CLUSTERING AND RECOMMENDATION SYSTEM ON TURKEY HOTEL DATASET

(TÜRKİYE OTEL VERİLERİ ÜZERİNDE KÜMELEME VE ÖNERİ SİSTEMİ)

by

ÖMER ARİFOĞULLARI, B.S.

Thesis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

in

SMART SYSTEMS ENGINEERING

in the

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

JUNE 2023

ACKNOWLEDGMENTS

I would like to express my deep gratitude to Assist. Prof. Dr. Günce Keziban Orman as my research supervisor, for her patient guidance, enthusiastic encouragement and useful critiques for this thesis.

In addition, I would like to thank my supportive mother, my beloved sister and my incredibly patient friends...

July 2023
Ömer ARİFOĞULLARI

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AMI	Adjusted Mutual Index
ARI	Adjusted Rand Index
B2B	Business to Business
C-H	Calinski-Harabasz
D-B	Davies-Bouldin
DBSCAN	Density-based Spatial Clustering of Applications with Noise
EFT	Explicit Feedback Technique
EM	Expectation Maximization
HAC	Hierarchical Agglomerative Clustering
HFT	Hybrid Feedback Technique
IDE	Integrated Development Environment
IFT	Implicit Feedback Technique
MI	Mutual Index
MSSQL	Microsoft SQL Server
NMF	Non-Negative Matrix Factorization
OPTICS	Ordering Points to Identify the Clustering Structure
PCA	Principal Component Analysis
RFM	Recency, Frequency, Monetary
RI	Rand Index
SPCA	Sparse Principal Component Analysis
UNWTO	The United Nations World Tourism Organization
VS	Visual Studio
WCSS	Within-Cluster Sum of Square

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

As in most sectors, the development of an intelligent recommendation system in tourism becomes an important issue. Tourism agencies are putting maximum effort into suggesting the best and most valuable hotels for their customers. With the help of B2B relations between agencies and hotels, tourism agencies hold large feature datasets about hotels. Summarizing or interpretation of huge amount of data, requires the implementation of data analysis methodologies. Also, the tourism data is unique in terms of geography and culture. Thus, every new dataset requires a dedicated analytical process. Furthermore, raw data is in the form of a sparse binary matrix of hotel features, it poses a technical challenge for any analytical process. This thesis presents a comparison of different clustering and dimension reduction methodologies for real-world hotel data of this nature. The dataset represents 61% of the hotels in Turkey. Hotel clustering is the first step to acquire the hotel recommendations. To generate matching recommendations for customers, multiple level system is designed. Another layer of this cascade structure is clustering of the users according to their previous hotels moreover clusters of hotels. Our approach has resulted with linkage of two type of collaborative filtering which called as hybrid recommendation system. Thereby the suggested system provides personalized hotel recommendations based on the hotel's amenities and visitor patterns.

The first challenge in this work is the nature of the raw data set. The hotel features that we work with are all binary variables. While there are plenty of metrics, algorithms, and techniques dedicated to discovering knowledge from numerical variables, the methods for processing binary ones are limited. Thus, one of our contributions is to propose an experimental methodology for discovering the best clusters for the hotels, which are explained with binary features. In this methodology, we transform the sparse binary data set into numerical ones by using different dimension reduction techniques. Then, well-known clustering algorithms are applied and evaluated by various success criteria metrics. The most succeeded clustering algorithm has been decided as OPTICS. Due to the nature of the algorithm, noise labeled hotels has been created by the results of the

algorithm. Cosine similarity between hotel features is calculated and the result has been used for elimination of the noise labeled hotels. Since this is the first step in the process, rather than focusing on the interpretation of preliminary clusters, we have focused on solving analytical problems such as determining the number of the best clusters, identifying the most distinguished clusters, and simply selecting the best algorithm for hotel clustering. During the recommendation engine design, we used collaborative filtering method which consist of item and user features. Thereby the engine can be considered as a hybrid system. According to designed similarity matrix between users, the target user (customer) receives a bunch of last visited hotels. Similarity between users has been decided based on their clusters of already visited hotels. Thereby every user has at least one similar user, and these users have at least one order in our analysis table. While creating recommendation arrays, the results checked with sales test table which consist only latest orders. If the test value for target customer matches with one of the members of recommendation array, the recommendation considered as success. Even though the binary labeled success criteria (success or fail) which is very loose method, the recommendation engine has not achieved the expected success ratio. At last, data preparation steps, used models and recommendation system results has been discussed and the reasons of the success rates are explained in this thesis.

Keywords : clustering, data dimension reduction, hotel clustering, dimension reduction, tourism recommender systems

ÖZET

Çoğu sektörde olduğu gibi, turizmde de akıllı bir öneri sisteminin geliştirilmesi önemli bir konu olmuştur. Turizm acenteleri, müşterilerine en iyi ve en uygun otelleri önermek için maksimum efor sarf etmektedirler. Turizm acenteleri, otellerle kurdukları kurumsal ilişkilerin de yardımı ile, otel özellikleri hakkında büyük veri kümeleri tutmaktadırlar. Büyük hacimli verinin özetlenmesi ya da açıklanması, veri analizi yöntemlerini gerektirmektedir. Ayrıca turizm verisi kültürel ve coğrafi olarak eşsizdir. Böylelikle, her bir veri seti özelleştirilmiş bir analitik süreç gerektirir. Ancak, seyrek karakterli ikilik matris şeklindeki otel özelliklerinin ham veri seti, analitik işlemler için teknik zorluklara yol açmaktadır. Bu tez, bu türden gerçek otel verileriyle farklı kümeleme metodolojilerinin kıyaslanmasını, boyut indirgeme tekniklerini sunmaktadır. Veri seti, Türkiye'deki otellerin %61'ini barındırmaktadır.

Otel önerisi elde etmek için, kümeleme ilk adımdır. Doğru bir öneri sistemi oluşturmak için çok katmanlı bir sistem tasarlanmıştır. Bu çok katmanlı yapının bir başka katmanı ise; müşterilerin, gittiği daha önceki otellere ve hatta otellerin tespit edilen kümelerine göre gruplanmasıdır. Yaklaşımımız, hibrit öneri sistemi olarak anılan, işbirlikçi filtrelemenin iki tipinin birleşimiyle sonuçlanmıştır. Böylelikle yaklaşımımız hem ziyaretçi şablonuna hem de otellerin özelliklerine dayanarak kişiselleştirilmiş bir otel öneri sistemiyle sonuçlanmıştır. İkilik değerlerle açıklanmış olan otel verilerinin en iyi otel kümeleme keşfini sunan deneysel bir yaklaşımın sunulması da çalışmamızdaki katkılardan birisi olmuştur.

Çalışmadaki teknik detaylara bakıldığında, bu çalışmadaki aşılması gereken ilk aşama ham veri setleridir. Çalıştığımız otel veri setinin tamamı ikilik değişkenlerden oluşmaktadır. Sayısal verilerin anlamsal keşfi için pek çok metrik, algoritma ve teknik varken, ikilik değerlerin işlenmesi için metodolojiler kısıtlıdır. Bu yöntemde farklı boyut indirgeme teknikleri kullanılarak, ikilik veri setini sayısal değerlere dönüştürülmüştür. Ardından iyi bilinen kümeleme algoritmaları uygulanmış ve çeşitli başarım kriterleri ile değerlendirilmiştir. En başarılı kümeleme algoritması olarak OPTICS'e karar verilmiştir. Algoritma, doğası gereği, gürültü etiketli oteller de yaratmıştır. Gürültü etiketli

otellerin elenmesi için otel özellikleri üzerinden hesaplanan kosinüs benzerliği kullanılmıştır. Bu, sürecin ilk adımı olduğundan, ön kümelerin yorumlanmasına odaklanmak yerine, en iyi küme sayılarını belirleme, en iyi ayrıştırılmış kümeleri belirleme ve basitçe otel kümelemesi için en iyi algoritmayı seçme gibi analitik problemleri çözmeye odaklanılmıştır. Öneri motorunun tasarımında ürün ve kullanıcı özelliklerini içeren işbirlikçi filtreleme metodu kullanılmıştır. Nihai olarak tasarlanan öneri motoru, hibrit bir öneri sistemi halini almıştır. Kullanıcılar (müşteriler) arasında tasarlanan benzerlik matrisine göre, hedef kullanıcı son ziyaret edilen otellerin oluşturduğu belirli sayıda öneriler almaktadır. Kullanıcılar arasındaki benzerlik ise onların ziyaret ettikleri otellere göre ölçülmüştür. Böylece her bir kullanıcı için en azından bir tane benzer kullanıcı tespit edilmiş ve bu tespit edilen kullanıcıların analiz tablomuzda da en az bir siparişi bulunmaktadır. Öneri dizileri yaratılırken; sonuçlar, yalnızca son satışı içeren satış test tablosundaki verilerle kontrol edilmiştir. Değerlendirme adımında ise: eğer hedef müşteri için ayrılan test değeri; önerilen dizinin üyelerinden birisi ise, öneri başarılı sayılmıştır. Tasarladığımız ikilik etiketli başarım kriteri (başarılı ya da başarısız) çok gevşek bir yöntem olmasına rağmen, öneri motorumuz beklenen başarım oranını sağlayamamıştır. Tezde son olarak, veri hazırlama, kullanılan modeller ve öneri sistemi sonuçları tartışılmış ve başarım oranlarının sebepleri açıklanmıştır.

Anahtar Kelimeler : kümeleme, veri boyutu indirgeme, otel kümeleme, boyut indirgeme, turizm öneri sistemleri

1 INTRODUCTION

After the long lockdown periods, the populations of some specific countries (e.g., India, Qatar and Saudi Arabia) increased their expenditures in the tourism sector compared to the pre-pandemic era. According to UNWTO (The United Nations World Tourism Organization), some countries, including Turkey, have already reached the same level of income compared between pre-pandemic and post-pandemic. The value of tourism has renewed itself and is still a big player in the countries' economies (*World Tourism Organization*, accessed : 28.11.2022). Customer tendencies for searching and buying have started to change with e-commerce and travel itenaries that published online. Therefore number of online reservations are passing number of physical store orders year by year. Via online systems, customer behaviours and order features easily collected by computer systems. Enormous volume of data that have been collected are used for recommendation systems by travel agencies to compete each other. Travel agencies' service quality should be quick, especially during peak seasons and when competition becomes a duel arena for companies. Time and options are limited, and customers easily access different sources to compare the best-matching hotels according to their needs.

Over time, different papers have been published for the tourism sector. Furthermore, it is possible to find tourism recommendation systems or hotel clustering research in other sectors that have a high volatile customer tendency and are highly competitive, like as retailing, streaming, and e-commerce platforms. Sánchez-Pérez, M. et. al. examine the effects of vertical and horizontal differentiation to explain hotel pricing decisions, considering the moderating roles of competition and location (Sánchez-Pérez Manuel, 2020).

There are two major analytical problems for recommendation systems and these are : Diversity and sparsity. Sparsity is mostly caused by rare interactions between users and items. Nevertheless there are many options to recommend to users. This sparse interaction table causes weak recommendations. On the other hand, diversity is hard to determine without users demographic values. The mentioned demographic features mostly are gender, socio-economic status, graduation level... But focusing only to di-

versity can cause repetition of recommendations which similar to previous transactions/sales/usage that are already obtained. One of the solutions to eliminate these problems is clustering of users or items and designing recommendation system for each clustered data. But mostly there is not enough user features on user tables whereas gathering of item features are generally easier. Therefore we have focused on hotel clustering of user and hotel data. Afterward this clustering stage is considered as preliminary step for recommendation system. However, clustering of hotel features has become out and out analytical problem for us. Thereby this thesis is dedicated to accurate clustering of dataset which binary explained hotel features. Some other related works for clustering can be found at literature e.g. Akyol, M. et. al. have been published hotel clustering and feature importance on their research (Akyol, 2021). At present, hotels provide many features to their customers, and the importance of the features can be a problem for the enterprise level recommendation algorithms. Also location-specific (dataset specific) research have been published in recent years. One of the related papers, published by Rodriguez-Victoria, O.E. et al. (Rodríguez-Victoria et al., 2017), investigated hotel clustering methodologies with Colombia hotels. Another paper has been published by Dağ, O. et. al. which focuses on Antalya, Turkey hotels in their paper (Dağ and KARAATLI, 2020). Natural, cultural, and political differences between tourism-economy territories cause different options and features regarding hotels and their customers.

Our work is done on the hotel feature data set provided by Setur Tech A.Ş. The registered touristic facility number is 4198 in Turkey (*Turistik Tesis Ve İşletmeler*, accessed : 10.01.2023) and our dataset covers more than 61% of those facilities. Therefore, our intact dataset may provide unexampled insight about hotel groups located in Turkey, and could be an example for the Mediterranean region. Thus, the first objective of this thesis has been assigned as clustering. Second and final objective of this thesis is to design a recommendation system for clients who have already travel agency's regular customers.

In this thesis, we propose an experimental setup for determining the best hotel clustering structure and recommendation model for use in the tourism industry. We have understood that previous research for hotel recommendation systems are not focused analytical explanation of hotel features and most of papers have been focused to propose a new recommendation system. On the contrary, like as e-commerce data, hotel and customer data do not tell its own tale. The information that is revealed by the de-

tailed analysis, can be more useful for hotel recommendation systems. Our work can be seen as an example of the data analysis process for finding the most proper clustering structure for the hotels.

The contributions of this thesis are as follow :

— A hotel clustering framework has been established to use on future research on smart tourism industry. The results have been evaluated by different dimension reduction methods and different clustering methods this framework and binary sparse hotel features dataset has been clustered with best resulted method.

— This framework has been applied to hotel dataset provided by Setur Tech A.Ş. that consist real values and more than half of the hotels in Turkey. The results of the experiment inspected and interpreted within a broad perspective.

— A simple recommendation system has been designed to evaluate the benefits of obtained hotel clusters.

In the next chapters, provide overview of recommender systems, explanation of methodology and experimental setup. After the theoretical part of the thesis, the outputs of the chosen or designed systems have been shared with the reader. After these steps, discussions has been placed in last section where the recommendation success rates criticised and also it's causes in the data preparation steps has been discussed.

2 OVERVIEW OF RECOMMENDER SYSTEMS

Each user has unique demands. But behind of these demands, there is a pattern in big perspective. Recommender systems have been designed to find these patterns to provide better results.

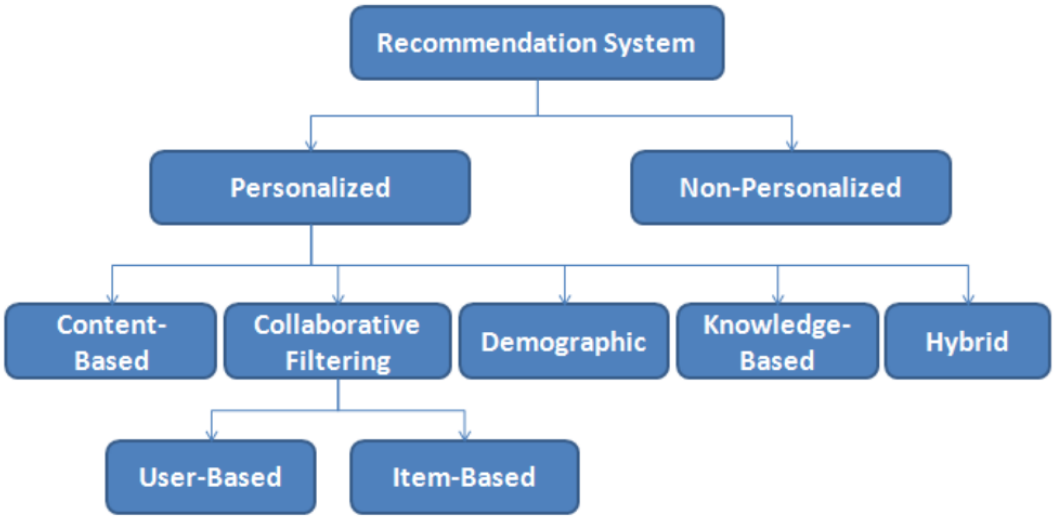


FIGURE 2.1 – Recommendation System

Image courtesy from (Das et al., 2017)

2.1 Personalized Recommendation System

This type of recommender systems have 3 different perspectives those are :

1. Interpersonal impact, which implies whom you would believe,
2. Interest circle derivation, which implies whose interest is similar to yours
3. User individual interests

2.1.1 Content Based Filtering :

These algorithms aim to suggest items according to user's past behavior.

Merits

— Other user's data is not required.

— No data sparsity as well as cold start.

Demerits

— Content analysis is essential to define the item features.

— The excellence of the product can't be estimated.

2.1.2 Collaborative Filtering

There are 2 types of method in this technique. User based collaborative filters group users according to their choices. Item based collaborative filters group items and try to find similar groups with its recommendations.

Merits

— The excellence of the item can be estimated through user ratings.

Demerits

— Cold start problem for different users and new products.

— Stability vs. plasticity issue.

2.1.3 RFM

One of the most popular, easy-to-use, and effective segmentation methods to enable marketers to analyze customer behavior is RFM analysis. RFM analysis is another user based collaborative filtering method.

2.1.3.1 Description

The RFM abbreviation stands for recency, frequency and monetary. RFM is a predictive model that can separate subject of research (i.e. customers, researchers, users) as good, average or inactive based on transactional data. The model can also predict the future probability and monetary value of subject, based on past transactions.

The concept of recency, frequency, monetary value (RFM) is thought to date from an article by Jan Roelf Bult and Tom Wansbeek, "Optimal Selection for Direct Mail," published in a 1995 issue of Marketing Science (Bult and Wansbeek, 1995).

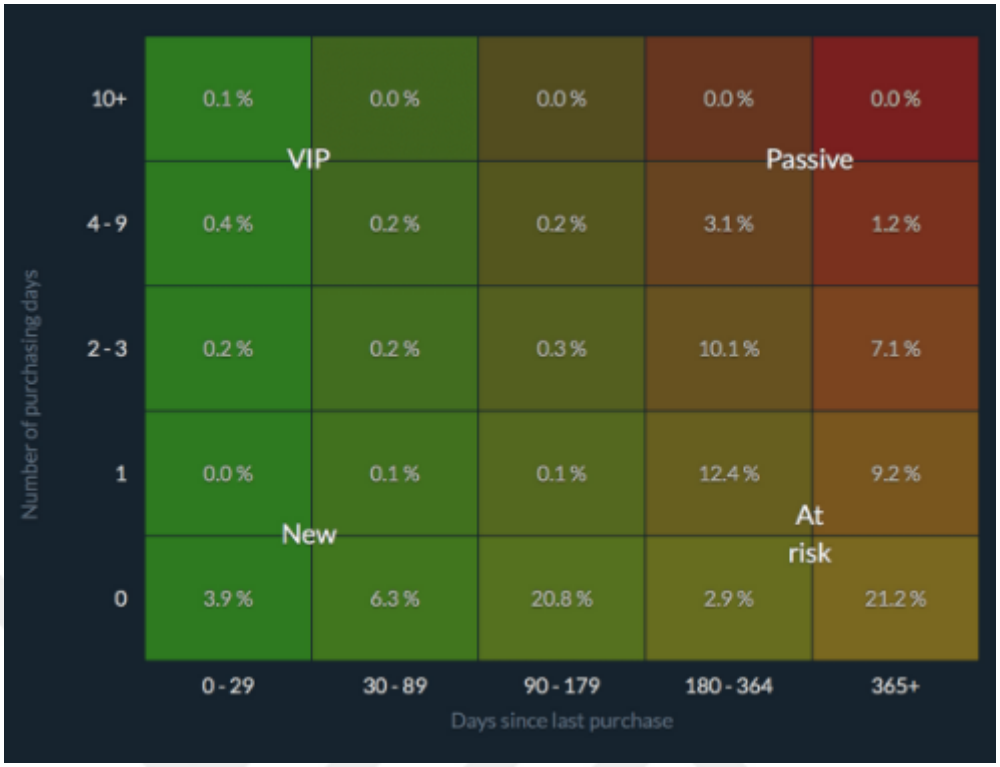


FIGURE 2.2 – RFM example of Categorization, Reduced 2 Dimension

Image courtesy from (Custobar, accessed : 10.10.2021)

2.1.3.2 Detailed Explanation

RFM technique relies on subject behaviour. According to characteristic of that behavioral attitude, subject can be categorized (Shown as figure 2.2). Thus, marketing/capacity planning will be modeled and that model will be used for predictive analysis.

