MEF UNIVERSITY

RFM BASED CUSTOMER SEGMENTATION FOR A MOBILE APPLICATION

Capstone Project

Ozan Barış Baykan

İSTANBUL, 2021

MEF UNIVERSITY

RFM BASED CUSTOMER SEGMENTATION FOR A MOBILE APPLICATION

Capstone Project

Ozan Barış Baykan

Advisor: Prof. Dr. Özgür Özlük

İSTANBUL, 2021

MEF UNIVERSITY

Name of the project: RFM BASED CUSTOMER SEGMENTATION FOR A MOBILE APPLICATION
Name/Last Name of the Student: Ozan Barış Baykan
Date of Thesis Defense: 01/09/2021

I hereby state that the graduation project prepared by Ozan Barış Baykan has been completed under my supervision. I accept this work as a "Graduation Project".

01/09/2021
Prof. Dr. Özgür Özlük

I hereby state that I have examined this graduation project by Ozan Barış Baykan which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

01/09/2021
Prof. Dr. Özgür Özlük

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof. Dr. Özgür Özlük	……
2. Dr. Tuna Çakar	……

ACADEMIC HONESTY PLEDGE

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Ozan Barış Baykan 01/09/2021

Name Date Signature

EXECUTIVE SUMMARY

RFM BASED CUSTOMER SEGMENTATION FOR A MOBILE APPLICATION

Ozan Barış Baykan

Advisor: Prof. Dr. Özgür Özlük

SEPTEMBER, 2021, 31 pages

In this project, customer segmentation was made for Doggo, a mobile application that brings together trained dog walkers for people who are not able to provide daily needs of their dogs. The data was organized by obtaining the columns of recency, frequency, monetary and tenure, and RFM-based customer segmentation was made using machine learning algorithms such as K-means and Gaussian Mixture Model (GMM). Then, the model was built with the part of the dataset that includes recency, monetary and tenure columns using K-means. In addition, with a function developed, the RFM and tenure will be repeated at intervals determined by the Doggo operation team, and this tool is used to monitor the customer condition changing. Various marketing campaigns have been proposed according to the current situation and the transitions they have made.

Key Words: Marketing, Customer Segmentation, RFM, Clustering, Machine Learning, K-means clustering, GMM clustering

ÖZET

MOBİL BİR UYGULAMA İÇİN GSP TABANLI MÜŞTERİ SEGMENTASYONU

Ozan Barış Baykan

Proje Danışmanı: Prof. Dr. Özgür Özlük

EYLÜL, 2021, 31 sayfa

Bu projede köpeklerinin günlük ihtiyaçları karşılama konusunda zorluk yaşayanlar için eğitimli köpek gezdiricilerini bir araya getiren mobil uygulama Doggo için müşteri segmentasyonu yapılmıştır. Veriler, güncellik, sıklık, parasal getiri ve kullanım süresi sütunları elde edilerek düzenlenmiştir. K-means ve Gaussal Karışık Model (GKM) gibi makine öğrenmesi algoritmaları kullanılarak GSP tabanlı müşteri segmentasyonu yapılmıştır. Daha sonra, veri kümesinin K-ortalamalar kullanılarak güncellik, parasal getiri ve kullanım süresi sütunlarını içeren kısmı ile model oluşturulmuştur. Ayrıca, geliştirilen bir fonksiyon ile GSP ve kullanım süresi Doggo operasyon ekibi tarafından belirlenecek periyotlarla tekrarlanacak ve kullanıcıların kümeler arasındaki geçişlerini izlenebileceği bir araç oluşturulacaktır. İçinde bulundukları mevcut duruma ve yaptıkları geçişlere göre çeşitli pazarlama kampanyaları önerilmiştir.

Anahtar Kelimeler: Pazarlama, Müşteri Segmentasyonu, GSP, Kümeleme, Makine Öğrenimi, K-ortalama kümeleme, GKM kümeleme

TABLE OF CONTENTS

ACADEMIC HONESTY PLEDGE v

EXECUTIVE SUMMARY vi

ÖZET vii

TABLE OF CONTENTS viii

TABLE OF FIGURES ix

1 INTRODUCTION 1

1.1. Literature Survey 2

2 ABOUT THE DATA 4

2.1 Data Cleaning 5

2.2 Exploratory Data Analysis 5

3 PROJECT DEFINITION 14

3.1 Problem Statement 14

3.2 Project Objectives 14

3.3 Project Scope 14

4 METHODOLOGY 15

4.1 Preprocessing 15

4.2 Generating RFM and Tenure Columns 15

4.3 K-means 17

4.4 Gaussian Mixture Model (GMM) 21

4.5 Repetitive RFM and Tenure Interval Analysis 22

5 RESULTS 24

5.1 The Labels of K-means Clustering with 8 Clusters 28

REFERENCES 30

TABLE OF FIGURES

Figure 1: Top 10 owners based on the number of walks purchased 6
Figure 2: Top 10 owners based on amount of money they spent 7
Figure 3: Distribution of Total Amount Grouped by Customer 8
Figure 4: Distribution of Amount 9
Figure 5: Walking Types 10
Figure 6: Top 10 Districts 11
Figure 7: Total Cumulative Monetary 12
Figure 8: The Last Year Cumulative Monetary 13
Figure 9: A Part of RFM and Tenure Dataset 16
Figure 10: Correlation Heatmap 16
Figure 11: Result of the Elbow Method 18
Figure 12: Result of the Silhouette Score 18
Figure 13: The Result of the K-means Clustering with 8 Clusters 19
Figure 14: An Example of Pair Plots 19
Figure 15: An Example of Descriptive Statistics Table 20
Figure 16: Similarity Heat Map 21
Figure 17: Repetitive RFM and Tenure Interval Analysis Graph 22
Figure 18: K-means Clustering with 8 Clusters 25
Figure 19: K-means Clustering with 8 Clusters' Cardinality Plot 26
Figure 20: K-means Clustering with 8 Clusters' Magnitude Plot 26
Figure 21: K-means Clustering with 8 Clusters' Magnitude vs Cardinality Plot 27
Figure 22: K-means Clustering with 8 Clusters' Magnitude per Cardinality Plot 27

1 INTRODUCTION

Doggo is a mobile application whose goal is to match dog owners and dog walkers according to their needs and by the price they can afford. Firstly, Doggo trains dog walkers which they made an agreement to provide contentment and reliability at the highest level, then a matching process takes place. Associates of Doggo would like to be very good at matching and customer satisfaction. One of the steps of improving customer satisfaction is defining the different customer segmentations. For example, if it is known which customer belongs to which cluster, a campaign can be organized directly to related customers. In order to reach the right customer in the right cluster, customer segmentation must be applied quite accurately. The RFM model is a typical technique for developing marketing strategy and is frequently used in database marketing (Wei. et al., 2010). For many years, RFM (Recency, Frequency, and Monetary) values have been used to determine which consumers are important to the firm, which customers require promotional efforts, and so on (Dogan et al., 2018). Based on the RFM values of the consumers, an enterprise's clients are successfully segmented into groups with comparable characteristics (Christy et al., 2018). According to the dataset or the actual case, recency, monetary and frequency that belong to customers are used for clustering and there is an additional and at least as important as any others that is Tenure, which means lifetime of customer, can be used as fourth criteria as in this project. RFM and tenure columns are generated using the part from last year. However, the analysis will be repeated for each weekly period and customers' movements between clusters can be monitored periodically. Clustering is carried out using K-means and GMM algorithms in this project. The models are run with different cluster numbers to specify most proper clustering conditions related to Doggo and clusters are labeled in the same point of view. In order to confirm the accuracy of the clusters, a similarity heatmap is created and information about the status of the members of the clusters is obtained by various distance calculations. The labels and features of the customer will be compared and migration between clusters is monitored based on the customer. According to the results of these studies, suggestions are made in accordance with the marketing structure of Doggo. The cornerstone for analyzing consumer behavior is provided by RFM. As previously said, scoring techniques differ significantly, and each has its own set of advantages and disadvantages. This conversation is intended to aid marketers in establishing a good foundation for measuring, analyzing, and implementing consumer segmentation (Miglautsch, 2000). With

these models and findings, Doggo can identify their current and new customers and organize some specific campaigns for each group.

1.1. Literature Survey

Nowadays, big data analytics is becoming more popular day by day. Most innovative companies would like to improve their methods and processes using big data analytics techniques to obtain the best results. In every way, businesses must better comprehend their consumers' data. Customer-company interaction has become more dependent on detecting similarities and variances among consumers, anticipating their actions, and offering better alternatives and possibilities to customers. In this case, segmenting clients based on their data became critical (Dogan et al., 2018). According to research from Ducange et al. (2017), for an efficient and successful strategic marketing effort, getting and evaluating the valuable insight buried behind the massive quantity of data available on social media is becoming a must. One of the most important analyses is customer segmentation in the marketing field. Each customer has similar features that identify marketing with the other customers. Thanks to these similar features like maturity, sex, interest, and customers' passions, clustering is carried out and it is called customer segmentation (Curto, n.d.). Even though there are a lot of types of customer segmentation such as demographic, behavioral segmentation, and RFM based segmentation in the marketing field, a few similar examples with customer segmentation carried out in this project will be reviewed in this section. For calculating consumer behavior, a straightforward framework is created as the goal of RFM (Miglautsch, 2000). In this capstone project, RFM based customer segmentation is used to determine clusters like the following applications. According to the research from Sarvari et al. (2016), various trials are carried out to select the best approach to customer segmentation. The authors claim that demographic data make the RFM based customer segmentation more powerful. Based on the different scenarios such as the combination of weighted RFM (WRFM), unweighted RFM values, and with or without demographic data, the results show if the demographic data is combined with WRFM, the model can divide the customer into cluster more accurate, and the contribution of combination effect is indisputable. In a similar approach used in this study, various conditions are tried to obtain the best clustering results. In research from Anitha and Patil (2019) The dataset values and parameters, which are spread across a certain period of business activities, give an organized insight into consumer purchasing habits and behavior across numerous areas. This

study employs dataset segmentation principles utilizing the K-means Algorithm and is based on the RFM (Recency, Frequency, and Monetary) model. Christy et al. (2018) proposed that as a first step, the authors generate recency, frequency, and monetary columns, and clusters are created by the K-means algorithm using features in a customer base manner. The client group that generates the most profit for the organization is determined by examining the attitudes of each cluster. The k-means algorithm is a common clustering approach that reduces clustering error (Likas et al., 2003). However, K-means has some disadvantages such as the number of clusters must be given as an input to the algorithm, and it can be affected easily by initial starting conditions. The number of clusters K in the database is known in advance is implied by the K-Means technique, like many other clustering algorithms, however it is obviously not the case in real-world applications (Lozano et al., 1999). Another clustering algorithm used in this project is Gaussian Mixture Model (GMM). Because of this adaptive and probabilistic approach to data modeling, we have soft cluster assignments rather than hard cluster assignments like k-means. This indicates that any of the distributions with a matching probability may have created each data point (Foley, 2021). According to research from Dogan et al. (2018), they discovered that the existing customer segmentation, which is based only on consumer expenditure, is insufficient. On the other hand, improved consumer knowledge, well-designed approaches, and more optimal solutions can be obtained by using the RFM models they suggested. According to literature reviews, using RFM based customer segmentation is an undeniable approach in the various types of customer segmentation. The details of RFM based customer segmentation using K-means algorithm are mentioned in the following sections.

