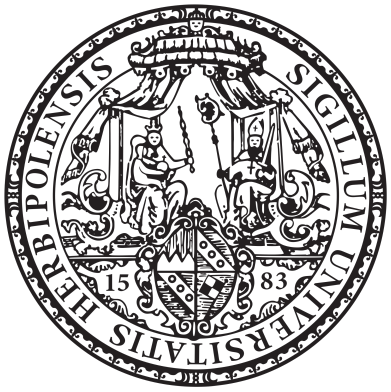
Analyzing Demand Patterns to Identify Similar Bike Stations in Chicago's Bike Sharing System

YAQIONG WANG

2876118



Seminararbeit

Lehrstuhl für Wirtschaftsinformatik und Business Analytics
Universität Würzburg

Betreuer: Prof. Dr. Gunther Gust
Assistent: Ignacio Úbeda

Würzburg, den 08.12.2022

Contents

List of Figures

List of Tables

Abstract

In this study, we aim to identify similar bike stations in Chicago's Bike-Sharing System (BSS) based on demand patterns. Our data analysis includes three phases: data understanding, where we explore potential factors influencing demand patterns and formulate research questions and hypotheses; data preparation, involving feature selection and extraction through visualization and statistical analysis; and data modelling, where we transform Origin-Destination (OD) data into hourly demand profiles for each station and apply the K-Means clustering algorithm. Our findings demonstrate a strong correlation between station demand patterns and the time of day, heavily influenced by the land use and geographic location of the stations. These insights are vital for improving urban mobility planning and the efficiency of BSS in Chicago.

1 Introduction

The concept of Bike-Sharing System (BSS), involving the shared use of a bicycle fleet, has already globally expanded since the 1960s. This form of urban mobility has spread across four continents – Europe, Asia, North America, and South America. Over time, nearly every bike-sharing program has transformed from the first generation of free bikes to the third or fourth generation, characterized by IT-based systems. The bike-sharing market has experienced significant growth (Shaheen, Guzman, and Zhang, 2010). As a market leader in Europe *Vélib' Métropole*, a pioneer in bike-sharing services based in Paris, launched in 2007. According to the data in October 2022, *Vélib' Métropole* announced 1,464 docking points and recorded 4.7 million trips in October alone (Vélib' Métropole, 2023). Transitioning to our research, the focus is on bike-sharing activities in the United States, specifically across the city of Chicago and its suburb, Evanston. *Divvy*, the bike share system we study, was initially launched in Chicago in 2013 and later expanded to Evanston in 2016. This system offers residents and visitors a convenient and affordable transportation option. As a program initiated by the *Chicago Department of Transportation (CDOT)*, *Divvy* features a fleet of specially designed bikes, accessible 24/7 throughout the year. These bikes can be accessed from and returned to any of the numerous docking stations spread across the network, supporting a variety of urban activities, including commuting, running errands, and city exploration (Divvy Bikes, 2023a).

Several studies have concentrated on understanding user behaviour and travel patterns in bike sharing, emphasizing insights into the development of bike-sharing mobility from a user perspective (Xing, Wang, and Lu, 2020). Our current study takes a different approach. Utilizing publicly available docked bike-sharing trip data provided by the company, we primarily focus on the start and end locations of bike trips, recorded in an Origin-Destination (OD) format. This method offers a more station-centric perspective, enabling us to glean insights from the standpoint of bike station demand and usage patterns.

By identifying similar bike stations based on their demand patterns and observing the usage capacity of each station, we can make more data-driven decisions regarding future bike-sharing station planning. This includes considerations such as combining stations, relocating, or expanding them. Moreover, by studying the demand patterns at bike stations, we can have a fundamental understanding of how to move and distribute bikes where they're needed most. For example, we can explore if the new electric bikes are more popular and whether more stations should be equipped with electric bike docks. This knowledge helps bike-sharing program improve user satisfaction and could encourage casual riders to become regular members. Based on these considerations, we have proposed the following research questions, which will be thoroughly explored in the subsequent methodology and result section:

- **RQ1:** How closely does the overall demand pattern align with the demand patterns observed within the cluster?
- **RQ2:** How do bike stations vary in terms of demand pattern performance?

- **RQ3:** Do bike stations with similar demand patterns tend to be geographically close to each other?

The structure of this paper is organized as follows: Initially, we explore the existing studies with a specific focus on the similarities among the bike-sharing stations. We then delineate the methodology adopted to address our research questions, beginning with the feature selection to distinguish demand patterns and characteristics, followed by the vectorization of every station as a preparation for applying unsupervised machine learning techniques. Our analytical approach incorporates the use of K-Means clustering to categorize bike stations through vectorized station data. By applying different metric standards, namely the elbow method and the silhouette score, we observe the clustering results, which are then represented on the city map. We compare the results from both k-values to respond to the research questions, assessing the effectiveness of the clusters. The paper also concludes with an acknowledgement of its limitations and suggestions for future research directions in urban mobility analytics.

2 Related Work

The bike-sharing programs exhibit diverse development globally. Therefore, it becomes crucial to conduct an in-depth analysis of local ride data to unearth specific insights in different urban contexts (Guo, Yang, and Chen, 2022). In the field of urban transport and BSS, research tends to diverge into two primary categories. The first focuses on understanding cycling behaviour, encompassing user behaviour and how trip purposes influence mobility services (Bordagaray et al., 2016; Builes-Jaramillo and Lotero, 2022; McKenzie, 2019). The second category is centered on the usage and demand patterns of bike stations and bicycles. Gaining insights into these patterns is crucial for bike-sharing companies and policymakers, as it significantly aids in enhancing mobility planning and the overall effectiveness of BSS. The specific research part of this study is to find the similarities among bike-sharing stations. namely, classify the bike stations. Within this scope, the following studies are taken into consideration:

Referring Zhao, Deng, and Song (2014), the research context centers on BSSs across various cities in China. The study classified the bike stations by evaluating station effectiveness and understanding station usage patterns through a turnover station ratio. The turnover ratio per bike per day and per station per day were combined to highlight the imbalances in bike distribution and station usage. The study uniquely employs supervised learning algorithms for cluster analysis of BSSs, differing from traditional unsupervised learning methods. This approach, combined with the use of OLS and PLS regression models and the innovative application of the Perceptually Important Points (PIP) process for time series data, provides a comprehensive understanding of bike-sharing station effectiveness, daily usage, and distribution patterns, all while effectively simplifying data without losing important structural details.

The research context of Chabchoub and Fricker (2014) focuses on BSS in Paris, France. The research classified the bike stations of *Velib* system into balanced, overloaded and underloaded categories based on resource ability (bikes and stations) through K-means and DTW (Dynamic Time Warping) metric for measuring station similarities, and DBA (Dynamic Barycenter Averaging) method for updating cluster centers.

The study from Xin et al. (2023) focuses on BSS in New York, United States. This newest study offers a fresh approach to bike-sharing research by focusing on bike mobility chains (BMCs) rather than raw OD data. It reconstructed bike-sharing OD trip data into BMCs and augmented these with spatial and temporal features. The application of word2vector embedding models and k-means clustering algorithms to BMCs reveals significant spatio-temporal inter-station connectivity and flow pattern, clustering bike stations based on a similar movement pattern. The research emphasized the necessity to shift focus from user mobility to bike mobility for effective management and planning which aligns with our current study goal.

Another study from Zhou (2015) focused specifically on the bike station clustering problem, and shares the similar background to the current study, both examining the dynamics within the same urban bike-sharing system in the United States. It begins with bike flow dynamics analysis, identifying neighbouring flows and grouping trips into communities with similar patterns. This is followed by a detailed spatio-temporal demand analysis for bikes and docks, using a two-hour time window to define the demand feature and hierarchical clustering to spot stations with comparable demand patterns. These insights are invaluable for enhancing the understanding of BSS and emphasize the importance of considering both the dynamics of bike flow and station demand in their planning and operation.

The existing research has predominantly focused on user-driven demand analysis in BSS, with fewer studies focusing deeply on station demand patterns from a data-centric perspective. Notably, some recent studies have started exploring station clustering. Our current study contributes to this emerging field by utilizing the latest data set, which encompasses an expanded network of bike stations. This expansion adds complexity and necessity to our analysis. By addressing the research questions outlined earlier, our study aims to gain a fundamental understanding of the characteristics of each cluster, the land use associated with demand patterns, and the similarities both within and across clusters. Although the analysis method is established, applying these to the latest data set updates and enriches our insights. Our study lays a foundational understanding of station demand patterns, crucial for addressing ongoing challenges in BSS, such as station relocation and resource balancing. These insights are valuable for bike-sharing programs in strategizing and implementing new operational approaches.

