ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

School of Information Sciences & Technology
Department of Informatics
Master of Science in Data Science

Thesis

Customer Spending Behavior

Aigli Kyriakoudi

Academic Supervisor: Professor Vasilios Vassalos

Athens, October 2019


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Abstract

The objective of this thesis it to analyze customer segmentation techniques and provide an implementation in real banking data from the National Bank of Greece. The provided datasets contain one month's transaction information of customers with credit and debit cards. In these datasets, customer segmentation is used to provide meaningful segments of customers based on their spending habits. Firstly, the theoretical framework for customer segmentation and clustering is provided. Then, the data are examined and processed in Apache Spark using Python API. This includes a categorization of the transactions, which is achieved based on the market segment the supplier belongs to and using the available Merchant Category Codes for each transaction, and the creation of the percentages of each customer's total spending in each category. Afterwards, using Python, dimensionality reduction is achieved with Principal Component Analysis (PCA) and then K-Means clustering is implemented on the extracted features. The best parameter k, which indicates the number of clusters, is selected with internal validation metrics suitable for clustering. Finally, all formed clusters are further examined based on their customer's spending, purchases and available demographical data to provide a customer profile for each cluster.

Keywords: data mining, clustering, customer segmentation, K-Means

Contents

Abstract 3
Index of Tables 6
Index of Figures 7
Chapter 1 Introduction 9
1.1 Overview 9
1.2 Related works 9
Chapter 2 Theoretical framework 15
2.1 Data mining in customer segmentation 15
2.2 Clustering algorithms 16
2.3 Preprocessing steps for clustering algorithms 17
2.4 Clustering evaluation 19
2.5 Cluster profiling 21
Chapter 3 Customer segmentation in banking data 23
3.1 Programming languages and software 23
3.2 Data description 23
3.3 Exploratory data analysis 26
3.4 Data aggregations 31
3.5 Feature selection 33
3.6 Clustering algorithm implementation and evaluation 36
Chapter 4 Results 39
4.1 Identification and characterization of clusters 39
4.2 Profiling of clusters 42
4.2.1 Credit dataset 43
4.2.2 Debit cards 52
4.2.3 Profiling based on region 62
4.3 Customers with both credit and debit cards 69
4.4 Debit customers with credit cards 70
Chapter 5 Conclusion 75
Bibliography 77

Index of Tables

Table 1 10 most popular codes for the credit dataset with their description and general category 26
Table 2 10 most popular codes for the debit dataset with their description and general category 27
Table 3 Statistics for credit and debit datasets 27
Table 4 10 codes with higher total spending for the credit dataset with their description and general category 28
Table 5 10 codes with higher total spending for the debit dataset with their description and general category 29
Table 6 Components matrix for the credit dataset 35
Table 7 Components matrix for the debit dataset 35
Table 8 Component interpretation for the credit dataset 35
Table 9 Component interpretation for the debit dataset 36
Table 10 Silhouette coefficient for each cluster of the two datasets 38
Table 11 Distribution of the derived clusters for each dataset 38
Table 12 Cluster centers for the credit dataset 39
Table 13 Cluster centers for the debit dataset 39
Table 14 Mean percentage of total spending by category for each cluster of the credit dataset 40
Table 15 Mean percentage of total spending by category for each cluster of the debit dataset 41
Table 16 Mean purchase amount per category for each cluster of the credit dataset 41
Table 17 Mean purchase amount per category for each cluster of the debit dataset 41
Table 18 Cluster labels 42
Table 19 Percentage of customers per region for each cluster of the credit dataset 62
Table 20 Percentage of customers per region for each cluster of the debit dataset 62
Table 21 Components matrix for debit customers with credit cards 70
Table 22 Silhouette coefficient for each cluster 71
Table 23 Cluster centers 72
Table 24 Percentage of customers for the debit dataset and the debit customers with credit cards 73

Index of Figures

Figure 1 Percentages of customers belonging in each category of the attributes 'Gender', 'Marital Status' and 'Age Category' for credit and debit datasets 29
Figure 2 Percentages of customers belonging in each category of the attributes 'Occupation' and 'Educational Level' for credit and debit datasets..... 30
Figure 3 Percentages of customers living in each of the 13 regions in Greece for credit and debit datasets 31
Figure 4 Percentage of number of purchases per category for credit and debit card users 32
Figure 5 Percentage of amount in Euro spent per category for credit and debit card users 32
Figure 6 Correlation plot of the 17 attributes for the credit dataset..... 33
Figure 7 Correlation plot of the 17 attributes for the debit dataset..... 34
Figure 8 Clustering evaluation graphs for the credit dataset for values of k from 2 to 10 37
Figure 9 Clustering evaluation graphs for the debit dataset for values of k from 2 to 10 37
Figure 10 Percentages of total spending and purchases per category for cluster 1 / Telcos 43
Figure 11 Percentages of the values of categorial attributes for cluster 1 / Telcos 44
Figure 12 Percentages of total spending and purchases per category for cluster 2 / Food 44
Figure 13 Percentages of the values of categorial attributes for cluster 2 / Food..... 45
Figure 14 Percentages of total spending and purchases per category for cluster 3 / All Categories 45
Figure 15 Percentages of the values of categorial attributes for cluster 3 / All Categories 46
Figure 16 Percentages of total spending and purchases per category for cluster 4 / Services 46
Figure 17 Percentages of the values of categorial attributes for cluster 4 / Services..... 47
Figure 18 Percentages of total spending and purchases per category for cluster 5 / Automotive 47
Figure 19 Percentages of the values of categorial attributes for cluster 5 / Automotive .. 48
Figure 20 Percentages of total spending and purchases per category for cluster 6 / Entertainment..... 48
Figure 21 Percentages of the values of categorial attributes for cluster 6 / Entertainment 49
Figure 22 Percentages of total spending and purchases per category for cluster 7 / Travel 50
Figure 23 Percentages of the values of categorial attributes for cluster 7 / Travel..... 50
Figure 24 Percentages of total spending and purchases per category for cluster 8 / Shopping 51
Figure 25 Percentages of the values of categorial attributes for cluster 8 / Shopping..... 51
Figure 26 Percentages of total spending and purchases per category for cluster 9 / Clothing..... 52
Figure 27 Percentages of the values of categorial attributes for cluster 9 / Clothing 52

Figure 28 Percentages of total spending and purchases per category for cluster 1 / Entertainment 53
Figure 29 Percentages of the values of categorial attributes for cluster 1 / Entertainment 53
Figure 30 Percentages of total spending and purchases per category for cluster 2 / Food 54
Figure 31 Percentages of the values of categorial attributes for cluster 2 / Food 54
Figure 32 Percentages of total spending and purchases per category for cluster 3 / Automotive 55
Figure 33 Percentages of the values of categorial attributes for cluster 3 / Automotive .. 55
Figure 34 Percentages of total spending and purchases per category for cluster 4 / Shopping 56
Figure 35 Percentages of the values of categorial attributes for cluster 4 / Shopping 56
Figure 36 Percentages of total spending and purchases per category for cluster 5 / All Categories 57
Figure 37 Percentages of the values of categorial attributes for cluster 5 / All Categories 57
Figure 38 Percentages of total spending and purchases per category for cluster 6 / Gambling 58
Figure 39 Percentages of the values of categorial attributes for cluster 6 / Gambling 58
Figure 40 Percentages of total spending and purchases per category for cluster 7 / Clothing 59
Figure 41 Percentages of the values of categorial attributes for cluster 7 / Clothing 59
Figure 42 Percentages of total spending and purchases per category for cluster 8 / Telcos 60
Figure 43 Percentages of the values of categorial attributes for cluster 8 / Telcos 60
Figure 44 Percentages of total spending and purchases per category for cluster 9 / Medical 61
Figure 45 Percentages of the values of categorial attributes for cluster 9 / Medical 61
Figure 46 Map of cardholders for cluster label ‘All Categories’ 63
Figure 47 Map of cardholders for cluster label ‘Automotive’ 64
Figure 48 Map of cardholders for cluster label ‘Clothing’ 64
Figure 49 Map of cardholders for cluster label ‘Entertainment’ 65
Figure 50 Map of cardholders for cluster label ‘Food’ 65
Figure 51 Map of cardholders for cluster label ‘Shopping’ 66
Figure 52 Map of cardholders for cluster label ‘Telcos’ 66
Figure 53 Map of cardholders for cluster label ‘Services’ 67
Figure 54 Map of cardholders for cluster label ‘Travel’ 67
Figure 55 Map of cardholders for cluster label ‘Gambling’ 68
Figure 56 Map of cardholders for cluster label ‘Medical’ 68
Figure 57 Percentage of total spending per category for customers with both cards 69
Figure 58 Clustering evaluation graphs for debit customers with credit cards for values of k from 2 to 10 71
Figure 59 Percentages of total spending per category 72

Chapter 1 Introduction

1.1 Overview

The evolution of technology resulted in companies being able to acquire huge amounts of data. Utilizing that data enables companies to become increasingly customer-centric and can provide them with useful insights so they could target specific groups of customers. An effective method to process the data and be able to identify different groups of customers that are similar in specific areas relevant to marketing, such as age, gender, interests and purchasing habits is customer segmentation.

The main goal of this thesis is to analyze the spending habits of customers that have credit and debit cards issued by the National Bank of Greece. This analysis offers insight on the main categories that customers tend to spend their money on and allows the bank, or a company with data about spending behavior, to divide customers into meaningful segments according to their behavior. Each segment is then further analyzed based on the demographic profile of the customers and then the general profile of each segment can be used to help the bank gain knowledge of how each customer should be targeted based on their profile. There are two datasets available, one for users with credit cards and one for users with debit cards and the results are compared to see if there are differences in the spending patterns of customers based on the type of card they use. Debit cards allow bank customers to withdraw money from the funds they have deposited at the bank for their transactions, while with credit cards, customers borrow money from the bank up to a certain limit in order to purchase items or withdraw cash.

The workflow of the thesis started with research of related works in customer segmentation in general and customer spending behavior in particular, as well as a review of the techniques used to obtain the segments in each case. Then, the data from the bank were explored and the most suitable technique to handle this type of data was implemented.

1.2 Related works

This chapter focuses on relevant previous research and reviews the algorithms implemented and the obtained results. Namvar et al. (2010) [1] implemented a two-phase clustering algorithm on banking data to segment customers based on their Recency, Frequency, Monetary scores (RFM) and their lifetime value (LTV). Recency is the period since the last purchase, with a lower value corresponding to a higher probability of the customer making another purchase, Frequency is the number of purchases made within a certain period with higher frequency indicating greater loyalty and Monetary is the money

spent during a certain period, with higher values indicating customers that spend more money in that period. LTV is defined as the present value of the future cash flows attributed to the customer during his/her entire relationship with the company. The first phase of the algorithm is a K-Means clustering to divide customers into segments based on their RFM. Then using demographic data, each cluster is partitioned to new clusters and finally a profile for each customer is created using their LTV. The dataset consisted of 491 customers with 25 attributes. A self-organizing map (SOM) clustering was used to measure the importance of demographic variables and resulted in keeping 'educational level', 'occupation level' and 'age' for the next step. The RFM values for each customer were calculated based on their transactions with the time period beginning six months earlier than the last existing transaction in the database. K-Means clustering was implemented on the RFM data and three clusters were obtained. Then each cluster was internally segmented based on the three important demographic variables resulting in nine clusters. The LTV of each customer was calculated based on their current value and their potential value. The current value was calculated using the balance sheet for the last six months and the potential value was calculated based on two things: the probability that a customer would use a service and the profit the company obtains from the customer using that service. The average profit was predicted by the bank's managers, while the probabilities were predicted using a neural network. The predicted LTV for each customer was used to calculate mean value of LTV within each cluster. Based on the obtained cluster profiles, the bank can understand customer value in each cluster and establish better customer relationship management strategies.

Shahadat Hossain (2017) [2] focused on the comparison of centroid based and density based clustering algorithms for customer segmentation. The dataset contains 440 clients of a wholesale distributor regarding their annual spending on six different commodities (Fresh, Milk, Grocery, Frozen, Detergents_paper and Delicassen). K-Means and DBSCAN were used as the centroid based and density based algorithms respectively. K-Means was implemented, after normalization of the data, with values of k from 2 to 5 based on Euclidean distance and Manhattan distance although in the end they do not show differences in the results. DBSCAN needs as input the radius of the neighborhood within which points are considered neighbor to each other and the number of minimum points needed to be considered as a cluster. In each obtained cluster the points are categorized as core, border or noise points. Noise points remain unclustered at the end of the algorithm and can be used to identify customers with unusual spending behavior. K-Means is faster than DBSCAN but DBSCAN detects anomalous customers with different spending habits which helps in ensuring customer satisfaction and optimal profit.

Tripathi et al. (2018) [3] implemented K-Means and agglomerative hierarchical clustering on a dataset of 200 mall customers. The attributes describing each customer are genre, age, annual income and spending score, which is calculated by the company based on the

customers' spending habits. For K-Means, the optimal number of clusters was chosen using the within cluster sum of squared errors (WCSS) with k=5 clusters being the number after which the change in WCSS is no longer visible. Hierarchical clustering is much slower but provides a better visualization since it shows exactly at which point the clusters merged or split. Both algorithms resulted in 5 clusters which show the difference on the users' spending score based on their annual income.

RFM analysis is also used by Aggelis & Christodoulakis (2005) [4] in active e-banking users to identify high-response customers. The period under study is January 1st to December 12th of the year 2002. The analysis is based on the pyramid model, which groups customers based on the revenue they generate in four categories: small, medium, big and top. Customers exhibiting high RFM score, calculated as the sum of the three variables, should normally conduct more transactions and result in higher profit for the bank and thus the obtained clusters should correspond to the four pyramid categories. Clustering was implemented using K-Means and Two-Step algorithm. Two Step is a cluster analysis of SPSS Clementine where the first step makes a single pass through the data, during which it compresses the raw input data into a set of sub-clusters and the second step uses a hierarchical clustering method to progressively merge the sub-clusters into larger and larger clusters, without requiring another pass through the data. Although hierarchical clustering does not have good performance with large datasets, Two-Step's initial pre-clustering makes hierarchical clustering fast even for large datasets. There are some differences in the two methods, but it is concluded that the knowledge of the RFM score of active e-banking users can rank them according to the four categories and thus, the bank can identify the most important customers.

According to R. Di Clemente et al. (2018) [5], credit card transactions combined with mobile phone records can be used to reveal spending habits or life styles that are consistent with the demographic data of the individuals. The time period under study is 10 weeks, starting at the 1st week of May 2015 and the dataset consists of 150,000 users who live in Mexico City. The features include a used identification string, the timestamp of the transactions, the amount and the Merchant Category Code (MCC) which provides a label for each transaction based on its category, e.g. 'Grocery Stores, Supermarkets'. The demographic data for each user consist of age, gender and residential zip code. For one-tenth of the users, Call Detailed Records (CDR) over a period of six months are also available. These records include time, duration, location of the calls and identification of the receiver. Analyzing the records at a district level resulted in an observed correlation between the median credit card expenditure and the average monthly wage. The monthly expenditure of card users is high in relation to their monthly wages, which indicated that users with higher wages in each district use their credit cards more often. The majority of shoppers use more frequently 20 transactions codes, among hundreds of possible MCCs (the most frequent ones are 'Grocery stores, supermarkets' followed by 'Eating places and

restaurants', 'Bridge and road fees, tolls' and 'Computer network/Information services'). Although slight variations occur when the population is divided by wealth, age and gender, the number of top transactions is dominated by merchant codes related to food, mobility and communication. The Call Detailed Records are used to provide information about the mobility and social network of each cluster's users. Clustering the users by their similarity in purchase sequences resulted in six clusters. One cluster contains the uncategorized users while the other five depict distinct patterns on how the users spend their money, move and contact other users. For each cluster, one transaction type surpasses the others in the spending habits of its users, with 90% of the users repeating it in the time period. The 1st cluster is labeled 'Commuters' and the main characteristics of its users are that they live far from the city center, spend the most, travel long distances and are mostly male. Users of cluster 2 are labeled 'Homemakers', are mostly women and they represent the oldest group with least expenditure and mobility. Cluster 3, labeled as 'Youths' has younger users with high mobility while cluster 4, labeled as 'Tech', also has younger users but with computer networks and information services as a core transaction with higher than average expenditure and higher diversity in their social network. Finally, cluster 6 users are labeled as 'Diners', spend mostly in restaurants, are middle aged and have higher mobility diversity and higher expenditure.

The socioeconomic conditions of urban areas by analyzing their inhabitants spending habits is the focus of Sobolevsky et al. (2016) [6]. Bank card records are used to measure individual purchase activity and reveal patterns of economic activity. The dataset contains 4.5 million active customers of Banco Bilbao Vizcaya Argentaria (BBVA) in Spain 2011. The number of transactions is more than 178 million, with a cumulative spending exceeding 10.3 billion euro. The business classification includes 76 categories such as restaurants, gas stations, supermarkets or travels. The three characteristics related to the economic dimension of each transaction are: the activity of each customer defined as the total number of transactions performed during a year, the average value of a single transaction and the spending diversity, which is measured as the inverse of the sum of squared visit frequencies to each business category. The two characteristics based on customers' mobility are: distant mobility, measured as the percentage of transactions executed over 200 km from home and local mobility, measured as the average distance between each customer's home location and the retail points (calculated based on transactions made within 100 km from home). For these quantities, customers are required to use their bank card frequently enough, thus customers who performed at least 50 transactions in 2011 were considered and then restricted to those who performed transactions during at least nine different months. Customer spending behavior is different according to age and gender. The spending activity increases between 18 and 30 years old and then remains stable until 40 years old, followed by a steady decrease. Women make on average more transactions than men. The average amount per transaction increases linearly with the customer's age. This could be related to the ability of older people to spend more

or buy more expensive products or in general it could show a habit of fewer transactions with largest amounts of products bought each time. Women appear to spend more often but for smaller amounts. As men grow older, they tend to have more distant purchases, while women follow this behavior after they are 40 years old, showing small and decreased mobility until that age. Examining local mobility, it can be observed that for men it remains stable after 25-30 years old with a slight decrease with age, while for women a significant and stable decrease is exhibited. Women are mostly involved in basic shopping needs and tend to shop closer to home as they grow older. To examine if higher spending activity is exhibited in bigger cities, the average values of the bank card usage characteristics and their dependence on city size are analyzed. In bigger cities people seem to engage more easily in social activities while basic needs, like grocery stores, supermarkets and gas stations are independent of city size. Spanish cities were then segmented into three clusters using K-Means clustering. The size of the city of residence had a noticeable impact on all the characteristics of individual spending habits. The average value of bank transactions and household expenditures were highly correlated. Also, the clustering results captured meaningful economic patterns beyond the considered data. While the clustering was based solely on the spending behavior of city residents, the obtained categories corresponded with the attractiveness of those cities to foreign visitors.

Zhou et al. (2016) [7] studied how income affects the spending behavior of customers with prepaid cards and especially if there exists a 'payday effect', i.e. if payroll day has an impact on customers' spending. The data used involved 41,610 customers from a large national bank in the United States. The time period under study was 28 months, with each customer performing transactions for at least 11 days. Prepaid cards are different from charge cards since no credit is extended and customers can reload them and spend up to the balance of their account. The customer can also use the card to withdraw cash from ATMs, directly deposit paychecks, pay bills online and write paper checks that draw funds on the card. The attributes for each customer were card key, transaction amount in dollars, transaction date, transaction post time, Merchant Category Code (MCC), merchant name and transaction description. The dataset included customers that had a payroll deposit at the account within 45 days of the account opening and at least six payroll deposits in the time period under study. The transactions were divided to spending and deposit transactions. Spending transactions had the description of 'Purchase of Goods or services' and were selected based on the available MCCs. Six categories were obtained from MCCs which included 'restaurant', 'auto-related spending', 'grocery', 'drug store/alcohol' and 'wholesale/department store', while all other transactions formed a separate category called 'other spending'. The four deposit categories included 'Payroll', 'Manual deposit', 'Purchase return' and 'other income'. Four customer groups were observed based on their median payroll deposits frequency: 'weekly paid', 'bi-weekly paid', 'monthly paid' and customers that do not fall in one of these categories, referred to as 'others'. In all groups, people tend to spend more on the weekend than weekdays. Linear regression was used to

estimate the daily total spending in a particular category based on the following features: day of the week, month of the year, year, holiday and an individual variable associated with each customer. The ‘payday’ features take the form of a 15-dimensional indicator vector that starts from ‘payday -7 days’ and ends at ‘payday +7 days’. Control variables derived from the ‘account balance’ include balance in dollars, logarithm of balance +1 and an indicator vector that shows in which of the bins [0, 50), [50, 100), [100, 500), [500, 2000] and [2000, ∞) the balance falls in. The results show that there are strong payday effects in all six spending categories. The prepaid card users often wait until their salary is in their bank account to spend it, with increased spending on payday or within the following couple of days. Related to how people spend according to their balance, small balances decrease the chance of spending, and people seem to be more comfortable to spend when they have more money in their bank account.

Chapter 2 Theoretical framework

2.1 Data mining in customer segmentation

Data mining is the process of discovering patterns in large datasets. It is the analysis step that is used to extract useful information from databases and transform it into a comprehensible structure for further analysis and use. A data mining procedure involves extensive data management, followed by the application of a statistical or machine learning algorithm on the data and the development of an appropriate model. Data mining models can be divided in two categories based on their goal: supervised or predictive models and unsupervised models. Supervised or predictive models aim to predict the occurrence of an event or the values of a numeric continuous attribute and are divided in Classification and Estimation models respectively based on these goals. The goal of unsupervised models is to uncover patterns based on the input attributes. They are further separated in Clustering models, which discover groups in the data that have similar characteristics, and Association rule learning, which searches for relationships between variables. An example of Association rule learning is determining which products are frequently used together and use this information for marketing purposes. Data mining also involves anomaly detection, which is the identification of unusual data records, and summarization, which provides a more compact representation of the data set with visualizations and reports.

Data mining can be used in customer segmentation and targeted marketing campaigns [8]. There are three different types of customer segmentation:

- Value-based: customers are ranked and segmented based on their current and expected value for the company
- Behavioral: behavioral attributes, e.g. spending behavior, are used to segment customers
- Value-at-risk: customers are segmented into groups according to their value and estimated churn scores

The targeted marketing campaigns, on the other hand, involve:

- Churn modeling and estimation of the likelihood of churn
- Estimation of the probability that a customer will choose a new product, switch to more profitable product or increase his/her usage of a current product
- Estimation of the customer's lifetime value (LTV)

2.2 Clustering algorithms

Each modeling technique has different applications and different algorithms can be used to accomplish the requested result. For segmentation, clustering algorithms are used with the most popular of them being K-Means.

K-Means is an efficient and fast distance-based clustering algorithm. It tries to separate data points in k groups or clusters of equal variances. The number of clusters is defined by the user and the algorithm's steps are:

1. Pick k points, which are named centroids, one per cluster
2. Assign remaining points to closest centroid, based on the minimum Euclidean distance between the point and the centroids
3. In each cluster, update the location of its centroid as the mean of the points in the cluster
4. Reassign points if necessary
5. Repeat steps 3-4 until the clusters stabilize

K-Means seeks to minimize the sum of squared distances from the centroids. [9] This criterion is called the Within Cluster Sum of Squares (WCSS) or inertia. Inertia measures the internal coherence of the clusters. It is a metric that assumes that clusters are convex and isotropic and although it is not a normalized metric, lower values are better and zero is optimal [10].

Other clustering algorithms that can be used for segmentation are agglomerative hierarchical clustering algorithms, density-based algorithms and Kohonen Networks / Self-Organizing Maps (SOM's).

Agglomerative hierarchical clustering starts with the points as individual clusters and at each step merges the closest pair of clusters. To find which clusters are close, a measure of cluster proximity must be defined. Cluster proximity can be defined as the proximity of the closest two points between different clusters, the proximity between the farthest two points or by taking the average of the pairwise proximities of all pairs of points from different clusters. An alternative technique used to merge two clusters is assuming that each one is represented by its centroid, like in K-Means, and attempt to minimize the sum of the squared distances of points from their cluster centroids. Hierarchical clustering is useful if a hierarchy is to be obtained from the data, since the results can be visualized in the form of a dendrogram. However, it is expensive in terms of computational and storage requirements and is not suitable for large datasets. [11]

Density-based clustering locates regions of high density which are separated by regions of low density. The most popular density-based clustering algorithm is DBSCAN (Density

Based Spatial Clustering of Applications with Noise). Density is defined as the number of points within a specified distance. DBSCAN classifies points into three groups. Core points are at the interior of a cluster and have more than a specified number of points within the specified distance. Border points are within the neighborhood of a core point but have less than the specified number of points within the specified distance. Points that are neither core nor border are defined as noise points. The steps of the algorithm are [11]:

1. Label all points as core, border or noise points
2. Discard noise points
3. Put an edge between all core points that are within the specified distance of each other
4. Form a separate cluster for each group of connected core points
5. Assign each border point to one of the clusters of its associated core points

DBSCAN has high computational cost for high-dimensional data. However, it has low space requirements since it only needs to keep a small amount of data for each point, i.e. the cluster label and the identification of whether it is a core, border or noise point.

Kohonen networks are neural networks that are used for clustering and typically contain two layers. The input layer includes all clustering fields while the output layer is a two-dimensional grid map which contains the output neurons which will form the clusters. The input neurons of the input layer are connected to each of the output neurons with weights. These weights are initially set at random and change as the model is trained and are analogous to the cluster centers of the K-Means algorithm. Each record is assigned to the output neuron with the most similar pattern characteristics based on the Euclidean distance between the record's input values and the centers of the output neurons. The weights of all neighboring neurons are then adjusted so that they better match the pattern of the assigned record. Because of this neighborhood adaptation, the topology of the output map has a practical meaning, with similar clusters appearing close together as nearby neurons. If another record with similar characteristics is presented to that output neuron, it will have a greater chance of being assigned to it. Many iterations and weight adjustments are involved in the training of Kohonen networks making them slower than the K-Means algorithm.

2.3 Preprocessing steps for clustering algorithms

A recommended preprocessing step before the implementation of a clustering algorithm is the application of a data reduction technique [8]. This can simplify and enhance the segmentation process by removing redundant information. The most popular technique for feature reduction is Principal Component Analysis (PCA), which is a statistical technique used to reduce the original input fields to a limited number of components while retaining most of the original information. The components derived from PCA are uncorrelated with

each other, but they are associated or correlated with a specific set of the original fields. The components are extracted in decreasing order of importance so that the loss of information is minimal. The first component carries as much of the total variability of the input fields as possible, i.e. it explains most of their information, and each succeeding component is constructed to account for the remaining information while being uncorrelated to its predecessors. The resulting components are linear combinations of the original variables and are uncorrelated and orthogonal to each other. The key issues that need to be addressed in the application of PCA are the optimal number of components, the relations between the original fields and the components, and the meaning of each component.

The optimal number of PCA components should be selected based on one of the following criteria:

1. The eigenvalue criterion: The eigenvalue is a measure of the variance that each component accounts for. Components with eigenvalues below 1 are not retained.
2. The percentage of variance criterion: The number of components is determined by a desired level of the total explained percentage of variance. The threshold value depends on the specific situation, but it should not fall below 60-65%. Since this percentage relates to the information being explained by the components, values of 80-90% are preferred if it is necessary to keep as much of the original information as possible.
3. The interpretability and business meaning of the components: Components should be interpretable, useful and should have clear business meaning.
4. The scree test criterion: The eigenvalues decrease as the number of components increases. A large drop followed by a plateau in the eigenvalues indicates a transition from large to small values with the unique variance of each component dominating the common variance. This point indicates the optimal number of components.

