# A HYBRID RECOMMENDATION SYSTEM APPLICATION BY RFM SEGMENTATION

(RFM SEGMENTASYONU ILE BIR HIBRIT ÖNERI SISTEM UYGULAMASI)

by

**B.S. Begüm UYANIK**

**Thesis**

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

**MASTER OF SCIENCE**

**in**

**INTELLIGENT SYSTEMS ENGINEERING**

**in the**

**GRADUATE SCHOOL OF SCIENCE AND ENGINEERING**

**of**

**GALATASARAY UNIVERSITY**

Jun 2023

**ACKNOWLEDGEMENTS**

I would like to convey my genuine appreciation to my thesis advisor, Asst. Prof. Dr. Günce Keziban Orman, for the priceless guidance, encouragement, and support she provided me during the entirety of my research, in particular for her constructive criticism and helpful suggestions during my data analysis. Her insights and feedback played a crucial role in shaping my ideas and improving my methodology.

I am also profoundly thankful to the faculty members of the Department of Engineering and Technology at Galatasaray University for providing me with stimulating academic environment and opportunities for professional development.

I would like to extend my gratitude to my colleagues and friends who supported me during my research journey. Their discussions and encouragement were constant source of my motivation.

Finally, I would like to convey my deep gratitude to my family for their constant support, affection, and understanding. Their sacrifices and belief in me made this achievement possible.

Jun 23
Begüm UYANIK

# TABLE OF CONTENTS

**LIST OF SYMBOLS**

**CBF** **:** Content Based Filtering
**CF** **:** Collaborative Filtering
**App** **:** Appendix
**RFM** **:** Recency, Frequency, Monetary

## LIST OF FIGURES

## LIST OF TABLES

**ABSTRACT**

A recommendation system is a type of information filtering system established to suggest items or content to users based on their preferences and behavior by using data analysis techniques. The system is designed to predict what users might like based on their past interactions with the system or other users like them.

To build a recommendation system, commonly used data analysis techniques are collaborative filtering, content-based filtering, and hybrid filtering.

Collaborative filtering analyzes the past behavior of users and the items they have interacted with to identify patterns and similarities between users.

Content-based filtering uses similar items, such as type or keywords, that the items have their own characteristics. Hybrid filtering combines these two approaches to provide more accurate recommendations.

To implement a recommendation system, data is collected from various sources such as user ratings, user behavior, user demographics, and item characteristics. This data is then analyzed to identify patterns and relationships between users and items, and the results are used to suggest personalized recommendations to users.

Overall, recommendation systems are a powerful tool for businesses to increase user engagement and improve customer satisfaction by providing personalized and relevant recommendations to users.

Many online service providers use a recommendation system to assist their customers' decision-making by generating recommendations. Accordingly, this thesis proposes a new recommendation system for tourism customers to make online reservations for hotels with the features they need, saving customers time and increasing the impact of personalized hotel recommendations. This new system combined collaborative and content-based filtering approaches and created a new hybrid recommendation system.

Two datasets containing customer information and hotel features were analyzed by Recency, Frequency, Monetary (RFM) method in order to identify customers according to their purchasing nature.

The main idea of the recommendation system is to establish correlations between users and products and make the decision to choose the most suitable product or information for a particular user.

As a result of the exponential growth of online data, this vast amount of information for use in the tourism industry can be leveraged by decision-makers to make purchasing decisions.

Filtering, prioritizing, and beneficially presenting relevant information reduces this overload. There are following three main ways that recommendation systems can generate a recommendation list for a user; content-based, collaborative-based, and hybrid approaches. This thesis describes each category and its techniques in detail.

RFM Analysis is used to identify customer segments by measuring customers' purchasing habits. It is the process of labeling customers by determining the Recency, Frequency, and Monetary values of their purchases and ranking them on a scoring model. Scoring is based on how recently they bought (Recency), how often they bought (Frequency), and purchase size (Monetary).

Experimental results show that the accuracy of behavior analysis using Manhattan distance-based hybrid filtering is greatly improved compared to collaborative and content-based algorithms.

## ÖZET

Öneri sistemi, veri analizi teknikleri kullanarak kullanıcıların tercih ve davranışlarına göre öğeler veya içerikler önermek için kurulan bir bilgi filtreleme sistemidir. Sistem, kullanıcıların sisteme veya onlara benzer diğer kullanıcılara geçmiş etkileşimlerine dayanarak kullanıcıların ne gibi şeylerden hoşlanacağını tahmin etmek için tasarlanmıştır.

Öneri sistemi oluşturmak için, genellikle kullanılan veri analizi teknikleri arasında işbirlikçi filtreleme, içerik tabanlı filtreleme ve hibrit filtreleme bulunur.

İşbirlikçi filtreleme, kullanıcıların geçmiş davranışlarını ve etkileşimde bulundukları öğeleri analiz ederek kullanıcılar arasındaki kalıpları ve benzerlikleri belirlemek için kullanılır.

İçerik tabanlı filtreleme, öğelerin kendi özellikleri olan tür veya anahtar kelimeler gibi benzer öğeleri kullanır. Hibrit filtreleme, bu iki yaklaşımı birleştirerek daha doğru öneriler sağlar.

Öneri sistemi uygulamak için, kullanıcı derecelendirmeleri, kullanıcı davranışı, kullanıcı demografisi ve öğe özellikleri gibi çeşitli kaynaklardan veriler toplanır. Bu veriler daha sonra kullanıcılar ve öğeler arasındaki kalıpları ve ilişkileri belirlemek için analiz edilir ve sonuçlar kullanıcılara kişiselleştirilmiş öneriler sunmak için kullanılır.

Genel olarak, öneri sistemleri, kullanıcılara kişiselleştirilmiş ve ilgili öneriler sağlayarak işletmelerin kullanıcı etkileşimini artırmasına ve müşteri memnuniyetini artırmasına yardımcı olan güçlü bir araçtır.

Birçok çevrimiçi hizmet sağlayıcı, tavsiyeler üreterek müşterilerinin karar vermelerine yardımcı olmak için bir öneri sistemi kullanır. Bu tez, turizm müşterilerinin ihtiyaç duydukları özelliklere sahip otellere, çevrimiçi rezervasyon yapmaları için müşterilere zaman kazandıran ve kişiselleştirilmiş otel önerilerinin etkisini artıran yeni bir öneri sistemi önermektedir. Bu çalışmada sunulan yeni sistem, işbirlikçi ve içerik tabanlı filtreleme yaklaşımlarını birleştirmektedir ve yeni bir hibrit öneri sistemi oluşturulmuştur.

Müşteri bilgilerini ve otel özelliklerini içeren iki veri seti, müşterileri satın alma özelliklerine göre tanımlamak için Recency, Frequency, Monetary (RFM) yöntemi ile analiz edilmiştir. Öneri sisteminin ana fikri, kullanıcılar ve ürünler arasında korelasyonlar kurmak ve belirli bir kullanıcı için en uygun ürün veya bilgiyi seçme kararı vermektir.

Çevrimiçi verilerin katlanarak büyümesinin bir sonucu olarak, turizm endüstrisinde kullanılacak bu büyük miktardaki bilgi, karar vericiler tarafından satın alma kararları vermek için kullanılabilir. İlgili bilgileri filtrelemek, önceliklendirmek ve faydalı bir şekilde sunmak bu aşırı yükü azaltır.

Öneri sistemlerinin bir kullanıcı için bir öneri listesi oluşturabilmesinin üç ana yolu vardır; içerik tabanlı, işbirlikçi tabanlı ve hibrit yaklaşımlar. Bu makale, her kategoriyi ve tekniklerini ayrıntılı olarak açıklamaktadır.

RFM Analizi, müşterilerin satın alma alışkanlıklarını ölçerek müşteri segmentlerini belirlemek için kullanılır. Müşterilerin satın alımlarının Recency, Frequency ve Monetary değerlerini belirleyerek bir puanlama modelinde sıralayarak etiketleme işlemidir. Puanlama, ne kadar yakın zamanda satın aldıklarına (Recency), ne sıklıkta satın aldıklarına (Frequency) ve satın alma boyutuna (Monetary) dayalıdır.

Deneysel sonuçlar, Manhattan Distance tabanlı hibrit filtreleme kullanan davranış analizinin doğruluğunun, işbirlikçi ve içerik tabanlı algoritmalara kıyasla büyük ölçüde iyileştirildiğini göstermektedir.

# 1. INTRODUCTION

In the last 15 years, the Internet has evolved from a group work tool to support the work of scientists at CERN, but rather a global knowledge space with more than a billion users. With the spread of the internet, many opportunities have emerged, such as sharing information and ideas with other users (Faulhaber, 1995).

Nevertheless, this time, users encountered a new problem with the internet. The amount of data and units has increased greatly, leading to data overload. Finding out what the user is actually looking for has become a big problem. With the need to filter and sort items and information, developers found recommendation systems as a solution.

Due to the fast-paced growth of e-commerce, online tourism websites have become a prevalent method for booking tourism services (Lv, 2021). Nowadays, people widely use online reservation systems to plan their holidays. A large number of options makes it difficult for hotel customers to decide when and where to go. In addition, due to the wealth of information available in online reservation systems, customers may miss out on a more suitable option for them. In this sense, recommendation systems play a major role in customers' choices (Berker, et al, 2020).

Recommendation systems are useful for service providers and users. They decrease transaction costs for finding and choosing products in an online shopping environment (Vidmar

Recommender systems also have some benefits for businesses.

Firstly, Revenue Algorithm studies using recommendation systems to increase revenue by increasing the number of sales for companies with online customers such as Amazon sites.

Secondly, Personalization - Data collected indirectly can be used to ensure that the services of the website are suitable for the user's preference.

Thirdly, Discovery - giving recommendations to people like shopping, movies, songs, etc. increases the chances of revisiting a web page when they find it (Williams, 1993).

Online reviews of hotels, which can be found on platforms like Booking.com, TripAdvisor, and Expedia, hold the same level of significance for people seeking accommodation as recommendations or suggestions from their friends.

Given the nature of online tourism and user behavior, it is important to first cluster online user behavior in order to identify the factors that influence their choices. By analyzing these factors, exploring user behavior patterns, and predicting future actions, we can provide a technical foundation for different user groups to discover their specific requirements. Hotels should also respond accordingly based on the clustering results.

Recommender systems have the following different types of filtering to create an effective recommendation engine; content-based, collaborative-based, and hybrid (Williams, 1993).

There are many programming languages and tools that can be used for building recommendation systems. Here are some commonly used programming languages and libraries:

**Python:** Python is a popular programming language extensively utilized in the field of data analysis and machine learning. Robust frameworks like scikit-learn, TensorFlow, Keras, and PyTorch offer efficient resources for constructing recommendation systems.

**R:** R is another programming language for data analysis and machine learning. The recommenderlab package provides tools for building recommendation systems in R.

**Java:** Java is a widely used programming language, and libraries such as Mahout and LensKit provide tools for building recommendation systems in Java.

**Scala:** Scala is a programming language frequently employed in constructing recommendation systems and operates on the Java Virtual Machine. Tools like Apache Spark libraries offer the necessary resources for developing recommendation systems in Scala.

**MATLAB:** MATLAB is a programming language commonly used for numerical analysis and data processing. The MATLAB Recommender System Toolbox provides tools for building recommendation systems in MATLAB.

These are just a few examples of the programming languages and libraries that can be used for building recommendation systems. The specific requirements and objectives of the project are taken into account for the selection of languages and tools.

While it is possible to use Excel for basic data analysis tasks, it is not typically used as the primary tool for building recommendation systems.

Excel has some built-in functions that can be used for basic calculations, filtering, and sorting, but it does not have the advanced machine learning algorithms and techniques required for building sophisticated recommendation systems. However, it is possible to use Excel as part of the data analysis process in conjunction with other tools and techniques. For example, Excel can be used for data cleaning and preparation, and the results can be exported to another tool such as Python or R for building the recommendation system.

Data analysis includes a systematic process of examining and interpreting data to extract insights, identify patterns, and make informed decisions. Here are the common steps involved in data analysis:

**Data collection:** This refers to the procedure of acquiring data from various sources like databases, social media platforms, surveys, and websites.

**Data cleaning:** Data cleaning is a crucial stage in the data analysis process that entails identifying and rectifying errors or inconsistencies in the data to enhance its quality and accuracy. The following are steps involved in data cleaning:

**Remove duplicates:** Detecting and eliminating any duplicate entries in the dataset, as they can distort the analysis outcomes.

**Handling missing values:** Recognizing any missing values in the dataset and determining how to manage them. Depending on the extent of missing data, you can opt to either eliminate the records with missing values or employ an appropriate technique such as mean imputation or regression imputation to fill in the missing values.

**Data type conversions:** Ensure that the data types are consistent and appropriate for the analysis. For example, numerical data should be in the correct format, and dates should be converted to a standardized format.

**Outlier detection and handling:** Identify any outliers or extreme values in the dataset and decide how to handle them. Outliers can be removed, or they can be corrected by applying an appropriate transformation.

Handling inconsistent or incorrect data: Identify any inconsistent or incorrect data in the dataset and correct it. For example, inconsistent data can be corrected by applying standardization techniques, or incorrect data can be corrected manually or using automated tools.

**Data normalization:** Normalize the data to ensure that it is on the same scale and to eliminate the effect of different units of measurement.
These are some of the common steps involved in data cleaning. The specific steps will depend on the nature of the data and the analysis requirements. It is important to have a systematic approach to data cleaning to ensure that the data is accurate and reliable for analysis.

