Report - Preliminary understanding of the bixi 2021 season

In a rapidly transforming urban landscape, sustainable and efficient modes of transport are not only desirable but essential. BIXI, Montreal's pioneering bike share initiative, has emerged as a key player in this transformative movement, offering residents and visitors an environmentally friendly, convenient and health-friendly alternative to traditional means of transportation.

We Chike Odenigbo, Charles Julien, Atul Sharma and Gabriel Jobert, a leading consulting team have undertaken a meticulous exploration of BIXI's operations for the 2021 season. Our goal is to highlight the opportunities hiding in BIXI's data. Such an examination is vital, not only for the stakeholders directly associated with BIXI, but also for urban planners, policy-makers, environmentalists and businesses who see the potential of this revolutionary mode of transport.

As you read through this report, you will discover a harmonious blend of quantitative rigour and qualitative analysis, designed to facilitate informed decision-making by all the stakeholders involved.

To extract more information from the station number, we enhance the dataset with the actual name and geographical coordinates of each station.

General Summary

After setting up the data, we can quickly visualize general trends by plotting histograms of all variables and calculating basic descriptive statistics on our dataset.

Data summary

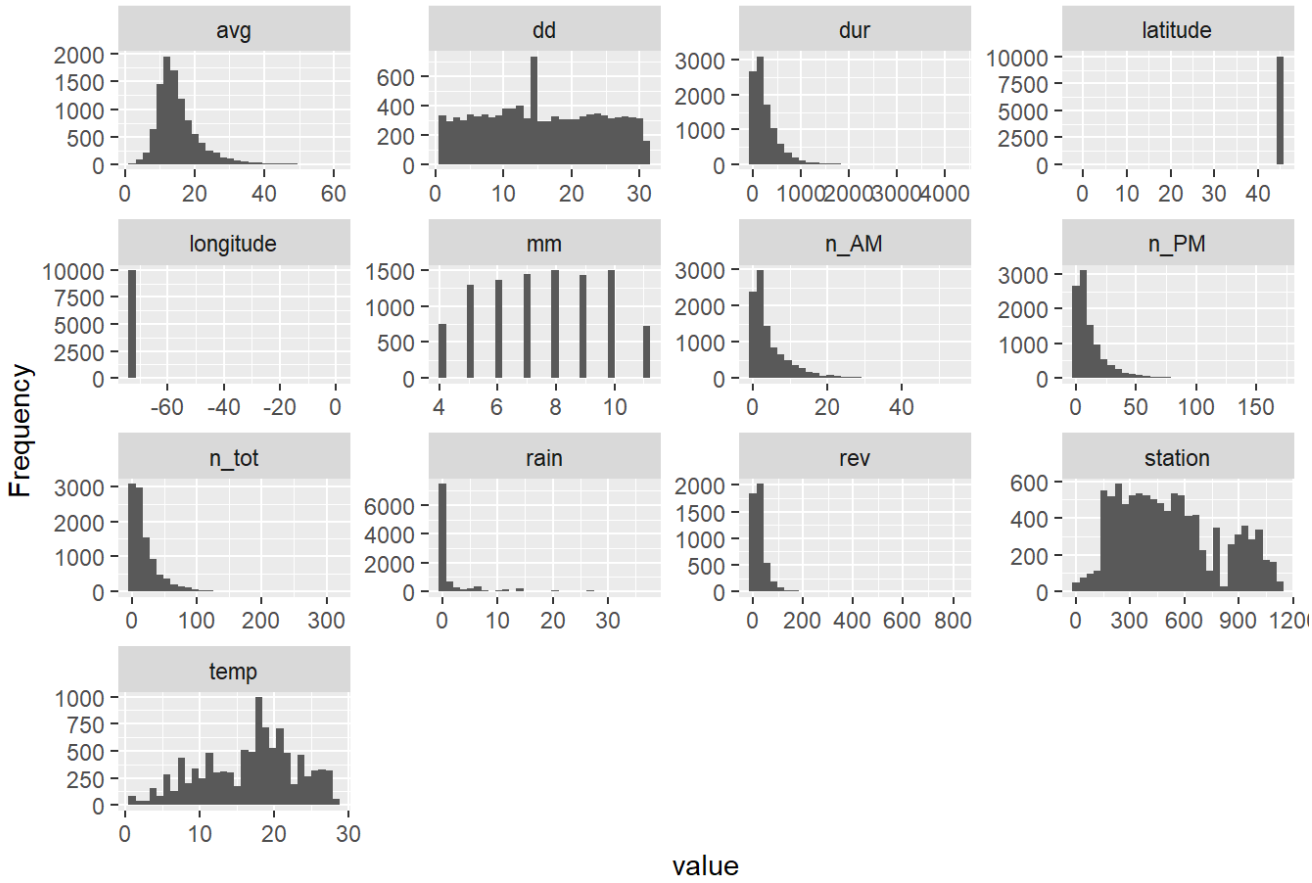
Name	df_main
Number of rows	10000
Number of columns	17
_____	
Column type frequency:	
character	2
numeric	15
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
wday	0	1	6	9	0	7	0
name	0	1	8	74	0	793	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
station	0	1.00	521.54	277.80	10.00	294.00	484.50	693.00	1138.00	
mm	0	1.00	7.56	2.06	4.00	6.00	8.00	9.00	11.00	
dd	0	1.00	15.64	8.72	1.00	8.00	15.00	23.00	31.00	
mem	0	1.00	0.53	0.50	0.00	0.00	1.00	1.00	1.00	
holiday	0	1.00	0.02	0.15	0.00	0.00	0.00	0.00	1.00	
dur	0	1.00	271.24	307.61	1.02	67.26	170.86	368.38	4253.30	
avg	0	1.00	15.30	6.59	1.02	11.22	13.82	17.72	59.65	
rev	5266	0.47	28.08	36.69	1.40	9.23	18.69	35.10	816.74	
n_AM	0	1.00	4.32	5.59	0.00	1.00	2.00	6.00	55.00	
n_PM	0	1.00	11.51	14.28	0.00	2.00	6.00	15.00	170.00	
n_tot	0	1.00	19.93	23.80	1.00	4.00	11.00	27.00	316.00	
temp	0	1.00	16.79	6.37	0.80	11.80	17.90	21.20	28.20	
rain	0	1.00	2.01	5.22	0.00	0.00	0.00	0.70	37.00	
latitude	0	1.00	45.44	1.92	-1.00	45.50	45.52	45.54	45.65	
longitude	0	1.00	-73.46	2.99	-73.75	-73.61	-73.58	-73.57	-1.00	



