Statistical Analysis of BIXI Montréal Bike Rentals - Part II

Introduction

In this report, we will present the selected model and explain the rationale behind its choice. The report is organized into three parts. First, we outline the preprocessing steps required to enhance the data quality, ensuring that our data is clean, consistent, and suitable for modeling. Next, we assess which model family is best suited for modeling the proportion. Finally, we determine which covariates should be included to model the proportion effectively, examining if these chosen covariates successfully capture the underlying distribution. We also evaluate whether any adjustments are necessary to better align the model with the true characteristics of the proportion, ensuring robustness and accuracy in our final model.

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.

Response Variable

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 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, possibly due to regular commuter behavior or other predictable factors.

Response vs. Predictor Variables

We can also assess the response variable in relationship to its predictors. This can allow us to make certain assumptions about the data, such as identifying potential linear or non-linear relationships, detecting multicollinearity among predictors, and evaluating the strength and direction of associations. By examining these relationships, we can gain insights into which predictors are likely to have the most significant impact on the response variable, guide our model selection process, and improve our understanding of any underlying patterns or interactions within the data.

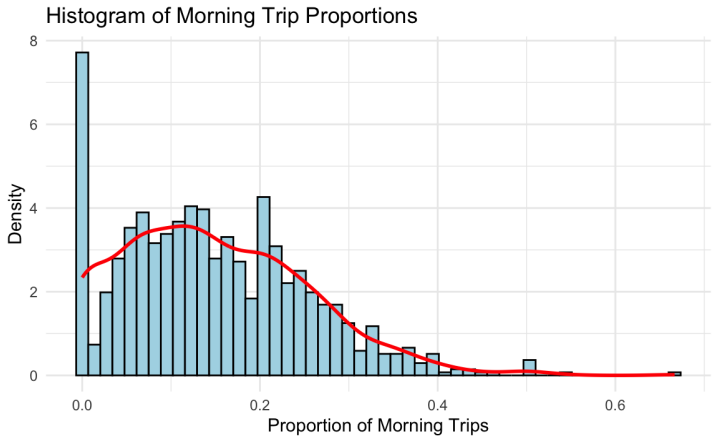

Figure 2: This histogram illustrates the distribution of the proportion of morning trips across all stations. The density curve overlays the histogram, highlighting a right-skewed pattern with a notable peak at zero, indicating many stations have no morning trips on certain days.

In Figure 3, we reveal an interesting pattern in the proportion of morning trips. There is a noticeable reduction in morning trips on weekends, with a slightly smaller drop at the start and end of the workweek, specifically Mondays and Fridays. In contrast, the middle days of the week tend to show a higher proportion of morning trips, suggesting a stronger weekday commuting trend.

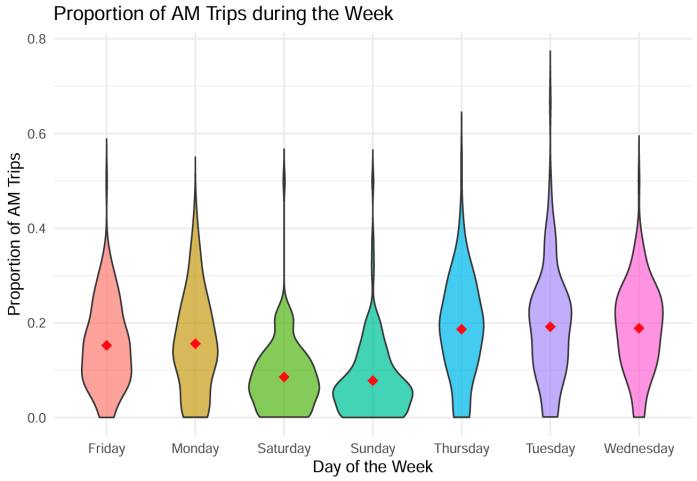

Figure 3: This violin plot displays the distribution of the proportion of morning trips (`AM_proportion`) for each day of the week. The red diamonds represent the mean proportions for each day, highlighting variations in morning trip patterns across weekdays and weekends. The wider sections indicate higher density of proportions, while narrower sections show less common values.

Our monthly analysis shown in Figure 4, shows variation in the proportion of morning trips throughout the summer. This pattern appears to align with the school calendar, with a decrease in morning trips toward the end of the school year, followed by an increase when school resumes in the fall.