**Data exploration:** Data exploration step includes exploring the data to acquire a deeper understanding of its characteristics, such as its distribution, variability, and relationship between variables.

**Data transformation:** Transformation step includes transforming the data to make data suitable for analysis. This step includes normalizing the data, reducing the number of variables, or creating new variables based on existing ones.

**Data modeling:** This involves selecting an appropriate model or technique for analyzing the data. This may include regression analysis, clustering, or machine learning algorithms.

**Data visualization:** Visualization of data includes presenting the results of the analysis in a visual format, such as charts or graphs, to make it easier to understand and interpret. Interpretation: This is the process of conclude and making decisions based on the analysis results.

**Communication:** This involves presenting the results of the analysis to stakeholders in a clear and concise manner to facilitate decision-making.

These are the general steps involved in data analysis. The specific steps may vary depending on the nature of the data and the analysis requirements. It is important to have a systematic approach to data analysis to ensure that the results are accurate and reliable.

In this thesis, real data from SeturTech, a technology company that develops systems that offer hotels and holidays to users, are used. These data are made with a hybrid approach to hotel recommendations for users. The hotel recommendation system has been enhanced in real-life performance by merging two recommendation systems, one utilizing content-based filtering and the other employing collaborative filtering techniques.

This thesis differs from previous studies by utilizing a comprehensive hotel attribute list in the content-based method. Additionally, the collaborative filtering approach involves creating an interaction matrix that resembles the user-product score matrix, incorporating preference amounts of the users. There are different distance metrics to calculate the similarity of customers;

The primary purpose of utilizing the Cosine distance and Cosine Similarity metric is to identify similarities between two data points. As the cosine distance between the data points increases, indicating a larger angle between them, the cosine similarity, or the degree of similarity, decreases, and vice versa. Consequently, data points that are closer to each other are considered more similar than those that are farther apart. The cosine similarity is calculated using the formula Cos $\theta$, while the cosine distance is obtained by subtracting the cosine similarity from 1.

On the other hand, the Manhattan Distance metric, also known as city block distance or taxicab geometry, is employed when calculating the distance between two data points in a grid-like path. When determining customer profiles, the RFM (Recency, Frequency, Monetary) method is used. Recency indicates the time elapsed since the user's last consumption, with a higher score assigned for more recent consumption. The time is measured in days. Frequency reflects how often the user visits within a specific time period. Monetary represents the amount of money a user has spent during a given period. The score is influenced by the transaction amount, with higher amounts receiving a higher score.

To present the details of the work done, the rest of this thesis is structured as follows. In section II, the developed hotel recommendation system and the three types of recommendation system filtering methods are introduced. In section III, RFM analysis of the hotel data is reviewed. In section IV, the results of the tests are given. Finally, we conclude in Section V.

## 2. LITERURE REVIEW

The recommender system is the biggest subfield of data mining and there are two main approaches:

**1) Non-customized**
**2) Customized.**

Each approach has different techniques in machine learning. The non-customized recommender system gives the same item recommendation to all system users, rather than individual user data. Users' interests are not considered.

In contrast, the personalized recommendation system considers the preference or interest of each user, thus recommending certain items to the user more effectively.

There are three basic approaches in the customized recommender system:

**1) Content-based filtering approach:** Characteristics are derived from information items.

**2) Collaborative filtering approach:** Characteristics derived from the user's environment.

**3) Hybrid filtering approach:** Each content-based filtering approach and collaborative filtering approach has its own advantages and disadvantages. In order to address these limitations, a hybrid approach that combines both methods is employed (Poor, 1985).

## 2.1 Content Based Filtering

Content-based filtering method, in other words, "cognitive filtering", works according to user profiles that were created in the beginning. The creation of such profiles is provided by the user creating an account for himself and logging into the system.

A profile contains information about a user and their tastes. To establish user profiles, it is essential for recommendation systems to gather initial information about the users. For this purpose, surveys are prepared and administered.

The recommendation engine compares the items that users have rated positively with each other and determines their similarities (Faulhaber, 1995). The more a user interacts with the system, the stronger the user profile is created. In this content-based filter, only the recorded information of the user is sufficient instead to other similar users (Poor, 1985). For instance, if a user liked a website containing the words "car", "engine" and "petrol", the pages proposed by CBF will be relevant to the automotive world (Bobadilla, et al., 2013).

Content-based recommendation systems are formed of three basic parts in terms of high-level architecture. Firstly, the system preprocesses units with a content analyzer. After that, a professional profile learner gets information concerning the users.

Eventually, the filtering component reveals several suitable suggestions (Seyednezhad, et al, 2018). The three sections mentioned are detailed below:

**Content Analyzer:** The task of the content analyzer is to prepare the available data for the next process and convert the information from its original state into more abstract and feasible. For instance, this converter has the ability to accept a web page as input and convert it to a keyword vector (Seyednezhad, et al., 2018).

**Profile Learner:** Profile learner is a module specially designed for the user. It obtains preprocessed information from the content analyzer and generalizes them to structure the user's preferences (Seyednezhad, et al., 2018).

**Filtering Component:** The final stage of the content-based filtering method involves the filtering component, which suggests relevant items to the user by considering their user profile (Seyednezhad, et al., 2018).

Creating a model for the user's preference from user history is a classification learning style. It is divided into binary categories such as "user likes" and "user dislikes". For example, if a user buys a product, it is a sign that the user likes that product.

However, if the user buys the product and returns it, this is a sign that the user does not like the product. Generally, implicit methods can collect large amounts of data with some uncertainty as to whether the user likes the item (Pazzani, Billsus, 2007).

In this thesis, the features of the hotels' customers visited before are taken as the basis to create the profiles of the users. For example, in Table 2.1- User-Hotel Features Example, user 1 has gone to hotels A and B.

The capacities of the hotels that User 1 visits are 300 and 400 people, respectively, their distance to the sea is 200 and 100 meters, and breakfast is served in these two hotels. Therefore, when calculating User 1's profile, this content information is used and the average values of the relevant rows in Table 2.1 are taken. In this case, the capacity of the hotels where User 1 goes is 350 people on average, and their distance from the sea is 150 meters.

*Table 2.1 User-Hotel Features Example*

| User | Hotel | Capacity | Distance to the Sea | Breakfast |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| 1 | A | 300 | 200 | 1 |
| 1 | B | 400 | 100 | 1 |
| 2 | C | 100 | 600 | 0 |
| 2 | D | 150 | 1250 | 0 |
| 2 | E | 125 | 1500 | 1 |

It also seeks breakfast service at the hotels that User 1 visits. On the other hand, User 2 displayed a different profile by choosing hotels with less capacity and hotels farther from the sea. As a result, the profiles of these users will be as in Table 2.2- User Profiles with CB.

*Table 2.2 User Profiles with CB*

| User | Capacity | Distance to the Sea | Breakfast |
|---|---|---|---|
| 1 | 350 | 150 | 1 |
| 2 | 125 | 1100 | 0.33 |

Using Microsoft Excel, the hotels selected by the customer and the features of the hotels are combined in a single table.

Hotel list data has been cleaned by preprocessing steps. Some hotels selected by customers are not in the hotel feature list, they were detected using excel and removed from the list.

These pre-processing steps consist of correcting missing data (Ex: Hotel names) and removing inconsistent data (Ex: Number of rooms). Currency code and foreign currency sales amounts columns, which will not be used in the setup of the recommendation system, are deleted and a comparison was made in Turkish Lira.

The names of districts and towns were removed, and the cities remained.

Branch name, code, and type; sales, entry, and exit dates, and how many people stayed overnight are removed. Then, the most important 8 features (pool, beach, breakfast, etc.) of the hotels are determined by subtracting the features with less than 600 selections of customers. By reducing the data size, the calculation time for determining the similarity between user profiles and hotels is minimized, resulting in shorter processing times.

The average of the features in the hotels selected by the customers is calculated using the pivot table in excel. Customers with a binary value of 0 for the features of all selected hotels are removed from the thesis and the data set was simplified by removing them from the data set.

### 2.1.1 Content-Based Filtering Method Limitations

**I. New user issue:**

It is a problem caused by a lack of information. When a new user enters the system, the system does not have enough data about the user profile and preferences, so a suitable product profile cannot be created accordingly. As a result, the advice may not be good enough (Poor, 1985).

**II. Excessive specialization:**

It is caused by the recommendation system recommending similar types of products based on available historical data. When the user wants to try a new product, it may cause problems. For example, even if a user will like fiction movies if they are suggested; Since he only liked comedy movies in the past, the system will only suggest new comedy movies to him. Therefore, excessive specialization can narrow the scope of recommendation (Poor, 1985).

## 2.2. Collaborative Filtering

Collaborative filtering is a popular user recommendation approach. The collaborative filtering approach was first described and explained in 1997 by Paul Resnick and Hal Varian (Iskinye, et al., 2015).

If two users have common items with similar ratings, it is assumed that they have similar tastes. Such users form a group or a so-called neighborhood.

A user receives recommendations for items that they have not previously rated but are already rated positively by users in their neighborhood (Sun, Luo, 2010).

In CF systems, a recommendation is made to a user collectively based on past ratings of all users. The Grundy system is the first recommendation system to propose the use of stereotypes as a mechanism to create user models based on limited information about each user.

The system creates the model of the individual user and relevant books are recommended for each user (Poor, 1985). Video Recommender (Hill, et al., 1995) and amazon.com (Chien, George, 1999) are some examples of collaborative filtering.

Collaborative filtering suggestions improve their accuracy as the amount of data on items increases (Shani, Gunawardana, 2011).

Collaborative filtering method approaches can be divided into three subgroups:

### I. User-based approach:

This approach was proposed by University of Minnesota Professor Jonathan L. Herlocker in the late 1990s (Sun, Luo, 2010). In this approach, users take the main role. In this filtering, the subset of users is selected based on their similarity to active users. Customers

who have the same taste construct a group. The user is given suggestions based on the items evaluated by other users in his group (Faulhaber, 1995). The weighted combination of their ratings is then used to estimate the rating for the user (Poor, 1985).

**II. Item-based approach:**

This approach was proposed by University of Minnesota researchers in 2001(Kaya, Kaleli, 2022). As a system grows, the number of users increases and so does the complexity of finding similar users. Therefore, a new approach to item-item collaborative filtering was proposed rather than finding similar users (Poor, 1985). The system creates neighborhoods according to the tastes of the users. The system then generates suggestions with items found in a user's preferred neighborhood (Faulhaber, 1995).

**III. Model-based approach:**

The system makes recommendations for users by estimating the parameters of statistical models for user ratings. In this approach, a pre-calculated model is a design based on available data. This model-based approach quickly responds to the user's preference when the user query appears. Thanks to this approach, the system can be visualized more accurately and can also reduce errors. The most commonly used methods are MF (Matrix factorization), and SVD (Singular value decomposition) (Poor, 1985).

In this thesis, while using the CF method, the number of times a customer went to each hotel is accepted as the score the customer gave to the hotel. In order to simplify the number of hotels, hotels with less than a total of 10 selection by the customers are removed from the dataset.

### 2.2.1 Limitation of Collaborative Filtering

**I. Cold-start problem:**

This problem refers to insufficient information to give the user a recommendation. Collaborative filtering is entirely dependent on the similar neighbor in the system, but these similar neighbors are not present in the system in the first stage known as the cold start issue and they are not known by the system. This reduces the performance of the recommendation system (Iskinye, et al., 2015). This problem can be avoided with the hybrid approach (Poor, 1985).

**II. First-rater problem:**

The system cannot recommend an item that has not been previously rated. As new items are entered into the system, many users did not refer to the items, so there are not enough ratings for these items. The problem can be solved with a hybrid approach.

**III. Sparsity**

The sparsity problem is an important issue. Collaborative recommendation systems often build users' neighborhoods using their profiles. Sparsity occurs when the user does not rank these items (Faulhaber, 1995). If a user has only rated a few items, it is quite difficult to determine their taste, and may be in the wrong neighborhood (Poor, 1985). Sparsity is a problem of lack of information.

**IV. Popularity Bias**

The system cannot recommend products to someone with unique tastes. Sometimes the user has a unique taste compared to all other users on the system. This problem is known as the "popularity bias" issue. This problem can be solved with a hybrid approach (Poor, 1985).

## 2.3 Hybrid Approach

For better results, some recommendation systems use a combination of collaborative and content-based approaches to take advantage of each of them. By using the hybrid approach, the limitations of the content-based and collaborative approaches such as cold-

started problems can be avoided. The combination of these two approaches can be achieved in different ways:

• Applying both methods separately and combining the results.

• Combining some content-based features with a collaborative approach.

• To include some collaborative features in the content-based approach.

Babodilla, et al have classified collaborative filtering and content-based filtering combined into four different groups as in figure 2.1 to make a hybrid method (Seyednezhad, et al., 2018).

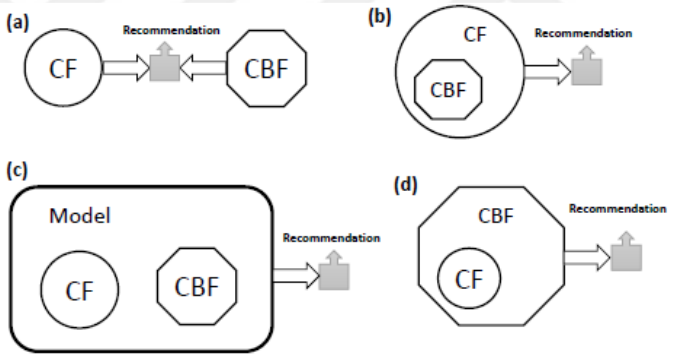


*Figure 2.1 Different methods of Combining CF and CBF*

Figure 2.1. a shows the hybrid method combining CF and CBF with a weighting method.

