane, and the density is between 200,000 kg and 500,000 kg.

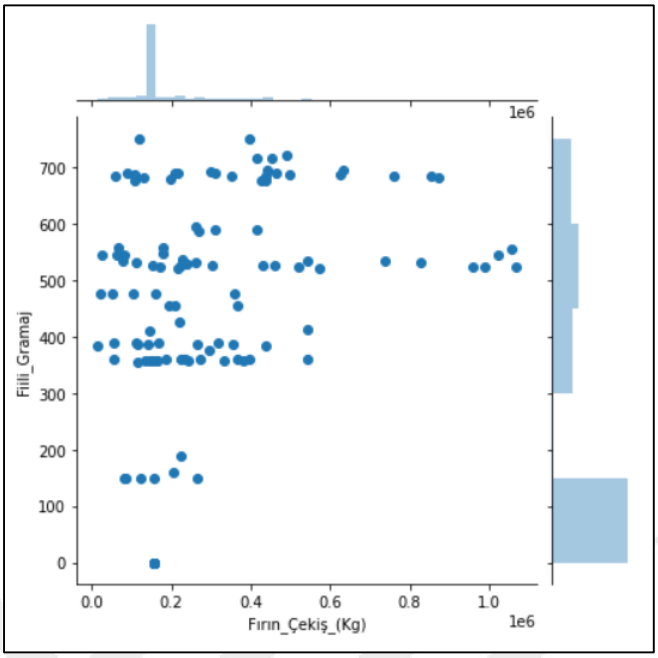


Figure 2.81. Furnace traction(kg) versus actual weight(kg)

The visual of the two important values used to calculate the ratio of the" *Fırın_Çekiş*" and the "*Üretim(kg)*"(Production) data, where the production efficiency for the wine bottle is calculated, is as follows.

Due to the missing values in the "*Üretim(Kg)*" values, the density at 0(zero) point in the graph draws attention. However, if these points were entered, it should be said that the density in "*Üretim(kg)*" column would be between 250,000 and 750,000 in the same linear table by considering the "*Fırın_Çekiş(kg)*" points. Likewise, the missing values in "*Fırın_Çekiş(kg)*" column show the noise at the starting point of the graphic.

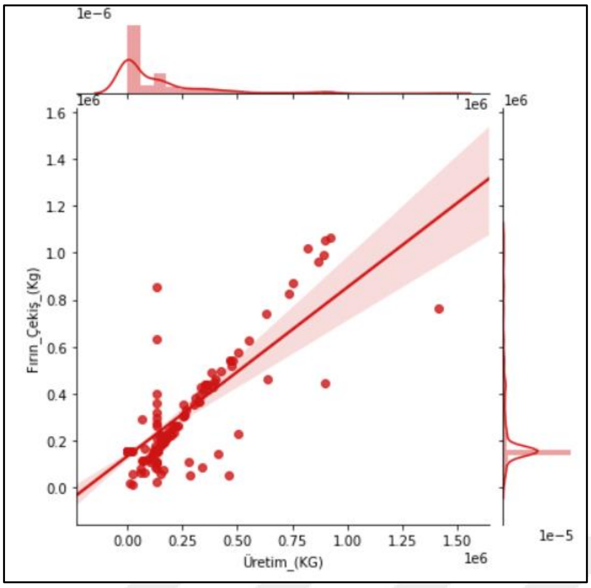


Figure 2.82. Production(kg) versus furnace traction(kg)

To understand some details using movable graphs, Holoviews is utilized in some specific features. It is an open-source python library that makes data visualization easier. Holoviews works on conveying the message that data is trying to tell rather than focusing on how to plot visualizations. Holoviews works on Numpy and Params, and for visualization, it supports ‘Bokeh’ and ‘Matplotlib’.

It is created visualization with a bar plot of “*Kabul_(Adt)*” and “*Ret_(Adt)*” according to “*Fırın_Çekiş_(Kg)*” feature by utilizing Holoviews. It is able to see the change in the number of acceptance and rejection according to the furnace traction values, via the bar on the side.

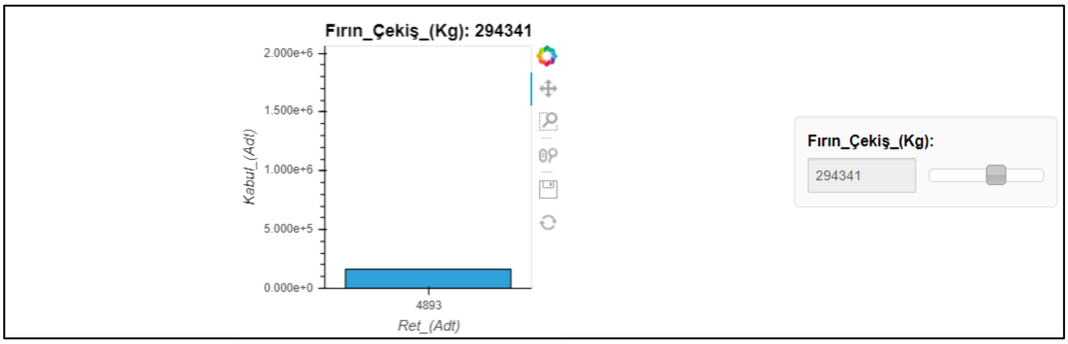


Figure 2.83. Acceptance and rejection according to furnace traction(kg)

It is used plotting a scatter plot between our target variable '*Fırın_Çekiş_(Kg)*' and "*Üretim_Adet*" and "*Brüt_Üretim_(Kg)*" feature variables. For creating the visualizations, it is started by defining our feature variables and then creating a dataset using Holoviews. Thus, the difference can be seen easily at the same interval.

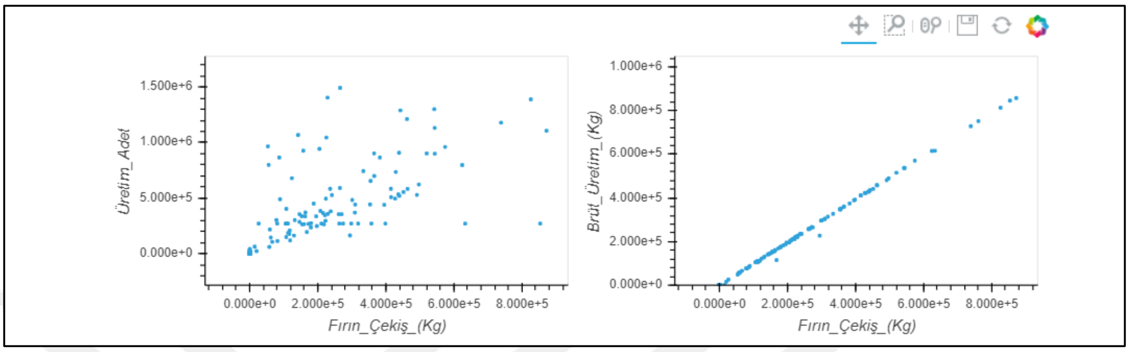


Figure 2.84. Production quantity and gross production(kg) according to furnace traction(kg)

Looking at the production data, the wine bottle colors in production are divided into two. It is Emerald Green and colorless. The color with the highest gross production is emerald and the highest production is realized in December. If I examine the furnace traction according to the gross production on a monthly basis, the furnace traction also increases according to the abundance of production. As the furnace traction increases, the net working time of the furnace also increases. They are in a linear relationship with each other. However, there is no correlation between furnace traction and actual weight. Actual weight standards take values. In order to calculate the production efficiency, the furnace traction and production columns are taken into account in the graph drawn. Efficiency is calculated as 84%. As the furnace traction increases, the amount of acceptance and rejection also increases, as the production will increase. Likewise, there is a linear relationship between furnace traction and production quantities.

3. CORRELATION ANALYSIS

Correlation analysis; is a statistical method that gives information about the relationship between variables.

The correlation coefficient is a value that shows the strength of the relationship between the dependent variable and the independent variables. In this part of this thesis, correlation tables are created with datasets of quality laboratory, glass furnaces, bubbles, production, energy, and glass components. To define the correlation, for example, whether there is a relationship between natural gas consumption and the rate of cullet or the relationship between the temperature in the furnace Bottom 820 region and the temperature in the furnace Bottom 817 region can be examined with the correlation coefficient.

