ANALYSIS OF OUT-OF-TOWN EXPENDITURES AND TOURIST TRIPS USING CREDIT CARD TRANSACTION DATA

by
GERGELY BUDA

Submitted to the Graduate School of Management
in partial fulfilment of
the requirements for the degree of Master of Science in Business Analytics

Sabancı University
December 2019

ANALYSIS OF OUT-OF-TOWN EXPENDITURES AND TOURIST TRIPS USING CREDIT CARD TRANSACTION DATA

Approved by:

Prof. Cenk Koçaş .....
(Thesis Supervisor)

Assoc. Prof. Raha Akhavan Tabatabaei .....

Assoc. Prof. Enes Eryarsoy .....

Date of Approval: December 5, 2019

ABSTRACT

ANALYSIS OF OUT-OF-TOWN EXPENDITURES AND TOURIST TRIPS USING CREDIT CARD TRANSACTION DATA

GERGELY BUDA

BUSINESS ANALYTICS M.Sc THESIS, DECEMBER 2019

Thesis Supervisor: Prof. Cenk Koçaş

Keywords: transaction data, credit card transactions, human mobility, tourist expenditure, tourist trips, purpose of travel

Credit card transaction data contains a vast amount of valuable information that can indicate consumer behaviour patterns and mark out human mobility. In this study we analyse the transactions carried out by a sample of 10.000 Istanbul-based customers of a Turkish bank to scrutinize expenditures incurred out of Istanbul. In our preliminary descriptive analysis, we examine the relation between demographic attributes and spending measures, as well as investigate the extent to which the population and the number of points of interest imply higher or lower credit card expenditure by visitors. We develop a methodology to extract tourist trips from consecutive credit card transactions. Subsequently, we implement a hierarchical clustering method to evaluate what the purpose of these trips might have been. Our results indicate 5 clusters of purpose: 'Leisure', 'Business', 'Acquisition', 'Visiting Friends and Relative' and 'Package Holiday'. The same clustering method is applied to segment provinces of Turkey based on which product and service categories visitors prefer. We deploy a number of predictive models to estimate tourist expenditure and whether a person would embark on a trip in the upcoming months. The predictive power of these models are generally moderate; nevertheless, several of the most useful predictors are behavioural or are related to previous trips, factors that have not been considered in literature.

ÖZET

ŞEHIR DIŞI HARCAMALARIN VE TURIST GEZILERININ KREDI KARTI IŞLEMSEL VERILERI KULLANILARAK ANALIZI

GERGELY BUDA

İŞ ANALİTİĞİ YÜKSEK LİSANS TEZİ, ARALIK 2019

Tez Danışmanı: Prof. Dr. Cenk Koçaş

Anahtar Kelimeler: işlemsel veriler, kredi kartı işlemleri, insan hareketliliği, turizm harcaması, turizm gezileri, seyahat amacı

Kredi kartı işlemsel verileri, tüketici davranış şekillerini gösterebilecek ve insan hareketliliğini belirleyebilecek çok miktarda değerli bilgi içermektedir. Bu çalışmada, bir Türk bankasının İstanbul'a kayıtlı 10.000 müşterisi tarafından gerçekleştirilen ve İstanbul dışından yapılan harcamalar analiz edilmektedir. Demografik özellikler ve harcama arasındaki ilişkinin yanı sıra, nüfus ve cazibe merkezlerinin sayısı ile ziyaretçilerin kredi kartı harcamalarının arasında bir ilişki olup olmadığı ilk betimsel analiz ile irdelenmektedir. Kredi kartı işlemlerinden turist seyahatlerini çıkaran bir metodoloji geliştirilmiştir. Daha sonra, bu seyahatlerin amacının ne olabileceğini değerlendirmek için hiyerarşik bir kümeleme yöntemi uygulanmıştır. Seyahat amaçları beş kümeye ayrılmıştır: "Keyfi Seyahatler", "İş Seyahatleri", "Alışveriş Amaçlı Seyahatler", "Arkadaş ve Akraba Ziyaretleri" ve "Tatil Paketleri". Türkiye illerinin ziyaretçilerin hangi ürün ve hizmet kategorisini tercih ettiğine bağlı olarak kümelenmesinde de aynı kümeleme yönteminden yararlanılmıştır. Turist harcamalarını ve bir kişinin önümüzdeki aylarda seyahate çıkıp çıkmayacağını tahmin etmek için bir dizi öngörücü model kullanılmıştır. Bu modellerin öngörü gücü genellikle ortalama olmakla beraber, en etkili değişkenlerin bir kısmının, literatürde pek göz önünde bulundurulmamış olsa da, önceki seyahatlerle ilgili ve davranışsal olduğu tespit edilmiştir.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and gratitude to my thesis supervisor, Prof. Cenk Koçaş, and my co-advisor, Prof. Burçin Bozkaya for having guided me through my journey of composing the present thesis, contributing with their expertise knowledge and original ideas. I am very grateful for them appreciating some of my unconventional approaches and leading the way to harmonising these ideas in a scientific way.

I would like to thank the conscientious and hard work of other faculty members and guest lecturers, namely to Assoc. Prof. Raha Akhavan Tabatabaei, Assoc. Prof. Abdullah Daşcı, Assoc. Prof. Ali Doruk Günaydın and Dr. Mustafa Hayri Tongarlak. Their passion for the field has been considerably impactful for my academic development. Likewise, I am thankful to Ms. Ekin Basat for the administrative support I frequently needed during my studies.

I would like to thank Burcu Sarı and Eda Eylül Akdemir from my cohort for having treated me with their friendship and having helped me to integrate into the Turkish culture. Also, I am eternally grateful to Ömer Sarıgül for motivating and supporting me with respect to my academic activities and to Guillermo Gómez González for being the first to spark my interest in Machine Learning and algorithms with his interesting projects.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

These days, more and more human activities involve the use of some sort of technology that produce vast amount of data. Together with the rise of importance given to systematic data analysis and the proliferation of data mining techniques and algorithms, researchers have ventured to mine large databases in order to gain valuable insights. The worth of these explorations is multi-faceted. They make a contribution on a theoretical level to many disciplines, such as sociology, economics, behavioral finance or marketing management; they give awareness of different human behavioral patterns and trends. Beyond the theory, the outcome of this bulk of studies can help corporations understand the customer better, surpass intuition-based decision making and eventually, convert these insights into profit.

Beside data collected via the use of mobile phones, GPS or social media, credit card transactions have constituted a major source of data for research purposes. This trend derives from the fact that each transaction with credit cards gets registered on several platforms, and amounts to a large database that may be combined with other relational databases, such as customer demographics or other transaction types. Early research was more focused on the credit card system itself, including fraud detection (Chan, Fan, Prodromidis & Stolfo, 1999). The deployment of data mining tools on transaction data to investigate human behavior gained ground in the new millennium. Geo-located credit card data has been used to analyze human mobility and networks (Sobolevsky, Sitko, R. T., Arias & Ratti, 2014b), social-economic conditions of cities and regions (Sobolevsky, Sitko, R. T., Arias & Ratti, 2015) and well-being (Lathia, Quercia & Crowcroft, 2012), to name a few.

Information about human mobility has been used as a predictor for other variables, such as financial difficulties (Singh, Bozkaya & Pentland, 2015). More seldom, mobility was taken as the unit of measurement, e.g. in the form of daily trips with the aim of unraveling motifs of mobility (Schneider, Belik, Couronne, Smoreda & Gonzalez, 2013). On the other hand, up to our knowledge, big credit card transactional data has not been relied upon to scrutinize non-daily, unique trips realized outside of the home city of customers.

The present study has several goals. On a more descriptive level, we aim to deploy some of the approaches used in literature with the goal of marking out regional differences in touristic expenditure. Later, we aim to relate the results to regional socio-economic and touristic macro data, as well as unveil demographic differences.

Secondly, we set as an objective to establish a reasonable framework to extract out-of-home trips from a large credit card transaction data, as well as to derive meaningful features describing them.

Our final objective is to make an educated clustering based on the potential motives of these highlighted trips and to eventually set up predictive models which would assist us to predict a customer's tendency to use their credit card outside of their home urban area. Additionally, we intend to estimate the amount of expenditure customers incur during their trips.

The present thesis proceeds as follows. In Chapter 2, we delve into the relevant literature to acquire knowledge about available methods to approach tourism, tourist trips and expenditures. This way, we will be able to define what our contribution is to the pool of academic studies.

In order to accomplish the aforementioned objectives, we will use a database compiled by a private bank in Turkey, consisting of a sample of 10.000 customers - taken from a larger database of 100.000 customers at random - their respective 1,176,929 credit card transactions, with a time span of one year. The data source will be described in Chapter 3 in detail; then we will proceed with a description of the data preprocessing techniques we use, followed by a preliminary descriptive analysis in Chapter 4. In this chapter, we will also outline our approach to extract the out-of-home trips.

Chapter 5 will present the methodological tools we use to address our research questions. We will finalize this section with considerations on feature extraction and modification, beyond the ones given in the database. In Chapter 6, we will present our results and interpret them in order to reach the objectives we initially set out. Our final summary and remarks will be presented in Chapter 7.

2. LITERATURE REVIEW

In this chapter, we will review the available studies under the umbrella of four main topics. First, we will seek to define tourism and trips, then present some of the many studies describing the different spending behaviors people demonstrate during their time away from home, as well as their preference for means of payment. Secondly, we will explore the literature on human mobility analyzed based on credit card transaction data along with relevant features. Thirdly, we will examine how trips differ based on travelers' purpose and what traveler profiles match with these motives. Then, we will present the research conducted on predicting tourist expenses, with an eye for the methods and features used. Finally, we will introduce some of the statistics published on domestic tourism in Turkey, followed by our remarks on how we expect to contribute to the literature with our research.

2.1 Tourism, Trips and Differences in Purchasing Behaviour

Tourism management evolved into a self-standing academic discipline throughout the past century. One of the earliest and most acknowledged endeavors to conceptualize tourism was compiled and put to paper by Leiper (1979). The article investigates the evolution of the conceptualization of tourism; the word 'tourism' originates from Greek and stands for a circular tool – the meaning it subsequently attained refers to the notion that tourism starts and terminates at the same point, similar to a circle. Early definitions focused on the services provided rather than spatial and temporal elements (McIntosh, 1977). Later definitions differ mainly on three variables: the distance traveled, purpose of trip and the duration. While required distance is loosely defined, a 'visitor' is generally regarded as a 'tourist' if their stay exceeds a time interval of 24 hours – otherwise we could name them as excursionists (I.U.O.T.O., 1963).

A common feature of tourists is that they are net consumers within the region visited

- they consume more than they earn; nevertheless they may be remunerated by an entity located in the region of origin (Burkart & Medlik, 1974). Based on this view, business trips are also considered a type of tourism, and form the second major type of tourism according to purpose. The authors argue that business trips are discretionary acts and constitute a departure from normal day-to-day activities.

Generally academicians do not draw a splitting line for the distance, beyond which a trip is considered a touristic trip; nevertheless, trips have been defined in terms of three geographic components: the tourist generating region, the tourist destination region and the transit area (Gunn, 1972). The tourist generating region stands for the permanent residence of the traveler, the destination is the location which counts with the attraction or attractions inciting the traveler to stay temporarily. Transit routes are linking regions that are stopover points, potentially offering some points of interest but not constituting the main destination for the tourist. From a data analytics point of view, all three stages extend potential variables, as all three involve several actions made with corresponding behavioral patterns: the planning and organization before the trip, interaction with facilities during the transit and interaction with attractions and services at the destination.

Literature offers insight on the preferences and demographic segmentation of people going on trips. Shopping is one of the most popular activities tourists engage in during their trip. Since tourists temporarily break free from work and home-related duties, their shopping behavior is different from that displayed at home. (LeHew & Wesley, 2007).

While early research was more focused on purchases of souvenirs, other product and service types emerged as being favored by travelers, such as clothing and jewelry, books or even electronics (Timothy, 2005). The same research shows that with the increasing popularity of self-catering accommodations and visits to friends or family, grocery shopping has come forth as a significant category. Different demographic groups exhibit different patterns in shopping among categories of products and services. Daily goods, clothes and jewelry were found to be more preferred by female travelers, while males opted for spending more on dining out, tobacco and alcoholic beverages (Jansen-Verbeke, 1987) (Oh, Cheng, Lehto & O'Leary, 2004). Income and age were also found to be a point of difference: travelers with lower income surprisingly spent more during trips, primarily on clothing – a product category that was also favored by younger age groups (Lehro, Cai, O'Leary & Huan, 2004). Irrespective of the product category, Turner & Reisinger (2001) claimed that domestic tourists seek out products that are unique in some way – such as, those only available at the destination or supplied with higher quality or more affordable price.

Beyond demographic differences and preferences for certain commodity types, it is relevant for the present study to contemplate the different means of payments that tourists can make use of. Clearly, the bank data at our disposition indicates any sort of transaction with the credit card, whether it is an online payment, a credit card payment via the EFTPOS system or a cash withdrawal or deposit carried out at an automated teller machine. From here on we will refer to credit card transactions simply as card transactions since we do not have debit, gift or any other form of card transactions in our data set. In recent years, use of cash has declined, while card transactions have shown a substantial rise in popularity – these two payment methods are in fact asserted to be complements, indicating negative correlation between themselves (Scholnick, Massoud, Saunders, Carbo-Valverde & Rodriguez-Fernandez, 2008). This result was later replicated by Carbo-Valverde & Rodriguez-Fernandez (2014). The study of El-haddad & Almahmeed (1992) showed only the first signs of ATM usage for cash withdrawal spreading from educated young people to lower income, older masses; a quarter decade later it is sound to claim that ATMs are used by all demographic groups. Considering ATM withdrawals might be a valuable addition to credit card transactions, especially while predicting expenditure.

Regarding card usage, literature argues that women have a higher likelihood to use bank card and they generally spend more (Hayhoe, Leach, Turner, Bruin & Lawrence, 2005). Sobolevsky et al. (2015) found that the average value per credit card transaction was higher for males, denoting a higher concentration of economic activity for this gender; whereas women realized a higher number of transactions and demonstrated higher spending diversity. According to Borzekowski, Kiser & Ahmed (2008), the propensity to use bank cards declines with age. These details may be compared further on to the descriptive statistics of our bank data.

2.2 Human Mobility, Networks and Regional Indicators

Analysis of mobility patterns based on credit card transactions was the focus of a series of articles drawing on a large database acquired from the Spanish bank BBVA.

Sobolevsky et al. (2014b) examined the mobility networks of bank customers, drawing up the strength of the edges between regions and municipalities in Spain in order to reflect the money circulation. Based on this modularity optimization algorithm, they concluded that neighbouring regions are cohesive and spatially connected in

terms of money flow; that is, people choose to spend most of their money in close areas. Furthermore, administrative borders between provinces were also coinciding with the general geographic spread of individuals' spending, in a sense, showing that regional borders are also psychological borders in terms of spending.

The same group of researchers (2014a) found that the spending activity of tourists increases in a superlinear fashion with a larger population size of a region. Provinces were subsequently clustered based on the residuals to the trendline; these clusters exhibited considerable differences in yearly temporal patterns and relative deviations for spending on different categories of products and services.

A further study based on the BBVA data conducted (Lenormand, Louail, Cantu-Ross, Picornell, Herranz, Arias, Barthelemy, Miguel & Ramasco, 2016) established a thorough descriptive analysis of credit card transactions. The authors' findings reiterated a higher spending concentration for men but higher volumes of spending for women. As for an age and occupation breakdown, young people and students spent the least both in volume and amount per transaction, middle-aged workers had the highest total expense, while elderly and retired customers spent a moderate amount of money distributed through relatively few transactions.

Mobility measures developed from bank card transaction data have not only been used as tools to be related to demographic variables or to describe and cluster administrative areas; recent studies have also built predictive models with mobility features as input variables. Singh et al. (2015) made a valuable contribution to financial analysis systems by finding that human mobility measures, namely diversity, loyalty and regularity, are better predictors for financial well-being indicators than demographic ones. Krumme, Llorent, Cebrian, Pentland & Moro (2013) set up Markov models to predict shopping choices based on recent mobility and shopping entropy.

Schneider et al. (2013) used surveys and mobile call data from two cities to extract ubiquitous mobility networks, that they denominated as 'motifs'. They found that each person has a typical mobility motif that repeat over several days, visiting few locations and often in the same order and same starting and ending point. This approach, however, is hardly applicable for rare and infrequent events, such as long-distance trips.

2.3 Trips and Motives

As previously mentioned, literature generally categorizes the purpose of touristic trips into 'leisure' and 'business' holidays. Many studies investigate the profile of tourists that set off with either of the two motives. Cai, Lehto & O'Leary (2001) collected survey data from Chinese outbound travelers and found that more than 50% of leisure-related trips were visits to relatives in the US. Business trips were mostly set out by males in managerial positions and showed an inverse U-shaped income distribution due to the income gap between workers in public and private sphere. People embarking on leisure trips started planning and making airlines reservation much earlier than the business group. Significant difference in spending categories were found for entertainment and lodging – the former being higher for leisure trips, the latter for business trips. Organized package trips were more favoured by people with leisure purposes, who also stayed for longer periods of time and manifested more diversity in activities. Moll-de Alba, Prats & Coromina (2016) reaffirmed that people on business trips spend shorter time than leisure travelers, and also discovered that they spend more on a daily basis and are more likely to travel by plane and stay in hotels. In this article, business trips were proportionally more predominant for middle-aged and male demographics.

According to the indications of the World Tourism Organization, 'leisure trips' could be instead defined as 'personal trips', which incorporates leisure and recreation, education-related trips and visits to relatives and friends – this agglomerated broad type together with 'business trips' represent the two main motives (UN, 2008). In literature on tourism management we often come across the term 'VFR' tourism, an acronym that stands for 'visiting friends and relatives'. VFR first started to gain recognition in the 1990s due to its growth and being the principal type of tourism in some regions. VFR tourists have relatively low expenses due to being provided with accommodation and food in many cases; these tourists, however, spend more on shopping and services (Seaton & Palmer, 1997). The same study revealed that VFR destinations are either large urban areas or smaller towns with relatively fewer holiday tourists and attractions. VFR type of travel was also characteristic for the 15-34-year-old younger group and singles. In the case of Turkey, VFR travels could be assumed to occur during the two religious holidays, counting with a vast outflow of people from metropolitan areas.

2.4 Predictive Models for Tourist Expenditure

Predicting the amount of expenditure of tourists at a destination has been a principal theme of many academic papers on tourism management. Mok & Iverson (2000) measured total expenditure including both expenditures during the trip and prepaid expenses, such as accommodation, airfare or other transportation, meals and package tours. Evidently, this interpretation of expenditure was feasible for a research based on surveys, but would pose a difficulty for unlabeled transaction data where it is unclear which expenses incurred prior to a trip can be marked as trip-related expenses. Nevertheless, it is sensible to consider expenses both at destination and outside the destination – e.g. during the transit – similarly to the study of Fredman (2008).

There is a noteworthy overlap between the independent variables researchers found to be significant in predicting tourist expenditure. The hypotheses of Lee, Var & Blaine (1996) based on economic theory were validated by the outcome of their study, implying that personal income is a main contributing determinant in such predictions. Fredman (2008) showed that household income and length of stay have the largest positive impact on expenditure. Additionally, the researcher's results indicated that people travelling from longer distances had higher overall expenditures, generally due to increased transportation costs. Another study on Taiwanese tourists suggests that the heavy spender segment was generally younger, stayed at the destination for a longer time and were either travelling alone or with a small group of people (Mok & Iverson, 2000). The 'must-have' list of determinants suggested by Thrane (2014) also includes travel party size, type of accommodation and destination beside the above-mentioned regressors; the researcher observed the desirable level of R2 at 40% for OLS models aiming to explain the variation of travel expenditure.

Duration of stay was found to be a significant determinant even for one-day trips, expressed in hours (Downward & Lumsdon, 2000). The size and the composition of the travelling group also emerged as a significant factor for such short outings. The extension of the study for longer than one-day trips extended the circle of significant variables with income (Downward & Lumsdon, 2003). This finding raises the question whether more or less static demographic variables are more reliable in predictions for longer trips, whereas spending on short trips may be more dependent on trip-related variables.

The review article of Brida & Scuderi (2013) compiled 86 papers and 354 models in

total on determinants of tourist expenditure. The authors categorized the common regressors under four groups: constraints, socio-demographic, psychographic and trip-related variables. Constrains include income, financial difficulties and assets, latter of which has been estimated based on expenditures at home, e.g. in the paper of Jang, Ham & Hong (2007), using spending for food at home to predict food-away-from-home expenditures. Socio-demographic determinants include age, education, gender, marital status, occupation and the like. For the most part, these variables were found to be significant in close to half of the papers that used them to test models. Psychographic variables relate to opinions about the trip. Trip-related regressors encompass variables linked to the specific trip rather than the traveler. A good majority of these variables, such as destination, duration, accommodation type, travel distance, time of the holiday, means of transportation and purpose, were significant in most reviewed articles. It is important to note that these papers conducted surveys; a study dependent on transaction data may face hardship at estimating some of the psychographic and trip-related variables.

The literature review of Brida & Scuderi (2013) also revealed that researchers have approached the phenomenon from two interconnected perspectives. One of them is the strive to explain the variance in expenditure with the above-mentioned four variable groups; this constitutes a regression problem. The other objective has revolved around a binary discreet choice of whether or not a person would set out on a trip, or alternatively, whether the person would or would not purchase tourist goods. This latter endeavour calls for an approach of classification. The authors of the review article pointed out that the classification problem was generally dealt with logistic regression models, while scientists used simple OLS linear regression for expenditure predictions – approach which the authors regarded as deficient due to unrealistic assumptions about a normal distribution of tourist expenditure.

2.5 Domestic Tourism in Turkey

According to the reports of the Turkish statistical government agency, TÜIK (2018), in 2018, close to 80 million trips were registered, with an average number of 8.1 nights spent there and an average of 513 TL spending per trip. Clearly, these are aggregate statistics for the total population of Turkey. It has been also observed that the average spending per trip and the number of trips was highest in the third quarter of 2018, with lower average spending in the first two quarters and lowest number of trips in the fourth quarter TÜIK(2019a). The same report claimed that

67% of the surveyed stated that the purpose of their trip was to visit friends or family, with 17.9% of respondents having traveled for leisure. Most people stayed at their family or friends' home, in second place they stayed at their own property – such as, a summer cottage – with only around 7% of the nights in the analyzed first quarter of year spent in hotels. We must note that these figures vary based on the quarter of year investigated in the reports. In the third, summer quarter of 2018 saw 36.4% of travelers going on trips for leisure, and around 12% of the nights spent in hotels TÜIK(2019b).

