Statistical Analysis of BIXI Montréal Bike Rental Patterns

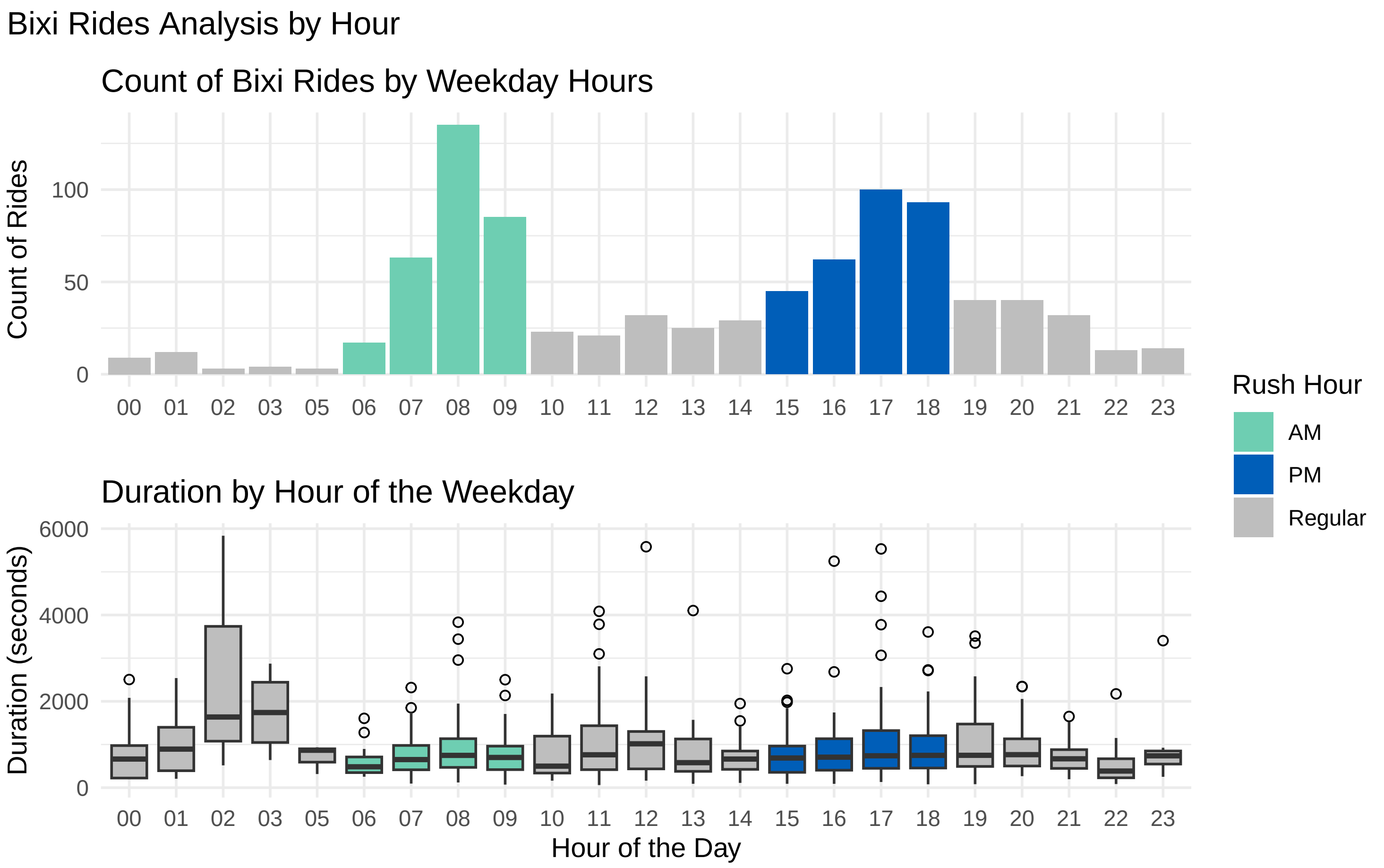
Introduction

BIXI Montréal is a non-profit organization that manages a bike sharing system in the metropolitan area. For the first part of this project, we will conduct an analysis of open access data on BIXI bike rentals.

Explanatory Data Analysis

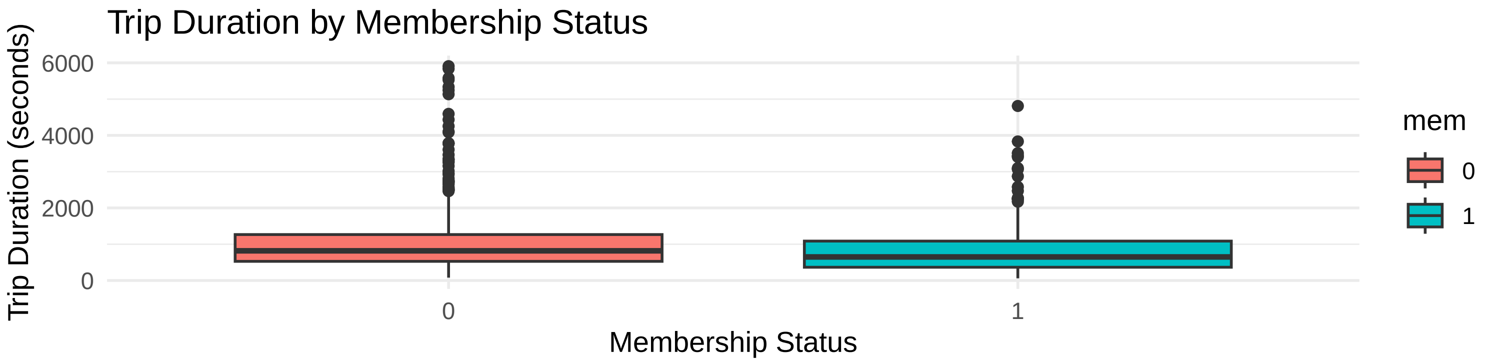
In our journey to uncover patterns in BIXI bike-sharing data, we begin by asking a fundamental question: How do users interact with the service throughout the day? In the plot below, we illustrate the distribution of rides taken at various hours, revealing a compelling story. It indicates that the majority of the utilization made by BIXI users falls within the conventional morning and evening rush-hour windows. This peak utilization period reflects the travel patterns of Montreal commuters, as bike-sharing emerges as a practical alternative to other transportation modes during times of heavy congestion. Outside of these peak times, bike usage tapers off and would likely see a shift towards more leisure usage during off-peak hours.

```
bixi_ride_analysis_plot
```



The exciting aspects of the BIXI service have to do with the difference in behavior amongst members and non-members. By comparing the trip durations amongst the two groups, we notice that non-members take longer trips. The *Membership Distribution* plot shows that members, many of whom likely use BIXI as part of their commute, take shorter, more efficient rides. Non-members seem to use BIXI much more casually, taking longer trips, which may be good for recreation or sightseeing. This presents an important divide between casual and committed users regarding membership plans and incentives. (Further breakdown for AM, PM, and rush hour duration are appended to Business Question 1 Analysis.)

