Titre: Title:	Data-Driven and Multidimensional Clustering of Bike Sharing Stations and Travel Patterns
Auteur: Author:	Fatemeh Zaferani
Date:	2024
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Zaferani, F. (2024). Data-Driven and Multidimensional Clustering of Bike Sharing Stations and Travel Patterns [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/59281/

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Data-Driven and Multidimensional Clustering of Bike Sharing Stations and Travel Patterns

FATEMEH ZAFERANI

Département des génies civil, géologique et des mines

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie Civil

Août 2024

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Data-Driven and Multidimensional Clustering of Bike Sharing Stations and Travel Patterns

présenté par **Fatemeh ZAFERANI**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Martin TRÉPANIER, président

Francesco CIARI, membre et directeur de recherche

Nicolas SAUNIER, membre et codirecteur de recherche

Luis F. MIRANDA-MORENO, membre

RÉSUMÉ

Cette thèse présente une analyse du système de vélo-partage Bixi à Montréal, Canada, avec un accent particulier sur la compréhension du comportement des utilisateurs et des caractéristiques des stations. Le système Bixi est devenu un élément essentiel du réseau de transport de la ville, offrant une solution de transport flexible et écologique. Cette recherche adopte une approche basée sur les données en utilisant des données ouvertes pour obtenir des informations détaillées sur le fonctionnement du système.

La recherche commence par une analyse descriptive des données, permettant d'identifier les tendances clés et les caractéristiques de la demande de Bixi. Cette étape préliminaire fournit une vue d'ensemble générale des modèles d'utilisation du système, servant de base pour des investigations plus détaillées. En employant des techniques telles que la régression, la visualisation, le regroupement et l'analyse spatiale, nous décomposons les données sous différents angles, révélant des détails complexes sur le fonctionnement du système.

L'étude examine également les facteurs influençant la demande de voyages à vélo, en tenant compte de variables telles que les conditions météorologiques, le moment de la journée, l'emplacement et l'utilisation des terres. Nos résultats montrent que la demande de Bixi est façonnée par une combinaison de ces facteurs, certains ayant un impact plus significatif que d'autres. Nous avons recours à des analyses de regroupement, telles que DBSCAN, pour explorer la répartition spatiale et la densité des stations, fournissant des informations sur la diversité et la complexité des stations Bixi.

Un aspect essentiel de notre recherche est l'examen de la durée de vie des stations de 2015 à 2020. Cette analyse révèle que toutes les stations ne conservent pas un emplacement permanent en raison de facteurs comme la faible demande ou les coûts opérationnels élevés. Cependant, les stations qui ont fonctionné sans interruption pendant cinq ans ont probable-

ment bénéficié d'une forte demande ou d'un positionnement stratégique.

L'analyse de régression a mis en évidence plusieurs facteurs clés influençant la demande de voyages à vélo, notamment l'emplacement des stations par rapport aux points d'intérêt (POIs) et la densité des stations voisines. L'étude utilise également les méthodes de regroupement de DBSCAN, de propagation d'affinité et de K-means pour classifier les stations en fonction de leurs caractéristiques d'utilisation et de voyage. Ces méthodes offrent une vision détaillée des caractéristiques des stations, de leur utilisation et de leur distribution spatiale.

Les résultats de l'étude soulignent le potentiel des données ouvertes pour une analyse approfondie et la prise de décision. Ils mettent en évidence l'importance des stratégies basées sur les données pour optimiser les systèmes de vélo-partage, améliorer l'allocation des ressources, la planification de la maintenance et les stratégies de marketing. En fin de compte, cette analyse complète contribue à une meilleure compréhension du système Bixi, ce qui peut conduire à une amélioration du service, une augmentation de l'utilisation et une rentabilité accrue. L'étude propose également des recommandations pour des recherches futures, telles que l'incorporation de données supplémentaires, l'exploration des tendances à long terme et l'analyse comparative avec d'autres systèmes de vélo-partage, afin de continuer à optimiser les services de vélo-partage et répondre efficacement aux besoins des populations urbaines.

ABSTRACT

Bixi, a bike-sharing service based in Montreal, Canada, has become an integral part of the city's transportation system. It offers a flexible and eco-friendly mode of transport, allowing users to rent bikes from various stations across the city and return them at their convenience. Particularly popular during the warmer months, Bixi serves a broad range of users, including both residents and tourists.

Bike-sharing systems (BSSs) like Bixi are not just a convenient means of travel; they also promote a healthier lifestyle and contribute to the reduction of carbon emissions in urban areas. However, the optimization of such systems is crucial for their success and sustainability. This involves a deep understanding of user behavior and station characteristics, which can be achieved through comprehensive data analysis.

In our study, we delve into the Bixi system in Montreal, aiming to uncover the factors that influence the daily number of trips at each station. Our approach is data-driven, relying on a variety of methods to extract meaningful insights from the available data.

Our initial step involves a descriptive analysis of the data, which helps us identify the primary trends and characteristics of BSS demand. This process provides a general overview of the system's usage patterns, laying the groundwork for more detailed investigations. To further explore the patterns and trends of Bixi trips, we employ a range of techniques, including descriptive statistics, visualization, clustering, and regression. These methods allow us to dissect the data from multiple angles, revealing intricate details about the system's operation.

One of our key areas of investigation is the impact of various variables on Bixi demand. We examine factors such as weather conditions, time of day, location, and land use, all of which are known to affect the usage of bike-sharing systems. Our findings show that Bixi's demand is influenced by a combination of these factors, with some having a more significant impact

than others. Through cluster analysis, we group stations based on user and travel features, examining their spatial distribution and density. This analysis provides valuable insights into the diversity and complexity of Bixi stations, contributing to a more nuanced understanding of the system.

Furthermore, our study highlights the importance of leveraging open data for optimizing bike-sharing systems. By analyzing the spatial distribution of stations and their usage patterns, we can inform better resource allocation, maintenance scheduling, and marketing strategies. Our findings suggest that tailored approaches to station management and user engagement can enhance service efficiency and user satisfaction. The insights gained from this study also offer a foundation for future research, including the incorporation of additional data sources, exploration of long-term trends, and comparative analysis with other cities. Ultimately, this research contributes to the broader goal of improving the effectiveness and sustainability of bike-sharing systems, ensuring they continue to meet the needs of urban populations effectively.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

IETF	Internet Engineering Task Force
OSI	Open Systems Interconnection
AMI	Adjusted Mutual Information
AP	Affinity Propagation
ARI	Adjusted Rand Index
BSS	Bike Sharing System
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
MAE	Mean Absolute Error
OLS	Ordinary Least Squares
POIs	Points of Interest
RMSE	Root Mean Squared Error
RSE	Residual Standard Error

CHAPTER 1 INTRODUCTION

Bike-sharing is a convenient and eco-friendly way to navigate the city and explore its numerous attractions and activities. Cities that implement bike-sharing programs encourage citizens to engage in greater physical activity, which has significant positive effects on their health. Studies have shown that increased physical activity can lead to improved cardiovascular health, reduced risk of chronic diseases, and enhanced mental well-being [1, 2].

For users who do not want the hassle of maintaining their own bikes and prefer the flexibility of returning them practically anywhere in the city, Bike-Sharing Systems (BSSs) offer a highly convenient solution. However, the operation and maintenance of these systems are extremely challenging due to the growing number of users and the varying demand at different times and on different days [3, 4].

Research has shown that bike-sharing systems can have a positive impact on public health by promoting physical activity and reducing air pollution [5, 6]. Additionally, studies have highlighted the importance of equitable access to bike-sharing services, ensuring that all segments of the population can benefit from these programs [7].

Studying bike-sharing stations is crucial in BSS because the location of each station has profound implications for both operators and users. Each station must meet a variety of requirements, such as being easily accessible, ensuring bike availability when used as an origin, and ensuring the availability of empty docks when used as a destination. The placement of the stations is one of the most critical factors in the success of these systems [8,9]. Municipalities or public-private partnerships are typically responsible for introducing BSSs. The primary concern for public investment in bicycle mobility is to design and implement the system in a way that maximizes its benefits, as public investment in bicycle mobility is always subject to budget constraints [4,10]. These factors are prompting an increasing number of communities

to develop BSSs. As of 2010, BSSs were available in over 125 locations across four continents, totaling around 140,000 shared bicycles worldwide [5, 6].

One notable example of a BSS is Bixi, which is available in Montreal and the surrounding region. The name "Bixi" is derived from the words "bike" and "taxi," reflecting its role as a public self-service bicycle-sharing program that is both affordable and environmentally friendly. Bixi operates as a station-based BSS, where users can pick up a bike at one station and drop it off at another station run by the same operator.

Bixi stations offer a dock-based service, with specified spaces to store bikes. Over time, each dock at each station is either vacant or filled. Numerous new stations are opened every year, and many others are decommissioned either temporarily or permanently depending on demand. Trucks are used to physically re-balance Bixi bikes, as the number of bikes that can be borrowed and returned is limited by the capacity of docking stations. The system operator has two main objectives: to correct this imbalance and to satisfy demand. The primary challenge is that demand must be addressed every hour of every day [11]. Therefore, conducting an accurate study of stations helps prevent empty or full docking stations when returning or borrowing bikes and assists the business in finding the best locations for each station. This ultimately improves the service quality for both the business and the users.

The primary objective of this study is to gain an understanding of the demand and station characteristics of the Bixi BSS in Montreal, Canada. To achieve this, we aim to explore the diversity of Bixi stations, the complexity of the network they form, and the factors that influence their usage.

Our approach is data-driven, leveraging Bixi transaction data from 2015 to 2019 to analyze different stations based on their characteristics and usage patterns. We examine factors such as trip characteristics (including distance, travel time, peak morning and afternoon number of trips, etc.), weather variables, and land use POI (Points of Interest) data.

We use clustering techniques to group stations based on user and travel features and to analyze the spatial distribution and density of stations. This allows us to identify patterns and trends that can inform decision-making for the operation and expansion of the BSS.

Furthermore, we incorporate the effects of weather into our analysis and examine how weather conditions, proximity to facilities (POIs), the number of nearby stations, and temporal factors affect the demand for bike-sharing. We fit a linear regression model, using the number of hourly trips as the response variable and the predictor variables related to the weather, the POIs, and the time of the trip.

In summary, this study aims to contribute to the existing literature by providing a comprehensive, data-driven understanding of the Bixi BSS. The insights gained will not only help in understanding the usage patterns of Bixi stations but also guide decision-making for the operation and expansion of the BSS, ultimately contributing to improving the service quality for both the business and the users.

As we conclude this introduction, we transition into the main body of this thesis, which is structured into several key sections, each designed to provide an exploration of the research topic.

- **Introduction:** This initial chapter sets the stage for the entire study, outlining the research problem and objectives.
- **Literature Review:** Chapter 2 provides a critical review of existing research relevant to the study, identifying gaps that the current research aims to fill.
- **Methodology:** Chapter 3 explains the data sources (Bixi data, weather data, points of interest), the time frame, and the data preprocessing steps. It also describes the analytical techniques used, including regression and cluster analysis. Lastly, it discusses the evaluation metrics used.

- **Data Analysis:** Chapter 3 presents a detailed examination of the data used in the study, including Bixi data, weather data, and points of interest.

- **Influence of Weather and Proximity Factors on Bike-Sharing Demand** Chapter 5 presents a rigorous statistical analysis of the impact of weather and proximity factors on bike-sharing demand.

- **Result and Discussion:** Chapter 6 presents the findings of the research and discusses their implications in terms of station expansion, resource allocation, service improvement, and marketing strategies.

- **Conclusion:** The final chapter provides a summary of the research findings and their implications, suggesting areas for future research. research.

CHAPTER 2 LITERATURE REVIEW

The advent of BSS has significantly transformed urban mobility, offering a sustainable, healthy, and convenient mode of transportation. These systems have proliferated globally, driven by advancements in technology and a growing awareness of environmental sustainability. However, the successful operation and management of BSS pose significant challenges, and extensive research is still necessary in various domains. This literature review provides an overview of some of the key areas of research related to BSS. It covers studies on predicting demand for bike-sharing services, the impact of weather and temporal factors on bike-sharing usage, service quality, infrastructure, and land use, user behavior, clustering techniques, modeling techniques, benefits of bike-sharing programs, rebalancing strategies, and the role of BSS in urban mobility. Each of these areas offers valuable insights into the complexities of BSS and contributes to the ongoing efforts to improve and optimize these systems.

2.1 Predicting Demand for Bike-Sharing Services:

Several studies have focused on predicting demand for bike-sharing services. For example, the study by [12] estimated the hourly demand for rentals and returns at each station in Montreal Bixi BSS, using temporal and meteorological data over two years. Their model predicts demand mean and variance using machine learning and statistical inference techniques. Similarly, the study by [13] proposes a place representation based bike demand prediction framework, which uses people movement data to analyze bike demand for for New York City's CitiBike system. A recent study [14] developed a probabilistic time-series forecasting model for BSS demand using a deep-learning model called DeepAR. This model captures complex demand patterns and correlations between stations, eliminating the need

for individual station models. Using data from Seoul Metropolitan City, the study applied DeepAR to estimate parameters of various distributions. The results showed that DeepAR outperforms other models in both district- and station-level forecasts.

2.2 Impact of Temporal Factors on BSS:

Temporal factors, such as time of the day and day of the week, also affect bike-sharing demand. [15] found that bike-sharing usage is higher during peak hours and on weekdays. [16] showed that bike-sharing demand varies significantly throughout the day, with peaks in the morning and evening. A study conducted in Hamburg, Germany, used an attention-based Temporal Fusion Transformer (TFT) model to identify key factors influencing bike-sharing activity, especially in terms of temporal and spatial contexts [17]. Additionally, [18] explored the seasonal variations in bike-sharing demand and found that weather conditions, such as temperature and precipitation, significantly impact usage patterns. They used a regression analysis to quantify the effects of different weather variables on bike-sharing demand. [19] conducted a comprehensive analysis of bike-sharing data from multiple cities and highlighted the importance of considering both temporal and spatial factors in predicting demand. Their study used clustering techniques to identify distinct usage patterns and provided insights into the factors driving these patterns.

2.3 Service Quality of BSS:

Service quality has also been a topic of interest in the literature. The study by [20] proposes a SERVQUAL-based approach to measure BSS service quality, investigating five dimensions of service quality: tangibility, reliability, empathy, assurance, and responsiveness. They conducted a questionnaire study with users of Bixi and received 313 responses. The authors used SERVQUAL and compliment/complaint analysis to identify areas of improvement or negative gaps where customer perceptions were lower than expectations. A study by [21] proposed a

target-based stochastic distributionally robust optimization (TSDRO) model that addresses both the efficiency and equity of the service level in docked BSS under demand uncertainty. Furthermore, [22] examined the impact of service quality on user satisfaction and loyalty in BSS. Their findings suggest that improving service quality dimensions, such as bike availability and station maintenance, can significantly enhance user satisfaction. [23] developed a user-centric evaluation framework for assessing the service quality of BSS, incorporating user feedback and operational data.

2.4 Infrastructure and Land Use:

Infrastructure and land use have also been examined in relation to bike-sharing usage. The study by [24] analyzes how BSS infrastructure and installation process affect BSS usage in Montreal. The study uses an econometric framework and data from the BIXI system to show that previous studies have over-estimated the impact of BSS infrastructure. The land use and points of interest (POIs) around bike-sharing stations have been found to influence the demand for bike-sharing. [25] found that bike-sharing stations located near commercial areas and transit stations have higher usage. [26] showed that the presence of POIs, such as restaurants and shopping centers, increases bike-sharing demand. A study conducted in Toronto, Canada, found that station proximity to high job density and food serving enterprises was correlated with higher ridership levels [27]. Moreover, [28] investigated the role of urban design and infrastructure in promoting bike-sharing usage. Their study highlights the importance of dedicated bike lanes and safe cycling infrastructure in increasing ridership. [29] analyzed the impact of land use diversity on bike-sharing demand and found that mixed-use developments with residential, commercial, and recreational facilities attract more users.

2.5 User Behavior:

User behavior has also been studied in relation to BSS. The study by [30] developed a structural equation model to explore the factors influencing the intention to use BSSs. They introduced key concepts from consumer psychology and behavior, such as decision-making involvement, customer participation, and perceived value. The study was conducted among bike-sharing users in Xi'an, China, with a total of 622 responses collected. The researchers discovered that all the factors they considered, including the level of involvement in the decision-making process for utilizing bike-sharing services, the extent of traveler participation, and the perceived value of the service by travelers, were significantly and positively correlated with the intention to use bike-sharing. A study conducted in Seoul explored the differences in the usage patterns of BSS and the impact of explanatory factors on the demand depending on the type of pass [31].

2.6 Clustering Techniques:

Several studies have used clustering techniques to analyze bike-sharing data. These studies demonstrate the usefulness of clustering techniques in analyzing bike-sharing data and providing insights into user behavior, demand patterns, system expansions, and imbalances in bike distribution. In the study by [32], a three-level clustering approach is used to estimate the demand pattern for New York City bikes. [33] used a three-level clustering approach to estimate the demand pattern for New York City bikes. [33] used a k-medoid clustering analysis to understand the behavior of bikeshare-metro-bikeshare users. A study conducted in Ningbo, China, used k-means clustering to explore the spatiotemporal activities pattern of BSS [34].Additionally, [35] conducted a comprehensive cluster analysis of public BSS worldwide, categorizing them into distinct clusters based on various characteristics. Their study identified four main clusters: public, private, mixed, and other, providing a framework for evaluating and comparing different systems. [36] explored the impact of weather on bike-

sharing usage through clustering methods, using k-means to identify three clusters corresponding to different weather conditions in Washington D.C. Their findings highlighted the significant influence of temperature and precipitation on bike usage patterns.

2.7 Modeling Techniques in BSS:

Various modeling techniques have been used to predict bike-sharing demand. [37] used machine learning and statistical inference techniques to predict demand mean and variance. [38] proposed a hierarchical prediction model that uses a bipartite clustering algorithm to cluster bike stations and a Gradient Boosting Regression Tree (GBRT) to predict the total number of bikes that will be rented in a city. A survey by [39] reviewed the latest studies about bike-sharing usage prediction with deep learning, with a classification for the prediction problems and models. Furthermore, [40] conducted a systematic literature review on machine learning approaches to BSS, highlighting the contributions of various algorithms in improving demand prediction. Their review identified key factors influencing BSS and categorized machine learning algorithms into classification and prediction groups, focusing on studies conducted in various cities worldwide. [41] introduced a novel hybrid deep learning model using graph convolutional neural networks to predict travel demand at the station level in Chicago. Their model demonstrated significant improvements in forecasting accuracy by integrating trajectory, weather, and access data.

2.8 Benefits of Bike-Sharing Programs:

Finally, several studies have discussed the benefits of bike-sharing programs. The studies by [42] and [43] provide an overview of bike-sharing programs and their benefits. A study by [44] estimated quantitatively the potential of bike-sharing to promote transport resilience. [45] reviewed recent literature on BSS and highlighted their positive impacts on public health, environment, and urban mobility.

2.9 Rebalancing Strategies in BSS:

The study by Yi, Huang, and Peng [46] proposed a customer-oriented rebalancing strategy for BSS. They used a one-dimensional Random Walk Process with two absorption walls to calculate the optimal state of each station. Another study by Freund et al. [47] reported on a collaborative effort with Citi Bike to develop and implement real data-driven optimization to guide their rebalancing efforts. Furthermore, Beigi et al. [48] conducted a case study on the Capital Bikeshare system in Washington D.C., proposing a deterministic integer programming model for station reallocation and rebalancing. Cipriano, Colomba, and Garza [49] introduced a dynamic rebalancing methodology based on frequent pattern mining, demonstrating its effectiveness using data from the Barcelona bike-sharing system.

2.10 BSS and Urban Mobility:

Rojas-Rueda and Clockston [50] studied the global activity of users in BSS operating in the cities of Chicago and New York. They explored the temporal and spatial characteristics of the mobility of cyclists. Another study by Kon et al. [51] introduced a novel analytical method that can be used to process millions of bike-sharing trips and analyze bike-sharing mobility. Additionally, [52] analyzed the structure of local and non-local dynamics in BSS in Chicago and New York, using origin-destination matrices to characterize spatial displacements.

2.11 BSS and Environmental Sustainability:

A study by Martinez and Tapia [53] discussed the positive externalities of BSS, including providing urban residents with a convenient and time-saving travel mode. Another study by Fishman [54] reviewed recent literature on BSS and their impacts on the economy, energy use, the environment, and public health. Furthermore, [55] quantitatively explored the externalities of BSS in China, finding significant positive impacts on traffic reduction, en-

ergy consumption, and public health. [56] discussed the design, simulation, and management of BSS to promote sustainability, highlighting the benefits of dockless systems and electric bicycles.

2.12 BSS and Public Health:

Reichel [57] reviewed research examining the health effects of bike sharing, distribution of access to these services, use of helmets, and more. Another study by Kille [58] provided an overview of the state of research on bike-sharing programs. Additionally, [59] quantified the public health benefits of BSS in the U.S., estimating savings of over $36 million annually in healthcare costs. [60] reviewed the potential of bike-sharing during public health crises, such as COVID-19, highlighting its role in maintaining transport resilience and supporting essential workers.

2.13 Economic Impact of BSS:

The study by Sergi Martínez et al. [61] analyzed the socioeconomic dimension of BSS and found that every euro invested in bike-sharing public services generates an average impact that ranges from 79 cents to €1.14. Another study by Li and Wang [62] explored how bike-sharing platforms achieve sustained growth in the digital era.

However, the optimal planning and management of BSS depend on the understanding of the travel demand and behavior of bike-sharing users. Therefore, we study the BSS stations based on factors such as travel time, number of trips, distance, the type of days (week-end and holidays), peak hours, and different categories of POIs. These factors can help us identify the usage and performance of BSS stations, as well as the preferences and needs of bike-sharing users.