Figure 2.1. b shows the methods using CBF methods to extract the features and send the recommendation to CF.

Figure 2.1. c shows a combined model using CF and CBF to obtain the outputs of another classifier such as the probability model.

Figure 2.1. d shows a model for CBF using output from CF. For example, user ratings can help CBF better identify users (Seyednezhad, et al., 2018).

Probability methods are used in hybrid filtering. Examples such as genetic algorithms, neural networks, and Bayesian networks can be given (Bobadilla, et al., 2013).

## 3. MATERIALS AND METHODS

The SeturTech data set for recommendations is prepared based on the schema presented in the Figure 3.1-Hybrid approach to SeturTech Data Set. The input set is composed of a user-feature set.

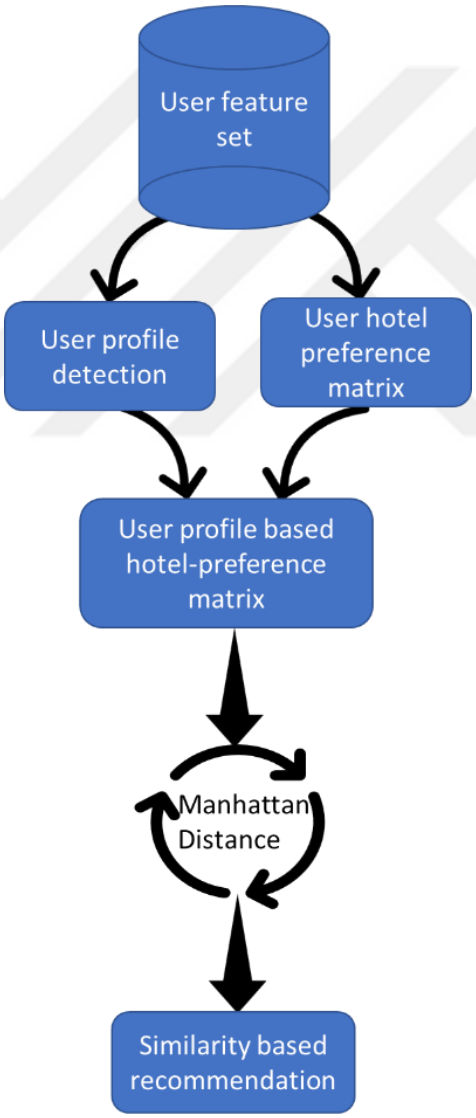


*Figure 3.1 Hybrid approach to SeturTech Data Set*

### 3.1 User Profile Detection

The user profiles have been analyzed based on the GSU Sales dataset and the Hotel Features dataset obtained from SeturTech company.

The following features, which will not affect the results of the analyzes in the sales dataset, are not considered in this thesis;

1) The columns for currency code and sales amounts in foreign currency have been removed, and comparisons have been made in Turkish Lira.

2) The data has been analyzed based on cities instead of districts and towns.

3) The branch's name, code, and type of sale are not used in this thesis, so they have been extracted from the dataset.

4) The sales date, customer check-in and check-out dates, and information about the number of people staying overnight at the hotel are not relevant to classification and RFM Methods and have been excluded from the dataset.

### 3.2 User-Hotel Preferences & User Profile Based Hotel Preferences

To construct the content-based filtering, collaborative filtering, and Manhattan distance-based similarity matrices, two key matrices are needed: the User-Hotel Preferences matrix and the User Profile Based Hotel Preferences matrix. These matrices serve as the foundation for generating the similarity matrices in the respective filtering approaches.

In this section, how hotel preferences based on user profiles are analyzed is explained.

#### 3.2.1 Content Based Filtering

In order to obtain content-based filtering, the average value of the features selected by the customers in the hotels is taken with the pivot table, and the values given by the customers to each hotel feature are assigned as this average value. The brief section of the result matrix is given in Table 3.1- A Brief Example of Result Matrix of Content Based Filtering.

*Table 3.1 A Brief Example of Result Matrix of Content Based Filtering*

| Customer ID | Average-Child-Baby Friendly | Average-Pool | Average-Spa-Thermal Hotel | Average-Sauna-Hamam | Average Fitness |
|---|---|---|---|---|---|
| 13 | 1 | 1 | 1 | 1 | 1 |
| 31 | 1 | 1 | 1 | 1 | 1 |
| 141 | 0 | 0 | 0 | 0 | 0 |
| 166 | 1 | 1 | 1 | 1 | 1 |
| 300 | 1 | 1 | 1 | 1 | 1 |
| 318 | 1 | 1 | 1 | 1 | 1 |
| 329 | 1 | 1 | 1 | 1 | 1 |
| 430 | 0,6667 | 1 | 1 | 1 | 1 |
| 446 | 1 | 1 | 1 | 1 | 0 |
| 458 | 1 | 1 | 1 | 1 | 0 |
| 644 | 1 | 1 | 0,5 | 0,5 | 0 |
| 732 | 1 | 1 | 1 | 1 | 1 |
| 825 | 1 | 1 | 1 | 1 | 0 |
| 982 | 1 | 1 | 1 | 1 | 1 |
| 1032 | 1 | 1 | 1 | 1 | 0,667 |
| 1080 | 0,75 | 0,75 | 0,75 | 0,75 | 0,5 |
| 1090 | 1 | 0 | 0 | 0 | 0 |
| 1994 | 1 | 1 | 1 | 1 | 0 |
| 2007 | 0 | 1 | 1 | 1 | 1 |
| 2021 | 1 | 1 | 0,666666667 | 0,666666667 | 0 |
| 2026 | 1 | 1 | 1 | 1 | 0 |
| 2049 | 1 | 1 | 1 | 1 | 0 |
| 2101 | 1 | 1 | 0,5 | 0,5 | 0,5 |
| 2207 | 1 | 1 | 0 | 0 | 0 |
| 2242 | 0 | 1 | 0 | 1 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 2269 | 1 | 1 | 0,5 | 0,5 | 0 |
| 2277 | 1 | 1 | 0 | 0 | 0 |
| 2299 | 1 | 1 | 1 | 1 | 0 |
| 2385 | 1 | 1 | 1 | 1 | 0 |
| 2403 | 1 | 1 | 1 | 1 | 0 |
| 2434 | 1 | 1 | 1 | 1 | 0 |
| 2489 | 1 | 1 | 1 | 1 | 1 |
| 2554 | 1 | 1 | 0,5 | 1 | 0 |
| 2599 | 1 | 1 | 0 | 0 | 0 |
| 2611 | 1 | 1 | 1 | 1 | 1 |
| 2614 | 0 | 1 | 1 | 0 | 1 |
| 2672 | 1 | 1 | 0,5 | 1 | 0 |
| 2712 | 1 | 1 | 1 | 1 | 0 |
| 1994 | 1 | 1 | 1 | 1 | 0 |
| 2007 | 0 | 1 | 1 | 1 | 1 |

It has been observed that some hotels in our existing dataset have all their features set to "0". The customers for whom all values are 0 have been removed from the dataset as they hinder accurate analysis results and do not contribute to the implementation of Manhattan Similarity. Some of these customers are shown in Table 3. 2-Combination of User-Hotel Matrices and Hotels with No Features in the Dataset.

*Table 3.2 Combination of User-Hotel Matrices and Hotels with No Features in the Dataset*

| Customer ID | Average-Child-Baby Friendly | Average-Pool | Average-Spa-Thermal Hotel | Average-Sauna-Hamam | Average Fitness |
|---|---|---|---|---|---|
| 16193 | 10498 | 0 | 0 | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| 17359 | 2401 | 0 | 0 | 0 | 0 |
| 18820 | 2855 | 0 | 0 | 0 | 0 |
| 20351 | 2855 | 0 | 0 | 0 | 0 |
| 22934 | 1090 | 0 | 0 | 0 | 0 |
| 22934 | 1090 | 0 | 0 | 0 | 0 |
| 23024 | 1057 | 0 | 0 | 0 | 0 |
| 24373 | 1057 | 0 | 0 | 0 | 0 |
| 34582 | 2855 | 0 | 0 | 0 | 0 |
| 34640 | 10498 | 0 | 0 | 0 | 0 |
| 39142 | 1057 | 0 | 0 | 0 | 0 |
| 39558 | 1057 | 0 | 0 | 0 | 0 |
| 39597 | 1057 | 0 | 0 | 0 | 0 |
| 39597 | 1057 | 0 | 0 | 0 | 0 |
| 65885 | 10498 | 0 | 0 | 0 | 0 |
| 74000 | 534 | 0 | 0 | 0 | 0 |
| 74165 | 534 | 0 | 0 | 0 | 0 |
| 74494 | 534 | 0 | 0 | 0 | 0 |
| 77936 | 534 | 0 | 0 | 0 | 0 |
| 83140 | 534 | 0 | 0 | 0 | 0 |
| 86171 | 534 | 0 | 0 | 0 | 0 |
| 87204 | 1090 | 0 | 0 | 0 | 0 |
| 88900 | 534 | 0 | 0 | 0 | 0 |
| 90060 | 10498 | 0 | 0 | 0 | 0 |
| 92313 | 534 | 0 | 0 | 0 | 0 |
| 92991 | 534 | 0 | 0 | 0 | 0 |
| 95499 | 1090 | 0 | 0 | 0 | 0 |
| 96779 | 534 | 0 | 0 | 0 | 0 |

### 3.2.2 Collaborative Filtering

For collaborative filtering, due to the need to calculate the points given by each customer to each hotel, the number of visits by the customer to the hotels is calculated as the score given by that customer to the relevant hotel, and the result matrix is obtained.

Additionally, it is aimed to increase the reliability by removing the customers who are included in the data set and whose features of the selected hotels are not included in the "hotel features" dataset (whole row is 0). It has been identified that there are hotels in the GSU Sales dataset that are not existed in the Hotel Features dataset, which indicates that some hotels visited by customers are not included in the hotel features dataset.

Examples of customers who are not common in both datasets are shown in the table 3.3- Users With No Common Hotel Features in the Dataset And The GSU Sales Dataset.

*Table 3.3 Users With No Common Hotel Features in the Dataset And The GSU Sales Dataset*

| Customer ID(Hotesl Features Dataset) | Customer ID (GSU Sales Dataset) | HotelID |
|---|---|---|
| - | 141 | 9831 |
| - | 191 | 373 |
| - | 217 | 396 |
| - | 755 | 2832 |
| - | 1274 | 2457 |
| - | 1274 | 848 |
| - | 1562 | 414 |
| - | 1855663 | 102503 |
| - | 1855664 | 102503 |
| - | 1855670 | 102503 |
| - | 1855672 | 102503 |

| | | |
|---|---|---|
| - | 1855672 | 102503 |
| - | 1855672 | 102503 |
| - | 1855673 | 102503 |
| - | 1855674 | 102503 |

While applying this collaborative filtering method, the total number of selections of the hotels is calculated with the pivot table and hotels with less than 10 visits by customers are excluded from the matrix in order to obtain clearer and more reliable results.

To separate the hotels that have been chosen less than 10 times from the dataset, all hotels and their total number of selections were calculated as shown in Table 3.4-A Brief Example of the Rankings of Hotel Selection Counts.

*Table 3.4 A Brief Example of the Rankings of Hotel Selection Counts*

| Hotel ID | Total Number of Preferences |
|---|---|
| 370 | 327 |
| 371 | 165 |
| 373 | 649 |
| 377 | 39 |
| 378 | 70 |
| 379 | 81 |
| 381 | 16 |
| 383 | 97 |
| 384 | 68 |
| 385 | 45 |
| 386 | 314 |
| 388 | 11 |
| 389 | 25 |

| | |
|---|---|
| 390 | 189 |
| 391 | 42 |

At the end, a matrix is obtained with customers in the rows and hotels in the columns, and the scores given by the customers to each hotel as seen in Table 3.5-A Brief Example of Result Matrix of Collaborative Filtering.

*Table 3.5 A Brief Example of Result Matrix of Collaborative Filtering*

| | **HOTEL ID SCORES** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Customer ID | 370 | 383 | 384 | 385 | 386 | 419 | 445 | 466 |
| 430 | 2 | | | | | | | |
| 732 | | | | | 1 | | | |
| 821 | | | | | | | | 1 |
| 825 | | | | | | 1 | | |
| 1080 | | 1 | | | | | | |
| 1134 | | | | | | | 1 | |
| 16230 | | | | | | | | 1 |
| 16234 | | | | | | | 1 | |
| 16247 | | | | | | | | 1 |
| 16278 | | | | | 1 | | | |
| 16378 | | | | | | | | 3 |
| 16379 | | | | | | | | 1 |
| 16381 | | | | | | | | |
| 16383 | | | | | | | | 2 |
| 16385 | | | | | 1 | | | |
| 16440 | 1 | | | | | | | |
| 16465 | | | | | | | | 5 |

| | | | | |
|---|---|---|---|---|
| 16467 | | | | 1 |
| 16468 | | | | |
| 16469 | | | | 1 |
| 16516 | 1 | | | |
| 16546 | | | | 1 |
| 16559 | | | | 1 |
| 16583 | | | | 1 |
| 16584 | | | | 10 |

### 3.3 Similarity Based Recommendation

SeturTech dataset includes user properties and user-hotel preferences. First of all, the user profiles are analyzed with RFM in a single data with an engine. A user-hotel preference matrix is created from the same data. Then, a user profile-based hotel preference matrix is created. While using the Manhattan distance method, a similarity-based recommendation system is created. For example, calculating user-2 that is most similar to user-1, the hotel that user-1 did not go to but user-2 went to is recommended to user-1.

