INVESTIGATING PUBLIC TRANSIT DAILY TRAVEL BEHAVIOR USING SMART CARD DATA: A CASE STUDY OF KONYA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MAJED AL KRDY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
CIVIL ENGINEERING

SEPTEMBER 2022

Approval of the thesis:

INVESTIGATING PUBLIC TRANSIT DAILY TRAVEL BEHAVIOR USING SMART CARD DATA: A CASE STUDY OF KONYA

submitted by **MAJED AL KRDY** in partial fulfillment of the requirements for the degree of **Master of Science** i**n Civil Engineering, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Erdem Canbay
Head of the Department, **Civil Engineering** _____

Prof. Dr. Hediye Tüydeş Yaman
Supervisor, **Civil Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Ela Babalık
City and Regional Planning, METU _____

Prof. Dr. Hediye Tüydeş Yaman
Civil Engineering, METU _____

Prof. Dr. Murat Güler
Civil Engineering, METU _____

Asst. Prof. Dr. Funda Türe Kibar
Civil Engineering, Başkent University _____

Asst. Prof. Dr. Oruç Altıntaşı
Civil Engineering, İzmir Kâtip Çelebi University _____

Date: 02.09.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Majed Al krdy

Signature :

ABSTRACT

INVESTIGATING PUBLIC TRANSIT DAILY TRAVEL BEHAVIOR USING SMART CARD DATA: A CASE STUDY OF KONYA

Al krdy, Majed
Master of Science, Civil Engineering
Supervisor: Prof. Dr. Hediye Tüydeş Yaman

September 2022, 71 pages

As urban diversity grows in cities, so does the heterogeneity of public transit (PT) travel behavior. To accommodate for usage heterogeneity, a deeper understanding of travel behavior beyond descriptive analysis is necessary, which can be performed using smart card data (SCD), if available. In this study, K-means clustering algorithm combined with a data mining technique that includes unsupervised learning clustering algorithm is proposed to detect different demand segments among PT users based on their daily boarding activity from the SCD. The numerical results are obtained for the PT usage in the city of Konya, Turkey. Descriptive statistics of PT behavior from the SCD serve as an introductory analysis, while daily travel patterns at user levels are searched for detection of various travel patterns, which can be used by local authorities to improve the PT services and their customization for local demand.

Keywords: Public Transit, Smart Card Data, Descriptive statistics, Unsupervised Learning Algorithm, Clustering

ÖZ

TOPLU TAŞIMA GÜNLÜK SEYAHAT DAVRANIŞLARININ AKILLI KART VERİLERİYLE ARAŞTIRILMASI: KONYA ÖRNEĞİ

Al krdy, Majed
Yüksek Lisans, İnşaat Mühendisliği
Tez Yöneticisi: Prof. Dr. Hediye Tüydeş Yaman

Eylül 2020, 71 sayfa

Şehirlerin kentsel çeşitliliği arttıkça toplu taşıma seyahat davranışının heterojenliği de artıyor. Toplu taşıma kullanımının heterojenliğine uyum sağlamak için seyahat davranışının derinlemesine anlaşılması bir zorunluluk haline geliyor. Bu nedenle bu çalışma, Konya toplu taşıma otomatik ücret toplama sisteminden üretilen akıllı kart verilerini kullanarak toplu taşıma davranışının tanımlayıcı istatistiklerini incelemektedir. Ayrıca, Konya ilçesi toplu taşıma ağındaki kullanıcıları günlük biniş aktivitelerine göre segmentlere ayırmak için gözetimsiz öğrenme kümeleme algoritmasını içeren bir veri madenciliği tekniği kullanılacaktır. Tanımlayıcı istatistikler, veri görselleştirme araçları aracılığıyla veri bileşimini, ana özellikleri ve parametreleri anlamak için bir giriş işlevi görür. K-ortalamalar kümeleme algoritması ise toplu taşıma kullanıcılarının günlük biniş modellerinin özelliklerini anlamaya yardımcı olur. Veri madenciliği yaklaşımı, toplu taşıma kullanıcılarının günlük seyahat modellerini akıllı kart verilerinden çıkarma yeteneğine sahiptir. Günlük seyahat modelleri, seyahat talebi modellemesini ve hizmet özelleştirmesini kolaylaştırdığından ulaşım yetkilileri için çok önemlidir.

Anahtar Kelimeler: Toplu Taşıma, Akıllı Kart Verileri, Tanımlayıcı istatistikler, Gözetimsiz Öğrenme Algoritması, Kümeleme

To my dear father who have always inspired me, Mohammed Osama Al krdy

ACKNOWLEDGMENTS

I would like to present my enormous gratitude to my advisor, Prof. Dr. Hediye Tüydeş Yaman for her vision and thoughtful insights, as well as her guidance and encouragement throughout my study period. Without her, the successful completion of this work would have not been possible.

I'm also very grateful to have the support of several fellow research assistants in the transportation engineering laboratory and PARABOL Consulting, Gülçin Dalgıç, Beyhan İpekyüzç and Elif karagümüş.

I would like to thank my dear mother Rim Bazboz, my father Mohammed Osama Al krdy, and all my friends for their motivation and encouragement throughout my study.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AHTV	Average Hourly Transaction Volume
AI	Artificial Intelligence
BAHTV	Bus Average Hourly Transaction Volume
GPS	Global Positioning System
ITS	Intelligent Transportation System
PT	Public Transit
SCAFC	Smart Card Automated Fare Collection
SCD	Smart Card Data
SLA	Supervised Learning Algorithm
TAHTV	Tram Average Hourly Transaction Volume
ULA	Unsupervised Learning Algorithm

CHAPTER 1

INTRODUCTION

With the rapid increase in private vehicle ownership, coupled with an accelerating rate of urbanization, public transit (PT) has become the primary countermeasure in providing sustainable urban transportation and fighting against various concerns such as traffic congestion, energy consumption, and air pollution. This has elevated the significance of PT systems in accommodating the growing urban mobility demand.

Urban mobility demand is often diverse in nature due to the heterogeneity of urban populations with various social and cultural backgrounds. Despite a great proportion of the urban population adopting the 9:00-17:00 commuting lifestyle, additional routines driven by home-based or part-time jobs, flexible working hours, self-employment, and education-based mobility constitute a significant portion of the urban population. With the growing diversity in daily routines and activities comes a higher level of variability in travel patterns and behaviors. Concurrently, Public transit agencies are burdened with changing demand patterns, fluctuating ridership, and erratic travel behaviors. A thorough understanding of such travel patterns is vital to keep pace with the urban travel demand development.

The typical approach for examining PT users' behavior is based on homogeneity assumptions. To meet the travel mobility needs of a diverse population, it must be viewed as a composition of various demand segments, with different travel behaviors and patterns. In other words, it is a necessity for PT authorities to shift from an aggregated approach to a user-based one. The distillation of travel patterns down to an individual level requires longitudinal data where every user is being observed over time.

PT authorities have long sought for strategies to attract urban population into favoring PT over other modes of transport. From marketing campaigns to easier fare

collection system, to service adjustment and optimization, all with the purpose of increasing ridership and improving user satisfaction. Having a thorough understanding of the heterogenic nature of user behavior is effective in attaining such purposes. This is because it allows PT authorities to better organize the service down to different demand segments. For instance, information regarding service disruption and in-vehicle crowding can be provided based upon every user's travel behavior with the help of interactive tools such as smartphone applications. Furthermore, understanding heterogenic travel behavior enables user segmentation based on similarities in travel behavior. Consequently, fare policies, marketing campaigns and PT planning can be performed separately for each segment, rather than being executed at an aggregate level. For example, a marketing campaign can be launched specifically to target a segment that uses PT less frequently in order to influence their behavior.

1.1 Motivation

Smart Card Automated Fare Collection Systems (SCAFCSs) provide an opportunity to investigate aggregated PT user behavior. Apart from its main objective being revenue collection, a SCAFCS collects and stores data corresponding to temporal and spatial information regarding the PT usage. With the increasing availability of Smart Card Data (SCD), numerous studies have investigated a variety of topics, from performance and accessibility assessment (Trépanier et al., 2009; Yun et al., 2021) to transit demand management (Halvorsen, 2015) and behavioral analysis (Cui et al., 2018).

In line with the evolution of Intelligent Transportation systems (ITS), studies have been focusing on Artificial Intelligence (AI) implementation in PT travel behavioral analysis. Techniques such as supervised (SLA) and unsupervised learning algorithms (ULA) including regression, classification, and clustering are utilized in travel behavior research.

However, some challenges arise when analyzing big data, such as smart card data. For instance, computational complexity, storage capacity, and computing power are some of the predominant limitations when handling big data. In addition, the raw data format is mostly unsuitable for heterogenic behavioral analysis. Thus, additional data processing is required prior to executing the analysis. Moreover, travel behavior is often governed by a variety of uncontrolled variables such as weather, social events, and holidays. Therefore, it's challenging to account for such variables when analyzing travel behavior variability.

In line with the benefits of heterogenic travel behavior analysis and big data handling issues, the aim of this study is to come up with an efficient yet simple machine learning algorithm that is capable of analyzing PT SCD. The output is intended to aid public transit authorities in developing a better understanding of travel behavior and execute relevant strategies accordingly.

1.2 Scope of The Study

This thesis aims to answer the following research questions:

- To what extent clustering algorithm is effective in analyzing PT daily travel behaviors?
- How disaggregate SCD can be utilized in travel behavior analysis?
- What SCD processing is essential to assess travel behavior analysis?

In this context, the scope of the study includes:

1. Performing data manipulation to extract descriptive statistics from raw SCD.
2. Utilizing an unsupervised learning clustering algorithm to extract daily travel patterns from SCD.

The application is done using Konya's PT SCD. First, data preprocessing is performed independently for extracting descriptive statistics and constructing clustering algorithm input. Descriptive statistics include daily and hourly transaction

volumes, usage per card type, weekdays and weekends volume, as well as daily usage frequency.

On the other hand, clustering preprocessing extracts daily boarding vectors that represent individuals' PT daily usage. Boarding vectors are then used as an input for K-means clustering. Finally, daily boarding profiles along with usage frequency and card types are examined.

1.3 Thesis Organization

This thesis is organized as follows. Chapter 2 provides a literature review of SCD, its applications in PT studies, as well as the implementation of data mining techniques in PT travel behavior analysis. In addition, a brief of the K-means clustering algorithm is introduced. Chapter 3 discusses the methodology implemented for analyzing PT travel patterns including details of data processing and the algorithm used. In Chapter 4, Konya's PT system, as well as SCD, are briefly described, followed by the results and findings of implemented clustering analysis. Finally, Chapter 5 includes the conclusion, contribution, as well as future recommendations.

CHAPTER 2

LITERATURE REVIEW

2.1 Smart Card Automated Fare Collection System (SCAFCS) for PT

Popularity of SCAFCSs has been increasing rapidly over the past few decades. Dated back in 1997, the first implemented SCAFC system for public transit use was in Hong Kong known as the Octopus Smart Card (Chau & Poon, 2003). Ever since lots of transportation authorities started implementing the SCAFC system as a revenue collection and management tool. Not only because it is more convenient for operators, but it also reduces boarding time, vehicle downtime, driver workload, as well as preventing fraud (Deschaintres et al., 2019).

There are two main types of smart cards, contact and contactless. A contact card has an embedded chip where its surface must be in direct contact with the card reader (Pelletier et al., 2011). A contactless smart card however has an embedded microchip that allows the user to complete a transaction by simply placing the card within proximity to the card reader (Bai et al., 2008). The transactions are then transferred from the card readers to a central database where it gets stored for accounting purposes. Additional information related to transactions such as time, location, and card type is also recorded.

The continuous stream of data collected by SCAFC had proven to be useful for research purposes (Li et al., 2018). However, some challenges arise when dealing with SCD analysis, such as the lack of passengers' demographic information, trip purpose, and destination as most SCARC systems only record the user's alighting point (Faroqi & Mesbah, 2021; Viallard et al., 2019). Moreover, privacy is a considerable concern when dealing with SCD since it is a record of a person's

movement through the transit network, which might impede access and use of such data (Agard et al., 2006).

2.2 Smart Card Data (SCD) in PT Research

SCD have been frequently utilized in PT research. These data were used to perform strategic studies, such as PT behavioral analysis, user loyalty and segmentation, and PT system long-term planning. In addition, tactical studies were performed addressing route use and adjustment, service customization, and travel patterns. The third category of research done using SCD is operational which includes service performance assessment, fraud detection, and error correction (Espinoza et al., 2018).

2.2.1 Strategic Studies

The strategic level of research includes studies that are related to long-term PT network planning (Pelletier et al., 2011). With the aid of smart card data, every user can be traced across the transit network where its travel behavior and pattern can be examined (Agard et al., 2006). Moreover, the availability of card identification numbers, date, and time of every transaction facilitates the determination of ridership volumes. With a focus on user classification and characterization, long-term planning studies as such lack users' personal socio-demographic information. Therefore, it's preferable to integrate the SCD with classic data collection approaches such as household surveys (Trépanier et al., 2009). Another use of SCD is PT user loyalty assessment. By looking at the dates on which the smart card was used, the lifespan of each user can be determined (Trépanier et al., 2012).

2.2.2 Tactical Studies

Tactical level research focuses on service adjustments based on SCD analysis. Despite PT ridership variations, many public transit authorities provide consistent schedules during weekdays (Utsunomiya et al., 2006). SCD can be utilized to determine the maximum daily loading point on every route, which can be used to adjust schedules accordingly (Trépanier et al., 2007).

Furthermore, most smart card systems record the boarding points, along with other corresponding features such as date, time, and location, yet the alighting point is mostly unknown. Nevertheless, SCD can be useful in determining the alighting points using behavioral regularities. For instance, a user's first boarding point of a day can be considered the last alighting point for the same day if a pattern of boarding points can be observed over a certain period. If the alighting points were determined, a detailed origin-destination matrix can then be generated (Alsger et al., 2015).

The study of PT transfers is quite common in the literature. Hofmann et al. (2009) implied that a better understanding of PT transfers is essential for PT network adjustments in terms of geometry and schedules. Such adjustments can be made in coordination with various means of PT to meet the needs of users (Munizaga et al., 2010).

2.2.3 Operational Studies

SCD at the operational level research can be utilized in determining service performance indicators on PT networks. Some of these indicators are schedule adherence, service coverage, reliability, and vehicle occupancy (Trépanier et al., 2009; Uniman et al., 2010). For instance, to calculate schedule adherence, boarding times of smart card transactions at every bus stop along a specific route must be compared with the route's schedule. However, because boarding usually takes several seconds, only the first transaction at every stop is considered to be the vehicle's arrival time. In addition, since the exact boarding time is known for every

transaction, the average boarding time can be calculated separately for different bus stops and routes (Hickman, 2002). Such statistics can also be integrated by card or transaction type depending on the desired level of analysis depth (Utsunomiya et al., 2006).

SCD can also help detect errors in the public transit automated fare collection system. These recurrent errors play a vital role in the ability to identify faulty equipment, human error, and fraud (Hussain et al., 2021). One of the most common errors is the mismatch between the transaction's recorded GPS location and the actual location along the planned route where the transaction occurred. Such errors can be corrected during the data preprocessing stage with the help of attribution techniques and data comparison (Chapleau & Chu, 2007).

2.3 Travel Behavioral Analysis Using SCD

Travel behavioral analysis studies users' movement across the transit network for any purpose (Axhausen & Zürich, 2007). Early concepts of travel behavior focused on optimizing mobility through aggregated approaches, giving little or no attention to individual behaviors. However, the recent burgeoning availability of big data for transport applications has motivated researchers to seek detailed studies of individuals' travel patterns, variations, and overall behavior (Briand et al., 2017). Humans in nature are quite regular in daily travel from a spatial and temporal point of view, which deems them highly anticipatable. Nonetheless, some people change their travel behavior at some point due to numerous factors such as weather condition, holidays, social events, and other day-to-day interactions. Overall, travel behavior is believed to be neither completely consistent nor totally variable (Espinoza et al., 2018).

2.3.1 Travel Behavioral Analysis Significance

Travel behavioral analysis studies started more than 40 years ago. Transportation authorities had always emphasized the importance of such studies due to various reasons. Egu and Bonnel (2020) wrote that a full understanding of PT users' behavior is vital for developing and evaluating ridership improvement strategies. Goulet Langlois et al. (2016) implied that user segmentation based on behavioral differences is essential for executing effective traffic demand management campaigns. It is also useful for service planning due to the availability of information regarding different groups of users traveling along different parts of the PT network. This will eventually help enhance users' experience and overall satisfaction (Halvorsen, 2015). Furthermore, the study of users' mobility patterns is required for improving transport demand forecast and service adjustment. This could result in a reduction in operation cost, as well as an optimization in vehicle allocation across the network (Deschaintres et al., 2019). Moreover, understanding daily travel patterns is important for a better travel demand management assessment (Morency et al., 2006).

