Statistical Analysis of BIXI Montréal Bike Rentals - Part II

Gilles Bou Dalha Ghoussoub *(11343945)*   Olivier Makuch *(11345395)*   Olivia Mirijello *(11345446)*   Pedram Peiroasfia *(11337867)*

Introduction

Bike-sharing systems like BIXI Montréal have become an essential part of urban transportation, providing commuters with a sustainable and flexible way to get around the city. This report delves into the factors that influence BIXI usage patterns, with a particular focus on morning departures and the total number of trips at each station. Using advanced statistical modeling, the analysis seeks to uncover how temporal, spatial, and environmental factors shape usage trends.

This report is organized to provide a clear and detailed analysis of BIXI Montréal's bike-sharing patterns. It starts with a **Data Preprocessing** section, where the focus is on cleaning and transforming the data to ensure it is accurate and ready for analysis. Next, the **Model Selection and Refinement** section explains the process of choosing the best statistical model. This section also highlights the importance of including key covariates and interaction terms to better capture the trends in the data. Finally, the **Analysis of Research Questions** uses the model to explore a variety of topics. This structure ensures a logical flow from data preparation to meaningful insights, offering a thorough and practical exploration of the factors that influence BIXI usage.

Exploratory Data Analysis

Before beginning the exploratory data analysis (EDA), we reviewed the summary statistics of the dataset to identify any potential issues. Figure 1 provides a data preview and a summary of its key features:

station	arrondissement	mm	dd
Bibliothèque du Vieux-St-Laurent (de l'Église / Filiatrault)	Saint-Laurent	October	15
Ropery / Augustin-Cantin	Le Sud-Ouest	June	7
Métro Langelier (Langelier / Sherbrooke)	Mercier - Hochelaga-Maisonneuve	July	2
5e avenue / Masson	Rosemont - La Petite-Patrie	May	18
Thimens / Alexis-Nihon	Saint-Laurent	September	20
de Bellechasse / de St-Vallier	Rosemont - La Petite-Patrie	September	17

station	arrondissement	mm	dd	wday
Length:1000	Length:1000	May :170	Min. : 1.00	Length:1000
Class :character	Class :character	June :154	1st Qu.: 9.00	Class :character
Mode :character	Mode :character	July :192	Median :16.00	Mode :character
NA	NA	August :162	Mean :16.04	NA
NA	NA	September:152	3rd Qu.:23.25	NA
NA	NA	October :170	Max. :31.00	NA

wday	AM	tot	temp	precip	AM_proportion
Sunday	0	7	9.5	0.0	0.0000000
Wednesday	16	49	13.7	3.7	0.3265306
Sunday	5	37	23.3	2.1	0.1351351
Thursday	15	100	7.6	0.0	0.1500000
Wednesday	2	10	15.7	0.0	0.2000000
Sunday	6	128	17.9	0.0	0.0468750

AM	tot	temp	precip	AM_proportion
Min. : 0.000	Min. : 1.00	Min. : 1.60	Min. : 0.000	Min. :0.00000
1st Qu.: 2.000	1st Qu.: 20.00	1st Qu.:15.20	1st Qu.: 0.000	1st Qu.:0.06593
Median : 6.000	Median : 45.00	Median :18.30	Median : 0.000	Median :0.13559
Mean : 7.986	Mean : 55.71	Mean :17.91	Mean : 2.769	Mean :0.14882
3rd Qu.:12.000	3rd Qu.: 73.25	3rd Qu.:21.90	3rd Qu.: 1.400	3rd Qu.:0.21832
Max. :55.000	Max. :335.00	Max. :27.70	Max. :63.000	Max. :0.66667

Figure 1: The first table on the left provides a preview of the first six rows of the dataset, showing variables such as the station, borough (`arrondissement`), month (`mm`), day (`dd`), weekday (`wday`), morning trips (`AM`), total trips (`tot`), temperature (temp), precipitation (`precip`), and the proportion of morning trips (`AM_proportion`). The table below displays additional sample data with numerical and categorical variables, highlighting their distributions. The table on the right summarizes the key statistics for each variable, including the minimum, maximum, mean, quartiles, and class type, giving an overall understanding of the dataset's characteristics.

We begin our analysis by exploring the response variable to gain a clearer understanding of how to approach model development. This involves examining its distribution, starting with an assessment of the quartiles, followed by a review of histograms for total duration.

The quartile analysis in Figure 2 left panel, reveals that most days have relatively low morning usage, with 95% of days showing a morning trip proportion below one-third. Days with a proportion exceeding 43.94% are particularly rare, making up only about 1% of the dataset. This suggests that high morning usage is unusual and could be driven by specific events or conditions. Figure 2 supports this observation, showing a right skew in the data. Notably, there are days with no morning trips, and in general, the proportion of trips taken in the morning tends to stay below 0.5. Additionally, there is a noticeable spike just below a proportion of 0.2, indicating that many days have a consistent pattern where around 20% of trips occur in the morning. This could represent a typical level of morning usage.