Manhattan distance is a metric that computes the distance between two points by summing up the absolute differences between their Cartesian coordinates. Put simply, it is the total sum of the differences between the x-coordinates and the y-coordinates.

$$\text{Manhattan Distance } [\{a, b, c\}, \{x, y, z\}]$$

$$||[a - x]|| + ||[b - y] + ||[c - z]|| \quad (1)$$

Using Manhattan Distance given in equation (1), it is intended to calculate the similarity between customers. The Python Programming Language program is used for this

calculation. Each group of 8 groups is saved as a CSV file and the Manhattan distance is calculated. Since customers whose distance from Manhattan is closer to each other are considered similar to each other, the distance of each customer from Manhattan to the other is obtained. Since customers whose Manhattan distance is "0" will have gone to the same hotels, 0's have been eliminated from the code, and outputs are saved.

### 3.3.1 RFM Analysis of Customer Features

RFM analysis is based on the following assumptions:

1. Customers who have made recent purchases are more likely to purchase again than customers who have not purchased recently.

2. Customers who make more frequent purchases are more likely to repurchase the company's products.

3. Customers with a higher total purchase amount are more likely to purchase again.

For RFM analysis in the SeturTech dataset, each customer's Invoice number, Sales Date, and Sales Amount information is used in this thesis. A customer may have made more than one purchase. Since RFM is about customers, sales amounts are aggregated by customer ID.

The data is grouped with the pivot table feature in the Excel program. With the pivot table, operations such as sorting, summing, and averaging can be performed.
In the pivot table, customers are added to the "rows" column. Thus, each customer appears only once.

For the "Recency" calculation; customers need to know when the last time they visited the site is.

The "maximum" of the data in the “Sales Date" column is selected from the pivot table, and the last time they visited is shown in the column.

For "Frequency", due to the need for information about how often he visits the website, it is calculated how many invoices belong to the same customer to find out how many times the customer has visited the website.

“Monetary” requires information on how much the customer has made a total purchase. The total amount spent by the same customer is calculated with the pivot table.

To calculate RFM scores, from 1 to 5 points are distributed (eg frequency value of a frequent customer should be 5). Recency, frequency, and monetary matrices are created with formulas in Excel. For recency, for example, the customer who visited the website the longest time takes the value 1.

The time elapsed since the last visit is divided into groups according to percentiles. The last visitor was given after the 80% slice and took the value of 5. The customer who spends the least money should get 1 as its monetary value. According to these calculations, the RFM scores of all customers are determined.

Since the values we will calculate with RFM analysis are recency, frequency, and monetary, the values we will use in our 2 separate data sets are as follows:

1) Customer code
2) Billing information
3) Sales date
4) Sales Price

While analyzing these values, the pivot table is used and the customer code is selected as the line, the date of the last sale, how many invoices are issued, and the total sales price are determined.

**Recency Scores:**

A recency matrix is created to find customers' recency scores. The recency matrix is divided into percentiles with the PERCENTILE.INC formula. The customer who visited us the longest time is determined and scored 1, then the matrix is divided into the following percentages, in order: 20, 40, 60, and 80. The recency scores are assigned as in Table 3.6-Recency Matrix.

*Table 3.6 Recency Matrix*

| Recency | Recency Score |
|---|---|
| 1.01.2019 | 1 |
| 12.04.2019 | 2 |
| 18.07.2019 | 3 |
| 18.12.2019 | 4 |
| 30.07.2020 | 5 |

According to the created recency matrix (Table 3.6-Recency Matrix), R- Scores are assigned to customers by the VLOOKUP formula.

**Frequency Scores:**

While assigning frequency values to customers, the PERCENTILE.INC formula is used. The customer who visits the company at least should get 1 as the frequency value. After

calculating the customer with the least visits to the company with the pivot table, the most recent visit dates are divided into the following percentages, in order: 20, 40, 60, 80, 90, and 95. The frequency matrix is given in Table 3.7-Frequency Matrix.

*Table 3.7 Frequency Matrix*

| Frequency | Frequency Score |
|---|---|
| 1 | 3 |
| 2 | 4 |
| 3 | 5 |

As seen in Table 3.7-Frequency Matrix, the frequency value of the customer is 1 up to the 80% slice. This value means that at least 80% of customers shopped from the company only once.

According to the created frequency matrix (Table 3.7-Frequency Matrix), F- Scores are assigned to customers by the VLOOKUP formula.

**Monetary Scores:**

According to the total sales price amount created with the pivot table in the RFM table, the monetary matrix is created with the PERCENTILE.INC formula and monetary scores are assigned. The results are given in Table 3.8 Monetary Matrix.

*Table 3.8 Monetary Matrix*

| Monetary | Monetary Score |
|---|---|
| 0 | 1 |
| 5572800 | 2 |
| 11696000 | 3 |

| | |
|---|---|
| 18920000 | 4 |
| 30100000 | 5 |

# 4. RESULTS

In this section, the outputs obtained after performing the above-mentioned analyzes are shared.

## 4.1 Datasets and Descriptive Analysis

This thesis utilizes two datasets obtained from SeturTech company. One of them is the "Hotel Features" data set which includes which of the 27 features (rows of the first data set) such as bed & breakfast, spa-thermal hotel, casino hotel, all-inclusive, half board, pool, sandy beach of 2562 different hotels meet.

Another is a "GSU Sales Data" data set containing 40600 sales data (rows of data set) of a company and 32 sales information (columns of the data set) such as Hotel ID, the code of the customer from whom the sale was made, the gender of the customer, the date of sale, the name of the branch where the sale was made.

Hotel features and sales data features that are not needed for this thesis are cleared from these two data sets.

In these two data sets, only the data to be used in this thesis is obtained, the sales data are sorted on the basis of the customers' code, and the hotel features data are kept to be integrated with the customers' choices.

After determining the hotels selected by the customer codes, the value of 1 or 0 assigned to the hotel selected by the customer from the data set with the characteristics of the hotels positioned with "=VLOOKUP()" in Microsoft Excel Program.

It is determined by the IFERROR formula that some of the hotels selected by the customers are not in the hotel feature list and are deleted from the list.

### 4.1.1 Hotel Features Dataset

The features of the hotels are listed below:

• Child-Baby Friendly

• Ski Hotel

• Pool

• Pool-Summer

• Sandy Beach

• Sandy Sea

• Next to Sea Shore

• Pebble Beach

• Pebble Sea

• Other Beach

• Business Friendly Hotel

• Casino Hotel

• Honeymoon Hotel

• Aquapark

• Golf Hotel

• Spa-Thermal Hotel

• Sauna-Turkish Bath

• Water Sports

• Fitness

• All Inclusive (Without Alchohol)

• All-Inclusive

• Room-Breakfast

• Room Only

• Full Board

• Full Board Plus

• Ultra All-Inclusive

• Half Board

In Table 4.1-A Brief Example from the Hotel Features Data Set, a partial representation of the data from the hotel features dataset is shown.

*Table 4.1 A Brief Example from the Hotel Features Data Set*

| HotelID | Child-Baby Friendly | Pool | Pool-Summer | Beach | Spa-Termal Otel |
|---|---|---|---|---|---|
| 370 | 1 | 1 | 1 | 1 | 1 |
| 371 | 1 | 1 | 1 | 1 | 1 |
| 374 | 1 | 1 | 1 | 1 | 0 |
| 377 | 1 | 1 | 1 | 1 | 1 |
| 378 | 1 | 1 | 1 | 1 | 1 |
| 379 | 1 | 1 | 1 | 0 | 1 |
| 380 | 1 | 1 | 1 | 0 | 1 |
| 381 | 1 | 1 | 1 | 1 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| 383 | 1 | 1 | 1 | 1 | 1 |
| 384 | 0 | 1 | 1 | 1 | 1 |
| 385 | 1 | 1 | 0 | 0 | 1 |
| 386 | 1 | 1 | 1 | 1 | 1 |
| 388 | 1 | 1 | 1 | 1 | 1 |
| 389 | 1 | 1 | 1 | 0 | 1 |
| 390 | 1 | 1 | 0 | 0 | 1 |
| 391 | 1 | 1 | 1 | 1 | 0 |
| 393 | 1 | 1 | 1 | 1 | 1 |
| 395 | 0 | 1 | 1 | 1 | 0 |
| 396 | 0 | 0 | 0 | 0 | 1 |
| 397 | 1 | 1 | 1 | 0 | 1 |
| 398 | 1 | 1 | 1 | 1 | 0 |
| 401 | 1 | 1 | 1 | 1 | 0 |
| 402 | 1 | 1 | 1 | 0 | 1 |
| 403 | 1 | 1 | 1 | 1 | 1 |
| 404 | 1 | 1 | 1 | 1 | 1 |
| 405 | 1 | 1 | 1 | 0 | 1 |
| 407 | 1 | 1 | 1 | 1 | 1 |
| 411 | 1 | 1 | 1 | 1 | 0 |
| 412 | 1 | 1 | 1 | 1 | 1 |
| 413 | 1 | 1 | 1 | 1 | 1 |
| 415 | 1 | 1 | 1 | 1 | 0 |
| 417 | 0 | 1 | 1 | 0 | 1 |
| 418 | 1 | 1 | 1 | 1 | 1 |
| 419 | 1 | 1 | 1 | 0 | 1 |
| 420 | 1 | 1 | 0 | 0 | 1 |
| 421 | 1 | 1 | 1 | 0 | 1 |
| 422 | 1 | 1 | 1 | 0 | 1 |
| 428 | 1 | 1 | 1 | 1 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 430 | 0 | 0 | 0 | 0 | 1 |
| 433 | 1 | 1 | 1 | 1 | 1 |
| 434 | 1 | 1 | 1 | 1 | 0 |
| 436 | 1 | 1 | 1 | 1 | 1 |
| 438 | 1 | 1 | 1 | 1 | 1 |
| 439 | 1 | 1 | 1 | 1 | 1 |
| 440 | 0 | 1 | 1 | 0 | 0 |
| 441 | 1 | 1 | 1 | 1 | 1 |
| 442 | 1 | 1 | 1 | 1 | 1 |
| 443 | 1 | 1 | 0 | 0 | 0 |
| 444 | 1 | 1 | 1 | 0 | 1 |
| 445 | 1 | 1 | 1 | 1 | 1 |
| 448 | 1 | 1 | 1 | 0 | 0 |
| 449 | 1 | 1 | 1 | 1 | 1 |
| 471 | 1 | 1 | 1 | 1 | 1 |
| 472 | 1 | 1 | 1 | 0 | 1 |
| 473 | 1 | 1 | 1 | 1 | 1 |
| 428 | 1 | 1 | 1 | 1 | 1 |
| 430 | 0 | 0 | 0 | 0 | 1 |
| 433 | 1 | 1 | 1 | 1 | 1 |
| 434 | 1 | 1 | 1 | 1 | 0 |

There are 27 binary features in this data set.

Explanations of this data are given in the following items.

- Out of 2562 data, 1230 of them are Child-Baby friendly, while 1332 of them are not.
- Out of 2562 data, 10 of them are Golf Hotel, while 2552 of them are not.

- Out of 2562 data, 1071 of them are Spa-Thermal Hotel, while 1491 of them are not.
- Out of 2562 data, 1099 of them are Fitness Hotel, while 1463 of them are not.
- Out of 2562 data, 180 of them are All Inclusive, while 2382 of them are not.
- Out of 2562 data, 1449 of them have Sauna-Turkish Bath, while 1113 of them are not.
- Out of 2562 data, 130 of them are Ultra All Inclusive, while 2432 of them are not.
- Out of 2562 data, 1490 of them have Room-Breakfast, while 1072 of them are not.
- Out of 2562 data, 27 of them are Full Board Plus, while 2535 of them are not.
- Out of 2562 data, 438 of them have Water Sports, while 2124 of them are not.
- Out of 2562 data, 152 of them are Room Only, while 2410 of them are not.
- Out of 2562 data, 27 of them are Full Pension, while 2535 of them are not.
- Out of 2562 data, 240 of them are Half Board, while 2322 of them are not.
- Out of 2562 data, 10 of them are Casino Hotel, while 2552 of them are not.
- Out of 2562 data, 148 of them are Business Friendly Hotel, while 1414 of them are not.
- Out of 2562 data, 609 of them have Other Beach, while 1953 of them are not.
- Out of 2562 data, 14 of them have Sandy Sea, while 2548 of them are not.
- Out of 2562 data, 2517 of them are Ski Hotel, while 45 of them are not.
- Out of 2562 data, 2425 of them have Pebble Beach, while 137 of them are not.
- Out of 2562 data, 1890 of them have pool, while 672 of them are not.
- Out of 2562 data, 389 of them have Sandy Beach, while 2173 of them are not.
- Out of 2562 data, 416 of them are Next to Sea Shore, while 2548 of them are not.
- Out of 2562 data, 956 of them have Summer Pool, while 1606 of them are not.
- Out of 2562 data, one of them has Pebble Sea, while 2561 of them are not.
- Out of 2562 data, 357 of them have Aquapark, while 2205 of them are not.
- Out of 2562 data, 83 of them are Honeymoon Hotel, while 2479 of them are not.

- Out of 2562 data, 25 of them are All Inclusive (Without Alchohol), while 2537 of them are not.