2 ABOUT THE DATA

Doggo collects the different types of data related to owners, dogs, walkers, etc. Walks Customer Segment and Owner datasets are the most proper two to cluster the customer. The Walks Customer Segment and the Owners datasets are merged, and it consists of 49 columns and 61,275 transactions.

Walks Customer Segment data includes id related columns dogid, walkingid, paymentid, promocodeid, walking related columns walkingtype, walkstatus, time related columns ordercreatedtime, checkintime, confirmtime, matchtime, starttime, finishtime, location related columns district, neighborhood, payment related columns paymentstatus, paymenttype, amount, discount, walkerincome.

Explanation of features that are used for this project are listed as follows:

- ownerid: Identification code of owner
- dogid: Identification code of dog
- paymentid: Identification code of payment
- walkingType: (Planned/AdHoc/Customize/Package) type of the service
    i. Planned: A walk that is planned at least 2 hours before the walking time
    ii. AdHoc: A walk that a customer needs suddenly. Appropriate walker is provided within one hour
    iii. Package: A bundle of walks for 1 month
    iv. Customize: It could be considered as a package, but the difference is customers can shape the package based on their wishes.
- walkstatus: (Finished/Cancelled) type of the walk status
    i. Finished: A walk that is carried out by walker
    ii. Cancelled: A walk that is cancelled by owner
- ordercreatedtime: Time of created order for a walking
- checkintime: Check in time for a walking
- district: The district of owner address
- paymentstatus: (Free/Approved) type of the payment status
    i. Free: A walk that is gifted from the company to the customer for free
    ii. Approved: A walk that is paid by customer
- amount: The money charged from owner for the walk

Even though each feature is essential to understand data or get some insight from data or wrangling operations, the amount and the checkintime are most important for this paper. The amount is the payment carried out from customers in TL currency, and the checkintime is when the dog walker takes the dog from the owner, which means the walking will be complete. The higher importance of these two columns is that amount is used to create monetary values. The checkintime is used to generate recency and frequency values for RFM based clustering models.

Owners dataset contains id, mgapplytime, signuptime, firstdogaddedtime, and these features gives some insights and provides a better understanding of customers and their behaviours.

2.1 Data Cleaning

As the first step of this part, missing values are detected because missing values can affect the RFM creating processes and customer segmentation. Each variable is not necessary to create RFM based models and exploratory data analysis. The most important columns are ownerid, walkingid, checkintime and amount. They will be used to create recency, frequency, and monetary columns. There is no missing value in those related variables.

2.2 Exploratory Data Analysis

The dataset is investigated to understand it, and data wrangling processes are carried out to get the dataset ready for exploratory data analysis. It provides insights from the dataset such as errors or miscoding of variables, correlations between variables before modelling processes.

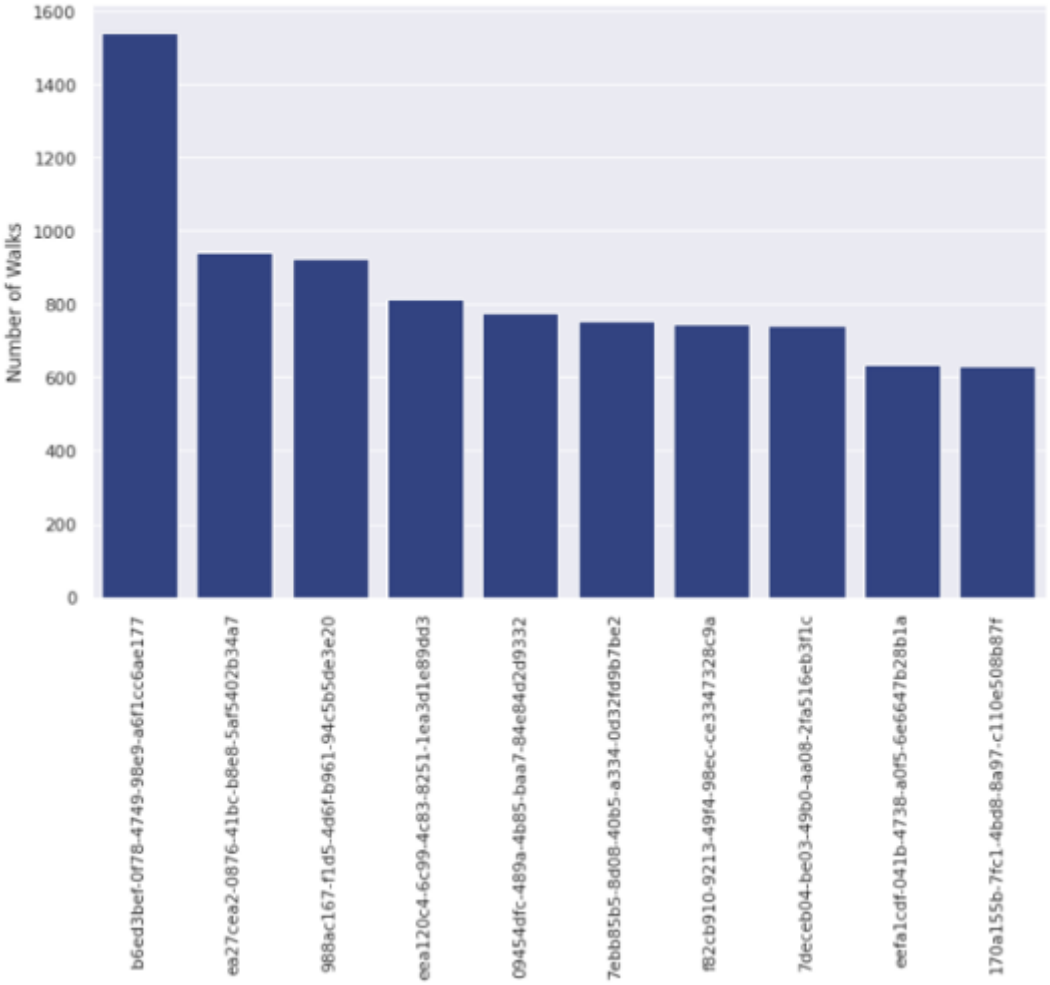


Figure 1: Top 10 owners based on the number of walks purchased

Figure 1 shows the top ten owners based on the number of walks purchased. The number of walks purchased by the top ten customers constitutes 13.2% of the total number of walks. There is a significant difference between the first customer and the second customer. The first customer could be a loyal customer. On the other hand, this value could be an outlier.

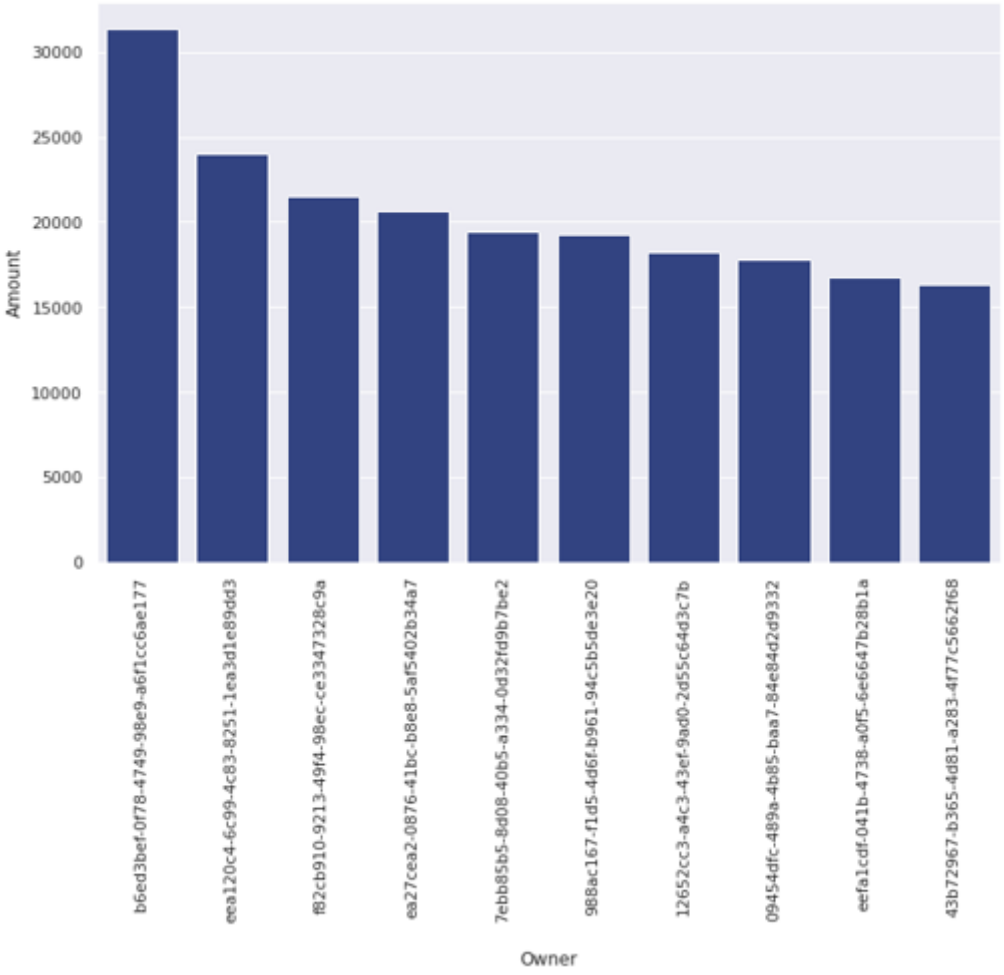


Figure 2: Top 10 owners based on amount of money they spent

Figure 2 presents that the results similar to the analysis made according to the number of walks purchased by the customers were obtained. Again, the first customer provided more returns than the other customers. Amount has a positive correlation with the number of walks.