For a marketing model, higher points in each plane, mean most valued customers. We can create those planes as frequency-monetary, frequency-recency, recency-monetary.

Cluster analysis uses mathematical models to discover groups of similar customers based on the smallest variations among customers within each group. In the context of customer segmentation, cluster analysis is the use of a mathematical model to discover groups of similar customers based on finding the smallest variations among customers within each group. These homogeneous groups are known as “customer archetypes” or “personas”.

The goal of cluster analysis in marketing is to accurately segment customers in order

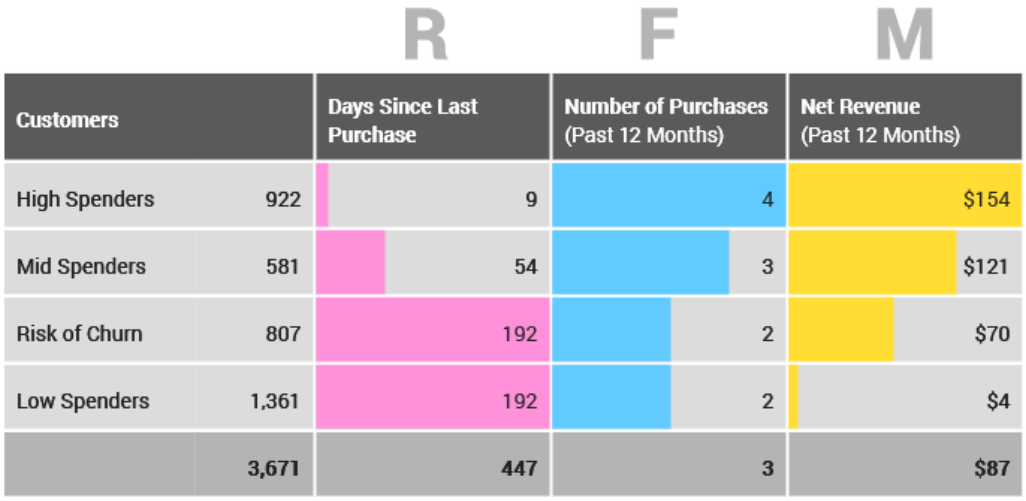


FIGURE 2.3 – RFM Example of Categorization

to achieve more effective customer marketing via personalization.

The figure 2.3 shows the results of a three-dimension cluster analysis performed on the customer base of an e-commerce site. This analysis resulted in the discovery of four customer personas i.e. high spenders, mid spenders, risk of churn, low spenders.

2.1.4 Demographic filtering

Demographic recommendation technique uses only demographic data of population.

Merits

— The technique is domain independent because Item feature is not needed.

Demerits

— Collection of demographic information give rise to privacy issue.

— Plasticity vs. stability issue.

2.1.5 Knowledge-Based recommender system

This technique is for low frequent user interaction systems such as home or car classified websites. In this type of system, the algorithm considers the knowledge about the item and its feature, user preference (asked explicitly), recommendation criteria before giving

the recommendation.(Bobadilla et al., 2013) The accuracy of the model is judged in view of how helpful the prescribed thing is to the client.

Types are :

— Constraint Based : Pre-described characterized set of recommendation principles.

— Case-based : Retrieval of comparable things on the premise of various sorts as likeness measures.

2.1.6 Hybrid Recommender System

Hybrid recommender systems are built by joining different techniques. This type of recommender systems design to diminish demerits and enhance merits of each individual systems. The engineering question is which type of systems to be considered and which order. (Arthur et al., 2019)

2.2 Recommender System Feedback Techniques

All recommender systems need data as a feedback from user. Those feedback techniques are can be (Das et al., 2017) :

— Implicit Feedback Technique : Implicit Feedback Technique (IFT) gather data from user without user's consent. IFT can be search history, mouse movements, web consumption history etc.

— Explicit Feedback Technique : Explicit Feedback Technique (EFT) involves the users for assigning either numeric or score rating (mostly discrete numbers) for evaluating the product.

— Hybrid Feedback Technique : Hybrid Feedback Technique (HFT) is the combination of both IFT and EFT. HFT is mostly expensive and computationally intensive.

3 METHODOLOGY

The following subjects will be covered in this chapter of the thesis :

— the types of datasets that we have,

— the data preparation steps to use these datasets,

— dimension reduction methods to cluster hotels that we used on this research,

— clustering algorithms for hotel clustering

Following these explanations, the 4 chapter contains the specifics of the experiments, including the setup of the recommendation engine and cluster success criteria. The arrangement of the methodology chapters follows mainly that of the experimental setup. Although various data cleaning techniques, such as the 7^{th} step on the 3.1.3.2 section, have been found while we are working on the recommendation engine. Thereby, we have revised and updated some steps during experiments.

3.1 Data Discovery and Preparation

We started our project by getting two separate datasets from Setur Tech A.Ş. The first dataset is a collection of hotel features, while the second is an anonymised sales dataset that includes information from the sales department for the two years beginning in 2019 and ending in 2022. We will discuss the hotel feature dataset, sales dataset, and data preparation process in the following sections.

3.1.1 Hotel Feature Dataset

The hotels that Setur customers have at least once visited are included in the hotel feature dataset. There are 2562 distinct hotels, each with a unique set of 27 features. 2562 distinct hotels are included in the dataset, but only 1281 hotelIDs can combine with the sales dataset. These features are all displayed in table 3.1. Each of these features are binary, indicating whether the relevant hotel has the relevant feature or

	HotelID	Çocuk-Bebek Dostu	Kayak Oteli	Havuz	Havuz-Yaz	Kumlu Plaj	Kumlu Deniz	Denize Sıfır	Çakıllı Plaj	Çakıllı Deniz
1	370	1	0	1	1	1	0	1	0	0
2	371	1	0	1	1	1	0	1	0	0
3	374	1	0	1	1	0	0	0	1	0
4	377	1	0	1	1	1	0	1	0	0
5	378	1	0	1	1	1	0	1	0	0
6	379	1	0	1	1	0	0	0	0	0
7	380	1	0	1	1	0	0	0	0	0
8	381	1	0	1	1	1	0	0	0	0
9	383	1	0	1	1	0	0	1	0	0
10	384	0	0	1	1	1	0	1	0	0

FIGURE 3.1 – First 10 Rows and First 9 Features from Hotel Features Table

not. Every hotel is uniquely identified by a single primary key called HotelID. In this study, the HotelID column has been deidentified to safeguard the interests of B2B clients. However, the data provider can use a reverse algorithm to connect anonymized HotelID values to genuine hotel IDs.

Small sample of hotel feature table can be seen at figure 3.1.

The features on figure 3.1 explain whether the corresponding hotel is child-friendly, ski hotel, pool, summer pool, sandy beach, sandy sea floor, is next to the shore, gravel beach, gravel sea floor, and so on. In the tourism sector, most of the hotels can provide only limited features at once. Therefore, most of the rows are filled with a 0 value, which 0 value means there is no option for customers at the hotel in question. Our raw dataset is binary, sparse and consists 2562 rows and 27 columns. Sparse and binary datasets present some difficulties for clustering and Mao Y. et. al. propose different approach for this type of datasets (Ye et al., 2018). The clustering process will be more detailed on section 3.2.

3.1.2 Sales Dataset

Before the data preparation steps, sales dataset consist 32 columns and 40599 rows. Some of the features and sample rows can be seen at figure 3.2. All of the columns are explained on table 3.2. Hotel ID, Customer ID, Agent ID, Branch ID have been kept anonymous for us. As a result, no information regarding agent sales or customer profiles and activities can be shared.

TABLE 3.1 – Original Features and Translations of Hotel Feature Dataset

Columns	**Explanations**
HotelID	Unique Hotel ID
Çocuk-Bebek Dostu	Baby Toddler Friendly
Kayak Oteli	Ski Hotel
Havuz	Swimming Pool
Havuz-Yaz	Summer Pool
Kumlu Plaj	Sandy Beach
Kumlu Deniz	Sandy Sea Floor
Denize Sıfır	On Shore
Çakıllı Plaj	Gravel Beach
Çakıllı Deniz	Gravel Sea Floor
Diğer Plaj	Other Type of Beach
İş Dostu Otel	Work Friendly Hotel
Casino Oteli	Casino Hotel
Balayı Oteli	Honeymoon Hotel
Aquapark	Aquapark
Golf Oteli	Golf Hotel
Spa-Termal Otel	Spa-Thermal
Sauna-Hamam	Sauna-Hamam
Su Sporları	Water Sports
Fitness	Fitness
Alkolsüz Her Şey Dahil	All Included Except Alcohol
Her Şey Dahil	All Included
Oda Kahvaltı	Room and Breakfast Included
Sadece Oda	Only Room
Tam Pansiyon	Full Board
Tam Pansiyon Plus	Full Board Plus
Ultra Her Şey Dahil	Ultra All Included
Yarım Pansiyon	Half Board

TABLE 3.2 – Columns and Their Explanations of Sales Dataset

Columns	**Explanations**
IslemTipi	Order Type
HotelID	Hotel ID
HotelType	Hotel Type
CityName	City Name
ParentRegionName	Region Name
BolgeAdi	Geographical Region Name
HotelCategoryName	Hotel Category Name
MusteriKodu	Customer ID
TklMasterID	Agent ID
MusterininYasi	Customer Age
MusterininCinsiyeti	Customer Gender
OtelMusteriSegmenti	Hotel Customer Segment
SubeAdi	Branch Name
SubeKodu	Branch ID
SubeTipi	Branch Type
SatisTarihi	Order Date and Time
GirisTarihi	Check-in Date and Time
CikisTarihi	Check-out Date and Time
Pax	Pax
OdaSayisi	Room Count
AdultCount	Adult Count
ChildCount	Child Count
GunSayisi	Number of Booked Days
SatisTarihiID	Order Date
GirisTarihiID	Check-in Date
CikisTarihiID	Check-out Date
KisiGece	Number of Person x Days
OtelAdultCount	Booked Adult Count
OtelChildCount	Booked Child Count
CurrencyCode	Currency Code
SatisTutari	Payment Amount as TRY
DövizliSatisTutari	Payment Amount as Order Currency

	SubeAdi	SubeTipi	SatisTarihi	GirisTarihiID	CikisTarihiID	Pax	OdaSayisi	AdultCount	ChildCount	GunSayisi	KisiGece
1	Web Şube	Online	2020-07-26 20:55:00.000	20200801	20200803	3	1	2	1	2	4
2	Call Center	SUBE	2019-08-23 12:54:00.000	20190923	20190926	2	1	2	0	3	6
3	Setur Merkez	SUBE	2019-06-28 21:46:00.000	20190815	20190817	2	1	2	0	2	4
4	Web Şube	Online	2020-01-07 09:01:00.000	20200120	20200125	3	1	2	1	5	12,5
5	Mobile	Online	2020-09-11 15:34:00.000	20200912	20200913	3	1	2	1	1	2
6	Call Center	SUBE	2020-08-28 13:53:00.000	20200829	20200830	3	1	2	1	1	2,5
7	Web Şube	Online	2019-08-24 15:50:00.000	20190901	20190906	2	1	2	0	5	10
8	Mobile	Online	2019-06-22 14:56:00.000	20190719	20190722	2	1	2	0	3	6
9	Setur Nişant...	SUBE	2019-09-04 13:35:00.000	20190916	20190918	3	1	3	0	2	6
10	Setur Bağd...	SUBE	2020-07-24 19:04:00.000	20200731	20200802	2	1	2	0	2	4

FIGURE 3.2 – First 10 Rows and Selected 11 Features from Sales Table

3.1.3 Data Preparation

Data preparation is one of the major steps in data analysis/mining works. Thereby, we have deeply analyzed the datasets and prepared for our MSSQL tables to use for clustering and recommendation. On the next sections, hotel feature dataset and sales dataset preparations has been explained.

3.1.3.1 Hotel Feature Dataset Preparation

Unmodified raw dataset imported into an MSSQL database. The HotelID column contains no null values, and each binary feature column contains at least one "1" value in every table row. Therefore, no feature columns in the feature table need to be deleted. The HotelID column has also had its uniqueness verified. Recurring rows from the table have been removed at this stage. Regarding these, a hotel feature table with 2562 rows and 27 columns has been implemented.

3.1.3.2 Sales Dataset Preparation

The preparation of sales data involves several phases. Unmodified raw dataset imported into an MSSQL database. Next the import process, the following list of data preparation tasks is explained. Figures 3.3 and 3.4 explain and illustrate these stages in the proper order. Figure 3.3 displays the first through eighth entries in the list, and Figure 3.4 displays the final item in the list. Applied steps on these figures are :

1. String stored datetime values converted to date time column type.

2. String stored "null" values converted null data type.
3. 3 rows have been deleted that have null payment amounts.
4. 1400 rows have been removed that could be inserted with wrong business flows or irregular orders sold by same agents.
5. Suspicious order or staying activity investigated on sales table. 1545 rows have been removed. These records show same agent sold different hotels for same days with different number of person/room types etc.
6. Some of the SatisTarihi column values were null in the table. These null values have been filled by SatisTarihiID values and "00:00:00.000" values. Thereby, 10645 rows have been updated.
7. We came to the conclusion that some of the sale records we have on the sales table don't have entries on the hotel features table when we compared the HotelID column of the sales table to the HotelID of the hotel features table. In relation to it, we eliminated 9212 rows from the sales table.
8. Our latest version of sales table have 28439 records. We have grouped and counted of each customers orders by CustomerID values. We have needed at least 2 orders by each customer to understand success criteria of our recommendation system. Therefore, we have removed single order records that have taken by customers. Within this step, sales table have only 6620 records.
9. Some of customers have more than 2 orders, thereby we have split latest order of each customer to sales test table as test data of our recommendation system. The previous sale records have gone to analyze table which is equal to 3997 rows. As a result of this step ; "Sale Test" table consists 2623 records and "Sale Analyze" table have 3997 rows.

According to the examines we performed to prepare the data, we lost about 84% of the sales records we imported from the raw sales dataset. Losing thus much data and how it affected the outcomes will be covered in the " 6 Conclusion" chapter on the following pages of this thesis. Furthermore, during the data preparation procedures, none of the columns were deleted or rearranged.

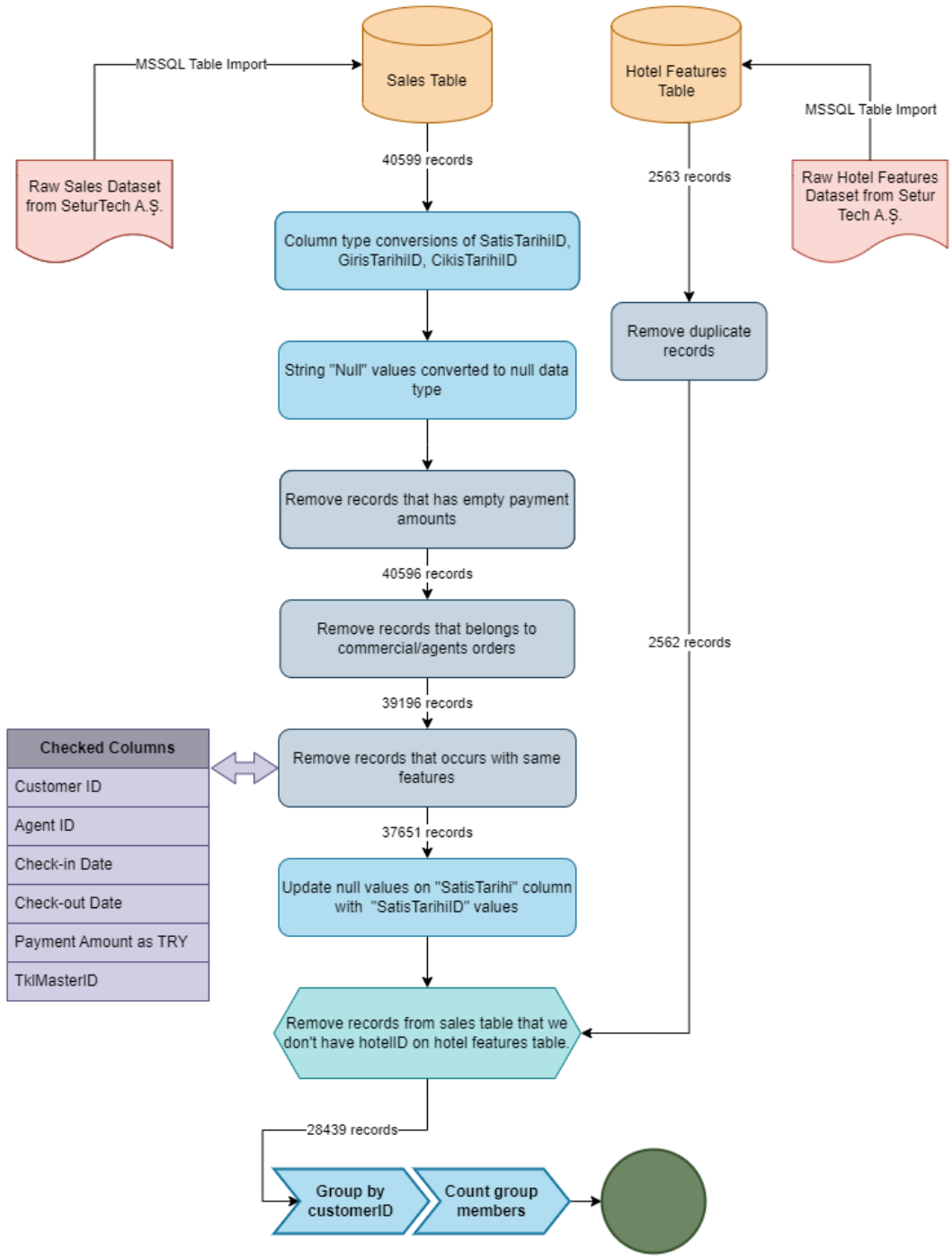


FIGURE 3.3 – Sales and Hotel Features Tables Data Preparation Step-1

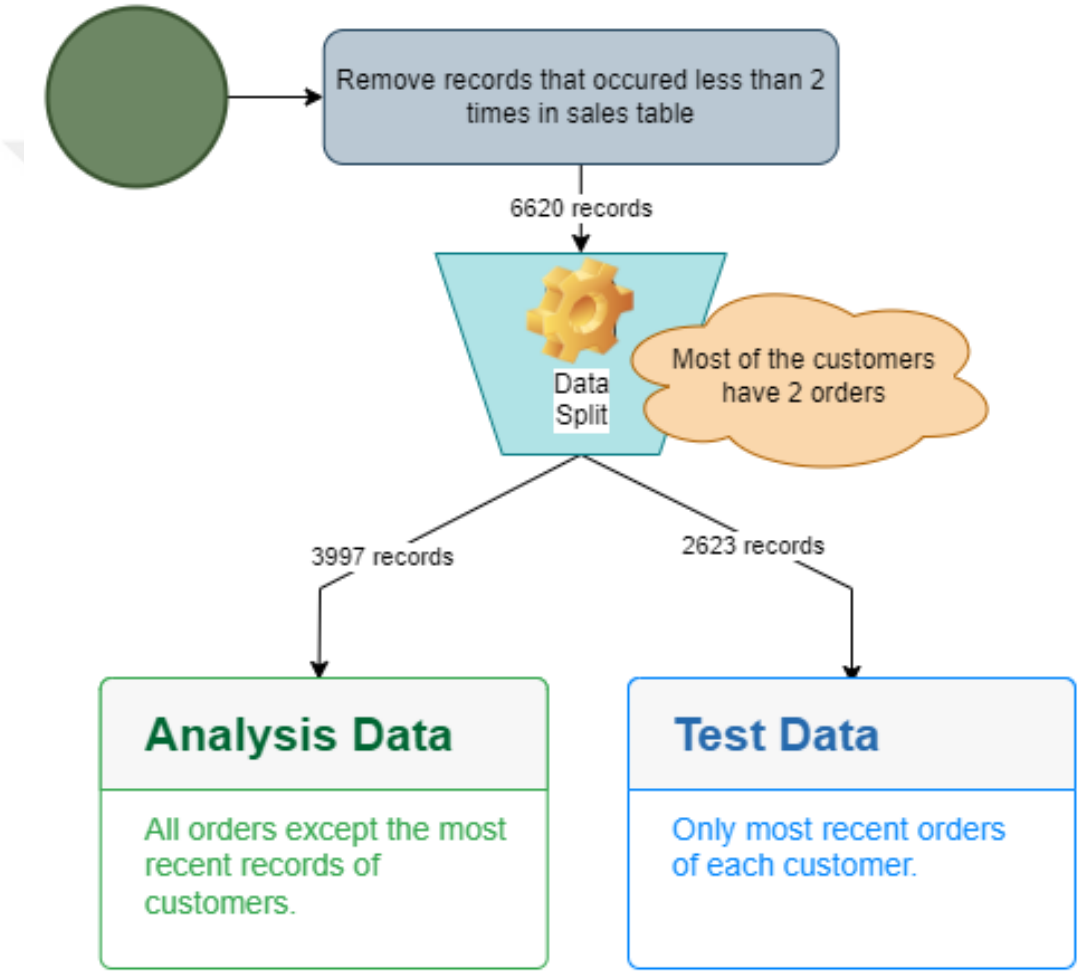


FIGURE 3.4 – Sales and Hotel Features Tables Data Preparation Step-2

3.2 Dimension Reduction Methods

As mentioned on previous chapters, the hotel feature dataset has 2562 rows and 27 different attributes, which have been called dimensions in data analysis. The first step is to plot all of the features together to see if there is any correlation between them. As a result of this step, there is no straightforward correlation between features. This means, business analysts and travel agents can not understand or estimate the hotel groups with simple deterministic methods. Even if there is not a straightforward correlation between features, we can clearly understand that if there is a sandy or gravel seabed, there should also be a sandy or gravel coastline. These kinds of low-level correlations between features, determine the informational value of columns. To simplify analysis and increase the efficiency of computing, different dimension reduction methods have been proposed in different studies over time.