The right panel in Figure 2 shows the histogram of precipitation levels. Most of the precipitation data is concentrated at 0 mm, suggesting that many days experience no rain. Given this, it may be useful to summarize the data by creating a binary variable to indicate whether it is raining or not. As a cyclist, what likely matters most is whether it is raining at all, rather

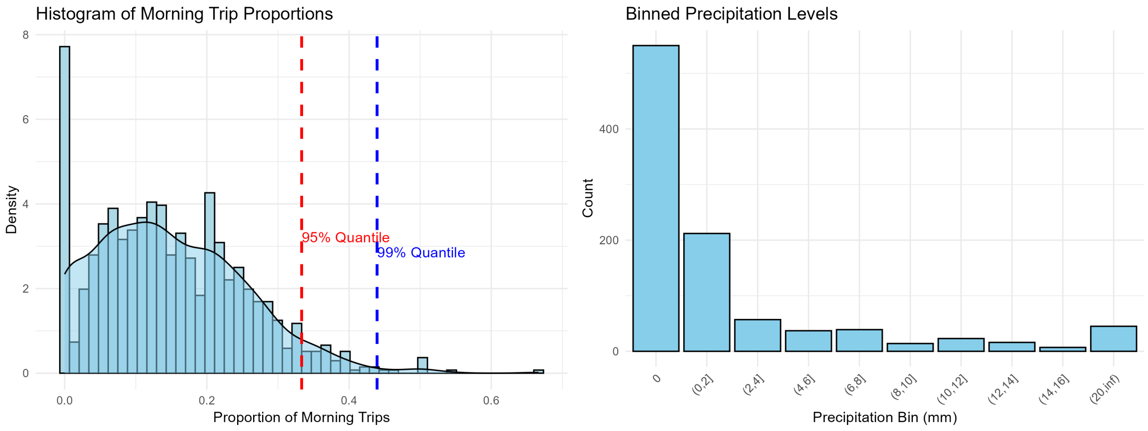

Figure 2: The left panel presents the histogram of the proportion of morning trips across all stations, overlaid with a density curve. The distribution is right-skewed with a prominent peak at zero, suggesting that many stations have no morning trips on certain days. Together, these panels provide insight into precipitation patterns and their potential relationship with morning trip proportions. The right panel shows the histogram of precipitation levels, highlighting that most of the data is concentrated around 0 mm, indicating minimal or no precipitation on the majority of days.

than the specific amount of precipitation. This perspective aligns with our proposed transformation, which simplifies the data while retaining its practical relevance.

In Figure 3, we present the distribution of morning trip proportions across both days of the week and months. The left panel highlights a noticeable reduction in morning trips on weekends, with smaller decreases observed on Mondays and Fridays. Midweek days show a higher proportion of morning trips, reflecting a stronger weekday commuting trend. The right panel illustrates monthly variations in morning trips from May to October, with a decline toward the end of the school year in early summer and a subsequent rise in the fall as schools reopen. Together, these patterns suggest temporal dynamics influenced by both weekly work routines and seasonal changes.

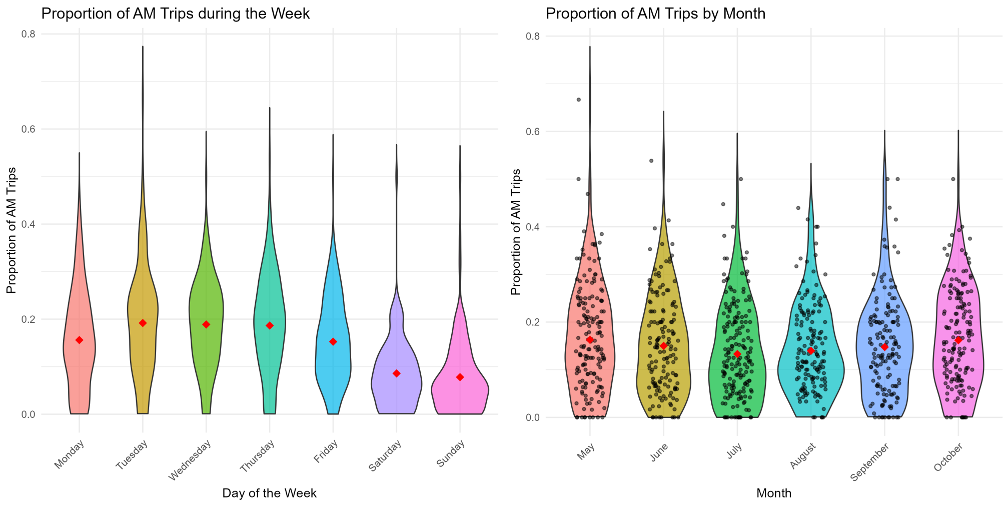

Figure 3: This combined violin plot illustrates the distribution of the proportion of morning trips (`AM_proportion`) across days of the week and months (May to October). The left panel shows weekday variations and the right panel highlights monthly trends, where individual data points are shown as black dots, and similar density patterns are observed. In both panels, red diamonds mark the mean proportions.

