UNDERSTANDING SHOPPING BEHAVIOR OF CUSTOMERS USING TRANSACTIONAL DATA

by

MINE TUNA

Submitted to the Graduate School of Management

in partial fulfillment of

the requirements for the degree of

Master of Science

Sabancı University

July 2018

UNDERSTANDING SHOPPING BEHAVIOR OF CUSTOMERS USIING TRANSACTIONAL DATA

APPROVED BY:

Prof. Dr. Burçin Bozkaya ....
(Thesis Supervisor)

Prof. Dr. Cenk Koçaş ....

Assoc. Prof. Dr. Naime Meltem Çakıcı ....

DATE OF APPROVAL: 30/07/2018

İŞLEMSEL VERİ KULLANILARAK İNSANLARIN ALIŞVERİŞ DAVRANIŞLARININ ANLAMLANDIRILMASI

Mine Tuna

İş Analitiği, Yüksek Lisans Tezi, 2018

Tez Danışmanı: Prof. Dr. Burçin Bozkaya

Anahtar Kelimeler: Alışveriş Merkezleri, Davranış Analizi, Kümeleme, Büyük Veri, Alışveriş Deneyimi

Özet

Dijital teknolojiler çok miktarda veri üreterek insan davranışlarını izlememize olanak sağlarlar. Bu çalışmada özel bir Türk bankasının 60 bin müşteri ve 2 milyon kredi kartı işlemi içeren verisi, bireylerin alışveriş davranışlarını incelemek için kullanılmıştır. Online alışverişin arttığı bir çağda olmamıza rağmen, insanlar yine de alışveriş yapma duygusunu yaşamak için alışveriş merkezlerini ve ana caddelerde yer alan mağazaları tercih etmektedirler. İnsanlar genellikle alışveriş yapacakları yere karar verirken alışveriş yapacakları yerin mağaza çeşitliliğini, ulaşılabilirliğini, konforunu ve sosyal yönlerini dikkate alırlar. Bu çalışmada insanların çeşitlilik arama davranışları alışveriş merkezleri ve alışveriş kategorileri bağlamında incelenmiştir. Bireylerin alışveriş davranışlarını ayırt etmek için kredi kartı harcamalarından elde edilen davranışsal özelliklerin yer aldığı, K-medyan kümeleme algoritması kullanılmıştır. Ayrıca bireyleri oluşturulan kümelerden birine, kümelerle olan demografik benzerliğini ölçerek atayan bir metot önerilmiştir. Sonuçlarımıza göre demografik özellikler ile alışveriş davranışları arasında bir bağlantı olduğu saptanmıştır. Bulgular ayrıca kadınların alışveriş yaparken çeşitli alışveriş merkezleri ve kategorileri aramaya meyilli olduğunu, dolayısı ile alışverişi eğlenceli ve sosyal aktivite olarak algıladığını göstermektedir. Diğer taraftan erkeklerin ise, ihtiyaca dayalı harcamalar için belirli alışveriş merkezlerini tercih ettiğini, böylece alışveriş için zaman ve enerji harcamayi tercih etmediklerini göstermektedir. Yaptığımız çalışmanın, pazarlamacılara doğru müşteri gruplarına doğru stratejiler ile iletişime geçmelerine yol göstereceğini ummaktayız.

UNDERSTANDING SHOPPING BEHAVIOR OF CUSTOMERS USING TRANSACTIONAL DATA

Mine Tuna
Business Analytics, Master's Thesis, 2018
Thesis Supervisor: Prof. Dr. Burçin Bozkaya

Keywords: Shopping Malls, Behavioral Analytics, Clustering, Big Data, Shopping Experience

Abstract

Digital technologies allow us to trace human behaviors by generating large amounts of data. In this study a private Turkish bank data containing 60 thousand customers and 2 million credit card transactions are used to analyze the shopping behaviors of individuals. Even though we are in an age of growing online shopping, people still prefer to visit shopping malls, or the stores placed in high streets to experience shopping. They usually make their shopping place decisions according to store variety, accessibility, comfort, and social aspects. In this study, we investigate people's variety seeking behavior in the context of shopping malls and shopping categories to assess their shopping experience. We use K-means clustering algorithm to distinguish between customers' shopping behaviors by using the behavioral features we extract from their credit card spending. In addition, we propose a method to assign individuals to one of the segments by measuring the demographic property similarity with segments. Our results indicate that there is an association between demographic properties and shopping behavior. The findings also suggest that females are more likely to search for variety of shopping malls and categories, and hence perceive shopping as an entertaining and social activity, whereas men prefer to shop in particular shopping malls for need-driven purchases indicating that they do not wish to lose time and energy for shopping. We hope that our research will guide the marketers to communicate the right group of customers with the right strategy.

To my beloved family

ACKNOWLEDGEMENTS

I would like to thank my advisor Prof. Dr. Burçin Bozkaya for his invaluable guidance in all phases of my thesis. Also, I would like to thank Prof. Dr. Cenk Koçaş for his valuable suggestions and mentoring in my study. I would like to thank all the professors that I had a chance to meet and work with at Sabancı University for all their contributions in me.

I am grateful to my parents, Müge Tuna and Mehmet Eser Tuna for their endless support, patience, and guidance throughout my undergraduate and graduate studies. They always believe in me and support me all the time; I am lucky to have them. Also, I would like to thank all my family members for their peerless support during this process. My special thanks to my precious grandfather who brought me up and had big contributions in me. I hope you are watching and getting proud of me. I dedicate this thesis to my dear family.

Also, I would like to thank my dearest friends Tuğçe and Hilal for their support and friendship. One of the greatest contributions of Sabancı University has been to meet with my lovely friend Tuğçe. I am thankful for your support and help. I have shared the journey starting from high school to master's degree at the same schools with my dearest friend Hilal. Thank you so much for your friendship and support, doing master's would not be enjoyable without you.

Finally, I would like to thank all business analytics graduate students for their friendship and kindly helps.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The digital technologies allow tracing human daily activities like the places they shop, the things they eat, the people they call and the products they buy, and by doing so, make it possible to collect large amounts of data. These collected data provide the most important input for analyzing human behavior patterns. One of the data producing technologies is mobile payment systems, which provide banks to collect large amounts of data of its credit card users. Mobile payment systems enable to identify spatial-temporal patterns of shopping activities (Yoshimura et al., 2016). In order to develop good marketing strategies, understanding the shopping behavior of customers is very important. For instance, it can help develop personalized campaigns or make it easier for identifying the potential customers (Yen et al., 2018).

Although online shopping is growing, consumers still prefer to make their purchases in brick-and-mortar stores by either going to shopping malls or stores located in high streets. Customers usually want to get shopping experience, which online shopping cannot provide and so they are faced with the decision of selecting shopping areas considering store choices and location. People usually make their shopping place decisions according to store variety, accessibility, comfort, atmosphere and social aspects. According to Huff (1964), people are more likely to shop areas close to their home or workplace. However, with the conveniences in transportation, they may choose attractive locations which include more variety of store options. However, we cannot expect all individuals to behave in the same way. Therefore, in this thesis, we aim to identify the shopping behavior differences of consumers in the context of shopping malls by looking at the behavioral features extracted from Big Data.

In this study, data from a private Turkish bank containing 60 thousand customers and 2 million credit card transactions located in Turkey are used. In addition, we collect data containing shopping mall coordinates in Istanbul. Each transaction is assigned to one of the shopping malls if the determined distance criterion is satisfied and we continue our analysis with the reduced data containing only shopping mall transactions. In order to detect customers' variety seeking behavior in the context of shopping malls and shopping categories, we use two behavioral features: diversity and loyalty. Diversity refers to the notion that customers' shopping behavior can vary over various shopping malls or shopping categories and loyalty measures how much a customer is loyal to their particular shopping mall(s) or shopping categor(ies). We approach our problem by considering these two different types of behavioral features.

In the first step of our study, we segment customers into four groups by K-means clustering algorithm using diversity and loyalty as clustering dimensions. Our aim is to differentiate the customers according to their variety seeking intentions with shopping malls and shopping categories. Then, we provide insights on demographic profiles, transactional and shopping category characteristics of the constituted segments to find out the distinguishing differences among segments. In the end, we associate our findings with the shopping experience discussed in the literature.

In the second step, we split our data into training and test sets, and use the trained model to assign customers in the test set to the segment they belong to by considering their demographic information. Then, we check the coherence of actual and assigned segments to understand how efficient the demographic information is for distinguishing the shopping behavior of consumers. In cases where marketers know only the demographic information of individuals, we expect, based on our results, that marketers can better communicate with them by offering the right shopping places and product types or take better actions on potential customers.

This thesis is organized as follows. In Chapter 2, we review the literature on behavioral analysis and feature extraction, shopping place choice and patronage, and gender differences in shopping experience. In Chapter 3, we explain the data used in the study along with data preprocessing steps and the features extracted to understand the shopping behavior of customers. In Chapter 4, we give brief information about the methodologies we used. In Chapter 5, we present the results that are obtained in the computational analysis and discuss the inferences from these computational analyses. Finally, in Chapter 6, we provide our concluding remarks along with a summary of the thesis outcome achieved and its contributions.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we present the literature review under three topics: behavioral analysis and feature extraction, the consumers' shopping place choice and patronage, and gender differences in shopping experience. We conclude the chapter with the discussion of our contributions to the literature.

2.1 Behavioral Analysis and Feature Extraction

Storing large amounts of data about customers enables extracting information about their behavioral properties. For instance, datasets containing coordinate information of transactions allow in many studies the extraction of customer mobility behavior in relation to financial wellbeing (Singh et al., 2015; Srivastava et al., 2014). Singh et al. (2013) try to predict the spending behavior of people using spatio-temporal behavior measurements. Krumme et al. (2013) study the prediction of store visitation patterns. Clemente et al. (2017) and Guidotti et al. (2018) investigate customers' regularities of temporal purchasing behavior. Features like diversity, loyalty and regularity are constituted in many studies to identify the customers' behavioral characteristics.

In order to extract the spatial and temporal behaviors of the customers in a transactional data set, Singh et al. (2015) propose diversity, loyalty and regularity features. Diversity measures how the customers' shopping experience vary in time and location. Loyalty is the percentage of purchases occurred in the three most frequently visited locations or shopped day of the week. Regularity calculates the similarity in customers' shopping behavior over shorter and longer periods. These three features enable to predict the financial difficulties, which are defined as overspending, late payment and financial trouble of customers with an improvement from 30% to 49% compared to models that contain only demographic variables.

On the other hand, Srivastava et al. (2014) extract behavioral features of customers from a bank data set to analyze the financial well-being of merchants. They measure total revenue and consistency in revenue to provide information about merchants' credit riskiness to the banks. They measure the behavioral features: diversity and propensity according to the time of the day, age groups, distance, day of the week, educational status, gender, transaction amounts and transactions by loyal customers for each merchant. Their results show that customers who belong to a specific age group and visit merchants in specific times of the day enable merchants to generate high revenues. On the other hand, diversity in age groups and visitation times of the customers provide stability in merchants' revenue. In addition, their research also has a positive impact on merchants to reach the potential target of customers like correct age groups, in correct days or hours for further marketing campaigns.

Singh et al. (2013) use a data set of 52 adults forming 26 couples and consisting of their self-reported spending data and social interaction patterns including phone calls, SMS logs and face-to-face interaction. They use diversity and loyalty features, and also introduce overspending that measures the ratio of actual monthly spending to the self-reported income in a survey. Naïve Bayes method is used to predict the spending behavior of couples. The study shows that social behavior measurements extracted from face-to-face interaction, call, and SMS logs have predictive power on spending behavior for couples with regard to exploring various businesses, becoming loyal customers, and overspending. The results also indicate that mobile phone based social interactions can provide more predictive power on spending behavior compared to other features.

Dong et al. (2017) argue that social bridges between communities result in similarity in their purchase behavior. The authors define social bridge as people who live in different communities but work at closer locations. Therefore, people who work in same or close workplaces possibly interact and exchange information between one another. They define three behavioral indices which are the number of unique co-visited stores by customers in different pre-defined communities, the similarity of temporal distributions of purchases made by customers from different communities and the sum of absolute differences in median spending amount in the merchant categories. Their results show that social bridges indicate a much stronger similarity in purchasing behaviors compared to sociodemographic and income. They also find out that females are affected more by social bridges compared to males.

Clemente et al. (2017) group the consumers into five different segments according to their similarity in purchasing sequences to identify their lifestyles. They use credit card transaction data set that contains age, gender, and residential zip code of the consumers combined with mobile phone data. The results show that the segments are also differentiated in terms of demographic properties like gender and age.

Like the study of Clemente et al. (2017), Guidotti et al. (2018) also examine the regularities of temporal purchasing behavior of consumers. The authors extract purchasing profiles of customers from a retailer data set to distinguish them according to shopping behavioral patterns. The behavioral characteristics are grouped under two titles. Regular customers that do not involve large number of temporal purchasing behavior patterns and changing customers that involve several types of behavioral patterns. The target of the study is to offer personalized services to the customers based on their temporal purchasing behavior.

Krumme et al. (2013) investigate the predictability of customers' store visitation patterns using transactional data. The authors compute temporally uncorrelated entropy, which takes into account only the frequencies of store visits not containing sequences, and sequence dependent entropy, which takes into account sequence of store visitations. Their results suggest that although predicting the next shopping place of customer involves too much uncertainty, over the long run, the behaviors show regularities and become predictable.

Eagle et al. (2010) use communication network data and national census data in their study to examine the relation between social networks and socioeconomic opportunities. They measure individuals' social and spatial diversity with the Shannon entropy formula. Their findings show that there is a positive association between peoples' diversification in relationships and the economic development of communities.

The study of Song et al (2010) shows that individual mobility can be predicted with a 93% potential. They use a mobile phone data set containing 50,000 users. They introduce three entropy measurements: Random entropy, Temporal-uncorrelated entropy and Actual entropy to find a relation between the random and the regular human mobility. The authors point out that predicting human actions can guide urban planning and traffic engineering, thus it may have positive impact on societies' well-being.

2.2 Shopping Place Choice and Patronage

People determine where to shop according to factors like atmosphere of the shopping place, physical characteristics of the shopping place such as parking lots, distance to travel, variety of the stores or products offered, or the services such as those involving entertainment. In this section, we present the studies in the literature related to shopping place choice and consumers patronage intentions.

Huff (1964) develop a gravity- based model commonly known as the Huff model, which attempts to describe customers' patronage behavior. He comes up with three significant results. The first result shows that people patronize shopping areas that are closer to their home or work places. The second one indicates that for the different types of goods offered, how far consumers are willing to travel changes. Lastly, people tend to shop more in shopping places with a variety of merchants.

Hart et al. (2007) investigate the shopping experience of customers' impact on their re-patronization behavior to the specific shopping malls. The authors examine the relation between perceived image of shopping malls and enjoyment of shopping experience, between enjoyment of shopping experience in a particular mall and the re-patronization behavior, and also between the gender differences and re-patronization behavior. They conduct a questionnaire in the United Kingdom to test these three relationships. They present four elements which are thought to have an impact on shopping enjoyment experience: accessibility, environment, atmosphere, and service personnel. Their results show that the enjoyment of shopping experience is related with these four elements and people are willing to re-patronize the shopping places where they enjoy. In addition, the results indicate that men are more loyal to their specific shopping places and for women shopping enjoyment is more related with browsing a variety of shopping locations and comparing different alternatives while choosing a shopping place.

One of the main reasons that consumers care about store choice is the variety offered. Consumers are more likely to patronize specific stores which offer more varied assortments and thus make it easy to find what they plan to buy. Therefore, retailers try to figure out how consumers perceive the variety and how this perception impacts store choice and satisfaction of consumers. Hoch et al. (1999) develop a mathematical model of perceived variety depending upon spatial locations of objects and their multi-attribute structure by adding psychological set of restrictions on the variety model. The findings show that people mostly choose stores which are perceived as offering high variety of assortments. They conclude that perceived variety will influence store choice when there is uncertainty in the preferences of consumers such as when they do not know which product to buy or which store sells it.

Hozier and Stem (1985) use a dataset which is generated from a mail-based survey to examine the association between outshopping behavior and retail patronage loyalty. Their results show that loyalty has a stronger relationship with outshopping behavior compared to attitudes toward local retailer attributes or demographic variables. They conclude that the unexplained part in outshopping behavior may come from the services offered such as entertainment or gravity related variables.

Sit et al. (2003) study the impact of entertainment attributes on shopping mall patronage and aim to identify the entertainment seeking shopper segment. First, the authors explore the elements that are important of shopping mall image for shoppers. In the second part, they cluster the shoppers into six segments according to determined attributes: merchandising, macro-accessibility, micro-accessibility, personal service, amenities, ambulance, atmospherics, specialty entertainment, special event entertainment, food, and security. The identified six segments are the entertainment shopper, serious shopper, demanding shopper, convenience shopper, apathetic shopper, and service shopper. They introduce entertainment shoppers who see shopping as a leisure activity that can do entertainment activities, socializing or browsing.

Haj-Salem et al. (2016) investigate the elements that lead to shopping place loyalty in the context of shopping malls and different perceptions among genders that lead to mall loyalty. They conduct a questionnaire in two shopping malls located in North

America. Their results indicate that males are affected by atmosphere, prices, and identification with the place. On the other hand, females are affected by the atmosphere, physical design of the place and the quality of products and services.

2.3 Gender Differences in Shopping Experience

Kruger and Byker (2009) argue that gender difference in shopping experience is influenced by foraging strategies adapted in humankind. Although the environment that hunters and gatherers were living and the challenges they faced have changed, they are operating in a same way using same behavioral repertoire (Hantula, 2003). The context of foraging has been shifted to grocery stores, shopping-malls, and websites instead of hunting and gathering due to the cultural evaluation (Hantula, 2003). In this section, we present the articles that study the gender differences in shopping experience.

Bakewell and Mitchell (2004) investigate the shopping decisions of males. In order to do this, they use the Consumer Style Inventory developed by Sproles and Kendall (1986) that profiles consumers according to their decision-styles by categorizing them using eight factors (price/ value consciousness, perfectionism, brand consciousness, novelty/ fashion consciousness, habitual/ brand-loyal, recreational shopping consciousness, impulsive/careless and confused by over choice). The authors also add four factors: store-loyal/low-price seeking, time-energy conserving, confused time restricted and store-promiscuity to the Consumer Style Inventory. They conduct a survey with 245 male undergraduate students. Their findings indicate that there are differences in decision-making styles among males and females. Although some male customers perceive shopping as a leisure activity, the majority of them perceive shopping as a time- and energy-consuming event. Therefore, the majority of male customers shop from the same stores or are indifferent to the stores to spend less time for shopping. In addition to that, they observe brand consciousness in male shoppers. As a result, they conclude that since the male shoppers have different shopping decisions compared to females, a consumer style inventory specific to males is needed.

The study of Teller and Thomson (2012) suggests that there is a gender difference in the perception of accessibility-related attributes of shopping locations like parking and infrastructure. Males care more about the logistics of the shopping effort since they do not prefer to spend time in shopping. However, their findings show that the agglomeration’s attractiveness which constitutes of atmosphere and store variety is perceived the same by females and males.

Noble et al. (2006) examine which factors influence loyalty to local merchants. They use survey data on consumers’ choice on where to shop without having restrictions to local merchants. Their findings indicate that women seek wide assortments and are motivated by the opportunities for browsing products, whereas men search for convenience during shopping. Therefore, women enjoy shopping experience and socially interact during that time. On the other hand, men try to spend less time while browsing or they interact less socially. As a result, the findings show that gender difference has an impact on shopping motivation, which effects local merchant loyalty. In addition to that, authors highlight that women are more likely to be loyal to local merchants that is explained by their dependence on the community where they live.

Alreck and Settle (2002) introduce two different shopping styles that consumers practice while purchasing goods. One is concerned with spending large amounts of time and energy to shop for the best alternative while enjoying the experience. The other one is concerned with shopping for only the required goods while minimizing the shopping time and effort without having pleasure from the shopping experience. The authors conduct a survey on adults involving questions about shopping attitude, shopping style, image profile and demographic status. Their findings show that women are more likely to enjoy shopping experience and perceive shopping as a social activity. On the other hand, men tend to prefer stores which enable them to find their required goods easily without wasting time and effort.