2.6 Our Contribution to the Literature

In our present research, we intend to fuse traditional approaches that are long-established in literature – such as predictions of tourist expenditure – with innovative data sources and approaches.

As per the preceding literature review, credit card transaction data has not been used for predictions about tourist expenditure. Surveys have been the primary source of data; however, we deem this method vulnerable to systematic errors, such as recall bias: it may be difficult for a person to remember all the product categories and the corresponding amount spent during a trip. Furthermore, there may be psychological biases in recall due to subconscious motives and attitudes. For instance, one might hardly recall purchases with little emotional value to the consumer, but these clearly appear in transactional data. Exact dates and locations of purchases are hard to keep in mind and recall during a survey, whereas time-stamped transactions are able to convey such information. Also, in a survey people may not want to admit having made purchases of some products, which appear explicitly in transactional data.

Additionally, we aim to introduce independent variables that are utilized by literature rarely or on a small scale: we explore deeper whether spending behaviour of people at their home region could be an estimator for their spending patterns on trips. We depart from the traditional modeling techniques of linear regression and logistic regression and implement more contemporary algorithms.

Another novel endeavor we aim to embark on is the attempt at generating a methodology for extracting long-distance trips from transaction data, rather than focusing on local urban mobility. This objective does not allow for a sequence or network analysis frequently used in research; instead, we rely upon descriptive literature and critical judgment to make a guess at the purpose of the extracted trips.

Finally, we aspire to contribute to the bulk of statistics available for domestic tourism in Turkey. We expect to elicit a more detailed breakdown of domestic tourist spending, both in terms of geological location and category of products.

3. DATA AND PREPROCESSING

In this chapter, we will introduce the dataset we obtained for the purpose of this study, along with its size and features. We will discuss the data preprocessing measures we take in order to proceed with a clean and functional variant of the database.

3.1 Data Collection

Our study is based on secondary data recorded by one of the major private banks in Turkey, counting with over 17.4 million customers. A sample of 10.000 customers was selected randomly and was handed over to our research team. The customers in the sample opened accounts with corresponding bank cards supplied in one of the bank's branches within the borders of Istanbul. This means, the data is limited in a sense that it does not contain data about customers who opened an account outside of Istanbul. Nevertheless, we deem that this bias allows a more extensive analysis, taking into account that residents of large cities travel in the country on a wider scale (Sobolevsky et al., 2015). The data covers a period of 1 year, starting from July 1, 2014 up till June 30, 2015.

The dataset obtained encompasses 22 different tables, out of which we regarded 4 tables as valuable for our study. These tables are titled as Customer Demographics table, Credit Card Transactions table, ATM Transactions Table and Risk Scores table. We listed the features these tables comprise of in **Table 3.1**.

The Customer Demographics table evidently contains 10.000 rows, each marking a different customer and their indicated 11 self-explanatory attributes. The 'Income' variable was either self-reported or estimated based on the customers' statement of income. Apart from 'Income', 'Age' and 'Years as customer', all variables are categorical, while the table also includes the customers' home and work coordinates.

Customer Demographics	Credit Card Transactions	ATM Transactions	Risk Scores
Masked Customer ID	Masked Customer ID	Masked Customer ID	Masked Customer ID
Gender	Date	Date	Risk Score
Marital status	Time	ID of ATM	
Education	Amount	Withdrawal/Deposit	
Job status	Merchant Type	Amount	
Income	Masked Merchant ID	Currency	
Age	Online transaction	Coordinates (X and Y)	
Home Coordinates (X and Y)	Expense type	Flag for unknown coordinate	
Work Coordinates (X and Y)	Currency		
Years as customer	Coordinates (X and Y)		

Table 3.1 Data tables and features

The Credit Card Transactions table is made up of 1.176.929 rows or transactions with 10 features. In this table, 'Merchant type' refers to an MCC code or spending category. The bank gave us assistance to interpret these four-digit codes with an additional table. The 1078 different MCC codes, out of which a good quantity appears in the transactions table, are grouped into 24 major Merchant categories, as seen in **Table 3.2**.

Merchant categories		
Other	Jeweler	Services
Car rental	Health and Cosmetics	Insurance
Vehicle Sale, Maintanance	Food	Building Materials, Hardware
Gas station	Apparel and Accessory	Direct Marketing
Airline	Supermarket	Various Grocery
Travel Agent and Carrier	Furniture and Decoration	Clubs and Organizations
Accomodation	Electrics and Electronics	Education, Stationary
Casino	Telecommunication	Contractors

Table 3.2 Merchant categories

While some merchant categories are self-explanatory, some are a little more implicit. The 'Other' merchant category, despite having several subcategories, in the dataset it is principally represented by retail sales and commercial equipments, and has a relatively high average amount per transaction, at 412.89 Turkish Liras. 'Health and Cosmetics' incorporates pharmacies, hospital and dentist expenses as well as cosmetics products. The category 'Food' stands for restaurants and caterings; not to be confused with the category 'Various Grocery', which are specialized stores for a specific food product, or 'Supermarket', under which the totality of a shopping basket at a supermarket or grocery store falls regardless of its composition. 'Services' range from home utilities, gardening services to cleaning, photography, funerary services, to name a few. 'Direct Marketing' refers to telesales services, while 'Education, Stationary' comprises both of tuition expenses and stationary or office small equipment.

Still in the Credit Card Transactions data table, the feature of 'Online transaction' is a dummy variable, taking the value of 1 if the transaction is made via an online payment, rather than a direct on-site use of the credit card. 'Expense type' could in theory take the value of 'cash advance' but our sample only contains the value 'shopping'.

The ATM Transactions data table includes both cash withdrawals and cash deposits, designated by the feature 'Withdrawal/Deposit'. The Risk Scores table shows how financially risky the bank assesses each customer to be.

In order to guarantee the confidentiality of personal data, the complete database is anonymized. Instead of names, the 'Masked Customer ID' serves as a marker to differentiate between customers. This feature is also the primary key in the Customer Demographics table, and is a foreign key in the other two tables, linking the three data tables together.

3.2 Data Preprocessing

3.2.1 Initial Data Arrangements

The programming language used for data cleaning and the majority of the subsequent analysis is Python, coded on the Anaconda Spyder environment. Before any descriptive analysis, we made the database undergo a cleaning process.

In the Credit Card Transactions table, we ordered the transactions chronologically for each of the customers to facilitate the extraction of trips later on. We substituted the 'Merchant Type' values with the corresponding 24 'Merchant Category' discussed before; where the 'Merchant Type' had a missing value, the 'Merchant Category' was marked as 'unknown'. In the ATM Transactions table, the column for 'Amount' was replaced by the additive inverse – it was signed negative – if and only if the transaction was a money deposit to the ATM. Hence, both in this table and the Credit Card Transactions table, transactions that we eventually expect to be expenses are with positive sign. Finally, all the static variables of the Customer Demographics table were merged to the other two data tables, based on the 'Masked Customer ID'.

3.2.2 Giving Meaning to Coordinates

The coordinate variables, expressed in X-coordinates for Earth longitude and Y-coordinates for latitude, define specific locations within the borders of whole Turkey, not solely Istanbul; fact which facilitates our further analysis on out-of-town trips. The coordinates of the dataset were projected on so-called shapefiles – geo-spatial vector data – representing Turkey and the country's administrative areas: its 81 provinces and further dissection into districts. We used QGIS, an open-source platform for geographic information system applications. **Figure 3.1** shows the projected coordinates from the Credit Card Transactions table. Coastal areas in the West, South and North, as well as large urban areas, such as Ankara appear with greater density of transactions, apart from Istanbul.

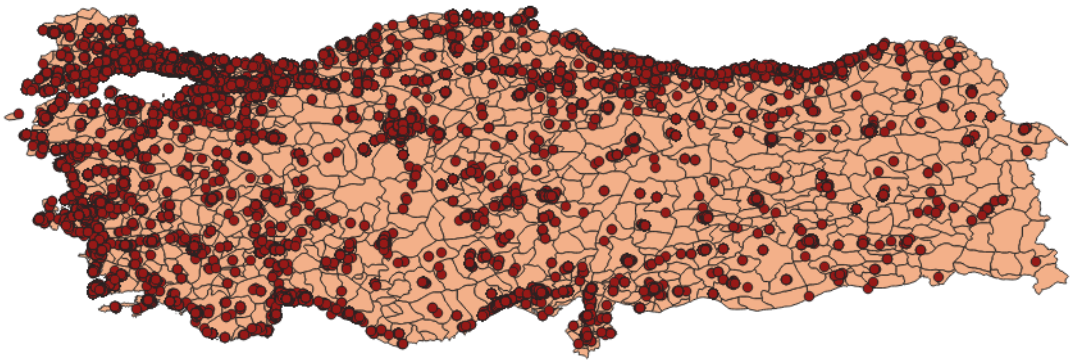

Figure 3.1 Transaction coordinates projected onto Turkey's map

Subsequently, we joined the coordinate layers and the shapefile based on location, and transferred the corresponding shapefile attributes. In this manner, we endowed each coordinate with the corresponding province and district. We repeated the process for all three data tables and joined the 'Province' and 'District' variables to the tables. For those few cases, where the GIS platform could not recognize the corresponding province and district – e.g. a transaction realized on the sea - we manually added the closest area names.

In order to measure geographic distances, we used the Haversine formula. This formula takes into consideration the spherical surface of the Earth, and is formulated as below:

$$a = \sin^2(\delta\phi/2) + \cos(\phi) * \cos(\phi) * \sin^2(\delta\lambda/2)$$

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

$$d = R * c$$

where ϕ is latitude, λ is longitude, R is earth's mean radius of 6.371 km

Formula 3.1 Haversine formula

We deployed the formula to two different ends. Firstly, we reckoned that a transaction taking place at an airport could be a good indicator that the means of transportation was by air, provided that it precedes a trip. We sought out the central coordinates of the two airports operating at the time when the sample was drawn: Istanbul Atatürk Airport and Sabiha Gökçen International Airport. We used the Haversine formula to calculate the distance in kilometers from each of the airports and added two new feature to the Credit Card Transactions table, namely 'Distance From Atatürk' and 'Distance From Sabiha'. Finally, after studying the size and layout of the airports, we regarded a transaction as one within the premises of an airport if it took place within 1 kilometer from its central coordinates – marked by the variable 'Airport'.

Secondly, we had in view to mark the distance of each transaction from the home and work address of each and every corresponding customer. Since the demographics variables have already been merged to the Credit Card Transactions table based on the 'Masked Customer ID', we implemented of the Haversine distance measure, creating the variables 'Distance From Work' and 'Distance From Home'.

3.2.3 Dealing with Missing Coordinates

Our methodology to derive trips requires that subsequent transactions indicate the real location the customer was positioned at the time. Coordinates corresponding online transactions pinpoint the location of the merchant rather than the whereabouts of the customer; thus, we set the online transactions aside, making up 104.917 data instances.

The downside of the Credit Card Transactions data table obtained from the bank is the relatively significant number of transactions with missing XY coordinates. The underlying reason is that transactions' geo-coordinates are only conveyed to the bank's data center on condition that the EFTPOS machine used by the merchant had been provided by the bank. Although the bank prides itself in having distributed

more than 580.000 EFTPOS terminals, 31.8% or 375.225 rows of transactions in the dataset contain missing coordinates. Needless to say, we cannot state with full certainty where these transactions took place. Nevertheless, in pursuance of possessing a close-to-complete record of out-of-town transaction so that the expenditure estimations are more accurate and no significant transactions are omitted, we applied a heuristic approach to fill in the missing data. Substituting missing values either in the coordinates or in the 'District' variables may be too optimistic, therefore, our replacements took place only for the 'Province' variable. After a detailed exploration of the dataset and trial-and-error attempts, we decided to fill in a missing value with a specific province, provided that both the chronologically preceding and succeeding non-missing transactions were indicating the same province, and at least one of them was recorded within less than 2 days from the materialization time of the transaction. **Table 3.3** shows an extracted example where the missing value could be filled in with 'Istanbul'.

After all these procedures, we were left with 984.343 data points in our Credit Card Transactions table.

Customer ID	Date	Time	Province
1584872	14/08/2018	13:05:02	Istanbul
1584872	14/08/2018	13:27:51	Istanbul
1584872	14/08/2018	15:16:45	nan
1584872	21/08/2018	16:46:42	Istanbul

Table 3.3 An example for cases when a missing province was filled in

4. DATA TRANSFORMATIONS AND DESCRIPTIVE ANALYSIS

In this Chapter, we will present the transformations we have put the raw data tables through. In the first place, we demonstrate the algorithm we devised in order to derive trips from the transactions.

4.1 Creating Trips Table

As a first step, the Credit Card Transactions was subject to a grouping transformation. The premise of this process is that the data table had to be ordered primarily by the 'Masked Customer ID', and secondarily by the merged 'Date and Time'. Then, our algorithm grouped together those consecutive transactions, that not only belonged to the same customer, but also were recorded in the same province. We called each of these groups a 'stay', referring to the continued presence of a person in a certain place, whether in Istanbul or out of home. This phase of the transformation is demonstrated with an example in **Figure 4.1**.

The new Stays table created via this process preserves the chronological order. Based on the transactions within each group, we appended new features along with the 'Masked Customer ID' and the 'Province' for the Stays table. These attributes are listed on the left-side panel of **Table 4.1**.

***Credit Card Transactions* table**

Masked Cust. ID	Date and Time	Amount	Province	Merchant Category
1605587	2015-02-26 11:42:43	129.94	Istanbul	Supermarket
1605587	2015-03-09 06:10:47	203.18	Istanbul	Gas station
1605587	2015-03-11 16:17:51	19.99	Trabzon	Apparel and Accessory
1605587	2015-03-11 16:45:21	99	Trabzon	Apparel and Accessory
1605587	2015-03-13 16:14:55	235	Trabzon	Apparel and Accessory
1605587	2015-03-13 16:30:15	80	Trabzon	Apparel and Accessory
1605587	2015-03-13 16:54:16	36.85	Trabzon	Supermarket
1605587	2015-05-06 12:13:15	165	Istanbul	Apparel and Accessory

***Stays* table**

Masked Cust. ID	Province	First date	Last Date	No. Days	No. Transactions	Total Amount
1605587	Istanbul	2014-10-04 17:17:32	2015-03-09 06:10:47	156	20	7030.09
1605587	Trabzon	2015-03-11 16:17:51	2015-03-13 16:54:16	3	5	470.84
1605587	Istanbul	2015-05-06 12:13:15	2015-06-29 09:59:39	54	12	1181.55

Figure 4.1 Grouping Credit Card Transactions into Stays

Beside the most fundamental statistics, such as minimum or mean, we also calculated the 'per transaction' and 'per day' amounts and the central location of the coordinates – the average of the X and Y coordinates. Likewise, the Stays table indicates a Merchant Category breakdown of the Total Amount by means of 24 variables.

Stays	*Trips*	
Masked Customer ID	Masked Customer ID	Total Amount at destination (TL)
Province	Duration	Minimum amount (TL)
First Date	Places	Minimum amount at destination(TL)
Last Date	Transits	Maximum amount (TL)
Number of Days	Destination	Maximum amount at destination (TL)
Number of Transactions	Duration at Destination	Average amount per transaction
Total amount (TL)	First Date	Average amount per transaction at destination
Minimum amount (TL)	Last dDate	Average amount per day
Maximum amount (TL)	No. Of transactions	Average amount per day at destination
Median amount (TL)	No. of transactions at destination	Distance
Standard Deviation of Amounts	Total amount (TL)	Expense in Merchant Categories (24 variables)
Average amount per transaction		
Average amount per day		
Average Coordinates (X and Y)		
Expense in Merchant Categories (24 variables)		
Exact duration		

Table 4.1 Features created for Stays and Trips data tables

The second core step of extracting trips from the dataset required us to determine some guiding principles for what we can designate as tourist trip. The algorithm we

used to transform the Stays table into Trips table is exhibited in **Figure 4.2**. The cornerstones of the algorithm are the following considerations:

- As discussed in the literature review, a trip consists not only of the activities at the destination but also the transit that leads there (Gunn, 1972). A transaction taking place at a transit location is secured a separate row in the Stays table, but should be considered as part of the trip for the Trips table.
- Regarding those customers, who opened their bank account by providing a home or work address located in a province other than Istanbul, we cannot make a certain statement whether their stay in that province is temporary and not revenue-generating, therefore we marked them as 'Not part of a trip'.

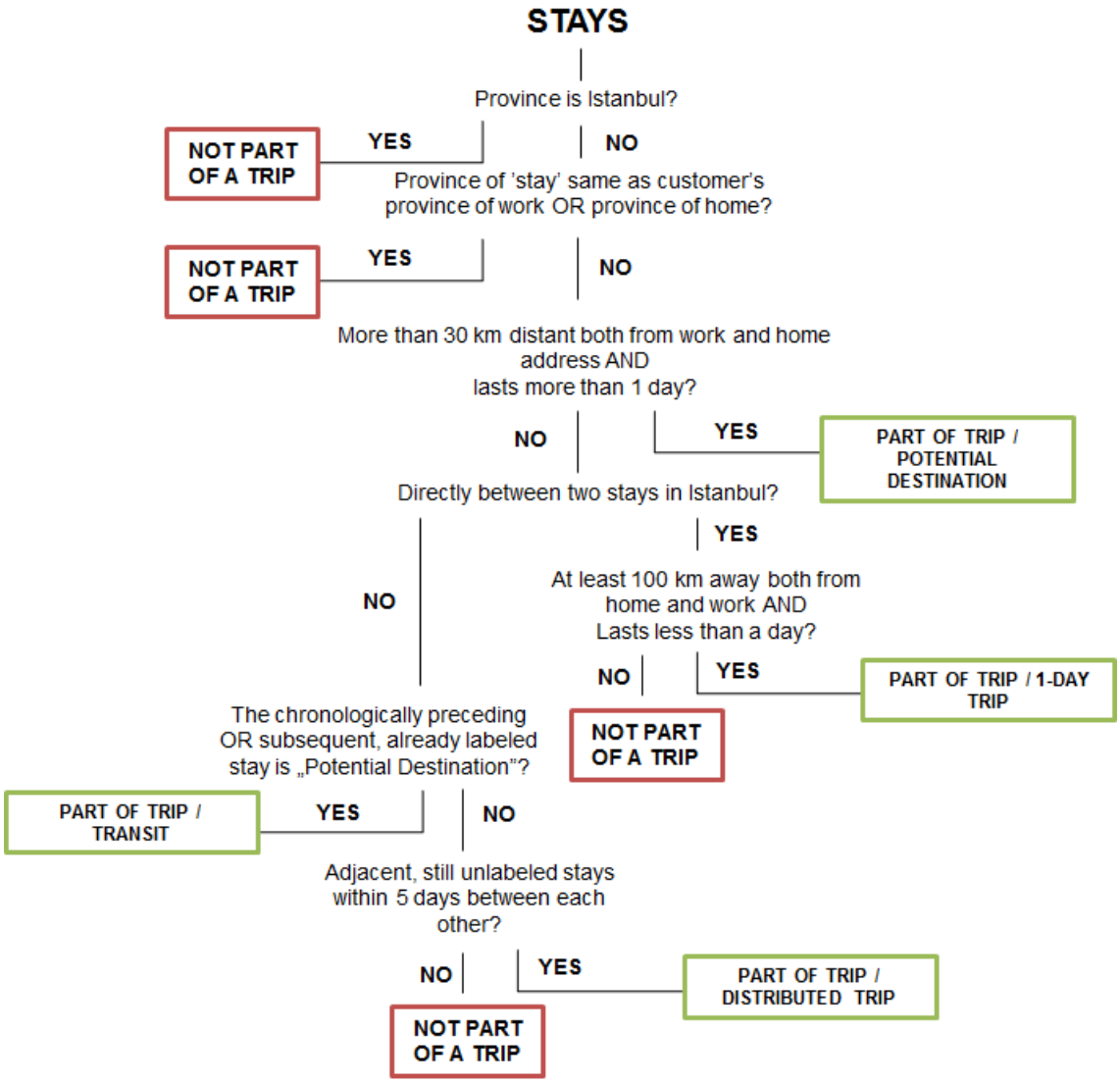

Figure 4.2 Algorithm for converting 'stays' into 'trips'

- We wanted to rule out stays that are in reality linked to regular commuting

to and from work outside of Istanbul, as the work address registered could have altered or be wrong. In terms of minimal distance both from work and home address, we found the threshold of 30 kilometers to be a reasonable marker for longer than 1 day stays and 100 kilometers for 1 day stays for them to be considered part of a trip. Leiper (1979) considered 100 kilometers to distinguish between 'local' and 'non-local' transactions. We applied this limit for less than 1 day short stays and also called for them to be directly between two stays in Istanbul, otherwise 'Transit' might fit them better. We decided on the 30 kilometer threshold for longer stays after observing that higher limits would rule out stays at holiday resorts located close to the edge of Istanbul. Needless to say that for future applications of the algorithm, one must consider the city and its surroundings in question for the precise determination of these parameters.

- At the lower end of **Figure 4.2**, we aimed to tag the transit areas, provided that from the chronologically adjacent, already labeled stays there is one marked with 'Potential Destination'. Finally, those stays that are not labeled by the end of the algorithm, are adjacent chronologically and are within 5 days distance, we considered to be road trips or 'Distributed Trips'; these show several stops in different provinces with short time spent at each.

To further illustrate how the algorithm transforms the 'stays' into 'trips', we included four examples in **Figure 4.3** for each of the possible trip types: 'Trip with transit', 'Trip without transit', 'One day trip' and 'Distributed trip'.

Three different variables keep track of the locations visited during the trips. The 'Places' variable lists all provinces touched upon; the 'Transit' variable, if non-empty, marks the provinces that we assume the traveler to pass across. To the third variable, 'Destination', we assigned the province where most transactions occurred in the case of 'trips with transit' and 'trips without transit'; we set as 'Destination' the province with the transaction furthest away from Istanbul. The rest of the variables coincide with the ones in the Stays table, except that separate variables were created for destination-related statistics.

4.2 Descriptive Analysis by Province

Departing from the Stays table, we created the Province Summary data table with the 81 Turkish provinces being the primary key.