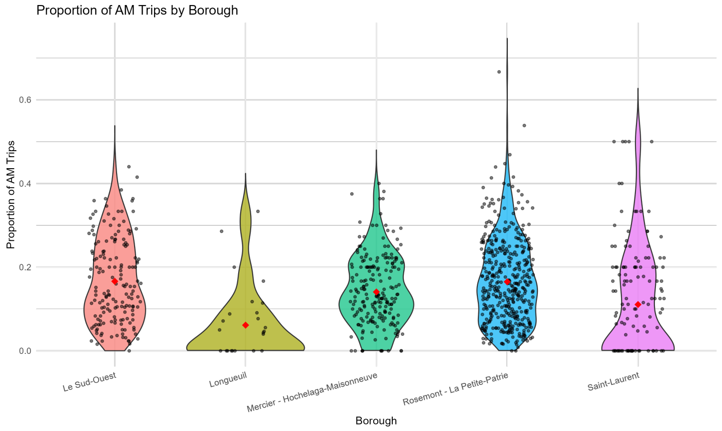

Figure 4: This violin plot shows the distribution of morning trip proportions (`AM_proportion`) for each borough. The red diamonds represent the mean proportions, while individual data points are displayed as black dots.

In Figure 4, we reveal noticeable differences in the proportion of morning trips across various boroughs. This variation may reflect underlying demographic factors that influence biking habits. For instance, certain boroughs might have a higher concentration of students, office workers, or other demographic groups more inclined to use bikes for transportation. Additionally, proximity to economic centers could play a role; boroughs located farther from major office areas might exhibit lower morning biking rates, as residents may choose other forms of transportation for longer commutes. It is also worth noting that the sample size for Longueuil is smaller compared to the other boroughs, which could introduce challenges in making robust inferences for this borough.

Model Selection Process

Data Preprocessing

In this section we will discuss what preprocessing steps were applied to our initial dataset to have a higher quality data:

1. *Factorization of Categorical Variables:* The following columns are converted to factors to treat as categorical variables: `Month, Day, arrondissement, wday`.
2. *Converting Precipitation to a Binary Variable:* `precip` (precipitation) is first converted to numeric format and then, it is transformed into a binary factor. The rationale behind this approach is that a large number of records show zero precipitation. As discussed in the first part of the project, cyclists are likely more concerned with whether it will rain at all, rather than the exact amount of rainfall, when deciding whether to take a bike. This binary perspective—rain or no rain—more accurately reflects how precipitation impacts a cyclist's decision-making process.
   - `0` when there's no precipitation (`precip` = 0).
   - `1` when there's precipitation (`precip` > 0).
3. *Creating weekend variable:* Another new variable, `weekend`, is created as an indicator variable (binary). This variable is `1` when it's a weekend (Saturday or Sunday) and `0` otherwise. The justification of this variable is explained later.
4. *Removing Unnecessary Columns:* The `Day` column is removed.

Choosing the Model

The choice of a binomial model with a logit link function (logistic regression) for analyzing the proportion of AM trips, `AM_proportion` = $\frac{\texttt{AM}}{\texttt{tot}}$, is driven by the nature of the data and the statistical properties required for valid inference.

1. *Nature of the Response Variable:* The response variable, `AM_proportion`, represents the proportion of trips that depart in the morning. The type of response is:
   - Bounded between 0 and 1: The data cannot exceed these bounds, making standard linear regression inappropriate as it can predict values outside this range.
   - Derived from Counts: The proportion is based on two discrete counts (`AM` and `tot`), which naturally leads to a binomial distribution framework.

2. *The Binomial Model for Proportion Data:* The binomial model is well-suited for scenarios where the response variable represents the number of ”successes” (e.g., morning trips) out of a fixed number of trials (e.g., total trips).
3. *Logit Link Function - Addressing the Bounded Nature of Proportions:* The logit link function is used in logistic regression to transform the proportion p from the bounded range [0,1] to the real line $(-\infty, \infty)$: $logit(p) = log(\frac{p}{1-p})$. This transformation ensures that the model's predictions for the proportion remain within the valid range (0 to 1). In addition, the logit transformation allows for a linear relationship between the transformed response (log-odds) and the predictors, making model estimation efficient.

Selection of Covariates

First, we discuss what covariates we should include in the model due to the questions being asked in question 1:

- We should include `arrondissement` as we are directly being asked about how it effects the mean odds ratio in (a).
- We should include `mm` (month). It might not seem like months play a significant role (right panel of Figure 3), but due to question (b), we will be including it.
- `wday` also plays a significant role, as shown in the first plot, and is specifically addressed in this part of the analysis. Therefore, it should certainly be included in the model. Additionally, the second part of question (c) explores whether rain has a compounding effect on the weekend versus weekday variable. In statistical terms, this translates to examining whether there is a significant interaction between these two covariates. So we ill also include `precip`.
- `temp` plays a crucial role in influencing the mean odds of morning departures. Additionally, question (d) prompts us to examine whether its effect differs between weekends and weekdays, indicating the need to include an interaction term between these two variables in the model.