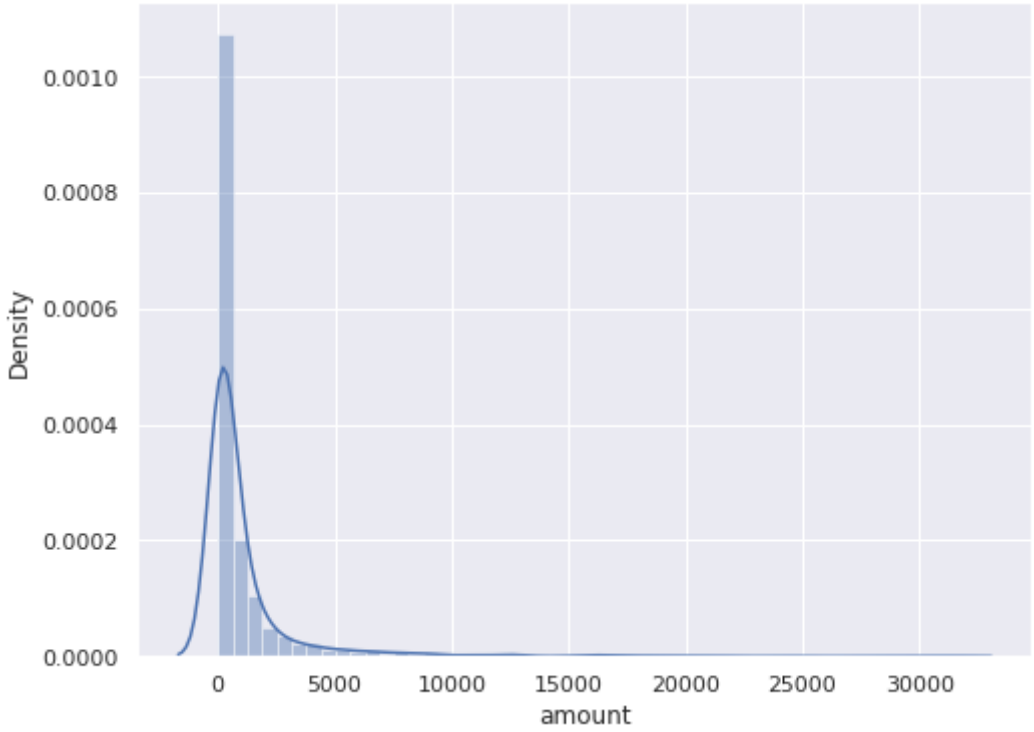


Figure 3: Distribution of Total Amount Grouped by Customer

As further analysis of the amount, the distribution of the total amount grouped by customers is shown in Figure 3. It is a highly right-skewed distribution and not a normal distribution due to its long tail and outliers. Doggo presents the first walk to its customer, but it reflects its system as a customer transaction, so there is an accumulation between 0-50 ₺.

In Figure 4 below, the distribution of the amount is plotted to explain the Doggo transaction system and support it with visualisation. Each transaction of a customer is recorded one by one separately. In other words, the payment process is realised when walking is completed. Thus, each transaction has a unique walkid even if a customer purchased a package of walking.

The average price of walking is around 30 ₺ if customers buy some package program or customise their schedule in a week or a month. Some discounts can be applied according to service. These are the reasons for batching between 20-40 ₺. The higher prices in the graph are related to the other services provided by Doggo, such as sitting, which means home care, boarding, which represents nursing home care.

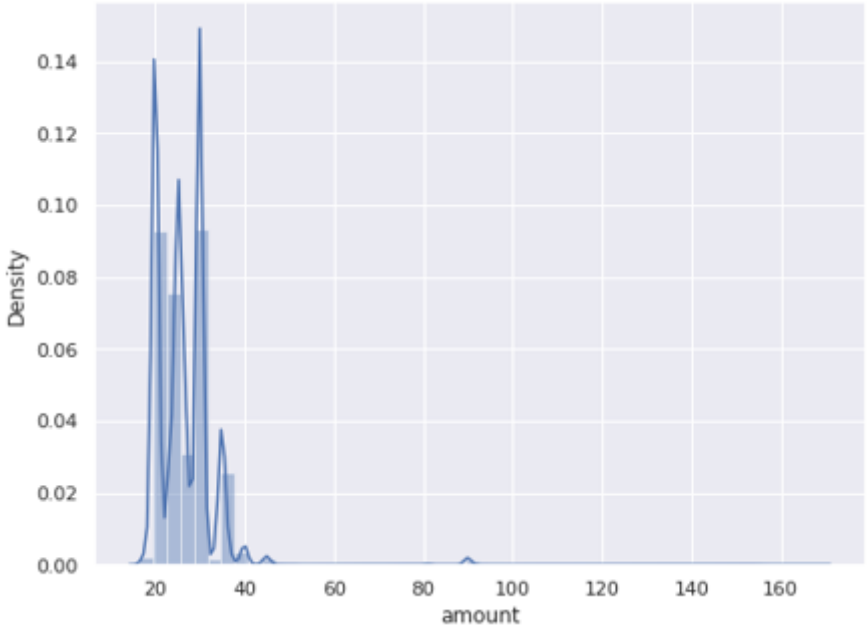


Figure 4: Distribution of Amount

After the analysis above, some errors are realized at the end of the amount column sorted by descending order. The numbers less than 16.48 can be misleading amounts and they affect the mean, and they are replaced with the median of the amount.

Secondly, the customer that has the highest monetary value is dropped from data, because it leads to higher standard deviation, and it is commented as an outlier.

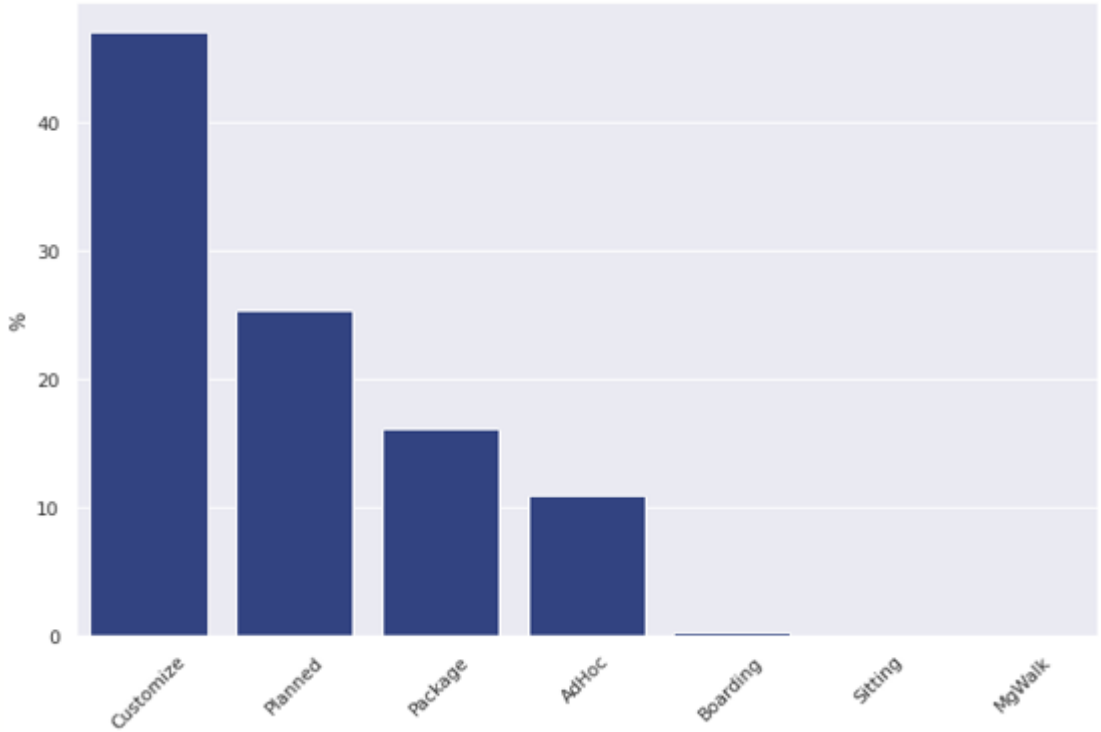


Figure 5: Walking Types

According to walking type analysis, almost half of all transactions are customized walking generated by customers based on their schedules or choices. Planned walking is second, and it means 25% of walks were scheduled before the walking, but this category includes only single walks.