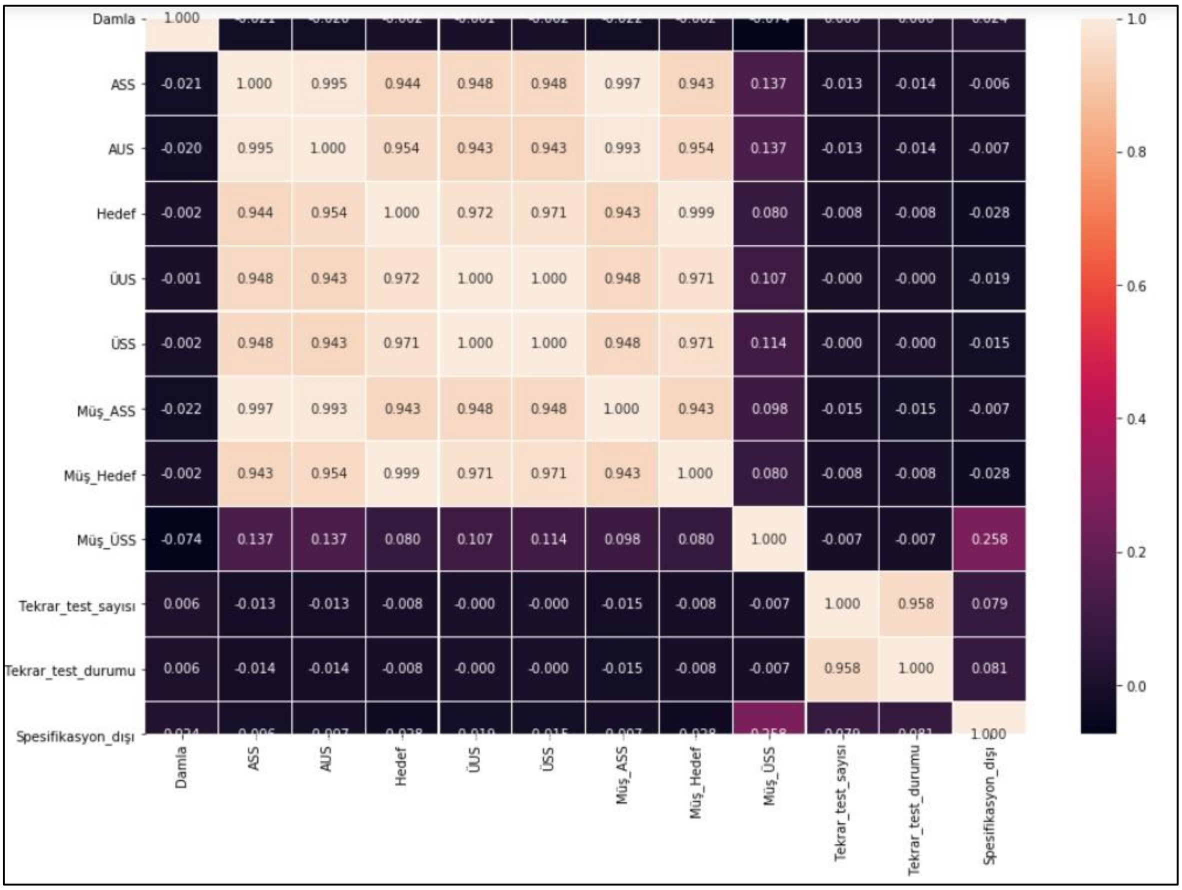
The correlation coefficient gives information about the direction of the variables and how the interactions are. It is defined as the ratio of explained variance and unexplained variance. It is the measure of the linear relationship between two variables and is independent of the units of the variables studied between -1≤ correlation coefficient ≤1. When the coefficient approaches zero, it indicates that there is a weak relationship between the variables. It is said that there is a positive relationship if the variables increase or decrease together, and if one variable increases and the other decreases, there is a negative relationship.[19]

3.1. CORRELATION BETWEEN QUALITY LAB DATA VARIABLES

In this section, the parameters to be controlled during the quality analysis of the samples will be evaluated. The spec values of the products produced with the customer's demands are examined by correlation and the relationship between them is observed.

"*ASS*" (Lower spec limit), "*AUS"* (Lower warning limit), *"ÜUS"* (Upper warning limit), *"ÜSS"* (Upper spec limit), *"Hedef"*(Target), *"Müşteri Hedef"*(Customer Target) and *"Müşteri ASS"*(Customer Lower Spec limit) features are highly positive correlation with each other. *"Müş_ÜSS"*(Customer Upper Spec Limit) is not related with other features. There is a negative correlation between *"Spesifikasyon_dışı"* columns with "*tekrar test sayısı"*and *"tekrar test durumu"* but they are positively correlated with each other.

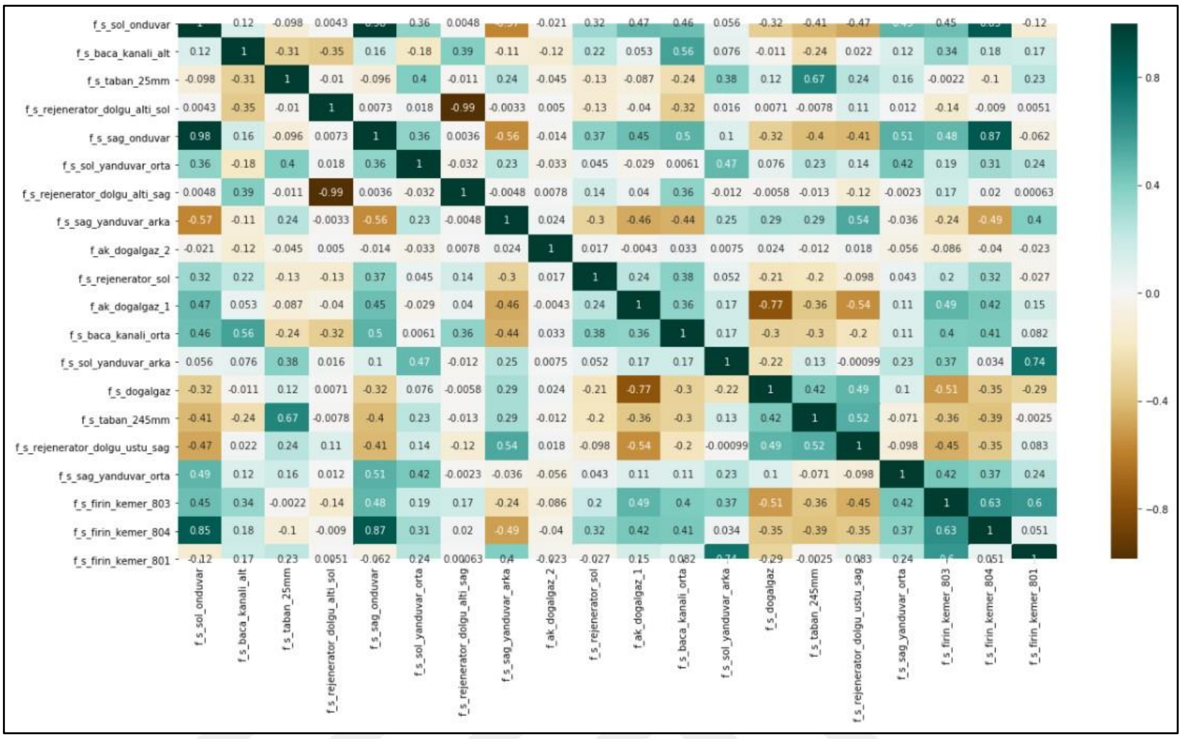
Table 3.1. Quality lab. data correlation



3.2. CORRELATION BETWEEN ENERGY DATA VARIABLES

Table 3.2. Energy data correlation shows the correlation of the part of the energy data related to temperatures. The interaction of furnace temperatures in different regions with each other can be easily understood. For example, there is a high positive correlation between the furnace front left wall and the furnace right wall. There is also a significantly high positive correlation between the right front wall of the furnace and the furnace belt 804 region. While the data are generally positively correlated, there is also a high negative correlation between the right region under the regenerator filling and the left region almost -1 value. It is also shown that other features have negative and positive correlations with small differences with each other.

Table 3.2. Energy data correlation



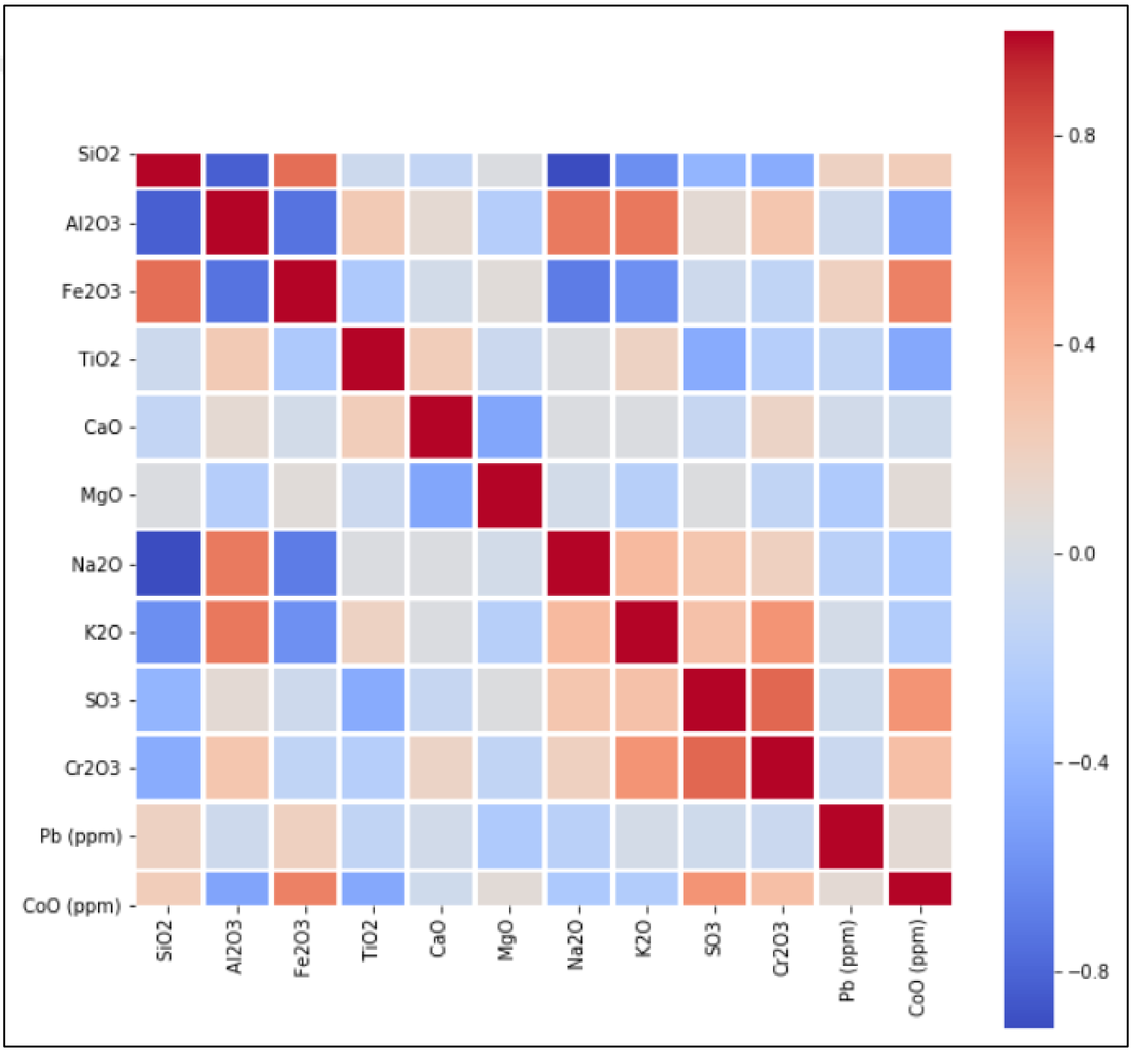
3.3. CORRELATION BETWEEN GLASS COMPONENTS DATA VARIABLES

Glass is melted in large refractory-lined tank furnaces, with capacities up to 2000 ton, using natural gas at temperatures in excess of 1500° C. The typical raw materials consist of silica (in the form of sand), soda ash, limestone, dolomite, and other minor ingredients.