2.3.2 Travel Behavioral Analysis Dimensions

Travel behavior studies mostly relied on cross-sectional data where information was gathered directly from users through an active solicitation (Chen et al., 2016). Nevertheless, cross-sectional data collection techniques lack the daily variation factor and tend to be unstable in terms of travelers' behavior. Therefore, multiday data (such as smart card) is advantageous when dealing with travel behavior and its day-to-day variability (Egu & Bonnel, 2020).

Travel behavior across the transit network varies from one user to another, as well as for the same user over time. Interpersonal variability refers to the diversity in travel patterns among transit users. Intrapersonal, on the other hand, demonstrates the change in a user's travel behavior over time. Egu and Bonnel (2020) investigated interpersonal and intrapersonal variability simultaneously using a combination of a

similarity metric and a clustering technique. Deschaintres et al. (2019) used 51 weeks of SCD to analyze interpersonal and intrapersonal transit user behavior. Espinoza et al. (2018) assessed the stability of transit users' behavior by studying their intrapersonal variability.

The level based on which the travel behavior variability is examined varies based on data availability, convenience, and aim of the study. Some studies investigate hourly (Zhao et al., 2014), daily (Cats & Ferranti, 2022), weekly (Viallard et al., 2019), or in some cases a combination of various levels of analysis (Deschaintres et al., 2019). Furthermore, travel behavior can be analyzed with respect to spatial, temporal, or spatiotemporal variability. Cats & Ferranti (2022) implemented two different clustering approaches on SCD to segment PT users based on their temporal travel behavior. Tu et al. (2018) used a combination of two regression analyses to investigate spatial variations in travel behavior using SCD and GPS trajectories. Liu et al. (2022) utilized machine learning algorithms in travel segmentation based on weekly spatiotemporal travel behavior using SCD.

2.3.3 Travel Segmentation

Travel segmentation is one of the common approaches when studying PT travel behavior. It includes categorizing users into distinct groups with similar travel patterns based on a pre-determined similarity metric. Travel segmentation helps PT operators understand user behavior and therefore provide a better service based on actual mobility patterns. In addition, it allows operators to customize the provided service based on the segmented groups of users. Despite most users having a typical 5-days per week AM-PM commutes, nowadays a lot of businesses are adapting various working days and hours schemes. Therefore, travel segmentation as a tool has become more relevant than ever (Deschaintres et al., 2019).

2.4 SCD and Machine Learning Algorithms

Data mining is an interdisciplinary process of extracting and discovering patterns from a large dataset (Hastie et al., 2009). It involves the use of various machine learning techniques, such as clustering, regression, and dimensionality reduction. It also aims at finding correlations between multiple features in a dataset. With the increase in access to PT big data, numerous studies were done using data mining techniques for various purposes (Li et al., 2018; Ma et al., 2013).

Machine learning algorithms utilize artificial intelligence in executing human-like tasks. These algorithms leverage big data in such a way that it improves the performance, accuracy, and efficiency of such tasks (Yuan et al., 2021). Figure 2.1 summarizes machine learning techniques for data mining applications. Supervised learning techniques utilize a sample dataset to predict relationships and dependencies in the studied dataset. In other words, the algorithm approximates a function based on the sample data set and uses it to predict the output of another dataset.

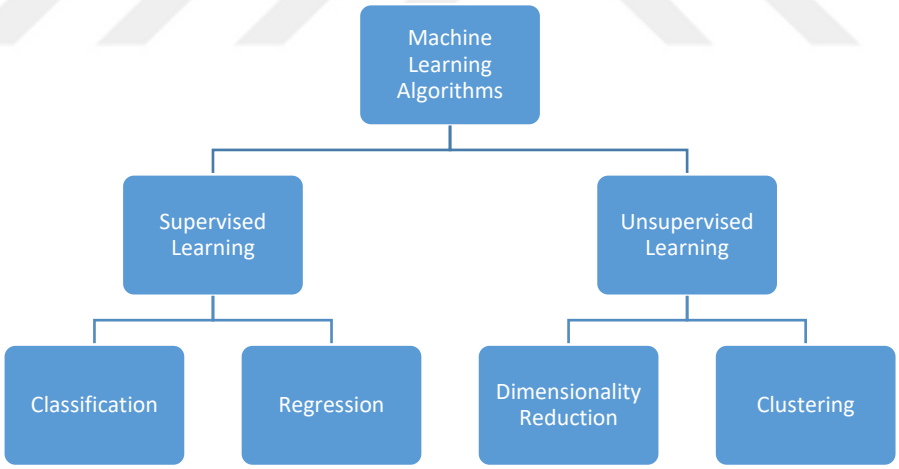


Figure 2.1: Main Machine Learning Algorithms

There are two main categories for supervised learning algorithms (SLA), classification, and regression. Classification refers to the process of predicting the category of a given data point based on the training dataset. There are various

applications of classification algorithms. A straightforward example is identifying weather a received email should be classified as spam or non-spam. The algorithm will use a training dataset where emails are already labeled as spam or non-spam. Based on the analysis of the training dataset, the classification algorithm will be able to predict the class (spam, non-spam) of newly received emails.

Regression however is used for understanding the relationship between features and outcomes using a training dataset. Once the relationship is determined, the outcome of the target dataset can be predicted (Martín et al., 2014). For example, Darshana Abeyrathna et al. (2021) utilized a regression algorithm in predicting passenger counts during a possible pandemic scenario using COVID-19 passengers transport data. The aim was to mitigate economic losses causes by unpreparedness of public transit authorities against pandemic related events.

Unsupervised learning techniques on the other hand can detect structures or patterns in a dataset via observation without the need for a sample dataset. Unsupervised machine learning methods can be categorized as clustering and dimensionality reduction (de la Torre et al., 2021). Clustering is the process of grouping data points into segments that have similar characteristics with respect to a certain similarity measure. The similarity between data point can be measured using various parameters such as shortest distance and density. Dimensionality reduction on the other hand refers to transforming a data set from a higher dimensional space into a lower one while maintaining the data's informational integrity (Van Der Maaten et al., 2009). For example, Kim et al. (2018) studied the correlation between spatiotemporal travel patterns and local environmental characteristics. There were 142 environmental variables that were used in the analysis. Thus, a dimensionality reduction technique was used to represent the data using far less components (variables) while maintaining informational integrity. This had helped reduce the computational complexity as well as improve parameter interpretation.

As mentioned earlier, supervised machine learning algorithms requires a training dataset based on which the algorithm can predict relationships or classes. To recall,

the aim of this study was to extract travel patterns from SCD directly without the need for a training dataset. Thus, choosing a supervised machine learning algorithm would not be effective. This narrows the available techniques down to clustering and dimensionality reduction. However, it's important to note that dimensionality reduction algorithms are often executed as an additional step prior to using other machine learning algorithms in order to transform the data into a more convenient structure. Therefore, clustering techniques will be further investigated to choose a suitable algorithm for this study.

2.4.1 Clustering Algorithms in Travel Behavior Analysis

Clustering algorithms has been extensively used as a data mining tool in travel behavior related studies. Various clustering approaches had been utilized in analyzing different aspects of travel behavior such as daily travel variability (Egu & Bonnel, 2020) , travel behavior consistency (Espinoza et al., 2018) , and behavior evolution (Viallard et al., 2019). Despite having a plenty of clustering algorithms, this section will focus on the ones that had been used for travel behavior analysis purposes.

Table 2.1 illustrates the main clustering algorithm utilized in travel behavior analysis in the literature. It's noticeable that transit bus networks are the most recurrent in behavioral analysis research. This is due to the heterogenetic nature of bus systems and its significance within the PT network. Metro and subway system come second, whereas ferries are less significant to travel behavior studies.

Table 2.1: Travel behavior analysis studies using clustering algorithms

Author(s)	**Transit modes**	**Objective**	**Clustering Algorithm**
(Cats & Ferranti, 2022)	bus, tram, metro, commuter train, ferry.	Mobility temporal patterns investigation	K-means, Hierarchical agglomerative clustering (HAC)
(Egu & Bonnel, 2020)	Metro, Tramway, Bus	Investigate daily variability	K-means, HAC
(Deschaintres et al., 2019)	Bus, Metro	User behavior analysis	K-means
(Briand et al., 2017)	Bus	Analyzing annual travel behavior variability	Gaussian Mixture Models (GMM)
(Ma et al., 2017)	Bus, Subway	Investigating commuting patterns	Spectral
(Zhao et al., 2014)	Bus, Metro	Understanding spatiotemporal travel patterns	K-means
(Ma et al., 2013)	Bus, Subway	Analyzing transit users' travel patterns	Density-based spatial clustering of applications with noise (DBSCAN)
(Agard et al., 2006)	Bus	Extracting Travel behavior regularity and daily patterns	K-means, HAC

Furthermore, some authors tend to combine two or more different clustering algorithms. This is usually done to eliminate the drawbacks of a certain clustering algorithm by integrating it with another. For example, Cats & Ferranti (2022) implemented a two-step approach were travel patterns are clustered using a K-means algorithm with a large number of clusters. The clusters are then grouped into a small number of profiles using a hierarchical clustering algorithm. This approach was used to overcome the disadvantageous computational complexity of hierarchical clustering. Therefore, having a thorough understanding of clustering algorithms' strengths and weaknesses is essential for choosing a suitable one.

Computational complexity is considered one of the most significant aspects of clustering algorithms. It represents the amount of recourses required to run an algorithm, particularly time and memory usage. Time complexity refers to the amount of time required to run an algorithm whereas space complexity represents the amount of memory space that an algorithm needs to run.

Table 2.2 shows time and space complexity of various clustering algorithms where n refers to the number of data points, k is the number of clusters and d is the number of dimensions or features. In addition, O which is also known as the big O notation refers to upper bond or the worst-case-scenario of running an algorithm. K-means clustering has the lowest time and space complexity with $O(nkd)$ and $O(n)$ respectively. HAC and spectral clustering have a second order space complexity and a second and third order time complexity for HAC and spectral respectively. DBSCAN on the contrary have a low time complexity of $O(n \log n)$. However, when the number of features (dimensions) increases, the time complexity increases up to $O(n^2)$. GMM clustering's computational complexity is strongly dependent on the number of features with a cubic and square time and space complexity respectively. Therefore, the algorithm computational cost increases dramatically with higher dimensional data. Nevertheless, apart from computational complexity, other aspects of clustering algorithms must be considered while selecting a suitable one such as data compatibility, sensitivity to outliers, performance with big data, and flexibility.

Table 2.2: Computational complexity of clustering algorithms (Bawane, 2017; Dunhan, 2006; Xu & Wunsch, 2005)

Algorithm	Time complexity	Space Complexity
K-means	$O(nkd)$	$O(n)$
HAC	$O(n^2)$	$O(n^2)$
Spectral	$O(n^3)$	$O(n^2)$
DBSCAN	$O\ (n \log n)$*	$O(n^2)$
GMM	$O(nkd^3)$	$O(nkd^2)$

* $O(n^2)$ for higher dimensional data

Table 2.3 summarizes the advantages and disadvantages of the clustering algorithms. K-means clustering surpasses all other algorithms in terms of simplicity and compatibility with big data (Sehgal & Grag, 2014). However, the number of clusters must be predefined, which may be difficult in some cases. Furthermore, K-means clustering is sensitive to centroid initialization which may lead to inconsistency in the output (Deepti et al., 2015). HAC algorithm produces a hierarchy of clusters represented by a dendrogram, which is helpful for data visualization. It also does not require predefining the number of clusters which is a significant advantage (Abu Abbas, 2008). However, HAC does not perform well with big data, and it has a very high computational complexity as show in Table 2.2. Spectral clustering has a flexibility of working with different cluster shapes. It also does not necessarily require the full dataset to get executed, since it can be performed with the similarity matrix only(Ding et al., 2010). Nonetheless, similar to K-means it requires predefining the number of clusters and it is sensitive to outliers (Rodriguez et al., 2019). The DBSCAN algorithm has an advantage of performing well with arbitrary shaped clusters. In addition, it can detect noise within the data making it robust to outliers (Deepti et al., 2015). However, DBSCAN does not perform well when clustering high dimensional data (Rodriguez et al., 2019). Similar to DBSCAN, GMM clustering performs well with various cluster shapes. It is also considered a soft clustering technique where data points can belong to more than one cluster. Yet, it has a slow convergence rate compared to other clustering algorithms, and it is sensitive to centroid initialization (Shireman et al., 2017).

Table 2.3: Advantages and disadvantages of clustering algorithms

Clustering algorithm	Advantages	Disadvantages
K-means	• Simple to implement and run. • Performs well with big data. • Produce tight clusters. • Results are easy to interpret.	• Number of clusters must be predefined. • Sensitive to centroid initialization. • Performs well only with numerical variables.
HAC	• Does not require predefining the number of clusters. • Produces a dendrogram that helps visualizing the data. • Shows the hierarchical relationship between clusters.	• Does not work well with big data. • Lack of making corrections to clusters splitting/merging. • Result interpretation may be subjective.
Spectral	• Performs well with different cluster shapes. • Does not require the dataset, can be performed with the similarity matrix.	• Number of clusters must be predefined. • Does not perform well with noisy datasets.
DBSCAN	• Does not require predefining the number of clusters. • Performs well with arbitrary shaped clusters. • Ability to specify noise within the data (robust to outliers).	• Does not work well with varying density clusters. • Does not perform well with high dimensional data.
GMM	• Flexibility of cluster shapes. • Soft clustering technique (Data points' membership of multiple clusters).	• Sensitive to centroid initialization. • Slow convergence rate.

As mentioned earlier, K-means outperforms other clustering algorithms in terms of time and space complexity. It has also been proven that K-means is advantageous when dealing with big data, such as SCD. In line with the aim of the study being the construction of a simple and efficient clustering algorithm for the purpose of behavioral analysis using SCD, K-means clustering will be chosen as the primary data mining tool. The following section will briefly describe K-means clustering algorithm, with an emphasis on relevant travel behavioral analysis research.

2.4.1.1 K-means Clustering

K-means Clustering refers to an unsupervised machine learning algorithm where an n number of observations are partitioned into K number of pre-defined clusters for a certain dataset. A cluster is defined as a collection of observations grouped based on certain similarities (Rodriguez et al., 2019). The similarity measure for K-means clustering is the distance between points. The algorithm works as follows:

1- **Setting K:** The number of clusters based on which the data is partitioned is defined as K. This parameter must be pre-defined prior to running the algorithm.
2- **Centroid Initialization:** There are various methods of selecting the initial clusters' centroids. The traditional approach is the random initialization of centroids. However, other smart methods are used such as K-means++ where the centroids are distributed as far as possible from each other.
3- **Data Points Assignment:** The distances between all data points and all centroids are computed, and then each data point is assigned to the closest centroid. There are different types of distance measurement approaches, some of which are:
    A. Euclidean: The straight-line distance between two points in a Euclidean space, which is given by the following formula:

$$ED_{ab} = \sqrt{\sum_{k=1}^{m}(a_{ik} - b_{ik})^2}$$

B. Manhattan: The sum of the absolute coordinates between two points given by the formula:

$$MD_{ab} = \sum_{k=1}^{m}|a_{ik} - b_{ik}|$$

C. Chebyshev: The maximum difference in coordinates between two points, which is represented as:

$$CD_{ab} = max\left(\sum_{k=1}^{m}|a_{ik} - b_{ik}|\right)$$

4- **Centroid Re-assignment:** The initially assigned centroids are recalculated based on the average of points in each cluster previously assigned in step 3.

5- **Iteration:** Steps 3 and 4 are repeated until convergence, a point where the centroids are no longer changing (Figure 2.2).