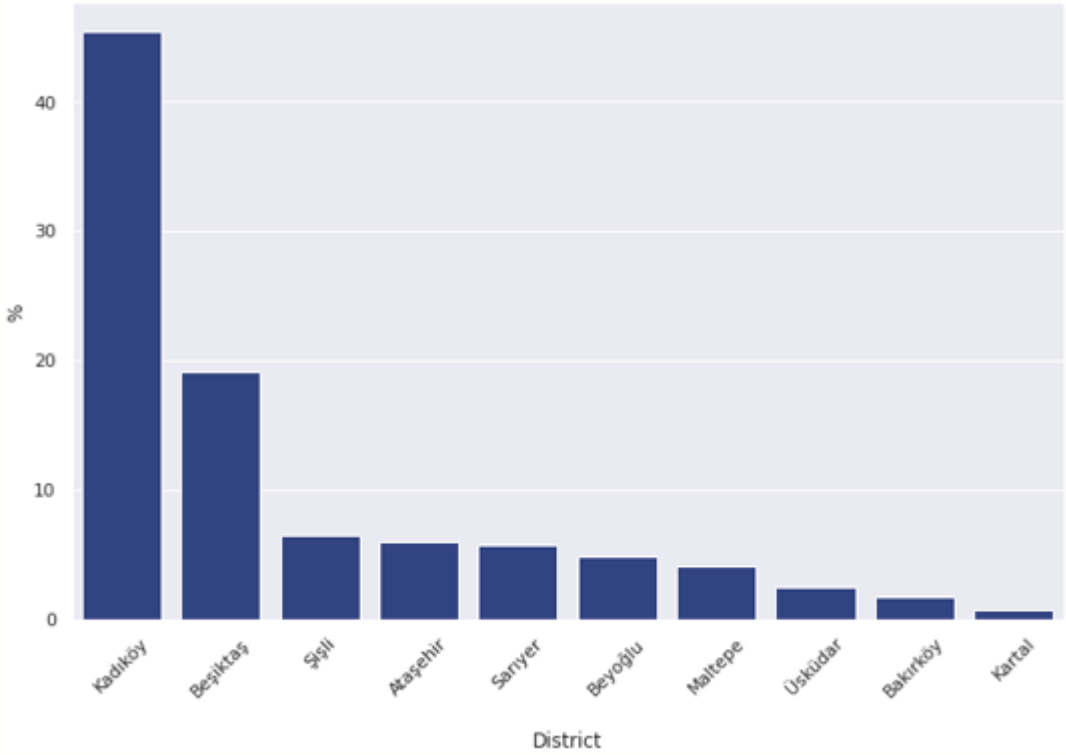


Figure 6: Top 10 Districts

In the top ten districts, Kadıköy is the number one, and there is a significant difference from the others. The customers from Kadıköy are almost half of the total customers (45%). The customers from Kadıköy and Beşiktaş constitute 65% of the total customers. The district variable can be used as a supporting variable of RFM in the following step to separate customer groups. On the other hand, another customer segmentation can be done for Kadiköy particularly.

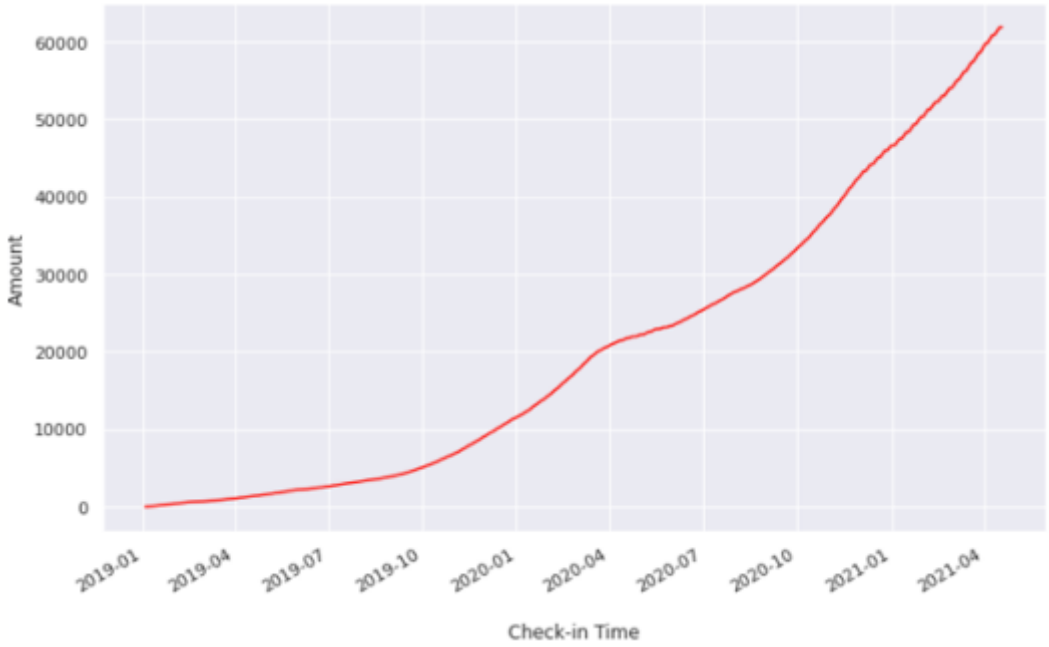


Figure 7: Total Cumulative Monetary

According to the dataset, the first transaction was realized January'19 and the last transaction was realized April'21. The data includes almost two and a half years of transaction records. During the first six-month period, company income progressed at a slower pace. However, during the last four-month period, the slope of the graph has drawn a steeper angle, which may be an indication that the company is moving forward successfully.

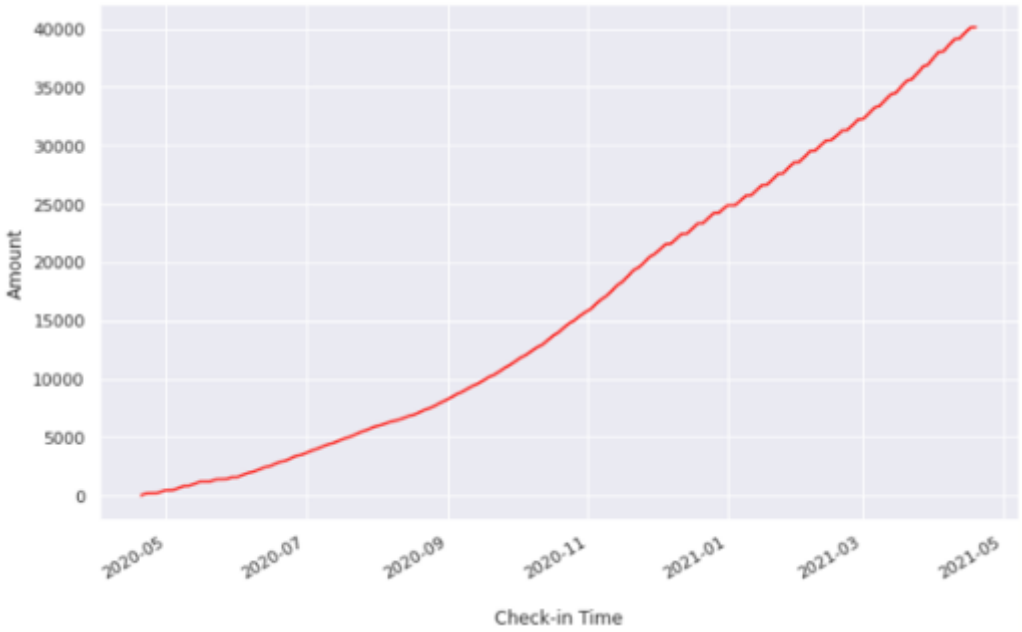


Figure 8: The Last Year Cumulative Monetary

The last one year starts on 2019-01-03 and ends on 2021-04-19. A similar situation is observed in this graph while comparing the graph above. The transactions have occurred more accelerated recently. However, there is no big deviation during the last year's slope, and it is more stable than the total cumulative monetary.

3 PROJECT DEFINITION

This section explains the problem statement, project objectives and project scope.

3.1 Problem Statement

Since dog owners do not have time to walk their dogs, they need a walker who they can trust. It is beneficial for marketing activities to learn about people's needs and habits using this application according to their preferences. They have trouble finding suitable walkers in the time intervals they seek, their proximity, the price they seek, etc. Knowing and understanding customers is the key to the success of any company. If companies have some insights about their customers, they can provide faster and clearer solutions to customers' needs. Thus, they can increase customer satisfaction and brand loyalty. One of the most popular methods is customer segmentation to analyse customer needs and to monitor customer conditions. Segmenting customers to develop better marketing strategies and making specific marketing moves for each customer group will increase the company's returns.

3.2 Project Objectives

The most important project objective is to cluster the customers to proper segments according to their features using machine learning algorithms and get some useful insights about customer segments. The second step of our project is providing a repetitive customer monitoring system and following up customer activity and behavior.

3.3 Project Scope

In this project, the last year of data is used for customer segmentation. At the first stage, customer segmentation is made with this piece of data, and then the movements of the customers are observed by re-examining them on a specified interval that is determined by the Doggo operation team using a special function. As mentioned in the previous part, RFM related features which are ownerid, walkingid, checkintime and amount are used in this project. Our model is not specific to Doggo company, it can be applied to any company.

4 METHODOLOGY

This study is carried out on Google Colab platform using Python 3.6.9 programming language and its libraries such as pandas, NumPy, matplotlib, seaborn, datetime, statistics, sk-learn.

4.1 Preprocessing

In this study, two different datasets are merged, and unnecessary columns are dropped. Even though there are some columns that include some missing values, there is no missing value in the columns that were used for this study. Exploratory data analysis is carried out to get better understanding.

After the exploratory data analysis, some errors are realized at the end of the amount column sorted by descending order. The numbers less than 16.48 can be misleading amounts and they affect the mean, and they are replaced with the median of the amount. As a next step, the customer that has the highest monetary value is dropped from data, because it leads to higher standard deviation, and it is commented as an outlier.

Two walk statuses are finished and canceled. Finished walks are used to calculate RFM because canceled walks are not proper transactions. There are some problematic rows in checkintime and ordercreatedtime such as check in time is not recent than order created time. Order creation is the previous step of check in. Therefore, a new column is generated which is called checkin_revized, and the latest date is selected from these two columns. Then, checkin_revized is normalized to drop hour information.

4.2 Generating RFM and Tenure Columns

After preprocessing, the next step is generating RFM and Tenure columns. Firstly, a new dataset that includes the last year of the whole dataset using datetime functions. A snapshot date is created, it is the next day of the last transaction. It is important because RFM values generally start with 1. Then, the new dataset which is called df_lastyear is grouped by ownerid and RFM columns are created using the Pandas library. Recency and tenure are generated by checkin_revized, frequency is generated by checkintime, monetary is generated by amount column. As a quick review, recency shows how many days before a customer made the last transaction, frequency indicates how many transactions are made in a given time period, monetary shows that a customer spends how much money in a given time period, tenure is a

lifetime of a customer, in other words, how many days has a customer been using this application?

ownerid	Recency	Frequency	MonetaryValue	Tenure
0062f2a6-878d-4100-9bea-2b50e266a56e	212	1	29.90	212
008fabab-f868-4019-8b50-4022025b51b2	226	27	807.40	415
00d2941f-7e9f-4422-a662-2d9012607d35	66	2	179.80	67
012069f5-41a9-472c-ad99-1d72ceca4966	316	6	169.90	347
0122b307-340f-41e4-88ac-bdd9e14eb26d	39	5	146.51	415

Figure 9: A Part of RFM and Tenure Dataset

After the data frame is generated, correlation between features is checked and Frequency and monetary are highly correlated, so frequency is excluded from RFM analysis. The analyses are completed using Recency, Monetary and Tenure features.