In **Figure 4.4**, we displayed the 15 provinces where the highest total amount of credit card transactions took place, and showed the number of transactions corresponding to these expenditures. We notice that the provinces adjacent to Istanbul – Tekirdağ and Kocaeli - record the largest total amount of expenditures by residents of Istanbul. This finding is in accordance with Sobolevsky et al. (2014b), who found that close-by provinces are connected with bigger cash flows. Some of the explanations might be the presence of commuters, people doing a portion of their shopping activities in proximate places outside the administrative borders of Istanbul, or the necessity to cross these provinces if the person travels by car towards inner parts of the Anatolian peninsula. Beside neighbouring provinces, those with the most populated urban areas, such as Ankara or Izmir, and well-known seaside holiday destinations, such as Muğla and Antalya, figure among the provinces with highest spending. The line indicating the number of transactions made roughly follows the trendline for total expenditures, indicating relatively small deviations in terms of average spending per transaction.

We analyzed the total number of days we assume the sample of customers to have spent in each provinces in **Figure A.1** in the Appendix. Indubitably, the duration of stays extracted from the transaction data are only approximations. The figure puts forth a similar scenery as the one with expenditure levels in terms of province rankings.

In the Province Summary table, we queried the expenditures on each of the Merchant Categories with the objective of examining the proportion among them and its variation from province to province. To avoid clutter in our figures, we disposed of the categories with the lowest – less than a million Turkish Lira – expenditure levels, namely, 'Car rental', 'Airline', 'Casino', 'Direct marketing', 'Clubs and organizations' and 'Contractors'. The remaining Merchant Categories were projected on the map of Turkey, with pie charts indicating the category breakdown of expenditure over each province. The resulting **Figure A2-8** are presented in Appendix A.

Excerpt from ***Trips*** **datatable**

Ex.	Masked Cust. ID	Length (Days)	Places	Transits	Destination	First Date	Last Date	No. Of Trans
1	1680319	16	['Manisa']	[]	['Manisa']	2014-08-26 09:01:28	2014-09-11 17:42:12	6
2	1984913	10	['Giresun']	['Trabzon', 'Bolu']	['Giresun']	2014-07-24 17:47:37	2014-08-03 15:01:21	15
3	2866872	7	['Afyon', 'Isparta', 'Antalya', 'Sakarya']	[]	['Antalya']	2014-08-17 08:49:07	2014-08-24 19:13:11	4
4	3171082	0	['Sivas']	[]	['Sivas']	2014-08-08 15:40:55	2014-08-08 15:40:55	1

Examples in ***Stays*** **datatable**

Example 1: Trip without transit

Masked Cust. ID	Province	First Date	Last Date	No. Of Days	No. Of Transactions	Total Amount
1680319	Istanbul	2014-07-10 11:49:09	2014-08-17 15:35:50	39	3	962.05
1680319	Manisa	2014-08-26 09:01:28	2014-09-11 17:42:12	17	6	5634
1680319	Istanbul	2014-10-10 17:23:34	2014-11-23 15:43:04	44	4	15255.07

Example 2: Trip with transit

Masked Cust. ID	Province	First Date	Last Date	No. Of Days	No. Of Transactions	Total Amount
1984913	Istanbul	2014-07-03 23:15:16	2014-07-24 12:22:35	21	15	2262.08
1984913	Trabzon	2014-07-24 17:47:37	2014-07-24 17:47:37	1	1	340
1984913	Giresun	2014-07-25 10:26:51	2014-08-03 06:58:24	9	13	2262.76
1984913	Bolu	2014-08-03 15:01:21	2014-08-03 15:01:21	1	1	200.02
1984913	Istanbul	2014-08-20 14:19:41	2015-02-07 05:36:22	171	26	14351.74

Example 3: Distributed trip

Masked Cust. ID	Province	First Date	Last Date	No. Of Days	No. Of Transactions	Total Amount
2866872	Istanbul	2014-07-19 19:30:02	2014-08-03 19:39:33	16	13	2162.4
2866872	Kocaeli	2014-08-12 10:15:37	2014-08-16 19:59:20	5	2	431.37
2866872	Afyon	2014-08-17 08:49:07	2014-08-17 08:49:07	1	1	53.5
2866872	Isparta	2014-08-17 11:29:01	2014-08-17 11:29:01	1	1	260
2866872	Antalya	2014-08-24 12:58:12	2014-08-24 12:58:12	1	1	220.02
2866872	Sakarya	2014-08-24 19:13:11	2014-08-24 19:13:11	1	1	283
2866872	Istanbul	2014-08-31 21:40:52	2015-01-25 19:55:39	147	18	7240.8

Example 4: One-day Trip

Masked Cust. ID	Province	First Date	Last Date	No. Of Days	No. Of Transactions	Total Amount
3171082	Istanbul	2014-08-05 14:27:41	2014-08-07 15:59:36	3	3	4258.8
3171082	Sivas	2014-08-08 15:40:55	2014-08-08 15:40:55	1	1	103.4
3171082	Istanbul	2014-08-09 11:45:38	2014-08-19 15:33:20	11	11	2765.4

Figure 4.3 Examples of 'stays' grouped into 'trips'

Istanbul demonstrates a very diverse composition of expenditure with respect to categories, with supermarkets, clothes and accessories, fuel, insurance and electronics being the top classes. In neighboring regions, credit cards are mostly used for fuel, with supermarket and clothes expenses coming behind. People may visit these areas occasionally to do shopping, potentially for lower prices. The province Bilecik stands out with its considerable proportion of expenditure on fuel, implying that the province is presumably a transit region; in Edirne, payments in supermarkets are prevalent.

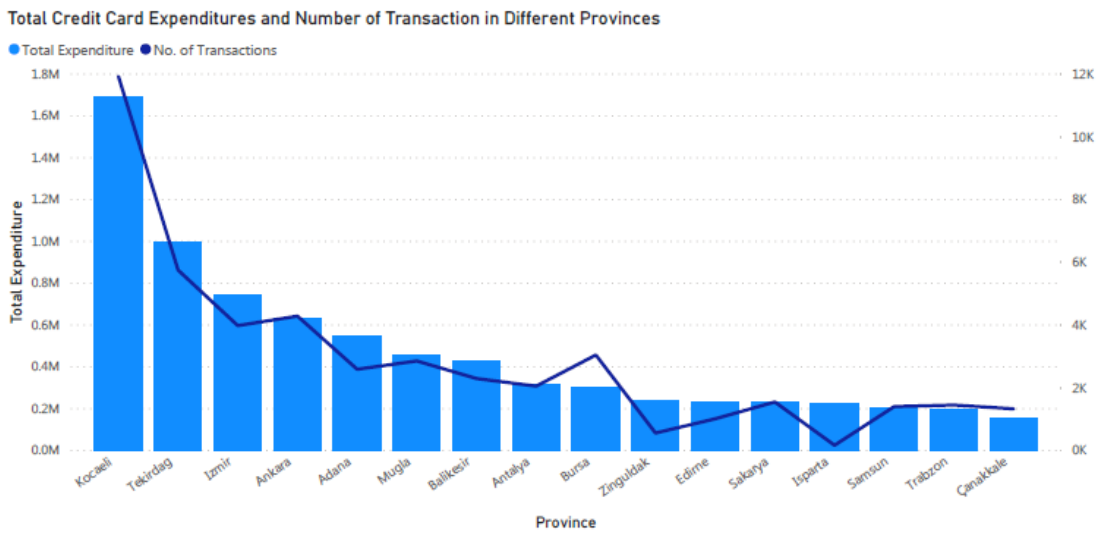

Figure 4.4 Total expenditures and number of transactions in each province

In the Aegean region, provinces without a coastline see expenditure on fuel to be significant - they may be transit regions for tourists who are driving toward the coast. Izmir and Muğla, two of the principal summer resorts in Turkey, demonstrate a similarly diversified proportion in categories like Istanbul. Expenditure on accommodation is more apparent in these regions, the region being a popular resort area for residents of Istanbul. Interestingly, purchase and maintenance of vehicles is more salient in Izmir, similar to accommodation costs in Afyon and 'other' wholesale purchases in Uşak.

The eastern regions of the Mediterranean appear to be transit areas, given that fuel expenditure is high. From Antalya to Hatay, areas accounting for a great share of domestic and international tourism, a varied mixture of spending categories are represented, with electronics more in the foreground relative to other regions, as well as building materials and hardware in Antalya or healthcare and cosmetics in Adana.

Most western regions at the Black Sea appear as transit areas with fuel expenses prevailing. Eastern urban areas like Trabzon, Rize, Samsun and Giresun show both higher total expenditures and a broader variety of spending classes. Here, transactions for building materials, hardware and electronics are more manifest.

In the Central Anatolian region, Ankara and Konya see a similar distribution of expenses to that in Istanbul, aside from the findings that transactions on apparel and accessories gain a higher fraction in Ankara, similar to building materials and hardware in Konya. The touristic region of Cappadocia in Nevşehir and Kayseri provinces marks elevated expenditure on accommodation, given that they are touristic regions.

Southeast Anatolia and Eastern Anatolia are seemingly less frequented by residents of Istanbul, at least within the obtained sample of 10.000 customers. Due to this fact, some of the provinces in these regions present a homogeneous distribution of credit card expenditure.

These descriptive statistics reveal that tourists may visit some out-of-Istanbul areas like Konya or Northern regions to purchase equipments and building materials - these areas may provide them at a lower price given its nearby natural resources. Many landlocked provinces appear to be mostly transit areas, with little attractiveness with respect to other products and services.

4.3 Population, POIs and Expenditure in Different Provinces

Sobolevsky et al. (2014a) observed a superlinear relationship between population of urban areas with volumes of domestic and foreign transactions. We conducted a similar analysis with our dataset on the basis of the 81 provinces in Turkey. We downloaded the province-based population statistics from the website of the Turkish Statistical Institute TÜIK(2019c) and extracted those corresponding to 2014, the year of the first transactions in our database. Following the example of Sobolevsky et al. (2015), we plotted the total expenditure against its population on a log-log scale, as seen in **Figure 4.5**. The purpose of this transformation is to defy skewness – the predominance of provinces with large cities – and to allow for an insight into the rate of change in expenditure, on the basis of population.

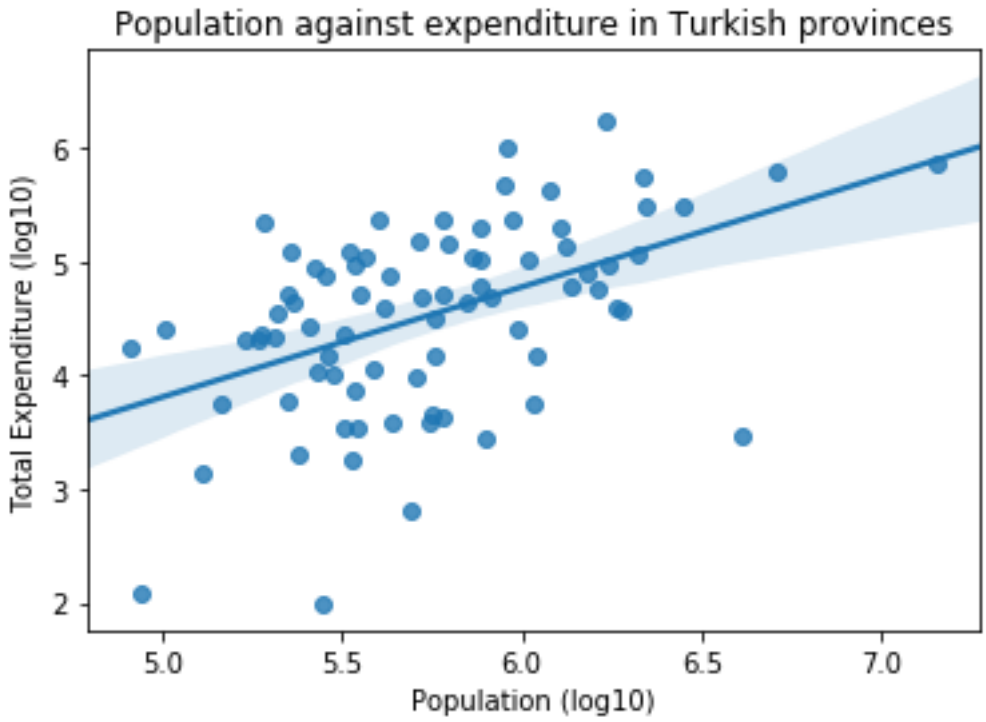

Figure 4.5 Total expenditure against population on a log-log scale

The logarithm of population with base 10 results to be statistically significant with a p-value of 5.7401E-06 in the explanation of the expenditure registered in provinces, with the base 10 logarithm applied for the dependent variable as well. The R2 or coefficient of determination is 23.3%, indicating that this fraction of the variation in the dependent variable can be explained by the variation in the independent variable.

We obtained datasets of geographic coordinates of different points of interest (POI) within the borders of Turkey in the fourth quarter of 2014, projected them onto the Turkey shapefile, and joined the layers by location. Eventually, we gathered the number of points of interest in each province. From the set of points of interest, we kept the ones related to 'Business', 'Entertainment', 'Restaurant', 'Shopping' and 'Travel', and assigned the corresponding number of data points to each variable.

As a second endeavor, we aligned the five POI variables with the population and expenditure data, and fitted a multiple linear regression - with the number of POI and population being the independent variables and expenditure being the dependent one for the provinces. After recursively eliminating the variables with higher than 5% individual p-values, our final model consists of only 'Business' and 'Shopping' related points of interest, as well as population size. The output of the regression is showed in **Table 4.2**.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,817350
R Square	0,668060
Adjusted R Square	0,654957
Standard Error	0,488488
Observations	80

ANOVA

	df	SS	MS	F	Significance F
Regression	3	36,49867944	12,16623	50,98572785	3,64753E-18
Residual	76	18,13513803	0,23862		
Total	79	54,63381747			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	4,96768	1,04922	4,73465	0,00001	2,87798	7,05738	2,87798	7,05738
Business_POI	0,82518	0,18950	4,35456	0,00004	0,44776	1,20260	0,44776	1,20260
Shopping_POI	0,74224	0,15710	4,72446	0,00001	0,42933	1,05514	0,42933	1,05514
Population	-0,69454	0,21453	-3,23751	0,00179	-1,12181	-0,26727	-1,12181	-0,26727

Table 4.2 Output of multiple linear regressions for expenditure in provinces

The coefficient of determination has increased considerably to 66.8%, so the model is better at explaining the variation in expenditure within each province. The two POI categories that remained in the model are establishments where one could assume to encounter considerable amount of cash flow. The number of these establishments emerge to be a fairly good indicator of how much residents of Istanbul would be spending in the province; however, we cannot make assumptions of causality by nature of the methodology. Interestingly, with the three dependent variables, population seems to negatively impact dependent variable; hence, we found it crucial to examine the correlations between dependent variables alongside with the regression output.

	Business_POI	*Entert_POI*	*Restaurant_POI*	*Shopping_POI*	*Travel_POI*	*Population*
Business_POI	1,0000					
Entert_POI	0,7729	1,0000				
Restaurant_POI	0,8399	0,8746	1,0000			
Shopping_POI	0,8142	0,8944	0,9207	1,0000		
Travel_POI	0,8218	0,8663	0,8327	0,7892	1,0000	
Population	0,7703	0,6601	0,6905	0,7208	0,6626	1,0000

Table 4.3 Correlations between variables

Based on the output exhibited in **Table 4.3**, population has the lowest correlation with expenditure. The two POI variables remaining in the model, 'Business' and 'Shopping', are strongly correlated, a warning sign that the model suffers from multicollinearity. Hence, in general, any of the POI variables could be separately used to get a sense about the total expenditure, but should be used with caution together, as they may have an impact on each other. Out of the three, 'Shopping POI' appears to be the predictor with the highest individual coefficient of determination, namely

55.1%. A model with 'Shopping POI' and 'Population' as independent variable is seen in **Table 4.4**. The coefficient the Population gets slightly closer to zero, but remains to be negative. All in all, the number of Points of Interest in a region is a better indicator of tourist expenditure than the population of a region.

Regression Statistics	
Multiple R	0,76501
R Square	0,58524
Adjusted R Square	0,574467
Standard Error	0,54248
Observations	80

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	31,97392	15,98696	54,32485	1,92826E-15
Residual	77	22,6599	0,294284		
Total	79	54,63382			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	2,262066	0,938913	2,409239	0,018375	0,392451001	4,131681675	0,392451001	4,131681675
Shopping_POI	1,143102	0,14138	8,085331	6,97E-12	0,861579087	1,424625199	0,861579087	1,424625199
Population	-0,26878	0,212058	-1,26746	0,208811	-0,691037219	0,153486842	-0,691037219	0,153486842

Table 4.4 Output of multiple linear regressions for expenditure in provinces - 2 independent variables

4.4 Customers' Expenditures in and out of Istanbul

With the objective of exploring the expenditure of customers in and outside of Istanbul, we extended the Customer Demographics table with features for fundamental statistics – mean, median of transactions, estimated number of days, average amount per day and per transaction – separately for transactions registered in and out of Istanbul. Apart from these variables, the expenditures on each of the merchant categories are forged into additional variables. The primary key in this table remains to be the anonymized customer identifiers, which constitutes the unit of analysis in this section. To assist the analysis, we created bins for the variables 'Age' and 'Income', with bin sizes of 15 years after 30 years of age and 2000 Lira, respectively. For instance, the first age bin is that of between 18 and 30 years, then 30-45, followed by 15 years of increments for each bin. The first income bin is that between 0 and 2000 Liras, with other bins also having a spread of 2000 TL. We also computed the proportion of out-of-Istanbul spending over the total overall expenditure, for each person.

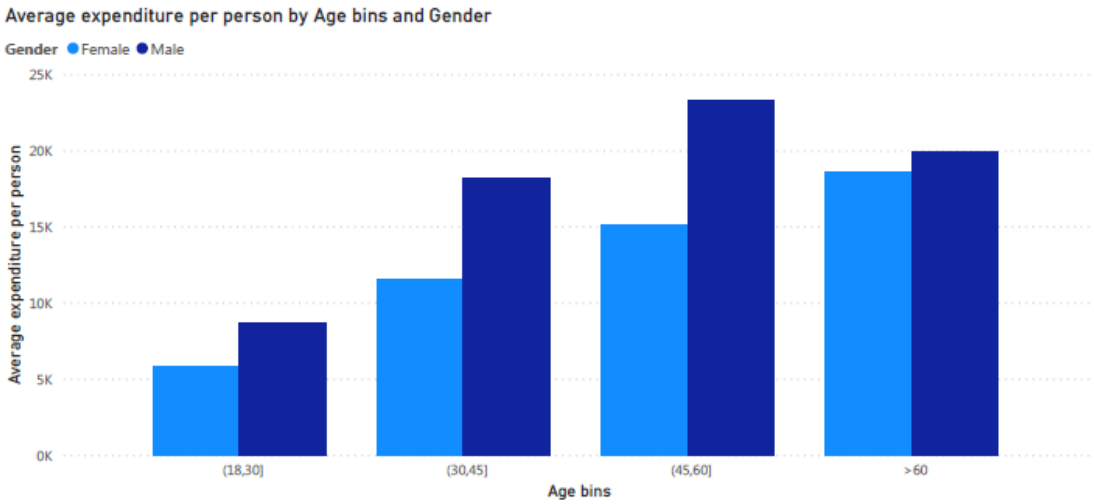

Figure 4.6 Credit Card Expenditures Per Person, By Age and Gender

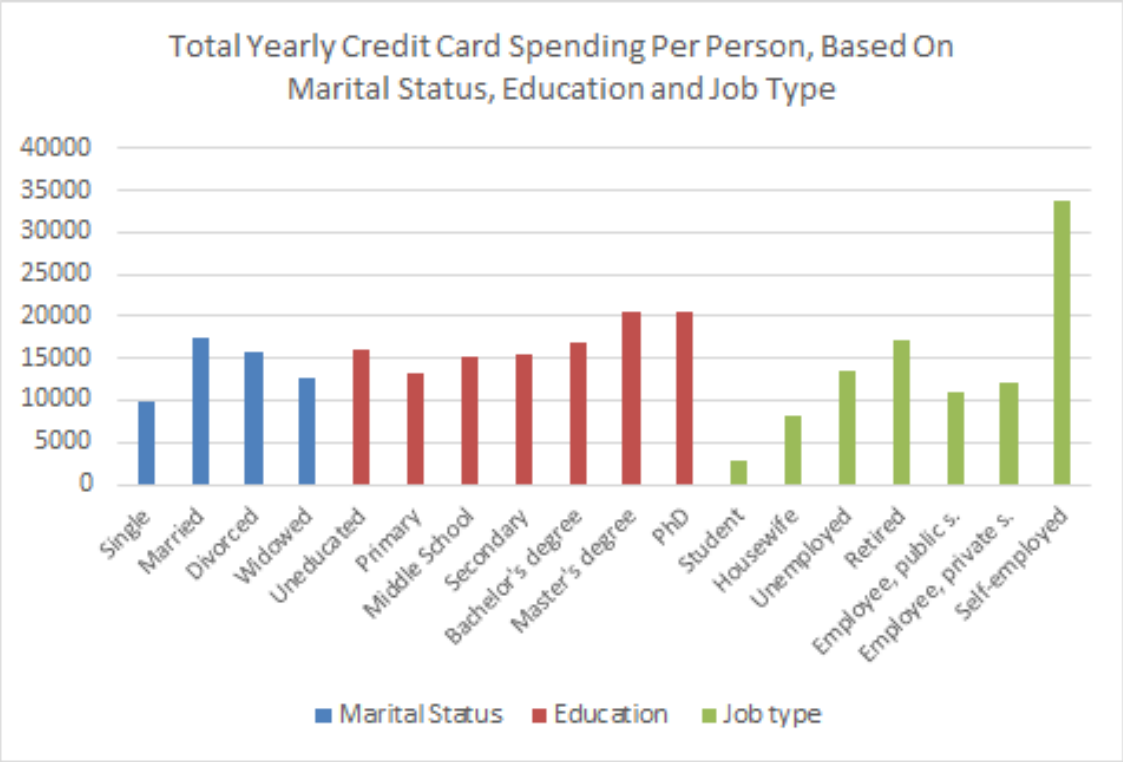

Figure 4.7 Credit Card Expenditures Per Person, By Marital Status, Education and Job Type

Figure 4.6 suggests that male customers' average total spending per person exceeds that of women, with the gap slightly closing at higher age groups. The distribution for male customers follows an left-skewed U-shape; for females, the expenditure monotonously increases. We can conclude that these findings are in accordance with theories of economic life-cycle, with income peaking for the middle-aged; the related

expenditure, however, seems to reach its highest point later for women, potentially because they still have the ability to do such physical activities as shopping.