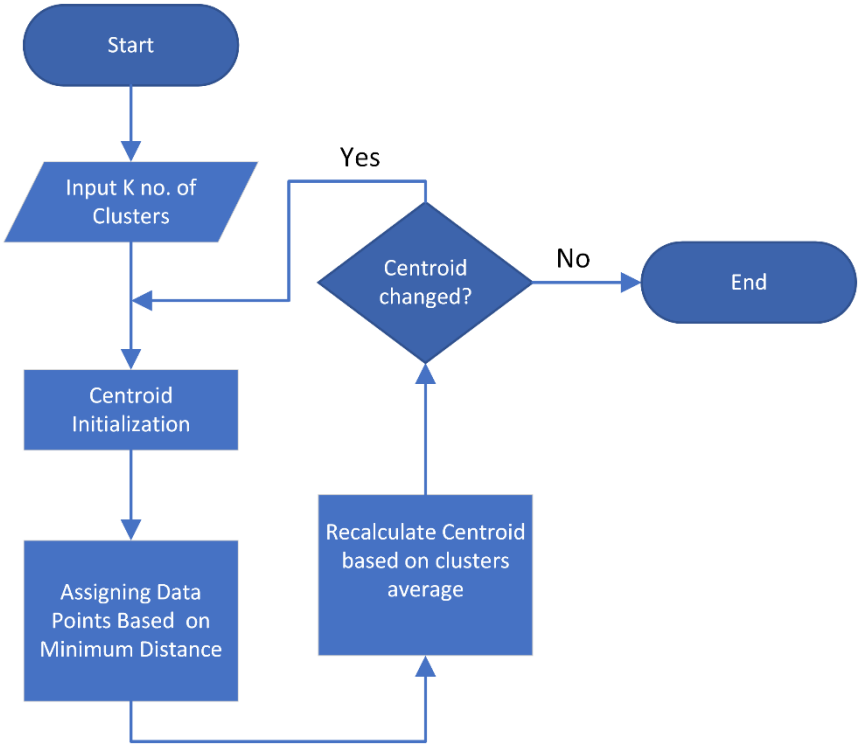


Figure 2.2: K-means Clustering Algorithm Flowchart

The K-means clustering algorithm had been extensively utilized in travel behavioral analysis studies using SCD. Deschaintres et al. (2019) used K-means clustering to cluster card-week vectors and examine intrapersonal PT use variability. Viallard et al. (2019) constructed weekly based transaction profiles and clustered them using a K-means algorithm to further investigate travel patterns. Furthermore, Zhao et al. (2014) extracted temporal and spatial features of users and then clustered them using K-means to understand in-depth the various groups of PT users. In addition, Briand et al. (2017) used a mixture of K-means clustering and gaussian mixture model algorithms to analyze PT user behavior over 5 years of SCD

CHAPTER 3

CASE STUDY

The city of Konya is located in the south-central region of Turkey and is considered the largest province in terms of surface area at approximately 38,257 km^2 (Uyan, 2014). It consists of 31 districts with a population of 2.27 million by the year 2022 according to the Turkish Statistical Institute (TÜİK). Konya has been inhabited for thousands of years and ruled by numerous civilizations, which gives it a great historical significance that reflects on the city's current demography, economy, and infrastructure. Despite Konya being known for its agriculture and archeology in modern Turkey, nowadays it's being steered towards industrialization. Therefore, the metropolitan municipality of Konya has sought to develop a robust public transport system that meets the needs of its inhabitants.

3.1 PT System in Konya

The public transit of Konya consists mainly of Tram, Bus, and Minibus networks. Right around the mid-1940s, the first two PT buses were purchased by the municipality and started operating on two different routes. By 1992, the first tram line was opened operating between Alaaddin and Cumhuriyet with a 10.5 km long route. Nowadays, the PT fleet consists of 663 buses, 4 electric buses, and 112 trams over a 54 km long tramway. Figure 3.1 shows Konya province and Konya's metropolitan municipality rejoin (Konya city).

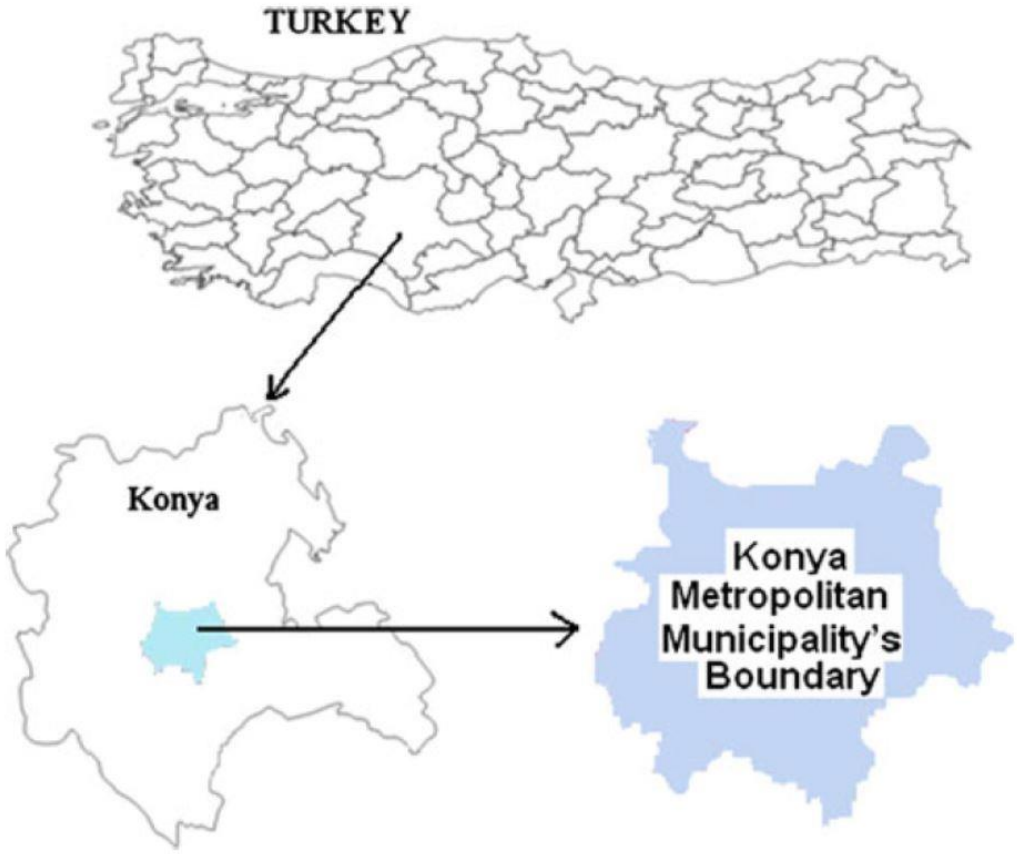


Figure 3.1: The geographical position of Konya District and Konya's Metropolitan Municipality boundary (Uyan, 2014)

As mentioned earlier, the province of Konya consists of 31 districts three of which are part of Konya metropolitan municipality's boundary which are Meram, selçuklu, and Karatay. The PT bus system's coverage extends to all districts of Konya province whereas the tram system is limited to Konya city. Within the city of Konya, the majority of PT trips along the northern axis are made by tram and minibus. The city center however shows a concentration of PT bus usage along with private vehicles which aggravate congestion and impede mobility (Uyan et al., 2017).

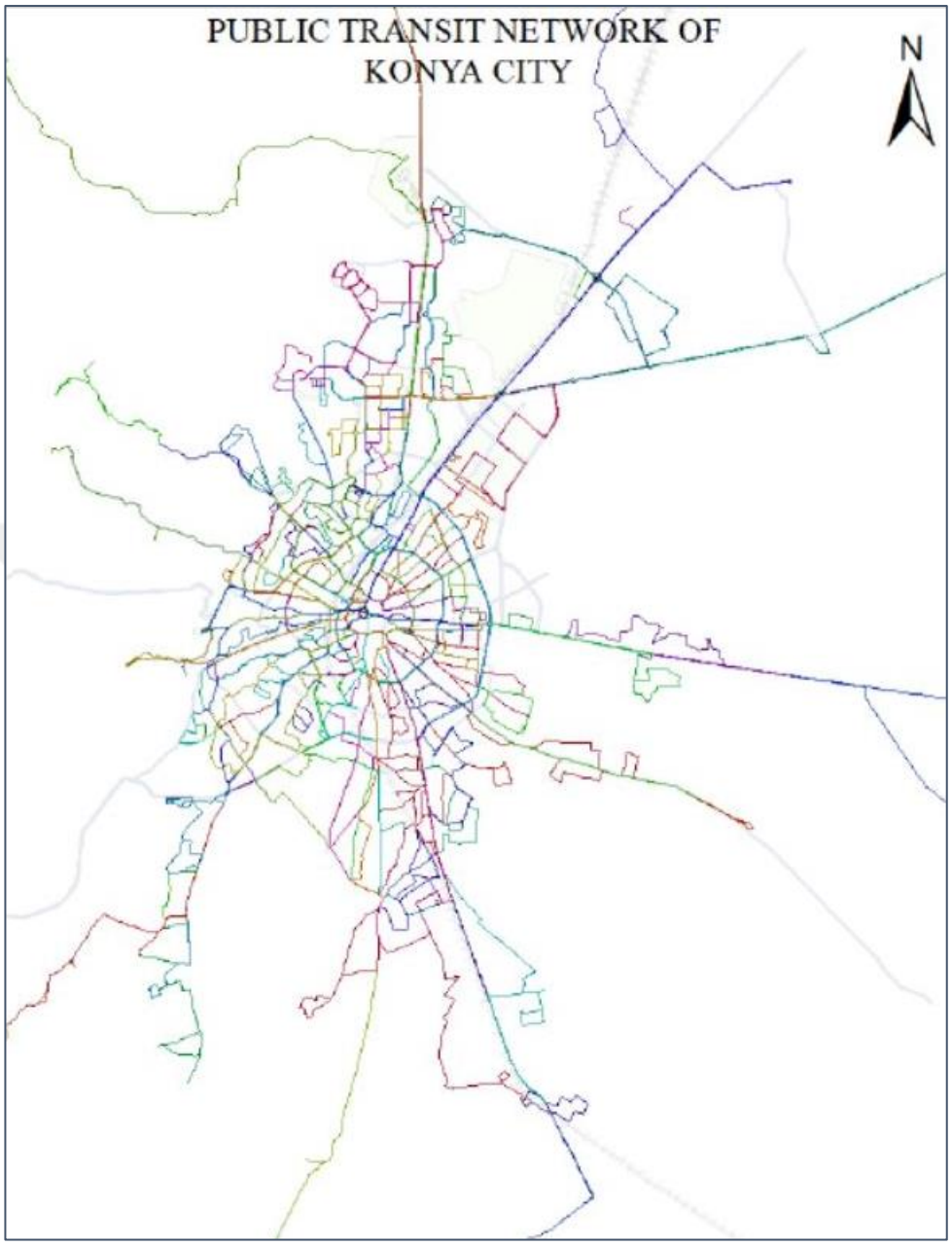


Figure 3.2: Konya city's PT Network

3.2 Konya PT SCD

ATUS refers to the Intelligent PT System developed by Konya Metropolitan Municipality to provide PT users with a convenient service. The system does not only provide Smart Cards for fare payment, but also assistance with regards to Bus

and Tram schedules, expected waiting time, and navigation through various platforms such as website, mobile app, and in-vehicle display monitors.

With the increasing popularity of PT Smart Cards in Konya, data collection and management came into the picture. SCD used in this study is provided by the Metropolitan Municipality of Konya. The data consists of 7,550,849 Smart Card transactions during May 2018. Thus, the duration of this study will be 1 month (31 days).

Figure 3.3 shows the PT Smart Card transactions across Konya province. Bus transactions are recorded in various districts across the province, unlike tram transactions which are concentrated in the city of Konya. Moreover, a high density of bus transactions is noticeable in Konya city, as opposed to other districts, which have a significantly lower number of transactions.

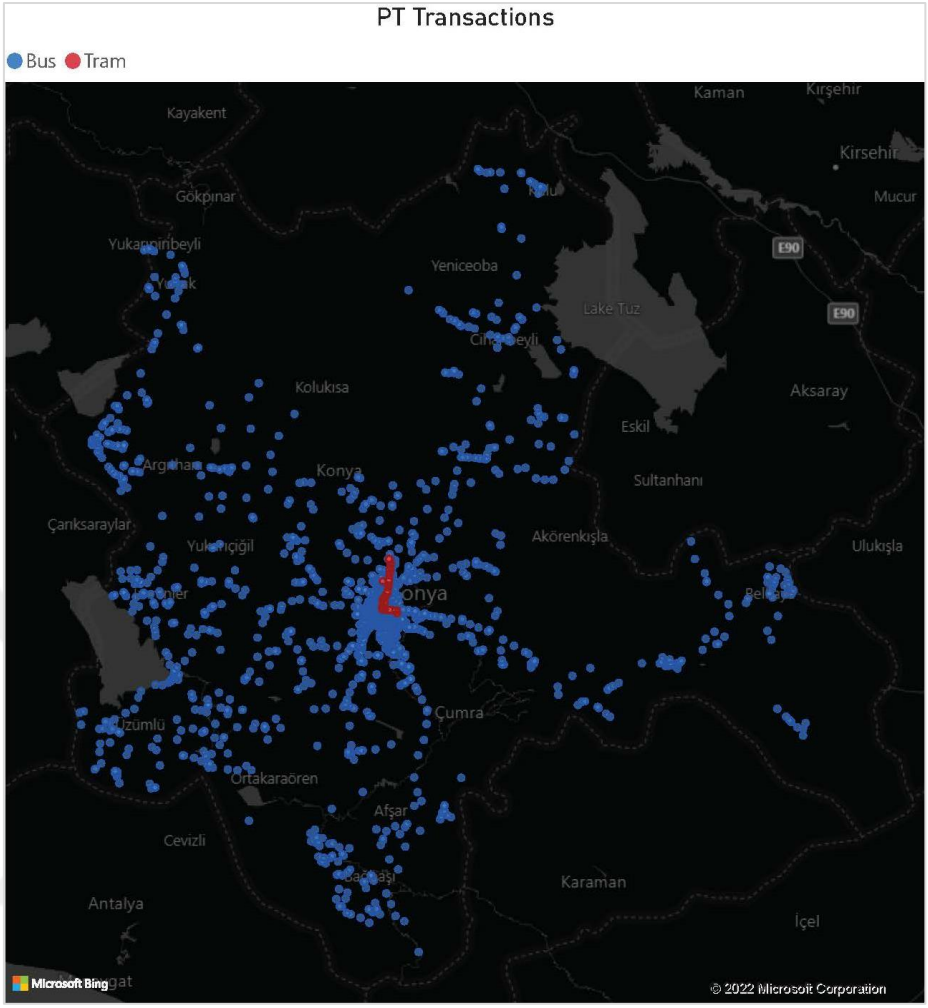


Figure 3.3: Konya District Map with PT Smart Card Transactions During May 2018

When a Smart Card is used, a number of attributes are recorded in Konya municipality's database. Table 3.1 shows the attributes of the Smart Card transactions and their description. TXN_TYPE, CARD_ID, CARD_TYPE, and TIMESTAMP are the main parameters used in this study.

Table 3.1: SCD Attributes

Attribute	Description
VEH_ID	The ID of the vehicle on which the transaction was recorded
TXN_TYPE	Transaction type: Standard use, Daily card use, Blocked card, and Transfer use.
CARD_ID	32-character long Smart Card serial number
CARD_TYPE	Smart Card type: Standard card, Discount card, 65+ citizen card, Subscription, Personnel and Free.
LATITUDE/ LONGITUDE	GPS location at which the transaction occurred.
TIMESTAMP	Date and time of the transaction record.
LINE_ID	Service line identification number.
SUB_LINE_ID	Sub-line identification number.

3.3 Descriptive Statistics

Prior to discussing users' travel behavior through PT SCD, it's vital to examine some of the descriptive statistics features. The following section presents daily and hourly transaction volumes, usage per card type, weekdays and weekends volume, as well as daily usage frequency.

3.3.1 Hourly and Daily Transaction Volume

Average hourly transactions were calculated separately for bus and tram during weekdays, Saturdays, and Sundays using a python code. This code extracts the corresponding transactions for each hour in a day for all days of the study period from the raw data. As shown in Figure 3.4, Weekday's average hourly transaction volume (AHTV) comprises a higher value during the morning peak. However, Saturday and Sunday's AHTV surpasses right after the morning peak up till

midnight. In addition, Weekday's profile displays a more conspicuous two AM-PM peaks, unlike Saturdays and Sundays. As for the tram's AHTV, the peak starts at 5:00 and increases dramatically till 7:00 and 8:00, for weekdays and weekends respectively. The bus's peak, however, starts at 5:00 on weekdays and Saturdays, and at 6:00 on Sundays.

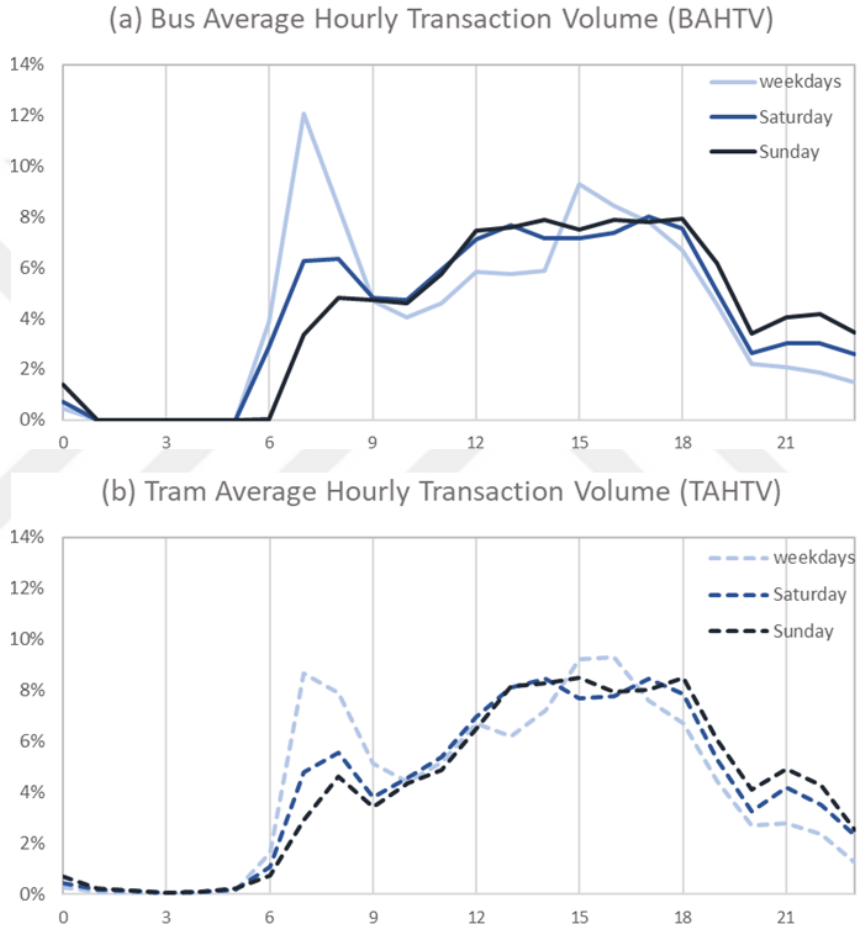


Figure 3.4: Average hourly transaction volume for (a) bus and (b) tram during weekdays, Saturdays, and Sundays

The number of total daily transactions was calculated for each day of the month including weekends and weekdays. This was done to demonstrate the variation in smartcard daily usage during the study period. As shown in Figure 3.5, bus usage is almost twice as much as the tram. In addition, a considerable drop in PT use is noticeable during public holidays. Moreover, there is an overall higher level of usage

during weekdays and a lower one on Saturdays and Sundays respectively. Furthermore, a slightly higher than usual usage is observed during the 14th and 15th of May 2018. This is because those two days were right before the beginning of the holy month of Ramadan (16th of May 2018).

