Share Bike Station Clustering and Usage Prediction

by

Sijia (Nancy) Li, 1004073904

April 2022

An undergraduate thesis (MIE498Y1) submitted for
the Bachelor of Applied Science in Industrial Engineering
Department of Mechanical and Industrial Engineering
University of Toronto

Abstract

Bike sharing systems have become increasingly popular around the world for its benefits in public health, environment, and urban mobility. Managing the operations of such systems, however, requires a good understanding of bike usages. In fact, the prediction of bike demands is a critical component of many operational problems, including downstream bike rebalancing efforts. In recent years, there is a growing interest in adopting the cluster-then-predict approach, whereby bike stations are grouped into clusters first and bike demand predictions are made later for each cluster. This thesis project has two main objectives: first, examining how various station-to-cluster assignments may impact the bike check-out and check-in predictions, and second, understanding how transition information may influence the bike check-in predictions. We present a five-step cluster-then-predict framework that generates station clusters using a state-of-the-art clustering algorithm, makes check-out predictions, computes cluster-to-cluster transition matrices, makes check-in predictions, and evaluates the predictions. Computational experiment results on Bike Share Toronto's 2019 ridership data show that strategic clustering can produce clusters that will lead to better bike demand predictions than clustering by Forward Sortation Areas. In addition, it is discovered that the check-in prediction errors are reduced when the number of transition matrices used for the predictions increases (i.e., greater granularity in transition information). This work motivates further explorations of bike station clustering algorithms and establishes a connection between check-out and check-in predictions using cluster-to-cluster transition information. The cluster-then-predict method presented can also serve as a good starting framework to be extended and improved in the future.

Acknowledgements

I would like to first express my sincere gratitude to Professor Timothy Chan for providing me with the wonderful opportunity to work on an undergraduate thesis project under your supervision. Without this opportunity, I would not have been able to be connected with and be mentored by your amazing PhD students.

In addition, I would also like to express my many thanks to my PhD supervisors Ian Zhu and Bo Lin. Your continuous support, encouragement, and idea sharing not only made this thesis possible, but also made the past year enjoyable. I also really appreciate all the advices that both of you have given me about graduate schools and graduate programs. I will certainly miss our weekly meetings, and I hope that our professional paths will cross again in the future.

Finally, I am very grateful to my family and all my friends who have believed in me and supported me along the way. I want to especially thank two influential people in my life: my best friend Lizzy for always being on my side when I was going through emotional roller coasters, and my sister Siyun for always knowing exactly when to push me to get through challenging times.

Contents

1 Introduction

Bike sharing systems have gained popularity in recent years. Cities around the world have developed their own bike sharing systems, such as Capital Bikeshare in Washington D.C., Citi Bike in New York City, Vélib in Paris, Bicing in Barcelona, Ofo in Beijing, and YouBike in Taiwan. The popularity of bike sharing systems is driven by its positive impacts on public health, environment, and urban mobility [De Hartog et al., 2010, Maizlish et al., 2013, Woodcock et al., 2013]. Toronto has launched a bike sharing system in 2011 named Bike Share Toronto. By 2022, Bike Share Toronto has expanded their operations to provide users access to $6,850$ bikes and 625 stations across the city [Bike Share Toronto, 2022].

Along with the rapid growth of bike sharing systems, there is a growing interest in studying operational problems surrounding such systems. These operational problems include, but not limited to: How many bikes should each station have in stock at the start of each day? How many docks should be allocated to each station? Where should the bike stations be located in the city? On which days of the week and at what hours of the day will there be more demands at certain stations compared to other stations? When and how should the bike sharing systems operators initiate bike reallocation efforts to ensure that there will be enough bikes for users to rent and enough docks for users to return their bikes at each station? In order to answer these questions, it is crucial to obtain information about bike demands, thereby showing that making accurate bike usage predictions is an important first step in tackling operational problems in bike sharing systems.

However, predicting bike usages is not a trivial task since bike usages are typically imbalanced in both time and space. Take one bike station as an example. From the temporal perspective, there may be many more riders renting a bike from a station in the morning than in the afternoon. From the spatial perspective, not all riders who rented a bike from a station will return the bike to the same station. In addition, there are also two different but

closely related types of bike usages: bike rentals and bike returns. In an ideal state, a user looking to rent a bike should be able to check-out a bike from the bike station of interest without finding it empty, while a user looking to return a bike should be able to check-in the bike at the target station without finding it full. This means that two types of bike usage predictions should be made: the number of check-outs and the number of check-ins. In particular, the check-out and check-in predictions can be made separately by developing separate check-out and check-in prediction models, or the two types of predictions can be tied together using the transition patterns between different bike stations. In other words, the latter approach suggests that if the predicted number of check-outs and the probabilities of transition between stations are known, then the predicted number of check-ins can also be computed. Since the number of bikes checked out from some stations will likely affect the number of bikes checked into some other nearby stations, it may be better to relate the check-out and check-in predictions, such as by using transition patterns in the latter approach.

In addition to the different types of bike usages, predictions for bike usages can also be made in three levels of granularity: station-level, cluster-level (aggregate predictions over subsets of stations), and system-level (aggregate predictions over all stations). Station-level predictions may be more irregular and have random fluctuations, causing such predictions extremely difficult to make. This motivates researchers to shift attention to cluster-level and system-level predictions, since these predictions are more robust, more regular, and easier to predict [Kim, 2021]. Although system-level predictions can be the easiest to make, information regarding bike demands at different geographical regions within the whole system is lost. This suggests that system-level predictions may not be very useful in solving downstream operational problems such as the reallocation of bikes between various stations. On the other hand, cluster-level predictions have both sufficient robustness as well as granularity. Not only more regular bike usages can be observed at the cluster-level, cluster-level

predictions can also be useful for supporting studies of bike reallocation between different station clusters. As a result, it has become increasingly popular to examine the cluster-then-predict approach for making bike usage predictions. In such approach, bike stations are first grouped into clusters and then bike usage predictions are made for each cluster. In the clustering step, it may be helpful to group stations with high geographical proximity and similar station-to-station transition patterns into the same cluster. Since bike usages are imbalanced in both time and space, bike usage predictions should take into account of both time and cluster assignment. Lastly, it is also worth to note that bike usage predictions may vary depending on the station-to-cluster assignments, with some assignments leading to smaller prediction errors than other assignments.

While some previous studies have already explored the cluster-then-predict approach, there still remains some gaps in interpreting the quality of bike clustering results as well as understanding the impact of cluster-to-cluster transitions on bike demand predictions. Given these gaps, this thesis project attempts to explore two research questions: 1) how different bike station clusters impact the bike check-out and check-in predictions, and 2) how does transition information between clusters affect bike check-in predictions. To achieve this, we first perform a preliminary bike usage analysis on 2019 Bike Share Toronto data. We then present a bike demand prediction framework consisting of five components: 1) bike station clustering, 2) check-out predictions, 3) transition matrices computation, 4) check-in predictions, and 5) check-out and check-in predictions evaluation. Lastly, we apply the framework on Bike Share Toronto and analyze the prediction results to derive relevant insights.

To the best of our knowledge, this is the first study on bike share usage prediction in the City of Toronto. The results can provide some preliminary insights on bike sharing systems usages and serve as a basis for the development of more sophisticated algorithms to support bike sharing system infrastructure planning and operational decision-making, such as improving

the allocation of docks to stations [Freund et al., 2019].

The rest of the thesis is structured as follows. Section 2 reviews previous studies on bike sharing systems around the world. Section 3 analyzes Bike Share Toronto data. Section 4 presents the bike station clustering and usage prediction framework. The framework is then applied to Bike Share Toronto with experiments and results summarized in Section 5. Lastly, Section 6 provides a summary of the key conclusions from the thesis project and Section 7 recommends some potential future directions.

2 Literature Review

Previous literature studied various operational problems associated with bike sharing systems around the world. These studies predominantly focused on bike station clustering, bike usage predictions, traffic flows analysis, bike-to-station reallocation, dock-to-station reallocation, and bike sharing systems incentive programs development. In particular, this thesis project examines the check-out and check-in predictions when bike stations are clustered using the state-of-the-art clustering algorithm and attempts to relate check-out and check-in predictions.

In a bike sharing system, predicting bike usages (or demands) accurately is an especially important first step in helping decision-makers to better plan system operations. Bike demands directly relate to the allocation (or supply) of bikes at each station as well as customer experience. If a bike-sharing system can foresee bike demands with high accuracy, then the system's management team can better respond and adapt quicker to the changing demands. This may be developing good operational strategies to minimize any under-utilization of bikes and any over-demand of bikes at stations. The most obvious approach is to predict bike demands for each bike station (station-level predictions). However, there is also reason to believe that it may be effective to make demand predictions at the cluster-level. For example, Yuan et al. [2014] and Etienne and Latifa [2014] studied urban functional zones (e.g., residential, commercial, recreational zones) in which trip purposes are similar within the same functional zones, and different across different functional zones. It may be inferred from this knowledge that the bike demands at stations in the same zone are similar, and different in different zones. Hence, it is sensible to group bike stations into clusters and predict bike demands at the cluster-level.

In fact, previous studies have already shown some promising evidence that cluster-level demand predictions are more accurate than station-level demand predictions. Both Li et al.