An important consideration is whether to simplify `wday` to capture only the distinction between weekdays and weekends, or to retain its full detail across all seven days. As shown in left panel of Figure 3, there appear to be notable differences in means across specific weekdays, such as Monday versus Tuesday. This hypothesis could be further examined by comparing models that use `wday` versus `weekend` as covariates, using AIC or BIC to determine which approach provides a better fit (although they aren't official statistical tests but rather measures used for model selection). Note that we chose not to include both `wday` and `weekend` in the model due to perfect collinearity.

> H_0: The simpler model (using `weekend`) is preferred, indicating that using `wday` instead of `weekend` does not significantly improve the fit.
>
> H_1: The more complex model (using `wday`) is preferred, suggesting that including `wday` as a covariate (and excluding `weekend`) enhances the model fit.

The model with the lower BIC value is considered better, hence, if $\texttt{BIC}_{wday} < \texttt{BIC}_{weekend}$, then the model using `wday` is preferred. The result of this implementation is as follows:

- BIC model with `weekend`: 5932.643
- BIC model with `wday`: 5885.785

Therefore, the model with `wday` is preferred. However, we have opted to exclude `wday` (a factor with seven levels) from the model, and rather use `weekend` instead. Including `wday` would offer more detailed insights into weekday differences in AM trip proportions, which might align better with observed patterns, but given the specific focus of our analysis, particularly for questions 1c and 1d, our primary interest is in distinguishing between weekend and weekday trends. This choice represents a trade-off between model interpretability and complexity: using `wday` could enhance the model's reflection of real-world variations but might make it harder to interpret.

In addition, we also excluded the `station` variable from the model. With 148 levels, including this variable would overcomplicate the model and reduce interpretability. Also, the limited data for each station level would make it difficult to accurately estimate the coefficients, and it doesn't directly contribute to answering any of our research questions.

To summarize, until now we must include `arrondissement, Month, temp, precip, weekend, weekend:temp`, and `weekend:precip` in the model. We add another covariate, following our analysis of the first project, namely, `temp:precip`:

$$\texttt{logit}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\texttt{Month} + \beta_2\texttt{arrondissement} + \beta_3\texttt{temp} + \beta_4\texttt{precip} \\ + \beta_5\texttt{weekend} + \beta_6(\texttt{weekend}:\texttt{precip}) + \beta_7(\texttt{temp}:\texttt{weekend}) + \beta_8(\texttt{temp}:\texttt{precip}) \quad (1)$$

Model Refinements and Adjustments

When using binomial modeling, we typically assume that the dispersion parameter (ϕ) equals one. However, it is crucial to check this assumption, as a deviation from $\phi = 1$ would indicate overdispersion or underdispersion, which can undermine the model's validity. To check for dispersion (ϕ) in our model, we estimate the dispersion, which is found to be `2.705`. This statistic helps assess whether the dispersion parameter differs from one. ϕ is estimated using the scaled deviance:

$$\hat{\phi} = \frac{D(y; \hat{\mu})}{n - p - 1}, \tag{2}$$

where:

- $D(y; \hat{\mu})$ is the deviance,
- y_i represents the observed count,
- $\hat{\mu}_i$ is the predicted mean from the model,
- n is the total number of observations, and
- p is the number of model parameters.

Since $\phi \geq 1$ indicates signs of overdispersion, we proceed with necessary adjustments to address this issue in the subsequent steps. The diagnostic plots for the initial model are shown in Figure 5.

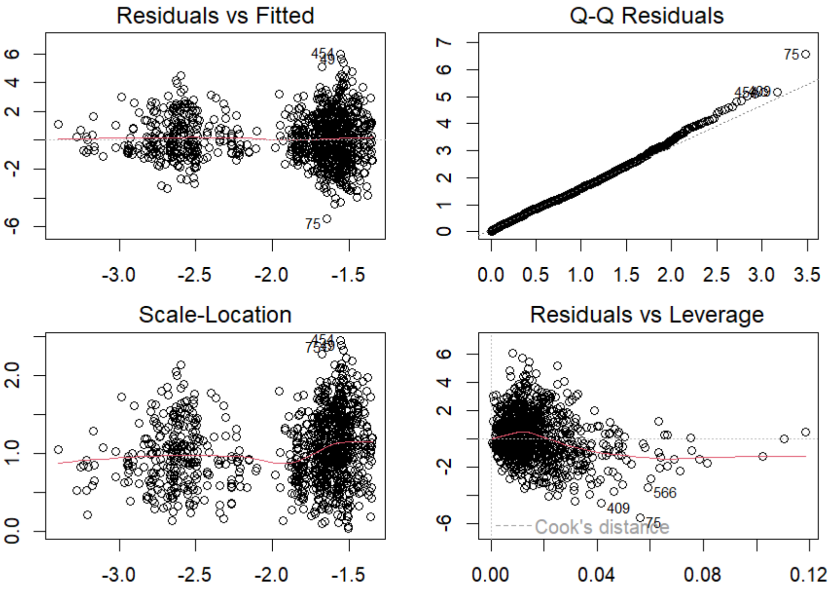

Figure 5: Diagnostic plots for the binomial model before addressing overdispersion. The plots show residuals vs fitted values, Q-Q residuals, scale-location, and residuals vs leverage, providing insights into the model's performance.