In this study, we tackle the clustering of a data set with binary values. Since most of the clustering algorithms are dedicated to segmenting the data sets having numerical features, we first transform this data set into a numerical one. Indeed, there are several different methods for addressing this issue. In this work, we concentrate on two specific feature selection methods, Sparse Principal Component Analysis (SPCA) (Johnstone and Lu, 2009) and Non-Negative Matrix Factorization (NMF) (Dhillon and Sra, 2005) since they are not only assisting to transform a binary dataset into a numerical one but also reducing the dimension if it is necessary.

3.2.1 Sparse Principal Component Analysis

Principal component analysis (PCA) is described by Jolliffe et. al. (Jolliffe and Cadima, 2016) as a simple reduction of the dimension of a dataset while preserving as much statistical variability as possible. This means finding new variables that are linear functions of those in the original dataset, that successively maximize variance, and that are uncorrelated with each other. In our data set, all features consist of binary data. But, as explained in section 3.1, our dataset has a sparse characteristic. For this reason,

we have chosen a more specialized version of PCA, which is *sparse* PCA (SPCA). In traditional PCA, the principal components are linear combinations of all the original features, whereas in SPCA, the principal components are linear combinations of a sub-

set of the original features. The goal of SPCA is to identify a small set of features that explain the most variance in the data, while ignoring the noise or irrelevant information. This can be particularly useful in high-dimensional datasets, where many of the original features may be redundant or uninformative.

3.2.2 Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF), like PCA, is a dimension reduction technique. In contrast to PCA, NMF models are easy to understand and interpret. However, NMF can not be applied to every dataset. It is required that the sample features be "non-negative", so greater than or equal to 0. Non-negative matrix factorization (NMF) is a linear algebra method used for matrix decomposition and data analysis. Given a non-negative matrix V, NMF seeks to factorize it into two non-negative matrices, W and H, such that $V \approx WH$. The columns of W represent a set of basis vectors, or patterns, that are used to linearly combine the columns of H to approximate the original matrix V. The columns of H represent the weights that are assigned to each basis vector for each column of V. In other words, the matrix W defines a set of features or patterns that are common to the data, and the matrix H defines how much of each feature is present in each data point.

3.3 Clustering Algorithms

In this chapter, we will examine four different clustering algorithms. The interior logic of algorithms is explained to the reader in the following paragraphs. Further experimental details are shared in chapter 4.

3.3.1 K-Means

K-means is an iterative partition algorithm. Every cluster is represented by its centroid. Around these centroid points, according to neighborhood function, each cluster forms globular, non-overlapping shapes. The designer of the system should specify the number of clusters (K). The fundamental steps of the algorithm are as follows : first, K different centroids are chosen randomly from the data set as the representatives of K clusters.

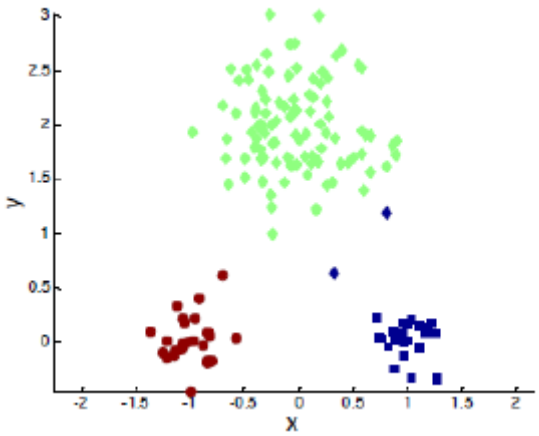


FIGURE 3.5 – K-means example on 2-D Space

Image courtesy from (Orman, 2021)

Second, each single data point is assigned to a cluster according to its closeness to cluster centroids. Third, the cluster centroids are reassigned. Fourth, the algorithm converges if the cluster centroids do not change between two consecutive iterations ; otherwise, the second and third steps are repeated. Different K-means versions can use different distance metrics (e.g., Euclidean, Chebychev) and different algorithms (e.g., Lloyd, Elkan) for expectation maximization (EM). Figure 3.5 is an example of k-means clustering on 2 dimension space.

3.3.2 Hierarchical Agglomerative Clustering

Hierarchical agglomerative clustering (HAC)is also an iterative algorithm. It arranges samples into a hierarchy of clusters. The number of clusters begins with the same number of rows as same as number of hotels, and each cluster contains one instance. Clusters merge iteratively until all nodes are labeled by a cluster. Figure 3.6 illustrates steps of HAC methodology. These merging processes use different distance (linkage) metrics. The most commonly used linkage metrics are single, complete, average, and ward measures showed at figure 3.7. Each algorithm has its pros and cons. In a nutshell, these linkages are :

— 'Single' uses the minimum of the distances between all observations in the two sets.

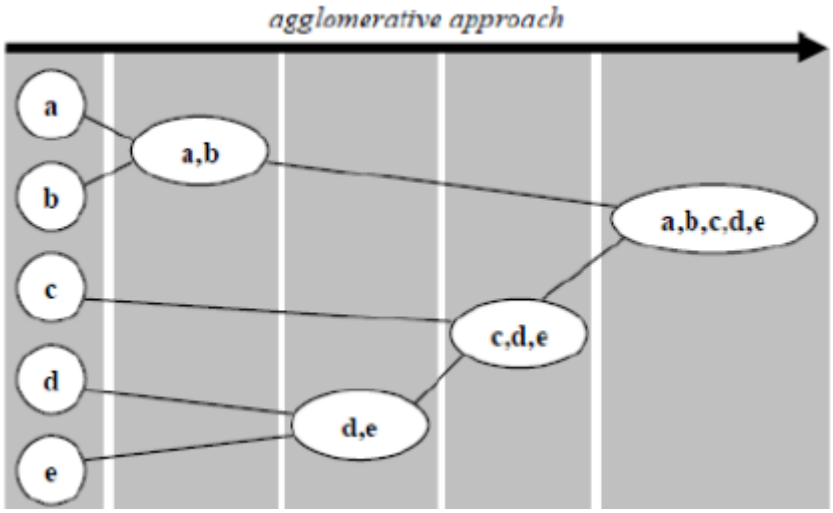


FIGURE 3.6 – Hierarchical Agglomerative Clustering Example

Image courtesy from (Orman, 2021)

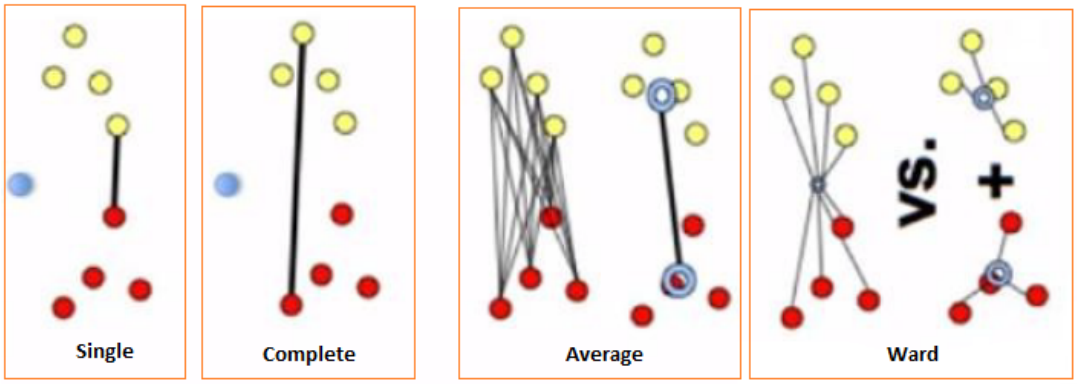


FIGURE 3.7 – Clustering Distance Measurement Example

Image courtesy from (Orman, 2021)

— 'Complete' or 'maximum' linkage uses the maximum distances between all observations in the two sets.

— 'Average' uses the average of the distances of each observation in the two sets.

— 'Ward' minimizes the variance of the clusters being merged.

3.3.3 Density-based Spatial Clustering of Applications with Noise

Density-based Spatial Clustering of Applications with Noise (DBSCAN) is a well-known clustering algorithm. But there's no free launch and relying on DBSCAN to find the right number of clusters completely on its own. The system designer should be aware of two parameters of DBSCAN. Firstly, the epsilon parameter (ε) defines the

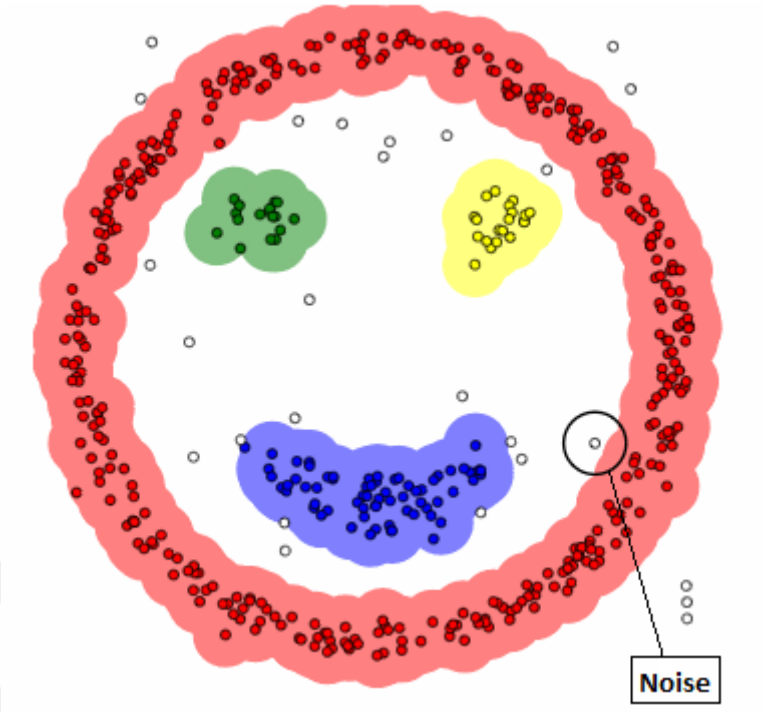


FIGURE 3.8 – Illustration of the DBSCAN Cluster Model

maximum distance between points within the same cluster. Second parameter is the minimum sample value for clusters, or "How many points can be called a cluster at a minimum ?". There can be some points with a wider distance than the epsilon value and lesser size than the minimum sample setting, which are called noise points.

3.3.4 Ordering Points to Identify the Clustering Structure

Ordering points to identify the clustering structure (OPTICS), which was founded in June 1999 by Ankerst, Breunig, Kriegel, and Sander, is another density-based algorithm (Ankerst et al., 1999). OPTICS is a more advanced version of the DBSCAN algorithm that finds the best epsilon value by ordering points based on their spatial values. The OPTICS algorithm permits a fluctuating density of the clusters, whereas the DBSCAN algorithm assumes a fixed density of the clusters.

OPTICS extends the idea of the DBSCAN algorithm by two additional terms, namely :

— Core Distance

— Reachability Distance

Core distance is the smallest radius that a given point must have in order to be designated as a core point. The supplied point's core distance is not defined if it is not the

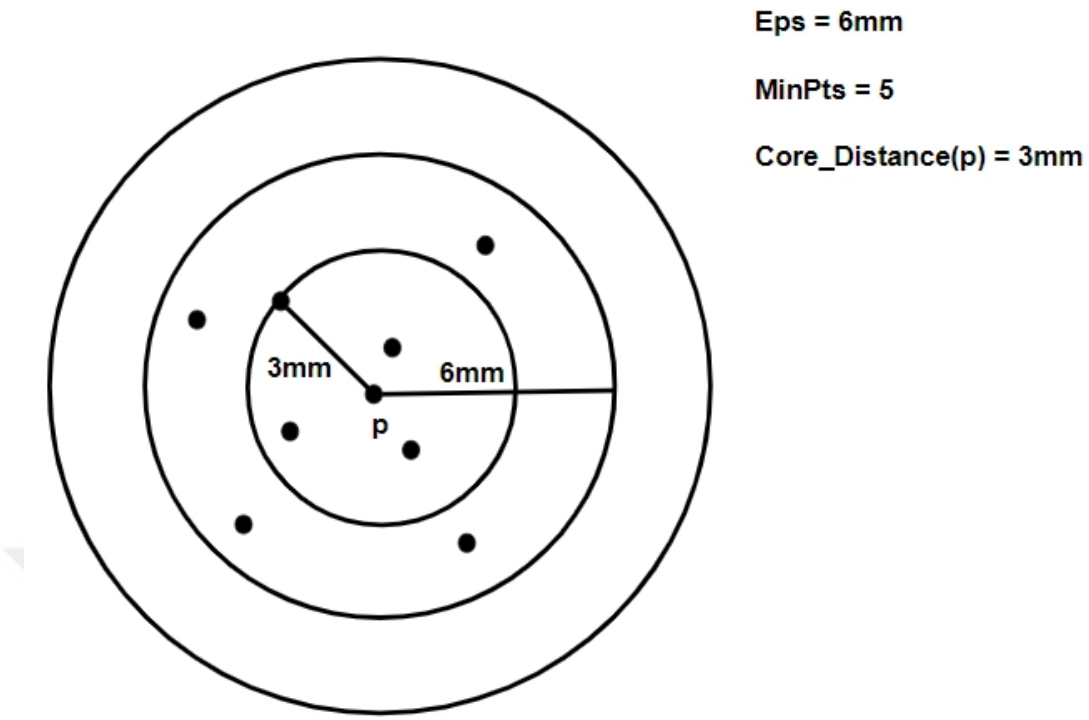


FIGURE 3.9 – Core Distance Determination on OPTICS

Image courtesy from (Dhanyaja, accessed : 28.05.2023)

core point. Reachability distance is calculated in respect to q, another piece of data. The reachability distance between two points p and q is the maximum of the Euclidean Distance (or any other distance metric) between the two points and the Core Distance of p. The reachability distance is not stated if q is not a core point.

4 EXPERIMENTAL SETUP

After proposed objectives of this thesis and data preparation, we have prepared an experiment setup with the knowledge that we have took from literature research. In this section, we aim to explain these experimental steps. All these steps summarised at figure 4.1.

Although next section explains all steps in detail, the summary of experiments like as :

1. After data preparation, NMF and SPCA dimension reduction methods applied to hotel feature table.
2. All of the hotel clustering methods that we applied explained at section 3.3.
3. According to performance metrics of clusters, we have decided the best dimension methodology and its applied clustering algorithm.
4. Cosine similarity applied to hotel features table to eliminate noise labelled hotels.The hotel clusterID's updated according to result of cosine similarity matrix.
5. Noise labeled hotels have been appended to matched clusterID's to result of decided clustering model.
6. These clusterID's have been matched on sales analysis table.
7. Designed recommendation engine has taken these results as input and recommended 10 hotel to each customer. The detail of recommendation system engine have been explained on section 4.2.
8. Recommendation output compared with sales test table.
9. Analysis of the results.

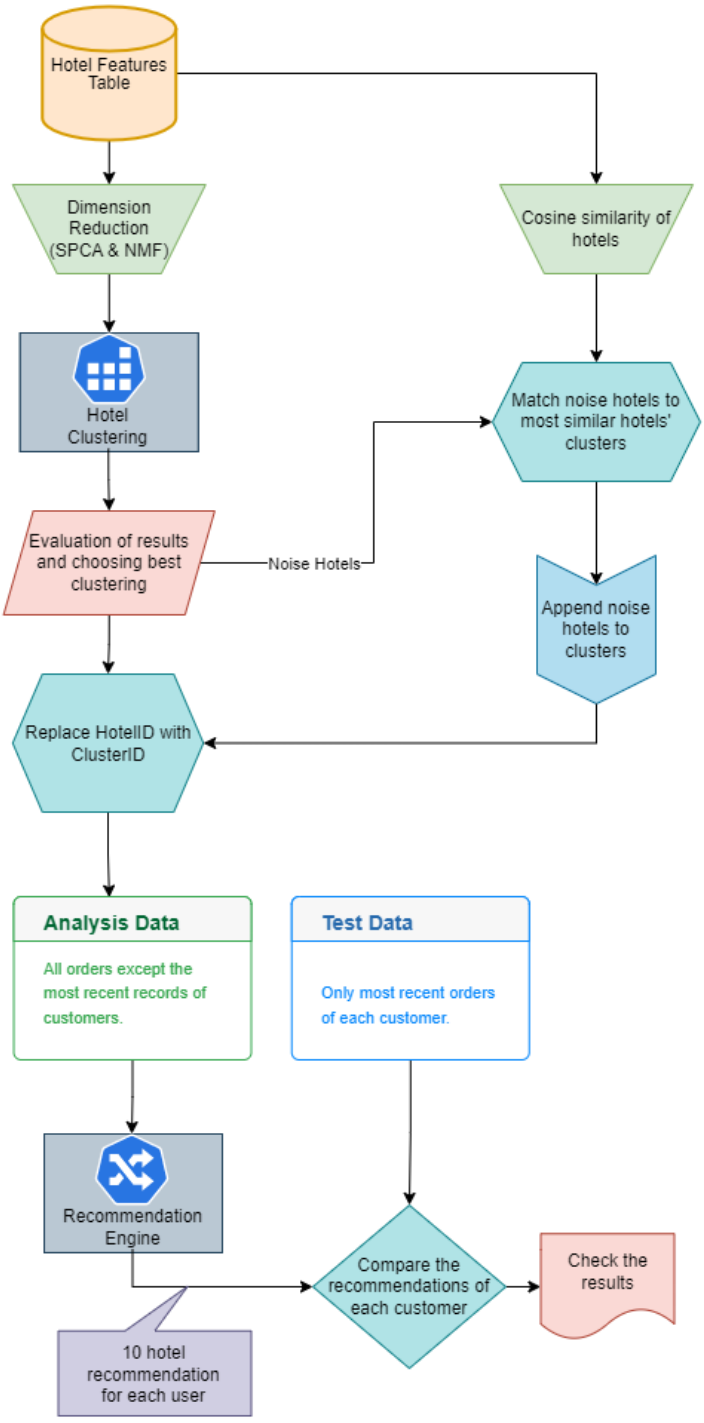


FIGURE 4.1 – Experimental Setup Flow Chart

4.1 Hotel Clustering

Main objective of this experiment was setting up the most well-separated clustering of hotel data, which was difficult to cluster in its raw form. For this reason, we have applied two different dimension reduction techniques and four clustering algorithms that explained on chapter 3. We have focused on two important criteria of the final cluster structure when choosing both the most proper dimension reduction and the most effective clustering algorithm. These criteria are as follows : first, the clustering structure with the most cohesive clusters with high intra-similarity, and second, the clustering structure with the most well-separated clusters with high inter-distance. Calculation of inter distance and intra similarity methods and different distance metrics have been explained by Solen J. et. al. (Soler et al., 2013).

The two dimension reduction methods, SPCA and NMF are applied with the feature number from 2 to 26. The minimum dimension count has been chosen as 2, since with less than 2 dimensions, the variance of the data will be lower than 0.7 which means a high loss from the variability information of the data set. The maximum value must be less than all feature numbers (i.e. number of feature columns).