[2015] and Kim [2021] studied the cluster-then-predict approach. In their studies, they proposed hierarchical prediction models to cluster bike stations first, and then predict future bike demands. Li et al. [2015] proposed a Bipartite Clustering Algorithm to iteratively cluster bike stations first based on geographical locations, and then based on temporal transition patterns. Given the station clusters, the demand prediction for each station cluster was made by first training four different models and then allocating bike demands proportionally to each cluster from the entire system-level demand prediction. In the paper, they compared the cluster-level predictions from their proposed method to the station-level predictions, demonstrating that cluster-level predictions are more robust, accurate, and insightful than station-level predictions. However, the iterative process of alternating between spatial and temporal K-Means clustering steps (a heuristic approach) might lead to non-convergence in the algorithm. In addition, performing a large number of iterations and tuning hyperparameters in the algorithm also required high computational time and resources. Another limitation was that the paper did not explore how different clustering results impact or correlate to demand prediction errors. Moreover, Kim [2021] proposed an alternative clustering algorithm that could group bike stations into clusters deterministically given information about bike station locations as well as hourly number of check-outs and check-ins. The main contributions of this clustering algorithm were reducing the computational cost and circumventing the risk of non-convergence. One downside with this algorithm was that the check-out and check-in patterns of the stations did not relate to other stations. In other words, the algorithm did not consider station-to-station transition patterns when clustering stations, hence, potentially overlooking some useful transitioning behaviours that could help generate better station clusters leading to more accurate bike usage predictions.

In addition to making more accurate bike demand predictions, bike station clustering can also be useful in other use cases. Schuijbroek et al. [2017] proposed a cluster-first route-second heuristic to design optimal truck routes within each cluster of self-sufficient stations such

that bike pickups and deliveries could be made to satisfy target bike inventory levels. The clustering algorithm used in this paper, however, is another greedy algorithm that does not guarantee convergence. Additionally, Chen et al. [2016] grouped bike stations into clusters based on time, weather, and special events features for predicting over-demand probabilities. In this paper, the demands were explicitly modelled as probability distributions, which did not take into account the spatial relationships between stations. Given bike station clusters, the number of bike rentals and returns at the cluster-level were estimated and used in Monte Carlo simulation to predict the probability of over-demand for each cluster. Although neither clustering approaches in these two studies were perfect, the use of station clusters beyond traffic demand predictions further motivated the need for improving and developing more efficient and robust bike station clustering algorithms.

To summarize, bike station clustering is useful in various tasks including bike demand prediction, truck route optimization for bike re-allocation, and prediction of bike over-demand probabilities. However, there remains gaps in evaluating the quality of station clusters, and understanding how temporal and spatial transition patterns influence cluster-level check-in predictions. This thesis project attempts to explore these gaps.

3 Bike Share Toronto Data

In this section, a preliminary data analysis is conducted on two data sets from Bike Share Toronto: 2019 ridership data and 2021 station location data. The following sub-sections will describe the temporal, spatial, and spatial-temporal bike usage patterns discovered in the data.

3.1 Temporal Bike Usage Analysis

Bike Share Toronto's ridership data is publicly available on the City of Toronto's Open Data website [Open Data Toronto, 2021]. Since the covid-19 pandemic has likely impacted bike usages across the City of Toronto in 2020 and 2021, the most recent data that is representative of regular bike usage patterns in Toronto would be the ridership data from 2019. For this reason, the focus of this project will be on bike usages in 2019. In particular, the 2019 ridership data includes records with trip duration, start and end station IDs, start and end station names, trip start and end times, bike ID, and user type (either annual member or casual member).

After removing trip records with missing information, the total number of trips recorded in 2019 is 2,439,047. In addition, there are a total of 4,901 bikes and 469 bike stations in operation. The percentages of total trips taken by annual members and casual members are 76% and 24%, respectively. In 2019, the average daily number of uses per bike is around 1.2 times. In terms of trip duration, the riding time is around 12.5 minutes on average and 17 minutes for 75% of the riders. Figure 1 shows the average number of daily uses per bike by month in 2019. It is observed that the average number of daily uses per bike is the highest in the summer months and the lowest for the winter months, and somewhere in between for the spring and fall months.

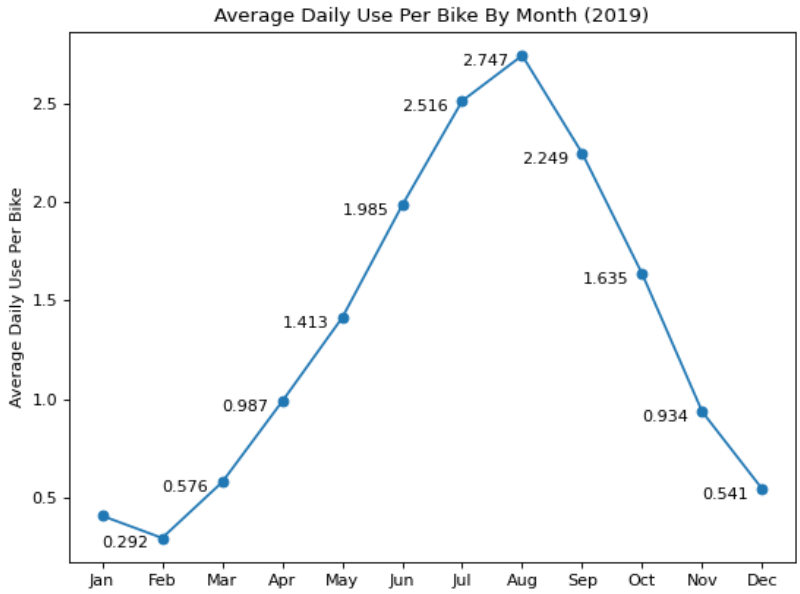


Figure 1: Average Daily Uses per Bike by Month

In addition to the monthly usage patterns, weekly patterns can also be observed in the data, as shown in Figure 2. The 75th percentiles of number of trips taken on weekdays (Monday to Friday) are higher than those on weekends (Saturday and Sunday). However, the mean number of trips taken on weekdays and on weekends are similar.

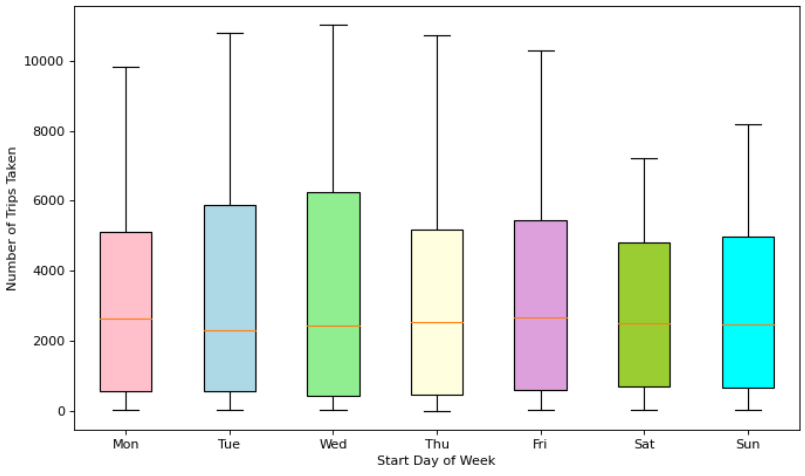


Figure 2: Number of Trips Taken by Day of Week

Lastly, there are also daily bike usage patterns in the data. Figure 3 demonstrates the average number of trips taken by hour, separated by weekdays, weekends, and holidays. For weekdays and weekends, the highest demands are during morning and evening rush hours (7AM-10AM and 3PM-7PM), the lowest demands are in early morning and late night hours (10PM-7AM), and the demands during the remaining hours are somewhere in between. On holidays, early morning and late night hours also have the lowest demands, while the highest demands are during remaining hours. Additional temporal bike usage insights can be found in Appendix A.

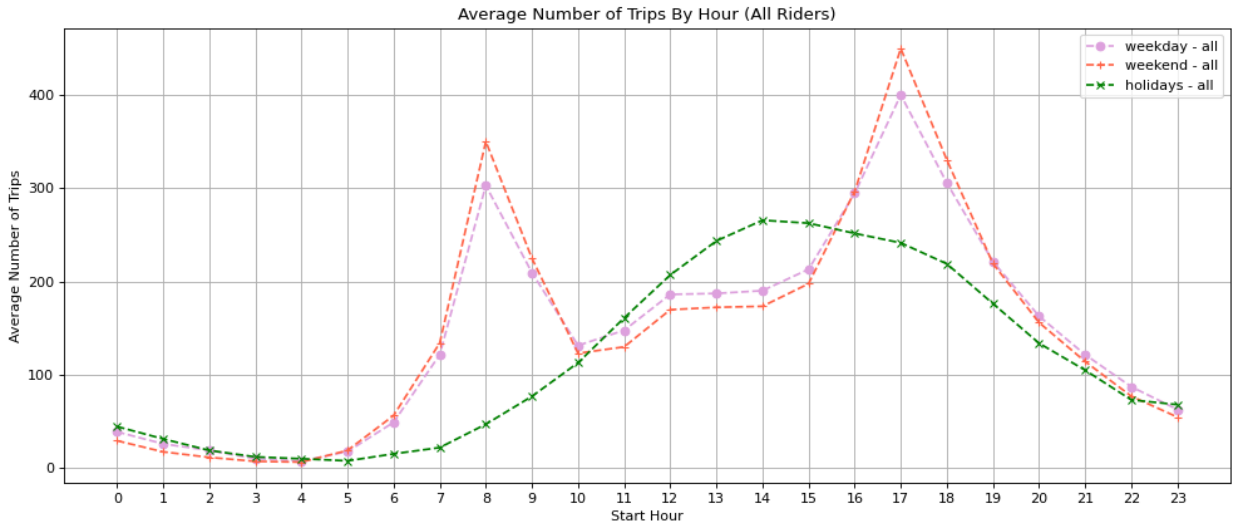


Figure 3: Average Daily Number of Trips by Hour

3.2 Spatial Bike Usage Analysis

Bike Share Toronto updates the station locations periodically and only the most recent up-to-date station locations data is publicly available. For this project, the information about bike station locations is from September 2021. The station location data set contains records of station ID, station name, latitude, longitude, capacity, and address. Since the ridership data is from 2019 while the station location data is from 2021, there are some discrepancies in matching ridership trip records with station location records (i.e., new stations added, old stations removed, station names altered). These discrepancies are resolved by manually matching trip records and station location records based on the station names (i.e., the sta-

tions' nearest traffic intersections), filling in missing latitude and longitude information based on the station names, and removing trip records that cannot be matched. After matching bike station location information with ridership data as well as resolving the discrepancies, the final number of bike stations and number of trip records used in the analysis is reduced from 469 to 464 and from 2,439,047 to 2,428,443, respectively.