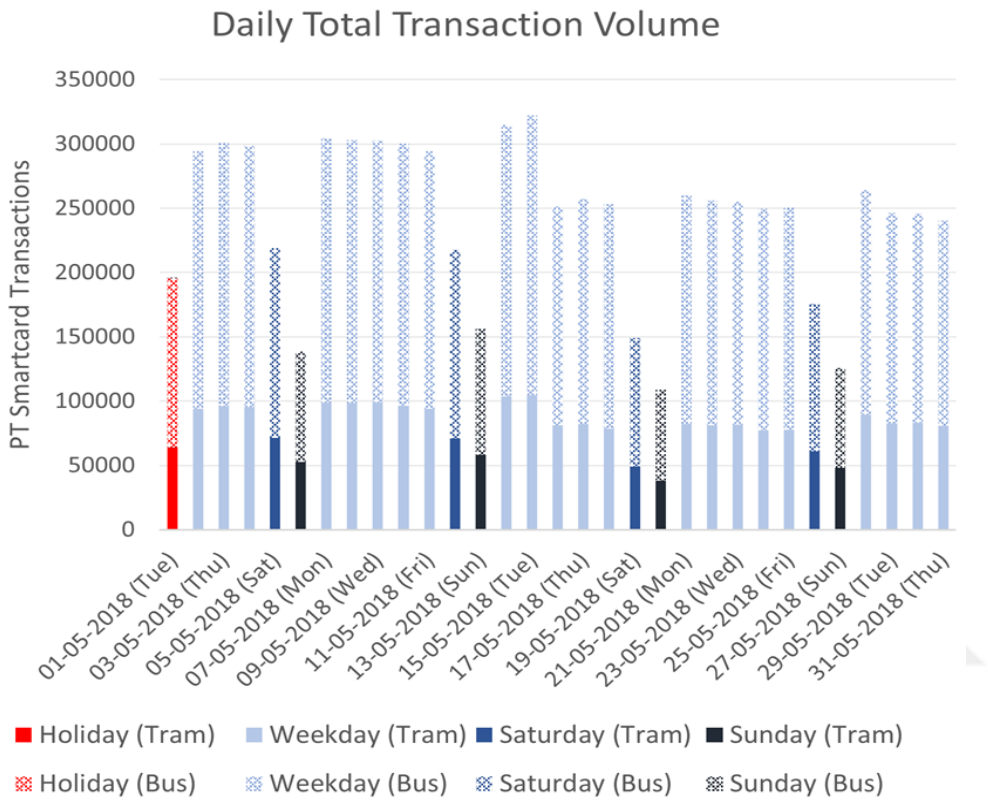


Figure 3.5: Daily total transaction volume

3.3.2 PT Usage per Card Type

The counts of transactions corresponding to a specific card type were extracted from the raw data and used to demonstrate the percentage of card type usage during the study period. Discount cards that are given to students and teachers comprise almost 50% of the transactions as shown in Figure 3.6. Standard card type comes second with 30%, followed by subscription and 65+ citizens with almost 9% and 8% respectively. The rest of the ticket types contribute to a far less fraction of usage.

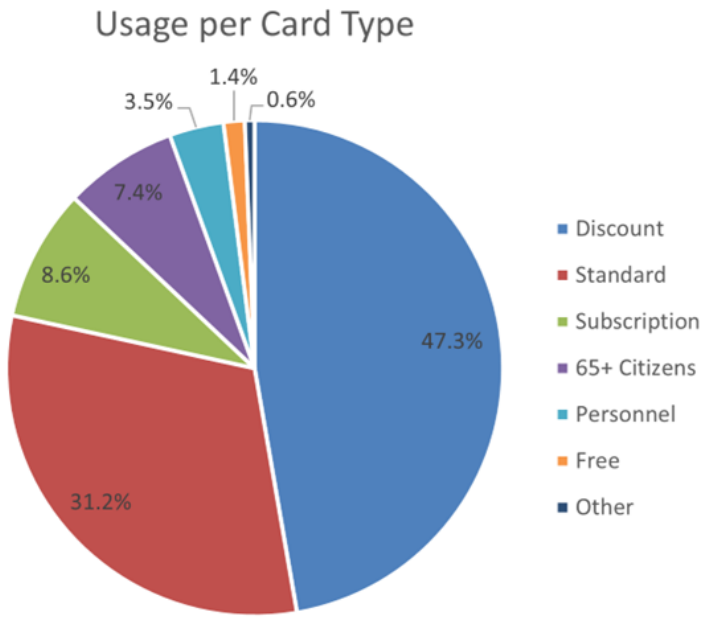


Figure 3.6: PT usage per card type

Figure 3.7 demonstrates the temporal variation in smartcard usage with regard to card types during weekdays and weekends. During weekdays, the discount card type contributes to a higher portion of the transactions than any other card type, especially during peak times. Nevertheless, a slight drop in discount card usage is noticed during weekends with an increase in standard and subscription usage. Furthermore, 65+ citizen card type usage is concentrated around midday during weekdays and stretched further towards evening during weekends.

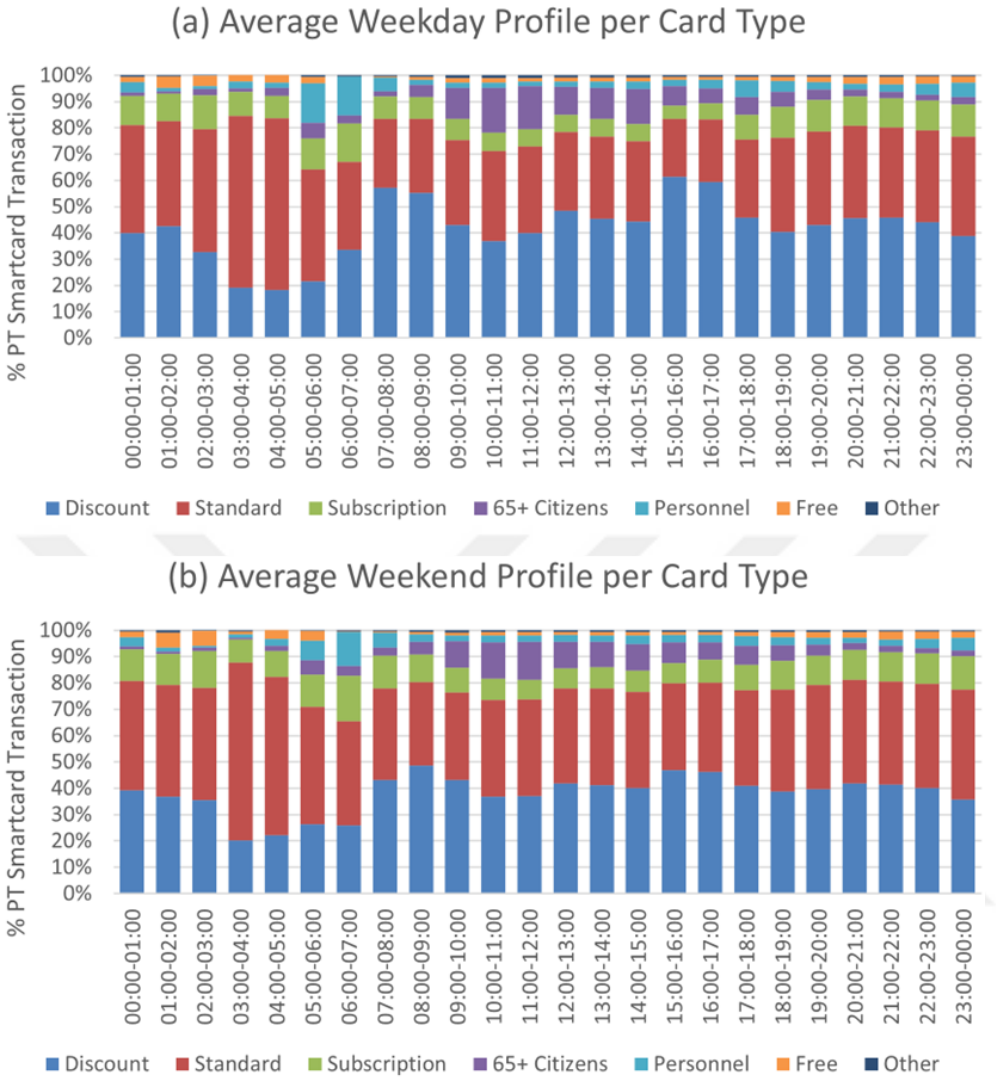


Figure 3.7: (a) Average weekday profile per card type (b) Average weekend profile per card type

3.3.3 PT Usage Frequency

A python code has been developed to calculate the card usage frequency. The code calculates the number of times each cardholder uses PT every day of the study period. Figure 3.8 represent the daily transaction frequency. Almost 40% of the card holders use their cards twice a day, whereas 11%, 10%, and 5% use their cards 3, 4, and 5 times or more respectively. Once per day usage is considerably high with 34% of transactions.

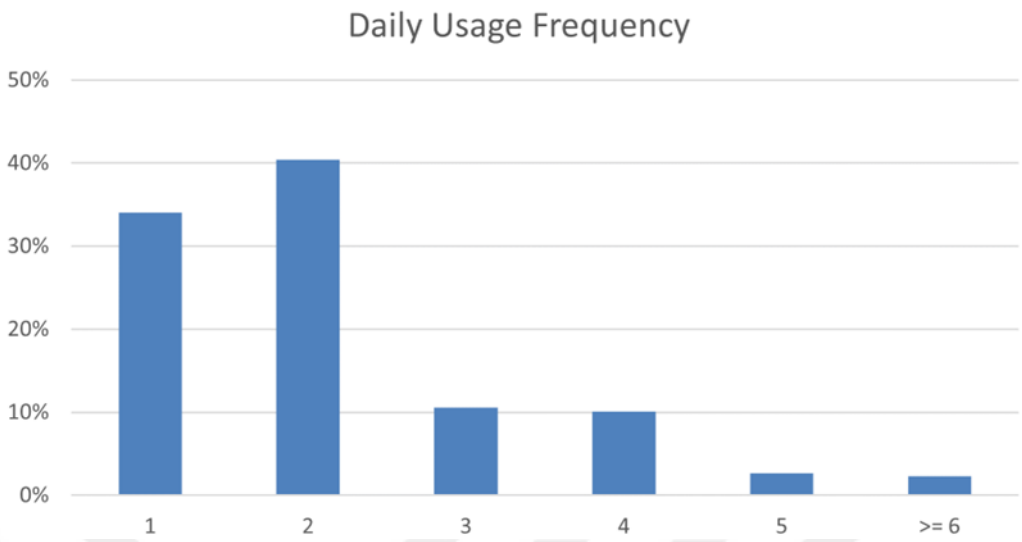


Figure 3.8: Daily usage frequency

CHAPTER 4

METHODOLOGY

4.1 Framework

The methodology presented in this study consist of 2 major sections, data preprocessing and K-means clustering. The preprocessing step include data cleaning were missing or duplicate values are removed. Data re-formatting where raw data is transformed into a suitable format for clustering algorithm. Daily boarding vectors formation which will be the input to the K-means clustering function. Data normalization which transforms the features into a desirable scale. The K-means clustering section includes label assignment were daily boarding vectors get assigned into clusters. It also includes storing results and exporting it where further data visualization is performed. All data-related tasks were performed using python programming language, whereas visuals were prepared using Microsoft Excel. Figure 4.1 summarized the methodology framework of this study.

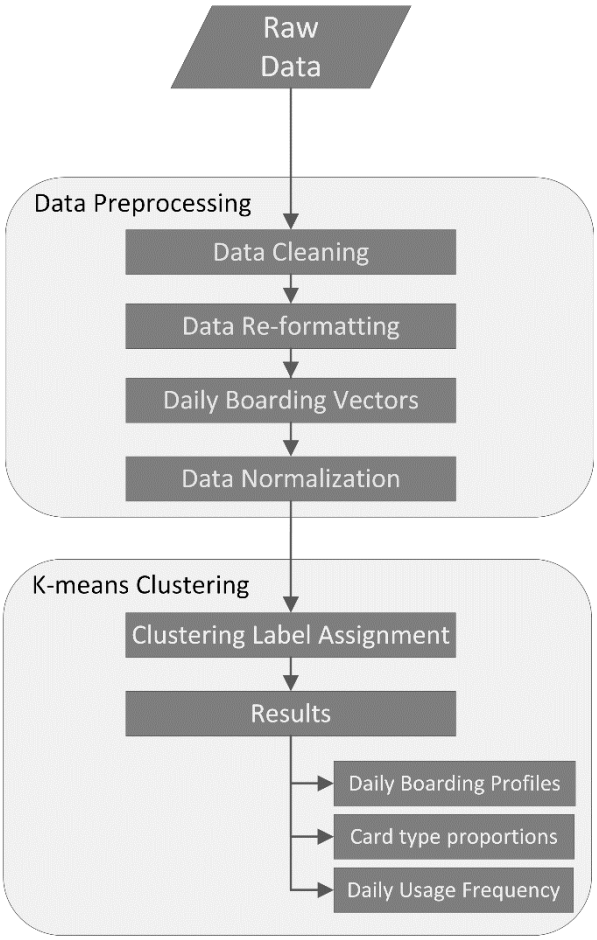


Figure 4.1: Methodology Framework

4.2 Clustering

4.2.1 Data Preprocessing

Data preprocessing is a data mining technique where raw data is transformed into a useful and efficient format. The SCD which was retrieved from Konya metropolitan municipality is illustrated in Table 4.1.

Table 4.1: Smart Card Raw Data received from Konya Metropolitan Municipality

VEH_ID	TXN_TYPE	CARD_ID	CARD_TYPE	LATITUDE	LONGITUDE	TIMESTAMP	LINE_ID	SUB_LINE_ID
844	BILET	4A73…B8F	32	379245177	325021050	2.05.2018 02:17	53	0
172	BILET	4215…21C	0	378720600	324976650	5.05.2018 16:54	10	2
469	BILET	D3E6…414	0	378119883	325203767	13.05.2018 19:11	124	0
715	ABILET	6029…3D8	2	378737350	324888233	21.05.2018 22:31	2	1

Data cleaning: The first step was to check for missing or duplicate values. This check can be easily done using the Pandas library in python. Luckily, the only missing values found were latitudes and longitudes for some transactions, which are irrelevant for this study. Furthermore, transactions recorded by cards that have more than one card type throughout the study period were removed. This usually occurs when a user changes their card type (e.g., standard card to subscription card) without purchasing a new card. The total number of transactions dropped were 5841, which corresponds to 0.07% of the data.

Data re-formatting: The second step included changing the format of the features in such a way that it would be suitable for the data analysis. The following table represents the new data format.

Table 4.2: Smart Card Re-Formatted Data

INDEX	CARD_ID	CARD_TYPE	DATE	TIME	DAY_TYPE
0	4A73…B8F	1	2	2	2
1	4215…21C	0	5	16	5
2	D3E6…414	0	13	19	6
3	6029…3D8	5	21	22	0

It’s noticeable that some features were excluded since they are extraneous to this study. In addition, the timestamp was split into date and time, whereas card type was given different notations to ease its illustration later on. Moreover, two more columns were added, one representing the serial number of transactions or INDEX, the other

however is a new feature named DAY_TYPE representing the day of the week on which the transaction occurs.

The next step transforms the instantaneous transactions of all users to daily boarding profiles as follows:

1. CARD_TYPE and DAY_TYPE columns were dropped temporarily.
2. TIME feature was converted into an indicator variable with 24-time slots as shown in the table below.