Given the scope of the course, we won't be conducting an in-depth analysis of the diagnostic plots. However, the Q-Q plot and evidence of overdispersion suggest that identifying outliers could be beneficial.

Using this approach, we identified 67 outliers (6.7% of our data) based on the standardized Pearson residuals, which is a common method for identifying outliers in GLMs [1]. The key question is whether removing these outliers affects the model coefficients and our interpretation of the model. Notably, the dispersion changes significantly to `1.862`, a reduction of 31.16%. Despite this reduction, the Q-Q plot still suggests some issues, though we will not delve further into them as it is beyond the scope of this course. The updated diagnostic plot after removing outliers is shown in Figure 6.

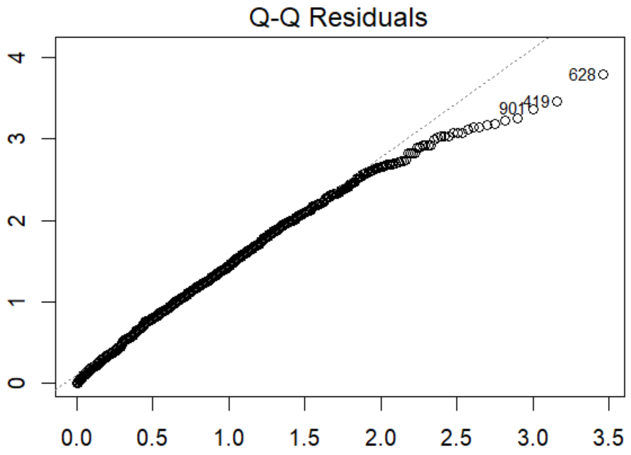

Figure 6: Diagnostic plot (Q-Q plot) for the binomial model after removing outliers. Although the dispersion value reduced significantly, the Q-Q plot still suggests deviations from normality.

The primary objective of removing outliers can be summarized with the following question: Does the interpretation of our coefficients change? Specifically, do the coefficients that were significant remain significant, and do those that were not significant stay unchanged?

Table 1: Comparison of Coefficients Before and After Outlier Removal (Significance level of $\alpha = 5\%$)

Term	**Estimate Before**	**p-value**	**Estimate After**	**p-value**	**Significance Status**
(Intercept)	-1.3496401	0.0000000	-1.2840867	0.0000000	Both significant
Month6	-0.0020949	0.9641895	0.0064251	0.8994807	Both non-significant
Month7	0.0098580	0.8532493	-0.0160359	0.7767274	Both non-significant
Month8	0.0067644	0.8863580	-0.0203086	0.6869174	Both non-significant
Month9	0.0174264	0.7003272	-0.0142314	0.7690958	Both non-significant
Month10	0.0802753	0.0863671	0.0468566	0.3505707	Both non-significant
arrondissementLongueuil	-0.5922612	0.0028905	-0.5757848	0.0038350	Both significant
arrondissementMercier - Hochelaga-Maisonneuve	-0.0646882	0.1360649	-0.0805641	0.0772132	Both non-significant
arrondissementRosemont - La Petite-Patrie	0.0688745	0.0375435	0.0838066	0.0183941	Both significant
arrondissementSaint-Laurent	-0.2164590	0.0094820	-0.2176011	0.0104181	Both significant
temp	-0.0153876	0.0001087	-0.0177618	0.0000434	Both significant
precip1	-0.1324061	0.2027315	-0.2068703	0.0620049	Both non-significant
weekend1	-0.7728499	0.0000023	-0.8797220	0.0000002	Both significant
precip1:weekend1	-0.0179471	0.8002175	-0.0087653	0.9057505	Both non-significant
temp:weekend1	-0.0123033	0.1482544	-0.0067507	0.4468773	Both non-significant
temp:precip1	0.0085750	0.1135791	0.0099913	0.0825929	Both non-significant

As observed in Table 1, the coefficients that were initially significant remain significant even after accounting for the outliers, while those that were not significant remain unchanged and stay insignificant. This suggests that removing the outliers does not alter the overall significance of the model coefficients and, therefore, we can proceed with our initial model (without outliers). It is also worth noting that the data we are dealing with comprises valid records of people using BIXI. From this perspective, there may not be any actual outliers in the dataset; instead, we might be observing some extreme (but valid) data points. This reinforces the decision to use the model without outlier detection.