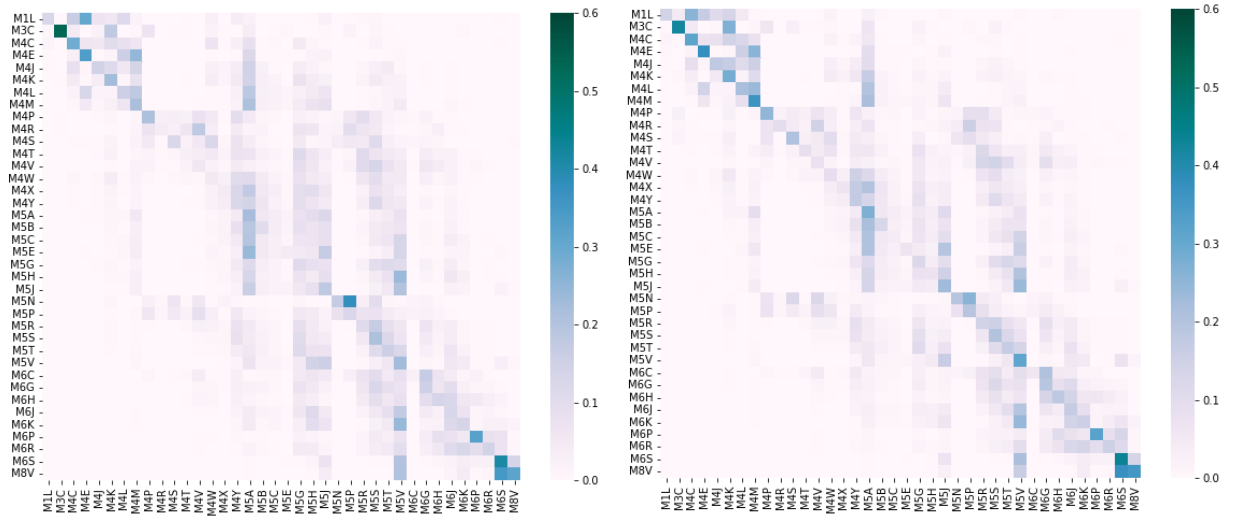
In addition, the City of Toronto can be classified into 38 Forward Sortation Areas (FSAs), given by the first three characters of the postal codes. Each of the first three characters of the postal codes (or an FSA) designates a postal delivery area within Canada. Figure 4 shows the location of the bike stations in the City of Toronto (left) and an example of the geographical distribution of log normalized average daily number of check-outs by FSAs (right). It is observed that during day hours (11AM-3PM) on weekdays, the city-centre has much higher number of check-outs (dark green) than the peripheral areas (dark pink). By visual inspection, similar check-out patterns are observed for other hours of day on weekdays (i.e., morning rush hours, evening rush hours, and night hours). There is also little difference in the geographical distributions between the number of check-outs and the number of check-ins. The same spatial bike usage patterns are observed for weekends.

(a) (b)

Figure 4: (a) Bike Stations in the City of Toronto (b) Log Normalized Average Daily Number of Check-outs (Weekday Day Hours From 11AM To 3PM, Grouped by FSAs): the scale goes from dark pink (smallest numbers) to dark green (largest numbers)

3.3 Spatio-temporal Bike Usage Analysis

Combining the information about trips taken in 2019 and bike station locations, the transition patterns between different FSAs in the City of Toronto can be examined. Figure 5 shows the transition patterns between the 38 FSAs on weekdays and on weekends. Figure 6 shows the transition patterns between the FSAs by hour. The transition patterns are described by the average daily number of check-outs (normalized to between 0 and 1). For example, the normalized average daily number of check-outs from "M1L" (one FSA) to "M4C" (another FSA) on weekends is 0.26. Comparing the heatmaps, the transition patterns are similar between weekdays and weekends, but different for various hours of day. Some pronounced differences in transition patterns between different hours of day are highlighted in red in Figure 6. For instance, the transition patterns in the top left corner vary for Hours 0, 6, and 12. In addition, transitions appear random in Hours 4 through 9, but a distinguishable diagonal line can be observed in the heatmaps for before Hour 3 and after Hour 10.



(a) (b)

Figure 5: (a) Transition Patterns Between FSAs On Weekdays (b) Transition Patterns Between FSAs On Weekends

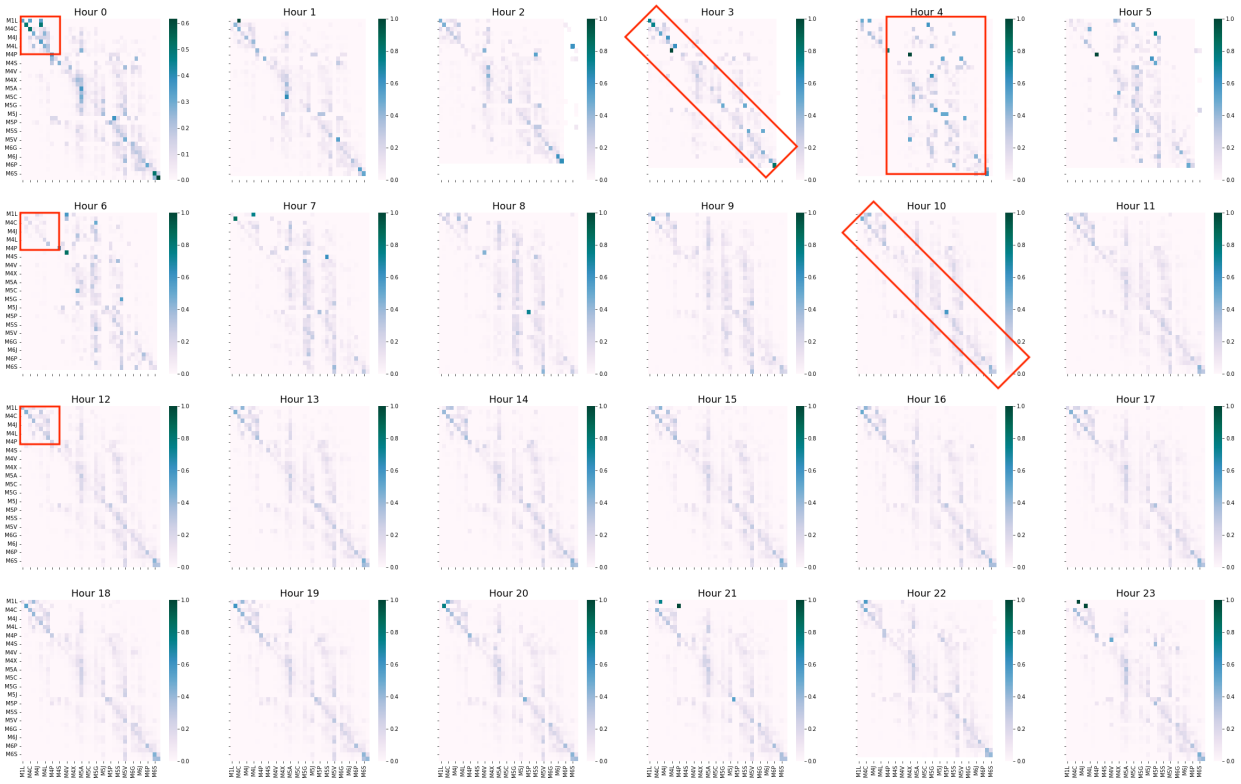


Figure 6: Transition Patterns Between FSAs By Hour

4 Bike Station Clustering and Usage Prediction

This section presents a cluster-then-predict framework for bike usage predictions. As presented in Figure 7, this framework consists of five steps: 1) bike station clustering, 2) cluster-level check-out prediction, 3) cluster-to-cluster transition matrix computation, 4) cluster-level check-in prediction, and 5) evaluation. In the following sub-sections, each of these steps is introduced in detail.

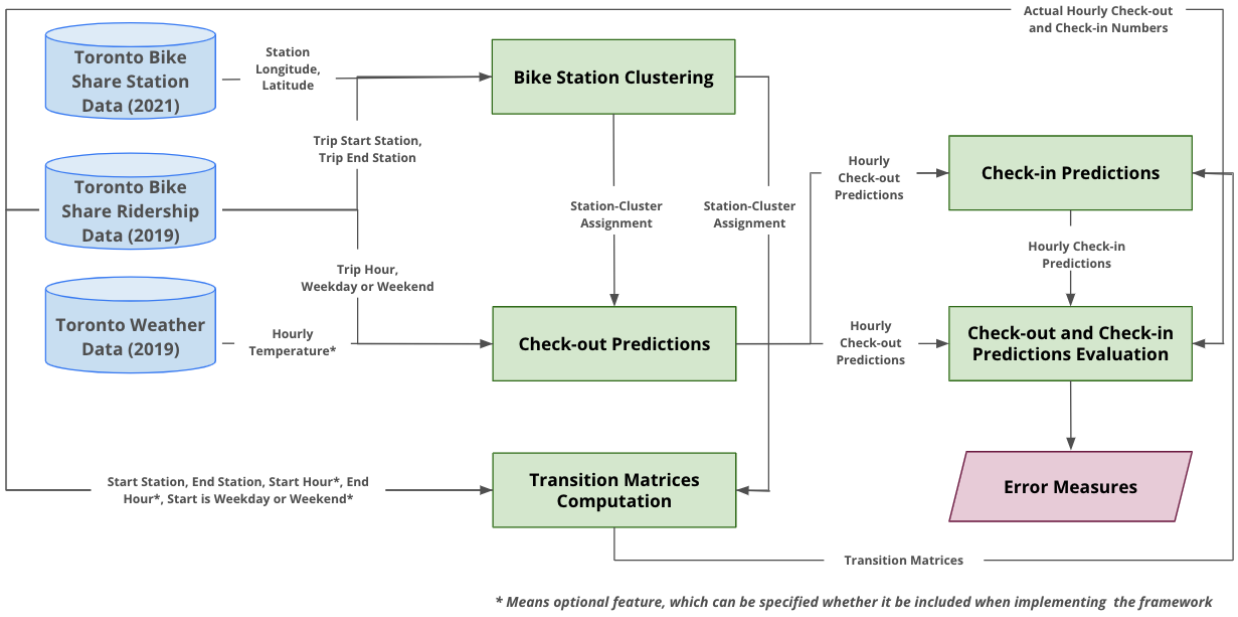


Figure 7: Cluster-Then-Predict Framework

4.1 Bike Station Clustering

From literature review, one major motivation for clustering bike stations is to predict bike demands for clusters of stations. This is because cluster-level bike demand predictions are more robust and accurate than station-level bike demand predictions. To generate bike station clusters, two factors should be considered: 1) the geographical locations of the bike stations, to ensure that the bike stations within the same cluster are close to one another geographically for users' convenience, and 2) the transition patterns of the bike stations, to