The component matrix presents the linear correlations between the original attributes in the rows and derived components in the columns. These correlations are called loadings and are used for the interpretation and labeling of the derived components. Loadings above 0.4 in absolute value are considered significant and denote the original attributes that are represented in each component. To make the interpretation of the loadings easily understandable, the loadings matrix should be sorted according to the original attributes in descending order, so that attributes associated with the same component appear closer together. Examining the matrix, labels for the components can be derived with names that appropriately summarize their meaning. Positive loadings correspond to positive correlations between the attribute and the component while negative loadings correspond to negative correlations. The labels of the components will help in the analysis of the

clusters obtained by a clustering algorithm to see which component has higher value in each cluster and then accordingly, provide a label for each cluster.

Another necessary preprocessing step in clustering is the standardization or normalization of the data. Since clustering models are based in the differences between records, different measurement scales can lead to biased solutions since some fields may be measured in larger values and they could dominate the solution. Thus, all attributes should be scaled to ensure that large values do not affect the solution. A common scaling method is standardization of the features by removing the mean and scaling to unit variance. The standardized field is created as: $\frac{x-\mu}{\sigma}$, where x is the record value, μ is the mean value of the attribute and σ is the standard deviation of the attribute. Other scaling methods frequently used involve scaling features to lie between a given range. Min-Max scaling approach rescales features as: $\frac{x-min}{max-min}$, where x is the record value and min, max correspond to the minimum and maximum values of the attribute respectively. The transformed features are in the range of $[0,1]$. Max-Abs scaler works like Min-Max scaler, but the data are scaled to lie in the range $[-1,1]$ based on the absolute maximum. It is best suited for already centered at zero or sparse data. Sparsity is not destroyed since it does not center the data. [10]

2.4 Clustering evaluation

The evaluation of the performance of a clustering algorithm should be examined in terms of the number and relative size of the clusters, as well as their cohesion and separation. Cohesion measures how closely related the objects in the cluster are, while separation measures how distinct or well separated a cluster is from other clusters. Silhouette coefficient is a measure which combines both the internal cohesion and the external separation of a clustering solution. [9] The silhouette coefficient for each record i is defined as:

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}},$$

where a_i is the average Euclidean distance of record i to all other records of the same cluster and b_i is the minimum of the average distances of i from members of another cluster.

The silhouette coefficient takes values between -1 and 1. As a_i measures how dissimilar record i is to its own cluster, a small value indicates a good solution, while large b_i implies that it is strikingly matched to its neighboring cluster. Thus, values of the silhouette coefficient close to 1 indicate that the data points are appropriately clustered, values near

zero show that the point is borderline and values close to -1 indicate wrong assignment of the data point.

By averaging over all points in a cluster, its average silhouette coefficient is calculated. The overall silhouette coefficient which is a measure of the goodness of the clustering solution is calculated by taking the average over all records. Values of the silhouette coefficient above 0.5 indicate reasonable partitioning, while values less than 0.2 denote problems in the clustering solution.

There are two more metrics that can be used to evaluate the clustering results: the Calinski-Harabasz index and the Davies-Bouldin index. [10] Better defined clusters are related to a higher Calinski-Harabasz score which is defined as:

$$s(k) = \frac{SS_B}{SS_W} \times \frac{N-k}{k-1},$$

where k is the number of clusters, N is the total number of observations, SS_W is the overall within-cluster variance which is equivalent to the total within sum of squares and SS_B is the overall between-cluster variance. The between-cluster variance measures the variance of all cluster centroids from the dataset's grand centroid. Large values indicate that the centroids of each cluster will be spread out and are not close to each other and thus the ratio of $\frac{SS_B}{SS_W}$ should be largest at the optimal number of clusters.

Lower Davies-Bouldin indices relate to models with better separation between the clusters. The index is defined as the average similarity between each cluster C_i, $i = 1,2, \dots k$ and its most similar one C_j. Defining:

- s_i as the average distance between each point of cluster i and the centroid of the cluster
- d_{ij} as the distance between the centroids of clusters i and j

the similarity is a nonnegative and symmetric measure defined as: $R_{ij} = \frac{s_i + s_j}{d_{ij}}$.

The Davies-Bouldin index is defined as: $DB = \frac{1}{k}\sum_{i=1}^{k} max\, R_{ij}$, for $i \neq j$ with zero being the lowest possible score. Values of the index closer to zero indicate a better partition.

2.5 Cluster profiling

After the evaluation of the clustering algorithm and the selection of the optimal number of clusters, each cluster needs to be analyzed and its characteristics should be recognized.

The cluster profiling involves two steps [8]:

1. **Examination of the cluster centers:** In this step, the input data patterns that distinguish each cluster are identified. The cluster labels will be interpreted according to their differentiating characteristics and the uniqueness of each cluster is determined according to the input fields. Each cluster's center, or centroid, is defined by taking the averages of each input field over all records of the cluster and it is the most representative member of the cluster. Each cluster should be checked individually since high or low mean values indicate the behavior of this cluster's data records.
2. **Comparing clusters with respect to other attributes:** Clusters should also be examined with respect to attributes that are not involved in the cluster formation, such as key performance indicators (KPIs) and demographic information. For continuous fields, the means for each cluster should be evaluated, while categorical attributes should be compared using frequencies and percentages.

Chapter 3 Customer segmentation in banking data

3.1 Programming languages and software

Due to the large amount of records, all preprocessing steps were implemented in Apache Spark using Python API [11]. The implementation of the K-Means algorithm on the final data was made with Python's [12] scikit-learn library [10] while all figures provided were made with Tableau Software.

3.2 Data description

The datasets under study contain transaction records for customers with debit and credit cards in the National Bank of Greece (NBG). There are two datasets, one for each card type, referred to as *'credit dataset'* and *'debit dataset'*, for transactions of one month. Each dataset originally contained 465,030 records.
Some rows contained mistakes in their records, such as having the names of the columns instead of actual values in each column. There were also records in the credit dataset which had negative numbers in the amount of the transaction. These records were removed, resulting in 460,537 records for the credit dataset and 465,000 records for the debit dataset. The credit dataset had 9 attributes, the timestamp of the transaction, the amount in Euro, the amount in the currency of the terminal, the country of the terminal, a country index (taking values of 0 if country is Greece and 1 otherwise), the currency code, the original currency code, the Merchant Category Code (MCC) and the card id. The original currency code attribute only had the value '978' which corresponds to Euro and matches the original currency of each card. There are 7,774 transactions in which currency code is different than 978, indicating transactions paid in a foreign currency. Since for each transaction, its amount in Euro was available, columns regarding the currency and the countries were discarded. The MCC is a four-digit number used to classify suppliers into market segments. There are approximately 881 codes used to denote various types of businesses, e.g. 5411 Grocery Stores, Supermarkets, 8011 Doctors, etc. A file containing all available MCCs and their descriptions was downloaded from:
https://www.ogs.ny.gov/purchase/snt/awardnotes/7900822712rf018MerchantCategoryCodeList.xls. MCCs were further grouped into 17 general categories based on their descriptions. The categories, as well as a description with some examples for each one, are listed below:

1. Automotive: includes codes related to cars and other moving vehicles, e.g. Car and Truck Dealers, Fuel Dealers, Parking Lots and Garages, Car Washes, etc.

2. Business to Business: includes codes related to transactions between businesses. These codes are characterized by the label of (Business to Business MCC) in their description.
3. Clothing: includes codes for clothing, accessory and shoe stores for women, men and children.
4. Education: includes transactions for schools, colleges and universities.
5. Entertainment: includes codes for music related stores, restaurants, drinking places, galleries, sports, recreational activities etc.
6. Food: includes codes for supermarkets, bakeries and food stores.
7. Gambling: includes lotteries, online gambling, horse/dog racing and betting.
8. Government: includes codes for fines, tax payments, bond payments and other government services.
9. Home/Garden: includes codes for everything related to home and garden such as furniture, appliances, garden supply stores, repair stores and contractors.
10. Medical: includes codes for ambulance services, doctors, pharmacies, hospitals, etc.
11. Personal Care: includes codes for beauty shops, massage parlors, spas, laundry and dry-cleaning services and shoe repair shops.
12. Services: includes courier services, insurance services, payments, advertising services, legal services and all other codes labeled as services that are not in another category.
13. Shopping: includes all stores not already contained in other categories, such as department stores, duty free stores, cosmetic stores, book stores, cigar stores, florists, pet shops and retail stores in general.
14. Telcos: includes all codes related to telecommunication services such as telephone charges, network/information services (e.g. e-mail, web-hosting services) and cable, satellite and radio services.
15. Transportation: includes costs for commuter passenger transportation, trains, ferries, taxicabs, limousines, buses and tolls.
16. Travel and Leisure: includes codes for airlines, car rentals and hotels with each one having its own code (resulting in a total of 572 different codes in this category).
17. Utilities: includes the separate code for utilities involving electric, gas, water and sanitary.

The debit dataset had 5 attributes, the timestamp of the transaction, the Merchant Category Code (MCC), the amount in Euro, the card id and the account id.
The two datasets were merged with the file containing the descriptions for the MCCs. It was observed that there were two codes used for the description ‘Opticians, Optical Goods and Eyeglasses’, 8043 and 8044, and thus only 8043 was kept and transactions having MCC 8044 were changed to have MCC 8043. There were also some codes in the dataset

that did not match with any of the available codes and further research revealed that these codes were reserved for private use. Since it was not possible to describe them, the transactions involving these codes were removed resulting in 460,534 transactions in the credit and 464,996 in the debit dataset.

The last dataset that was provided, contained information about the customers, i.e. demographic data. There were 10 attributes in this dataset: card id, account id, type of the card ('CCRD' for credit cards and 'DCRD' for debit cards), gender, date of birth, marital status, educational level, occupation, prefecture and zip code. Using the date of birth, the age for each client was calculated and due to some mistakes in the data, ages below 18 years and above 100 years were characterized as missing, so that the other attributes for data with wrong ages are not affected. A new variable was created, named *age_category* that assigned age to one of 6 groups labeled as: '18-24', '25-34', '35-44', '45-54', '55-64' and '65+'. From the prefectures, another variable named *Region* was created, which assigned each prefecture to its region. There are 13 regions in Greece: Attica, Central Greece, Central Macedonia, Crete, Eastern Macedonia and Thrace, Epirus, Ionian Islands, North Aegean, Peloponnese, South Aegean, Thessaly, Western Greece and Western Macedonia. Then, using the available zip codes, maps from Google APIs was used to find the coordinates, i.e. latitude and longitude, for each zip code. The acquired table contained the zip code, its longitude and latitude, as well as the name of that area. To cross-check that the results were correct, the zip codes were compared to a dataset containing all Greek zip codes, acquired from:
https://www.taxheaven.gr/acforum/files/file/1480-%CF%84%CE%B1%CF%87%CF%85%CE%B4%CF%81%CE%BF%CE%BC%CE%B9%CE%BA%CE%BF%CE%B9-%CE%BA%CF%89%CE%B4%CE%B9%CE%BA%CE%B5%CF%83-%CE%B5%CE%BB%CE%BB%CE%B1%CE%B4%CE%BF%CF%83/. This dataset also added information about the City and Street or village for each zip code. Some locations with negative longitude were removed, since the minimum value for longitude in Greece is about 19.5. Although this dataset provided demographic information about the customers, it cannot be accepted as a general truth since cards may be used by other people and not just their owner. The final dataset with the coordinates was merged with the customer information dataset and then according to the type of the card, the customer information dataset was merged with the credit and debit datasets respectively.

3.3 Exploratory data analysis

This section focuses on exploring the attributes of each dataset and extracting information from them.

The number of customers in the credit and debit datasets are 163,036 and 318,850 respectively. Some customers may appear in both datasets. These numbers were acquired after the merge with the customer information data. After the merge, some cards appeared twice due to the fact that information about two persons were connected to the same card. Probably these people are husband and wife or parent and child. Since it was not possible to determine which person was using the card during the transaction, to avoid any mistakes in the profiling, all demographic information about these cards was erased and one record for each card was kept.

All general spending categories but not all MCCs were present in both datasets, with the credit dataset having 356 distinct MCCs and the debit dataset having 322 distinct MCCs. Tables 1 and 2 show the 10 most popular codes, according to the total number of transactions involving each code in the time period under study, with their descriptions and general category for the credit and debit datasets:

MCC	**Description**	**Category**
5411	Grocery Stores and Supermarkets	Food
5541	Service Stations	Automotive
5968	Continuity/Subscription Merchants	Services
5812	Eating Places and Restaurants	Entertainment
4814	Telecommunication Services, Including Local and Long Distance Calls, Credit Card Calls, Call Through Use of Magnetic-Strip-Reading Telephones, and Fax Services	Telcos
5462	Bakeries	Food
4899	Cable, Satellite, and Other Pay Television and Radio Services	Telcos
5499	Miscellaneous Food Stores-Convenience Stores and Specialty Markets	Food
5912	Drug Stores and Pharmacies	Medical
5813	Drinking Places (Alcoholic Beverages) - Bars, Taverns, Nightclubs, Cocktail Lounges, and Discotheques	Entertainment

Table 1 10 most popular codes for the credit dataset with their description and general category

MCC	Description	Category
5411	Grocery Stores and Supermarkets	Food
5541	Service Stations	Automotive
5812	Eating Places and Restaurants	Entertainment
7995	Betting, including Lottery Tickets, Casino Gaming Chips, Off- Track Betting, and Wagers at Race Track	Gambling
5691	Men's and Women's Clothing Stores	Clothing
5813	Drinking Places (Alcoholic Beverages) - Bars, Taverns, Nightclubs, Cocktail Lounges, and Discotheques	Entertainment
5814	Quick Payment Service-Fast Food Restaurants	Entertainment
4814	Telecommunication Services, Including Local and Long Distance Calls, Credit Card Calls, Call Through Use of Magnetic-Strip-Reading Telephones, and Fax Services	Telcos
5499	Miscellaneous Food Stores-Convenience Stores and Specialty Markets	Food
5912	Drug Stores and Pharmacies	Medical

Table 2 10 most popular codes for the debit dataset with their description and general category

Transactions at Grocery Stores and Supermarkets, Service Stations, Eating Places and Restaurants, Telecommunication Services, and Drug Stores and Pharmacies appear in both datasets in the top 10 of the MCCs. It is highly notable that in the debit dataset, Betting is quite high on the list, while in the credit dataset it does not appear in the first 10 most frequently purchased codes.

The distribution of the transaction amount in Euro differs in the two datasets. The maximum amount in the credit dataset is 21,381.42 € while in the debit dataset it is 7,725.0 €. The minimum amount is the same in the two datasets but the mean and the standard deviation differ as it can be seen in Table 3.

	Credit dataset	Debit dataset
Mean	33.79	28.87
Standard deviation	111.55	60.05
Median	11.99	15.0
Minimum	0.01	0.01
Maximum	21,381.42	7,725.0

Table 3 Statistics for credit and debit datasets

The standard deviation in both datasets is quite large, the mean is greater than the median and the difference between the minimum and maximum value is large. All these indicate that the distributions of amount are right skewed and there are only a few transactions with large amounts in Euro in both datasets.

Since the objective is to extract information about the spending behavior of customers, it is interesting to see which MCCs correspond to higher total spending with respect to the sum of all transactions' amount. The results, shown in Table 4 for the credit dataset and Table 5 for the debit dataset, display that some of the MCCs with large number of transactions also correspond to higher total spending. Also, in the credit dataset, codes corresponding to spending in hotels, utilities and travel agencies appear in the first 10 codes with high total spending. On the other hand, in the debit dataset, household appliance stores, payments, travel agencies and cosmetic stores have high total spending although they did not appear in the first 10 most frequently purchased codes.

MCC	**Description**	**Category**
5411	Grocery Stores and Supermarkets	Food
5541	Service Stations	Automotive
5968	Continuity/Subscription Merchants	Services
7011	Lodging - Hotels, Motels, and Resorts	Travel and Leisure
6300	Insurance Sales and Underwriting	Services
4900	Utilities-Electric, Gas, Water, and Sanitary	Utilities
5812	Eating Places and Restaurants	Entertainment
4722	Travel Agencies	Travel and Leisure
4814	Telecommunication Services, Including Local and Long Distance Calls, Credit Card Calls, Call Through Use of Magnetic-Strip-Reading Telephones, and Fax Services	Telcos
5999	Miscellaneous & Specialty Retail Stores	Shopping

Table 4 10 codes with higher total spending for the credit dataset with their description and general category

MCC	**Description**	**Category**
5411	Grocery Stores and Supermarkets	Food
5541	Service Stations	Automotive
5812	Eating Places and Restaurants	Entertainment

7995	Betting, including Lottery Tickets, Casino Gaming Chips, Off-Track Betting, and Wagers at Race Track	Gambling
5691	Men's and Women's Clothing Stores	Clothing
5722	Household Appliance Stores	Home/Garden
6011	Financial Institutions--Automated Cash Disbursements	Services
4722	Travel Agencies	Travel and Leisure
4814	Telecommunication Services, Including Local and Long Distance Calls, Credit Card Calls, Call Through Use of Magnetic-Strip-Reading Telephones, and Fax Services	Telcos
5977	Cosmetic Stores	Shopping

Table 5 10 codes with higher total spending for the debit dataset with their description and general category

To conclude the data exploration part, customer information was evaluated for both datasets. The credit dataset has 163,036 customers and the debit dataset has 318,850 customers. Figure 1 shows the percentages of customers belonging in each category of the attributes 'Gender', 'Marital Status' and 'Age Category' for each dataset.

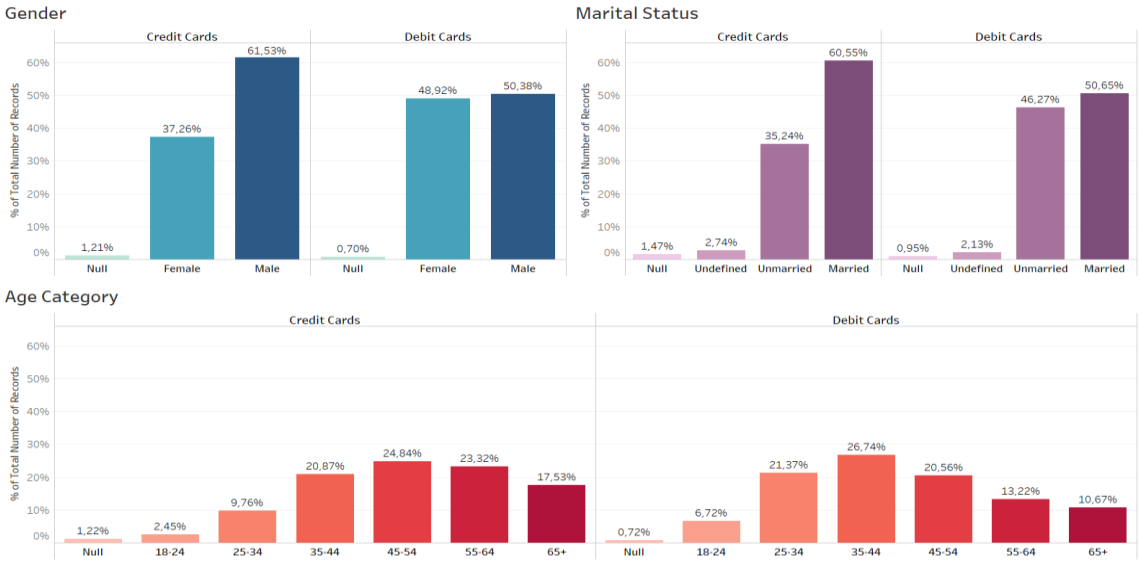

Figure 1 Percentages of customers belonging in each category of the attributes 'Gender', 'Marital Status' and 'Age Category' for credit and debit datasets

The category named 'Null' that appears in all attributes corresponds to the customers for whom no information is available for that attribute. In the credit dataset, there are more male than female customers while in the debit dataset the difference between them is less than 2%. The difference between Married and Unmarried customers is also bigger in the

credit dataset. Considering customers based on their age, the largest percentages for the credit dataset are in the categories '35-44', '45-54' and '55-64', while in the debit dataset there are more people in the category '25-34' than in '55-64'. It seems that debit cards are more popular among younger people.

Figure 2 shows the percentages of customers belonging in each category of the attributes 'Occupation' and 'Educational Level' for each dataset. There are a lot of missing data for the 'Educational Level' attribute in both datasets. According to their occupation, most cardholders are Employed. Unemployed people, pensioners and freelancers seem to prefer credit cards, while debit cards are more popular than credit cards for university students which was also observed with a higher percentage in age category '18-24' in debit cards (Figure 1). In both datasets, large percentages in educational level appear in High School Education, followed by Higher Education Institutes (HEIs) and Technical Educational Institutes (TEIs) and Basic Education.

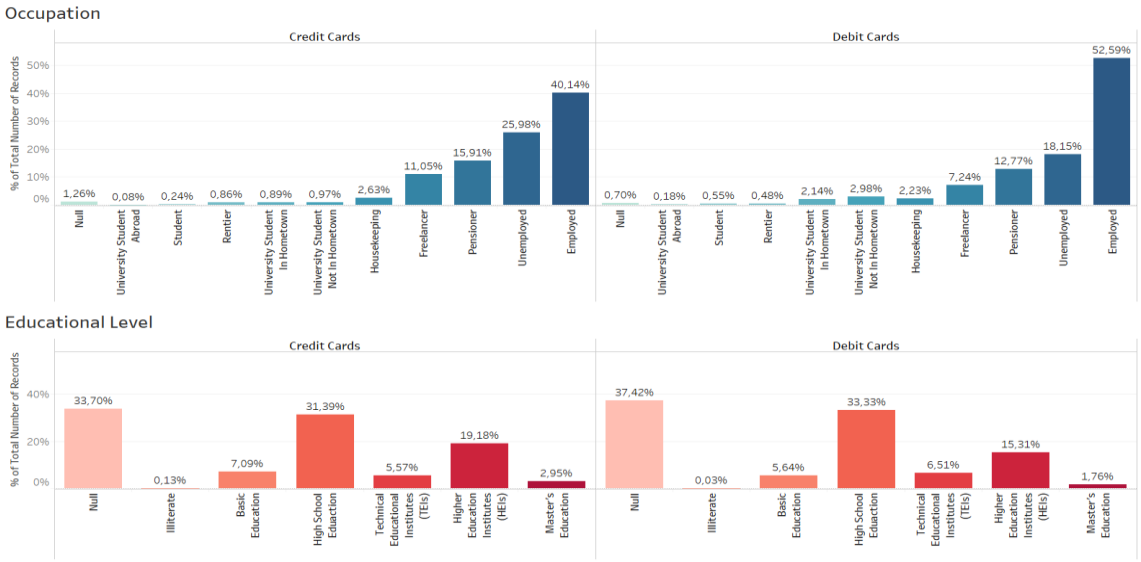

Figure 2 Percentages of customers belonging in each category of the attributes 'Occupation' and 'Educational Level' for credit and debit datasets

Figure 3 shows the percentage of customers living in each of the 13 regions in Greece. There is less than 2% data missing for this attribute in both datasets. The highest percentages are in Attica and Central Macedonia which is expected since Greece's largest cities, Athens and Thessaloniki, are in those two regions respectively.

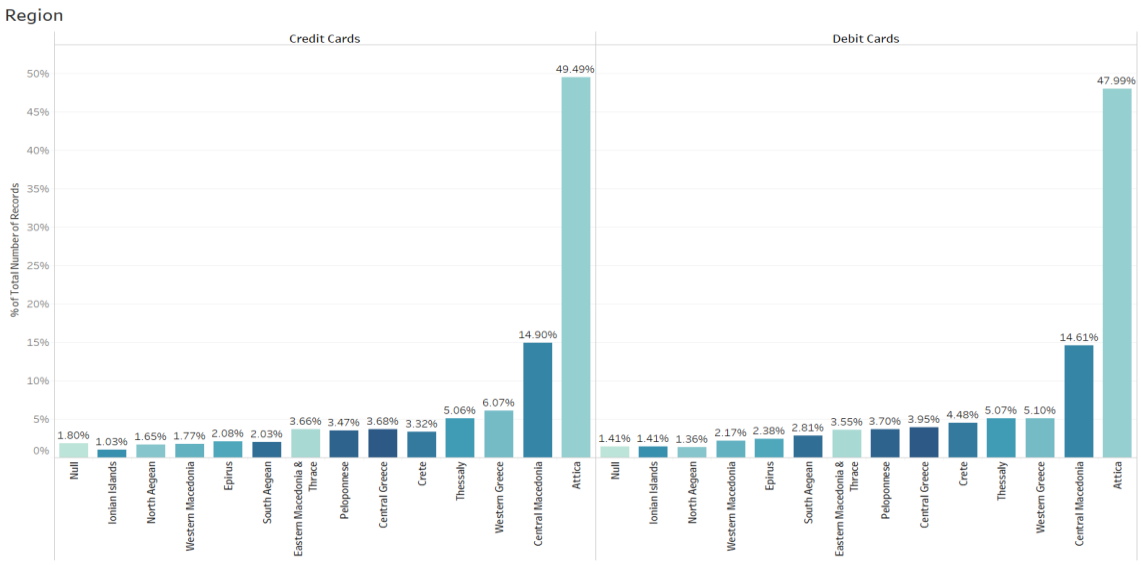

Figure 3 Percentages of customers living in each of the 13 regions in Greece for credit and debit datasets

3.4 Data aggregations

This section focuses on the aggregations that were used in the data to prepare them for the clustering algorithm. Since the number of records is large, K-Means algorithm will be used for clustering as it is more efficient than the other algorithms.

For the behavioral segmentation of customers, obvious segments of data, e.g. inactive customers, should be removed, data should be smoothed using monthly averages, percentages, ratios or other summarizing KPIs and demographic attributes should be avoided. It is best to use demographic attributes at the cluster profiling stage. All data are from transactions performed during the time period thus there were no inactive customers to exclude. Since the time period involved only a month of data, it was decided that percentages of total spending per category would be used as the input fields for the algorithm.

To create the variables needed for the clustering algorithm, the data in each dataset were aggregated according to the card id. The total purchases, the total spending, the number of purchases per category and the total spending per category were calculated for each customer. Then, the percentage of total spending per category was calculated by dividing the total spending per category and the total spending for each customer. This resulted in 17 attributes to be used for clustering, each one corresponding to one of the general categories described in 3.1 with values the percentage of total spending each customer has spent in that category.

Figures 4 and 5 describe the behavior that people with credit and debit cards exhibit for each general category of spending.