Analysis of Research Question 1

Analysis of 1a)

The objective of this analysis is to determine whether the odds of a BIXI trip departing in the morning vary significantly across different boroughs in Montréal. In logistic regression, each coefficient estimate is tested against the null hypothesis that the coefficient is zero ($H_0 : \beta_j = 0$) using a Wald test. The results are as follows (note that the coefficients are reported in Table 1 and hence not repeated here):

- *Interpretation for* `arrondissement`: Longueuil ($p = 0.002891$), Rosemont - La Petite-Patrie ($p = 0.037543$), and Saint-Laurent ($p = 0.009482$) have statistically significant coefficients. This implies that the odds of an AM departure are significantly lower in these boroughs compared to the reference borough (Le Sud-Ouest), with the exception of Rosemont - La Petite-Patrie, where the odds of an AM departure are higher (but not significant in significance level of $\alpha = 1\%$) than the reference level due to the positive coefficient. Mercier - Hochelaga-Maisonneuve ($p = 0.136065$) is not statistically significant ($p < 0.01$), indicating no strong evidence that it differs from the reference borough.

- *Exponentiated Coefficients and Interpretation*: The exponentiated coefficients (odds ratios) for the significant boroughs show the multiplicative effect on the odds of an AM departure:
  - Longueuil $\approx e^{\beta_{\text{Longueuil}}} = 0.55307$
  - Saint-Laurent $\approx e^{\beta_{\text{Saint-Laurent}}} = 0.8053$

  These odds ratios indicate how the proportion of AM departures decreases in these boroughs compared to the reference level holding everything else constant. To summarize, there is significant variation in the odds of an AM departure across different boroughs, particularly for Longueuil and Saint-Laurent, where the odds are lower.

Next, we perform the analysis of deviance using LRT. The Likelihood Ratio Test (LRT) compares nested models to determine if adding a specific predictor significantly improves the model fit for `arrondissement`:

H_0: `arrondissement` does not improve the model (no significant effect).

H_1: `arrondissement` improves the model (significant effect).

Or similarly:

H_0: $\beta_{Longueuil} = \beta_{Mercier-Hochelaga-Maisonneuve} = \beta_{Saint-Laurent} = \beta_{Rosemont-LaPetite-Patrie} = 0$

H_1: At least one of the coefficients is not zero.

The results of this test concluded that the deviance change when including `arrondissement` is 55.73 with a p-value of $2.287e - 11$, which is highly significant ($p < 0.001$). This implies that adding `arrondissement` significantly improves the model, confirming that there is substantial variation in AM departures across boroughs. The LRT result aligns with the Wald test results, indicating that `arrondissement` is a significant predictor of AM departure proportions, and its inclusion in the model is justified.

To summarize, the tests show strong evidence that the odds of an AM departure vary across different boroughs. Both the Wald tests for individual coefficients and the Likelihood Ratio Test indicate significant variation, particularly in Longueuil and Saint-Laurent (under the significance level of 1%). This confirms the substantial effect of `arrondissement` on morning departure patterns, aligning with the expected interpretation framework where significant p-values indicate a meaningful effect of the predictor.

For the business insights, the analysis reveals significant variation in the odds of morning departures across boroughs, highlighting important implications for BIXI's operations. Boroughs like Longueuil and Saint-Laurent, which show lower odds of morning trips, may require targeted interventions, such as improved station placement or enhanced bike availability during peak hours, to encourage usage. Conversely, boroughs like Rosemont - La Petite-Patrie, with higher odds of morning departures, suggest successful biking patterns that could serve as models for other areas. These findings emphasize the need for geographically tailored strategies to optimize bike-sharing usage and better address the unique needs of each borough.

Analysis of 1b)

Now, we examine which month has the highest and lowest odds of a morning BIXI departure, using the monthly coefficients in the logistic regression model to compare each month to a baseline reference month. In logistic regression, coefficients for categorical variables, like `Month` in our model, show how the log-odds of the outcome differ from a baseline or reference category. Here, May is the reference month, so the coefficients for other months indicate how their log-odds of an AM departure compare to May's. A positive coefficient suggests higher odds than May, while a negative one suggests lower odds.

The model's coefficients for each month are:

- June: $-0.0021, p = 0.9642$
- July: $0.0099, p = 0.8532$
- August: $0.0068, p = 0.8864$
- September: $0.0174, p = 0.7003$
- October: $0.0803, p = 0.0864$ (significant at $p < 0.1$)

October has the highest coefficient (0.0803), indicating a marginally significant increase in the odds of an AM departure compared to May. Exponentiating this coefficient gives an odds ratio of $e^{0.0803} \approx 1.0836$, suggesting that the odds of an AM departure in October are approximately 8.36% higher than in May. For the other months (June through September), the coefficients and corresponding odds ratios are near zero or close to 1, showing no strong evidence of differences in AM departure odds relative to May. This reinforces the finding that October exhibits the highest odds of an AM departure, albeit with marginal significance.

As for business insights, the lack of significant differences for other months (other than October) suggests that seasonal patterns do not strongly influence morning departures, which might be surprising given the expectation of increased usage in summer months like June or July. This counterintuitive result could reflect unique factors influencing biking behavior in October, such as cooler but still favorable weather, reduced summer travel, or increased commuting due to the resumption of regular work and school schedules. These findings highlight the need to further investigate temporal factors affecting usage to optimize seasonal planning and promotion strategies.