ensure stations within the same cluster share similar transitions patterns to other clusters. In this thesis, we adopt the Bipartite Clustering Algorithm proposed by Li et al. [2015], which takes both of these two factors into consideration. It is an iterative heuristic clustering algorithm that alternates between 1) clustering stations using K-means based on the stations' geographical locations, and 2) clustering stations using K-means based on stations' transition patterns to clusters obtained in the previous step. The algorithm terminates when either station-to-cluster assignments no longer change (the algorithm has converged) or the maximum number of iterations (W) has reached. The pseudocode for the Bipartite Clustering Algorithm is presented in Algorithm 1 with relevant notations in Table 1.

During implementation, the longitudes and latitudes of the bike stations are used as features for the geographical clustering portion of the algorithm. In addition, one transition matrix is computed for each station (t-matrix), and the transition matrices from all stations are used as features for the transition patterns clustering portion of the algorithm. Each transition matrix ($M_i \in \mathbb{R}^{p \times m}$) represents a bike station's transition pattern to m clusters (generated from clustering with stations' geographical locations) during p time periods (i.e., weekday morning peak hours, weekday day hours, weekday evening peak hours, weekday night hours, weekend morning peak hours, weekend day hours, weekend evening peak hours, and weekend night hours).

Notation	**Description**
S_i	The i^{th} station
n	Number of stations
Tr	Trip, Tr = (S_o, S_d, τ_o, τ_d)
S_o / S_d	Origin / Destination station, consist of latitude and longitude
τ_o / τ_d	The time when the bike is checked out at S_o / checked in at S_d
$C_{2,j} \in \mathbb{R}^{n \times m}$	The j^{th} cluster

Table 1: Notations for Algorithm 1

Algorithm 1 Bipartite Clustering Algorithm (BC Algorithm)

Input: Stations $\{S_i\}_{i=1}^{n}$, historical trips $\{Tr_i\}_{i=1}^{H}$, iteration threshold W, parameters $K_1 > K_2$

Output: K_1 clusters: $C_{1,1}, C_{1,2}, ..., C_{1,K_1}$

Cluster $\{S_i\}_{i=1}^{n}$ into K_1 clusters: $C_{1,1}, C_{1,2}, ..., C_{1,K_1}$ by K-means clustering based on stations' geographical locations;
Initialize $w = 0$;
while $w < W$ **do** ▷ Continue to increment k until iteration threshold W
    for i = 1:n **do** ▷ Iterate through n stations
        Generate t-matrix M_i of station S_i; ▷ Describes a station's transition pattern
    Cluster $\{S_i\}_{i=1}^{n}$ into K_2 clusters: $C_{2,1}, C_{2,2}, ..., C_{2,K_2}$ by K-means based on $\{M_i\}_{i=1}^{n}$
    for j = 1:K_2 **do**
        Cluster stations in $C_{2,j}$ into $[\frac{N_j K_1}{n}]$ groups; ▷ Geo-clustering proportionally
    Obtain K_1 updated clusters: $C_{1,1}, C_{1,2}, ..., C_{1,K_1}$
    if $C_{1,1}, C_{1,2}, ..., C_{1,K_1}$ **then** do not change ▷ K_1 clusters converge
        Break;
    w = w+1;
return K_1 clusters: $C_{1,1}, C_{1,2}, ..., C_{1,K_1}$

4.2 Cluster-level Check-out Prediction

In the bike demand prediction framework, the number of check-outs from each cluster at each hour is predicted by linear regression, as shown in equation (1).

$$\hat{y}_{(t,c)} = \beta_0 + \beta_1 x_{1,(t,c)} + \beta_2 x_{2,(t,c)} + ... + \beta_k x_{k,(t,c)} \tag{1}$$

where $x_{i,(t,c)}$ variables are features corresponding to time period t and cluster c, $\hat{y}_{(t,c)}$ are target variables representing the predicted number of check-outs for time period t and cluster c, and $y_{(t,c)}$ are the actual number of check-outs (ground truths). Each time period t is a tuple of (h, w), where h represents the hour and w represents whether it is a weekday or a weekend.

Initially, three factors are considered in the check-out predictions: hour of the day, weekday or weekend, and cluster assignment. In particular, the number of trips that users take will differ throughout the day depending on what hour of day it is and differ throughout the week

depending whether it is a weekday or a weekend. The number of trips for different clusters also varies as the clusters of stations are located in different geographical locations within the city. For instance, clusters in the city centre may have more number of check-outs than clusters outside of the city centre. During implementation, categorical features (hour of the day, weekday or weekend, and cluster assignment) are converted to dummy variables using one-hot encoding.

In addition, it was discovered previously in Section 3.1 that bike demands varied throughout the year. Bike demands are the highest in the summer months, the lowest for winter months, and somewhere in between for the spring and fall months. This motivates the need for an additional feature in the model to capture the seasonality trend in bike demands. After analyzing Toronto's weather data in 2019 [Government of Canada, 2021], it is found that the varying bike demand patterns (Figure 1) match with the temperature patterns throughout the year (Figure 8). There will be higher bike share usages when it is warmer (summer months) and lower bike share usages when it is colder (winter months). We therefore introduce temperature as the fourth feature for bike usage predictions.

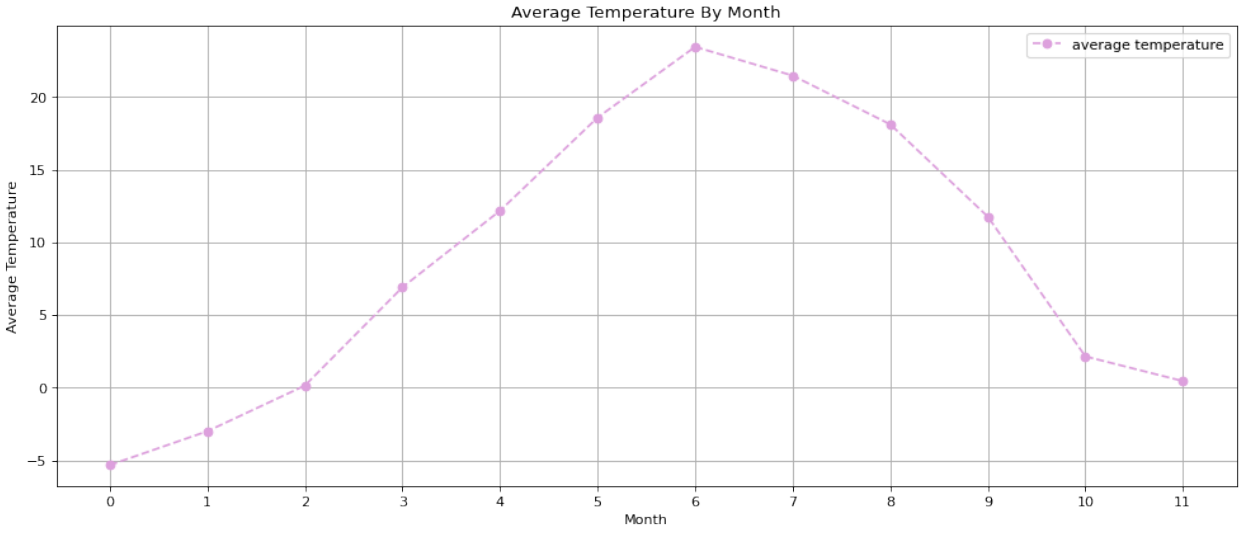


Figure 8: Average Temperature in Toronto by Month (2019)

4.3 Cluster-to-Cluster Transition Matrix Computation

After predicting the number of check-outs, we relate the number of check-outs to the number of check-ins using transition patterns between different bike station clusters, modelled by transition matrices. However, generating one transition matrix for each hour of each day in a year will be computationally expensive and unnecessary, as individual hour and day transition matrix can be unstable and less generalizable to unseen data. Similar to the idea of predicting bike demands in clusters, the transition patterns of bike station clusters can also be aggregated and used as a summary statistic to model the transitions between clusters. There is a trade-off between the degree of aggregation, the computational cost, and the accuracy of downstream check-in predictions. A higher degree of aggregation (generating fewer transition matrices) will be easier and faster to compute, but may be too simple for capturing sufficient transition behaviours between clusters, and vice versa. For instance, we can aggregate all training data by hour and compute 24 transition matrices representing the transition patterns for each hour of the day. This ensures that the time of day is considered as part of the transition patterns, but information about different transition patterns between weekday and weekend may be lost.