Table 4.3: SCD Preprocessing (TIME as an indicator variable)

INDEX	CARD_ID	DATE	TIME_00	TIME_01	TIME_02	…	TIME_22	TIME_23
0	4A73…B8F	2	12	0	1	…	0	0
1	4215…21C	5	16	0	0	…	0	0
2	D3E6…414	13	19	0	0	…	0	0
3	6029…3D8	21	22	0	0	…	1	0

All transactions occurring on a specific date for a certain user CARD_ID are summed up, resulting in a daily boarding profile for all users on every day of the study period. The equation below provides the mathematical formulation of daily boarding profiles.

$$v_{i,d} = \left[\sum t_i(h_1) \ , \sum t_i(h_2) \ , \ldots , \sum t_i(h_{24}) \right] \quad (1)$$

Where:

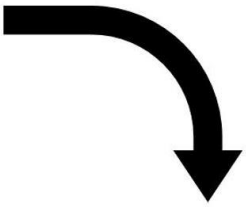
$v_{i,d}$ = Daily Boarding Profile for user i at day d.

$t_i(h)$ = Transaction of user i at hour h of the day d.

To better understand how this works, say for instance the person holding CARD_ID no. X had used public transit 7 times throughout the study period over 4 different

days at certain times. In that case, Figure 4.2 illustrates the transformation of this person's usage into daily boarding profiles.

No.	DATE	TIME
1	2	2
2	2	2
3	2	23
4	6	10
5	15	4
6	15	15
7	29	22



INDEX	DATE	TIME_00	TIME_01	TIME_02	TIME_03	TIME_04	TIME_05	…	TIME_22	TIME_23
0	2	0	0	2	0	0	0	…	0	1
1	6	0	0	0	0	0	0	…	0	0
2	15	0	0	0	0	1	0	…	0	0
3	29	0	0	0	0	0	0	…	1	0

Figure 4.2: Example of data transformation into daily boarding profiles

The above highlighted daily boarding profiles are the input to the K-means clustering algorithm. Other features such as date, card type, and day type were excluded because the algorithm is meant to segment users' behavior based on their daily usage, regardless of what card type they use or what day of the week they travel on.

Data normalization: The last step before running the algorithm is normalization. Data normalization refers to the process of changing the numeric values of columns in the data to a common scale. This ensures equal consideration by the algorithm for all the features. Furthermore, minimizing the features' range with normalization reduces the overall cost of the algorithm, leading to a higher convergence rate (Mohamad & Usman, 2013). There are various normalization techniques, such as Z normalization, Min-Max normalization, and Unit Vector normalization. For methodological simplicity, the Min-Max normalization will be used. For dataset N with i number of rows and j columns, the normalized value is calculated as follows:

$$N(X_{ij}) = \frac{X_{ij} - \min(X_j)}{\max(X_j) - \min(X_j)} \quad (2)$$

4.2.2 Number of Clusters Selection Criteria

To run a K-means clustering algorithm, the number of clusters must be pre-determined. Various methods are traditionally used to determine the optimal number of clusters, such as the Elbow Method, the Silhouette Score, and the Gap Statistic. For this study, Scikit Learn's silhouette score function had been used to plot the silhouette score for values of K number of clusters ranging from 2 to 15. The silhouette score is defined as:

$$S_i = \frac{b_i - a_i}{max(b_i, a_i)} \quad (3)$$

Where:

a = Mean distance between point i and all other points in the same cluster.

b = Mean distance between point i and all other points in all other clusters.

The score is then calculated by averaging S for all data points, with a value that ranges from -1 to 1. The silhouette score defines how tightly the data points are grouped in a cluster and how separated the data points in a certain cluster are from other clusters. Thus, it is crucial to note the following:

- A value near 1 implies well-defined distinguishable clusters.
- A value of 0 means that clusters are significantly close to each other.
- A value near -1 means that data points are assigned to clusters incorrectly.

However, due to a computational complexity limitation, 4 random samples of 15% each were used to calculate the silhouette score as shown in Figure 4.3.

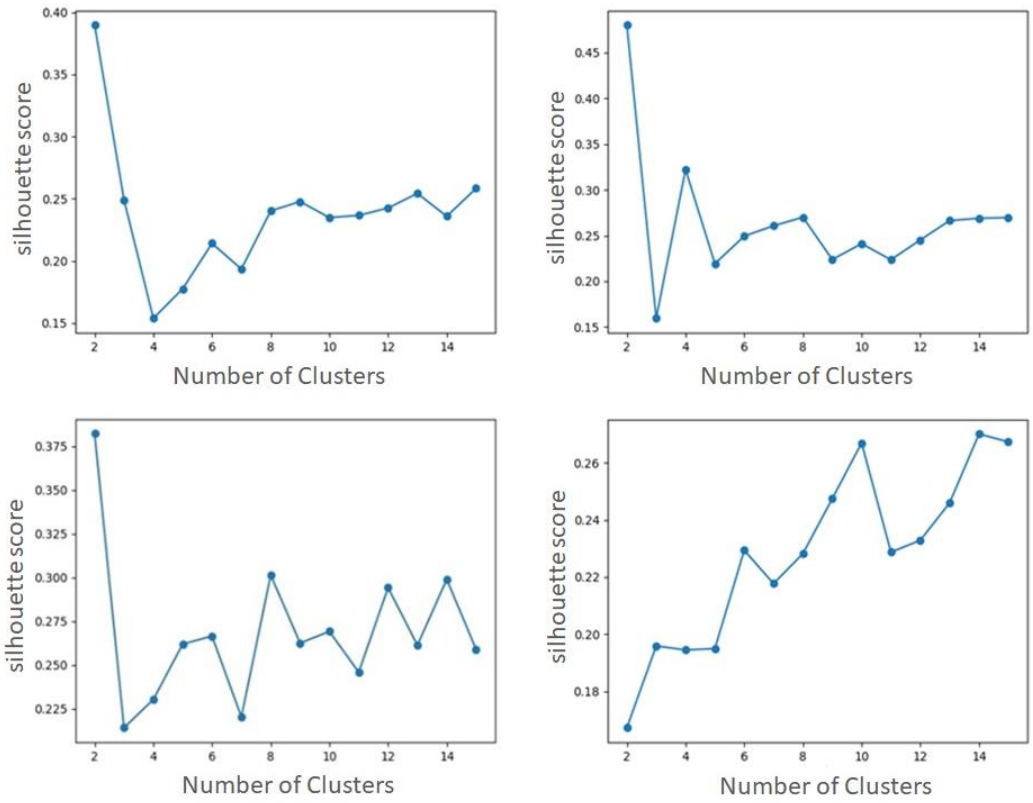


Figure 4.3: The Silhouette Score of 4 Samples.

The fluctuating pattern in the above graphs is indisputable. Despite having an overall direct proportionality in the 4th sample, it's clear that the silhouette score for that sample value is relatively low compared to the other samples. Therefore, this study is going to aim at examining the evolution of data segmentation as the number of clusters changes, rather than choosing an optimal number of clusters. The K-mean clustering algorithm will be executed 5 times over 2, 4, 6, 8, and 12 clusters. While doing so, the change in users' membership within clusters over the study period will be investigated, taking into account the user's card type, and daily usage frequency.

4.2.3 K-means Clustering Algorithm

As mentioned earlier, the programming language used for running the K-means clustering algorithm is python. Therefore, Scikit-learn, which is a machine learning library for python, is used to implement the unsupervised machine learning

clustering algorithm. Table 4.4 summarizes the set of parameters deemed necessary to run the algorithm. The centroid initialization method is chosen to be “K-means++” since it eliminates the drawback of K-means being dependent on centroid initialization. The parameters “init” and “n_init” are sat to ensure convergence before the algorithm terminates. In addition, the “random_state” parameter is essential for the consistency of the results through K = 2 to K = 12 runs. Furthermore, the chosen algorithm method is “Lloyd” since it is less memory intensive than other available methods such as “Elkan”.

Table 4.4: K-means clustering algorithm parameters

Parameter	Description	Value
n_clusters	The number of clusters and centroids to generate.	2,4,6,8,12
init	The centroid initialization method.	“K-means++”
n_init	The number of K-means algorithm runs with different centroid arrangements. The final results correspond to the best output in terms of inertia.	10
max_iter	The algorithm’s maximum number of iterations for a single number of clusters run.	300
random_state	The lot number of the initial randomly generated centroids.	470
algorithm	The type of K-means clustering to be used	“Lloyd”

CHAPTER 5

RESULTS AND FINDINGS

5.1 K-means Clustering Results

The K-means clustering algorithm was executed for 2, 4, 6, 8, and 12 clusters. On each run, the Average Hourly Transaction (AHT) was calculated for all clusters at a 1-hour interval. The unique pattern resulting is referred to as the Daily Boarding Profile (DBP). In addition, the cluster's share is included to demonstrate the proportion of daily boarding vectors associated with a certain pattern. Moreover, the user's card type percentage and daily usage frequency were determined for each cluster. This section discusses in detail the resulting DBPs, as well as the relations between travel behavioral patterns, card type, and daily usage frequency.

5.1.1 K = 2 Clusters

The first algorithm run has a K number of clusters equal to two. As shown in Figure 5.1, cluster 1's daily boarding profile shows a sharp AM peak around 6:00 with an average hourly transaction of 1.2, and a significantly smaller peak from 15:00 to 18:00 with an almost 0.2 average hourly transaction rate. This cluster's share of the daily boarding vectors is roughly 5% which corresponds to a very small proportion of the PT users. On the other hand, cluster 2 represents around 95% of the daily boarding profiles with a lower average hourly transaction rate of 0.1 to 0.2. Similarly, a slight AM peak at 7:00 is noticeable with an almost consistent rate throughout the day up until 16:00, where the rate starts dropping gradually till towards the end of the day.

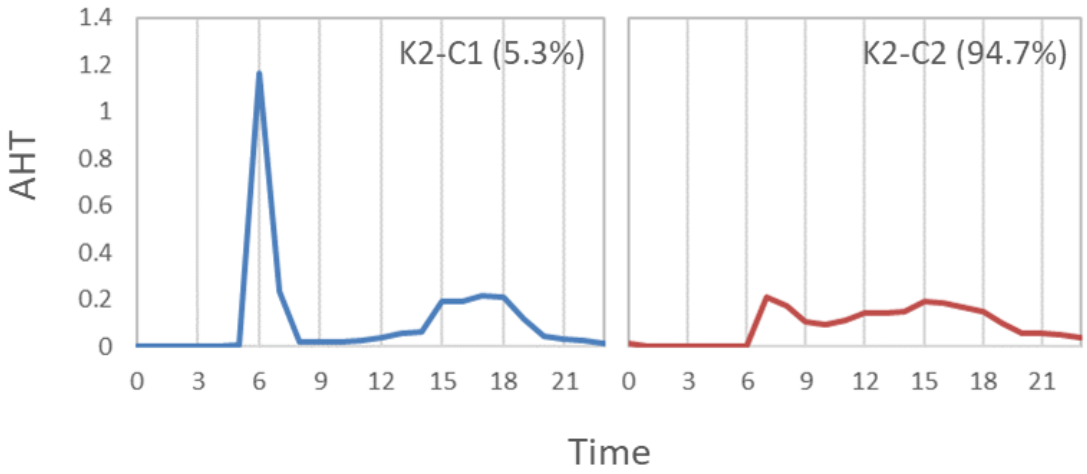


Figure 5.1: Daily Boarding Profiles for K = 2 Clusters

The composition of daily usage frequency for each cluster is shown in Figure 5.2(a). As for cluster 1, up to 41% of the users use public transit twice a day. Despite being unreasonable at first glance since cluster one's daily boarding profile shows one dominant AM peak, the reason behind this is that users belonging to this cluster mostly commute at 7:00 yet perform their second trip within a wider period from 15:00 to 18:00, causing a lower hourly transaction rate that spreads over a longer time period. In addition, cluster 1 has a higher 4, 5, and 6 times per day usage relative to cluster 2. However, cluster 2 has a higher once-per-day usage at approximately 35% compared to 18% for cluster 1.

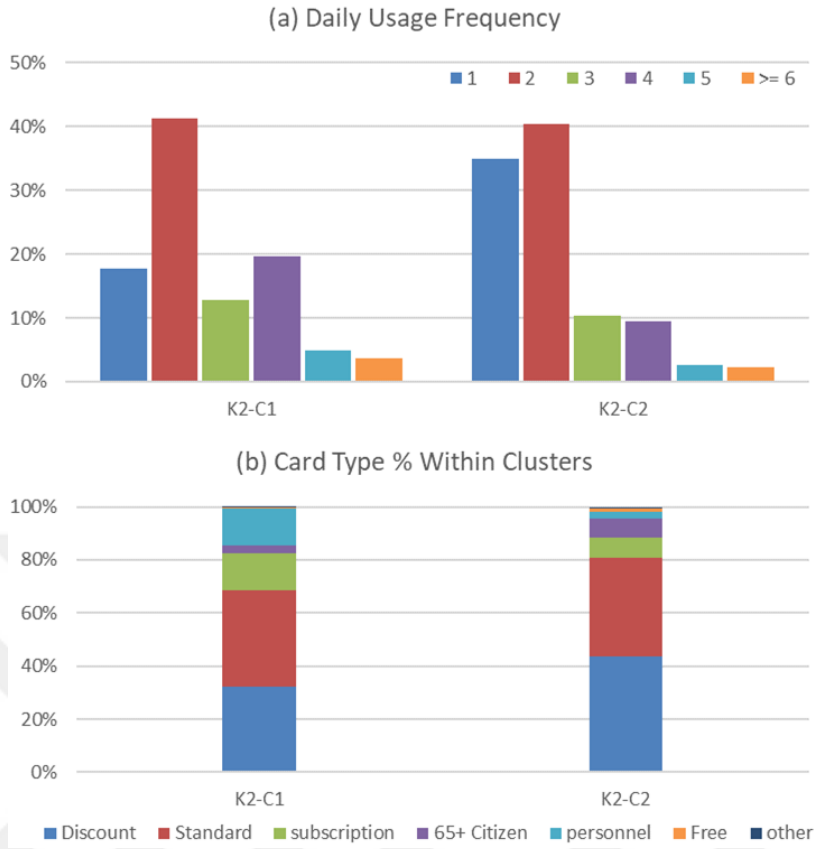


Figure 5.2: (a) Daily usage frequency and (b) card type distribution within clusters (k = 2)

Figure 5.2(b) illustrates the distribution of users into clusters based on their card type. Cluster 2 has a slightly higher proportion of standard and discount card types than cluster 1. Moreover, cluster 2 has a significantly higher proportion of 65+ citizens and free card types. Whereas cluster 1 has 60% and 80% of subscription and personnel card type proportions respectively.

5.1.2 K = 4 Clusters

As the number of clusters increases to 4, new clusters with different patterns start to form. Clusters 1 and 2 are similar to the previous run, except for a drop in cluster 2's average hourly transaction rate at 15:00, as well as a drop in the cluster's share from 95% to 63%. Cluster 3 represents late commuters with an average hourly transaction

of 1.2 at 23:00. However, it only corresponds to roughly 3% of the daily boarding profiles. Cluster 4 represents typical commute travel with two sharp AM and PM peaks around 7:00 and 15:00 respectively. This pattern includes 29% of the daily boarding profiles.

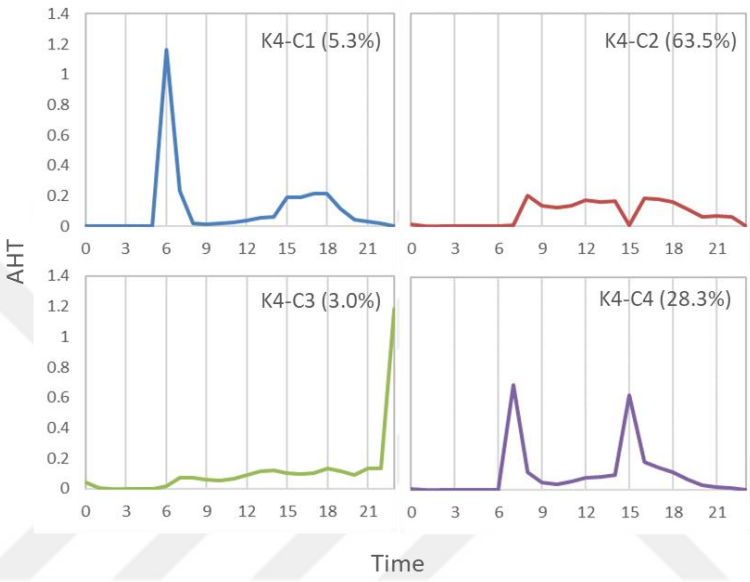


Figure 5.3: Daily Boarding Profiles for K = 4 Clusters

With regard to daily usage frequency, clusters 1 and 2 show a similar pattern to k = 2 run as shown in Figure 5.4(a). As for cluster 3, despite having transactions concentrated at 23:00, almost 35% of the boarding profiles contained two transactions per day. However, this group of users has one transaction in common at 23:00, and the other transaction is distributed throughout the rest of the day. Cluster 4 on the other hand has a 46% twice-per-day use, which is consistent with the daily boarding profile of the two peaks.