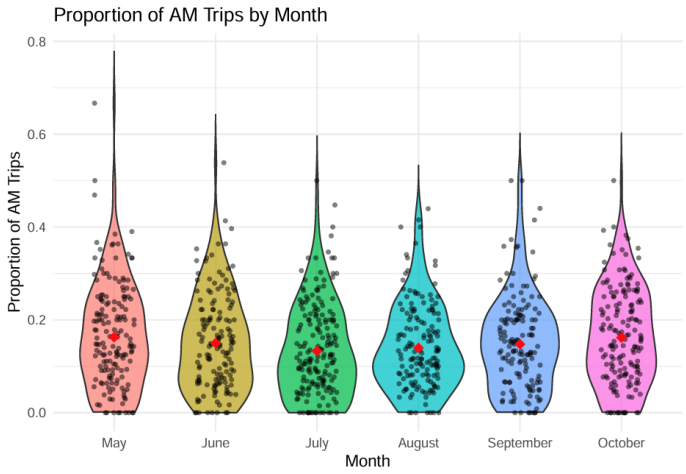

Figure 4: This violin plot illustrates the distribution of morning trip proportions (`AM_proportion`) for each month from May to October. The red diamonds indicate the mean proportion for each month, while individual data points are shown as black dots. The plot highlights monthly variations in morning trip proportions, with wider sections indicating higher density and narrower sections showing less frequent values.

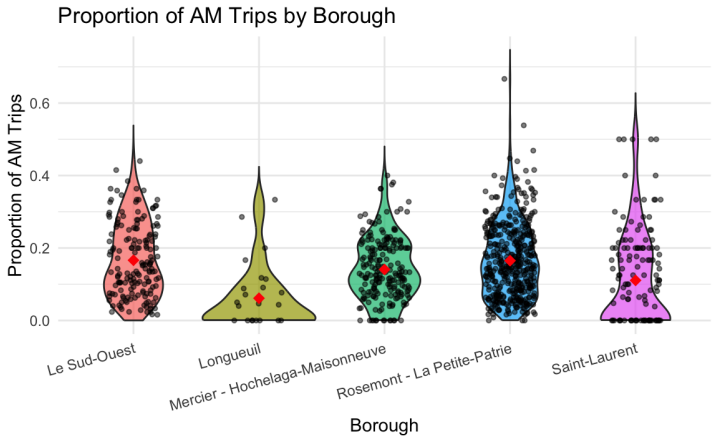

Figure 5: This violin plot shows the distribution of morning trip proportions (`AM_proportion`) for each borough. The red diamonds represent the mean proportions, while individual data points are displayed as black dots. The plot highlights differences in morning trip patterns across boroughs, with variations in density reflecting differing levels of morning trip activity.

In Figure 5, 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.

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. *Calculating the Proportion of AM Trips:* A new variable `AM_proportion` is calculated as the ratio of morning trips (`AM`) to the total number of trips (`tot`): $\texttt{AM\_proportion} = \frac{\texttt{AM}}{\texttt{tot}}$. Another new variable, `weekend`, is created and is 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 Family

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

1. *Nature of the Response Variable: Proportions Derived from Counts:* 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.
- We should include `mm` (month). It might not seem like months play a significant role (as shown in the plot below), but due to the question, 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 the question 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.
- `temp` plays a crucial role in influencing the mean odds of morning departures. Additionally, this question 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 the plot below, 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.

The BIC results for both models are presented below. We chose to use BIC because it accounts for the number of parameters in each model, helping to balance model complexity with fit.

H_0: The simpler model (using `weekend`) is preferred, indicating that `wday` 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.

We chose not to include both `wday` and `weekend` in the model due to perfect collinearity. 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.

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

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

In addition, we also excluded the `station` variable from the model. With 148 levels, including this variable would over-complicate 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.

Model Refinements and Adjustments

To check for dispersion (ϕ) in our model, we calculate the overdispersion test statistic, which is found to be `2.705`. This statistic indicates the degree of dispersion within the model, with ϕ estimated using the Pearson chi-squared statistic:

$$\hat{\phi} = \frac{\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{\sqrt{\hat{V}_i}}\right)^2}{n-p}, \tag{2}$$

where:

- y_i represents the observed count,
- $\hat{y}_i$ is the predicted count from the model,
- $\hat{V}_i$ denotes the model-based variance (calculated as $\hat{V}_i = n_i \cdot \hat{p}_i \cdot (1 - \hat{p}_i)$),
- 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. We begin with the diagnostic plots, as shown below:

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

Using this approach, we identified 67 outliers based on the standardized Pearson residuals. The key question is whether removing these outliers affects the model coefficients and our interpretation of the model. Notably, the disperson changes significantly. The initial dispersion value is `1.862`, which shows overdispersion. After removing outliers, the dispersion reduced by approximately 31% (dispersion value is `0.3117`). Despite this reduction, the QQ plot still suggests some issues, though we will not delve further into them as it is beyond the scope of this course.