Assume that we are interested in using D transition matrices in the cluster-then-predict framework, the transition patterns between clusters at period $d \in D$ are represented by a square transition matrix (A_d). In A_d, each entry ($a_{i,j}$) represents the number of trips that starts from cluster i and ends at cluster j. These number of trips are normalized across the rows such that the sum of the values in each row will be one, as illustrated in equation (2), where I is the set of all starting clusters and J is the set of all ending clusters. In other words, each entry in the transition matrix can be considered as the probability of starting a trip in cluster i and ending the trip in cluster j.

$$\sum_{j \in J} a_{i,j} = 1 \qquad \forall i \in I \tag{2}$$

4.4 Cluster-level Check-in Prediction

Furthermore, we make a check-in prediction for each time period $t \in T$ and for each cluster $j \in J$, where time period t is a tuple of (h, w) with h representing the hour and w representing whether it is a weekday or a weekend. T is the set of all time periods and J is a set of all ending clusters or all the clusters in which bikes can be checked into. Specifically, for each t (hour, weekday or weekend) and j (cluster) combination, the number of check-ins is predicted based on: 1) $\hat{y}_{(t,c)}$, the check-out predictions from Section 4.2, and 2) $a^d_{i,j}$, the transition patterns given by A_d from Section 4.3. A_d denotes the transition patterns between clusters at period d. For instance, $d = 0, 1, ..., 23$ for using 24 transition matrices (one for each hour of the day) to relate the number of check-outs and check-ins. When predicting the number of check-ins for any cluster at Hour 5, for example, the transition patterns from A_5 with entries of $a^5_{i,j}$ will be used. Equation (3) shows the computation of check-in predictions, denoted by $\hat{z}_{t,j}$.

$$\hat{z}_{t,j} = \sum_{c \in C, i=c} \hat{y}_{(t,c)} a^d_{i,j} \tag{3}$$

As an illustration, Figure 9 shows an example of check-in prediction using one transition matrix ($d = 1$) following steps labelled from 1 to 3. In this example, there are 38 clusters in total, labelled from 0 to 37. In the final step (step 3), the sum of the products circled in blue is the predicted number of check-ins for Cluster 0 at (Hour 0, Weekdays).

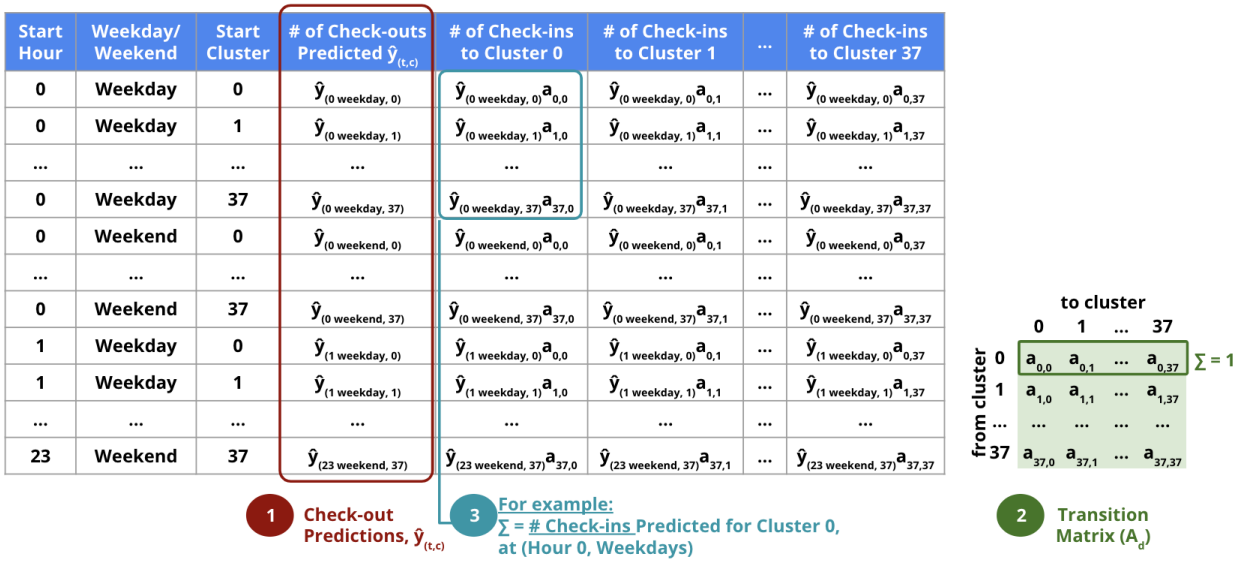


Figure 9: Example of Predicting Number of Check-ins with One Transition Matrix

4.5 Check-out and Check-in Predictions Evaluation

Finally, four error measures are computed to evaluate the check-out and check-in predictions from Sections 4.2 and 4.4, namely mean squared error (MSE), root mean squared error (RMSE), maximum residual error (MRE), and mean absolute error (MAE). The equations for the error measures can be found in (4), where r_i is the actual value and $\hat{r}_i$ is the predicted value. In particular, n denotes the total number of actual versus predicted values in evaluation. For check-out predictions, each i represents a tuple of (t, c) with $t = (h, w)$. For check-in predictions, each i represents a tuple of (t, j) with $t = (h, w)$.

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2 \quad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2} \\ MRE &= \max|r_i - \hat{r}_i| \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}|r_i - \hat{r}_i| \end{aligned} \tag{4}$$

5 Application to Bike Share Toronto

This section applies the Bike Station Clustering and Usage Prediction framework from Section 4 on Bike Share Toronto data from Section 3. Computational experiments are conducted to analyze: 1) the impact of different station-to-cluster assignments on the check-out and check-in predictions, and 2) the relationship between cluster-to-cluster transitions and check-in predictions.

5.1 Computational Experiments

Bike Share Toronto's 2019 ridership data is used to conduct computational experiments. First of all, the data is split into training and testing sets, where the training set consists of data from January to September and the testing set consists of data from October to December. In particular, the training set is used to train the check-out prediction model and generate the transition matrix (or matrices). The testing set is used to predict the number of check-outs and the number of check-ins, as well as evaluating the predictions using the four error measures described in Section 4.5.

The cluster-then-predict framework from Section 4 is ran using different combinations of parameter specifications in terms of $W \in \{1, 5, 10, 15, 20\}$ (the maximum number of iterations in Algorithm 1), $K2 \in \{8, 12, 16\}$ (the number of intermediary clusters to generate in Algorithm 1), and temperature feature (include or exclude in check-out and check-in predictions). It is worth to note that W can be considered as the number of iterations instead because it is found that the clustering algorithm will not converge after completing 20 iterations. Moreover, the number of check-ins also depends on the transition patterns between different clusters, thus, the number of transition matrices to use is also specified (1 for a single transition matrix, 24 for hourly transition matrices, 48 for hourly and weekday/weekend transition matrices). Table 2 summarizes the parameters used in the computational exper-

iments. In addition, a brief description of the code for the computational experiments is included in Appendix B.

Parameters	Possible Values				
W	1	5	10	15	20
K2	8	12	16		
n Transition Matrices*	1	24	48		
Temperature Feature	Yes	No			

Table 2: Parameters Specifications For Computational Experiments

5.2 FSA Clusters and BC Algorithm Clusters Comparison

As a baseline clustering method, bike stations are grouped into 38 clusters by Forward Sortation Areas (FSAs). Figure 10 shows the MSE values for check-out and check-in predictions with 1, 24, and 48 transition matrices. The closer it is to the bottom left corner on the scatter plot, the better the check-out and check-in predictions. It is observed that regardless of the number of transition matrices and regardless of whether the temperature feature is included in the prediction models, clusters generated using the Bipartite Clustering (BC) Algorithm (Algorithm 1) always result in better performance in terms of MSE compared to clusters based on FSAs. Similar results are obtained for other error measures (RMSE, MRE, and MAE). By comparing the prediction errors from using FSA clusters and using BC Algorithm clusters, it can be concluded that different station-to-cluster assignments will lead to different prediction performances. There is also evidence that strategic clustering by using both geographical locations of stations and transition patterns can be help improve check-out and check-in predictions. This motivates the need to develop clustering algorithms (e.g., the BC Algorithm) that can group bike stations such that the check-out and check-in prediction errors can be reduced or even minimized. One limitation with the BC Algorithm, however, is that it takes approximately 6 minutes to complete one iteration of the algorithm with 464

bike stations, which is very computationally expensive. Therefore, it is also important to watch out for the computational cost when developing alternative clustering algorithms.