Data transformation/features engineering/imputation

Next, we enhance the data quality by performing various data transformations. We identify and impute missing values, sort ordinal values, derive seasons from months, and identify long weekends, among other enhancements.

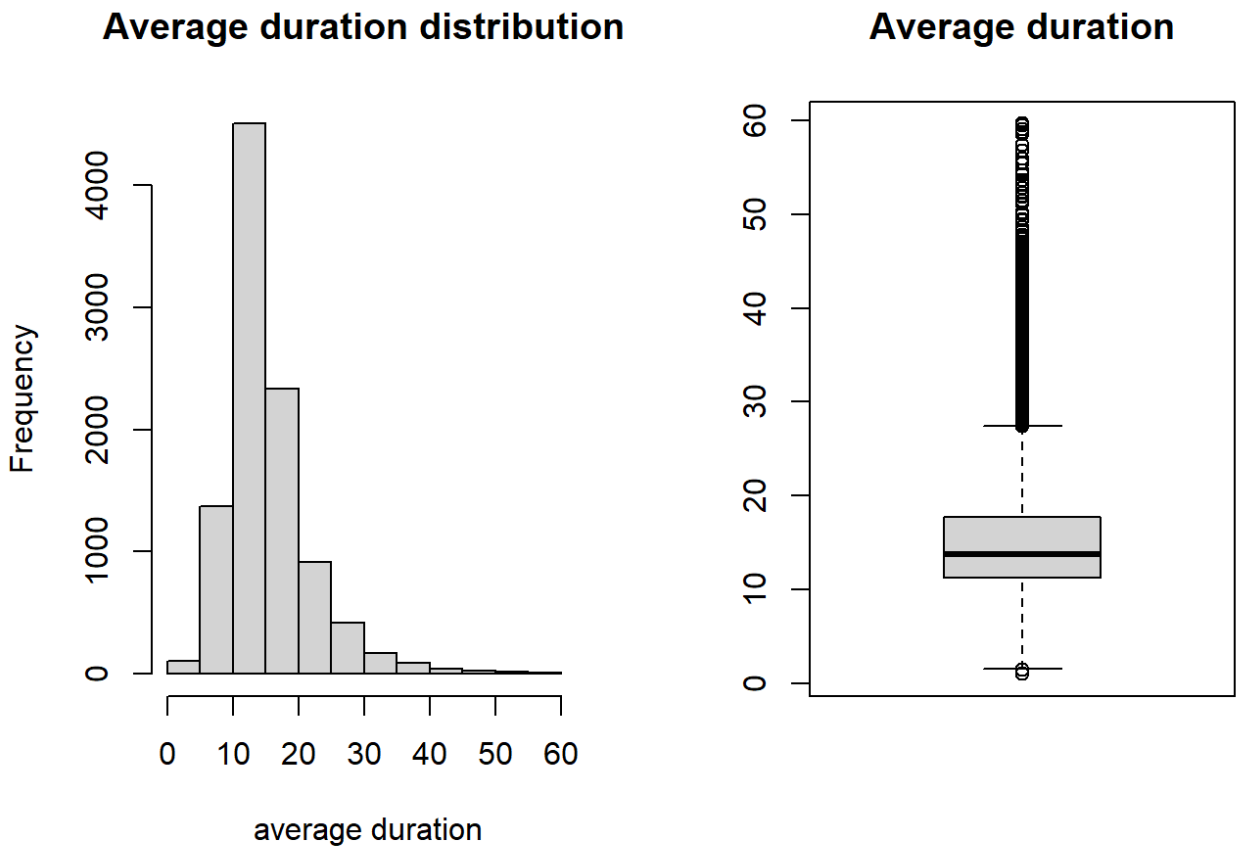
Univariate exploration

Let’s begin by exploring our dataset variable by variable.

Interest variables

Our study aims to better understand three main components of the Bixi system: trip duration, revenue generation, and everyday’s usage. Thus, we will start by exploring these variables individually.

