Analysis of Montreal Bike-Share (BIXI) Open Data

Aiden Dever
School of Computing and Data Science,
Wentworth Institute of Technology

This paper presents the results of an investigation into Montreal Bike-Share (BIXI) Data. The BIXI system, one of the largest in North America, faces challenges in bike distribution across its network. This paper explores a method for predicting BIXI traffic using weather, population, and station location information as a step towards solving the bike allocation problem. The bike allocation problem arises when a disproportionate number of bikes leaving a station vs bikes returning results in no bikes being available at a given station.

I. Introduction (*Heading 1*)

Montreal is a large city in Quebec, Canada. While most cities have a public bike-share system akin to "Blue Bikes" in Boston, or "Citi Bikes" in New York City, Montreal's system (*BIXI*) is exceptionally large. *BIXI* is North America's first bike-share system, modeled after similar systems in Europe [4]. *BIXI* is also remarkably open about their data, making them an ideal candidate for data exploration and prediction projects. In recent years, *BIXI* has sought to expand its services, and started a pilot program during the winter of 2023/2024 to test winter bike-shares. These bike-share networks tend to have a few core problems, related to how bikes distribute over the course of the day, and whether bicycle stations are placed well enough to keep up with demand.

Many existing research projects have sought to analyze *BIXI* data, and a few have sought to predict bike distribution. One project from 2021 sought to apply the traveling salesman problem to solving unequal bike distribution among *BIXI* stations [5]. However, none have combined weather, population, and *BIXI* data to attempt to solve the problem.

The unequal bike distribution problem is a common problem in public bike-share systems where a disproportionate number of bikes leaving a station vs bikes returning results in no bikes being available at a given station. Typically, addressing unequal bike distribution requires expensive interventions that involve manually moving bikes from one station to another. Efficiently predicting *BIXI* traffic could significantly reduce these costs and improve the network.

II. Datasets

A. Source of dataset

The first dataset, referred to as "trip data" was obtained directly from *BIXI*'s open data program. While many years' worth of data was available, 2023 was chosen for several reasons. 2023 was the first year the winter pilot program was running, allowing for additional analysis. 2023 was chosen instead of 2024 so there would be a complete year, and a season-to-season analysis could be done. This dataset has been used by the Canadian government and can be counted on as being a reliable source.

The second dataset, containing population by arrondissement, was obtained from a map created by the Montreal city government, based on data from the 2021 Canadian census [1]. Some BIXI stations are outside of Montreal "proper" and are not covered in this dataset. To correct for this, additional data was pulled from the 2021 Canadian census and added to the population dataset obtained from the Montreal government [2].

Additionally, weather data was pulled in via a Python library, *Meteostat*. *Meteostat* has hourly weather data around the world for the past decade. *Meteostat* uses data pulled together from a variety of sources, including the NOAA in the U.S and the Meteorological Service of Canada [3].

B. Character of the datasets

The granular trip dataset from *BXI* was very large. The dataset contained start/stop information for every trip on the *BIXI* service in 2023 which amounted to around 11 million rows or 1.29 GB. No cleaning was done to this data for much of the analysis, except for removing a few outliers to make for easier graphing. These outliers had invalid location data, invalid weather data, or other irregular data.

The population datasets did not contain any irregularities. However, it is important to note differences in how the population data was defined. For stations within Montreal, the "Start Station Arrondissement" field from *BIXI* referred to the arrondissement. For stations outside Montreal, the field referred to the name of the city. This results in population having different definitions for different stations, which can lead to some irregularities.

The weather data from the *Meteostat* library was very well managed and did not require any cleaning. The only additional step needed was to match up the hourly timestamp from *Meteostat* to the more granular timestamp from the *BIXI* trip data. To solve this problem, the closest weather entry was assigned to each *BIXI* trip.

Ultimately, all three datasets were combined into one master dataset that contained all the features from each dataset above. This dataset was aggregated by several different columns to analyze different themes in the data.