We used the other categorical demographics attributes, namely 'Marital Status', 'Education' and 'Job type' to visualize further sample statistics for credit card expenditure. **Figure 4.7** presents some trends that could be expected to appear; however, one could consider as remarkable the fact that the retired has more outflow from their account than workers both in private and public sector, and uneducated people spend nearly as much on average as customers holding a bachelor's degree. Self-employed people are out front in terms of spending, suggesting that they might use their credit cards for purchases in relation to their enterprises.

Figure 4.8 presents a heatmap for different demographic groups' expenditure on the various merchant categories. Once again, categories with limited volume of transactions are removed, and so is the highest income group, due to their expense on insurance being an outlier, hindering the interpretability of the rest of the figure. Most of the categories follow the bell-shaped economic life-cycle with relation to age; however, expenditures on 'Health and Cosmetics', 'Apparel and Accessories' and 'Electrics and Electronics' do not start to decline after a certain age. No unexpected findings come from the income breakdown, as higher income seems to concur with higher spending on each category. Male customers appear to have a higher average spending on all categories, with the exception of 'Apparel and accessories'. In relation to marital status, singles appear to be light spenders with only 'Various grocery' being higher than some of the other demographic groups. Married people turn out to spend more in gas stations and supermarket, or opt for an insurance; divorced people, on the other hand spend more on health and cosmetics and clothing. Finally, the widowed appear to be the highest spenders on furniture and services.

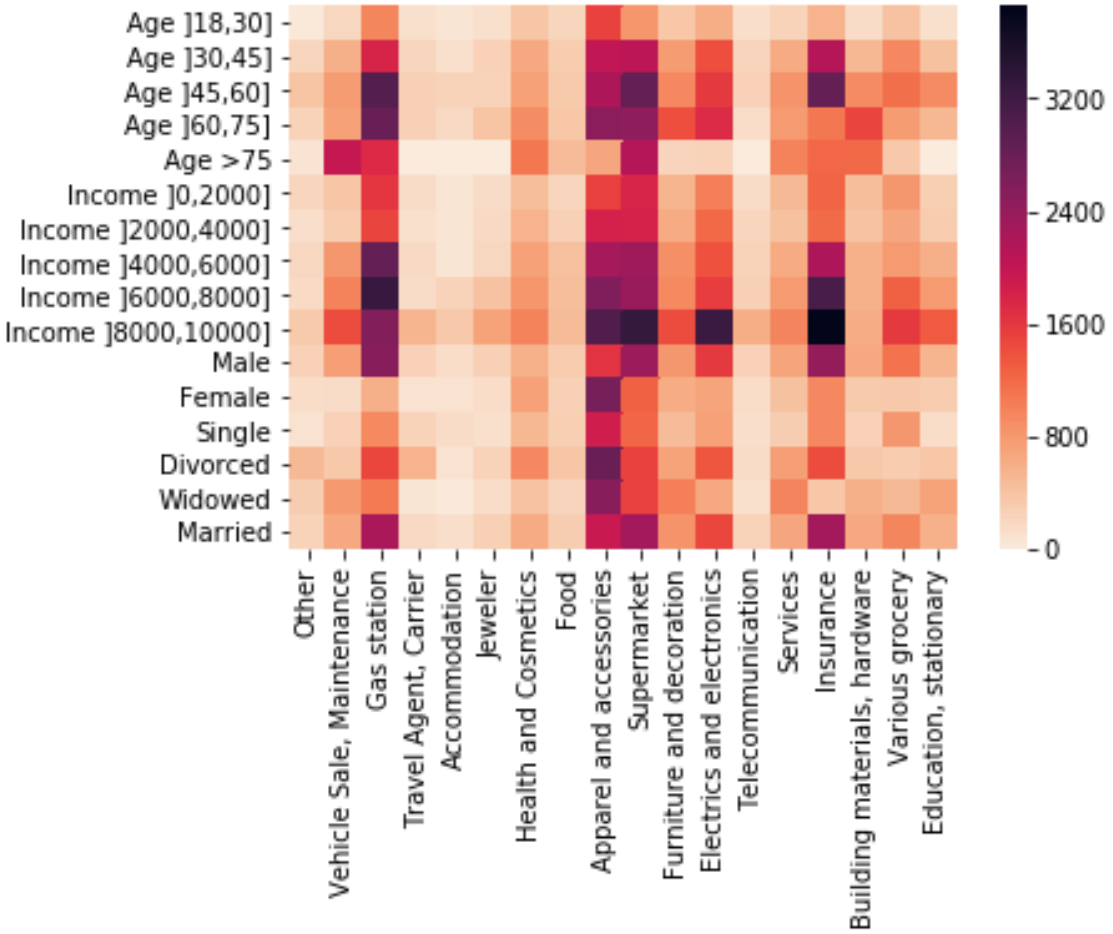

Figure 4.8 Average credit card expenditure per person, over various demographics

As a last point in question, we examined for each person the proportion of credit card expenditures outside of Istanbul over all expenditures, and examined the divergence of this proportion over various demographic segments. The corresponding bar graphs **Figure A9 - A11** are included in Appendix A.

The proportion of expenditures outside of Istanbul over the total overall expenditure seems to increase as the age increases, with the youngest age group spending a little higher proportion than the early-middle-age group. The increasing trend then slumps after 75 years, when we can assume the people to retire and travel less, hence the lower proportion.

As for the distribution with age as the x-axis, if we ignore the first bin, which may include non-workers, people falsely declaring or not declaring their income, we can observe a U-shaped distribution. The low 2000-4000 TL and the highest income bins had the largest proportions spent outside, while the middle-class was more prone to spend their money in Istanbul. One possible explanation could be that many of the blue-collar workers in Istanbul are migrants from other regions and they may

visit their family from time to time. Middle-class may afford some trips to abroad, decreasing the corresponding ratio, while the higher class might be able to afford several travels both in and out of Turkey.

Quite interestingly, students had the highest proportion of out-of-Istanbul expenditure, perhaps explained by the fact that they might live at their families' home in Istanbul and/or have their expenses covered by parents, but not out of Istanbul. Housewives, at the other extreme, preferred to spend their money mostly in Istanbul. Out of all groups determined by marital status, single people had the highest proportion spent out of Istanbul. Perhaps, these people had more opportunities to travel, not having any family related commitments. For education levels higher than high school, the proportion of outside-of-Istanbul expenses increases significantly, reaching almost as high as 12% for doctorate graduates. This observation is also in accordance with patterns in daily expenditures examined previously.

5. METHODOLOGY

The first two sections of this chapter comprises the methodology for the analysis based on unsupervised learning. As the name suggests, rather than making predictions about an output, we aim to infer patterns and discover segments or clusters based a selected array of features. In the subsequent two sections, we will introduce the methods we use to deal with a regression and a classification problem – i.e. problems requiring supervised learning. Our approach to these questions involves creating models to make predictions about an output variable of interest.

5.1 Clustering of Provinces

5.1.1 Hierarchical Clustering

Clustering refers to a machine learning technique of grouping data points together that are similar based on some features. Hierarchical clustering is a subtype of similarity or distance-based clustering (Murphy, 2012). While a top-down variant exists as well, we applied a bottom-up agglomerative approach, that is, the starting units of the algorithm are each and every data observation separately. Based on a predetermined dissimilarity measure, the bottom-up clustering algorithm keeps merging similar data points into joint clusters until only one large single group is created. In order to arrive at meaningful clusters, this process may be terminated at any desired level of inter-cluster and intra-cluster difference or based on another condition. While this algorithm fails to compete with K-Means in terms of time complexity, it is equally easy to implement and provides a more explicit understanding and interpretation through the dendrogram chart, presented later in this thesis.

5.1.2 Features and parameters of clustering

Our aim with the first clustering endeavor is to examine how differently people allocate their expenses over various product and service companies when they are travelling out of Istanbul. To this end, our basis of comparison is the percentages associated with each merchant category out of the total expenditures registered in Istanbul, totaled up for all customers in the sample. The units of analysis are the 81 provinces. For each province, the sum of expenditure on each merchant category was queried; these aggregate numbers were then converted into percentages, such that the total for each and every province is 100%. Thus, the input to the clustering analysis are the 24 merchant categories, expressed in percentage breakdowns for each province, which constitute the data observations.

In order to prevent any of the categories dominating over the other in terms of assigning data points to cluster, we normalized the whole dataset so that all variables have the same scale on which values can vary. Even though taking percentages could be a way of normalizing, by considering the deviations from the proportions in Istanbul, some of the categories with higher percentage deviation could be dominant in the clustering algorithm, case which is unwanted. This process is considered important in the cases when the researcher does not possess any knowledge about any of the variables being intrinsically more important (Kaufman & Rousseeuw, 1990).

As for the distance measure selected for our clustering, we opted for the default and widely used Euclidean distance. The criterion for merging is called the ward method. At each step, the algorithm aims to guarantee a minimal increase in the total within-cluster variance, presenting an alternative for the single-linkage algorithm (Ward, 1963). The chosen method is in line with our objective to achieve compact clusters with provinces being similar to each other within them.

In order to determine the number of clusters, which eventually sets the termination criterion for the hierarchical clustering algorithm, we use the elbow method. This approach involves plotting the variances – in our case, the 'Ward' variances – related to different number of clusters. The variance evidently falls as we increase the number of clusters. The ideal number of clusters is loosely defined to be located at the 'elbow' of the plotted chart: where the cost drops dramatically (Thorndike, 1953). Although the method is highly heuristic and vulnerable to criticism, it is strikingly intuitive and allows for a flexible analysis for all points that could be considered as 'elbow'.

Figure 5.1 presents the plotted elbow-method analysis. The blue line indicates

the Ward distances corresponding to each number of clusters. The orange-coloured line was plotted to indicate the acceleration of decrease of the blue line; that is, it is obtained through the second derivative of the distance line. The most dramatic drop in variance appears to be at k=2. This, however, might result in a very simplistic clustering output. Thus, we will examine the second peak in the acceleration curve at k=4 as well, which appears both to curb the total variance and to carry better interpretability than other further options at k=7 or k=11.

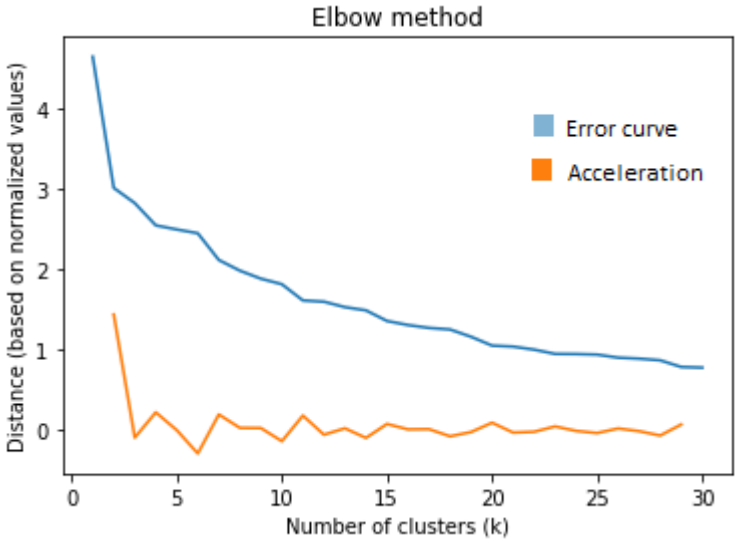

Figure 5.1 The elbow method executed for the clustering of provinces

5.2 Clustering of Trips by Purpose

5.2.1 Input variables of the algorithm

The second objective that involves a clustering-based solution is to group the extracted trips based on what we assume to be their associated purpose: leisure, business, or any of their subgroups, such as VFR. Up to our knowledge, no researcher has previously undertaken a similar analysis; only descriptive statistics of demographics related to different tourist motives are available. Hence, we rely partly on some non-discriminatory demographics variables as well as rationality to create new ones as inputs into the clustering algorithm. The input variables and their justifications are as follows:

X_1 - Average expenditure per day at destination: Moll-de Alba et al. (2016) found that business travelers had higher daily expenditure than leisure travelers; nonetheless, some of these expenses may be prepaid or covered by the sending organization, so as Brida & Scuderi (2013) suggests, the purpose in relation to expenditure is significant, but can be heterogeneous concerning the sign of the purpose coefficient and correlation. VFR spending, on the other hand, is clearly lower due to patronage from the visited family or friend and generally no accommodation cost (Seaton & Palmer, 1997).

X_2 - Trip duration: This variable appears to take lower values for business trips than leisure-related trips (Moll-de Alba et al., 2016), while VFR trips show variance country by country (Seaton & Palmer, 1997). In Turkey, we could expect a good many of VFR trips to last a bit longer, just around the length of religious holidays together with connecting days, that is, 7-10 days.

X_3 - Religious holiday: We attached a new variable to the Trips table, indicating whether the trip overlapped with any of the two main Islamic religious holidays in 2014. Eid Al-Fitr, or 'Ramadan Bayramı' fell between July 27 – July 30; Eid Al-Adha or 'Kurban Bayramı' was celebrated between October 4 – October 7 in 2014. If on any single day of the two holidays a transaction of the trip was registered, the variable took the value of 1. This variable is chiefly important for recognizing VFR trips, which were found to coincide mostly with national holidays (Seaton & Palmer, 1997).

X_4 - Summer: Perhaps arguable to an extent, we assumed trips happening in the summer to have a relatively higher probability to be undertaken with leisure purposes than, for instance, a trip in January to Antalya. The 'Summer' variable added to the Trips table takes the value of 1, if a part or the totality of the trip was registered between May – September, bounds included.

X_5 - Seaside: Similar to Summer, we cannot state in absolute terms that a trip with destination at a province with a seashore is related to leisure; still, it can be assumed to be related to the purpose of the trip, and trips that do not match this assumption would be assigned to the right cluster based on the remaining input variables.

X_6 - Airline expense 1 month prior to trip: We regard the means of transportation to the destination as a good indicator of the purpose, as light spender customers with families going on a VFR trip may prefer to travel by car. This variable was attached by querying not only the Credit Card Transactions table, but also the table set aside for online transactions. The expenses on airline companies were identified by the respective merchant category.

X_7 - Travel agent expense 1 month prior to trip: Using the same method as for X_6, we identified the trips with expenditure on the Travel Agent merchant category. We assume organized tours to be more related to leisure than business.

X_8 - Transit province recognized: Still in connection with the mean of transportation, this variable indicates if the trip has been marked as a 'Trip with transit', that is, if any province crossed towards the assumed destination has been detected.

X_9 - Number of 'Travel POI' at destination: One could assume that a considerable number of leisure-related holidays might involve the visit of an attraction or attractions. This input variable indicates the number of registered touristic points of interests at the destination.

X_{10} – Working – While we refrained from using possibly discriminatory demographic variables, such as gender, we aggregated a variable indicated whether we can assume the person to be working. A trip was only associated with the value '0' for this variable, if the person embarking on the trip was a student, unemployed, retired or housewife. This variable could assume to differentiate and assist to identify business-related trips.

5.2.2 Outliers and method

After a deep scrutiny of the Trips data table, we encountered some trips that we suspected not to be in line with the definition of tourism – they appeared to be rather permanent stays at an out-of-Istanbul location. Their commonality was their long duration and high total expenditure. This ascertainment led us to detect and delete outliers from the Trips data table.

We used the method of Standard Deviation (SD) to recognize outliers. This approach consists of labeling a data point as outlier if it falls beyond the limits of $\bar{x}$ $\pm$ 3 SD, or three times the standard deviation below and above the sample mean (Olewuezi, 2011). We applied this method on the data table for four variables that we found to be compelling:

- Trip duration (in days)
- Average expenditure per day
- Average expenditure per day at destination
- Average expenditure per transaction

The outlier elimination process reduced the 10 848 rows in the Trips table to 8 881 trips, separately saved as the Filtered Trips data table.

Once again, we normalized the input variables and used the agglomerative hierarchical clustering algorithm. To decide on the number of clusters, we relied on the graph for the Elbow method, as seen in **Figure 5.2**. The acceleration in decrease of the blue distance curve reaches its peak at k=5, thus, we will terminate the hierarchical clustering algorithm ones we reach five clusters, which are presented in Chapter 6.

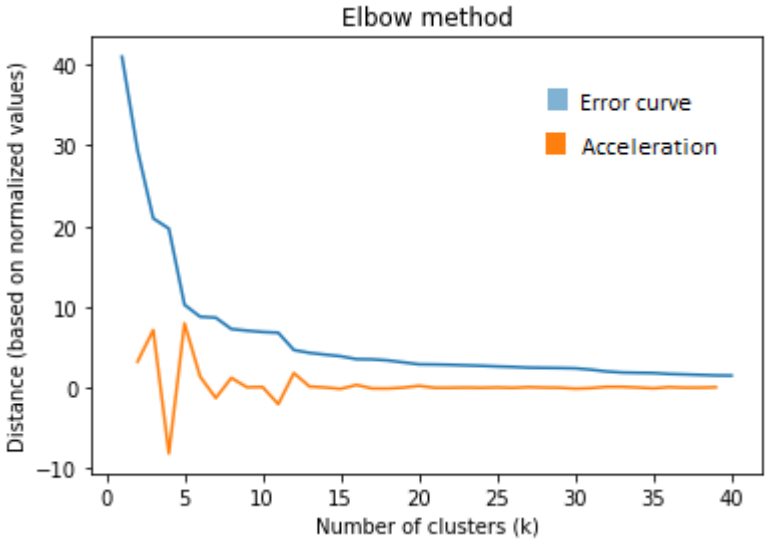

Figure 5.2 The elbow method executed for the clustering of trips

5.3 Predicting the Occurrence of a Trip

5.3.1 A Sliding Window Method

As our first objective involving predictions, we aim to find the best model to predict whether a customer would engage in a trip in the upcoming months. The knowledge of this may be useful to the bank, as they may direct related advertisements and promotions. They may also be inclined to provide information about branches and ATMs outside the regular living area of the customer. Finally, the bank may want to make sure that the customer is aware of any travel-related loans the bank may have to offer.

In practice, the bank or any other institution possessing credit card transaction data can only rely on past data to predict the future. Thus, a part of our data for a one-year period would not be known at some point in time by the bank, were we at any date within the range except for the last day. To simulate this practicality, we used a sliding window technique, that is a common approach in image processing and array-related operations, but could be reinterpreted and applied to time series data, like the one we possess.

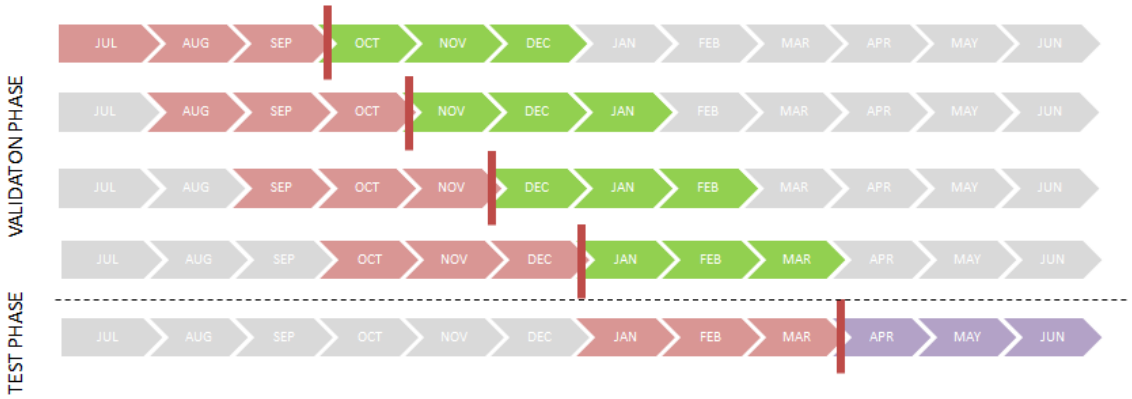

Figure 5.3 Sliding window method for predicting the occurrence of a trip

Figure 5.3 helps to describe the method used. In each scenario, represented by the horizontal timelines, we imagine being at the time that the red vertical lines mark. We reckon that predicting for the following three months is a sufficiently long time period for a good model to make better-than-random predictions. In each of these timelines, the red months represent the period, based on which behavioural predictors are collected; the green months represent the time frame, where the dependent variable either takes on 0 or 1. For each of these scenarios, the 10.000 customers are split into 75%-25% train and test sets.

In the validation phase, we deploy different models with different parameter settings, and use independent variables of three kinds:

1.1 Demographics, which are static for a customer

1.2 Financial behavior-related variables, counted based on the segments marked by red

1.3 Information about previous trips in the red segments

For the second and third type of variables, **Figure 5.3** indicates a 3 month-long period of observation, but we also experiment with time frames of 6 month.

After deploying several models with different settings in the validation phase, we

will obtain adequate knowledge to select the best predictors and high-performing algorithms to select the 'top 3' of models, which we would then implemented on the isolated test-phase timeline. According to this application, the best performing model and its performance measure is reported.

5.3.2 New features of prediction

In addition to the demographic variables described above, we also converted the letters marking customer risk in the Risk Scores table into numbers. Furthermore, we projected the home and work coordinates onto a map marking the different districts of Istanbul. We coupled the corresponding districts with the customers to form two new variables for the predictive analysis.

The nuance that differentiates the analysis in this and the following section is that we examine and quantify the behavioral patterns the customers demonstrate in Istanbul, and use them to predict their engagement in trips to other parts of Turkey. Firstly, we adopt the measure of diversity, used by Singh et al. (2015) through the following entropy equation:

$$D_i = \frac{-\sum_{j=1}^{N} p_{ij} * \log(p_{ij}))}{\log(M)}$$

Formula 5.1 Diversity formula

The formula takes different temporal or spatial bins into consideration, such as different days of the week or different districts. The term p_{ij} stands for the proportion of transactions that fall in bin j, counted for customer i. The numerator serves as an entropy measure, while the denominator ensures that the diversity is normalized by the base 10 logarithm of M, which denotes the number of non-empty bins. The resulting numbers fall between 0 and 1, with 1 indicating perfect diversity.

The diversity formula serves to indicate how equally customers distributed their expenses over different bins. The bins we will use to form diversity variables are as follows:

- Spending category diversity: Bins are represented by the 24 merchant groups.
- Spatial diversity: Bins are the 39 districts of Istanbul, so the diversity measure expresses spatial distribution of transactions
- Time of day diversity: The corresponding 24 bins are the hours of the day,

grouped with a bin width of one hour. The variable quantifies the level of equal distribution over different times of a day.

- Day of week diversity: The bins are the 7 days of the week, hence, the variable shows how equally the customer spends over different parts of the week.