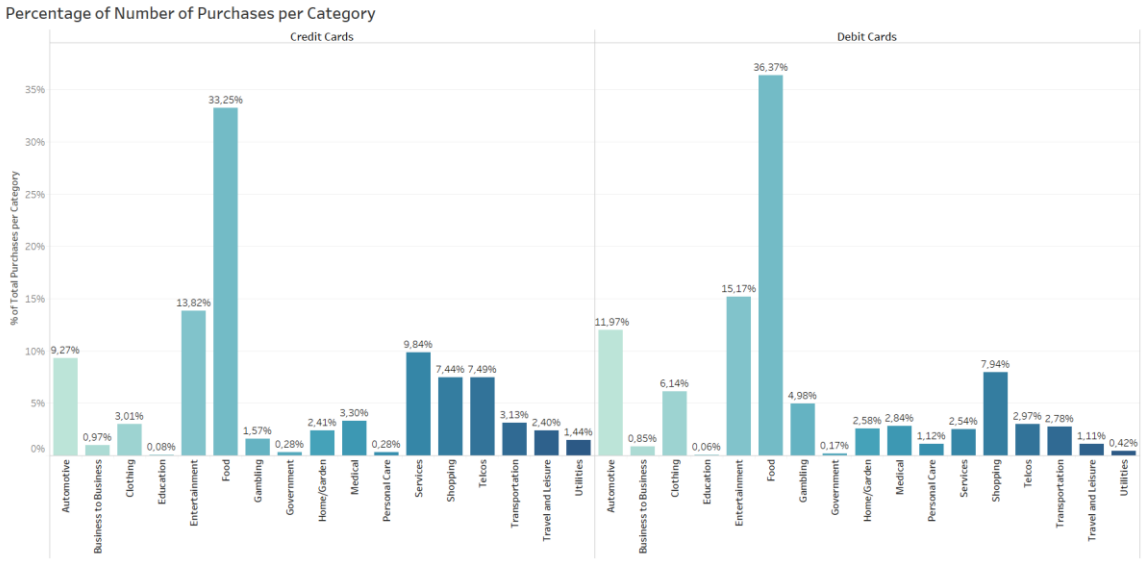

Figure 4 Percentage of number of purchases per category for credit and debit card users

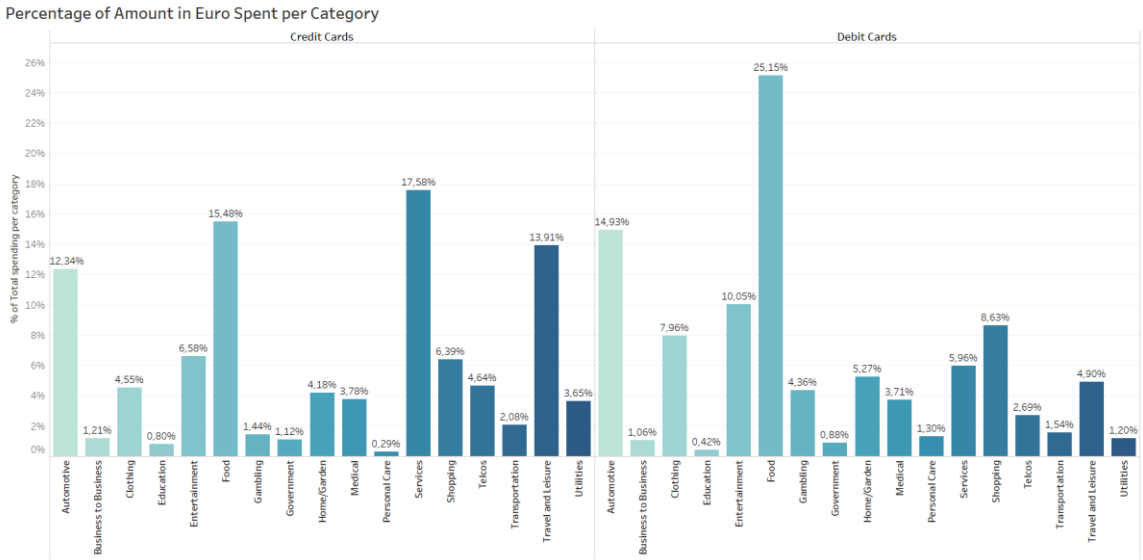

Figure 5 Percentage of amount in Euro spent per category for credit and debit card users

Since there is quite a difference in the number of credit and debit card transactions, percentages were used to be able to compare them. According to the number of purchases per each category, customers spend mostly on Food. Purchases in the Entertainment category follow in both datasets. There are more purchases made in the Automotive category in the debit dataset than in the credit dataset, but the categories of Services and Telcos show larger percentages of purchases in the credit dataset. Debit card customers though, spend more frequently on Gambling and Clothing than credit card customers.

Considering the percentage of money spent in each category for the two datasets, Food is still the top category in the debit dataset, but in the credit dataset people spent a bit more on Services than on Food. Travel and Leisure is a category with high percentage of spending in the credit dataset. More money is spent at Utilities and Telcos in the credit dataset, while debit card users spent more money in the categories of Clothing, Gambling and Shopping.

3.5 Feature selection

This section focuses on the selection of the features that will be used in the clustering algorithm. Since demographical attributes should be avoided in behavioral segmentation and should only be used in the cluster profiling, for each dataset the card id and the attributes corresponding to the percentage of total spending in each of the 17 general categories were selected. Figures 6 and 7 show the correlations between the 17 attributes for the two datasets.

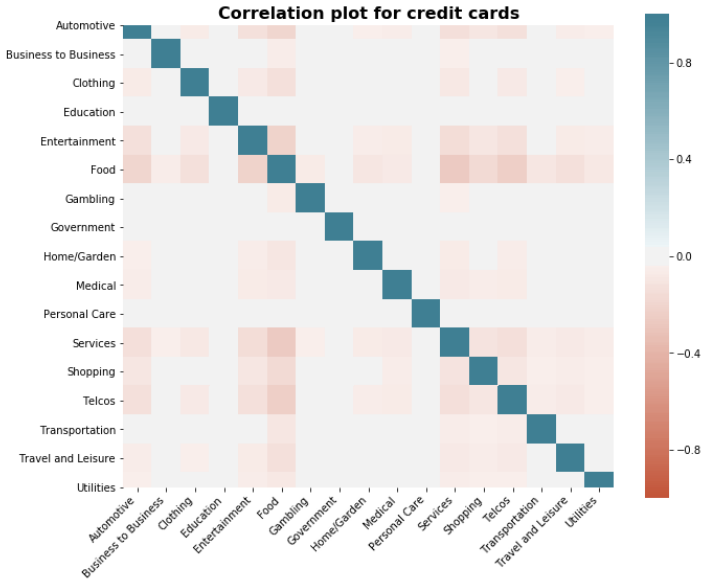

Figure 6 Correlation plot of the 17 attributes for the credit dataset

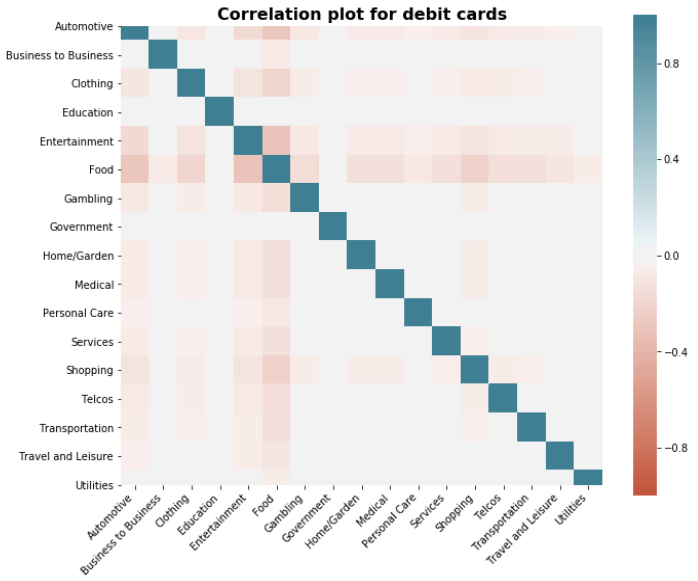

Figure 7 Correlation plot of the 17 attributes for the debit dataset

The attributes for both datasets do not have strong correlations with each other. All correlations are below absolute 0.4. Spending in the Food category seems to have low correlations with some of the other categories for both datasets, but they are not large enough to show a distinct pattern.

Then the 17 attributes of the percentage of total spending by category were used as inputs to a PCA model for data reduction and identification of the underlying dimensions. According to the cumulative explained variance criterion and setting the threshold to 0.85, eight components were extracted for each dataset. For the credit dataset, the eight components had a cumulative explained variance of 89.8%, while for the debit dataset they explained 88.35% of the original information.

For each dataset, the extracted components were interpreted in terms of their correlations (loadings) with the original attributes by creating the components matrix. Loadings above 0.4 in absolute value denote the original attributes that are represented in each component while values below that threshold are omitted. The matrices are sorted in descending order according to the original attributes, with values close to 1 in absolute value signifying strong correlations. Tables 6 and 7 show the components matrix for the credit and debit datasets respectively. Attributes without strong correlations with the extracted components are omitted for a clearer representation.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Food	0.9166							
Services		0.8522						
Telcos			0.7346					
Entertainment			-0.6761	-0.4432				
Automotive				0.8002				
Shopping					0.8259	-0.4359		
Travel and Leisure						0.5457	-0.6913	
Clothing						0.4628	0.7189	
Medical								0.7222

Table 6 Components matrix for the credit dataset

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Food	0.9173							
Entertainment		-0.7324	-0.4930					
Automotive		0.6783	-0.5805					
Shopping			0.4980	-0.6995				
Clothing				0.6928	-0.5500			
Gambling					0.5348	0.7652		
Telcos						-0.4482	0.7785	
Medical							-0.4868	0.7202
Services								-0.5991

Table 7 Components matrix for the debit dataset

The interpretations and labels of each component according to the component matrix for each dataset are summarized in tables 8 and 9.

Component	Label and Description
1	**Food**: Component 1 represents purchases in the Food category
2	**Services**: Component 2 measures the spending in the Services category
3	**Telcos, negative Entertainment**: Component 3 is positively correlated with spending for telecommunications services and negatively correlated with spending in the Entertainment category
4	**Automotive, negative Entertainment**: Large positive scores indicate spending in Automotive category while large negative scores suggest increased spending in the Entertainment category
5	**Shopping**: Component 5 represents spending in retail stores and other stores in the Shopping category
6	**Clothing, Travel, negative Shopping**: Component 6 is positively correlated with spending in clothing stores and travel expenses, but it is also negatively correlated with spending in the Shopping category
7	**Clothing, negative Travel**: Large positive scores indicate spending in clothing stores while large negative scores suggest increased travel expenses
8	**Medical**: Purchases for health services are highly and positively correlated with component 8

Table 8 Component interpretation for the credit dataset

Component	Label and Description
1	**Food**: Component 1 represents purchases in the Food category
2	**Automotive, negative Entertainment**: Large positive scores indicate spending in Automotive category while large negative scores suggest increased spending in Entertainment category
3	**Shopping, negative Automotive, Entertainment**: Component 3 is positively correlated with spending in the Shopping category and negatively correlated with spending in the Automotive and Entertainment categories
4	**Clothing, negative Shopping**: Component 4 is positively correlated with spending in clothing stores, but it is also negatively correlated with spending in the Shopping category
5	**Gambling, negative Clothing**: Component 5 represents a positive correlation with spending in the Gambling category and a negative correlation with spending in clothing stores
6	**Gambling, negative Telcos**: Component 6 is positively correlated with gambling and negatively correlated with spending in telecommunication services
7	**Telcos, negative Medical**: Large positive scores indicate spending in telecommunication services while large negative scores suggest increased medical expenses
8	**Medical, negative Services**: Purchases for health services are highly and positively correlated with component 8, while spending in the Services category is negatively correlated with this component

Table 9 Component interpretation for the debit dataset

3.6 Clustering algorithm implementation and evaluation

For each dataset, its extracted components were used as input in a K-Means algorithm to reveal groups of customers with similar spending behavior. Since K-Means needs the number of clusters *k* as input, values of *k* from 2 to 10 were evaluated according to the elbow curve, the silhouette coefficient, the Calinski-Harabasz index and the Davies-Bouldin index. The optimal number of clusters according to the elbow curve is the point in which the line breaks like an elbow. The silhouette coefficient and Calinski-Harabasz index need to be high at the optimal number, with values of silhouette coefficient close to 1 indicating better cohesion and separation among the clusters. Davies-Bouldin index needs to be closer to zero to have the optimal *k*. Figures 8 and 9 show the graphs of these four measures for the different values of *k* for each dataset. For both datasets the optimal number of clusters according to the four measures is 9.

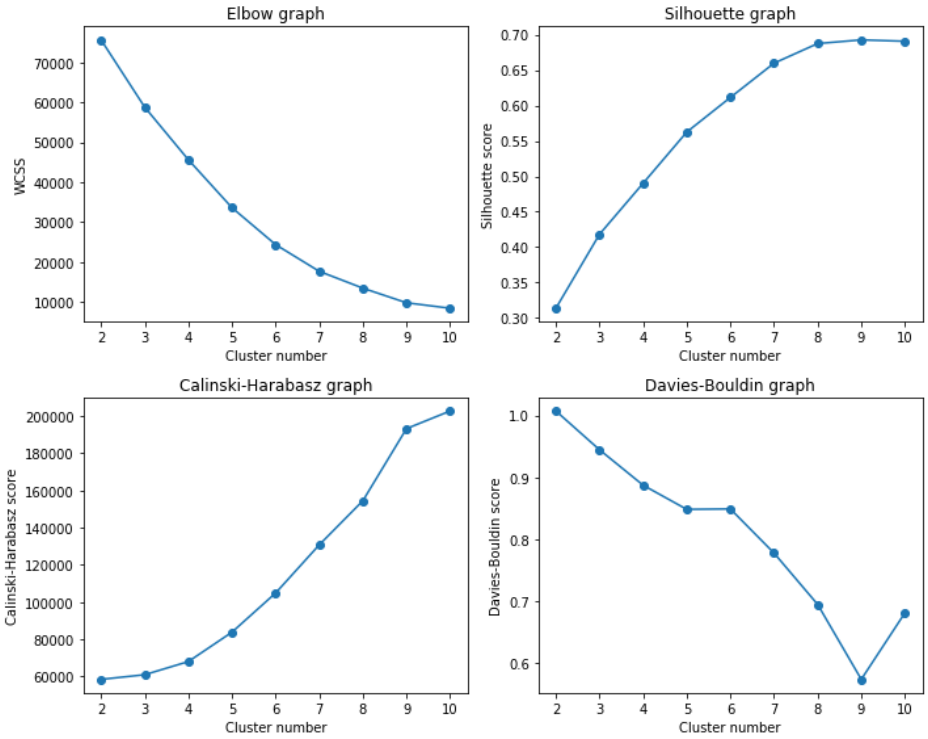

Figure 8 Clustering evaluation graphs for the credit dataset for values of k from 2 to 10

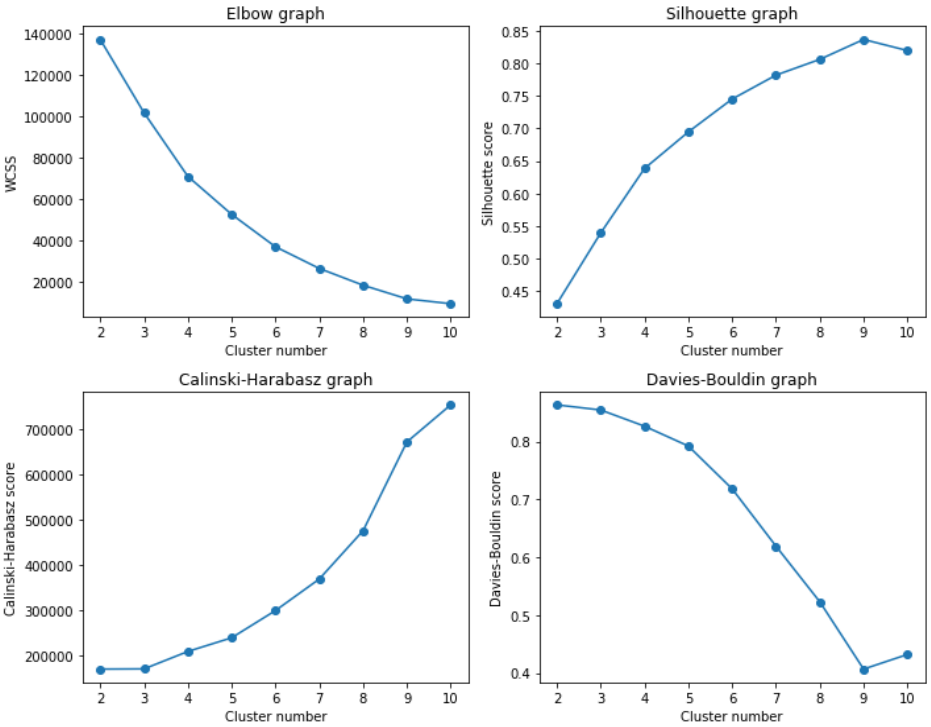

Figure 9 Clustering evaluation graphs for the debit dataset for values of k from 2 to 10

K-Means clustering algorithm was implemented with *k=9* clusters for each dataset. Customers were segmented into nine groups according to their PCA scores. To evaluate the solution for each dataset, the silhouette coefficient of each of the 9 clusters was calculated. Values above 0.5 indicate clusters that are far from neighboring clusters. Values close to 0 show clusters that are not compact and are not far from other clusters. Table 10 shows the silhouette coefficients for the 9 clusters of the credit and debit datasets. Clusters for the debit dataset seem to have higher silhouette coefficients than clusters for the credit dataset. For most clusters, their silhouette scores are above 0.6. The only exceptions are cluster 3 for the credit dataset and cluster 5 for the debit dataset which have small silhouette coefficients, indicating that they do not have good cohesion or separation from the other clusters.

	Silhouette coefficient	
Cluster Number	**Credit dataset**	**Debit dataset**
1	0.8248	0.8596
2	0.7868	0.9001
3	0.3925	0.8399
4	0.8221	0.8055
5	0.6419	0.5906
6	0.7505	0.9249
7	0.7034	0.8140
8	0.6649	0.8586
9	0.6800	0.7992

Table 10 Silhouette coefficient for each cluster of the two datasets

Table 11 shows the distribution of the revealed clusters for each dataset. For each cluster the number of customers is shown, as well as the percentage of customers from the total population of the dataset belonging to that cluster. For the credit dataset, cluster 2 is the largest one with clusters 3 and 4 following, while for the debit dataset cluster 2 is the largest one with clusters 1, 3 and 5 following. Clusters 7 and 9 for the credit dataset and 6, 8 and 9 for the debit dataset are the smallest ones.

	Number of customers (% of customers)	
Cluster Number	**Credit dataset**	**Debit dataset**
1	16,762 (10.28%)	45,822 (14.37%)
2	41,868 (25.68%)	112,944 (35.42%)
3	24,399 (14.97%)	43,318 (13.59%)
4	21,947 (13.46%)	24,712 (7.75%)
5	16,566 (10.16%)	39,066 (12.25%)
6	17,469 (10.71%)	12,118 (3.80%)
7	6,460 (3.96%)	21,523 (6.75%)
8	11,533 (7.07%)	10,019 (3.14%)
9	6,032 (3.70%)	9,328 (2.93%)

Table 11 Distribution of the derived clusters for each dataset

Chapter 4 Results

4.1 Identification and characterization of clusters

The next step of the segmentation process is to examine the structure of each cluster according to the PCA scores. This is accomplished by taking the mean values of the component scores for each cluster (Tables 12 and 13). Since the component scores were calculated using percentages, their overall population mean is absolute 0.5. Therefore, values above 0.5 in absolute value indicate scores above the average and increased spending, whereas values below 0.5 denote below average scores and lower spending.

	1	2	3	4	5	6	7	8	9
Food	-0.30	0.69	-0.09	-0.43	-0.19	-0.26	-0.18	-0.19	-0.15
Services	-0.37	0.07	-0.06	0.73	-0.17	-0.36	-0.08	-0.11	-0.08
Telcos, negative Entertainment	0.67	0.01	-0.01	0.00	-0.05	-0.60	-0.01	-0.02	-0.00
Automotive, negative Entertainment	-0.29	-0.04	0.05	-0.10	0.70	-0.36	0.08	0.11	0.06
Shopping	-0.15	-0.03	0.07	-0.08	-0.26	-0.15	0.12	0.77	0.16
Clothing, Travel, negative Shopping	-0.09	-0.03	0.14	-0.06	-0.12	-0.09	0.58	-0.26	0.51
Clothing, negative Travel	-0.00	-0.00	0.02	-0.00	-0.00	-0.00	-0.58	-0.02	0.63
Medical	-0.03	-0.01	0.24	-0.02	-0.03	-0.03	-0.21	-0.05	-0.26

Table 12 Cluster centers for the credit dataset

	1	2	3	4	5	6	7	8	9
Food	-0.49	0.64	-0.44	-0.31	-0.24	-0.27	-0.29	-0.26	-0.24
Automotive, negative Entertainment	-0.66	-0.02	0.66	0.04	0.03	0.04	0.04	0.03	0.03
Shopping, negative Automotive, Entertainment	-0.32	-0.05	-0.40	0.60	0.18	0.24	0.44	0.21	0.20
Clothing, negative Shopping	-0.04	-0.01	-0.05	-0.60	0.06	0.11	0.68	0.08	0.08
Gambling, negative Clothing	-0.07	-0.02	-0.07	-0.23	0.22	0.65	-0.39	0.35	0.28
Gambling, negative Telcos	0.02	0.01	0.02	0.06	-0.17	0.68	0.08	-0.49	-0.29
Telcos, negative Medical	0.01	0.00	0.01	0.02	-0.15	0.09	0.03	0.71	-0.48
Medical, negative Services	0.00	0.00	0.00	0.01	-0.22	0.04	0.01	0.11	0.65

Table 13 Cluster centers for the debit dataset

For the credit dataset, cluster 1 is associated with spending in telecommunication services. Cluster 2 is characterized by increased spending in the Food category while customers of cluster 3 spend in all categories. Cluster 4 relates to spending in the Services category and

cluster 5 has customers who spend mostly in the Automotive category. Cluster 6 has high negative correlation with the 3rd and 4th principal components, corresponding to high spending in the Entertainment category. Cluster 7 shows increased usage for traveling expenses. Clusters 8 and 9 are associated with spending in the Shopping category and clothing stores respectively.

The obtained clusters for the debit dataset have some similarities with the clusters of the credit dataset, but they also have some definite differences. More specifically, for the debit dataset, cluster 1 is associated with spending in the Entertainment category and cluster 2 is characterized by increased spending in the Food category. Customers of clusters 3 and 4 tend to spend more money in the Automotive and Shopping categories respectively. Cluster 5 contains customers who spend in all categories. Cluster 6 relates to spending in the Gambling category while cluster's 7 customers tend to make purchases at clothing stores. Telecommunication fees is the main characteristic for customers of cluster 8 whereas cluster 9 is associated with spending in health services.

These results were also evaluated by observing the mean percentage of total spending by category for each cluster and the mean purchase amount by category for each cluster.

Clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Telcos	**93.35**	0.67	2.99	1.23	1.33	1.49	1.22	1.53	1.35
Entertainment	1.83	1.50	3.74	0.93	1.94	**90.25**	2.06	2.37	1.38
Food	1.39	**92.17**	10.24	1.29	5.65	3.30	1.74	3.70	3.48
Services	0.85	0.57	3.16	**93.16**	1.61	0.84	1.85	1.87	1.20
Shopping	0.77	1.09	3.42	0.88	1.24	1.29	1.14	**86.10**	2.66
Automotive	0.44	1.34	4.99	1.00	**84.49**	0.97	2.88	1.11	1.03
Transportation	0.27	0.19	10.46	0.28	0.78	0.53	0.90	0.31	0.29
Home/Garden	0.19	0.48	14.04	0.23	0.63	0.22	0.37	0.68	0.69
Utilities	0.19	0.13	11.89	0.12	0.25	0.05	0.23	0.19	0.15
Clothing	0.15	0.46	1.76	0.26	0.43	0.23	0.57	1.09	**86.40**
Gambling	0.15	0.09	6.31	0.05	0.11	0.11	0.07	0.11	0.06
Medical	0.14	0.93	17.06	0.22	0.63	0.25	0.38	0.41	0.65
Business to Business	0.11	0.23	5.01	0.09	0.36	0.14	0.17	0.26	0.28
Travel and Leisure	0.08	0.05	0.88	0.20	0.41	0.24	**86.23**	0.15	0.18
Personal Care	0.03	0.07	1.37	0.02	0.06	0.06	0.07	0.08	0.15
Government	0.03	0.03	1.82	0.03	0.07	0.02	0.09	0.04	0.03
Education	0.01	0.00	0.87	0.01	0.01	0.01	0.02	0.01	0.03

Table 14 Mean percentage of total spending by category for each cluster of the credit dataset

Clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Entertainment	**94.75**	0.98	1.04	1.49	1.58	0.62	1.20	1.04	1.15
Food	2.49	**96.45**	3.08	2.90	2.84	1.04	2.64	1.55	3.37
Shopping	0.61	0.53	0.50	**92.52**	1.04	0.21	1.27	0.50	0.73
Automotive	0.56	0.67	**93.86**	0.63	1.05	0.32	0.57	0.37	0.69
Transportation	0.34	0.15	0.24	0.24	17.63	0.06	0.19	0.21	0.22
Clothing	0.32	0.29	0.25	0.86	0.68	0.06	**92.73**	0.33	0.59
Medical	0.17	0.25	0.19	0.30	0.37	0.04	0.31	0.16	**92.36**
Telcos	0.17	0.12	0.15	0.20	0.36	0.15	0.22	**94.67**	0.12

Home/Garden	0.13	0.15	0.17	0.22	21.01	0.06	0.25	0.15	0.21
Gambling	0.11	0.11	0.18	0.13	0.16	**97.30**	0.07	0.30	0.10
Personal Care	0.10	0.10	0.08	0.13	9.29	0.02	0.22	0.08	0.23
Services	0.10	0.07	0.09	0.18	21.43	0.06	0.11	0.17	0.09
Business to Business	0.07	0.09	0.09	0.12	6.37	0.04	0.15	0.08	0.09
Travel and Leisure	0.06	0.03	0.05	0.05	10.62	0.02	0.05	0.03	0.05
Utilities	0.01	0.01	0.01	0.01	3.54	0.00	0.01	0.33	0.01
Government	0.00	0.00	0.01	0.01	1.37	0.00	0.01	0.02	0.00
Education	0.00	0.00	0.00	0.01	0.65	0.00	0.00	0.00	0.00

Table 15 Mean percentage of total spending by category for each cluster of the debit dataset

Clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Telcos	**28.00**	0.61	4.78	1.26	1.55	0.94	3.20	1.36	1.47
Entertainment	0.79	0.79	6.06	1.34	2.20	**39.25**	8.19	1.71	1.87
Food	0.78	**41.35**	16.71	2.02	5.93	1.97	5.48	3.03	3.65
Services	0.69	0.75	7.94	**107.59**	2.72	0.98	8.03	2.09	2.15
Automotive	0.60	1.95	11.74	2.21	**82.71**	1.42	11.74	1.78	1.93
Shopping	0.53	0.94	6.73	1.21	1.65	1.05	4.74	57.40	3.66
Utilities	0.44	0.27	21.31	0.38	0.52	0.08	1.14	0.36	0.33
Travel and Leisure	0.22	0.12	3.79	1.02	1.38	0.84	**310.64**	0.47	0.55
Transportation	0.19	0.21	9.84	0.50	1.52	0.48	3.27	0.35	0.45
Clothing	0.18	0.56	3.86	0.56	0.69	0.41	2.22	1.63	**87.30**
Home/Garden	0.17	0.50	23.47	0.45	1.15	0.22	1.61	0.80	0.98
Medical	0.11	0.97	20.50	0.41	0.82	0.27	1.52	0.49	0.89
Business to Business	0.11	0.19	6.36	0.16	0.48	0.11	0.61	0.34	0.39
Gambling	0.07	0.04	8.82	0.06	0.08	0.07	0.37	0.06	0.06
Government	0.06	0.05	6.56	0.18	0.16	0.03	0.53	0.06	0.05
Personal Care	0.02	0.07	1.42	0.03	0.09	0.05	0.26	0.11	0.18
Education	0.01	0.01	4.98	0.02	0.01	0.03	0.13	0.02	0.10

Table 16 Mean purchase amount per category for each cluster of the credit dataset

Clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Entertainment	**25.88**	0.40	0.60	0.72	1.10	0.29	0.77	0.41	0.65
Food	1.02	**27.14**	1.75	1.61	2.12	0.46	1.67	0.77	1.78
Automotive	0.41	0.45	**42.88**	0.54	1.08	0.24	0.47	0.30	0.54
Shopping	0.27	0.29	0.32	**41.81**	0.92	0.11	0.97	0.28	0.52
Clothing	0.21	0.20	0.20	0.74	0.75	0.04	**45.19**	0.29	0.49
Transportation	0.10	0.05	0.14	0.11	4.71	0.02	0.11	0.08	0.10
Medical	0.09	0.12	0.13	0.22	0.33	0.02	0.24	0.11	**48.09**
Home/Garden	0.08	0.09	0.13	0.17	17.32	0.07	0.19	0.10	0.12
Telcos	0.08	0.06	0.10	0.13	0.37	0.08	0.16	**32.26**	0.08
Services	0.07	0.04	0.08	0.14	19.99	0.04	0.09	0.14	0.08
Personal Care	0.06	0.06	0.05	0.12	3.92	0.01	0.18	0.04	0.24
Gambling	0.05	0.04	0.08	0.10	0.13	**46.63**	0.04	0.15	0.04
Travel and Leisure	0.05	0.02	0.05	0.08	16.55	0.04	0.08	0.06	0.04
Business to Business	0.03	0.05	0.07	0.08	3.24	0.02	0.10	0.06	0.07
Utilities	0.01	0.01	0.01	0.02	3.97	0.01	0.01	0.40	0.01
Government	0.00	0.00	0.01	0.01	2.97	0.00	0.01	0.04	NaN
Education	0.00	0.00	0.00	0.00	1.44	NaN	0.00	0.00	NaN

Table 17 Mean purchase amount per category for each cluster of the debit dataset

The results from Tables 14, 15, 16 and 17 agree with the results of the significance of each component in each cluster. For the credit dataset, customers of cluster 3 who exhibit no specific behavior, tend to spend more in the categories of Food, Transportation, Utilities, Home/Garden and Medical, while they also spend some money for their vehicles. For the debit dataset, the customers of cluster 5 tend to spend more in the categories of Home/Garden, Services and Travel and Leisure while also spending an average of 17.63% of their total spending in the Transportation category.