III. Methodology

A. Method A

Through exploratory data analysis, it was revealed that stations and arrondissements farther away from "downtown" Montreal were correlated with longer trip times and fewer trips. A feature was created by using Euclidean distance to calculate the distance between a station and "downtown" Montreal, defined as the historical center, Notre-Dame de Montreal. This was done using the following equation:

$$\sqrt{(111 * (a - x1))^2 + (111 * cos(b - x2))^2} = y \quad (1)$$

Here, a is the latitude of the station, and b is the longitude of the station. *x1 and x2* are the latitude and longitude of Notre-Dame, and y is the estimated distance to downtown Montreal in kilometers. This follows the general assumption that at most places on the Earth, degrees latitude can be translated to kilometers by multiplying latitude by 111.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

Degrees longitude can also be translated to kilometers by multiplying 111 by the cosine of the longitude [6].

B. Method B

To solve the bike allocation problem described in the introduction, it is necessary to predict the number of trips leaving a station each day. An *SKLearn SGDRegressor* model was used to predict the number of trips. This model was chosen mostly because of the volume of data available. While many features were available in the dataset provided, ultimately three categorical features and four numerical features were found to be the most highly correlated and useful for this model. The dataset was cleaned by dropping rows with null or missing values and split on an 80/20 training/testing split. The categorical columns were one hot encoded to be usable for this model and preprocessed to fit the *SKLearn* requirements [7]. The features and response variable are outlined below:

- **Day/Month:** Day/Month the trip occurred.
- **Weather Category:** The weather category for the day (cloudy, rainy, snowy, etc.) defined by Meteostat.
- **Day of the Week:** 1-7 number showing day of the week.
- **Starting Arrondissement:** Arrondissement the trip started in.
- **Number of Trips:** The response variable. Defined as the number of trips, aggregated daily.

To compare the efficacy of different models, a gradient boosting regression model and a random forest model were both used as well.

C. Method C

General exploratory analysis was done to gain insight into various patterns in the BIXI trip data. Several bar charts were made to examine the relationship between categorical variables and the two response variables, trip length and number of trips. Additionally, a *Plotly* map was created to examine the distribution of BIXI stations throughout Greater Montreal and examine the relationship between distance to downtown and the number of trips.

D. Method D

Another common problem when managing bike-shares is trying to understand how long a customer will need a bike for, so the bike-share can manage the stock. To solve this problem, an *SGDRegressor* model is used to predict the trip length in minutes based on the available features in the combined dataset. An *SGDRegressor* model was chosen to handle the large amount of data. The features in the model and the response variable are outlined below:

- **Start Station Name/ End Station Name:** The beginning and end BIXI station for the trip.
- **Weather Category:** The weather category for the day (cloudy, rainy, snowy, etc.) defined by Meteostat.
- **Month:** Month the trip occurred.
- **Population:** The population of the arrondissement associated with the starting BIXI station.
- **Distance to Center:** The distance to the center of Montreal, method outlined in *Method A.*
- **Trip Length (minutes):** The response variable. Defined as the length of the trip, from checking out the bike to returning the bike.

IV. Results

A. Result A

Exploratory data analysis using the calculations outlined in *Method A* revealed that stations further away from the city center had less traffic, and the trips that left those stations had higher trip times.

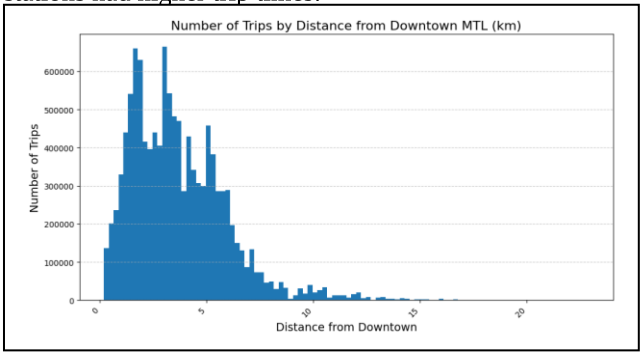

Fig. 1. Number of trips by distance from downtown Montreal (km).