Apart from recreational differences among gender in shopping experience, Grewal et al. (2003) examine the waiting duration expectations and store atmosphere differences in store patronage. They examine customers’ behavior in a jewelry store where the participants in the experiment are unfamiliar with the store. Jewelry store is selected as a place where a special service from the employers is required, unlike the retail stores such as supermarkets and discount stores where customers mostly experience shopping by themselves while browsing for products or trying them. Their findings show that men have less tolerance to waiting compared to women, so long waiting times during shopping will likely decrease their patronization of the store.

Mitchell and Walsh (2004) investigate the usability of the consumer style inventory and determine the differences in decision-making styles of German female and male consumers. Their results show that only four factors are significant in both gender, which are consciousness, perfectionism, over choice and impulsiveness. For the male consumer, the distinctive characteristics like satisfying, time restriction and economy indicate that males are more likely to minimize their shopping time. Also, fashion sale seeking characteristic shows that males are more responsive to sale and they track sale times. On the other hand, for female customers, characteristics such as recreation indicate that they gain pleasure while shopping and perceive shopping as a leisure time activity. In addition to that, females tend to show more variety seeking behavior to shop for new goods. The authors conclude that there is a need to modify the customer style inventory due to the gender differences in consumer decision making.

Otnes and McGrath (2001) offer a different perspective compared to the previous studies in the literature presented so far. They argue that the shopping behavior differences among genders are not as distinct as indicated in other studies. In addition, they state that men's behaviors are stereotyped in the previous studies which is not an accurate reflection of reality. In this study, they analyze the validity of the three male shopper stereotypes: Grab and Go, Whine and/or Wait, Fear of Feminine. Grab and Go refers to the need-driven purchasing behavior of men and not perceiving shopping as a social or recreational activity. Whine and/or Wait suggests that men get bored while accompanying their partners and young men get unhappy with shopping experience. Fear of Feminine suggests that men stay away from shopping behavior or products which connote feminism. The authors make personal interviews to understand men's shopping experience in detail. According to the Jump and Haas (1987), the demographic properties such as high level of education and income associate with the less traditional gender roles. Otnes and McGranth (2001) argue that gender transcendence among men, which can consist of different demographics, help to understand male shopping attitudes.

In addition to the gender difference, Zeithaml (1985) considers demographic features such as female working status, income, age, and marital status to find difference in supermarket shopping behavior. Female or housewife-mothers are seen as the target group of household purchases by the marketers. However, he states that changing roles in the family like the increasing number of working female or divorces, and differentiated demographic profiles affect individuals' supermarket shopping behavior. Therefore, he argues that there is a need to adapt to these changes.

Raajpoot et al. (2008) conduct a questionnaire in three different shopping malls in Montreal to analyze both gender differences and differences of working status of women in shopping center patronage. The authors find the following three major differences between women and men, which they consider as not much significant: a) better product assortments make women's shopping experience more exciting; b) accessible places increase women's shopping experience; c) men pay more attention to employee behavior in stores. The differences in shopping experience among housewife and working women are found much more significant. Housewives care about the accessibility of shopping location compared to working women. On the other hand, working women pay attention to employee behavior in stores and tend to re-patronize more if they are satisfied with the overall shopping experience.

Evans et al. (1996) argue that social and economic influences change the gender roles in shopping behavior discussed in the previous studies. For instance, men start to get involved in shopping activity or women's shopping habits change with their involvement in the workforce or with the increasing number of single mothers. They divide people into three shopper segments: male, working women and female homemakers. They analyze the social impact on these three shopper segments. Their results show that female homemakers perceive shopping as an important role in their lives due to the social norms. On the other hand, working women enjoy shopping experience and perceive it as a social interaction opportunity. The authors foresee that since the patronage intentions are affected by the social referents, males may also perceive shopping as a socialization opportunity and their involvement in shopping will increase in the future.

The studies about gender differences in shopping experience in the literature indicate that women and men perceive shopping differently. Women mostly perceive shopping as leisure time and social activity, whereas men perceive it as a time and energy consuming activity. Besides women seek out wide assortments and are willing to browse for products more compared to men. However, some studies show that due to the change in role of women and men in society and family life, distinguishing shopping experience according to gender only may not be accurate and one should consider other demographics and dimensions as well.

2.4 Our Contribution to the Literature

In the literature, although behavioral features such as diversity and loyalty are used in some studies to understand spatio-temporal behavior of individuals, none of the studies uses them in the shopping mall context. Secondly, past studies use survey data in order to distinguish shopping experience among genders and also associate with demographic features like working status. However, we think that individuals may not give consistent answers in surveys, especially men may hide their interest in shopping since shopping is perceived as a feminine activity. In our study, we explore the real actions of consumers using Big Data for the first time in the literature in this context rather than relying on surveys to find out their shopping experience. Lastly, we constitute consumer segments according to their diversity and loyalty behaviors in shopping malls and shopping categories. In our findings, we obtain some segments which have not been identified before in previous studies.

CHAPTER 3

DATA AND PREPROCESSING

In this chapter, we present the dataset used in this study and explain the data preprocessing methods we used, which prepare our study for the detailed analyses we have conducted. In addition, we propose the use of two behavioral features; diversity and loyalty that give information about customers shopping behaviors in shopping malls. Finally, we demonstrate our analysis of the generated behavioral features.

3.1 Data Collection

In this study, secondary data are used which were collected by one of the leading private banks in Turkey. This bank has over 15 million customers, 4.8 million credit card users and 800 branches in total. A randomly selected sample of 62,392 customers and associated attributes are supplied by the bank for analysis. The time frame for the dataset is one year starting from July 1, 2014 to June 30, 2015. The dataset consists of 20 tables, 269 columns and 28,075,313 rows in total. The explanations and details of each table can be seen in the Table 3.1. Each customer is assigned a unique anonymous ID by the bank and each table has these unique IDs as primary key.

	TABLE	# of columns	# of rows
1	CUSTOMER DEMOGRAPHICS	28	62,392
2	CREDIT CARD INFORMATION	8	61,629
3	BRANCH TRANSACTION	7	339,329
4	CALL CENTER	4	165,029
5	AUTO PAYMENT	4	143,334
6	RISK SCORE	3	728,541
7	CREDIT CARD TRANSACTION	11	4,254,652
8	CREDIT CARD RECEIPT PAYMENT	5	931,100
9	CREDIT CARD RECEIPT	8	811,786
10	ACCOUNT BALANCE	14	748,704
11	ATM TRANSACTION	9	1,428,180
12	ELECTRONIC FUNDS TRANSFER	5	301,454
13	REMITTANCE	7	164,838
14	MOBILE & INTERNET	8	14,340,122
15	RESPONSE SCORE	3	1,019,506
16	CAMPAIGN BATCH	9	819,013
17	CAMPAIGN	7	133,511
18	PRODUCT OWNERSHIP & ACTIVITY	102	748,705
19	CUSTOMER ACTIVENESS	3	811,096
20	CHURN	24	62,392
	TOTAL	**269**	**28,075,313**

Table 3.1: Tables Received from the Bank

In our study, we use the Customer Demographics table and the Credit Card Transaction table only. Apart from the data received from the Bank, we have also collected shopping mall center coordinates located in Istanbul using Google Maps. Since new shopping malls open in Istanbul at a fast rate, all collected shopping malls opening dates are checked and the ones whose opening date is later than 6/30/2015 is removed from the analysis. In total, 66 shopping malls are selected for the study. Table 3.2 shows the details of the data used in the analysis.

Customer Demographics Data
Customer Masked ID
Customer Segment
Branch Code
Branch Coordinates (X and Y)
Customer Home Coordinates (X and Y)
Customer Workplace Coordinates (X and Y)
Gender
Marital Status
Education
Job Type
Income
Age
Bank Age
Risk Code

Credit Card Transactions Data
Customer Masked ID
Date
Time
Amount
Merchant Type
Merchant Masked ID
Online Transaction
Expense Type
Currency
Coordinates (X and Y)

Shopping Mall Data
Shopping Mall Name
Latitude
Longitude

Table 3.2: Collected Data for Study

3.2 Data Preprocessing

Our study focuses on transactions located in Istanbul which has population over 15 million and is the 8th largest city in the world ("Population of Cities in Turkey (2018)"). Istanbul is a large metropolitan consisting of people having a variety of purchasing power and shopping behavior. The data received from the bank consists of credit card transactions distributed across entire Turkey. The first preprocessing step therefore involves selecting credit card transactions located in Istanbul only. The QGIS software is used in order to extract transactions which are located in Istanbul. A fitting rectangle is drawn around Istanbul borders and the data points inside the rectangle are extracted for further analysis. The rectangle involves some data points located in neighboring provinces close to the borders, which are also considered. Figure 3.1 shows the distribution of the extracted credit card transactions. After this step, 2,733,293 credit card transactions remain out of 4,254,652 in total. Reducing the number of rows also allows better computational efficiency for further analysis in the study.

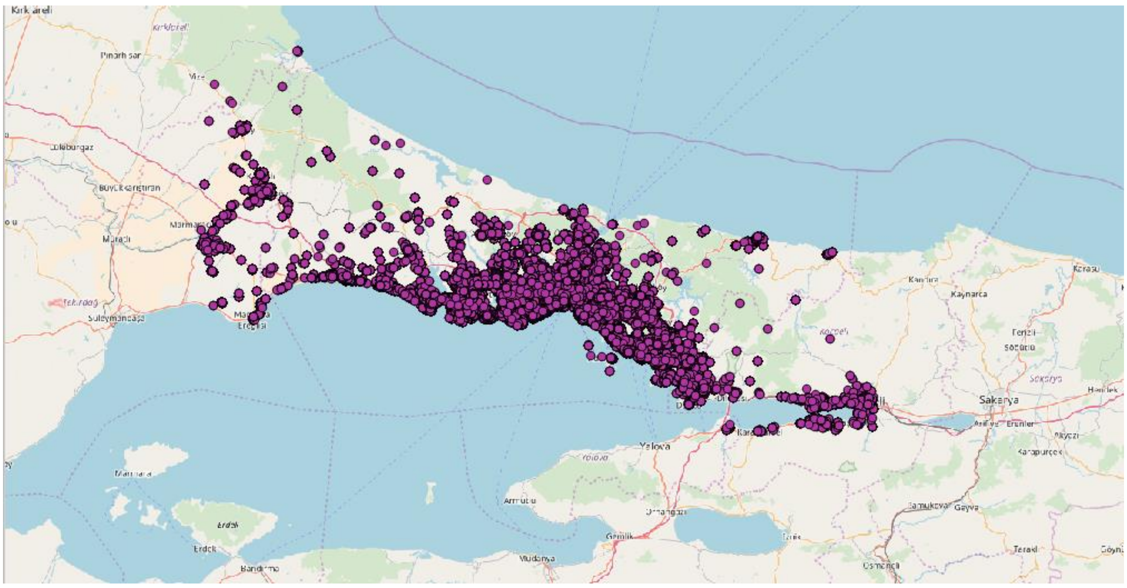

Figure 3.1: Credit Card Transaction Distribution across Istanbul

In the second step, the distance between each transaction and each shopping mall is calculated. If the distance is 200 meters or less, then the transaction is assigned to that shopping mall and is assumed to have taken place at that mall. If multiple shopping malls satisfy the distance criteria, then the transaction is assigned to shopping mall which is closest. A sample of shopping mall locations and transactions which satisfy the criteria can be seen in the Figure 3.2. The red dots represent the locations of the shopping malls and the green dots represent the transactions that are counted within the assigned shopping malls. The Haversine formula given below is used to calculate the great-circle distance between shopping mall locations and transaction locations:

$$dlon = lon2 - lon1$$

$$dlat = lat2 - lat1$$

$$a = \sin^2\left(\frac{dlat}{2}\right) + \cos(lat1) * \cos(lat2) * \sin^2(\frac{dlon}{2})$$

$$c = 2 * atan2(\sqrt{a}, \sqrt{1-a})$$

$$d = R * c \quad \textit{(where R is the radius of the Earth 6,371km)} \tag{3.1}$$

The calculations are done in the R programming language and the geosphere package is used for applying Haversine formula. The results are transferred to the credit card transaction data table in a new column. For each transaction, the assigned shopping mall name (or NA, if no shopping mall is assigned) is entered in this column.

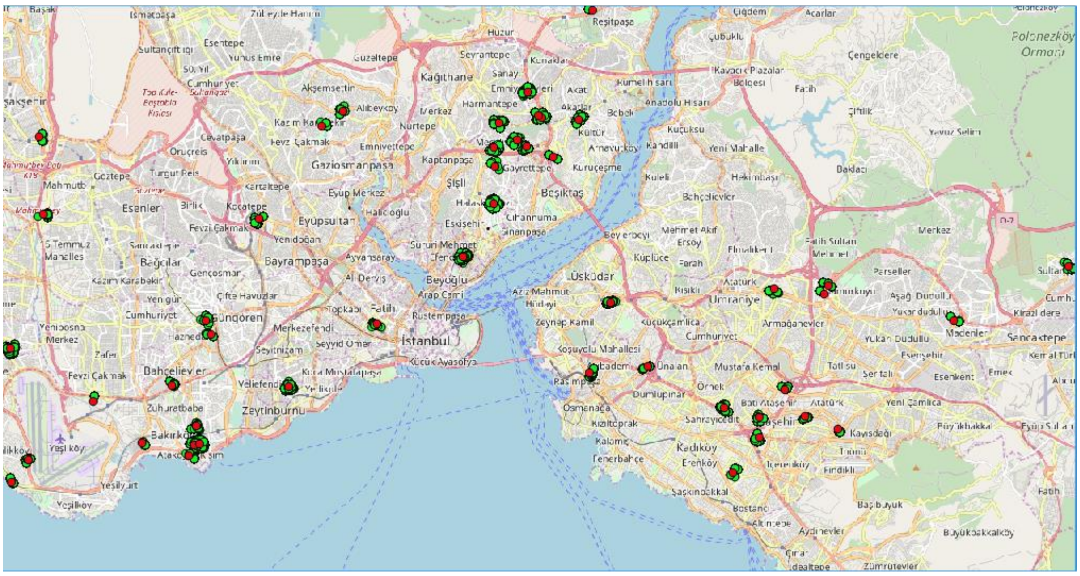

Figure 3.2: Shopping Mall and Transaction Locations

The transaction data table has a merchant type column, which is coded with numeric values. The descriptions of merchant types are provided in a different table which consists of category, merchant type, category name and description. In the third step, some categories are removed, and some new categories are generated for our further analysis. Since our analysis focuses on shopping malls, the categories which cannot occur in shopping malls like car rental, gas station, accommodation and airways are removed. In addition, some merchant types are also removed from categories. For instance, school payments are eliminated from Education / Stationery / Office Equipment category and hospital payments are eliminated from Health / Healthcare Products category. On the other hand, some new categories such as Cosmetics and Entertainment are generated. Cosmetics is separated from Health / Healthcare Products category and Entertainment is separated from Service Sectors category according to category names. Out of 24 categories which are supplied by the bank, we produce 14 categories, which we assume could take place in shopping malls. The list of categories is shown in Table 3.2. In this table, Entertainment corresponds to places such as cinemas, amusement parks, aquariums, Food corresponds to restaurants and fast-food restaurants, Education / Stationary / Office Equipment corresponds to bookstores, hobby stores, stationaries gift shops, Health / Healthcare Products corresponds to pharmacies, Service Sectors correspond to locations like dry cleaning, flower stores, pet shops, photo studios and Various Food corresponds to places like bakeries, confectioners and tobacco shops.

Category
Clothing and Accessory
Electronic Appliance, Computer
Cosmetics
Construction Materials, Hardware Store
Furniture and Decoration
Entertainment
Food
Education/ Stationery/ Office Equipment
Health / Healthcare Products
Supermarket
Goldsmiths
Service Sectors
Various Food
Telecommunications

Table 3.3: Categories Used in the Study

After finalizing the category table, in our fourth step, data integration is performed to merge different tables into a single table. All customers are represented with a unique anonymous customer ID common to all tables. Customer Demographics and Credit Card

Transactions tables are merged based on the customer ID. In addition, the Category Table is merged along with the merchant types into the newly constituted table. After this integration, data cleaning is performed to prepare the data for further analysis.

The last step of our data preprocessing is data cleaning. After merging the data tables, transactions with missing value are deleted. Online transactions, which are indicated by a binary variable with value 1 are deleted from the analysis. The customers who have a total number of transactions less than 12 and have at most 2 shopping mall transactions in one year are considered as inactive and eliminated from the analysis. In addition, the transactions which are not assigned to any shopping mall are removed from the dataset. In the end, 4,254,652 transactions are reduced to 150,828 transactions and 62,392 customers are reduced to 14,843 customers.

3.3 Feature Extraction

In our analysis, we generate behavioral features for each customer. We use their credit card transaction information in order to extract these behavioral features.

3.3.1 Diversity

We calculate diversity which means that a customer's shopping behavior can vary over various "bins". In our case, bins are defined as shopping malls, shopping categories and shopping days of the week. p_{ij} refers to the fraction of transactions that fall within bin *j* for each customer *i*. That is, p_{ij} is calculated for each customer for each bin. Then we calculate the diversity of each customer *i* by normalizing the entropy of transactions counted in all bins by *N*, where *N* denotes the total number of bins. The diversity formula is given below:

$$D_i = \frac{-\sum_{j=1}^{N} p_{ij} log p_{ij}}{log N} \tag{3.2}$$

Due to the normalization, the resulting values D_i are between 0 and 1. Numbers closer to 1 mean higher diversity values for customers. For instance, when a customer transacts equally in almost every different shopping mall, the diversity value becomes almost 1.

Singh et al. (2015) use the same diversity formula in their study with a single difference. They use *M* for normalization instead of *N*, which denotes the total number of non-empty bins instead of the total number of bins. In this case, when a customer spreads his or her transactions almost equally across different bins, then the diversity value becomes high. In our case, however, we prefer the diversity value to be high when the transactions are diversified equally to <u>all</u> bins, so we use a modified version of the Shannon entropy formula used by Singh et al. (2015).

3.3.2 Behavioral Features Generated Using Diversity Formula

<u>Shopping Mall Diversity:</u> The bins are taken as shopping malls and the transaction diversity across shopping malls for each customer is calculated. Values of the shopping mall diversity close to 1 indicate that a customer does her credit card transactions in a large variety of shopping malls.

<u>Category Diversity:</u> The bins are taken as shopping categories and the transaction diversity across shopping categories for each customer is calculated. Values of the category diversity close to 1 indicate that a customer does her credit card transactions in a large variety of categories.

<u>Shopping Mall Diversity for each Individual Shopping Category:</u> Shopping malls are again used as bins similar to shopping mall diversity, however, transactions are filtered according to shopping categories and 14 different diversity scores are calculated for each category for each customer.

<u>Day Diversity</u>: Days of the week are used as bins and for each customer, the shopping day diversity is calculated for the purchases made during the one-year period. Values of the day diversity close to 1 indicate that a customer makes purchases equally in various days of the week.

3.3.3 Loyalty

Loyalty is defined as the percentage of a customer's transactions that take place in his or her *k* most frequented bins. Let f_i be the combined fraction of all transactions of customer *i* that occur in the top *k* most frequented bins. The loyalty of each customer *i* is calculated by the formula given below:

$$L_i = \frac{f_i}{\sum_{j=1}^{N} p_{ij}} \tag{3.3}$$

Loyalty values are between 0 and 1. Larger loyalty values indicate high loyalty behaviors of a customer towards given bins.

3.3.4 Behavioral Features Generated Using Loyalty Formula

Shopping Mall Loyalty: The bins are taken as shopping malls and the value *k* is taken as 2. The value two typically indicates the shopping malls, one close to customer's working place and one close to customer's house. Larger values in shopping mall loyalty score show that a customer makes most of the transactions in the top two visited shopping malls.

Category Loyalty: The bins are taken as shopping categories and the value *k* is taken as 2. Larger values in shopping category loyalty score show that a customer makes most of her transactions in two most preferred categories out of 14 categories.

Shopping Mall Loyalty for Individual Category: Shopping malls are used as bins like the shopping mall loyalty, however, transactions are filtered according to shopping categories and 14 different loyalty scores are calculated for each customer.

Day Loyalty: The bins are taken as days of the week and the value *k* in the equation is taken as 2. The loyalty scores closer to 1 show that a customer has made most of her purchases in two days of the week.

Table 3.4 indicates the demographical, behavioral, and financial features used in our study. The dataset consists of 6 demographic features. The first feature X1 is the unique ID of customers, and the remaining features X2-X6 are the age, the education status, the gender, the marital status, and the job type of the customer, respectively. X7 and X8 are the shopping mall diversity and the category diversity calculated for each customer according to the shopping mall bins and shopping category bins of transactions. X9 to X22 show the diversity for each shopping category according to the shopping mall bins. X23 and X24 are the shopping mall loyalty and the category loyalty calculated for each customer according to the shopping mall bins and shopping category bins of transactions.