The 19 features, which contain relatively little data in the hotel features dataset, are not considered in this thesis, and the following 8 features are used:

1) Child-Baby Friendly
2) Pool
3) Pool-Summer
4) Other Beach
5) Spa-Thermal Hotel
6) Sauna-Turkish Bath
7) Fitness
8) Room Breakfast

In Figures 4.1-4.27, the distributions of all features are shown within a pie chart. The distributions of these features within the dataset are shown in the following pie chart graphics.

### 4.1.2 GSU Sales Dataset

In the GSU Sales dataset, there are 40,600 sales records. When the purchases made by the same customer are consolidated, it is observed that there are 18,685 unique customers.

The GSU Sales dataset has been exclusively used for RFM (Recency, Frequency, Monetary) analysis and will be discussed in detail in the next section.

In Table 4.2- A Brief Example from the Sales Data Set, a partial example of the data from the GSU Sales dataset is shown.

*Table 4.2 A Brief Example from the Sales Data Set*

| Hotel ID | City Name | Customer ID | Tkl Master ID | Number of Rooms | AdultCount |
|---|---|---|---|---|---|
| 1816 | Antalya | 40288 | 49821 | 1 | 2 |
| 1907 | Bafra | 35463 | 38235 | 1 | 2 |
| 1580 | Muğla | 35181 | 38123 | 1 | 2 |
| 1248 | Antalya | 35493 | 38249 | 1 | 2 |
| 378 | İzmir | 31575 | 28333 | 1 | 2 |
| 2143 | Eskişehir | 29495 | 24384 | 1 | 2 |
| 9704 | Muğla | 35709 | 38351 | 1 | 2 |
| 739 | Bolu | 9629 | 51791 | 1 | 2 |
| 380 | Muğla | 111383 | 97093 | 1 | 2 |
| 646 | Antalya | 143305 | 69709 | 1 | 2 |
| 9704 | Muğla | 131596 | 342077 | 1 | 2 |
| 1648 | Antalya | 19021 | 238994 | 1 | 3 |
| 2750 | Gaziantep | 19023 | 169941 | 1 | 3 |
| 9595 | Antalya | 3598 | 79952 | 1 | 2 |
| 512 | Antalya | 19425 | 280154 | 1 | 2 |
| 497 | Girne | 97058 | 89594 | 1 | 2 |
| 786 | Antalya | 25961 | 324861 | 1 | 2 |
| 786 | Antalya | 25965 | 324863 | 1 | 2 |
| 786 | Antalya | 34867 | 38008 | 1 | 1 |
| 786 | Antalya | 18130 | 51575 | 1 | 2 |
| 786 | Antalya | 18130 | 51575 | 1 | 2 |
| 786 | Antalya | 32289 | 28596 | 1 | 2 |
| 661 | Antalya | 32300 | 142346 | 1 | 2 |
| 786 | Antalya | 25994 | 324875 | 1 | 2 |
| 2533 | İzmir | 35771 | 38385 | 1 | 1 |
| 3094 | Adana | 25338 | 321722 | 1 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 904 | Antalya | 33641 | 34527 | 1 | 2 |
| 786 | Antalya | 35068 | 38083 | 1 | 2 |
| 3244 | Muğla | 35858 | 38437 | 1 | 2 |
| 443 | Bolu | 26188 | 324937 | 1 | 3 |
| 1248 | Antalya | 18666 | 278717 | 1 | 2 |
| 9704 | Muğla | 34939 | 38027 | 1 | 3 |
| 1137 | Antalya | 35009 | 38059 | 1 | 2 |
| 1405 | Antalya | 31394 | 28261 | 1 | 3 |
| 1405 | Antalya | 31410 | 28269 | 1 | 4 |
| 644 | Muğla | 132379 | 304471 | 1 | 2 |
| 1907 | Bafra | 67713 | 57876 | 1 | 2 |
| 769 | Antalya | 136871 | 329396 | 1 | 3 |
| 1539 | Ankara | 18370 | 190875 | 1 | 2 |
| 1182 | Antalya | 148791 | 154082 | 1 | 2 |
| 1907 | Bafra | 38822 | 194534 | 1 | 2 |
| 1248 | Antalya | 38833 | 45485 | 1 | 2 |
| 2075 | Ankara | 16925 | 235278 | 1 | 2 |
| 905 | Antalya | 31273 | 28225 | 1 | 2 |

## 4.2 Data Analysis and Distribution Results

In this thesis, the RFM method has been predominantly used in determining customer profiles during data analysis.

RFM (Recency, Frequency, Monetary) analysis is a data analysis technique commonly used in marketing and customer segmentation. It helps businesses understand and categorize customers based on their purchasing behavior.

By analyzing the RFM components, customers can be segmented into different groups, such as:

**High-Value Customers:** Those who have made recent purchases, frequent purchases, and have spent a significant amount of money.

**Potential Loyal Customers:** Those who have made recent purchases but haven't spent a significant amount of money yet.

**Churning Customers:** Those who haven't made purchases recently and have low frequency and monetary value.

RFM analysis helps businesses identify their most valuable customers, develop targeted marketing strategies, improve customer retention, and maximize profitability.

In the methodology section (Chapter 3), a detailed explanation of how the RFM scores of the examined users were calculated is provided in this thesis.

According to the created recency matrix (Table 3.6), frequency matrix (Table 3.7), and monetary matrix (Table 3.8), M- Scores are assigned to customers by the Pyhton Programming Language.

Examples of the result of the R-Score, F-Score, and M- Scores processed into the RFM matrix are shown in table 4.3-Recency, Frequency, and Monetary Scores of Some Customers.

*Table 4.3 Recency, Frequency, and Monetary Scores of Some Customers*

| Customer ID | The Last Sale Date | Number of Master ID | Total Selling Price | R Score | F Score | M Score |
|---|---|---|---|---|---|---|
| 13 | 15.07.2020 | 2 | 84,753.00 | 4 | 4 | 5 |
| 31 | 17.07.2020 | 3 | 19,541.30 | 4 | 5 | 4 |
| 166 | 19.12.2019 | 2 | 314,717.00 | 4 | 4 | 5 |
| 171 | 29.09.2020 | 3 | 38,562.10 | 5 | 5 | 5 |
| 1080 | 30.07.2020 | 4 | 40290700 | 5 | 5 | 5 |
| 1285 | 30.12.2019 | 2 | 86103200 | 4 | 4 | 5 |
| 2021 | 14.02.2020 | 3 | 44160800 | 4 | 5 | 5 |
| 2242 | 25.08.2020 | 2 | 36397300 | 5 | 4 | 5 |
| 2385 | 19.06.2020 | 2 | 39477200 | 4 | 4 | 5 |
| 2554 | 22.10.2020 | 2 | 61718600 | 5 | 4 | 5 |
| 2611 | 9.07.2020 | 2 | 62356900 | 4 | 4 | 5 |
| 2672 | 7.12.2020 | 2 | 42551100 | 5 | 4 | 5 |
| 2887 | 28.07.2020 | 2 | 34506600 | 4 | 4 | 5 |
| 2933 | 16.07.2020 | 5 | 33505900 | 4 | 5 | 5 |
| 2939 | 4.09.2020 | 2 | 34276600 | 5 | 4 | 5 |
| 3731 | 7.02.2020 | 2 | 36735800 | 4 | 4 | 5 |
| 3784 | 23.12.2019 | 3 | 44654600 | 4 | 5 | 5 |
| 4066 | 24.07.2020 | 2 | 66496100 | 4 | 4 | 5 |
| 4081 | 11.04.2020 | 2 | 56413400 | 4 | 4 | 5 |
| 4249 | 25.09.2020 | 12 | 47454200 | 5 | 5 | 5 |
| 4320 | 13.07.2020 | 3 | 72937900 | 4 | 5 | 5 |
| 4998 | 5.06.2020 | 4 | 72771500 | 4 | 5 | 5 |
| 5108 | 10.09.2020 | 2 | 41505800 | 5 | 4 | 5 |
| 5333 | 18.11.2020 | 2 | 24148800 | 5 | 4 | 4 |
| 5870 | 23.10.2020 | 2 | 25348500 | 5 | 4 | 4 |
| 6666 | 28.09.2020 | 4 | 77488300 | 5 | 5 | 5 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 6736 | 21.10.2020 | 3 | 30736400 | 5 | 5 | 5 |
| 6758 | 20.07.2020 | 2 | 58137400 | 4 | 4 | 5 |
| 6929 | 6.08.2020 | 3 | 30879200 | 5 | 5 | 5 |
| 7204 | 23.12.2019 | 2 | 58678600 | 4 | 4 | 5 |
| 7427 | 23.10.2020 | 11 | 27670800 | 5 | 5 | 4 |
| 217 | 8.10.2019 | 1 | 1654400 | 3 | 3 | 1 |
| 294 | 24.07.2019 | 1 | 6579000 | 3 | 3 | 2 |
| 296 | 3.05.2019 | 1 | 16522500 | 2 | 3 | 3 |
| 318 | 5.08.2019 | 1 | 17028000 | 3 | 3 | 3 |
| 430 | 9.09.2019 | 3 | 21954400 | 3 | 5 | 4 |
| 446 | 7.06.2019 | 1 | 26993300 | 2 | 3 | 4 |
| 458 | 2.07.2019 | 1 | 12829100 | 2 | 3 | 3 |
| 644 | 5.11.2019 | 2 | 12792500 | 3 | 4 | 3 |
| 755 | 22.11.2019 | 1 | 4042000 | 3 | 3 | 1 |
| 825 | 26.08.2019 | 1 | 13996500 | 3 | 3 | 3 |
| 982 | 27.05.2019 | 1 | 25877400 | 2 | 3 | 4 |
| 1032 | 29.06.2019 | 3 | 45584800 | 2 | 5 | 5 |
| 1090 | 22.05.2019 | 1 | 2409200 | 2 | 3 | 1 |
| 1154 | 11.07.2019 | 1 | 39027000 | 2 | 3 | 5 |
| 1168 | 13.06.2019 | 1 | 33260500 | 2 | 3 | 5 |

### 4.1.1 Customer Segmentation

To achieve its business goals, a company can use customer segmentation to target its marketing efforts and resources to valuable and loyal customers (Alsayat, 2022).

To prepare a recommendation system using collaborative filtering, customers are divided into specific groups. Since applying this system to those in the lost customer class would not yield efficient results, it is needed to determine focus groups. For this purpose, primarily customer profiles are determined. These focus groups are selected according to

the RFM scores assigned by RFM analysis, using frequency, monetary, and recency scores. Thus, 6 different customer groups emerged.

Customers in SeturTech data are divided into the following six different segmentations:

1) Loyal
2) Potential Loyal
3) Promising
4) Hesitant
5) Need Attention
6) Detractors

This customer segmentation is determined according to RFM scores given in Table 4.4-RFM Scores of Customer Profiles.

*Table 4.4 RFM Scores of Customer Profiles*

| **Customer Profile** | **RFM SCORES** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Loyalists** | 555 | 545 | 544 | 444 | 454 | 455 | 445 | 554 | |
| **Potential Lotalists** | 332 | 434<br>333 | 443<br>342 | 344<br>343 | 442<br>334 | 244<br>412 | 424<br>413 | 441<br>414 | 431 |
| | 421 | 422 | 423 | 424 | 433 | 441 | 432 | | |
| **Promising** | 233 | 234 | 241 | 311 | 312 | 313 | 314 | 321 | 322 |
| | 331 | 341 | 324 | 323 | | | | | |
| **Hesitant** | 124 | 133 | 134 | 142 | 143 | 144 | 214 | 224 | 234 |
| | 232 | 243 | 242 | | | | | | |
| **Need Attention** | 122 | 123 | 131 | 132 | 141 | 212 | 213 | 221 | 222 |
| | 223 | 231 | | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Detractors** | 111 | 112 | 113 | 114 | 121 | 131 | 211 | 311 | 411 |

According to the sales data dataset of GSU, RFM analyses have been conducted and RFM scores have been provided. Customers who underwent RFM analysis were categorized into 6 groups: Loyalists, Potential Loyalists, Promising, Hesitant, Need Attention, and Detractors. Among them, the loyalist customers with 8 different RFM scores have been identified as the focus group.

The distribution of these RFM groups across the entire dataset is shown in the histogram graph in Figures 4.28-4.32.