As the distance from downtown increases, the number of trips decreases significantly, spiking around five kilometers from downtown. This can also be seen in *figure 2*:

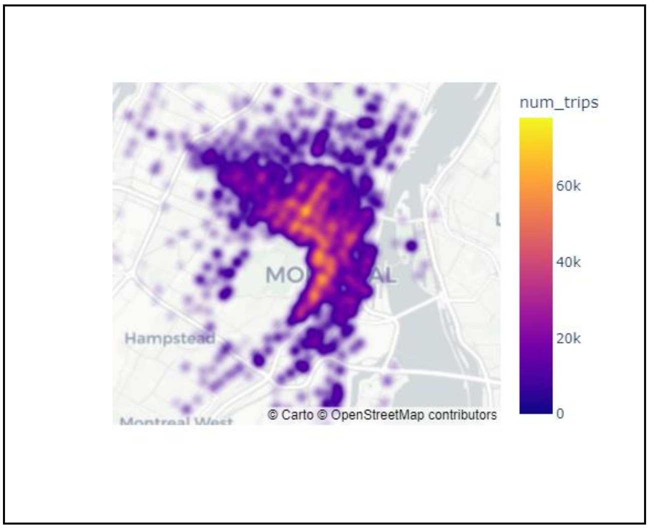

Fig. 2. Number of trips plotted on a *Plotly* map of Montreal.

Downtown is in the middle of the high density area outlined on the map, supporting the claim that stations farther out from the center have fewer trips.

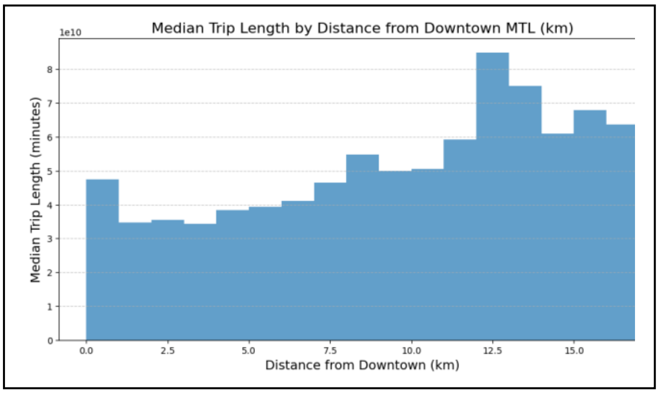

Fig. 3. Median trip length by distance from downtown Montreal (km).

As the distance from downtown increases, the trip length also increases, indicating that most trips are to the downtown area.

B. *Results B*

Machine learning models for predicting daily trip numbers achieved R^2 scores of 0.71 – 0.87. The random forest model performed best (R^2 = 0.86), highlighting the importance of feature interaction.

TABLE I. PERFORMANCE BY MODEL TYPE (NUMBER OF TRIPS)

Performance	**Model Type**		
	SGD Regressor	***Gradient Boosting Regressor***	***Random Forests***
R^2	.79	.72	.86
Mean Absolute Error	19	24	14.7

Fig. 4. Performance by Model Type (Predicting Number of Trips by Station)

C. *Results C*

The number of trips did not correlate directly with population, suggesting additional factors are more important in predicting trip numbers.

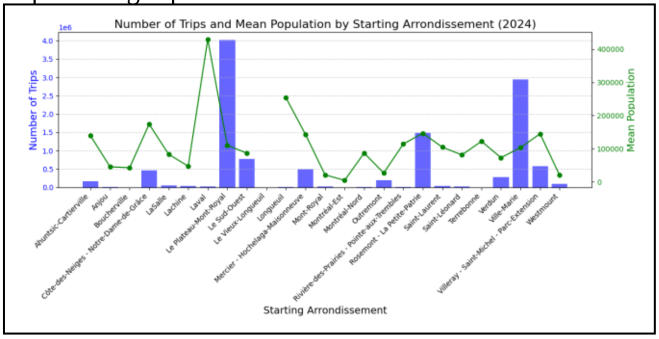

Fig. 5. Number of trips and mean population by starting arrondissement.

D. *Results D*

A machine learning model outlined in *Method D* was created to predict the length of a trip based on the features available in the trip dataset. A variety of different regression models were used, but performance was poor. The best iteration was an *SGDRegressor* that achieved an R^2 value of .0002, indicating limited utility. This suggests that the approach outlined in *Results B* may be more useful for solving the bike allocation problem.

V. DISCUSSION

While many of the features used in the prediction model exhibited strong correlations with the response variable, including an additional variable capturing the number of trips from the prior year would likely have enhanced the model significantly. However, the substantial size of the *BIXI* dataset presented significant computational challenges. For instance, the runtime for processing weather data alone required nearly eight hours. Extending this process to multiple years of data was deemed impractical without access to more resources.