Cluster numbers, K or n, have been selected between 2 and 50 for K-Means and HAC Single/Average/Complete/Ward methods according to the empirical experiments. The best cluster number is chosen by applying well-known elbow method. Figure 4.2 is an example of 15-dimensioned k-means clustering with WCSS scores. This figure shows that WCSS (Within-Cluster Sum of Square) values are becoming flat for $n > 50$ values. Regarding to these experiments, k-means cluster numbers chosen between 1 and 50 ($50 \geq k > 1$).

As explained in DBSCAN section, epsilon (ε) variable defines the distance between nodes in the same cluster. In our work, we have determined $\varepsilon_{min} = 0.02$, $\varepsilon_{max} = 0.3$ and $\Delta = 0.001$.

Even though, the selected clustering methods have best performance metrics, SPCA-OPTICS model creates noise hotels. We have prepared another step for noise hotels to increase recommendation quality. Otherwise, customers who have bought only noise labeled hotels, couldn't get any recommendation. Thereby, we have calculated cosine

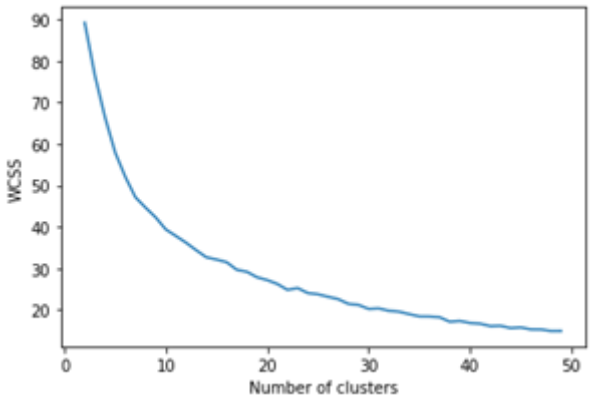


FIGURE 4.2 – WCSS Score - Cluster Numbers

similarity for each noise hotels, within hotel feature table where included only clustered hotels. As a result, we have gotten hotel clusters by OPTICS and neighbour noise hotels to a member of these clusters.

We have designed code blocks with Python programming language on Jupyter notebook of VS Code IDE. Mainly there were many for loops to iterate the parameters of dimension reduction methods and clustering models. When we obtained clustering results, features of hotels have been analyzed with SandDance extension of Azure Data Studio. Obtained visuals (APPENDIX) have been analysed on section 5.1.1 in regards of feature homogeneity and heterogeneity.

4.1.1 Used Performance Metrics on Clustering

All of the clustering methods and both dimension reduction methods that used in this thesis have been aimed at maximizing the silhouette score. A couple of performance metrics, which are explained in the next sub-sections, have been used to compare these clustering and dimension reduction methods. Some of these performance metrics (i.e., Rand Index Score, Adjusted Rand Index Score, Mutual Information Score, Adjusted Mutual Information Score) requires ground truth labels which the labels obtained from field specialists. And the rest (i.e., Silhouette Score, Calinski Harabasz Score, Davies Bouldin Score) are internal performance metrics of the clustering model, which does not require ground truth labels.

4.1.1.1 Silhouette Score

With simple terms, silhouette score is a measurement that compare the point distance to neighbour cluster's points. Rousseeuw, Peter explains as : "Each cluster is represented by so-called silhouette, which is based on the comparison of its tightness and separation. This silhouette shows which objects lie well within their cluster and which ones are merely somewhere in between clusters." (Rousseeuw, 1987). The result of the nth iteration of the clustering model that achieves the maximum silhouette score has been recorded.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, if|C_I| > 1 \quad (4.1)$$

where $|C_I|$ is the number of points belonging to cluster i.

4.1.1.2 Calinski-Harabasz Score / Index

This internal metric was introduced by Calinski and Harabasz in 1974. The C-H index is an evaluation method based on the degree of dispersion between clusters and within-cluster dispersion. A higher score on the index means better clustering dispersion. The C-H index for K number of clusters on a dataset D,

$$C - H = [\frac{\sum_{k=1^K} n_k||c_k - c||^2}{K-1}]/[\frac{\sum_{k=1^K} \sum_{i=1} n_k||d_i - c_k||^2}{N-K}] \quad (4.2)$$

where, n_k and c_k are the number of points and centroid of the k^{th} cluster respectively, c is the global centroid, N is the total number of data points.

4.1.1.3 Davies Bouldin Score / Index

Another internal metric that we have used in our evaluation method is the Davies Bouldin Score. The scoring algorithm is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Thus, clusters that are farther apart than the others and less dispersed will result in a better score. On the D-B score, Lower scoring shows a higher quality of clustering. Similarity is defined as a measure R_{ij} that trades off :

— s_i, the average distance between each point of cluster i and the centroid of that cluster – also know as cluster diameter.

— d_{ij}, the distance between cluster centroids i and j.

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (4.3)$$

Then the Davies-Bouldin index is defined as :

$$D - B = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \quad (4.4)$$

where k is number of cluster.

4.1.1.4 (Adjusted) Rand Index Score

The Rand index score determines the degree of similarity between calculated and ground truth cluster labels. Since our dataset does not contain the true label as a column, we can not compare the true clusters and predicted clusters. Therefore, we have taken SPCA cluster labels as true labels and compared them with NMF cluster labels. Adjusted Rand Index (ARI) improved algorithm from Rand Index. It's proposed to avoid higher cluster numbers that cause a higher index score trap. Higher scoring means identical cluster sequences in the compared arrays for the adjusted rand index score. If C is a ground truth class assignment and K the clustering, let us define a , b as :

— a, the number of pairs of elements that are in the same set in C and in the same set in K

— b, the number of pairs of elements that are in different sets in C and in different sets in K

The unadjusted Rand index is then :

$$RI = \frac{a + b}{C_2^{n_{samples}}} \quad (4.5)$$

where $C_2^{n_{samples}}$ is the total number of possible pairs in the dataset.

$$ARI = \frac{RI - E_{xpected}[RI]}{max(RI) - E_{xpected}[RI]} \quad (4.6)$$

4.1.1.5 (Adjusted) Mutual Information Score

Mutual information score is a measurement of the similarity between two different labels of the same data. Even though this performance metric requires ground-truth

labels, it is also symmetric, which means that switching predicted labels and true labels results in the same score. This feature allows us to use labels as a second label array that is gathered with SPCA, instead of ground truth labels. Like adjusted rand index scoring, an adjusted version of the mutual information score (AMI) helps us avoid the same trap caused by a higher number of clustering labels. Higher scoring means a higher chance of mutual information within same-labeled clusters. For two clusterings U and V, the mutual information is given as :

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i,j) log \frac{P(i,j)}{P(i)P'(j)} \quad (4.7)$$

$$AMI(U,V) = \frac{[MI(U,V) - E_{xpected}(MI(U,V))]}{[avg(H(U),H(V)) - E_{xpected}(MI(U,V))]} \quad (4.8)$$

4.2 Recommendation System Engine

As we explained in introduction chapter of this thesis, the main objective is proposing a hotel recommendation system for tourism sector. To establish this, we have designed a recommendation system engine after decision of best clustering model. Schema of recommendation engine can be seen at figure 4.3 Already, we have separated sales table to two different tables that explained at section 3.1.3.2. And after cluster process, we have been replaced hotelID values with ClusterID values. Therefore, sales analysis table will be referred as "revised sales analysis table" on next steps.

We have created customer cosine similarity matrix with using of ClusterID values from revised sales analysis table. According to that new similarity matrix, we had a list of most similar customers of targeted customer. This creates arrays where values are other userID's for each targeted customer. Within these similarity arrays of customers, we have kept the similarity weight of each customer inside of dynamically created Python dictionaries.

For each targeted customer, we design a calculation method to estimate the number of ballot for each closest customers. Thereby, each targeted customer will take recommendation starting from the most closest customer. Even though, there can be some close customers that have more than one ballot right, total ballot number limited to 10 regarding to number of recommendation that we would suggest. We have tried different

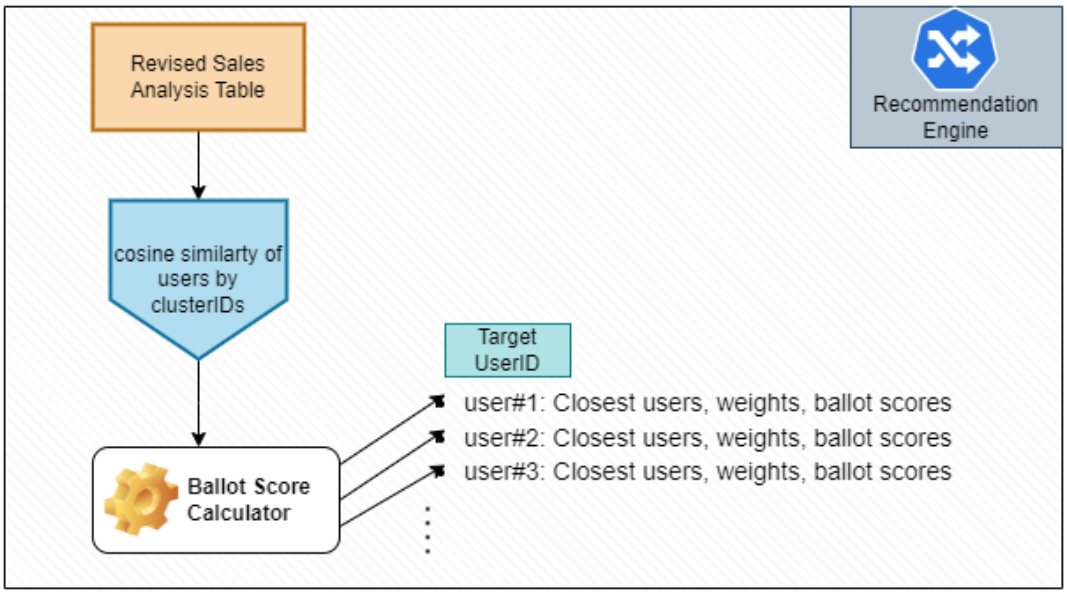


FIGURE 4.3 – Recommendation Engine Diagram

TABLE 4.1 – Example of Closest Customer Dictionaries

Rank	CustomerID	Weight	Ballot Score
1	47617	0.81649658092773	2.0
2	17208	0.73029674334022	1.0
3	129804	0.70710678118655	1.0
4	17132	0.70710678118655	1.0
5	17907	0.70710678118655	1.0
6	19228	0.70710678118655	1.0
7	19133	0.70710678118655	1.0
8	17082	0.70710678118655	1.0
9	75469	0.70710678118655	1.0
10	23012	0.70710678118655	NaN

multiplier factor to reach maximum number of matched recommendation in sales test table. The weight parameter will be discussed in section 5.2. An example of results ; 10 most similar customers of userID :171, their weights and their calculated ballot counts have been shown at table 4.1.

5 RESULTS

The findings of the thesis could serve as a basis for developing hotel recommendation systems for academics and professionals in the private sector. Even though accuracy of the recommendation is low, better performance metrics can be obtained with un-anonymised, clean datasets on future works. We have succeeded to obtain 10 different hotel recommendations to each user as in our purpose that was explained on the beginning of the thesis.

As we explained on the introduction chapter, our first aim was to get the best clustering methodology and then a recommendation system experiment that connected to the chosen clustering methodology. The results of the first part of the research may be considered independently. However second part of the research has affected the achievement of the first part which is hotel clustering. Overall, personalized, hybrid recommendation approach is proposed in this thesis which consist all types of collaborative filtering to handle noise labeled hotels and similar user preferences.

5.1 Hotel Clustering Results

Within the maximum silhouette scores of different clustering models, different number of clusters have been recorded for NMF and SPCA dimension reduction methods. Figure 5.1 shows the maximum silhouette scores and cluster numbers obtained.

The complexity of the experimental setup increases the complexity of the result sets. We represent all the result scores related to our experiments in Table 5.1. The data owner does not have any approved clustering labels except salesman inquiries such as ‘most liked, easiest to sell, most chosen. . . ’ to bring to light the hotel groups (i.e., clusters) therefore, we had to take one dimension reduction method as true labels for rand index, mutual information and their adjusted versions. Following paragraphs explain clustering methods versus dimension reduction methods and overall high-performing clustering decisions. The dilemma of "performance scores vs. cluster numbers" has been left for the discussions section.

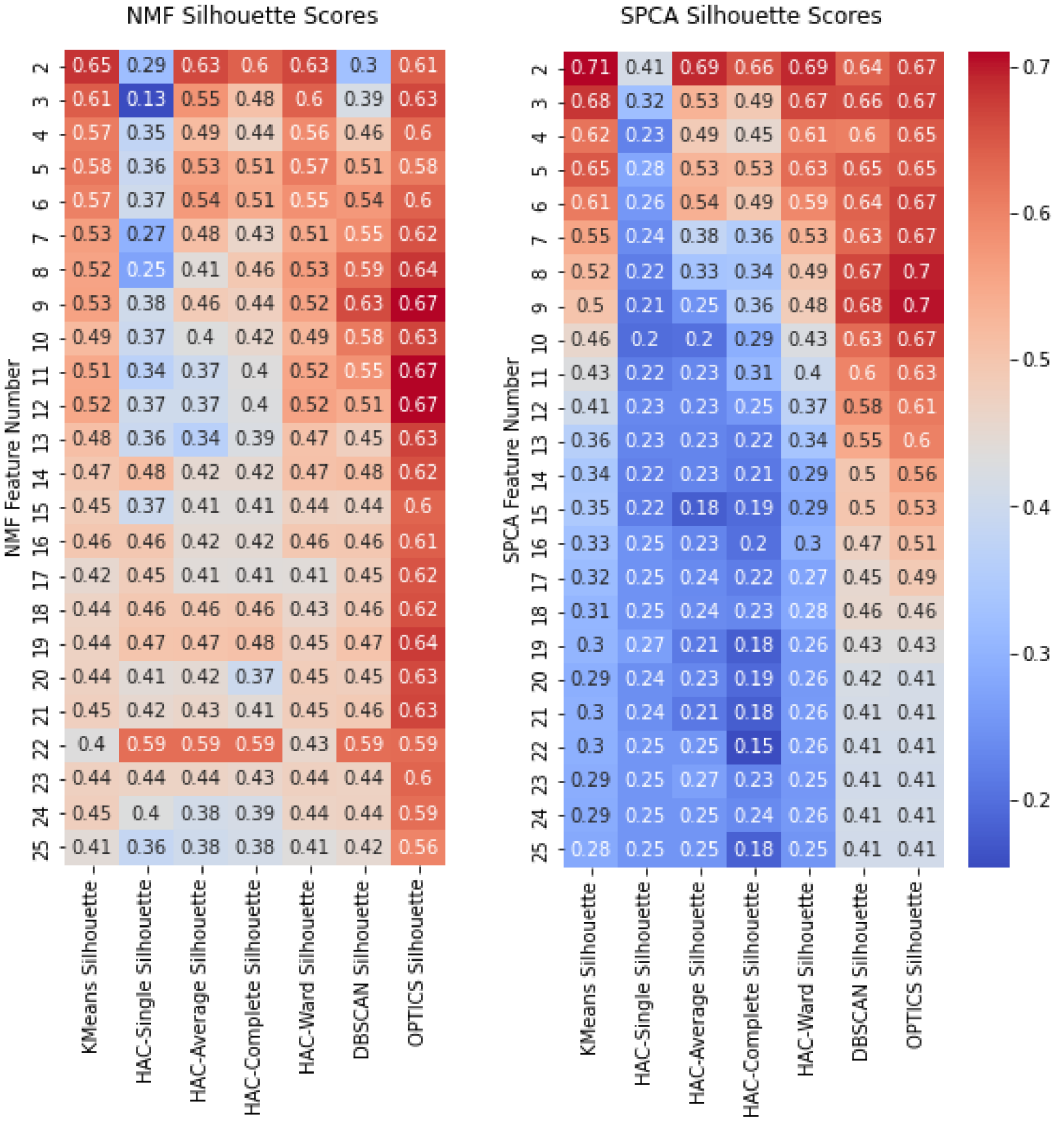


FIGURE 5.1 – NMF and SPCA Silhouette Scores

K-means and HAC (Hierarchical Agglomerative Clustering) require the cluster numbers in advance before the analysis. Akyol, Mert (Akyol, 2021) worked with a similar dataset (hotel feature set from Turkey) and the researcher used 20 clusters for their K-means model. Furthermore, we have calculated WCSS values without applying any dimension reduction method to our dataset. With help of previous work and our analysis, we took cluster values between 2 and 50. Despite of reached highest silhouette scores at 2-dimensional space, cluster numbers are quite low with comparison of different methods. Therefore, SPCA reduced space has a higher C-H score and a lower D-B score. Since we have different cluster numbers on each reduction model, we have only compared ARI and AMI scores. Since these metrics are comparisons of two different clustering label arrays, we have only one score for both dimension reduction methods that is calculated as : "NMF Cluster Arrays/Members" divided by "SPCA Cluster Arrays/Members". ARI and AMI score almost as high as the same values on HAC models but lower than the OPTICS and DBSCAN models.

As we mentioned in section 3.3.2, HAC has four different linkage methods. According to our silhouette scores, we can order the HAC linkage types as (weakest to strongest) : single $<$ complete $<$ average $=$ ward for NMF and SPCA reductions. C-H and D-B scores follow the same order as the given order. Single linkage has average RI and lowest AMI scores. Despite having the same cluster numbers, NMF generates ten times the dimension space as the SPCA method. Randomness and non-mutuality could have been caused by dimension difference. On the contrary, complete linked models have fewer space dimensions (only 2) at maximum silhouette scores. Almost the same ARI scores were obtained with single linkage's. Higher AMI scores have been obtained with complete linked HAC. Average and ward linked HAC models show the same patterns between NMF and SPCA methods. Overall evaluation of C-H, ARI and AMI scores ; shows that ward linkage has higher clustering quality over dimension reduction methods except D-B scoring. For that exception, the difference is only 0,009.

DBSCAN provides the second highest value of silhouette score on summary table 5.1.

Despite ranking second in silhouette scoring, the DBSCAN model has the highest value in RI, ARI, MI, AMI scores. This proves that 9th dimensional space and 112 different clusters have the same order and contain the same nodes. Even though the calculated epsilon values (0.047 on NMF, 0.282 on SPCA) are quite different, the noise counts are the same for both methods.

On the overall models explained thus far, the OPTICS model has the highest silhouette scores over both dimension reduction methods. We have obtained different numbers of dimension space and cluster numbers, C-H and D-B scores differ too. ARI score getting close to k-means and HAC ward/average models' ARI scores. OPTICS models have been calculated to have lower noise points (in orderly 226, 244 at maximum silhouette scores) than the DBSCAN models (400 for both methods at maximum silhouette scores). Both of the OPTICS and DBSCAN models have been take minimum cluster member parameter as 5. Therefore, less than 5 neighbours considered as noise in our dataset. Lower values decrease the number of noise but increase the number of clusters.

As a final statement, both of dimension reduction methods can be applied on our dataset and OPTICS models will provide high quality clusters. Thereby, we have used 9 dimension SPCA reduction and OPTICS clustering results in our recommendation system.