Feature Num.	Future Name	Data Type	Feature Type
X1	Customer ID	Integer	Demographic
X2	Age	Double	Demographic
X3	Education Status	Text	Demographic
X4	Gender	Text	Demographic
X5	Marital Status	Text	Demographic
X6	Job Type	Text	Demographic
X7	Shopping Mall Diversity	Double	Behavioral
X8	Category Diversity	Double	Behavioral
X9	Clothing and Accessory Diversity	Double	Behavioral
X10	Electronic Appliance, Computer Diversity	Double	Behavioral
X11	Cosmetics Diversity	Double	Behavioral
X12	Construction Materials, Hardware Store Diversity	Double	Behavioral
X13	Furniture and Decoration Diversity	Double	Behavioral
X14	Entertainment Diversity	Double	Behavioral
X15	Food Diversity	Double	Behavioral
X16	Education/ Stationery/ Office Equipment Diversity	Double	Behavioral
X17	Health / Healthcare Products Diversity	Double	Behavioral
X18	Supermarket Diversity	Double	Behavioral
X19	Goldmiths Diversity	Double	Behavioral
X20	Service Sectors Diversity	Double	Behavioral
X21	Various Food Diversity	Double	Behavioral
X22	Telecommunications Diversity	Double	Behavioral
X23	Shopping Mall Loyalty	Double	Behavioral
X24	Category Loyalty	Double	Behavioral
X25	Clothing and Accessory Loyalty	Double	Behavioral
X26	Electronic Appliance, Computer Loyalty	Double	Behavioral
X27	Cosmetics Loyalty	Double	Behavioral
X28	Construction Materials, Hardware Store Loyalty	Double	Behavioral
X29	Furniture and Decoration Loyalty	Double	Behavioral
X30	Entertainment Loyalty	Double	Behavioral
X31	Food Loyalty	Double	Behavioral
X32	Education/ Stationery/ Office Equipment Loyalty	Double	Behavioral
X33	Health / Healthcare Products Loyalty	Double	Behavioral
X34	Supermarket Loyalty	Double	Behavioral
X35	Goldmiths Loyalty	Double	Behavioral
X36	Service Sectors Loyalty	Double	Behavioral
X37	Various Food Loyalty	Double	Behavioral
X38	Telecommunications Loyalty	Double	Behavioral
X39	Day Diversity	Double	Behavioral
X40	Day Loyalty	Double	Behavioral
X41	Average Transaction Amount	Double	Financial

Table 3.4: Feature Properties

X25 through_X38 are the loyalty features for each shopping category according to the shopping mall bins. X39 and X40 indicate the shopping day diversity and loyalty that are calculated according to the days of the week as bins. X41 indicates the average transaction amount for each customer.

3.4 Descriptive Statistics

In this part, we report the descriptive statistics of our features in Table 3.3 including their minimum (Min), maximum (Max), mean, median, first quartile (1st QU) and third quartile (3rd Qu), standard deviation (Std Dev) and the number of missing attributes (NA's) for numeric features and number of occurrences for categorical features.

Demographic Features					
Age		**Education Status**		**Job Type**	
Min:	19	Primary School:	441	Private Sector Employee:	10297
1st Qu:	31	Middle School:	684	Public Servant:	1282
Median:	37	High School:	5540	Retiree:	1060
Mean:	38.56	College:	1458	Self-Employed:	1784
3rd Qu:	45	University:	5756	Non-Employed:	93
Max:	83	Master:	734	Housewife:	212
Std Dev:	9.56	PhD:	63	Other:	115
		Uneducated:	157		
		Unknown:	10		
Gender		**Marital Status**			
Female:	6946	Single:	3527		
Male:	7897	Married:	10115		
		Divorced:	778		
		Unknown:	423		

Table 3.5: Descriptive Statistics of Demographic Features

Behavioral and Financial Features					
Shopping Mall Diversity		**Category Diversity**		**Clothing and Accessory Diversity**	
Min:	0.0000	Min:	0.0000	Min:	0.0000
1st Qu:	0.0979	1st Qu:	0.1554	1st Qu:	0.0000
Median:	0.1632	Median:	0.2412	Median:	0.1519
Mean:	0.1757	Mean:	0.2536	Mean:	0.1283
3rd Qu:	0.2620	3rd Qu:	0.3871	3rd Qu:	0.2250
Max:	0.5517	Max:	0.7434	Max:	0.5122
Std Dev:	0.1183	Std Dev:	0.1680	Std Dev:	0.1162
NA's:	0	NA's:	0	NA's:	1273
Electronic Appliance, Computer Diversity		**Cosmetics Diversity**		**Construction Materials, Hardware Store Diversity**	
Min:	0.0000	Min:	0.0000	Min:	0.0000

1st Qu: 0.0000 Median: 0.0000 Mean: 0.0220 3rd Qu: 0.0000 Max: 0.3640 Std Dev: 0.0585 NA's: 10881	1st Qu: 0.0000 Median: 0.0000 Mean: 0.0310 3rd Qu: 0.0000 Max: 0.3840 Std Dev: 0.0691 NA's: 11045	1st Qu: 0.0000 Median: 0.0000 Mean: 0.0010 3rd Qu: 0.0000 Max: 0.1650 Std Dev: 0.0144 NA's: 13949
Furniture and Decoration Diversity	**Service Sectors Diversity**	**Education/ Stationery/ Office Equipment Diversity**
Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0360 3rd Qu: 0.0000 Max: 0.3840 Std Dev: 0.0730 NA's: 10101	Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0020 3rd Qu: 0.0000 Max: 0.1650 Std Dev: 0.0197 NA's: 14703	Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0030 3rd Qu: 0.0000 Max: 0.3180 Std Dev: 0.0265 NA's: 14570
Goldsmiths Diversity	**Entertainment Diversity**	**Supermarket Diversity**
Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0038 3rd Qu: 0.0000 Max: 0.1654 Std Dev: 0.0244 NA's: 14562	Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0030 3rd Qu: 0.0000 Max: 0.3180 Std Dev: 0.0265 NA's: 14570	Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0100 3rd Qu: 0.0000 Max: 0.2760 Std Dev: 0.0388 NA's: 11048
Food Diversity	**Telecommunications Diversity**	**Health / Healthcare Products Diversity**
Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0310 3rd Qu: 0.0000 Max: 0.3720 Std Dev: 0.0657 NA's: 11588	Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0020 3rd Qu: 0.0000 Max: 0.1650 Std Dev: 0.0158 NA's: 14702	Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0100 3rd Qu: 0.0000 Max: 0.2620 Std Dev: 0.0404 NA's: 13120
Various Food Diversity	**Shopping Mall Loyalty**	**Category Loyalty**
Min: 0.0000 1st Qu: 0.0000 Median: 0.0000 Mean: 0.0080 3rd Qu: 0.0000 Max: 0.2620 Std Dev: 0.0365 NA's: 13980	Min: 0.2857 1st Qu: 0.7500 Median: 0.9565 Mean: 0.8699 3rd Qu: 1.0000 Max: 1.0000 Std Dev: 0.1581 NA's: 0	Min: 0.0000 1st Qu: 0.8000 Median: 1.0000 Mean: 0.8980 3rd Qu: 1.0000 Max: 1.0000 Std Dev: 0.1327 NA's: 0

Clothing and Accessory Loyalty		**Electronic Appliance, Computer Loyalty**		**Cosmetics Loyalty**	
Min:	0.2870	Min:	0.5380	Min:	0.4000
1st Qu:	0.8571	1st Qu:	1.0000	1st Qu:	1.0000
Median:	1.0000	Median:	1.0000	Median:	1.0000
Mean:	0.9206	Mean:	0.9960	Mean:	0.9910
3rd Qu:	1.0000	3rd Qu:	1.0000	3rd Qu:	1.0000
Max:	1.0000	Max:	1.0000	Max:	1.0000
Std Dev:	0.1357	Std Dev:	0.0350	Std Dev:	0.0497
NA's:	1274	NA's:	10882	NA's:	11045
Furniture and Decoration Loyalty		**Construction Materials, Hardware Store Loyalty**		**Service Sectors Loyalty**	
Min:	0.4000	Min:	0.5380	Min:	1.0000
1st Qu:	1.0000	1st Qu:	1.0000	1st Qu:	1.0000
Median:	1.0000	Median:	0.0000	Median:	1.0000
Mean:	0.9910	Mean:	0.9960	Mean:	1.0000
3rd Qu:	1.0000	3rd Qu:	1.0000	3rd Qu:	1.0000
Max:	1.0000	Max:	1.0000	Max:	1.0000
Std Dev:	0.0505	Std Dev:	0.0350	Std Dev:	0.0000
NA's:	10102	NA's:	10882	NA's:	14703
Education/ Stationery/ Office Equipment Loyalty		**Various Food Loyalty**		**Entertainment Loyalty**	
Min:	0.5000	Min:	0.6670	Min:	0.6000
1st Qu:	1.0000	1st Qu:	0.0000	1st Qu:	1.0000
Median:	1.0000	Median:	1.0000	Median:	1.0000
Mean:	0.9930	Mean:	0.9980	Mean:	0.9980
3rd Qu:	1.0000	3rd Qu:	1.0000	3rd Qu:	1.0000
Max:	1.0000	Max:	1.0000	Max:	1.0000
Std Dev:	0.0424	Std Dev:	0.0206	Std Dev:	0.0242
NA's:	12504	NA's:	13981	NA's:	14570
Goldsmiths Loyalty		**Food Loyalty**		**Supermarket Loyalty**	
Min:	1.0000	Min:	0.4000	Min:	0.6670
1st Qu:	1.0000	1st Qu:	1.0000	1st Qu:	1.0000
Median:	1.0000	Median:	1.0000	Median:	1.0000
Mean:	1.0000	Mean:	0.9860	Mean:	0.9990
3rd Qu:	1.0000	3rd Qu:	1.0000	3rd Qu:	1.0000
Max:	1.0000	Max:	1.0000	Max:	1.0000
Std Dev:	0.0000	Std Dev:	0.0654	Std Dev:	0.0179
NA's:	14563	NA's:	11590	NA's:	11048
Telecommunications Loyalty		**Health / Healthcare Products Loyalty**		**Shopping Day Loyalty**	
Min:	1.0000	Min:	0.3330	Min:	0.3077
1st Qu:	1.0000	1st Qu:	1.0000	1st Qu:	0.5833
Median:	1.0000	Median:	1.0000	Median:	0.7000
Mean:	1.0000	Mean:	0.9740	Mean:	0.7243

3rd Qu: 1.0000 Max: 1.0000 Std Dev: 0.0000 NA's: 14702	3rd Qu: 1.0000 Max: 1.0000 Std Dev: 0.1056 NA's: 13120	3rd Qu: 0.8571 Max: 1.0000 Std Dev: 0.1872 NA's: 0
Shopping Day Diversity	**Average Transaction Amount**	
Min: 0.0000 1st Qu: 0.4615 Median: 0.5929 Mean: 0.5917 3rd Qu: 0.7733 Max: 0.9967 Std Dev: 0.2294 NA's: 0	Min: 3.787 1st Qu: 58.487 Median: 94.322 Mean: 165.080 3rd Qu: 162.290 Max: 19037.33 Std Dev: 366.7205 NA's: 0	

Table 3.6: Descriptive Statistics of Behavioral and Financial Features

3.5 Explanatory Data Analysis

In this section, we present the explanatory analysis we have done, before constructing models.

3.5.1 Dispersion of Customers' Transaction Counts

Figure 3.3 shows the histogram of customers' yearly total transaction count in shopping malls located in Istanbul. The minimum number of transactions is 3 and the maximum number of transaction is 611. The average transaction count of customer's in the data set is 10.17. 80% of the transaction counts are between 3 and 40.

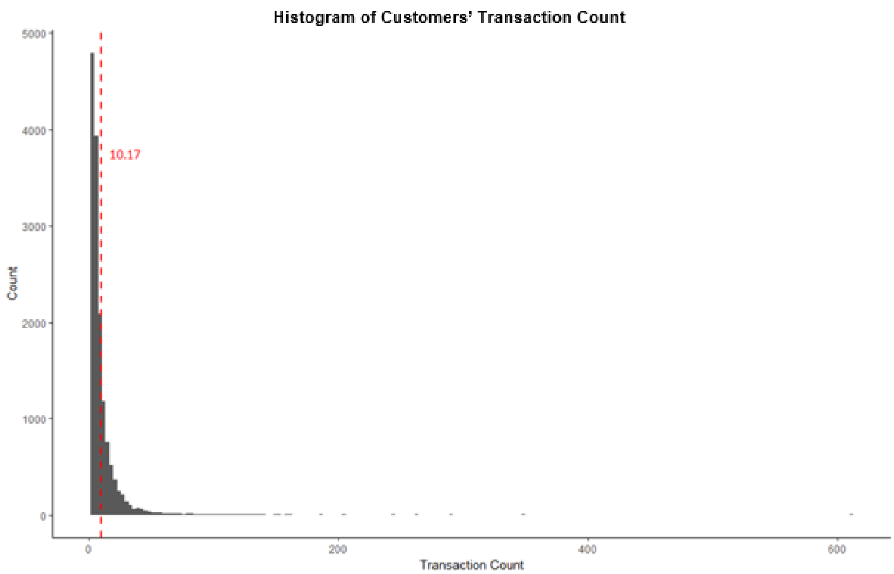

Figure 3. 3: Histogram of Customers' Yearly Transaction Count in Shopping Malls

3.5.2 Diversity and Loyalty Analysis

The cumulative density function (CDF) of diversity in Figure 3.4 shows that the customers are more diverse in terms of their shopping categories for transactions made in shopping malls than the shopping malls they visited. In addition, 20% of the customers have a diversity score of 0 for either location or category, which means that they have done their purchases on the same shopping mall or on the same category.

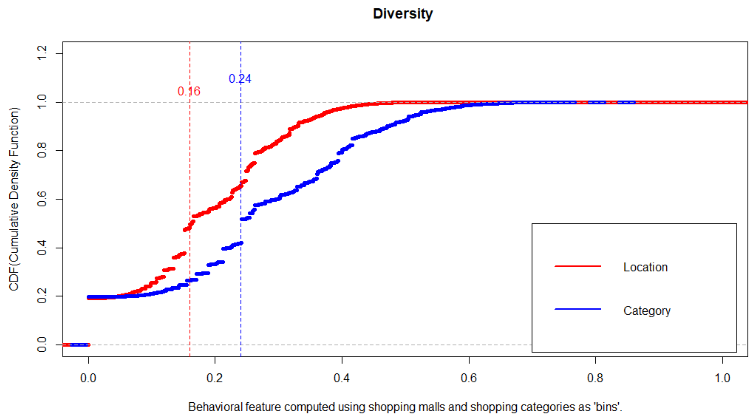

Figure 3. 4: Cumulative Density Function (CDF) of Diversity

The cumulative density function (CDF) of diversity in Figure 3.5 indicates that approximately 95% of the purchases are done in the two most preferred shopping malls by customers. In addition, nearly half of the customers make all their purchases on their most preferred two categories in shopping malls.

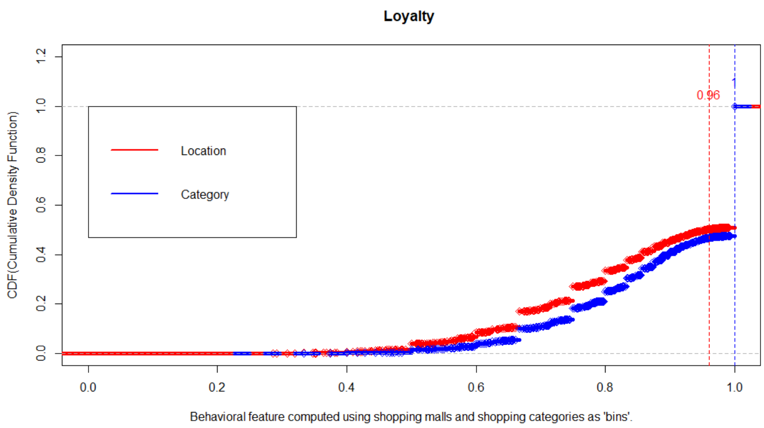

Figure 3. 5: Cumulative Density Function (CDF) of Loyalty

3.5.3 Category Diversity and Category Loyalty Analysis

Distinct customer numbers that are calculated by counting the number of customers who make a purchase in a given shopping category can be seen in Figure 3.6. The top six category that have the highest number of distinct customer counts are selected for the analysis. These are Clothing and Accessory, Furniture and Decoration, Electronic Appliance and Computer, Cosmetics, Supermarket, and Food.

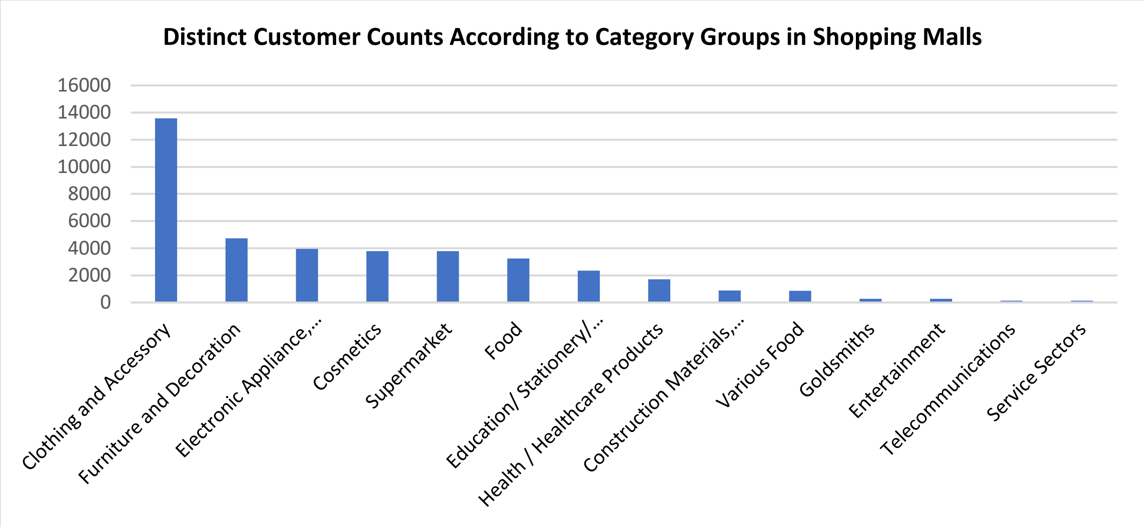

Figure 3. 6: Distinct Customer Counts by Category

Figure 3.7 shows that people are more diverse in Clothing and Accessory purchases in shopping malls. In other words, people prefer to visit various shopping malls and make purchases for Clothing and Accessory. The second most diverse category is Furniture and Decoration and the least diverse category among the selected six categories is Supermarket purchases.

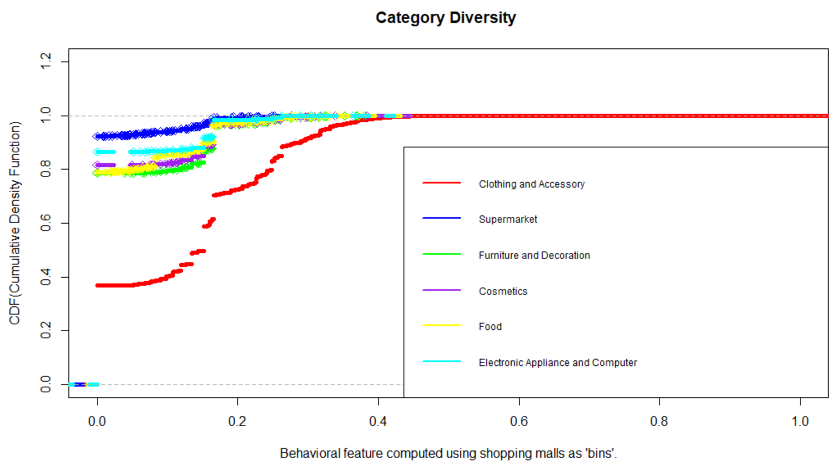

Figure 3. 7: Cumulative Density Function (CDF) of Category Diversity

People prefer to shop in same shopping malls they used to go for categories: Supermarket, Furniture and Decoration, Cosmetics and Electronic Appliances and

Computer. On the other hand, Figure 3.8 shows that people are more likely to do their Clothing and Accessory purchases in different shopping malls rather than their most preferred two shopping malls they used to go.