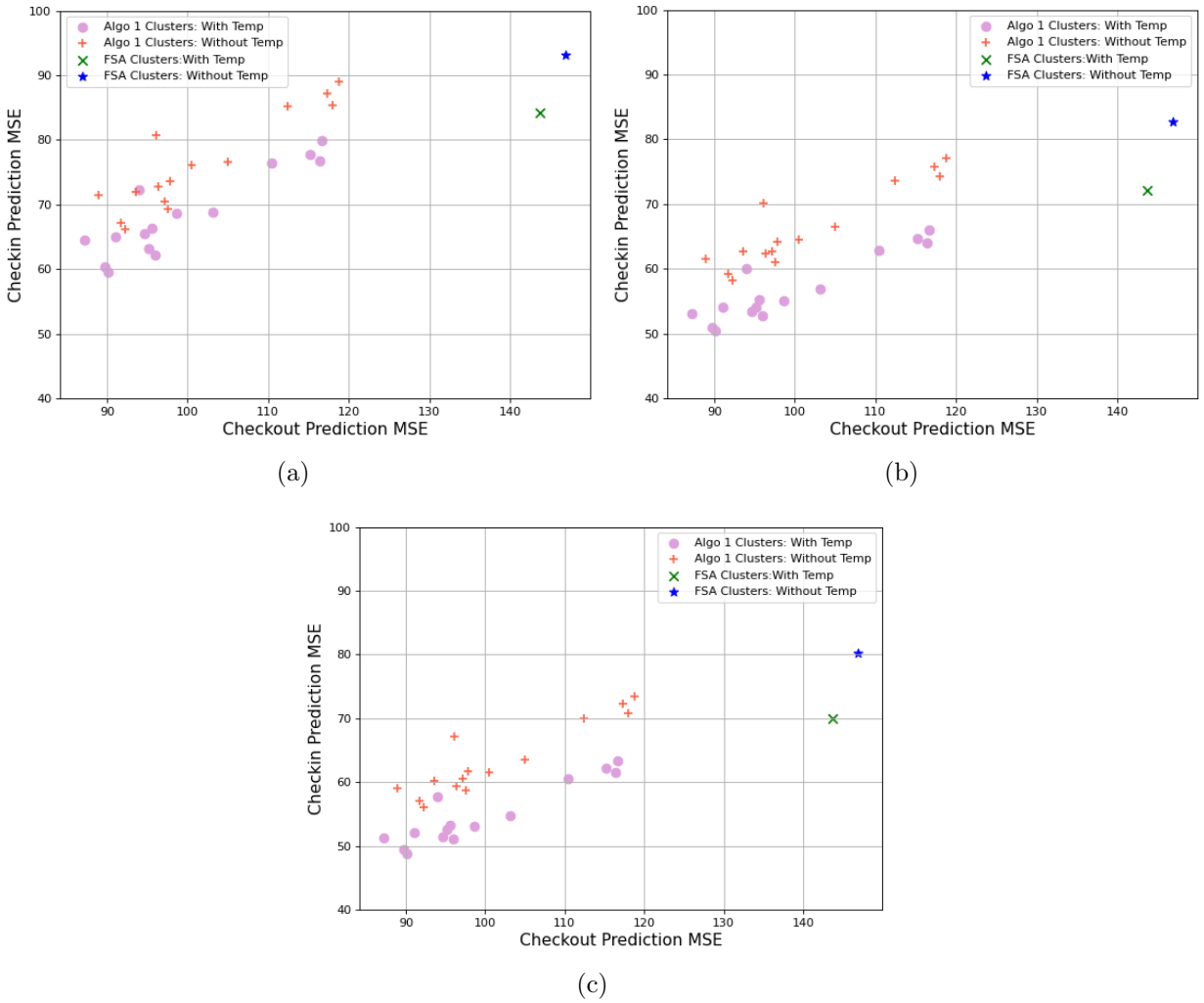


Figure 10: Comparison of MSE Values When Using FSA Clusters versus BC Algorithm Clusters with (a) 1 Transition Matrix, (b) 24 Transition Matrices, (c) 48 Transition Matrices

5.3 BC Algorithm Hyperparameters

This subsection examines the check-out and check-in prediction results separately, and compares the check-out and check-in prediction errors with respect to varying $K2$ (the number of intermediary clusters to generate) and W (the number of iterations) in Algorithm 1.

5.3.1 Check-out Predictions

First, Figure 11 shows the check-out prediction errors (MSE, RMSE, MRE, and MAE) for various W and various $K2$ values, with temperature feature included in the prediction models. Removing the temperature feature leads to the same observations as below.

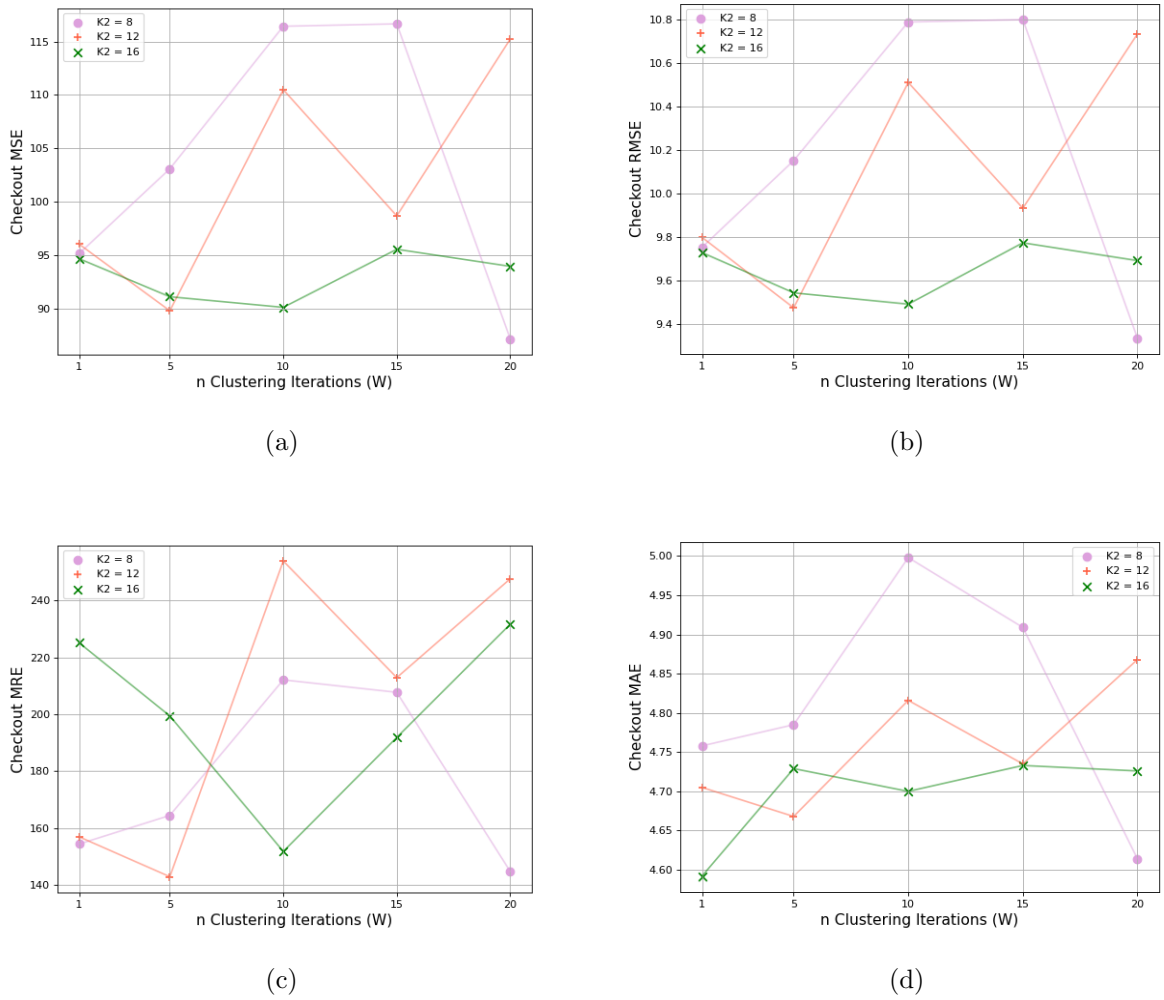


Figure 11: Comparison of Check-out Prediction Performance in Terms of (a) MSE, (b) RMSE, (c) MRE, and (d) MAE with Various BC Algorithm Hyperparmeters, with Temperature Feature

Comparing the check-out prediction errors for fixed W and varying $K2$, there is no correlation between $K2$ and the prediction errors (MSE, RMSE, MRE, or MAE). In other words, increasing $K2$ neither monotonically increase nor monotonically decrease the check-out pre-

diction errors. Take check-out MSE values for $W = 5$ and $W = 10$ as examples. If $W = 5$, then the ranking of $K2$ values in the order of worst to best performance in terms of MSE is $K2 = 8, 16, 12$. If $W = 10$, then the ranking of $K2$ values in the order of worst to best performance in terms of MSE is $K2 = 8, 12, 16$.

Comparing the check-out prediction errors for varying W and fixed $K2$, there is no correlation between W and the prediction errors (MSE, RMSE, MRE, or MAE) either. Take $K2 = 8$ in the MSE figure as an example. As W increases, the check-out MSE values neither monotonically increase nor monotonically decrease.

5.3.2 Check-in Predictions

Likewise, Figure 12 shows the check-in prediction errors (MSE, RMSE, MRE, and MAE) for various W and various $K2$ values, with 24 transition matrices and temperature feature included in the prediction models. There is also no correlation between the check-in prediction errors and W or $K2$. In addition, check-in predictions also depend on the number of transition matrices used in the cluster-then-predict framework. Similar to the results from using 24 transition matrices, the check-in prediction errors do not correlate to W or $K2$ when making predictions with 1 or 48 transition matrices. The conclusions remain the same when the temperature feature is removed from the prediction model.

Intuitively, it might be expected that the quality of station clusters will improve as the number of iterations (W) increases, hence, leading to better predictions in terms of both the number of check-outs and check-ins. It might also be expected that the quality of station clusters will improve when there is either an increase or a decrease in the number of intermediary clusters generated ($K2$) in the BC algorithm. However, neither of these hypotheses are true, providing evidence that the quality of station clusters generated by the BC algorithm may be unstable and more investigation is needed to improve the clustering algorithm.