3 Methodology

3.1 Dataset Introduction

Our study focuses on the time-series demand patterns of BSS stations operated by *Divvy* in Chicago, specifically analyzing data from July 2022, while July obtains the most records. The BSS network according to our data set had expanded to encompass 1,147 unique bike stations, strategically distributed throughout Chicago's diverse communities. Our dataset, sourced from *Kaggle* (Kaggle, 2023), provides a comprehensive OD-style view of the system's operations.

The dataset comprises 12 columns, each offering specific information: **rider_id** identifies the individual trip. **rideable_type** distinguishes between traditional bikes and electronic bikes. **started_at, ended_at** record the check-in and check-out times at the bike stations, detailed down to the hour, minute, and second. Each station is uniquely identified by a **start_station_name(start_station_ID)** or an **end_station_name (end_station_ID)**. This information is critical for our study as it represents the station profile IDs, encompassing vectors of varying lengths. The last four columns provide the geographical coordinates **start_lat, start_lng, end_lat, end_lng** for the start and end stations. Column **member_causual** indicates whether the trip was made by a member or a casual user.

3.2 Data Preparation

In the data preparation phase, we commenced with a preliminary examination of the raw data, keeping our research objectives in focus. We identified the start station ID and the frequency of trip occurrences as key indicators for extracting demand patterns. Additionally, the geographical coordinates associated with each station allow us to visualize the distribution of bike stations within the city of Chicago, through those that demonstrate similar demand patterns.

To facilitate the analysis, we initially converted the trip occurrence records into a standardized date format, capturing year, month, hour, minute, and second. This transformation is crucial for subsequent steps, such as splitting the data into time slots and aggregating demand values. During the data collection phase, we observed that some records lacked station ID information. To maintain data integrity and consistency, records with missing station IDs were excluded from the data set. Additionally, we enhanced the data set by adding 'weekday' and 'calendar week' columns to each trip occurrence, enabling us to examine demand trends across different days of the week over several weeks. To gain an overarching view of the bike stations, we extracted a subset of the data, creating a separate data frame that includes only the station IDs and their geographical coordinates. This subset will be instrumental in the final step of our spatial analysis.

3.3 Exploratory Data Analysis and Feature Selection

"EDA techniques are used to interactively discover and visualize trends, behaviours, and relationships in data" (Chowdhury, Apon, and Dey, 2017). These techniques are crucial in guiding effective feature selection. EDA consists of three primary steps in data analysis: presentation, exploration, and discovery (Chowdhury, Apon, and Dey, 2017). Visualisation is an important aspect across these steps. In our study, the target of the presentation is to find intuitive insights within the dataset. By prioritizing key features such as station ID and trip frequency, we aim to analyze hourly demand trends across weekdays within July. Through visual inspection of the overall demand patterns, we gain the initial relationship in the time series dataset. The exploration phase involves developing preliminary hypotheses based on these visual observations and subsequently using statistical tests to verify the assumptions. Finally, the discovery step is the path paving the way for feature selection.

Feature selection is an effective and efficient data pre-processing strategy, especially for high-dimensional data in data mining and machine learning. Its main goals are to create simpler and more understandable models, enhance data mining performance, and prepare data that is both clean and easily interpretable (Li et al., 2017). This approach ensures that the chosen features are most representative of the underlying demand trends we seek to uncover and also to improve the analysis performance.

3.3.1 Demand Pattern Analysis

In our extensive dataset, individual trips are not used as the primary index. Instead, we undertake a process of re-engineering the data to explore how time influences demand on different days. Therefore, we proposed the assumption **bike station usage demand fluctuates within a single day but follows a similar trend across weekdays and weeks.**

To investigate this assumption, we reformat the 'started_at' column that separates the time information into 0-23 hourly segments, in conjunction with the newly added 'weekday' and 'calendar_week' columns. The goal is to unearth the patterns of demand across the 24 hours of weekdays and observe their progression throughout the four weeks of the month, using visualization and plotting for in-depth analysis.

In **Figure 1** presents the hourly demand for all bike stations' demand across different weekdays in July. Each sub-plot corresponds to a day of the week, illustrating the number of trips initiated from bike stations at each hour of the day. The vertical axis represents the count of trips started, the shade represents the demand difference among weeks. while the horizontal axis depicts the time of day, ranging from 0 to 23 hours.

After visual inspection of the subplots, it becomes evident that demand fluctuates not only within the hours of a day but also across different days of the week. A detailed examination of the plots reveals that, despite some acceptable variations in demand during the same hourly slots throughout the weeks, a consistent trend emerges from Monday to Friday, which are the typical working days. The overall demand exhibits a similar pattern in working days, whereas the weekends demonstrate a distinctively different trend.

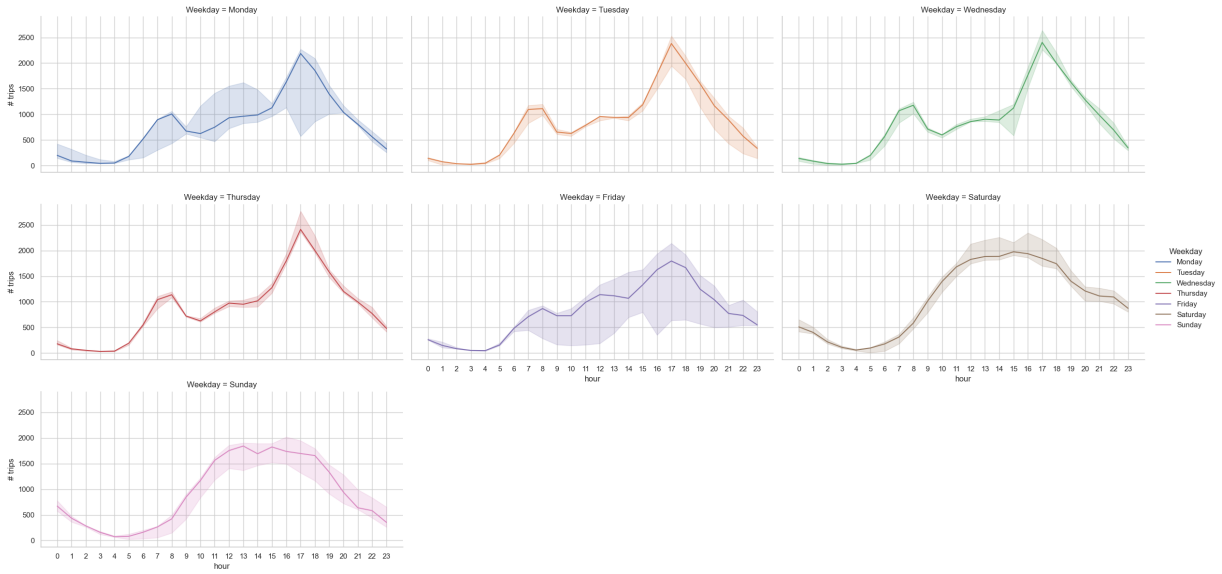


Figure 1: Hourly Demand across Weekdays in July

Drilling down to an hourly level during the working days, we observe a period of low activity from 0-5 AM, indicating a time when the bike stations are least active. Following this low active time, there is a gradual uptick in usage that peaks around 9 AM. This early morning peak is succeeded by a slight decrease, only for the demand to ascend once more, culminating in the highest peak at 5 PM. After the evening peak, there is a discernible decline in demand that continues until it reaches a low point until midnight.

Thus, we can preliminarily conclude that excluding weekends, the demand during working days follows a consistent pattern, irrespective of the specific workday or calendar week.

While visual inspection of the data offers preliminary insights, it can sometimes lead to misleading conclusions. Therefore, our next step involves a statistical analysis to solidify the findings inferred from our visual plots. A key aspect of this analysis is examining stationarity within the time series data, an essential factor in time series analysis. We aim to determine whether there is stationarity across different weeks and working days, which would significantly influence our feature selection decisions. Stationarity across weeks would imply that the demand demonstrates consistent statistical properties throughout various calendar weeks. Conversely, stationarity across working days would suggest a regular pattern of fluctuation in demand, characterized by consistent morning and evening peaks.

We must carefully note that after excluding weekend demand—due to our focus on similar patterns observed exclusively on weekdays—the count of unique bike stations considered in our analysis drops from 1,147 to 1,127. This reduction indicates that 20 bike stations are active only during weekends and will be also excluded from our further analysis.