While numerous studies have explored various aspects of BSS, there remains a significant

gap in understanding BSS at the station level using factors related to trips. Many studies have focused on predicting demand for bike-sharing services or understanding the impact of weather and temporal factors on BSS usage. However, there is limited research on the specific characteristics of individual stations and how these characteristics influence the performance and usage of BSS. Furthermore, while clustering techniques have been used to analyze BSS data, these studies often focus on general usage patterns or system-level insights rather than station-specific analysis. This thesis aims to fill this gap by adopting a data-driven approach to understand user behavior and station characteristics in the Bixi BSS in Montreal, Canada. By utilizing open data and creating variables from historical Bixi data, this study employs methods such as regression, visualization, and clustering to identify key trends and factors influencing the number of trips at each station. The analysis of station lifespan from 2015 to 2020 further adds a temporal dimension to the study, providing insights into the evolution of Bixi stations over time. Ultimately, this work emphasizes the importance of data-driven strategies in optimizing BSS operations and improving user satisfaction. Future work could explore predictive models for bike demand and the application of clustering methods to other datasets, expanding the understanding of BSS at the station level and beyond

CHAPTER 3 METHODOLOGY

This chapter outlines the methodology employed in our study of the Bixi bike-sharing system in Montreal. We use different data sources and analytical techniques to uncover patterns and trends in bike usage, and how these correlate with points of interest (POIs) and weather conditions. The diagram 3.1 outlines the steps involved, from data collection to spatial clustering, regression modeling, and final clustering based on usage and characteristics. This visual representation provides a clear overview of the methodology and workflow.

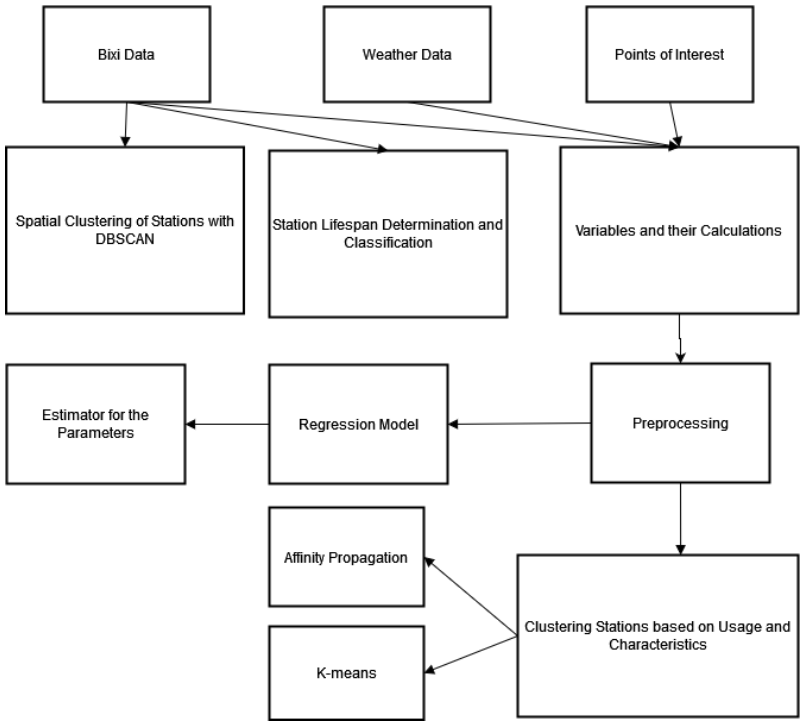


Figure 3.1 Flowchart of Analysis Process

3.1 Data

Certainly, the emergence of multi-source big data opens up a new paradigm for improving bike demand analysis. We use data from Bi bike-sharing services, POI data, and weather data to construct station-level bike demand solutions. Specifically, for this study, we used the three following data sets:

- Bixi historical data and station status [63]
- Weather data set [64]
- Point Of Interest (POI) [65]

3.1.1 Bixi data

Bixi Montreal releases trip data monthly and yearly through its Open Data website. Bixi's open data includes Origin and destination (OD) and station data from 2014 to the present. The OD data is released monthly which includes for each trip the start and end station code, the start and end date and time, and the travel time duration in seconds. Data about each station that is released yearly includes the longitude and latitude of each station, each station code, and each station name. Our geographical location primarily influenced our choice of Montreal as the focus for our data analysis. We selected the time frame from 2015 to 2020 to ensure a robust and substantial dataset for our analysis. It is important to note that our research commenced in 2022, at which point the data for 2021 was not sufficiently comprehensive for inclusion in our study.

3.1.2 Weather data

The weather dataset utilized in this study is sourced from the Open Data website of Climate Canada [64], which provides historical weather data on an hourly basis. Specifically, the

Table 3.1 Bixi Data: Trip Records

Variable	Description	Type
start_date	The date and time when the trip started	Continuous - DateTime
start_station_code	The code of the station where the trip started	Categorical - Nominal
end_date	The date and time when the trip ended	Continuous - DateTime
end_station_code	The code of the station where the trip ended	Categorical - Nominal
duration_sec	The duration of the trip in seconds	Continuous - Interval
is_member	A binary value indicating whether the user is a member	Categorical - Binary

Table 3.2 Bixi Data: Station Status Data

Variable	Description	Type
code	The unique code of the station	Categorical - Nominal
name	The name of the station	Categorical - Nominal
latitude	The latitude coordinate of the station	Continuous - Ratio
longitude	The longitude coordinate of the station	Continuous - Ratio

dataset employed for this research is derived from the MONTREAL/PIERRE ELLIOTT TRUDEAU INTL station. In our preliminary analysis, we observed that several datasets from different stations contained variables with over 90 percent missing values. Given the significance of precipitation amounts in influencing ride-sharing patterns, as suggested by prior studies and our understanding, we supplemented our dataset with precipitation amounts from the MONTREAL MIRABEL INTL weather station in Montreal. This approach allowed us to create a more comprehensive dataset for our study.

- Temperature: The temperature of the air in degrees Celsius (°C). At most principal stations the maximum and minimum temperatures are for a day beginning at 0601 Greenwich (or Universal) Mean Time, which is within a few hours of midnight local standard time in Canada.

- Precipitation: The total hourly precipitation is the total precipitation amount for minutes 00 through 60, inclusive, computed as the sum of the four 15-minute precipitation amounts. Precipitation amounts are stored in mm with a resolution of 0.1 mm [64].

- Wind Speed: The speed of motion of air in kilometres per hour (km/h) usually observed at 10 metres above the ground. It represents the average speed during the one-, two- or ten-minute period ending at the time of observation. In observing, it is measured in nautical miles per hour or kilometres per hour [64].
- Visibility: Visibility in kilometres (km) is the distance at which objects of suitable size can be seen and identified. Atmospheric visibility can be reduced by precipitation, fog, haze or other obstructions to visibility such as blowing snow or dust [64].
- Wind Chill: Wind chill is an index to indicate how cold the weather feels to the average person. It is derived by combining temperature and wind velocity values into one number to reflect the perceived temperature.
- Weather: Categorical variable describing the overall weather conditions (e.g., Clear, Partly Cloudy, Mostly Cloudy, Overcast, Rain, Snow, Fog, Thunderstorm) [64].
- Precipitation: The amount of precipitation (rain, snow, etc.)
- Station Pressure (kPa): The atmospheric pressure in kilopascals (kPa) at the station elevation. Atmospheric pressure is the force per unit area exerted by the atmosphere as a consequence of the mass of air in a vertical column from the elevation of the observing station to the top of the atmosphere [64].

After analyzing the weather variables, we identified a set of variables that showed significant correlations. Therefore, we included the following variables in the regression analysis:

- Precipitation Amount (mm)
- Clear (from weather feature)
- Fog (from weather feature)

- Rain (from weather feature)
- Snow (from weather feature)

3.1.3 Points of Interest

The riders of BSSs frequently commute between specific functional areas of the city [66]. To account for these tendencies, we incorporated Points of Interest (POIs), which are defined as specific locations that could be of interest or utility, situated near each bike station. The dataset encompasses different types of POI such as Cultural, Commercial, Recreational/Sports, and Public Service. These POIs are extracted from the 'Famille' variable in the POI dataset, which was employed in this study. Table 3.3 illustrates the various categories and the 'Famille' variable (category) and the frequency distribution of each Category-Sub_Category Pair in the raw data. The dataset was obtained from the Open Government website for Montreal [65].

Table 3.3 The frequency distribution of POI in different categories.

Category	**Sub_Category**	**Category Sub_Category Frequency**
Commercial	Tourist Attraction	1
Commercial	Building and Place of Interest	1
Commercial	Convention / Exhibition Center	3
Commercial	Accommodation	173
Commercial	Group of Shops	65
Cultural	Public Art	108
Cultural	Tourist Attraction	7
Cultural	Library	50
Cultural	Building and Place of Interest	139
Cultural	Circuit and Route	53
Cultural	Place of Diffusion	153
Cultural	Cultural Establishment	47
Recreational / Sports	Tourist Attraction	10
Recreational / Sports	Building and Place of Interest	1
Recreational / Sports	Park and Other Green Space	955
Recreational / Sports	Community Equipment	260
Recreational / Sports	Sports and Recreational Equipment	162
Recreational / Sports	School Establishment	2
Public Service	Tourist Information	4
Public Service	Emergency Service	82
Public Service	Health Service	77
Public Service	Governmental Service	80
Public Service	Municipal Service	95
Public Service	Transport	205
Public Service	School Establishment	498

3.2 Station to Station Distance

A BSS allows people to use bicycles whenever they need them, without incurring the cost and obligations that come with owning a bicycle. Because of bike-sharing flexibility, this mode of transportation is ideal for short distances and one-way excursions [5]. To be acceptable to bike-sharing users, the distance between stations, the start and destination of the trip, should be kept to a minimum [67]. The distance between bike stations must also be addressed when deciding where to put bike stations. In Paris, for example, bike stations are positioned every four blocks, or 300 meters, providing convenient access [68].

In this study, we used the Python library `geopy` to calculate the geodesic distance between bike-sharing stations. The geodesic distance, calculated using the `geopy.distance.geodesic` function, is the shortest distance between two points on the Earth.

The `geopy.distance.geodesic` function takes as input two tuples, each representing the latitude and longitude of a point. It returns the geodesic distance between these two points in kilometers.

The `geopy` library is widely used in geographic and geospatial analysis due to its accuracy and ease of use. It is well-documented and maintained, making it a reliable choice for calculating distances. For example, in a study by [69], the `geopy` library was used to calculate distances between various geographic points to analyze travel patterns. Similarly, [70] utilized `geopy` to determine the proximity of service locations in urban planning research.

In the context of Bixi, distance and travel time are inherently linked. When users rent a bike, they travel a certain distance over a specific period. Therefore, understanding the relationship between these two variables can provide valuable insights into user behavior and the efficiency of the ride-sharing system.

3.3 Variables and their Calculations

In order to conduct a comprehensive study of the Bixi BSS, we have performed feature engineering on our dataset. This involved extracting valuable information from the transaction data to describe the stations, thereby expanding our dataset beyond the limited set of variables provided in the Bixi open data. The additional variables we calculated capture various aspects of the system's usage and characteristics.

In this section, we will provide a definition and clarification of their calculation process.

Notations

- Let S be the set of all stations, where each station $s \in S$ has a specific latitude and longitude.
- Let C be the set of categories of POIs: ['Cultural', 'Recreational/sporting', 'Public service', 'Commercial'].
- Let P be the set of all Points of Interest (POIs), where each POI $p \in P$ belongs to a category $c \in C$.
- Let D be the maximum distance considered for counting POIs.
- Let $d(s, p)$ be the distance between a station s and a POI p.
- Let $d(s_1, s_2)$ be the geodesic distance between two stations s_1 and s_2.
- Let R be the radius within which we are counting nearby stations (0.4 km in this case).
- Let T represent the number trips of Bixi bike users.
- Let M represent the number of trips made by members.
- Let H be the hourl number of hours during the day.

- Let h_i be the hour of the day for the i-th row in the dataset.
- Let w_i indicate whether the i-th row is a weekend (1 if yes, 0 otherwise).
- Let $I_{s,w}$ be the set of indices for each station s and whether it's a weekend w.
- Let $N_{s,w}$ be the number of elements for each station s and whether it's a weekend w, represented as $|I_{s,w}|$.
- Let $N_{h,s,d}$ be the number of trips at hour h, station s, and day d, where d belongs to weekends.
- Let $M_{s,w}$ be the sum of trips done by members for each station s and whether it's a weekend w.
- Let TT represent the hourly travel time.
- Let $PTM_{s,w}$ be the percentage of trips done by members for each station s and whether it's a weekend w.
- Let $\bar{T}_{s,w}$ be the mean value of the hourly number of trips for each station s and whether it's a weekend w.
- Let V represent the set of variables: ['duration_sec', 'distance', 'peak_morning_distance', 'peak_afternoon_distance', 'peak_morning_travel_time', 'peak_afternoon_travel_time', 'weekend_peak_afternoon_travel_time', 'travel_time', 'Percentage of travel time weekend', 'Percentage of travel time weekdays'].
- Let $\bar{V}_{s,v,w}$ be the mean value of each variable v in V for each station s.
- Let $PTMW_{s,w}$ The percentage of trips done with members for each station and whether it's weekend.

Calculations

- The function Number of POIs can be represented as:

$$\text{Number of POIs(s, c)} \ = \sum_{p \in P_c} [d(s,p) < D] \tag{3.1}$$

where P_c is the set of POIs that belong to category c, and $[d(s,p) < D]$ is an indicator function that equals 1 if the condition is true and 0 otherwise. This function calculates the number of POIs of a certain category that are within a certain distance from a station. This helps understand the environment around a station and its potential attractiveness to users.

- The function for the number of nearby stations can be represented as:

$$\text{Number of Nearby Stations}(s) = \sum_{s' \in S, s' \neq s} [d(s,s') \leq R] \tag{3.2}$$

where $[d(s,s') \leq R]$ is an indicator function that equals 1 if the condition is true (i.e., the distance from station s to station s' is less than or equal to R (400m in this study) and 0 otherwise. In the context of BSS, the distance between stations plays a crucial role in the system's efficiency and user satisfaction. According to [71], the likelihood of a person using a bike-sharing service decreases significantly with increasing distance from a station. Specifically, someone roughly 300m from a station is 60% less likely to use the bike-sharing service than someone right next to it. The chances that someone will use the service decrease with increasing distance and at around 500m away it is highly unlikely that a person will consider using it. The same study concluded that decreasing walking distance to a station has a much higher impact than increasing bike availability.

Moreover, a study conducted in Washington D.C. found that bike share stations attract

more businesses because riders are more likely to spend within a four-block radius as a byproduct of convenience [72]. This highlights the economic benefits of strategically placing bike-sharing stations.

Furthermore, a systematic review of station location techniques for bicycle-sharing systems planning and operation mentions that the criteria to select the best place for the station were: 300–500 m as a minimum distance to important origin or destination bicycle trip generators [73]. This suggests that the optimal distance between stations can vary based on several factors, including the specific characteristics of the area, user behavior, and the goals of the bike-sharing system.

This function calculates the number of other stations within a certain radius of a station.

- The mean number of trips(hourly) during weekends is given by:

$$\mu_{s,\text{weekend}} = \frac{1}{H} \sum_{h=1, d \in \text{weekend}}^{H} N_{h,s,d} \tag{3.3}$$

$$\text{Peak Hours (Weekend)} = \{h : N_{h,s,d} = \max_{h,d \in \text{weekend}} (N_{h,s,d})\} \tag{3.4}$$

These functions calculate the average number of trips at each hour of each day in the weekend for a station and identify the hours during the weekend when the number of trips is at its maximum.

- For each row i in the data frame, if the hour h_i is within the peak hours, the corresponding metric m_{h_i} (travel time, distance, or number of trips) is recorded in the new column. Otherwise, a zero is recorded. This is done for both weekdays and weekends

and for morning and afternoon peak hours. The functions are:

$$\text{Peak Morning\_metric}_i = \begin{cases} m_{h_i} & \text{if } h_i \in \{7, 8, 9\} \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

$$\text{Peak Afternoon\_metric}_i = \begin{cases} m_{h_i} & \text{if } h_i \in \{16, 17, 18\} \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

$$\text{Weekend Peak Afternoon\_metric}_i = \begin{cases} m_{h_i} & \text{if } h_i \in \{14, 15, 16, 17\} \text{ and } w_i == 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

These functions record the metric (travel time, distance, or number of trips) for each row in the data frame if the hour is within the peak hours.

- The percentage of trips done by members for each station is calculated as:

$$PTM_s = \frac{M_s}{\sum_{i \in S} M_i} \times 100 \tag{3.8}$$

where:

  - PTM_s is the percentage of trips made by members at station s.
  - M_s is the number of trips made by members at station s.
  - $\sum_{i \in S} M_i$ is the total number of trips made by members at all stations.

This function calculates the percentage of trips made by members at each station.

- The percentage of trips done with members for each station and whether it's weekend is calculated as:

$$PTMW_{s,w} = \frac{M_{s,w}}{\sum_{i \in I_{s,w}} M_i} \times 100 \tag{3.9}$$

This function calculates the percentage of trips made by members at each station during weekends and weekdays.

- The percentage of travel time for each station's trip and whether it's weekend is calculated as:

$$PTT_{s,w} = \frac{TT_{s,w}}{\sum_{i \in I_{s,w}} TT_i} \times 100 \tag{3.10}$$

This function calculates the percentage of travel time at each station during weekends and weekdays.

- For each variable v in V and each station s and whether it's a weekend w, the function calculates the mean value as:

$$\bar{V}_{s,v,w} = \frac{1}{N_{s,w}} \sum_{i \in I_{s,w}} V_{i,v} \tag{3.11}$$

where $I_{s,w}$ is the set of indices in the group and $N_{s,w} = |I_{s,w}|$ is the number of elements in the group. This function calculates the mean value of each variable for each station during weekends and weekdays.

- Let T represent the Trips. For each station s and whether it's a weekend w, the function calculates the mean value of trips as:

$$\bar{T}_{s,w} = \frac{1}{N_{s,w}} \sum_{i \in I_{s,w}} T_i \tag{3.12}$$

where $I_{s,w}$ is the set of indices in the group and $N_{s,w} = |I_{s,w}|$ is the number of elements in the group. This function calculates the average number of trips from each station during weekends ($w = 1$) and weekdays ($w = 0$). It's calculated by summing up the number of trips T_i made from the station for all indices i in the group $I_{s,w}$, and then dividing by the number of elements in the group $N_{s,w}$.

- The percentage of trips for each station and whether it's a weekend is calculated as:

$$PT_{s,w} = \frac{T_{s,w}}{\sum_{i \in I_{s,w}} T_i} \times 100 \tag{3.13}$$

  where $PT_{s,w}$ is the percentage of trips at station s during the weekend w, $T_{s,w}$ is the number of trips at station s during the weekend w, and $\sum_{i \in I_{s,w}} T_i$ is the total number of trips at all stations during the weekend w. n calculates the percentage of trips at each station during weekends and weekdays.

These calculations are used to extract valuable information from the transaction data to describe the stations, thereby expanding the dataset beyond the limited set of variables provided in the Bixi open data. The additional variables capture various aspects of the system's usage and characteristics. This feature engineering process is crucial for conducting a comprehensive study of the BSS. It underscores the importance of considering the specific characteristics and demands of each station when making operational decisions.

3.4 Spatial Clustering of Stations with DBSCAN

By clustering stations based on their geographic coordinates, we aim to understand the spatial distribution and density of stations across the city. Clustering provides a quantitative measure of the density and distribution of stations, identifying areas of high density (many stations close together) and low density (few stations far apart), which might not be immediately apparent from a map. For large datasets covering a wide geographic area, visual inspection can become challenging. Clustering can handle these large datasets and provide insights that are difficult to obtain visually. The results of clustering can inform decision-making processes. For example, if a cluster of stations has significantly higher usage, additional resources could be allocated there. This approach allows for data-driven decision making, optimizing resource allocation based on actual usage patterns.

To do so, we employed Density-Based Clustering on their geographic coordinates. This unsupervised learning approach identifies distinct clusters in the data, where a cluster is defined as a contiguous region of high point density separated from other clusters by regions of low point density.

We used the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm for this purpose. Unlike partitioning-based clustering methods such as k-means, DBSCAN does not require the user to specify the number of clusters in advance. Instead, it discovers clusters based on the density of data points in a region.

The data points in this case are the geographic coordinates of the stations, specifically their latitude and longitude. These two features form the basis of our clustering task. Each station is represented as a data vector with two dimensions - latitude and longitude.

DBSCAN works by defining a neighborhood around each data point. If there are at least a minimum number of points (MinPts) of 6 within a given radius (ϵ) 0.38, that data point is considered a core point. Points that are within the (ϵ) radius of a core point, but do not have MinPts within their own (ϵ) radius, are considered to be border points. All other points are considered noise. The algorithm starts with an arbitrary data point, and if there are at least MinPts within a radius of (ϵ) from that point, a new cluster is created. The cluster then grows by adding all directly reachable points to the cluster. This process continues until no more points can be added to the cluster.

In our implementation, we divided 'eps' by the Earth's radius to convert it into radians before passing it to the DBSCAN algorithm. This adjustment was necessary because our data is in latitude and longitude format. We set 'min_samples' to a specific value that best suited our dataset. For the 'algorithm' parameter, we used the 'ball_tree' algorithm. For the 'metric' parameter, we used the 'haversine' metric because our data is in latitude and longitude format, and the haversine formula gives great-circle distances between two points on a sphere from their longitudes and latitudes.

3.5 Station Lifespan Determination and Classification

The lifespan of a station is defined as the number of unique years the station appears in the dataset in our study it is from 2015 to 2020. This approach is used to understand the longevity and usage patterns of different bike-sharing stations over time. By defining the lifespan of a station as the number of unique years it appears in the dataset, researchers can gain insights into the duration of each station's activity. Additionally, they can determine the number of stations that have been added and their respective locations.

Let $G = \{g_1, g_2, ..., g_n\}$ be the set of groups formed by grouping the data by station code, where n is the number of unique station codes. For each group g_i, let $Y(g_i)$ represent the set of unique years the station appears in the dataset.

The lifespan of a station s belonging to group g_i is given by $L_s = |Y(g_i)|$, where $|Y(g_i)|$ denotes the cardinality of the set $Y(g_i)$, i.e., the number of unique years the station appears in the dataset.

Based on the lifespan, we classify the stations using the following function:

$$Class(L_s) = \begin{cases} \text{Class 1} & \text{if } L_s = 1 \\ \text{Class 2} & \text{if } L_s = 2 \\ \text{Class 3} & \text{if } L_s = 3 \\ \text{Class 4} & \text{if } L_s = 4 \\ \text{Class 5} & \text{if } L_s = 5 \\ \text{Class 6} & \text{if } L_s = 6 \\ \text{Class Unknown} & \text{otherwise} \end{cases}$$

3.6 Regression model

We examine how weather conditions, points of interest (POIs), the number of nearby stations, and temporal factors affect the demand for Bixi BSS. A linear regression model (Ordinary Least Squares (OLS)) was fitted, using the number of hourly trips as the response variable and the predictor variables related to the weather, the POIs, and the time of the trip. The Ordinary Least Squares (OLS) model is a widely used statistical method in linear regression. The primary objective of OLS is to find the hyperplane that minimizes the sum of the squared differences between the observed and predicted values of the dependent variable. This is achieved by estimating the parameters that minimize the sum of the squared residuals.

