Statistical Analysis of BIXI Montréal Bike Rentals - Part III

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

Introduction

This project explores BIXI Montréal's bike-sharing data from the 2023 season, focusing on understanding the factors influencing trip durations at the station level. The dataset includes individual trip records filtered to trips under two hours between May and October. Key variables such as borough location, departure month, day of the week, and average trip duration provide the foundation for analyzing patterns and trends.

The primary goal is to investigate whether trip durations vary across boroughs and between weekends and weekdays using statistical modeling approaches. Through a series of linear regression models, including both fixed and random effects, this analysis aims to quantify relationships, assess variability, and evaluate the assumptions of independence and station-level effects. The findings contribute to understanding usage patterns, offering actionable insights for operational and strategic decisions related to BIXI's service planning and optimization.

Exploratory Data Analysis

Before conducting an exploratory data analysis, the following preprocessing steps were applied to clean the dataset:

- Column Renaming: The `mm` and `dd` columns, representing the departure month and day, were renamed to Month and Day for better readability and alignment with analysis conventions.
- Handling Categorical Variables: Variables such as `Month`, `Day`, `arrondissement`, and `wday` were converted to factors to appropriately reflect their categorical nature.
- Creating a Weekend Indicator: A binary variable `weekend` was created to distinguish weekends (Saturday and Sunday, coded as 1) from weekdays (Monday to Friday, coded as 0).
- Removing Unnecessary Columns: The `Day` column was removed after being factorized, as it was not required for the subsequent analysis.

The BIXI dataset was verified for missing values, and no missing observations were found, ensuring a clean dataset for analysis. A total of 100 distinct stations were identified. The dataset includes 1000 observations, reflecting trips taken between May and October 2023. The number of observations was identified through the number of rows in the dataset. Table 1 summarizes the descriptive statistics for the number of observations at different stations. The minimum number of observations per station was 3, while the maximum reached 18. The median value of 10 and the mean value of 10 indicate a fairly balanced distribution of trips across stations, with slight variability as shown by the first and third quartiles.

Table 1: Descriptive Statistics for Observations Across Stations

Statistic	Value
Minimum	3.00
1st Quartile	8.00
Median	10.00
Mean	10.00
3rd Quartile	11.25
Maximum	18.00

Statistical Modeling

In this section, we analyze the factors affecting BIXI trip durations using linear regression models. The models include borough and weekend status as predictors, with adjustments for station-level effects to account for differences between stations. This allows us to explore how trip durations vary across locations and between weekdays and weekends.

Model 1: Linear Regression with Independent Observations

In Question 1, a simple linear regression model is used to explore the relationship between trip durations and key predictors, including borough (`arrondissement`) and `weekend` status. This model assumes that all observations are independent and does not account for variability at the station level.

$$\text{dur}_i = \beta_0 + \sum_{k=1}^{K-1} \beta_{1k} \cdot \text{arrondissement}_{ik} + \beta_2 \cdot \text{weekend}_i + \epsilon_i$$

where:

dur_i is the trip duration for observation i,

β_0 is the intercept (mean trip duration for the reference borough on weekdays),

β_{1k} are the coefficients for borough k (relative to the reference borough, assuming we have K boroughs),

$\text{arrondissement}_{ik}$ is the dummy variable for borough k for observation i,

(1 if the trip is in borough k, 0 otherwise),

β_2 is the coefficient for weekends (difference between weekend and weekday trips),

weekend_i is an indicator variable for weekends (1 if the day is a weekend, 0 otherwise),

ϵ_i is the residual error term for observation i, assumed to follow $\epsilon_i \sim N(0, \sigma^2)$.

Table 2: Coefficients of the Linear Regression Model (Model 1).

Covariate	**Value**	**Std. Error**	**t-value**	**p-value**
(Intercept)	15.566262	0.4627001	33.64223	0.0000
arrondissementLe Plateau-Mont-Royal	-4.087081	0.4998713	-8.17627	0.0000
arrondissementLe Sud-Ouest	-3.703873	0.7647247	-4.84341	0.0000
arrondissementMercier - Hochelaga-Maisonneuve	0.116690	0.5943238	0.19634	0.8444
arrondissementRosemont - La Petite-Patrie	-1.554990	0.5389884	-2.88502	0.0040
arrondissementVille-Marie	-0.671287	0.5075214	-1.32268	0.1862
arrondissementVilleray - Saint-Michel - Parc-Extension	-2.513639	0.7455455	-3.37154	0.0008
weekend	1.897106	0.2597424	7.30380	0.0000

The intercept represents the average trip duration for the reference borough (Côte-des-Neiges - Notre-Dame-de-Grâce), on weekdays. Based on the model summary shown in Table 2, the intercept is 15.566, indicating that the average trip duration for trips starting in the reference borough on a weekday is approximately 15.57 minutes. The regression parameter for the `weekend` covariate is 1.897, which indicates that trips taken on weekends are, on average, approximately 1.90 minutes longer than trips taken on weekdays, holding the borough constant. This result is statistically significant ($p < 0.001$), meaning the weekend effect on trip duration is unlikely to be due to randomness.