TABLE 5.1 – Summary table of NMF and SPCA Dimension Reduction methods over clustering methods

Feature	**NMF**	**SPCA**
Reduced Dimension Number on Kmeans	2	2
K-means Cluster Number	28	43
K-means Silhouette Score	**0,65**	**0,71**
K-means Calinski Harabasz Score	8111,64	15980,60
K-means Davies Bouldin Score	0,52	0,47
K-means Rand Index Score	0,97	
K-means Adjusted Rand Index Score	0,57	
K-means Mutual Information Score	2,71	
K-means Adjusted Mutual Information Score	0,77	
Reduced Dimension Number on HAC (Single)	22	2
HAC (Single) Cluster Number	2	2
HAC (Single) Silhouette Score	**0,59**	**0,41**
HAC (Single) Calinski Harabasz Score	212,42	1734,61
HAC (Single) Davies Bouldin Score	0,71	1,07
HAC (Single) Rand Index Score	0,50	
HAC (Single) Adjusted Rand Index Score	0,00	
HAC (Single) Mutual Information Score	0,00	
HAC (Single) Adjusted Mutual Information Score	0,01	
Reduced Dimension Number on HAC (Average)	2	2
HAC (Average) Cluster Number	29	38
HAC (Average) Silhouette Score	**0,63**	**0,69**
HAC (Average) Calinski Harabasz Score	7062,89	12576,80
HAC (Average) Davies Bouldin Score	0,51	0,44
HAC (Average) Rand Index Score	0,96	
HAC (Average) Adjusted Rand Index Score	0,51	
HAC (Average) Mutual Information Score	2,52	
HAC (Average) Adjusted Mutual Information Score	0,75	
Reduced Dimension Number on HAC (Ward)	2	2
HAC (Ward) Cluster Number	29	42

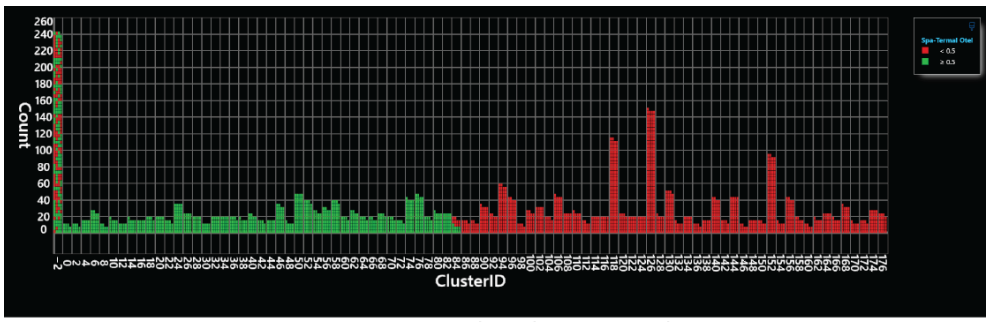
Table 5.1 continued from previous page

Feature	NMF	SPCA
HAC (Ward) Silhouette Score	**0,63**	**0,69**
HAC (Ward) Calinski Harabasz Score	7670,26	15087,40
HAC (Ward) Davies Bouldin Score	0,53	0,45
HAC (Ward) Rand Index Score	0,97	
HAC (Ward) Adjusted Rand Index Score	0,57	
HAC (Ward) Mutual Information Score	2,74	
HAC (Ward) Adjusted Mutual Information Score	0,79	
Reduced Dimension Number on DBSCAN	9	9
DBSCAN Cluster Number	112	112
DBSCAN Silhouette Score	**0,63**	**0,68**
DBSCAN Calinski Harabasz Score	105,79	106,13
DBSCAN Davies Bouldin Score	1,03	1,04
DBSCAN Rand Index Score	1,00	
DBSCAN Adjusted Rand Index Score	1,00	
DBSCAN Mutual Information Score	4,06	
DBSCAN Adjusted Mutual Information Score	1,00	
Reduced Dimension Number on OPTICS	12	9
OPTICS Cluster Number	182	178
OPTICS Silhouette Score	**0,67**	**0,70**
OPTICS Calinski Harabasz Score	76,27	93,03
OPTICS Davies Bouldin Score	1,17	1,12
OPTICS Rand Index Score	0,98	
OPTICS Adjusted Rand Index Score	0,54	
OPTICS Mutual Information Score	4,12	
OPTICS Adjusted Mutual Information Score	0,79	

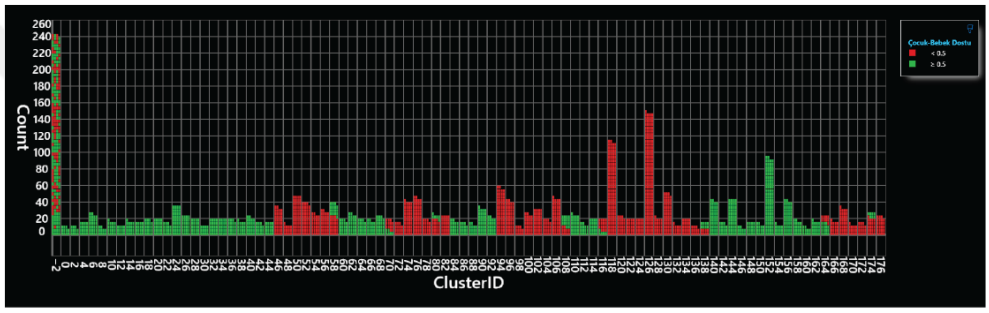
5.1.1 Feature Homogeneity of Hotel Clusters

In this chapter some features have been explained that we used on hotel features table, regarding to homogeneity and heterogeneity distributions over clusters and rare features that we have observed. Homogeneity of feature can be explained like as the served

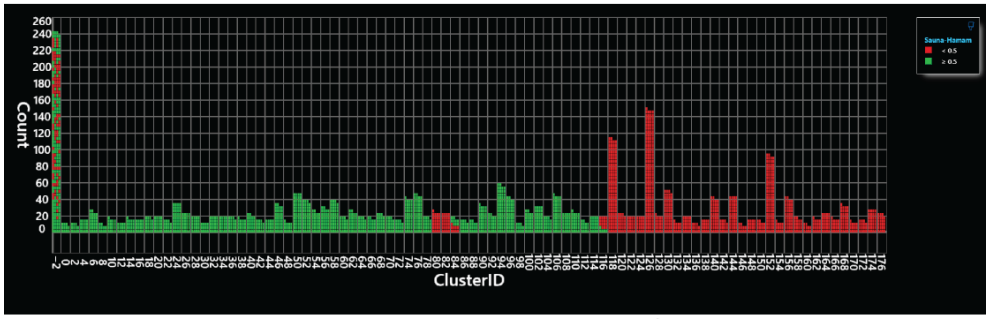
feature of hotels are kept in some cluster borders. This means that all of the cluster members have that feature in specific cluster, or vice versa. "Spa Termal Hotel", "Çocuk Bebek Dostu", "Sauna Hamam", "Oda Kahvaltı", "Fitness" features have homogenise form over cluster and the first 4 features have been shown at figure 5.2. Heterogeneity of features can be explained as the feature served in some of the hotels over different clusters. Therefore there is no stricst border lines between clusters on figure 5.3. Some of these features are "Kumlu Plaj", "Havuz", "Yarım Pansiyon", "Su Sporları". The hotel feature table was examined during the data preparation process ; this shows us, all features of table show on at least one hotel. But some of these features are too rare, therefore on the illustrations of clusters these features only show up in a few clusters. These features can be listed as "Çakıllı Deniz", "Casino Hoteli", Golf Hoteli", "Kumlu Deniz", "Tam Pansiyon". Figure 5.4 shows the first 4 features that we listed in previous sentence. Overall, as seen in the graphs, the features reflect heterogeneous distribution, and the uncommon ones do not assist our clustering approach. The bigger version of these graphs for both covered in this section and the other features that we haven't discussed about, have been appended on APPENDIX chapter of this thesis.



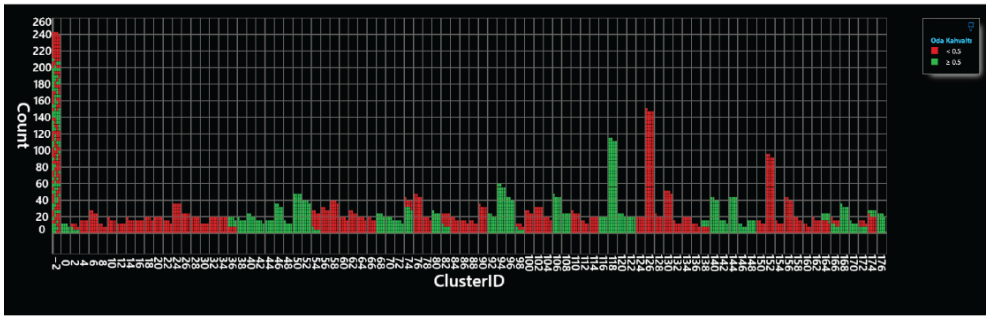
Feature "Spa Termal Hotel": Green 1, Red 0



Feature "Çocuk Bebek Dostu": Green 1, Red 0

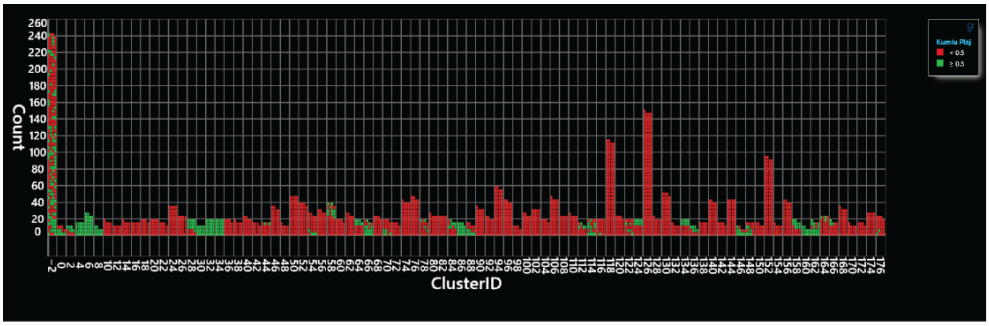


Feature "Sauna Hamam": Green 1, Red 0

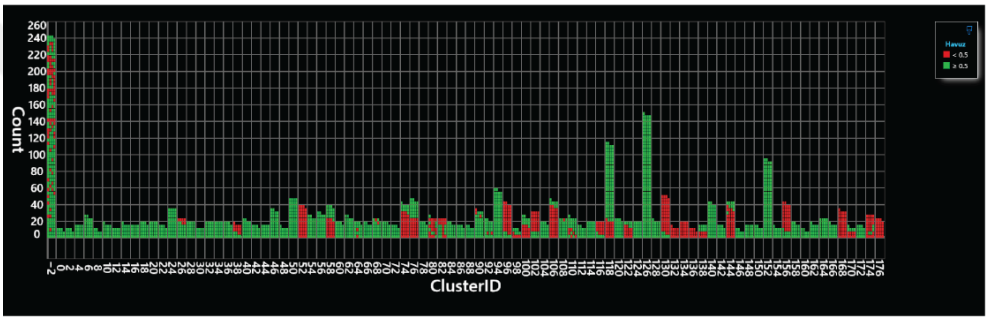


Feature "Oda Kahvaltı": Green 1, Red 0

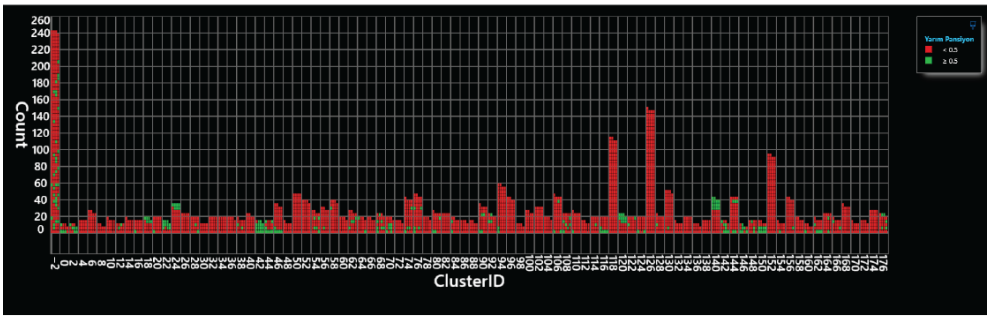
FIGURE 5.2 – Some Features That Shows Homogenise Distribution over Clusters



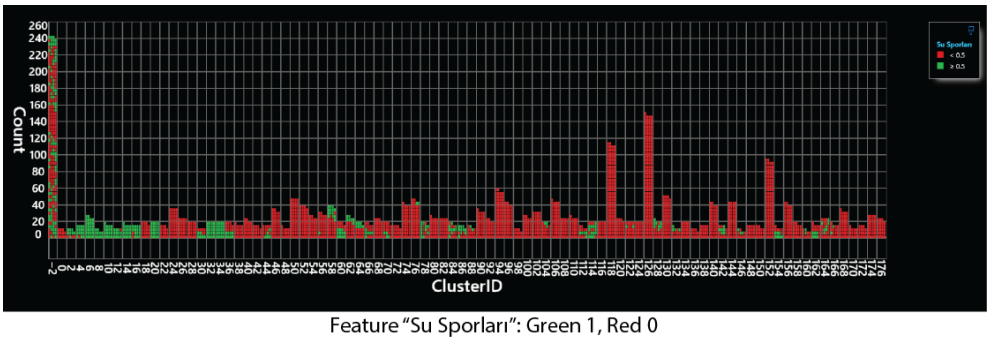
Feature "Kumlu Plaj": Green 1, Red 0



Feature "Havuz": Green 1, Red 0

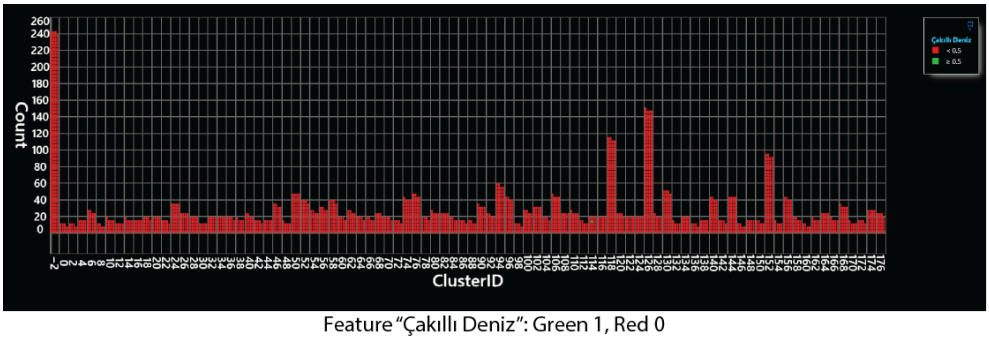


Feature "Yarım Pansiyon": Green 1, Red 0

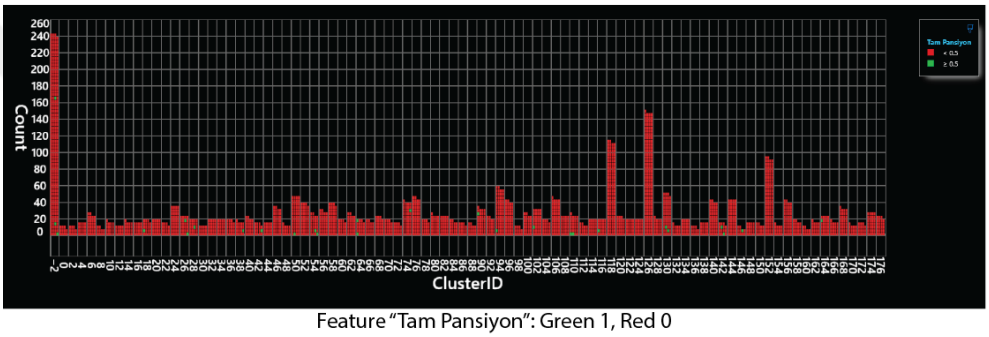


Feature "Su Sporları": Green 1, Red 0

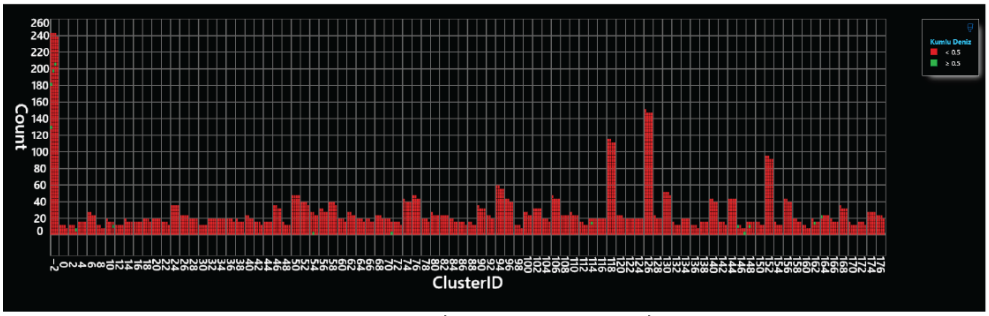
FIGURE 5.3 – Some Features That Shows Heterogeneous Distribution over Clusters



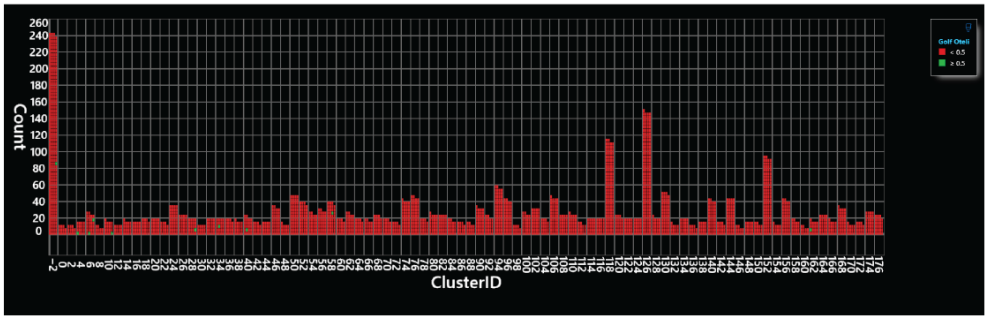
Feature "Çakıllı Deniz": Green 1, Red 0



Feature "Tam Pansiyon": Green 1, Red 0



Feature "Kumlu Deniz": Green 1, Red 0



Feature "Golf Hotel": Green 1, Red 0

FIGURE 5.4 – Rare Features

5.2 Recommendation System Results

The weight that we mentioned on recommendation system engine setup, is similarity multiplier for calculation of ballot numbers of similar customers. Thereby, decimal similarity scores between customer to customer have been converted to integer numbers. But it's is not simply rounding mechanism for decimal numbers. As we were performing the calculations, the ballot score procedure was used as a weight component to determine whether to give priority to the customer or customers who are the most similar. The distribution of similarity calculation, ballot scores of similar customers over two different weight multipliers have been shown at table 5.3. Some consumers may receive less than 10 products when choosing a weight less than 2. Since some of the targeted customers have less than 10 similar customers e.g. target customer "30941" only has 7 similar customers. As seen in the second portion of the table 5.3 recommendation order column for 8, 9, 10 ; do not have any similarity for the targeted customer.

Our recommendation algorithm has been tested repeatedly with various weights that affect the number of votes provided. Of course, the outcomes of the tests have been impacted by varying weights for similarity points.Binary success criteria can be explained as ; the values are checked whether the one of the recommended hotel matches at sales test table for that targeted customer. If previous statement positive, it is considered as success, if it is not ; it is considered as failure. Due to the fact that 85% of our clients only have two orders in the sales table, we haven't taken the ratio of our matched recommendations to all of our recommendations into consideration. Results for 2 different weights are summarised on table 5.2. According to table 5.2, ratio of succeeded recommendations to total recommendations are 15,9% and 14,1% as respectively selected weight 2 and 5.

TABLE 5.2 – Number of Succeded/Failed Recommendations over Different Weights

Similarity Weight	**2**	**5**
Succeded Recommendation	418	370
Failed Recommendation	2205	2253
Total Recommendation	2623	2623

TABLE 5.3 – Recommendation Example with Different Weights

Target Customer ID : 171			**Weight=2**	**Weight=5**
Recommendation Order	**Similar Customer**	**Similarity**	**Ballot Score**	**Ballot Score**
1	47617	0.8164966	2.0	4.0
2	17208	0.7302967	1.0	4.0
3	129804	0.7071068	1.0	2.0
4	17132	0.7071068	1.0	NaN
5	17907	0.7071068	1.0	NaN
6	19228	0.7071068	1.0	NaN
7	19133	0.7071068	1.0	NaN
8	17082	0.7071068	1.0	NaN
9	75469	0.7071068	1.0	NaN
10	23012	0.7071068	NaN	NaN
Target Customer ID : 30941			**Weight=2**	**Weight=5**
Recommendation Order	**Similar Customer**	**Similarity**	**Ballot Score**	**Ballot Score**
1	35054	1.0000000	2.0	5.0
2	47613	1.0000000	2.0	5.0
3	39012	1.0000000	2.0	NaN
4	29857	1.0000000	2.0	NaN
5	41081	0.7071068	1.0	NaN
6	47234	0.7071068	1.0	NaN
7	50902	0.5773503	NaN	NaN
8	13	0.0000000	NaN	NaN
9	53601	0.0000000	NaN	NaN
10	53377	0.0000000	NaN	NaN

6 CONCLUSION

In this study, we focused on the tourism data set and aimed at revealing an experimental procedure for finding the best cluster structure by using binary hotel features and creation a recommendation engine for similar customers. We also aimed to understand how different dimension reduction methods affect the clustering process. These results have sparked off discussion about the importance of dimension reduction methods and clustering algorithms, which lie behind hotel recommendation system designs. We have applied two dimension reduction techniques ; SPCA and NMF and four clustering algorithms ; K-means, HAC (with different linkages), DBSCAN and OPTICS. The results reveal that OPTICS can be the best clustering algorithm for this dataset especially when the number of features are reduced to between 9 and 12. Moreover, clustering performance metrics of the DBSCAN model prove that ; we can have the exact same clusters with the same dimensional space that are reduced by different dimension reduction methods. An important future step in this work could be to focus on data noise. DBSCAN and OPTICS result in high performance metrics but they find many noisy points. Understanding the dynamics of such noises can give us an idea about the "unique" hotels, which can be on the other side of the medallion for marketing strategies. The dilemma of "higher cluster numbers cause higher silhouette scores" may be another discussion topic about this thesis. Silhouette scores may not be the only evaluation score of high-quality clustering process. As we examined in this thesis, other performance metrics might be chosen in a bunch of iteration steps to understand hotel clustering. Our dataset covers more than 61% of registered touristic facilities. Thus, this clustering results could provide an overview to Turkey's hotels. With these insights, tourism recommendation systems may improve and increase the satisfaction of customers and hotel owners in the further steps.