The mathematical representation of the OLS model is given by:

$$Y = X\beta + \epsilon \tag{3.14}$$

where:

- Y represents the dependent variable.
- X denotes the matrix of independent variables.
- β is the vector of parameters to be estimated.
- ϵ is the error term.

3.6.1 OLS Estimator for the Parameters

The OLS estimator for the parameters is given by:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3.15}$$

where:

- $\hat{\beta}$ represents the estimated parameters.
- X' is the transpose of the matrix X.
- Y is the dependent variable.

3.7 Clustering Stations based on Usage and Characteristics

In our study, we harness the power of clustering techniques to categorize bike-sharing stations, taking into account their usage patterns and distinct characteristics. This method illuminates the underlying structures within the data, offering valuable insights that can guide strategic decision-making

For this analysis, we considered numerous variables encompassing aspects such as the distance, members' percentage, peak morning and afternoon distances, travel times during different periods of the day and week, the number of trips, and the percentage of trips over the weekend.

We experimented with different clustering techniques, including agglomerative clustering, hierarchical clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Gaussian Mixture Models. After evaluating the performance and suitability of these methods, we selected affinity propagation and k-means clustering as the most effective techniques for our analysis.

By grouping stations with similar characteristics, we can identify trends, optimize operations, and enhance the efficiency of the BSS. we renamed the clusters to maintain uniformity between the two clustering techniques, AP and k-means. This approach facilitated easier comparison and analysis.

In the preprocessing steps, we opted to use robust scaling to normalize the data. We chose

robust scaling based on its performance when compared with other normalization methods. These comparisons were conducted as part of our preliminary analysis, the details of which are presented in the transformations section of this thesis.

By applying robust scaling, we ensured that all features contribute equally to the model performance and prevented any one feature from dominating others due to its scale. This step was crucial in preparing our data for the subsequent clustering analysis using the DBSCAN algorithm.

In the following sections, we will delve deeper into the specifics of these clusters, providing a granular view of the station clusters and their defining characteristics. This preliminary explanation aims to provide a roadmap for the detailed exploration that follows, setting the stage for a deep and nuanced understanding of the station clusters and their implications.

3.7.1 Affinity propagation

Affinity propagation (AP) is a clustering algorithm that was proposed that Frey and Dueck proposed in the Science journal in 2007 [74]. It is based on the idea of finding representative data points, called exemplars, that have high affinity with other data points in the same cluster. Unlike other clustering methods that require specifying the number of clusters in advance, AP can automatically determine the optimal number of clusters by maximizing a global criterion called net similarity.

The algorithm operates by exchanging messages between data points. Each message represents the suitability of one data point to be the exemplar of another. There are two types of messages: responsibility and availability.

Responsibility, denoted as $r(i, j)$, reflects how well-suited a data point x_j is to serve as the exemplar for another point x_i, considering all other potential exemplars. The self-responsibility $r(j, j)$ represents how well-suited data point x_j is to serve as its own exemplar.

Availability, denoted as $a(i, j)$, communicates how appropriate it would be for a data point x_i to choose another point x_j as its exemplar, taking into account the support from other points for this choice.

These messages are updated iteratively until the algorithm converges, at which point the final exemplars and clusters are identified. The similarity between data points, denoted as $s(i, j)$, is a key factor in these calculations. It quantifies how alike two data points are, influencing both the responsibility and availability calculations. The more similar two data points are, the more likely they are to be in the same cluster.

AP has been applied to various domains, such as image segmentation, face recognition, gene expression analysis, and text clustering [75, 76]. AP has several advantages over other clustering algorithms, such as robustness to noise and outliers, flexibility to handle different types of data and similarity measures, and scalability to large datasets. However, AP also has some drawbacks, such as high computational complexity, sensitivity to parameter settings, and difficulty in handling non-convex clusters.

The equations for updating the responsibility and availability matrices are as follows:

$$r(i, j) \leftarrow s(i, j) - \max_{j' \neq j} \{a(i, j') + s(i, j')\}$$

$$a(i, j) \leftarrow \min\{0, r(j, j) + \sum_{i' \notin \{i,j\}} \max\{0, r(i', j)\}\}$$

where $r(i, j)$ is the responsibility of data point x_j to be the exemplar for data point x_i, $s(i, j)$ is the similarity between data points x_i and x_j, and $a(i, j)$ is the availability of data point x_j to be the exemplar for data point x_i.

3.7.2 k-means

The k-means clustering algorithm aims to partition a set of data points into k disjoint clusters such that the within-cluster variance is minimized. Each cluster is represented by the mean of its members, also known as the cluster centroid. The optimization in terms of this algorithm as follows:

$$\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{3.16}$$

where C is the set of clusters, k is the number of clusters, x is a data point, μ_i is the centroid of cluster C_i, and $\| \cdot \|$ is the Euclidean norm.

3.8 Handling Cyclical Variables through Sine Transformation

The sine transformation is used in the regression analysis to handle cyclical variables such as time of day and time of year. These variables are cyclical in nature because they repeat after a certain period. For example, the time of day repeats every 24 hours and the time of year repeats every 12 months.

The sine transformation is particularly useful for these types of variables because it can capture the cyclical pattern in a way that linear variables cannot. A linear variable assumes that a unit change in the variable has the same effect, regardless of the value of the variable. However, for cyclical variables, a unit change can have different effects depending on the current value of the variable.

For instance, consider the time of day variable. The difference between 23:00 and 01:00 is the same as the difference between 01:00 and 03:00 in terms of hours passed. However, if we were to model time of day as a linear variable, the model would interpret the difference between 23:00 and 01:00 as -22 hours, which is not accurate.

By applying the sine transformation, we can correctly model the cyclical nature of these variables. The sine function has a range between -1 and 1, and it repeats every 2π units. So, we scale our variable to match this cycle. For example, for hour of the day, we would transform it using $\sin\left(\frac{2\pi \cdot \text{hour}}{24}\right)$, and for month of the year, we would transform it using $\sin\left(\frac{2\pi \cdot \text{month}}{12}\right)$.

This transformation allows the model to understand that the time of day and time of year are cyclical, and it can therefore make more accurate predictions.

3.9 Preprocessing

Different scaling and transformation methods can have varying effects on the results of machine learning algorithms, depending on the characteristics of the data and the algorithm being used. In the case of Bixi, we discovered that Robust Scaling transformation was particularly effective in preprocessing the data for clustering. This could be due to a number of factors, such as the distribution of the trips or the presence of outliers, since the data is also really skewed. For this study, we employed Affinity Propagation and K-means clustering techniques in our analysis.

- **MinMaxScaler**: This method scales the data by transforming it to a specific range, usually between 0 and 1. It does this by subtracting the minimum value of the feature from each data point and then dividing by the range of the feature (maximum value - minimum value). x is the original value of a data point. The formula for this is:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.17}$$

- **Robust Scaling**: This method scales the data by using the first quartile (Q_1) and third quartile (Q_3) values, which makes it more robust to outliers compared to methods like Min-Max scaling.

In this method, Q_1 represents the first quartile, which is the value below which 25% of the data falls. Similarly, Q_3 represents the third quartile, which is the value below which 75% of the data falls.

Given a feature x, the scaled value of an individual data point x_{scaled} is computed as follows:

$$x_{\text{scaled}} = \frac{x - Q_1(x)}{Q_3(x) - Q_1(x)} \tag{3.18}$$

In this formula, x is an individual data point from the scaling feature . The scaled value x_{scaled} is computed by subtracting the first quartile of the feature $Q_1(x)$ from x, and then dividing by the interquartile range $Q_3(x) - Q_1(x)$. This results in a new feature where the interquartile range is between 0 and 1.

- **Standard Scaler**: In this method, given a feature x, the standardized value of an individual data point, denoted as z, is computed as follows:

$$z = \frac{x - \mu}{\sigma} \tag{3.19}$$

In this formula: - z is the standardized value of the data point. - x is the original value of the data point. - μ is the mean of the feature. - σ is the standard deviation of the feature.

This method standardizes the data by subtracting the mean of the feature from each data point $(x - \mu)$, and then dividing by the standard deviation (σ).

- **Log Transform**: This method applies a logarithmic transformation to the data. This can be useful when dealing with data that has a skewed distribution. The formula for this is:

$$x_{scaled} = \log(x) \tag{3.20}$$

3.10 Evaluation metrics

In this study, we aim to group BSSs based on their usage and characteristics. To achieve this, we employ two different clustering methods: k-means and Affinity Propagation. Given the distinct nature of these methods, it is crucial to understand how similar or different the results they produce are. This is where the role of evaluation metrics comes in.

The chosen evaluation metrics allow us to quantitatively compare the clusters formed by the two methods. They provide a measure of similarity between the cluster assignments, ensuring that the comparison is objective and not influenced by the inherent differences in the clustering methods.

By comparing these metrics for K-Means and Affinity Propagation, we can assess the consistency between these methods in the context of our specific task. If the metrics yield similar scores for both methods, it suggests that despite their different approaches, both methods are identifying similar patterns in the data. On the other hand, significant differences in the scores could indicate that one method may be more suited to this particular task than the other.

- **F-statistic**: This statistic tests the overall significance of the model. The null hypothesis is that all of the regression coefficients are equal to zero. The F-statistic is calculated as follows:

$$F = \frac{\text{Explained variance}}{\text{Unexplained variance}} \tag{3.21}$$

- **Prob (F-statistic)**: This is the probability that the null hypothesis for the F-statistic is true (i.e., all of the regression coefficients are zero).

- **Log-Likelihood**: This statistic measures the log of the likelihood that the model would produce the observed values of the dependent variable. The higher the log-likelihood, the better the model is at predicting the observed values.

- **RMSE (Root Mean Squared Error)**: This statistic measures the average magnitude of the residuals or prediction errors. It is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.22}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

- **Residual Standard Error (RSE)** :In the analysis of the linear regression model, the Residual Standard Error (RSE) was computed to assess the quality of the model fit. The RSE provides a measure of the standard deviation of the residuals, indicating the typical difference between the observed and predicted values. It is calculated using the following formula:

$$RSE = \sqrt{\frac{\Sigma(y - \hat{y})^2}{df}}$$

where: - y is the observed value, - $\hat{y}$ is the predicted value, - df is the degrees of freedom, calculated as the total number of observations minus the total number of model parameters.

- **MAE (Mean Absolute Error)**: This statistic measures the average magnitude of the residuals or prediction errors, without considering their direction. It is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.23}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

- **Adjusted Rand Index (ARI)**: The Adjusted Rand Index (ARI) is a measure used to compare two different cluster assignments, regardless of the actual labels. An ARI

score of 1 signifies perfect concordance between two clusterings, implying that they are identical. Conversely, an ARI score of 0 indicates that the observed agreement between the two clusterings is equivalent to what would be expected by random chance.

In the ARI formula:

- a_i denotes the number of data points in the i-th cluster in the first clustering assignment.
- b_j denotes the number of data points in the j-th cluster in the second clustering assignment.
- n_{ij} denotes the number of data points that are in the i-th cluster in the first clustering assignment and in the j-th cluster in the second clustering assignment.
- n is the total number of data points.

The equation for ARI is:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{3.24}$$

- **Homogeneity**: In the context of the homogeneity score calculation:

  - C and K represent two different cluster assignments of the same set of data points.

  - The entropy $H(C)$ of a clustering C is a measure of its uncertainty or randomness. It's calculated as:

$$H(C) = -\sum_{i=1}^{n} p(c_i) \log p(c_i) \tag{3.25}$$

where $p(c_i)$ is the proportion of data points in cluster c_i in the clustering C, and the sum is over all clusters in C.

- The conditional entropy $H(C|K)$ of a clustering C given another clustering K is a measure of the uncertainty of C after K is known. It's calculated as:

$$H(C|K) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(c_i, k_j) \log \frac{p(c_i, k_j)}{p(k_j)} \tag{3.26}$$

where $p(c_i, k_j)$ is the proportion of data points in both cluster c_i in C and cluster k_j in K, $p(k_j)$ is the proportion of data points in cluster k_j in K, and the sums are over all clusters in C and K.

The homogeneity score h is then calculated as:

$$h = 1 - \frac{H(C|K)}{H(C)} \tag{3.27}$$

- **V-measure**: The V-measure is the harmonic mean of homogeneity h and completeness c, given by $v = 2 \cdot \frac{h \cdot c}{h+c}$. It quantifies the effectiveness of the clustering by considering both the homogeneity and completeness of the clustering. A V-measure score close to 1 indicates that both homogeneity and completeness are high.

- **Completeness**: The completeness score c is given by $c = 1 - \frac{H(K|C)}{H(K)}$, where $H(K|C)$ is the conditional entropy of one clustering given the other, and $H(K)$ is the entropy of the first clustering. A completeness score close to 1 indicates that all the data points that are members of a given cluster in one clustering are elements of the same cluster in the other.

$$c = 1 - \frac{H(K|C)}{H(K)} \tag{3.28}$$

- **Adjusted Mutual Information (AMI)**:The Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for chance. It is defined as:

$$AMI(U,V) = \frac{MI(U,V) - E[MI(U,V)]}{\max\{H(U), H(V)\} - E[MI(U,V)]} \tag{3.29}$$

where: - U and V are the two clusterings, - $MI(U,V)$ is the mutual information between U and V, - $E[MI(U,V)]$ is the expected mutual information, - $H(U)$ and $H(V)$ are the entropies of U and V respectively.

An AMI score close to 1 indicates that the two clusterings are very similar.

Mutual Information (MI) between two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" obtained about one random variable, by observing the other random variable. The mutual information between two discrete random variables X and Y can be calculated as follows:

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{3.30}$$

where: - $p(x,y)$ is the joint probability mass function of X and Y, - $p(x)$ and $p(y)$ are the marginal probability mass functions of X and Y respectively.

- **Silhouette**

The silhouette coefficient is a measure used to assess the quality of a clustering. It gives an indication of how well each data point fits into its assigned cluster. The silhouette coefficient for each data point is calculated using the following formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.31}$$

Here:

- $s(i)$ is the silhouette coefficient for data point i. - $a(i)$ is the average distance from data point i to all other data points in the same cluster. This measures the cohesion or how close data point i is to other points in its cluster. - $b(i)$ is the smallest average distance from data point i to all points in any other cluster, of which i is not a member. This measures the separation or how far data point i is from points in other clusters.

CHAPTER 4 DATA ANALYSIS

In this chapter, we will explore the Bixi BSS in detail. The objective of our analysis is to understand the dynamics of BSS. We aim to uncover patterns and trends in the usage of Bixi bikes, and how these correlate with various POIs and weather conditions.

4.1 Bixi data

4.1.1 Distance between each pair of stations

The Figure 4.3 shows that the number of bike-sharing trips reaches a peak during the summer, slowly decreases during the fall, and slowly increases again in the spring. All BIXIs are taken from the network on November 15—the end of the season—and it takes Bixi's employees around two days to put the entire fleet back together. The stations also need to be stored over the winter in order to protect the equipment from snow and ice and to be able to clean the streets of Montreal after a snowfall. Bixi needs a big area that is simple to access for the personnel in charge of maintenance because they are significantly larger than bikes. All the stations with their associated docking ports are stored in Olympic Park's parking garages [77].

For the analysis, we enriched our dataset by generating additional features from the existing ones. This strategy enabled us to uncover deeper insights. We get time-related features such as day, month, year, day of the week, and weekend. We then generated additional features such as the distance between each pair of stations, the number of stations in close proximity to each station, the distance traveled during peak hours, travel time, and so on. Considering that Bixi operates on a seasonal basis, it experiences a higher number of travels during the warmer months, as illustrated in Figure 4.3.

By calculating the distance (Geodesic Distance) for each trip, as explained in 3.2, we can

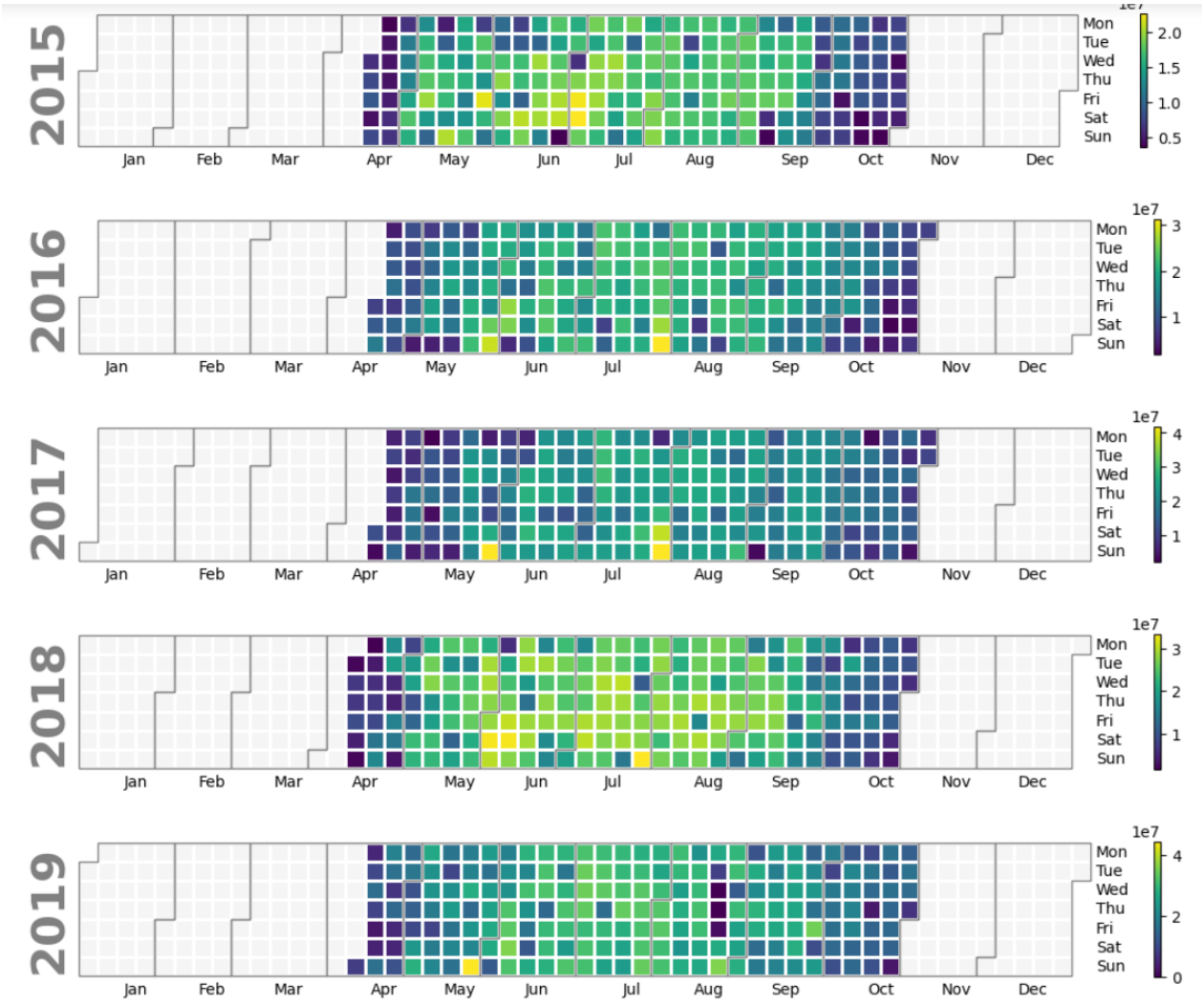


Figure 4.1 heat-map of Total(Sum) travel time per day from 2015 to 2019

better understand these patterns and trends. It allows us to quantify the relationship between travel time and distance, observe how this relationship changes over time, and identify any anomalies or interesting patterns.

The relationship between travel distance and trips is provided in Figure 4.2, which represents Bixi trips in 2019 for different months.

The scatter plots 4.2 illustrate a clear relationship between the distance traveled and the elevation gained during bike-sharing trips. The varying densities of points across the graphs suggest that certain distances are associated with higher elevation gains, which could be indicative of the terrain where the bike-sharing system is used.

As the distance increases, we notice a decrease in the frequency of trips. This suggests that users are less inclined to use bike-sharing for longer distances, highlighting the importance of station placement to optimize usage and efficiency.

Each month's graph provides insights into how the distance-elevation relationship changes with the seasons, potentially reflecting changes in user behavior or environmental factors. While most trips are short, there is a segment of users who consistently travel more than 5 km. This diversity in user behavior adds to the complexity of predicting bike-sharing demand and optimizing station locations.

These insights can inform strategic decisions on station placement, bike-sharing demand prediction, and system expansion to better serve the community's transportation needs. Station-to-station distance and elevation gain emerge as key factors in understanding and optimizing BSS. The scatter plot provides a clear visual representation of these trends and patterns. This analysis can help in making data-driven decisions for the expansion and improvement of bike-sharing services, ensuring they meet the users' needs effectively.