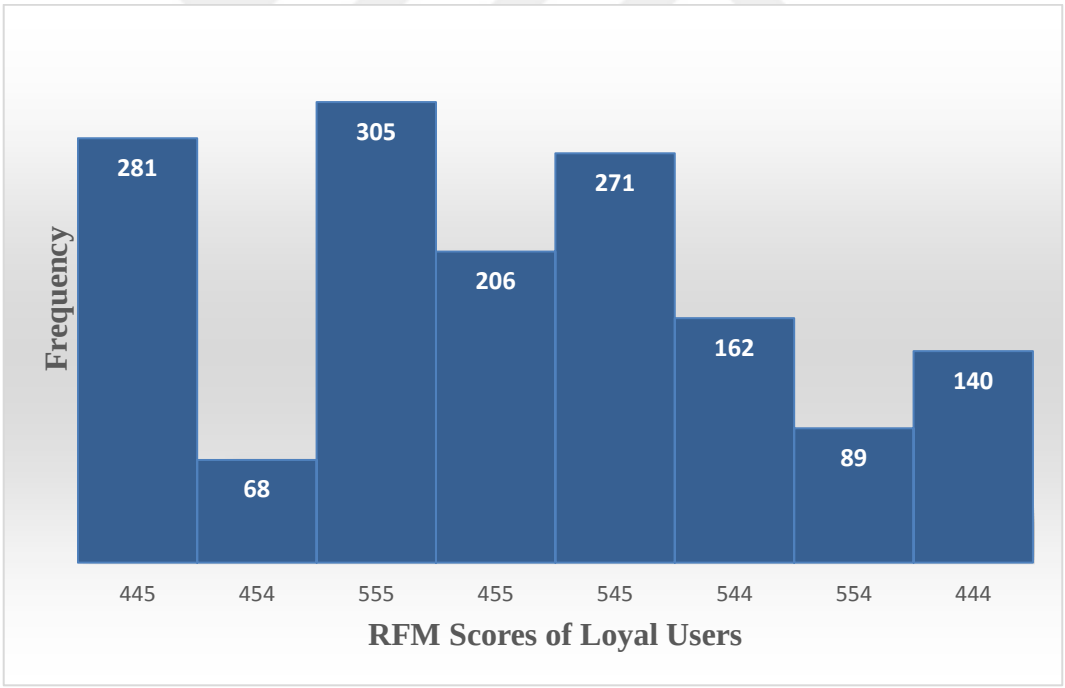


*Figure 4.1 RFM Score Distribution of Loyal Users*

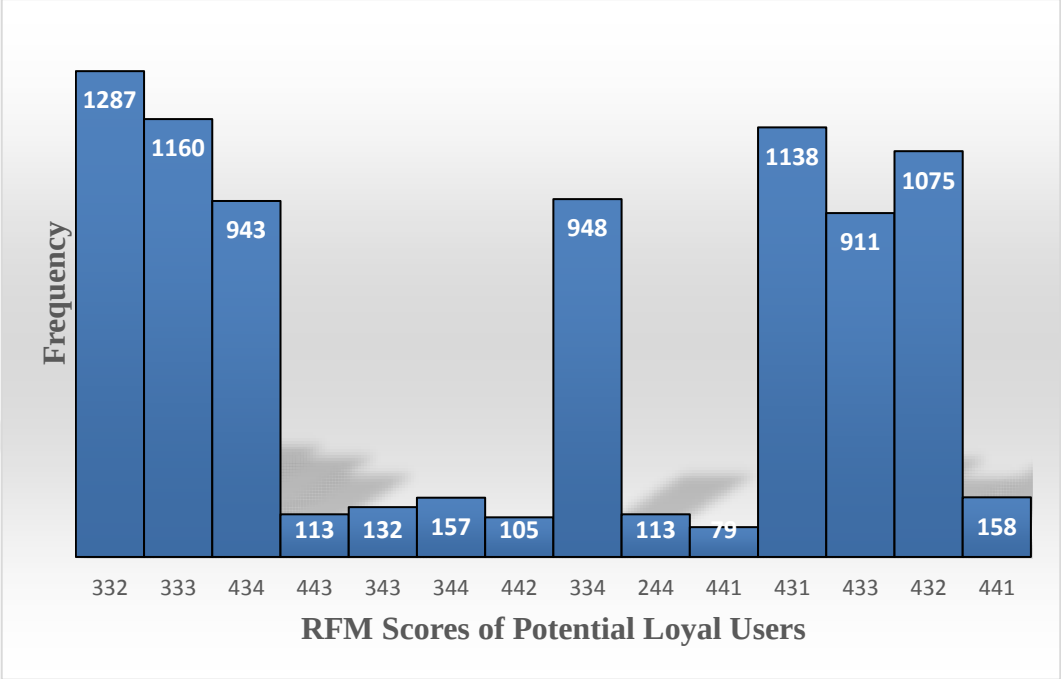


*Figure 4.2 RFM Score Distribution of Potential Loyal Users*

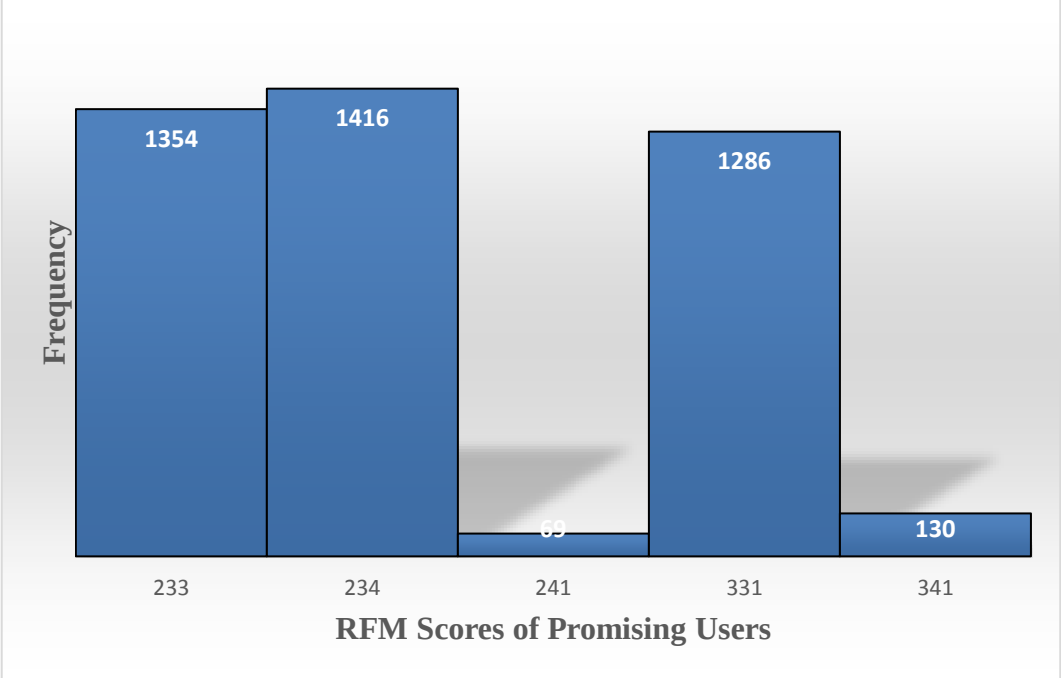


*Figure 4.3 RFM Score Distribution of Promising Users*

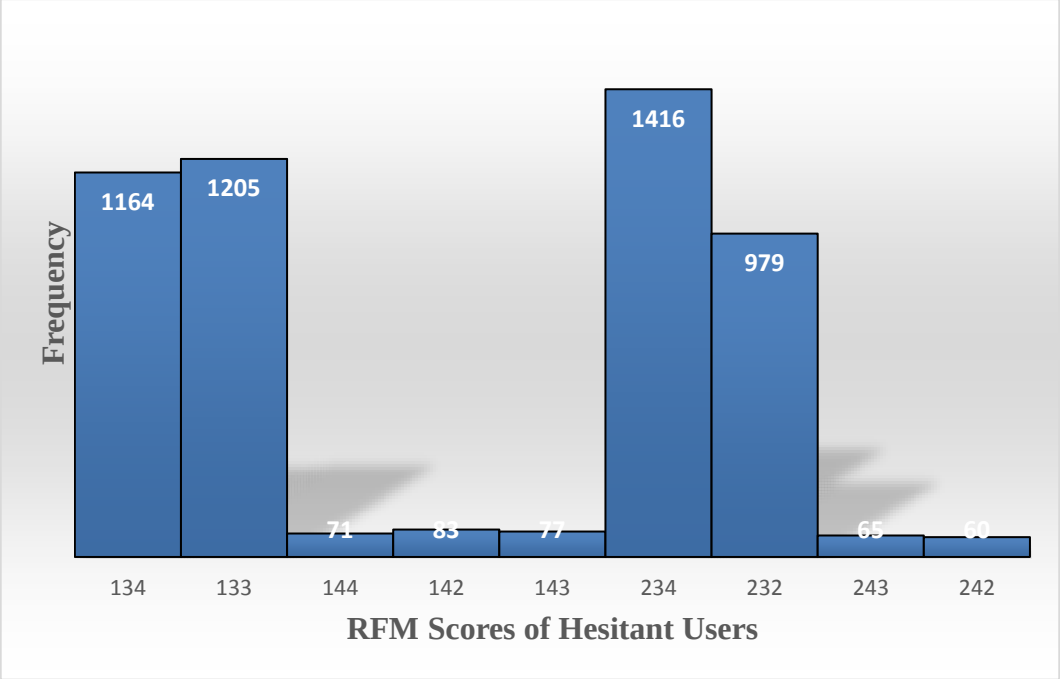


*Figure 4.4 RFM Score Distribution of Hesitant Users*

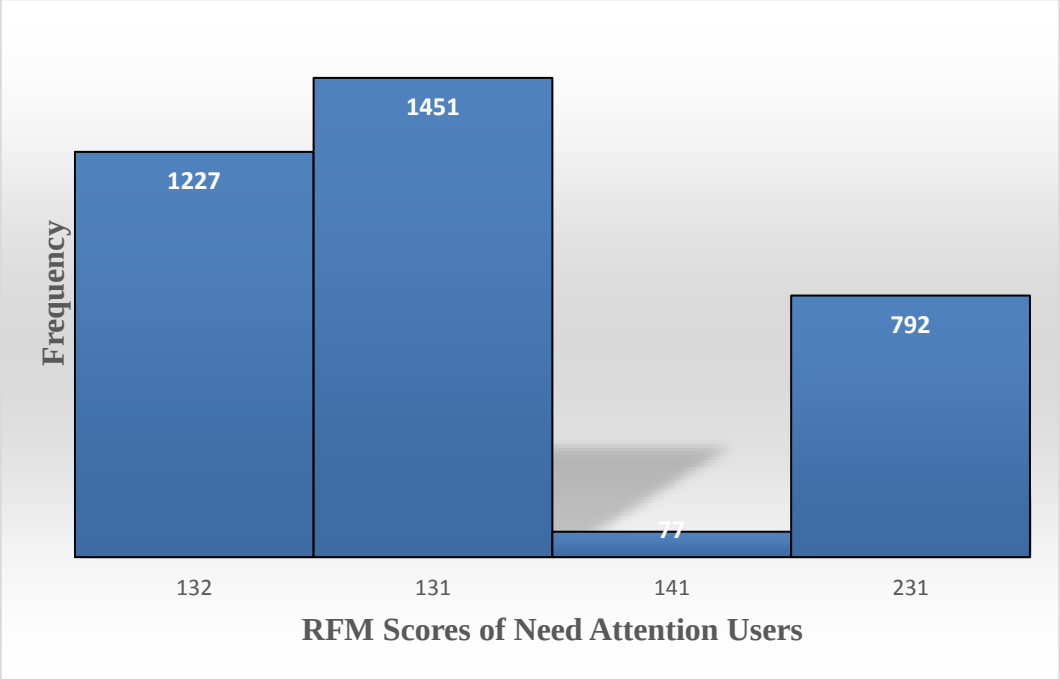


*Figure 4.5 RFM Score Distribution of Need Attention Users*

Loyal, potential loyal, and promising groups consist of active customers who visited the company last recently, shop from the company frequently, and spend much when shopping.

It is considered that attract active customers' attention is a more sensible strategy than focusing on detractor groups. Retaining existing customers and ensuring their loyalty is of great importance. It may be more valuable for the business in the long run to enhance the satisfaction and loyalty of existing customers. Such a strategy can increase the satisfaction of existing customers, encourage them to make more purchases, and attract the attention of potential customers.

Since enhancing customer satisfaction and creating positive experiences are important both for retaining existing customers and attracting potential customers, loyal customers were determined as the focus group in this thesis.

The RFM scores, recency, frequency, and monetary respectively, of the loyal customers selected as the focus group are as follows;

- 555
- 554
- 545
- 544
- 444
- 454
- 455
- 445

Hotel scores created by collaborative filtering are added to customers in each group.

### 4.2 Customer Similarities

Manhattan distance similarity, which is also referred to as city block distance or L1 distance, is a similarity measure used to assess the similarity between two vectors in a multi-dimensional space. It calculates the sum of absolute differences between corresponding elements of the vectors.

In data analysis and recommendation systems, Manhattan distance similarity can be employed to evaluate the similarity between two items or users by considering their feature vectors. It proves to be particularly valuable when working with categorical or binary data.

The formula for computing Manhattan distance similarity between two vectors A and B, both of length n, is as follows:

$$\text{Manhattan}(A, B) = |A_1 - B_1| + |A_2 - B_2| + \ldots + |A\square - B\square| \quad (2)$$

Here, $A_i$ and $B_i$ represent the elements of vectors A and B respectively at index i.

Manhattan distance similarity provides a numerical value that indicates the dissimilarity between two vectors. A lower value indicates higher similarity, while a higher value indicates greater dissimilarity.

By computing the Manhattan distance similarity between items or users, it is possible to identify those that are more similar to each other and make recommendations based on their similarity scores.

The methodology of applying Manhattan Distance to the users included in this thesis is explained in detail in Chapter 3.

According to focus groups, Manhattan distance is calculated using Python Programming Language and the closest customers to each other are determined. Here, customers who have made the same choices are removed, and outputs are obtained.

Table 4.5-Output of Manhattan Distances shows an example of the customers who most closely resemble each customer.

*Table 4.5 Output of Manhattan Distances*

| Customer ID | Compatible Customer | | |
|---|---|---|---|
| 31 | 4066 | | |
| 13 | 2887 | | |
| 166 | 171 | 2385 | 4066 |
| 171 | 166 | 2554 | 2887 |
| 1080 | 13 | 1285 | 2021 |

As seen in Table 4.5, the user most similar to user with ID number 166 is user ID 171. Moving towards the columns on the right in the same row, the Manhattan Similarity decreases for this user, indicating that user ID 4066 has less similarity to user ID 166 compared to user ID 2385 and user ID 171.