As mentioned at previous chapters in this thesis, real-world datasets have been implemented. A loss of 84% of the sales dataset was caused by the application of anonymization methods and other problematic scenarios that we have encountered during data preparation steps at section 3.1.3.2. Thereby our sales data could have been biased or lost of fundamental parts of itself. Additionally, lack of expert insights caused finding an analytically acceptable solution for noise labeled hotels but we do not have

any verification of these noise labeled hotels and their clusters by cosine similarity. The purpose of finding at least one similar customer to target customer, directed us to usage of visited hotels' clusters. This solution creates big hotelID pool for recommendation engine but usage of second last visited of these hotels on sales analyze table was the another questionable part of this study. Because, the last and most up-to-date data separated for sales test dataset. Also this separation process of sales data has been created unusual ratio of analysis and test datasets. Since, most of the customers have only 2 order records in our sales table. Thereby, the number of analyze and test records were almost equal to each other. Whereas related papers have higher amount of analyze records than test records (Ramzan et al., 2019), (Türker et al., 2020). Regarding to these explanations, the result of recommendation was quite weak scores than we have expected at first starting point of this thesis.

Even though our hotel dataset covers more than half of the registered hotels in Turkey and usage of contemporary clustering and recommendation techniques accepted by academicians and experts, the overall results showed us, the methodology of this recommendation design has still can be improved. Nevertheless, used dimension reduction methods and extensively compared clustering methods have been created comparison and decision step point to other researches in our research team.

The related papers on published in tourism field have higher accuracy than our recommendation results (Ramzan et al., 2019), (Türker et al., 2020). Ramzan et. al, have huge amount of records their datasets and they studied different methods on their papers. Their F1 score reaches 95% on average. The paper of Türker et. al. have applied similar methods like as our thesis work but their analyze and test datasets consist higher amount of records than ours. The study of Türker et. al. is reached 21% recall score at max. Nevertheless last study shows us, improvement of data source could lead more accurate results for tourism recommender systems.

REFERENCES

Akyol, M. (2021). Clustering hotels and analyzing the importance of their features by machine learning techniques, *Journal of Computer Science and Technologies - Mersin Üniversitesi* **2**(1) : p. 16–23.

Ankerst, M., Breunig, M. M., Kriegel, H.-P. and Sander, J. (1999). Optics : Ordering points to identify the clustering structure, *SIGMOD Rec.* **28**(2) : p. 49–60.

Arthur, J. K., Doh, R. F., Mantey, E. A. and Osei-Kwakye, J. (2019). Complex network based recommender system, **178**(40).

Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A. (2013). Recommender systems survey, *Knowledge-Based Systems* **46** : p. 109 – 132.

Bult, J. R. and Wansbeek, T. (1995). Optimal selection for direct mail, *Marketing Science* **14** : p. 378–394.

Custobar (accessed : 10.10.2021). Rfm example of categorization.
URL: *www.custobar.com/blog/rfm-matrix/*

Das, D., Sahoo, L. and Datta, S. (2017). A survey on recommendation system, *International Journal of Computer Applications* **160** : p. 6–10.

Dağ, O. and KARAATLI, M. (2020). Resort otellerin kümeleme analizi ile incelenmesi : Antalya ili Örneği, *Journal of Suleyman Demirel University Institute of Social Sciences* (36) : p. 200–232.

Dhanyaja, N. (accessed : 28.05.2023). Optics.
URL: *www.youtube.com/watch ?v=9Zy-JfgAHRcab*$_c$*hannel = DhanyajaN*

Dhillon, I. S. and Sra, S. (2005). Generalized nonnegative matrix approximations with bregman divergences, **18** : p. 283–290.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions, *J. Am. Stat. Assoc.* **104**(486) : p. 682–693.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis : a review and recent developments, *Philos. Trans. A Math. Phys. Eng. Sci.* **374**(2065).

Orman, G. K. (2021). Data mining - clustering, University Lecture.

Ramzan, B., Bajwa, I. S., Jamil, N., Amin, R. U., Ramzan, S., Mirza, F. and Sarwar, N. (2019). An intelligent data analysis for recommendation systems using machine learning, *Sci. Program.* **2019** : p. 1–20.

Rodríguez-Victoria, O. E., Puig, F. and González-Loureiro, M. (2017). Clustering, innovation and hotel competitiveness : evidence from the colombia destination, *International Journal of Contemporary Hospitality Management* **29**(11) : p. 2785–2806.

Rousseeuw, P. J. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* **20** : p. 53–65.

Soler, J., Tencé, F., Gaubert, L. and Buche, C. (2013). Data clustering and similarity, *The Twenty-Sixth International FLAIRS Conference*, FLAIRS.

Sánchez-Pérez Manuel, Illescas-Manzano María Dolores, M.-P. S. (2020). You're the only one, or simply the best. hotels differentiation, competition, agglomeration, and pricing, *International Journal of Hospitality Management* **85** : p. 102362.

Turistik Tesis Ve İşletmeler (accessed : 10.01.2023).
URL: *http ://www.tursab.org.tr/istatistikler/turistik-tesis-isletmeler*

Türker, B. B., Tugay, R., Oguducu, S. G. and Kizil, I. (2020). Hotel recommendation system based on user profiles and collaborative filtering.

World Tourism Organization (accessed : 28.11.2022).
URL: *www.unwto.org/news/international-tourism-back-to-60-of-pre-pandemic-levels-in-january-july-2022*

Ye, M., Zhang, P. and Nie, L. (2018). Clustering sparse binary data with hierarchical bayesian bernoulli mixture model, *Computational Statistics Data Analysis* **123** : p. 32–49.

APPENDIX CLUSTERID - COUNT TABLES OVER HOTEL FEATURES

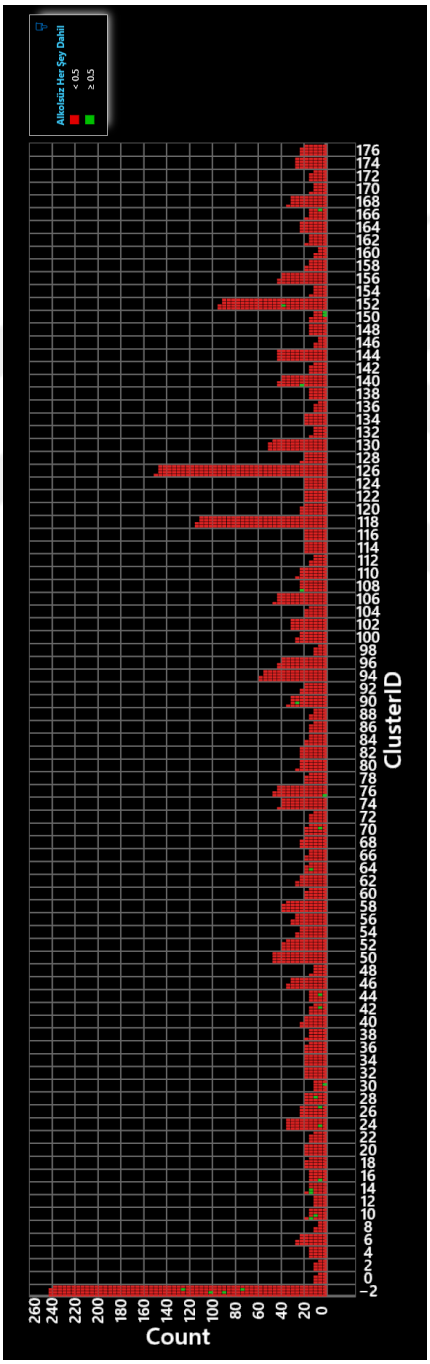


Figure APPENDIX.1 – ClusterID vs. Count on "Alkolsüz Her Şey Dahil"

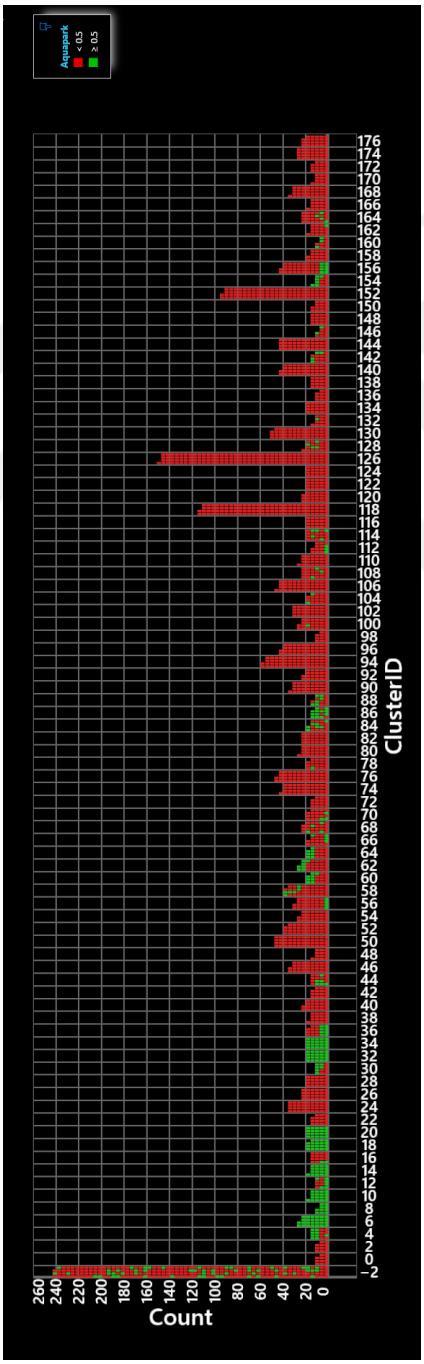


Figure APPENDIX.2 – ClusterID vs. Count on "Aquapark"

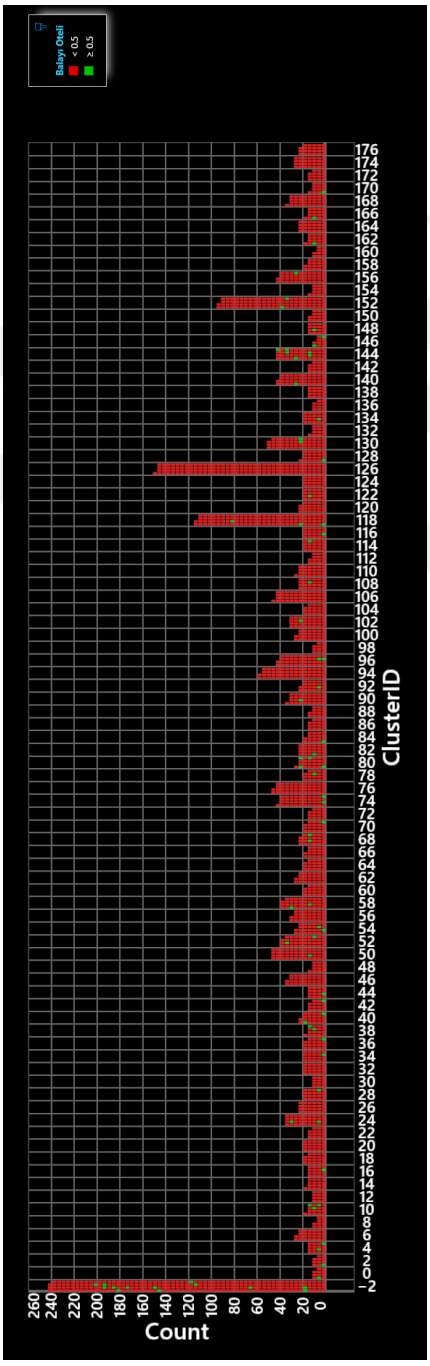


Figure APPENDIX.3 – ClusterID vs. Count on "Balayı Oteli"

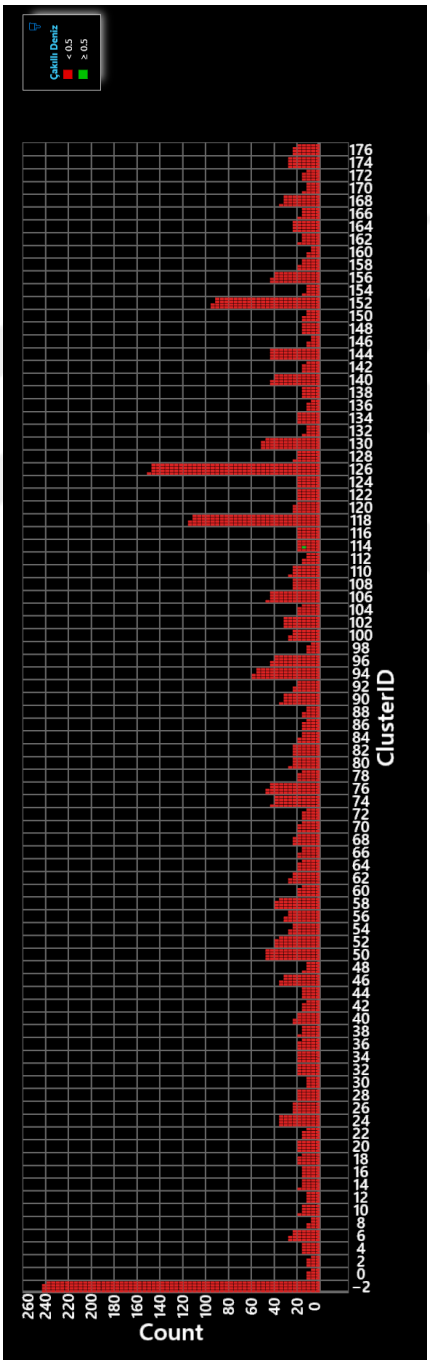


Figure APPENDIX.4 – ClusterID vs. Count on "Çakıllı Deniz"

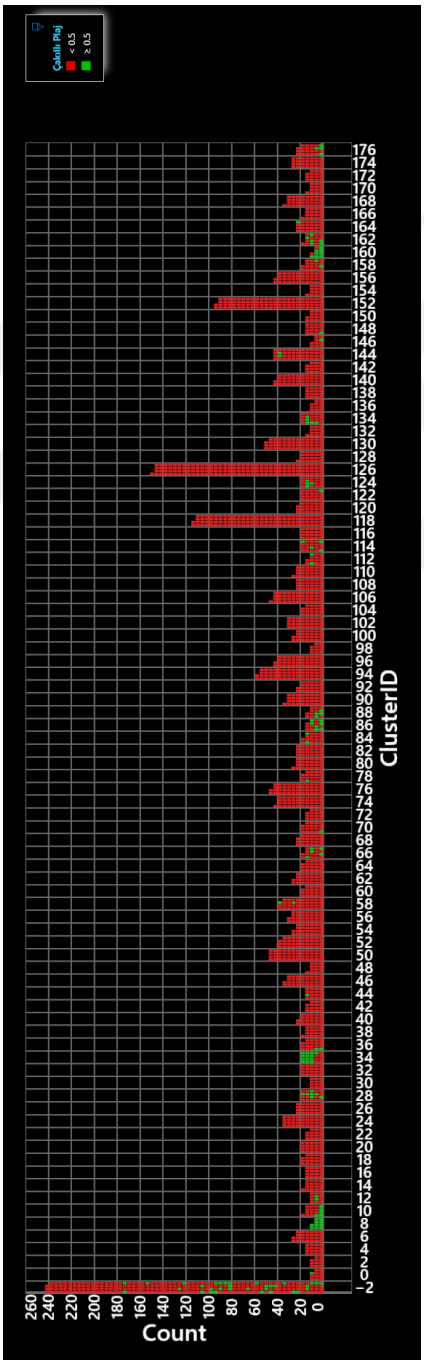


Figure APPENDIX.5 – ClusterID vs. Count on "Çakıllı Plaj"

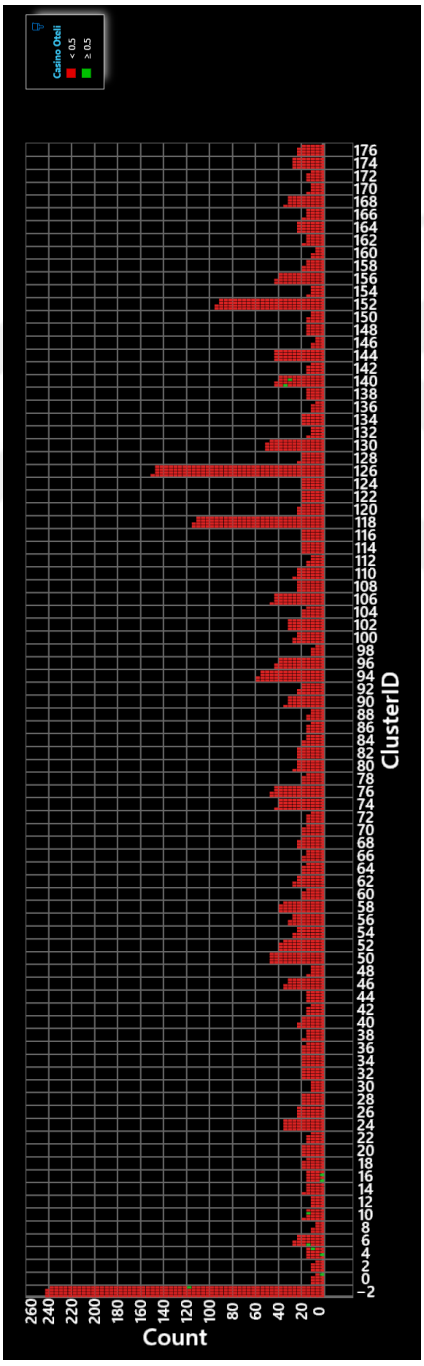


Figure APPENDIX.6 – ClusterID vs. Count on "Casino Oteli"

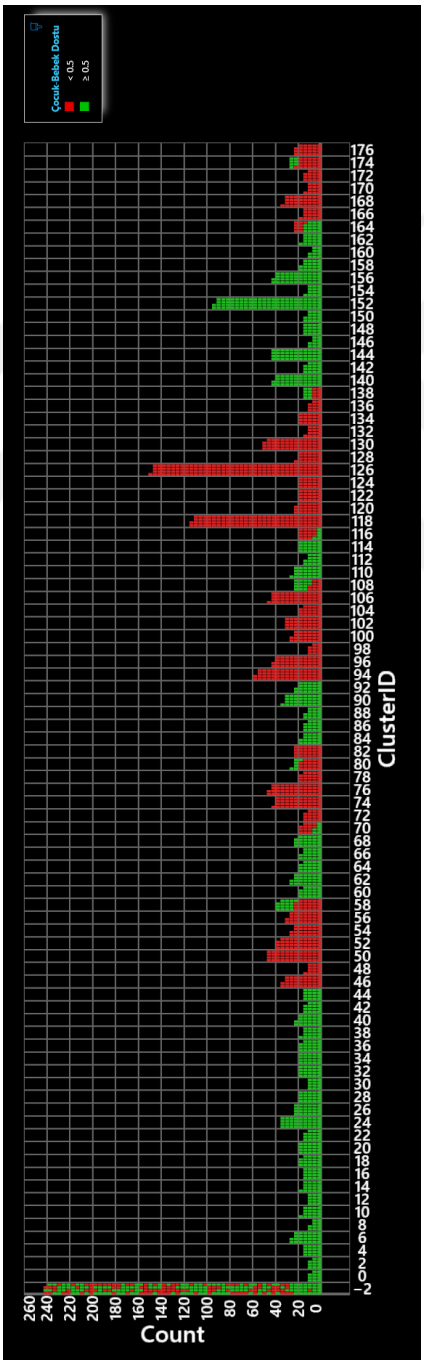


Figure APPENDIX.7 – ClusterID vs. Count on "CocukBebekDostu"