Distribution of average duration

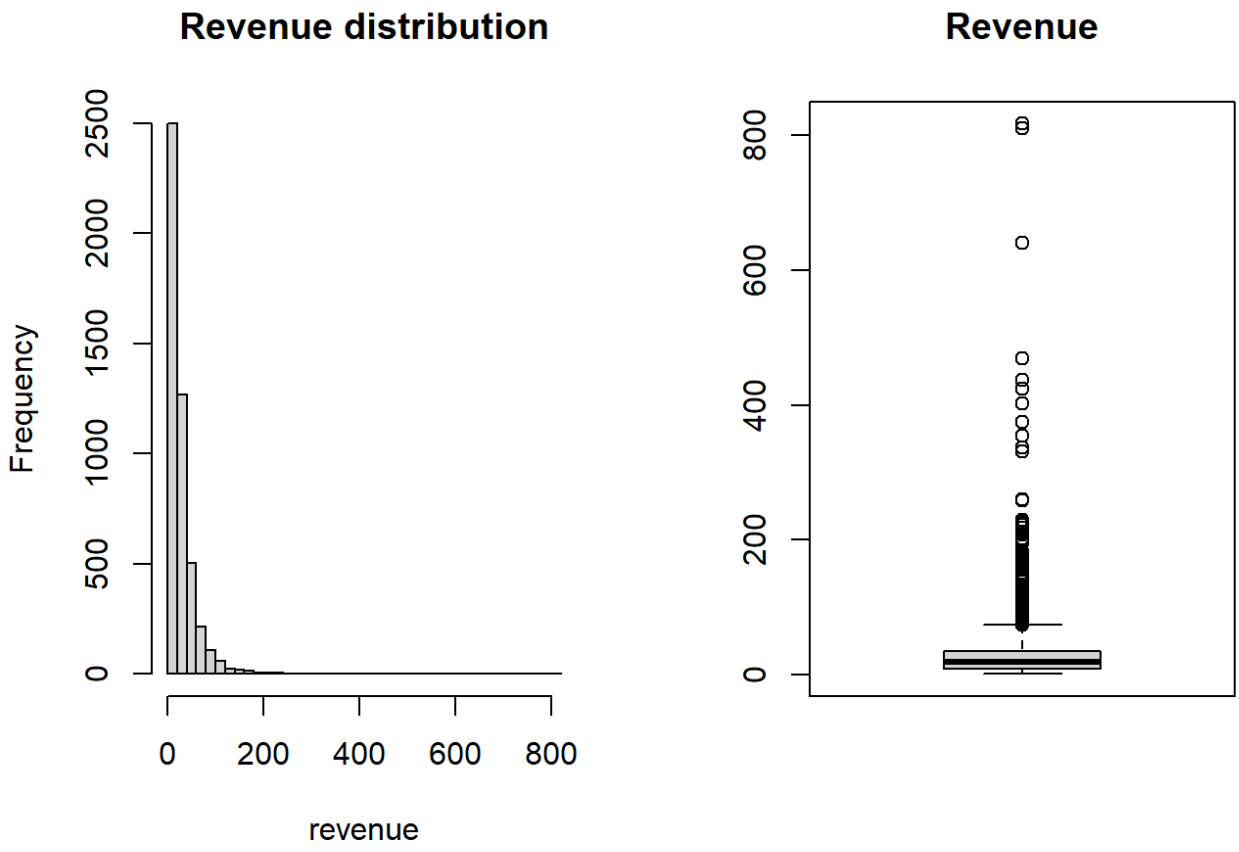


Observations:

The average trip duration is less skewed than the total duration. It's important to note that only short trips under 60 minutes are represented in the dataset. The mean of average trip duration is around **15 minutes** and median **14 minutes**.

It would be intriguing to investigate the relationship between trip duration and weather, as well as weekday versus weekend.