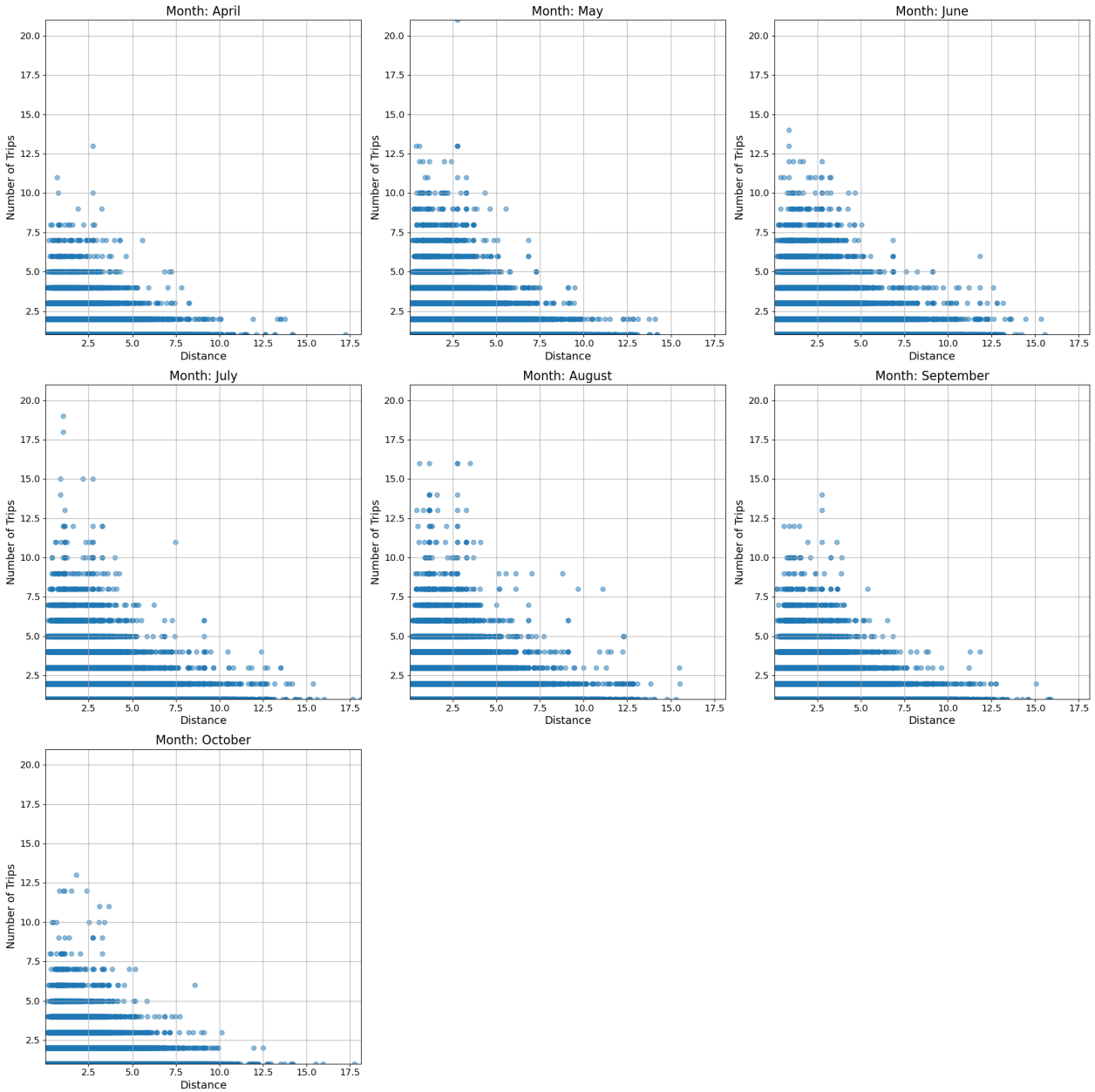


Figure 4.2 Scatter plots of hourly number of trips over distances(km) in 2019 for different months.

4.1.2 Approximate Speed of Trips

In the table 4.1, we have summarized the distribution of trip speeds for the bike-sharing users. In this study, the speed of each trip was calculated based on the direct distances between stations, as detailed in Chapter 3. The table provides a clear breakdown of the percentage of trips that fall within various speed ranges.

From the table, we can observe that the majority of trips, approximately 57.88%, occur at speeds between 0 and 10 km/h. The next significant speed range is 10-20 km/h, accounting for around 39.83% of trips.

As we move to higher speed ranges, the percentage of trips significantly decreases. For instance, only about 0.19% of trips occur at speeds between 20 and 30 km/h. This trend continues with an even smaller percentage of trips occurring at speeds above 30 km/h.

Table 4.1 Percentage of Trips for Different Speed Ranges

Speed Range (km/h)	**Percentage of Trips (%)**
0-10	57.88
10-20	39.83
20-30	0.19
30-35	0.02
35-40	0.01
40-50	0.02
50-60	0.01
60-70	0.005
70-80	0.003
80-90	0.002
90-100	0.001

Continuing from the table analysis, it's important to note that while the majority of trips occur at lower speeds, there are several factors that could contribute to the instances of higher speeds observed in the data:

- **Bike Redistribution**: Bixi redistributes bikes throughout the day to ensure availability where and when they're needed. This redistribution is typically done using trucks,

can cause the error in the starta and end trips locations and as aresult the calculated speed.

- **Indirect Trips**: Users might not always take the most direct route between their start and end points. They could stop for errands, take scenic routes, or visit multiple places during a single trip. This could result in longer trip durations, even if the start and end points are close together.
- **Bike Type**: Bixi offer electric bikes in addition to traditional pedal bikes. Electric bikes can reach higher speeds, especially if the user is also pedaling.
- **User Behavior**: Some users might be professional cyclists or simply enjoy riding at high speeds. Alternatively, users might be rushing to get somewhere quickly.

In our analysis, we considered trips with speeds above 35 km/h as outliers and removed them from the dataset. These outliers represent less than 0.05% of the total trips.

4.1.3 Trips distributions

The number of Bixi stations has increased in the past years, reflecting the growing popularity and expansion of the bike-sharing system in Montreal. Table 4.2 details the year-by-year increase in the number of stations from 2015 to 2020. Figure 4.3 shows the number of travels for each of these years, indicating usage trends.

Table 4.2 Total number of stations each year.

year	number of stations
2018	552
2019	619
2020	641

Moreover, the distribution of Bixi stations across Montreal has evolved from 2015 to 2020, as depicted in Figure 4.4. The maps for each year show the spatial growth of the network,

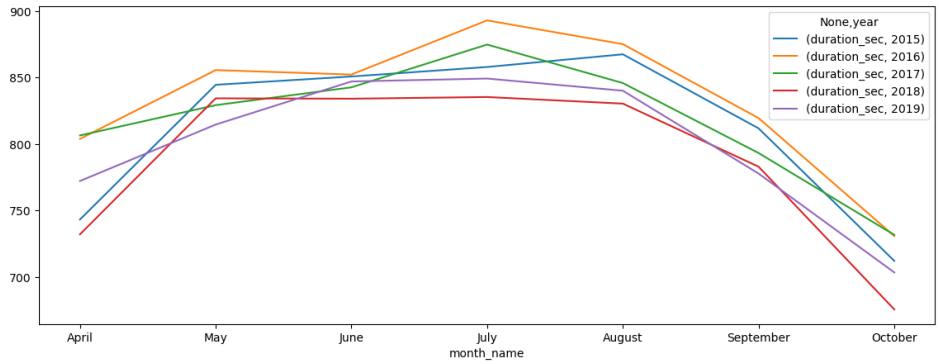


Figure 4.3 Monthly travel time from 2015 to 2020

highlighting areas of new station development.

The number of trips and travel time at each station differ across the days of the week. Specifically, the patterns observed on weekdays are not the same as those on weekends, Figure 4.5.

Table 4.2 shows that the number Bixi station has expanded through the years (more details, and graph in the classification section).The figure 4.8 illustrates the increased spatial coverage of Bixi BSS during the years since its opening Figure 4.8.

The figure 4.6 represents the Bixi trips from 2017 to 2019. The left graph shows the number of trips per year. An upward trend is observed, indicating an increase in the number of trips over the years. The right graph depicts the travel time in seconds for each corresponding year. The consistency in the height of the bars across all five years suggests that the average travel time has remained relatively stable. This data implies that while Bixi ride-sharing usage has increased over the years, the average travel time has not seen a significant change. This could be indicative of enhancements in efficiency or infrastructure, enabling a higher number of trips without an increase in the average travel time.

The heat maps figures 4.7 offer a comprehensive overview of the usage patterns of Bixi bike-sharing system across different hours of the day and days of the week. The top-left heat map, representing the number of trips, reveals a bimodal distribution with peaks during weekday

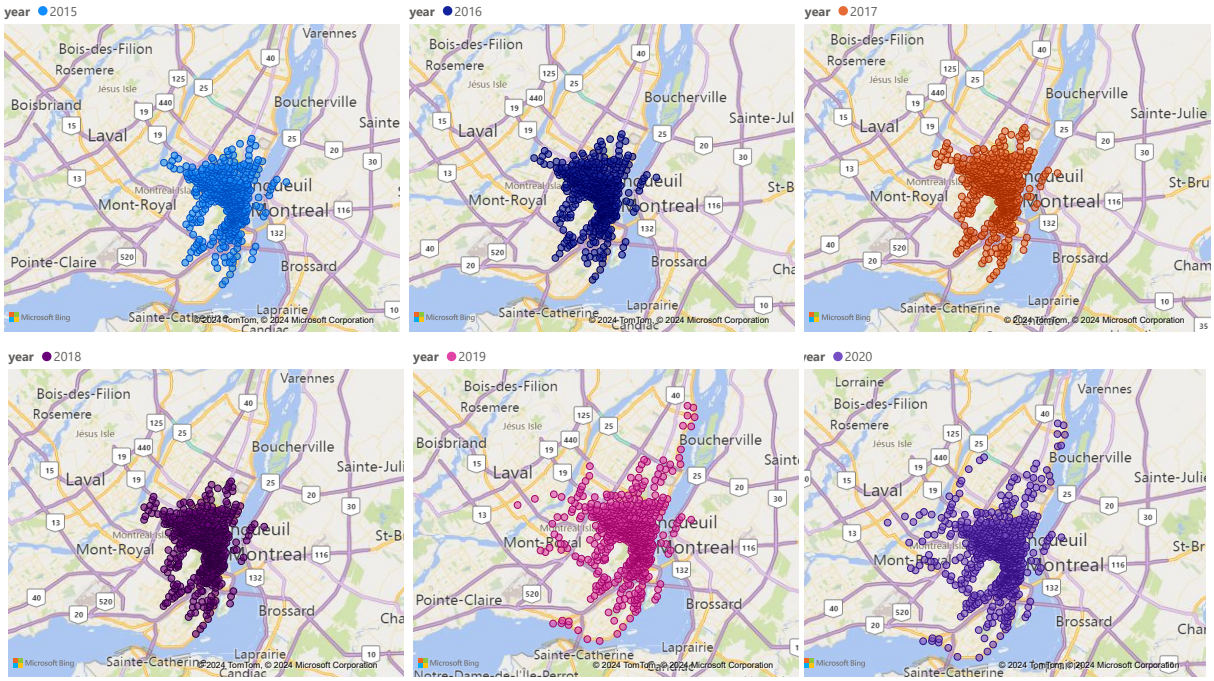


Figure 4.4 Distribution of Bixi stations in Montreal from 2015 to 2020

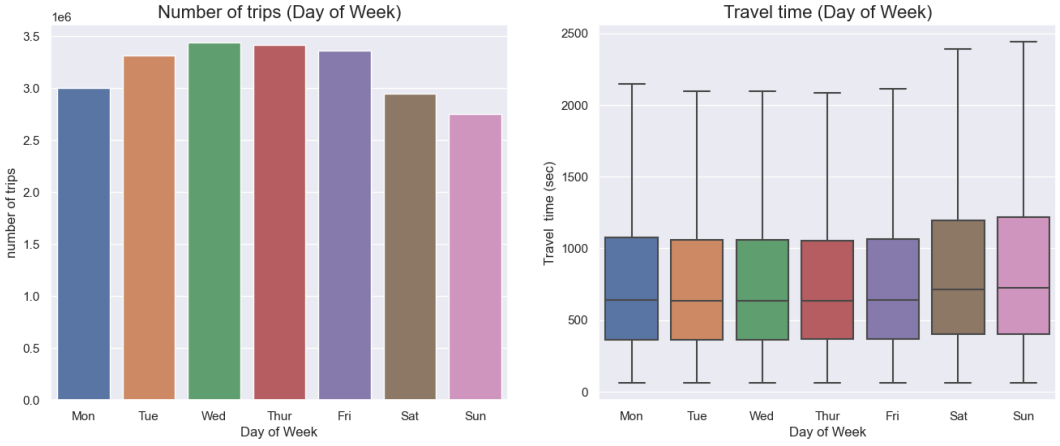


Figure 4.5 Bixi travel time and hourly number of trips for each week over 5 years(2015 to 2019)

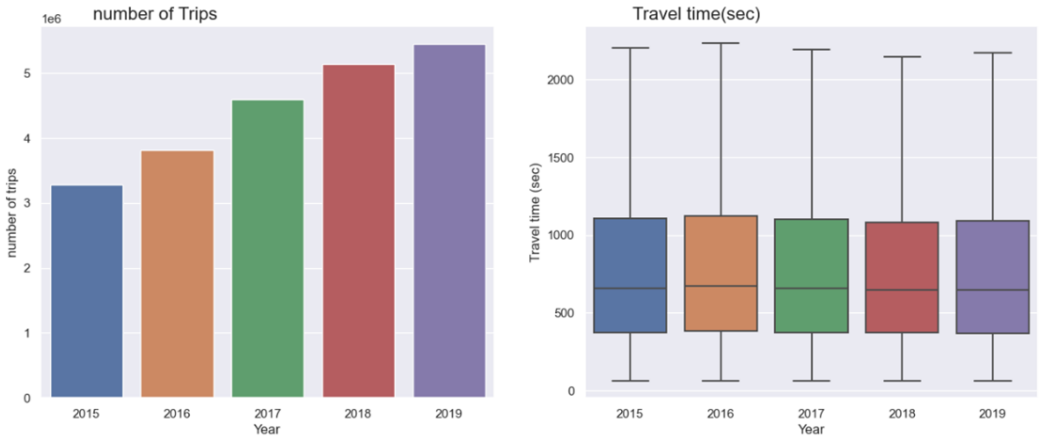


Figure 4.6 Bixi travel time and hourly number of trips for each year from 2015 to 2019)

mornings and evenings. This pattern is indicative of a strong correlation with commuter traffic, suggesting that Bixi bikes are a popular choice for work-related travel. The top-right heat map, which illustrates the distance covered in these trips, shows a different pattern. Longer trips are more frequent in the midday and on weekends. This could be interpreted as a reflection of leisure activities, where users might be taking longer, more scenic routes or using the bikes for sightseeing purposes.

Moving to the bottom-left heat map, it represents the travel time. The longer travel times during rush hours could be attributed to increased traffic congestion. This suggests that despite the convenience of BSS, they are not immune to the effects of heavy traffic.

Lastly, the bottom-right heat map presents the travel speed. Interestingly, higher speeds are achieved during off-peak hours, especially late nights and early mornings. This could be due to less traffic and fewer obstacles, allowing users to travel faster.

These analysis provide valuable insights into the usage patterns of the Bixi BSS. They suggest that Bixi bikes are not only used for commuting but also for leisure activities, with clear patterns that correlate with typical workday schedules and weekend activities. These findings could be instrumental in planning and managing BSS to better serve the needs of the users.

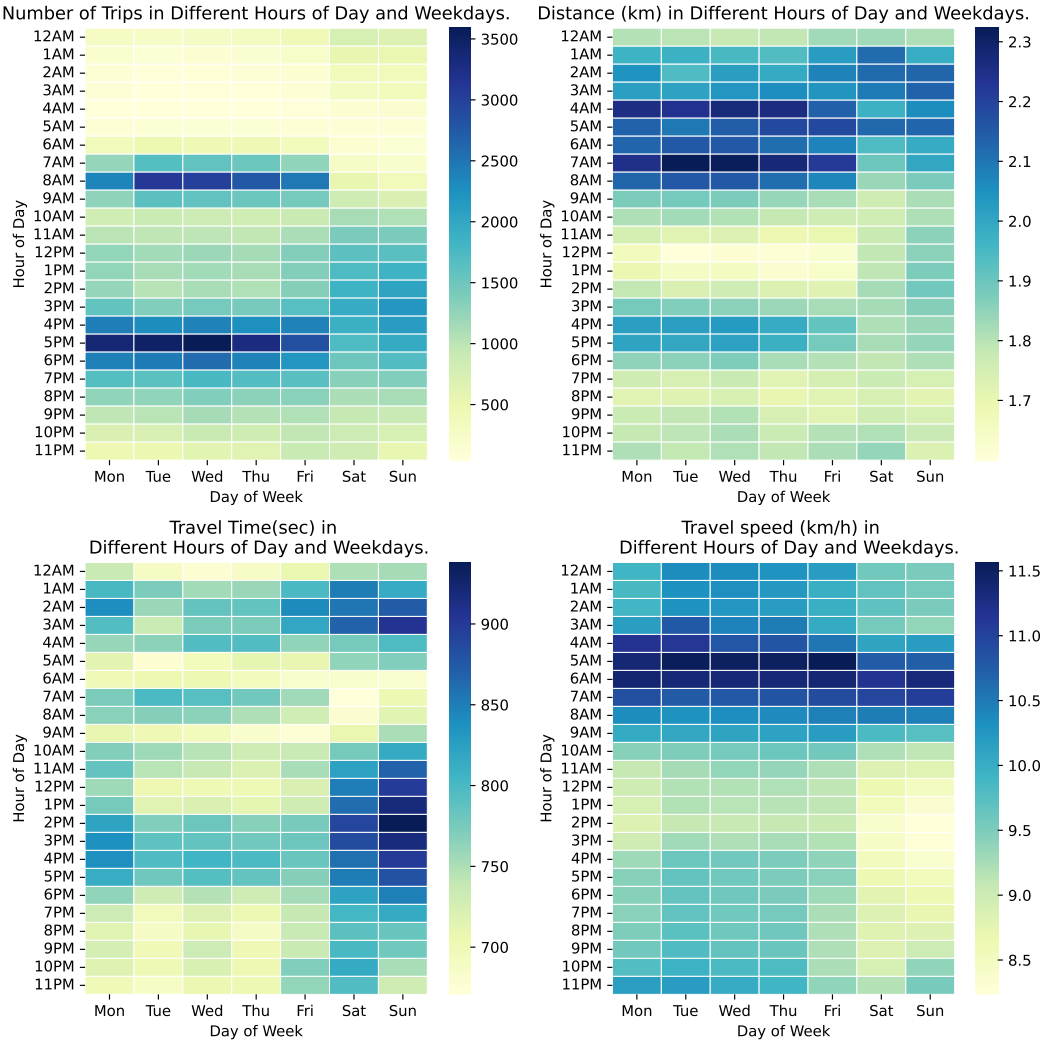


Figure 4.7 Trip Frequency, Distance, Duration, and Speed Across Different Hours and Days of the Week

4.2 Analysis of Bixi Stations' Extensions and Lifespan

In this study, we classified Bixi stations into distinct classes based on their operational lifespan within the BSS from 2015 to 2020. The operational lifespan of a station refers to the number of years it has been in service.

The criteria for a station to belong to a specific class:

- **Class 1**: Stations that appear in the dataset for only one year. This includes stations that were launched in the final year of the study period, 2020.
- **Class 2**: Stations that appear in the dataset for two different years.
- **Class 3**: Stations that appear in the dataset for three different years.
- **Class 4**: Stations that appear in the dataset for four different years.
- **Class 5**: Stations that appear in the dataset for five years.
- **Class 6**: Stations that appear in the dataset for all six years.

To identify whether a station is the same between two successive years, we used the station's unique identifier provided in the dataset. If the identifier of a station appears in the data for two consecutive years, we consider it as the same station.

Table 4.3 provides a summary of the number of stations for each class, while Figure 4.8 illustrates their geographical distribution. This classification methodology helps us understand the patterns associated with the relocation and persistence of stations within the BSS, which is crucial for planning and managing the service.

- **Class 1**: This class, with 22 stations, has an average of 5612 trips per station per hour and an average travel time of 1063.4 minutes per station per hour. Despite their short

lifespan, these stations have a relatively high usage rate. The longer travel time could be attributed to their locations or the specific demand patterns of these stations.

- **Class 2**: Comprising 69 stations, this class has a significantly higher number of trips per station per hour (140352) and a slightly higher average travel time (1199.56 minutes). These stations, operational for two years, are quite popular and heavily used.

- **Class 3**: This class includes 6 stations with 41294 trips per station per hour and an average travel time of 833 minutes. The lower number of trips and shorter travel time suggest these stations primarily serve local or short-distance trips.

- **Class 4**: With 82 stations, this class shows a high number of trips (710786 per station per hour) and an average travel time of 837.87 minutes. These stations, operational for four years, seem to be integral parts of the BSS, serving a large number of trips.

- **Class 5**: This class has 7 stations with 59927 trips per station per hour and an average travel time of 919.71 minutes. Despite the small number of stations, the relatively high number of trips and longer travel time suggest these stations might be located in areas of high demand.

- **Class 6**: Representing the most stable part of the BSS, this class has 457 stations that have been operational for all six years. These stations have the highest number of trips (6952076 per station per hour) and an average travel time of 807.32 minutes, indicating that they efficiently serve a large number of short to medium-length trips.

This analysis further supports the findings of the study. It reveals the dynamic nature of the BSS, where stations can have different lifespans and roles within the system. Understanding these dynamics is crucial for the effective planning and management of the service. It also underscores the importance of considering the specific characteristics and demands of each station when making decisions about station placement and relocation.

Table 4.3 Operational characteristics of Bixi stations classified by lifespan

	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**
Num Trips per Station per Hour	5612	140352	41294	710786	59927	6952076
Avg. Travel Time per Station per Hour (s)	1063.4	1199.56	833.01	837.87	919.71	807.32
Num Stations	22	69	6	82	7	457

Lastly, each class exhibits a unique spatial distribution and coverage. Figure 4.8 presents the location of each station per year, categorized by class. Moreover, the analysis reveals

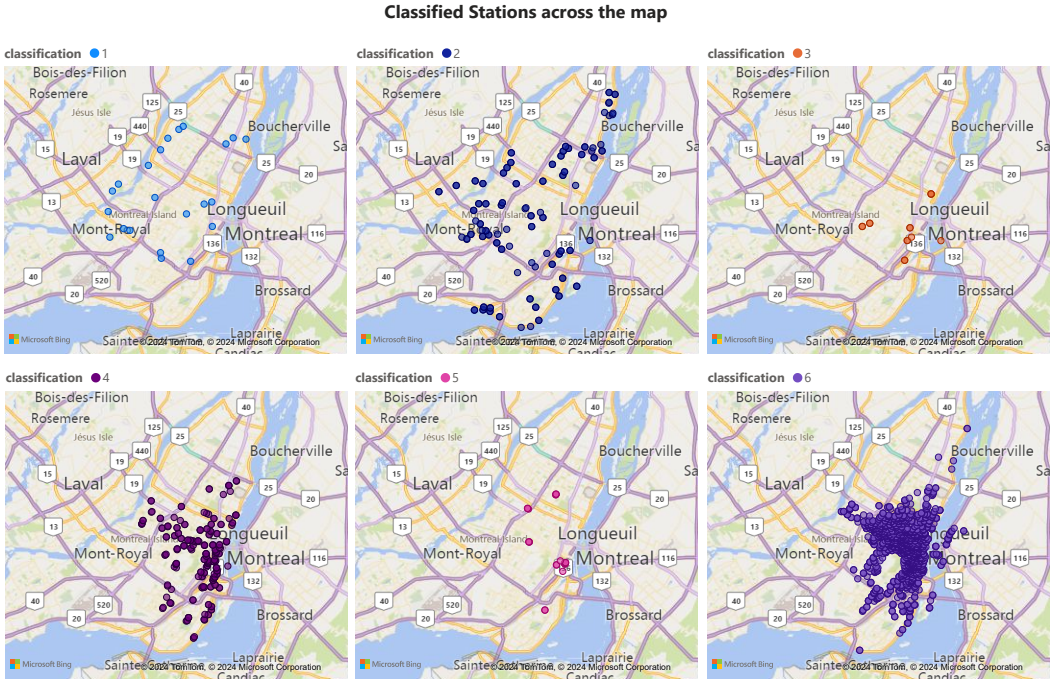


Figure 4.8 Station distribution of each class from 2015 to 2020

the dynamic nature of BSSs, where stations can have different lifespans and roles within the system. This dynamism is crucial to consider when planning for the expansion or contraction of the system, as it affects the system's capacity, coverage, and service quality. The maps illustrate the geographical spread of stations within the city of Montreal, categorized by their operational lifespan. Each classification is represented by a unique color, providing a clear visual differentiation between the classes. This spatial analysis is essential for understand-

ing the accessibility and coverage provided by the BSS across urban areas. It also aids in identifying potential gaps in the service and opportunities for expansion or optimization.