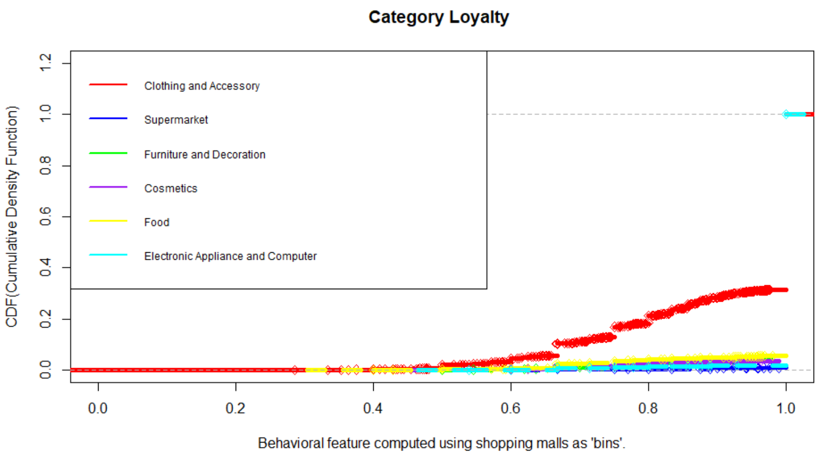

Figure 3. 8: Cumulative Density Function (CDF) of Category Loyalty

CHAPTER 4

METHODOLOGY

In this chapter, we explain the algorithms that we use in this study. First, in order to segment customers according to their shopping behavior, we use K- means clustering algorithm, which is a part of unsupervised learning. Secondly, we provide explanations of how we classify customers into the segments that we constitute with the demographic information of customers.

4.1 Unsupervised Learning

Data mining approaches can be categorized into two: supervised learning and unsupervised learning. In unsupervised learning the data is unlabeled, and the target is to detect unknown patterns and recognize relationships among input measurements unlike predicting an outcome in supervised learning. Since the data points do not have the associated ground truth values, it is not possible to measure the accuracy of the outcome of the models in unsupervised learning. Unsupervised learning includes association rules, cluster analysis and principal component analysis.

4.1.1 K-means Clustering

K-means clustering algorithm developed by MacQueen et al. (1967) aims to partition the observations in a data set into k number of groups. The desired number of clusters k is determined beforehand. In the first step of algorithm, *k* number of data points are randomly chosen from the dataset to become the initial set of cluster "centers". In the second step, all data points are assigned to the closest center. Then each center, which is a vector of the feature means of the data points within its corresponding cluster, is recalculated for each cluster. The data points are then reassigned to the new closest center, and the algorithm continues to iterate until the centers of the clusters remain unchanged from one iteration to the next.

In our study, we apply the K-means clustering algorithm in order to identify the groups of similar customers according to their shopping behaviors. We propose two different K-means clustering models containing either shopping mall diversity and category diversity or shopping mall loyalty and category loyalty.

4.1.1.2 Determining Number of Clusters

In our study, to determine number of clusters, we use the Elbow Method. For the number of clusters, *k*, ranging from 1 to 10, the total Within Clusters Sum of Squares (WCSS) is calculated. WCSS is defined as the sum of the squared distance between each data point in the cluster and the cluster center. The optimal number of clusters is found by the Elbow Method where the number of clusters *k* is chosen when the dramatic decrease of total WCSS stops at a value *k*. When the total WCSS is plotted against the number of clusters, an angle can be seen at value *k* and after *k*, it reaches a plateau (Bholowalia & Kumar, 2014). After this value *k*, increasing the number of clusters will not provide better modelling of the data.

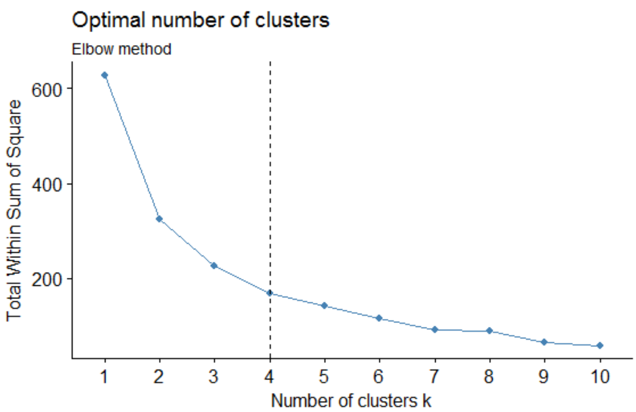

Figure 4.1: Optimal Number of Diversity Based Clusters

In the two models constructed based on diversity and loyalty scores, we applied the Elbow Method in order to select the optimal number of clusters. In Figure 4.1 and Figure 4.2, it can be seen that 4 is the optimal number of clusters for both diversity based segments and loyalty based segments.

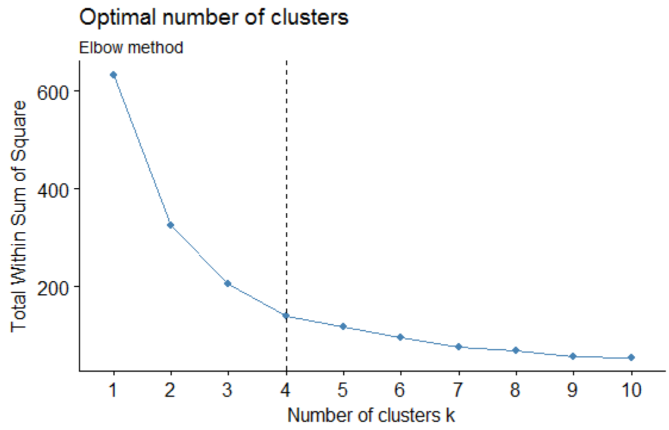

Figure 4.2: Optimal Number of Loyalty Based Clusters

4.2 Classification Algorithm

We are not normally able to calculate diversity and loyalty of new customers, and hence assign them to existing segments, since we do not have the past data of their transactions. One way to assign them to existing clusters is through the use of their demographic features, and here we propose an approach to do that by measuring the cosine similarity between a customer’s demographic profile and the segment’s demographic profile. In the second part of our study, our aim is to foresee new customers' shopping behavior by considering demographics while obtaining high accuracy. In order to do this, we split our data into train and test sets constituting 70% and 30% of the customers, respectively. We apply our classification algorithm for both diversity and loyalty based segments.

In the first step, we apply K-means algorithm on the train set and constitute four different clusters. Then, we measure the distribution of the demographic properties and the centers of each resulting cluster. The information on demographic profiles and centers of each cluster can be found in Appendix A. In the third step, we use Euclidian distance to assign each customer in the test set to the nearest cluster using the train set centers. The distance between test set’s behavioral features including mall and category based values ($test_m$, $test_c$) and the train set centers ($train_m$, $train_c$) are calculated with the following formula (4.1). The customer is assigned to the cluster that gives the minimum Euclidian distance. Therefore, the assigned labels for each customer become the ground truth.

$$\text{dist}((\text{test}_m, \text{test}_c), (\text{train}_m, \text{train}_c)) = \sqrt{(test_m - train_m)^2 + (test_c - train_c)^2} \qquad (4.1)$$

In the fourth step, we measure the cosine similarity between train set's demographic profile and the customers in test set. Before doing that, in order to include significant demographic features in our model, we calculate chi-square statistics for each demographic property in the train set and check the significance by taking the significance level when *p*-value is less than 0.05. Our findings show that job type is not a distinguishing characteristic among clusters, so we exclude it in our model. We generate vectors composed of proportions of gender, age, marital status, and education status within its cluster for each cluster. In addition, for each customer, a binary vector with the customer's gender, age, marital status, and education status is generated. The sample of generated vectors for both segments and customers can be seen in Table 4.1. The segment names will be explained in Chapter 5.

	Male	Female	Married	Single	Divorced	Unknown	19-30	31-42	43-54	55-66	67-83	Primary School	Middle School	High School	College	University	Master	PHD	Uneducated	Unknown
Recreational Shoppers	0.53	0.47	0.73	0.18	0.07	0.02	0.10	0.54	0.29	0.07	0.01	0.01	0.03	0.29	0.10	0.47	0.08	0.01	0.01	0.00
Category Hunters	0.58	0.42	0.74	0.18	0.06	0.02	0.13	0.49	0.29	0.08	0.01	0.03	0.06	0.38	0.10	0.37	0.05	0.00	0.01	0.00
Mall Squatters	0.61	0.39	0.78	0.14	0.06	0.02	0.10	0.51	0.30	0.08	0.01	0.03	0.05	0.37	0.11	0.36	0.05	0.01	0.01	0.00
Mission Shoppers	0.69	0.31	0.78	0.15	0.04	0.02	0.09	0.51	0.32	0.07	0.01	0.05	0.06	0.48	0.08	0.27	0.03	0.00	0.01	0.00
Customer 1	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
Customer 2	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
Customer 3	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0

Table 4. 1: Sample of Demographic Property Vectors

The cosine similarity measures the cosine of the angle between of each cluster vector ($\vec{a}$) and customer vector ($\vec{b}$). The formula used to calculate cosine similarity can be seen in (4.2). The outcome ranges between 0 and 1, where the values closer to 1 indicate high similarity between the two vectors. Then the customer is assigned to the cluster which gives the maximum similarity measure in terms of demographic profile.

$$cos\theta = \frac{\vec{a}.\vec{b}}{\|\vec{a}\|\|\vec{b}\|} \qquad (4.2)$$

In the final step, we calculate the matching rate of the cosine similarity cluster assignments and the ground truth. The results are presented in Chapter 5.

CHAPTER 5

RESULTS AND DISCUSSION

In this chapter, we give the description of constructed segments (i.e. clusters) and their demographic, transactional and shopping category profiles. In addition, we clarify the relationship between the two cluster models comprising of diversity and loyalty, and their constructed segments. Lastly, we present the results of our segment classifications.

5.1 Consumer Segments

The K-means clustering method is used to implement two different models based on diversity scores and loyalty scores regarding the consumers' shopping behavior in the context of shopping malls and shopping categories. The following two sections give the detailed information about the consumer segments we identify with our dataset.

5.1.1 Diversity Based Segmentation

Shopping Mall Diversity scores and Category Diversity scores are used for clustering in order to segment consumers according to their shopping behaviors. Four different consumer segments are identified and labeled based on the average values of the two diversity scores. Figure 5.1 shows the resulting four clusters on a scatter plot. One-way analysis of variance (ANOVA) test shows that the average values of each feature are statistically different between each pair of segments at the significance level 0.05. (Table 5.1)

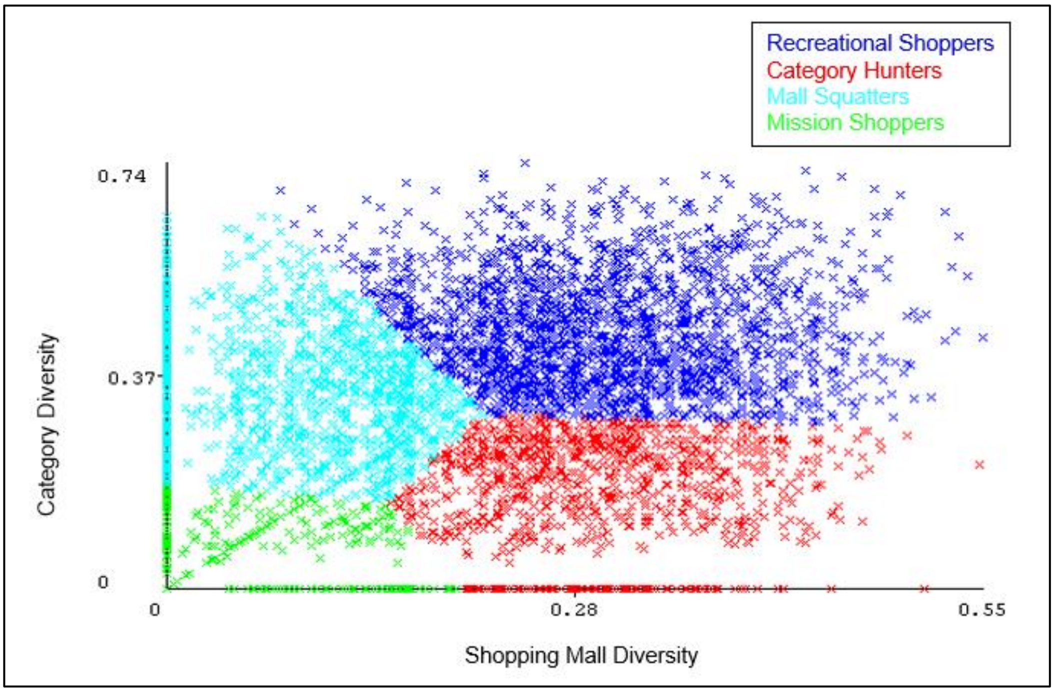

Figure 5.1: Diversity Based Segments

Ruiz et al. (2004) construct four segments according to the activities consumers perform in a shopping mall. In our study, we named the first segment as Recreational Shoppers, which is a name similar to the one by Ruiz et al. (2004) given to the group of consumers who see shopping as a social interaction and leisure activity. This segment comprises of 3826 customers that equals 26% of all customers. Post-hoc Tukey test indicates that this segment contains, on average, customers with high Shopping Mall Diversity scores as well as high Category Diversity scores. This segment consists of people who prefer to make their purchases in diversified shopping malls and diversified categories in the shopping malls. It can be said that these people like to walk around in malls to shop for various categories and also visit shopping malls in various locations, which is a sign that they like to enjoy the shopping experience.

The second segment is named as Category Hunters. This segment comprises of 5089 customers, which is 34% of all customers. The people that take part in this segment have, on average, high Shopping Mall Diversity scores and low Category Diversity scores. This segment consists of people who make their purchases in a variety of shopping malls focusing on their particular categories in those shopping malls. These people have willingness to search for their targeted categories in various shopping malls.

The third segment is named as Mall Squatters. This segment comprises of 3114 customers, which corresponds to 21% of all customers. The people who are part of this segment have, on average, low Shopping Mall Diversity scores and high Category Diversity scores. This segment consists of people who make their purchases in their preferred shopping malls and on various categories in those shopping malls. These shopping malls satisfy the needs of the customers in this segment by offering variety of category alternatives, therefore they do not need to visit any other malls.

Ruiz et al. (2004) identified one group of customers as Mission Shoppers who only go to the malls to buy products they already planned to purchase. We also named our last segment as Mission Shoppers. This segment comprises of 2814 customers, which is 19% of all customers. According to the Post-hoc Tukey test, people have, on average, low Shopping Mall Diversity scores and low Category Diversity scores compared to the other segments. The people who are part of this segment prefer to visit few numbers of shopping malls and make purchases in few categories in those shopping malls. We argue that people who fall within this segment are target oriented and they visit particular shopping malls to buy particular products.

	Recreational Shoppers	**Category Hunters**	**Mall Squatters**	**Mission Shoppers**	**Total**	**F**	**p-value**
Shopping Mall Diversity Score	0.28	0.27	0.10	0.06	0.18	10724	0.000
Category Diversity Score	0.44	0.16	0.30	0.03	0.25	14099	0.000

Table 5.1: Consumer Segments and Their Average Diversity Scores

5.1.1.1 Demographic Profiles of Diversity Based Segments

Statistics of the diversity based segments’ demographic profiles can be seen in Table 5.2. In order to find out statistically significant differences between consumer segments, Chi-square test or one-way analysis of variance (ANOVA) test are used. Differences are accepted as statistically significant when p-value is less than 0.05.

Chi-square tests are performed to examine the relation between consumer segments and demographic features such as gender, job type, marital status and education status. The relation between gender and consumer segments are significant, X^2 (3, N = 14843) = 263.50, p <.000. A significant relation between marital status and consumer segments is found, X^2 (9, N = 14843) =79.91, p <.000. The relation between education status and consumer segments are found to be statistically significant, X^2 (24, N = 14843) =607.060, p <.000. On the other hand, there is not a statistically significant relation between job type and consumer segments, X^2 (21, N = 14843) =27.728, p = .0663.

A one-way ANOVA is conducted to compare the difference of age across all consumer segments. The analysis shows that there is such a difference and it is significant, F (3, 14839) = 10.76, p <.000. Post hoc comparisons using the Tukey test are carried out. The post hoc comparisons reveals that Category Hunters are the youngest segment.

Pairwise comparison of proportion test shows that the proportion of females in Category Hunters is significantly larger compared to Mall Squatters and Mission Shoppers. On the other hand, the proportion of males in Mission Shoppers is significantly larger compared to Recreational Shoppers and Category Hunters.

Pairwise comparison of proportion test indicates that the proportion of married people in Mission Shoppers and Mall Squatters is significantly larger compared to Category Hunters and Recreational Shoppers. On the other hand, the proportion of single people in Category Hunters is significantly larger compared to Mission Shoppers and Mall Squatters.

We generate two groups named as low education level and high education level in order to test the significance of proportion difference among segments. Low education level comprises of uneducated, primary school, middle school, and high school categories. On the other hand, high education level comprises of college, university, Master and PhD. Our results show that Recreational Shoppers have the highest education level with high education proportion of 0.67. On the other hand, Mission Shoppers have the lowest education level with low education proportion of 0.57.

In Appendix B, charts associated with segment properties can be found.

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total	Test Statistic	p-value
Gender						X^2	
Male	47.4%	44.7%	58.9%	60.2%	53.2%		
Female	52.6%	55.3%	41.1%	39.8%	46.8%	263.50	0.000
Marital Status						X^2	
Married	67.1%	62.9%	71.2%	69.9%	68.1%		
Single	24.8%	28.5%	21.1%	21.9%	23.8%		
Divorced	5.7%	5.7%	4.8%	5.0%	5.2%		
Unknown	2.5%	2.9%	2.9%	3.2%	2.8%	79.91	0.000
Age Group						F	
19-30	21.1%	26.7%	19.9%	21.4%	21.9%		
31-42	49.2%	43.7%	47.8%	46.1%	47.0%		
43-54	23.2%	23.7%	25.3%	26.1%	24.5%		
55-66	5.9%	5.6%	6.2%	5.7%	5.9%		
67-83	0.7%	0.4%	0.9%	0.6%	0.7%	10.76	0.000
Education Status						X^2	
Primary School	1.2%	2.5%	3.4%	5.2%	3.0%		
Middle School	2.6%	4.6%	5.3%	6.3%	4.6%		
High School	28.1%	34.8%	41.6%	44.9%	37.3%		
College	9.7%	10.1%	10.1%	9.0%	9.8%		
University	49.7%	41.7%	34.1%	29.3%	38.8%		
Master	7.3%	5.0%	3.9%	3.6%	4.9%		
PhD	0.5%	0.4%	0.4%	0.4%	0.4%		
Uneducated	0.9%	0.9%	1.2%	1.3%	1.1%		
Unknown	0.1%	0.1%	0.1%	0.0%	0.1%	607.060	0.000
Job Type						X^2	
Private Sector Employee	68.6%	68.2%	70.3%	70.0%	69.4%		
Public Servant	9.8%	9.3%	7.7%	8.0%	8.6%		
Self-Employed	11.3%	12.4%	12.0%	12.5%	12.0%		
Retiree	7.5%	7.0%	7.1%	6.8%	7.1%		
Housewife	1.4%	1.6%	1.5%	1.3%	1.4%		
Non-Employed	0.4%	0.6%	0.7%	0.8%	0.6%		
Other	0.9%	0.9%	0.7%	0.6%	0.8%	27.728	0.0663

Table 5.2: Demographic Profiles of Diversity Based Segments

5.1.1.2 Transactional Characteristics of Diversity Based Segments

Here we first calculate the average transaction amount per customer for each diversity based segment, which can be seen in Table 5.3. A one-way ANOVA test is conducted to compare the difference in average transaction amount per customer across diversity based segments. The results show that there is a statistically significant difference of average transaction amount per customer among consumer segments, $F(3, 14839) = 3.475$, p =

0.0153. Afterwards, post hoc Tukey test is carried out and the result of the test reveals that the average transaction amounts per customer of Recreational Shoppers and Mission Shoppers are different significantly at $p < .05$. Recreational Shoppers have the lowest average transaction amount and Mission Shoppers have the highest.