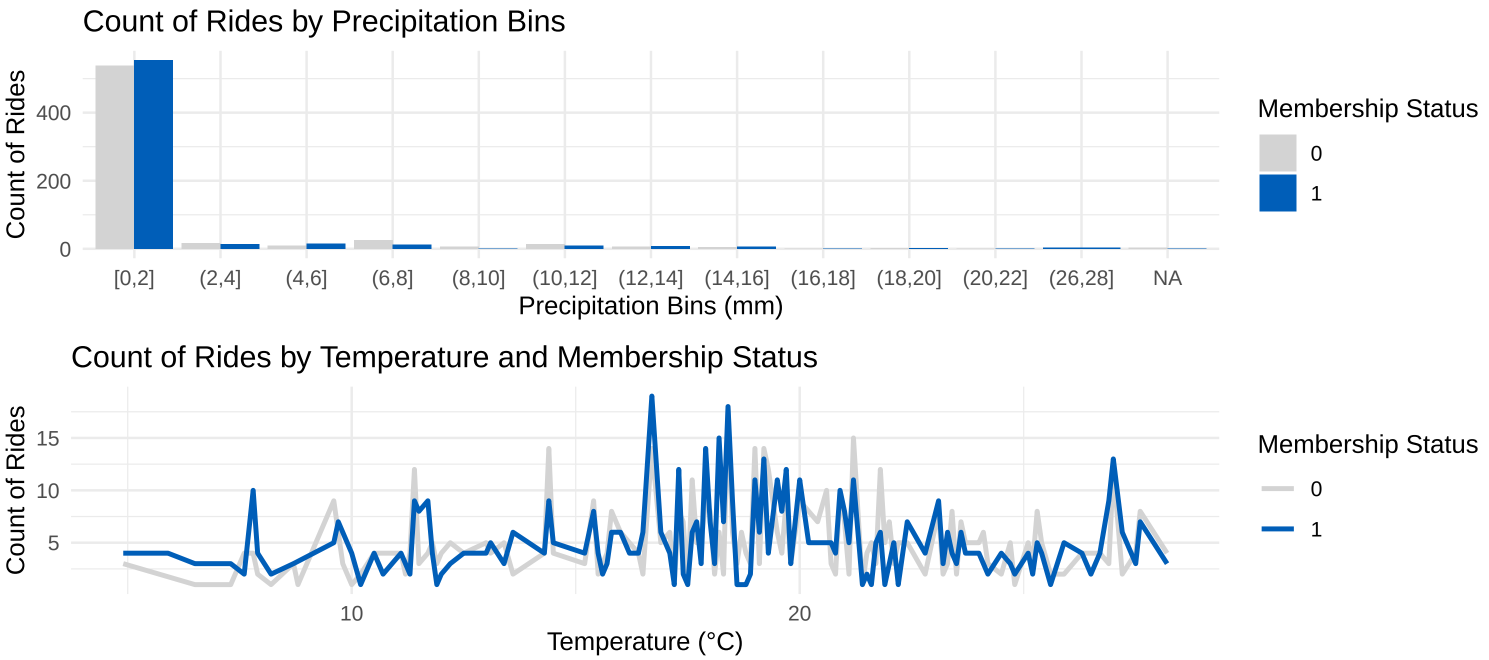
```
dur_mem_box_plot
```



We further analyze the explanatory variables by checking the balance of categorical variables such as the number of rides across other days of the week. As can be seen from Appendix A, the number of rides is quite well distributed across all weekdays. Appendix C also describes the raw data in detail with the added holiday dates for the year 2021. These insights ensure that no single day or set of variables dominates the data and therefore provide a very sound basis for statistical modeling.

The other factor to take into consideration, as we dig deeper, revolves around weather. Montreal is an unpredictable city in terms of climate, with this having the potential to have a big impact on how and when people will make use of BIXI. By again linking temperature and precipitation with the frequency of trips, it becomes very easy to draw a pattern from it. Most trips happen in comfortable temperatures—between 15°C and 25°C. The *Count of Rides by Precipitation Bins* plot shows the distribution of temperature during trips. Below or above this comfortable range, bike usage really declines, with the fewest trips taken during spikes of hot or cold. The *Count of Rides by Temperature and Membership Status* plot also shows that even light precipitation discourages bikers, as rainfall over 5 mm is associated with substantially fewer trips. Such weather-based comprehension implies that BIXI users are sensitive to environmental conditions. This calls for a correlation analysis of weather variables with trip duration in detail, found in Appendix D.

```
grid.arrange(count_prec_plot, count_temp_mem_plot, ncol=1)
```



In general, the exploratory analysis provides a rich qualitative description of how BIXI is used. Each of membership status, weather, and time of day play distinct roles in shaping how and when people use the service. The insights provided will not only guide our next steps in statistical modeling but also have valuable implications for BIXI's operational strategy—optimizing bike availability during peak hours or tailoring marketing efforts to different users. This data seems to tell a story of consistency and adaptability, where at any given time of day, convenience and weather dictate the rhythm at which BIXI operates.

Preprocessing and Model Preparation

Before diving into the business questions, we decided to preprocess the data and decide which variables are most important for modelling. Starting with `dep`, it represents the departure time of the BIXI users and is initially stored in datetime format. We extracted the date partition using the `as.Date` function for further processing. The `mem` variable indicates membership status and is a binary variable. We converted it to a categorical factor with two levels: `Non-Member` and `Member`. The `wday` variable represents the day of the week and is categorized accordingly (e.g., Monday, Tuesday, . . .). We transformed it into a categorical variable for easier manipulation. With regards to the weather variables (`temp` and `prec`), these are both numerical. We converted `prec` into a binary variable because it contained many zero entries, indicating no precipitation. This is shown in the plot found in Appendix D. This conversion results in `prec=1` when `prec>0` and 0 otherwise. The rationale is that users are likely more concerned about whether it is raining rather than the specific amount of rain. Moving onto `rushhour`, there is no concept of rush hour on holidays. With that, when `weekend=1`, we set `rushhour=NonPeak`. For weekdays, `rushhour` is classified as `AM` for morning rush hour (value of 1), `PM` for evening rush hour (value of 2), and `NonPeak` for all other times (value of 3). The `dur` variable represents the trip duration in seconds and is retained as a numeric variable.

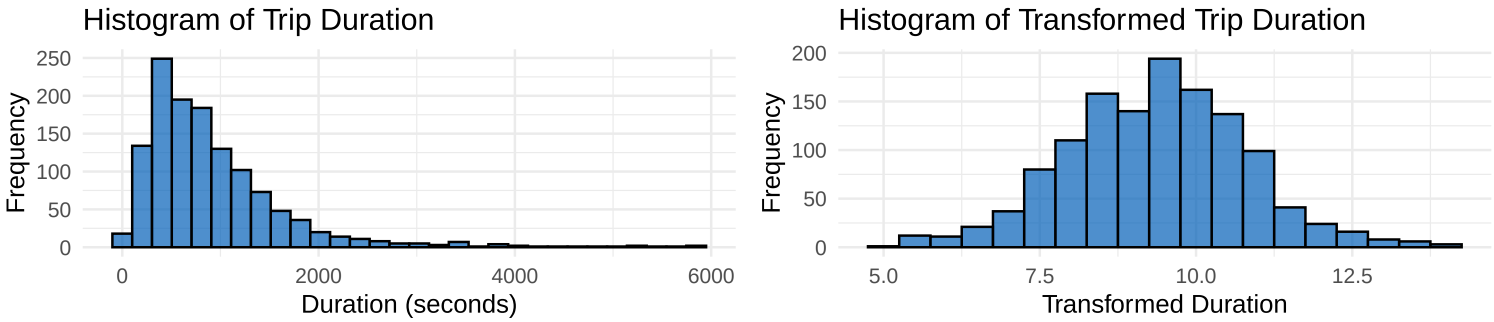
Then, we created a new binary variable called `weekend` to identify if a particular date is a holiday or falls on a weekend. This transformation was implemented to capture user behavior patterns, as people may use BIXI differently on weekends or holidays compared to regular weekdays. For instance, usage during weekdays might be higher due to the large number of users commuting to work. As a last step for the preprocessing, we removed the variables that we deemed as unnecessary for modelling. We removed `wday` and `date` as they were not required.

```r
base_model <- lm(dur ~ mem + temp + prec + rushhour + weekend + mem:prec + temp:prec, data = bixi_data)
```

Now that the data is preprocessed and we have removed redundant variables, let us briefly explain the base model. The base model (`base_model`) serves as the initial reference for understanding the relationship between the target variable (`dur`) and several explanatory variables. We included `mem`, `temp`, `prec`, `rushhour`, `weekend`, and relevant interaction terms to capture the potential influence of these factors on trip duration. The interaction terms considered are `mem:prec` and `temp:prec`. For `mem:prec`, it explores the influence of precipitation on members and non-members. For example, members might be less deterred by rain, given their commitment to biking, whereas non-members may opt for other forms of transportation when it rains. For `temp:prec`, it is justified since the likelihood of precipitation varies with temperature. Understanding how these two variables interact can provide insights into user behavior under different weather conditions.