Distribution of revenue

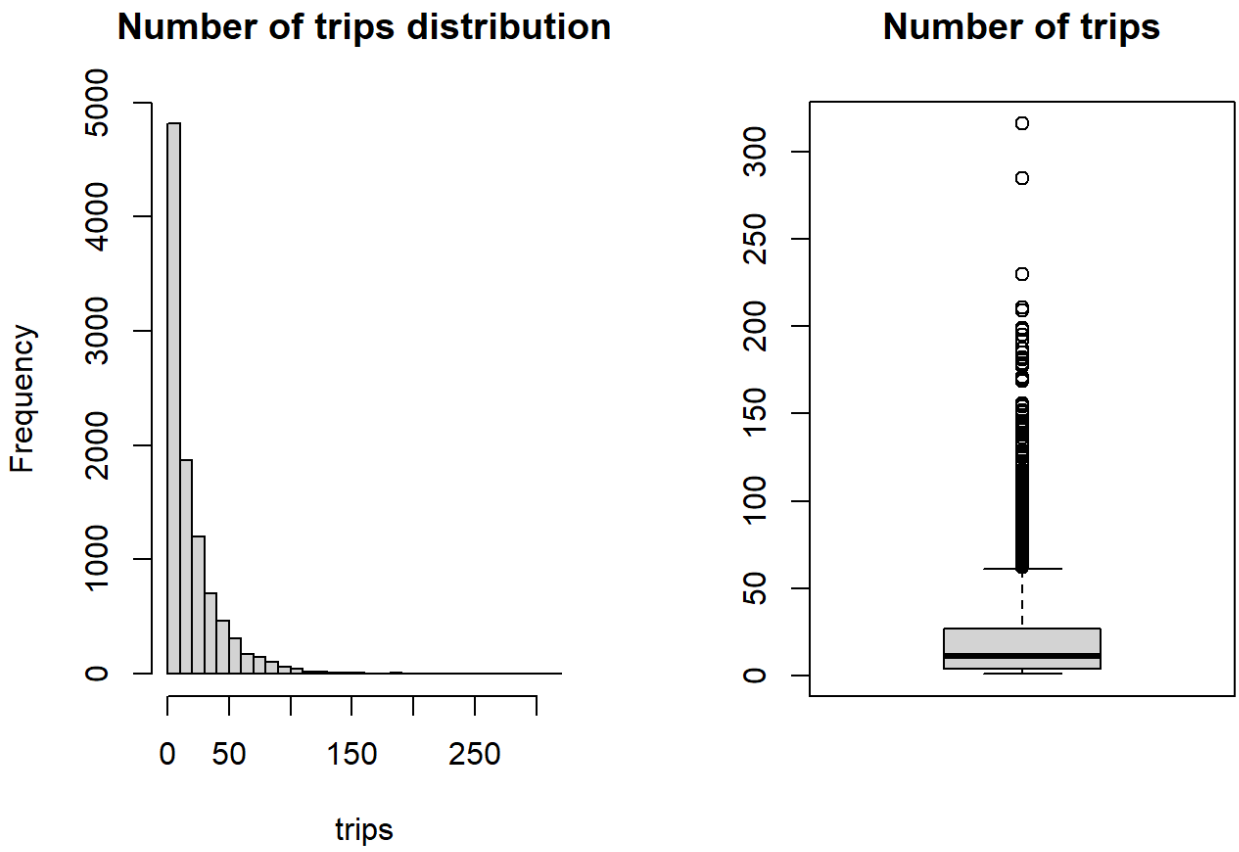


Observations:

The revenue distribution is also asymmetric and skewed to the right, as expected. It has a median of **18 dollars** and mean of **28 dollars** for a specific station on a specific day.

It would be intriguing to investigate if some stations or days generate more revenue.

Distribution of number of trips



Observations:

The distribution shape of the number of trips is highly similar to that of the duration. It has a mean of **20 trips per day** for a station and a median of **11 trips**. The maximum value is **316 trips**.

It would be intriguing to investigate if there are more trips at certain stations, if they occur more frequently in the morning or afternoon, and if members use the system more than non-members.

Most of our variables of interest would require a log transformation if used as target variables in a regression framework.

Support variables

Now that we have a good understanding of the distribution of our interest variables, let's ensure we have a basic understanding of our support variables.

Stations

How many station is there in the dataset and how often do they appear on average?

```
## [1] 793
```

```
## [1] 12.61034
```

Observations:

Given that there are over 700 stations in the dataset, it would be more effective to group them into subgroups based on their respective regions.

Each station occur on average 12 times in our dataset.

Membership

Is the number of rows for member and non-member roughly equal?

```
##
##    0    1
## 4734 5266
```

Observations:

Our hypothesis was that there would be one row for members and one row for non-members per station per day, resulting in an equal representation of members and non-members. Our analysis found that this hypothesis was roughly correct.

Time of day

Are there more trips in the morning or afternoon?

```
## [1] 43215
```

```
## [1] 115140
```

Observations:

In our dataset, the number of trips in the afternoon is nearly **three times** greater than in the morning.

Temperature

We know that temperature ranges between 0 and 28 degree Celsius with mean and median around 17. This makes sense since the month interval is between April and November inclusively.

How often does it rain and when it does, what is the mean precipitation?

```
## [1] 0.3828
```

```
## [1] 5.260293
```

Observations:

In our dataset, **38%** of the observations indicate the presence of rain, and when it does rain, the average precipitation is approximately **5mm**.

Bivariate exploration

Now that we have a good idea of what each variable is like, let's explore how the interest variables are affected by each other and by our support variables. Our goal is to understand the factors that may affect our variables of interest.