4.2 Profiling of clusters

The next step involves the creation of the demographic profile of each cluster. To make this easier, labels were given to each cluster according to its most popular spending category, as described in 4.1. Table 18 shows the cluster labels for both datasets.

	Credit dataset	**Debit dataset**
Cluster 1	Telcos	Entertainment
Cluster 2	Food	Food
Cluster 3	All Categories	Automotive
Cluster 4	Services	Shopping
Cluster 5	Automotive	All Categories
Cluster 6	Entertainment	Gambling
Cluster 7	Travel	Clothing
Cluster 8	Shopping	Telcos
Cluster 9	Clothing	Medical

Table 18 Cluster labels

Spending in the categories of Telcos, Food, Automotive, Entertainment, Shopping and Clothing appear in both datasets. There is also a cluster in each dataset with customers with unspecific behavior who spend in all categories. Spending in Services and travel expenses appear in the credit dataset, while Gambling and medical expenses are distinct clusters in the debit dataset.

In the following sections, each cluster from each dataset is going to be evaluated based on the demographical attributes and the spending behavior of its customers. The spending behavior is going to be evaluated according to the percentages of total spending and total purchases in each general category for each cluster. For the demographical attributes, percentages for each value will be given for each cluster. Also, for each cluster, the dataset from which it comes from, its label, its number of customers and the percentage of total customers of the dataset will be shown.

4.2.1 Credit dataset

Cluster 1 / Telcos: 10.28% of all credit card customers are in cluster 1. They make many purchases and spend more money in the Telcos category as opposed to other categories. Their secondary most popular categories are Entertainment, Food, Services, Shopping and Automotive. For the first four categories, there are more purchases made than money spent. For the Automotive category, although the purchases are only 1.38% of their total purchases, they spend 1.83% of their money in these purchases. They are at a percentage of 63.32% men, most of them are over 25 years old and there are 10% more married people than unmarried. People over 65 years old correspond to a small percentage of 7.64% in this cluster. Considering their educational level and occupation, the vast majority has at least High School Education and they are employed.

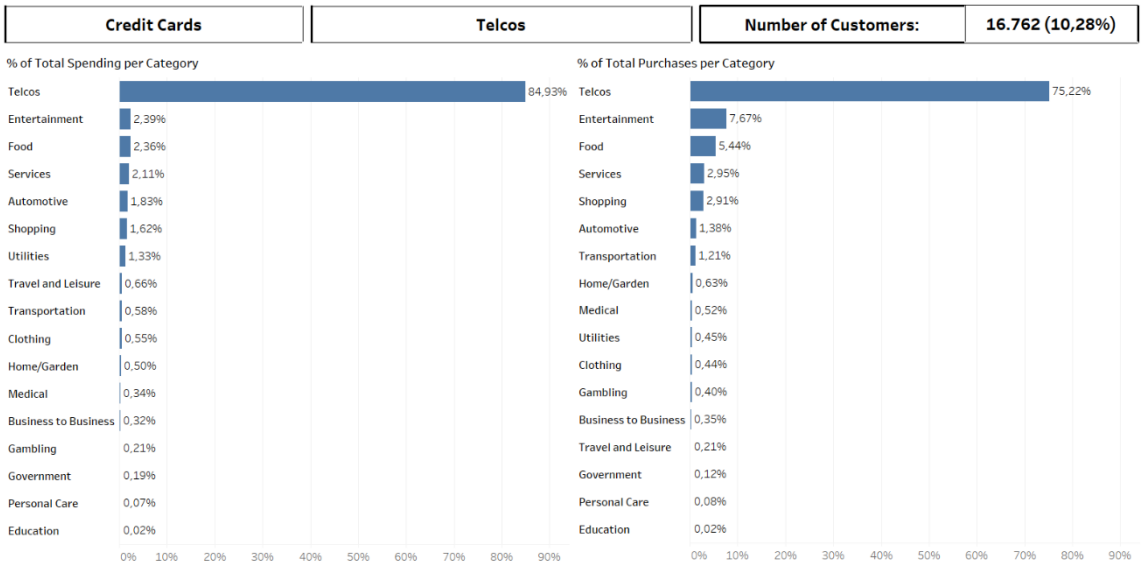

Figure 10 Percentages of total spending and purchases per category for cluster 1 / Telcos

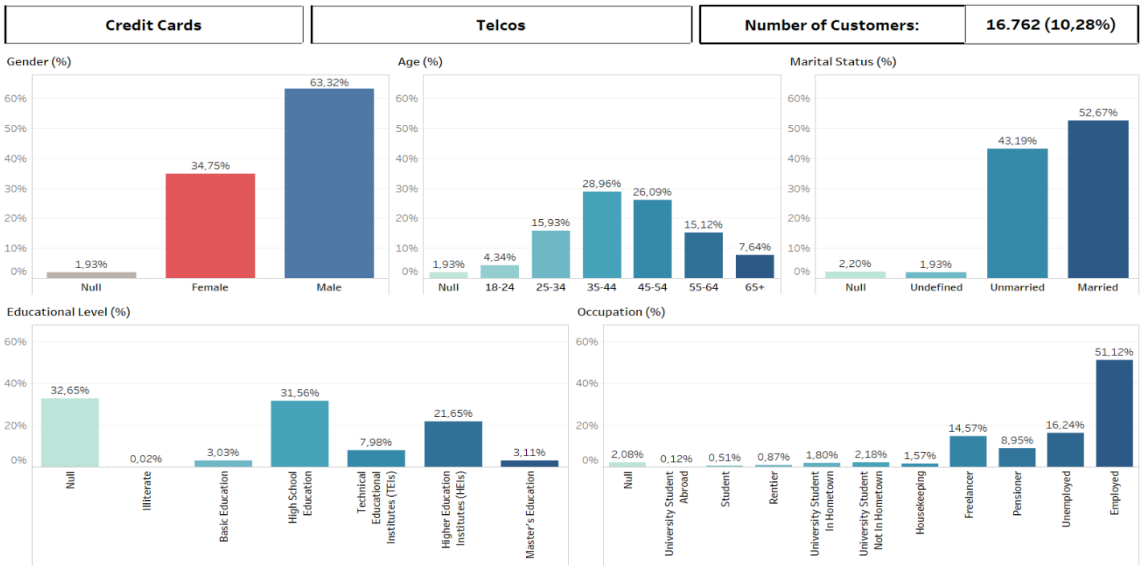

Figure 11 Percentages of the values of categorial attributes for cluster 1 / Telcos

Cluster 2 / Food: This cluster's customers comprise the 25.68% of total customers and in addition to spending in Supermarkets and other convenience stores, they also make purchases and spend money in the Entertainment, Shopping, Automotive and Medical categories. They are by a percentage of 61.42% married, are mostly above 45 years old and are both men and women. There is also a 16.09% in the age category 35-44 while younger people appear in smaller percentages in this cluster. There is a significant percentage of customers with basic education in this cluster and regarding their occupation, 48.89% of them are unemployed, 24.13% employed and there is a 14.12% of pensioners.

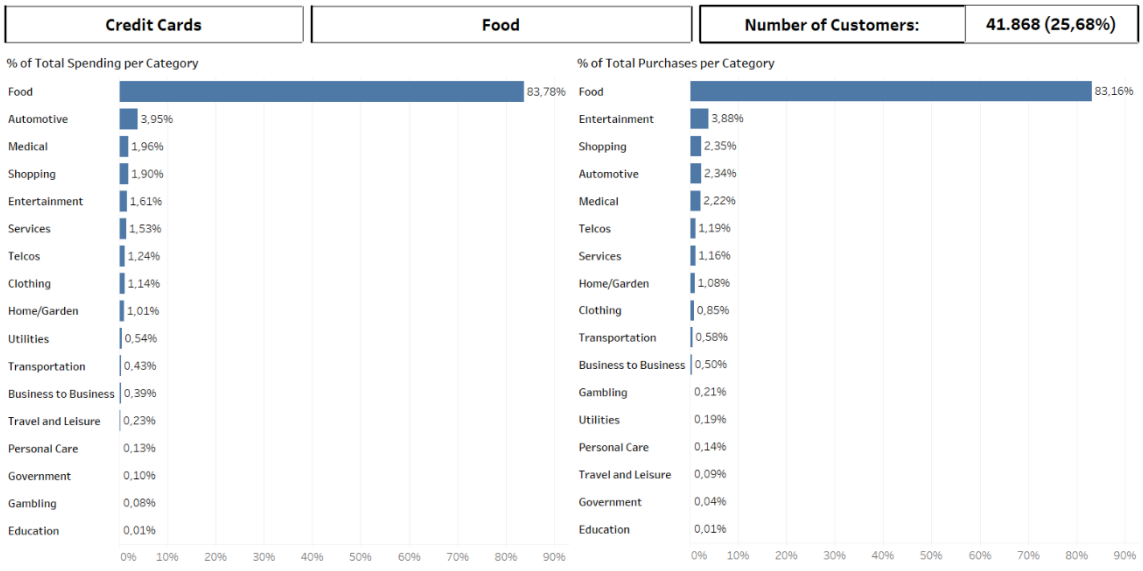

Figure 12 Percentages of total spending and purchases per category for cluster 2 / Food

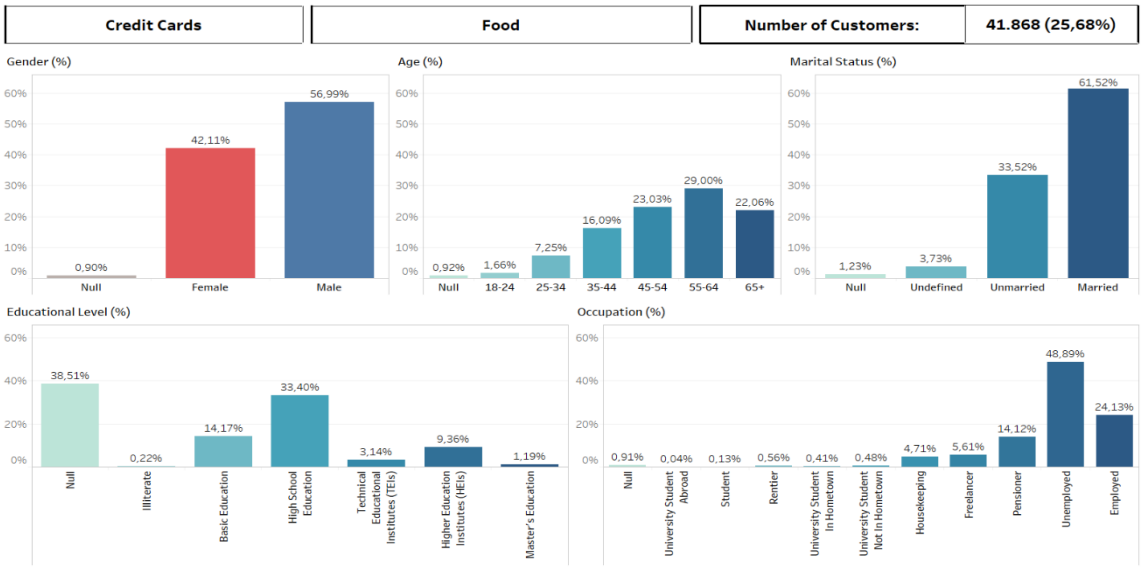

Figure 13 Percentages of the values of categorial attributes for cluster 2 / Food

Cluster 3 / All Categories: 14.97% of the credit cards customers belong in this cluster. These customers do not follow a specific behavior but spend in many categories. They make more purchases in the Food and Entertainment categories and spend more money in the Home/Garden and Utilities categories. They also make about 15% of their total purchases and spend 13% of their total spending in expenses related to their vehicles and transportation. They are mostly male, married and above 35 years old. Most of them have at least High School Education and are at 40.55% Employed, 21.95% Pensioners and 21.18% of them are Unemployed.

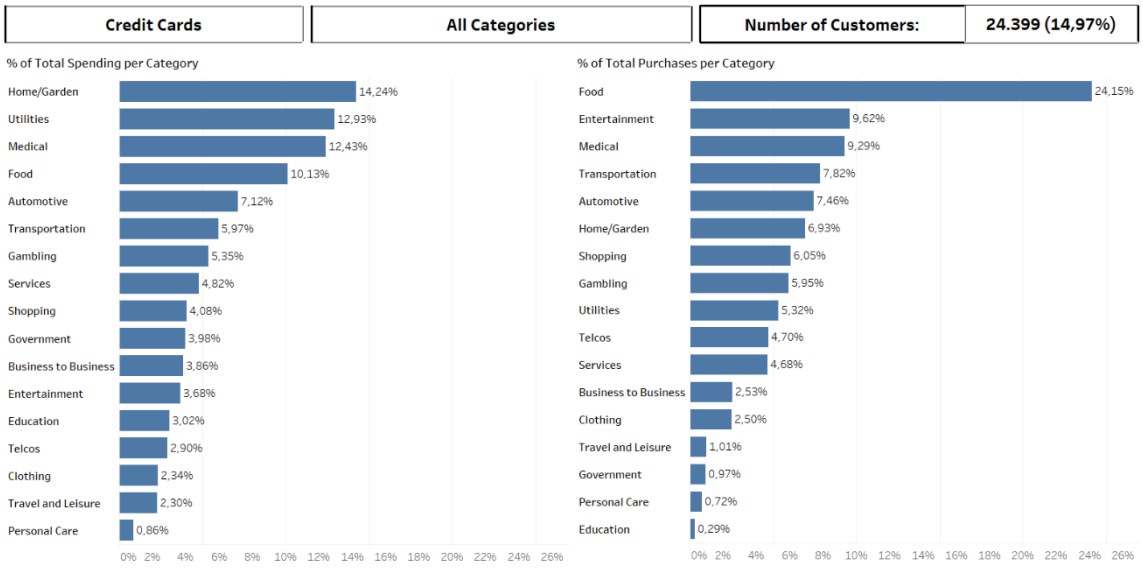

Figure 14 Percentages of total spending and purchases per category for cluster 3 / All Categories

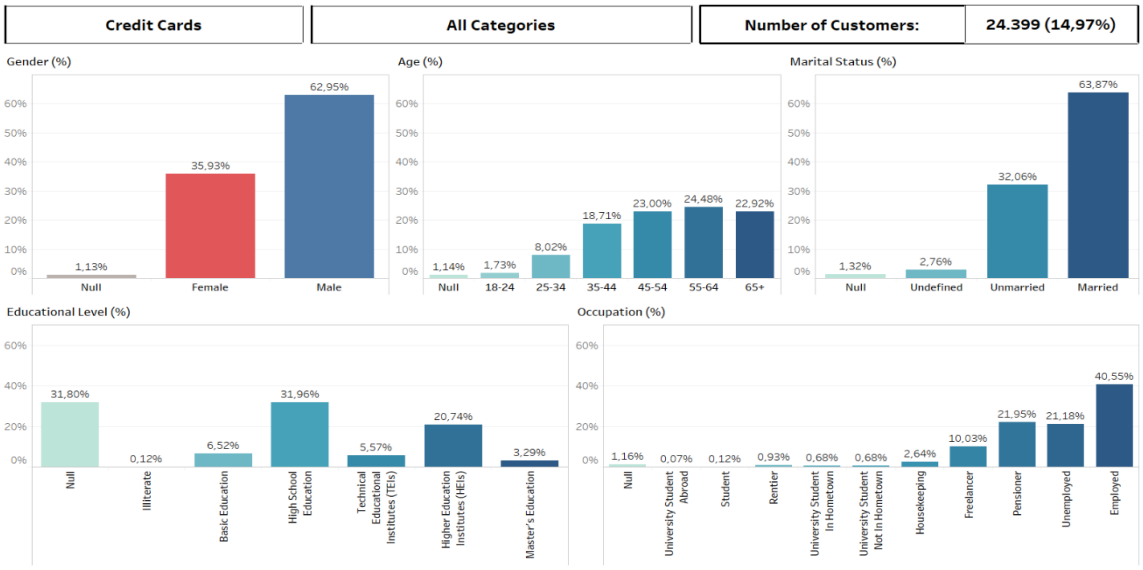

Figure 15 Percentages of the values of categorial attributes for cluster 3 / All Categories

<u>Cluster 4 / Services:</u> There are 13.46% of customers who spend more money in the Services category than the other categories. They also make a few purchases in the categories of Food, Entertainment, Telcos, Shopping and Automotive. 69.03% of them are married, 65.01% are male and more than 90% of this cluster's customers are older than 35. They are mostly employed, freelancers or pensioners. Regarding their Educational level, 27.84% of them have a High School Education, while a percentage of 27.08% has finished a Higher Education Institute (HEI).

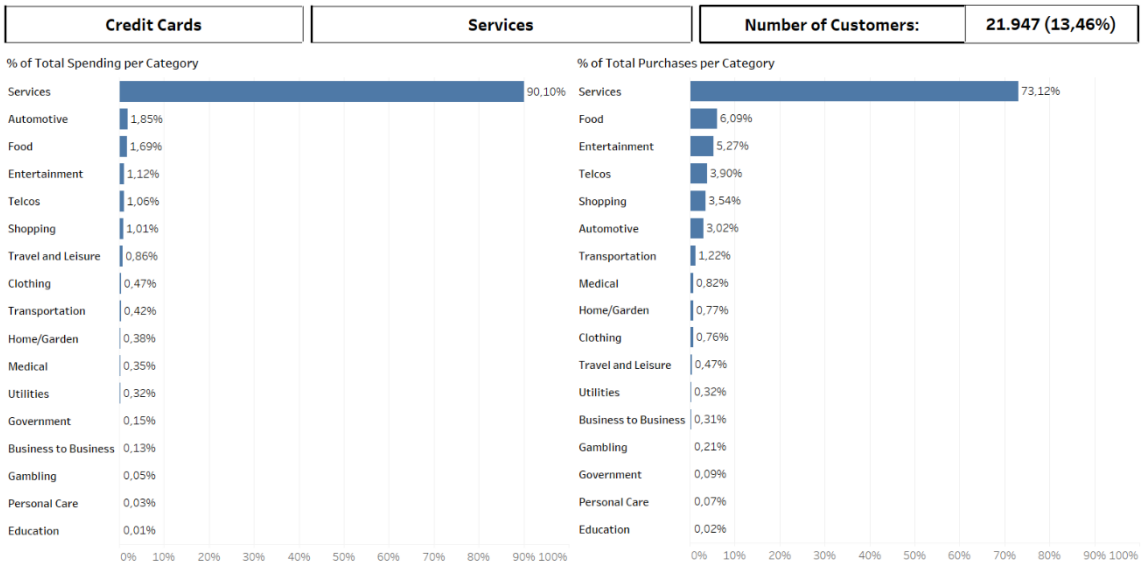

Figure 16 Percentages of total spending and purchases per category for cluster 4 / Services

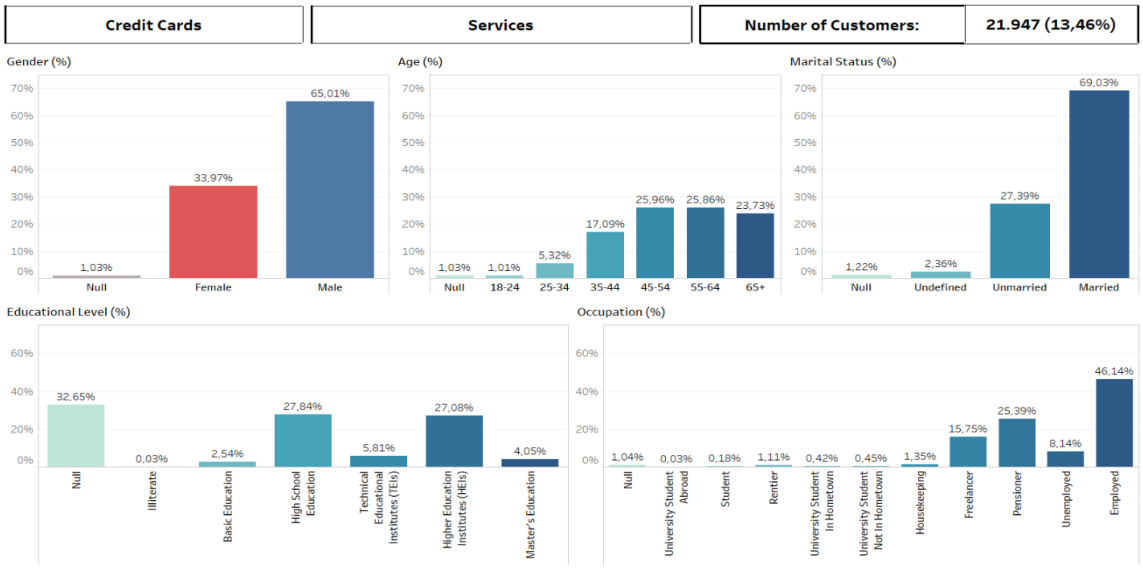

Figure 17 Percentages of the values of categorial attributes for cluster 4 / Services

<u>Cluster 5 / Automotive:</u> 10.16% spend mostly in the Automotive category. Customers of this cluster also make a significant amount of purchases in the Food, Entertainment and Transportation categories. There are 33% more married customers than unmarried ones. 70% of the customers are aged between 35 and 64 years old. This cluster also has the smallest percentage of women compared to the other clusters, at 23.66% while male customers comprise 75.43% of this cluster's population. Most of them are employed, but there are also significant percentages of unemployed (21.18%), pensioners (18.17%) and freelancers (12.29%).

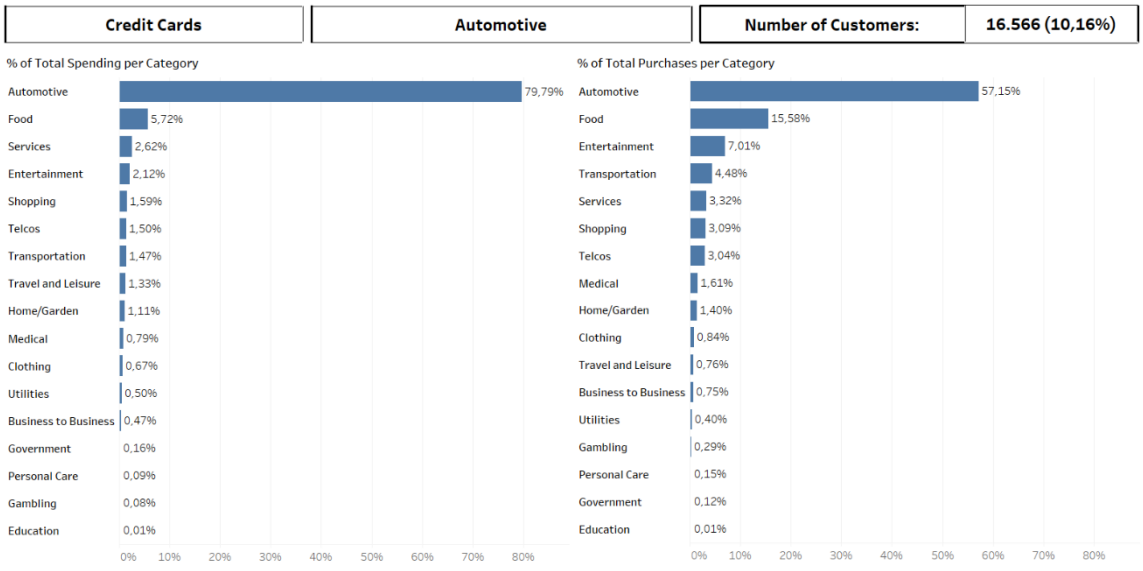

Figure 18 Percentages of total spending and purchases per category for cluster 5 / Automotive

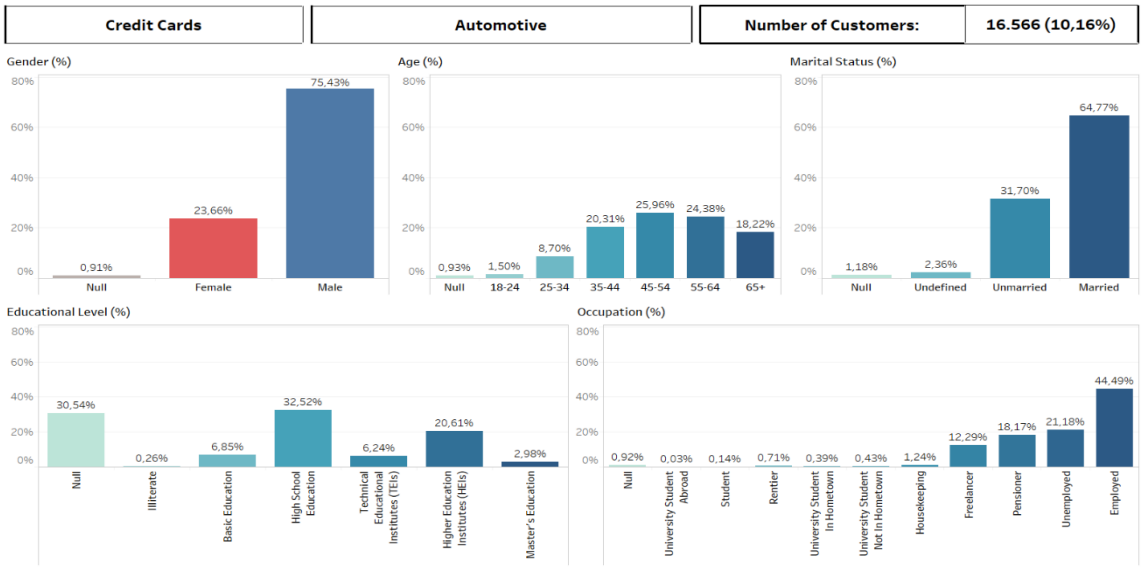

Figure 19 Percentages of the values of categorial attributes for cluster 5 / Automotive

Cluster 6 / Entertainment: This cluster's customers comprise the 10.71% of total customers and except from the Entertainment category which is their main spending category, they also spend money and make purchases in the Food category. There are about 25% more men than women but there is nearly the same percentage of married and unmarried people. Also, there is only a percentage of 6.8% older people of ages 65+, while most of them are from 25 to 64 years old. They are at a percentage of 44.53% employed, with unemployed, freelancers and pensioners following.