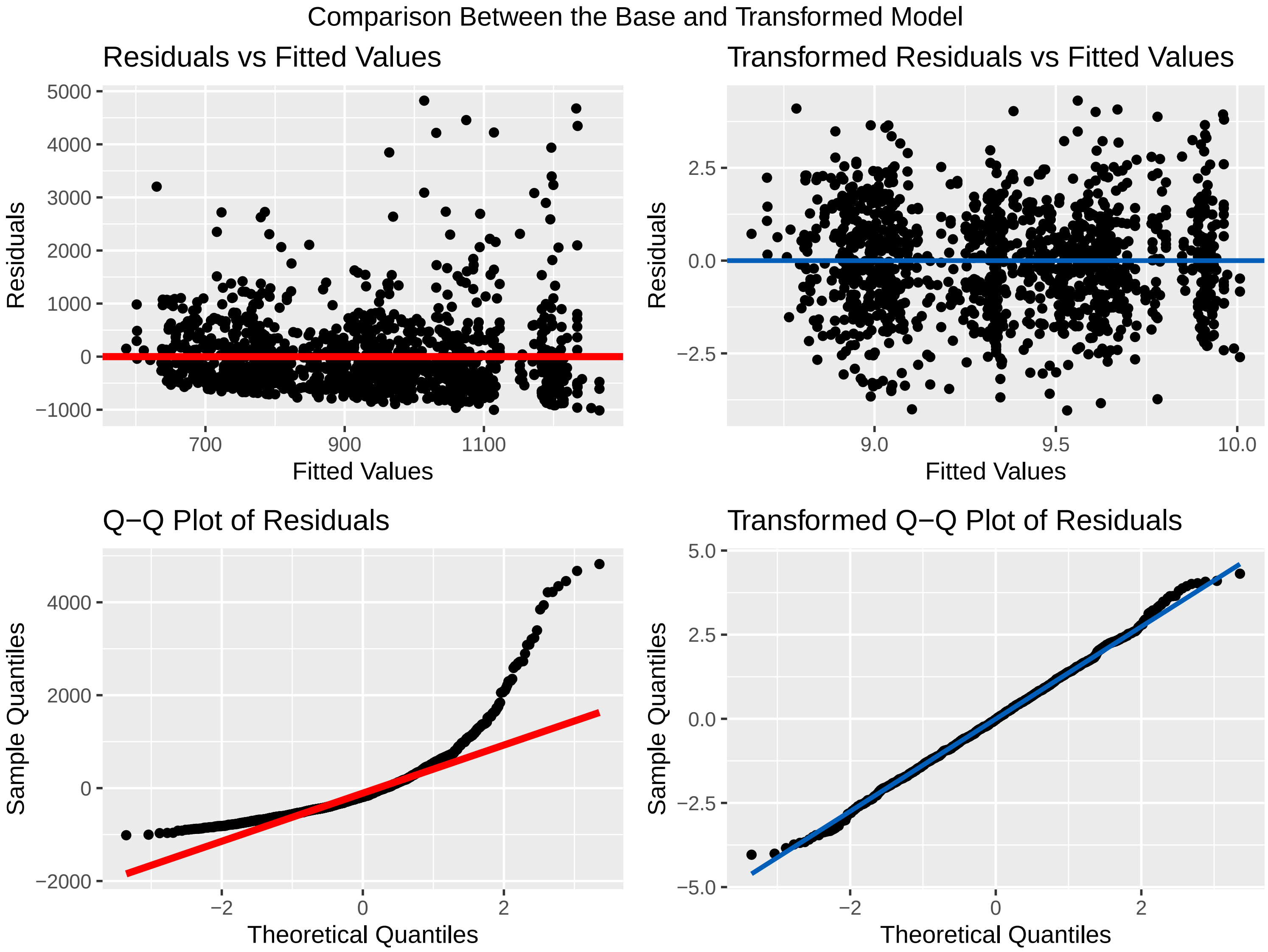
Looking at the table in Appendix B, we see that the sample size for each combination is enough and therefore we can rely on the estimates of betas. Now, in order to assess how well the base model fits the data and whether it meets the assumptions of the linear regression model, we plotted Residuals vs. Fitted as well as the Q-Q plot. As mentioned previously and by observing these plots we can clearly see that the model does not fit the data well, nor does it meet the assumptions. Applying the Box-Cox transformation method enhances our model, as demonstrated in the `dur` histogram below.

```r
grid.arrange(dur_hist_plot, dur_boxcox_hist_plot, ncol=2)
```



Now, looking at the Residuals vs. Fitted plot and Q-Q plot of the transformed base model below, we can see that it now roughly satisfies the assumptions of the linear regression model and fits the data much better!

```r
grid.arrange(residual_fitted_plot, transformed_residual_fitted_plot, qq_plot,
             transformed_qq_plot, ncol=2,
             top = "Comparison Between the Base and Transformed Model" )
```



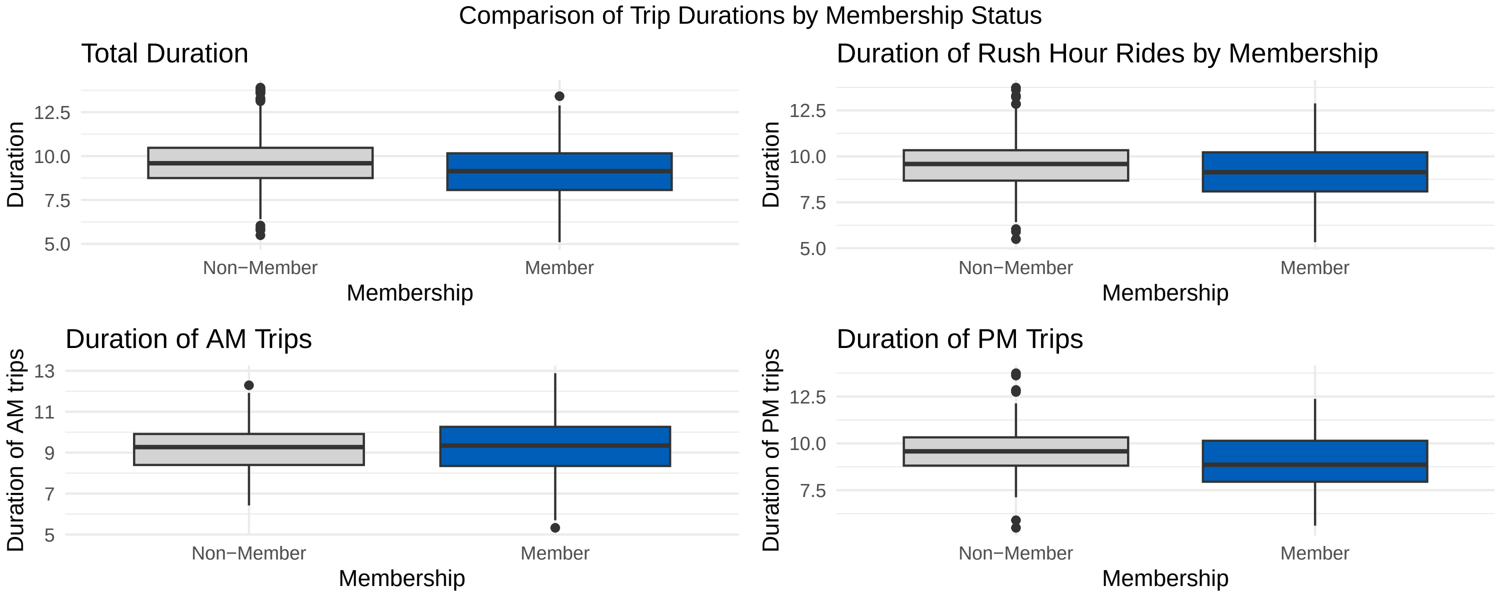
Going forward, we will be using the transformed model, hence our new target variable is `dur_boxcox`. **In our testing, whenever we mention "duration of trip", we are referring to the transformed duration.**

Business Question 1: On average, do BIXI members have shorter trips than non-members? Are the results the same if one adjusts for weekend vs. non-weekend usage?