The formula renders that the diversity measure will fall between 0 and 1. The diversity formula, on the downside, does not report on how many different bins the customer decided to spend within, e.g. how many different merchant categories the customer spends on. In order to introduce a variable with this purpose, we separately use the M measure from **Formula 5.1**, indicating the number of different bins that the customer engaged in within the selected time frame. We named this variable 'Variety'.

Along with these variables, we used summary variables related to some statistics of expenditures, as exhibited in **Table 5.1**.

Initial variables for prediction the occurrence of a trip

Demographic	**Financial Behaviour in Istanbul***	**Trips before***
Gender (Dummy)	Total expenditure*	Occurrence of a previous trip (binary)*
Marital status (Dummy)	Mean expenditure*	Total spending on previous trip*
Education (Dummy)	Median expenditure*	Purpose of previous trip*
Job status (Dummy)	Daily average expenditure*	
Income	Spending proportions (%) on the 24 merchant categories (24 variables)*	
Age		
Home district (Dummy)	Spending category diversity*	
Work District (Dummy)	Spending category variety*	
Years as customer	Spatial diversity*	
Risk level	Spatial variety*	
	Time of day diversity*	
	Time of day variety*	
	Day of week diversity*	
	Day of week variety*	

** Calculated over a specific period of time, according to the Sliding Window method*

Table 5.1 Features for predicting the occurrence of a trip

As expressed in **Table 5.1**, the three final input variables to our first predicting endeavor entails information about a customer's past trips, including the purpose of it, presumed based on the results of the clustering algorithm.

5.3.3 Algorithms, Feature Selection and Parameter Tuning

In order to find a well-performing model for our purpose, we intend to encounter the most suitable combination of settings for four different aspects: the size of the sliding window for observation, feature selection, the algorithm or statistical model and the value of hyperparameters in the model, if applicable. **Table 5.2** summarizes the alternatives for each decision that we consider. To reach the optimal model for the dataset, we try out and test every combination of them. We experiment both with 3- and 6- month long windows, that is, the variables in **Table 5.1** - marked with an asterisk - are to be calculated for both 3 months and 6 months prior to the time of prediction.

Settings for prediction of occurence of trips	
Size of window for observation	**Feature selection method**
3 months	Variable Importances (Random Forest)
6 months	Principle Component Analysis (PCA)
Algorithms & Paramteres to tune	
Logistic Regression	(1) C (2) Penalty term
Support Vector Machines (SVM)	(1) Kernel type (2) C
Random Forest	(1) Criterion (2) Number of estimators (3) Minimum samples (4) Maximum depth
ANN Deep Learning	-

Table 5.2 Settings for prediction of occurrence of trips

Feature selection is an essential process if one wants to assure that irrelevant variables are not included in the final model, decreasing the complexity and increasing the interpretability thereof. Selection of variables may be carried out through subset selection, such as the one we will apply based on variable importances, or through dimension reduction, like PCA (James, Witten, Hastie & Tibshirani, 2013).

Variable Importances (VI) are ingrained components of tree-based classifiers. The value of them is calculated considering the decrease in error when split by the given variable; it is expressed relative to the highest variable importance so that the values fall between 0 and 1. In other words, the VI of a feature indicates how much information we gain about the natural underlying separation of classes provided that we separate them based on that variable. While Variable Importances can be used in many arbitrary ways to decrease the number of features, we will persevere with selecting the variables up to the 95th percent level of cumulative variable importance,

added up based on features in a decreasing order for their importance.

Principle Component Analysis (PCA) consists of a process of reducing the dimensionality of the data by recurrently creating so-called principle components, along with the direction where the data points vary the most (James et al., 2013). The newly created axes are not as easy to interpret as axes strictly connected to a variable; nonetheless, the process reduces the complexity of the final model and has the potential to increase its accuracy. We will choose the number of components so that 95% of the variance remains to be explained by the new composition.

The research objective on hand presents a classification problem, and the dependent variable is binary: the customer either goes on a trip (1) or does not (0). The statistical models are picked and adjusted in accordance. For the two methods, namely SVM and ANN, we performed feature scaling, based on distance and weight-based algorithms.

Logistic Regression models the probability of the dependent variable belonging to either of the two classes. It is similar to linear regression but goes through a logit transformation. The two hyperparameters we will tune are related to a penalty or regularization term λ that discourages overfitting. L1 or Lasso regularization allows unimportant variables' coefficients to shrink to zero, while L2 or Ridge regularization does not. The value of C is inversely positioned with respect to λ, which determines the strength of regularization.

Support Vector Machine (SVM) finds a hyperplane to separate the two classes based on a specific number of keystone observations, known as support vectors. The C parameter controls the size of the margin and poses a penalty to data points violating the margin constraint – it is tuned to find the balance between variance and bias. The kernel, on the other hand, allows for a more accurate model while dealing with non-linear problems, as an RBF kernel maps the dataset to a higher dimensional space to look for separating hyperplanes that are potentially nonlinear in the normal dimensions.

Random Forests are ensemble models consisting of several decision trees that are decorrelated by choosing a different random set of parameters for each tree and deciding based on majority vote. The two possible criteria are the Gini-index and the Entropy, helping to decide on the variable, based on which the next split would be executed. The number of estimators parameter refers to the number of trees used in the model, with higher number of trees presenting the opportunity for lower error. The minimum samples and maximum depth parameters are responsible for the bias-variance tradeoff by controlling how many observations there must be in a

leaf minimally, and how deep the tree can grow and fit to the given dataset.

Artificial Neural Networks (ANN) uses neural nets, with hidden layer(s) between input and output layers, recurrently giving importance weights to neurons and transforming the inputs via an activation function. While many aspects could be changed in ANNs for optimization, the decision mechanism is often heuristic and computationally expensive. In our analysis, we persevere with an ANN structure of 2 hidden layers with 6 neurons each with Rectified Linear Unit – 'relu' - as its activation function in hidden layers and sigmoid function in the output layer, such that the output is a probability. We iterate through 200 epochs.

We conduct the tuning of hyperparameters on a grid basis – GridSearchCV from the SciKit library - with the algorithm fitting every combination of values and deciding based on the cross-validated accuracies.

Within a loop, a separate data table is established for each window of time period. Evidently, the dependent variable will be either 1 or 0 based on whether a trip occurred in the following 3 months after the time of prediction. Subsequently, the customers are randomly split into a train and test set with a proportion of 75%-25%, respectively. We found it unnecessary to create a separate validation set, given that the methodology of sliding windows incorporates a validation phase.

After removing 4 customers with no transactions registered at all, 4.676 customers appear to have engaged in trips over a year, and 5.320 did not. This implies a relatively balanced dataset on the whole, however, for the three-month periods of prediction in the sliding window method, the number of people not embarking on trips is considerably higher. For instance, in the first time band with 3 months of observation - that is, in the months of October, November and December - only 1.183 people went on a trip out of 10.000. To address this issue, we applied a simple oversampling method to the minority class, where actual data points are re-sampled several times to correct the imbalance. This sampling technique is deployed on the train set, such that the test set remains the same. In this way, the algorithms do not incline toward a majority class.

5.4 Predicting the Expenditure During Trip

5.4.1 Sliding Window Method

Our second predictive goal is the topic profoundly explored in various research papers: predicting tourist expenditure. The novelty of our analysis is the use of transaction data to estimate not only the dependent variable, trip expenditure, but also the dependent variables, such as financial behaviour displayed in Istanbul. Inherently, many of the estimations may be inaccurate, e.g. we cannot track transactions done with cards of other banks or the merchant categories the person spent cash on, rather than swiping a bank card.

To this end, we turn to the sliding window method once more, with the modification that there is no need to observe subsequent months. Rather than compiling a separate data table at the end of different months, we operate starting with one single data table of all extracted trips, and the window of observation refers to the 3 or 6 months prior to the starting date of each trip. The method is described in **Figure 5.4**.

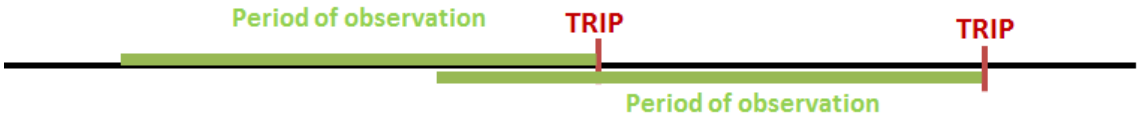

Figure 5.4 Sliding windows for prediction of trip expenditure

5.4.2 Input features

This second predictive venture aims to evaluate whether data analysis based on transaction data could be at odds with the predicting power gained from surveys, which have been the fundamental data sources in all related research papers (Brida & Scuderi, 2013). Our compiled data table includes some of the independent variables that these traditional survey methods lean on; we also introduce behavioral variables that these surveys rarely resort to or could not possibly measure without considerable bias. The complete list of variables used is exhibited in **Table 5.3**. Compared to the first prediction, we adjoined variables already clarified in the clustering method

for the purpose of a trip. Additionally, we query the ATM Transactions data table for net withdrawals in the last 3 days before the trip, as well as during the trip. We also add trip duration, among other variables that frequently appear in survey-based research.

Initial variables for prediction the occurrence of a trip

Demographic	**Financial Behaviour in Istanbul***
Gender (Dummy)	Total expenditure*
Marital status (Dummy)	Mean expenditure*
Education (Dummy)	Median expenditure*
Job status (Dummy)	Daily average expenditure*
Income	Spending proportions (%) on the 24 merchant categories (24 variables)*
Age	
Home district (Dummy)	Spending category diversity*
Work District (Dummy)	Spending category variety*
Years as customer	Spatial diversity*
Risk level	Spatial variety*
	Time of day diversity*
Trip-related	Time of day variety*
ATM withdrawals 3 days before	Day of week diversity*
Airline spending 1 month before	Day of week variety*
Travel Agent spending 1 month before	
Car Rental 1 week before	**Financial Behaviour out of Istanbul***
ATM withdrawals during trip	Total expenditure*
Trip duration	Mean expenditure*
Religious holiday	Median expenditure*
Summer	Daily average expenditure*
Seaside	Spending proportions (%) on the 24 merchant categories (24 variables)*
Number of 'Travel POI' at destination	
Neighbouring province	
Destination province (81 Dummy)	
Province Cluster (from Clustering 1, Dummy)	
Distance to destination	
Transit	
Number of previous trips*	
Total expense in previous trips*	

** Calculated over a specific period of time, according to the Sliding Window method*

Table 5.3 Features for predicting trip expenditure

5.4.3 Algorithms, Feature Selection and Parameter Tuning

Unlike the first case of prediction, the dependent variable being a continuous one renders a regression-type problem. The outline of methods to address this problem is given in **Table 5.4**.

There are methods that can be reused for a regression problem too, with small reinterpretation. Linear regression is similar to Logistic Regression, without a Sigmoid function converting continuous scores into probabilities. As for the Random Forest model, the variable importances are calculated based on decreasing variances, rather than changes in the Gini-index, used in the classification problem. Accordingly, the Random Forest model does not require a splitting criterion to be scpecified anymore.

The ANN model is set up with the same settings as in the classification problem: 2 hidden layers with 6 neurons in each of them, iterated through 200 epochs. Activation function in the hidden layers remains to be the 'relu' function; however, we do not use a binary activation function or any other function at the output layer as we are looking for continuous outputs. The independent variables were scaled with the library 'StandardScaler' before using them as inputs; the dependent variable was not normalized and de-normalized, as it does not present a similar bias as independent variables on different scales.

Settings for prediction of trip expenditure	
Size of window for observation	**Feature selection method**
3 months	Variable Importances (Random Forest)
6 months	Principle Component Analysis (PCA)
Algorithms & Paramteres to tune	
Linear Regression	-
Random Forest	(1) Number of estimators (2) Minimum samples (3) Maximum depth
ANN Deep Learning	-

Table 5.4 Settings for predicting trip expenditure

6. RESULTS AND DISCUSSION

In this chapter, we provide our findings in relation to the research objectives stated before and for which we have described the applicable machine learning methods. Hence, the first two sections present the outcome of the clustering undertakings, followed by the two sections of predictive analysis on classification and regression problems.

6.1 Clustering of Provinces

After a heuristic investigation on what results we would achieve with the number of clusters being either 2 or 4, we opted for reporting the results for a k=4 setting, due to the additional two clusters being endowed with valuable information load.

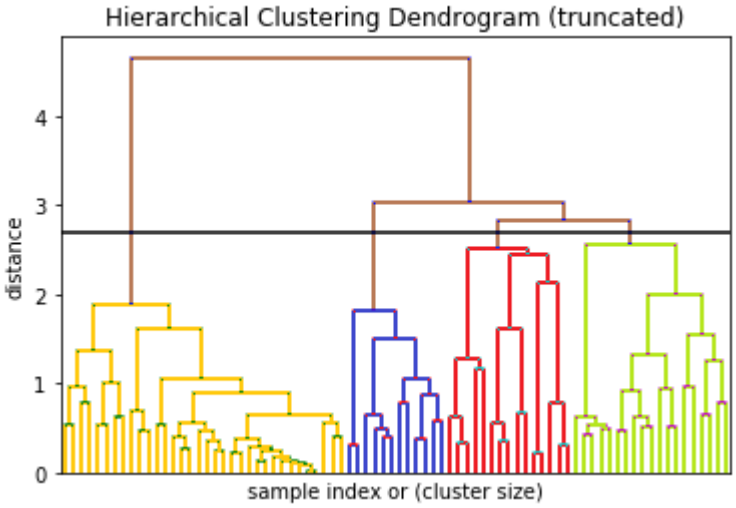

Figure 6.1 Dendrogram for clustering of provinces

We exhibited the process that the hierarchical clustering algorithm follows in **Figure 6.1**. This type of plot is referred to as a dendrogram; studying the figure from

the bottom towards the top, we can follow how the algorithm merges the data observations - i.e. provinces - one by one at each step. The grey horizontal line slightly above the distance of 2.6 cuts across four vertical line segments, indicating that four clusters are created, which are also illustrated with different colours. The dendrogram gives an early insight about the size and distance of clusters. For instance, the cluster marked with green appears to consist of the largest number of observations, while the red cluster seems to be the 'tightest', it having the lowest within-cluster variance.

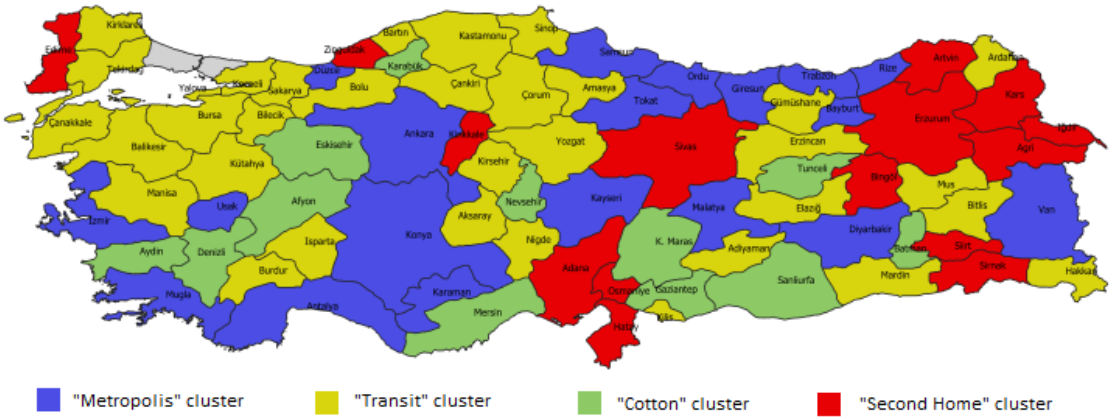

Figure 6.2 Clusters of provinces based on category expenditures

After accordingly labeling the provinces, we projected the result onto the map of Turkey, as seen in **Figure 6.2**. As a reminder of the method leading to this result, it should be highlighted that the provinces in each cluster are not primarily similar because of similar proportions spent on categories; we observe these clusters, because their members differ from the expenditure allocation of Istanbul in a similar way.

Table 6.1 summarizes the percentages of expenditure on each merchant category for the four clusters. Only merchant categories corresponding to significant transaction volumes are included. Additionally, we compiled **Table 6.2** to demonstrate the distribution of various demographics within each cluster. The unit of measurement in this latter table is the related expenditures; in other words, a 17.03% at the intersection of 'Single' and 'Metropolis cluster' indicates that 17.03% of the expenditures in the cluster can be attributed to single people.

Based on the segmentation output of the clustering algorithm, we named the resulting four clusters as follows:

- „Metropolis" cluster (in blue in **Figure 6.2**): The cluster comprises provinces with the largest cities in Turkey apart from Istanbul, such as Izmir, Ankara and Antalya. These large urban areas generally lie at the Aegean and Black

Sea coasts, with a few exceptions like Konya or Ankara. In these provinces, the use of credit cards is rather diverse, similar to that in Istanbul. Payments on vehicle repair, furniture, healthcare and cosmetics or education and stationery are more significant here than in other clusters, once again, similarly to Istanbul.

More educated people choose to visit provinces in this cluster compared to other province groups. Also, compared to the other clusters, we can observe a more diverse mixture of professions, marital statuses and job types. Big cities offer opportunities of business and leisure for different types of people. However, these urban areas might happen to be more expensive – the cluster's visitors have the highest mean income.

- „Second home" cluster (in red in **Figure 6.2**): The cluster includes provinces, such as Sivas or Erzurum, from where the highest number of migrants come to Istanbul (CNN, 2018). This peculiarity is also legitimate for some of the Black Sea provinces, however, provinces in the "second home" cluster are less metropolitan than the Black Sea area. Supermarket and electronics shopping, as well as telecommunication and general services constitute a remarkable proportion of spending compared to other clusters. Based on all of these observations, we could assume that there may be a higher probability of spotting VFR travelers in this area.

  Most of the expenses are incurred by people who possess lower education level and are single; a relatively higher proportion of them compared to other cluster statistics are unemployed, student or retired, or work in the private sector. We could assume that some of these people relocated to Istanbul for better opportunities in jobs or education, supporting the hypothesis that these areas are the homeland of many residents of Istanbul, but not as metropolitan and with less infrastructure than the Black Sea provinces.

- „Cotton" cluster (in green in **Figure 6.2**): The cluster merges provinces that can be associated with more than 50% of Turkey's cotton production, namely Şanlıurfa in the Southeast Anatolian region and Aydın in the Aegean region (TAGEM, 2018). The predominant purchase category is clothing; interestingly, accommodation is also a remarkable category, which could be potentially explained by residents of Istanbul visiting these areas particularly in order to shop around. Additionally, the category of educational and stationary products and services prevail in the cluster.

  The first guess may be that this cluster elicits visits with business purposes due

to its high production capacity. Indeed, among self-employed people, this area is highly favored compared to other categories, but there is also a significant presence of workers in the public sphere.

- „Transit" cluster (in yellow in **Figure 6.2**): The prevalent spending category in these regions is fuel. We can assume that these provinces are used as transit areas to reach other destinations. Their locations being inland or close to Istanbul confirms this assumption.

  This segment of provinces is mostly visited by the married and men, working in the private sector. The spenders in these clusters having the lowest income average and being mostly married could indicate that they may prefer to embark on trips by car, rather than buying costly air tickets for each and every family member.

	"Metropolis" cluster (blue)	"Second home" cluster (red)	"Cotton" cluster (green)	"Transit" cluster (yellow)
Vehicle Sale, Maintanance	14,40%	2,28%	0,90%	2,43%
Gas station	14,11%	10,92%	16,63%	48,30%
Travel Agent and Carrier	5,68%	0,28%	0,78%	0,52%
Accomodation	1,99%	0,68%	7,90%	1,99%
Jeweler	2,90%	0,09%	2,74%	2,70%
Health and Cosmetics	2,17%	1,83%	2,14%	1,21%
Food	1,20%	1,05%	1,57%	2,13%
Apparel and Accessory	11,32%	8,38%	33,23%	7,16%
Supermarket	9,17%	16,91%	6,26%	10,09%
Furniture and Decoration	11,56%	2,65%	1,87%	2,50%
Electrics and Electronics	6,31%	15,29%	9,52%	6,15%
Telecommunication	0,49%	6,94%	0,24%	0,73%
Services	2,83%	12,60%	0,80%	2,54%
Insurance	0,34%	0,14%	0,18%	0,10%
Building Materials, Hardware	9,03%	0,89%	3,21%	2,54%
Various Grocery	1,24%	13,50%	4,15%	3,48%
Clubs and Organizations	0,14%	0,02%	0,02%	0,02%
Education, Stationary	1,57%	0,57%	1,31%	0,31%

Table 6.1 Province clusters and category proportions

We present the goodness of fit statistics in **Table 6.3**. The within-cluster distances are based on the average diameter distance, that is, the average of distances between each and every pair of observations included. The between-cluster distances are presented in terms of single linkage, or the closest distance between two observations in two different clusters. On average, the 'Transit' cluster appears to be the densest one and also the most dissimilar to other clusters. On the other hand, the 'Cotton' cluster and the 'Metropolis' cluster are more similar to one another.

	"Metropolis" cluster (blue)	"Second home" cluster (red)	"Cotton" cluster (green)	"Transit" cluster (yellow)
Single	17,03%	24,61%	17,32%	10,20%
Divorced	6,57%	0,80%	6,05%	2,23%
Widowed	0,15%	0,07%	0,20%	0,39%
Married	76,25%	74,52%	76,43%	87,19%
Male	74,90%	82,37%	74,43%	87,53%
Female	25,10%	17,63%	25,57%	12,47%
PhD	1,37%	0,13%	1,29%	0,36%
Uneducated	0,72%	0,80%	1,39%	0,86%
Primary school	4,21%	13,98%	4,06%	3,84%
Master's degree	6,87%	4,51%	8,37%	3,71%
Secondary school	36,25%	42,71%	34,01%	39,71%
Middle school	5,52%	9,70%	8,57%	8,83%
Bachelor's degree	38,37%	25,03%	34,51%	37,07%
Retired	7,05%	1,73%	9,55%	5,94%
Housewife	0,50%	0,02%	0,95%	1,36%
Self-employed	26,44%	18,31%	28,04%	22,11%
Retired but working	1,72%	12,94%	1,68%	4,18%
Unemployed	0,39%	1,15%	0,94%	0,21%
Student	0,01%	0,14%	0,00%	0,00%
Employee, public s.	12,19%	7,14%	16,16%	9,29%
Employee, private s.	49,94%	52,35%	41,62%	55,43%
Age (mean)	40,69	40,02	40,09	41,12
Income (mean)	5611,87	5039,92	5113,34	4566,23

Table 6.2 Province clusters and demographics

Cluster sizes		**Intra-cluster distances**	
Second Home	15	Second Home	1,2839
Transit	34	Transit	0,6538
Cotton	12	Cotton	0,8868
Metropolis	19	Metropolis	1,0173

Inter-cluster distances

	Second Home	Transit	Cotton	Metropolis
Second Hom	0			
Transit	0,6589	0		
Cotton	0,7115	0,6886	0	
Metropolis	0,7769	0,6657	0,5756	0

Table 6.3 Goodness of fit for province clusters

6.2 Clustering of Trips by Purpose

A simplified dendrogram for the hierarchical clustering algorithm deployed on the trip-based data is seen in **Figure 6.3**. Most of the merges take place under the distance of 1, which explains why only a few observations are visible to the naked

eye in the dendrogram. Differences between the final 5 clusters are remarkably big despite having normalized the data – this may indicate a good separation of dense clusters.