The extended model includes a fixed effect for stations, represented as:

$$\text{dur}_i = \beta_0 + \sum_{k=1}^{K-1} \beta_{1k} \cdot \text{arrondissement}_{ik} + \beta_2 \cdot \text{weekend}_i + \sum_{l=1}^{L-1} \beta_{3l} \cdot \text{station}_{il} + \epsilon_i$$

where:

dur_i is the trip duration for observation i,

β_0 is the intercept (mean trip duration for the reference borough at the reference station on weekdays),

β_{3l} are the fixed effect coefficients for station l (relative to the reference station, having L stations),

station_{il} is the dummy variable for station l for observation i,

(1 if the trip starts at station l, 0 otherwise),

ϵ_i is the residual error term for observation i, assumed to follow $\epsilon_i \sim N(0, \sigma^2)$.

By including `station` as a fixed effect, the model attempts to account for station-specific variability in trip durations. However, this approach leads to perfect collinearity in the dataset, as each station belongs to exactly one borough. This collinearity results in a singular design matrix, preventing the model from being fit. The error message generated in R:

```
Error in glsEstimate(glsSt, control = glsEstControl) :
  computed "gls" fit is singular, rank 102
```

This issue arises because `arrondissement` already captures the variability attributed to `station`. Including both variables in the model introduces redundant information, making the design matrix non-invertible. To confirm the perfect collinearity, an algorithm was applied to check the relationship between `station` and `arrondissement`:

1. Group the dataset by `station` and `arrondissement`.
2. Count the number of unique boroughs for each station.
3. Verify whether any station is associated with more than one borough.

The results show that each station is uniquely associated with exactly one borough, confirming that `station` and `arrondissement` are perfectly collinear.

The result of this analysis implies that one of the collinear variables (either `station` or `arrondissement`) must be removed from the model:

- If station-specific effects are of interest, include `station` and omit `arrondissement`.
- Alternatively, if borough-level effects are of interest, include `arrondissement` and omit `station`.

Model 2: Random Intercept Model

In Question 2, a linear mixed-effects model is employed to analyze the factors influencing average trip durations, accounting for the hierarchical structure of the data. The model includes boroughs (`arrondissement`) and weekend status as fixed effects to examine how these factors impact trip durations. Additionally, a random intercept is introduced for each station (b_i) to capture unobserved heterogeneity across stations. By incorporating the random intercept, the model accounts for the dependency among observations within the same station, ensuring more reliable and accurate estimates of the fixed effects. This approach allows us to investigate whether borough-level differences and weekend effects significantly explain variations in trip durations.

$$\text{dur}_{ij} = \beta_0 + \sum_{k=1}^{K-1} \beta_{1k} \cdot \text{arrondissement}_{ijk} + \beta_2 \cdot \text{weekend}_{ij} + b_i + \epsilon_{ij}$$

where:

dur_{ij} is the trip duration for the j-th observation at station i,

β_0 is the fixed intercept (mean trip duration for the reference borough on weekdays across all stations),

β_{1k} are the fixed effect coefficients for borough k (relative to the reference borough),

$\text{arrondissement}_{ijk}$ is the dummy variable for borough k for the j-th observation at station i,
(1 if the trip is in borough k, 0 otherwise),

β_2 is the fixed effect coefficient for weekends (difference between weekend and weekday trips),

weekend_{ij} is an indicator variable for weekends (1 if the day is a weekend, 0 otherwise),

b_i is the random intercept for station i, which captures station-specific deviations from the mean,
with $b_i \sim N(0, \sigma_b^2)$, representing station-specific deviations from the population mean,

ϵ_{ij} is the residual error term for the j-th observation at station i, assumed to follow $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$,
where ϵ_{ij} and b_i are independent.