Figure 5.4: (a) Daily usage frequency and (b) card type distribution within clusters (k = 4)

The card type proportion of every cluster is shown in Figure 5.4(b). Standard card type users seem to form 30%-40% of every cluster. In addition, cluster 1 has a higher proportion of personnel card type at approximately 14%. Moreover, the discount card type represents more than half of cluster 4. This is reasonable since discount cards are given to teachers and students, a group of users who are more likely to travel during typical AM-PM peaks as the daily boarding profile of cluster 4 shows.

5.1.3 K = 6 Clusters

When observing the daily boarding profiles of the six clusters, it is clear that clusters 1 and 3 are consistent with the previous K = 4 run as shown in Figure 5.5. However,

cluster 2 shows a triple-peak pattern at around 7:00, 13:00, and 17:00. This group of users tend to either have a trip for lunch during the typical break time, or simply perform a separate activity before commuting back home at 15:00. Moreover, the peaks of other clusters such as 1, 5, and 6 corresponds to a drop in cluster 2 at 6:00, 12:00 and 15:00. Therefore, it appears that the K-means algorithm did not label daily boarding profiles with transactions at these peak times as cluster 2, since it is closer to other clusters in term of Euclidean distance. In addition, cluster 4 seems to have split into 3 different patterns when K was increased to 6 as shown in clusters 4, 5, and 6. Cluster 4 has a slight peak at 7:00, with a consistent low average hourly transaction rate up till 21:00 when a steep peak occurs. Cluster 5 represents midday travel behavior with the trips being concentrated around noon. Cluster 6 on the other hand shows a typical AM-PM peak pattern at 7:00 and 15:00 respectively. Users in this cluster, as opposed to cluster 1, tend to travel during the PM peak more often than AM. Nevertheless, cluster 2 still represents the majority of daily boarding vectors at around 63%.

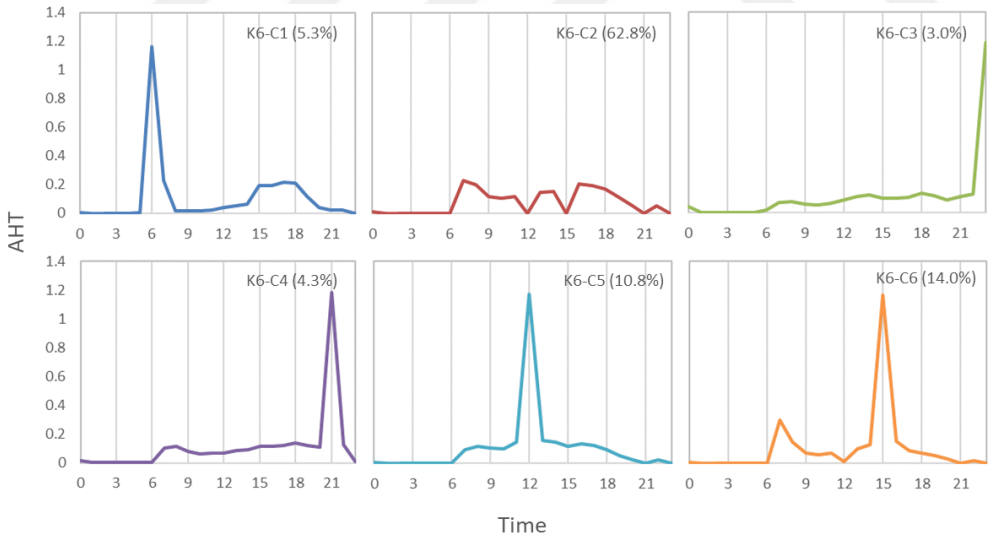


Figure 5.5: Daily Boarding Profiles for K = 6 Clusters

The daily usage frequency for clusters 1, 2, and 3 are similar to K = 4 run which coincides with the matching daily boarding profiles. Nevertheless, cluster 2 has relatively high once and twice-per-day usage, despite having a boarding profile with

3 peaks. As mentioned earlier, cluster 4 was split into 3 clusters (4, 5, and 6), and thus some slight changes were noticeable. For instance, cluster 4's twice-per-day usage dropped from over 47% to 33%. Clusters 5 and 6 have similar daily usage patterns with twice-a-day usage dominating at 42% and 45% respectively.

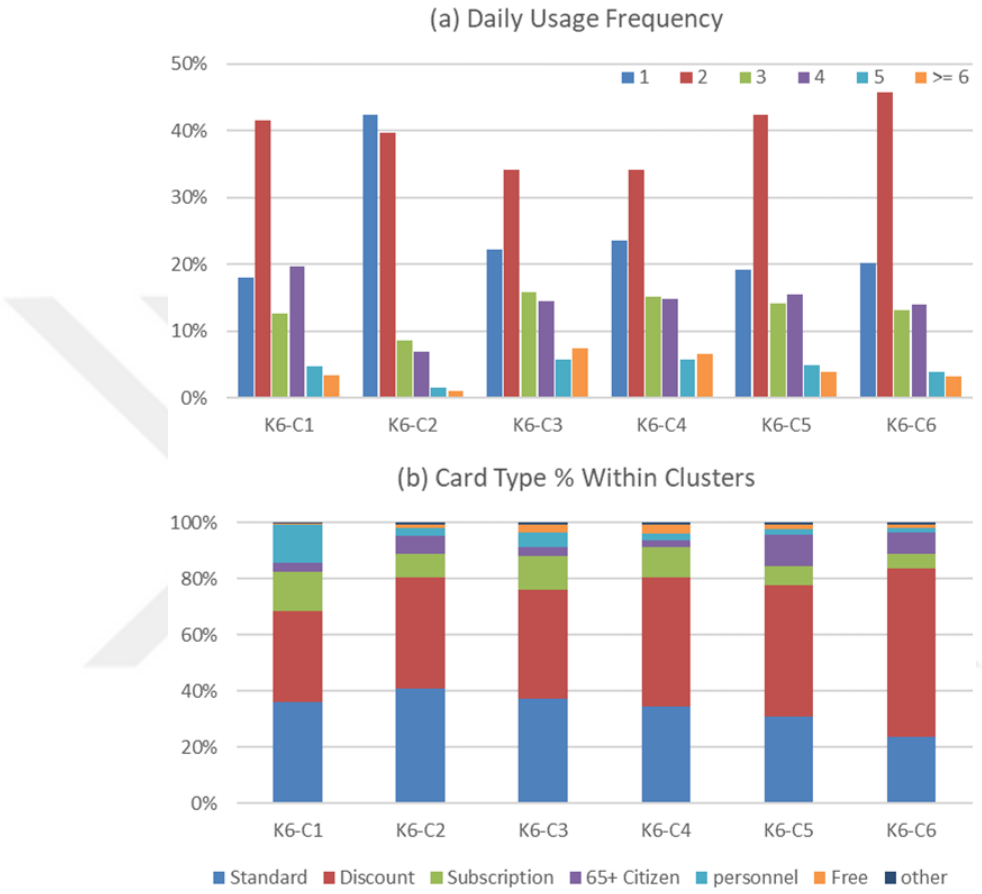


Figure 5.6: (a) Daily usage frequency and (b) card type distribution within clusters (k = 6)

As the number of clusters increases, card type proportions within clusters changes. Discount card type forms 35%-45% of all clusters except for cluster 6 where it accounts for around 60% of the daily boarding vectors. Therefore, a significant portion of discount card type users are more likely to use PT while commuting back home, and not the opposite since cluster 1 has the least proportion of discount card type. Moreover, the 65+ citizen card type had its highest proportion within cluster 5 where most of the transactions are performed between 9:00-15:00. This pattern

matches the characteristic of a 65+ card holder where no commute-like pattern is present.

5.1.4 K = 8 Clusters

When the number of clusters exceeds 6, the algorithm stops producing unique patterns, but rather daily boarding profiles that are similar to the previous runs. As shown in Figure 5.7, clusters 1 through 6 are identical to the clusters obtained in K = 6 run in terms of patterns and shares. This excludes cluster 2 where a 9% decrease in share is detected as well as some slight changes to the daily boarding profile. For instance, all average hourly transactions after 19:00 seem to have dropped to a value close to zero. Thus, cluster 2 for K = 8 run represents a smaller group of users who perform their trips between 6:00 and 19:00 at a very low average hourly transaction rate. Furthermore, clusters 7 and 8 are similar to clusters 4 and 6 except for the back commute timing, which occurs at 19:00 and 22:00 for clusters 7 and 8 respectively. The morning peak occurs at around 7:00 for clusters 6 and 7, with cluster 6 having a higher hourly transaction rate at approximately 0.3 compared to 0.2 for cluster 7. Clusters 4 and 8 on the other hand have a subtle AM peak with a steady average hourly transaction rate throughout the day.

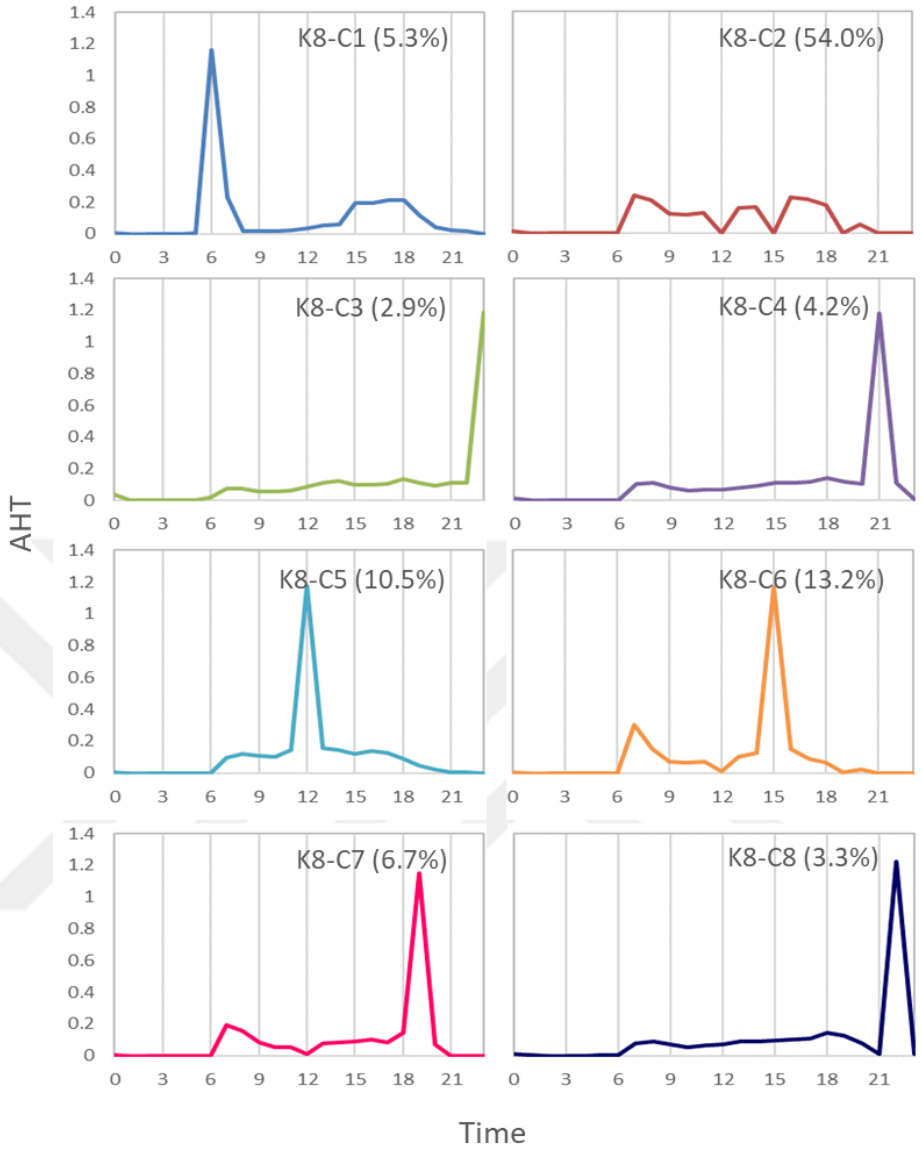


Figure 5.7: Daily Boarding Profiles for K = 8 Clusters

Figure 5.8(a) illustrates the daily usage frequency for each cluster. Similar to daily boarding profiles, daily usage frequency for clusters 1 to 6 are identical to the previous run. Clusters 7 and 8 however have their unique composition of daily usage frequency, with the twice-a-day being the highest portion at 41% and 38% for clusters 7 and 8 respectively. In addition, once-a-day usage is relatively high compared to clusters with similar composition at around 28%.

With a number of clusters equal to 8, the card type proportion within clusters 1 to 6 is consistent with the previous K = 6 run as shown in Figure 5.8(b). As for clusters 7 and 8, the composition is similar to cluster 3 despite having different daily boarding profiles. This suggests that card type proportions within clusters are independent of the cluster's daily boarding profile.

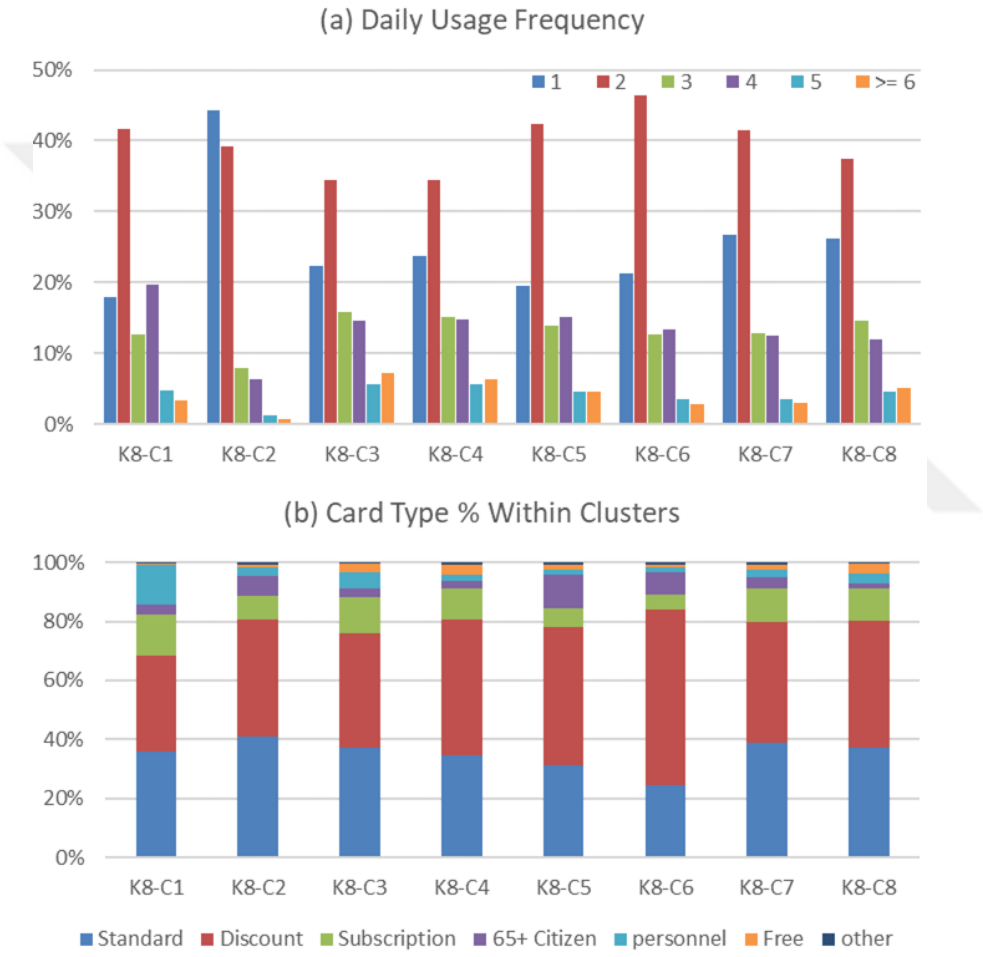


Figure 5.8: (a) Daily usage frequency and (b) card type distribution within clusters (k = 8)

5.1.5 K = 12 Clusters

Figure 5.9 shows the daily boarding profiles of 12 clusters from the algorithm's final run. Clusters 1 to 8 have had slight changes in terms of both pattern and share. For instance, a considerable drop can be observed in cluster 2's share from 54% to 29%. Most of this was shifted to other clusters causing cluster 2 to consist of a mixture of daily boarding profiles that could not be assigned to other clusters. In addition, cluster 6 experienced a modest drop in share, as well as a decrease in the average hourly transaction rate at 10:00. This was caused by the emergence of cluster 9 which has a peak around that time. Despite the repeated patterns, one new pattern emerged at the K = 12 run represented by clusters 10 and 11. This pattern includes two typical AM-PM peaks as well as a fluctuating average hourly transaction rate in between. This indicates a group of users that tend to use PT multiple times a day for both commuting and other activities.