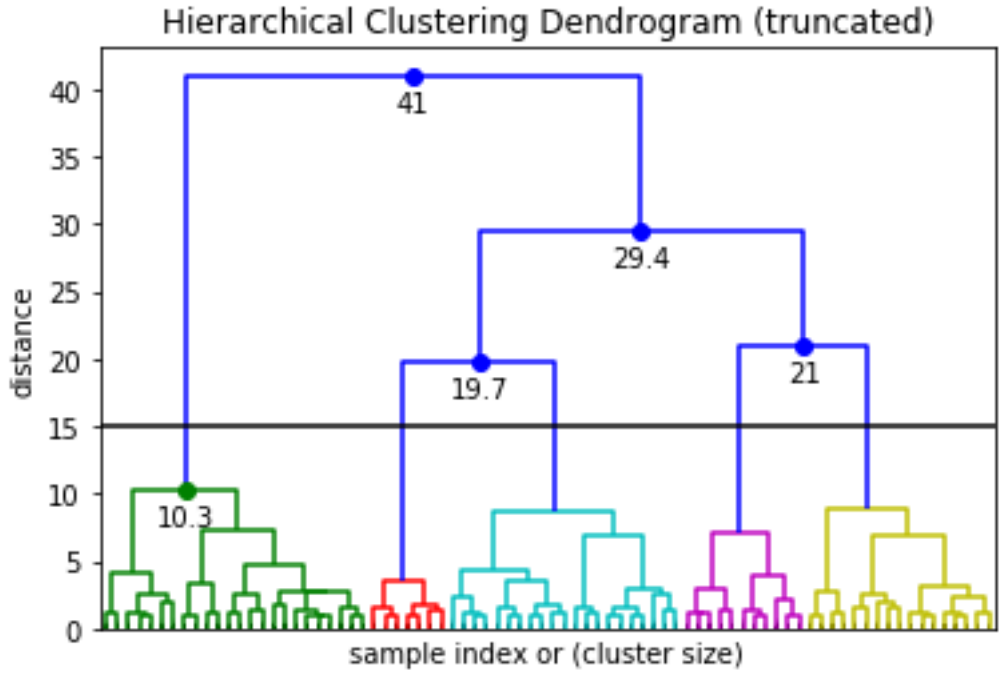

Figure 6.3 Dendrogram for clustering of trips

The resulting cluster labels are attached to the Filtered Trips data table and cluster statistics are extracted and exhibited in **Table 6.4**. Out of these non-demographic attributes, we marked in bold the ones that have been used as input variables in the clustering algorithm. The table contains average amounts and proportions for fundamental trip-related aspects, expenditures on merchant categories, attributes describing the time of trip, the destination and the mean of transportation, in addition to other aspects. **Table 6.5** presents a breakdown for demographic variables related to the five clusters, with education, job type, marital status and gender as categorical attributes, in addition to the mean age and mean income, which are continuous.

Accordingly, we named the five clusters as follows:

- "Leisure Holiday": Compared to other clusters, these trips occurred more often during the summer – along with 'Package Holiday' - and the destination of many had the most touristic points of interest. Additionally, this cluster had the highest proportion of seaside municipalities as destination. A good proportion of the travelers appear to have used cars, but some also had airline costs. People going on these kinds of trips stayed the longest and spent less

on a daily basis, than business travelers, in accordance with the indications of Moll-de Alba et al. (2016). The proportion of singles is relatively high compared to other trip clusters, and the leisure tourists had comparatively high expenditure on restaurants. Interestingly, the average of income is the lowest along with VFRs.

		VFR	Business	Acquisitions	Leisure	Package holiday
Fundamental statistics	**Length (days)**	4,34	2,91	2,42	3,19	2,82
	Expenditure (TL)	216,48	267,31	496,58	288,27	266,77
	Expenditure per day (TL)	76,78	143,34	261,54	136,04	116,34
	Expenditure per day at destination (TL)	79,71	145,53	271,85	142,33	120,38
	Number of transactions	3,04	2,53	2,41	2,41	2,85
	Expenditure per transaction (TL)	76,54	119,76	216,97	129,09	105,95
	Expenditure per transaction at destination(T	76,30	119,76	216,94	127,93	106,10
	Distance (km)	350,64	634,25	563,53	438,29	552,02
Merchant categories	Other / Commercial equipments etc.	0,44	0,19	4,18	2,83	0,35
	Vehicle Sale, Maintenance (TL)	5,17	5,94	18,02	4,42	0,13
	Gas Station (TL)	59,59	71,20	200,81	96,81	78,73
	Travel Agent and Carrier (TL)	3,51	3,88	7,16	4,37	4,05
	Accommodation (TL)	8,28	10,90	13,26	7,46	5,41
	Food (TL)	10,66	12,41	8,12	8,77	8,94
	Apparel and Accessory (TL)	34,95	60,17	54,26	44,65	47,99
	Supermarket (TL)	35,91	37,08	41,80	42,17	38,09
	Furniture and Decoration (TL)	8,12	7,97	16,49	13,71	10,03
	Electrics and Electronics (TL)	9,88	14,59	34,02	13,88	22,24
	Services (TL)	5,01	4,81	17,17	6,82	8,55
	Building Materials, Hardware (TL)	8,48	8,72	41,74	13,63	1,67
	Clubs and Organizations (TL)	0,54	0,00	0,10	1,00	0,00
	Education, Stationary (TL)	3,79	3,02	1,98	5,07	2,27
Transp.	**Transit (%)**	7,34%	4,81%	9,48%	8,47%	5,91%
	Airline expense 1 month prior to trip (TL)	19,79	647,46	3,98	0,93	187,06
Time	**During religious holiday (%)**	6,00%	4,34%	4,93%	6,90%	4,84%
	During summer season (%)	60,58%	53,33%	54,39%	59,60%	68,82%
Destination	**Destination with coastline (%)**	83,63%	64,03%	54,14%	71,94%	58,60%
	Number of touristic POI	844,57	378,74	116,89	222,08	246,74
	Destination in 'Metropolis cluster' (%)	40,02%	38,60%	23,50%	23,57%	25,27%
	Destination in 'Second Home cluster'(%)	5,46%	16,74%	13,14%	11,78%	17,20%
	Destination in 'Cotton cluster' (%)	4,55%	11,94%	10,68%	7,35%	11,83%
	Destination in 'Transit cluster' (%)	49,97%	32,71%	52,68%	57,30%	45,70%
Other	**Working person (%)**	91,93%	93,02%	91,79%	90,97%	94,62%
	Travel Agent expense 1 month prior to trip (	26,62	87,10	11,45	1,18	1667,68

Table 6.4 Trip clusters and non-demographic statistics

- "Business Trip": A good amount of these trips can be assumed to have been undertaken by airplane, a convenient mean of transportation for business people. A large proportion of these trips had the "Metropolis cluster" as their destination, but also relatively high proportion of them travelled to the 'Cotton cluster', which we assume to attract business people for certain industries. People embarking on business trips stayed for few days and had moderate spending, which could be explained by the observations, that some of the expenses are prepaid, such as the air ticket and perhaps accommodation. The

percentage of working people on these trips is one of the highest; they possess the highest average income and are the youngest group, with a considerable fraction of them being single. They are also the most educated, and nearly a quarter of them are self-employed.

- "Acquisitions": These trips seem to be similar to Business Trips, however, the expenditures appear significantly higher. The proportion of money spent on vehicles, building materials and hardware, as well as other commercial equipment is higher than in other clusters. There is evidence for land transportation being the main mode of transportation with high spending on fuel. Despite the high expenditure, the number of transactions is not higher than in other categories. Hence, we deduce that the trip may be purchase-oriented, with the high amounts indicating that some of these may be some sort of business acquisition. A good fraction of them are committed by the self-employed with strong dominance of males, but many of them have been in fact enrolled only in lower education.

- "Visiting Friends and Relatives (VFR)": Compared to other clusters, these trips happened more frequently during religious holidays. These travellers seem to have often moved around by car, and beside visiting the nearby transit areas, they ended up in all three other clusters of provinces, given that migrants in Istanbul have roots from all over the country. These trips were rather short in accordance with the observation of Seaton & Palmer (1997), and expenditure was moderate. Spending in supermarkets appears to be comparatively high, indicating a more similar spending attitude to what the tourist might have at home. VFR travellers are among the people with lowest income, are less educated and generally are married.

- "Package Holiday": The amount spent on travel agents beforehand is relatively high and some of the travellers flew by plane. The length of the trips is moderate. This cluster includes trips to destinations that are not by the coastline but have a relatively good number of touristic points of interest. This cluster demonstrates the highest age average, as may be the case for older people who prefer to have their holidays pre-arranged via some kind of travel agency.

Similar to the clustering of provinces, we scrutinized the goodness of fit of the emerging clusters with average within-cluster distance and between-cluster distance on a single-linkage basis in **Table 6.6**.

As previously deduced from statistics (TÜIK, 2019a), more than half of domestic

	VFR	Business	Acquisitions	Leisure	Package holiday
Marital Status					
Single	18,87%	19,84%	10,49%	14,25%	16,13%
Unknown	2,11%	2,17%	1,90%	1,57%	0,54%
Divorced	4,70%	3,57%	3,79%	2,81%	2,15%
Widowed	0,88%	0,16%	0,44%	0,45%	0,54%
Married	73,45%	74,26%	83,39%	80,92%	80,65%
Education					
PhD	0,66%	1,86%	0,69%	0,95%	5,38%
Uneducated	1,34%	0,62%	0,88%	1,12%	1,61%
Primary School	3,31%	4,34%	7,77%	4,71%	1,08%
Master's degree	5,98%	9,61%	3,60%	3,93%	8,06%
Secondary school	33,32%	30,85%	44,09%	40,35%	32,80%
Middle school	7,07%	4,19%	9,10%	7,41%	9,68%
Bachelor's degree	39,02%	40,78%	26,47%	32,55%	38,17%
College	9,31%	7,75%	7,39%	8,98%	3,23%
Gender					
Male	70,31%	74,57%	83,39%	78,62%	78,49%
Female	29,69%	25,43%	16,61%	21,38%	21,51%
Job type					
Retired self-employe	1,11%	1,24%	2,08%	1,01%	1,08%
Retired employed	2,65%	1,40%	3,22%	2,24%	1,61%
Unemployed	0,38%	0,31%	0,51%	0,51%	0,00%
Other	0,85%	0,93%	0,38%	0,51%	1,08%
Retired	6,81%	5,89%	7,14%	8,02%	4,84%
Housewife	0,70%	0,62%	0,44%	0,45%	0,54%
Student	0,17%	0,16%	0,13%	0,06%	0,00%
Self-employed	15,65%	23,88%	20,47%	18,41%	22,04%
Unknown	0,11%	0,16%	0,32%	0,06%	0,54%
Employee (public)	9,95%	12,56%	7,58%	11,28%	12,37%
Employee (private)	61,62%	52,87%	57,74%	57,46%	55,91%

Table 6.5 Trip clusters and demographic statistics

travellers could be expected to visit friends and family. This statistic was measured for the whole population of Turkey, rather than for the residents of Istanbul. Nevertheless, it is clear that the Leisure Holiday and VFR clusters we found are very near to each other in terms of within-cluster distance and there might be some overlap that balances out the dominance of Leisure Holiday in number. These clusters are not only closely adjacent but also the densest. Acquisition-related trips are not very distant from VFR trips either. Business and Organized Trips account for the smallest clusters in number of data observations.

We must acknowledge the limitations of the analysis in this section as we did not possess any labels to audit the accuracy of the clusters. The framework is therefore

Cluster sizes		Intra-cluster distances	
Leisure Holiday	4685	Leisure Holiday	0,1845
VFR	1782	VFR	0,1858
Acquisition	1583	Acquisition	0,2735
Business	645	Business	0,5084
Organized Trip	186	Organized Trip	0,3280

Inter-cluster distances

	Acquisition	Leisure Holiday	VFR	Business	Organized Trip
Acquisition	0				
Leisure Holiday	0,056	0			
VFR	0,002	0,003	0		
Business	0,049	0,0281	0,2856	0	
Organized Trip	0,0764	0,032	0,433	0,2182	0

Table 6.6 Goodness of fit for trip clusters

yet to be put to the test for further research and may be modified accordingly based on the findings. Yet, we find that many of the statistics concur with common sense and indications of previous descriptive research.

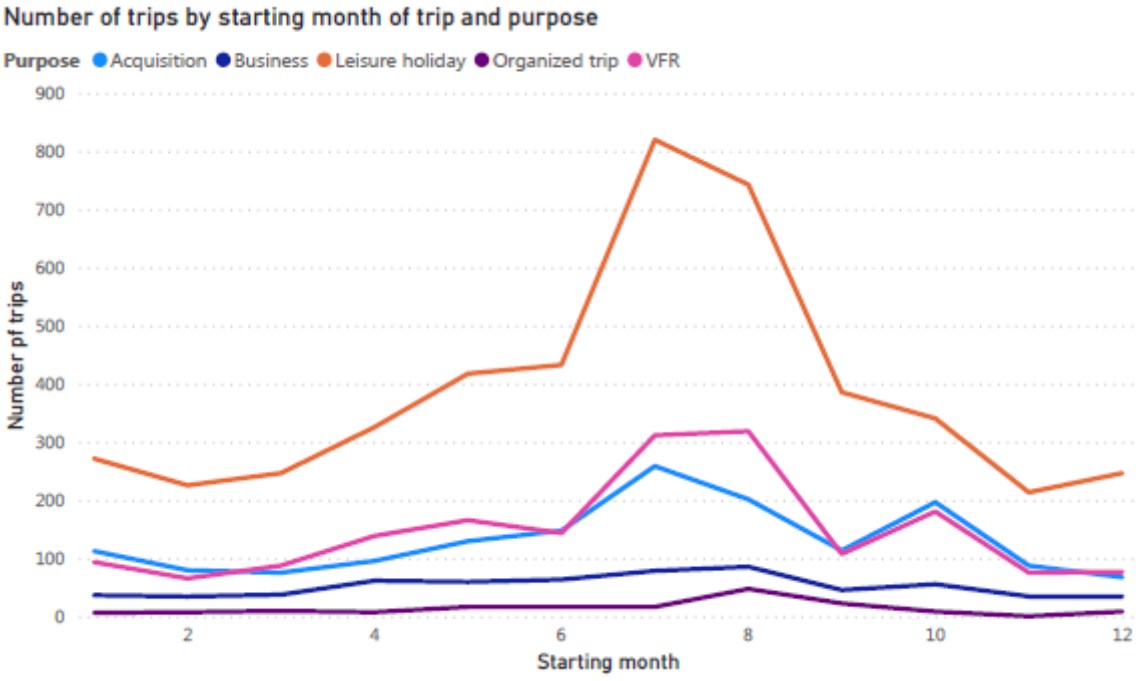

Figure 6.4 Number of trips by month of departure

Figure 6.4 shows the distribution of trips initiated in different months of the year, according to purpose of trip. Not surprisingly, most trips started in the summer, season that shows elevated number of trips for leisure, acquisition and VFR-related

trips. August also shows a higher number of organized trips; the line for business trips appears to be more levelled throughout the year. In October, when leisure-related trips appear to be on a decline, VFR and acquisition-related trips rebound slightly. Trips with all purposes slump to its minimum in February.

6.3 Predicting The Occurrence of A Trip

We evaluated our prediction models based on three metrics. Accuracy measures the proportion of correctly predicted observations over all data points. Precision indicates the fraction of correct predictions out of the observations the model predicted as 1 – or 'will go on a trip'. In a practical sense, if the bank approaches the customers who are shown to go on a trip by the model, precision will indicate the fraction of them that actually embarks on a trip. This measure could be valuable if the monetary cost of approaching a customer, the benefit of convincing a customer to do a certain action, and the expected rate of success of a campaign are known. Recall, on the other hand, refers to the proportion of correct predictions over all cases whose actual outcome is 1 – who actually go on a trip. If the banks goal is to reach out to as many actual travelers as possible, given that campaign costs are insignificant, recall plays an important role.

We ran all combination of model types and settings over the validation phase as mentioned in Section 5.3.1. Combinations with feature selection done through PCA underperformed compared to the Variable Importance method. All models performed better in terms of all metrics, when the window of observation was 6 months, instead of 3 months. Out of all algorithms, Logistic Regression, Random Forest and ANN demonstrated better predictive power than SVM, thus these models with these settings were deployed on the test time frame. The performance of the top three models in the test phase are seen in **Table 6.7**. The notes below the three models point out the hyperparameter settings that the models were given according to the tuning results in the validation phase.

Accuracies seem promising, but they must be interpreted together with the other metrics due to the imbalance of class labels in each time-band. Precision and Recall rates are moderate both in the training and test set, implying that the models struggle to identify some of the observations with positive label. The gap between train and test metrics is tolerable, indicating that no overwhelming overfitting occurred. It must be noted that for all three models, the gap between Recall and Precision can

Test phase metrics (TOP 3 models)

	Train*			Test		
	Accuracy	*Precision*	*Recall*	*Accuracy*	*Precision*	*Recall*
Logistic Regression (IV) *IV Selection, 6 months* *regularization: L1, C: 1*	84,11%	47,41%	33,50%	83,19%	41,15%	25,52%
Random Forest (IV) *IV Selection, 6 months* *# trees: 70, max depth: 4*	86,52%	62,08%	31,07%	84,99%	50,65%	20,42%
ANN (IV) *IV Selection, 6 months* *L2 kernel regularizer (0.1), bias regularizer (0.1)*	84,67%	50,00%	36,55%	84,13%	46,12%	29,97%

Table 6.7 Performance of top three models on the test phase time-band

be balanced out by choosing a different threshold than the standard p = 0.5; **Table 6.7** reflects the results gained with a threshold of 0.3 after observing the models' difficulty to predict positive labels.

The ROC curves for the three models are shown in **Figure 6.5**. The curves appear approximately midway between the diagonal line indicating random choice and the upper left corner of perfect prediction. If we relied solely on the ROC curves, we would declare the models to perform moderately.

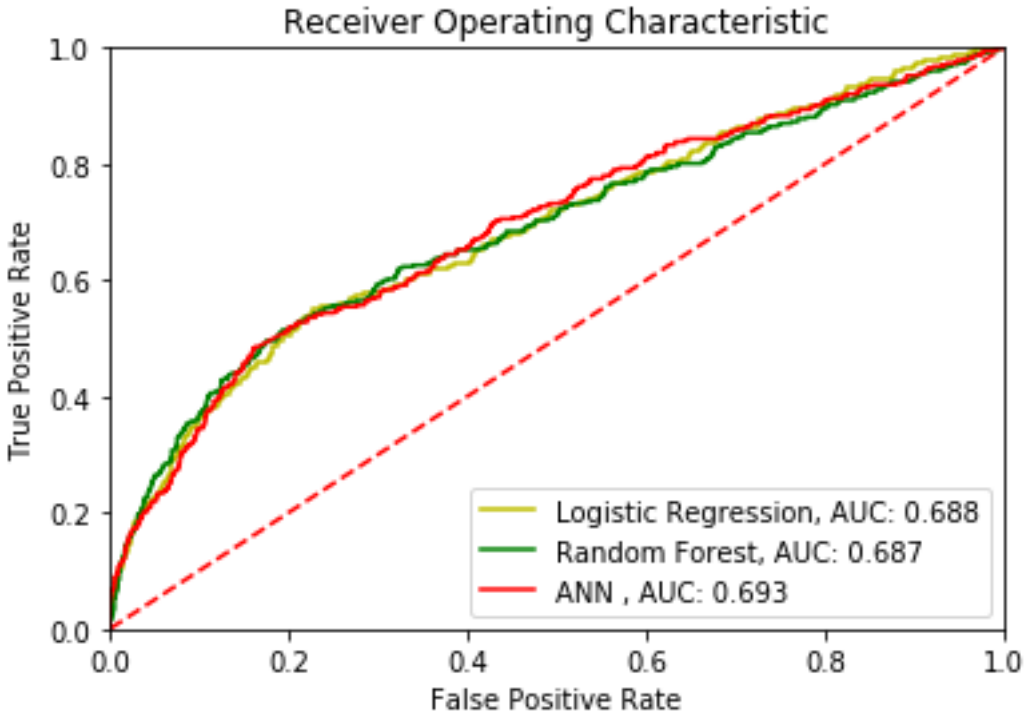

Figure 6.5 ROC curve for classification results

In order to make an unbiased evaluation of usefulness of the models, we plot the Precision-Recall Curve beside the ROC curve – this latter one could be deluding the same way as the accuracy rate. The Precision-Recall Curve for the three models in the final test phase is presented in **Figure 6.6**. The horizontal baseline indicates the proportion of positive cases in the test set out of all test data observations, amounting to 12.76%. This represents the 'no-skill' line, or the performance we would have achieved with completely random labeling decisions. Evidently, the models can achieve a higher Precision than what random guesses would, but since the curves lean toward the left-lower edge of the graph, the predicting performance is low to moderate. Apart from small differences – e.g. the Random Forest model performing worse with low recall or higher threshold – the three models perform similarly, with one or the other being slightly better at different threshold levels.