Estimated Within-Station Correlation

The within-station correlation measures the proportion of the total variance in trip durations that is attributable to differences between stations. It is calculated as:

$$\text{Within-Station Correlation} = \frac{\sigma_{\text{b}}^2}{\sigma_{\text{b}}^2 + \sigma_\epsilon^2}$$

Based on the model results:

$$\sigma_{\text{b}}^2 = 9.0014 \quad \text{(random intercept variance)}$$
$$\sigma_{\epsilon}^2 = 5.1465 \quad \text{(residual variance)}$$

The resulting within-station correlation is:

$$\text{Within-Station Correlation} = \frac{9.0014}{9.0014 + 5.1465} \approx 0.636$$

This indicates that approximately 63.6% of the variability in trip durations can be attributed to differences between stations, while the remaining 36.4% is due to variability within stations. The relatively high within-station correlation underscores the importance of station-level grouping in explaining trip duration variability, justifying the inclusion of a random intercept for station in the model.

Assumption of Independence

To assess whether the assumption of independence is reasonable, we compared the linear regression model (Model 1) with the linear mixed-effects model (Model 2) using a likelihood ratio test.

H_0: The independence assumption holds; adding a random intercept does not improve model fit (Model 1 is sufficient) ($\sigma_{\text{b}}^2 = 0$).

H_1: The independence assumption does not hold; adding a random intercept improves model fit significantly (Model 2 is better) ($\sigma_{\text{b}}^2 > 0$).

The results of the likelihood ratio test are summarized in Table 3.

Table 3: Comparison of Model 1 (Linear Regression) and Model 2 (Linear Mixed-Effects) Using a Likelihood Ratio Test.

Model	**df**	**AIC**	**BIC**	**Log-Likelihood**	**Test**	***p*-value**
Model 1	9	5455.845	5499.942	-2718.922		
Model 2	10	4765.218	4814.216	-2372.609	1 vs 2	< 0.0001

Based on the results of the likelihood ratio test, the test statistic is:

$$L.Ratio = -2 \times (-2718.922 - (-2372.609)) = 692.626$$

The p-value associated with the test ($p < 0.0001$) is far below the significance threshold of $\alpha = 0.01$. This strongly rejects the null hypothesis (H_0) that the independence assumption holds. It is important to note that, theoretically, the p-value for a likelihood ratio test should be divided by 2 to reflect the one-sided nature of the hypothesis test. However, given that the p-value is already extremely small, applying this correction would not change the outcome of the test. For practical purposes, this step is omitted here.

Including a random intercept for stations in Model 2 leads to a highly significant improvement in model fit. This indicates that the independence assumption in Model 1 is not reasonable, as station-level grouping explains a substantial portion of the variability in trip durations. The mixed-effects model (Model 2) is therefore more appropriate, as it accounts for the hierarchical structure of the data and the dependency of observations within stations. This conclusion aligns with the earlier observation of a high within-station correlation (≈ 0.636), which suggests that station-level effects contribute substantially to the variability in trip durations.

Significant Variation in Average Trip Duration Across Boroughs

The hypothesis test to determine whether there is significant variation in average trip duration across boroughs is conducted using a likelihood ratio test. The models compared are:

- The **full model**, which includes `arrondissement` (borough) and `weekend` as fixed effects, with a random intercept for stations.
- The **reduced model**, which includes only `weekend` as a fixed effect, with a random intercept for stations.

The hypotheses for the test are:

> H_0: There is no significant difference in average trip duration across boroughs ($\beta_k = 0$ for all k, where k corresponds to the boroughs).
>
> H_a: There is a significant difference in average trip duration between the reference borough and at least one other borough ($\beta_k \neq 0$ for at least one k).

The results of the likelihood ratio test are summarized in Table 4.

Table 4: Comparison of Reduced and Full Models Using a Likelihood Ratio Test.

Model	**df**	**AIC**	**BIC**	**Log-Likelihood**	**Test**	**p-value**
Reduced Model	4	4790.485	4810.116	-2391.242		
Full Model	10	4775.141	4824.218	-2377.570	1 vs 2	1×10^{-4}

The likelihood ratio statistic is calculated as:

$$L.Ratio = -2 \times (\text{log-likelihood of reduced model} - \text{log-likelihood of full model})$$

Substituting the values:

$$\begin{aligned} \text{Log-likelihood of reduced model} &= -2391.242 \\ \text{Log-likelihood of full model} &= -2377.570 \\ L.Ratio &= -2 \times (-2391.242 - (-2377.570)) = 27.344 \end{aligned}$$

The resulting p-value is 1×10^{-4}, which is well below the significance threshold of $\alpha = 0.01$. This strongly rejects the null hypothesis (H_0), indicating that there is significant variation in average trip duration across boroughs. The results demonstrate that including `arrondissement` as a fixed effect significantly improves model fit, indicating that borough-level differences influence average trip durations. This supports the idea that trip durations vary systematically across boroughs, highlighting the importance of incorporating borough-level predictors to enhance the model's explanatory power.