When looking at the correlation between glass components, *SiO2* and *Fe2O3* components are positively closely related and highly correlated with *Al2O3* and *Na2O*, and negatively associated with *K2O*. It has low correlation with other glass components especially *MgO* and *TiO2*. Then, *Cr2O3* has a relation with *SO3* positively. The others haven't got a closer relation between each other. The values are roughly zero so it can be accepted as not effective on correlation. The graphic distribution of glass components related to bottle quality can be seen with figures in section 2.1.3.

Table 3.3. Glass component data correlation

	SiO2	Al2O3	Fe2O3	TiO2	CaO	MgO	Na2O	K2O	SO3	Cr2O3	Pb (ppm)	CoO (ppm)
SiO2	1.000000	-0.825603	0.704506	-0.072002	-0.129642	0.027245	-0.907666	-0.608747	-0.400396	-0.448076	0.182052	0.220509
Al2O3	-0.825603	1.000000	-0.735315	0.245273	0.099071	-0.214272	0.661745	0.666823	0.091985	0.274136	-0.065380	-0.493521
Fe2O3	0.704506	-0.735315	1.000000	-0.248957	-0.031246	0.069670	-0.693189	-0.600014	-0.061729	-0.146592	0.193190	0.629962
TiO2	-0.072002	0.245273	-0.248957	1.000000	0.220336	-0.080144	0.030670	0.179940	-0.453819	-0.206820	-0.138816	-0.475127
CaO	-0.129642	0.099071	-0.031246	0.220336	1.000000	-0.486041	0.026446	0.027025	-0.108013	0.159088	-0.036086	-0.053931
MgO	0.027245	-0.214272	0.069670	-0.080144	-0.486041	1.000000	-0.032958	-0.196784	0.038090	-0.137967	-0.237632	0.076444
Na2O	-0.907666	0.661745	-0.693189	0.030670	0.026446	-0.032958	1.000000	0.350816	0.273413	0.194351	-0.182213	-0.257057
K2O	-0.608747	0.666823	-0.600014	0.179940	0.027025	-0.196784	0.350816	1.000000	0.300642	0.549831	-0.021722	-0.225964
SO3	-0.400396	0.091985	-0.061729	-0.453819	-0.108013	0.038090	0.273413	0.300642	1.000000	0.734441	-0.054153	0.549939
Cr2O3	-0.448076	0.274136	-0.146592	-0.206820	0.159088	-0.137967	0.194351	0.549831	0.734441	1.000000	-0.084108	0.317710
Pb (ppm)	0.182052	-0.065380	0.193190	-0.138816	-0.036086	-0.237632	-0.182213	-0.021722	-0.054153	-0.084108	1.000000	0.093954
CoO (ppm)	0.220509	-0.493521	0.629962	-0.475127	-0.053931	0.076444	-0.257057	-0.225964	0.549939	0.317710	0.093954	1.000000



Within the specification values, the glass that is determined as good is blue, and the bad and alarming values are red. As can be seen from the graph, it is seen that the data mostly shows a homogeneous distribution of the specification status.

As a result of the correlation, when the specification status of *Na2O, CaO, Cr2O3, SiO2* components, which are closely related, in the chemical of glass, Sodium Oxide takes a value between 69.5% and 70%. Calcium oxide is also about 10 percent, Chromium oxide is dense at 0.24 to 0.28 percent. Sodium oxide, on the other hand, is around 14-15% and when it is desired to observe with spec values, the spec values are not at certain points and show a homogeneous distribution to a large extent. There are other factors affecting the quality in these distributions and the quality results differ due to different features when the proportions of closely related chemical components decrease or increase in a certain range.

If you want to see the interaction of glass components with each other, see Table 3.3. Glass component data correlation.