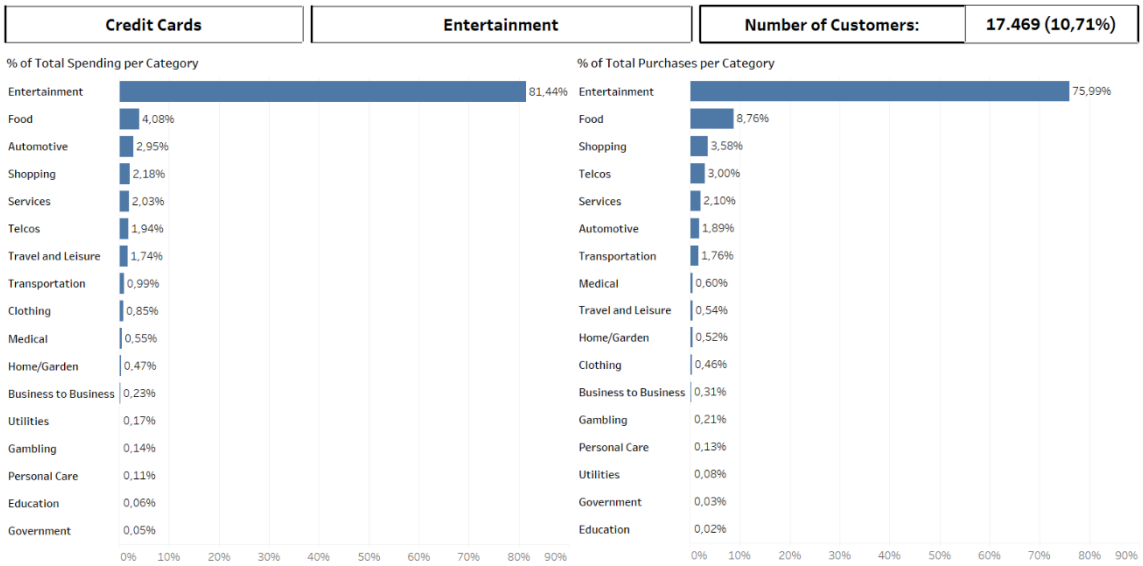

Figure 20 Percentages of total spending and purchases per category for cluster 6 / Entertainment

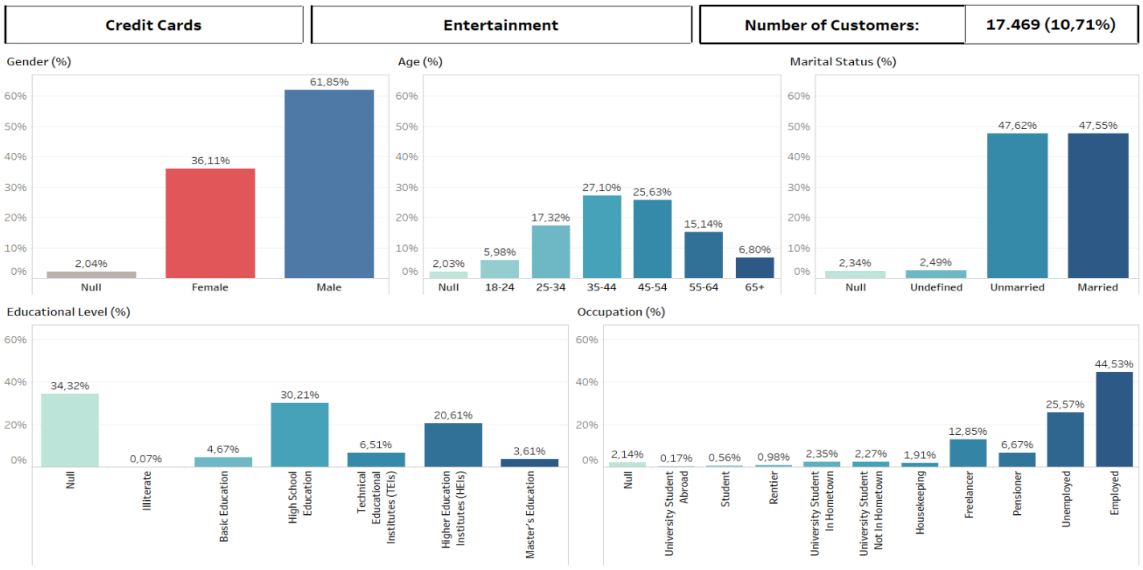

Figure 21 Percentages of the values of categorial attributes for cluster 6 / Entertainment

Cluster 7 / Travel: 3.96% of the credit dataset's customer use their cards mostly in travel expenses. They also spend some money in the Automotive, Entertainment, Services and Food categories. The purchases they make in the Travel and Leisure category is less than 50% of their total purchases. This indicates that they make few purchases but spend a lot of money in each one. The percentages of men and women in this cluster are 64.46% and 34.58% respectively while the percentages of married and unmarried people are 63.31% and 33.7% respectively. Travel expenses are more popular in people above 35 years old, but not quite popular in people above 65. 58.39% of this cluster's customers are employed, 17.23% are freelancers and there are smaller percentages of pensioners and unemployed people. It is also interesting to observe that this cluster is the first that shows a difference in the educational level, with 34.01% of its customers having finished a Higher Education Institute.

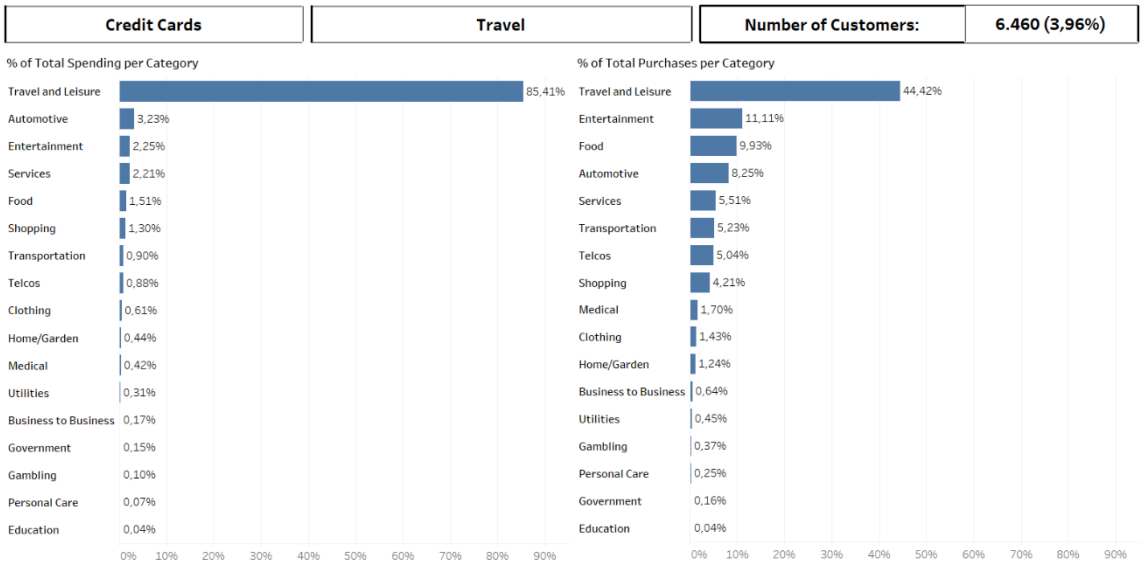

Figure 22 Percentages of total spending and purchases per category for cluster 7 / Travel

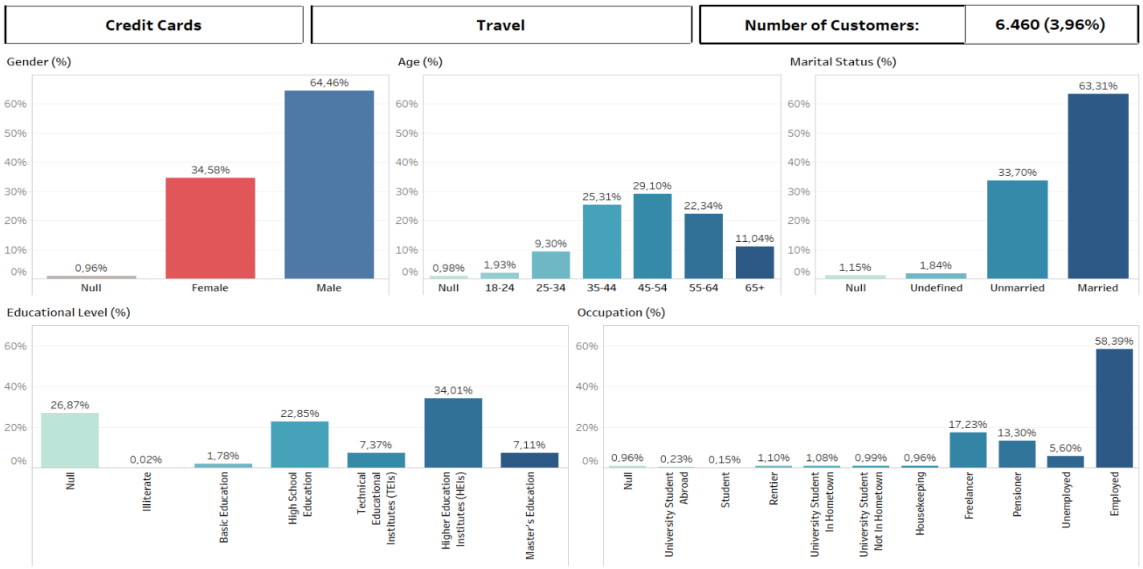

Figure 23 Percentages of the values of categorial attributes for cluster 7 / Travel

Cluster 8 / Shopping: 7.07% of customers have Shopping as their main category of spending. They also make purchases and spend money in the categories of Food, Entertainment and Automotive. There are about 15% more men than women in this cluster, mostly above 35 years old. 54.86% of the customers are married and 42.95% of them are employed.

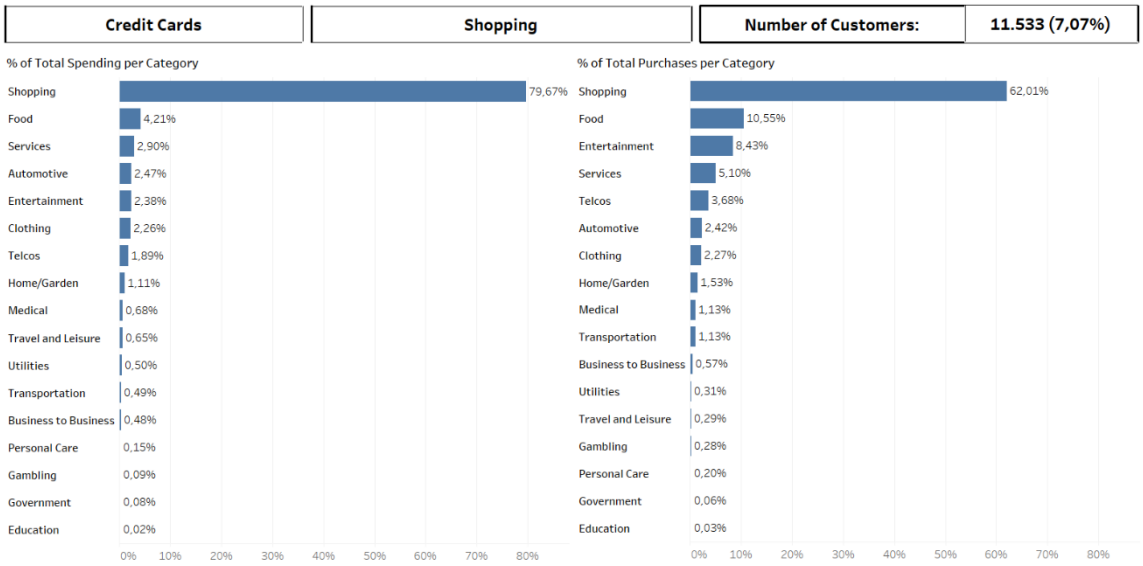

Figure 24 Percentages of total spending and purchases per category for cluster 8 / Shopping

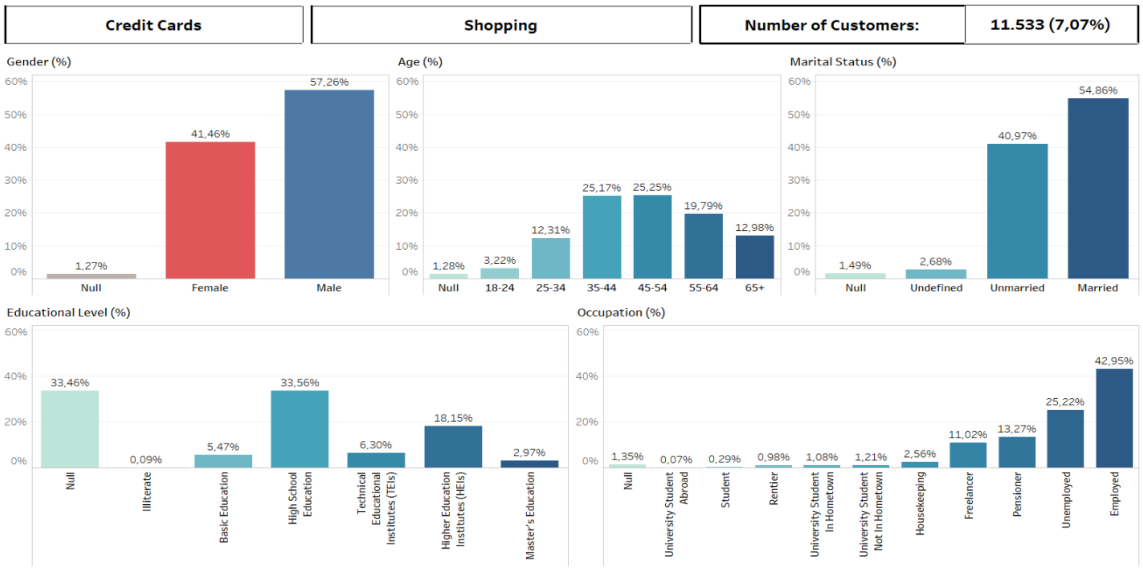

Figure 25 Percentages of the values of categorial attributes for cluster 8 / Shopping

<u>Cluster 9 / Clothing:</u> The final small percentage of 3.7% customers spend more money in clothing stores. They also make purchases and spend some money in the Food and Shopping categories. It is the first cluster with a larger percentage (63.36%) of women than men (35.61%). They are also mostly married (65.3%), employed (47.61%) and the majority of them (87%) are above 35 years old.

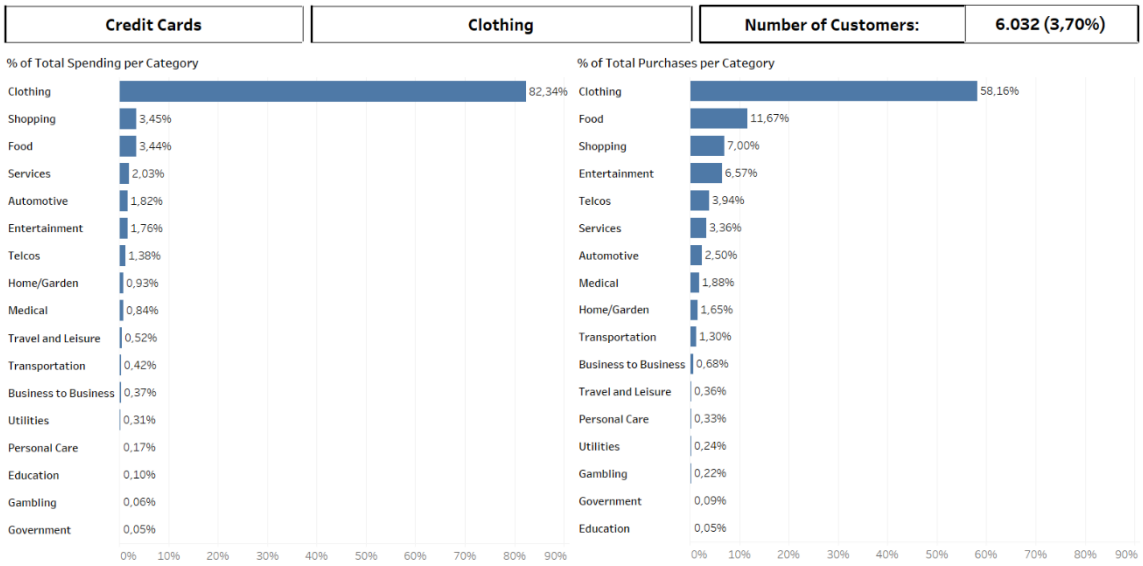

Figure 26 Percentages of total spending and purchases per category for cluster 9 / Clothing

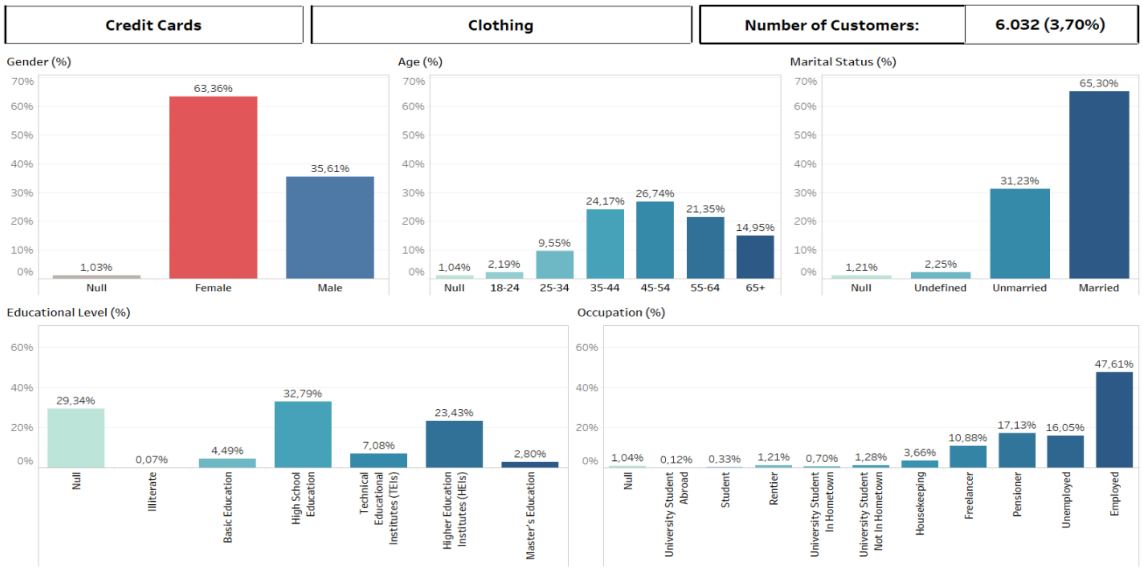

Figure 27 Percentages of the values of categorial attributes for cluster 9 / Clothing

4.2.2 Debit cards

Cluster 1 / Entertainment: 14.37% of the debit dataset's customers spend money in the Entertainment category, while they also make some purchases and spend a small amount of money in the Food category. They are mostly unmarried (56.95%) and 25 to 54 years

old (70%). The percentages of male and female customers are quite close and most of the customers are employed (53.19%).

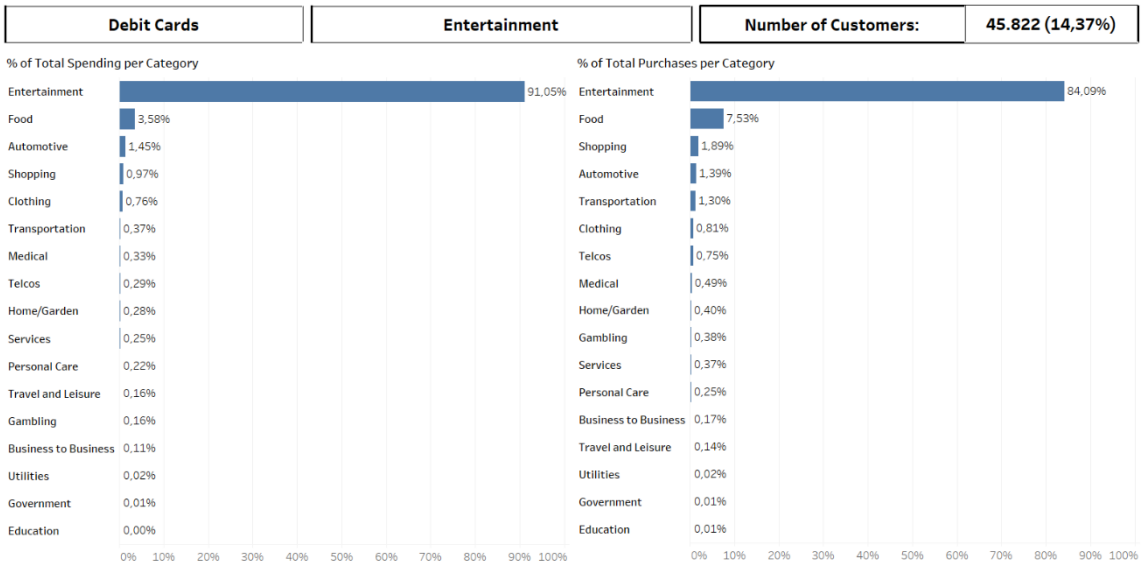

Figure 28 Percentages of total spending and purchases per category for cluster 1 / Entertainment

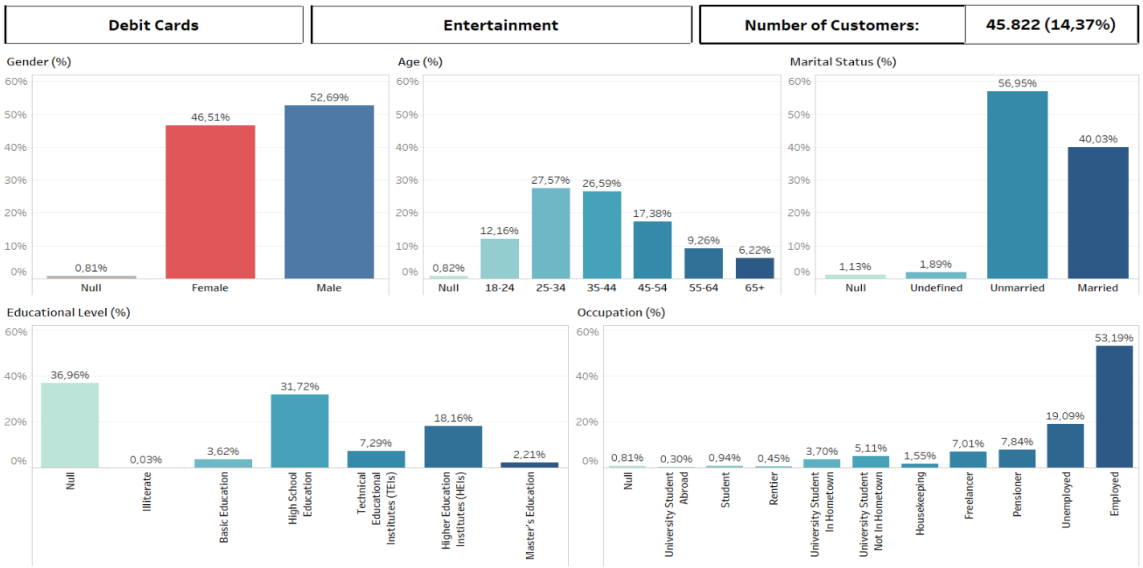

Figure 29 Percentages of the values of categorial attributes for cluster 1 / Entertainment

Cluster 2 / Food: The largest percentage of customers (35.42%) belongs in this cluster. They make very few purchases in other categories and spend mostly in the Food category. The percentages of men and women are quite close with 5.38% more women than men. Customers younger than 25 years old are only 4.29% of the customers in this cluster. The

majority of customers are older than 25 with the category of 35-44 having the largest percentage of 26.13%. 57.71% of customers are married and 51.27% employed. Unemployed people comprise the 18.43% of customers, while in the corresponding cluster of the credit dataset, there were more unemployed than employed people.

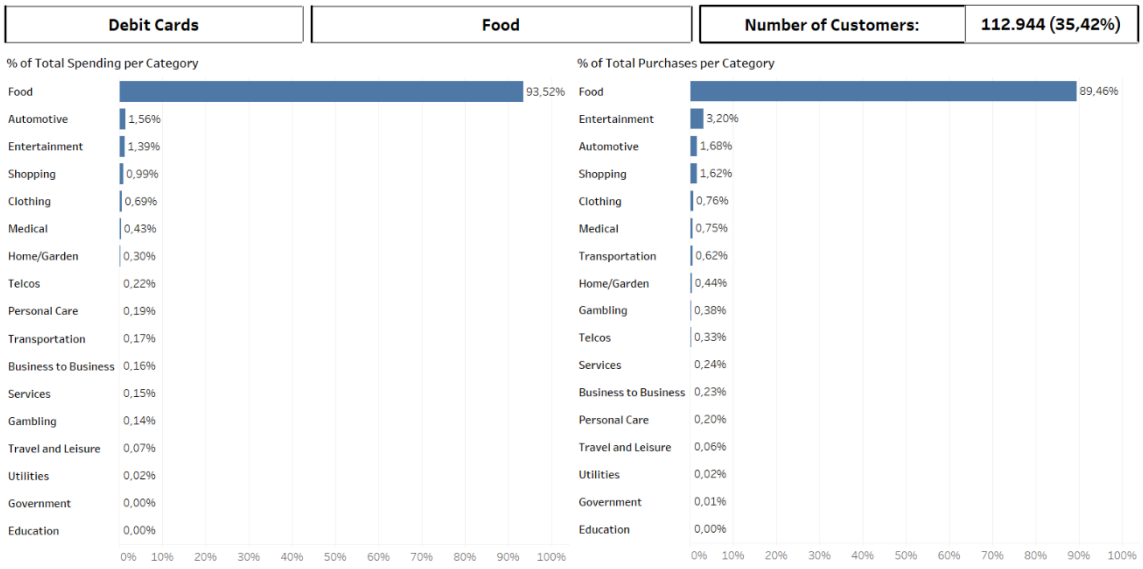

Figure 30 Percentages of total spending and purchases per category for cluster 2 / Food

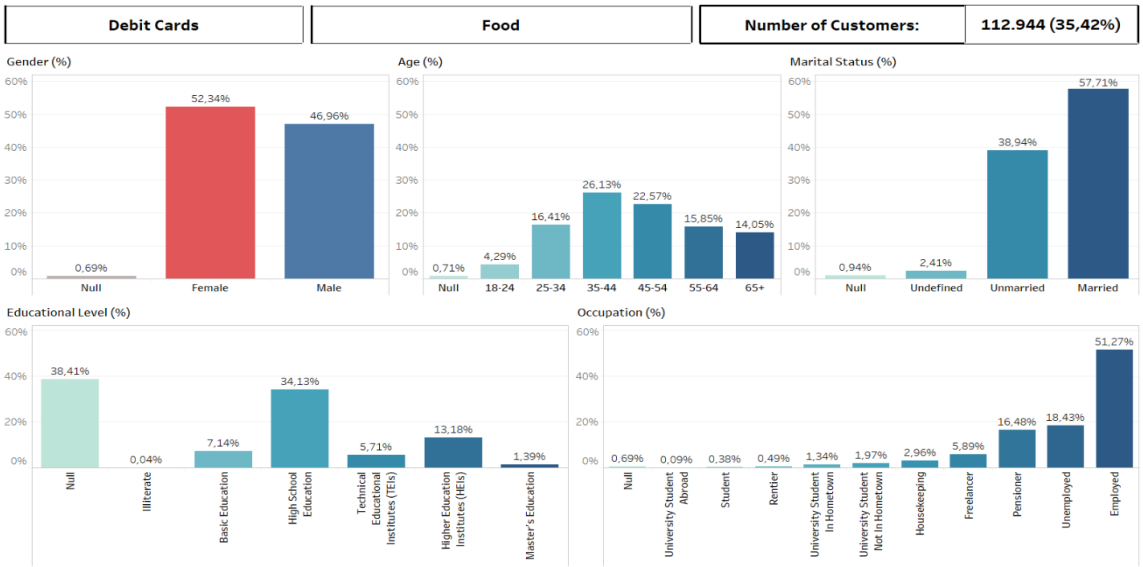

Figure 31 Percentages of the values of categorial attributes for cluster 2 / Food

Cluster 3 / Automotive: 13.59% of customers belong in this cluster and they spend more money in the Automotive category, while also making some purchases in the Food and

Entertainment categories. They are mostly men, above 25 years old and employed. Married people comprise 57.35% of this cluster.