Additional data sources mentioned in the project proposal, such as information on Montreal's bike infrastructure, could have provided valuable insights into the popularity of specific areas [8]. However, the format of this data proved too complex to effectively integrate into the existing workflow. Another unique challenge in Montreal is the seasonality of bike lanes. Incorporating these changes and aligning them with *BIXI* trip data would have been highly resource-intensive and difficult to execute.

A potentially more effective strategy for addressing bike-share imbalances could involve analyzing the number of bike trips paired with corresponding return trips. Return trips inherently mitigate imbalances without requiring external intervention. Unfortunately, the absence of customer identification numbers in the dataset made it impossible to differentiate between return and one-way trips, limiting the scope of this analysis.

Regarding trip length prediction, the methodology outlined in *Method D* proved suboptimal. Other models, such as random forests, would likely have achieved superior performance due to their ability to capture complex interactions between features, as highlighted in *Results B*. However, the computational demands of a random forest on a dataset of this scale rendered it impractical within the constraints of this project. In summary, while the project leveraged as many features as possible, additional resources could have improved the efficacy of this project.

VI. CONCLUSION

The investigation into Montreal's *BIXI* bike-share system provided valuable insight into the challenges of predicting bike traffic and solving the bike allocation problem. By integrating population statistics, weather data, and trip data, this revealed key factors influencing bike usage patterns. These patterns included proximity to downtown Montreal, weather patterns, and time of year. The machine learning models developed, particularly the random forest regressor showed much potential in predicting the number of trips at individual stations on a given day, offering a potential tool in the process to solve the bike allocation problem.

However, this investigation also highlighted the many limitations present in both data and methodology. The inability to incorporate large scale historical trip data from prior years, challenges with large datasets, and missing features like detailed bike lane infrastructure all limited the potential of these models. These constraints underscore the complexity of predicting bike-share usage in a dynamic large city like Montreal.

Future research could address these limitations by leveraging additional computational resources, integrating historical data, and exploring the seasonal nature of the bike lane infrastructure in Montreal.

Overall, while the models presented in this investigation offer a solid foundation for understanding and addressing the bike allocation problem, there is significant room for improvement. By building upon these findings, future efforts can contribute to a more sustainable and efficient bike-share system for Montreal and other similar cities. Solving problems like these and ensuring that large scale bike-share systems like *BIXI* are profitable and useful is essential to growing bicycling infrastructure and addressing the climate crisis.

References

[1] "Population Totale et Superficie des Arrondissements de Montreal (2022)," Ville de Montreal, Feb. 2022. Accessed: Dec. 08, 2024. [Online].Available:https://ville.montreal.qc.ca/pls/portal/docs/PAGE/MTL_STATS_FR/MEDIA/DOCUMENTS/CARTE_POPULATION%20ET%20SUPERFICIE%202021.PDF.

[2] Statistics Canada, "Census Program," *Statcan.gc.ca,* 2021. https://www12.statcan.gc.ca/census-recensement/index-eng.cfm.

[3] Meteostat, "Data Sources | Meteostat Developers," *dev.meteostat.net*. https://dev.meteostat.net/sources.html.

[4] Y. Freemark, "Bixi Close to Launching First Ambitious North American Bike-Share in Montréal," *The Transport Politic*, Apr. 23, 2009. https://www.thetransportpolitic.com/2009/04/23/bixi-close-to-launching-first-ambitious-north-american-bike-share-in-montreal/ (accessed Dec. 08, 2024).

[5] D. W, "A Mixed-Integer Optimization Approach to Rebalancing a Bike-Sharing System," *Medium*, Feb. 11, 2021. https://towardsdatascience.com/a-mixed-integer-optimization-approach-to-rebalancing-a-bike-sharing-system-48d5ad0898bd (accessed Dec. 08, 2024).

[6] H. Veregin, "How Big is a Degree?," *State Cartographer's Office*, Jan. 21, 2022. https://www.sco.wisc.edu/2022/01/21/how-big-is-a-degree/.

[7] "SGDRegressor," *scikit-learn*, 2024. https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.SGDRegressor.html (accessed Dec. 08, 2024).

[8] Ville de Montréal, "Montréal's bicycle network," *Montreal.ca*, 2024. https://services.montreal.ca/en/maps/bike-paths.