The same situation applies to all the users shown in Table 4.5. For instance, the user with User ID 13 has only one user, 2887, who has visited similar hotels. Therefore, when recommending suggestions to user ID 13, only the data of user ID 2887 can be taken into consideration.

### 4.3 Recommendation System Creation

After calculating customer similarities using the Manhattan Distance method through a Python program, a recommendation system has been designed for the users in the dataset obtained from SeturTech company and the success rate of the recommendation system designed for this thesis has been calculated.

For calculating the success rate, the users who have visited the greatest number of hotels were selected in the recommendation system to ensure a more reliable outcome.

Table 4.6 and Table 4.7 shows a brief example of the number of hotel visits from customers who have visited the greatest number of hotels according to customer profiles

and have RFM groups classified as "Loyalist" and "Potential Loyalist." The selection includes customers with both the highest frequency and diversity of hotel visits.

The dataset has been converted to the CSV (Comma-Separated Values) format to be used in the Python programming language.

*Table 4.6 Potential Loyalist Users-Total Number of Hotels Matrix*

| Customer ID | Total Number of Visited Hotels |
|---|---|
| 19310 | 53 |
| 17052 | 9 |
| 22358 | 8 |
| 22943 | 8 |
| 89751 | 8 |

*Table 4.7 Loyalist Users-Total Number of Visited Hotels Matrix*

| Customer ID | Total Number of Visited Hotels |
|---|---|
| 19662 | 146 |
| 89347 | 40 |
| 19683 | 34 |
| 129777 | 28 |
| 17532 | 20 |

After calculating users with similar characteristics as shown in Table 4.8- Users with Similar Characteristics to Potential Loyalist Customers, which are determined using the Manhattan Similarity code, test groups have been created from the users who have visited the greatest number of hotels.

Table 4.8 shows the results of calculating the users with the most similar characteristics to the users in the test groups prepared for Potential Loyalist users using the Manhattan Similarity code written in the Python programming language. This is just a sample table illustrating users with similar characteristics to the Potential Loyalist customer profiles, calculated using the Manhattan Similarity code in Python. The actual table includes more customer similarities and their corresponding user IDs.

*Table 4.8 Users with Similar Characteristics to Potential Loyalist Customers*

| Customer ID | Compatible Customer ID | | | | | | |
|---|---|---|---|---|---|---|---|
| 2933 | 19097 | 23574 | 52427 | 89382 | 106492 | 127782 | |
| 2021 | 19337 | 21050 | | | | | |
| 3784 | 19097 | 23574 | 52427 | 89382 | 106492 | | |
| 17052 | 23046 | | | | | | |
| 17127 | 19097 | 23574 | 26926 | 52427 | 89382 | 106492 | |
| 19622 | 16038 | 18802 | 19097 | 19858 | 20491 | 22036 | 22180 |
| 19310 | 28195 | | | | | | |
| 20091 | 19097 | 23574 | 52427 | 89382 | 106492 | | |
| 20166 | 18150 | 19097 | 23574 | 41217 | 49571 | 52427 | 89382 |
| 20199 | 19097 | 23574 | 41223 | 52427 | 53014 | 89382 | 93005 |
| 20206 | 19337 | | | | | | |
| 21050 | 2021 | 19097 | 23574 | 41223 | 52427 | 89382 | 106492 |
| 21299 | 17703 | 19097 | 23574 | 52427 | 89382 | 106492 | |
| 21353 | 18250 | 34668 | | | | | |
| 21381 | 19858 | | | | | | |
| 21854 | 19097 | 20030 | 23574 | 52427 | 89382 | 106492 | |
| 21906 | 30303 | 127833 | | | | | |
| 21952 | 17792 | 19097 | 20030 | 23574 | 52427 | 89382 | 106492 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 30252 | 16260 | 127848 | | | | | |
| 30282 | 19097 | 23574 | 52427 | 64725 | 89382 | 106492 | |
| 30303 | 127833 | | | | | | |
| 30601 | 25918 | | | | | | |
| 30702 | 19097 | 23574 | 52427 | 89382 | 106492 | | |
| 30869 | 30098 | | | | | | |
| 32514 | 19097 | 23574 | 26152 | 30098 | 52427 | 89382 | 106492 |

Test groups have been formed from users who have visited the most diverse and highest number of hotels.

In the test group for potential loyalist customer profiles, users with user IDs 17052 and 19310 have been selected. These two users are the ones who have visited the most diverse and highest number of hotels between potential loyalist customer profile users.

The table 4.9 presents the brief list of hotels visited by the two users with user IDs 17052 and 19310, who were selected for the test group in the potential loyalist customer profile.

*Table 4.9 Some of the Hotels Visited by the Potential Loyalist Customers in the Test Group*

| | **HOTEL ID** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Customer ID** | **373** | **384** | **409** | **442** | **457** | **466** | **493** | **528** | **560** | **584** | **600** | **756** |
| 17052 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 19310 | 1 | 1 | 0 | 1 | 3 | 1 | 2 | 3 | 0 | 1 | 7 | 2 |

For these two users given in Table 4.9-Some of the Hotels Visited by the Potential Loyalist Customers in the Test Group, in order to calculate the success rate of the recommendation system, a random 30% of the hotels they visited were removed from the dataset and marked as "0" in terms of the number of visits to those hotels.

For the user with ID 17052, the randomly selected hotel is hotel ID 1730.

For the user with ID 19310, the following hotel IDs have been removed from the list of 8 hotels: "373, 466, 850, 1268, 1400, 493, 9552, 786".

After adding "0" to the selected hotels in the user-hotel preference matrices, the updated list has been converted back to CSV format and re-added to the algorithm prepared for Manhattan similarity in Python. As a result, similar users with similar characteristics to the test users have been recalculated.

Table 4.10-Simulated Data of Hotels Not Visited by Test Group Users for Potential Loyalists presents a table where it is assumed that the two users in the test groups have not visited a random 30% of the hotels they actually visited, and these hotels have been recorded elsewhere.

*Table 4.10 Simulated Data of Hotels Not Visited by Test Group Users for Potential Loyalists*

| Customer ID | Compatible Customer ID | | | | | | |
|---|---|---|---|---|---|---|---|
| 16978 | 93005 | | | | | | |
| 17052 | 22358 | 22703 | 23046 | 104102 | | | |
| 17127 | 26926 | | | | | | |
| 19310 | 19806 | 23766 | 64741 | 127833 | | | |
| 18150 | 2021 | 19337 | 20166 | 21001 | 23163 | 49571 | 85552 |
| 18184 | 22118 | 125809 | | | | | |
| 18250 | 34668 | | | | | | |
| 18313 | 2201 | | | | | | |
| 18464 | 2021 | 2933 | 3784 | 4998 | 7922 | 13624 | 16038 |
| 18464 | 2021 | 2933 | 3784 | 4998 | 7922 | 13624 | 16038 |
| 18802 | 16038 | 82963 | | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 18985 | 2933 | 17706 | 18250 | 19097 | 19337 | 19622 | 20206 |
| 19008 | 22118 | 125809 | | | | | |
| 19097 | 23574 | 52427 | 89382 | 106492 | | | |
| 19310 | 19806 | 23766 | 64741 | 127833 | | | |
| 19337 | 20206 | | | | | | |
| 19366 | 22118 | 125809 | | | | | |
| 19622 | 2933 | 16038 | 18250 | 18802 | 18985 | 19097 | 19337 |
| 19657 | 7922 | | | | | | |
| 19675 | 28286 | 66293 | | | | | |
| 19709 | 16498 | 18133 | 18313 | 18802 | 20491 | 22703 | 23066 |
| 19794 | 2021 | | | | | | |
| 19806 | 64741 | | | | | | |
| 19858 | 16038 | 21381 | | | | | |
| 19933 | 16173 | 26030 | 30252 | 103919 | 104102 | 125540 | 126413 |

According to the new user-hotel visit matrix specified in Table 4.10, the user with ID 17052 now has the user with ID 22358 as the most similar user in terms of characteristics.

Similarly, for the user with ID 19310, the most similar user ID has changed to 19806.

Afterwards, the hotels that were randomly set aside for the users in the test groups were presented as recommendations to the new similar users in the updated similarity matrix provided in Table 4.10- Simulated Data of Hotels Not Visited by Test Group Users for Potential Loyalists. If the recommended hotels are already visited by the new similar users, it will demonstrate the success of the recommendation system. The success rate is obtained by calculating the ratio of unvisited hotels as well.

For the user with ID 17052, in the previous step, the hotel that was set aside was hotel ID 1730.

According to the new similarity matrix, the user with ID 22358, who is the closest user to user ID 17052, has also visited hotel ID 1730. Therefore, in this specific case, the success rate for this user is 100%.

Similarly, for another user in the test group, for the user ID 19310, in the previous step, the hotel that set aside were hotel ID 373, 466, 850, 1268, 1400, 493, 9552, 786.

According to the new similarity matrix, the user with ID 19806, who is the closest user to user ID 19310, has also visited hotel ID 373. User with ID 19806 has visited the hotels with ID numbers 373 and 457 in total.

Since only one of these two hotels is included in the recommendation system mentioned in this thesis, the success rate for the user ID 19310 is 50%.

Table 4.11 shows the results of calculating the users with the most similar characteristics to the users in the Loyalist classification using the Manhattan Similarity code written in the Python programming language. This is just a sample table illustrating users with similar characteristics to the Loyalist customers, calculated using the Manhattan Similarity code in Python. The actual table includes more customer similarities and their corresponding user IDs.

*Table 4.11 Users with Similar Characteristics to Loyalist Customers*

| Customer ID | Compatible Customer ID | | |
|---|---|---|---|
| 89347 | 17387 | 30572 | 101902 |
| 19662 | 16584 | | |
| 19683 | 21727 | 18149 | |
| 129777 | 19998 | | |
| 17532 | 6929 | 21789 | 132229 |
| 16372 | 132126 | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 18130 | 21727 | | | | | | |
| 89200 | 104256 | 16465 | | | | | |
| 17222 | 35853 | | | | | | |
| 16584 | 104256 | | | | | | |
| 22945 | 60334 | | | | | | |
| 4249 | 17487 | 22126 | 36661 | 66886 | 30572 | 101902 | |
| 19836 | 104184 | | | | | | |
| 86555 | 138605 | | | | | | |
| 17208 | 99674 | | | | | | |
| 30472 | 20332 | 30572 | 101902 | | | | |
| 104256 | 16584 | 16465 | | | | | |
| 35391 | 18149 | | | | | | |
| 36661 | 30572 | 101902 | | | | | |
| 36966 | 23032 | 132418 | 138605 | | | | |
| 37073 | 39415 | 53601 | | | | | |
| 37378 | 20332 | | | | | | |
| 39415 | 37073 | 53601 | | | | | |
| 40231 | 18257 | | | | | | |
| 40354 | 18369 | | | | | | |
| 40464 | 17647 | 23032 | 36966 | 74952 | 125705 | 132418 | 138605 |
| 42042 | 45751 | 77372 | 30572 | 101902 | | | |
| 42737 | 1080 | 23055 | 133884 | 30572 | 101902 | | |
| 43091 | 19957 | 34480 | | | | | |
| 22859 | 6929 | 16278 | 16666 | 16692 | 17387 | 17806 | 34979 |

In the test group for loyalist customer profiles, users with user IDs 89347, 19662, 19683, 129777, 17532, 16372, 18130, 89200, 17222, 16584, 22945, 4249, 19836, 86555, 17208, 30472, 104256, 16205, 16869, 17603, 18369, 19699, 19957, 20896, 34151, 22859, 18029, 16692, 143213, 48805, 23970, 20156, 48805 have been selected.

These users are the ones who have visited the most diverse and highest number of hotels between potential loyalist customer profile users.

The table 4.12 presents the brief list of hotels visited by the users who were selected for the test group in the loyalist customer profile.

*Table 4.12 Some of the Hotels Visited by the Loyalist Customers in the Test Group*

| | HOTEL ID | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CUSTOMER ID** | **371** | **379** | **386** | **388** | **390** | **396** | **407** | **409** | **419** | **422** | **424** | **428** |
| 19662 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 89347 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19683 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 129777 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 |
| 17532 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 16372 | 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 18130 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 89200 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17222 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 1 | 0 | 0 | 0 | 0 |
| 16584 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22945 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 |
| 4249 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19836 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 86555 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17208 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30472 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 104256 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

For these users given in Table 4.12-Some of the Hotels Visited by the Loyalist Customers in the Test Group, in order to calculate the success rate of the recommendation system, a random 30% of the hotels they visited were removed from the dataset and marked as "0" in terms of the number of visits to those hotels.

For the user with ID 89347, the randomly selected hotel is hotel ID 2883.
For the user with ID 19662, the randomly selected hotels are hotel ID 786, 887.
For the user with ID 19683, the randomly selected hotels are hotel ID 786, 466, 528.
For the user with ID 129777, the randomly selected hotels are hotel ID 1816, 1855, 1877, 466.
For the user with ID 17532, the randomly selected hotels are hotel ID 887, 2522.
For the user with ID 16372, the randomly selected hotels are hotel ID 733, 386, 1058.
For the user with ID 18130, the randomly selected hotels are hotel ID 786, 1066.
For the user with ID 89200, the randomly selected hotel is hotel ID 468.
For the user with ID 17222, the randomly selected hotel is hotel ID 696.
For the user with ID 16584, the randomly selected hotel is hotel ID 466.
For the user with ID 22945, the randomly selected hotel is hotel ID 379 and 1716.
For the user with ID 4249, the randomly selected hotels are hotel ID 390, 2599.
For the user with ID 19836, the randomly selected hotel is hotel ID 733.
For the user with ID 86555, the randomly selected hotel is hotel ID 1551.
For the user with ID 17208, the randomly selected hotel is hotel ID 466.
For the user with ID 30472, the randomly selected hotel is hotel ID 1251.
For the user with ID 104256, the randomly selected hotel is hotel ID 466.