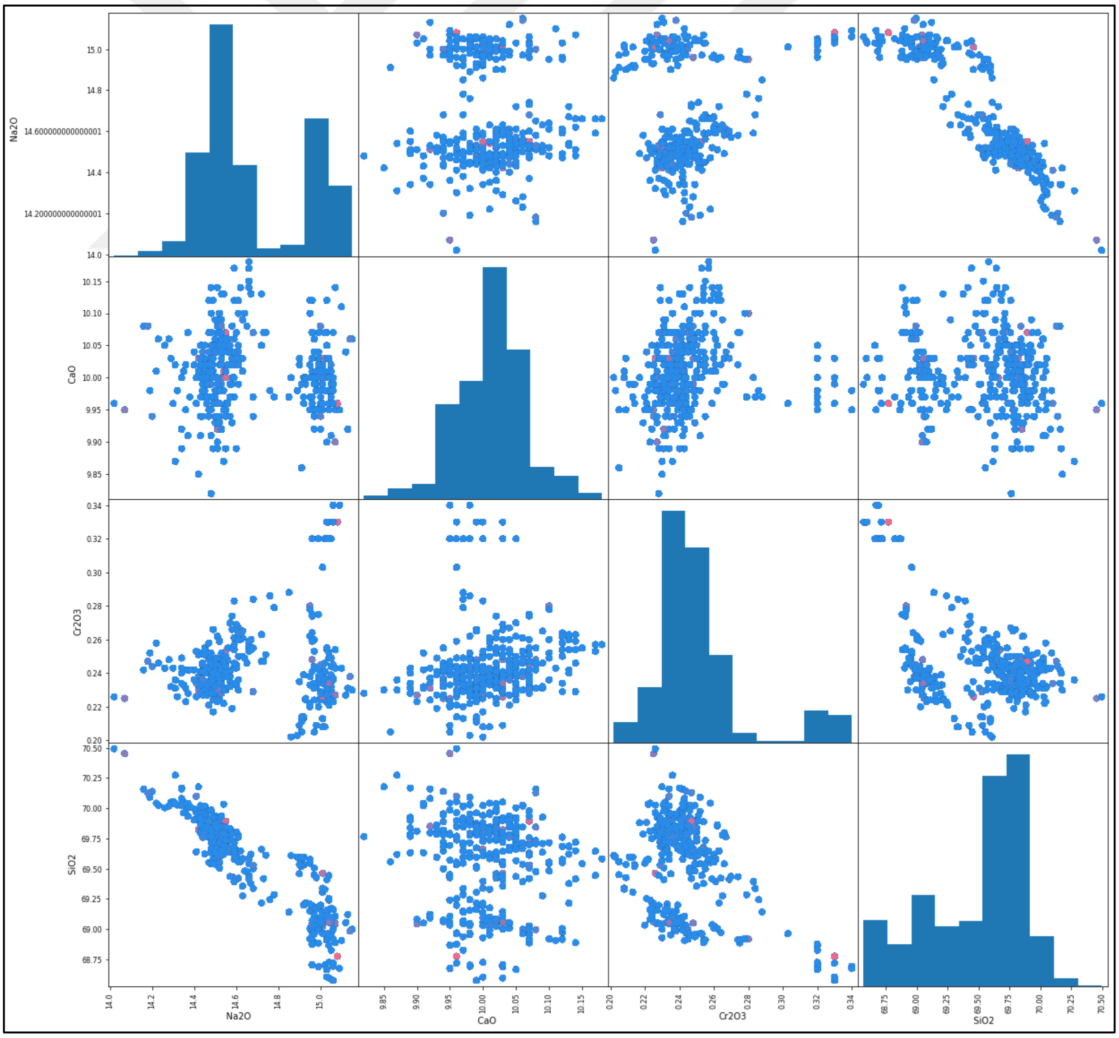


Figure 3.1. Relation between glass components by spec status

3.4. CORRELATION BETWEEN FURNACE DATA VARIABLES

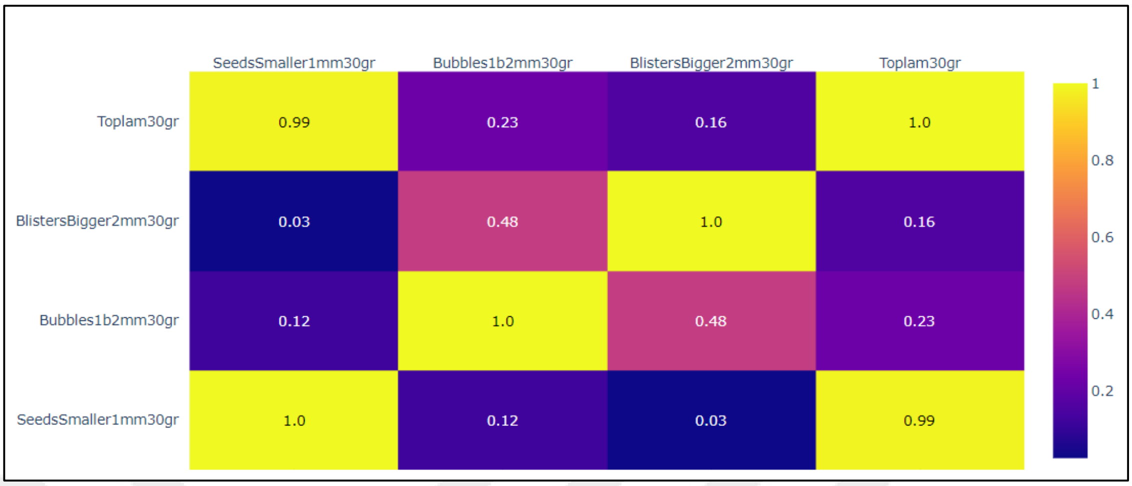
Cullet is negatively associated with *“NGas consumption kcal/kg"* and *“N.GAS consumption Sm3 @15C SM3 @15 C”* and *“AIR/FUEL RATIO*” but positively correlated with “*Fe+2*” and “*Electric kcal/kg"*. At the same time, *“SiO2”, “Al2O3”, “Fe2O3”, “TiO2”, “K2O”, “SO3”, “Cr2O3”* are correlated positively each other and negatively correlated with *“AIR/FUEL RATIO”* and *“CaO”, “MgO” , “Na2O”* are shown that scattered negatively but positively correlated with *“AIR/FUEL RATIO”*. If it is focused on glass components, they have such close relations negatively or positively. “*Electric Boost Zone1”, “Electric Boost Zone2 ", “Total boost”, “Electric kcal/kg "* columns are seen to be positively associated with each other. It can be said that *“NGas consumption kcal/kg "* and *“Energy Cons kcal/kg glass”* are also very closely related to each other as expected.

Table 3.4. Furnace data correlation

3.5. CORRELATION BETWEEN BUBBLE DATA VARIABLES

After definitions such as weight and height of the bottle are made in the measuring device, it gives a result for 1 kg in the number of bubbles or seed. Reduced to 30gr with the formulation, the *Toplam30gr* feature is written in the column. It is the part that is compared with the spec value. In correlation analysis of bubble data obtained from all product lines, it can be said that "*Toplam30gr"* feature has positive correlation with "*SeedsSmaller1mm30gr"* feature.

Table 3.5. Bubble data correlation



3.6. CORRELATION BETWEEN PRODUCTION DATA VARIABLE

It has a positive correlation with the "*Üretim_adet*" (Production quantity) column in the production data, and with the other columns except from "*Yeniden_ayırma*" (The Reallocation) and "*Std_gr*" columns. The values of "*Üretim_ (Kg)*"(Production_kg), "*Brüt_Üretim_ (Adt)*"(Gross_quantity) and "*Fırın_Çekiş_(Kg)*"(Furnace traction), which are observed to be highly correlated, have positive correlations. There is a positive correlation between "*Net_Çalışma_Süresi*"(Net Run Time) and "*Fırın_Çekiş_(Kg)*"(Furnace traction), "*Brüt_Üretim_(Kg)*"(Gross production(kg)), "*Brüt_Üretim_(Adt)*"(Gross production quantity), "*Kabul*"(Acceptance_ (Unit)). "*Çekiş_Süresi_dak*"(Traction Time minute) and "*Net_Çalışma_Süresi*" (Net Run Time), " *Çalışma_Sür_(dak)*" (Traction Time minute) are on the same line. It can be seen in the correlation table below that there is a negative correlation between " *Yeniden_Ayırma*" (Reallocation) column and the "*Std_g*" column in the data set with the other columns.

Table 3.6. Production data correlation

	Sipariş	Fiili_Dev._(damla/dk	Fiili_Gramaj	Fırın_Çekiş_(Kg)	Brüt_Üretim_(Kg)	Üretim_(KG)	Brüt_Üretim_(Adt)	Kabul_(Adt)	Ret_(Adt)	Yeniden_Ayırma	Çalışma_Sür._(dak)	Çekiş_süresi_dak	F/H_Açık_dakika	Net_Çalışma_Süresi	Üretim_Adet	Std_gr
Std_gr	-0.1	-0.32	0.16	0.07	0.08	0.08	-0.15	-0.13	-0.09	-0.15	-0.0	-0.0	-0.12	0.01	-0.13	1.0
Üretim_Adet	0.26	0.61	0.54	0.81	0.81	0.93	0.87	1.0	0.39	-0.11	0.79	0.79	0.35	0.8	1.0	-0.13
Net_Çalışma_Süresi	0.25	0.52	0.71	0.95	0.95	0.8	0.92	0.8	0.37	-0.14	1.0	1.0	0.42	1.0	0.8	0.01
F/H_Açık_dakika	0.23	0.53	0.44	0.46	0.44	0.33	0.46	0.35	0.24	-0.1	0.5	0.5	1.0	0.42	0.35	-0.12
Çekiş_süresi_dak	0.26	0.55	0.72	0.95	0.95	0.8	0.92	0.79	0.37	-0.14	1.0	1.0	0.5	1.0	0.79	-0.0
Çalışma_Sür._(dak)	0.26	0.55	0.72	0.95	0.95	0.8	0.92	0.79	0.37	-0.14	1.0	1.0	0.5	1.0	0.79	-0.0
Yeniden_Ayırma	-0.04	-0.16	-0.16	-0.13	-0.13	-0.1	-0.14	-0.12	-0.06	1.0	-0.14	-0.14	-0.1	-0.14	-0.11	-0.15
Ret_(Adt)	-0.03	0.32	0.24	0.33	0.33	0.35	0.37	0.39	1.0	-0.06	0.37	0.37	0.24	0.37	0.39	-0.09
Kabul_(Adt)	0.26	0.61	0.55	0.81	0.81	0.93	0.87	1.0	0.39	-0.12	0.79	0.79	0.35	0.8	1.0	-0.13
Brüt_Üretim_(Adt)	0.3	0.69	0.6	0.93	0.93	0.79	1.0	0.87	0.37	-0.14	0.92	0.92	0.46	0.92	0.87	-0.15
Üretim_(KG)	0.24	0.47	0.62	0.86	0.86	1.0	0.79	0.93	0.35	-0.1	0.8	0.8	0.33	0.8	0.93	0.08
Brüt_Üretim_(Kg)	0.28	0.54	0.7	1.0	1.0	0.86	0.93	0.81	0.33	-0.13	0.95	0.95	0.44	0.95	0.81	0.08
Fırın_Çekiş_(Kg)	0.29	0.55	0.7	1.0	1.0	0.86	0.93	0.81	0.33	-0.13	0.95	0.95	0.46	0.95	0.81	0.07
Fiili_Gramaj	0.41	0.64	1.0	0.7	0.7	0.62	0.6	0.55	0.24	-0.16	0.72	0.72	0.44	0.71	0.54	0.16
Fiili_Dev._(damla/dk	0.45	1.0	0.64	0.55	0.54	0.47	0.69	0.61	0.32	-0.16	0.55	0.55	0.53	0.52	0.61	-0.32
Sipariş	1.0	0.45	0.41	0.29	0.28	0.24	0.3	0.26	-0.03	-0.04	0.26	0.26	0.23	0.25	0.26	-0.1

When the correlation in the quality data set is examined, lower spec values, upper spec values, upper warning and lower warning spec values, customer lower spec value, and target are positively related to each other. However, they are not related to the customer upper spec. values. In the temperature values in the energy data, the features that are generally related to each other are limited. Regenerator underfill right temperature and Regenerator underfill left have a significant negative correlation with each other. The left anterior wall and the right anterior wall are positively correlated with each other. When the correlation between glass components is examined, Fe_2O_3 and SiO_2, Al_2O_3 and Na_2O and K_2O values and Cr_2O_3 and SO_3 values have positive correlations above the average. In addition, Al_2O_3 and Na_2O have a significant negative correlation with SiO_2. When the spec values in the components were examined, it was determined that poor-quality products were not found at certain points. Looking at the furnace data, there is a high positive correlation between the electricity consumption data. There is an above- average correlation between Air/Fuel Ratio values and glass components. CaO, MgO, Na_2O are positively correlated, while SiO_2, Al_2O_3, Fe_2O_3, TiO_2, K_2O, SO_3, Cr_2O_3 are negatively correlated. In bubble dataset, total 30 gr has a positive correlation with the SeedsSmaller1mm30gr variable. Looking at the correlation in production data, all variables are generally positively correlated with each other. Production numbers, actual weight, working hours, acceptance and rejection amounts have a positive correlation with each other.

4. REGRESSION ANALYSIS

4.1. REGRESSION METHODS

Regression analysis is the statistical technique for estimating and investigating the relationship between a dependent variable and one or more independent variables.

If the analysis will be done using a single variable during the analysis, it is called linear regression. However, Multiple regression is used in the anlaysis of relationships between dependent and multiple independent variables.

Linear Regression Analysis is a very popular and simplest form of regression. The relationship between the dependent variable and independent variables is assumed to be linear in nature. Then, Linear Regression fits the data in the line according to a mathematical formula. The most common method used to find a regression line is the least squares method. Remember that the equation of a line is:

$$y = bx + a \tag{4.1}$$

where b equals the slope of the line and a is where the line crosses the y-axis.

Multiple Linear Regression, also called Multiple Regression is a method to predict the dependent variable with the help of two or more independent variables. The main purpose of this analysis finds out the relation between the dependent variable and independent variables. To express it as a mathematical formula, the multiple slopes are typically labelled with the β symbol or letter and numbered as β_1, β_2, etc... The intercept is expressed with β_0. Then, the equation can be written:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \tag{4.2}$$

where the different x_1, x_2, etc... denote different variables.