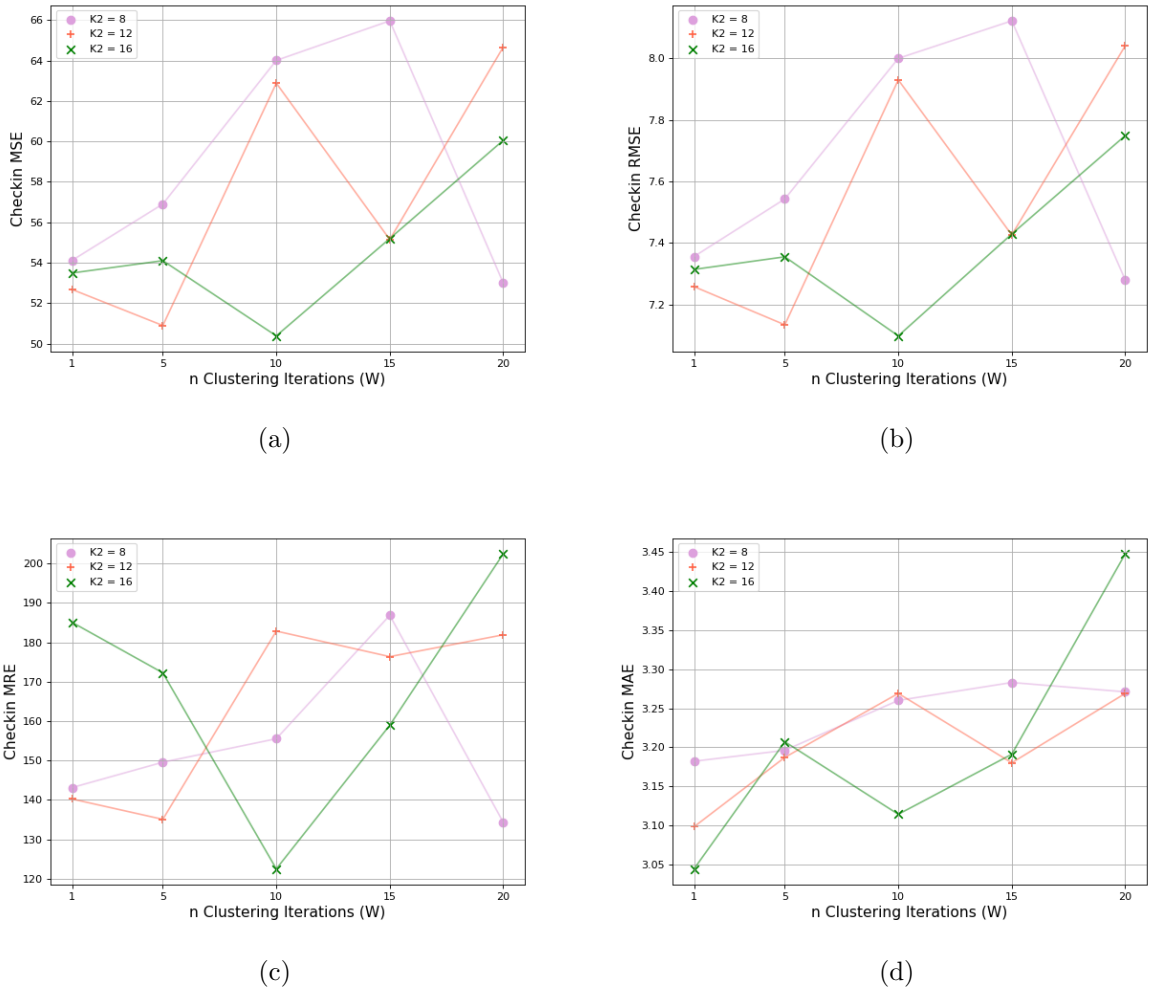


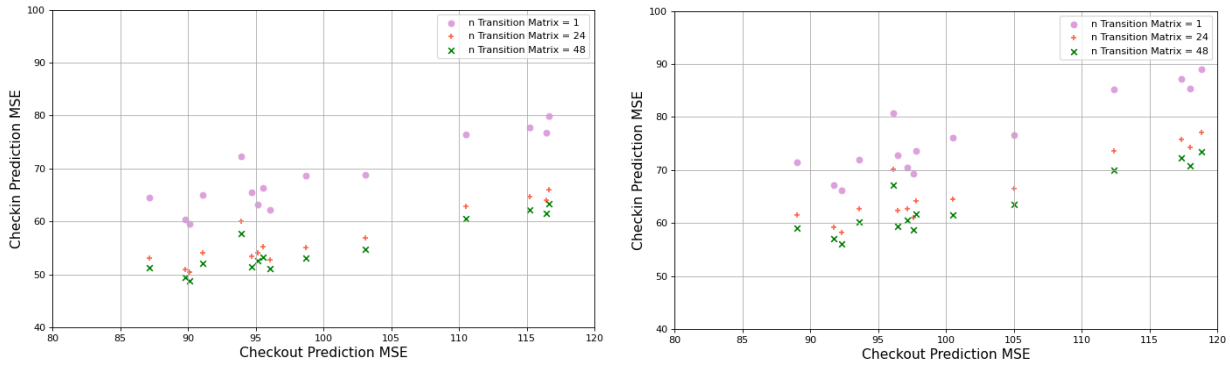
Figure 12: Comparison of Check-in Prediction Performance in Terms of (a) MSE, (b) RMSE, (c) MRE, and (d) MAE with Various BC Algorithm Hyperparmeters, with 24 Transition Matrices and Temperature Feature

5.4 Number of Transition Matrices

In addition to examining the impact of different station clusters to bike usage predictions, we also attempt to relate the number of check-outs and the number of check-ins by leveraging transition patterns between different clusters. Figure 13 shows the MSE for check-out and check-in predictions with different number of transition matrices. Similar results are obtained for other error measures (RMSE, MRE, and MAE). For both with and without the

temperature feature, MSE decreases as the number of transition matrices increases. There is a large decrease in MSE when the number of transition matrices is increased from 1 to 24 (one for each hour), but there is only a very small decrease in MSE when the number of transition matrices is increased from 24 to 48 (one for each hour and weekday/weekend). This is consistent with the results from the spatio-temporal bike usage analysis in Section 3.3, where similar transition patterns are observed during weekdays and weekends (Figure 5), and different transition patterns are observed during various hours of day (Figure 6). To balance between higher computational cost and lower check-in prediction errors from generating more transition matrices, it may be the most suitable to use hourly transition patterns to relate the number of check-outs and check-ins. Hourly transition patterns provide sufficient granularity in capturing important transition behaviours between clusters while ensuring that it would not be too costly to compute.

Furthermore, there is a distinguishable diagonal pattern between check-out and check-in MSE values. The positioning of MSE values along the diagonal indicates that the check-in predictions are positively correlated with the check-out predictions. This means that better check-out predictions will lead to better check-in predictions. Lastly, it is also observed that there is a shift from the bottom left to the upper right for the error values when the temperature feature is removed. This demonstrates that having the temperature feature in the prediction models will reduce the prediction errors for both the number of check-outs and check-ins, validating that the addition of this feature is useful, as described in Section 4.2.



(a) (b)

Figure 13: Check-out and Check-in Prediction MSE Values For Various Number of Transition Matrices (a) With Temperature Feature, (b) Without Temperature Feature

6 Conclusion

This thesis project explores two problems of interest: 1) the impact of different station-to-cluster assignments on check-out and check-in predictions, and 2) the impact of cluster-to-cluster transition patterns on check-in predictions. In particular, a cluster-then-predict framework is developed, which consists of five components: 1) bike station clustering using the Bipartite Clustering (BC) Algorithm, 2) cluster-level check-out prediction, 3) cluster-to-cluster transition matrix computation, 4) cluster-level check-in prediction, and 5) predictions evaluation. Computational experiments are conducted by applying the framework on Bike Share Toronto's 2019 ridership data. In these experiments, various combinations parameters values are tested to derive relevant insights.

First, it is discovered that strategic clustering of bike stations will lead to better check-out and check-in predictions. Comparing the check-out and check-in predictions based on clusters created by Forward Sortation Areas (FSAs) and those generated from the BC Algorithm (strategic clustering), BC Algorithm clusters outperform FSA clusters in terms of prediction errors. This motivates the need to study and develop strategic clustering algorithms that take into account of not only the stations' geographical locations, but also their transition patterns. For the BC Algorithm, it is also discovered there is no correlation between the algorithm's hyperparameters (i.e., W, $K2$) and the check-out (or check-in) predictions. This provides evidence that the quality of clusters generated by the BC Algorithm may be unstable. In other words, random fluctuations are observed in the check-out and check-in predictions from various station-to-cluster assignments given by the BC Algorithm. In addition, the computational time for one iteration of the BC Algorithm is approximately 6 minutes, showing that the algorithm can be computationally expensive. Therefore, more investigations may be needed to improve upon the state-of-the-art bike station clustering algorithms (e.g., the BC Algorithm).

Furthermore, check-in predictions are also influenced by cluster-to-cluster transition patterns. It is observed that when the number of transition matrices used for check-in predictions increases (i.e., greater granularity in transition information), the check-in prediction errors decrease. One watch-out is that there may be a point of diminishing returns (e.g., from 24 to 48 matrices), where every additional transition matrix results in a smaller reduction in the check-in prediction errors. Moreover, it is found that better check-out predictions will lead to better check-in predictions. This suggests that the number of check-outs and the number of check-ins are related. In fact, the use of transition patterns (modelled by transition matrices) in the cluster-then-predict framework is effective in establishing a clear connection between the number of check-outs and the number of check-ins. Finally, it is also observed that adding temperature feature in the demand prediction models will improve both the check-out and check-in predictions.

In conclusion, this thesis project examines the quality of bike station clusters generated from a state-of-the-art clustering algorithm and establishes the motivation for further exploration of alternative clustering algorithms. There is also evidence that information about cluster-to-cluster transitions will directly affect check-in predictions, which in turn establishes a clear connection between check-out and check-in predictions. These results provide some preliminary insights for supporting further investigations of bike station clustering and usage prediction methods. In addition, the cluster-then-predict approach presented in this project can also serve as a starting framework to be improved in the future by integrating alternative clustering algorithms and/or prediction models.

7 Future Work

Although the cluster-then-predict framework presented in this thesis project is useful in making cluster-level bike usage predictions and relating check-out and check-in predictions using cluster-to-cluster transition patterns, more investigations can be made on the basis of this framework in the future.