As observed, 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.

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:

- *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, with the exception of Rosemont - La Petite-Patrie, where the odds of an AM departure are significantly higher due to the positive coefficient. Mercier - Hochelaga-Maisonneuve ($p = 0.136065$) is not statistically significant ($p < 0.05$), 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 $\approx e^{\beta_{\text{Longueuil}}}$
  - Rosemont - La Petite-Patrie $\approx e^{\beta_{\text{Mercier}}}$
  - Saint-Laurent $\approx e^{\beta_{\text{Saint-Laurent}}}$

  These odds ratios indicate how the proportion of AM departures decreases in these boroughs compared to the reference level. To summarize, there is significant variation in the odds of an AM departure across different boroughs, particularly for Longueuil, Rosemont - La Petite-Patrie, 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).

The results of this test concluded that the deviance change when including `arrondissement` is 52.39 with a p-value of $1.144e-10$, 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, Rosemont - La Petite-Patrie, and Saint-Laurent. 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.

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: $0.0803, p = 0.0864$ (marginally significant at $p < 0.1$)

October has the highest coefficient (0.0803), indicating a potential increase in the odds of an AM departure compared to May, though this result is only marginally significant. This suggests that October may have slightly higher odds of morning departures. For the other months (June through September), coefficients are near zero and not statistically significant ($p > 0.1$), indicating no strong evidence of differences in AM departure odds compared to May.

We can interpret the results more clearly by exponentiating the coefficients to obtain the odds ratios. The odds ratio for October is $e^{0.0803} \approx 1.0836$. This implies that the odds of an AM departure in October are approximately 8.36% higher than in May. Odds ratios for other months are close to 1, showing no notable change in odds from May. With that, the model suggests that October has the highest odds of an AM departure, about 8% higher than May, though the effect is only marginally significant. The other months do not exhibit significant differences in AM departure odds relative to May.

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 ($\beta_5 = -0.7729, p = 2.3e-06 < 0.001$). This result strongly supports the idea that 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 odds of a morning departure on weekends are about 54% lower than on weekdays, all else held constant.
- *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.

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 according to the Wald test.

Switching to the likelihood ratio test, the reduced model (without the interaction term) 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).

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.

BIXI usage is inherently tied to **temporal and seasonal patterns**, which introduces a significant challenge when attempting to model total daily trips. Factors such as the day of the week, whether it's a weekend or holiday, and seasonal changes in weather can dramatically affect bike-sharing activity. For example, weekdays might see higher usage during commuting hours, while weekends might attract more recreational users. Seasonal weather variations, such as warmer summers or harsher winters, also play a crucial role, as adverse weather conditions typically lead to lower usage. The impact of ignoring these effects is substantial. Failure to account for temporal and seasonal variability can result in a model that oversimplifies or misrepresents the factors driving BIXI usage. For instance, a station might appear underutilized during winter months without recognizing that this is part of a broader seasonal pattern affecting all stations. Similarly, omitting time-specific trends can lead to inaccurate predictions, such as overestimating weekend demand based on weekday patterns. These temporal factors are also likely to interact with other covariates, like weather or station location, further complicating the modeling process.

Daily trip counts at BIXI stations often exhibit **overdispersion**, where the observed variance in trip counts is greater than what a simple Poisson distribution assumes. This is driven by numerous factors, such as differences in station characteristics, variability in daily demand due to external factors (e.g., weather or events), and unobserved heterogeneity among stations. For example, some stations might experience sporadic surges in usage during special events or redistribution efforts, leading to extreme variability in trip counts. Overdispersion can severely affect the reliability of statistical models if not properly addressed. In a standard Poisson regression, the assumption of equality between the mean and variance leads to underestimated standard errors, resulting in overly confident parameter estimates and p-values. This can give the false impression that certain variables are significant predictors when, in fact, the variability in the data has not been properly accounted for. Ignoring overdispersion can also reduce the predictive accuracy of the model, especially for stations with atypical or highly variable usage patterns. This issue becomes particularly problematic when scaling the model for operational planning or resource allocation, where small errors can have large downstream effects.

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, 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.