4.1.1. Linear Regression

Linear regression analysis examines the relationship between a dependent variable and an independent variable. A line equation that aims to show the linear relationship between these variables is formulated and analyzed. It is applied formula below to figure out the slope of the line.[20]

The sum of the products;

$$\sum_{i=1}^{N}\left(x_i - \overline{X}\right) \tag{4.3}$$

The sum of squares of x axes;

$$\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2 \tag{4.4}$$

The sum of squares of the x and y axes are squared before being multiplied together because the denominator must be given a positive result. And, the numerator has a clue in terms of positive or negative relations.

R squared score stands for Coefficient of determination and evaluates the performance of a linear regression model. It works by measuring the amount of variance in the predictions explained by the dataset. The R squared score is calculated to check how well-observed results are reproduced by the model, depending on the ratio of total deviation of results described by the model. It can be calculated by utilizing a mathematical formula that is equal to 1 minus the sum of squares of the residual errors divided by the total sum of the errors. In this research, it is found with the sklearn.metrics library by importing r2_score in Python.

It is clarified that both batch humidity and bubbler per minute columns have 4.804.968 observations in the 2019 dataset and 259.618 observations in the 2020 dataset. It can be analyzed with the Python library easily by utilizing sklearn.linear_model.LinearRegression is a linear regression model. While running machine learning with Regressor_OLS code, 20% of the dataset is used as test data, 20% as validation data, and, 60% as training data and it is trained by .fit() code.

When the dependent-independent relationship between batch humidity and bubbler per minute variables is questioned, it is seen that the batch humidity changes between 2.0 and 4.0 by ignoring the outliers. On the other hand, bubbler per minute has some specific values such as 30, 40, 45, 50, and 55 in the 2019 dataset. Data stack scatters with high values on the total batch humidity and bubbler per minute vary in a certain range.

It is seen that there is no linear relationship between the two variables. All values remain above the slope line.

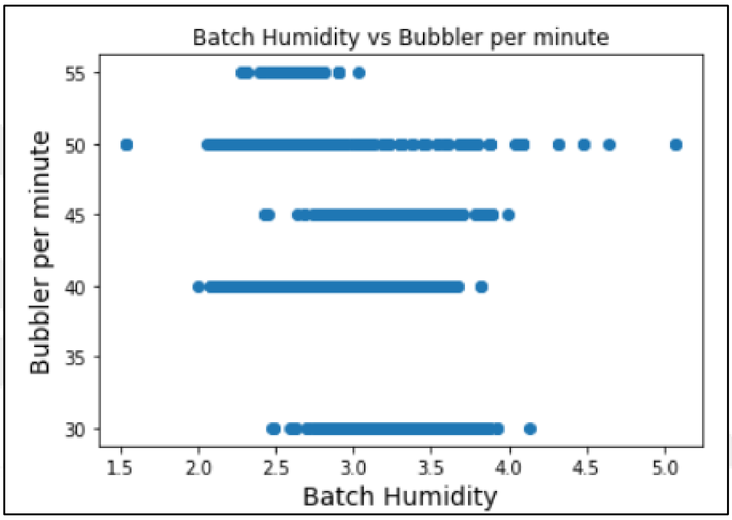


Figure 4.1. Batch humidity versus bubbler per minute in 2019

By the way, it is seen that the batch humidity changes between 2.4 and 3.8 while bubbler per minute has 20, 30, 40, 45, and 55 in the 2020 dataset. It is evaluated the model numerically using R squared, which is nothing but a measure of how close the data points fit the regression line. An R squared value lies between 0 to 1. A model is considered good enough if the R squared value nears 1. Simply using the score() method gives the R Squared value.

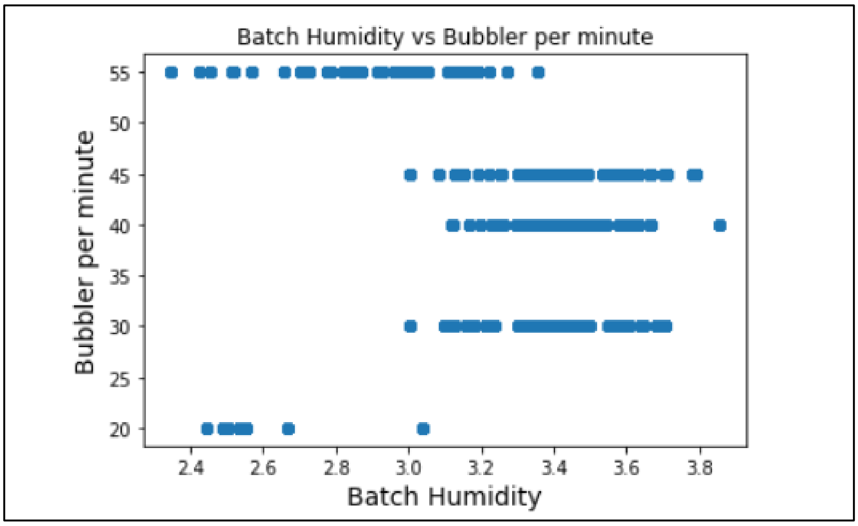


Figure 4.2. Batch humidity versus bubbler per minute in 2020

Table 4.1. R-squared

	2019	2020
R-squared score test data	0.093	0.309
R-squared score validation data	0.094	0.303

The R squared value is close to zero, hence, it can be said that the model is not a good fit. Moreover, R squared also tells the percentage of variation in the dependent variable that is accounted for by its regression on the independent variable. Hence, in this case, it can be said that approximately 10% of the total variation in bubbler per minute is explained by batch humidity in the 2019 dataset while 30% of the variation in the 2020 dataset is seen in Table 4.1 R-Squared.

4.1.2. Multiple Linear Regression

Regression models that are treated with one dependent variable and more than one independent variable are called multiple linear regression analysis. In this model, the independent variables also give information about the dependent variable. The analysis is both similar and different from linear regression analysis in terms of interpretation and calculation. In summary; the relationship between the change in a group of independent variables considered together with the dependent variable is expressed. The mathematical equation is mentioned in the Regression Methods part.

It can be rewritten the n observation as below, [21]

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \tag{4.5}$$

Then it turns into the matrix;

$$\mathrm{Y} = X\beta + \varepsilon \tag{4.6}$$

Calculating manuals causes a hard workload but Python, the OLS (Ordinary Least Squares) class of statistical models library is used to achieve this calculation. OLS is the method that minimizes the sum of the squares of the distances of all points from the regression line. In brief, it compares the difference between individual points in your data set and the predicted best fit line to measure the amount of error produced. Thus, it can best represent all points in the line's attributes matrix. In OLS, the significance level of each independent variable is examined. If the significance level(p-value) is more significant than 0.05, which is the value determined for the model, this independent variable is removed from the model. If there is more than one, this process is continued by applying the one with the largest p value. The model is complete if all p values are less than the limit value.

The summary of OLS for this study is that Dependent Variables are used in specification status and wine glass color. In the OLS Regression tables, Date and Time are created automatically. Then Number of Observations is added. Df Residuals is another name for Degrees of Freedom in the model. This is calculated in the form of [n(number of observations)-k(number of predicting variables)-1]. The Df Model means predicting variables. Covariance Type is listed as non-robust. Covariance is a measure of how two variables are linked positively or negatively, and a robust covariance is calculated in a way to minimize or eliminate variables.

R-squared is possibly the most critical measurement produced by this summary. It measures how much of the independent variable is explained by changes in the dependent variables. In percentage terms, it is exemplified in Table 4.2. 2019 OLS Regression Table for Spec Status, 0.001 would mean the model explains 1% of the change in *Spesifikasyon_dışı* the variable. Adjusted R-squared is important to observe multiple dependent variables' effect on the model. Linear regression has the quality that the model's R-squared value does not go down with additional variables, and it is only equal or higher. Therefore, the model is more accurate with multiple variables, even if poorly contributing. The adjusted R-squared penalizes the R-squared formula on the number of variables. Thus a lower adjusted score may inform some variables are not contributing to the model's R-squared properly.

The F-statistic in linear regression compares a linear model for variables against a model that replaces variables' effect to 0 to find out if the group of variables is statistically significant. To interpret this number correctly, using a chosen alpha value and a f_table is necessary. Prob (F-Statistic) uses this number to explain the accuracy of the null hypothesis or whether it is accurate that the variables' effect is 0. Log-likelihood is a numerical signifier of the likelihood that the model produced the given data. It is used to compare coefficient values for each variable in creating the model. AIC and BIC are both used to compare the effect of models in the process of linear regression, using a penalty system for measuring multiple variables. These numbers are used for the feature selection of variables.

Omnibus gives a clue about the normality of the distribution of our residuals using skew and kurtosis as measurements. 0-zero would indicate perfect normalcy. Prob(Omnibus) is a statistical test measuring the probability the residuals are normally distributed. 1-one would mean a perfectly normal distribution. Skew is a measurement of symmetry in the data, with 0-zero being perfect symmetry. Kurtosis measures the peakiness of the data or its concentration around 0-zero in a standard curve. Higher kurtosis implies fewer outliers.

Durbin-Watson is a measurement of homoscedasticity or an even distribution of errors throughout the data. Heteroscedasticity would imply an uneven distribution. For example, as the data point grows higher, the relative error grows higher. Ideal homoscedasticity lies between 1 and 2. Jarque-Bera(JB) and Prob(JB) are alternate methods of measuring the same value as Omnibus and Prob(Omnibus) using skewness and kurtosis. They are used values to confirm each other. The condition number is a measurement of the model's sensitivity compared to the size of changes in the data it is analyzing. A high condition number strongly implies multicollinearity. It is a term to describe two or more independent variables that are strongly related to each other and incorrectly affect the predicted variable by redundancy.[22]

After all definitions, by choosing the OLS Regression model, the p-values are examined and compared easily with the values thought to be closely related to quality and the wine bottle color. In creating a machine learning model with Regressor_OLS code, 20% of the dataset is used as test data, 10% as validation data, and 70% as training data, and it is trained by .fit() code. Then, R squared scores are calculated for both test and validation data for verification.