	Recreational Shoppers	**Category Hunters**	**Mall Squatters**	**Mission Shoppers**	**Total**	**F**	**p-value**
Average Transaction Amount per Customer	151.3	163.7	168.1	180	165.08	3.475	0.0153

Table 5.3: Average Transaction Amount per Customer of Diversity Based Segments

Secondly, average transaction count per customer for each diversity based segment is calculated and the details can be seen in Table 5.4. The results of the ANOVA test show that there is a significant difference of average transaction count per customer among consumer segments, $F(3, 14839) = 103.7$, $p < .000$. The post hoc Tukey results indicate that Recreational Shoppers, on average, have relatively more transaction counts in shopping malls compared to other segments.

	Recreational Shoppers	**Category Hunters**	**Mall Squatters**	**Mission Shoppers**	**Total**	**F**	**p-value**
Average Transaction Count per Customer	13.5	8.8	9.4	8.6	10.2	103.7	0.000

Table 5.4: Average Transaction Count per Customer of diversity Based Segments

Thirdly, Table 5.5 shows the average of total transaction amount per customer values for each consumer segment. The results of the ANOVA test indicate that total transaction amount of per customer is significantly different among customer segments, $F(3, 14839) = 17.92$, $p < .000$. According to the Post hoc Tukey test results, the total transaction amount of Recreational Shoppers is the highest of all segments.

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total	F	p-value
Total Transaction Amount per Customer	1798.9	1433.7	1209.7	1363.6	1437.8	17.92	0.000

Table 5.5: Total Transaction Amount per Customer of Diversity Based Segments

Lastly, The ANOVA test results show that there are significant differences between consumer segments and their Shopping Day Diversity and Shopping Day Loyalty scores. Findings in Table 5.6 show that Recreational Shoppers prefer to shop in diversified days of the week. Since their visit days are varied, they are more willing to see shopping as a leisure activity. This also promotes the findings stated by Ruiz et al. (2004) that Recreational Shoppers are looking for fun, leisure, and social interaction. On the other hand, Mission Shoppers have high Shopping Day Loyalty scores. They are more likely to shop their most preferred two days of the week compared to the other segments. Since they are task-oriented shoppers, they visit shopping malls on specific days of the week to buy the items they planned to purchase.

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total	F	p-value
Shopping Day Diversity	0.69	0.60	0.56	0.50	0.59	438.4	0.000
Shopping Day Loyalty	0.66	0.71	0.75	0.79	0.72	304.1	0.000

Table 5.6: Shopping Day Behavioral Scores of Diversity Based Segments

5.1.1.3 Shopping Category Profile of Diversity Based Segments

In Table 5.7 we report the percentage of total transaction counts according to the shopping categories for each segment. We calculate the variation of each segment’s shopping category from total shopping category by subtracting total category percentage from segment category percentage and dividing the result by total category percentage. The results are shown on the heat map in Figure 5.2. The blue colors shown on the heat map indicate that the shopping category percentage for a segment deviates from the overall shopping category percentage in a positive direction, whereas the red colors indicate a deviation in the negative direction.

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total
Clothing and Accessory	47.40%	80.03%	47.12%	52.46%	54.06%
Construction Materials, Hardware Store	1.34%	0.28%	1.06%	0.44%	0.92%
Cosmetics	7.70%	3.39%	6.11%	1.01%	5.35%
Education/ Stationery/ Office Equipment	4.55%	1.46%	2.71%	0.38%	2.74%
Electronic Appliance, Computer	5.07%	2.44%	6.96%	2.05%	4.71%
Entertainment	0.46%	0.08%	0.36%	0.13%	0.31%
Food	7.74%	2.61%	4.50%	2.15%	4.89%
Furniture and Decoration	10.21%	4.71%	7.30%	3.37%	7.19%
Goldsmiths	0.36%	0.12%	0.27%	0.18%	0.26%
Health / Healthcare Products	2.50%	0.93%	1.97%	0.29%	1.69%
Supermarket	10.86%	3.45%	19.92%	35.43%	16.29%
Service Sectors	0.26%	0.08%	0.15%	0.07%	0.16%
Telecommunications	0.11%	0.04%	0.26%	0.16%	0.15%
Various Food	1.46%	0.37%	1.32%	1.89%	1.28%

Table 5.7: Percentage of Total Transaction Count by Shopping Categories

It can be observed that Category Hunters mainly target several shopping malls, purchase clothing and accessory products that constitute 80.03% of their transactions and have a variation size of 0.48. The other categories apart from clothing and accessory stay below the total distribution. These people do not adhere to specific shopping malls and search for clothing and accessory related products in various locations.

The two main purposes of Mission Shoppers in visiting shopping malls are clothing and accessory purchases as well as supermarket purchases that constitute 52.46% and 35.43% of all their transactions, respectively. Their supermarket transactions have divergence size of 1.17, which is the highest value among all categories in all segments. Since they are more target-oriented people, they mostly acquire their household purchases in specific shopping malls in addition to the clothing and accessory purchases. They observe shopping as a have-to-do task rather than a joyful activity.

Recreational Shoppers' transactions are spread across all categories having positive divergence from total dispersion for most of the categories that can be seen in the Figure 5.2. They visit many shopping malls to buy various kinds of product groups.

They see shopping malls as places where they can socialize and entertain while buying some products.

Mall Squatters mainly visit some specific shopping malls which contain merchants of various categories and obtain most of their needs in the same places. In addition, the distribution of Mall Squatters transactions counts according to categories show similarity with the overall transaction distribution.

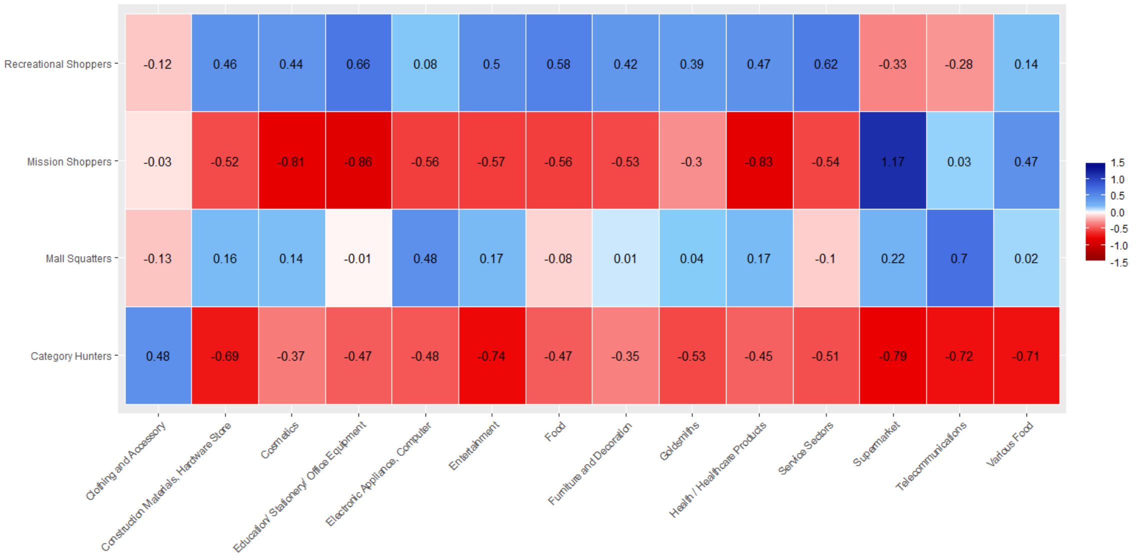

Figure 5.2: Category Dispersion of Diversity Based Segments

5.1.2 Loyalty Based Segments

Apart from diversity scores, in the second model we segment consumers according to their Shopping Mall Loyalty scores and Category Loyalty scores. The four clusters constructed can be seen in a scatter plot in Figure 5.3. One-way ANOVA test results in Table 5.8 show that the average values of each feature are statistically different across segments at significance level 0.05.

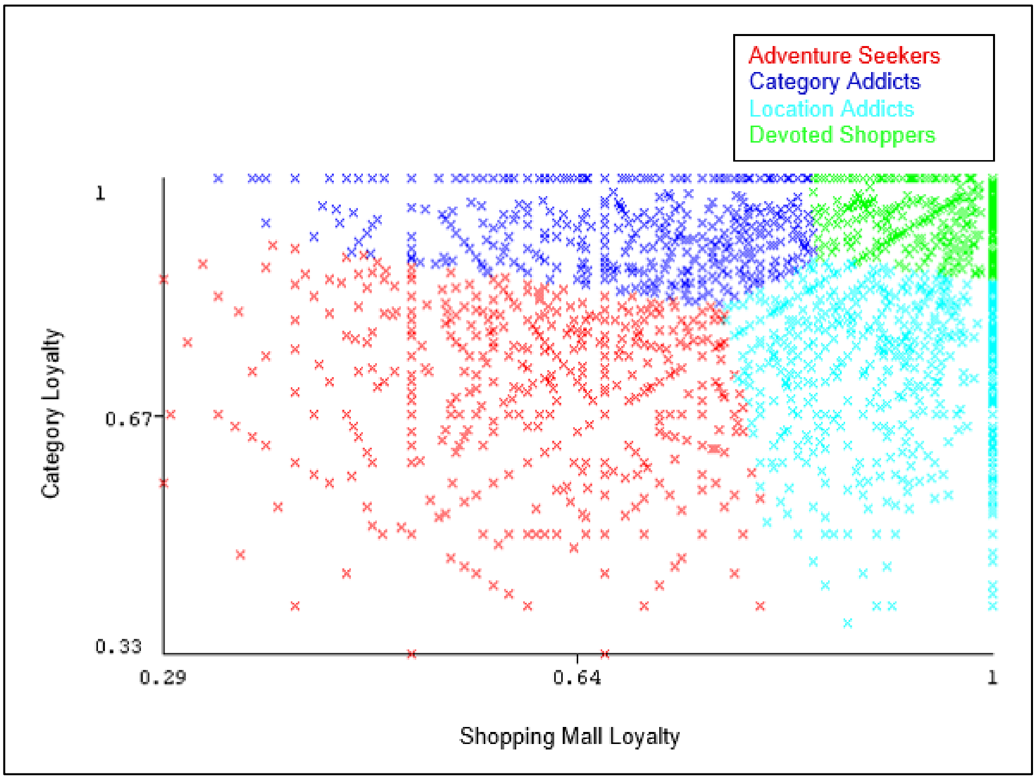

Figure 5.3: Loyalty Based Segments

We name the first segment of the four as Adventure Seekers. Post-hoc Tukey test results indicate that the consumers who fall into this segment have, on average, both low Shopping Mall Loyalty scores and low Category Loyalty scores. This segment comprises of 1845 customers, which is 12% of all customers. People who are part of this segment are not loyal to any specific shopping mall or specific shopping category. In other words, like the Recreational Shoppers, these people like to browse new places to shop as well as a variety of categories.

The second segment we name as Category Addicts. This segment comprises of 3130 customers, which corresponds to 21% of all customers. The people who are in this segment have, on average, low Shopping Mall Loyalty scores and high Category Loyalty scores. These people are not loyal to their two most preferred shopping malls and on the other hand, shop only for particular products that they are willing to purchase.

The third segment we name as Location Addicts. This segment comprises of 3095 customers, which is 21% of all customers. The people who are in this segment have, on average, high Shopping Mall Loyalty scores and low Category Loyalty scores. These are people who prefer to make their purchases in their preferred shopping malls for a variety of categories.

We name the last segment as Devoted Shoppers. This segment is comprised of 6773 customers, which is 46% of all customers. According to the Post-hoc Tukey test, people have, on average, high Shopping Mall Loyalty scores and high Category Loyalty scores. These people mostly shop in their two most preferred shopping malls and are willing to make purchases in their most preferred category groups in that shopping malls.

	Adventure Seekers	**Category Addicts**	**Location Addicts**	**Devoted Shoppers**	**Total**	**F**	**p-value**
Shopping Mall Loyalty Score	0.63	0.70	0.93	0.99	0.87	18764	0.000
Category Loyalty Score	0.72	0.99	0.75	0.98	0.90	15646	0.000

Table 5.8: Consumer Segments and Their Average Loyalty Scores

5.1.2.1 Demographic Profiles of Loyalty Based Segments

Table 5.9 shows the statistics of loyalty based segments' demographic profiles. Similar to the diversity based segment profile analysis, Chi-square test or one-way analysis of variance (ANOVA) test is conducted to find out statistically significant differences between consumer segments. Differences are accepted as statistically significant when p-value is less than 0.05.

Chi-square tests are performed to examine the relation between consumer segments and demographic features such as gender, marital status, education status and job type. The relation between gender and consumer segments are found to be significant, X^2 (3, N = 14843) = 205.17, p <.000. A significant relation between marital status and consumer segments is also found, X^2 (9, N = 14843) =75.105, p <.000. The relation between education status and consumer segments are found to be statistically significant, X^2 (24, N = 14843) =484.01 p <.000. On the other hand, there is no statistically significant relation between job type and consumer segments, X^2 (21, N = 14843) =23.352, p = .1774.

A one-way ANOVA is conducted to compare the differences in average age across consumer segments. Here we find a significant difference of age in consumer segments with the ANOVA analysis, F (3, 14839) = 11.970, p <.000. Post hoc comparisons using the Tukey test are also carried out and they reveal that Category Addicts are younger than Mall Addicts and Devoted Shoppers.

Pairwise comparison of proportions test shows that the proportion of females, which equals to 0.553 in Category Addicts is significantly the largest among all segments. On the other hand, the proportion of males in Devoted Shoppers, which equals to 0.59 is also significantly the largest among all segments.

Pairwise comparison of proportions test further indicates that the proportion of married people in Devoted Shoppers and Location Addicts are significantly larger compared to that in Category Addicts and Adventure Seekers segments. On the other hand, the proportion of single people in Category Addicts, which is 0.285, is significantly largest among all segments.

We further generate two groups named as low education level and high education level in order to test the significance of proportion for the education attribute. Low education level comprises of uneducated, primary school, middle school, and high school, whereas high education level comprises of college, university, Master, and PhD. Our results show that, Adventure Seekers have the highest education level with high education proportion of 0.69 compared to other segments. On the other hand, Devoted Shoppes has the lowest education level with low education proportion of 0.54.

In Appendix B, additional charts associated with segment properties can be found.

	Adventure Seekers	**Category Addicts**	**Location Addicts**	**Devoted Shoppers**	**Total**	**Test Statistic**	**p-value**
Gender						X^2	
Male	48.0%	44.7%	52.2%	59.0%	53.2%		
Female	52.0%	55.3%	47.8%	41.0%	46.8%	205.17	0.000
Marital Status						X^2	
Married	66.6%	63.1%	69.7%	70.2%	68.1%		
Single	25.9%	28.5%	21.6%	22.0%	23.8%		
Divorced	5.4%	5.7%	5.7%	4.8%	5.2%		
Unknown	2.2%	2.7%	2.9%	3.1%	2.8%	75.105	0.000
Age Group						F	
19-30	21.2%	26.4%	20.3%	20.8%	21.9%		
31-42	50.0%	44.5%	47.9%	46.9%	47.0%		
43-54	22.7%	23.2%	25.2%	25.4%	24.5%		
55-66	5.4%	5.6%	6.0%	6.1%	5.9%		
67-83	0.7%	0.4%	0.7%	0.8%	0.7%	11.970	0.000
Education Status						X^2	
Primary School	1.0%	2.2%	2.6%	4.0%	3.0%		
Middle School	2.4%	4.0%	4.2%	5.7%	4.6%		
High School	26.1%	33.8%	34.6%	43.3%	37.3%		

College	9.4%	9.8%	11.6%	9.1%	9.8%		
University	52.4%	43.2%	40.3%	32.3%	38.8%		
Master	6.9%	5.9%	5.3%	3.8%	4.9%		
PhD	0.7%	0.3%	0.5%	0.4%	0.4%		
Uneducated	1.0%	0.8%	0.8%	1.3%	1.1%		
Unknown	0.1%	0.1%	0.1%	0.0%	0.1%	484.01	0.000
Job Type						X^2	
Private Sector Employee	68.6%	68.1%	70.0%	69.9%	69.4%		
Public Servant	9.9%	9.6%	8.7%	7.8%	8.6%		
Self-Employed	11.5%	12.3%	11.1%	12.4%	12.0%		
Retiree	7.2%	6.9%	7.6%	7.1%	7.1%		
Housewife	1.5%	1.5%	1.5%	1.4%	1.4%		
Non-Employed	0.5%	0.7%	0.4%	0.7%	0.6%		
Other	0.9%	0.9%	0.8%	0.7%	0.8%	23.352	0.1774

Table 5.9: Demographic Properties of Loyalty Based Segments

5.1.2.2 Transactional Characteristics of Loyalty Based Segments

Table 5.10 shows the average transaction amounts per customer of each loyalty based consumer segment. A one-way ANOVA test is conducted to compare the difference between average transaction amounts per customer across all four loyalty based segments. The result shows that the difference is statistically significant, $F(3, 14839) = 6.365$, $p < .000$. Afterwards, post hoc Tukey test is carried out, which shows that Devoted Shoppers and Location Addicts are differed significantly at $p < .05$. Devoted Shoppers have the highest average transaction amount and Location Addicts have the lowest average transaction amount.

	Adventure Seekers	Category Addicts	Location Addicts	Devoted Shoppers	Total	F	p-value
Average Transaction Amount per Customer	153.4	163.1	145.7	178	165.08	6.365	0.000

Table 5.10: Average Transaction Amount per Customer of Loyalty Based Segments

We further calculate the average transaction count per customer of each segment. The results can be seen in Table 5.11. The results of ANOVA test show that the average transaction count of customers is significantly different among customer segments, F (3, 14839) = 43.6, p < .000. The post hoc Tukey results indicates that Location Addicts have relatively more transaction counts in shopping malls compared to other segments. On the other hand, Devoted Shoppers have the least average transaction count compared to other segments.

	Adventure Seekers	**Category Addicts**	**Location Addicts**	**Devoted Shoppers**	**Total**	**F**	**p-value**
Average Transaction Count per Customer	11.2	9.7	12.3	9.1	10.2	43.6	0.000

Table 5.11: Average Transaction Count per Customer of Loyalty Based Segments

We also calculate the average of total transaction counts per customer in each consumer segment. The results of ANOVA test show that the total transaction count per customer is significantly different across customer segments, F (3, 14839) = 3.606, p < .000. The post hoc Tukey results indicate that Devoted Shoppers have relatively less total transaction amounts compared to other segments.

	Adventure Seekers	**Category Addicts**	**Location Addicts**	**Devoted Shoppers**	**Total**	**F**	**p-value**
Total Transaction Amount per Customer	1538.0	1535.5	1524.8	1325.5	1437.8	3.606	0.000

Table 5.12: Total Transaction Amount per Customer of Loyalty Based Segments

Finally, the ANOVA tests show that there are significant differences between consumer segments in terms of their Shopping Day Diversity and Shopping Day Loyalty scores. Findings in Table 5.13 show that Adventure Seekers prefer to shop in diversified days of the week. Since their visit days are varied, they are more willing to see shopping as a leisure activity. On the other hand, Devoted Shoppers have high Shopping Day Loyalty scores. They are more likely to shop on their most preferred two days of the week compared to the other segments.

	Adventure Seekers	Category Addicts	Location Addicts	Devoted Shoppers	Total	F	p-value
Shopping Day Diversity	0.66	0.62	0.64	0.54	0.59	260.2	0.000
Shopping Day Loyalty	0.68	0.70	0.69	0.76	0.72	189.8	0.000

Table 5.13: Shopping Day Behavioral Scores of Loyalty Based Segments

5.1.2.3 Shopping Category Profile of Loyalty Based Segments

In Table 5.14 we report the percentage of total transaction counts according to the shopping categories for each segment. As we do in diversity based segments, the variation of each segments' shopping category from total shopping category is calculated by subtracting total category percentage from segment category percentage and dividing the result by total category percentage. The results are shown on the heat map in Figure 5.4.

	Adventure Seeking Shoppers	Category Addicts	Location Addicts	Devoted Shoppers	Total
Clothing and Accessory	46.68%	73.28%	43.91%	53.35%	54.06%
Construction Materials, Hardware Store	1.56%	0.36%	1.44%	0.65%	0.92%
Cosmetics	7.98%	4.29%	7.99%	3.34%	5.35%
Education/ Stationery/ Office Equipment	5.31%	1.80%	4.55%	1.22%	2.74%
Electronic Appliance, Computer	5.56%	2.78%	6.84%	4.04%	4.71%
Entertainment	0.44%	0.17%	0.43%	0.25%	0.31%
Food	8.54%	3.83%	6.56%	3.16%	4.89%
Furniture and Decoration	11.10%	6.27%	8.58%	5.48%	7.19%
Goldsmiths	0.34%	0.15%	0.48%	0.15%	0.26%
Health / Healthcare Products	2.78%	1.16%	2.74%	0.95%	1.69%
Supermarket	7.54%	5.24%	14.50%	25.79%	16.29%
Service Sectors	0.33%	0.10%	0.24%	0.09%	0.16%
Telecommunications	0.10%	0.06%	0.19%	0.19%	0.15%
Various Food	1.75%	0.51%	1.57%	1.34%	1.28%

Table 5.14: Percentage of Total Transaction Counts by Shopping Categories

Category Addicts, not being loyal to specific shopping malls, mainly search for clothing and accessory products, which constitutes 73.28% of their total transactions and has a variation size of 0.36 in other malls. Only clothing and accessory category has a positive variation in this segment, the other categories remain below the total distribution.