Hotels to be used for testing the recommendation system for users in this test group have been replaced with a "0" in the hotel-user preferences matrix.

After adding "0" to the selected hotels in the user-hotel preference matrices, the updated list has been converted back to CSV format and re-added to the algorithm prepared for

Manhattan similarity in Python. As a result, similar users with similar characteristics to the test users have been recalculated.

Table 4.13-Simulated Data of Hotels Not Visited by Test Group Users for Loyalists presents a table where it is assumed that the two users in the test groups have not visited a random 30% of the hotels they actually visited, and these hotels have been recorded elsewhere.

*Table 4.13 Simulated Data of Hotels Not Visited by Test Group Users for Loyalists*

| Customer ID | Compatible Customer ID | | | | |
|---|---|---|---|---|---|
| 89347 | 17387 | 30572 | 101902 | | |
| 19662 | 16584 | | | | |
| 19683 | 18130 | | | | |
| 129777 | 99674 | | | | |
| 17532 | 21789 | 132229 | | | |
| 16372 | 16278 | 16692 | 17487 | 68138 | |
| 18130 | 21727 | | | | |
| 89200 | 18138 | 46368 | 130274 | | |
| 17222 | 37378 | | | | |
| 16584 | 104256 | | | | |
| 22945 | 60334 | | | | |
| 4249 | 17487 | 36661 | 30572 | 101902 | |
| 12320 | 30572 | 101902 | | | |
| 19836 | 104184 | | | | |
| 86555 | 138605 | | | | |
| 17208 | 19063 | 35601 | 127042 | 30572 | 101902 |
| 30472 | 20332 | 30572 | 101902 | | |
| 104256 | 16584 | | | | |
| 16205 | 77111 | 22126 | 127346 | 30572 | 101902 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 16869 | 6929 | 30572 | 101902 | | | | |
| 17603 | 9635 | 16701 | 17046 | 18429 | 21434 | 24065 | 36427 |
| 18369 | 22932 | 40354 | 45751 | 30572 | 101902 | | |
| 19699 | 16701 | 16834 | 17806 | 19998 | 24886 | 90464 | 118642 |
| 19957 | 171 | 17046 | 18138 | 18149 | 43091 | 128664 | 30572 |
| 20896 | 171 | 16908 | 66886 | 102104 | 30572 | 101902 | |
| 34151 | 16234 | 18029 | 21434 | 21789 | 22811 | 26986 | 36966 |
| 22859 | 6929 | 16278 | 16666 | 16692 | 17387 | 17806 | 34979 |
| 18029 | 16234 | 16951 | 21434 | 21789 | 26986 | 34151 | 36966 |
| 16692 | 16278 | 16666 | 17387 | 22859 | 68138 | 132126 | 30572 |
| 143213 | 17647 | 23032 | 36966 | 40464 | 74952 | 125705 | 132418 |
| 48805 | 16278 | 17487 | 18429 | 19063 | 143709 | 30572 | 101902 |
| 23970 | 1080 | 18284 | 23055 | 42737 | 68138 | 127042 | 143672 |
| 20156 | 16701 | 23032 | 36966 | 132418 | 138605 | 20156 | 16701 |
| 48805 | 16278 | 17487 | 18429 | 19063 | 143709 | 30572 | 101902 |

According to the new user-hotel visit matrix specified in Table 4.13, the user with ID 89347 now has the user with ID 17387 as the most similar user in terms of characteristics.

Similarly, for the user with ID 19662, the most similar user ID has changed to 16584.

For all customers given in Table 4.13, the same customer-customer similarity has been used to match each customer with the closest customer in the adjacent column.

Afterwards, the hotels that were randomly set aside for the users in the test groups were presented as recommendations to the new similar users in the updated similarity matrix provided in Table 4.10- Simulated Data of Hotels Not Visited by Test Group Users for Potential Loyalists. If the recommended hotels are already visited by the new similar users, it will demonstrate the success of the recommendation system. The success rate is obtained by calculating the ratio of unvisited hotels as well.

For the user with ID 89347, in the previous step, the hotel that was set aside was hotel ID 2883. According to the new similarity matrix, the user with ID 17387, who is the closest user to user ID 89347, has also visited hotel ID 2883. Therefore, in this specific case, the success rate for this user is 100%.

Similarly, for another user in the test group, for the user ID 19662, in the previous step, the hotel that set aside were hotel ID 786, 887. According to the new similarity matrix, the user with ID 16584, who is the closest user to user ID 19662, has also visited hotel ID 786. Since only one of these two hotels is included in the recommendation system mentioned in this thesis, the success rate for the user ID 19662 is 50%.

For the user with ID 19683, in the previous step, the hotels that were set aside were hotel ID 786, 466, 528. According to the new similarity matrix, the user with ID 21727, who is the closest user to user ID 19683, has also visited hotel ID 466, 786. Therefore, in this specific case, the success rate for this user is 66.67%.

For the user with ID 129777, in the previous step, the hotel that set aside were hotel ID 1816, 1855, 1877, 466. According to the new similarity matrix, the user with ID 19998, who is the closest user to user ID 129777, has also visited hotel ID 466. Therefore, in this specific case, the success rate for this user is 20%.

For the user with ID 17532, in the previous step, the hotel that set aside were hotel ID 887, 2522. According to the new similarity matrix, the user with ID 6929, who is the closest user to user ID 17532, has also visited hotel ID 887. Therefore, in this specific case, the success rate for this user is 50%.

For the user with ID 16372, in the previous step, the hotel that set aside were hotel ID 733, 386, 1058. According to the new similarity matrix, the user with ID 132126, who is the closest user to user ID 16372, has also visited hotel ID 386. Therefore, in this specific case, the success rate for this user is 33.3%.

For the user with ID 18130, in the previous step, the hotel that set aside were hotel ID 786, 1066. According to the new similarity matrix, the user with ID 21727, who is the closest user to user ID 18130, has also visited hotel ID 786. Therefore, in this specific case, the success rate for this user is 50%.

For the user with ID 89200, in the previous step, the hotel that was set aside was hotel ID 468. According to the new similarity matrix, the user with ID 104256, who is the closest user to user ID 89200, has also visited hotel ID 468. Therefore, in this specific case, the success rate for this user is 100%.

For the user with ID 17222, in the previous step, the hotel that was set aside was hotel ID 696. According to the new similarity matrix, the user with ID 35853, who is the closest user to user ID 17222, has also visited hotel ID 696. Therefore, in this specific case, the success rate for this user is 100%.

For the user with ID 16584, in the previous step, the hotel that was set aside was hotel ID 466. According to the new similarity matrix, the user with ID 104256, who is the closest user to user ID 16584, has also visited hotel ID 466. Therefore, in this specific case, the success rate for this user is 100%.

For the user with ID 22945, in the previous step, the hotel that was set aside was hotel ID 379 and 1716. According to the new similarity matrix, the user with ID 60334, who is the closest user to user ID 22945, has also visited hotel ID 379. Therefore, in this specific case, the success rate for this user is 50%.

For the user with ID 4249, in the previous step, the hotel that set aside were hotel ID 390 and 2599. According to the new similarity matrix, the user with ID 17487, who is the closest user to user ID 4249, has also visited hotel ID 390. Therefore, in this specific case, the success rate for this user is 50%.

For the user with ID 19836, in the previous step, the hotel that was set aside was hotel ID 733. According to the new similarity matrix, the user with ID 104184, who is the closest user to user ID 19836, has also visited hotel ID 733. Therefore, in this specific case, the success rate for this user is 100%.

For the user with ID 89555, in the previous step, the hotel that was set aside was hotel ID 1551. According to the new similarity matrix, the user with ID 138605, who is the closest user to user ID 89555, has also visited hotel ID 1551. Therefore, in this specific case, the success rate for this user is 100%.

For the user with ID 17208, in the previous step, the hotel that was set aside was hotel ID 466. According to the new similarity matrix, the user with ID 99674, who is the closest user to user ID 17208, has not visited hotel ID 466. Therefore, in this specific case, the success rate for this user is 0%.

For the user with ID 30472, in the previous step, the hotel that was set aside was hotel ID 1251. According to the new similarity matrix, the user with ID 20332, who is the closest user to user ID 30472, has also visited hotel ID 1251. Therefore, in this specific case, the success rate for this user is 100%.

For the user with ID 104256, in the previous step, the hotel that was set aside was hotel ID 466. According to the new similarity matrix, the user with ID 16584, who is the closest user to user ID 104256, has also visited hotel ID 466. Therefore, in this specific case, the success rate for this user is 100%.

When the average success rates of test users in the loyalist and potential loyalist customer segments are calculated, the success rate becomes 69.47%.

## 5. CONCLUSION

As a conclusion, the recommendation system is an area in the industrial system that has multiple content-based, collaborative and hybrid approaches to increase companies' growth and productivity.

Recommendation systems provide access to personalized information on the web and have progressed in the last 10 years.

Recommendation systems created new options for information search and filtering. For instance, online shopping stores have increased their profits and music lovers discovered new songs.

Besides the positive effect of the recommendation system on customers, there are some limitations and deficiencies. This thesis has reviewed three different recommendation approaches in detail. Additionally, this thesis has reviewed Recency, Frequency, and Monetary analysis in detail.

## REFERENCES


Alsayat, A. (2022). Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. Neural Computing and Applications, 1-22.

Berker Türker, B., Tugay, R., Öğüdücü, Ş., & Kızıl, İ. (2020). Hotel Recommendation System Based on User Profiles and Collaborative Filtering. arXiv e-prints, arXiv-2009.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. Knowledge-based systems, 46, 109-132.

Seyednezhad, S. M., Cozart, K. N., Bowllan, J. A., & Smith, A. O. (2018). A review on recommendation systems: Context-aware to social-based. arXiv preprint arXiv:1811.11866.

Chien, Y. H., & George, E. I. (1999, January). A bayesian model for collaborative filtering. In AISTATS.

G. R. Faulhaber, "Design of service systems with priority reservation," in Conf. Rec. 1995 IEEE Int. Conf. Communications, pp. 3–8.

H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985.

Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995, May). Recommending and evaluating choices in a virtual community of use. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 194-201).

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 16(3), 261-273.

J. Williams, “Narrow-band analyzer (Thesis or Dissertation style),” Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

Kaya, T., & Kaleli, C. (2022). A novel top-n recommendation method for multi-criteria collaborative filtering. Expert Systems with Applications, 198, 116695.

Lv, X. (2021, March). Analysis and Optimization Strategy of Travel Hotel Website Reservation Behavior Based on Collaborative Filtering. In 2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (pp. 362-365). IEEE.

Motorola Semiconductor Data Manual, Motorola Semiconductor Products Inc., Phoenix, AZ, 1989.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In The adaptive web (pp. 325-341). Springer, Berlin, Heidelberg.

R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. IEEE Trans. Plasma Sci. [Online]. 21(3). pp. 876—880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar

S. Chen, B. Mulgrew, and P. M. Grant, “A clustering technique for digital communications channel equalization using radial basis function networks,” IEEE Trans. Neural Networks, vol. 4, pp. 570–578, July 1993.

Seyednezhad, S. M., Cozart, K. N., Bowllan, J. A., & Smith, A. O. (2018). A review on recommendation systems: Context-aware to social-based. arXiv preprint arXiv:1811.11866.

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In Recommender systems handbook (pp. 257-297). Springer, Boston, MA.

Sun, Z., & Luo, N. (2010, August). A new user-based collaborative filtering algorithm combining data-distribution. In 2010 International Conference of Information Science and Management Engineering (Vol. 2, pp. 19-23). IEEE.

Wu, J., Liu, C., Wu, Y., Cao, M., & Liu, Y. (2022). A novel hotel selection decision support model based on the online reviews from opinion leaders by best worst method. International Journal of Computational Intelligence Systems, 15(1), 1-20.

## BIOGRAPHICAL SKETCH

Begüm Uyanık is recognized as a successful engineer with completed education and gained experience. She completed her undergraduate education in Aeronautical Engineering Department at University of Turkish Aeronautical Association. After starting to work as a project engineer at Turkish Airlines Technic Inc., she was also accepted into the design engineering position at Turkish Aerospace, then worked as a system engineer. Here, she has been involved in many projects and offered creative and efficient solutions. Particularly in the field of system installation design in military aircraft projects, she has achieved successful results with different design solutions. She is currently pursuing her master's degree in Intelligent Systems Engineering at Galatasaray University. Begüm presented her paper titled "A Recommendation System Study on User-Product data" at the International Conference on Engineering Technologies (ICENTE'22) and her article titled "Manhattan Distance Based Hybrid Recommendation System" was published in the "International Journal of Applied Mathematics Electronic and Computers.

## PUBLICATIONS

- A Recommendation System Study on User-Product Data (International Conference on Engineering Technologies (ICENTE'22))
- Manhattan Distance Based Hybrid Recommendation System (International Journal of Applied Mathematics Electronic and Computers)