Analysis of 1c)

Here, our goal is to determine whether the odds of a morning BIXI departure decrease on weekends compared to weekdays and whether this decrease becomes even more pronounced in rainy conditions. The model that we have chosen (Equation (1)), lets us examine not only the effect of weekends on morning departures but also whether this effect is intensified by rain through the interaction term β_6.

BIXI's hypothesis is that odds of a morning departure are lower on weekends ($\beta_5 < 0$) and this decrease in odds is even greater on rainy weekends ($\beta_6 < 0$). Below, we dive into key findings from the model results:

- *Weekend Effect (β_5)*: The coefficient for `weekend` is negative and highly significant ($\beta_5 = -0.7729, p = 2.3 \times 10^{-6} < 0.001$). This indicates that, in the absence of other modifying factors (such as rain or temperature), the odds of a morning departure are lower on weekends compared to weekdays. When exponentiated, this coefficient gives an odds ratio of $e^{-0.7729} \approx 0.462$, meaning that the baseline odds of a morning departure on weekends are approximately 46.2% of the odds on weekdays when `temp = 0` and its weekdays (all else held constant). However, given the presence of interaction terms in the model (e.g., `weekend:precip`), the true effect of `weekend` on morning departure odds depends on the values of these interacting variables. Therefore, this result should be interpreted with caution, as the interaction terms may modify the weekend effect in specific contexts.
- *Interaction Effect - Weekend and Rain (β_6)*: The coefficient of the interaction term for `weekend:precip` is negative but not statistically significant ($\beta_6 = -0.017947, p = 0.8002 > 0.001$). Although the negative coefficient suggests that rain might slightly lower the odds of a morning departure on weekends, the lack of significance means there's no strong evidence to confirm this effect. The data does not provide enough support to conclude that rain further accentuates the weekend decrease in morning departures.

Hence, the model provides clear evidence that BIXI's hypothesis about weekends holds: the odds of a morning departure are significantly lower on weekends. However, it does not support the second part of the hypothesis, as there's no statistically significant evidence that rain further decreases the odds of a morning departure on weekends.

As for business insights, the analysis confirms that the odds of morning departures are significantly lower on weekends, supporting BIXI's hypothesis about weekend behavior. This insight is critical for operational planning, as it suggests reduced demand on weekends, allowing BIXI to optimize bike redistribution and potentially reduce staffing or operational costs during these periods. However, the model does not support the hypothesis that rain further exacerbates this weekend decrease, indicating that weather might not have a compounding effect on reduced weekend usage. This finding suggests that weekend users may already account for the possibility of rain in their decisions, and therefore, their decision is not really dependent on the day of the week, or more specifically, whether it's a weekend or not.

Analysis of 1d)

For this question, we aim to determine if the impact of temperature on the odds of a morning departure differs significantly between weekends and weekdays. This is done by testing the significance of the interaction term between `weekend` and `temperature` in our logistic regression model.

We have the following hypotheses:

H_0: The effect of temperature on the odds of a morning departure is the same on weekends and weekdays ($\beta_7 = 0$).

H_1: The effect of temperature on the odds of a morning departure differs between weekends and weekdays ($\beta_7 \neq 0$).

We will conduct the test at a significance level of $\alpha = 0.01$. To assess the interaction, we can use two approaches:

1. Wald Test: Evaluates the significance of the interaction term's coefficient in the model.
2. Likelihood Ratio Test (LRT): Compares a full model (with the interaction term) against a reduced model (without the interaction term) to see if the inclusion of the term significantly improves model fit.

With regards to the Wald Test, we obtained $\beta_7 = -0.0123$ with a standard error of 0.00851, z-value of -1.446, and p-value of 0.1483. The p-value is greater than our significance level, , so we fail to reject the null hypothesis. This suggests that there is no significant difference in the impact of temperature on the odds of a morning departure between weekends and weekdays.

Switching to the Likelihood Ratio Test, the reduced model (without the interaction term of `temp` and `weekend`) obtained a residual deviance of 2698.2 with 985 degrees of freedom. The full model (with the interaction term) obtained a residual deviance of 2696.1 with 984 degrees of freedom. For the test statistic, the difference in deviance is 2.0774 with a p-value of 0.1495. The p-value of 0.1495 is again greater than $\alpha = 0.01$, indicating that the inclusion of the interaction term does not significantly improve the model fit. This further supports the conclusion that the effect of temperature on morning departure odds does not vary significantly between weekends and weekdays.

Therefore, both the Wald test and the Likelihood Ratio Test (LRT) provide consistent evidence that the impact of temperature on the odds of a morning departure does not differ significantly between weekends and weekdays. At the 1% significance level, we do not have sufficient evidence to conclude a differential effect of temperature based on day type (weekend vs. weekday).

The analysis indicates no significant difference in the impact of temperature on morning departure odds between weekends and weekdays, suggesting that temperature consistently influences user behavior regardless of the day type. This complements the finding that morning departures are significantly lower on weekends, highlighting a consistent reduction in demand across varying weather conditions. From an operational perspective, this suggests that weekend users may already account for temperature effects in their planning, much like how rain does not further reduce weekend demand.