To assess the stationarity of the demand pattern, we apply the Augmented Dickey-Fuller (ADF) test, a widely recognized statistical method. The ADF test focuses on the null hy-

potheses associated in the two scenarios:

- **H01:** The time series has a unit root, meaning it is non-stationary across weeks.
- **H02:** The time series has a unit root, meaning it is non-stationary across working days.

We computed the ADF test using the `adfuller` function from the `Statsmodels` package in Python. We observe the key factors of ADF Statistics and P-value in this test outcome.

The ADF Statistic of -8.5055, being highly negative, suggests a strong basis for rejecting the null hypothesis. Additionally, the P-value of 1.2057062947358326e-13, being well below the common threshold of 0.05, further supports the rejection of the null hypothesis of non-stationarity. Consequently, we can conclude with a rejection of H01, affirming stationarity in the demand pattern across different weeks.

On the other hand, the ADF test for evaluating H02, which concerns stationarity across working days, was conducted following the same procedure. The results of the ADF test, including both the test statistics and p-values for Monday through Friday, are summarized in **Table 1**. Based on these results, we conclude with the rejection of H02, thereby confirming stationarity in the demand pattern throughout the working days.

Table 1: ADF Test Results for Working Days

Weekday	ADF Test Statistic	P-value	Conclusion
Monday	-3.6273	0.00526	Stationary
Tuesday	-6.0171	<0.0001	Stationary
Wednesday	-4.2886	0.00047	Stationary
Thursday	-6.4417	<0.0001	Stationary
Friday	-3.9406	0.00176	Stationary

3.3.2 Station-Level Analysis

To verify if the stationarity factors are also confirmed in the station-level analysis, we posit another assumption that the demand at individual stations also reflects the stationarity across weeks and working days which we observed in the overall demand analysis. To gain initial insights, we first focus on the bike stations with the highest usage levels. We select the top five stations with the greatest frequency of use, according to our records. These stations are identified by their 'start_station_id' and corresponding occurrence times, which are as follows: station "13022" with 13,877 trips, station "LF-005" with 8,177 trips, station "13042" with 7,429 trips, station "13300" with 7,366 trips, and station "TA1308000001" with 6,553 trips. The significant demand at these stations provides a crucial foundation for gaining intuitive insights in our station-level analysis.

In **Figure 2**, it is evident that the usage patterns of individual bike stations also vary distinctly between weekdays and weekends. In examining the hourly trends, we capture a general pattern like demand pattern analysis: inactivity during the early morning hours, a

morning peak around 9 AM and an evening peak, followed by a gradual decrease towards midnight. However, it's notable that individual bike stations vary in their activity levels, displaying different performances across various time slots. From a visual standpoint, it's challenging to claim that these stations strictly adhere to the overall demand pattern. However, an obvious distinction in demand patterns between working days and weekends can be captured. To obtain more precise results, a quantitative analysis at the station level is essential.

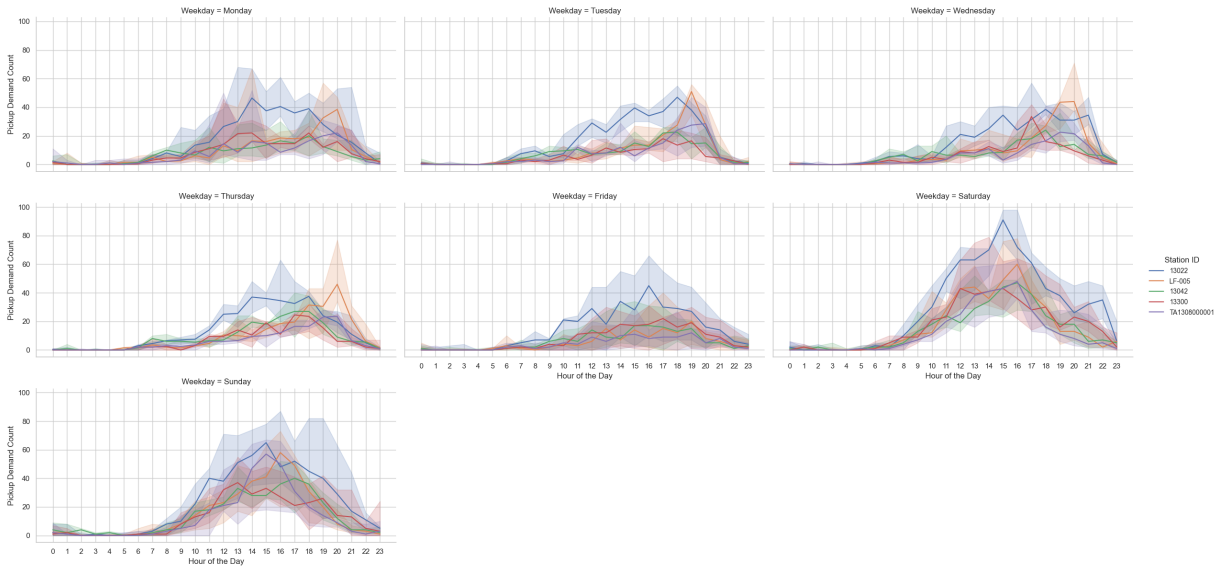


Figure 2: Top 5 Stations Hourly Demand across Weekdays in July

Given our demand pattern analysis, we next aim to assess stationarity at the individual station level through the ADF test. The null hypotheses for each bike station are as follows:

- **H01:** The time series of an individual bike station has a unit root, suggesting non-stationarity across weeks.
- **H02:** The time series of an individual bike station has a unit root, suggesting non-stationarity across working days.

As previously noted in the overall demand analysis, this station-level analysis also focuses exclusively on working days. We continue to exclude the 20 stations that are only active on weekends, thereby concentrating on the remaining 1,127 stations in our dataset.

Considering the extensive number of stations, presenting all testing results for each individual would be impractical. Therefore, we focus on summarizing the outcomes of our hypothesis testing. A station is considered to exhibit stationarity only if it meets both of the following criteria: a) The ADF statistic is lower than the varying critical value for different time series, which implies stronger evidence against the null hypothesis and b) The P-value is below the threshold of 0.05, indicating statistical significance.

Our results show that 1,118 out of the 1,127 bike stations meet these criteria, thereby

rejecting H01. This overwhelming majority suggests that, at the individual station level, there is a prevalent trend of stationarity across weeks.

Continuing with the same approach for H02, we conduct the ADF test for each station on every working day. For each day, we evaluate the hypothesis, determining whether it is rejected or not. If a station shows rejection of the null hypothesis across all five working days, we then conclude that the station's demand pattern is stationary across the working days. The results indicate that all 1,127 stations reject H02, leading us to conclude that the stationarity demand pattern is indeed present at the station level throughout the working days.

Based on our examination of stationarity, both in the overall demand pattern analysis and at the individual station level show stationarity, we can discard factors of calendar week and working days, instead aggregating them for a more focused comparison of hourly demand.

3.4 Feature Extraction

Having established the presence of stationarity across both weeks and working days, we also observed notable variance in demand throughout a 24-hour period. Consequently, we have decided to represent each station as a 24-dimensional vector, where each dimension corresponds to an hour's demand of the day. This approach allows us to capture the unique hourly demand characteristics of each station. The final bike station profile provides a comprehensive view of their usage patterns.

3.4.1 Median as central tendency

Mean, Median, and Mode are the three most commonly used measurements to present the central tendency. It is important to understand and choose the proper measurement to represent the hourly demand character of each bike station. The mean is known as the average and is computed by adding all the values in the data set and dividing by the number of observations. The median refers to the value located in the middle position of each dataset. The mode is known as the most frequent value that occurs within the dataset (Khorana et al., 2023). We took the median as the final central tendency representation in terms of the following considerations (Manikandan et al., 2011).

- **Presence of Extreme Scores:** Our dataset indicates that the demand patterns vary very differently in different hour slots, which can skew the mean. The median is more robust in such a scenario, providing a more representative measure of central demand.
- **Undetermined Values:**Not all stations have complete data across 24h meanwhile the big gap of hourly difference indeed shows the demand trend in time series. under this case, using the median as representation, being less sensitive to such outliner compared to the mean offers a more accurate reflection of the typical demand.
- **Open-Ended Distribution:**The hourly demand distribution of bike-sharing demand is open-ended, in practice, there is no upper limit on the number of bike station usage. The median is well-suited for such distributions, ensuring that the central tendency

is not disproportionately influenced by unusually high usage periods.

In **Figure 3**, we observed the hourly demand distribution both on all bike stations and one example bike station which is the busiest bike station "13022". By looking at both boxplots, we can conclude they both indicate the right skew. This skewness confirms that the median is the right choice for capturing what an hour demand pattern typically looks like.