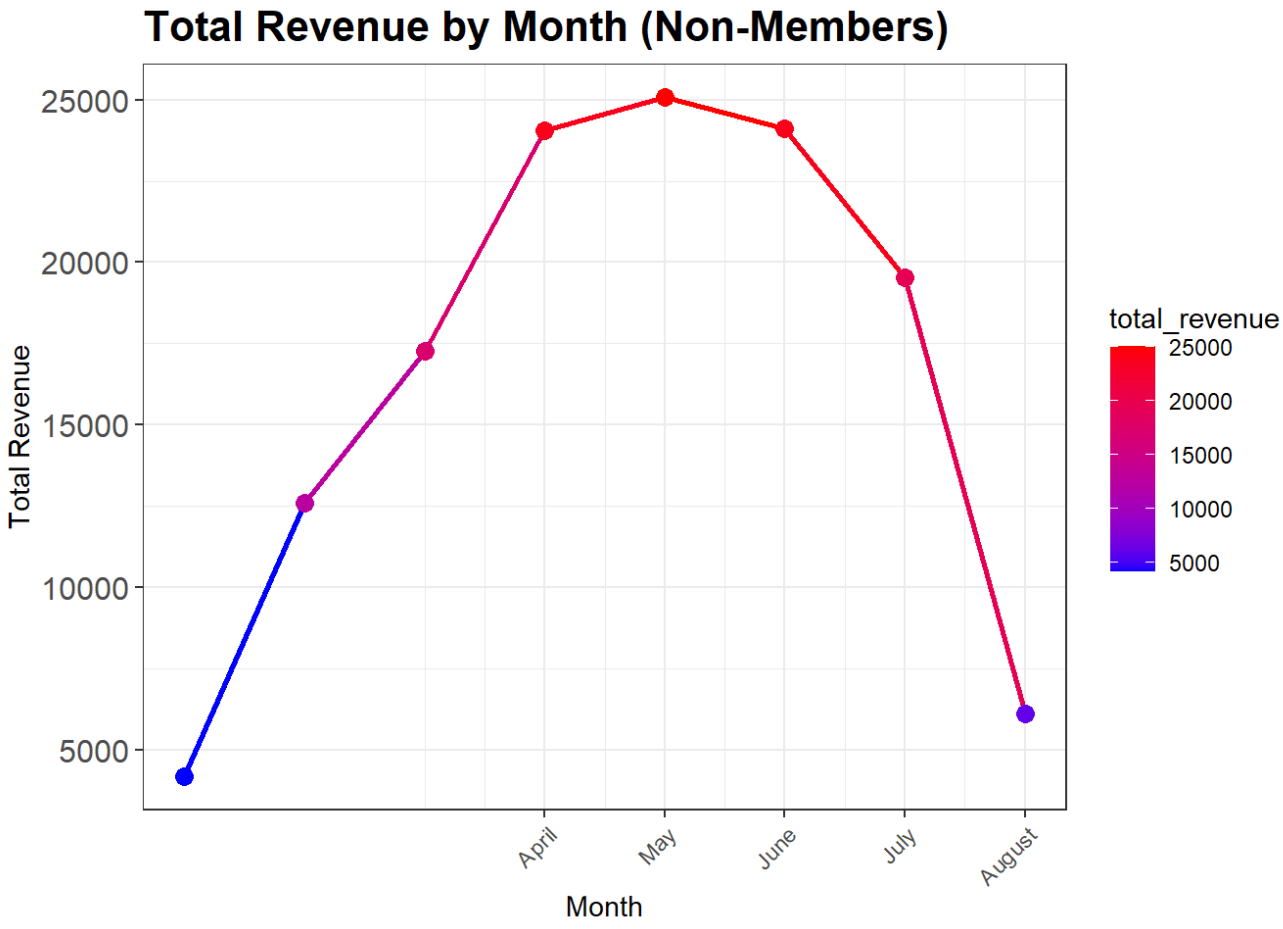
Revenue exploration

Let's begin our exploration with revenue.

Revenue in function of time

The relationship between revenue and the time of the day, week, or year is not so intuitive. One could wonder if Bixi generates more revenue when people are on vacation and use Bixi for leisure or when they go to work and use it for transportation. It's important to keep in mind that in our dataset, revenue is only generated by non-members.