First, more explorations can be made on each of the five parts in the framework. For bike station clustering, clustering algorithms from other papers could be validated. Alternative clustering algorithms could also be developed to minimize the computational time, and to improve the stability and quality of clustering results. After bike station clustering, the current framework uses linear regression for making check-out predictions. In the future, other linear and non-linear prediction models should be explored to compare and contrast the prediction results as well as discovering whether alternative models are more suitable for fitting the Bike Share Toronto data. In addition, the cluster-to-cluster transition patterns are described by transition matrices, which are computed based on the number of check-outs and check-ins in the training set. Future work can further examine the impact of varying the number of transition matrices in addition to the ones tested in this thesis project (1, 24, and 48). Alternative methods of representing the transition matrices can also be explored. Moreover, the framework relates the check-in predictions with check-out predictions by using a simple relationship of multiplying the predicted number of check-outs by the cluster-to-cluster transition probabilities. Other potential relationships between check-out and check-in predictions could be studied. Lastly, the check-out and check-in predictions for the entire testing set are evaluated using error measures of MSE, RMSE, MRE, and MAE. Additional performance measures can be developed to measure the prediction errors in either greater granularity or in alternative ways, for example, using error matrices that represent the difference between the check-in predictions from Day x (e.g., Day 1) and the mean check-in predictions over the entire time horizon of the testing data set.

In addition to further investigations on the five components of the cluster-then-predict framework, it might also be helpful to use additional publicly available data sets in bike sharing systems studies, such as population data and public transits data. Population data may be useful in creating user profiles as well as establishing possible connections between the user profiles and bike share usage behaviours. Public transits data (e.g., subway data, car sharing data) may be useful in discovering potential underlying reasons that people may choose one mode of transport over another. It may also be helpful to analyze the impact of jointly having multiple modes of transport in urban cities. For instance, a study combining data about population, a bike sharing system, and other modes of transport could be conducted to derive urban mobility insights and support urban planners' decision-making.

References

Bike Share Toronto. Bike share toronto. https://bikesharetoronto.com, 2022. Accessed: 2022-03-30.

Longbiao Chen, Daqing Zhang, Leye Wang, Dingqi Yang, Xiaojuan Ma, Shijian Li, Zhaohui Wu, Gang Pan, Thi-Mai-Trang Nguyen, and Jérémie Jakubowicz. Dynamic cluster-based over-demand prediction in bike sharing systems. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 841–852, 2016.

Jeroen Johan De Hartog, Hanna Boogaard, Hans Nijland, and Gerard Hoek. Do the health benefits of cycling outweigh the risks? *Environmental health perspectives*, 118(8):1109–1116, 2010.

Côme Etienne and Oukhellou Latifa. Model-based count series clustering for bike sharing system usage mining: a case study with the vélib'system of paris. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–21, 2014.

Daniel Freund, Shane G Henderson, Eoin O'Mahony, and David B Shmoys. Analytics and bikes: Riding tandem with motivate to improve mobility. *INFORMS Journal on Applied Analytics*, 49(5):310–323, 2019.

Government of Canada. Historial weather data. https://climate.weather.gc.ca/historical_data/search_historic_data_e.html, 2021. Accessed: 2022-01-30.

Kyoungok Kim. Spatial contiguity-constrained hierarchical clustering for traffic prediction in bike sharing systems. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2015.

Neil Maizlish, James Woodcock, Sean Co, Bart Ostro, Amir Fanai, and David Fairley. Health cobenefits and transportation-related reductions in greenhouse gas emissions in the san francisco bay area. *American journal of public health*, 103(4):703–709, 2013.

Open Data Toronto. Bike share toronto ridership data. https://open.toronto.ca/dataset/bike-share-toronto-ridership-data, 2021. Accessed: 2021-09-20.

Jasper Schuijbroek, Robert C Hampshire, and W-J Van Hoeve. Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research*, 257 (3):992–1004, 2017.

James Woodcock, Moshe Givoni, and Andrei Scott Morgan. Health impact modelling of active travel visions for england and wales using an integrated transport and health impact modelling tool (ithim). *PloS one*, 8(1):e51462, 2013.

Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):712–725, 2014.

A Additional Bike Share Toronto Insights

Two additional interesting insights are obtained when conducting temporal bike usage analysis of the Bike Share Toronto 2019 ridership data. In particular, there are two types of bike users: annual members who have signed up for an annual subscription and casual members who have purchased a single-trip, a one-day pass, or a three-day pass. Figure 14 shows the average number of trips by hour for annual members and casual members during weekdays, weekends, and holidays. It is discovered that on weekdays and weekends for annual members, the demands are the highest during morning and evening rush hours, the demands are the lowest during early morning and late night hours, and the demands are somewhat in between during the afternoon hours. During holidays for annual members and during any time (weekdays, weekends, and holidays) for casual members, no observation of morning and evening rush hours peak demands is made. The afternoon hours demands are the highest while early mornings and late nights have the lowest demands. For casual members, the demands during weekdays and weekends are lower than the demands during holidays.

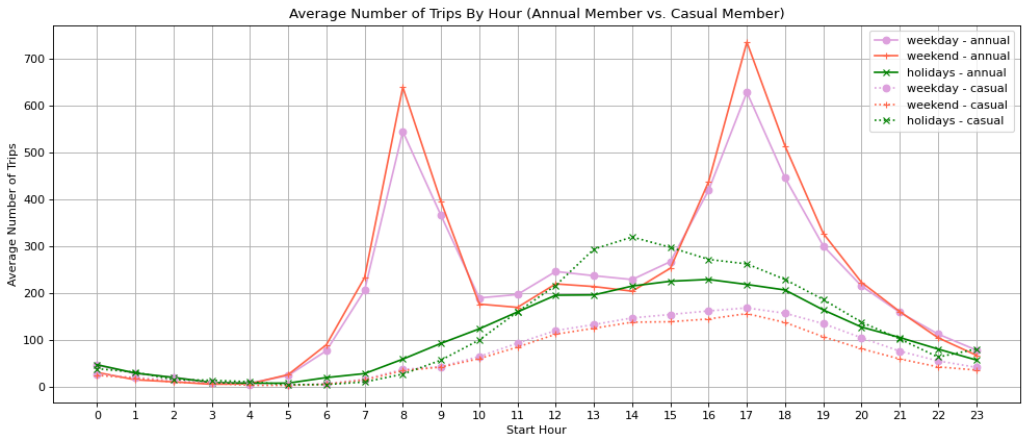


Figure 14: Average Number of Trips By Hour (Annual Members vs. Casual Members)

In addition, Bike Share Toronto promotes their business by having Free Ride Wednesdays in the month of July [Bike Share Toronto, 2022]. This has led to an interesting pattern in bike usages, suggesting that different operating policies and rules can influence bike demands.

Figure 15 shows the total number of trips taken on Wednesdays by month in 2019. It is observed that on Wednesdays, July has the highest total number of trips taken for casual member, leading to the highest total number of trips taken for all members as well. This provides evidence that having "Free Ride Wednesdays" may have led to an increase in bike demands, especially for casual members, on Wednesdays during the month of July.

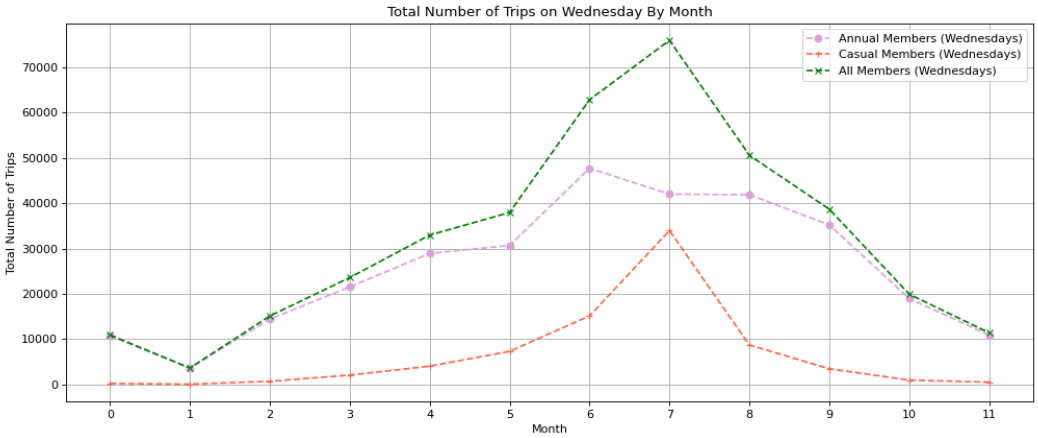


Figure 15: Total Number of Trips on Wednesdays by Month

B Code Description

The code for the thesis project can be found in this GitHub repository. It includes:

1. Code for bike usage analysis in the folder **1_Exploratory Analysis**:
    - Temporal bike usage analysis: 1 Exploratory Analysis (Bikeshare Ridership 2019).ipynb
    - Spatial bike usage analysis: 3 Spatial Visualizations (2019 Ridership) V2.ipynb
    - Spatio-temporal bike usage analysis: 3 Spatial Visualizations (2019 Ridership) V3 Heatmaps Only.ipynb
    - Ridership records and bike station locations matching: 2 Bike Station Locations Mapping (2019 Ridership).ipynb
2. Code for cluster-then-predict framework (including the Bipartite Clustering algorithm implementation) is in the folder **2_Computational Experiments**
    - Bipartite Clustering algorithm: generating_clusters folder
    - Cluster-then-predict framework using clusters grouped by FSA: predicting_demand/Checkout_Checkin_Predictions_FSA.ipynb (implementation referenced code from here)
    - Cluster-then-predict framework using clusters generated by Bipartite Clustering algorithm: predicting_demand/Checkout_Checkin_Predictions_V3.ipynb
3. Generated bike station clusters and check-out and check-in prediction errors are stored in the folder **3_Results**