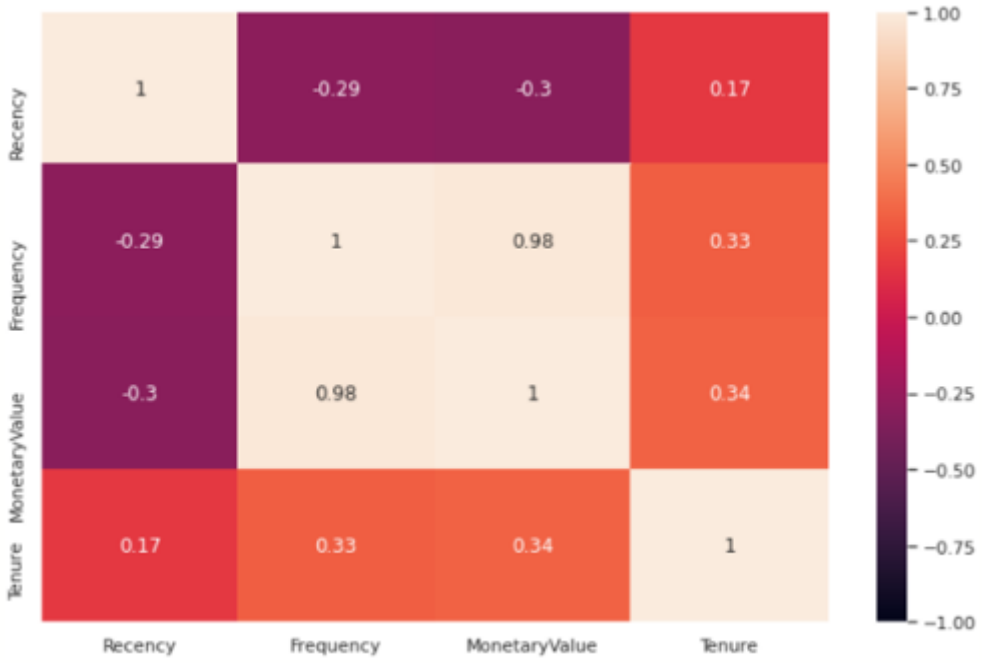


Figure 10: Correlation Heatmap

4.3 K-means

K-means is one of the most popular customer segmentation algorithms, so in this part application of k-means is explained. Various trials are done to find optimum solutions for customer segmentation of Doggo.

First step is normalization, it is assumed that a data is in Euclidean space. Each column is checked whether it is normally distributed or not. Shapiro Wilk Test is used for normality test in this study. According to results of Shapiro Wilk Test, Recency, Monetary and Tenure features are not normally distributed, and they are tried to normalise using four different transformation methods: square root, reciprocal, log, and boxcox one by one. All process failed, but the nearest values to normal distribution are obtained thanks to boxcox method based on Shapiro Wilk Test statistic values.

Secondly, scaling must be applied, because features have different unit such as recency unit is day, monetarvalue unit is Turkish lira. Therefore, scaling provides opportunities to compare different units. Two scaling methods which are Minmax scaler and Robust scaler are tried to get better results.

One of the important disadvantages of K-means is that the algorithm does not define the number of clusters. Thus, completing various trials with different numbers of clusters which is called k and defining a better k that provides maximum adaptivity with Doggo is a common approach. Two approaches are used to define cluster number, they are the Elbow method and Silhouette score. After the preprocessing of K-means, the Elbow method is applied with 25 options which are from 1 to 25. 8, 9 or 10 seem like a good option, but the silhouette score graph can be clearer to determine the number of clusters.

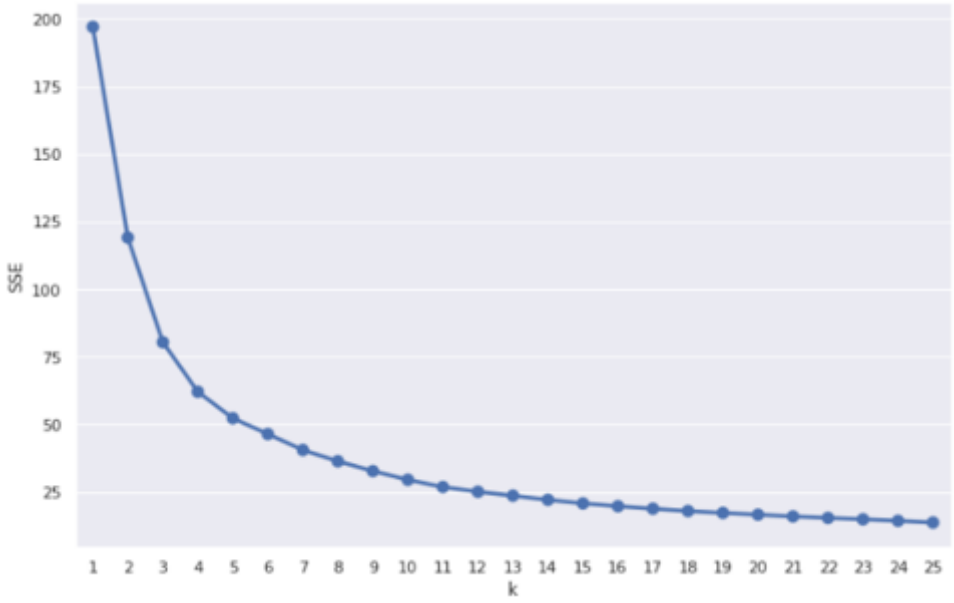


Figure 11: Result of the Elbow Method

A silhouette score graph is plotted, and it includes the cluster numbers from 2 to 25 like the Elbow method. There is a peak on 3, and the graph increases on 7, 8, 10, 11, 15. Then, 3, 8, 11 and 15 are determined as important clusters for this project, and they are checked by visualizing on different plots.

Figure 12: Result of the Silhouette Score

Each number in the important cluster numbers is tried using the k-means algorithm and its result is plotted in two dimensions. PCA is used to convert from 3 to 2 dimensions. Pair plots that include two features such as recency and monetaryvalue are generated on two dimensions to analyze the results of different cluster numbers and determine the best cluster number.

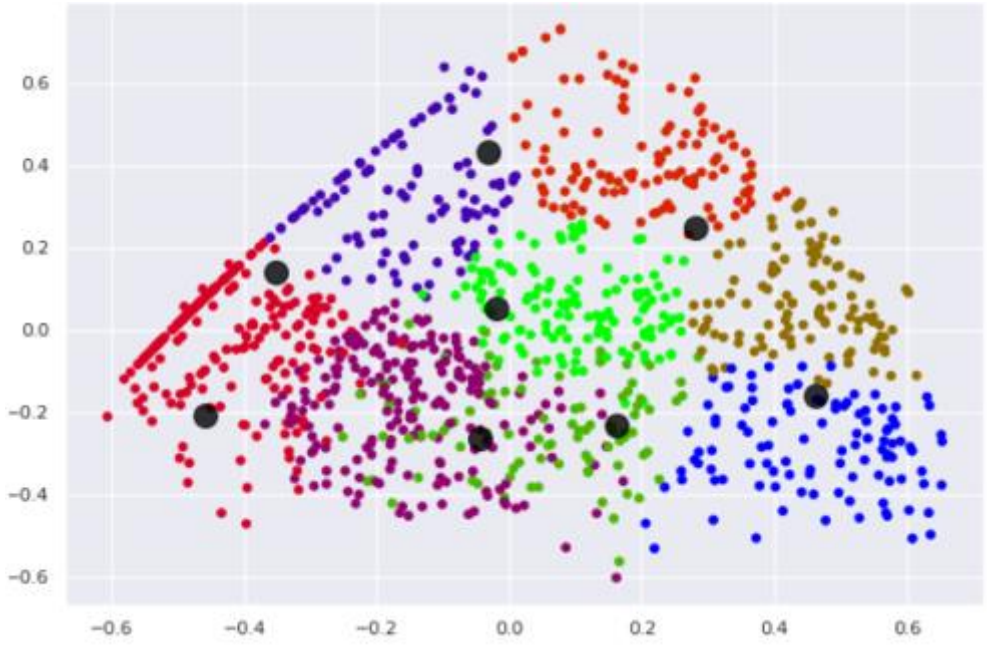


Figure 13: The Result of the K-means Clustering with 8 Clusters

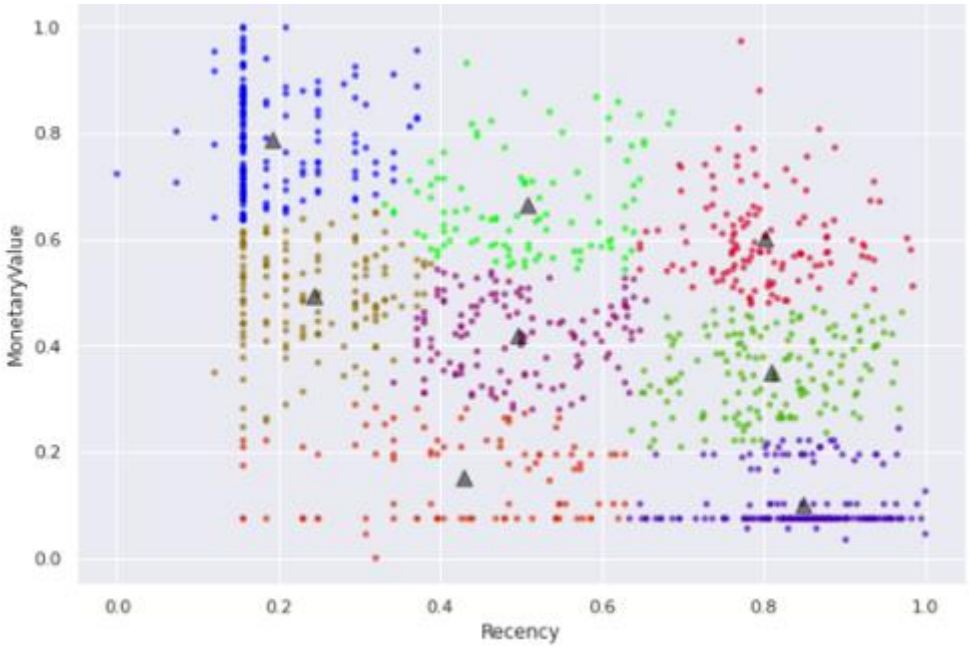


Figure 14: An Example of Pair Plots

As following step, descriptive statistics are calculated for each k and each cluster. Descriptive statistics includes each feature's mean, standard deviation and count information and maximum value of monetaryvalue is calculated. It provides insights to determine cluster

label which will be used in marketing campaigns or projects. There is an example of descriptive statistics for k=8.