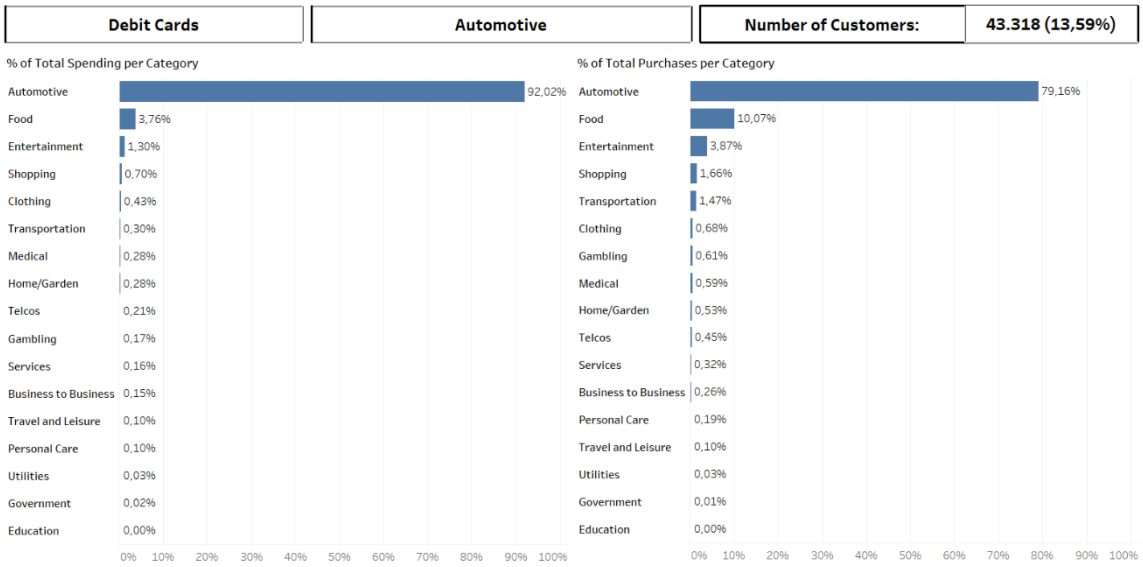

Figure 32 Percentages of total spending and purchases per category for cluster 3 / Automotive

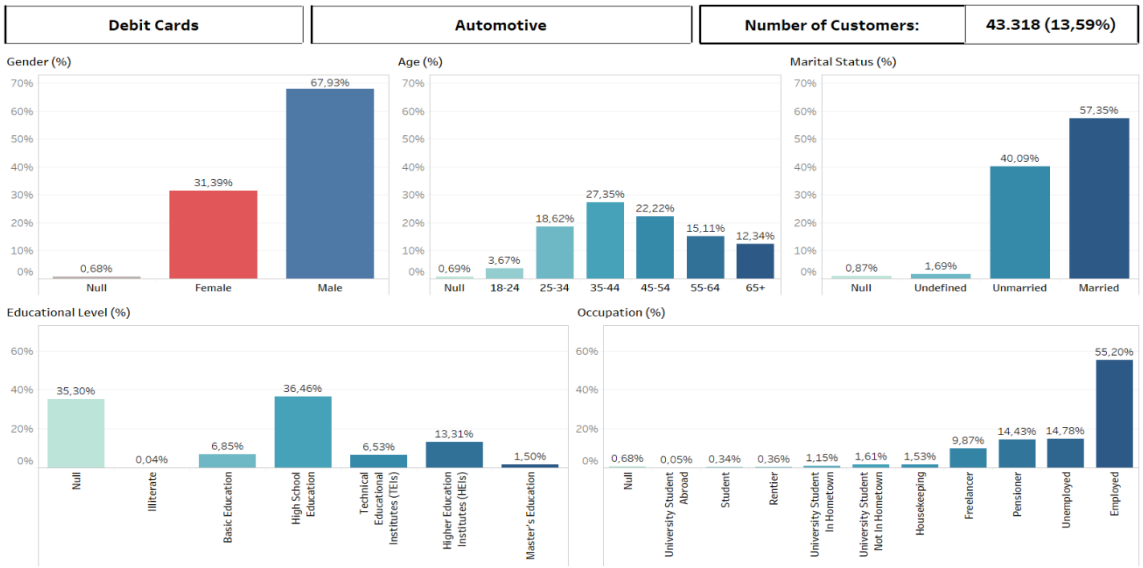

Figure 33 Percentages of the values of categorial attributes for cluster 3 / Automotive

Cluster 4 / Shopping: A smaller percentage of 7.75% customers spend mostly in the Shopping category with some purchases in the Food and Entertainment categories. The percentages of men and women are 40.62% and 58.67% respectively with more than 70%

of them aged between 25 and 54 years old. There are more unmarried (51.09%) people than married (45.72%) in this cluster. Most of the customers (54.64%) are employed.

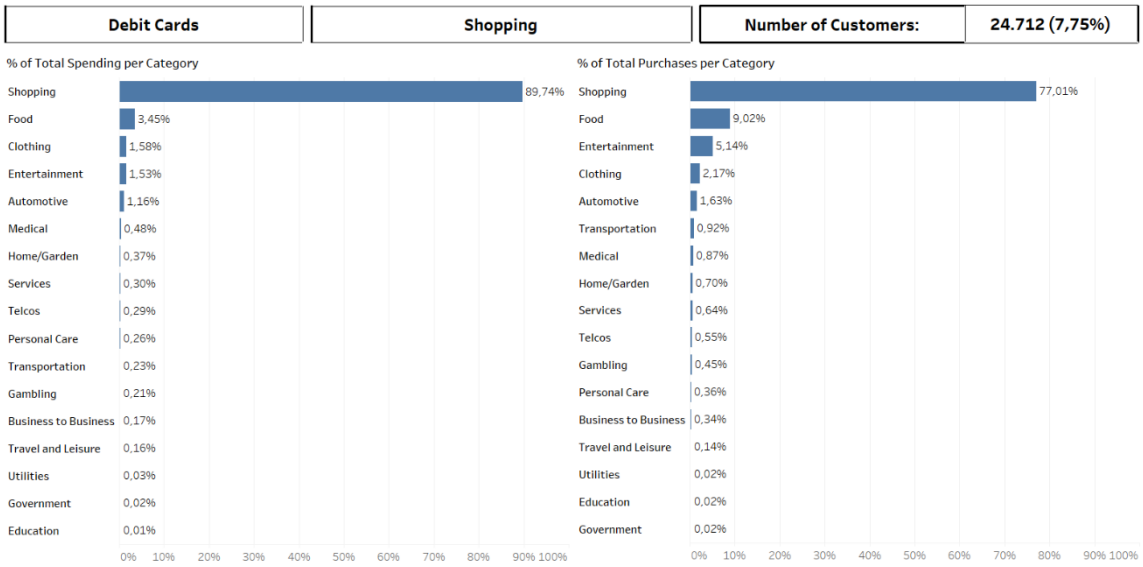

Figure 34 Percentages of total spending and purchases per category for cluster 4 / Shopping

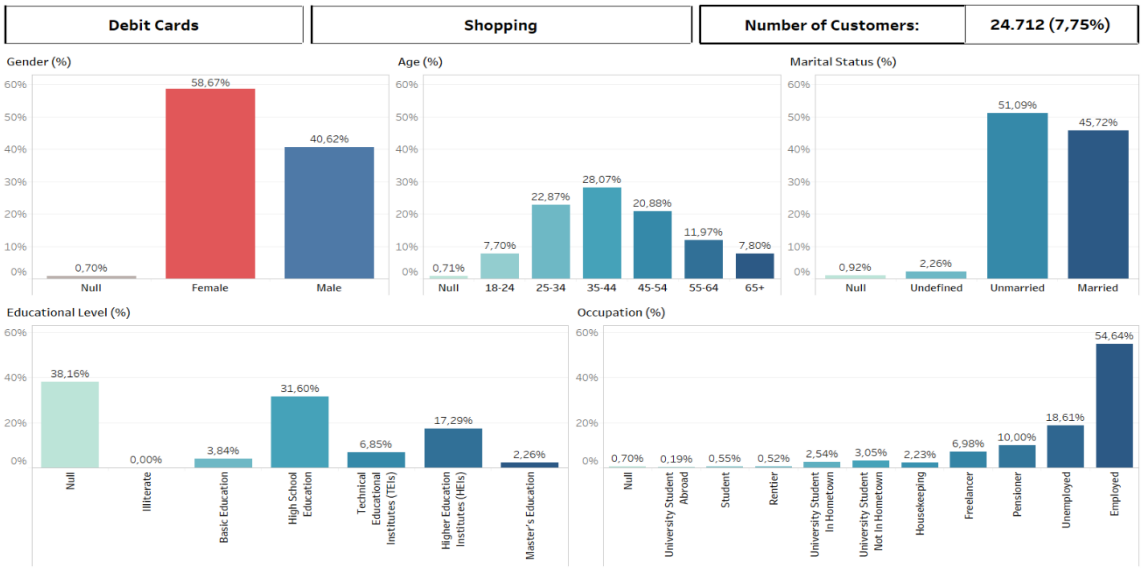

Figure 35 Percentages of the values of categorial attributes for cluster 4 / Shopping

Cluster 5 / All Categories: The cluster that corresponds to spending in all categories is a significant one, containing 12.25% of total customers of the debit dataset. They spend mostly in the categories of Services, Home/Garden and Travel and Leisure, while also making 15.34% of their total purchases in the Transportation category. They also make

9.95% of their total purchases in the Food category, but these purchases only amount to 2.61% of their total spending. They are at nearly the same percentage men (50.62%) and women (48.69%). They are mostly from 25 to 54 years old (70%). The percentages of unmarried and married people are 50.08% and 46.75% respectively. Employed people comprise 51.86% of customers, while there is a percentage of 17.15% unemployed customers.

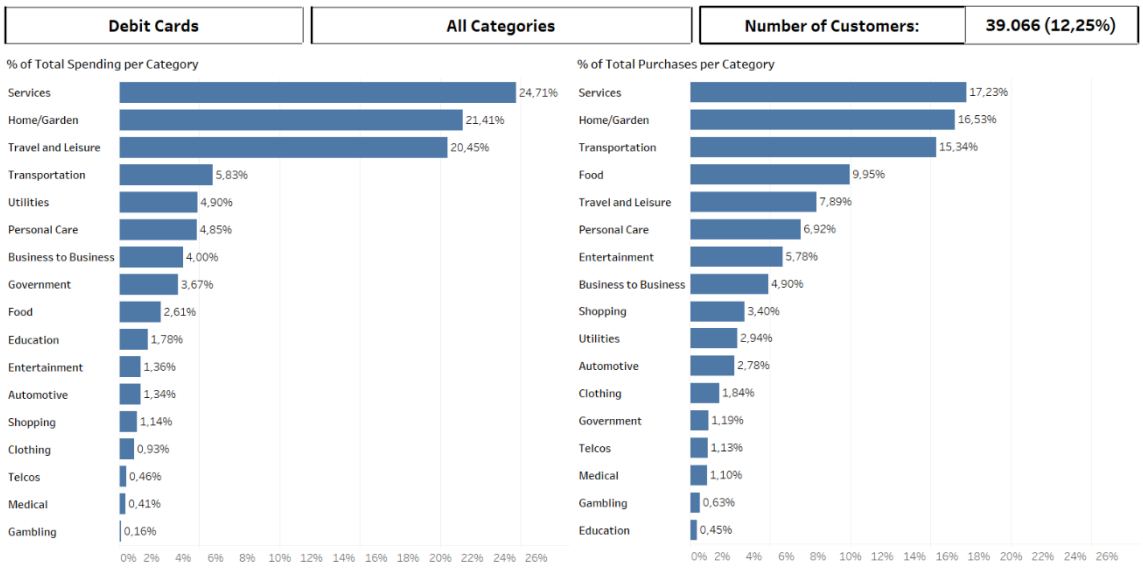

Figure 36 Percentages of total spending and purchases per category for cluster 5 / All Categories

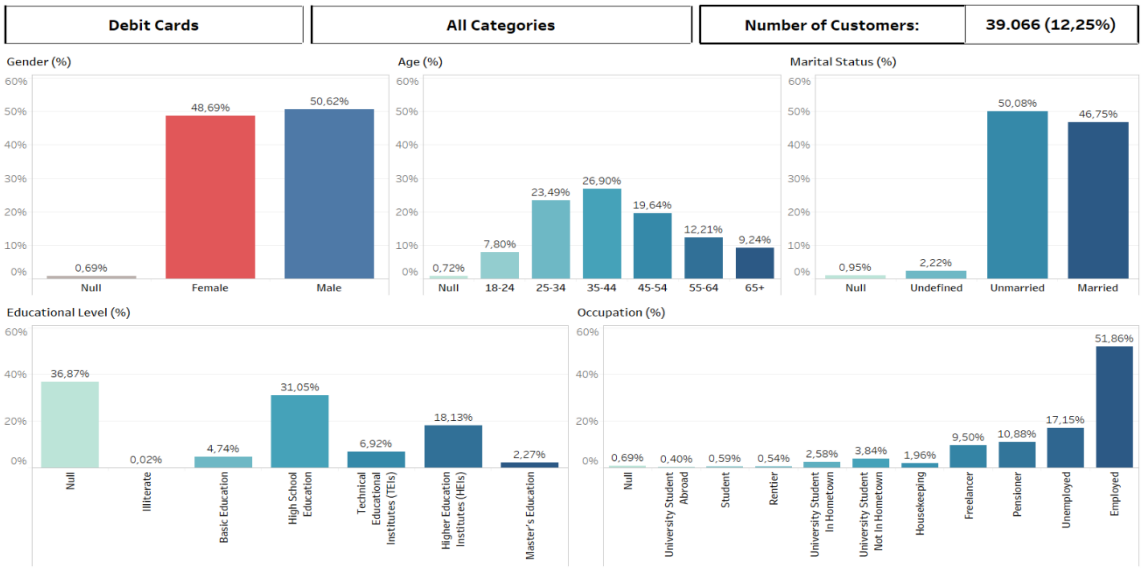

Figure 37 Percentages of the values of categorial attributes for cluster 5 / All Categories

Cluster 6 / Gambling: 3.8% of customers use their debit cards mostly for Gambling. This cluster's customers spend 96.94% of their total money in the Gambling category. This means that the primal usage of their debit cards is spending in the Gambling category while they spend little amount of money in other categories. Men comprise the majority of this cluster (81.5%) while there is a very small percentage of women (17.82%). More than 50% of them are aged between 25 and 34 years old. There are nearly 20% more unmarried customers than married ones. 50.73% of them are employed while there is also a percentage of 24.06% unemployed people.

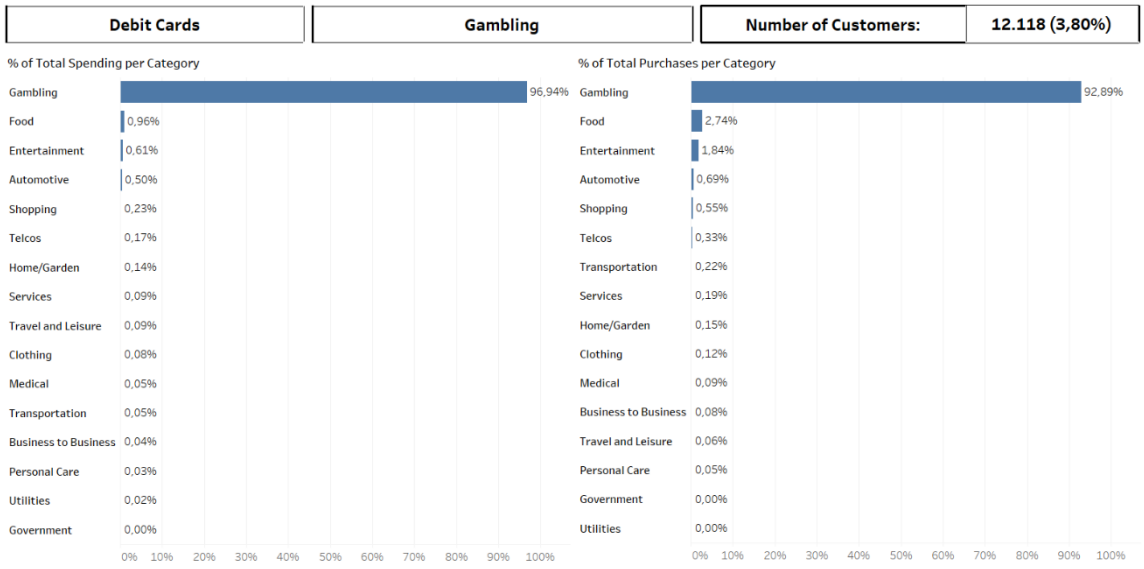

Figure 38 Percentages of total spending and purchases per category for cluster 6 / Gambling

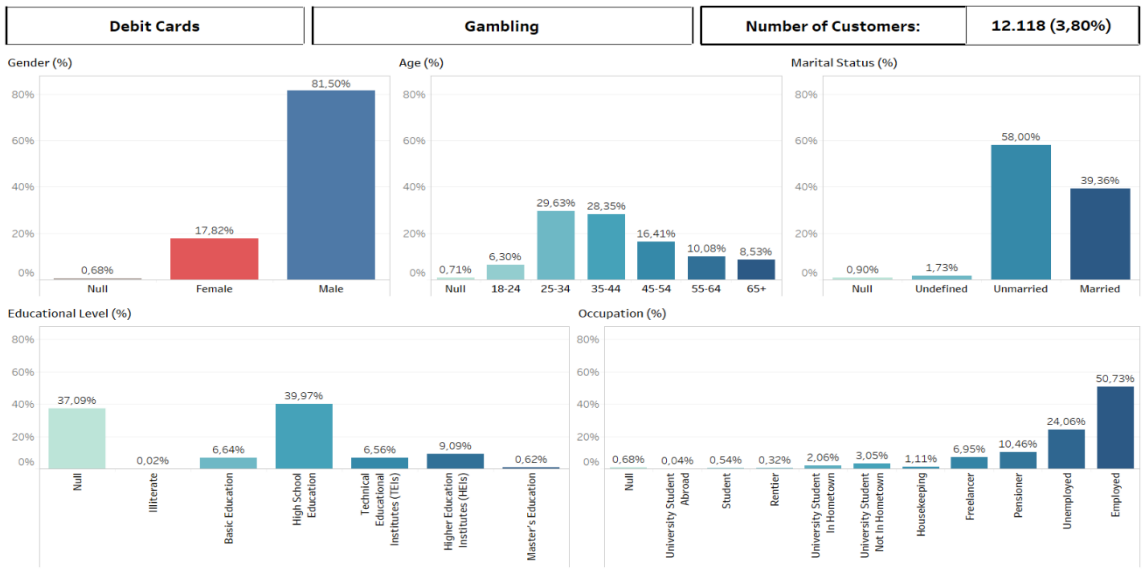

Figure 39 Percentages of the values of categorial attributes for cluster 6 / Gambling

Cluster 7 / Clothing: 6.75% of customers spend most of their money in clothing stores. They also spend some money and make purchases in the Food, Entertainment and Shopping categories. The percentage of men (24.22%) is significantly lower than the percentage of women (75.16%). More than 50% of the customers in this cluster belong in the age category 25-44. 52.15% are unmarried and 45% of customers in this cluster are married. Most customers (54.95%) are employed while there is a percentage of 19.28% unemployed customers.

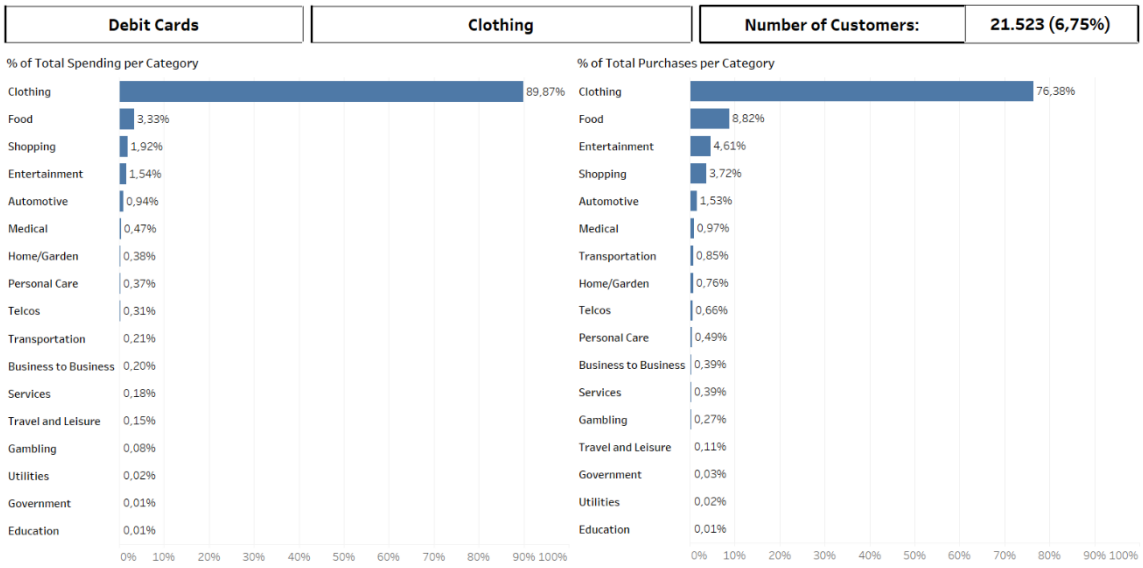

Figure 40 Percentages of total spending and purchases per category for cluster 7 / Clothing

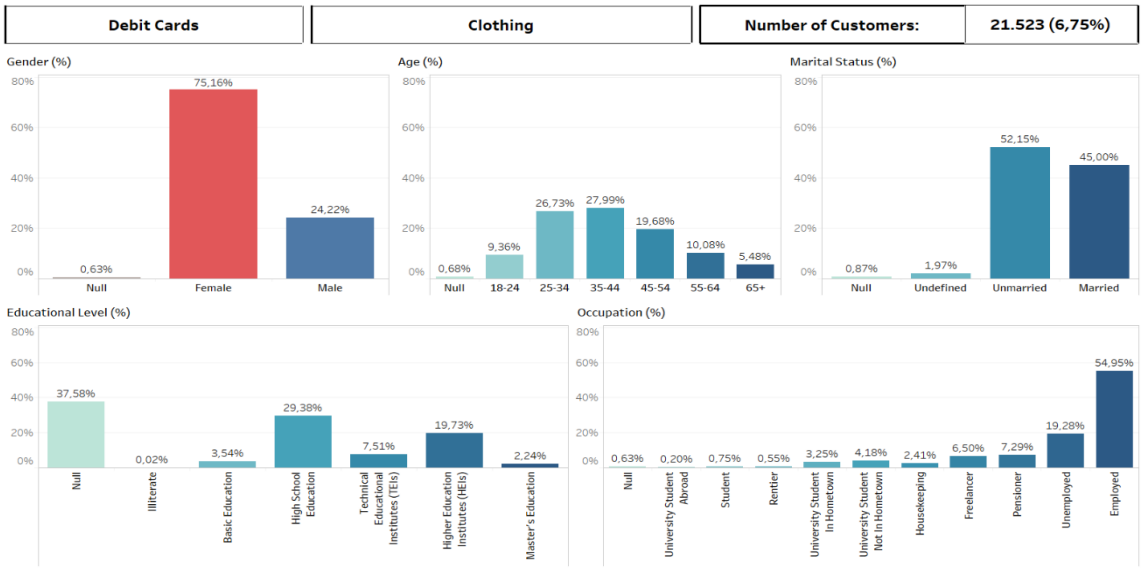

Figure 41 Percentages of the values of categorial attributes for cluster 7 / Clothing

Cluster 8 / Telcos: 3.14% of customers belong in this cluster and are characterized by increased spending in telecommunication fees. They also make few purchases and spend some money in the Food and Entertainment categories. They are mostly unmarried (57.67%) and young, with the largest percentage (30.68%) in the age category 25-34. There are 10% more men than women in this cluster and nearly half of the customers are employed.