Devoted Shoppers are addicted to the same shopping places and mainly two shopping categories: clothing and accessory purchases, and supermarket purchases, which constitute 53.35% and 25.79% of all of their transactions, respectively. Their supermarket transactions have divergence size of 0.58, which is the highest supermarket variation value from total dispersion among all segments. They visit their two most preferred shopping malls to visit supermarket and clothing and accessory stores.

Adventure Seekers transactions are spread across all categories with positive divergence from total dispersion for most of the categories. They visit many shopping malls to buy various kinds of product groups. Service sectors have a variation size of 1.06 which is the highest value among all categories in all segments, from the total dispersion. In other words, Service sector purchases are made in large proportions by the Adventure Seekers compared to other segments.

Location Addicts are loyal to specific shopping malls with merchants of various categories and obtain most of their needs in the same places. Their transactions are diversified to various category groups.

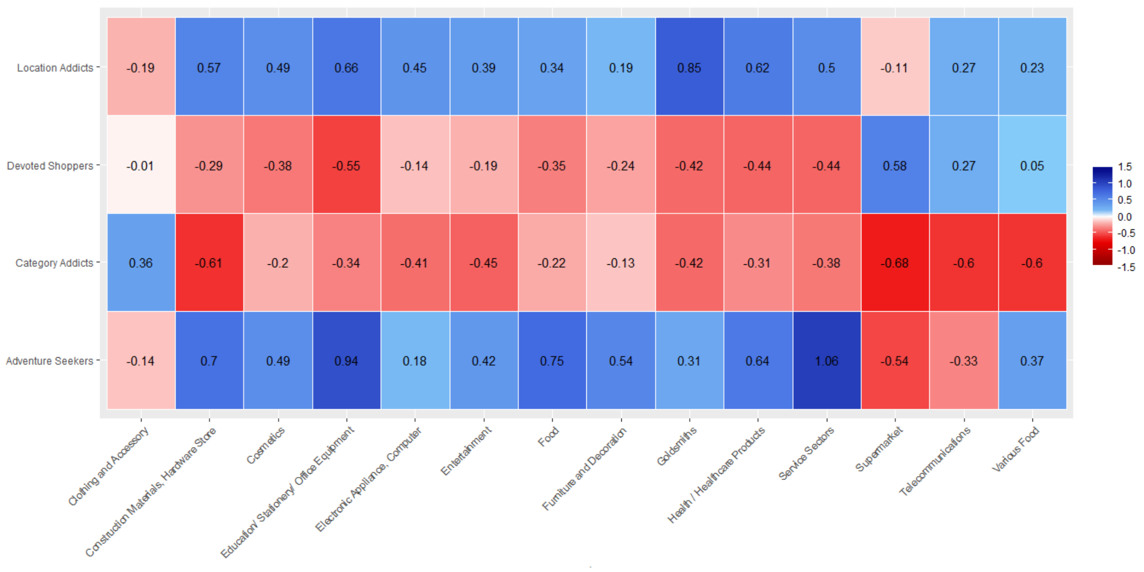

Figure 5.4: Category Dispersion of Loyalty Based Segments

5.1.3 Relationship Between Diversity and Loyalty Based Segments

In order to understand the relationship between diversity and loyalty features, the correlation coefficient between Category Loyalty and Category Diversity ($p = .00$, $r = -.85$) and also between Shopping Mall Loyalty and Shopping Mall Diversity ($p = .00$, $r =$

-.88) features are calculated. The results indicate that there is a strong negative correlation between diversity and loyalty for both shopping mall and category. The strong negative correlation between diversity and loyalty and having similarities in demographic, transactional and categorical profiles among diversity and loyalty based segments may suggest that Recreational Shoppers and Adventure Shoppers, Category Hunters and Category Addicts, Mall Squatters and Location Addicts, Mission Shoppers and Devoted Shoppers are equivalent segments. However, when we analyze the transition of individual customers between segments considering the two models, we observe that 60% of the customers stay in the anticipated segment in the other model and the 40% of the customers move to other segments. The transitions on flow chart and numerical results can be seen in the Figure 5.5 and Table 5.14 respectively.

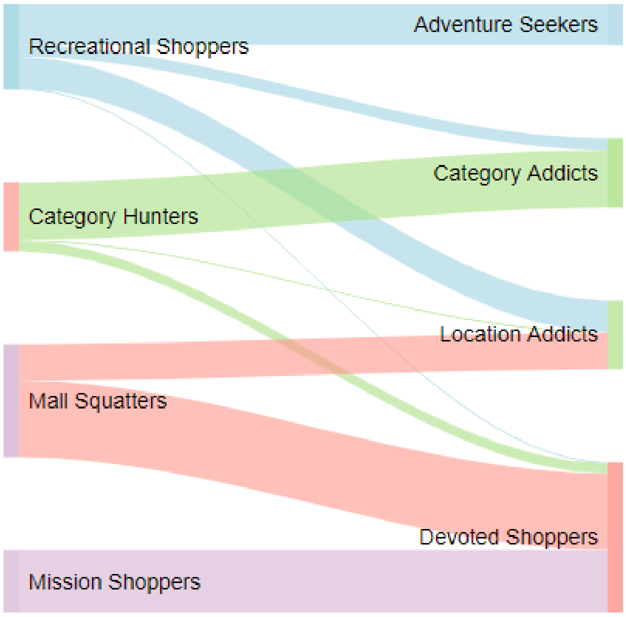

Figure 5.5: Transition Among Diversity and Loyalty Based Segments

From Figure 5.5 and Table 5.14, we observe that the largest number of transitions are between Recreational Shoppers and Adventure Seekers, and also Mall Squatters and Devoted Shoppers. In the Loyalty based clustering model, most of the customers are gathered in Devoted Shoppers segment, while Mission Shoppers are the smallest segment. As a result, the analysis shows that the segments constructed with two different behavioral features are not identical in composition.

	Adventure Seekers	Category Addicts	Location Addicts	Devoted Shoppers	Total
Recreational Shoppers	1827	536	1414	49	**3826**
Category Hunters	18	2591	43	462	**3114**
Mall Squatters		3	1638	3448	**5089**
Mission Shoppers				2814	**2814**
Total	**1845**	**3130**	**3095**	**6773**	**14843**

Table 5.15: Transition Among Diversity and Loyalty Based Segments

5.1.4 Silhouette Analysis

In our study, we use silhouette analysis to measure how the constructed models are well separated. Here we measure the similarity of an observation to its assigned cluster compared to neighboring clusters. The value ranges from −1 to +1, when a value close to 1 shows that the observation is well matched to the assigned cluster and a value close to -1 indicates that a neighboring cluster of an observation is much similar compared to its assigned cluster. The average silhouette width which is computed by taking the average silhouette values of all constituted clusters provides an assessment of clustering validity (Rousseeuw, 1986). Figure 5.6 and Figure 5.7 show the plot of silhouette for diversity and loyalty segments.

The average silhouette width of Mission Shoppers is the minimum among all segments, which means this segment is not as separated as the others. The average silhouette with of the diversity based model is 0.4.

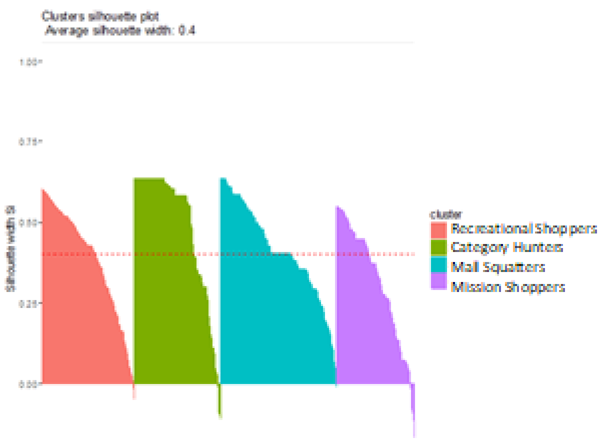

Figure 5.6: Silhouette Plot of Diversity Based Segments

The average silhouette width of Devoted Shoppers is the maximum with a value of 0.80 compared to other segments, which indicates that the cluster is well separated. The average silhouette with of the loyalty based model is 0.56. When we compare the average silhouette width of the two models, loyalty based model is much more well separated compared to diversity based model.

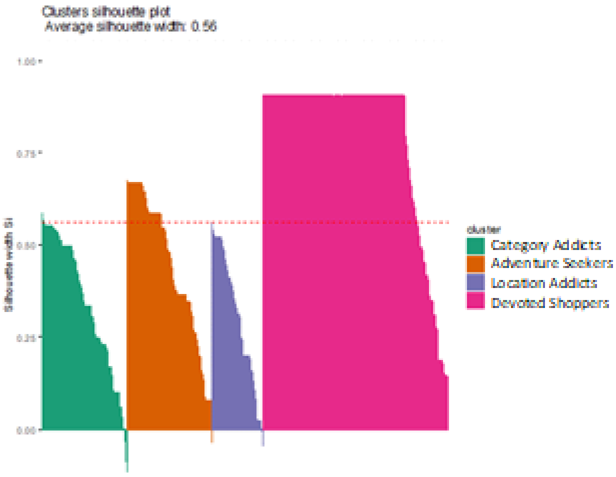

Figure 5.7: Silhouette Plot of Loyalty Based Segments

5.1.5 Summary of Segments

We approach our research target by constituting two clustering models containing diversity or loyalty scores of each customer. Our findings through segments give insight on peoples' variety seeking behavior in shopping in the context of shopping malls and shopping categories.

For the diversity dimension, we constructed four different segments: Recreational Shoppers, Category Hunters, Mall Squatters, and Mission Shoppers. We can infer from the Category Hunters segment that young and single women are more likely to shop for mainly clothing and accessory products in various shopping malls. Mission Shoppers mainly comprise of low educated married men who prefer to perform their purchases in few numbers of shopping malls and they mainly go shopping malls for supermarket, and clothing and accessory related products. In addition to that, they make on average fewer number of purchases and spend less money throughout the year compared to other segments. Also, they are more likely to make purchases in specific days of the week. We come with a notion that these people perceive shopping as a time and energy consuming event and they make need-driven purchases. Our findings with Mall Squatters show that married and male customers mostly do their purchases in the specific malls where they can find a variety of shopping categories. Again, their time and effort consumptions are low while searching for various category types, since they are more willing to shop in malls which contain a variety of merchant types. Besides, compared to Mission Shoppers the male consumers in this segment search a lot more for categories so we can say that they have bigger intentions to perceive shopping as an enjoyable activity. In addition, there is a demographic difference between these segments; the education level of Mall Addicts is higher than that of Mission Shoppers. This situation explains that generalizing shopping behavior only for a particular gender is not truly effective since the position of men and women in society changes and moves away from traditional perceptions. Lastly, Recreational Shoppers consist of educated females who prefer to shop in various shopping malls and for various categories. These people have the largest number of purchases on average and the largest total transaction amount across all segments. In addition, they go shopping malls on various days, which means they do not set aside specific days of the week for shopping and are generally more flexible for choosing a shopping day. Our findings suggest that these people are more likely to see shopping as a leisure time, social and recreational activity, and so they enjoy the shopping experience.

For the loyalty, we constitute four different segments: Adventure Seekers, Category Addicts, Location Addicts, and Devoted Shoppers. Our findings in these four segments show some similarity with diversity based segment characteristics. Category Addicts mainly consist of young and single female shoppers who are willing to make purchases at a variety of shopping malls. Their main consideration for shopping is buying clothing and accessory products. Our findings indicate that low educated and married men who appear in Devoted Shoppers prefer to make purchases on only their most visited shopping malls and in their most preferred categories: clothing and accessory, and supermarket. They both have the smallest number of purchases on average and the smallest total transaction amount spent in a one-year period in comparison to other segments. These people visit fewer number of shopping malls and shop for fewer categories which compose basically their needs. Therefore, it can be said that people who are part of this segment perceive shopping malls as a place where they can meet their requirements without entertaining the shopping experience and wasting time. Location addicts prefer to shop in their favorite shopping malls that offer a variety of category alternatives for their needs. The proportion of males are slightly higher than females in this segment and they are mainly married. In addition, they have the largest number of transaction counts on average compared to other segments. Adventure Seekers are comprised of highly educated female shoppers. They are more likely to visit new shopping malls and shop for diversified categories. Besides, they are not loyal to specific day(s) of the week for mall shopping and spend the largest amount of money over the one year period compared to other segments. We conclude that Adventure Seekers are the consumers that perceive shopping as an entertainment and leisure time activity.

Although we have some similar characteristics in diversity and loyalty segments, the distribution of people shows remarkable shifts of customers between segments corresponding to other feature as we discussed in section 5.1.3. However, our results show that demographic profiles show dissimilarity while distinguishing both diversity and loyalty segments. In addition, the segments of diversity and loyalty that are thought to be matching show appreciable similarities in demographic profiles.

5.2 Consumer Segments Prediction

We randomly split our data into 70% for the training set and 30% for the test set. We do this splitting process three times to generate three different train and test sets in order to validate our model. Then we predict which segment each customer will be placed in by cosine similarity of the demographic profiles. After that, we measure the prediction performance by calculating the matching rate of the segment assignments using cosine similarity with the ground truth.

Our results are presented in Figure 5.8. Since we have four clusters both in diversity and loyalty based models, our baseline is 25%. The results indicate that with only the demographic information of customers at hand, we can predict his or her segment with up to 33% accuracy and on average with 30% accuracy in diversity based models and with up to 36% accuracy and on average with 34% accuracy in loyalty based models. Therefore, we can predict better with the help of demographic knowledge compared to randomly assigning customers into segments. Our results show that demographic information of individuals is somewhat indicative of their shopping behavior.

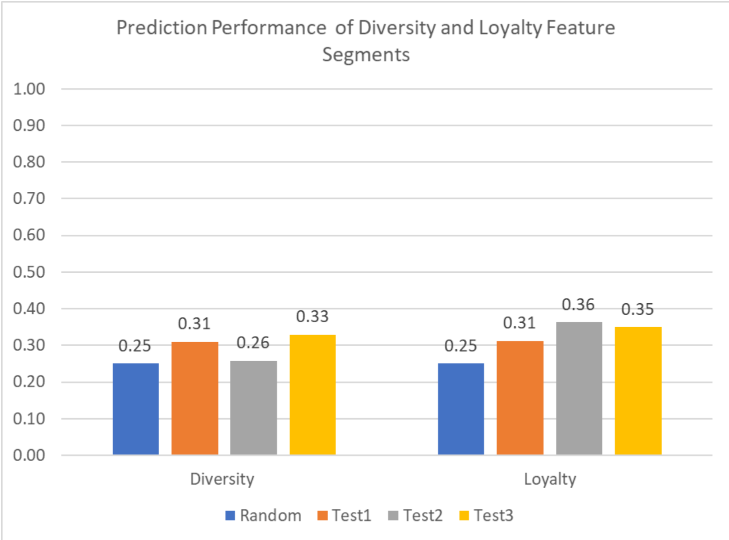

Figure 5.8: Prediction Performance of Diversity and Loyalty Based Segments

According to our segment assignments, assigning a customer to either one of the three wrong segments has the same impact. Our model does not take into consideration the opportunity costs and penalty costs associated with each misclassified segment. For instance, assigning a Mission Shopper to the Recreational Shoppers segment and treating him or her the same way as Recreational Shoppers by sending irrelevant campaigns on various category groups and services, or new shopping places that have opened, can be inappropriate and may have a negative impact leading to a negative “cost” for the company. On the other hand, assigning a Recreational Shopper to the Mission Shoppers segment may cause him or her to miss the relevant direct marketing campaigns and hence the opportunities of selling goods. In this study, we do not cover the subject that involves the differentiated costs of misclassification and we leave it for a future work.

CHAPTER 6

CONCLUSION

Understanding consumers' shopping behavior is important for marketers to develop campaigns and promotions relevant to the target group. In this study, we aim to identify the variety seeking behaviors of customers in the context of both shopping malls and shopping categories. In order to do this, we extract and calculate behavioral features for each customer such as diversity and loyalty from a large credit card transaction dataset provided by one of the largest Turkish banks.

Our approach includes two different types of clustering models containing features of diversity or loyalty. Using this approach, we construct four segments, which give insights about customers' shopping attitudes of variety seeking and shopping experiences, from both diversity and loyalty perspective. When we analyze the demographic profiles of segments, we obtain consistency with the gender differences in shopping attitudes discussed in the literature as women perceive shopping as a leisure time activity, social and entertaining experience and they seek variety of assortments whereas men perceive shopping as a time and energy consuming activity and prefer to purchase their need as soon as possible. On the other hand, some studies in the literature discuss that considering only gender differences will not be realistic and the gap between difference have started to decrease since the gender roles in the society are changing with the social and economic influences. Although we find our results to support the findings about gender difference in literature, we find the impact of demographic features on some group of customers. In our study, we discover a segment closely related to this change, which is comprised of educated males who prefer to go specific shopping malls that offer a variety of category alternatives to satisfy their needs. Our results show that with the increasing education level, men start to much more enjoy the shopping experience. In addition, we introduce a segment of customers, not previously reported in the literature, composed of young females who are willing to seek a variety of shopping malls for only clothing and accessory products.

Apart from the presented results, we propose a classification method that assigns customers into one of the four segments according to their demographic properties in relation to the demographic properties of the segments. We assign the individual into the segment which has the most demographic similarity with him or her. Our classification results show that we can assign the customer into the right segment with up to 36% accuracy which is higher than random guessing. This result implies that there is an association between customers' demographics and their shopping behaviors as well in the setting of shopping malls.

In conclusion, we hope that our research will guide the marketers to develop the right set of actions and approach the right group of customers with the right strategy.

BIBLIOGRAPHY

Alreck, P., & Settle, R. B. (2002). Gender effects on Internet, catalogue and store shopping. *Journal of Database Marketing & Customer Strategy Management, 9*(2), 150-162. doi:10.1057/palgrave.jdm.3240071.

Bakewell, C., & Mitchell, V. (2004). Male Consumer Decision-Making Styles. *The International Review of Retail, Distribution and Consumer Research, 14*(2), 223-240. doi:10.1080/0959396042000178205.

Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications, 105*(9).

Di Clemente R, Luengo-Oroz M, Travizano M, Vaitla B, Gonzalez MC (2017) Sequence of purchases in credit card data reveal life styles in urban populations. arXiv:1703.00409.

Dong, X., Suhara, Y., Bozkaya, B., Singh, V. K., Lepri, B., & Pentland, A. ‘. (2017). Social Bridges in Urban Purchase Behavior. *ACM Transactions on Intelligent Systems and Technology, 9*(3), 1-29. doi:10.1145/3149409.

Eagle, N., Macy, M., and Claxton, R. (2010). Network diversity and economic development. Science, 328(5981):1029-1031.

Evans, K. R., Christiansen, T., & Gill, J. D. (1996). The Impact of Social Influence and Role Expectations on Shopping Center Patronage Intentions. Journal of the Academy of Marketing Science, 24(3), 208-218. doi:10.1177/0092070396243002.

Grewal, D., Baker, J., Levy, M., & Voss, G. B. (2003). The effects of wait expectations and store atmosphere evaluations on patronage intentions in service-intensive retail stores. *Journal of Retailing*, 79(4), 259-268. doi:10.1016/j.jretai.2003.09.006.

Guidotti, R., Gabrielli, L., Monreale, A., Pedreschi, D., & Giannotti, F. (2018). Discovering temporal regularities in retail customers' shopping behavior. EPJ Data Science, 7(1). doi:10.1140/epjds/s13688-018-0133-0.

Hackett, P. M., & Foxall, G. R. (1994). A factor analytic study of consumers location specific values: A traditional high street and a modern shopping mall. *Journal of Marketing Management, 10*(1-3), 163-178. doi:10.1080/0267257x.1994.9964266.