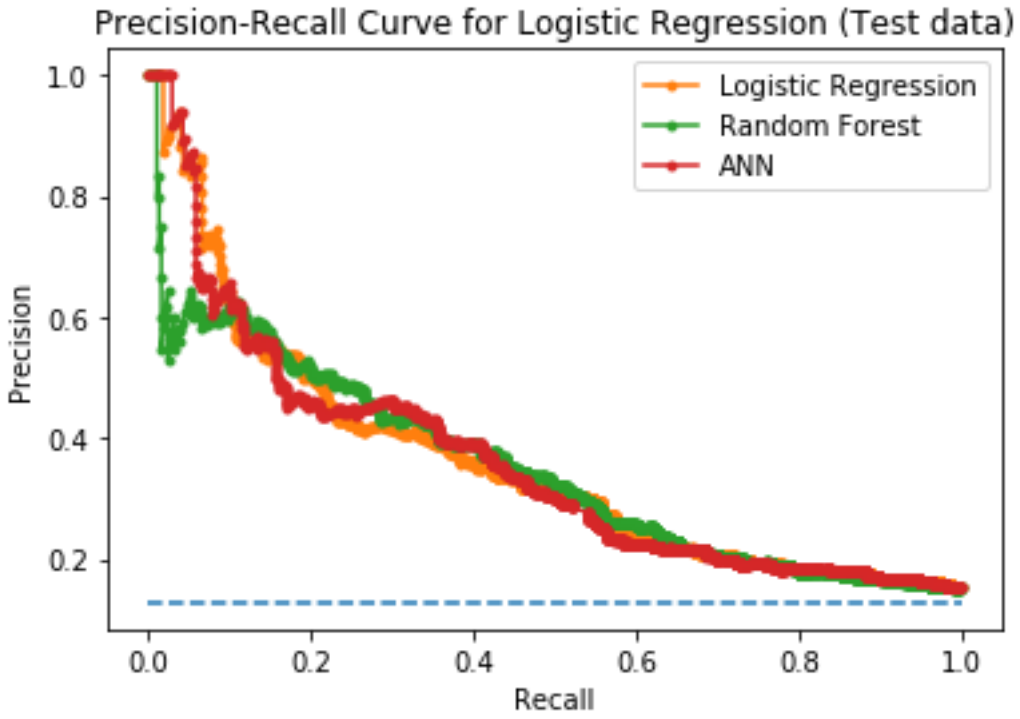

Figure 6.6 Precision-Recall Curve for the models in test phase, on test data

The moderate performance of the models could be explained by the unique, distinctive nature of out-of-town trips, in contrast with intra-urban mobility patterns. Potentially, a data set covering a longer period of time could account for seasonality, improving the Precision and Recall of the used models. Despite the limitations of the predictive model, the analysis assists us in examining the impact of different variables. To avoid confusion, we should note that after the validation phase, only 82 features were retained and we used all 82 in the test phase. The Variable Importance metrics for these variables in the test phase are displayed in **Table 6.8**. We

discovered that trips observed formerly and their related expenditure are in general the best features to split the data into purer subsets. Beside some demographic variables, like education or gender, some of the expense volumes on specific categories were also found to be significant and handy in the splitting algorithm. Finally, the spatial variety of consumers in Istanbul – i.e. how many districts of Istanbul they shopped in – appears important as well.

	Variable Importance
Number of previous trips	0,0665
Expense in previous trips	0,0658
Previous trip purpose: Leisure	0,0491
Istanbul expense on furniture	0,0414
Education: University	0,0382
Education: High school	0,0335
Gender: Male	0,0253
Spatial variety	0,0234
Job type: Employee, private sec.	0,023
Istanbul expense on services	0,0216
Istanbul expense on fuel	0,021

Table 6.8 Variable Importances for variables in the Test Phase (6 month bands)

6.4 Predicting Expenditure During A Trip

As a first step, we reduced the Filtered Trips data table to a subset, in which all trips have transactions recorded in the prior three and six months. In other words, for a window size of 3 months, the subset included 5,304 trips that started on September 1, 2014 or later, and for 6 month time-band of observation; 3,655 trips with starting date December 1, 2014, or later.

After compiling these two subsets of the data table, we calculated the correlations between the variables. For windows of observation of 3 months, the highest correlation with the dependent variable was with the Trip Duration, at 42,44%. This was followed by Total Expenditure outside of Istanbul and Expenditure on Previous Trips at 41,30% and 37,14% respectively – although these two variables have high multicollinearity of 90%, and one of them would be expected to be marked as insignificant during the feature selection process. Among destinations, Muğla's correlation with the tourist expenditure was the most significant, with a positive 10,93%, the province being a popular holiday resort with elevated prices in recent

years. The dummy variable for an assumed Transit was also above 30% for both time window settings. For the 6-month time-band, expenditure out of Istanbul and on previous trips showed lower correlation at 26,47% and 25,04%, respectively. On the other hand, some of the entropy measures reached more significant correlation with the target variable, e.g. 'Day of week variety' was at negative 13,98%.

As mentioned previously, Thrane (2014) set 40% of R2 as a desirable yardstick for future research on tourist expenditure, based on the outcome of authors' previous work. This performance metric is given in terms of fitting an OLS model to the totality of survey data, rather than some systematic predictive modeling. The R2 calculated for the dataset with 3 month of observation span is 45,37%, and for 6 months, 39.52%. In this sense, relying upon credit card transaction data appears to compete with survey-based methods in terms of explaining the variance in tourist expenditure. The feature selection based on Variable Importances reduced the data table down to 59 and 56 variables, respectively for 3 and 6 months. As seen in **Table 6.9**, one of the most remarkable details is that demographic variables did not appear to make good predictors for expenditure, and statistics and metrics about past expenditure domineer after the variable for trip duration. Secondly, metrics about out-of-town spending behavior are more significant closely before the trips, while for a time-band of observation of 6 months, median expenditure in Istanbul seems more suggestive.

Variable Importances for predicting Trip Expenditure		
	3-month windows	**6-month windows**
Trip duration	24,05%	25,80%
Out of Istanbul Mean Expenditure	7,89%	5,55%
Out of Istanbul Expense: Fuel	7,20%	2,28%
Out of Istanbul Total Expenditure	6,89%	7,53%
Out of Istanbul Median Expenditure	5,86%	2,85%
Total expense in previous trips	4,28%	0,58%
ATM withdrawal during trip	2,99%	0,39%
Risk score	2,82%	0,72%
Istanbul Median Expenditure	1,85%	11,49%
Spending category diversity	0,42%	1,87%

Table 6.9 Variable Importances for predicting Trip Expenditure

Table 6.10 summarizes the results of the best three models. In some cases, we needed to take further actions to reduce complexity to avoid overfitting. We applied L2 kernel and bias regularizers to the ANN model and reduced the number of hidden layers to 1 with 4 neurons. In the Random Forest algorithm, we defined the maximum depth to be 4. These modifications closed the gap between train and test performance metrics.

Overall, models with shorter, 3 month bands of observation performed better with the exception of Random Forest – also in terms of total variance explained with OLS - implying that the financial and mobility behavior surveilled closely before the trip is slightly more meaningful in terms of predicting expenditure during the trip.

Best 3 models, based on test R^2		
	Train R^2	Test R^2
Random Forest (IV) 6 month window	55,78%	40,19%
Linear Regression (IV) 3 month window	42,72%	38,09%
ANN (IV) 3 month window	42,65%	38,02%

Table 6.10 Best three models in predicting trip expenditure

The analysis has proved that variables extracted from transaction data have managed to be competitive with survey-based methods in terms of explaining the variation in expenditure. Additionally, the investigation revealed some of the most impactful variables both in terms of explaining variation and for predicting. We found that some of the statistics on spending behaviour from periods preceding the trip are good regressors as well, despite having been overlooked in survey-based approaches.

7. CONCLUSION

The literature is abundant on many of the issues touched upon in this thesis, such as prediction of tourist expenditure or analysis of human mobility on the basis of transactional data. Yet, transactional data has not been used for the analysis of human mobility for cases when people are away from their regular living area. In our study, we not only ventured to fill in this literature gap but also to extract statistics and models that could be handy for businesses and practitioners.

We utilized a credit card transaction data of 10.000 customers, within a period of one year to conduct our analysis. We found that out-of-town expenditures vary among different demographics segments, but generally follow what could be assumed based on people's life-cycle, interests and resources. We established a methodology to derive tourism or business-related trips out of big data, by following the geo-locations and the timeline of human credit card transactions. The extracted trips were then endowed with estimates of various features, e.g. duration, transit locations and distance traveled to the final destination.

The extracted trips were then subject to a clustering algorithm with the aim of finding what the traveler's purpose might have been, resulting in 5 segments: 'Leisure', 'Business', 'Acquisition', 'Visiting Friends and Relatives' and 'Package Holiday'. Demographic and financial behaviour-related breakdown of the elements in each cluster was well aligned with logic and intuition.

A similar clustering method was applied to distinguish between Turkey's 81 provinces based on the purpose of why people would visit or cross them. To this end, we calculated the proportions of total expenditures in different merchant categories, and related them to the corresponding statistics in Istanbul. The resulting 'Metropolis' cluster with large cities showed similar spending distribution to that in Istanbul, the 'Second Home' group exhibited a predominance of every-day staple purchases. The 'Cotton' cluster showed elevated proportions in categories related to Turkey's industry while the 'Transit' cluster comprised of provinces that most of the time did not serve as final destinations.

We endeavored to create a predictive model to foresee whether a customer would embark on a trip in the following three months. We applied a moving window approach, with different size of time bands for observing previous mobility and spending behavior of the customers. The models, despite having high overall accuracy, demonstrate low-to-moderate performance at finding the positive cases, but they pose an opportunity for a bank to deploy campaigns with higher Precision than random guesses, potentially leading to savings on marketing expenses.

Finally, we found that the analysis of trips based on transaction data, where features are created via estimations based on consecutive transactions, is competitive with survey-based methods to predict tourist expenditure in terms of variance explained. Both correlations and feature selection processes indicated that along with traditional survey questions regarding demographics and trip-related details, measures of past mobility and financial behavior are equally suggestive, in particular behavior displayed closely before the trip.

In conclusion, we hope that the present thesis has contributed to and managed to advocate the initiatives of using transaction data to analyse human mobility and financial behavior within the disciplines of sociology, management and tourism.

BIBLIOGRAPHY

Borzekowski, R., Kiser, E. K., & Ahmed, S. U. (2008). Consumers' use of debit cards: patterns, preferences, and price response. *Journal of Money, Credit and Banking*, *40*(1), 149–172.

Brida, J. G. & Scuderi, R. (2013). Determinants of tourist expenditure: A review of microeconometric models. *Tourism Management Perspectives*, *6*, 28–40.

Burkart, A. & Medlik, S. (1974). *Tourism: Past, Present and Future*. London: Heinemann.

Cai, L. A., Lehto, X. Y., & O'Leary, J. (2001). Profiling the us. -bound chinese travelers by purpose of trip. *Journal of Hospitality Leisure Marketing*, *7*(4), 3–16.

Carbo-Valverde, S. & Rodriguez-Fernandez, F. (2014). Atm withdrawals, debit card transactions at the point of sale and the demand for currency. *SERIEs*, 399–417.

Chan, P. K., Fan, W., Prodromidis, A., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems*, *14*(6), 67–74.

CNN (2018). İstanbul'da en çok nereli yaşıyor? Accessed: 2018-10-21.

Downward, P. & Lumsdon, L. (2000). The demand for day-visits: an analysis of visitor spending. *Tourism Economics*, *6*(3), 251–261.

Downward, P. & Lumsdon, L. (2003). Beyond the demand for day-visits: an analysis of visitor spending. *Tourism Economics*, *9*(1), 67–76.

El-haddad, A. B. & Almahmeed, M. A. (1992). Atm banking behaviour in kuwait: A consumer survey. *International Journal of Bank Marketing*, *10*(3), 25–32.

Fredman, P. (2008). Determinants of visitor expenditures in mountain tourism. *Tourism Economics*, *14*(2), 297–311.

Gunn, C. A. (1972). Vacationscape, designing tourist regions. Technical report, University of Texas, Austin.

Hayhoe, C. R., Leach, L. J., Turner, P. R., Bruin, M. J., & Lawrence, F. C. (2005). Differences in spending habits and credit use of college students. *The journal of consumer affairs*, *34*(1), 113–133.

I.U.O.T.O. (1963). The united nations' conference on international travel and tourism. Geneva. International Union of Official Travel Organizations.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer Science+Business Media.

Jang, S., Ham, S., & Hong, G. S. (2007). Food-away-from-home expenditure of senior households in the united states: A double-hurdle approach. *Journal of Hospitality Tourism Research*, *31*(2), 147–167.

Jansen-Verbeke, M. (1987). Women, shopping and leisure. *Leisure Studies*, *1*(1), 71–86.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience.

Krumme, C., Llorent, A., Cebrian, M., Pentland, A., & Moro, E. (2013). *The predictability of consumer visitation patterns*. Scientific Reports.

Lathia, N., Quercia, D., & Crowcroft, J. (2012). The hidden image of the city:

Sensing community well-being from urban mobility. *Pervasive*, *7319*, 91–98.

Lee, C. K., Var, T., & Blaine, T. W. (1996). Determinants of inbound tourist expenditures. *Annals of Tourism Research 23(3)*, *23*(3), 527–542.

LeHew, M. L. & Wesley, S. C. (2007). Tourist shoppers' satisfaction with regional shopping mall experiences. *International Journal of Culture, Tourism and Hospitality Research*, *1*(1), 82–96.

Lehro, X. Y., Cai, L. A., O'Leary, J. T., & Huan, T. C. (2004). Tourist shopping preferences and expenditure behaviours: The case of the taiwanese outbound market. *Journal of Vacation Marketing*, *10*(4), 320–332.

Leiper, N. (1979). The framework of tourism - towards a definition of tourism, tourist, and the tourist industry. *Annals of Tourism Research*, *6*(4), 390–407.

Lenormand, M., Louail, T., Cantu-Ross, O., Picornell, M., Herranz, R., Arias, J. M., Barthelemy, M., Miguel, M. S., & Ramasco, J. J. (2016). Influence of sociodemographic characteristics on human mobility. *Scientific Reports*, *5*.

McIntosh, R. W. (1977). *Tourism: Principles, Practices*. Columbus: Grid.

Mok, C. & Iverson, T. J. (2000). Expenditure-based segmentation: Taiwanese tourists to guam. *21*, 299–305.

Moll-de Alba, J., Prats, L., & Coromina, L. (2016). The need to adapt to travel expenditure patterns. a study comparing business and leisure tourists in barcelona. *Eurasian Bus Rev*, *6*, 253–267.

Murphy, K. P. (2012). *Machine Learning A Probabilistic Perspective*. Boston: MIT.

Oh, J. Y. J., Cheng, C. K., Lehto, X. Y., & O'Leary, J. (2004). Predictors of tourists' shopping behaviour: Examination of socio-demographic characteristics and trip typologies. *Journal Of Vacation Marketing*, *10*(4), 308–319.

Schneider, C. M., Belik, V., Couronne, T., Smoreda, Z., & Gonzalez, M. C. (2013). Unravelling daily human mobility motifs. *Journal of the Royal Society Interface*, *10*.

Scholnick, B., Massoud, N., Saunders, A., Carbo-Valverde, S., & Rodriguez-Fernandez, F. (2008). *The economics of credit cards, debit cards and ATMs: A survey*, volume 32. Journal of Banking Finance.

Seaton, A. & Palmer, C. (1997). Understanding vfr tourism behaviour: the first five years of the united kingdom tourism survey. *Tourism Management*, *18*(6), 345–355.

Singh, V. K., Bozkaya, B., & Pentland, A. (2015). Money walks: Implicit mobility behavior and financial well-being. *Plos One*, *10*(8).

Sobolevsky, S., Sitko, I., T. d. C. R., R. T., Hawelka, B., Arias, J. M., & Ratti, C. (2014a). Mining urban performance: Scale-independent classification of cities based on individual economic transactions. *ArXiv*.

Sobolevsky, S., Sitko, I., T. d. C. R., R. T., Hawelka, B., Arias, J. M., & Ratti, C. (2014b). Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain. *IEEE International Congress on Big Data*, 136–143.

Sobolevsky, S., Sitko, I., T. d. C. R., R. T., Hawelka, B., Arias, J. M., & Ratti, C. (2015). Cities through the prism of people's spending. *Plos One*, *11*(2).

TAGEM (2018). Tarım Ürünleri piyasaları: Pamuk. , Tarımsal Ekonomi ve Politika Geliştirme Enstitüsü.

Thorndike, R. L. (1953). Who belong in the family? *Psychometrika*, *18*(4).

Thrane, C. (2014). Modelling micro-level tourism expenditure: recommendations on the choice of independent variables, functional form and estimation technique. *20*(1), 51–60.

Timothy, D. J. (2005). *Shopping Tourism, Retailing, and Leisure*. Clevedon: Channel View Publications.

Turner, L. W. & Reisinger, Y. (2001). Shopping satisfaction for domestic tourists. *Journal of Retailing and Consumer Services*, *8*(1), 15–27.

TÜIK (2018). Yurtiçinde İkamet edenlerden yurtiçi seyahat yapanların seyahat ve geceleme sayısı ile harcamaları. , Türkiye Istatistik Kurumu.

TÜIK (2019a). Hanehalkı yurt İçi turizm, i. Çeyrek: Ocak - mart, 2019.

TÜIK (2019b). Hanehalkı yurt İçi turizm, iii. Çeyrek: Temmuz - eylül, 2018.

TÜIK (2019c). Nüfus istatistikleri / yıllara göre İl nüfusları, 2018.

UN (2008). *International recommendations for tourism statistics*. New York: United Nations Publications.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.