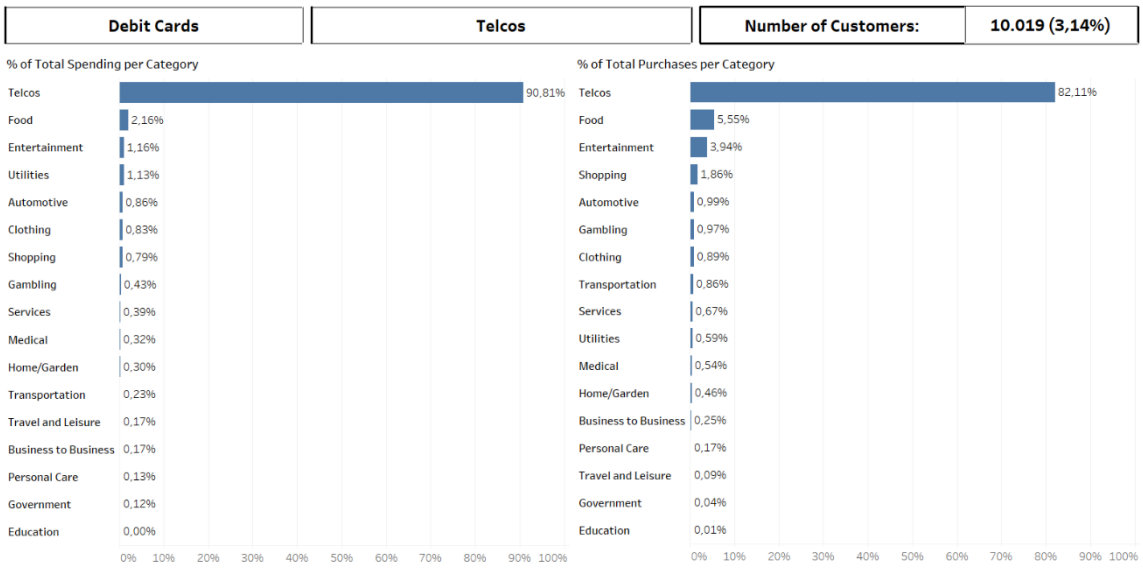

Figure 42 Percentages of total spending and purchases per category for cluster 8 / Telcos

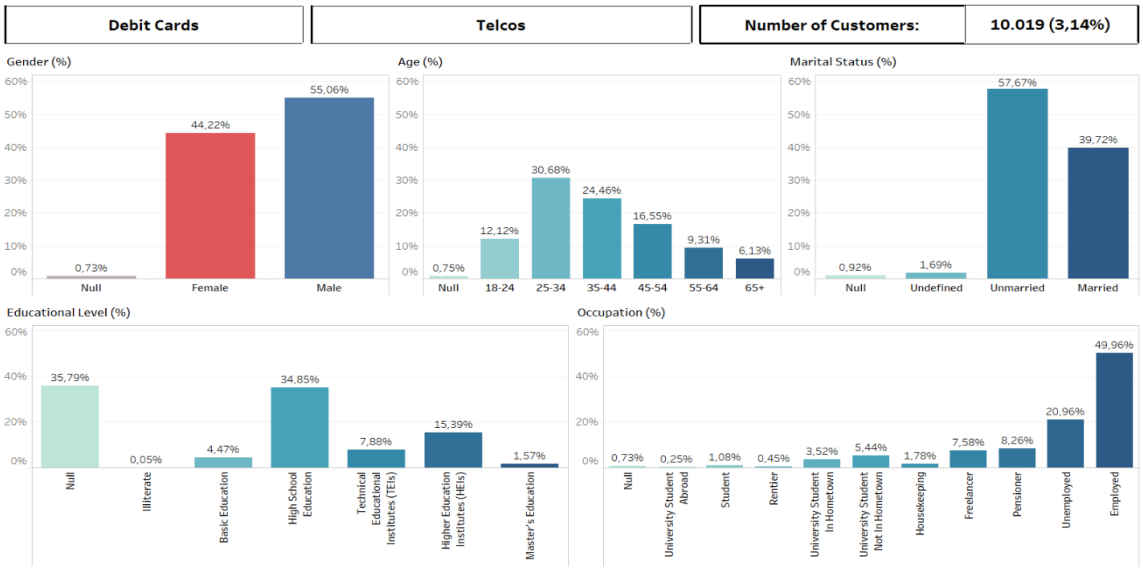

Figure 43 Percentages of the values of categorial attributes for cluster 8 / Telcos

Cluster 9 / Medical: The smallest percentage of customers (2.93%) belongs in this cluster. They spend mostly in health services and they also spend some money and make purchases in the Food and Entertainment categories. The percentages of women and men are 60.65% and 38.71% respectively. Customers of this cluster are mostly married (55.23%), employed (50.78%) and above 25 years old.

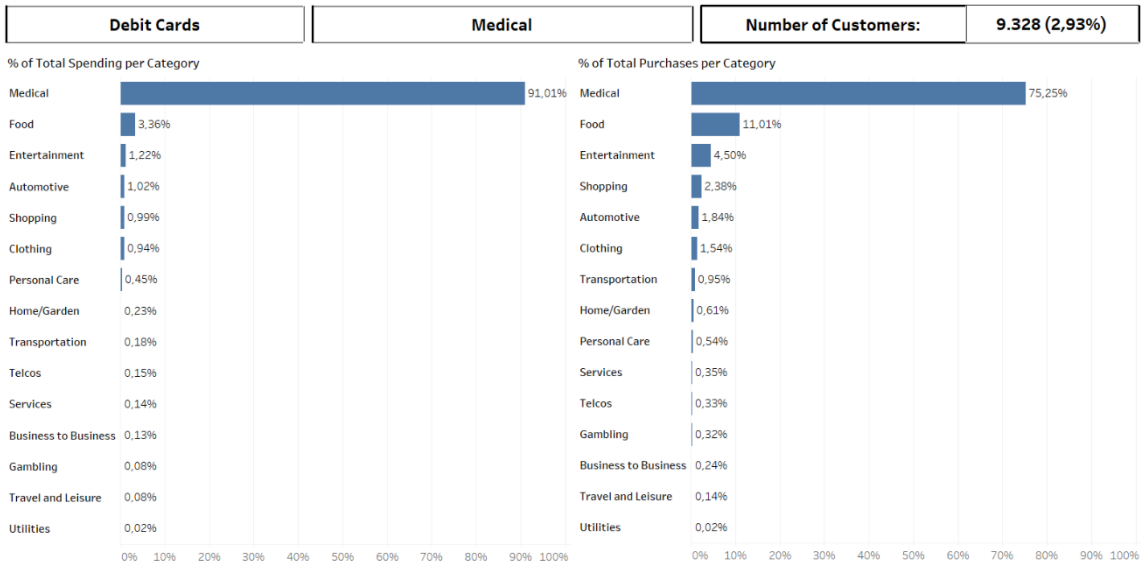

Figure 44 Percentages of total spending and purchases per category for cluster 9 / Medical

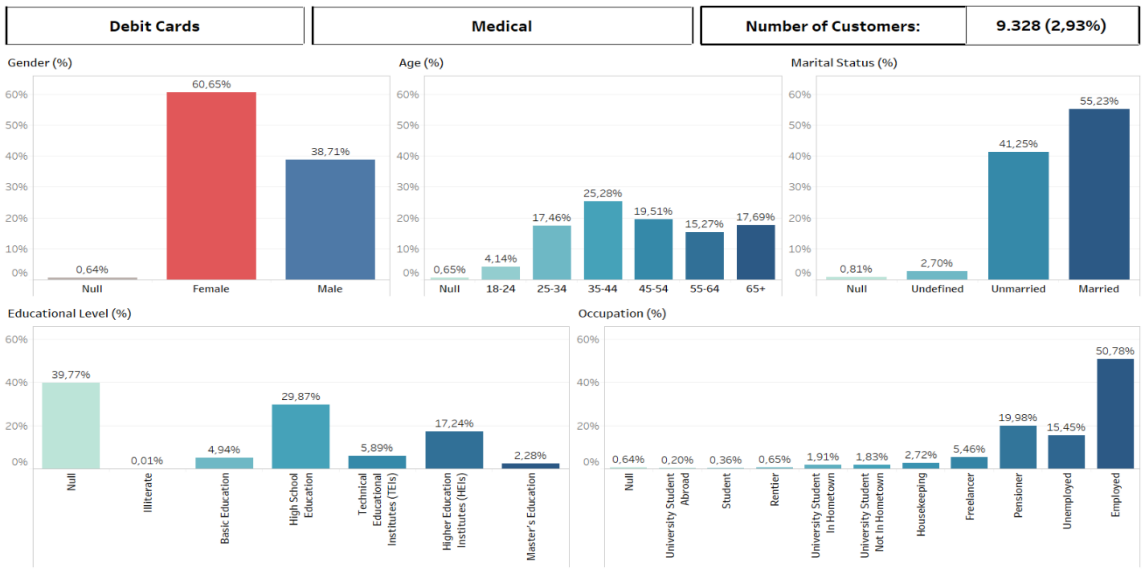

Figure 45 Percentages of the values of categorial attributes for cluster 9 / Medical

4.2.3 Profiling based on region

This section examines the characteristics of each cluster based on the region the customers live in. Tables 19 and 20 show the percentages of customers per Region for each cluster of the credit and debit datasets. In all clusters, most customers live in Attica or Central Macedonia which is expected since Athens and Thessaloniki are the largest cities of Greece. The regions that follow, i.e. Western Greece, Thessaly and Crete, have the three following cities with largest population size, i.e. Patras, Larissa and Heraklion. Although most clusters have more than 50% of their total customers living in the capital of Athens, there are a few clusters that do not follow this pattern. Namely, in the credit dataset, clusters Food and Automotive have more customers living in Western Greece and Thessaly than the other clusters. In the debit dataset, this happens in the clusters of Automotive, Gambling and Telcos.

% of Customers per Region

	Telcos	Food	All Categories	Services	Automotive	Entertainment	Travel	Shopping	Clothing
Attica	52,48%	42,61%	51,94%	59,47%	46,02%	54,98%	56,50%	51,57%	49,38%
Central Macedonia	13,21%	19,25%	14,26%	12,44%	16,22%	13,20%	15,40%	12,98%	12,77%
Western Greece	4,50%	8,42%	6,19%	3,85%	6,89%	5,70%	3,41%	6,32%	5,94%
Thessaly	4,95%	6,14%	4,84%	3,67%	6,13%	4,91%	3,52%	5,15%	5,21%
Central Greece	3,78%	4,06%	3,70%	2,85%	4,80%	3,35%	2,47%	3,75%	4,44%
Eastern Macedonia & Thrace	3,19%	4,90%	3,86%	2,59%	3,80%	2,55%	3,35%	3,83%	3,91%
Peloponnese	3,52%	3,97%	3,47%	2,85%	3,80%	3,62%	2,23%	3,49%	3,69%
Crete	3,87%	2,64%	3,28%	3,64%	4,00%	3,51%	4,15%	3,46%	3,64%
Epirus	2,20%	2,22%	2,08%	1,98%	2,47%	1,63%	1,92%	2,13%	2,55%
South Aegean	2,87%	1,41%	1,93%	2,54%	1,36%	2,56%	2,67%	2,24%	2,77%
Western Macedonia	1,91%	1,94%	1,83%	1,33%	2,11%	1,43%	1,71%	2,03%	1,99%
North Aegean	1,92%	1,50%	1,71%	1,71%	1,64%	1,47%	1,97%	1,83%	2,16%
Ionian Islands	1,59%	0,93%	0,91%	1,08%	0,76%	1,09%	0,69%	1,21%	1,55%

Table 19 Percentage of customers per region for each cluster of the credit dataset

% of Customers per Region

	Entertainment	Food	Automotive	Shopping	All Categories	Gambling	Clothing	Telcos	Medical
Attica	51,33%	47,59%	38,47%	56,97%	51,85%	42,30%	53,19%	45,36%	62,52%
Central Macedonia	14,52%	14,85%	15,71%	15,15%	13,66%	15,18%	15,43%	15,82%	12,79%
Western Greece	4,94%	5,31%	6,62%	4,00%	4,94%	5,79%	4,57%	4,69%	3,12%
Thessaly	5,01%	4,93%	6,29%	4,27%	5,02%	6,54%	5,06%	6,31%	3,10%
Crete	4,47%	4,75%	5,25%	3,86%	4,29%	4,12%	4,34%	4,16%	3,63%
Central Greece	3,53%	4,06%	5,98%	2,70%	3,44%	5,11%	2,85%	4,46%	3,06%
Peloponnese	3,42%	3,89%	4,98%	2,84%	3,52%	4,36%	3,01%	3,52%	2,62%
Eastern Macedonia & Thrace	3,19%	3,85%	4,58%	2,30%	3,53%	3,91%	3,08%	4,10%	1,98%
South Aegean	2,80%	2,95%	2,63%	2,75%	3,01%	3,20%	2,36%	3,35%	2,61%
Epirus	2,22%	2,47%	3,28%	1,58%	2,21%	3,05%	2,37%	2,02%	1,54%
Western Macedonia	1,90%	2,26%	3,10%	1,58%	1,86%	3,36%	1,70%	2,55%	1,29%
Ionian Islands	1,28%	1,67%	1,49%	1,04%	1,30%	1,62%	1,05%	1,67%	1,10%
North Aegean	1,40%	1,43%	1,65%	0,96%	1,36%	1,47%	0,97%	1,99%	0,66%

Table 20 Percentage of customers per region for each cluster of the debit dataset

Finally, for each cluster label, a map that shows the number of customers in each available area of data was drawn. Since some labels are the same for both datasets, it was decided to depict them in the same map, while choosing different colors for credit and debit cards. The total number of customers is depicted instead of percentages since as it was seen on Tables 19 and 20 the percentages are quite close for the two datasets and could not be easily displayed. In each map, which corresponds to each cluster label, the number of customers for each dataset is shown at the top, along with the percentage of cardholders in that cluster for each dataset. Credit cards are shown with blue color, while red color was used for debit cards.

The living areas of customers in the cluster 'All Categories' are nearly the same. Most customers for both datasets live in the capital of Athens. Thessaloniki, Patras, Larissa and Ioannina are the cities that follow for both datasets.

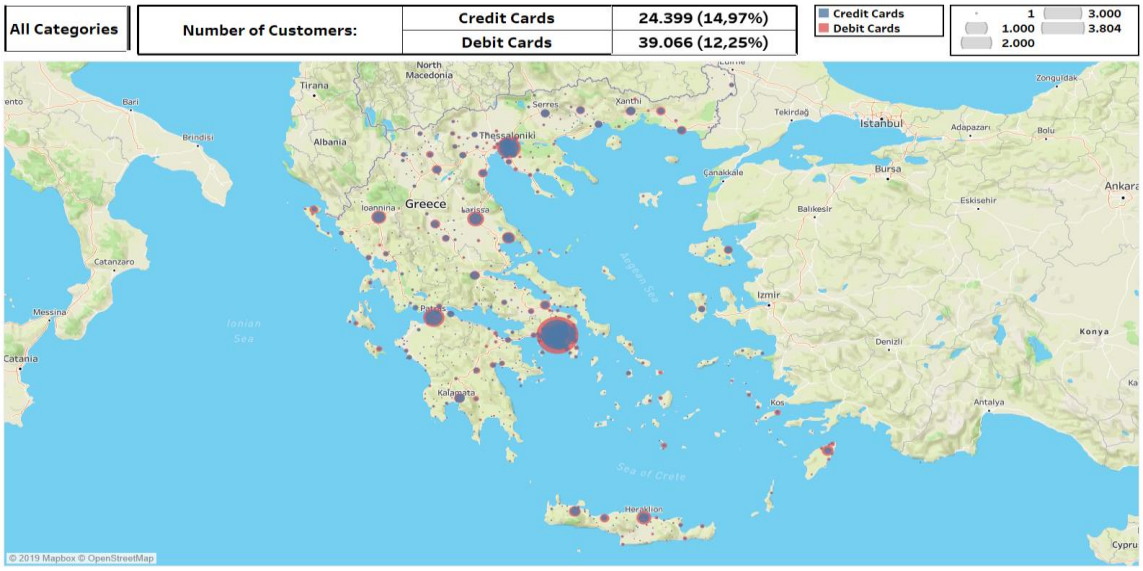

Figure 46 Map of cardholders for cluster label 'All Categories'

Customers with debit cards belonging in the 'Automotive' cluster are more than the customers with credit cards. Thus, it is reasonable that there are more customers of the debit dataset in all areas. Credit card customers belonging in this cluster mostly reside in the largest cities.

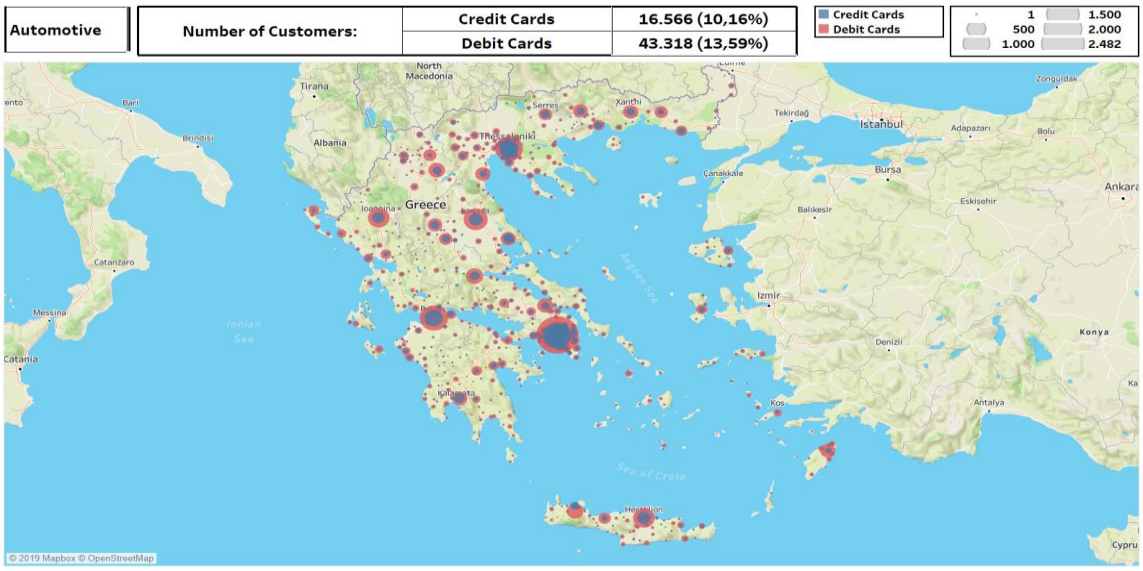

Figure 47 Map of cardholders for cluster label 'Automotive'

Customers who spend mostly in clothing stores, live in large cities like Athens, Thessaloniki, Patras, Larissa, Heraklion, Ioannina or Kalamata. This is true for both datasets, though there are more customers in the debit dataset's cluster as well.

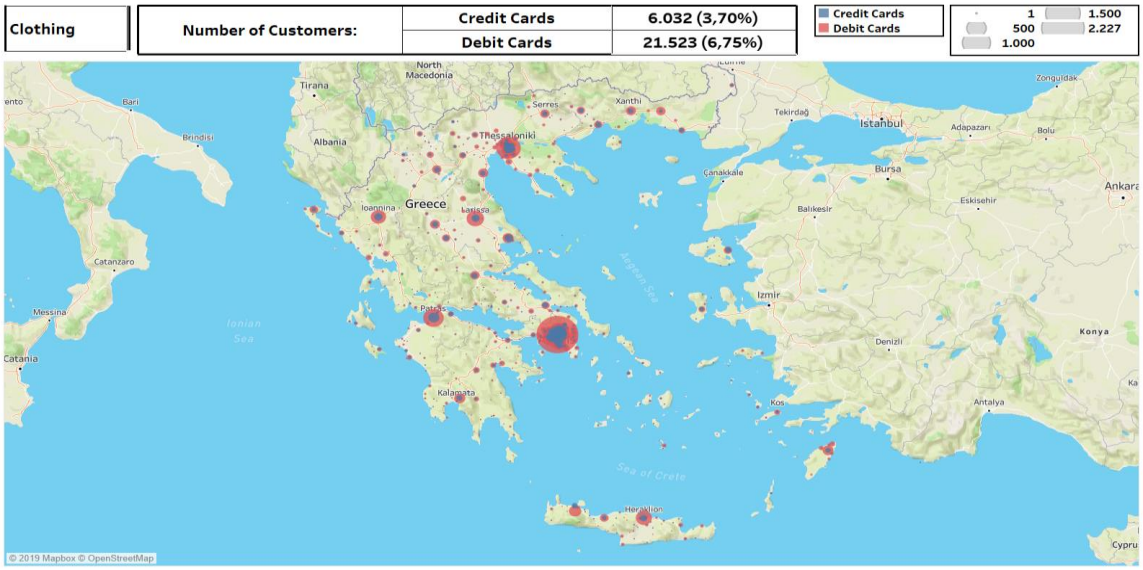

Figure 48 Map of cardholders for cluster label 'Clothing'

Customers that belong in the cluster labeled 'Entertainment' and spend more money in this category live in large cities in the main land or the islands.

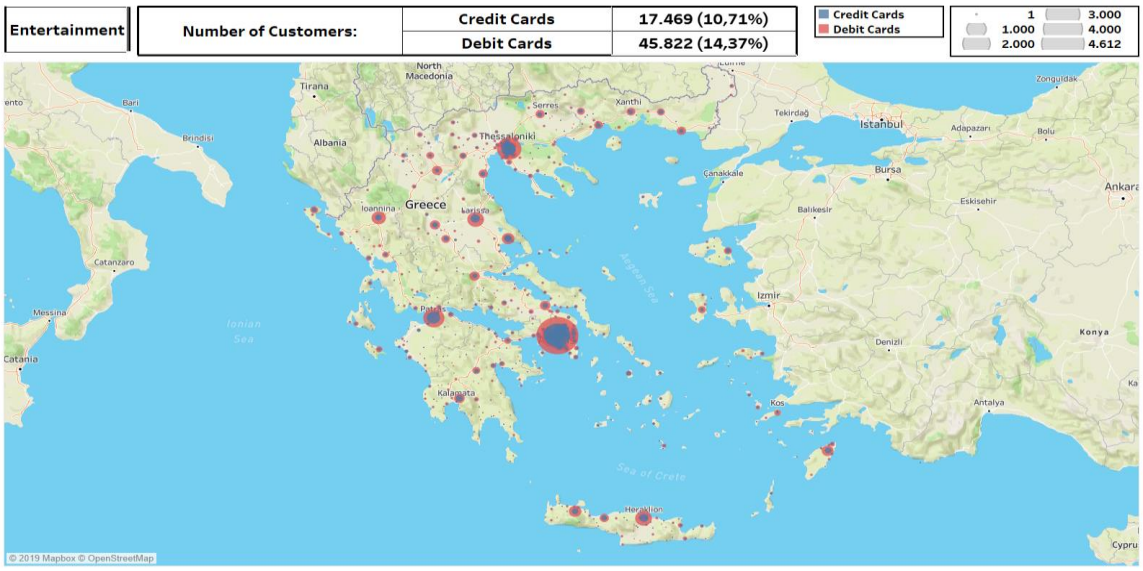

Figure 49 Map of cardholders for cluster label 'Entertainment'

The cluster related to spending in the Food category mostly has customers that live in large cities. This may also be due to the fact that people in urban areas may use their cards more often. Thessaloniki and Patras have many customers of the credit dataset living there.

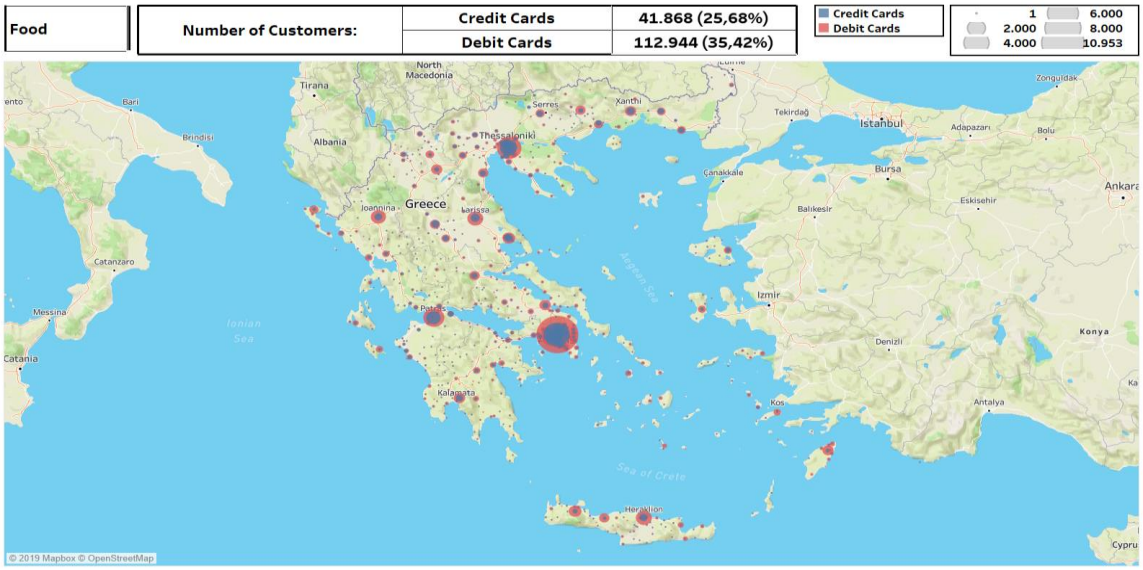

Figure 50 Map of cardholders for cluster label 'Food'

Cardholders that spend more money in the Shopping category live in the largest cities of Greece, with a very small number of customers living in villages. This is true for both

datasets with some islands, such as Lesvos, having more customers in the credit dataset than in the debit dataset belonging to this cluster.

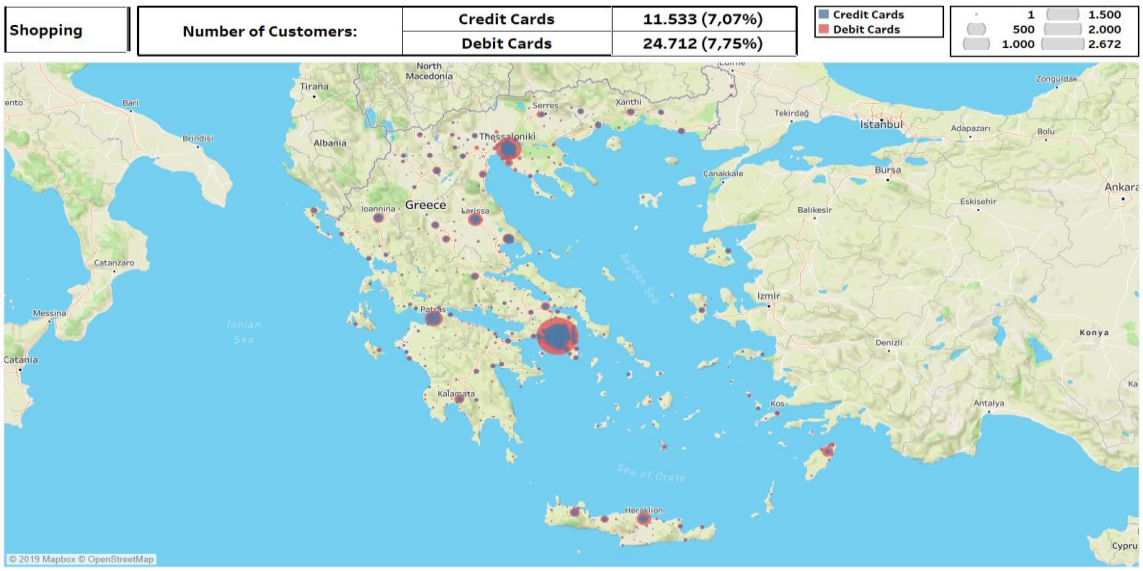

Figure 51 Map of cardholders for cluster label 'Shopping'

The cluster that is related to spending in telecommunication fees is the one that has more customers in the credit dataset than in the debit dataset. Thus, as expected, there are more customers belonging to the credit dataset in each area of Greece. The only exception seems to be the island of Rhodes, where people use their debit cards more to pay for their telecommunication fees.