Model 3: Random Intercept and Random Weekend Effect Model

In Question 3, a linear mixed-effects model is employed to analyze how the effect of weekends on average trip duration varies across stations. The model includes a fixed effect for weekends and random effects for both the intercept and the weekend variable at the station level. Importantly, the random effects are assumed to be independent, and residual errors are assumed to be independent and identically distributed.

This model captures the heterogeneity in average trip durations across stations (b_{i0}) and how the weekend effect varies between stations (b_{i2}). Moreover, the model assumes that the random intercepts (b_{i0}) and random slopes (b_{i2}) are independent of each other. Residual errors (ϵ_{ij}) are assumed to be independent and identically distributed, following a normal distribution with mean 0 and constant variance. Additionally, it is assumed that observations from different stations are independent, meaning that random effects and residual errors for one station do not influence another station.

$$\text{dur}_{ij} = \beta_0 + b_{i0} + \sum_{k=1}^{K-1} \beta_{1k} \cdot \text{arrondissement}_{ijk} + (\beta_2 + b_{i2}) \cdot \text{weekend}_{ij} + \epsilon_{ij}$$

where:

β_2 is the fixed effect coefficient for weekends (difference between weekend and weekday trips),

b_{i0} is the random intercept for station i, capturing station-specific deviations in average trip durations,

b_{i2} is the random effect coefficient for weekends at station i, capturing station-specific deviations in weekend effects, with $b_{i0} \sim N(0, \sigma_{b0}^2)$ and $b_{i2} \sim N(0, \sigma_{b2}^2)$,

ϵ_{ij} is the residual error term for the j-th observation at station i, assumed to follow $\epsilon_{ij} \sim N(0, \sigma^2)$, where b_{i0}, b_{i2}, and ϵ_{ij} are mutually independent.

Station with the Greatest Weekend Effect

Using the fitted model, the station-level predictions for the weekend effect were obtained by summing the fixed effect for weekends (β_2) and the random effect for weekends (b_{i2}). This approach allows for the identification of stations with the greatest deviations in weekend trip durations compared to weekdays.

The station with the greatest increase in average trip duration on weekends is **Bassin olympique (Chemin du Chenal le Moyne)**, with an increase of approximately **8.7414 minutes**. This result reflects the combined influence of the fixed effect and the station-specific random effect, indicating that trips originating from this station tend to last significantly longer on weekends than on weekdays.

Significant Variation in Weekend Effects Across Stations

To determine whether the effect of weekends on average trip durations varies significantly across stations, a likelihood ratio test was conducted. The test compares two models: one with random effects for both the intercept and the weekend variable at the station level where the random effects were assumed to be independent (**Model 3**), and another with only a random intercept at the station level (**No Weekend Random Effect**).

The hypotheses for the test are:

H_0: The effect of weekends does not vary across stations ($\sigma^2_{b2} = 0$).

H_a: The effect of weekends varies significantly across stations ($\sigma^2_{b2} > 0$).

The LRT is summarized in Table 5:

Table 5: Comparison of Reduced and Full Models Using a LRT for Weekend Effect.

Model	**df**	**AIC**	**BIC**	**Log-Likelihood**	**Test**	***p*-value**
Reduced Model	10	4765.218	4814.216	-2372.609		
Full Model	11	4728.600	4782.497	-2353.300	1 vs 2	1×10^{-4}

From the model outputs:

$$\begin{aligned} \text{Log-likelihood of Model 3} &= -2353.300, \\ \text{Log-likelihood of Model 3 (No Weekend Random Effect)} &= -2372.609, \\ L.Ratio &= -2 \times (-2372.609 - (-2353.300)) = 38.61779. \end{aligned}$$

The degrees of freedom for the test are $df = 2 - 1 = 1$, corresponding to the difference in the number of random effects. The resulting p-value is < 0.0001, which is well below the significance threshold of $\alpha = 0.01$ (Note that this was a non-standard test but as the p-value was very small, we didn't divide it by 2). This result strongly rejects the null hypothesis (H_0), indicating that the weekend effect varies significantly across stations. The inclusion of a random effect for weekends at the station level is therefore necessary to appropriately model the variability in weekend trip durations.