Haj-Salem, N., Chebat, J. C., Michon, R., & Oliveira, S. (2016). Why male and female shoppers do not see mall loyalty through the same lens? The mediating role of self-congruity. *Journal of Business Research*, *69*(3), 1219-1227. doi: 10.1016/j.jbusres.2015.09.011.

Hantula, D. A. (2003). Guest editorial: Evolutionary psychology and consumption. *Psychology and Marketing*, 20(9), 757-763. doi:10.1002/mar.10095.

Hart, C., Farrell, A. M., Stachow, G., Reed, G., & Cadogan, J. W. (2007). Enjoyment of the Shopping Experience: Impact on Customers Repatronage Intentions and Gender Influence. *The Service Industries Journal*, 27(5), 583-604. doi:10.1080/02642060701411757.

Hoch, Stephen J., Eric T. Bradlow and Brian Wansink (1999). "The Variety of an Assortment," *Marketing Science*, 18 (4), 527–546.

Hozier, George C. and Donald E. Stem (1985). "General Retail Patronage Loyalty as a Determinant of Consumer Outshopping Behavior," *Journal of the Academy of Marketing Science*, 13 (Winter), 32–46.

Huff, D. L. (1964). Defining and Estimating a Trading Area. *Journal of Marketing, 28*(3), 34. doi:10.2307/1249154.

Jump, Teresa L., and Linda Haas. (1987). "Dual-Career Fathers Participating in Child Care," Pp. 98–114 in Changing Men. Michael S. Kimmel (ed). *Newbury Park*, CA: Sage.

Kruger, D., & Byker, D. (2009). Evolved foraging psychology underlies sex differences in shopping experiences and behaviors. *Journal of Social, Evolutionary, and Cultural Psychology, 3*(4), 328-342. doi:10.1037/h0099312.

Krumme, C., Llorente, A., Cebrian, M., Pentland, A., & Moro, E. (2013). The predictability of consumer visitation patterns. Scientific Reports, 3(1). doi:10.1038/srep01645.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281297. Oakland, CA, USA.

Mitchell, V., & Walsh, G. (2004). Gender differences in German consumer decision-making styles. *Journal of Consumer Behavior*, 3(4), 331-346. doi:10.1002/cb.146.

Noble, S. M., Griffith, D. A., & Adjei, M. T. (2006). Drivers of local merchant loyalty: Understanding the influence of gender and shopping motives. *Journal of Retailing*, 82(3), 177-188. doi:10.1016/j.jretai.2006.05.002.

Otnes, C., & Mcgrath, M. A. (2001). Perceptions and realities of male shopping behavior. *Journal of Retailing*, 77(1), 111-137. doi:10.1016/s0022-4359(00)00047-6.

Population of Cities in Turkey (2018). (n.d.). Retrieved April 10, 2018, from

http://worldpopulationreview.com/countries/turkey-population/cities/.

Raajpoot, N. A., Sharma, A., & Chebat, J. (2008). The role of gender and work status in shopping center patronage. *Journal of Business Research*, 61(8), 825-833. doi:10.1016/j.jbusres.2007.09.009.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65. doi:10.1016/0377-0427(87)90125-7.

Ruiz, J., Chebat, J., & Hansen, P. (2004). Another trip to the mall: A segmentation study of customers based on their activities. *Journal of Retailing and Consumer Services*, 11(6), 333-350. doi: 10.1016/j.jretconser.2003.12.002.

Singh, V. K., Freeman, L., Lepri, B., & Pentland, A. S. (2013). Predicting Spending Behavior Using Socio-mobile Features. *2013 International Conference on Social Computing*. doi:10.1109/socialcom.2013.33.

Singh, V. K., Bozkaya, B., & Pentland, A. (2015). Money Walks: Implicit Mobility Behavior and Financial Well-Being. *Plos One, 10*(8). doi: 10.1371/journal.pone.0136628.

Sit, J., Merrilees, B., & Birch, D. (2003). Entertainment-seeking shopping centre patrons: The missing segments. *International Journal of Retail & Distribution Management, 31*(2), 80-94. doi:10.1108/09590550310461985.

Song, C., Qu, Z., Blumm, N., and Barabasi, A. L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018_1021.

Srivastava, K. A., Singh, V. K., Bozkaya, B., & Pentland, A. (2014) ". (n.d.). Assessing Financial Well-Being of Merchants by Analyzing Behavioral Patterns in Historical Transactions. *Geo-Intelligence and Visualization through Big Data Trends Advances in Geospatial Technologies*, 76-93. doi:10.4018/978-1-4666-8465-2.ch003.

Teller, C., & Thomson, J. A. (2012). Gender differences of shoppers in the marketing and management of retail agglomerations. *The Service Industries Journal*, 32(6), 961-980. doi:10.1080/02642069.2011.559725.

Wen, Y., Yeh, P., Tsai, T., Peng, W., & Shuai, H. (2018). Customer Purchase Behavior Prediction from Payment Datasets. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM 18*. doi:10.1145/3159652.3159707.

Yoshimura, Y., Sobolevsky, S., Hobin, J. N., Ratti, C., & Blat, J. (2016). Urban association rules: Uncovering linked trips for shopping behavior. Environment and Planning B: *Urban Analytics and City Science*, 45(2), 367-385. doi:10.1177/0265813516676487.

Zeithaml, V. A. (1985). The New Demographics and Market Fragmentation. *Journal of Marketing*,*49*(3), 64. doi:10.2307/1251616.

Appendix A: Center and Customer Demographics Information about Train Set 1, Train Set 2 and Train Set 3

Table A.1: Centers of Diversity Train Set 1 Clusters

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total	F	p-value
Shopping Mall Diversity Score	0.30	0.21	0.11	0.02	0.17	8923	0.000
Category Diversity Score	0.42	0.16	0.38	0.09	0.26	7175	0.000

Table A.2: Customer Demographics Profile of Diversity Train Set 1 Clusters

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total	Test Statistic	p-value
Gender						X^2	
Male	53.2%	58.5%	61.1%	69.4%	59.9%		
Female	46.8%	41.5%	38.9%	30.6%	40.1%	133.52	0.000
Marital Status						X^2	
Married	73.0%	74.3%	78.0%	78.3%	75.5%		
Single	18.4%	18.0%	14.0%	15.3%	16.8%		
Divorced	6.6%	5.8%	5.7%	4.3%	5.7%		
Unknown	2.0%	1.9%	2.3%	2.1%	2.0%	38.747	0.000
Age Group						F	
19-30	9.9%	12.9%	10.2%	8.9%	10.8%		
31-42	53.7%	49.5%	50.9%	51.5%	51.2%		
43-54	28.5%	29.2%	29.8%	31.5%	29.6%		
55-66	7.3%	7.6%	7.9%	7.2%	7.5%		
67-83	0.7%	0.9%	1.3%	1.0%	0.9%	3.804	0.009
Education Status						X^2	
Primary School	1.2%	2.9%	3.3%	5.2%	3.0%		
Middle School	2.9%	5.8%	5.4%	6.4%	5.1%		
High School	29.4%	38.3%	37.3%	48.2%	37.9%		
College	9.8%	9.8%	11.0%	8.4%	9.8%		
University	47.1%	36.6%	36.0%	26.6%	37.1%		
Master	8.1%	4.6%	5.3%	3.3%	5.4%		
PhD	0.7%	0.5%	0.6%	0.4%	0.5%		
Uneducated	1.0%	1.4%	1.0%	1.5%	1.2%		
Unknown	0.0%	0.1%	0.1%	0.0%	0.1%	416.74	0.000
Job Type						X^2	
Private Sector Employee	63.9%	63.4%	64.1%	65.5%	64.1%		
Public Servant	9.9%	9.2%	10.2%	8.3%	9.4%		
Self-Employed	13.4%	15.6%	13.1%	14.4%	14.3%		
Retiree	9.6%	8.6%	9.6%	8.9%	9.1%		
Housewife	1.5%	1.7%	1.4%	1.3%	1.5%		
Non-Employed	0.6%	0.5%	0.7%	0.8%	0.6%		
Other	1.2%	1.0%	1.0%	0.7%	1.0%	22.053	0.2296

Table A.3: Centers of Loyalty Train Set 1 Clusters

	Adventure Seekers	**Category Addicts**	**Location Addicts**	**Devoted Shoppers**	**Total**	**F**	**p-value**
Shopping Mall Loyalty Score	0.71	0.70	0.92	0.99	0.87	13051	0.000
Category Loyalty Score	0.63	0.97	0.75	0.98	0.90	10863	0.000

Table A.4: Customer Demographics Profile of Loyalty Train Set 1 Clusters

	Adventure Seekers	**Category Addicts**	**Mall Addicts**	**Devoted Shoppers**	**Total**	**Test Statistic**	**p-value**
Gender						X^2	
Male	54.1%	52.0%	58.6%	65.8%	59.9%		
Female	45.9%	48.0%	41.4%	34.2%	40.1%	143.900	0.000
Marital Status						X^2	
Married	72.8%	71.4%	76.5%	77.7%	75.5%		
Single	19.0%	20.2%	14.6%	15.5%	16.8%		
Divorced	6.3%	6.4%	6.6%	4.8%	5.7%		
Unknown	1.9%	2.0%	2.3%	1.9%	2.0%	52.719	0.000
Age Group						F	
19-30	9.8%	13.6%	9.8%	10.2%	10.8%		
31-42	54.8%	49.9%	51.1%	50.9%	51.2%		
43-54	27.9%	29.1%	30.3%	30.0%	29.6%		
55-66	6.6%	6.9%	7.9%	7.8%	7.5%		
67-83	0.9%	0.5%	0.9%	1.2%	0.9%	6.409	0.000
Education Status						X^2	
Primary School	1.0%	2.5%	2.6%	4.0%	3.0%		
Middle School	2.8%	4.8%	4.8%	6.0%	5.1%		
High School	27.4%	35.4%	34.8%	43.4%	37.9%		
College	9.8%	9.9%	11.0%	9.1%	9.8%		
University	49.2%	40.0%	39.3%	31.2%	37.1%		
Master	8.0%	6.1%	5.9%	4.0%	5.4%		
PhD	0.8%	0.4%	0.6%	0.4%	0.5%		
Uneducated	0.9%	0.8%	0.9%	1.7%	1.2%		
Unknown	0.1%	0.0%	0.1%	0.1%	0.1%	312.6	0.000
Job Type						X^2	
Private Sector Employee	64.8%	62.0%	64.3%	64.8%	64.1%		
Public Servant	9.9%	10.5%	9.8%	8.5%	9.4%		
Self-Employed	13.4%	15.1%	13.0%	14.8%	14.3%		
Retiree	8.8%	8.9%	9.9%	8.9%	9.1%		
Housewife	1.5%	1.7%	1.5%	1.4%	1.5%		
Non-Employed	0.5%	0.7%	0.5%	0.7%	0.6%		
Other	1.0%	1.0%	1.0%	0.9%	1.0%	25.56	0.1103

Table A.5: Centers of Diversity Train Set 2 Clusters

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total	F	p-value
Shopping Mall Diversity Score	0.30	0.21	0.10	0.01	0.17	8923	0.000
Category Diversity Score	0.42	0.15	0.39	0.09	0.25	7175	0.000

Table A.6: Customer Demographics Profile of Diversity Train Set 2 Clusters

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total	Test Statistic	p-value
Gender						X^2	
Male	40.5%	45.9%	49.1%	58.5%	47.6%		
Female	59.5%	54.1%	50.9%	41.5%	52.4%	133.52	0.000
Marital Status						X^2	
Married	60.8%	60.5%	66.2%	66.6%	62.8%		
Single	31.7%	30.4%	25.4%	25.1%	28.8%		
Divorced	4.6%	5.6%	4.6%	4.4%	4.9%		
Unknown	2.9%	3.5%	3.9%	3.9%	3.5%	38.747	0.000
Age Group						F	
19-30	31.9%	33.4%	28.0%	26.8%	30.8%		
31-42	48.7%	44.0%	48.8%	47.9%	46.8%		
43-54	16.1%	18.5%	19.0%	21.4%	18.6%		
55-66	3.2%	3.9%	3.7%	3.6%	3.6%		
67-83	0.2%	0.2%	0.5%	0.2%	0.2%	3.804	0.009
Education Status						X^2	
Primary School	1.1%	3.1%	3.4%	5.1%	3.1%		
Middle School	2.0%	4.0%	4.3%	5.7%	3.9%		
High School	27.4%	38.5%	39.0%	48.0%	37.7%		
College	9.8%	9.7%	11.5%	8.9%	9.9%		
University	51.8%	39.6%	36.9%	28.5%	40.0%		
Master	7.1%	4.2%	3.6%	3.0%	4.6%		
PhD	0.3%	0.3%	0.6%	0.3%	0.3%		
Uneducated	0.4%	0.5%	0.6%	0.3%	0.5%		
Unknown	0.1%	0.1%	0.1%	0.0%	0.1%	416.74	0.000
Job Type						X^2	
Private Sector Employee	75.1%	74.3%	75.9%	76.1%	75.2%		
Public Servant	9.4%	8.6%	6.8%	6.8%	8.1%		
Self-Employed	8.8%	10.0%	10.2%	10.3%	9.8%		
Retiree	3.9%	4.3%	4.8%	4.6%	4.4%		
Housewife	1.4%	1.5%	1.4%	1.0%	1.4%		
Non-Employed	0.5%	0.6%	0.4%	0.7%	0.5%		
Other	0.9%	0.6%	0.6%	0.4%	0.7%	22.053	0.2296

Table A.7: Centers of Loyalty Train Set 2 Clusters

	Adventure Seekers	Category Addicts	Location Addicts	Devoted Shoppers	Total	F	p-value
Shopping Mall Loyalty Score	0.65	0.69	0.94	0.99	0.87	13051	0.000
Category Loyalty Score	0.69	0.96	0.76	0.99	0.90	10863	0.000

Table A.8: Customer Demographics Profile of Loyalty Train Set 2 Clusters

	Adventure Seekers	Category Addicts	Location Addicts	Devoted Shoppers	Total	Test Statistic	p-value
Gender						X^2	
Male	43.4%	37.5%	47.2%	53.6%	47.6%		
Female	56.6%	62.5%	52.8%	46.4%	52.4%	143.900	0.000
Marital Status						X^2	
Married	61.9%	57.4%	64.9%	64.6%	62.8%		
Single	30.6%	33.9%	26.4%	26.9%	28.8%		
Divorced	4.7%	5.3%	4.9%	4.8%	4.9%		
Unknown	2.8%	3.4%	3.7%	3.6%	3.5%	52.719	0.000
Age Group						F	
19-30	30.7%	36.4%	28.7%	29.0%	30.8%		
31-42	49.3%	43.6%	48.6%	46.8%	46.8%		
643-54	16.9%	16.6%	18.8%	19.8%	18.6%		
55-66	2.9%	3.1%	3.6%	4.1%	3.6%		
67-83	0.2%	0.2%	0.3%	0.2%	0.2%	6.409	0.000
Education Status						X^2	
Primary School	1.2%	2.1%	2.7%	4.2%	3.1%		
Middle School	2.0%	3.2%	3.7%	4.8%	3.9%		
High School	27.1%	31.3%	35.3%	44.3%	37.7%		
College	10.0%	9.8%	11.7%	9.1%	9.9%		
University	52.7%	47.4%	41.0%	33.0%	40.0%		
Master	6.0%	5.3%	4.7%	3.8%	4.6%		
PhD	0.5%	0.3%	0.4%	0.3%	0.3%		
Uneducated	0.3%	0.6%	0.4%	0.5%	0.5%		
Unknown	0.2%	0.1%	0.1%	0.0%	0.1%	312.6	0.000
Job Type						X^2	
Private Sector Employee	75.3%	74.4%	75.6%	75.3%	75.2%		
Public Servant	9.2%	9.5%	7.8%	7.3%	8.1%		
Self-Employed	9.4%	9.2%	9.1%	10.5%	9.8%		
Retiree	3.3%	4.1%	4.9%	4.5%	4.4%		
Housewife	1.5%	1.4%	1.5%	1.3%	1.4%		
Non-Employed	0.5%	0.6%	0.3%	0.6%	0.5%		
Other	0.8%	0.8%	0.8%	0.5%	0.7%	18.503	0.4227

Table A.9: Centers of Diversity Train Set 3 Clusters

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total	F	p-value
Shopping Mall Diversity Score	0.29	0.27	0.10	0.06	0.18	7469	0.000
Category Diversity Score	0.44	0.16	0.30	0.02	0.25	9824	0.000

Table A.10: Customer Demographics Profile of Diversity Train Set 3 Clusters

	Recreational Shoppers	Category Hunters	Mall Squatters	Mission Shoppers	Total	Test Statistic	p-value
Gender						X^2	
Male	47.1%	45.0%	59.5%	61.1%	53.6%		
Female	52.9%	55.0%	40.5%	38.9%	46.4%	203.67	0.000
Marital Status						X^2	
Married	66.1%	63.0%	71.6%	70.8%	68.2%		
Single	25.9%	27.9%	21.0%	21.3%	23.8%		
Divorced	5.6%	6.0%	4.6%	4.9%	5.2%		
Unknown	2.4%	3.0%	2.7%	3.0%	2.8%	64.39	0.000
Age Group						F	
19-30	21.9%	26.6%	19.3%	21.3%	21.9%		
31-42	48.8%	43.3%	48.7%	46.0%	47.1%		
43-54	22.8%	24.2%	24.9%	26.3%	24.5%		
55-66	5.9%	5.5%	6.2%	5.5%	5.9%		
67-83	0.6%	0.4%	0.9%	0.9%	0.7%	7.555	0.000
Education Status						X^2	
Primary School	1.2%	2.6%	3.1%	5.1%	2.9%		
Middle School	2.5%	4.8%	5.1%	6.4%	4.6%		
High School	27.9%	35.0%	41.5%	44.4%	37.2%		
College	9.8%	9.9%	10.0%	9.1%	9.7%		
University	50.1%	41.6%	34.5%	29.6%	39.1%		
Master	7.1%	4.9%	3.9%	3.6%	4.9%		
PhD	0.5%	0.4%	0.4%	0.4%	0.4%		
Uneducated	0.9%	0.8%	1.3%	1.4%	1.1%		
Unknown	0.1%	0.1%	0.1%	0.0%	0.1%	412.44	0.000
Job Type						X^2	
Private Sector Employee	68.9%	68.8%	70.1%	69.7%	69.4%		
Public Servant	9.8%	9.0%	7.9%	7.5%	8.5%		
Self-Employed	11.1%	11.8%	11.9%	13.3%	11.9%		
Retiree	7.3%	7.2%	7.1%	7.0%	7.1%		
Housewife	1.6%	1.7%	1.5%	1.3%	1.5%		
Non-Employed	0.3%	0.6%	0.7%	0.7%	0.6%		
Other	0.9%	0.9%	0.8%	0.6%	0.8%	23.846	0.1464

Table A.11: Centers of Loyalty Train Set 3 Clusters

	Adventure Seekers	**Category Addicts**	**Location Addicts**	**Devoted Shoppers**	**Total**	**F**	**p-value**
Shopping Mall Loyalty Score	0.65	0.68	0.93	0.98	0.87	12210	0.000
Category Loyalty Score	0.69	0.96	0.76	0.99	0.90	11736	0.000