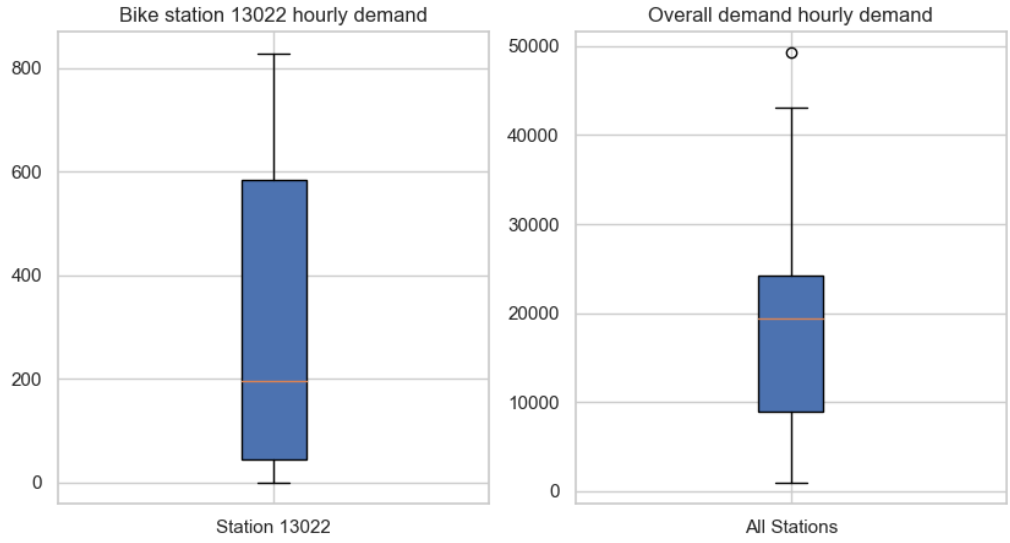


Figure 3: Hourly Demand Distribution

Finally, we have reshaped the data so that each station's activity is summarized as a 24-dimensionality vector. The value of each dimension is adopted from the median trip counts of that specific hour, taken across all working days within July. It is important to note that the number of data points contributing to these medians can vary significantly depending on the time of day. For example, during quieter hours, such as from 1-5 AM, the median may be adopted from few recorded demands (e.g. 1 or 2 records). In contrast, in peak hours, like 5 PM, present a richer dataset, with the median computed from the complete demand dataset (20 times) instances.

Adopting the median to represent hourly demand is especially advantageous for datasets with varying numbers of observations. This approach ensures that each station's hourly demand is captured without the bias that could result from different sample sizes. In contrast, applying the median to uniform-sized datasets does not account for this variability and may lead to an unrepresentative characterization of demand, especially in cases (low-demand stations, inactive timeslots) where the presence of many zero-demand intervals could skew the analysis. Therefore, using the median in different data sizes offers a more accurate reflection of hourly demand patterns.

Next, we turn our attention to the row-wise normalization. This normalization is a crucial

data pre-processing step to ensure the data falls into a common scale before applying unsupervised machine learning algorithms. Row-wise normalization involves adjusting the data in each row to a common scale without distorting differences in the ranges of values (each row represents an individual station and after the operation of row-wise normalization, the sum of each row should equal to 1). By doing so, we make sure that each station's activity pattern is considered equally.

3.5 K-Means Clustering Technique

Finding similarities among bike stations by analyzing demand patterns is a classic challenge in unsupervised learning, particularly within the field of clustering. The main distinction to supervised learning lies in the absence of pre-labelled bike stations. In supervised learning, models are trained on labelled data with known targets, enabling the machine to predict these outcomes for new data.

On the other hand, in the unsupervised learning approach, feature engineering is applied to the raw data to highlight demand patterns and thereby improve model performance, but all without any pre-assigned labels. The algorithms aim to explore this unlabeled data to detect structures, relationships, and similarities. Through this process, the machine learns the insights autonomously and groups the bike stations into clusters based on the similarity of their demand patterns. Choosing a machine learning algorithm was discussed in Chapter 1 of Lantz (2019).

In this study, we adopted the K-Means method, a method renowned for its simplicity, flexibility, efficiency, and wide applicability. The simplicity of K-Means lies in its straightforward approach to partitioning data into distinct clusters, making it easy to understand and interpret. K-Means is particularly efficient when handling large data sets; compared to other techniques like DBSCAN, it is computationally less demanding. Its broad applicability and flexibility make it well-suited to diverse research questions. This includes our analysis, which focuses on identifying similar bike stations based on demand patterns. It offers straightforward insights into understanding the complex data. Given these advantages, K-Means has been chosen for further analysis in this study.

K-Means technique involves into primarily two steps: First, the algorithm initializes by assigning data points to k initial clusters based on their Euclidean distance from randomly selected cluster centers. Then, it recalculates the centers of each cluster as the mean position of all the points assigned to that cluster and reassigns each data point to the cluster whose center is now closest. This process of updating the clusters is iterative. The reassigning and updating of cluster centers happened several times. The process continues until the reassignment of data points no longer significantly changes the cluster centers. At this point, when there are no significant changes in the cluster centers upon reassignment, the algorithm stops. This stop indicates that the clusters are determined and the process is complete. The advantages and process of the K-Means algorithm are fundamental for our analysis drawn from Chapter 9 in Lantz (2019).

In the clustering step, we fed the 24-dimensionality vector, representing each bike station's profile for the clustering method, we employed the KMeans algorithm from the `sklearn.cluster` module. To determine the optimal number of clusters k, we utilized the elbow method, using the calculation of the sum of squared distances from each point to its assigned center. The `silhouette_score` is used from `sklearn.metrics` to validate the consistency within each of the clusters. Running the K-Means algorithm involved with various k values, a process conducted by `sklearn`.

By observing the visualisation of the elbow method, we can capture how the homogeneity or heterogeneity within the clusters changes for various values of k. As illustrated in the **Figure 4**, we observe a continuous decrease in heterogeneity with more clusters. The goal is not to minimize heterogeneity but rather to find k, after this "elbow point" there are diminishing returns. In our plot, there is a bend around 4 clusters, suggesting that 4 might be the optimal number of k clusters for the bike stations.

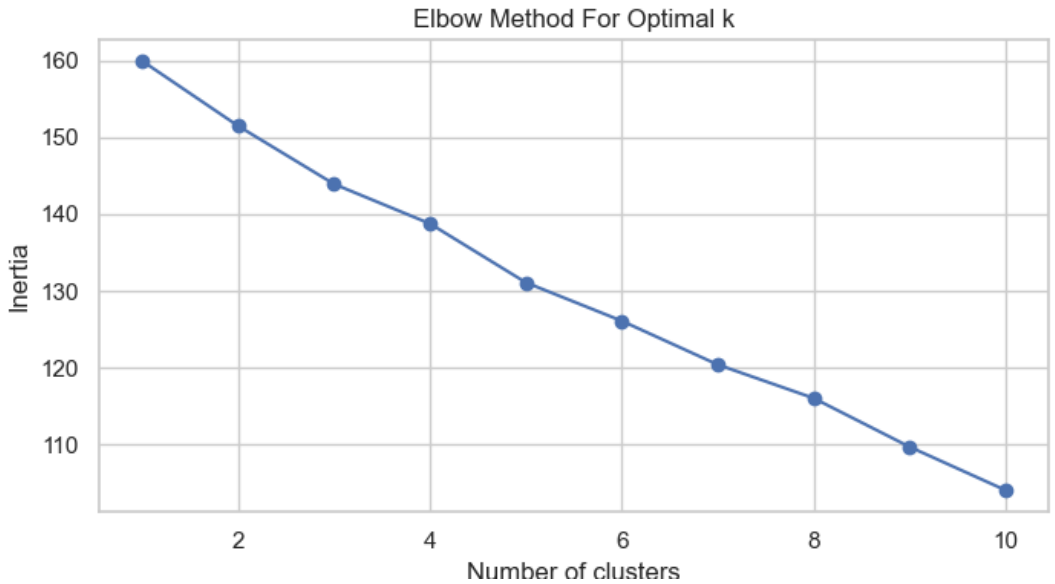


Figure 4: Number of clusters in Elbow Method

An average silhouette score is another way to find the optimal k clusters. The silhouette score range is from [-1,1], The value 1 indicates that the clusters are clearly clustered, the value of -1 indicates incorrect clustering, and the value 0 indicates the overlapped clusters. It measures how similar a bike station is to its cluster compared to other clusters. A higher silhouette score indicates better-defined and more distinct clusters (Rousseeuw, 1987; Rodriguez et al., 2019).