APPENDIX A

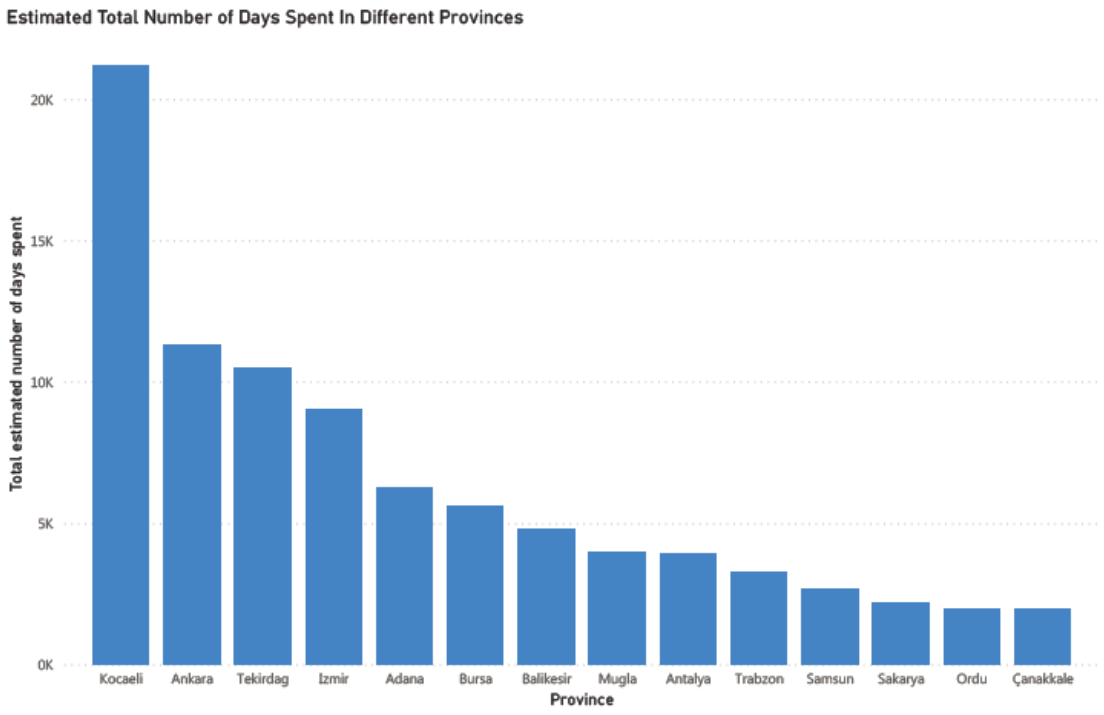

Figure A.1 Number of days recorded in each of the provinces

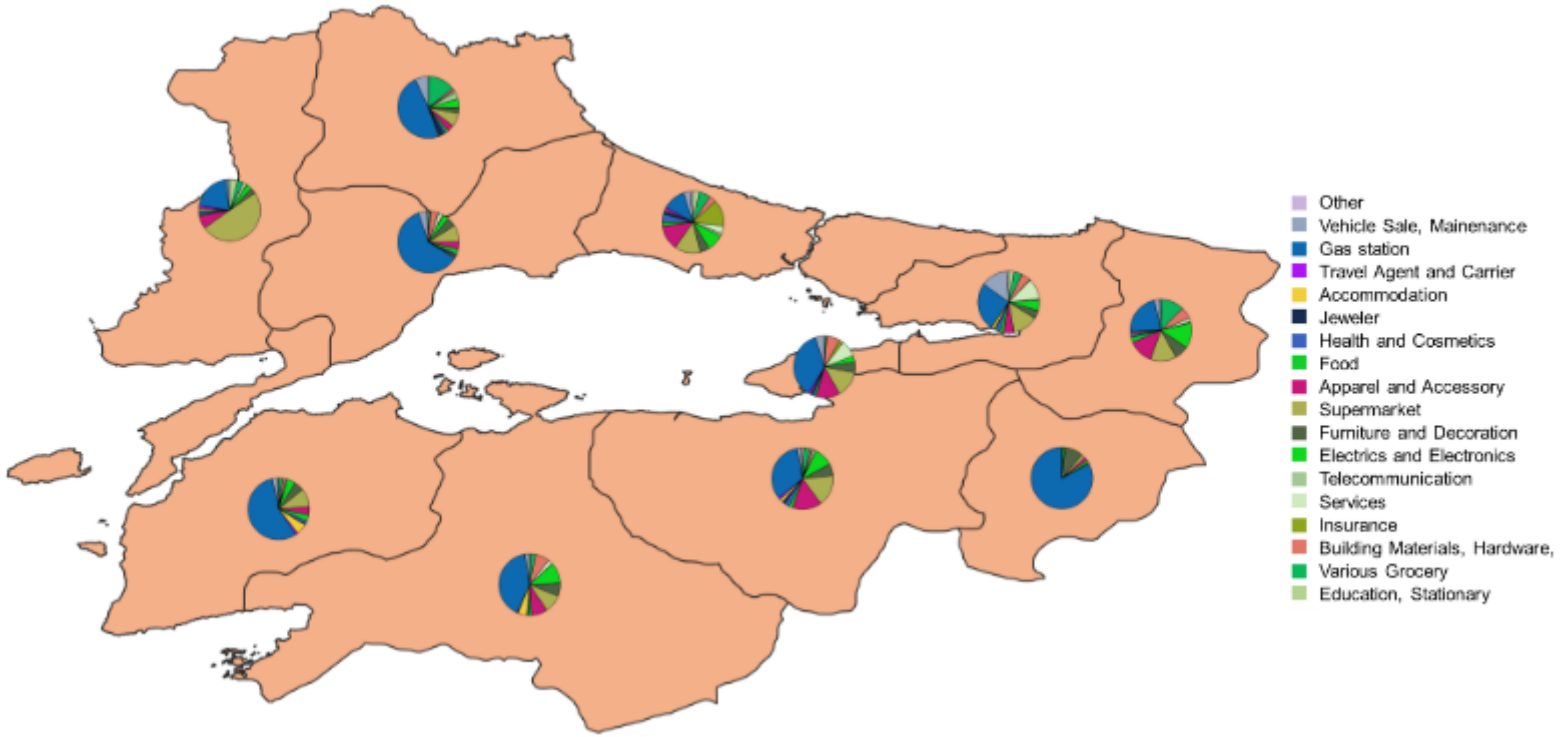

Figure A.2 Spending distribution in the Marmara region

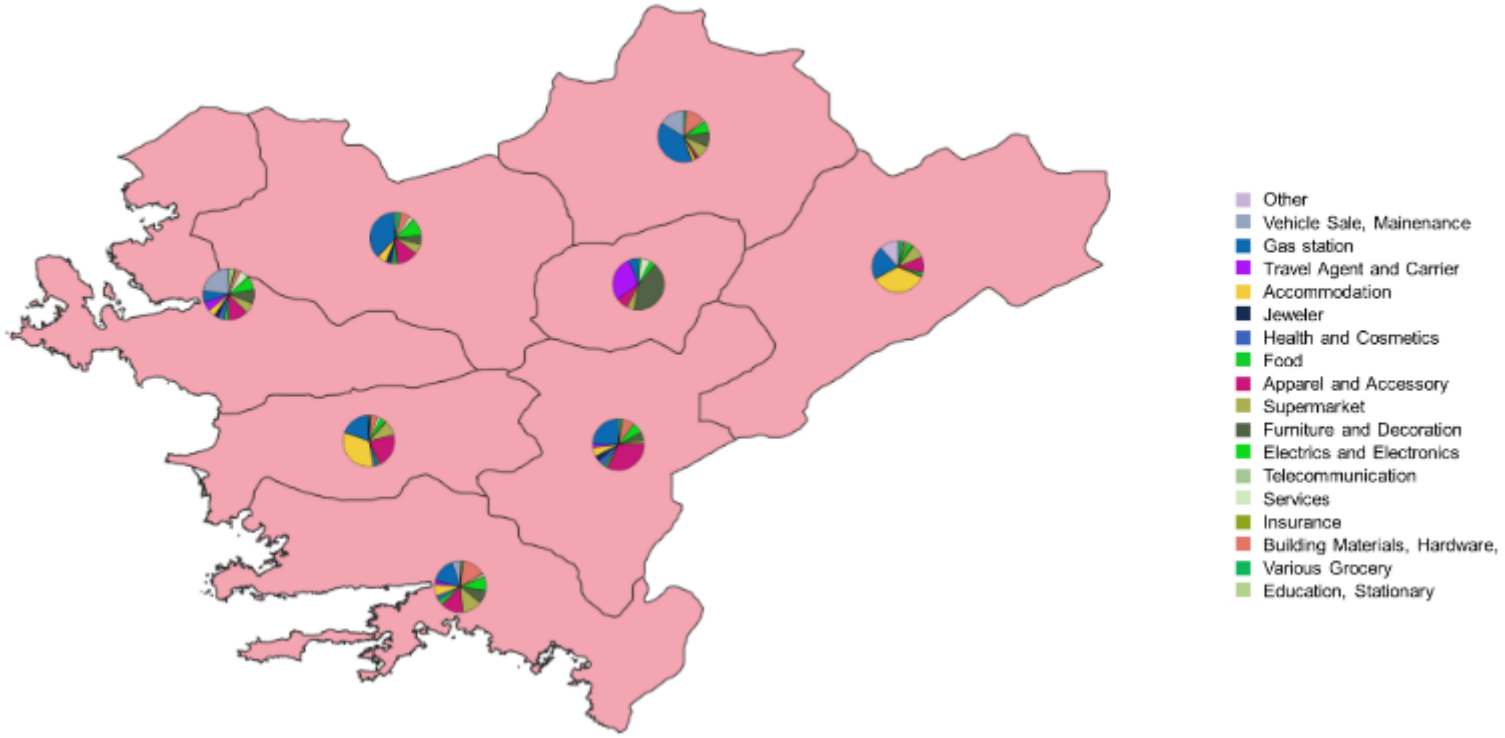

Figure A.3 Spending distribution in the Aegean region

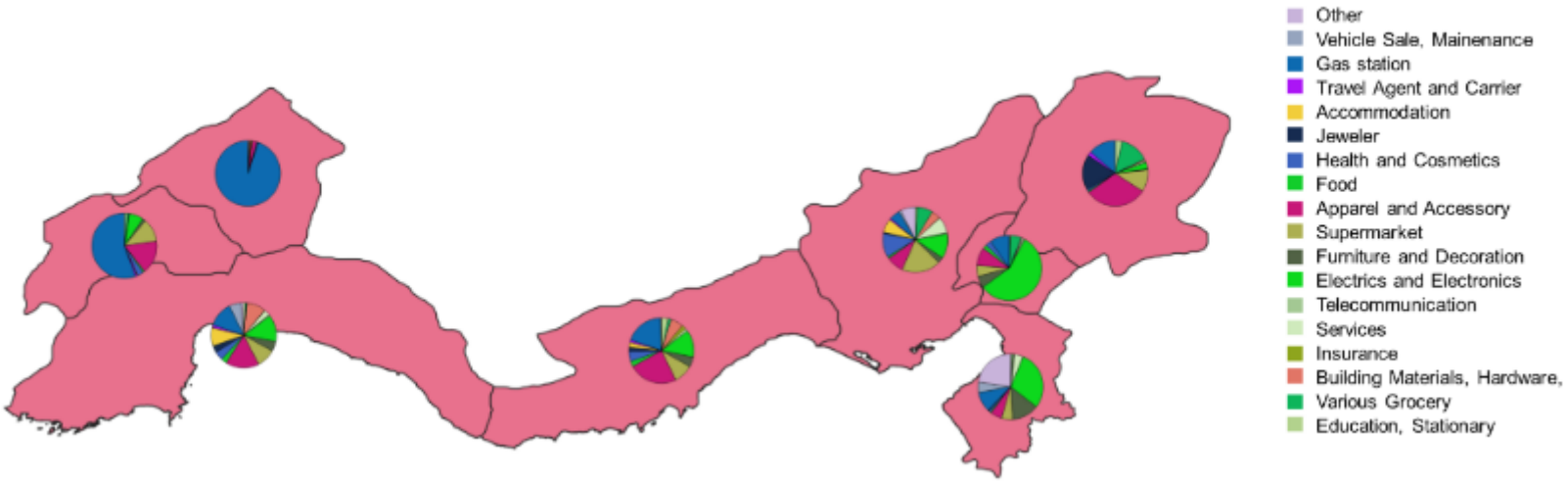

Figure A.4 Spending distribution in the Mediterranean region

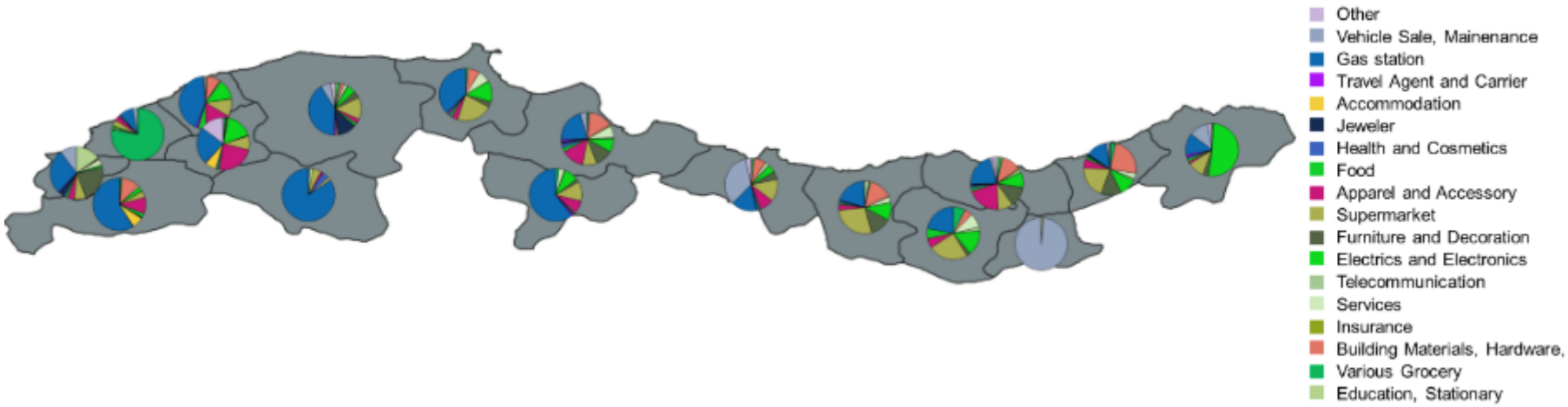

Figure A.5 Spending distribution in the Black Sea region

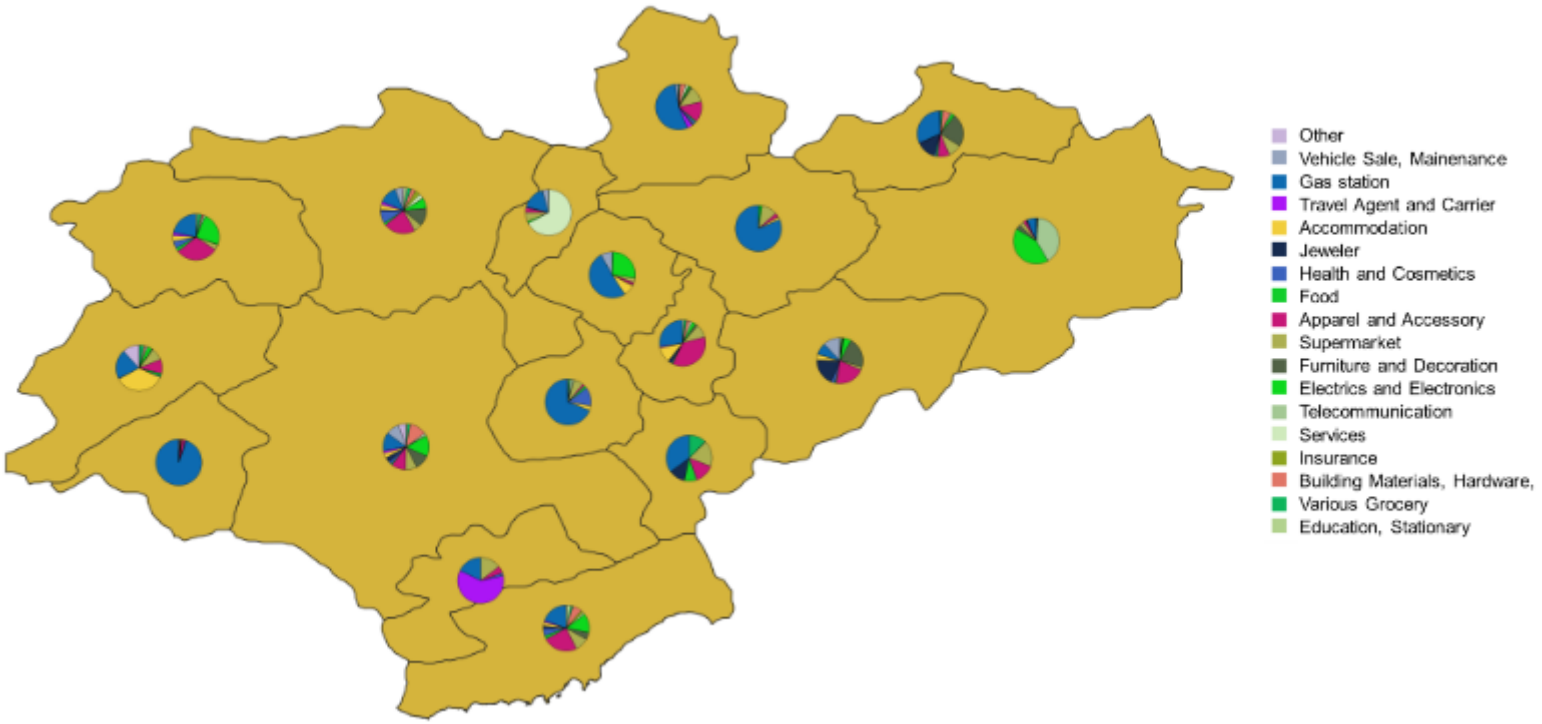

Figure A.6 Spending distribution in the Central Anatolian region

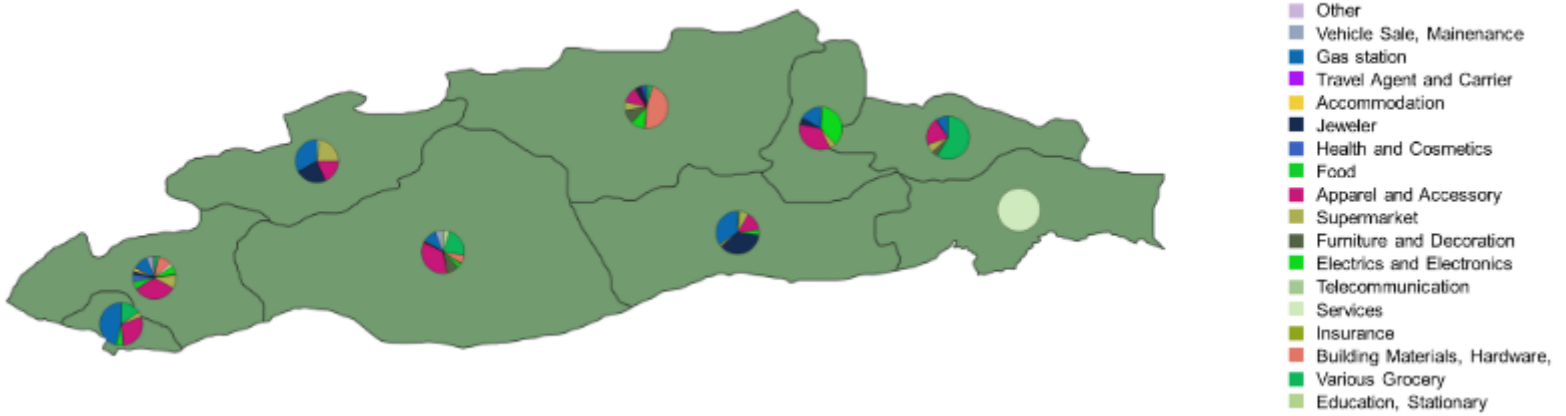

Figure A.7 Spending distribution in the Southeast Anatolian region

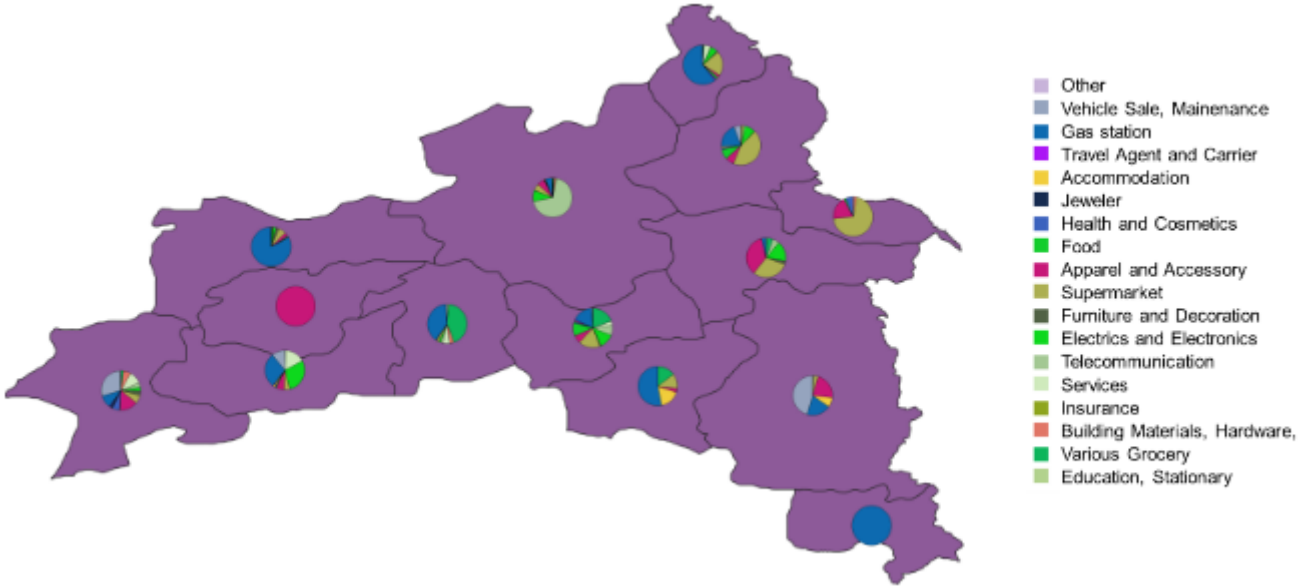

Figure A.8 Spending distribution in the Eastern Anatolian region

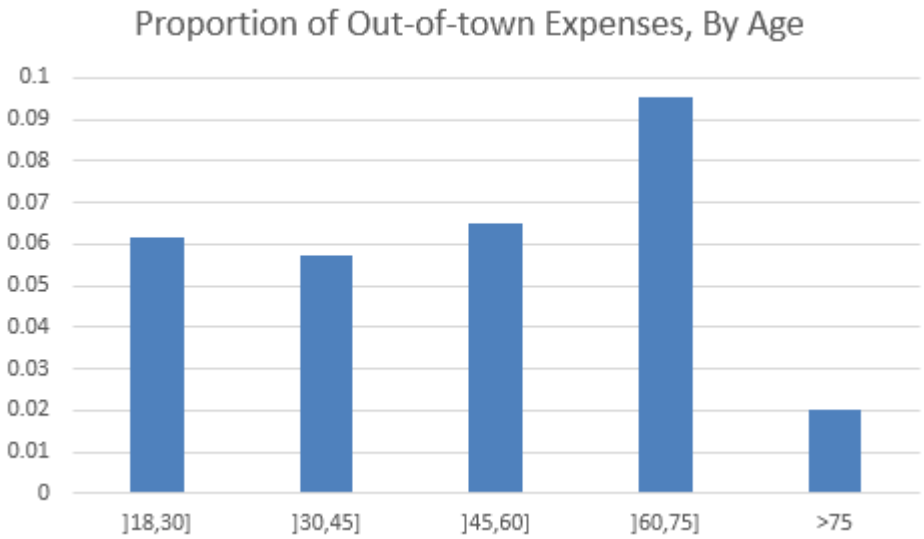

Figure A.9 Proportion of expenditures out of Istanbul out of total expenditures (age)

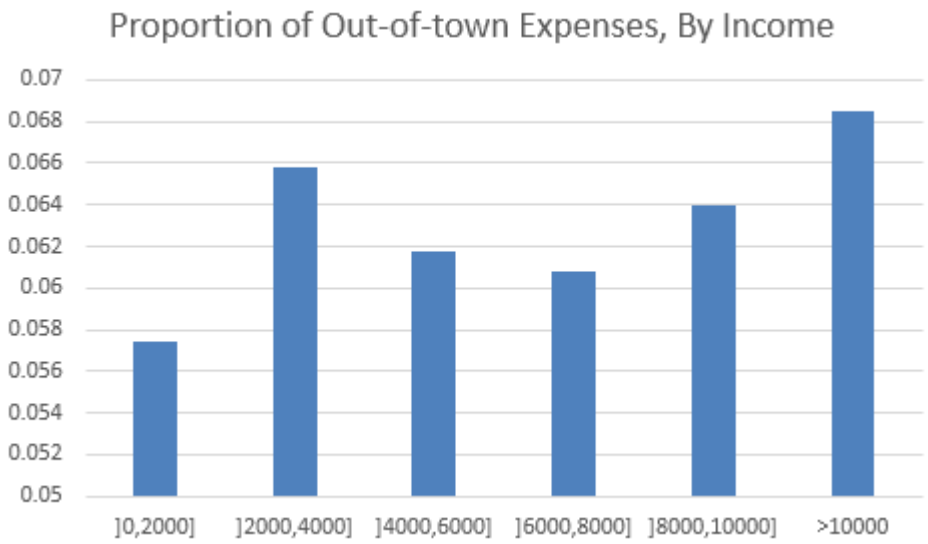

Figure A.10 Proportion of expenditures out of Istanbul out of total expenditure (income)

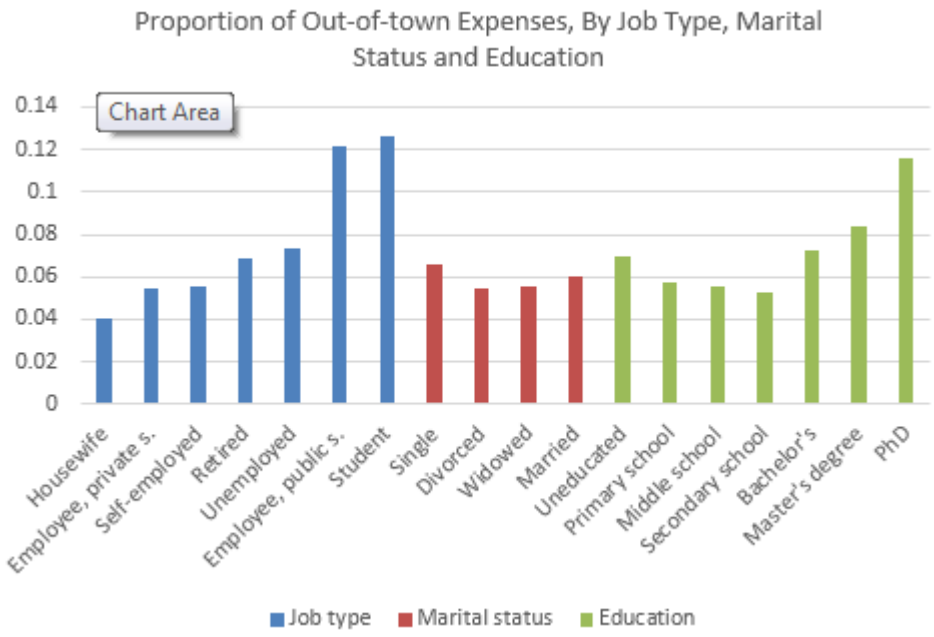

Figure A.11 Proportion of expenditures out of Istanbul out of total expenditure (Job type, Marital status, Education)