This analysis contributes to our understanding of the dynamics of BSSs and provides valuable insights for their planning and management. It underscores the importance of considering the specific characteristics and demands of each station when making operational decisions. This updated analysis includes the consideration that Class 1 stations could also include those launched in the final year of the study period, 2020, adding another layer of understanding to the study.

4.3 Points of interest

In this study, we incorporate Points of Interest (POI) data [65] to examine their influence on the usage patterns of Bixi stations. A POI is defined by its geographic coordinates and belongs to one or more categories. These categories, represented by the variable "Famille" in the POI data, provide information about the surrounding areas of each station. They are categorized as Cultural, Commercial, Recreational/Sports, and Public Service. We postulate that these categories may influence the demand for bike-sharing in diverse ways, contingent upon the users' objectives and preferences [78, 79].

The proximity of Points of Interest (POIs) was calculated using the Haversine distance formula. If the distance between a station and a POI is less than or equal to 450 meters, the POI is considered to be within the station's neighborhood zone. This choice of 450 meters as the threshold distance is supported by a comprehensive review of the literature and sensitivity analyses [78, 79]. It is noteworthy that a single POI may fall within the zones of multiple stations, and conversely, each station may be associated with multiple POIs within its neighborhood zone. This forms a many-to-many association between stations and POIs, based on a threshold distance approach [78, 79].

The number of POIs within a distance δ of each bike-sharing station, denoted as P_s, is computed using the following equation:

$$P_s = \sum_{p=1}^{M} I(d(s,p) \leq \delta) \tag{4.1}$$

Here, M represents the number of POIs, $d(s,p)$ is the distance between station s and POI p, and I is an indicator function that yields 1 if its argument is true, and 0 otherwise.

We chose four distinct types of POIs within the city. Certain POIs, such as public services, cultural, and sports facilities, exhibit a dispersed geographical distribution throughout the city. Conversely, commercial POIs are predominantly concentrated in the downtown area. Therefore, it is crucial to investigate how these diverse types of POIs influence the number of trips within the BSS.

Each POI category (Cultural, Sports, Commercial, Public Service) encompasses various types of interest points. Table 3.3 provides a summary of the POIs, along with their frequency distributions categorized by type.

4.4 Weather

This section is dedicated to exploring the relationship between weather conditions and bike-sharing patterns. The goal is to gain a deeper comprehension of the extent to which weather variables influence the utilization of the BSS. By examining the specific weather conditions and their degree of impact on BSS usage. In the end, we will select the proper features for analysis. A feature matrix has been created from the weather data set. At first, we had to deal with the missing values in some features such as weather, wind speed, temperature, visibility, pressure, humidity and wind chill, Precipitation, wind speed, real humidity, etc. since the number of missing values was small, we impute the missing values with nearest values. Features with more than 80 percent of missing values were removed from the matrix.

The weather feature had text values: rain, shower, snow, ice, cloudy, storm, sunny, clear, thunderstorm, shower fog, fog, etc. By using one hot encoding we transformed this feature into binary features. Weather conditions strongly impact bike-sharing trips. Therefore, the impact of weather conditions on BSS trip data was taken into consideration.

We performed some data preprocessing steps before building our models. We imputed the missing values with the nearest values, as they were few in number. We also dropped the features that had more than 80% of missing values. For the weather feature, which had categorical values such as rain, snow, fog, etc., we encoded them using one-hot encoding.

Higher temperatures are correlated with an increase in the number of bike-sharing trips 4.1. Figure 4.9 shows a clear trend where the total number (sum) of bike-sharing trips is significantly higher during cloudy weather, with nearly 800000 trips. This is followed by clear weather with around 600000 trips, indicating a preference for bike-sharing during different weather conditions. Fog and rain lead to a noticeable decrease in trips, with fog at approximately 100000 and rain at about 350000 trips, suggesting that adverse weather conditions have a deterrent effect on bike-sharing usage. Extreme weather conditions such as snow and thunderstorms result in the fewest number of trips, barely above zero for snow. This highlights the significant effect of different weather conditions on bike-sharing system (BSS) trips.

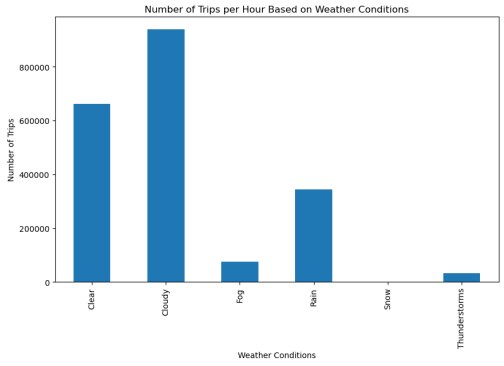


Figure 4.9 Portion of Departure in different weather conditions

CHAPTER 5 INFLUENCE OF WEATHER AND PROXIMITY FACTORS ON BIKE SHARING DEMAND

This section aims to examine how weather conditions, POIs, number of nearby stations, and temporal factors affect the demand for bike-sharing. A linear regression model was fitted, using the number of hourly trips as the response variable and the predictor variables related to the weather, the POIs, and the time of the trip.

Table 5.1 presents the descriptive statistics of the variables used in our regression model. These variables include environmental factors such as visibility, precipitation amount, and wind speed, as well as infrastructural factors like the number of nearby stations and the presence of cultural, sports, public service, and commercial facilities.

Table 5.1 Descriptive Statistics of Variables Used in the Regression Analysis

	mean	std	25%	50%	75%	max
Visibility (km)	51	20	40	40	72	72
Precip. Amount (mm)	0	1	0	0	0	28
Wind Spd (km/h)	14	7	9	13	17	45
Number of nearby stations	5	3	2	5	7	17
Number of cultural POI	5	7	0	2	6	39
Number of sports POI	5	5	2	4	7	41
Number of public service POI	6	8	2	4	7	45
Number of commercial POI	4	8	0	0	2	43
Number of hourly trips	4	4	1	2	4	157

The Figure 5.1 presents a Correlation Matrix Heatmap. This heatmap shows the correlation between various weather conditions, temporal factors, and other conditions with the number of trips.

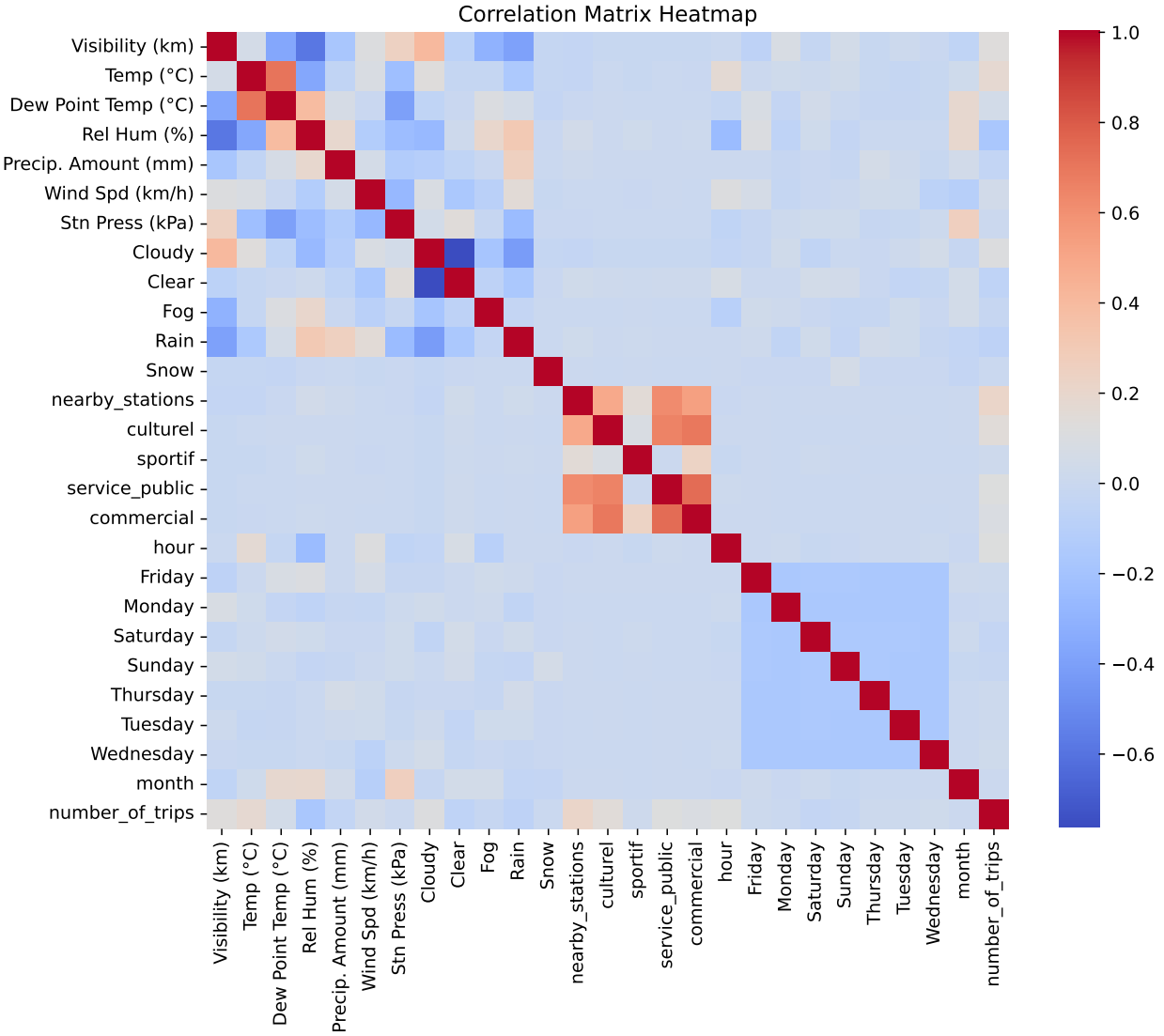


Figure 5.1 Correlation Heatmap of Weather, Temporal Factors, and Trip Numbers

This study aims to examine how weather conditions, POIs, the number of nearby stations, and temporal factors affect the demand for bike-sharing. A linear regression model was fitted, using the number of hourly trips as the response variable and the predictor variables related to the weather, the POIs, and the time of the trip.

To ensure the robustness of the results, the dataset was split into two subsets: a calibration (training) set and a validation (test) set. The calibration set, comprising 70% of the data, was used to estimate the model parameters, while the remaining 30% served as the validation set to assess the model's predictive accuracy. Highly correlated features were removed from this analysis to reduce multicollinearity and improve the model's performance.

The model's performance was further validated by calculating key metrics such as the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) on both the training and test sets. The similarity in these metrics across the two sets indicates that the model maintains consistent predictive accuracy and does not overfit the calibration data.

The model has two reference categories, 'Cloudy' for weather conditions and 'Wednesday' for days of the week. Detailed results are presented in Table 5.2. This model provides an understanding of the factors influencing bike-sharing demand by allowing for a nuanced interpretation of the coefficients. The coefficients in the model represent the change in the number of bike trips associated with a one-unit increase in the corresponding variable, holding all other variables constant. For binary variables, the coefficient represents the change in the response variable for a change from 0 to 1.

The results of the regression analysis provide valuable insights into the determinants of bike-sharing demand. The proximity to POIs and the density of nearby stations significantly influence the number of bike trips, underscoring the importance of strategic station placement for Bixi.

Weather conditions and temporal factors also play a critical role in shaping demand. Favor-

Table 5.2 Regression Results with Wednesday and Cloudy as Reference Categories

	coef	std err	t	P>|t|
const	0.0101	7.36e-05	137.650	0.000
Precip. Amount (mm)	-0.0414	0.001	-40.992	0.000
nearby_stations	0.0275	0.000	190.656	0.000
number of nearby cultural POIs	0.0133	0.000	75.736	0.000
number of nearby sport-related POIs	-0.0008	0.000	-4.209	0.000
number of nearby public_service POIs	-0.0079	0.000	-34.030	0.000
number of commercial POIs	-0.0073	0.000	-34.597	0.000
Clear	-0.0056	5.48e-05	-101.847	0.000
Fog	-0.0028	0.000	-15.903	0.000
Rain	-0.0071	8.56e-05	-83.377	0.000
Snow	-0.0085	0.001	-7.555	0.000
hour_sin	-0.0054	3.29e-05	-164.905	0.000
month_sin	-0.0016	3.77e-05	-42.442	0.000
Friday	-0.0003	8.43e-05	-3.755	0.000
Monday	-0.0015	8.37e-05	-17.941	0.000
Saturday	-0.0032	8.48e-05	-37.513	0.000
Sunday	-0.0027	8.46e-05	-31.684	0.000
Thursday	6.69e-07	8.41e-05	0.008	0.994
Tuesday	-0.0001	8.38e-05	-1.781	0.075

able weather and optimal visibility encourage increased bike trips, while unfavorable conditions such as high humidity and precipitation deter usage.

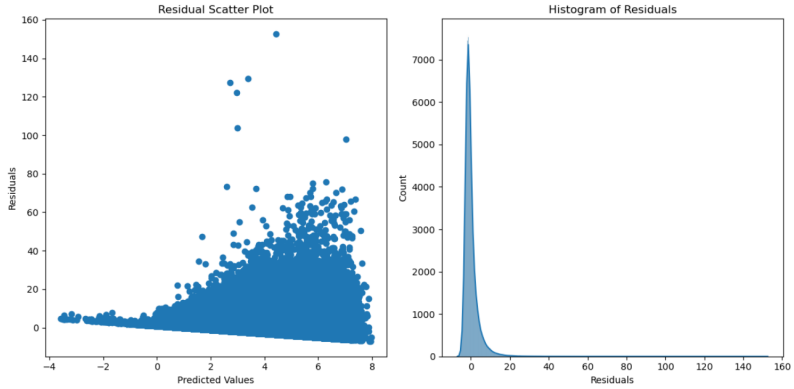
The influence of cultural POIs suggests that areas with cultural attractions may encourage recreational bike trips. Conversely, proximity to public services appears to have a negative impact on demand, potentially due to the availability of alternative transportation options or the nature of trips to these destinations.

This analysis highlights the multifaceted determinants of Bixi demand. By considering various reference categories, we have gained a nuanced understanding of how different factors influence bike-sharing usage, providing instrumental insights for the strategic development and optimization of Bixi services. The residual scatter plot (Figure 5.2) illustrates the residuals (the differences between observed and predicted values) as a function of the predicted values. In this plot, the residuals are partly centered around zero, which is a good sign.

However, there is a large increase in the spread of residuals as the predicted values increase, indicating some heteroscedasticity. This means that the variance of the residuals is not consistent across all predicted values. The spread of points above the zero line indicates variability in residuals as predicted values increase, which might suggest some deviations from normality.

The histogram of residuals (Figure 5.2) shows a sharp peak near zero, indicating that most residuals are small. However, the quick tapering off and the presence of a few large residuals suggest that there are some deviations from normality. These large residuals may need further investigation to determine if they are outliers.

Overall, while the residuals are mostly centered around zero, the increasing spread and presence of outliers suggest that there is room for improvement if future considerations involve demand prediction. Further investigation and refinement of the model are needed to address these issues and enhance its predictive accuracy.



(a) Residual Scatter Plot

Figure 5.2 Residual Analysis

Future research should look closely at each important variable. Using more data, like the number of bikes and docks, and other factors related to land use, could give us more useful insights.

Table 5.3 Summary of Model Metrics

Metric	**Value**
F-statistic	($p < 0.05$)
Prob (F-statistic)	0.00
Log-Likelihood	2.4817e+06
RMSE (Root Mean Squared Error)	0.0234
MAE (Mean Absolute Error)	0.0148
Residual Standard Error (RSE)	0.024

5.1 Statistical significance of the Model

The following is a detailed explanation of the key statistics derived from the regression model used to analyze Bixi BSS demand:

- **F-statistic**: This statistic tests the overall significance of the model. The null hypothesis is that all the regression coefficients are equal to zero. In this model, the F-statistic is significant ($p < 0.05$), indicating that the predictors as a set meaningfully contribute to the model and it is statistically significant.

- **Prob (F-statistic)**: This statistic tests the overall significance of the model. The null hypothesis is that all the regression coefficients are equal to zero. In this model, the Prob (F-statistic) is 0.00, indicating that the predictors as a set meaningfully contribute to the model and it is statistically significant. This means that at least one of the predictors' regression coefficient is not equal to zero in the model, hence the model is significant.

- **Log-Likelihood**: This statistic measures the log of the likelihood that the model would produce the observed values of the dependent variable. The higher the log-likelihood, the better the model is at predicting the observed values. In this model, the log-likelihood is approximately 2.4817e+06.

- **RMSE (Root Mean Squared Error)**: This statistic measures the average magni-

tude of the residuals or prediction errors. In this model, the RMSE is approximately 0.0234, indicating a good fit of the model.

- **MAE (Mean Absolute Error)**: This statistic measures the average magnitude of the residuals or prediction errors, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. In this model, the MAE is approximately 0.0148, indicating a good fit of the model.

- Residual Standard Error (RSE) In the analysis of the linear regression model, RSE was computed to assess the quality of the model fit. The RSE provides a measure of the standard deviation of the residuals, indicating the typical difference between the observed and predicted values. The computed RSE was found to be 0.024. This relatively small value suggests that the model provides a good fit to the data, capturing the underlying relationship effectively.

The regression analysis has identified precipitation amount, nearby stations, cultural points of interest, sports facilities, public services, commercial points of interest, weather conditions (clear, fog, rain, snow), time of day (hour_sin), time of year (month_sin), and days of the week (Friday, Monday, Saturday, Sunday, Thursday, Tuesday) as significant variables influencing Bixi bike-sharing demand. The analysis also suggests that cultural POIs positively affect demand, whereas proximity to public services and commercial POIs has a negative effect. It also points to further areas of study, including the use of more data and the effects of choosing different reference categories.

CHAPTER 6 RESULT AND DISCUSSION

6.1 Cluster Analysis

Clustering is a technique that groups data points based on their similarity and dissimilarity. It can help us analyze the travel behavior of the users and characteristics of bike-sharing stations in Montreal. We conducted bike-sharing station clustering using variables related to travel (such as trip duration, distance, Hourly number of trips, peak morning trips, peak afternoon trips, weekend trips, etc.) and POI such as the number of commercial attractions, public service, sports, and cultural point of interests. This way, we can discover different types of bike-sharing stations and their spatial distribution in a city. This can provide insights into the patterns and functions of bike sharing in Montréal, and support the planning and management of BSSs. For example, we can adjust bike-sharing stations' location, size, and service based on their cluster characteristics and needs. We can also design customized marketing strategies and incentives for different types of bike-sharing users based on their travel preferences and goals.

We performed two kinds of clustering analysis:

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**:
  - **Purpose**: This method is used for grouping the locations by their spatial proximity.
  - **Technique**: DBSCAN identifies clusters based on the density of data points. It groups together points that are closely packed together.
  - **Advantages**: DBSCAN can find arbitrarily shaped clusters and is robust to noise (outliers). It does not require the number of clusters to be specified beforehand.
  - **Application**: In our study, DBSCAN helps in identifying clusters of BSS based on

their geographical locations, which can reveal spatial patterns and distributions.

- **K-means Clustering**:
  - **Purpose**: This method is used for grouping the stations by their travel-related variables and points of interest.
  - **Technique**: K-means clustering partitions the data into K clusters, where each data point belongs to the cluster with the nearest mean. The algorithm iteratively adjusts the cluster centroids to minimize the variance within each cluster.
  - **Advantages**: K-means is efficient and works well with large datasets. It is easy to implement and interpret.
  - **Application**: In our study, K-means clustering helps in identifying clusters of BSS based on various attributes such as trip duration, number of trips, and points of interest. This can provide insights into the different types of stations and their usage patterns.
- **Affinity Propagation**:
  - **Purpose**: This method is used for grouping the stations based on their travel-related variables and points of interest.
  - **Technique**: Affinity Propagation identifies exemplars among the data points and forms clusters by sending messages between data points until a high-quality set of exemplars and corresponding clusters emerges.
  - **Advantages**: Affinity Propagation does not require the number of clusters to be specified beforehand and can handle large datasets efficiently.
  - **Application**: In our study, Affinity Propagation helps in identifying clusters of BSS based on various attributes such as trip duration, number of trips, and points

of interest, similar to K-means clustering. This provides additional insights into the different types of stations and their usage patterns.

By using DBSCAN, K-means clustering, and Affinity Propagation, we can gain a comprehensive understanding of the spatial and functional characteristics of BSS in Montreal. This multi-faceted approach allows us to tailor our strategies for station management and user engagement more effectively.

6.1.1 Dbscan result

In this section of the thesis, we present the results of the DBSCAN clustering analysis conducted on stations for the year 2019 based on their geographic coordinates.

The variables considered in this analysis show aspects such as the distance, members' percentage, peak morning and afternoon distances, travel times during different periods of the day and week, the number of trips, and the percentage of trips over the weekend. We also consider the cultural, sports, public service, and commercial characteristics of the stations.

This analysis allows us to capture a holistic view of the station clusters, revealing patterns and trends that would not be apparent if the variables were considered in isolation. The detailed analysis that follows will delve into the specifics of these patterns, providing a granular view of the station clusters and their defining characteristics. This preliminary explanation aims to provide a roadmap for the detailed exploration that follows, setting the stage for a deep and nuanced understanding of the station clusters and their implications.