In **Figure 5** , we observed the silhouette score change trend with different k values. Since 0.606 is the highest score in our results, it suggests that at $k = 3$, the clusters are the most distinct and well-separated compared to other k values being tested. For more precise information, we can find it in **Table 2**.

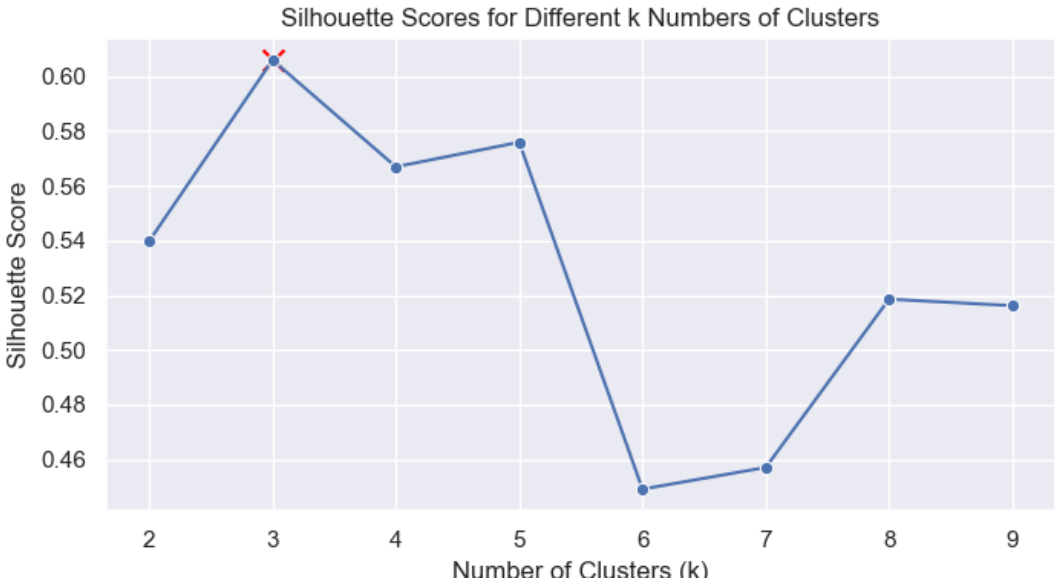


Figure 5: Number of clusters in Silhouette Score

Table 2: Silhouette Score for Different K clusters

Number of Clusters (K)	Silhouette Score
2	0.5399
3	0.6060
4	0.5670
5	0.5760
6	0.4492
7	0.4571
8	0.5186
9	0.5162

We can clearly observe that different methods yield varying optimal values for k. To validate the effectiveness of these differing clusters, The next step is to label each bike station, incorporating geographic data, and subsequently examining their spatial distribution on a map. Moreover, quantitative analysis will also be conducted to compare these clustering results. This approach will allow us to determine which k makes the most sense, aligning our findings with domain knowledge to ensure that the chosen clustering configuration not only adheres to data-driven insights but also resonates with practical, real-world understanding.

4 Results

4.1 Comparative Analysis

After assigning cluster labels to the bike stations based on geographic data and visualizing them on a map, we can observe the outcomes for $k = 3$ and $k = 4$ clusters. The bike stations distribution with $k = 4$ clusters is illustrated in **Figure 6**, and with $k = 3$ clusters in **Figure 7**.

Table 3 and 4 summarize the number of bike stations within each cluster, along with the associated colours used for their representation on the Chicago city map. The centroid latitude and longitude denote the central geographic coordinates for each cluster.

Regardless of whether we choose a k value of 4 or 3, we can address ***RQ3:*** *Do bike stations with similar demand patterns tend to be geographically close to each other?* By observing that bike stations with similar demand patterns tend to be geographically neighbouring. This indicates that, in most cases, stations nearby share similar demand patterns.

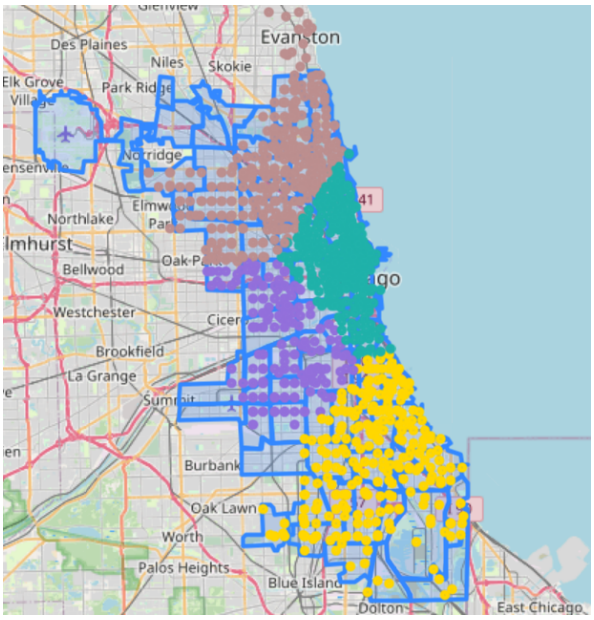


Figure 6: Bike Stations Distribution with clusters 4

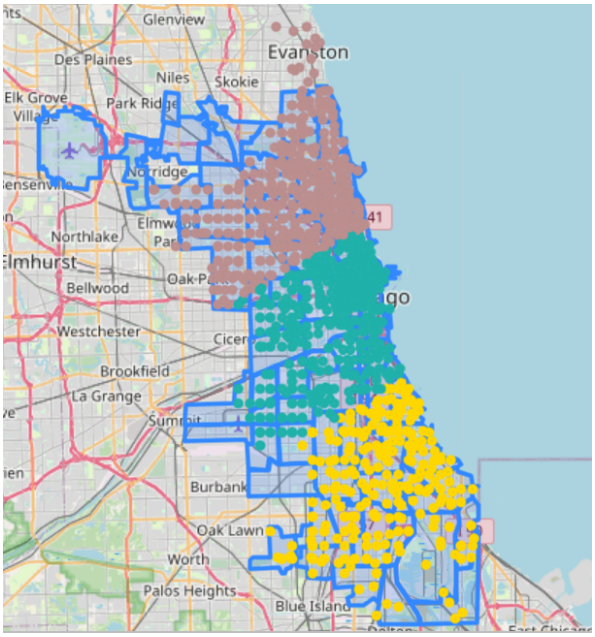


Figure 7: Bike Stations Distribution with clusters 3

To contextualize the clustering outcomes effectively, a preliminary grasp of the city's geography is essential. The city of Chicago spans across the geographic coordinates [41.881832, -87.623177]. It is located in the northeastern part of Illinois and is the third most populous city in the United States. Given its high population density and status as a popular tourist destination, the network of bike stations in the city has rapidly expanded, with many stations located close to one another. Chicago is notable for its large, flat land and its city grid

system, with the Chicago River and its system of bridges and canals running throughout the city. It's also characterized by its numerous parks, beaches, and marinas along the lakefront.

The result of the four-cluster model showed the following geographical distribution of bike stations: Cluster 0, consisting of 329 stations, is represented by rosybrown and covers the northwest region. In the southern zone, Cluster 1 with 283 stations marked in gold. Cluster 2, coloured green, includes 319 stations situated in the central area. Cluster 3, identified by purple, comprises 196 stations located in the southwest zone.

Examining the results from the three-cluster model, we observe the following distribution: Cluster 0, coloured rosybrown, emerges as the 411 stations, spanning the northwest parts of Chicago. Cluster 1, represented in gold, comprises 282 stations primarily situated in the city's southern region. Meanwhile, Cluster 2, marked in green and consisting of 434 stations, extends over the central areas.

Compared with the four-cluster model, some stations that were part of Clusters 2 and 3 (in the three-cluster model) have now transitioned into the rosybrown Cluster 0. Concurrently, the green Cluster 2 has absorbed all stations previously categorized under the purple Cluster 4 from the four-cluster model, forming now the largest Cluster 2 with 434 bike stations. Notably, the gold Cluster 1 maintains a similar size across both clustering models.

Table 3: Cluster Distribution and Characteristics for k=4

Cluster	Num Stations	Color	Centroid_[lat,lng]
0	329	rosybrown	[41.951843 -87.714938]
1	283	gold	[41.740469 -87.615015]
2	319	green	[41.889604 -87.645302]
3	196	purple	[41.827205 -87.701991]

Table 4: Cluster Distribution and Characteristics for k=3

Cluster	Num Stations	Color	Centroid_[lat,lng]
0	411	rosybrown	[41.945683 -87.706121]
1	282	gold	[41.739497 -87.616262]
2	434	green	[41.855786 -87.665216]

Adopting different cluster numbers can yield distinct perspectives on bike station usage patterns. To address ***RQ1****: How closely does the overall demand pattern align with the demand patterns observed within the cluster?* We turn to a comparative analysis of the 24-hour demand trends. This analysis is graphically represented in **Figure 8** for the four-cluster model, **Figure 9** for the three-cluster model, and contrasted against **Figure 1**.These visual comparisons enable us to discern the extent to which the demand patterns within each cluster align with the overarching demand trends captured in **Figure 1**.