Number of Cluster 8

	Recency		MonetaryValue			Tenure		
	mean	std	mean	std	max	mean	std	count
Cluster								
0	10.0	12.0	4713.0	3586.0	15907.0	529.0	153.0	107
1	52.0	27.0	67.0	41.0	189.0	69.0	43.0	105
2	179.0	60.0	820.0	1194.0	12731.0	285.0	123.0	205
3	209.0	58.0	53.0	32.0	140.0	281.0	147.0	220
4	10.0	7.0	276.0	195.0	1096.0	30.0	20.0	114
5	7.0	5.0	2181.0	1528.0	9106.0	160.0	64.0	110
6	48.0	33.0	503.0	533.0	3000.0	534.0	108.0	93
7	40.0	22.0	575.0	479.0	4069.0	158.0	59.0	131

Figure 15: An Example of Descriptive Statistics Table

After a plenty of trials, cluster labels are determined based on the visualizations above and statistical values. Further investigations are carried out to check the correction rate of clustering and distances are calculated for each cluster point to all cluster centroids. Because of that, the smallest distance and sum of squared distance are used for this analysis. If the distance between a point and its cluster centroids is smallest, it can be stated that the clustering is probably correct in a way of theoretical thinking. To visualize the distance analysis and measure quality of a model, four functions are used. The first one is used to calculate cardinality which means how many members are in a cluster. If the cardinality is similar among clusters, the model can be commented as a good model. The second one is used to calculate magnitude which means that total point-to-centroid distance. The best cluster has the lowest magnitude value. The third is used to draw the first two function plots. The last one is a shortcut of these three functions. This function will be used to check the quality of the most appropriate model.

Moreover, similarity analysis is completed to ensure each customer is in the correct cluster, the similarity between each point in a cluster. It can be helpful to realize a wrong cluster point. As a first step, a similarity function is generated. Next, similarities are calculated for each cluster and some statistical features are calculated. Similarity heat maps are plotted to visualize the analysis.

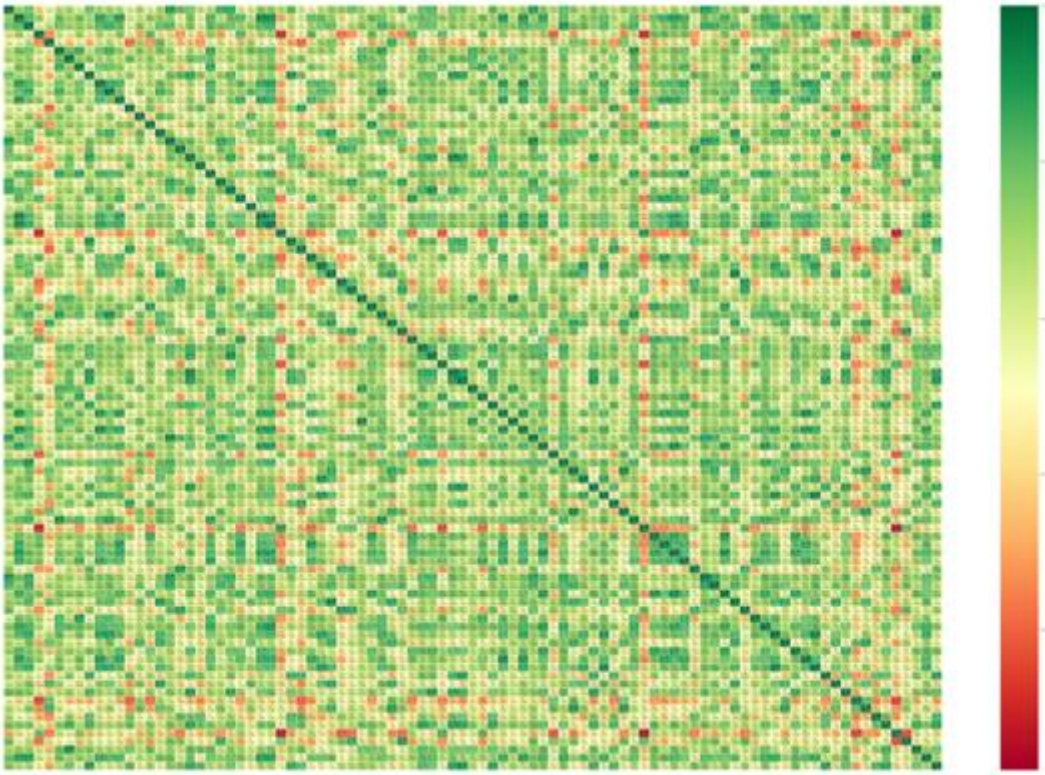


Figure 16: Similarity Heat Map

4.4 Gaussian Mixture Model (GMM)

GMM algorithm is used to find better convergence with Doggo customers as a different method. Three covariance types which are spherical, diagonal, and full are used to provide variety. Same steps that applied to K-means are applied to GMM as well. BIC and AIC methods are used to determine cluster numbers. Different cluster numbers are determined as important clusters and a GMM algorithm is run for each number of clusters. 2D and 3D plots are drawn, and descriptive statistics are calculated for each k.

Additionally, GMM provides a way to calculate the probability that a data point belongs to the relevant cluster. Mean, standard deviation, min and max values of each cluster probability are calculated. With threshold values such as less than 0.5 or 0.8, it is tried to make inferences about the clusters. The result of this analysis is to try to make the most appropriate customer segmentation with the support from the Doggo team.

4.5 Repetitive RFM and Tenure Interval Analysis

A function is created to monitor a customer activity during a specific period. The purpose of creating this function is to help the Doggo operation team. Thanks to it, a Doggo operation team member can access information about a customer's behavior with a table and a graph. The function goes back one day, starting from today, in a specified period, for example, in a 30-day period. In other words, the 30-day frame is shifted backwards in history one day each. The change in one of the metrics which are recency, frequency, monetary and tenure over time is reflected on the graph.

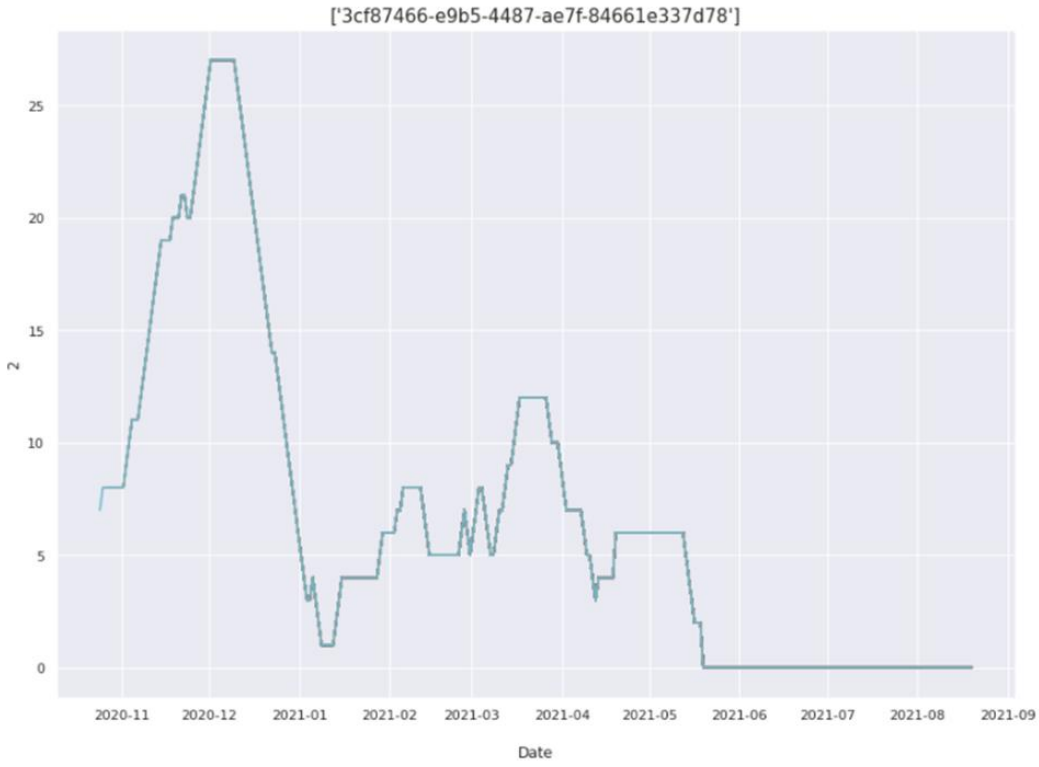


Figure 17: Repetitive RFM and Tenure Interval Analysis Graph

In the above graph, the frequency of the customer with the code "3cf87466-e9b5-4487-ae7f-84661e337d78" is plotted over time in 30-day periods. One day goes back from today's date (19/08/2021) by 300 days. In the function, the customer code, how many days the period will be, how many days to go back and which of the Recency, frequency, monetary and tenure metrics are selected are parameters, they can be shaped according to the investigation. The reason why the frequency is zero after 2021-05 in the graph is because the old data is used and there is no data in that date range.