Having looked at the regression table, it can be said that cullet rate, batch humidity, glass surface, air/fuel ratio, and bottom temperature of furnace columns are closely related to the values of bottle quality status in the 2019 dataset. However, in consequence of observation quantity, all features in the 2020 dataset have a significant relationship when the dependent value is chosen as spec status.

The multiple regression test for the variables predicted to be significant on the dependent variable spec status is shown in the table below. It is shown that the model is successful because it is clear that the features have significance between specification status.

Table 4.2. 2019 OLS regression table for spec status

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Spesifikasyon_dışı   R-squared:                       0.001
Model:                             OLS   Adj. R-squared:                  0.001
Method:                  Least Squares   F-statistic:                     547.1
Date:                 Mon, 10 Jan 2022   Prob (F-statistic):               0.00
Time:                         22:26:23   Log-Likelihood:            -3.7517e+06
No. Observations:              4804968   AIC:                         7.503e+06
Df Residuals:                  4804959   BIC:                         7.504e+06
Df Model:                            8
Covariance Type:             nonrobust
==================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                         -1.3701      0.043    -32.212      0.000      -1.453      -1.287
CULLET (%)                 -5.547e-05   1.93e-05     -2.868      0.004   -9.34e-05   -1.76e-05
BATCH HUMIDITY (%)             0.0109      0.001     15.463      0.000       0.010       0.012
GLASS SURFACE (oC)             0.0009   2.73e-05     34.137      0.000       0.001       0.001
AIR/FUEL RATIO                 0.0068      0.000     28.681      0.000       0.006       0.007
FURNACE BOTTOM TEMP 811 T/C    0.0006   3.62e-05     16.472      0.000       0.001       0.001
FURNACE BOTTOM TEMP 812 T/C   -0.0006   3.48e-05    -16.338      0.000      -0.001      -0.001
FURNACE BOTTOM TEMP 815 T/C -7.383e-05  4.57e-06    -16.161      0.000   -8.28e-05   -6.49e-05
FURNACE BOTTOM TEMP 820 T/C  5.851e-05  1.69e-06     34.542      0.000    5.52e-05    6.18e-05
==============================================================================
Omnibus:                   2510338.996   Durbin-Watson:                   0.076
Prob(Omnibus):                   0.000   Jarque-Bera (JB):         12401043.218
Skew:                            2.663   Prob(JB):                         0.00
Kurtosis:                        8.795   Cond. No.                     3.67e+05
==============================================================================
```

Table 4.3. 2020 OLS regression table for spec status

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Spesifikasyon_dışı   R-squared:                       0.106
Model:                            OLS   Adj. R-squared:                  0.106
Method:                 Least Squares   F-statistic:                     3847.
Date:                Mon, 10 Jan 2022   Prob (F-statistic):               0.00
Time:                        21:39:12   Log-Likelihood:            -2.4800e+05
No. Observations:              259648   AIC:                         4.960e+05
Df Residuals:                  259639   BIC:                         4.961e+05
Df Model:                           8
Covariance Type:            nonrobust
=================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                           3.2110      0.131     24.427      0.000       2.953       3.469
CULLET (%)                      0.0051      0.000     24.987      0.000       0.005       0.005
BATCH HUMIDITY (%)             -0.1890      0.008    -23.251      0.000      -0.205      -0.173
GLASS SURFACE (oC)             -0.0004   6.23e-06    -62.346      0.000      -0.000      -0.000
AIR/FUEL RATIO                  0.3738      0.007     52.182      0.000       0.360       0.388
FURNACE BOTTOM TEMP 811 T/C    -0.0031      0.000    -30.169      0.000      -0.003      -0.003
FURNACE BOTTOM TEMP 812 T/C     0.0069   7.11e-05     96.790      0.000       0.007       0.007
FURNACE BOTTOM TEMP 815 T/C    -0.0077   4.79e-05   -160.951      0.000      -0.008      -0.008
FURNACE BOTTOM TEMP 820 T/C    -0.0005   8.55e-06    -62.776      0.000      -0.001      -0.001
==============================================================================
Omnibus:                    40342.241   Durbin-Watson:                   0.326
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            61783.612
Skew:                           1.151   Prob(JB):                         0.00
Kurtosis:                       3.640   Cond. No.                     2.87e+05
==============================================================================
```

Table 4.4. R-squared

	2019	2020
R-squared score test data	0.001	0.107
R-squared score validation data	0.0009	0.111

In the models, R^2=0.002 in the 2019 dataset and R^2=0.106 in the 2020 dataset are deficient levels of explanatory values. Hence, there must be other variables in determining the glass specification value as is seen in Table 4.4. R-Squared.

The multiple regression test for the significant glass components on the dependent variable specification status is shown in the table below. All glass components are effective when evaluating the wine bottle quality, as seen in the regression tables.

Table 4.5. 2019 OLS regression table for spec status

OLS Regression Results

Dep. Variable:	Spesifikasyon_dışı	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	895.9
Date:	Wed, 05 Jan 2022	Prob (F-statistic):	0.00
Time:	08:54:44	Log-Likelihood:	-3.7490e+06
No. Observations:	4804968	AIC:	7.498e+06
Df Residuals:	4804956	BIC:	7.498e+06
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-61.5944	3.101	-19.863	0.000	-67.672	-55.517
TiO2	2.4791	0.088	28.158	0.000	2.307	2.652
SiO2	0.6187	0.031	19.997	0.000	0.558	0.679
Al2O3	0.4262	0.032	13.290	0.000	0.363	0.489
Fe2O3	0.9311	0.044	21.243	0.000	0.845	1.017
CaO	0.6141	0.031	19.646	0.000	0.553	0.675
MgO	0.7614	0.031	24.352	0.000	0.700	0.823
Na2O	0.5964	0.031	19.133	0.000	0.535	0.657
K2O	0.8853	0.038	23.207	0.000	0.811	0.960
SO3	0.8391	0.032	26.198	0.000	0.776	0.902
Cr2O3	0.1748	0.011	16.002	0.000	0.153	0.196
Pb (ppm)	-0.0015	4.04e-05	-37.006	0.000	-0.002	-0.001

Omnibus:	2510955.261	Durbin-Watson:	0.075
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12427384.732
Skew:	2.662	Prob(JB):	0.00
Kurtosis:	8.807	Cond. No.	9.50e+05