- **4 Clusters Model:** Cluster 0 experiences a modest rise in demand during the early morning up to 8 am, stabilizes throughout the day and then sees a slight increase in

the 5-6 pm timeslot. Cluster 0 broadly reflects the overall demand pattern. Clusters 1 and 3 maintain a consistent yet low level of demand during all hours, lacking any significant peaks or fluctuations. However, cluster 1 has more high demand than cluster 3. Cluster 2 is characterized by a noticeable surge in demand at 8 am and a progressive climb to another peak at 5 pm, highlighting the intensive use of bike stations in this cluster. Cluster 2 exhibits a demand pattern that closely mirrors the general demand.

- **3 Clusters Model:** Cluster 0 exhibits a moderate increase in demand starting at 8 am, followed by a steady rise that continues until reaching a peak at 5 pm, after which demand declines. This cluster shows a more defined demand curve throughout the day compared to the four-cluster result, likely due to incorporating some bike stations from the four-cluster model's Cluster 2, resulting in a demand pattern that more closely approximates the general demand trend. Cluster 1 shows low demand consistently over the day, reflecting the same pattern observed in the four-cluster result cluster 1, and does not exhibit the overall demand fluctuation. Cluster 2 shares similar demand patterns with both Cluster 0 in the same mode and Cluster 2 in the four-cluster mode, but with sharper peaks at 8 am and 5 pm. Notably, the total demand for this cluster, although it comprises 434 bike stations, is less than the demand for the 319 bike stations in Cluster 2 of the four-cluster results. Nevertheless, it effectively captures the essence of the general demand trend.

The graphical analysis also concludes that the elbow method, which identifies $k = 4$ as optimal, focuses on grouping bike stations in such a way that those within each cluster exhibit closely matched demand patterns. On the other hand, the silhouette score, which recommends $k = 3$, aims to achieve an optimal balance, ensuring that stations within a cluster are cohesive in their demand patterns while maintaining clear distinctions from stations in other clusters.

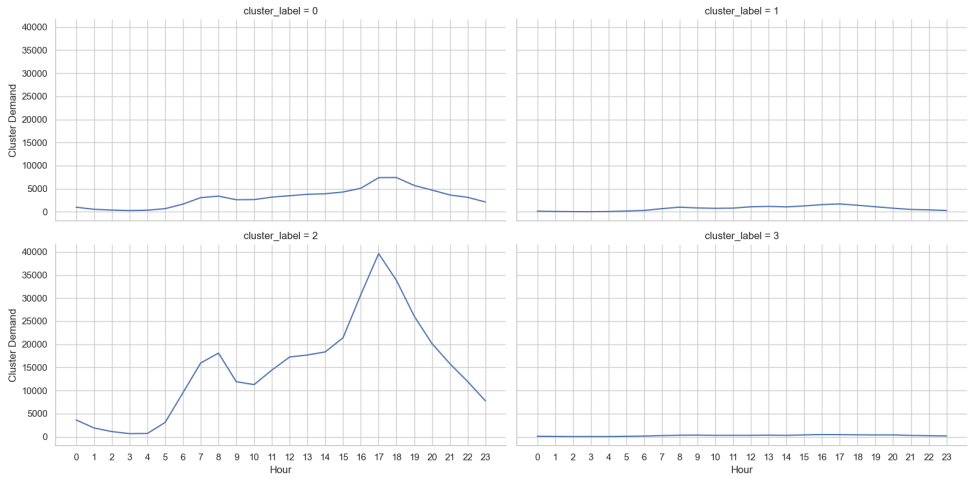


Figure 8: Cluster Demand Pattern with clusters 4

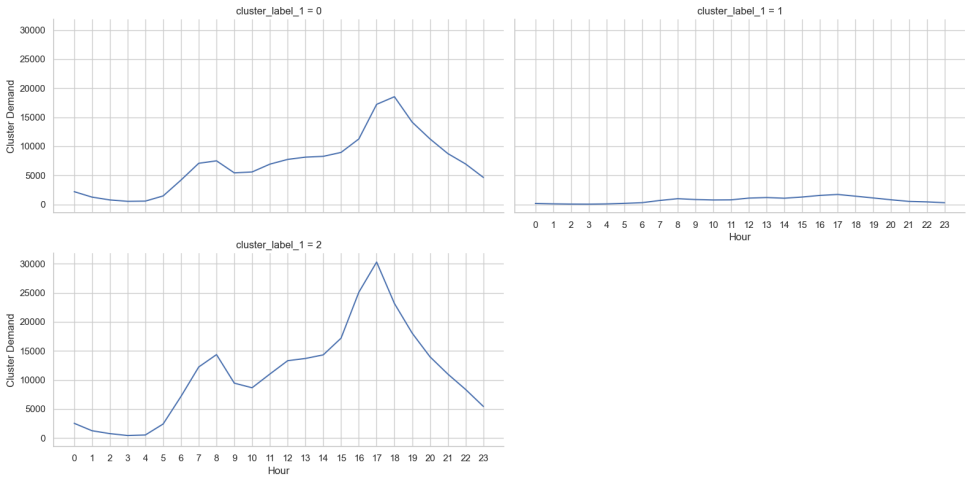


Figure 9: Cluster Demand Pattern with clusters 3

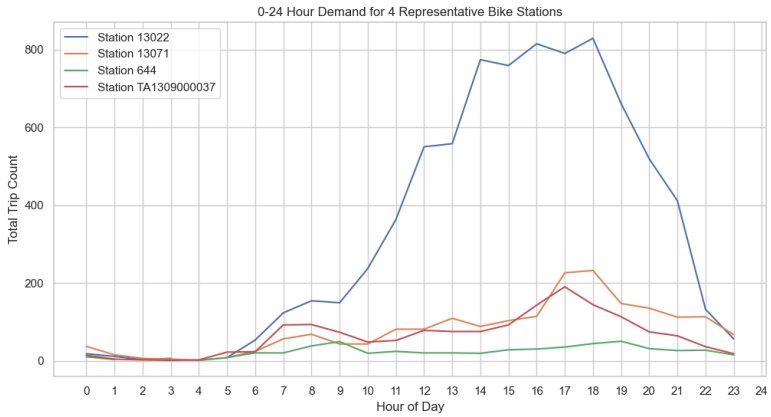


Figure 10: 4 Representative Bike Stations 0-24H Demand with 4 Clusters

To respond ***RQ2:****How do bike stations vary in terms of demand pattern performance?* We examine two key aspects: First, we assess the variations in demand patterns within individual clusters to determine their homogeneity. Second, we compare the demand patterns across different clusters to understand the distinctive characteristics of each cluster's demand profile. The idea is to first pick up the most representative bike station (max trips across July) in each cluster as the key stations. By visualizing the hourly usage trends of these key stations, we can determine the extent to which they conform to their cluster's overall demand trend, thus assessing the cluster's homogeneity. Subsequently, we will compare these trends across the clusters, integrating geographical information to uncover practical insights into station performance and usage.