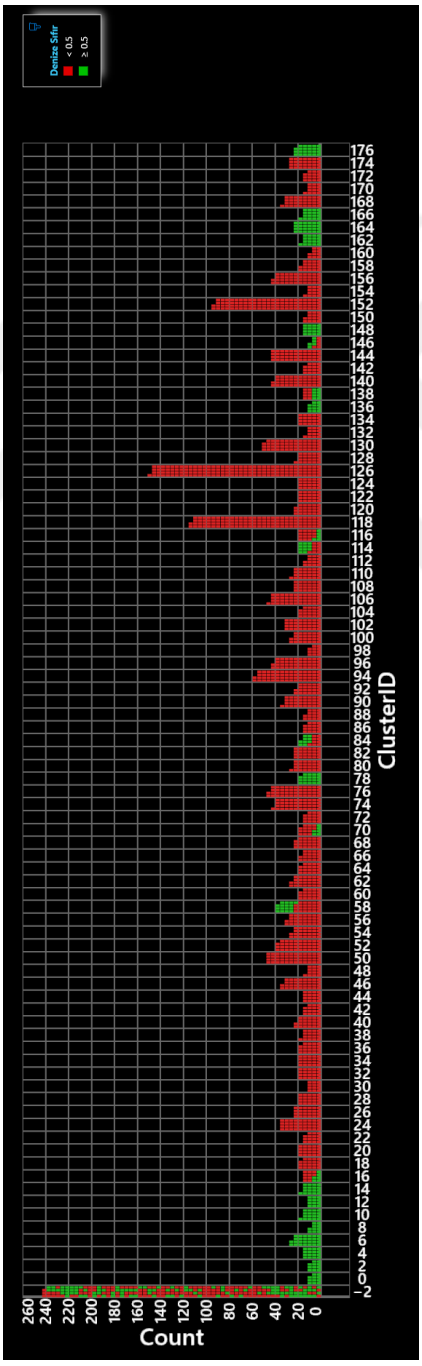


Figure APPENDIX.8 – ClusterID vs. Count on "Denize Sıfır"

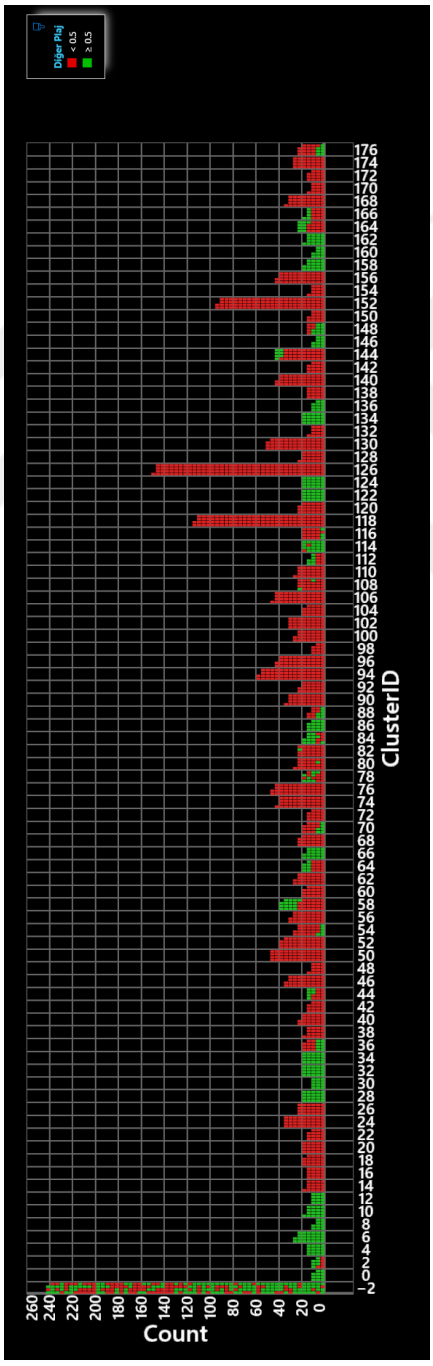


Figure APPENDIX.9 – ClusterID vs. Count on "Diğer Plaj"

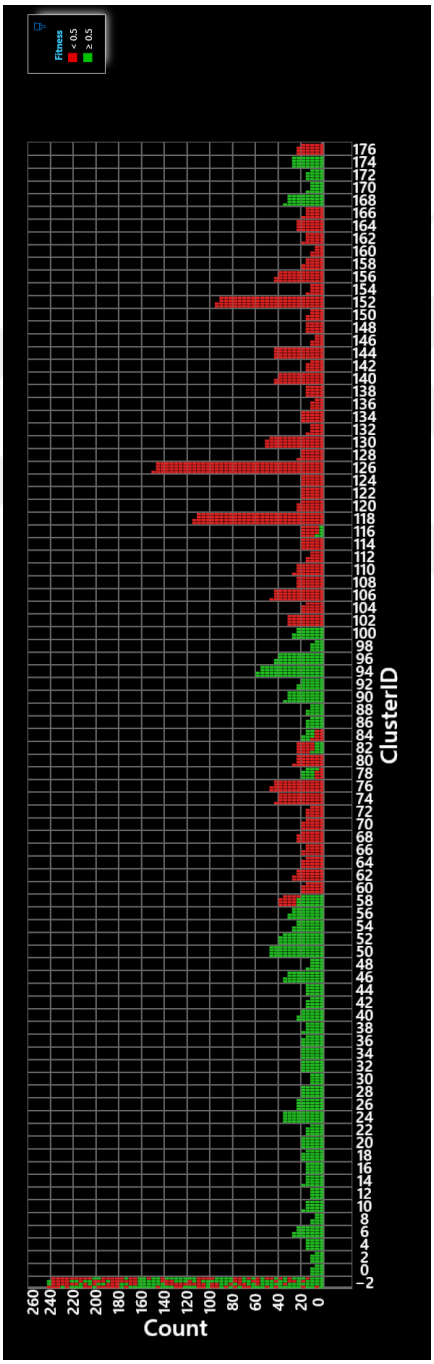


Figure APPENDIX.10 – ClusterID vs. Count on "Fitness"

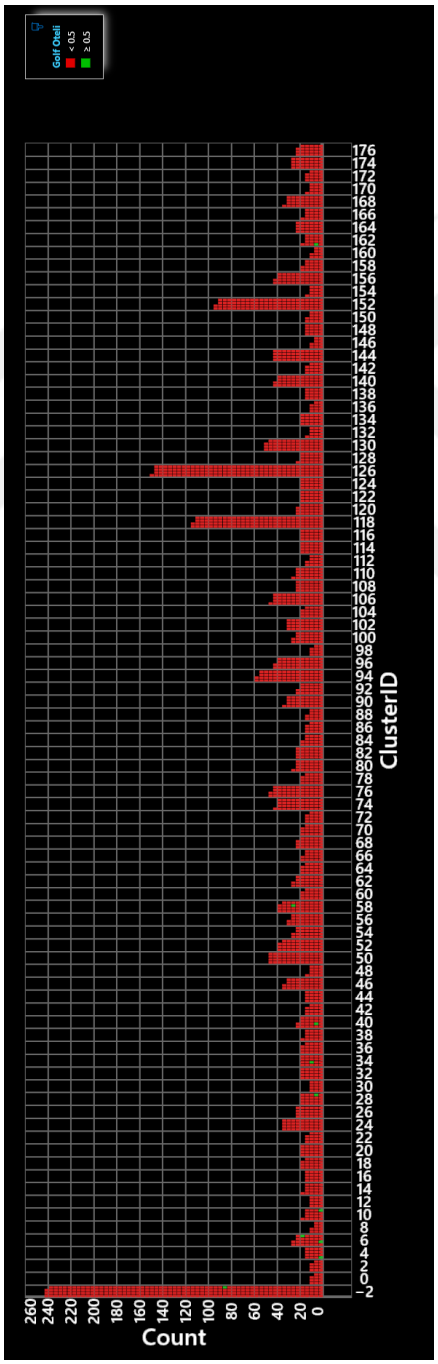


Figure APPENDIX.11 – ClusterID vs. Count on "Golf Oteli"

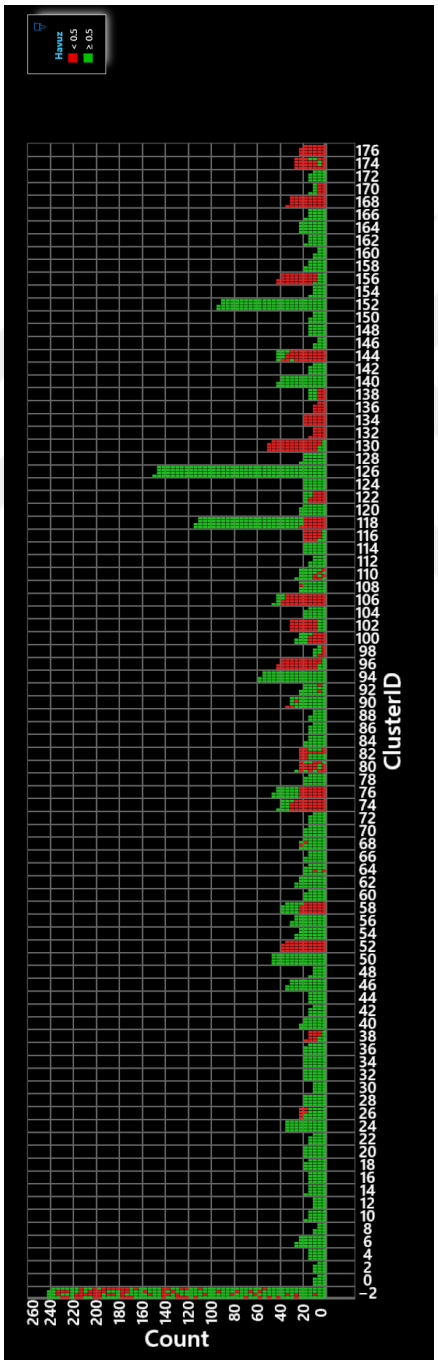


Figure APPENDIX.12 – ClusterID vs. Count on "Havuz"

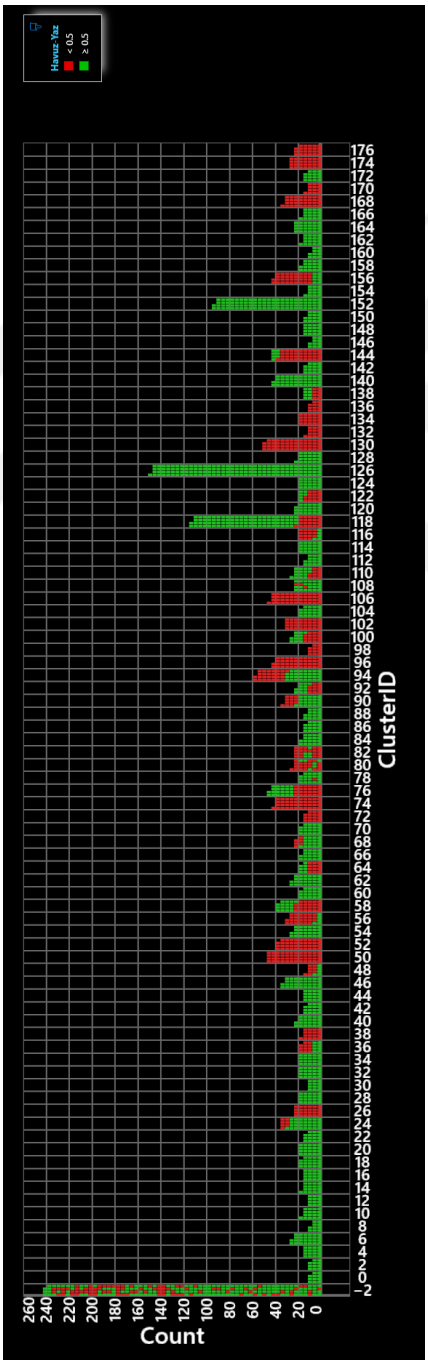


Figure APPENDIX.13 – ClusterID vs. Count on "Havuz Yaz"

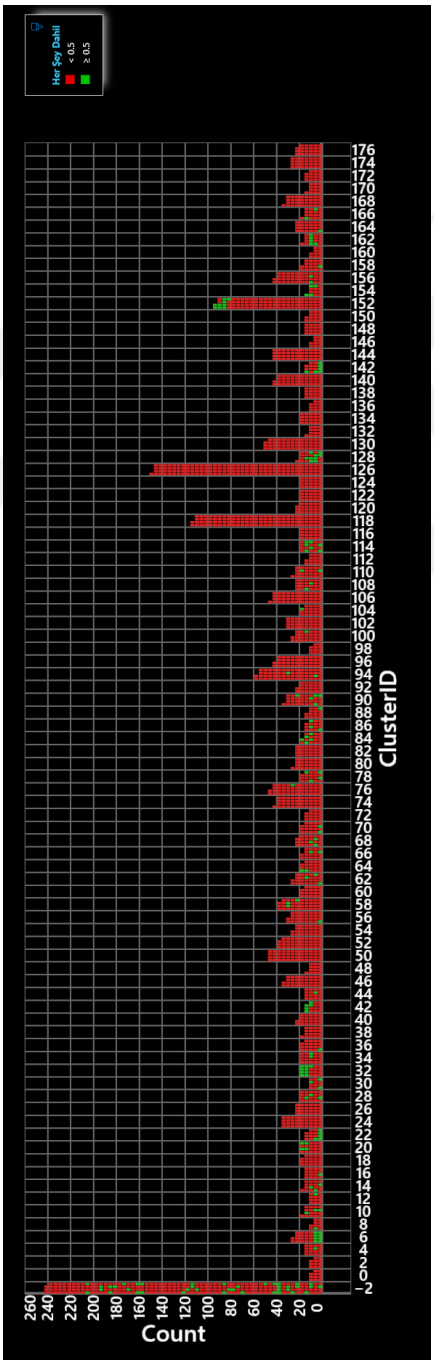


Figure APPENDIX.14 – ClusterID vs. Count on "Her Şey Dahil"

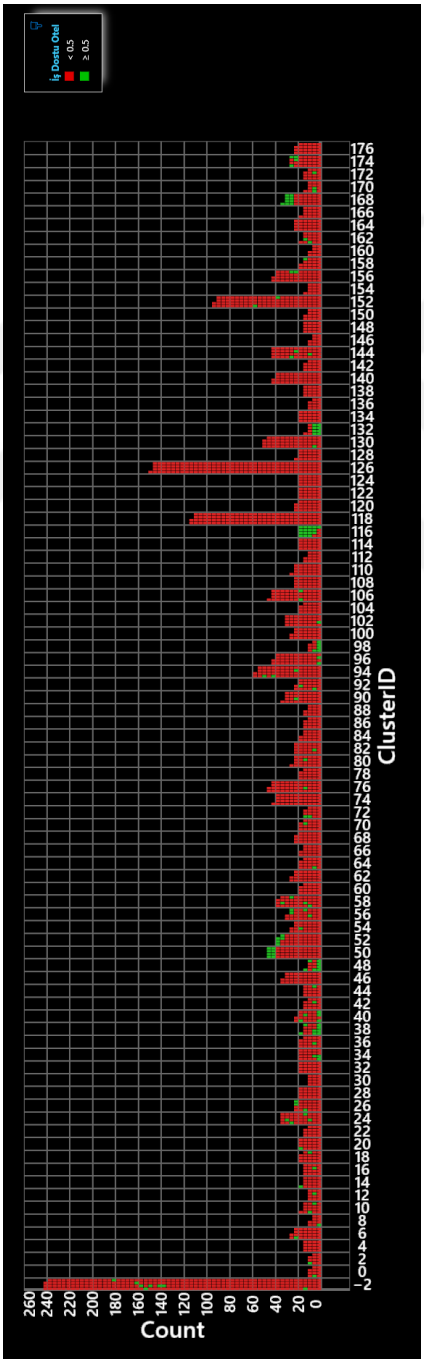


Figure APPENDIX.15 – ClusterID vs. Count on "İş Dostu Otel"

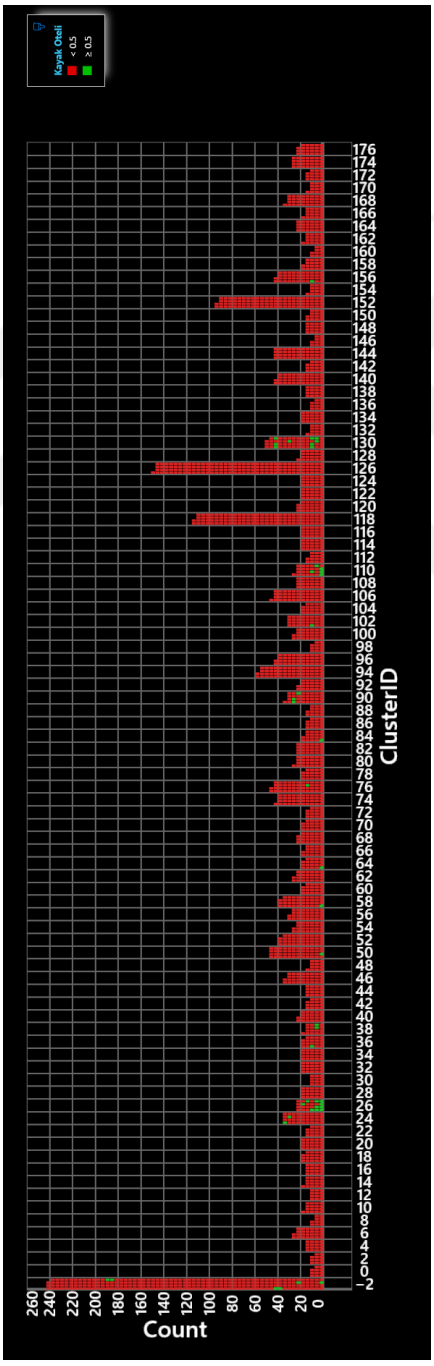


Figure APPENDIX.16 – ClusterID vs. Count on "Kayak Oteli"

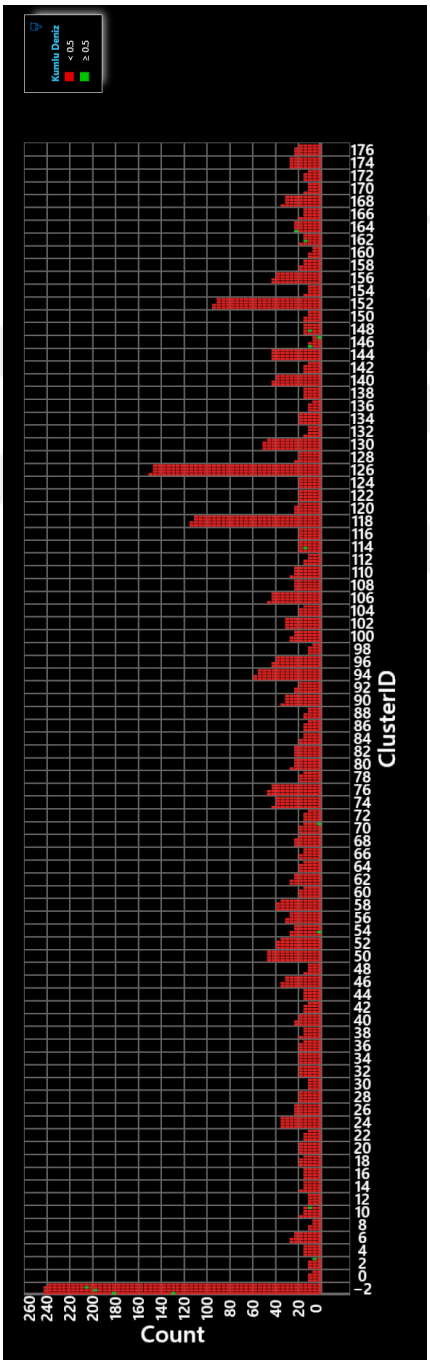


Figure APPENDIX.17 – ClusterID vs. Count on "Kumlu Deniz"