Here, we want to determine whether, on average, if BIXI members have shorter trips than non-BIXI members. The null hypothesis is that there's no difference in average trip duration between BIXI members and non-members ($H_o : \mu_{\text{members}} = \mu_{\text{non-members}}$). The alternative hypothesis is that BIXI members have a different average trip duration compared to non-members ($H_a : \mu_{\text{members}} \neq \mu_{\text{non-members}}$).

We assessed how `dur_boxcox` varies across `mem` and deduced that membership tends to be generally lower than non-members. However, in the morning rush hours, there seems to be slightly more variability in members than non-members.

```r
grid.arrange(p1, p4, p2, p3, ncol = 2, top = "Comparison of Trip Durations by Membership Status")
```



```r
summary(model1)
```

```
##
## Call:
## lm(formula = dur_boxcox ~ mem + temp + prec + rushhour + weekend +
##     mem:prec + temp:prec, data = bixi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0384 -0.9301 -0.0204  0.9206  4.3113
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.645940   0.201190  47.944  < 2e-16 ***
## memMember       -0.579676   0.097114  -5.969  3.1e-09 ***
## temp            -0.006851   0.009551  -0.717  0.47335
## prec            -0.561836   0.339260  -1.656  0.09796 .
## rushhourPM       0.107460   0.116974   0.919  0.35845
## rushhourNonPeak  0.103417   0.116915   0.885  0.37657
## weekend1         0.291659   0.108840   2.680  0.00747 **
## memMember:prec   0.067165   0.166774   0.403  0.68722
## temp:prec        0.020595   0.016669   1.236  0.21687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 1251 degrees of freedom
## Multiple R-squared:  0.05421,    Adjusted R-squared:  0.04816
## F-statistic: 8.963 on 8 and 1251 DF,  p-value: 5.166e-12
```

Trips from members are `0.58` units shorter than non-members. This is statistically significant since `3.1e-09 < 0.001`. With that, on average, non-members' trips are longer than members' trips, and this difference is statistically significant. Referring to the *Membership Distribution* plot shown in Explanatory Data Analysis, we can also see that non-members have a higher median trip duration compared to members, suggesting that, on average, non-members take slightly longer trips. Therefore, we reject the null hypothesis.

Now, we will determine if the results are the same if we adjust for weekend vs. non-weekend usage. The null hypothesis is that there's no interaction between membership and weekend usage ($H_o : \beta_{\text{memNon-member:weekend1}} = 0$). The alternative hypothesis is that there is an interaction between membership and weekend usage ($H_a : \beta_{\text{memNon-member:weekend1}} \neq 0$).

```r
summary(model2)
```

```
##
## Call:
## lm(formula = dur_boxcox ~ mem + temp + prec + rushhour + weekend +
##     mem:prec + temp:prec + mem:weekend, data = bixi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0771 -0.9430 -0.0253  0.9439  4.3818
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.586850   0.202313  47.386  < 2e-16 ***
## memMember           -0.446487   0.111700  -3.997 6.78e-05 ***
## temp                -0.007243   0.009534  -0.760 0.447570
## prec                -0.575498   0.338665  -1.699 0.089509 .
## rushhourPM           0.108014   0.116752   0.925 0.355064
## rushhourNonPeak      0.102821   0.116694   0.881 0.378422
## weekend1             0.490608   0.136663   3.590 0.000344 ***
## memMember:prec       0.057806   0.166503   0.347 0.728518
## temp:prec            0.021346   0.016641   1.283 0.199823
## memMember:weekend1 -0.404858   0.168736  -2.399 0.016570 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 1250 degrees of freedom
## Multiple R-squared:  0.05855,    Adjusted R-squared:  0.05177
## F-statistic: 8.637 on 9 and 1250 DF,  p-value: 1.191e-12
```

The estimate of `memMember:weekend1` is `-0.404858` which means that on average, members' trip durations are `0.404858` units shorter on weekends compared to non-weekends. By including the interaction term (`mem:weekend`), we see that it is statistically significant at the 0.05 level (`0.0166 < 0.05`). Therefore, the relationship between `mem` and `dur_boxcox` changes depending on whether the trip occurs on a weekend or not.

```r
summary(contrast_model2)
```

```
## weekend = 0:
##  contrast              estimate    SE   df t.ratio p.value
##  (Non-Member) - Member    0.418 0.099 1250   4.220  <.0001
##
## weekend = 1:
##  contrast              estimate    SE   df t.ratio p.value
##  (Non-Member) - Member    0.822 0.142 1250   5.788  <.0001
##
## Results are averaged over the levels of: prec, rushhour
```

For `weekend = 0`, Non-Members, on average, have longer trip durations than members on non-weekends by `0.418` units, and this difference is statistically significant since `p-value` < 0.0001. For `weekend = 1`, Non-Members, on average, have longer trip durations than members on weekends by `0.822` units, and this difference is statistically significant since `p-value` < `0.0001`.

These results illustrate that the difference in trip duration between non-members and members change on weekends and on non-weekends. This could imply that non-members use BIXI bikes more for leisurely, longer trips on weekends, while members might use them more for shorter trips. Therefore, we reject the null hypothesis since there is an interaction between between membership and weekend usage. The interaction term is statistically significant, and it provides valuable information about how the membership effect varies by weekend status, so it's worth keeping in the model.

Business Question 2: Are trip durations impacted by weather factors? In light of the results you obtain, should your initial model(s) be revisited?

The goal of this analysis is to determine whether weather variables, specifically temperature (`temp`) and precipitation (`prec`), have a significant impact on trip durations in the Bixi dataset. We initially included weather variables in our model and now aim to test whether simplifying the model by excluding these variables is justified. The hypothesis tests the significance of weather factors by comparing a full model (with weather variables) to a reduced model (without weather variables).

$$H_0 : dur_{\text{transformed}} \sim \text{mem} + \text{rushhour} + \text{weekend}$$

$$H_1 : dur_{\text{transformed}} \sim \text{mem} + \text{rushhour} + \text{weekend} + \text{mem:prec} + \text{temp:prec}$$

The initial, more complex model (H_1) was fitted, and the following results were obtained:

```r
summary(weather_model_transformed)
```

```
##
## Call:
## lm(formula = dur_boxcox ~ mem + temp + prec + rushhour + weekend +
##     temp:prec + mem:prec, data = bixi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0384 -0.9301 -0.0204  0.9206  4.3113
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.645940   0.201190  47.944  < 2e-16 ***
## memMember       -0.579676   0.097114  -5.969  3.1e-09 ***
## temp            -0.006851   0.009551  -0.717  0.47335
## prec            -0.561836   0.339260  -1.656  0.09796 .
## rushhourPM       0.107460   0.116974   0.919  0.35845
## rushhourNonPeak  0.103417   0.116915   0.885  0.37657
## weekend1         0.291659   0.108840   2.680  0.00747 **
## temp:prec        0.020595   0.016669   1.236  0.21687
## memMember:prec   0.067165   0.166774   0.403  0.68722
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.4 on 1251 degrees of freedom
## Multiple R-squared:  0.05421,    Adjusted R-squared:  0.04816
## F-statistic: 8.963 on 8 and 1251 DF,  p-value: 5.166e-12
```

Regarding the key findings from the table above, the coefficients for the weather variables (`temp`,`prec`) and their interaction terms (`mem:prec`,`temp:prec`) are not statistically significant (`p-value > 0.05`). Other factors, such as membership (`mem`) and weekend status (`weekend`), show significance, but weather-related predictors do not appear to meaningfully contribute to predicting trip duration.