- Cluster 0: This cluster has 236 stations. The high members percentage (82.87) indicates that these stations are frequently used by members. The high number of close POIs for cultural (11.48), sports (7.17), public service (12.47), and commercial (9.46) suggests these stations might be in close proximity to these types of locations. The number of trips per hour is 4.24, with peak morning trips at 4.38 and peak afternoon trips at 7.20.

Table 6.1 Multivariate Characteristics of Station Clusters based on the Spatial Clustering

Cluster	0	1	2	3	4	5	6
number_of_stations	236	15	18	23	16	8	303
distance	1.74	1.57	1.92	2.08	2.19	1.99	2.51
PTM	82.87	82.03	87.02	84.03	87.66	74.77	80.40
peak_morning_distance	1.70	1.46	2.27	2.54	2.71	2.09	2.87
peak_afternoon_distance	1.81	1.68	1.79	2.02	2.00	1.96	2.42
peak_morning_travel_time_hour	0.18	0.16	0.23	0.23	0.25	0.22	0.27
peak_afternoon_travel_time_hour	0.22	0.22	0.21	0.22	0.22	0.26	0.27
WPA(TT)_hour	0.24	0.25	0.23	0.25	0.24	0.31	0.30
travel_time_hour	0.21	0.20	0.22	0.22	0.23	0.26	0.28
PTT	26.55	29.41	26.95	29.45	28.23	35.55	28.60
PTMW	70.24	68.43	79.33	75.28	80.35	65.04	70.02
number_of_trips	4.24	2.60	2.90	3.48	2.40	3.15	2.22
peak_morning_trips	4.38	3.46	4.39	3.40	2.93	2.61	2.49
peak_afternoon_trips	7.20	3.18	3.59	5.64	3.06	4.45	2.86
weekend_peak_afternoon_trips	5.30	3.26	3.32	5.07	2.95	6.15	2.63
PT	24.48	25.12	26.26	27.64	27.39	32.46	26.94
culturel	11.48	2.93	2.06	2.13	0.69	4.62	1.30
sportif	7.17	11.33	7.00	5.39	4.06	12.00	4.06
service_public	12.47	3.47	5.72	3.78	3.12	6.75	3.07
commercial	9.46	0.60	0.72	1.00	0.00	0.75	0.27

- Cluster 1: This cluster has 15 stations with a high score for sports locations (11.33). The high percentage of travel time during the weekend (29.41) indicates these stations are particularly busy during weekends. The number of trips per hour is 2.60, with peak morning trips at 3.46 and peak afternoon trips at 3.18.

- Cluster 2: This cluster has 18 stations with a high members percentage (87.02). The high weekend peak afternoon travel time hour (0.23) could indicate these stations are particularly busy during weekend afternoons. The relatively high score for sports locations (7.00) could suggest these stations are located near sports facilities or areas popular for outdoor activities. The number of trips per hour is 2.90, with peak morning trips at 4.39 and peak afternoon trips at 3.59.

- Cluster 3: This cluster has 23 stations with a high members percentage (84.03). The high peak morning distance (2.54) and peak afternoon distance (2.02) suggest these

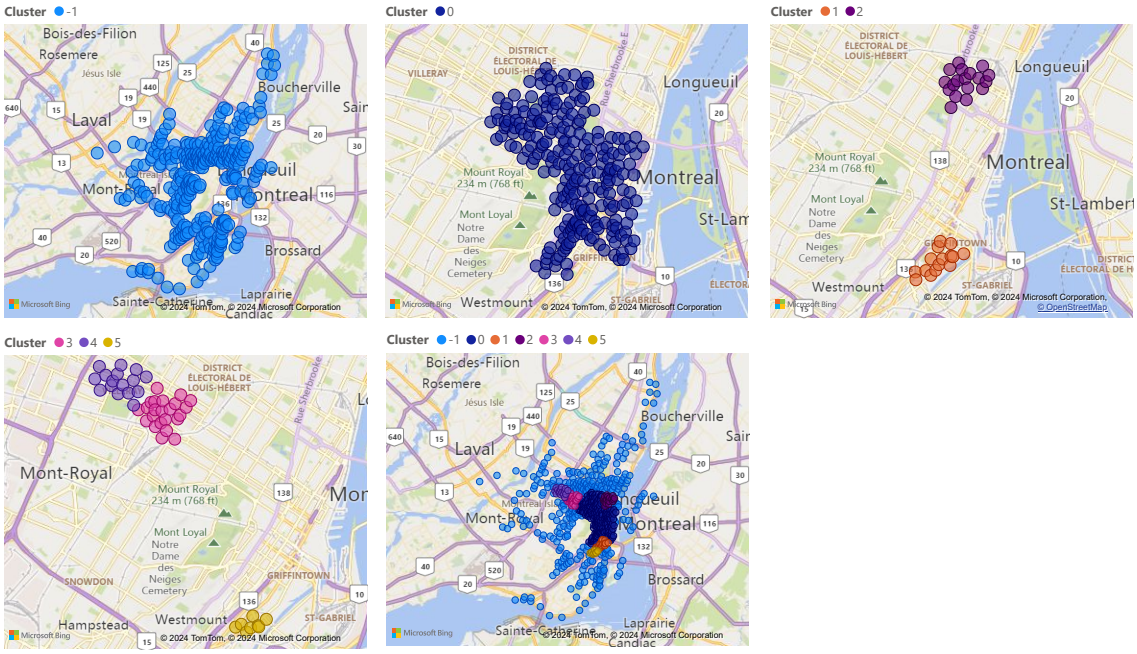


Figure 6.1 Density-based clustering of stations location

stations are used for longer trips, particularly during peak hours. The relatively high score for sports locations (5.39) could suggest these stations are located near sports facilities or areas popular for outdoor activities. The number of trips per hour is 3.48, with peak morning trips at 3.40 and peak afternoon trips at 5.64.

- Cluster 4: This cluster has 16 stations but the highest members percentage (87.66). The high percentage of travel time during the weekend (28.23) and the high percentage of members during the weekend (80.35) suggest these stations are located in areas popular during the weekends. The number of trips per hour is 2.40, with peak morning trips at 2.93 and peak afternoon trips at 3.06.

- Cluster 5: This cluster has 8 stations with the highest percentage of travel time during the weekend (35.55). The high score for sports locations (12.00) further supports this, suggesting these stations are located in areas popular for sports or recreational activi-

ties. The number of trips per hour is 3.15, with peak morning trips at 2.61 and peak afternoon trips at 4.45.

- Cluster 6: This cluster is the largest with 303 stations. The average distance of trips is 2.51. The high POI scores for cultural (1.30) and sports (4.06) locations suggest these stations are located in areas with a high density of these POIs. The number of trips per hour is 2.22, with peak morning trips at 2.49 and peak afternoon trips at 2.86.

We utilized DBSCAN to perform spatial clustering of BIXI stations based on their geographical coordinates. Spatial clustering is a powerful tool for analyzing the geographical distribution of BIXI stations. This location-based analysis can significantly contribute to BIXI's mission of providing an efficient and convenient transportation option for Montreal's residents by enhancing service efficiency, identifying potential areas for expansion, and informing strategic decisions.

- Identifying Areas for Expansion: The spatial distribution of the clusters can help BIXI identify potential areas for expansion. For example, the group (6) with 303 stations has an average distance of 2.51, which is highest among other clusters.

- Enhancing Service Efficiency: The average travel time during peak hours varies across the clusters, suggesting different usage patterns. For instance, cluster 5, which has only 8 stations, shows the second highest average travel time during peak hours (0.26 hours). This could indicate a high demand for BIXI services in this area, and resources could be reallocated to improve service efficiency.

- Informing Strategic Planning: The clusters also vary in terms of cultural, sports, public service, and commercial POIs. These POIs could provide valuable insights for strategic planning. For example, cluster 0, which has 236 stations, shows high values in all these categories. This suggests that these areas are popular destinations for BIXI users and could be prioritized in strategic planning.

6.1.2 Affinity Propagation

In our study, we conducted a comprehensive evaluation of the dataset using the AP clustering algorithm. We systematically calculated the silhouette scores for a range of preference values to determine the optimal number of clusters. Our analysis, which was corroborated by graphical representations and our domain expertise, revealed that a seven-cluster solution provided the most meaningful and coherent grouping of the data.

Table 6.2 Average Affinity Propagation results

Affinity propagation	0	1	2	3	4	5	6
number_of_stations	68	20	51	63	7	92	318
distance	3.13	3.64	2.28	1.86	2.19	1.71	2.00
PTM	80.01	52.25	70.82	84.31	51.48	82.78	85.74
peak_morning_distance	3.29	4.61	2.35	1.45	2.03	1.82	2.34
peak_afternoon_distance	3.17	3.53	2.31	2.13	2.25	1.73	1.88
peak_morning_travel_time_hour	0.31	0.44	0.24	0.16	0.26	0.19	0.23
peak_afternoon_travel_time_hour	0.34	0.41	0.30	0.24	0.34	0.21	0.22
WPA(TT)_hour	0.35	0.45	0.34	0.26	0.39	0.23	0.24
travel_time_hour	0.33	0.42	0.29	0.22	0.34	0.20	0.22
PTT	23.95	46.82	36.47	19.23	44.17	29.65	27.04
PTMW	67.40	44.80	55.97	64.58	39.83	72.70	76.78
number_of_trips	1.70	1.87	2.36	4.06	7.91	5.44	2.59
peak_morning_trips	1.53	1.22	1.98	2.92	3.29	5.89	3.42
peak_afternoon_trips	2.20	2.12	3.19	8.46	14.69	9.18	3.32
weekend_peak_afternoon_trips	1.77	2.72	3.59	3.54	19.07	7.75	2.97
PT	23.12	45.20	32.36	17.12	40.48	27.42	25.51
Number of POIs culturel	0.62	0.90	6.06	17.22	16.57	7.42	3.23
Number of POIs sportif	2.26	2.50	4.16	3.52	5.14	7.47	6.74
Number of POIs service_public	2.63	1.50	5.41	21.49	5.57	7.91	5.08
Number of POIs commercial	0.16	0.20	3.39	13.83	5.29	5.74	2.33

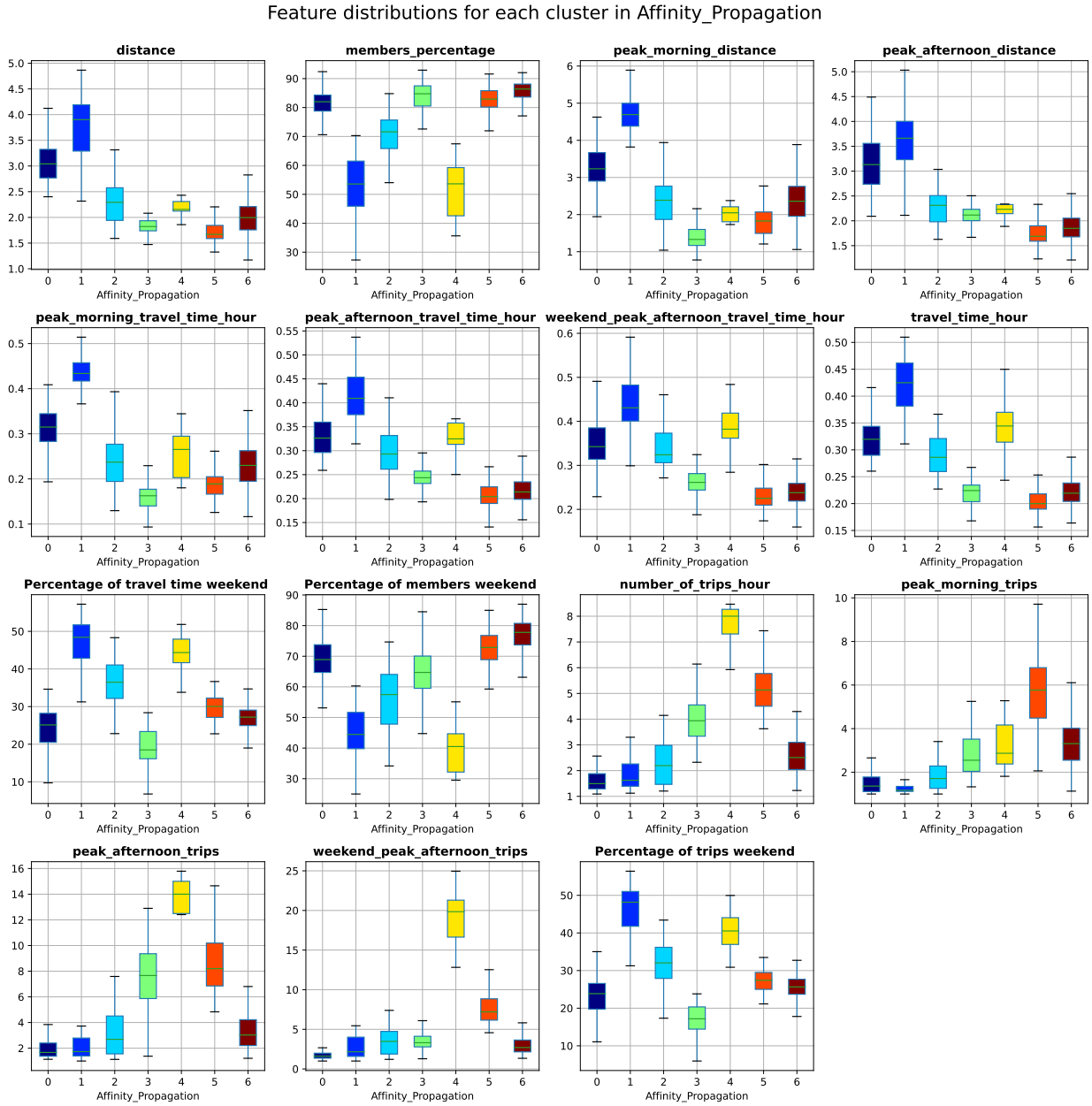


Figure 6.2 Affinity propagation clustering of stations

Table 6.3 Standard deviation values of BSS stations clustering using Affinity Propagation

Affinity propagation	0	1	2	3	4	5	6
number_of_stations	68	20	51	63	7	92	318
distance	0.48	0.93	0.39	0.25	0.20	0.19	0.33
PTM	8.68	11.33	7.41	4.90	11.76	4.24	3.33
peak_morning_distance	0.98	1.19	0.75	0.46	0.26	0.37	0.57
peak_afternoon_distance	0.57	0.96	0.39	0.25	0.24	0.23	0.31
peak_morning_travel_time_hour	0.07	0.04	0.06	0.04	0.06	0.03	0.05
peak_afternoon_travel_time_hour	0.06	0.07	0.04	0.03	0.06	0.03	0.03
weekend_peak_afternoon_travel_time_hour	0.06	0.09	0.05	0.04	0.06	0.03	0.03
travel_time_hour	0.05	0.05	0.04	0.02	0.06	0.02	0.03
PTT	5.74	6.63	5.35	4.59	6.07	3.97	3.29
PTMW	12.33	8.77	10.32	8.72	9.30	6.75	5.56
number_of_trips	0.60	0.66	0.93	1.34	1.36	1.40	0.68
peak_morning_trips	0.58	0.18	0.90	1.35	1.32	2.02	1.17
peak_afternoon_trips	1.31	0.92	1.74	4.10	5.34	3.68	1.34
weekend_peak_afternoon_trips	0.61	1.41	2.08	1.26	4.27	2.24	1.05
PT	5.22	9.09	5.36	4.17	6.43	3.67	3.03
Number of POIs culturel	1.08	1.48	11.84	11.90	13.21	7.88	5.28
Number of POIssportif	1.86	2.14	3.66	2.89	3.39	4.94	6.33
Number of POIs service_public	2.23	1.54	9.09	16.67	5.03	6.85	4.32
Number of POIs commercial	0.41	0.41	7.99	12.11	7.06	10.68	7.10

The table 6.2 shows the average values of various variables for each cluster of stations obtained by applying the AP clustering algorithm to the data of Bixi, a BSS in Montreal. The clusters are as follows:

- **Cluster 0** This cluster has 68 stations, with an average distance of 3.13. The number of trips in this cluster is 1.70. This relatively low number of trips might suggest that these stations are in less frequented areas or are used less frequently by the members. The stations are located in various areas in Montreal such as North St. Michel, Nicholas Ville, Hamstead, Jian Rosemount, District Electrical D Leuis Real, and areas around the river near HaShaga. This also includes areas above Park de Montreal, between Mount Royal and Park du mont royal. This suggests that these stations might be used for recreational activities such as cycling along the river, or for commuting from these neighborhoods into the city. The majority of trips in this cluster are made by members,

as indicated by the high membership percentage of 80.01%. The average travel time is 0.33 hours, with slightly longer travel times during peak hours. The percentage of travel time during the weekend is 23.95%, suggesting that the majority of travel time (76.05%) is during the weekdays. The POI data shows a higher number of sports-related and public service POIs, which could indicate that these stations are located near parks, recreational areas, or public service buildings.

- **Cluster 1** This cluster consists of 20 stations with an average distance of 3.64. The number of trips in this cluster is 1.87, which is slightly higher than Cluster 0, suggesting a slightly higher usage. These stations are mostly located sparsely in the southern part of the island of Montreal, around the border of the island, particularly in the areas of Lassalle and Lachine near the river. This suggests that these stations might serve both residential and commercial areas, possibly for commuting from these neighborhoods into the city. The membership percentage is lower at 52.25%, suggesting a more balanced usage between members and casual users. The average travel time is 0.42 hours, with longer travel times during peak hours. Interestingly, this cluster has a higher percentage of travel time during the weekend (46.82%), indicating a higher weekend usage compared to other clusters. The POI data shows a balanced distribution across all categories, suggesting a diverse range of attractions in these areas.

- **Cluster 2** This cluster comprises 51 stations with an average distance of 2.28. The number of trips in this cluster is 2.36, which is higher than the previous clusters, indicating a higher demand for the service in these areas. The stations are densely located in Plateau Mont-Royal and downtown. This suggests a high demand for the service in these areas, possibly for commuting to work or school. The membership percentage is 70.82%, indicating a high usage among members. The average travel time is 0.29 hours, suggesting that users might be using the service for short commutes within the city center. The percentage of travel time during the weekend is 36.47%,

indicating a balanced usage between weekdays and weekends. The POI data shows a high number of cultural POIs, indicating a possible interest in cultural activities among the users.

- **Cluster 3** Cluster 3 includes 63 stations with the shortest average distance of 1.86. The number of trips in this cluster is 4.06, which is significantly higher than the previous clusters, suggesting a high demand for the service in these areas. The stations are densely located in Plateau Mont-Royal and downtown mostly. This suggests a high demand for the service in these areas, possibly for commuting to work or school. The membership percentage is the highest among all clusters at 84.31%, suggesting a high demand for the service among members. The average travel time is 0.22 hours, the shortest among all clusters, supporting this inference. The percentage of travel time during the weekend is the lowest among all clusters at 19.23%, indicating the highest weekday usage. The POI data shows a high number of cultural and public service POIs, indicating a rich cultural scene and well-established public service infrastructure.

- **Cluster 4** This cluster, with the smallest number of stations at 7, has an average distance of 2.19. The number of trips in this cluster is 7.91, which is the highest among all clusters. This suggests that these stations, despite being fewer in number, are in high-demand areas. The stations are located in the east of Montreal, from Jarry Park to downtown, and from Parc Mont Royal to Avenue Papineau. This suggests that these stations might serve a mix of residential, commercial, and recreational areas. The membership percentage is the lowest among all clusters at 51.48%, suggesting a more balanced usage between members and casual users. The average travel time is 0.34 hours, with longer travel times during peak hours. The percentage of travel time during the weekend is the highest among all clusters at 44.17%, indicating the highest weekend usage. The POI data shows a high number of cultural POIs, suggesting a rich cultural scene in these areas.

- **Cluster 5** Cluster 5 has the second-highest number of stations at 92 and the lowest average distance of 1.71. The number of trips in this cluster is 5.44, which is also quite high, indicating a high demand for the service in these areas. The stations are located in the east of Montreal, from Jarry Park to downtown, and from Parc Mont Royal to Avenue Papineau. This suggests that these stations might serve a mix of residential, commercial, and recreational areas. The membership percentage is high at 82.78%, suggesting a high usage among members. The average travel time is 0.20 hours, the shortest among all clusters, showing that users might be using the service for short trips within these areas. The percentage of travel time during the weekend is 29.65%, indicating a balanced usage between weekdays and weekends. The POI data shows a balanced distribution across all categories, indicating a mix of commuting and leisure-related trips.

- **Cluster 6** This cluster has the highest number of stations at 318 and an average distance of 2.00. The number of trips in this cluster is 2.59, which, despite being lower than some of the smaller clusters, is still significant given the large number of stations. This suggests a high overall usage of the service in these areas. The stations are widely distributed in the same area as Cluster 5 but extend further to the north and east of Montreal, as well as in Verdun and St. Gabriel. This suggests that these stations might serve a wide range of areas, including both urban and suburban regions. The membership percentage is the highest among all clusters at 85.74%, suggesting a high usage among members. The average travel time is 0.22 hours, with slightly longer travel times during peak hours. The percentage of travel time during the weekend is 27.04%, indicating a balanced usage between weekdays and weekends. The POI data shows a higher number of sports-related POIs, which indicate that these stations are located near parks or recreational areas.

6.1.3 k-means result

In this section analysis using the k-means clustering algorithm. We systematically calculated the silhouette scores for a range of k values to determine the optimal number of clusters. Our analysis, which was corroborated by graphical representations and our domain expertise, revealed that a seven-cluster solution provided the most meaningful and coherent grouping of the data.