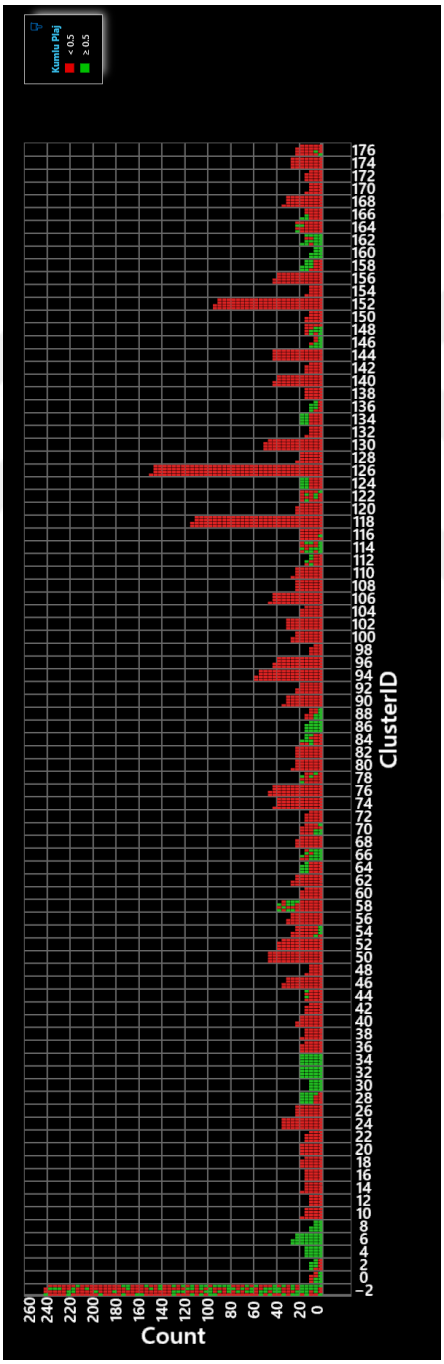


Figure APPENDIX.18 – ClusterID vs. Count on "Kumlu Plaj"

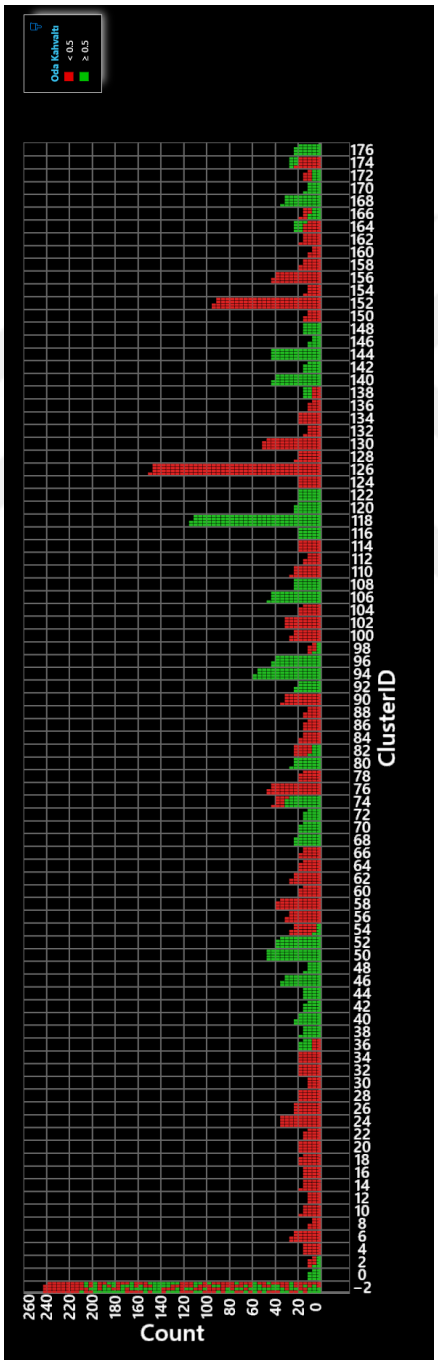


Figure APPENDIX.19 – ClusterID vs. Count on "Oda Kahvaltı"

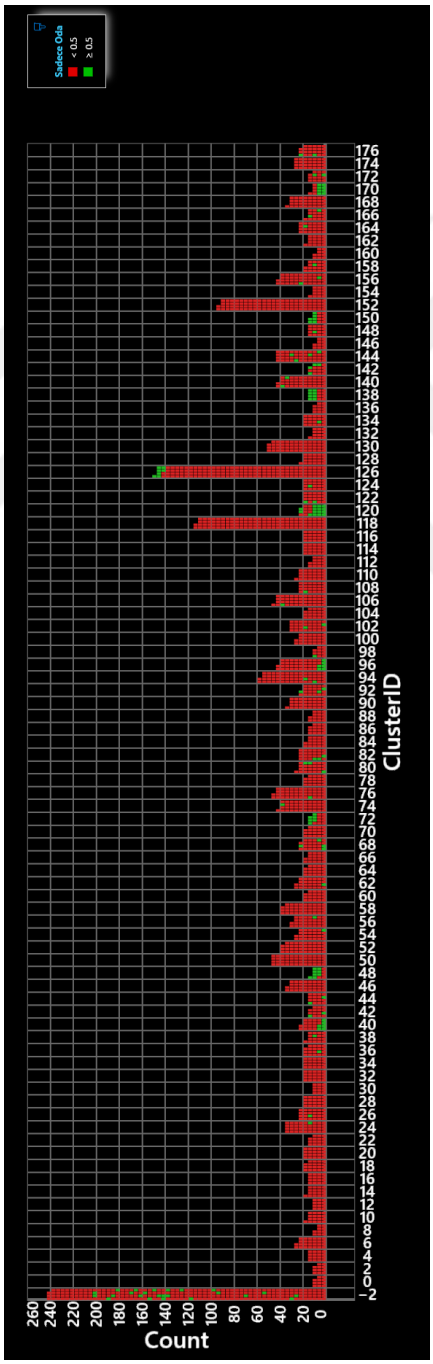


Figure APPENDIX.20 – ClusterID vs. Count on "Sadece Oda"

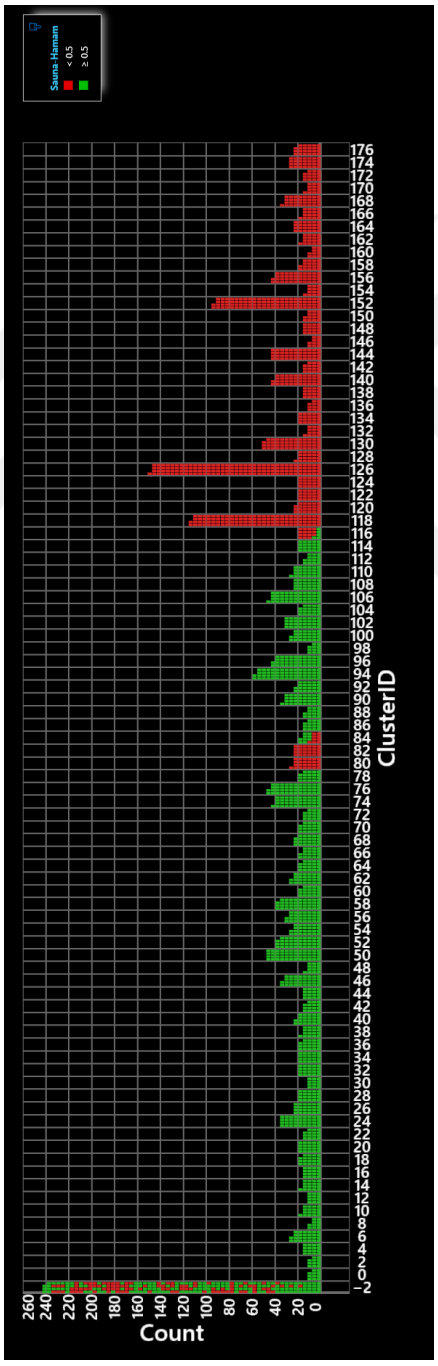


Figure APPENDIX.21 – ClusterID vs. Count on "Sauna Hamam"

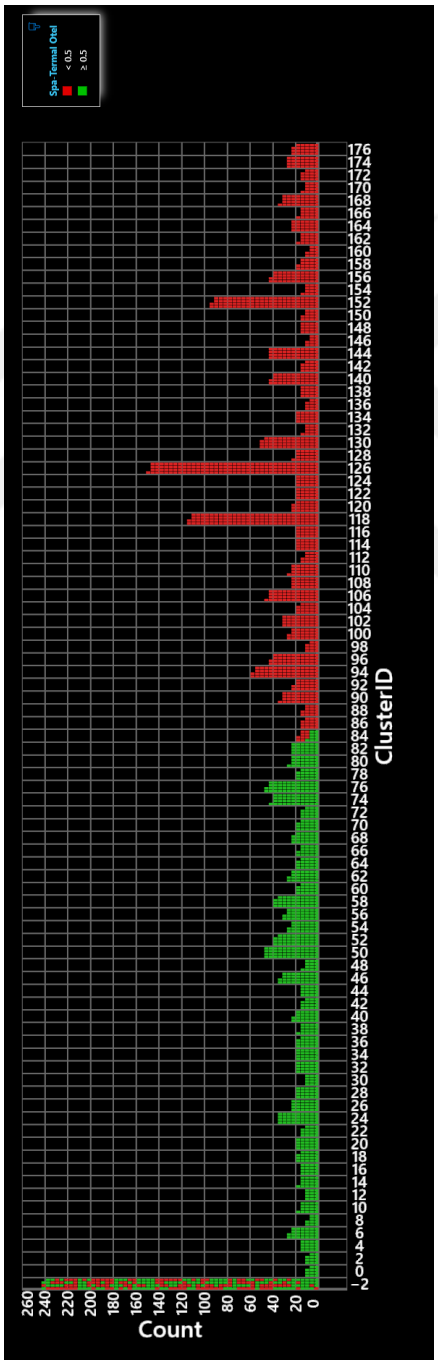


Figure APPENDIX.22 – ClusterID vs. Count on "Spa Termal Otel"

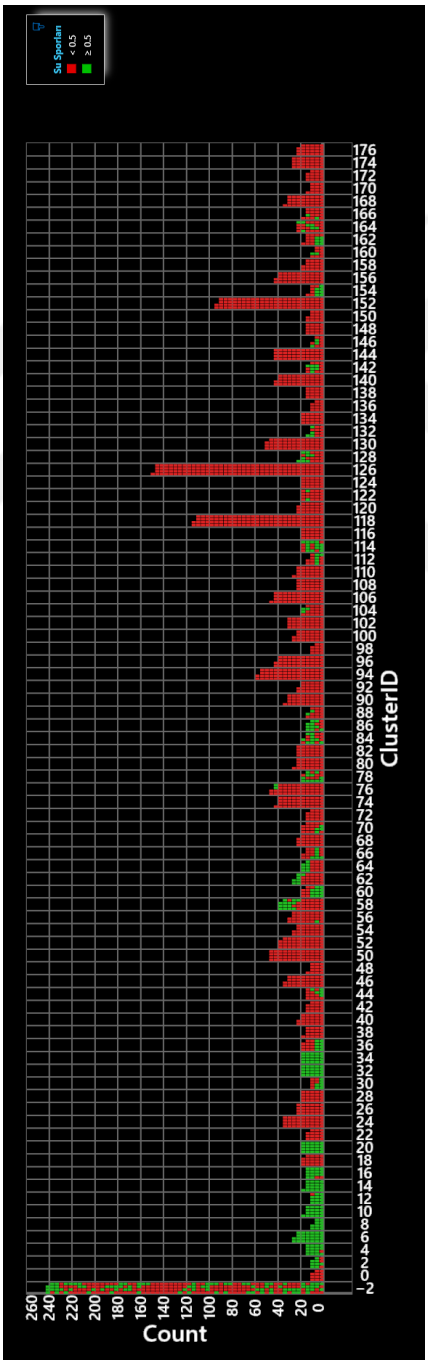


Figure APPENDIX.23 – ClusterID vs. Count on "Su Sporları"

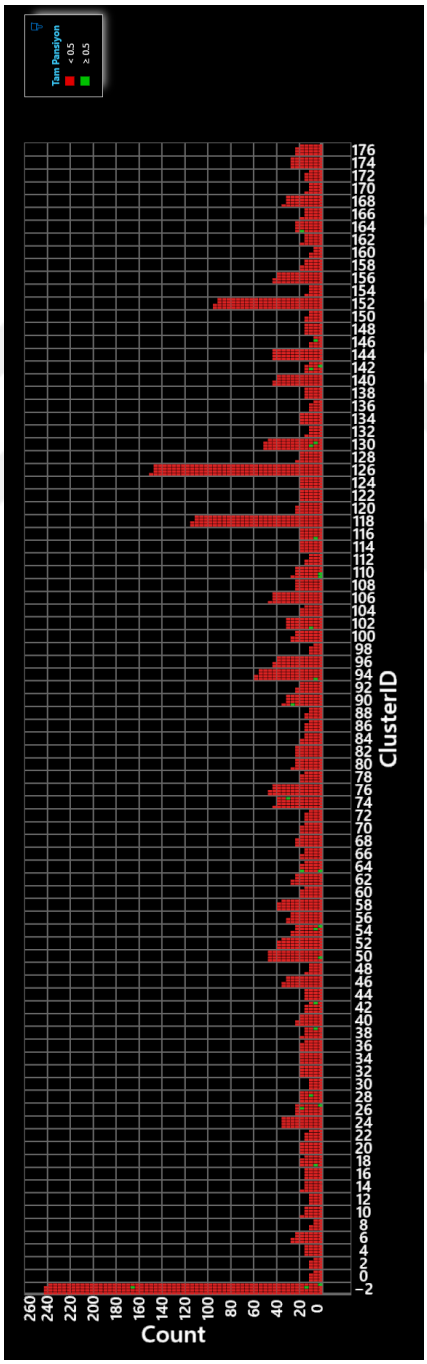


Figure APPENDIX.24 – ClusterID vs. Count on "Tam Pansiyon"

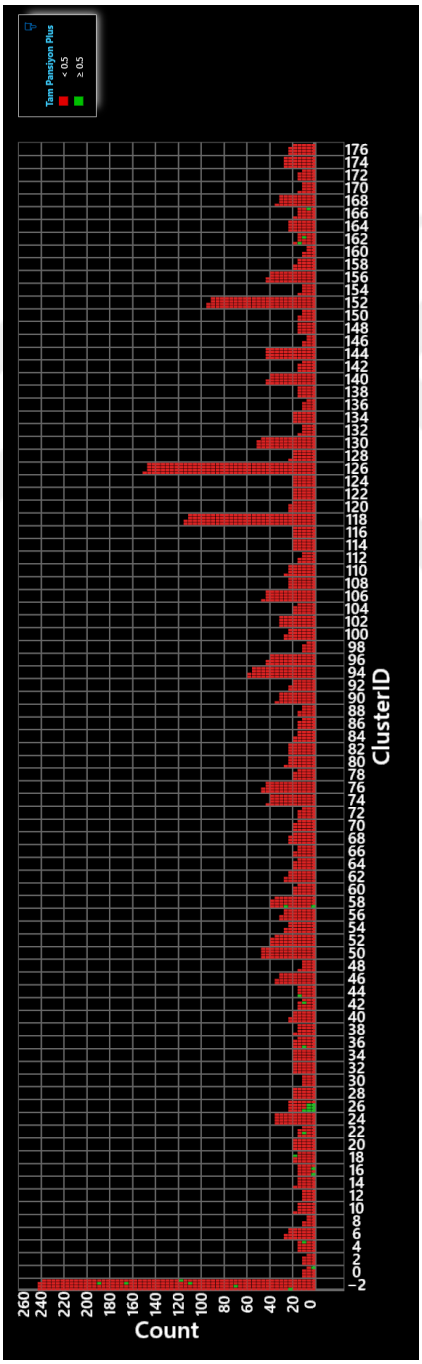


Figure APPENDIX.25 – ClusterID vs. Count on "Tam Pansiyon Plus"

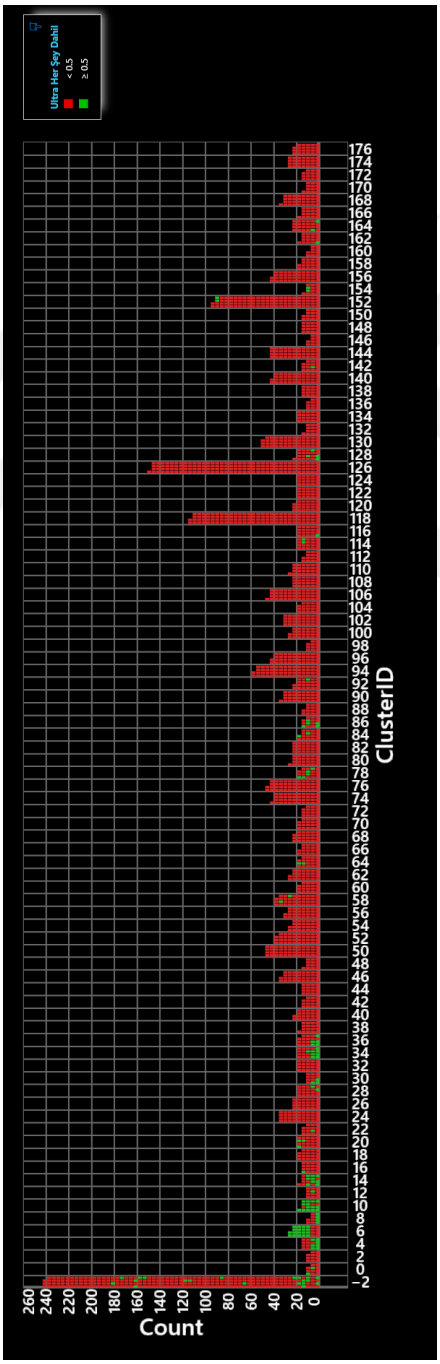


Figure APPENDIX.26 – ClusterID vs. Count on "Ultra Her Şey Dahil"

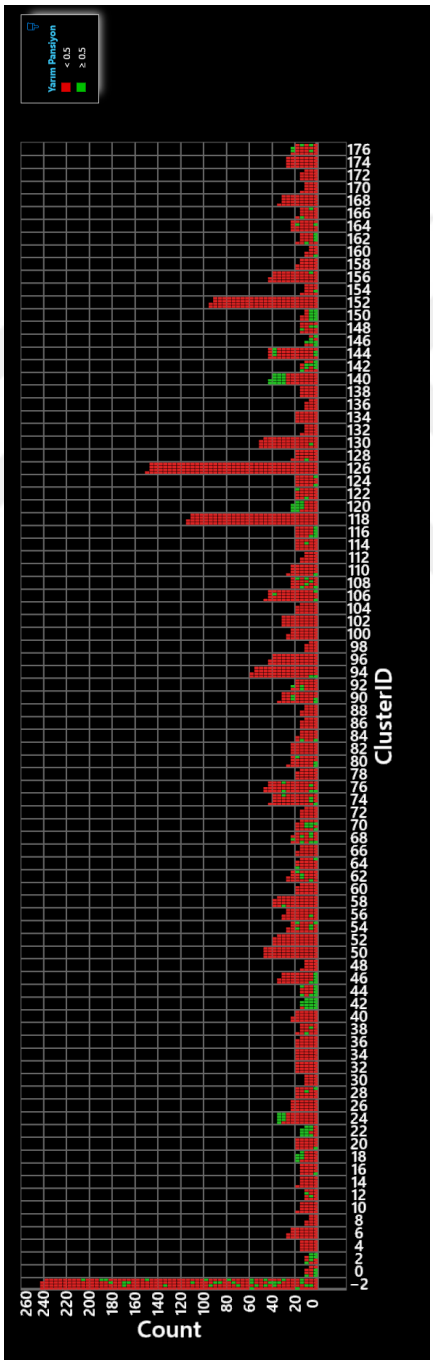


Figure APPENDIX.27 – ClusterID vs. Count on "Yarım Pansiyon"

BIOGRAPHICAL SKETCH

Ömer Arifoğulları has been received bachelor degree in Electrical and Electronics Engineering from the Engineering Faculty, İstanbul University, İstanbul, Turkey in Jan, 2017. He is master student in Smart Systems Engineering at Graduate School of Science and Engineering of Galatasaray University. In his professional life, he is working as database engineer since 2018 in banking and e-commerce sectors respectively. He is interested in data science, data prediction and optimisation for IT infrastructure systems and smart traffic systems.

PUBLICATIONS

— Arifoğulları, Ö. and Orman, G. K. (July, 2023). On Experimental Study of Hotel Clustering *(Approved)*, IMECS 2023