Correlation Values for Weekends and Weekdays

To evaluate the correlation between trip durations for two observations leaving from the same station on weekdays or weekends, we rely on the variance components extracted from the mixed-effects model in Model 3. As already stated, $b_{i0} \sim N(0, \sigma^2_{\text{b0}})$ is the random intercept for station i, $b_{i2} \sim N(0, \sigma^2_{b2})$ is the random slope for weekends, $\epsilon_{ij} \sim N(0, \sigma^2_{\text{residual}})$ is the residual, and the covariance between the random intercept and random slope is assumed to be zero, i.e., $\text{Cov}(b_{i0}, b_{i2}) = 0$. These assumptions ensure that the correlation structure depends entirely on the specified variance components.

For two weekday observations (`weekend` = 0) leaving from the same station, the correlation arises solely from the shared random intercept. The formula for this correlation is given by:

$$\text{Corr}_{\text{weekday, same station}} = \frac{\sigma^2_{\text{intercept}}}{\sigma^2_{\text{intercept}} + \sigma^2_{\text{residual}}} = \frac{7.640813}{7.640813 + 4.619891} = 0.6231953$$

This reflects the proportion of variability in trip durations explained by the random intercept relative to the total variability.

For two weekend observations (`weekend` $= 1$) leaving from the same station, both the random intercept and the random slope contribute to the correlation. The formula for the weekend correlation is

$$\text{Corr}_{\text{weekend, same station}} = \frac{\sigma^2_{\text{intercept}} + \sigma^2_{\text{weekend}}}{\sigma^2_{\text{intercept}} + \sigma^2_{\text{weekend}} + \sigma^2_{\text{residual}}} = \frac{7.640813 + 2.931141}{7.640813 + 2.931141 + 4.619891} = 0.6958967$$

This indicates that the variability in trip durations during weekends is influenced by the combined effects of the intercept and the weekend-specific random slope, relative to the total variability.

For two observations leaving from different stations, there are no shared random effects, as the random intercepts and slopes are station-specific. Consequently, the correlation between these observations is zero:

$$\text{Corr}_{\text{weekday, diff station}} = 0.$$

Based on the model assumptions, the correlation values do not vary across stations because the variance components $(\sigma^2_{\text{intercept}}, \sigma^2_{\text{weekend}}, \sigma^2_{\text{residual}})$ are constant across all stations. This homogeneity implies that the correlation structure is station-invariant. However, this conclusion depends on the assumption that the variance components are consistent for all stations. If the model were extended to allow station-specific variances or other hierarchical structures, the correlations might differ between stations. Under the current model specification, the correlation values are uniform across stations and determined solely by the fixed variance components.

Reflections

The analyses in the first two parts of the project relied on assumptions of independent observations, which neglected the hierarchical nature of the data and the substantial within-group correlation present at the station level. This oversight was critical because observations from the same station shared common characteristics and influences, such as location-specific factors, infrastructure, or user behavior patterns. By failing to account for these dependencies, the models in Parts 1 and 2 likely produced underestimated standard errors, biased coefficients, and unreliable significance tests, particularly for predictors affected by station-level grouping. Moreover, this independence assumption could have masked nuanced relationships, such as varying effects across stations. Addressing this structure in Part 3 with mixed-effects models provided a much-needed correction, ensuring robust and meaningful insights while highlighting the limitations of the earlier approaches.

Conclusion

This project explored the factors influencing BIXI Montréal bike trip durations using a combination of fixed and random effects models. By incorporating borough-level and station-level predictors, the analyses highlighted the importance of considering hierarchical data structures to ensure accurate statistical insights. The results demonstrated significant variations in trip durations across boroughs and weekends, as well as the critical role of station-specific variability.

Limitations

Despite the robustness of the mixed-effects models used, several limitations remain. Firstly, the analysis assumes that the variance components are constant across all stations, which may oversimplify station-specific dynamics. Additionally, potential interactions between boroughs and weekends were not explored, which could provide further insights into user behavior patterns. Finally, the models assume linear relationships and independence of random effects, which might not fully capture the complexities of real-world bike-sharing systems.

Contributions

Each team member contributed to different parts of the project. Gilles worked on Question 1 and also helped with reviewing the report. Olivia worked on Question 2 and compiled the final report to ensure it was well-organized and cohesive. Olivier handled the exploratory data analysis (EDA) and developed the R code. Pedram concentrated on Question 3 and meticulously rechecked all other sections of the project. This included identifying and correcting errors in the code and analyses for other questions to ensure accuracy and consistency throughout the report. These contributions ensured the project was completed efficiently and thoroughly.