To formally test whether excluding the weather variables from the model is justified, we perform an ANOVA test, which compares the nested models H_0 (simplified model without weather factors) and H_1 (initial model with weather factors). The F-statistic is `1.1138` with a p-value of `0.3484`. Since the p-value is much higher than the significance level of `0.05`, there is not enough evidence to reject the null hypothesis. In order to have a more concrete test, we will perform the test shown below.

```r
anova_results
```

```
## Analysis of Variance Table
##
## Model 1: dur_boxcox ~ mem + rushhour + weekend
## Model 2: dur_boxcox ~ mem + temp + prec + rushhour + weekend + temp:prec +
##     mem:prec
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1255 2459.1
## 2   1251 2450.3  4    8.7266 1.1138 0.3484
```

Based on the ANOVA test results, the weather variables (`temp` and `prec`) do not significantly improve the model when included. The simplified model (H_0, excluding weather variables) fits the data just as well as the more complicated model (H_1) that includes them. Therefore, revisiting the initial model is recommended, as excluding weather variables does not harm the model's performance. The simplified model (without weather factors) provides an equally good fit and is a more efficient choice for modeling trip durations in the Bixi dataset.

Business Question 3: Do rush hour trip durations differ from those during non peak hours (i.e., not rush hour) during weekdays? Are there differences between the AM and PM rush hour weekday usage, respectively?

In this question we should narrow down our analysis to only the partition of the dataset that is in the weekday (& not holiday). Based on our analysis from Question 2, we concluded that including weather covariates in the model don't do us any good, so the simpler model is better. Therefore, the model that we are dealing with is the following:

$$dur_{\text{transformed}} = \hat{\beta}_0 + \hat{\beta}_1 \text{mem} + \hat{\beta}_2 \text{rushhour}$$

In which `mem` and `rushhour` are respectively binary and categorical variables of whether the user is a member and what state of the day they used the BIXI bikes. By fitting the linear regression model, the result is as follows:

```r
summary(rushhour_transformed_model)
```

```
##
## Call:
## lm(formula = dur_boxcox ~ mem + rushhour, data = bixi_weekday)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0166 -0.9247 -0.0387  0.8908  4.3675
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.39571    0.09417  99.778  < 2e-16 ***
## memMember       -0.43036    0.09474  -4.542 6.36e-06 ***
## rushhourPM       0.11362    0.11564   0.983    0.326
## rushhourNonPeak  0.10836    0.11564   0.937    0.349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.385 on 851 degrees of freedom
## Multiple R-squared:  0.02506,    Adjusted R-squared:  0.02162
## F-statistic: 7.291 on 3 and 851 DF,  p-value: 7.868e-05
```

Meaning that:

$$dur_{\text{transformed}} = 8.96 + 0.43 \cdot \mathbb{1}_{\text{mem=Non-member}} + 0.1 \cdot \mathbb{1}_{\text{rushhour=Non-Member}} + 0.11 \cdot \mathbb{1}_{\text{rushhour=PM}}$$

We are asked to do a test for the following subjects (only for the subset of data that is non-holiday):

$$H_0 : \mu_{\text{rushhour}} = \mu_{\text{NonPeak}}$$

$$H_a : \mu_{\text{rushhour}} \neq \mu_{\text{NonPeak}}$$

$$H_0 : \mu_{\text{AM}} = \mu_{\text{PM}}$$
$$H_a : \mu_{\text{AM}} \neq \mu_{\text{PM}}$$

Note that for the first test, as it's given in the hint, we will follow contrast testing via calculating estimated marginal means. Our test would be transformed to:

$$H_0 : 0.5\mu_{\text{AM}} + 0.5\mu_{\text{PM}} - \mu_{\text{NonPeak}} = 0$$
$$H_a : 0.5\mu_{\text{AM}} + 0.5\mu_{\text{PM}} - \mu_{\text{NonPeak}} \neq 0$$

Test 2 also can be included in the scope of testing contrasts:

$$H_0 : \mu_{\text{AM}} - \mu_{\text{PM}} = 0$$
$$H_a : \mu_{\text{AM}} - \mu_{\text{PM}} \neq 0$$

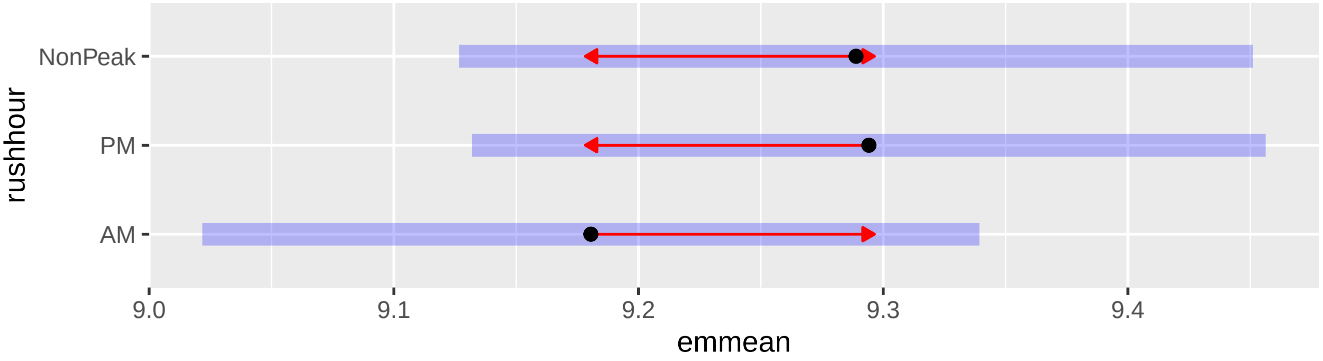
The results are:

```
summary(contrast_results)
```

```
##  contrast            estimate    SE  df t.ratio p.value
##  RushHour_vs_NonPeak  -0.0594 0.101 851  -0.589  0.5557
##  AM_vs_PM             -0.1084 0.116 851  -0.937  0.3490
##
## Results are averaged over the levels of: mem
```

According to the results, in the significance level of $\alpha = 5\%$, the differences aren't significant, meaning that we cannot find enough evidence to reject the null hypothesis. Consequently, with regards to the weekdays, we can deduce that there is no significant difference between duration time of people in the `rushhour` and in the `NonPeak` hours. Moreover, there is also not a significant difference in duration in AM and PM `rushhours`. We can also analyze with the estimated marginal means plot:

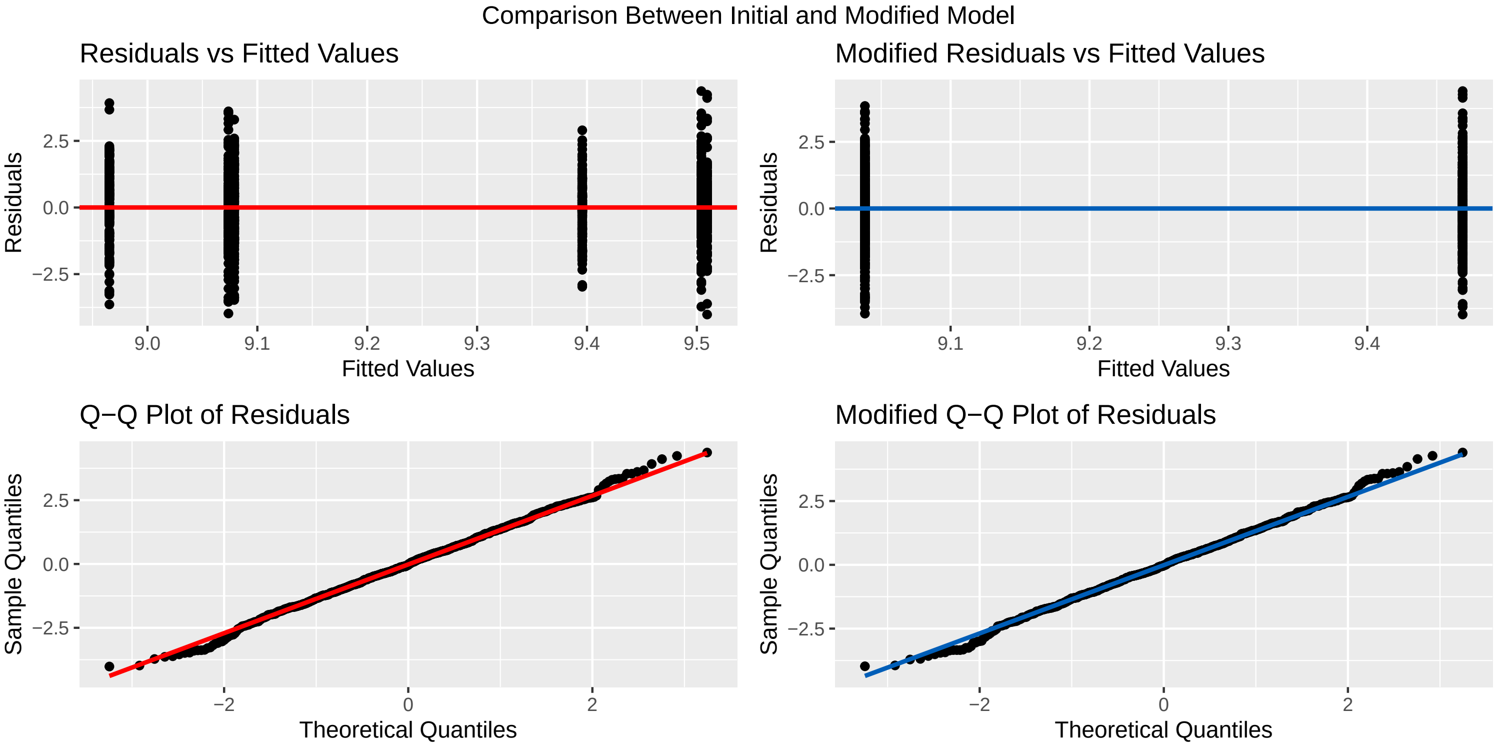
```
plot(emm_rushhour, comparisons = TRUE)
```



The red arrows connecting the categories indicate pairwise comparisons of trip durations. The overlap of confidence intervals suggests that the differences between the estimated means of these categories is not statistically significant. The AM and PM categories have almost identical estimated mean trip durations, as indicated by the significant overlap of their confidence intervals. The estimated mean of the transformed trip durations for AM and PM appear to be very similar, which aligns with the previous analysis where the contrast `AM_vs_PM` was not statistically significant. The `Non-Peak` category has a slightly higher mean trip duration compared to AM and PM, but the overlap of the confidence intervals suggests that this difference might not be significant.

In our initial model for this question and our modified model, the Q-Q plot and distribution of the residuals are as follows:

```
grid.arrange(rushhour_residual_fitted_plot,
             rushhour_transformed_residual_fitted_plot, rushhour_qq_plot,
             rushhour_transformed_qq_plot,  ncol=2,
             top="Comparison Between Initial and Modified Model")
```



Observing the initial model plots above, the hypothetical mean of zero we assume for the residuals fit their true mean. The residuals are fairly normal in the middle section, as most points lie close to the diagonal line. The residuals show heavy tails in both the left and right extremes, indicating the presence of more extreme values than expected under a normal distribution. The possible actions can be to apply a transformation to reduce the effect of heavy tails or further investigate the identified outliers to understand their impact on the model and consider potential remediation (e.g., robust regression, transformation, or exclusion). This analysis suggests that while the residuals seem mostly normal, there are deviations in the tails that could impact the overall model fit. Adjusting for these deviations might improve the model's performance and validity of inferences.

Conclusion

To summarize our solutions to the business questions, on average, non-members take slightly longer trips than members. If we were to adjust for weekend vs. non-weekend usage, non-members, on average, also take longer trips on weekends compared to non-weekends. With regards to weather factors, they do not significantly improve the model as trip durations are not impacted by them. With that, the initial model should be revisited and modified to not include weather factors. Lastly, there is no significant difference between duration time of people in the `rushhour` and in the `NonPeak` hours. Moreover, there is also not a significant difference in duration in AM and PM `rushhours`.

With regards to the limitations, according to Appendix B, the number of samples we have in each subcategory are not even roughly equal (we have more data on `prec=0` than `prec=1`), which makes the dataset imbalanced. Hence, the betas that we estimate are prone to overfitting to the case where `prec = 0`. Consequently, we would experience bias in our predictions with this data. In the Transformed Q-Q Plot of Residuals found in the Preprocessing and Model Preparation section, there are some deviations in the tails that might affect our linear regression assumptions that would be worth investigating and solving to have better and more reliable estimations.

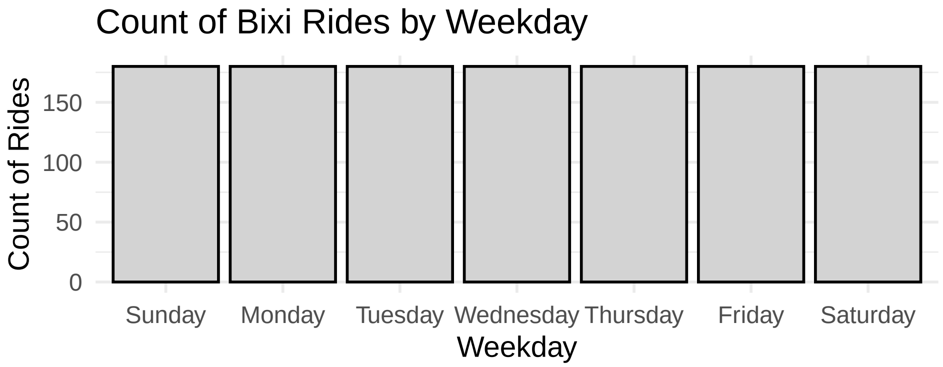
In our project, contributions were evenly distributed among team members, allowing everyone to engage with a specific question while also reviewing another. The tasks were distributed among team members to ensure efficient collaboration. **Olivia** played a key role in setting up the report, contributing to the Exploratory Data Analysis (EDA) report, and was responsible for the coding and analysis for Business Question 1, as well as the final review of the report. **Olivier** was tasked with the coding and analysis for Business Question 2 and the EDA analysis, while also assisting with the report's reproducibility review. **Gilles** was responsible for helping with the Box-Cox transformation and helping with EDA. He also contributed to the final review of the report. **Pedram** was responsible for the coding and analysis for Business Question 3, contributed to EDA analysis, and helped with the final review of the report. Team meetings were held at key stages to conclude the EDA validity and business questions, ensuring alignment across the group. All members played an essential role in bringing different aspects of the project to completion.