Table 6.4 Average K_means clustering results

K-means Clustering	0	1	2	3	4	5	6
number_of_stations	106	26	43	75	7	97	265
distance	2.85	3.63	2.20	1.82	2.19	1.73	1.95
PTM	82.10	53.93	69.65	83.86	51.48	83.08	86.19
peak_morning_distance	3.14	4.21	2.19	1.42	2.03	1.87	2.31
peak_afternoon_distance	2.83	3.58	2.27	2.07	2.25	1.72	1.82
peak_morning_travel_time	0.29	0.40	0.23	0.16	0.26	0.19	0.23
peak_afternoon_travel_time	0.30	0.43	0.29	0.24	0.34	0.21	0.21
WPA(TT)	0.33	0.44	0.34	0.26	0.39	0.23	0.24
travel_time	0.30	0.42	0.29	0.22	0.34	0.20	0.22
PTT	24.75	42.33	37.35	20.22	44.17	29.70	27.30
PTMW	70.27	43.86	55.02	64.86	39.83	73.37	77.62
number_of_trips	1.84	1.72	2.48	4.05	7.91	5.31	2.60
peak_morning_trips	1.89	1.14	1.97	3.04	3.29	6.04	3.45
peak_afternoon_trips	2.31	1.94	3.43	8.23	14.69	8.75	3.32
weekend_peak_afternoon_trips	1.94	2.44	3.86	3.69	19.07	7.57	2.98
PT	23.56	41.05	33.09	17.98	40.48	27.57	25.78
Number of POIs culturel	0.71	0.73	7.28	17.36	16.57	6.69	3.04
Number of POIs sportif	2.72	2.31	4.33	4.09	5.14	7.54	7.15
Number of POIs service_public	2.70	1.81	6.60	20.89	5.57	6.98	4.98
Number of POIs commercial	0.17	0.19	4.65	14.28	5.29	4.77	2.16

The table 6.4 shows the average values of various variables for each cluster of stations obtained by applying the k-means clustering algorithm to the data of Bixi, a BSS in Montreal. The clusters are as follows:

- **Cluster 0**: This cluster has 106 stations sparsely located in areas such as North Saint-Michel, Parc Nicolas-Viel, Hampstead, Rosemont, Louis-Riel district, and areas around the river near Parc du Mont-Royal, between Mount Royal and Park du mont royal. The

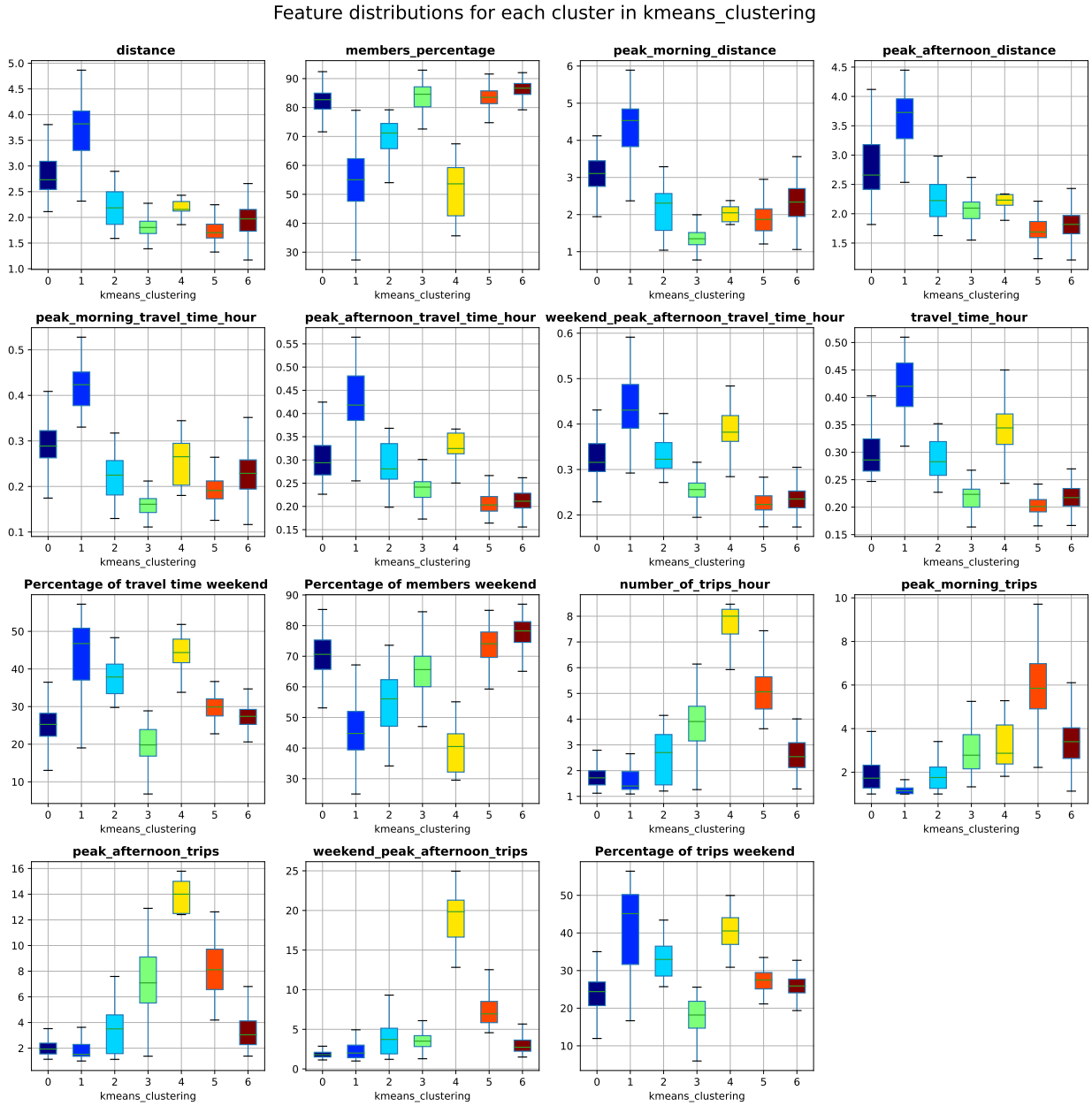


Figure 6.3 k-means clustering of stations

Table 6.5 Standard deviation values of BSS stations clustering using K-means Clustering

K-means Clustering	0	1	2	3	4	5	6
number_of_stations	106	26	43	75	7	97	265
distance	0.47	0.84	0.37	0.23	0.20	0.19	0.29
PTM	4.47	12.59	7.10	4.93	11.76	4.08	3.01
peak_morning_distance	0.78	1.46	0.73	0.39	0.26	0.39	0.53
peak_afternoon_distance	0.56	0.87	0.39	0.25	0.24	0.22	0.26
peak_morning_travel_time_hour	0.06	0.11	0.06	0.03	0.06	0.03	0.04
peak_afternoon_travel_time_hour	0.04	0.07	0.04	0.03	0.06	0.03	0.03
WPA(TT)_hour	0.05	0.09	0.05	0.04	0.06	0.03	0.03
travel_time_hour	0.04	0.05	0.04	0.02	0.06	0.02	0.02
PTT	5.08	6.63	4.83	4.88	6.07	3.67	3.21
PTMW	7.04	8.77	10.13	8.38	9.30	6.55	5.03
number_of_trips	0.58	0.64	0.98	1.34	1.36	1.35	0.61
peak_morning_trips	0.76	0.29	0.89	1.30	1.32	1.89	1.05
peak_afternoon_trips	1.22	0.87	1.85	4.31	5.34	3.40	1.26
weekend_peak_afternoon_trips	0.72	1.34	2.16	1.42	4.27	2.25	0.96
PT	4.73	11.61	4.98	4.41	6.43	3.42	2.89
culturel	1.14	1.34	12.64	11.88	13.21	7.55	4.27
sportif	2.29	2.04	3.73	3.65	3.39	4.79	6.65
service_public	2.12	2.21	10.34	16.02	5.03	4.85	3.83
commercial	0.40	0.40	7.99	12.11	7.06	10.04	6.72

average distance is 2.85. The percentage of members is 82.10%. The average distance traveled during peak morning hours is 3.14, while it's 2.83 during peak afternoon hours. The average travel time during these peak hours is 0.29 hours in the morning and 0.30 hours in the afternoon. The percentage of travel time during the weekend is 24.75%. This cluster has relatively low POIs across all categories, suggesting that these stations are not in close proximity to many points of interest. This could indicate that the stations in this cluster are primarily used for commuting rather than leisure or tourism.

- **Cluster 1**: This cluster consists of 26 stations located sparsely around the Montreal island border . The average distance is 3.63. The percentage of members is 53.93%. The average distance traveled during peak morning hours is 4.21, while it's 3.58 during peak afternoon hours. The average travel time during these peak hours is 0.40 hours in the morning and 0.43 hours in the afternoon. The percentage of travel time during the

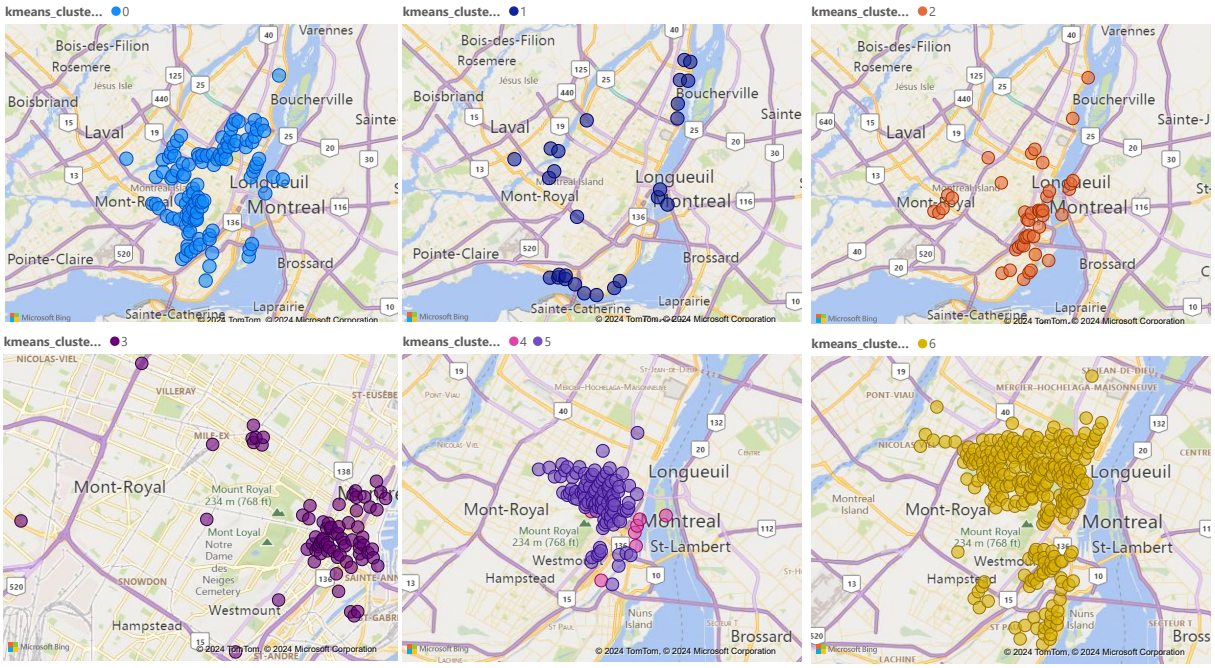


Figure 6.4 k-means clustering of stations location for 2019

weekend is 42.33%. This cluster has a relatively low number of cultural and commercial POIs, but a higher number of sports-related POIs. This indicates that these stations are located near recreational areas and are used more for leisure activities, especially on weekends.

- **Cluster 2**: This cluster comprises 43 stations mostly located in the southern part of downtown Montreal. The average distance is 2.20. The percentage of members is 69.65%. The average distance traveled during peak morning hours is 2.19, while it's 2.27 during peak afternoon hours. The average travel time during these peak hours is 0.23 hours in the morning and 0.29 hours in the afternoon. The percentage of travel time during the weekend is 37.35%. This cluster has a relatively high number of cultural and commercial POIs, demonstrating that these stations are located in a vibrant area with many attractions. This could explain the balanced usage between weekdays and weekends, as these stations might be used both for commuting and leisure activities.

- **Cluster 3**: This cluster includes 75 stations densely located in Plateau Mont-Royal and downtown.The average distance is 1.82. The percentage of members is high at 83.86%. The number of trips in this cluster is 4.05. This high number of trips suggests that these stations are in high-demand areas, possibly for commuting to work or school. The percentage of members is high at 83.86%. The average distance traveled during peak morning hours is 1.42, while it's 2.07 during peak afternoon hours. The average travel time during these peak hours is 0.16 hours in the morning and 0.24 hours in the afternoon. The percentage of travel time during the weekend is 20.22%. This cluster has the highest number of cultural and public service POIs among all clusters, indicating that these stations are located in a culturally rich area with many public services. This could explain the high weekday usage, as these stations might be used primarily for commuting to work or school.

- **Cluster 4**: This cluster, with the smallest number of stations at 7, is located sparsely around the Montreal border near the river. The average distance is 2.19. The percentage of members is the lowest among all clusters at 51.48%. The number of trips in this cluster is 7.91. Despite having the smallest number of stations, this cluster has the highest number of trips, suggesting that these stations are in high-demand areas. The average distance traveled during peak morning hours is 2.03, while it's 2.25 during peak afternoon hours. The average travel time during these peak hours is 0.26 hours in the morning and 0.39 hours in the afternoon. The percentage of travel time during the weekend is the highest among all clusters at 44.17%. This cluster has a relatively high number of cultural POIs and the highest weekend usage, suggesting that these stations are located near tourist attractions and are used more for leisure activities.

- **Cluster 5**: This cluster has 97 stations located in the east of Montreal, from Jarry Park to downtown, and from Parc Mont Royal to Avenue Papineau. The average distance is 1.73. The percentage of members is 83.08%. The number of trips in this cluster is 5.31.

This high number of trips indicates a high demand for the service in these areas. The percentage of members is 83.08%. The average distance traveled during peak morning hours is 1.87, while it's the lowest during peak afternoon hours at 1.72. The average travel time during these peak hours is 0.19 hours in the morning and 0.21 hours in the afternoon. The percentage of travel time during the weekend is 29.70%. This cluster has a balanced number of POIs across all categories, suggesting that these stations are located in a diverse area with a mix of cultural, sports-related, public service, and commercial points of interest. This could explain the balanced usage between weekdays and weekends, as these stations might be used for a variety of purposes.

- **Cluster 6**: This is the largest cluster with 265 stations widely distributed in the same area as Cluster 5 but extend further to the north and east of Montreal, as well as in Verdun and St. Garei. The average distance is 1.95.The average distance is 1.95. The percentage of members is very high at 86.19%. The number of trips in this cluster is 2.60. Despite being the largest cluster, the number of trips is not the highest, which suggests that the usage of the service is spread out across the many stations in this cluster. The average distance traveled during peak morning hours is 2.31, while it's 1.82 during peak afternoon hours. The average travel time during these peak hours is 0.23 hours in the morning and 0.21 hours in the afternoon. The percentage of travel time during the weekend is 27.30%. This cluster has a relatively low number of cultural POIs (3.04), but a higher number of sports-related POIs (7.15), showing that these stations are located near areas suitable for sports and recreational activities. The number of public service POIs is also relatively high (4.98), indicating the presence of services like tourist information, emergency services, health services, government services, municipal services, transport, and schools. The number of commercial POIs is low (2.16), suggesting that these stations are not in close proximity to many commercial points of interest like tourist attractions, buildings and places of interest, convention/exhibition

centers, accommodation, and groupings of shops.

6.1.4 Evaluation of results of K-means clustering and affinity

This section presents a comparative analysis of two clustering methods: Affinity Propagation and K-Means. These methods were employed to group BSS based on their usage and characteristics. The table 6.6 shows the results of this comparison. Each metric score signifies the degree of similarity between the clusters formed by both methods. For instance, an Adjusted Rand Index score of 0.74 implies a high similarity between the clusters produced by Affinity Propagation and K-Means. The scores presented in the table shows the consistency between

Table 6.6 Evaluation Metrics for clustering

Metric	**Score**
Adjusted Rand Index	0.74
Homogeneity	0.79
Completeness	0.74
V-measure	0.76
Adjusted Mutual Information	0.76

Affinity Propagation and K-Means. The high scores across all metrics suggest that both methods are categorizing the Bike Sharing Stations in a similar fashion.

Confusion Matrix

Table 6.7 presents a detailed comparison of the clustering results from the K-means and Affinity Propagation methods. Each cell in the matrix shows the number of stations assigned to the respective clusters by each method, highlighting both agreements and discrepancies. We renamed the clusters based on their similarity to each other in different methods(AP and k-means) to facilitate easier comparison.

The diagonal entries of the matrix represent instances where both methods agreed on the clustering. High values along the diagonal (e.g., 63, 20, 42, 61, 7, 89, 265) indicate signif-

Table 6.7 Confusion Matrix of Clustering Methods (K-means and AP)

	K-means Clusters						
Affinity Propagation Clusters	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Cluster 0	63	5	0	0	0	0	0
Cluster 1	0	20	0	0	0	0	0
Cluster 2	8	1	42	0	0	0	0
Cluster 3	2	0	0	61	0	0	0
Cluster 4	0	0	0	0	7	0	0
Cluster 5	0	0	1	2	0	89	0
Cluster 6	33	0	0	12	0	8	265

icant agreement between the K-means and Affinity Propagation methods, suggesting that both methods capture similar underlying structures in the data. For example, Cluster 0 in Affinity Propagation aligns well with Cluster 0 in K-means, with 63 instances agreeing on the clustering.

However, the off-diagonal entries (e.g., 5, 8, 1, 2, 33, 12, 8) represent instances where the methods disagreed. These discrepancies can be attributed to the inherent differences in the clustering algorithms. K-means relies on centroid-based clustering, while Affinity Propagation uses message passing between data points. These fundamental differences can lead to variations in cluster assignments. For instance, 33 instances from Cluster 6 in Affinity Propagation were assigned to Cluster 0 in K-means, indicating a disagreement.

The discrepancies also highlight the sensitivity of the clustering methods to their respective parameters. Small changes in parameters, such as the number of clusters for K-means or the preference parameter for Affinity Propagation, can result in different clustering outcomes.

The nature of the data, including the presence of noise, outliers, or varying cluster densities, can influence the clustering results. The confusion matrix reveal that different methods may respond differently to these data characteristics. For instance, the presence of 12 instances from Cluster 6 in Affinity Propagation being assigned to Cluster 3 in K-means indicates that the data characteristics might be affecting the clustering results.

Cluster 0 in K-means has 106 stations, while Cluster 0 in Affinity Propagation has 68 stations. 63 stations from Cluster 0 in K-means are also in Cluster 0 in Affinity Propagation, while 33 stations from Cluster 0 in K-means are in Cluster 6 in Affinity Propagation. Additionally, 33 stations from K-means Cluster 0, 12 from Cluster 3, and 8 from Cluster 5 are assigned to Cluster 6 in Affinity Propagation. This suggests that Cluster 6 in Affinity Propagation is more heterogeneous and includes stations with varying characteristics.

The following details provide insights into the travel times, distances, and number of trips for stations in different clusters. These characteristics help explain why certain stations are assigned to different clusters by the K-means and Affinity Propagation methods. For example, Stations in Cluster 0 for K-means have an average travel time of 0.30 hours and a distance of 2.85 km, while stations in Cluster 0 for Affinity Propagation have an average travel time of 0.33 hours and a distance of 3.13 km. Both clusters exhibit relatively short travel times, indicating a similarity in this aspect. However, the distance for Cluster 0 in Affinity Propagation is longer than that for Cluster 0 in K-means, highlighting a dissimilarity.

In terms of the number of trips, stations in Cluster 0 for K-means have an average of 1.84 trips, whereas stations in Cluster 0 for Affinity Propagation have an average of 1.70 trips. Both clusters have a moderate number of trips, showing a similarity. However, Cluster 0 in K-means has a slightly higher average number of trips compared to Cluster 0 in Affinity Propagation, indicating a difference in this characteristic.

The discrepancies between the clustering methods can be attributed to the inherent differences in the algorithms and the specific characteristics of the stations. The confusion matrix reveal a complex relationship between the clusters identified by the different methods. While there is significant agreement, some discrepancies also exist.

6.2 Discussion

Different analytical methods provide varied insights, and the selection of the appropriate method depends on the specific objectives of the analysis. For instance:

- **Descriptive Statistics and Visualization:** These methods are used initially to get an overall sense of the data, identifying key trends such as peak usage times. Visualization techniques help in presenting these findings in an easily interpretable format.
- **Regression Analysis:** This method is employed when the goal is to understand the relationships between demand and influencing factors such as weather, time of day, or proximity to points of interest (POIs). For example, regression can quantify how much temperature or proximity to a park influences the number of trips.
- **Cluster Analysis (DBSCAN, Affinity Propagation, K-means):** Clustering is particularly useful for segmenting stations based on usage patterns, which helps in tailoring strategies to specific clusters. Each clustering method has its strengths:
  - **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** This method is ideal for identifying clusters of stations based on geographical proximity and density, revealing how station location impacts usage. It's particularly useful in spatial analyses, helping Bixi identify areas where demand is naturally clustered.
  - **K-means Clustering and Affinity Propagation:** These methods group stations by similar characteristics, like the type and frequency of trips. This method is useful for clustering stations based on user features, such as membership status and usage frequency, allowing Bixi to understand and cater to different user segments. It's helpful for strategic planning, such as identifying underperforming stations or those with unique growth potential.

For instance, grouping stations by similar characteristics can help in identifying stations that are heavily used for commuting during peak hours, stations that are popular for leisure activities during weekends, or stations that are underutilized. These insights can guide a range of strategic decisions, from resource allocation and service planning to marketing and expansion. For example, Bixi can ensure adequate bike availability at stations with high demand, tailor marketing strategies to the interests and activities of users in different clusters, or identify promising areas for expansion. In essence, the cluster analysis provides an understanding of the current usage patterns and offers valuable insights that can help shape future strategies, ultimately leading to improved service and customer satisfaction. The main findings and recommendations of the analysis are presented in the following subsections.

6.2.1 Station Expansion

The cluster analysis provides valuable insights that can guide Bixi's station expansion strategy. By understanding the unique characteristics and needs of each cluster, Bixi can make informed decisions that enhance their service and cater to the needs of different user segments.

Cluster 4, which consists of only 7 stations, might require different strategies. This cluster has a balanced usage between members and casual users, and a higher percentage of travel time during the weekend, indicating a higher weekend usage compared to other clusters. Therefore, instead of adding more stations, it might be more beneficial to increase the number of bikes at each station to cater to the weekend crowd. Interestingly, despite having the smallest number of stations, this cluster has the highest number of trips (7.91) among all clusters, indicating a high utilization of the service.

6.2.2 Expansion Planning

The cluster analysis not only provides valuable insights for Bixi's current operations but also offers a data-driven approach to expansion planning. By identifying the characteristics of

successful group and applying these insights to potential expansion areas, Bixi can make informed decisions that maximize the chances of success in new locations.