Firstly, we examine the four-cluster model, the most active stations within each cluster. These are station "13071" in Cluster 0, leading with 1,913 trips; station "TA1309000037"

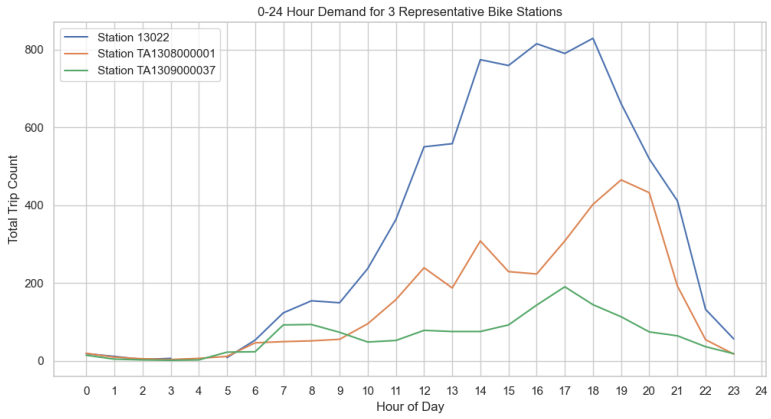


Figure 11: 3 Representative Bike Stations 0-24H Demand with 3 Clusters

in Cluster 1, with a total of 1,528 trips; station "13022" in Cluster 2, with a total demand of 7,979 trips; and station "644" in Cluster 3, accounting for 537 trips. These stations are identified as key due to their high usage within their respective clusters. By comparing **Figure 10** with **Figure 8**, we observe that, despite demand differences among key stations, each station exhibits a common pattern: a morning peak at 8 am, followed by a gradual increase in demand leading up to another peak around 5 pm, after which the demand decreases. However, station "644" in Cluster 3, despite being the most utilized station in its cluster, still displays very low demand. This aligns well with the overall low demand pattern observed in Cluster 3. Station "13022" in Cluster 2 stands out as the busiest, not just within its cluster but among all bike stations. It mirrors the cluster and overall demand patterns, with a significant observation: after 8 am, unlike the cluster and overall trends which show varied fluctuations, the demand at "13022" continues to rise steadily. The representative station "TA1309000037" in Cluster 1 also exhibits minor peaks around 7-8 am and 5 pm. However, the cluster's demand pattern is relatively stable compared to this station's trend. Station "13071" in Cluster 0, reflects the demand shape of its cluster, adhering to the observed trend of morning and evening peaks. In conclusion, the bike stations within the cluster have similar demand patterns in terms of peak usage times and demand trends.

In parallel, we observed the three-cluster model employing the same approach as before, by comparing the **Figure 11** with **Figure 9**. These key stations are "TA1308000001" in Cluster 0, leading with 3,562 trips; "TA1309000037" in Cluster 1, recording 1,528 trips; and "13022" in Cluster 2, registering a total of 7,979 trips. In cluster 0, the key station shifts from "1307" to "TA1308000001." This shift contributes to a clearer demand pattern within the cluster. However, individual station analysis reflects minor variations; for instance, the first peak occurs at 12 noon, followed by fluctuations and then another peak at around 7 pm, and demand starts to decline. This pattern is essentially a few hours delayed compared to the general cluster demand trend. Meanwhile, "TA1309000037" remains the key station in Cluster 1, where we can still observe the minor demand peaks at 8 am and 5 pm, consistent

with the cluster's demand pattern.

In conclusion, both the three-cluster and four-cluster models demonstrate homogeneity within their clusters. This is evident as the key stations in each cluster adhere to their respective cluster demand patterns at a very high level, which is also in line with the overall demand trends refer **Figure 1**. Although Cluster 3 in the four-cluster model exhibits a notably low activity level, it still follows the general trend of its cluster. This consistency across individual station trends and cluster-wide patterns suggests that the clustering is appropriately and effectively capturing the underlying demand dynamics.

5 Discussion

5.1 Interpreting Findings

Understanding the specific locations and surroundings of each cluster is key. These details help us see the deeper reasons behind each cluster's usage patterns. For a closer look at where these clusters' centroid lat,lng are located, check out the maps in Appendix A, adopted from Divvy Bikes (Divvy Bikes, 2023b). This geographical context is essential to understanding how and why people use bike stations differently in each area.

For the four-cluster model, in **Figure 12**, the centroid of cluster 0 is located in the northwest, surrounded by major streets such as W Elston Ave, N Central Park Ave, and N Drake Ave. This suggests a residential or mixed-use area. The observed demand pattern, characterized by low overall usage with minor peaks during commuting times, could indicate that residents perhaps use bike-sharing for short commutes within the area, like travelling to work or nearby destinations. This aligns with typical residential area commuting patterns.

In **Figure 13**, the centroid of cluster 1 is located in the southern part of the city, characterized by a grid pattern of streets, housing, and smaller local roads, suggesting the residential area setting. It aligns well with the consistently low demand observed. This suggests less frequent use of bike-sharing in this area. The presence of fluctuations during commuting times in individual bike stations, even with low overall demand, supports the residential character of the area and indicates some use of bike-sharing for commuting purposes, even on a smaller scale compared to more active areas.

In **Figure 14**, the centroid of cluster 2 along with the label "Downtown" and "River North" visible, this area is likely in the central part of Chicago, known for being a commercial and business hub with a mix of shopping, dining, entertainment, and residential spaces purpose. The concentration of landmarks, offices, and tourist attractions explains the dense distribution of bike stations and the high usage volume. The observation of significant bike station demand not only during commuting hours but also at other times of the day aligns with the diverse and constant activity typical of a central urban area.

In **Figure 15**, the centroid of cluster 3 is located southwest part, along with streets named W 35th St, W 36th St, and S California Ave. These communities are likely to be residential

areas. This correlates with the consistently low bike station demand observed in the cluster, aligning with typical patterns in residential neighbourhoods where bike-sharing usage is generally lower compared to commercial or mixed-used areas.

For the three-cluster model, in **Figure 16**, the centroid of cluster 0 across northwest Chicago, labelled "Portillo's" and "Jewel-Osco," along with a "Target" store nearby, suggests the commercial with retail and dining options. This aligns with the observed demand pattern, where high usage volume could be linked to increased bike station use during dining times, leading to peak demand.

In **Figure 17**, the centroid of cluster 1, situated in the southern region and close to the centroid of Cluster 1 from the four-cluster model, points to a residential area. This is consistent with the observed low-demand pattern,

In **Figure 18**, the centroid of Cluster 2 suggests a location not in the typical downtown area, but rather the centre of community Loop. This area, known for its artistic and cultural vibrancy, aligns with the large volume of bike stations in the cluster. The shift of the cluster centroid from a commercial area to an arts-focused neighbourhood like Pilsen, while still including the downtown area, supports the observed high bike station usage and the demand peaks during commuting times, reflecting the area's dynamic character.

These observations lead us to conclude that the demand patterns for bike-sharing stations are closely associated with the distinct characteristics of each area. Whether it is a commercial area, a residential area, or a cultural hotspot, the demand for bike stations varies differently, reflecting the land use of each community.

5.2 Limitations and Future Research

We acknowledge the limitations of our study. Post the EDA, our focus was on the hourly demand patterns during working days, overlooking factors such as weekend or holiday demand factors. Additionally, analyzing data from a single month limits our insight into monthly and seasonal variations. In the data pre-processing step, we employed a simple imputation method by discarding missing station ID records. This method could be enhanced by estimating the missing station IDs based on the shortest distance calculated from the Euclidean method to existing station ID records, ensuring a more comprehensive data set. Our study primarily examined pickup demand, thus potentially also missing out on the return demand patterns in bike stations.

Future research should aim for a more detailed analysis, applying a multi-dimensional feature vector representation that includes additional factors like weekend, holiday, monthly and seasonal demand character, thereby capturing a more detailed picture of bike station demand patterns. Additionally, exploring sophisticated analytical methods like various clustering techniques, Convolutional Neural Networks (CNNs), and advanced methodologies, such as word2vec combined with random walks to generate embedding matrices for each ride record, could be employed in machine learning models to gain deeper insights into the similarities of demand patterns within the network's structure and connectivity.

However, a significant challenge will be how to figure out meaningful relationships among these complex embeddings.

6 Conclusion

Our study, despite some limitations, has employed feature selection and extraction techniques to re-engineer OD data into a station's demand profile based on hourly time windows. Using the K-Means clustering algorithm, we have analyzed the characteristics of bike stations predicated on similar demand patterns. The findings suggest a strong correlation between individual bike station demand patterns and the time of day, with station usage volume being highly dependent on the land use of the community whether residential, commercial, or mixed and its geographic location.

In addressing our key research questions, we found:

- **RQ1:** Metrics used for identifying optimal cluster numbers consistently capture the general demand trend, despite highlighting different aspects.
- **RQ2:** Both three-cluster and four-cluster models demonstrate significant homogeneity within their clusters, which is also in line with the overall demand trends.
- **RQ3:** Geographical proximity correlates with similar demand patterns among bike stations, indicating pattern consistency in neighbouring areas.

This study establishes a fundamental understanding of the Chicago BSS. Building on this groundwork, future research should incorporate a multi-dimensional feature approach, utilizing advanced technologies to explore in-depth the complex relationships between bike stations. This deeper exploration will yield richer insights into the transport network's structure, crucial for enhancing city planning strategies.

References

Bordagaray, Maria, Luigi Dell'Olio, Achille Fonzone, and Ángel Ibeas (2016). "Capturing the conditions that introduce systematic variation in bike-sharing travel behavior using data mining techniques". In: *Transportation research part C: emerging technologies* 71, pp. 231–248.