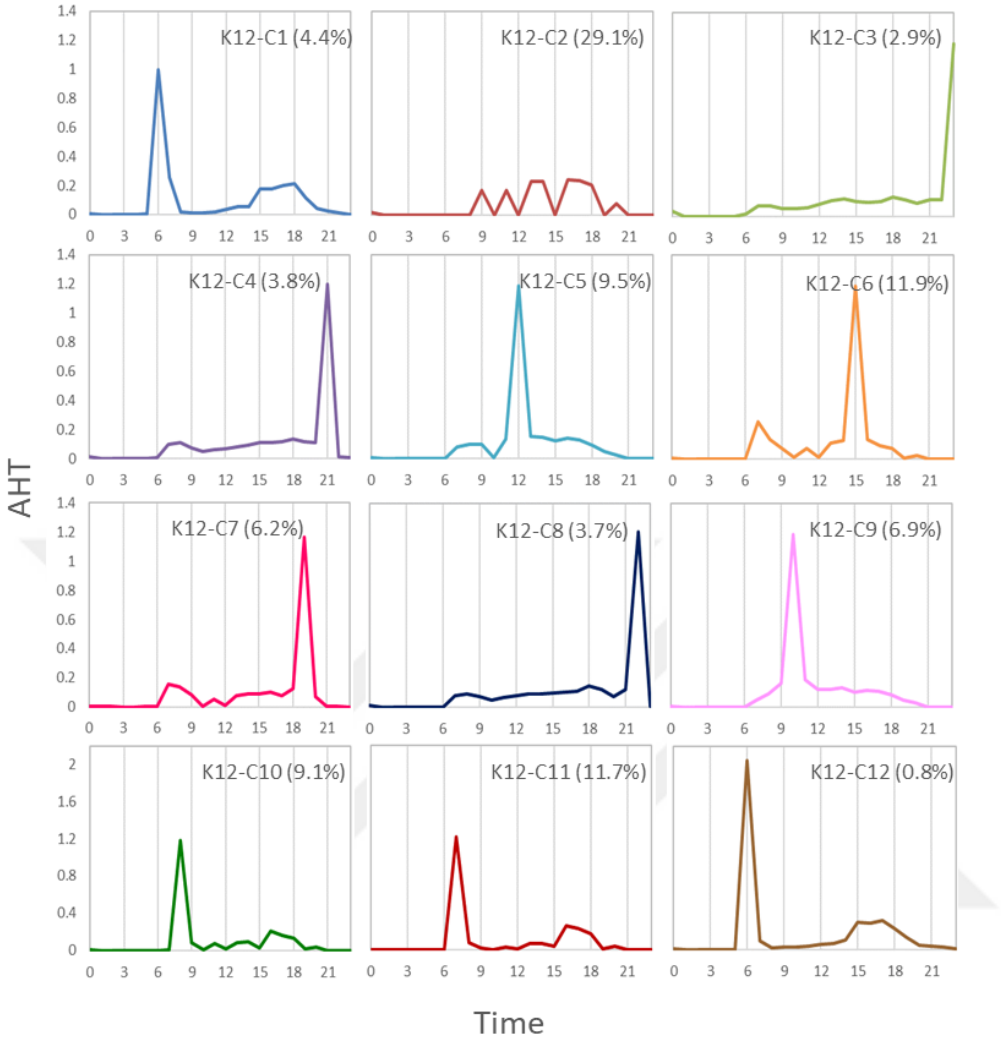


Figure 5.9: Daily Boarding Profiles for K = 12 Clusters

When looking at the daily usage frequency graph (Figure 5.10-b), two distinct features can be observed. First, the once-per-day usage for cluster 2 increased significantly up to 56%. This coincides with the fact related to cluster 2 being a mixture of random daily boarding vectors having transactions at various times of the day. In other words, since most of the daily boarding vectors contained one transaction only, the resulting average hourly transaction is fluctuating at a low rate from 0 to 0.2 for the daily boarding profile of cluster 2. The second distinguishable feature is that of cluster 10 where the four times a day usage represents over 40% of

the cluster. This group of users as shown in the daily boarding profile of cluster 10 perform two transactions for their commute trip as well as two more for other purposes between 10:00 and 15:00. Apart from that, the rest of the clusters have usual daily usage frequency with twice-per-day being the highest, followed by once-per-day, and the rest of the frequencies at a lower rate between 2% and 20%.

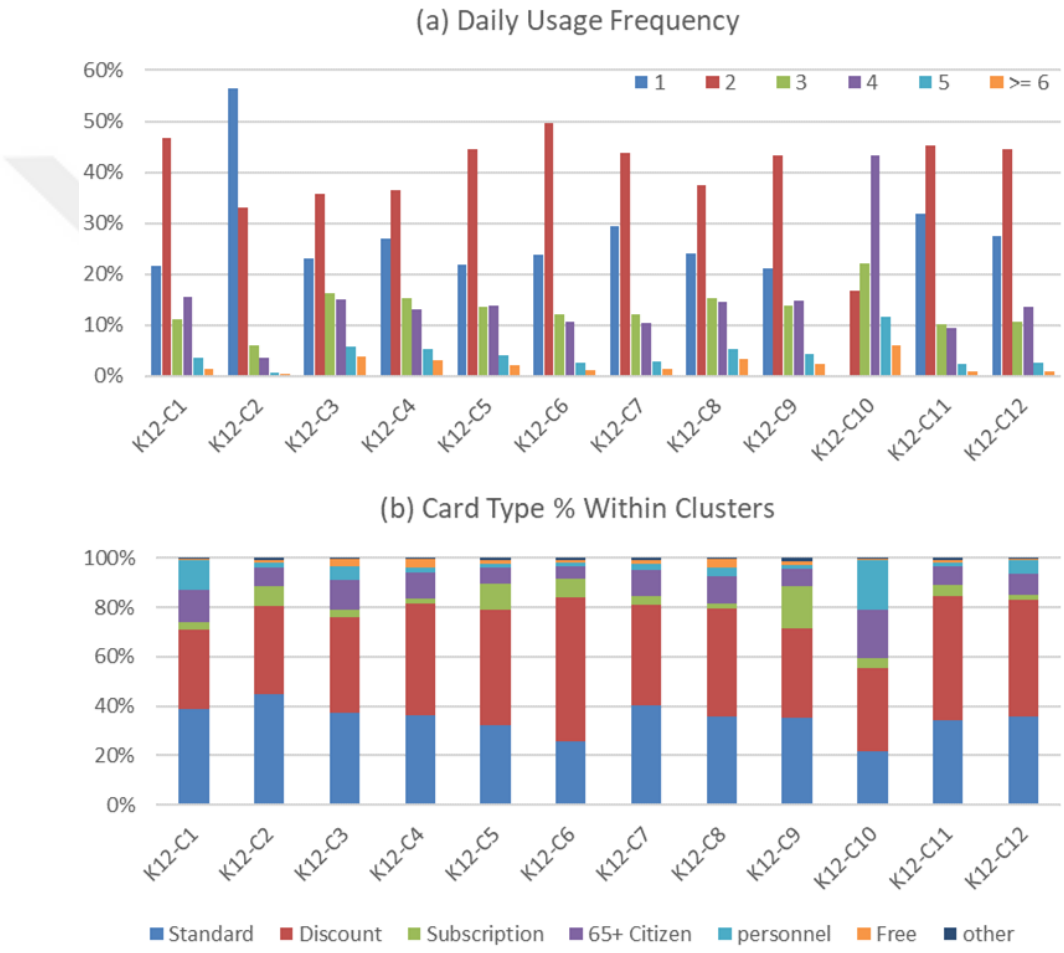


Figure 5.10: (a) Daily usage frequency and (b) card type distribution within clusters (k = 12)

Despite having some changes in the card type proportions within clusters compared to the K = 8 run, these changes fall in the range of 2%-5% which is insignificant. Furthermore, both clusters 11 and 12 are similar in card type composition to cluster

4, despite having a contrasting daily boarding profile. Once again, this implies that card type proportions within clusters are extraneous to daily boarding profiles. Moreover, cluster 10 consists of an unusual card type composition where 65+ citizen and personnel card types have their highest proportion whereas standard and discount card types have their lowest.

5.2 Clusters Evolution

The evolution of clusters as the number of K increases is based upon daily boarding profile and share.

Figure 5.11 demonstrates the change of daily boarding profile patterns for clusters 1 to 6. Clusters 1 and 3 are the most consistent with almost no change throughout all the runs. Cluster 4 started with a two-peak pattern then change at K = 6 run and remained consistent through K = 6 – 12. Moreover, clusters 5 and 6 remained unchanged until K = 12 run where a slight drop in average daily transactions is noticeable at 10:00. Cluster 2 on the other hand kept on changing throughout all the runs from K = 2 to K = 12. This is due to its significant share starting at 95% at K = 2 and gradually decreasing, allowing other clusters to be partitioned from it.

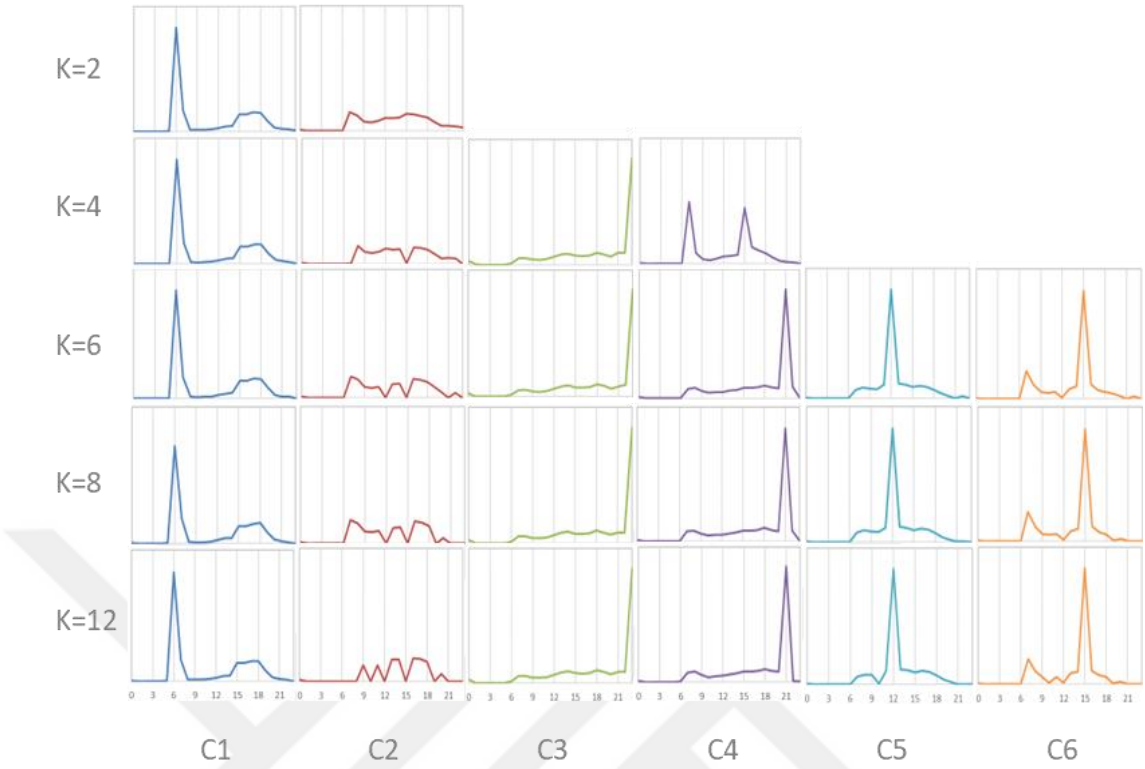


Figure 5.11: Evolution of Clusters 1 to 6 DBP (K = 2 to K = 12)

The second parameter governing clusters' evolution is the share. Table 5.1 displays the shares of clusters 1 – 6 throughout the algorithm's 5 runs. Similar to the daily boarding profiles, clusters 1 and 3 had a consistent share through all runs, whereas cluster 4 experienced a drop at K = 6 run. In addition, clusters 5 and 6's share decreased slightly from K = 6 to K = 12 runs. Moreover, cluster 2's share dropped significantly at K = 4 from 95% down to 64%, where it stayed almost constant till K = 6, and then dropped again at K = 8 and K = 12 to 54% and 29% respectively.

Table 5.1: Clusters 1-6 share through k=2 to k=12

	K2	**K4**	**K6**	**K8**	**K12**
C1	5.30%	5.30%	5.30%	5.30%	4.40%
C2	94.70%	63.50%	62.80%	53.90%	29.10%
C3		3.00%	2.90%	2.90%	2.90%
C4		28.30%	4.30%	4.20%	3.80%
C5			10.80%	10.50%	9.50%
C6			14.00%	13.20%	11.90%

Despite changing the number of clusters in each run, some clusters showed a significant level of consistency in terms of share and daily boarding profile. This in fact serves as a validation to the corresponding daily travel behavior, represented by the cluster's daily boarding profile. Furthermore, this reflects the robustness of the performed methodology as opposed to the classical K-means which is sensitive to outliers and centroid initialization as stated in chapter 2. In addition, the consistency of the output reflects data compatibility. In other words, it can be deduced that K-means clustering is suitable for SCD analysis.

5.3 Clusters Similarity

Despite having 12 different clusters at the last run, some of these clusters are similar in daily boarding profile patterns. As shown in Figure 5.12, group (a) clusters share a similar pattern with a concentration of usage at 10:00 and 12:00. Group (b) has a

slight peak at around 8:00, and a sharp PM peak at 15:00 and 19:00 respectively. Group (c) displays a slight usage between 6:00 and 9:00, with a late PM peak at 21:00 and 22:00 respectively. Moreover, group (d) has a pattern of a sharp AM peak, low usage between 9:00 and 15:00, and a PM peak around 15:00-18:00. Similarly, group (e) have an AM peak at 6:00 and a wider PM peak between 15:00 and 19:00. However, cluster 12 has a considerably high AHT at 6:00 with a value of 2. This indicates a group of users who are likely to perform trip-chaining while commuting to work. Nevertheless, this cluster's share is 0.8% which is insignificant. Finally, group (f) contains clusters that do not share any similarity with other clusters.

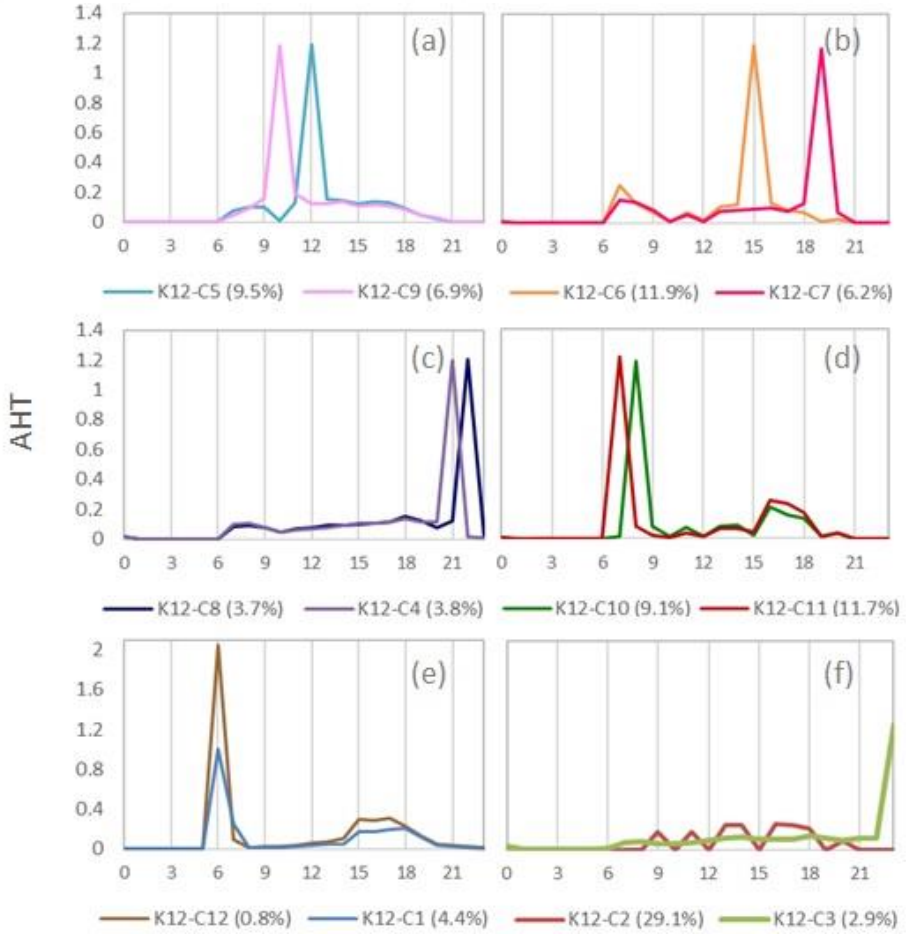


Figure 5.12: Clusters similarity assessment (K = 12)

5.4 Clusters Characteristics

All in all, points that can be deduced from the overall analysis of clusters' daily boarding profiles, daily usage and card type proportion are as follows:

1- Some features related to certain card type holders might change as the number of clusters increases. For example, at the K = 4 run, a significant portion of the discount card type holders followed a twice-per-day usage pattern within a typical AM-PM peak. However, as the number of clusters increased to 6, the pattern shifts toward one commuting trip around 7:00 and the possible use of other means of transport when commuting back home. Therefore, to draw a conclusion, the proportion following a certain pattern must be taken into account, where a larger proportion implies a stronger tendency towards a specific pattern.