Appendix

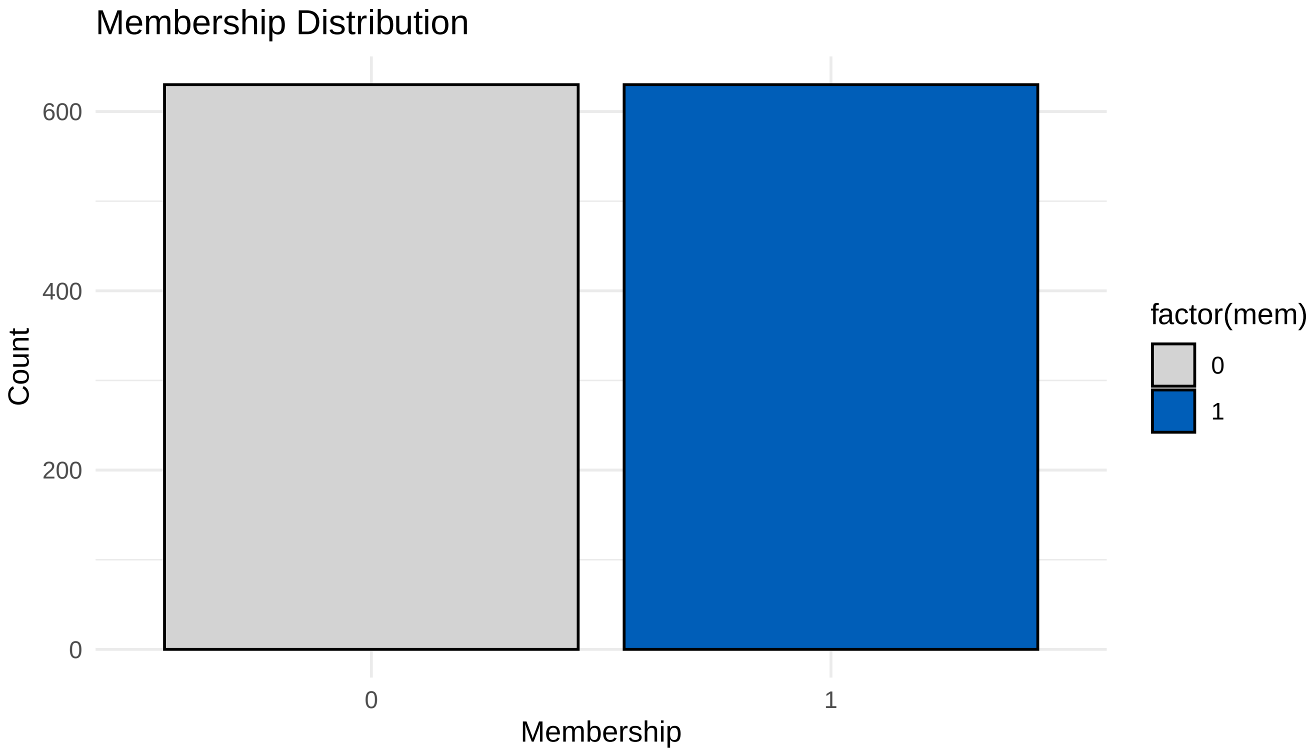
Appendix A

Verification of Balanced Data

```
rides_weekday_plot
```



```
mem_dist_plot
```



Appendix B

```r
# Display the resulting data frame
print(sample_size_interaction)
```

```
## # A tibble: 4 x 3
##   mem         prec sample_size
##   <fct>      <dbl>       <int>
## 1 Non-Member     0         416
## 2 Non-Member     1         214
## 3 Member         0         416
## 4 Member         1         214
```

Appendix C

Analyzing the Raw Data

The raw data consist of the records of every BIXI rental for the 2021 season. In particular, each observation consists in an individual trip and includes the following information: the start date and time, the start station, the end date and time, the end station, the total trip duration, and a variable indicating whether the user is a BIXI member. Only trips under 2 hours in the months extending from May to October, inclusively, are considered for the statistical analysis here. Note that in the 2021 season, members could use a regular BIXI for up to 45 minutes for free, and obtained rebates for electric bike rentals or longer trips. In addition to BIXI usage, weather information was merged with the BIXI data to provide the daily average temperature (in $\circ$C) and the daily cumulated amount of precipitation (in mm).

```r
bixi_data <- read.csv(here("Data", "MATH60604A-project-bixi_part1_team1.csv"))

# Get the head
kable(head(bixi_data))
```

dep	dur	mem	wday	temp	prec	rushhour
2021-05-21 07:46:06	703	0	Friday	25.3	0	1
2021-05-07 08:27:51	1284	0	Friday	9.7	0	1
2021-05-28 09:22:49	1486	0	Friday	7.8	0	1
2021-05-21 08:25:00	846	0	Friday	25.3	0	1
2021-05-14 07:44:10	1169	0	Friday	17.8	0	1
2021-05-28 14:40:01	623	0	Friday	7.8	0	3

```r
# Get the summary
kable(summary(bixi_data))
```

	dep	dur	mem	wday	temp	prec	rushhour
	Length:1260	Min. : 61.0	Min. :0.0	Length:1260	Min. : 4.90	Min. : 0.000	Min. :1
	Class :character	1st Qu.: 440.8	1st Qu.:0.0	Class :character	1st Qu.:15.70	1st Qu.: 0.000	1st Qu.:1
	Mode :character	Median : 738.0	Median :0.5	Mode :character	Median :18.80	Median : 0.000	Median :2
	NA	Mean : 919.9	Mean :0.5	NA	Mean :18.51	Mean : 1.346	Mean :2
	NA	3rd Qu.:1166.8	3rd Qu.:1.0	NA	3rd Qu.:21.80	3rd Qu.: 0.200	3rd Qu.:3
	NA	Max. :5908.0	Max. :1.0	NA	Max. :28.20	Max. :31.700	Max. :3

Holidays

Holidays that are incorporated for the 2021 calendar year:

- New Year's Day: January 1
- Good Friday: April 2
- Easter Monday: April 5
- National Patriots' Day: May 24
- Saint-Jean-Baptiste Day: June 24
- Canada Day: July 1
- Labour Day: September 6
- National Day for Truth and Reconciliation: September 30
- Thanksgiving: October 11
- Remembrance Day: November 11
- Christmas (Observed): December 27

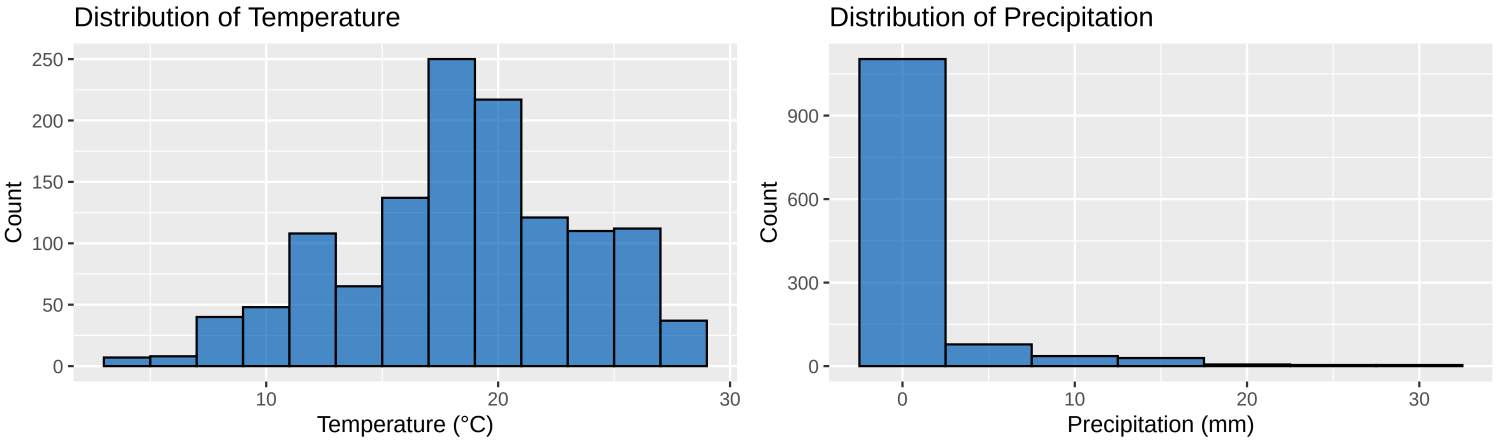
Appendix D

Analyzing Correlations between Precipitation and Temperature

In this section, we assess the contribution of two most significant weather variables that is, precipitation (`prec`) and temperature (`temp`), towards trip duration. Precisely, we want to know whether these variables will make a substantial improvement in the model.