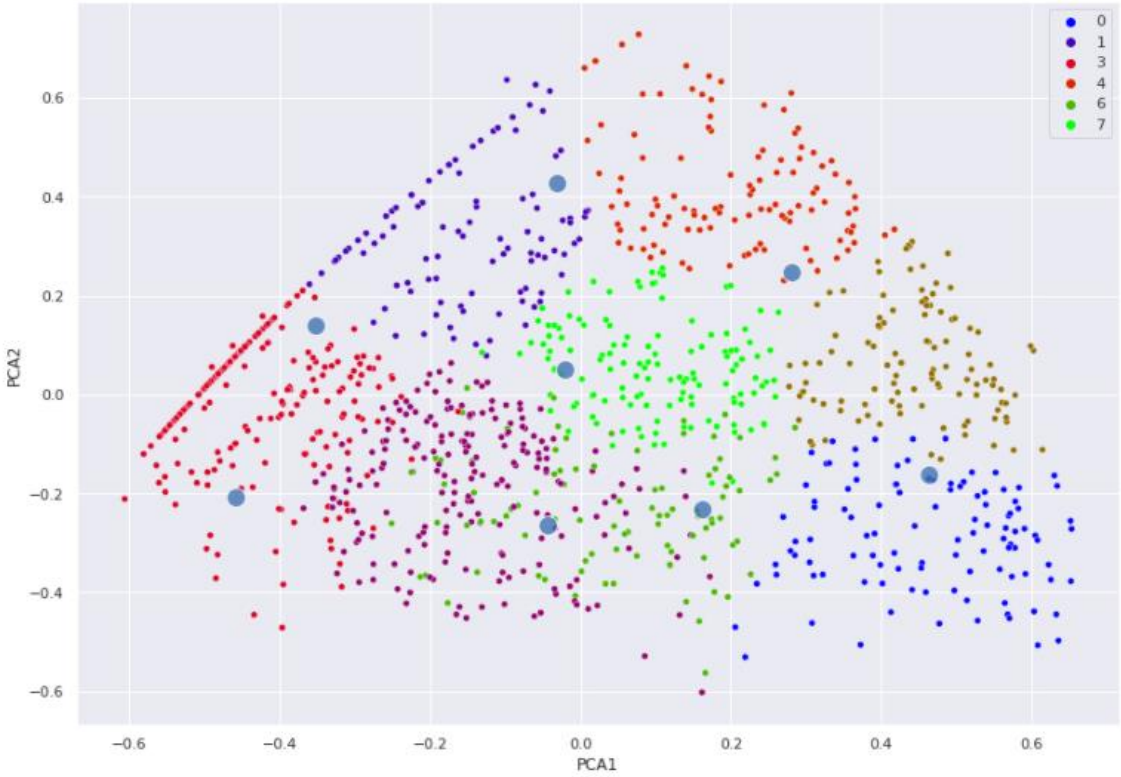
The function has four parameters which are P, R, F, and owner. One of the most important parameters is interval days which is called P and the function shows selected customer information going backwards periodically by the specified number of days. R provides to adjust that how many days to go back from today. Moreover, users are able to select any features of RFM and Tenure for a specified customer. F provides to select a feature such as frequency, and owner provides to select a customer. For example, if the Doggo operation team would like to see a 30-day period repetitively and frequency of a customer, they can monitor any changes and get useful insights about a customer's condition.

5 RESULTS

Customer segmentation is a crucial method to identify customers and generate appropriate marketing strategies. In this study, the optimum results are tried to get for Doggo which is a startup firm that has a mobile application. In accordance with this purpose, for scaling two different methods are used to prepare data to machine learning algorithms which are K-means and GMM. Four different cluster numbers are tried to find optimum result and three different covariance methods, and three different cluster numbers are tried to provide variety. Various trials are completed to find the best option that is fitted with Doggo customer profiles.

Some random customer id and their cluster numbers and features are selected from each model and the Doggo operation team analyzes these customers and results. On the other hand, meetings are carried out with Doggo in weeks and the project is adjusted based on Doggo requests and needs. According to discussions with Doggo, its operation team correction analysis, and distance, similarity analysis of this study, the most appropriate model is determined as k-means clustering with 8 clusters by common consent. Its two-dimensional plot and descriptive statistics are shown below. Labeling is completed using the Doggo jargon and marketing strategies will be managed based on these labels.

Here is the k-means clustering with eight clusters two-dimensional plot and its descriptive statistics table.



Number of Cluster 8

	Recency		MonetaryValue			Tenure		
	mean	std	mean	std	max	mean	std	count
Cluster								
0	10.0	12.0	4713.0	3586.0	15907.0	529.0	153.0	107
1	52.0	27.0	67.0	41.0	189.0	69.0	43.0	105
2	179.0	60.0	820.0	1194.0	12731.0	285.0	123.0	205
3	209.0	58.0	53.0	32.0	140.0	281.0	147.0	220
4	10.0	7.0	276.0	195.0	1096.0	30.0	20.0	114
5	7.0	5.0	2181.0	1528.0	9106.0	160.0	64.0	110
6	48.0	33.0	503.0	533.0	3000.0	534.0	108.0	93
7	40.0	22.0	575.0	479.0	4069.0	158.0	59.0	131

Figure 18: K-means Clustering with 8 Clusters

After the selection of the most appropriate model, the quality functions that are mentioned in chapter 4.3 are used to check K-means Clustering with 8 Clusters.

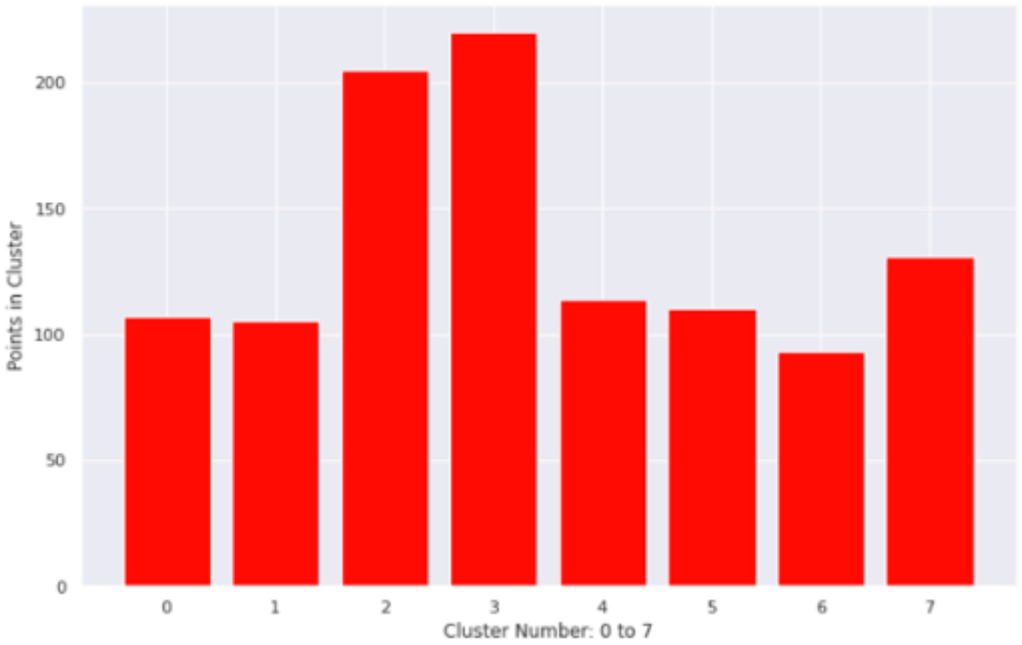


Figure 19: K-means Clustering with 8 Clusters' Cardinality Plot

According to the cardinality plot, except cluster 2 and 3, the other clusters' cardinalities are similar. In the literature, all clusters are expected to have approximately the same number of elements.

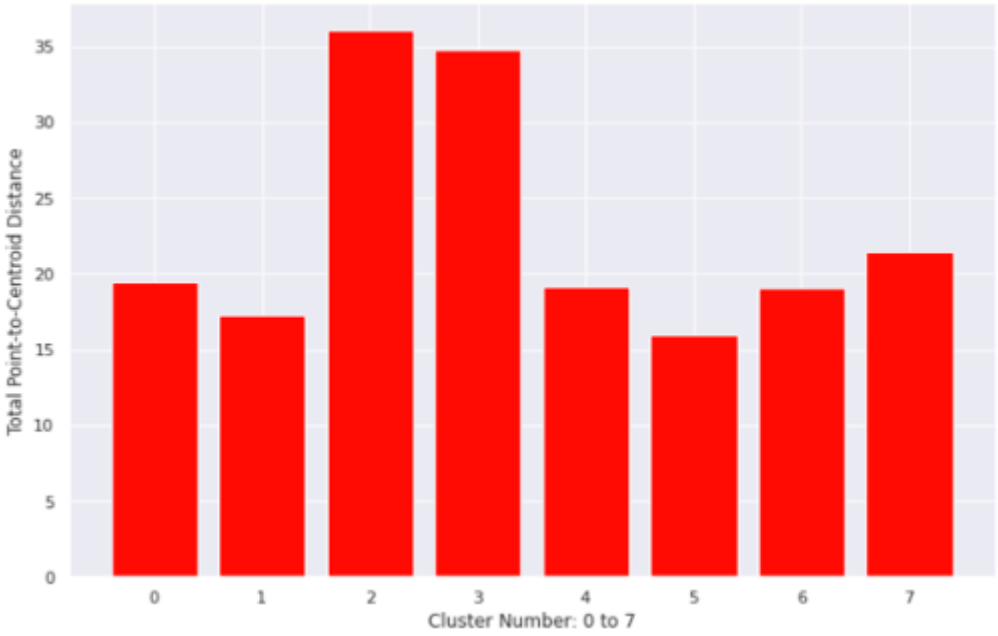


Figure 20: K-means Clustering with 8 Clusters' Magnitude Plot

According to the magnitude plot, except cluster 2 and 3, the other clusters' magnitudes are similar. Cluster 2 and cluster 3 have higher total point-to-centroid distance, but they have a higher number of members. To get better understanding, magnitude per cardinality plot is drawn below. Thanks to it, distribution can be checked in a clearer way.