2- Clusters' daily boarding profiles and shares are mostly consistent throughout the runs (k = 2 to k = 12). This excludes cluster 2 as its share keeps dropping at every run where it partitions further, and other clusters start forming. The same applies to cluster 4 where it starts partitioning at K = 6 run to form clusters 5 and 6.

3- The choice of K number of clusters depends on the level of depth which the analysis is supposed to achieve. The algorithm would keep on partitioning the sample into smaller detailed groups as the number of clusters increases. However, the run at which similar daily boarding profiles start showing up is a good indicator for a sufficient depth of analysis.

4- The average hourly transaction is significantly low during off-peak hours. This also applies to cluster 2 which held the highest share from k=2 to k=12. The reason behind this is that the majority of daily boarding vectors were filled with zeros since most of the daily usage would be concentrated at 1 or 2 time slots out of 24.

CHAPTER 6

CONCLUSION AND FUTURE RECOMMENDATIONS

6.1 Conclusion

As social and economic diversity grow in cities, so does the heterogeneity of PT travel behavior. In order to accommodate for PT usage changes, a thorough understanding of travel behavior is necessary. Thus, this study examined descriptive statistics of PT behavior using SCD generated from Konya's PT automated fare collection system. In addition, a data mining technique that includes an unsupervised learning clustering algorithm has been utilized to segment users based on their daily boarding activity within the PT network of Konya district. The descriptive statistics served as an introduction to understanding data composition, main features, and parameters through data visualization tools. The K-means clustering algorithm on the other hand helped understand the characteristics of PT users' daily boarding patterns.

The clustering algorithm has revealed the following daily travel patterns. A group of users who commute at different time intervals, which reflects the heterogeneity of working hours schemes. Late commuters who often perform at least one trip late at night. In addition, a group of users characterized with midday usage and a possible trip using other modes of transport.

Due to the heterogeneity of PT travel patterns, there exist a considerable number of distinguishable daily travel patterns. As the number of pre-determined clusters increases, so does the resulting daily travel pattern. However, some patterns are associated with an insignificant proportion of the population. Therefore, the choice of K number of clusters is dependent on the desired level of depth that the analysis is meant to achieve. Nevertheless, a reasonable indicator for a sufficient depth of

analysis is the algorithm's run at which similar daily patterns to previous runs start resulting.

The aim of this study had two parts: First, the development of a simple yet efficient machine learning algorithm, capable of analyzing PT SCD. Second, using the results obtained by the machine learning algorithm to aid PT authorities in developing a better understanding of travel behavior and execute relevant strategies accordingly.

The developed methodology used K-means clustering algorithm which is known for its superiority in computational complexity among other clustering algorithms. However, the algorithm has some drawbacks such as sensitivity to initialization and predefining number of clusters. To overcome these disadvantages, the study implemented K-means++ for centroid initialization and examined the evolution of patterns as the number of clusters increases, rather than choosing a fixed K value. The resulting daily boarding profiles were mostly consistent throughout different number of cluster runs. This reflects the compatibility of the proposed methodology with SCD behavioral analysis as discussed in section 5.2. Therefore, the same approach can be applied to any SCD by other transit systems, as long as it includes the card identification number, time, and date of the transaction.

Secondly, when examining the clusters' daily boarding profiles, it is noticeable that for most profiles there is a one dominant peak, either AM, midday, PM, or late at night. This indicates that a significant portion of users use the PT once a day. This is also backed up by the high percentage of once-a-day usage frequency as discussed in 3.3.3. In order for Konya's municipality to deal with this issue the following approaches are recommended:

1- The municipality can introduce a monthly subscription of once-per-day usage to facilitate PT usage for this segment of users. This would increase user satisfaction and possibly improve ridership as it adds up to the overall flexibility of the PT system of Konya.

2- Since a once-per-day usage suggests that at least 1 trip has been made by another mean of transport, service adjustments in such a way that it would accommodate for that other trip might encourage this segment of users to use PT and improve ridership. However, it is not a straightforward task to figure out what service adjustments are required. Therefore, the municipality should launch an investigation regarding the once-per-day usage. The aim would be to investigate this type of behavior including, the reason behind it, the required service adjustments, and its implications on PT usage.

Furthermore, it is recommended for the municipality of Konya to use the same methodology of this study in case it wishes to examine the travel behavior of a certain group. The methodology can be easily adjusted to run the analysis for a targeted segment of users. Such segments may be users of a certain card type, a certain bus route, or a particular geographical area of Konya district.

6.2 Contributions

The contribution of this study includes a data mining approach that is capable of extracting the daily travel patterns of PT users from SCD. It also includes travel pattern evolution with respect to the number of clusters. Daily travel patterns are essential for transport authorities since they facilitate travel demand modeling and service customization. Furthermore, travel pattern evolution is important for PT researchers seeking to comprehend the effect of clustering algorithm manipulation on the identified travel behavior. The study is also of interest to research bodies with limited computing apparatus. This is because the algorithm has a relatively low computation complexity, as well as the ability to process big data within a short run time.

6.3 Future Recommendations

There are several dimensions to which future studies can be steered towards. For instance, the GPS location of transactions can be included to account for the spatial variability of daily travel behavior. Also, intrapersonal variability can be examined over a longer study period by applying a similar algorithm to daily boarding vectors corresponding to an individual. Furthermore, PT transfer activities can be investigated using transaction type and GPS location where transfer spatiotemporal patterns can be identified.

Whereas the proposed data mining approach had extracted daily travel behaviors, another traditional data collection method such as household surveys can be utilized to validate the clustering algorithm's identified travel patterns. In addition, other data mining techniques can be compared to the proposed algorithm in order to identify the strength and weaknesses of data mining tools in travel behavior analysis.

Another future recommendation is adding an additional step to the clustering algorithm where it would be able to detect daily boarding profile repetition. This would help to automatically stop the execution of the algorithm at a k number of clusters where the main travel patterns had already been detected. For example, a mechanism of pattern comparison with the previously determined patterns can be added by the end of each iteration, such that the algorithm would stop if a certain level of similarity is achieved.

REFERENCES

Abu Abbas, O. (2008). Comparisons Between Data Clustering Algorithms. *The International Arab Journal of Information Technology*, *5*(3).

Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, *12*(PART 1). https://doi.org/10.3182/20060517-3-fr-2903.00211

Alsger, A. A., Mesbah, M., Ferreira, L., & Safi, H. (2015). Use of smart card fare data to estimate public transport origin-destination matrix. *Transportation Research Record*, *2535*, 88–96. https://doi.org/10.3141/2535-10

Axhausen, K., & Zürich, E. T. H. (2007). *Concepts of Travel Behavior Research*.

Bai, L., Kane, G., & Lyons, P. (2008). Open Architecture for Contactless Smartcard-based Portable Electronic Payment Systems. *IEEE Conference on Automation Science and Engineering*.

Bawane, V. S. (2017). *Study of Various Clustering Algorithm to Speed up Processing in Big Data Study of Various Clustering Algorithm to Speed up Processing in Big Data*. *January*.

Briand, A. S., Côme, E., Trépanier, M., & Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies*, *79*, 274–289. https://doi.org/10.1016/j.trc.2017.03.021

Cats, O., & Ferranti, F. (2022). Unravelling individual mobility temporal patterns using longitudinal smart card data. *Research in Transportation Business and Management*, 100816. https://doi.org/10.1016/j.rtbm.2022.100816

Chapleau, R., & Chu, K. K. A. (2007). Modeling Transit Travel Patterns from Location-Stamped Smart Card Data Using a Disaggregate Approach. *11th World Conference on Transport Research*.

Chau, P. Y. K., & Poon, S. (2003). Octopus: an e-cash payment system success

story. *Communications of the ACM*, *46*(9), 129–133.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C*, *68*, 285–299. https://doi.org/10.1016/j.trc.2016.04.005

Cui, Y., He, Q., & Khani, A. (2018). Travel Behavior Classification: An Approach with Social Network and Deep Learning. *Transportation Research Record*, *2672*(47), 68–80. https://doi.org/10.1177/0361198118772723

Darshana Abeyrathna, K., Rasca, S., Markvica, K., & Granmo, O. C. (2021). Public Transport Passenger Count Forecasting in Pandemic Scenarios Using Regression Tsetlin Machine. Case Study of Agder, Norway. *Smart Innovation, Systems and Technologies*, *231*, 27–37. https://doi.org/10.1007/978-981-16-2324-0_4/COVER

de la Torre, R., Corlu, C. G., Faulin, J., Onggo, B. S., & Juan, A. A. (2021). Simulation, optimization, and machine learning in sustainable transportation systems: Models and applications. *Sustainability (Switzerland)*, *13*(3), 1–21. https://doi.org/10.3390/su13031551

Deepti, S., Lokesh, S., Sisodia, S., & Khushboo Saxena. (2015). Clustering Techniques : A Brief Survey of Different Clustering Algorithms. *International Journal of Latest Trends in Engineering and Technology (IJLTET) Clustering*, *1*(3), 82–87. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1037.1127&rep=rep1&type=pdf

Deschaintres, E., Morency, C., & Trépanier, M. (2019). Analyzing Transit User Behavior with 51 Weeks of Smart Card Data. *Transportation Research Record*, *2673*(6), 33–45. https://doi.org/10.1177/0361198119834917

Ding, S., Zhang, L., & Zhang, Y. (2010). Research on spectral clustering algorithms and prospects. *ICCET 2010 - 2010 International Conference on Computer Engineering and Technology, Proceedings*, *6*, 149–153.

https://doi.org/10.1109/ICCET.2010.5486345

Dunhan, M. H. (2006). *Data Mining: Introductory and Advanced Topics*. Pearson Education India.

Egu, O., & Bonnel, P. (2020). Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon. In *Travel Behaviour and Society* (Vol. 19, pp. 112–123). https://doi.org/10.1016/j.tbs.2019.12.003

Espinoza, C., Munizaga, M., Bustos, B., & Trépanier, M. (2018). Assessing the public transport travel behavior consistency from smart card data. *Transportation Research Procedia*, *32*, 44–53. https://doi.org/10.1016/j.trpro.2018.10.008

Faroqi, H., & Mesbah, M. (2021). Inferring trip purpose by clustering sequences of smart card records. *Transportation Research Part C: Emerging Technologies*, *127*(November 2020), 103131. https://doi.org/10.1016/j.trc.2021.103131

Goulet Langlois, G., Koutsopoulos, H. N., & Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, *64*, 1–16. https://doi.org/10.1016/j.trc.2015.12.012

Halvorsen, A. (2015). *Improving Transit Demand Management with Smart Card Data: General Framework and Applications*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. In *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (Vol. 2). https://doi.org/10.1111/j.1467-985x.2010.00646_6.x

Hickman, M. (2002). Robust passenger itinerary planning using transit AVL data. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2002-Janua*(March), 840–845. https://doi.org/10.1109/ITSC.2002.1041329

Hofmann, M., Wilson, S. P., & White, P. (2009). Automated identification of

linked trips at trip level using electronic fare collection data. *Transportation Research Board 88th Annual Meeting*, *4*(August 2008), 1–18.

Hussain, E., Bhaskar, A., & Chung, E. (2021). Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. *Transportation Research Part C: Emerging Technologies*, *125*(February), 103044. https://doi.org/10.1016/j.trc.2021.103044

Kim, M., Kim, S., & Sohn, H. (2018). *Relationship between Spatio-Temporal Travel Patterns Derived from Smart-Card Data and Local Environmental Characteristics of Seoul , Korea*. https://doi.org/10.3390/su10030787

Li, T., Sun, D., Jing, P., & Yang, K. (2018). Smart card data mining of public transport destination: A literature review. *Information (Switzerland)*, *9*(1), 28–30. https://doi.org/10.3390/info9010018

Liu, S., Yamamoto, T., Yao, E., & Nakamura, T. (2022). Exploring Travel Pattern Variability of Public Transport Users Through Smart Card Data: Role of Gender and Age. *IEEE Transactions on Intelligent Transportation Systems*, *23*(5), 4247–4256. https://doi.org/10.1109/TITS.2020.3043021

Ma, X., Liu, C., Wen, H., Wang, Y., & Wu, Y. J. (2017). Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, *58*, 135–145. https://doi.org/10.1016/j.jtrangeo.2016.12.001

Ma, X., Wu, Y. J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, *36*, 1–12. https://doi.org/10.1016/j.trc.2013.07.010

Martín, L., Baena, L., Garach, L., López, G., & de Oña, J. (2014). Using Data Mining Techniques to Road Safety Improvement in Spanish Roads. *Procedia - Social and Behavioral Sciences*, *160*(Cit), 607–614. https://doi.org/10.1016/j.sbspro.2014.12.174

Mohamad, I. Bin, & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and*

Technology, *6*(17), 3299–3303. https://doi.org/10.19026/rjaset.6.3638

Morency, C., Trepanier, M., & Agard, B. (2006). Analysing the variability of transit users behaviour with smart card data. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 44–49. https://doi.org/10.1109/itsc.2006.1706716

Munizaga, M., Palma, C., & Mora, P. (2010). Public transport OD matrix estimation from smart card payment system data. *12th World Conference on Transport Research*, 1–16.

Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557–568. https://doi.org/10.1016/j.trc.2010.12.003

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. In *PLoS ONE* (Vol. 14, Issue 1). https://doi.org/10.1371/journal.pone.0210236

Sehgal, G., & Grag, K. (2014). Comparison of Various Clustering Algorithms. *International Journal of Computer Science and Information Technologies*, *5*(3), 3074–3076.

Shireman, E., Steinley, D., & Brusco, M. J. (2017). Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior Research Methods*, *49*(1), 282–293. https://doi.org/10.3758/s13428-015-0697-6

Trépanier, M., Habib, K. M. N., & Morency, C. (2012). Are transit users loyal? revelations from a hazard model based on smart card data. *Canadian Journal of Civil Engineering*, *39*(6), 610–618. https://doi.org/10.1139/L2012-048

Trépanier, M., Morency, C., & Agard, B. (2009). Calculation of Transit Performance Measures Using Smartcard Data. *Journal of Public Transportation*, *12*(1), 79–96. https://doi.org/10.5038/2375-0901.12.1.5

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *11*(1), 1–14. https://doi.org/10.1080/15472450601122256

Tu, W., Cao, R., Yue, Y., Zhou, B., Li, Q., & Li, Q. (2018). Spatial variations in urban public ridership derived from GPS trajectories and smart card data. *Journal of Transport Geography*, *69*(3688), 45–57. https://doi.org/10.1016/j.jtrangeo.2018.04.013

Uniman, D., Attanucci, J., Mishalani, R., & Wilson, N. (2010). Service reliability measurement using automated fare card data. *Transportation Research Record*, *2143*, 92–99. https://doi.org/10.3141/2143-12

Utsunomiya, M., Attanucci, J., & Wilson, N. (2006). Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation Research Record*, *1971*, 119–126. https://doi.org/10.3141/1971-16

Uyan, M. (2014). MSW landfill site selection by combining AHP with GIS for Konya, Turkey. *Environmental Earth Sciences*, *71*(4), 1629–1639. https://doi.org/10.1007/s12665-013-2567-9

Uyan, M., Sert, E., Osmanli, N., & Eruc, R. (2017). Determination of Transportation Networks Base on the Optimal Public Transportation Policy Using Spatial and Network Analysis Methods: a Case of the Konya, Turkey. *International Journal of Engineering and Geosciences*, *2*(1), 27–34. https://doi.org/10.26833/ijeg.286034

Van Der Maaten, L. J. P., Postma, E. O., & Van Den Herik, H. J. (2009). Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research*, *10*, 1–41. https://doi.org/10.1080/13506280444000102

Viallard, A., Trépanier, M., & Morency, C. (2019). Assessing the Evolution of Transit User Behavior from Smart Card Data. *Transportation Research Record*, *2673*(4), 184–194. https://doi.org/10.1177/0361198119834561

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141

Yuan, T., Da Rocha Neto, W., & Esteve Rothenberg, C. (2021). Machine learning for next-generation intelligent transportation system: A survey. *Transactions on Emerging Telecommunications Technologies*.

Yun, H., Lee, E. H., Kim, D. K., & Cho, S. H. (2021). Development of estimating methodology for transit accessibility using smart card data. *Transportation Research Record*, *2675*(11), 159–171. https://doi.org/10.1177/03611981211027562

Zhao, J., Tian, C., Zhang, F., Xu, C., & Feng, S. (2014). Understanding temporal and spatial travel patterns of individual passengers by mining smart card data. *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, 2991–2997. https://doi.org/10.1109/ITSC.2014.6958170