Table 4.6. 2020 OLS regression table for spec status

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Spesifikasyon_dışı   R-squared:                       0.118
Model:                            OLS   Adj. R-squared:                  0.118
Method:                 Least Squares   F-statistic:                     2675.
Date:                Tue, 04 Jan 2022   Prob (F-statistic):               0.00
Time:                        21:51:59   Log-Likelihood:            -2.4622e+05
No. Observations:              259648   AIC:                         4.925e+05
Df Residuals:                  259634   BIC:                         4.926e+05
Df Model:                          13
Covariance Type:            nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          26.3019      0.333     78.929      0.000      25.649      26.955
TiO2           -1.5978      0.091    -17.498      0.000      -1.777      -1.419
SiO2            0.0094      0.001     15.023      0.000       0.008       0.011
Al2O3           0.0214      0.001     35.003      0.000       0.020       0.023
Fe2O3          -0.7501      0.014    -52.637      0.000      -0.778      -0.722
CaO            -0.9379      0.024    -39.590      0.000      -0.984      -0.891
MgO            -5.3928      0.040   -134.059      0.000      -5.472      -5.314
Na2O           -0.0487      0.011     -4.278      0.000      -0.071      -0.026
K2O            -0.1184      0.023     -5.072      0.000      -0.164      -0.073
SO3            -1.0886      0.031    -34.847      0.000      -1.150      -1.027
Cr2O3           2.1839      0.041     52.643      0.000       2.103       2.265
Pb (ppm)       -0.0100      0.000    -68.662      0.000      -0.010      -0.010
Fe+2           -0.0060      0.000    -20.234      0.000      -0.007      -0.005
Fe+2 / Fe2O3   -0.0106      0.000    -41.004      0.000      -0.011      -0.010
==============================================================================
Omnibus:                    35243.399   Durbin-Watson:                   0.289
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            51343.651
Skew:                           1.064   Prob(JB):                         0.00
Kurtosis:                       3.468   Cond. No.                     2.77e+05
==============================================================================
```

Table 4.7. R-squared

	2019	2020
R-squared score test data	0.0019	0.117
R-squared score validation data	0.0022	0.118

In the models, $R^2=0.365$ in the 2019 dataset and $R^2=0.118$ in the 2020 dataset are very low levels of explanatory levels as is seen in Table 4.7. R-Squared. This indicates that the model is not sufficient to determine wine glass bottle specification value.

The raw materials in the glass compound are natural like other raw materials in nature and it is seen that they vary within the specification limits. Therefore, it is useful to analyze as it can affect the quality and color of the glass bottle. As stated in a research paper, one of the factors affecting the properties of glass are impurities such as chromium oxide (Cr2O3) and iron oxide (Fe2O3), and it is essential to look at their variability. Although the values are within specification limits, changes in the amount of impurities affect the glass color. [23]

In transparent glass, color occurs through the absorption of certain wavelengths of light and transmission of other wavelengths. In thicker sections of glass, more light is absorbed than in thinner sections, making the thicker sections appear darker and sometimes have a different hue.[24]

In this way, it is desired to examine glass components in the formation of glass color. When it is desired to form a hypothesis to evaluate this situation, an evaluation is made by labeling the glass colors in the furnace data and using the OLS Regression test method.

Table 4.8. 2019 OLS regression table for wine bottle colors

OLS Regression Results

Dep. Variable:	COLOR	R-squared:	0.365
Model:	OLS	Adj. R-squared:	0.365
Method:	Least Squares	F-statistic:	2.127e+05
Date:	Mon, 10 Jan 2022	Prob (F-statistic):	0.00
Time:	22:46:51	Log-Likelihood:	-4.4501e+06
No. Observations:	4804968	AIC:	8.900e+06
Df Residuals:	4804954	BIC:	8.900e+06
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-680.0972	3.612	-188.298	0.000	-687.176	-673.018
BATCH HUMIDITY (%)	0.7580	0.001	953.734	0.000	0.756	0.760
GLASS SURFACE (oC)	0.0171	2.15e-05	793.469	0.000	0.017	0.017
SiO2	6.5154	0.036	180.794	0.000	6.445	6.586
TiO2	-26.0272	0.105	-247.660	0.000	-26.233	-25.821
Al2O3	7.7642	0.037	207.417	0.000	7.691	7.838
Fe2O3	8.6253	0.051	168.928	0.000	8.525	8.725
CaO	6.7053	0.036	184.055	0.000	6.634	6.777
MgO	6.3401	0.036	174.325	0.000	6.269	6.411
Na2O	6.5674	0.036	180.967	0.000	6.496	6.639
K2O	5.2006	0.044	117.444	0.000	5.114	5.287
SO3	-2.1610	0.037	-57.764	0.000	-2.234	-2.088
Cr2O3	0.5371	0.013	42.289	0.000	0.512	0.562
Pb (ppm)	-0.0005	4.68e-05	-9.773	0.000	-0.001	-0.000

Omnibus:	60513.500	Durbin-Watson:	2.519
Prob(Omnibus):	0.000	Jarque-Bera (JB):	65784.964
Skew:	-0.248	Prob(JB):	0.00
Kurtosis:	3.287	Cond. No.	2.03e+07

Table 4.9. 2020 OLS regression table for wine bottle colors

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  COLOR   R-squared:                       0.977
Model:                            OLS   Adj. R-squared:                  0.977
Method:                 Least Squares   F-statistic:                 8.480e+05
Date:                Mon, 10 Jan 2022   Prob (F-statistic):               0.00
Time:                        22:43:55   Log-Likelihood:             1.9893e+05
No. Observations:              259648   AIC:                        -3.978e+05
Df Residuals:                  259634   BIC:                        -3.977e+05
Df Model:                          13
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  5.6529      0.063     90.026      0.000       5.530       5.776
BATCH HUMIDITY (%)     0.3888      0.002    251.887      0.000       0.386       0.392
GLASS SURFACE (oC)     0.0013   1.52e-05     83.606      0.000       0.001       0.001
SiO2                  -0.0069      0.000    -65.017      0.000      -0.007      -0.007
TiO2                  -1.9059      0.016   -116.269      0.000      -1.938      -1.874
Al2O3                  0.0051      0.000     47.432      0.000       0.005       0.005
Fe2O3                  0.0147      0.002      6.078      0.000       0.010       0.019
CaO                   -0.5250      0.004   -120.747      0.000      -0.533      -0.516
MgO                   -0.2400      0.007    -34.845      0.000      -0.253      -0.226
Na2O                  -0.0995      0.002    -48.897      0.000      -0.103      -0.095
K2O                   -0.3486      0.004    -84.459      0.000      -0.357      -0.341
SO3                   -0.0877      0.005    -17.435      0.000      -0.098      -0.078
Cr2O3                  4.4041      0.007    665.900      0.000       4.391       4.417
Pb (ppm)               0.0034   2.61e-05    130.027      0.000       0.003       0.003
==============================================================================
Omnibus:                    42509.992   Durbin-Watson:                   2.370
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           222743.531
Skew:                          -0.696   Prob(JB):                         0.00
Kurtosis:                       7.319   Cond. No.                     4.03e+05
==============================================================================
```

Table 4.10. R-squared

	2019	2020
R-squared score test data	0.365	0.976
R-squared score validation data	0.367	0.977

In the models, R^2=0.365 in the 2019 dataset is a medium level of explanatory values and R^2=0.979 in the 2020 dataset is a very high level of explanatory values as is seen in Table 4.10. R-Squared. Thus, this indicates that these variables are mostly sufficient to determine the glass color except for the cullet rate.

Another research topic is the relationship between glass color and cullet. To find the answer to the question of whether the rate of cullet has an effect on the formation of glass color, a comparison is made with the Anova One-Way statistical method as well as OLS Regression.

As known, when ANOVA is applied, the response is expected that it has a normal distribution and the variance is constant and it would be independent and randomly distributed.[25] One-way ANOVA test is a tool used to test whether there is a statistically significant difference between the means of independent groups. One-way analysis of variance can be used whenever it is desired to understand whether a variable differs for different groups as in this example. Wine bottle colors are separated into groups and used in the stats library for analysis.

Table 4.11. 2019 OLS regression table for wine bottle colors in cullet rate

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  COLOR   R-squared:                       0.351
Model:                            OLS   Adj. R-squared:                  0.351
Method:                 Least Squares   F-statistic:                 2.604e+06
Date:                Mon, 10 Jan 2022   Prob (F-statistic):               0.00
Time:                        22:51:43   Log-Likelihood:            -4.5019e+06
No. Observations:             4804968   AIC:                         9.004e+06
Df Residuals:                 4804966   BIC:                         9.004e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0640      0.001    -91.612      0.000      -0.065      -0.063
CULLET (%)     0.0215   1.33e-05   1613.538      0.000       0.022       0.022
==============================================================================
Omnibus:                  2896078.185   Durbin-Watson:                   2.550
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           573771.571
Skew:                           0.625   Prob(JB):                         0.00
Kurtosis:                       1.859   Cond. No.                         130.
==============================================================================
```

Table 4.12. Analysis of variance(anova) for 2019

```
stats.f_oneway(beyaz.CULLET,ir_green.CULLET,uv_green.CULLET)

F_onewayResult(statistic=3260.630635920627, pvalue=0.0)
```

Table 4.13. 2020 OLS regression table for wine bottle colors for cullet rate

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  COLOR   R-squared:                       0.278
Model:                            OLS   Adj. R-squared:                  0.278
Method:                 Least Squares   F-statistic:                 9.983e+04
Date:                Mon, 10 Jan 2022   Prob (F-statistic):               0.00
Time:                        22:53:24   Log-Likelihood:            -2.4851e+05
No. Observations:              259648   AIC:                         4.970e+05
Df Residuals:                  259646   BIC:                         4.970e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.1912      0.003    429.843      0.000       1.186       1.197
CULLET (%)     0.0170   5.37e-05    315.961      0.000       0.017       0.017
==============================================================================
Omnibus:                    10643.190   Durbin-Watson:                   2.742
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12005.084
Skew:                           0.525   Prob(JB):                         0.00
Kurtosis:                       3.081   Cond. No.                         116.
==============================================================================
```

As can be seen in both results, the p value is 0.00 and this means that there is statistically significant a difference between the cullet and the glass color groups.

Table 4.14. Analysis of variance(anova) for 2020

```
stats.f_oneway(amber.CULLET,flint.CULLET,ir_green.CULLET,uv_green.CULLET)
F_onewayResult(statistic=1720.1900186181495, pvalue=0.0)
```

5. DIFFICULTIES OF GLASS DATA

In this work, the glass data presented is very comprehensive and the wine bottle data is evaluated to run the study on a specific bottle but in the data obtained, it has been observed that no data record is kept for the products that are deemed to be of unquality wine bottles in the automation system of the glass production facilities. For this reason, an analysis is carried out only on the collected data, and we picked a visualization-oriented approach for analyzing the data. Since the data does not provide us with a fine-grained relation between the individual glass products and the corresponding production data, we have to come up with a total approach. If the data was prepared more systematically, evaluating the glass on the product and drawing meaningful results with machine learning would lead to clearer results.

For a glass manufacturer to obtain accurate results, it is necessary to record the glass data in a cloud system as integrity, from the preparation of the blend to the entry of the quality tests, and to repeat the work on this data hourly, daily and monthly. In this way, the factor causing the formation of unquality products can be easily seen and precautions can be taken. In many ways, the disorganization of data can result in negative results for glass manufacturers and reduce the reliability of data work.

From another point of view, when it comes to quality, electricity, and energy consumption, certain features such as glass components, the cullet rate, and, the number of beads come to mind, but the structure of the furnace in which the glass is produced and the maintenance frequency of the furnace is also very important. If good results cannot be obtained despite well-prepared glass components during production, it is necessary to consider other reasons. While evaluating and examining the dataset, an analysis is carried out assuming that everything went well. Rather than a study to reduce the high values in energy and electricity consumption, visuals related to the cause are added and tried to be interpreted.

6. CONCLUSION

In this work, certain inferences are obtained about glass production by using the data visualization method with the parameters considered valuable in glass science. We examined the correlation between the production parameters. In addition, using the regression analysis method, we examined whether the characteristics are effective in the color and quality of the wine bottle. Some of the important findings can be explained as follows:

- Dimensional measuring testers are used significantly more than other quality control machines. Hot and Weight Control Types are more frequently used. It can be seen in Figure 2.10. Control Machines versus Wine Bottle Types on Spec. The status that some wine bottle products pass some quality tests more and fail others. It can be seen that clearly, the probability of passing the dimensional measurements and weight tests in wine bottles as good quality is more difficult than others. "*AGR_Kafa_Kaplama_1"* and *"2", "AGR_Dik_Yük_2", "Yük",* and *"ElectroPhysik*" tests are running without problem.
- The stone-button average value is 3 as seen in Figure 2.14. Stone_button mean on spec. status. It is part of the production process. The type of glass is an important factor in the condition and appearance of the stones. Its value can be changed by the glass type. As seen in the graph's comment, each value has both quality and not quality wine bottle products.
- The cullet rate has to be lower than 85% for quality bottles. Because the foreign materials from the cullet will be discarded by the Mcal machine in the automation system. It means that as the glass cullet increases, the number of foreign matter will also increase dependently.
- Bubbles play an important role in the factors affecting glass quality. If the bubbles are above a certain value, the products are sent back to the furnace. It is seen using visualization that the maximum limit of wine bottles in the datasets is 64. Considering the main reason for this, the decrease in the furnace bottom temperatures causes the bubbles to increase. The way to minimize this is to keep furnace temperatures under control according to expert information.
- However, the cullet rate in each glass color has different rates when the spec is considered. For example, the IR-Green bottle color needs to have a higher cullet rate

than others. Hereby, it can be said that each wine bottle color has a different cullet rate.