Table A.12: Customer Demographics Profile of Loyalty Train Set 3 Clusters

	Adventure Seekers	**Category Addicts**	**Location Addicts**	**Devoted Shoppers**	**Total**	**Test Statistic**	**p-value**
Gender						X^2	
Male	48.5%	44.2%	53.9%	59.0%	53.6%		
Female	51.5%	55.8%	46.1%	41.0%	46.4%	148.23	0.000
Marital Status						X^2	
Married	65.6%	63.5%	70.5%	70.0%	68.2%		
Single	26.6%	27.9%	21.2%	22.3%	23.8%		
Divorced	5.2%	5.7%	5.4%	5.0%	5.2%		
Unknown	2.6%	2.8%	2.9%	2.8%	2.8%	45.09	0.000
Age Group						F	
19-30	21.5%	26.1%	20.3%	20.7%	21.9%		
31-42	50.7%	44.5%	47.8%	47.1%	47.1%		
43-54	22.2%	23.2%	25.3%	25.3%	24.5%		
55-66	5.2%	5.7%	5.8%	6.1%	5.9%		
67-83	0.4%	0.5%	0.8%	0.8%	0.7%	8.016	0.000
Education Status						X^2	
Primary School	0.8%	2.2%	2.7%	3.8%	2.9%		
Middle School	2.0%	4.0%	4.4%	5.6%	4.6%		
High School	26.0%	33.5%	34.5%	42.9%	37.2%		
College	9.8%	9.1%	11.4%	9.3%	9.7%		
University	52.8%	44.6%	40.3%	32.5%	39.1%		
Master	6.8%	5.8%	5.0%	4.0%	4.9%		
PhD	0.7%	0.3%	0.6%	0.3%	0.4%		
Uneducated	1.1%	0.5%	0.9%	1.4%	1.1%		
Unknown	0.2%	0.1%	0.1%	0.1%	0.1%	224.26	0.000
Job Type						X^2	
Private Sector Employee	68.6%	68.8%	70.3%	69.6%	69.4%		
Public Servant	10.0%	9.4%	8.7%	7.7%	8.5%		
Self-Employed	12.1%	11.4%	10.7%	12.7%	11.9%		
Retiree	6.1%	7.5%	7.1%	7.2%	7.1%		
Housewife	1.7%	1.5%	1.8%	1.4%	1.5%		
Non-Employed	0.4%	0.6%	0.4%	0.7%	0.6%		
Other	1.0%	0.8%	1.0%	0.7%	0.8%	24.023	0.1424

Appendix B: Segment Properties

Figure B.1: Gender Distribution of Diversity Based Segments

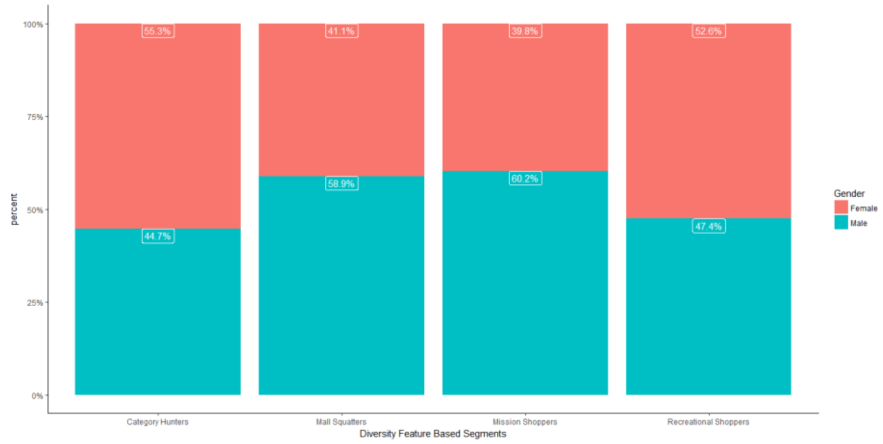

Figure B.2: Marital Status Distribution of Diversity Based Segments

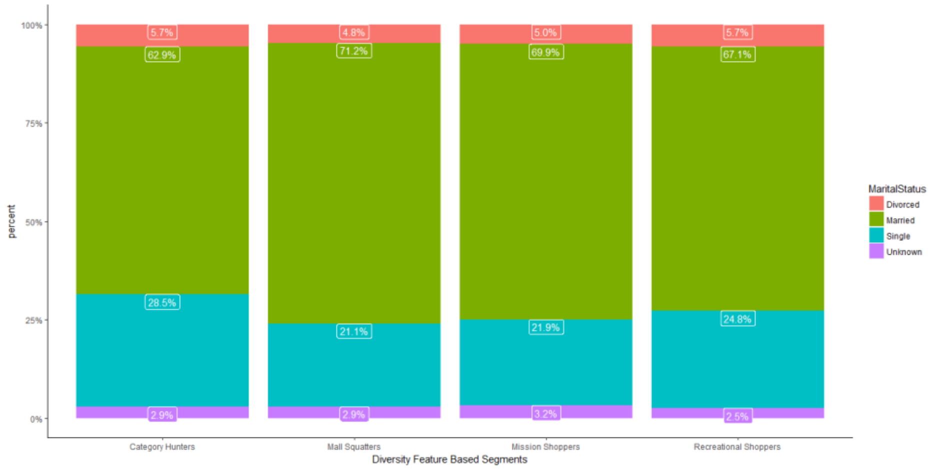

Figure B.3: Education Level Distribution of Diversity Based Segments

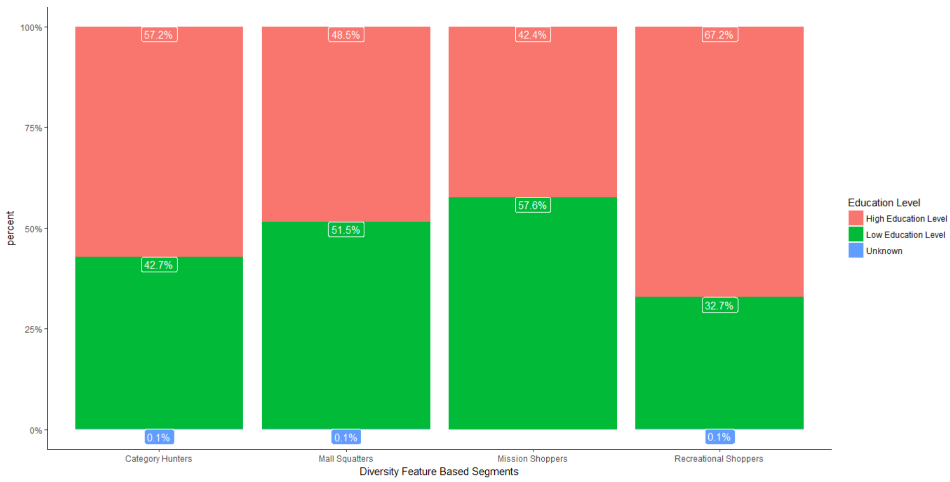

Figure B.4: Age Distribution of Diversity Based Segments

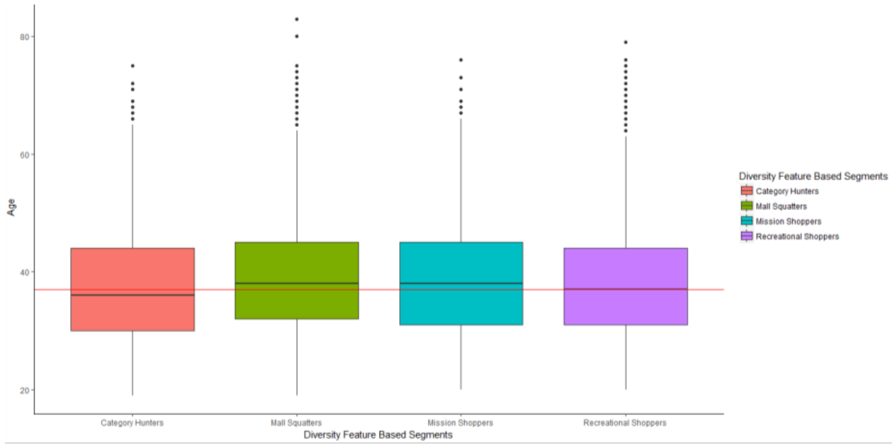

Figure B.5: Transaction Count Distribution of Diversity Based Segments

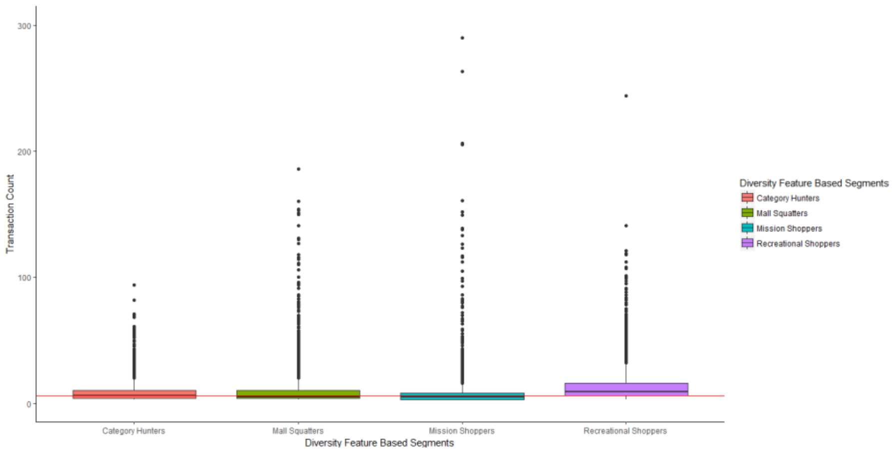

Figure B.6: Day Diversity Distribution of Diversity Based Segments

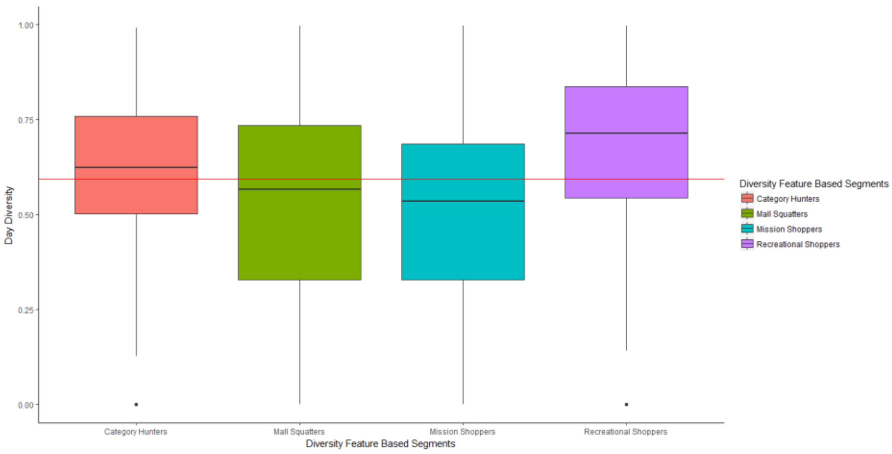

Figure B.7: Day Loyalty Distribution of Diversity Based Segments

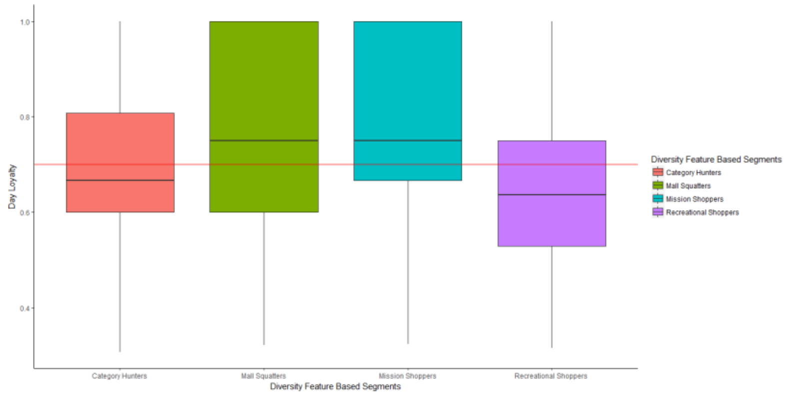

Figure B.8: Gender Distribution of Loyalty Based Segments

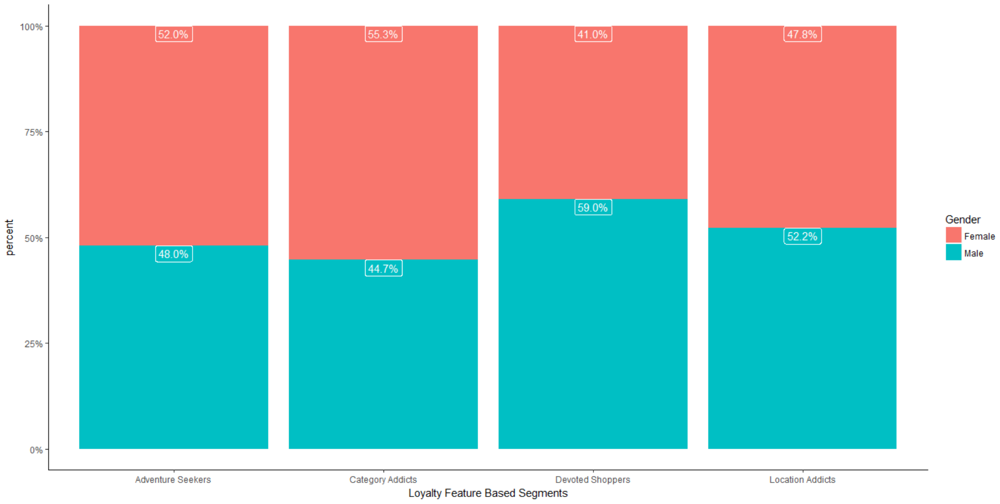

Figure B.9: Marital Status Distribution of Loyalty Based Segments

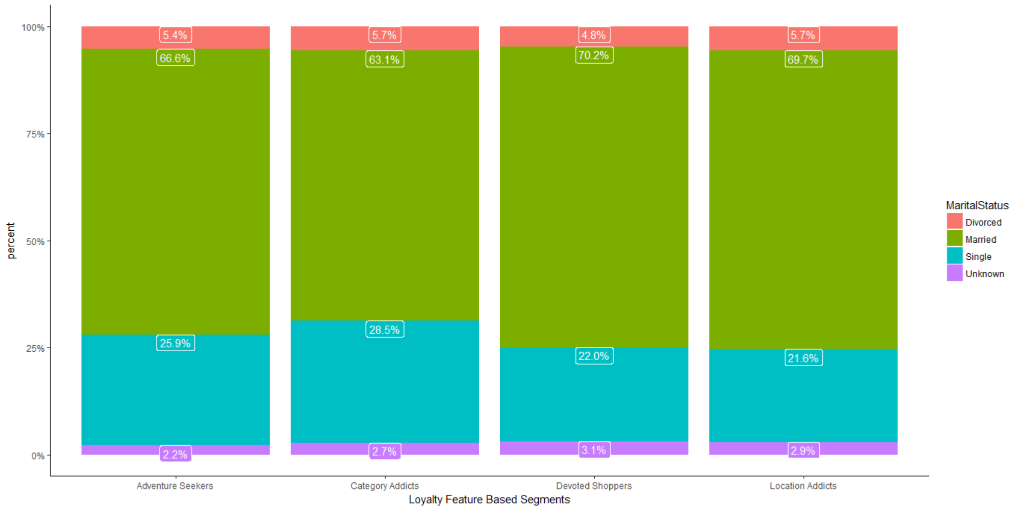

Figure B.10: Education Level Distribution of Loyalty Based Segments

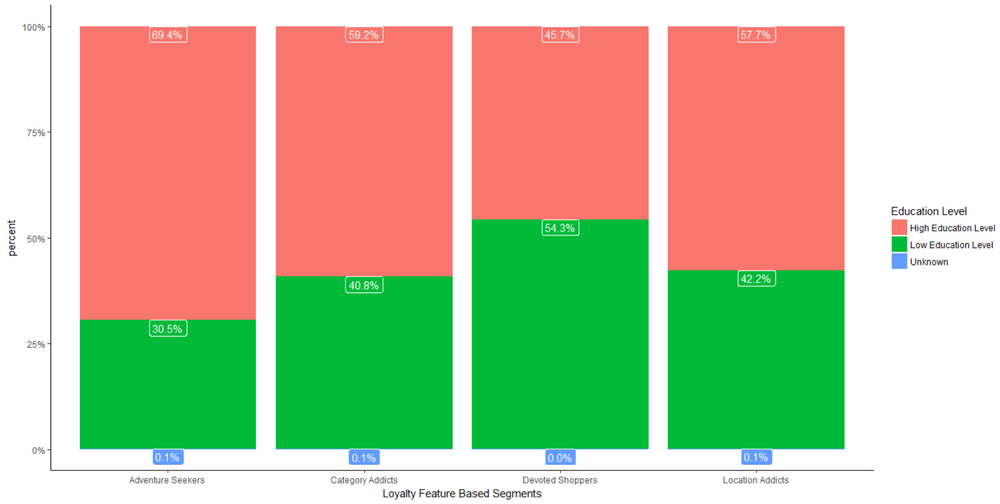

Figure B.11: Age Distribution of Loyalty Based Segments

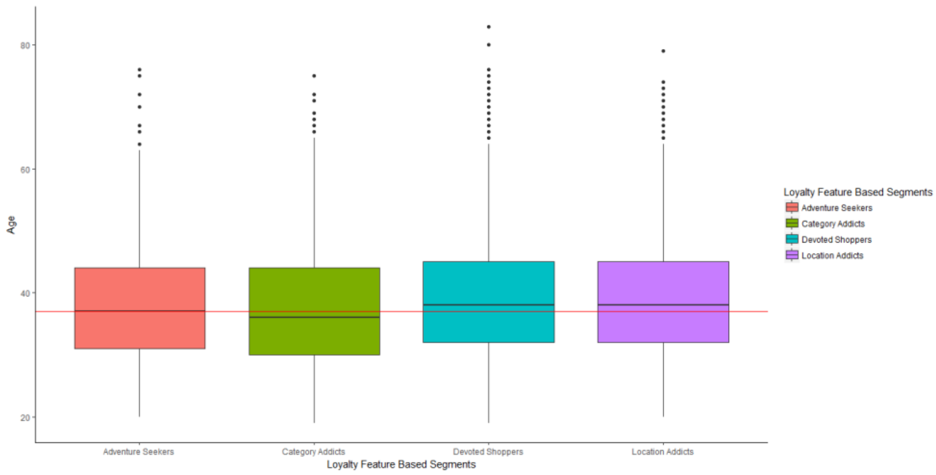

Figure B.12: Transaction Count Distribution of Loyalty Based Segments

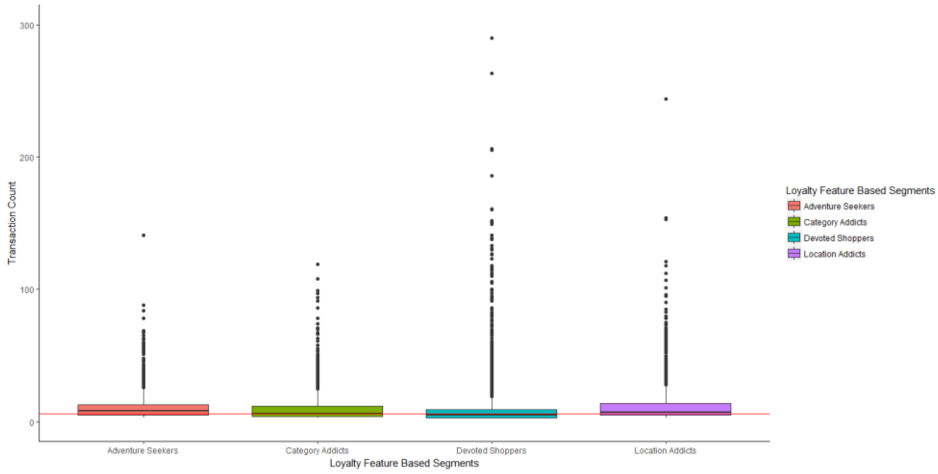

Figure B.13: Day Diversity Distribution of Loyalty Based Segments

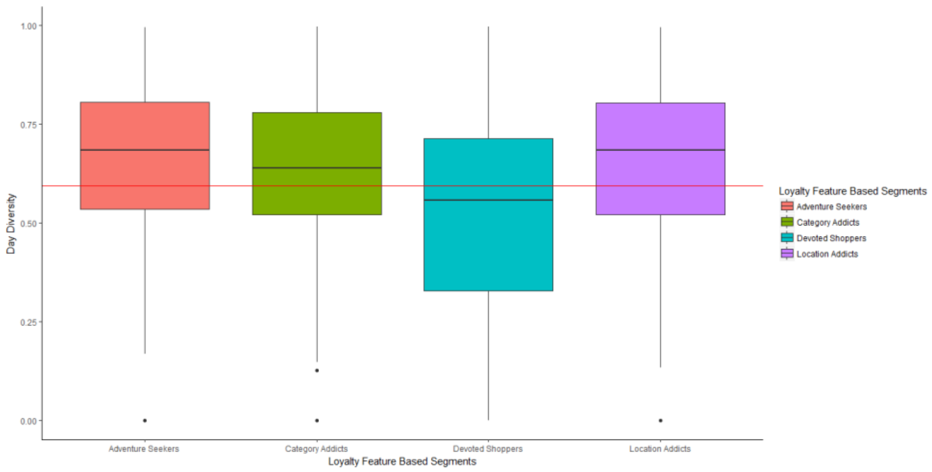

Figure B.14: Day Loyalty Distribution of Loyalty Based Segments