Analysis of Research Question 2

The second question focuses on modeling daily BIXI usage by examining the total number of trips departing from each station. The objective is to explore these difficulties and their implications for building an accurate and reliable model.

One of the primary challenges in modeling BIXI usage is the implicit **missing data problem**. The dataset only includes records for stations with at least one trip on a given day, meaning there is no information for stations that were not used on certain days. This absence of data is not equivalent to recording "zero trips" because it fails to distinguish between stations that had no demand and those that were unavailable due to maintenance, redistribution, or being at full capacity. This ambiguity makes it difficult to draw accurate conclusions about station-level usage patterns, as the underlying reason for the lack of trips is unknown. Implicit data gaps can introduce systematic bias by misrepresenting high-demand stations as underutilized in models, leading to incorrect conclusions about demand and redistribution. This inability to distinguish operational constraints from actual demand undermines the reliability of predictive models, particularly for strategic planning and estimating latent demand.

BIXI stations have **varying capacities** in terms of the number of bikes they can hold, and this capacity fluctuates dynamically over time due to redistribution operations, bike maintenance, and operational issues. These constraints directly affect the total number of trips a station can support on any given day, particularly in high-demand periods. This creates a fundamental challenge in interpreting trip counts. Without accounting for capacity, a station with low daily trips might be misclassified as underutilized when, in reality, its usage is capped by the limited number of bikes available. Conversely, stations with higher capacities might appear to have higher demand simply because they can accommodate more trips, even if the per-bike usage rate is comparable. Ignoring capacity constraints not only skews descriptive statistics but also leads to biased regression estimates, undermining the model's ability to explain true usage patterns or forecast future demand under different operational scenarios.

Daily trip counts at BIXI stations are interconnected, **influenced by both time-based and location-based patterns**. Temporal correlations occur because a station's usage on one day often depends on its activity on previous or following days, reflecting consistent commuting habits or neighborhood trends. Spatial correlations arise when stations close to each other affect each other's usage—if one station runs out of bikes, for instance, users are likely to turn to a nearby station. Redistribution operations, where bikes are moved between stations to balance supply, further add to these dependencies.

Ignoring these relationships can have significant consequences. Models that assume trips are independent can underestimate variability, leading to overly confident predictions about which factors are important. This can result in poor resource allocation or flawed insights into usage patterns. Additionally, overlooking shared influences can hide important trends in the data. Recognizing and accounting for these correlations is essential for accurate forecasting and effective policy-making that considers both individual stations and the system as a whole.

Daily trip counts at BIXI stations often exhibit **overdispersion**, where the variance exceeds what a Poisson model assumes. This arises from factors like station characteristics, weather, events, and unobserved heterogeneity, causing variability in trip counts. Overdispersion can undermine model reliability, leading to underestimated standard errors and overly confident p-values, which may falsely highlight certain predictors as significant. Addressing this issue is critical to improving model accuracy, especially for operational planning and resource allocation, where variability can significantly impact decisions.

Conclusion

In conclusion, this analysis highlights the challenges and insights involved in understanding BIXI trip patterns, focusing on both the proportion of morning departures and total station-level usage. The results show clear variations in morning departure odds across boroughs and reveal the impact of seasonal and temporal factors on usage. However, issues such as missing data, capacity limits, dependency of the stations, and overdispersion in trip counts complicate the modeling process and call for careful consideration. Despite these challenges, the models provide useful insights to support BIXI's operational planning and strategic decisions, while also pointing to areas where further refinements could improve predictions.

Limitations

- **Overdispersion:** The binomial model showed signs of overdispersion, with a dispersion parameter $\phi = 2.705$. A possible solution is to adopt a quasi-binomial or beta-binomial model to better account for overdispersion.
- **Lack of Station-Level Analysis:** The model excludes station-level variations, which could provide important insights into local factors influencing BIXI usage. Incorporating random effects through a mixed-effects model could capture these unobserved heterogeneities.

Contributions

The team collaborated effectively to complete the project, starting with a group meeting to discuss and determine the base model, during which everyone shared their thoughts and ideas. Olivier took the lead on the exploratory data analysis (EDA) and interpreting the insights, while also contributing to drafting and reviewing the final report. Olivia focused on addressing questions 1a) and 1b), compiling the report, and conducting a thorough review before submission. Pedram worked on questions 1c) and 1d), played a key role in defining the base model, supported Olivia and Olivier with assembling the report, and reviewed all code to ensure accuracy and consistency. Gilles focused on question 2 and provided valuable feedback by reviewing the entire report. The team's collaborative effort ensured a well-rounded and comprehensive final product.

References

[1] Midi, H., & Ariffin, S. B. (2013). Modified standardized Pearson residual for the identification of outliers in logistic regression model. *Journal of Applied Sciences, 13*(5), 828–836. `https://doi.org/10.3923/jas.2013.828.836`