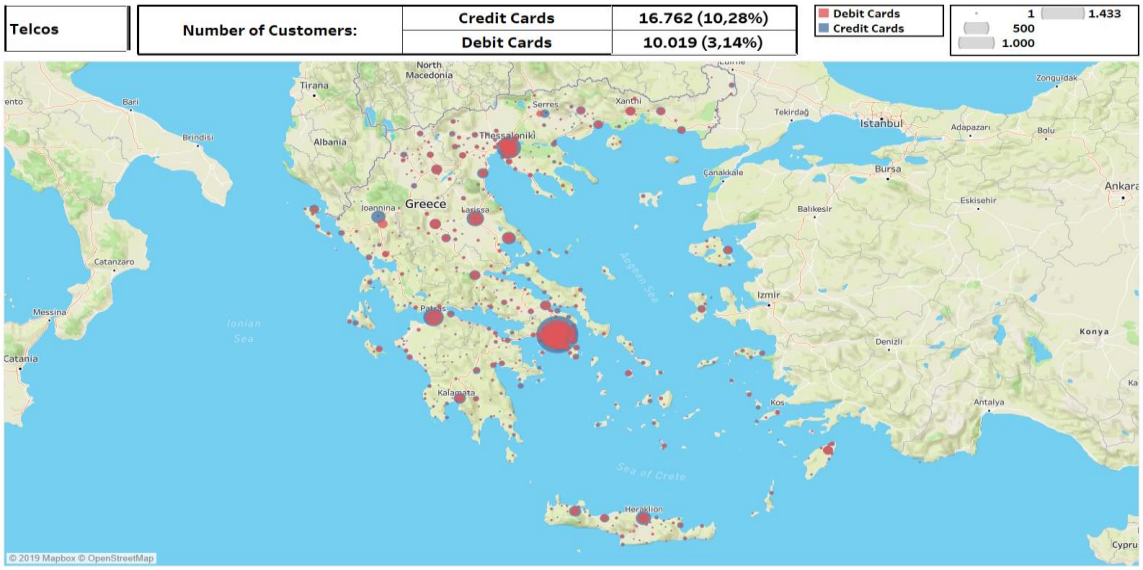

Figure 52 Map of cardholders for cluster label 'Telcos'

Most customers belonging in the credit dataset's 'Services' cluster live in cities rather than villages. Athens and Thessaloniki are still the two cities with the largest number of customers. The cities that follow are Patras, Ioannina, Volos, Larissa and Chania.

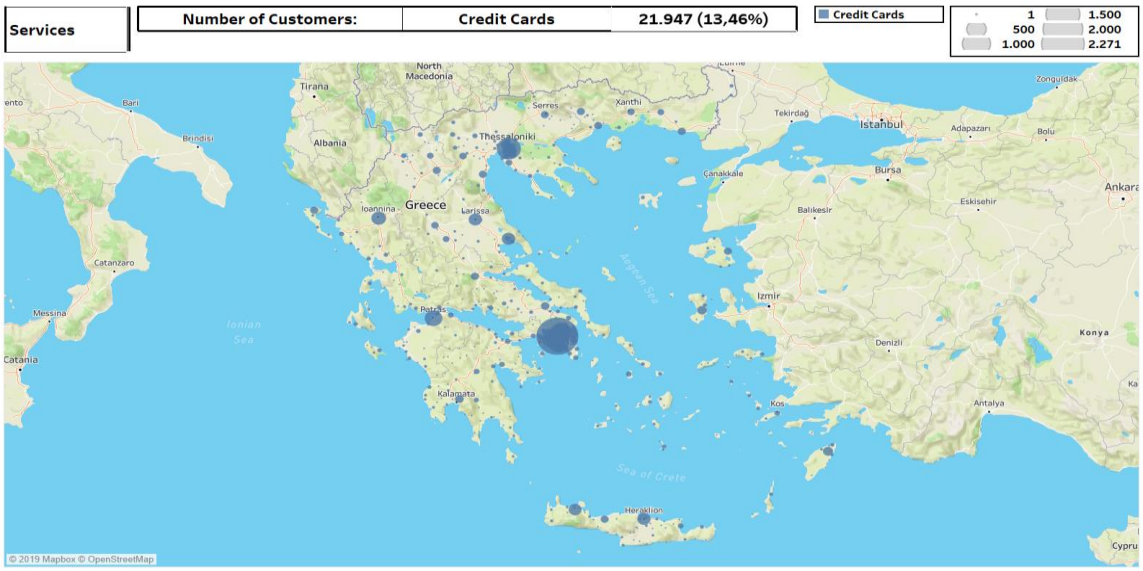

Figure 53 Map of cardholders for cluster label 'Services'

The customers from the credit dataset who spend more money in travel expenses mainly live in Greece's largest cities. Most of them live in Athens, Thessaloniki and Patras. Other cities in the main land and islands follow, while it seems that not many customers living in villages belong to this cluster.

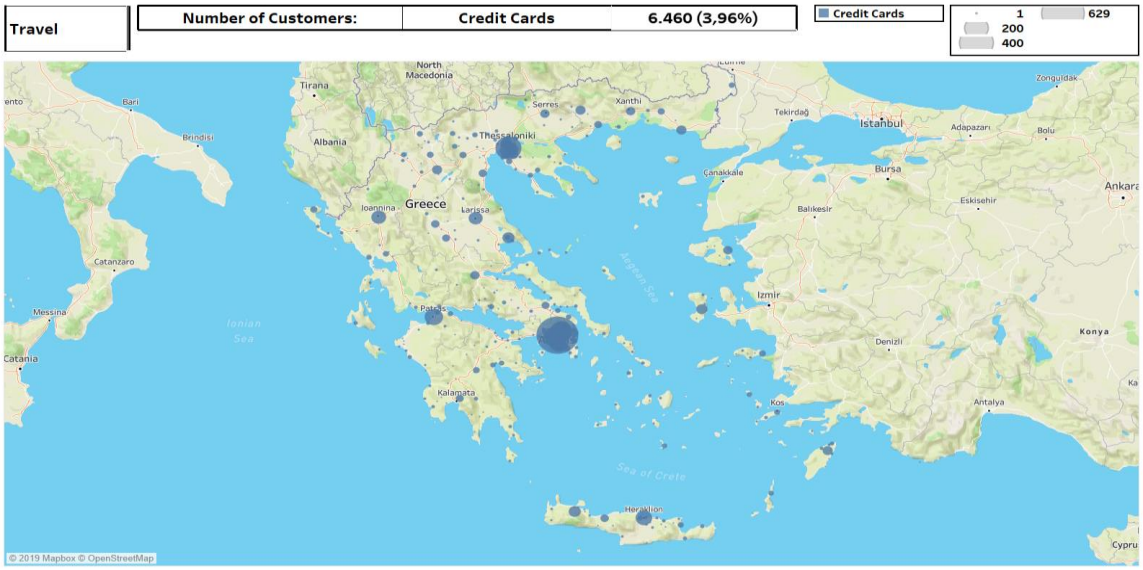

Figure 54 Map of cardholders for cluster label 'Travel'

The cluster named ‘Gambling’ only appears in the debit dataset. Despite the largest cities of Greece having many customers that belong in this cluster, smaller cities or villages also have more than 200 customers living there. This indicates that in nearly every available area there is at least one customer who belongs in this cluster.

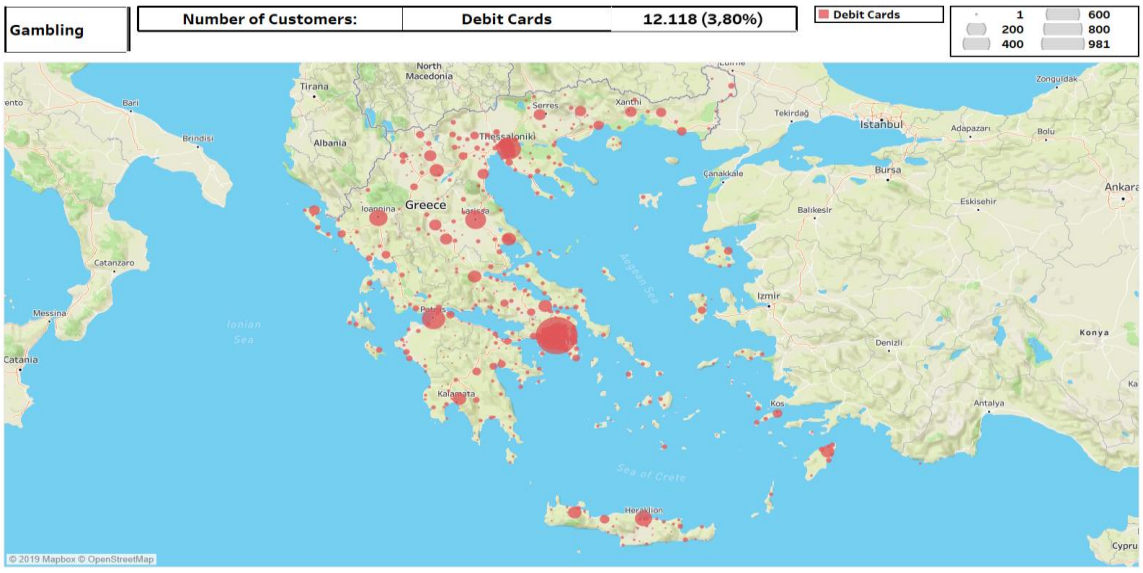

Figure 55 Map of cardholders for cluster label ‘Gambling’

The ‘Medical’ cluster has most of its customers living in big cities, mainly Athens, Thessaloniki, Patras, Heraklion, Ioannina and Larissa. The cities that follow are Chania, Volos, Lamia, Kalamata and the capitals of Rhodes and Corfu.

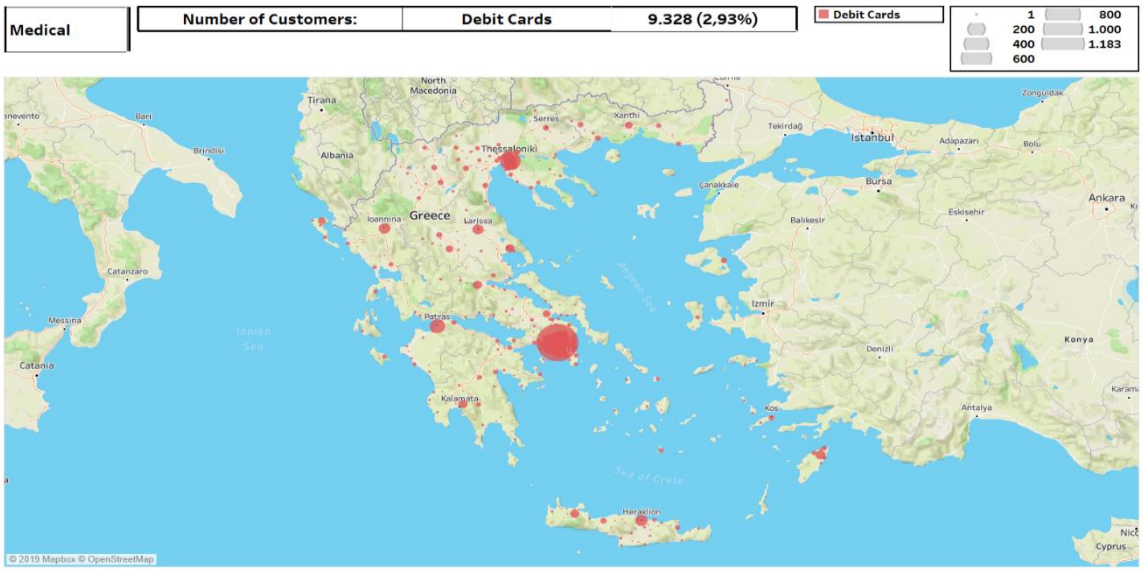

Figure 56 Map of cardholders for cluster label ‘Medical’

4.3 Customers with both credit and debit cards

There are customers who have both credit and debit cards. In the dataset provided by the bank, those customers were identified based on their account id. There are 14,493 customers who spent money with both their credit and debit cards in the time period under study. In these customers, more transactions are made with credit cards than with debit cards. Trying to implement K-Means clustering on those customers based on their total spending and disregarding the type of card they used, did not provide good results. The algorithm produced clusters without good cohesion and separation and the solution had very low silhouette coefficients despite experimenting with various values of *k*. The optimal number, although still low compared to the separate datasets occurred at 9 clusters. K-Means was implemented on this dataset with 9 clusters, but the acquired clusters had very small silhouette coefficients indicating that the algorithm could not find distinct patterns for these customers. Perhaps if more data of people with both cards were available, a separate clustering algorithm with good results could be implemented for them. Thus, it was decided to analyze the behavior of these customers based on how they spend money in the 17 original categories. Figure 57 depicts the percentage of money that these customers spend in each category depending on the card they use. Especially in the category of Food, they prefer to use their debit cards while when paying for Services, they mostly use their credit cards. In the other categories they follow the behavior of the total population although they seem to spend more money in Travel and Leisure and Education with their credit cards and less money in Utilities, opposed to the total population.

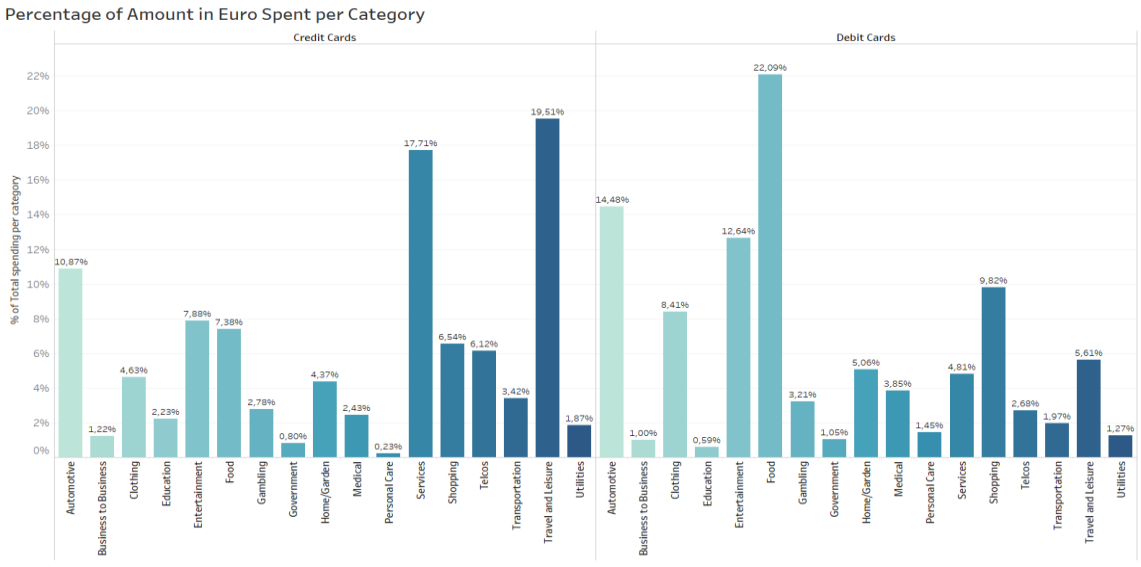

Figure 57 Percentage of total spending per category for customers with both cards

4.4 Debit customers with credit cards

In this section, transactions of customers with both cards are added on the debit dataset to see if the clustering solution changes. The same procedure as with separate datasets is followed. Firstly, the percentages of total spending per category are calculated for these customers and they are used as input in a PCA model. The percentage of cumulative explained variance is at 88.18% with 8 components. The PCA loadings are then evaluated as in the previous section to see which original attributes are correlated with each component. Table 21 shows the components matrix, where attributes with correlations below 0.4 in absolute value are omitted for clearer representation.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Food	0.9188							
Entertainment		0.7426	-0.4802					
Automotive		-0.6651	-0.6011					
Shopping			0.4956	-0.710				
Clothing				0.673	-0.5929			
Gambling					0.4144	0.7582		
Telcos						-0.6333	-0.5513	
Services							0.7321	-0.5146
Medical								0.6824

Table 21 Components matrix for debit customers with credit cards

The important attributes which are correlated with the components are the same as with the debit dataset. The next step is to evaluate the clustering solution and find the optimal number of clusters. Using the same evaluation metrics as before, it can be seen from Figure 58 that the optimal number of clusters is 9.

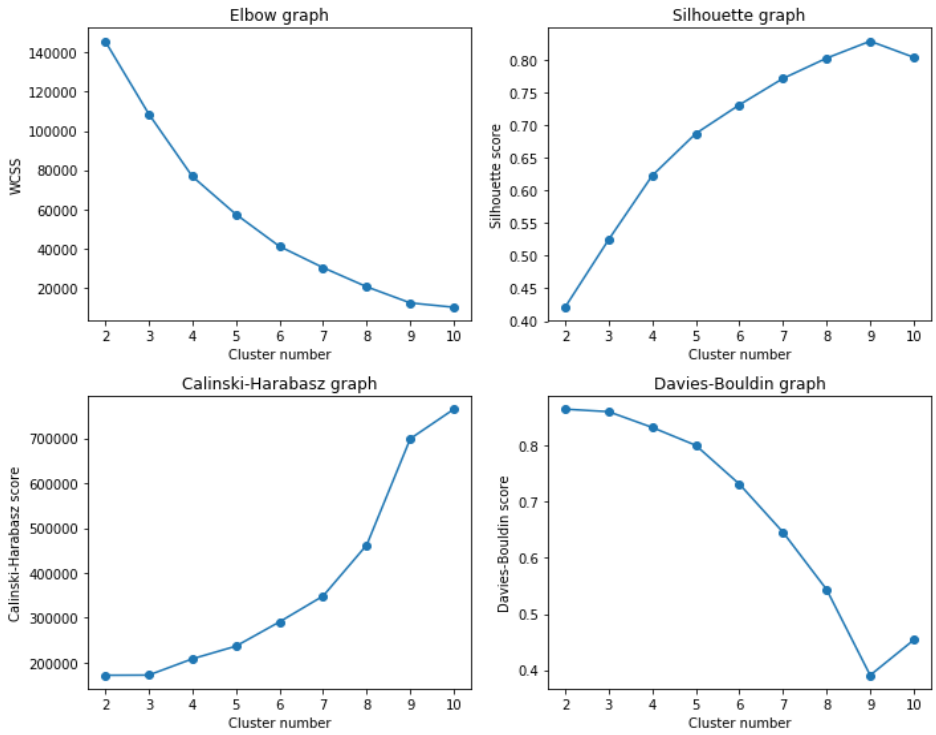

Figure 58 Clustering evaluation graphs for debit customers with credit cards for values of k from 2 to 10

K-Means was then implemented with 9 clusters to see if these clusters are different from the clusters of the debit dataset. In Table 22, the silhouette coefficients for each cluster are provided. Although cluster 2 has the lowest coefficient, it still very high, above 0.6. All other clusters have very high silhouette coefficients, indicating a good clustering solution and clusters with good cohesion and separation.

Cluster Number	Silhouette coefficient
1	0.8989
2	0.6001
3	0.8334
4	0.8534
5	0.8106
6	0.9214
7	0.8513
8	0.7989
9	0.8397

Table 22 Silhouette coefficient for each cluster

The next step is to examine the cluster centers based on the PCA scores. Table 23 shows the mean values of the component scores for each cluster. Values above 0.5 in absolute

value indicate scores above the average and increased spending, whereas values below 0.5 denote below average scores and lower spending.

	1	2	3	4	5	6	7	8	9
Food	0.65	-0.23	-0.42	-0.48	-0.28	-0.26	-0.26	-0.30	-0.25
Entertainment, negative Automotive	0.02	-0.03	-0.65	0.66	-0.06	-0.05	-0.04	-0.06	-0.04
Shopping, negative Automotive, Entertainment	-0.05	0.17	-0.42	-0.31	0.41	0.24	0.23	0.59	0.21
Clothing, negative Shopping	-0.01	0.07	-0.06	-0.05	0.67	0.12	0.11	-0.60	0.09
Gambling, negative Clothing	-0.07	-0.02	-0.07	-0.23	0.22	0.65	-0.39	0.35	0.28
Gambling, negative Telcos	0.00	-0.04	0.01	0.01	0.02	0.72	-0.61	0.02	-0.13
Services, negative Telcos	-0.00	0.11	-0.01	-0.01	-0.04	-0.27	-0.48	-0.03	0.72
Medical, negative Services	-0.01	0.29	-0.01	-0.01	-0.04	-0.15	-0.18	-0.03	-0.44

Table 23 Cluster centers

It seems that cluster 9, which had a larger score in the component related to Medical in the debit dataset, is now more correlated with the category of Services. To evaluate this, in Figure 59, the percentages of total spending per category for each cluster are depicted.

Percentages of total spending per category

	Clusters								
	1	2	3	4	5	6	7	8	9
Automotive	1,63%	1,68%	91,34%	1,53%	0,99%	0,51%	0,95%	1,19%	1,17%
Business to Business	0,17%	4,13%	0,19%	0,12%	0,21%	0,04%	0,17%	0,17%	0,17%
Clothing	0,70%	1,10%	0,43%	0,76%	89,54%	0,09%	0,81%	1,62%	0,49%
Education	0,00%	2,54%	0,00%	0,00%	0,01%		0,00%	0,01%	0,00%
Entertainment	1,43%	1,68%	1,35%	90,45%	1,55%	0,62%	1,48%	1,62%	0,98%
Food	93,29%	3,06%	3,73%	3,56%	3,23%	0,92%	2,02%	3,39%	1,45%
Gambling	0,14%	0,16%	0,17%	0,17%	0,08%	96,81%	0,40%	0,20%	0,11%
Government	0,00%	3,73%	0,02%	0,01%	0,02%	0,00%	0,10%	0,02%	0,03%
Home/Garden	0,32%	21,55%	0,34%	0,28%	0,43%	0,14%	0,33%	0,45%	0,24%
Medical	0,38%	14,71%	0,23%	0,25%	0,38%	0,04%	0,24%	0,36%	0,15%
Personal Care	0,19%	4,67%	0,10%	0,21%	0,35%	0,03%	0,10%	0,25%	0,03%
Services	0,23%	0,71%	0,33%	0,45%	0,30%	0,18%	0,68%	0,62%	92,97%
Shopping	1,01%	1,43%	0,75%	1,11%	2,02%	0,23%	0,85%	89,27%	0,69%
Telcos	0,25%	0,64%	0,29%	0,42%	0,43%	0,19%	90,00%	0,38%	0,42%
Transportation	0,17%	6,56%	0,43%	0,41%	0,24%	0,05%	0,40%	0,24%	0,29%
Travel and Leisure	0,07%	26,42%	0,27%	0,24%	0,18%	0,12%	0,35%	0,17%	0,64%
Utilities	0,02%	5,23%	0,04%	0,03%	0,03%	0,02%	1,14%	0,03%	0,18%

Figure 59 Percentages of total spending per category

Cluster 1 is associated with spending in the Food category. Cluster's 2 customers spend in all categories and do not show a distinct pattern. Customers of clusters 3 and 4 tend to spend more money in the Automotive and Entertainment categories respectively. Cluster

5 is related with spending in clothing stores. Customers of cluster 6 spend more money in the Gambling category and minimal amounts in all other categories. Cluster 7 is associated with telecommunication expenses while cluster 8 has customers that spend mostly in the Shopping category. Cluster 9 has the customers who spend more money in the category of Services, as opposed to other categories. After the addition of the transactions with credit cards for the debit customers, spending more money in Services became a distinct cluster, while spending in the Medical category became a part of the 'All Categories' cluster.

By assigning labels for the clusters based on the category that their customers spend more money on, this dataset provides the same labels as the debit dataset with the only difference being that instead of Medical, the last cluster should be labeled Services.

Table 24 shows the percentages of customers belonging in each same-labeled cluster for the debit dataset and the debit customers with credit cards. Also, the percentages of customers belonging in cluster 9 of both datasets is shown. The differences in the percentages are below $\pm 1\%$ for all clusters. The cluster which corresponds to spending in the Services category is a little bigger than the cluster of the debit dataset corresponding to medical expenses.

	Percentage of customers	
Spending habits	**Debit dataset**	**Debit customers with credit cards**
Entertainment	14.37%	14.43%
Food	35.42%	34.44%
Automotive	13.59%	13.35%
Shopping	7.75%	7.77%
All Categories	12.25%	12.69%
Gambling	3.80%	3.7%
Clothing	6.75%	6.62%
Telcos	3.14%	3.68%
Cluster 9	2.93% (Medical)	3.31% (Services)

Table 24 Percentage of customers for the debit dataset and the debit customers with credit cards

Chapter 5 Conclusion

Spending behavior is a customer characteristic that helps businesses segment their customers into distinctive groups. These segments can help target specific customers with offers that may be of interest to them. In the datasets from the bank, the segmentation process helped to highlight the most important categories that customers tend to spend their money on. These categories seem to be mostly the same for both credit card and debit card customers. Most customers with credit and debit cards seem to use their cards for specific purposes. There is only one cluster of customers in the credit and one in the debit dataset that uses their cards to cover various types of needs. These are the customers that belong in the clusters labeled 'All Categories'. There are 24,399 customers with credit cards and 39,066 customers with debit cards who use their cards for various purposes. Customers belonging to other clusters should be targeted according to their spending habits. For example, a card that partners with airlines, travel agencies, hotels and car rentals companies could be targeted to the customers with credit cards who spend mostly in the category of Travel and Leisure. For all customers, offers and possibly cards could be made according to their needs and habits. This can be executed by targeting the customers of each cluster with offers or cards that cover their differentiated needs.

This segmentation process could also be improved with the usage of more data. More transactions could be used and for larger periods of time to acquire more general clusters describing customers' spending behavior. Having more transactions for each customer could result in a more general understanding of his/her spending habits. Information about each customer's income or balance could also help in the segmentation process and provide valuable insight. Using more transactions for a larger period could also provide the necessary data to implement an RFM analysis to segment customers based on how often they make transactions and how much money they spend. This could also provide information about the value each customer has for the bank and the acquired groups of customers could be targeted based on the customers' value.

Bibliography

[1] M. Namvar, M. Gholamian, S. KhakAbi, "A Two Phase Clustering Method for Intelligent Customer Segmentation," in *Intelligent Systems, Modelling and Simulation (ISMS), 2010 International Conference*, Liverpool, 2010.

[2] A. S. Hossain, "Customer segmentation using centroid based and density based clustering algorithms," in *2017 3rd International Conference on Electrical Information and Communication Technology*, 2017.

[3] S. Tripathi, A. Bhardwaj, E. Poovammal, "Approaches to clustering in customer segmentation," *International Journal of Engineering & Technology*, vol. 7, no. 3.12, pp. 802-807, 2018.

[4] V. Aggelis, D. Christodoulakis, "Customer clustering using rfm analysis," in *Proceedings of the 9th WSEAS International Conference on Computers*, 2005.

[5] R. Di Clemente, M. Luengo-Oroz, M. Travizano, S. Xu, B. Vaitla, M.C. González, "Sequences of purchases in credit card data reveal lifestyles in urban populations," *Nature communications,* p. 9, 2018.

[6] S. Sobolevsky, I. Sitko, R. T. Des Combes, B. Hawelka, J. M. Arias, C. Ratti, "Cities through the prism of people's spending behavior," *Plos One*, 2016.

[7] S. Zhou, A. Montgomery, G. Gordon, "Exploring customer spending behavior and payday effect using prepaid cards transaction data," *Machine Learning Department, Carnegie Mellon University*, pp. 1-17, 2016.

[8] K. K. Tsiptsis, A. Chorianopoulos, Data mining techniques in CRM: inside customer segmentation, John Wiley & Sons., 2011.

[9] Y. Kotidis, "Clustering," *Data Mining Lecture, Athens University of Economics and Business*, 2019.

[10] Pedregosa et. al., "Scikit-learn: Machine Learning in Python," *JMLR*, vol. 12, pp. 2825-2830, 2011.

[11] P. N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Education, Inc, 2006.

[12] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, et al., "Apache spark: a unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56-65, 2016.

[13] G. Van Rossum, F. L. Drake Jr, Python tutorial, Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.