Builes-Jaramillo, Alejandro and Laura Lotero (2022). "Spatial-temporal network analysis of the public bicycle sharing system in Medellın, Colombia". In: *Journal of Transport Geography* 105, p. 103460.

Chabchoub, Yousra and Christine Fricker (2014). "Classification of the vélib stations using Kmeans, Dynamic Time Wraping and DBA averaging method". In: *2014 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*. IEEE, pp. 1–5.

Chowdhury, Mashrur, Amy Apon, and Kakan Dey (2017). *Data analytics for intelligent transportation systems*. Elsevier.

Divvy Bikes (2023a). *About Divvy*. (visited on:21.11.2023). URL: `https://divvybikes.com/about`.

— (2023b). *Explore Chicago*. (visited on: 04.12.2023). URL: `https://divvybikes.com/explore-chicago`.

Guo, Yuanyuan, Linchuan Yang, and Yang Chen (2022). "Bike share usage and the built environment: a review". In: *Frontiers in public health* 10, p. 848169.

Kaggle (2023). *Cyclistic Bike Share 2023*. (visited on: 01.11.2023). URL: `https://www.kaggle.com/datasets/godofoutcasts/cyclistic-bike-share-2023?select=202303-divvy-tripdata.csv`.

Khorana, Arjun, Ayoosh Pareek, Matthieu Ollivier, Sophia J Madjarova, Kyle N Kunze, Benedict U Nwachukwu, Jón Karlsson, Erick M Marigi, and Riley J Williams III (2023). "Choosing the appropriate measure of central tendency: mean, median, or mode?" In: *Knee Surgery, Sports Traumatology, Arthroscopy* 31.1, pp. 12–15.

Lantz, Brett (2019). *Machine Learning with R: Expert Techniques for Predictive Modeling*. Packt Publishing Ltd.

Li, Jundong, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu (2017). "Feature selection: A data perspective". In: *ACM computing surveys (CSUR)* 50.6, pp. 1–45.

Manikandan, Subramanian et al. (2011). "Measures of central tendency: Median and mode". In: *J Pharmacol Pharmacother* 2.3, pp. 214–215.

McKenzie, Grant (2019). "Spatiotemporal comparative analysis of scooter-share and bike-share usage patterns in Washington, DC". In: *Journal of transport geography* 78, pp. 19–28.

Rodriguez, Mayra Z, Cesar H Comin, Dalcimar Casanova, Odemir M Bruno, Diego R Amancio, Luciano da F Costa, and Francisco A Rodrigues (2019). "Clustering algorithms: A comparative approach". In: *PloS one* 14.1, e0210236.

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Shaheen, Susan A, Stacey Guzman, and Hua Zhang (2010). “Bikesharing in Europe, the Americas, and Asia: past, present, and future”. In: *Transportation research record* 2143.1, pp. 159–167.

Vélib’ Métropole (2023). *The Vélib’ service*. (visited on: 21.11.2023). URL: `https://www.velib-metropole.fr/en/service`.

Xin, Rui, Jian Yang, Bo Ai, Linfang Ding, Tingting Li, and Ruoxin Zhu (2023). “Spatiotemporal analysis of bike mobility chain: A new perspective on mobility pattern discovery in urban bike-sharing system”. In: *Journal of Transport Geography* 109, p. 103606.

Xing, Yingying, Ke Wang, and Jian John Lu (2020). “Exploring travel patterns and trip purposes of dockless bike-sharing by analyzing massive bike-sharing data in Shanghai, China”. In: *Journal of transport geography* 87, p. 102787.

Zhao, Jinbao, Wei Deng, and Yan Song (2014). “Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in China”. In: *Transport Policy* 35, pp. 253–264.

Zhou, Xiaolu (2015). “Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in Chicago”. In: *PloS one* 10.10, e0137922.

A Appendix

Note: All maps in this Appendix are adapted from "Explore Chicago" by Divvy Bikes, from `https://divvybikes.com/explore-chicago` on 04.12.2023

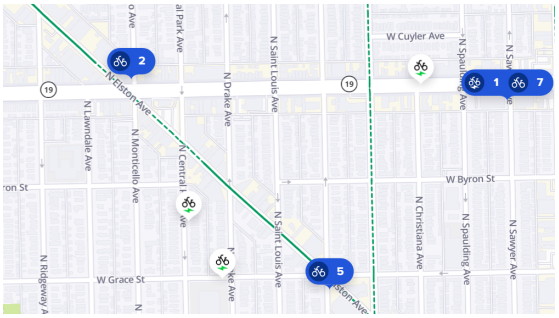


Figure 12: cluster 0 geo map with k=4

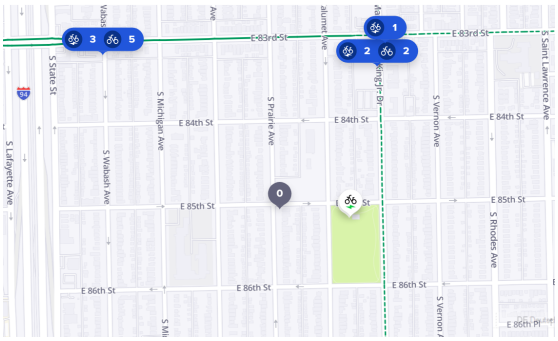


Figure 13: cluster 1 geo map with k=4

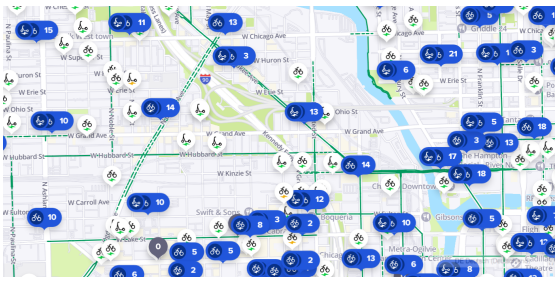


Figure 14: cluster 2 geo map with k=4

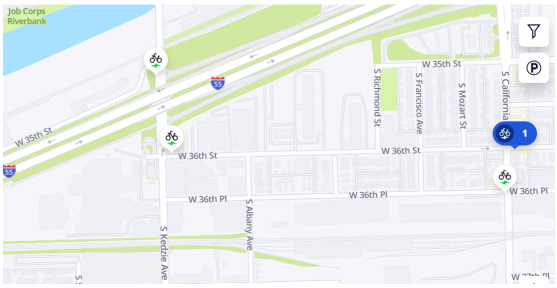


Figure 15: cluster 3 geo map with k=4

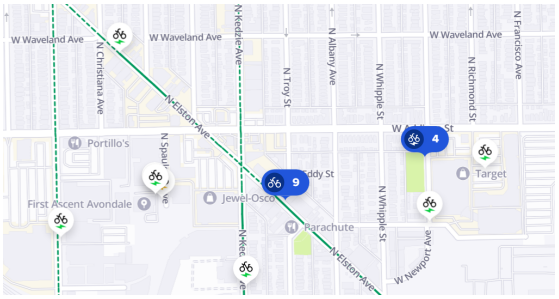


Figure 16: cluster 0 geo map with k=3

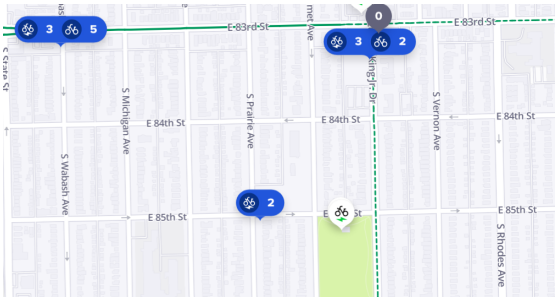


Figure 17: cluster 1 geo map with k=3

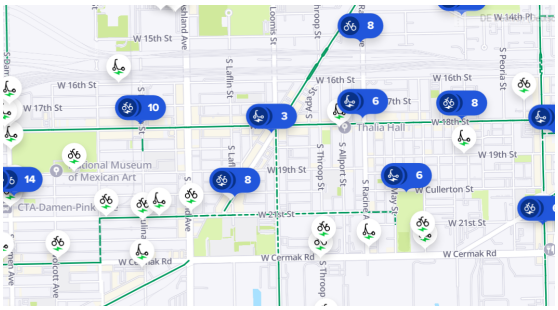


Figure 18: cluster 2 geo map with k=3

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die Zitate deutlich kenntlich gemacht zu haben.
Ich erkläre weiterhin, dass die vorliegende Arbeit in gleicher oder ähnlicher Form noch nicht im Rahmen eines anderen Prüfungsverfahrens eingereicht wurde.

Würzburg, den December 8, 2023 Yaqiong Wang