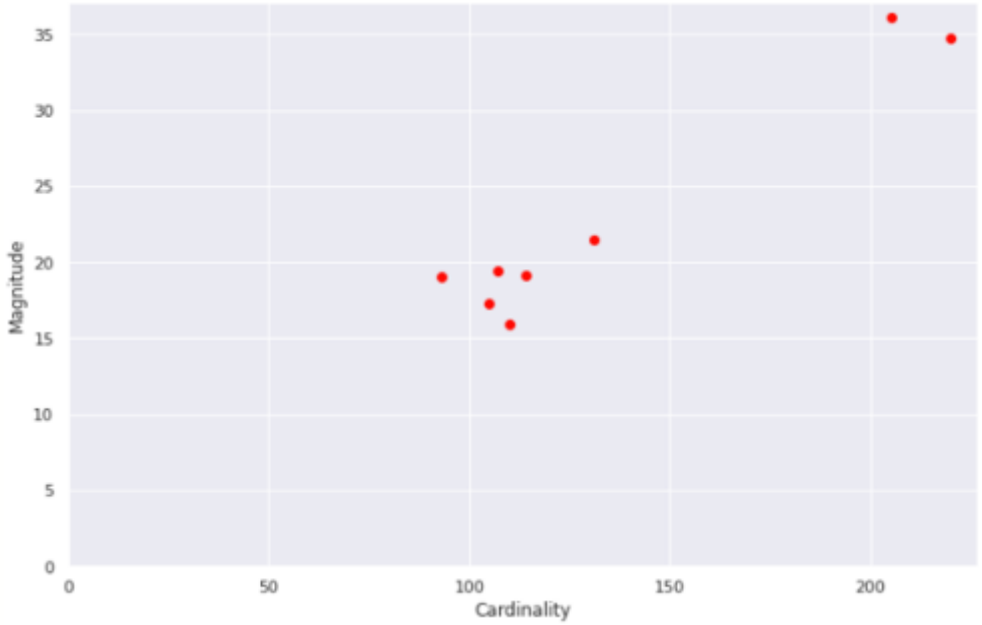


Figure 21: K-means Clustering with 8 Clusters' Magnitude vs Cardinality Plot

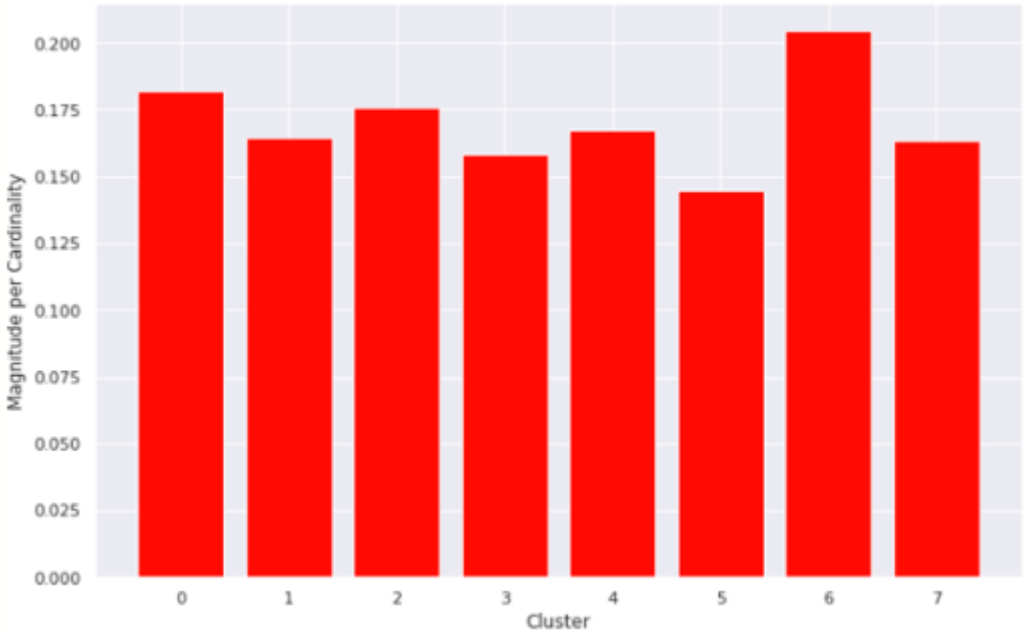


Figure 22: K-means Clustering with 8 Clusters' Magnitude per Cardinality Plot

According to the last two plots above, the clusters' cardinality and magnitude are shown in a scatter plot. Then, magnitude per cardinality is calculated and a bar plot shows the results. Based on the bar plot, cluster 5 has the best distance rate and the worst cluster is number 6. On the other hand, the similarity analysis supports this result and cluster 6 has lowest similarity means.

As a further study, customers that have lower similarity scores can be detected using the similarity heat map and these customers can be investigated in detail to find the right cluster for each cluster with Doggo operation team support.

5.1 The Labels of K-means Clustering with 8 Clusters

The labels of k-means clustering with 8 clusters

Cluster 0: VIP (LR-HM-HT)

Cluster 1: Newest Passive (HR-LM-LT)

Cluster 2: Most Expensive Churn (HR-HM-HT)

Cluster 3: Just First Walk (HR-LM-HT)

Cluster 4: Newcomers (LR-LM-LT)

Cluster 5: New VIP (LR-HM-LT)

Cluster 6: Old Passive (HR-LM-HT)

Cluster 7: New Passive (HR-LM-LT)

For above table, VIP, Newest Passive, Most Expensive Churn, Just First Walk, Newcomers, New VIP, Old Passive, New Passive are labels and the abbreviations such as (LR-HM-HT) are features of each cluster according to the recency, monetary and tenure. H corresponds high, L corresponds low.

Cluster 0 which label is VIP has low recency, high monetary, and high tenure. It means that these customers are active, old customers and they provide good returns to Doggo. Renewals, newer products may be offered to these cluster members.

Cluster 1 which label is Newest Passive has high recency, low monetary, and low tenure. It means that these customers are passive, they have just joined the Doggo family and have not yet made a good return. More discounts may be offered, and they get motivated to use the doggo app more.

Cluster 2 which label is Most Expensive Churn has high recency, high monetary, and high tenure. It means that these customers are very passive, old customers and they provided good returns in the past, but they do not use the Doggo application anymore. Therefore, they are the most expensive loss for the Doggo. A friendly and incentive message or notification that includes a special offer may be sent to them.

Cluster 3 which label is Just First Walk has high recency, low monetary, and high tenure. Doggo gives a welcome gift which is a free walk to each new customer. However, in the Doggo system, these free walks are not shown as free, so they have monetaryvalue. Most customers of this cluster have used the free walk then they have not used it anymore. It can be tough for them to get back in the door again. However, a friendly and incentive message or notification that includes a special offer may be sent to them as well.

Cluster 4 which label is Newcomers has low recency, low monetary, and low tenure. It means that these customers have just joined the Doggo family and the newest members. Renewals, newer campaigns, and products may be offered.

Cluster 5 which label is New VIP has low recency, high monetary, and low tenure. It means that these customers are active, new customers and they provide good returns to Doggo and have potential to be VIP in the future. Messages or notifications may be sent stating that our cooperation is going well. Renewals, newer campaigns, and products may be offered as well.

Cluster 6 which label is Old Passive has high recency, low monetary, and high tenure. It means that these customers are passive, old customers and they have not made good returns to Doggo. These customers are familiar with Doggo and they must be encouraged to buy new walks or other products. Some special and incentive campaigns may be effective.

Cluster 7 which label is New Passive has high recency, low monetary, and low tenure. It means that these customers are passive, older than cluster 1, but they are still new customers for the Doggo. Messages and notifications may be sent to help them get to know Doggo and increase the frequency of their use of the application. Welcome campaigns may also be offered.

REFERENCES

Wei, J., Lin, S., & Wu, H. (2010). A review of the application of RFM model. *African Journal of Business Management. 4*(19), 4199-4206. https://doi.org/10.5897/AJBM.9000026

Dogan, O., Aycin, E., Bulut, Z.A. (2018). Customer segmentation by using RFM model and clustering methods: A case study in the retail industry. *International Journal of Contemporary Economics and Administrative Sciences. 8*(1), 1-19. ISSN: 1925 – 4423

Christy, A.J., Umamakeswari, A., & Priyatharsini, L., & Neyaa, A. (2018). RFM ranking – An effective approach to customer segmentation. Journal of King Saud University - *Computer and Information Sciences, 32*(10), 1215. https://doi.org/10.1016/j.jksuci.2018.09.004.

Miglautsch, J. (2000). Thoughts on RFM scoring. *J Database Mark Cust Strategy Manag 8*(1), 67-72. https://doi.org/10.1057/palgrave.jdm.3240019

Dogan, O., Aycin, E., Bulut, Z.A. (2018). Customer segmentation by using RFM model and clustering methods: A case study in the retail industry. *International Journal of Contemporary Economics and Administrative Sciences. 8*(1), 1-19. ISSN: 1925 – 4423

Ducange, P., Pecori, R., & Mezzina, P. (2017). A glimpse on big data analytics in the framework of marketing strategies. *Soft Computing, 22*(1), 325–342. https://doi.org/10.100s7/s00500-017-2536-4

Curto, J. (n.d.). *Customer Segmentation · Customer Analytics with R.* Retrieved June 13, 2021, from https://josepcurtodiaz.gitbooks.io/customer-analytics-with-r/content/chapter7.html

Miglautsch, J. (2000). Thoughts on RFM scoring. *J Database Mark Cust Strategy Manag 8*(1), 67-72. https://doi.org/10.1057/palgrave.jdm.3240019

Sarvari, P.A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes, 45*(7), 1129-1157. https://doi.org/10.1108/K-07-2015-0180

Anitha, P., & Patil, M. M. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*. Published. https://doi.org/10.1016/j.jksuci.2019.12.011

Christy, A.J., Umamakeswari, A., & Priyatharsini, L., & Neyaa, A. (2018). RFM ranking – An

effective approach to customer segmentation. *Journal of King Saud University Computer and Information Sciences, 32*(10), 1215. https://doi.org/10.1016/j.jksuci.2018.09.004.

Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36*(2), 451–461. https://doi.org/10.1016/s0031-3203(02)00060-2

Lozano, J.A., Pena, J.M., Larranaga, P., (1999). An empirical comparison of four initialization methods for the k-means algorithm, *Pattern Recognition Lett*, *20*(10), 1027–1040. https://doi.org/10.1016/S0167-8655(99)00069-0

Foley, D. (2021, January 3). *Gaussian Mixture Modelling (GMM) - Towards Data Science*. Medium. https://towardsdatascience.com/gaussian-mixture-modelling-gmm-833c88587c7f

Dogan, O., Aycin, E., Bulut, Z.A. (2018). Customer segmentation by using RFM model and clustering methods: A case study in the retail industry. *International Journal of Contemporary Economics and Administrative Sciences. 8*(1), 1-19. ISSN: 1925 – 4423