Looking at the plots below, we see that both `prec` and `temp` do not seem to be symmetric, nor look like a normal distribution. We assessed the correlation between `prec` and `temp` at 0.05 significance level and obtained `p-value = 0.05088`, which is only slightly greater than α. We fail to reject the null hypothesis.

```
grid.arrange(temp_hist_plot, prec_temp_hist_plot, ncol=2)
```



Skewness in the distributions of temperature and precipitation can be observed from their respective histograms. Temperature exhibits a near-normal distribution with a slight positive skew, while precipitation is heavily right-skewed due to most days receiving little or no rainfall. A distribution as such gives reasonable indication that these variables may not behave in a linear manner with respect to trip duration and thus should be further investigated through correlation tests.

```
# Pearson correlation tests
pearson_test_temp_dur <- cor.test(bixi_data$temp, bixi_data$dur, method = "pearson")
pearson_test_prec_dur <- cor.test(bixi_data$prec, bixi_data$dur, method = "pearson")
pearson_test_temp_prec <- cor.test(bixi_data$temp, bixi_data$prec, method = "pearson")

# Output the results
cat("\nCorrelation P between Temperature and Duration:\n"); print(pearson_test_temp_dur)
```

```
##
## Correlation P between Temperature and Duration:


##
##  Pearson’s product-moment correlation
##
## data:  bixi_data$temp and bixi_data$dur
## t = -0.58385, df = 1258, p-value = 0.5594
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.07161915  0.03880170
## sample estimates:
##         cor
## -0.01645891
```

```r
cat("Correlation P between Precipitation and Duration:\n"); print( pearson_test_prec_dur)
```

```
## Correlation P between Precipitation and Duration:

##
##  Pearson's product-moment correlation
##
## data:  bixi_data$prec and bixi_data$dur
## t = -1.7456, df = 1258, p-value = 0.08113
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.104097806  0.006086803
## sample estimates:
##         cor
## -0.04915505
```

```r
cat("Correlation P between Temperature and Precipitation:\n"); print(pearson_test_temp_prec)
```

```
## Correlation P between Temperature and Precipitation:

##
##  Pearson's product-moment correlation
##
## data:  bixi_data$temp and bixi_data$prec
## t = 1.9543, df = 1258, p-value = 0.05088
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.0002086978  0.1099086654
## sample estimates:
##        cor
## 0.05501727
```

Tests of Pearson correlation have been conducted in order to check the linear relations concerning temperature, precipitation, and trip duration. The correlation between `temp` and `dur` is `-0.016`, the p-value is `0.559`, hence very weak and not statistically significant. Also, `prec` and `dur` is related at `-0.049`, correlating with a p-value of `0.081`. This indicates a weak negative correlation that is not statistically significant. Finally, the correlation between `temp` and `prec` was `0.055` (p-value = `0.051`), implying a very weak positive correlation between the two weather variables, though this result is borderline significant.

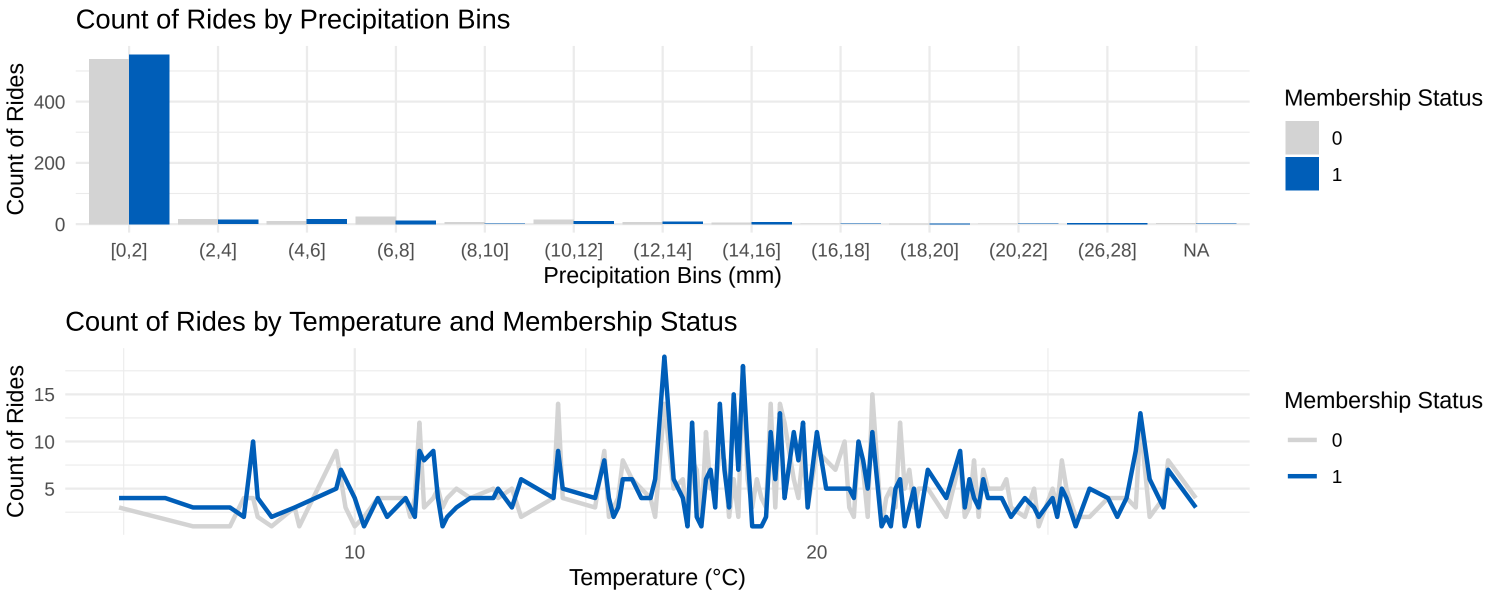
The low value of the correlations among the weather variables with the trip duration variable indicates that the weather factors (temperature and precipitation) do not linearly affect the duration of BIXI trips. Besides that, the marginal significance in the relationship between temperature and precipitation is not strong enough to prove the inclusion of these weather variables in the final model.

Moreover, based on Business Question 2, we further perform an ANOVA analysis on the model with and without weather variables. With the p-value equal to `0.348`, the ANOVA test shows that adding weather variables does not add much value to the model. Therefore, the inclusion of weather parameters in the model is redundant, as simplicity in the model would not be detrimental to the accuracy in the predictions.

To further investigate the association between the membership status and weather, we also analyzed whether BIXI members tend to ride under more extreme weather conditions and are more likely to ride regardless of the temperature or precipitation. To find the relationship between ride frequency with increasing levels of precipitation for members and non-members, we have binned our data based on precipitation. From the plot below, we can see that for increasing precipitation, members and non-members both have tendencies to reduce rides. The number of rides drops significantly at precipitation more than 10 mm. However, members seem a bit resilient, and during moderate precipitation of 5–10 mm, rides are higher compared to non-members.

Similarly, we looked at how ride frequency varies with temperature. Members and non-members both display a similar pattern, with a peak in bike usage occurring when temperatures are between 15–25°C, which represents the most favorable biking conditions. As temperatures drop below 10°C or exceed 25°C, the number of rides decreases for both groups. However, members again show slightly higher resilience, maintaining more frequent rides in both colder and hotter conditions compared to non-members.

```
grid.arrange(count_prec_plot, count_temp_mem_plot, ncol=1)
```



To conclude, while both members and non-members are affected by weather conditions, members show a slight tendency to ride under less favorable weather conditions, particularly in moderate rain and more extreme temperatures. This behavior could be attributed to members' commitment to using BIXI bikes as a primary mode of transportation, compared to non-members who might use the service more sporadically or leisurely, opting for other modes of transport under adverse weather conditions.