- Batch humidity has to be around 3.5%, and the air/fuel ratio has to be 9.2 to decrease energy and electric consumption. Furnace pull values have to be around 360.000 kg/day to reduce energy and electric consumption. However, The cullet rate has to be between 70% and 85% to decrease natural gas consumption, and the cullet rate has to be lower than 85% to reduce energy and electric consumption. For example, while *IR_GREEN* has the lowest energy consumption in wine bottle colors, *Flint* bottle color has the highest electric consumption value than others.
- In the energy dataset showing natural gas, mainly furnace pull values vary between 400 tons and 500 tons. When the natural gas consumption and furnace pull values are graphed, the density is seen as 400 tons. As the pull value and the cullet rate increase, the amount of natural gas consumption decreases. When the Air/Fuel Ratio values in natural gas consumption are examined, it is seen that there is a minimum natural gas consumption of 9.8 and maximum consumption of 10.2. Furnace humidity ranges from about 2.0% to 3.5%. When the humidity is at 3.5%, natural gas consumption decreases. When the days are grouped and analyzed throughout the year, the most natural gas consumption is the 18th day of the month.
- "SiO2" and "*Fe2O3"* components are positively closely to related and highly correlated "*Al2O3"* and "*Na2O"*, and negatively associated with "*K2O"*. It can be said that if one increases, other features also increase. *"SiO2"* and "*Fe2O3"* have a low correlation with other glass components, especially "*MgO*" and "*TiO2"*. "*Cr2O3*" has a relation with "*SO3"* positively. So, the features are dependent on each other to determine the values of the glass component's rate.
- Cullet rate is negatively associated with *natural gas consumption* and *air-fuel ratio.* Still, it is positively correlated with "Fe+2" and "Electric kcal/kg".
- "SiO2", "Al2O3*", "Fe2O3", "TiO2", "K2O", "SO3",* and *"Cr2O3"* are correlated positively with each other and negatively correlated with "*air-fuel ratio"* and *"CaO", "MgO", "Na2O"* is shown that scattered negatively but positively correlated with *air-fuel ratio*. Furthermore, *Furnace Humidity* features are closely related to energy consumption. The outcome related to humidity is really important to find out because of energy consumption.

- All glass components are closely related to the wine bottle's quality and color when viewed from multiple regression analyses. At the same time, it is understood by looking at the R-Square values that other reasons affect the quality and color of the wine bottles. For example, furnace cleaning, maintenance, and repair must be routine. Furnaces must be suitable for the production amount and must be controlled at regular intervals to handle quality products.
- According to the linear regression model, there is no linear relationship between Batch Humidity and Bubbler per minute. They have specific values for both batch humidity and bubbler per minute. Therefore, the hypothesis is not correct due to the result.
- *Cullet Rate, Batch Humidity, Glass Surface, Air-Fuel Ratio,* and furnace b*ottom temperatures* are evaluated with regression analysis for spec status but it is seen that they are not enough to understand the spec status. Each one is related to the quality of the wine bottle while glass component rates play an important role with spec status. Nonetheless, it can be concluded that there are other factors affecting the wine bottle quality. As it is mentioned, furnace maintenance and repair frequency may be important for quality.

Research on glass science is very important for furnace blend measurements, the proportions of glass components, cullet and stone ratios, furnace humidity, as well as furnace quality, and repair in determining the glass quality. By all means, it is very difficult to make definite judgments on glass quality but make an analysis by using correct data that can be made for each feature so the correct results and causes can be determined. In this work, which is conducted on anonymous data, the insufficient quality of the data led me to do more detailed research on this subject and to see the missing topics. In order not to disturb the balance of nature while producing glass, it is possible to conduct research with the information obtained from the data to reduce environmental pollution. At the same time, ways to reduce energy, electricity, and natural gas consumption in glass production can be examined. As a continuation of this work, in addition, with a higher quality dataset, multiple regression can be applied by taking a spec value as a dependent. In this case, it would be clear whether there is a change or not. In addition, the subject of glass color is a very valuable research topic. It is made on the different datasets in this work, and the color factor is taken as a dependent and evaluated with other features. The valuable result is the cullet rate which plays an important role in the color of the glass. It can be observed with different datasets whether there are additional different factors by applying machine learning on the subject of glass color.

REFERENCES

1. Şişecam Cam Eserleri - Camın Gelişimi. [cited 2021 30 April]. Available from: http://www.sisecamcameserleri.com/tr/camin-tarihcesi/camin-gelisimi.

2. Chopinet MH. The history of glass. InSpringer handbook of glass 2019 (pp. 1-47). Springer, Cham.

3. Danilenko PV. Who invented the glasses and when. Handbook of a General Practitioner. 2020(7):57-62.

4. Kaczmarczyk A. The source of cobalt in ancient Egyptian pigments. InProceedings of the 24th international Archaeometry symposium 1986 (pp. 369-376).

5. Rehren T. Aspects of the production of cobalt-blue glass in Egypt. Archaeometry. 2001 Nov;43(4):483-9.

6. Schmitz A, Kamiński J, Scalet BM, Soria A. Energy consumption and CO2 emissions of the European glass industry. Energy policy. 2011 Jan 1;39(1):142-55.

7. Pellegrino JL. Energy and environmental profile of the US glass industry. Energetics, Inc., Columbia, MD (United States); 2002 Apr 1.

8. “(PDF) A multiobjective optimisation system for a glass furnace.” https://www.researchgate.net/publication/228700022_A_multiobjective_optimisation_system_for_a_glass_furnace (accessed May 27, 2021).

9. Stormont R. Electric melting and boosting for glass quality improvement. Glass Worldwide. 2009(25).

10. Kraus MA, Drass M. Artificial intelligence for structural glass engineering applications—overview, case studies and future potentials. Glass Structures & Engineering. 2020 Nov;5(3):247-85.

11. Bagui S, Nandi D, White RJ. Machine learning and deep learning for phishing email classification using one-hot encoding. Journal of Computer Science. 2021;17(7).

12. Von Hippel PT. Biases in SPSS 12.0 missing value analysis. The American

Statistician. 2004 May 1;58(2):160-4.

13. P. T. von Hippel, “Biases in SPSS 12.0 Missing Value Analysis,” *Am. Stat.*, vol. 58, no. 2, pp. 160–164, 2004, doi: 10.1198/0003130043204.

14. Liu CH, Tsai CF, Sue KL, Huang MW. The feature selection effect on missing value imputation of medical datasets. applied sciences. 2020 Mar 29;10(7):2344.

15. Çizimleri BM. Seramik ve Cam Teknolojisi. Megep, Ankara. 2008.

16. Li L, Lin HJ, Han J, Ruan J, Xie J, Zhao X. Three-dimensional glass furnace model of combustion space and glass tank with electric boosting. Materials Transactions. 2019 Jun 1;60(6):1034-43.

17. Bakanlığı ME, Eğitim M, Projesi ÖS. Seramik ve cam teknolojisi cam üretimi ve şekillendirme yöntemleri.

18. KARASU B, BEREKET O, BİRYAN E, SANOĞLU D. The latest developments in glass science and technology. El-Cezeri. 2017;4(2):209-33.

19. Bakdash JZ, Marusich LR. rmcorr: Repeated measures correlation. R package version 0.3. 0.

20. Jan SL, Shieh G. Sample size calculations for model validation in linear regression analysis. BMC medical research methodology. 2019 Dec;19(1):1-9.

21. van Ginkel JR. Significance tests and estimates for R 2 for multiple regression in multiply imputed datasets: A cautionary note on earlier findings, and alternative solutions. Multivariate Behavioral Research. 2019 Jul 4;54(4):514-29.

22. Hooker SA, Schmiege SJ, Trivedi RB, Amoyal NR, Bekelman DB. Mutuality and heart failure self-care in patients and their informal caregivers. European Journal of Cardiovascular Nursing. 2018 Feb 1;17(2):102-13.

23. Analizleri BM. Cam Analizleri. Megep, Ankara. 2012.

24. Çiftçi F, AĞATEKİN M. Cam ve renk. Sanat ve Tasarım Dergisi. 2019;9(1):164-78.

25. Wahid Z, Latiff AI, Ahmad K. Application of one-way ANOVA in completely randomized experiments. InJournal of Physics: Conference Series 2017 Dec 1 (Vol.

949, No. 1, p. 012017). IOP Publishing.

APPENDIX A

Below is a table about the datasets and the columns used. Missing values are few in datasets. In addition, the number of columns that are worth creating a graph in the table and that can contribute to the analysis are indicated.

Table A. 1. Dataset Information

Dataset	Column Quantity	Used Column	Missing Value / Quantity of columns
Furnace	97	20	28
Furnace	36	32	Oxygen upper left, Oxygen lower left, Oxygen upper right, Oxygen lower right
Glass Components	21	21	No missing value
Quality Laboratory	41	39	*Mastör*, *Yorum*
Bubble	13	13	No missing value
Energy	206	20	No missing value
Energy	87	50	37
Production	28	6	No missing value