For instance, Cluster 3, which includes 63 stations densely located in Plateau Mont-Royal and downtown, stands out as a particularly successful cluster. The success of this cluster could be attributed to several factors. The high number of cultural and public service Points of Interest (POIs) suggests that these stations are located in culturally rich areas with many public services. These areas likely attract a large number of both residents and visitors, leading to high demand for bike-sharing services. Moreover, the high percentage of members in this cluster indicates a strong user base that regularly uses the service, contributing to its success. The number of trips in this cluster is 4.06, which is relatively high, further indicating the success of this cluster.

Cluster 6, the largest cluster with 318 stations, also has a high percentage of members. This cluster could provide valuable insights for expansion, especially in areas with a high number of cultural, sports, public service, and commercial Points of Interest (POIs). The number of trips in this cluster is 2.59, which is relatively low considering the size of the cluster, indicating potential for growth.

6.2.3 Resource Allocation

The cluster analysis provides detailed information about the usage patterns of different clusters, which can inform decisions about resource allocation. For instance, Cluster 0 and Cluster 2 have a high percentage of members and high usage during peak hours. This suggests that these stations are heavily used by members, possibly for commuting. In such cases, Bixi can ensure that there are enough bikes available in these areas during these times to meet the demand and provide a reliable service. This can lead to improved customer satisfaction and retention. On the other hand, clusters with longer travel distances, like Cluster 1, might need more durable bikes to withstand the wear and tear of longer trips.

6.2.4 Service Improvement

If a cluster has a low percentage of members, it might indicate that the service is not meeting the needs of potential users in these areas. For example, Cluster 4, which has the smallest number of stations and a lower percentage of members among all clusters, might be under-utilized. This cluster has a higher percentage of travel time during the weekend, indicating a higher weekend usage. Despite the high number of trips (7.91), the highest among all clusters, the membership is low in these areas, indicating a potential area for improvement. Bixi could conduct surveys or other forms of research to find out why the membership is low in these areas. Based on the findings, Bixi could take measures to increase membership in these areas, such as relocating stations, increasing marketing efforts, or offering special promotions targeted at weekend users and tourists.

6.2.5 Marketing and Promotion

The cluster analysis can also assist in targeted marketing and promotion. For example, Cluster 1, which has a higher percentage of travel time during the weekend and a significant number of trips (3.14), might be ideal for promoting Bixi's service for recreational activities. Given the high number of trips in this cluster, Bixi could run campaigns highlighting how their service can be used to reach parks, sports facilities, or other recreational areas. They could also offer special promotions during major sporting events. Additionally, considering the longer travel distances in this cluster, Bixi could promote the use of their service for longer, scenic bike rides in these areas.

Cluster 4, despite having the smallest number of stations, has the highest number of trips (7.91) among all clusters. This indicates that the service is highly utilized in these areas, especially during the weekends. Bixi could leverage this information to run targeted marketing campaigns in these areas, promoting the benefits of their service for weekend recreational activities.

Cluster 5, with 92 stations, a high percentage of members, and a substantial number of trips (5.44), shows a high demand for bike-sharing services. This cluster could be targeted for promotions aimed at regular users. Special offers or loyalty programs could be introduced to further encourage usage and retain members in these areas.

Cluster 6, being the largest cluster with 318 stations, could benefit from broad-based marketing campaigns that highlight the wide availability and convenience of Bixi's service across a large number of stations.

Overall, this cluster analysis provides a comprehensive and nuanced understanding of the usage patterns of the Bixi BSS in Montreal. It allows for evidence-based decision-making, which can lead to more efficient resource allocation, more effective marketing strategies, and ultimately, a better user experience. This understanding can also guide the system, ensuring that it aligns with existing usage patterns and user needs.

CHAPTER 7 CONCLUSION

In this study, we adopted a data-driven approach to understand user behavior and station characteristics in the Bixi BSS in Montreal, Canada. Utilizing open data and creating almost all variables from historical Bixi data allowed us to gain a comprehensive understanding of the system.

By employing methods such as regression, visualization, and clustering, we identified key trends and factors influencing the number of trips at each station. Factors such as weather, time, location, and land use were found to significantly impact the demand for Bixi.

Through cluster analysis, we grouped stations based on user and travel features. We examined their spatial distribution and density using DBSCAN, which provided insights into the diversity and complexity of Bixi stations.

Our approach underscores the potential of using open data for in-depth analysis and decision-making. It highlights the importance of data-driven strategies in optimizing BSSs and catering to the diverse needs of users.

Ultimately, this comprehensive analysis can inform various aspects of business operations, leading to improved service, increased usage, and higher profitability. By ensuring that bikes are available when and where they are most needed, the efficiency of the system can be improved, leading to greater user satisfaction.

The examination of Bixi stations' extensions and lifespan has provided valuable insights into the operational dynamics of BSSs. The classification of stations based on their operational lifespan revealed distinct patterns in station usage and performance. This analysis emphasized the importance of considering each station's specific characteristics and demands when making decisions about station placement, relocation, and removal.

The regression analysis of bike-sharing demand provided significant insights into the influence

of weather conditions, points of interest (POIs), the number of nearby stations, and temporal factors on bike-sharing demand. The strategic placement of stations, proximity to POIs, and density of nearby stations emerged as key factors influencing bike-sharing demand.

Weather conditions and temporal factors also play a pivotal role in shaping demand. Favorable weather and optimal visibility encourage increased bike trips, while unfavorable conditions such as high humidity and precipitation deter usage.

The DBSCAN Spatial Clustering Analysis applied on longitude and latitude data identified distinct clusters representing different patterns of station usage and locations. This analysis provides valuable insights into the spatial distribution and usage patterns of the stations, which can guide strategic decisions for improving services and user experience.

The Affinity Propagation and K-means clustering methods dissected the data into distinct clusters, each representing a unique group of stations sharing similar characteristics. These insights are invaluable for tailoring services to meet user needs, planning station expansions to accommodate growing demand, and implementing targeted marketing strategies to attract new users and retain existing ones.

This study has illuminated the complex dynamics of BSSs and the factors contributing to their success. It has highlighted the importance of strategic planning and management in ensuring the effectiveness and sustainability of these systems. Future research could further explore the factors influencing station lifespan and performance, contributing to the development of more efficient and sustainable BSSs.

7.1 Limitations

This section outlines the constraints encountered during the study of Montreal's BSS.

- **Geographical Limitation:** The study focuses on Montreal's Bixi BSS. While this provides valuable insights into BSS usage in Montreal, the findings may not be general-

izable to other cities or BSS due to differences in factors such as urban infrastructure, population density, and local transportation policies.

- **Data Limitations:** The open-source data used in this study does not provide information on the total demand for bike-sharing. For instance, it does not provide data on the number of bikes available at each station at any given time. This is a significant limitation as it prevents a complete understanding of the supply-demand dynamics of the Bixi system.
- **Assumption of Demand:** The study assumes that the recorded trips in the data represent the demand for the bss. However, this may not always be the case. For example, potential demand that was not met due to a lack of available bikes is not represented in the data.
- **Methodological Limitations:** While the clustering techniques used in this study provide valuable insights into the diversity and complexity of Bixi stations, they are dependent on the chosen parameters and may produce different results with different parameter settings.

7.2 Future work

The potential directions for future research to enhance our understanding of BSS can be incorporating additional data, analyzing long-term trends and external events, conducting comparative analyses, exploring policy implications, and conducting user surveys to gain deeper insights into user behavior and preferences.

- **Incorporating Additional Data:** Future research could benefit from incorporating additional data into the analysis. For instance, data on the number of bikes and docks available at each station at different times could provide a more complete picture of the supply-demand dynamics of the Bixi system. Similarly, data on unmet demand (e.g.,

instances where potential users wanted to use the system but could not due to a lack of available bikes or docking spaces) could also be valuable.

- **Long-Term Trends and External Events:** Analyzing data over a longer time period could help capture long-term trends in bike-sharing usage. Additionally, investigating the impact of external events (such as public transport strikes or extreme weather events) on bike-sharing demand could provide interesting insights.

- **Comparative Analysis:** Conducting a comparative analysis of BSS in different cities could provide insights into how factors such as urban infrastructure, population density, and local transportation policies influence bike-sharing usage.

- **Policy Implications:** The findings of this study and future research could be used to inform policy decisions related to BSS. For example, understanding the factors that influence bike-sharing demand could help policymakers and system operators make decisions about where to locate new stations or how to allocate resources.

- **User Surveys:** Conducting surveys of bike-sharing users could provide additional insights into user behavior and preferences, which may not be fully captured by the transaction data.

REFERENCES

[1] L.-Y. Qiu and L.-Y. He, "Bike sharing and the economy, the environment, and health-related externalities," *Sustainability*, vol. 10, no. 4, 2018. [Online]. Available: https://www.mdpi.com/2071-1050/10/4/1145

[2] J. Woodcock, "Health benefits of bike-sharing systems," *The Lancet*, vol. 388, pp. 456–467, 2016.

[3] E. Fishman, "Bike share: A synthesis of the literature," *Transport Reviews*, vol. 36, no. 1, pp. 92–113, 2016.

[4] P. Midgley, "The role of smart bike-sharing systems in urban mobility," *Journal of Urban Mobility*, vol. 5, pp. 23–35, 2011.

[5] S. Shaheen, S. Guzman, and H. Zhang, "Bikesharing in europe, the americas, and asia: Past, present, and future," *Institute of Transportation Studies, UC Davis, Institute of Transportation Studies, Working Paper Series*, vol. 2143, 01 2010.

[6] Y. Zhang, "The impact of weather on bike-sharing usage: A multi-city study," *Transportation Research Part D: Transport and Environment*, vol. 47, pp. 84–97, 2016.

[7] J. Resource, "Bike-sharing programs: Research roundup," *Journalist's Resource*, 2018. [Online]. Available: https://journalistsresource.org/studies/environment/transportation/bike-sharing-programs-research-roundup/

[8] P. DeMaio, "Bike-sharing: History, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, pp. 41–56, 2009.

[9] M. Ricci, "Bike-sharing: A review of evidence on impacts and processes of implementation and operation," *Research in Transportation Business & Management*, vol. 15, pp. 28–38, 2015.

[10] I. Frade and A. Ribeiro, "Bike-sharing stations: A maximal covering location approach," *Transportation Research Part A General*, vol. 82, p. 216–227, 12 2015.

[11] P. Vogel and D. C. Mattfeld, "Modeling of repositioning activities in bike-sharing systems," 2009.

[12] P. Hulot, D. Aloise, and S. D. Jena, "Towards station-level demand prediction for effective rebalancing in bike-sharing systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 378–386. [Online]. Available: https://doi.org/10.1145/3219819.3219873

[13] Y. Zhou and Y. Huang, "Place representation based bike demand prediction," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 1577–1586.

[14] H. Lim, K. Chung, and S. Lee, "Probabilistic forecasting for demand of a bike-sharing service using a deep-learning approach," *Sustainability*, vol. 14, no. 23, p. 15889, 2022. [Online]. Available: https://www.mdpi.com/2071-1050/14/23/15889

[15] S. Rühmann, S. Leible, and T. Lewandowski, "Interpretable bike-sharing activity prediction with a temporal fusion transformer to unveil influential factors: A case study in hamburg, germany," *Sustainability*, vol. 16, no. 8, p. 3230, 2024.

[16] "Discovering spatiotemporal usage patterns of a bike-sharing system by type of pass: a case study from seoul," *Transportation*, 2023.

[17] T. Schmidt, M. Scholz, and M. Fischetti, "Temporal fusion transformer for interpretable multi-horizon time series forecasting on bike-sharing system," *Transportation Research Part C: Emerging Technologies*, 2024.

[18] J. Li, "Seasonal variations in bike-sharing demand," *Journal of Transport Geography*, vol. 98, pp. 102–115, 2023.

[19] Y. Zhang, "Spatio-temporal analysis of bike-sharing data," *Urban Studies*, vol. 59, no. 4, pp. 789–805, 2022.

[20] M. Verma and A. Awasthi, "Evaluating bikesharing service quality: a case study for bixi, montreal," *International Journal of Productivity and Quality Management*, vol. 29, no. 1, pp. 45–61, 2020.

[21] Q. Chen, Z. Liu, J. Zhang, and X. Li, "Target-based stochastic distributionally robust optimization for docked bike-sharing systems under demand uncertainty," *Transportation Research Part B: Methodological*, 2024.

[22] M. Johnson, "Service quality and user satisfaction in bike-sharing systems," *Journal of Transportation Research*, vol. 56, no. 2, pp. 234–250, 2023.

[23] H. Wang, "User-centric evaluation of bike-sharing service quality," *Transportation Research Part C*, vol. 134, pp. 103–115, 2022.

[24] A. Faghih-Imani and N. Eluru, "Determining the role of bicycle sharing system infrastructure installation decision on usage: Case study of montreal bixi system," *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 685–698, 2016.

[25] Y. Guo, L. Yang, and Y. Chen, "Bike share usage and the built environment: A review," *Frontiers in Public Health*, vol. 10, 2022.

[26] B. Wei and L. Zhu, "Exploring the impact of built environment factors on the relationships between bike sharing and public transportation: A case study of new york," *ISPRS International Journal of Geo-Information*, vol. 12, no. 7, p. 293, 2023.

[27] K. Hosford, M. Winters, D. Juhlin, and D. Fuller, "The impact of station proximity to food serving enterprises on bike-sharing ridership," *Journal of Transport Geography*, 2024.

[28] J. Smith, "Urban design and bike-sharing usage," *Journal of Urban Mobility*, vol. 12, no. 3, pp. 45–60, 2023.

[29] K. Lee, "Land use diversity and bike-sharing demand," *Transportation Research Part C*, vol. 134, pp. 103–115, 2022.

[30] J. Wang, X. Long, and W. Li, "Understanding the intention to use bike-sharing system: A case study in xi'an, china," *PLOS ONE*, vol. 16, no. 12, p. e0258790, 2021.

[31] K. Kim, "Discovering spatiotemporal usage patterns of a bike-sharing system by type of pass: a case study from seoul," *Transportation*, 2023.

[32] J. Quach and R. Malekian, "Exploring the weather impact on bike sharing usage through a clustering analysis," *arXiv preprint arXiv:2008.07249*, 2020.

[33] N. Rennie, C. Cleophas, A. M. Sykulski, and F. Dost, "Analysing and visualising bike-sharing demand with outliers," *Discover Data*, vol. 1, no. 1, p. 1, 2023.

[34] Y. Zhang, S. Wang, and X. Li, "Exploring the spatiotemporal activities pattern of bike-sharing systems using k-means clustering: A case study in ningbo, china," *Journal of Cleaner Production*, 2024.

[35] T. Mátrai and J. Tóth, "Cluster analysis of public bike sharing systems for categorization," *Sustainability*, vol. 12, no. 14, p. 5501, 2020.

[36] Y. Xue and J. Li, "Exploring the weather impact on bike sharing usage through clustering in washington d.c." *Journal of Transport Geography*, vol. 68, pp. 102–115, 2018.

[37] K. Kumari, "End-to-end case study: Bike sharing demand prediction," *Analytics Vidhya*, 2023.

[38] C. Gao and Y. Chen, "Using machine learning methods to predict demand for bike sharing," *Information and Communication Technologies in Tourism*, pp. 282–296, 2022.

[39] M. S. Bahadori, A. B. Gonçalves, and F. Moura, "A systematic review of station location techniques for bicycle-sharing systems planning and operation," *ISPRS International Journal of Geo-Information*, 2021.

[40] V. Albuquerque, M. S. Dias, and F. Bacao, "Machine learning approaches to bike-sharing systems: A systematic literature review," *ISPRS International Journal of Geo-Information*, vol. 10, no. 2, p. 62, 2021.

[41] A. Behroozi and A. Edrisi, "Predicting travel demand of a bike sharing system using graph convolutional neural networks in chicago," *arXiv preprint arXiv:2408.09317*, 2024.

[42] C. Options, "The many benefits of bike sharing programs," https://www.commuteoptions.org/the-many-benefits-of-bike-sharing-programs/, 2022.

[43] U. of British Columbia, "Bike share | cycling in cities," https://cyclingincities.spph.ubc.ca/motivating-cycling/bikeshare-systems/, 2022.

[44] P. U. Solutions, "The 3 health impacts of a bike share system for cities," 2021.

[45] E. Fishman, "Bikeshare: A review of recent literature," *Transport Reviews*, vol. 36, no. 1, pp. 92–113, 2016.

[46] P. Yi, F. Huang, and J. Peng, "A rebalancing strategy for the imbalance problem in bike-sharing systems," *Energies*, 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/23/4588

[47] D. Freund, A. Norouzi-Fard, A. Paul, C. Wang, S. G. Henderson, and D. B. Shmoys, "Data-driven rebalancing methods for bike-share systems," *SpringerLink*, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s10203-020-00261-7

[48] P. Beigi, M. Khoueiry, M. S. Rajabi, and S. Hamdar, "Station reallocation and rebalancing strategy for bike-sharing systems: A case study of washington dc," *arXiv preprint arXiv:2204.07875*, 2022.

[49] M. Cipriano, L. Colomba, and P. Garza, "A data-driven based dynamic rebalancing methodology for bike sharing systems," *Applied Sciences*, vol. 11, no. 15, p. 6967, 2021.

[50] D. Loaiza-Monsalve and A. P. Riascos, "Human mobility in bike-sharing systems: Structure of local and non-local dynamics," *PLOS ONE*, 2019. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229616

[51] F. Kon, Éderson C. Ferreira, H. A. de Souza, F. Duarte, and P. Santi, "Abstracting mobility flows from bike-sharing systems," *MIT Open Access Articles*, 2021. [Online]. Available: https://dspace.mit.edu/handle/1721.1/130680

[52] D. Loaiza-Monsalve and A. P. Riascos, "Human mobility in bike-sharing systems: Structure of local and non-local dynamics in chicago and new york," *PLOS ONE*, vol. 14, no. 3, p. e0213106, 2019.

[53] L. Caggiani and R. Camporeale, "Toward sustainability: Bike-sharing systems design, simulation and management," *Sustainability*, 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/11/6182

[54] E. Fishman, "Bikeshare: A review of recent literature," *Transport Reviews*, 2016. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01441647.2015.1033036

[55] L.-Y. Qiu and L.-Y. He, "Bike sharing and the economy, the environment, and health-related externalities in china," *Sustainability*, vol. 10, no. 4, p. 1145, 2018.

[56] L. Caggiani and R. Camporeale, "Toward sustainability: Bike-sharing systems design, simulation, and management," *Sustainability*, vol. 13, no. 14, p. 7519, 2021.

[57] C. Reichel, "Bike sharing: Research on health effects, helmet use and equitable access," *JournalistsResource*, 2018. [Online]. Available: https://journalistsresource.org/studies/environment/transportation/bike-sharing-health-helmets-equity/

[58] L. W. Kille, "Bikeshare systems: Recent research on their growth, users' demographics and their health and societal impacts," *JournalistsResource*, 2015. [Online]. Available: https://journalistsresource.org/studies/environment/transportation/bike-sharing-programs-research-roundup/

[59] D. Rojas-Rueda and R. Clockston, "The public health benefits of bike share, quantified," *Streetsblog USA*, vol. 23, no. 7, pp. 36–50, 2021.

[60] J. F. Teixeira and C. Silva, "The potential of bike-sharing during public health crises: A review," *Journal of Transport & Health*, vol. 28, p. 100933, 2023.

[61] S. Martínez, A. Tapia, V. Bernardo, J. E. Ricart, and M. R. Planas, "The economic impact of bike sharing in european cities," *IESE Business School*, 2019. [Online]. Available: https://www.iese.edu/media/research/pdfs/ST-0505-E.pdf

[62] J. Li and W. Wang, "From renting economy to sharing economy: How do bike-sharing platforms grow in the digital era?" *Journal of the Knowledge Economy*, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s13132-023-01417-3

[63] B. Montréal, "Open data - bixi montréal," 2024. [Online]. Available: https://bixi.com/en/open-data/

[64] Historical weather data,government of canada. [Online]. Available: https://climate.weather.gc.ca/

[65] O. G. Portal, "Places of interest in the city of montreal." [Online]. Available: https://open.canada.ca/data/en/dataset/763fe3b8-cdc3-4b8a-bbbd-a0a9bc587c56/resource/abb993f7-b519-4de6-806b-d8f6b039d646

[66] J. Liu, L. Sun, Q. Li, J. Ming, Y. Liu, and H. Xiong, "Functional zone based hierarchical demand prediction for bike system expansion," 08 2017, pp. 957–966.

[67] J. Shu, M. Chou, Q. Liu, C. Teo, and I.-L. Wang, "Models for effective deployment and redistribution of bicycles within public bicycle-sharing systems," *Operations Research*, vol. 61, pp. 1346–1359, 11 2013.

[68] P. DeMaio, "Bike-sharing: History, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, pp. 41–56, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077291X22002600

[69] A. Author and B. Coauthor, "Travel pattern analysis using geospatial data," *Journal of Geospatial Analysis*, vol. 12, no. 3, pp. 123–135, 2020.

[70] C. Another Author and D. Collaborator, "Proximity analysis in urban planning using geopy," *Urban Planning Journal*, vol. 8, no. 2, pp. 98–110, 2019.

[71] J. Bachand-Marleau, B. H. Y. Lee, and A. M. El-Geneidy, "The impact of distance and station density on the attractiveness of bike sharing," *Transportation research part D: transport and environment*, vol. 17, no. 7, pp. 522–524, 2012.

[72] D. Buck, R. Buehler, P. Happ, B. Rawls, P. Chung, and N. Borecki, "The economic impact of bike share stations on local businesses," *Transportation Research Record*, vol. 2520, no. 1, pp. 92–99, 2015.

[73] Y. Zhang, T. Thomas, M. Brussel, and M. van Maarseveen, "A systematic review of station location techniques for bicycle-sharing systems planning and operation," *Transportation Research Procedia*, vol. 19, pp. 304–324, 2016.

[74] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[75] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[76] J. Zhang, X. Li, and J. Wang, "Affinity propagation clustering of spatio-temporal data based on kernel density estimation," *IEEE Access*, vol. 7, pp. 15 767–15 777, 2019.

[77] B. Montréal, "Where do all the bixi go in the winter? - bixi montréal," 2021. [Online]. Available: https://bixi.com/en/ou-vont-les-bixi-lhiver/

[78] F. M. Mohammad Sadegh Bahadori, Alexandre B. Gonçalves, "A systematic review of station location techniques for bicycle-sharing systems planning and operation," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 8, p. 554, 2021.

[79] S. Shu, Y. Bian, J. Rong, and D. Xu, "Determining the exact location of a public bicycle station—the optimal distance between the building entrance/exit and the station," *PLOS ONE*, vol. 14, no. 2, p. e0212478, 2019.