Revenue in function of months

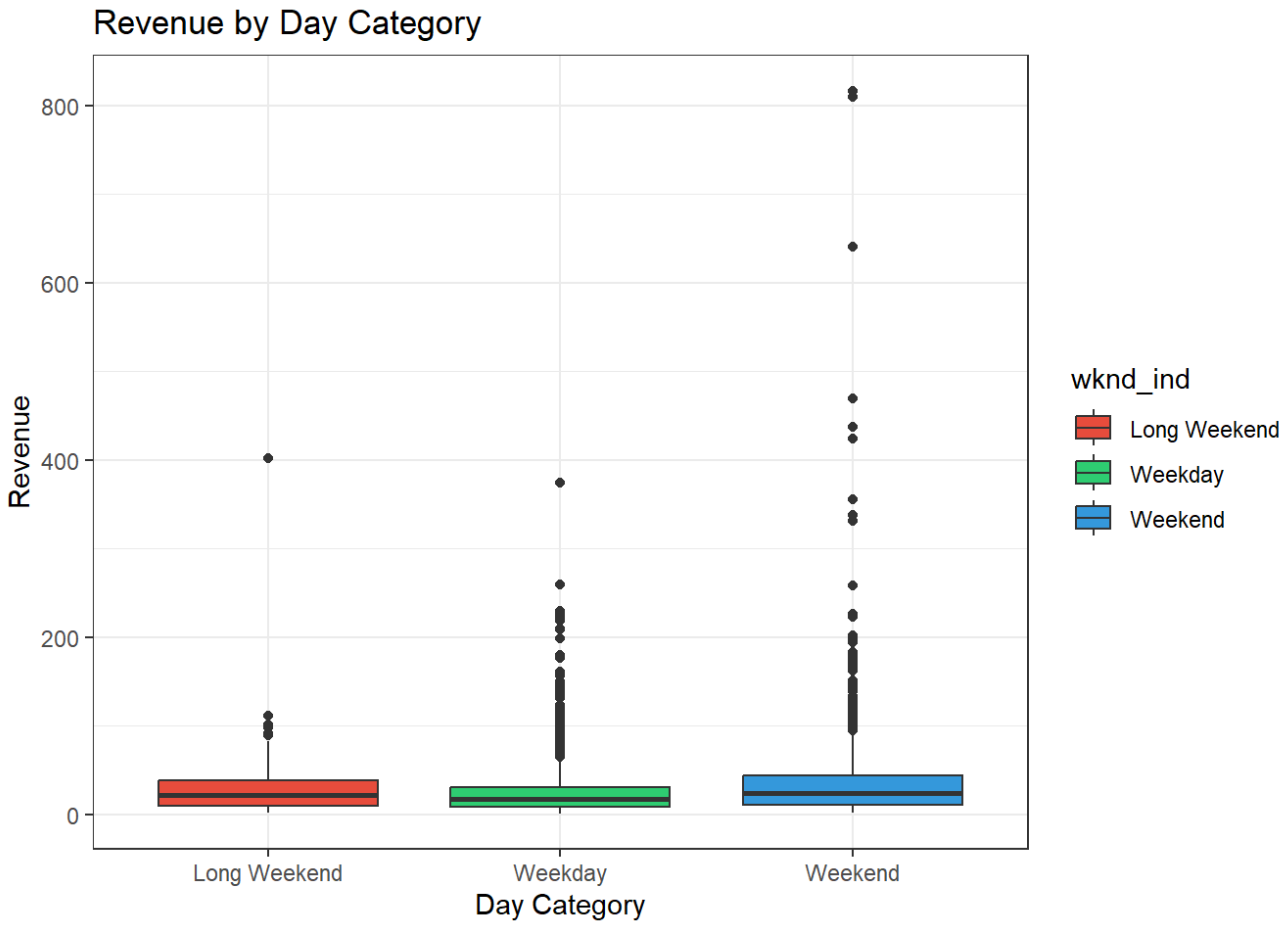


Observations:

There is here strong evidence of seasonal trends. Peaks in revenue coincide with warmer months (June to October).

Revenue in function of holiday

```
## Warning: Removed 5266 rows containing non-finite values (`stat_boxplot()`).
```



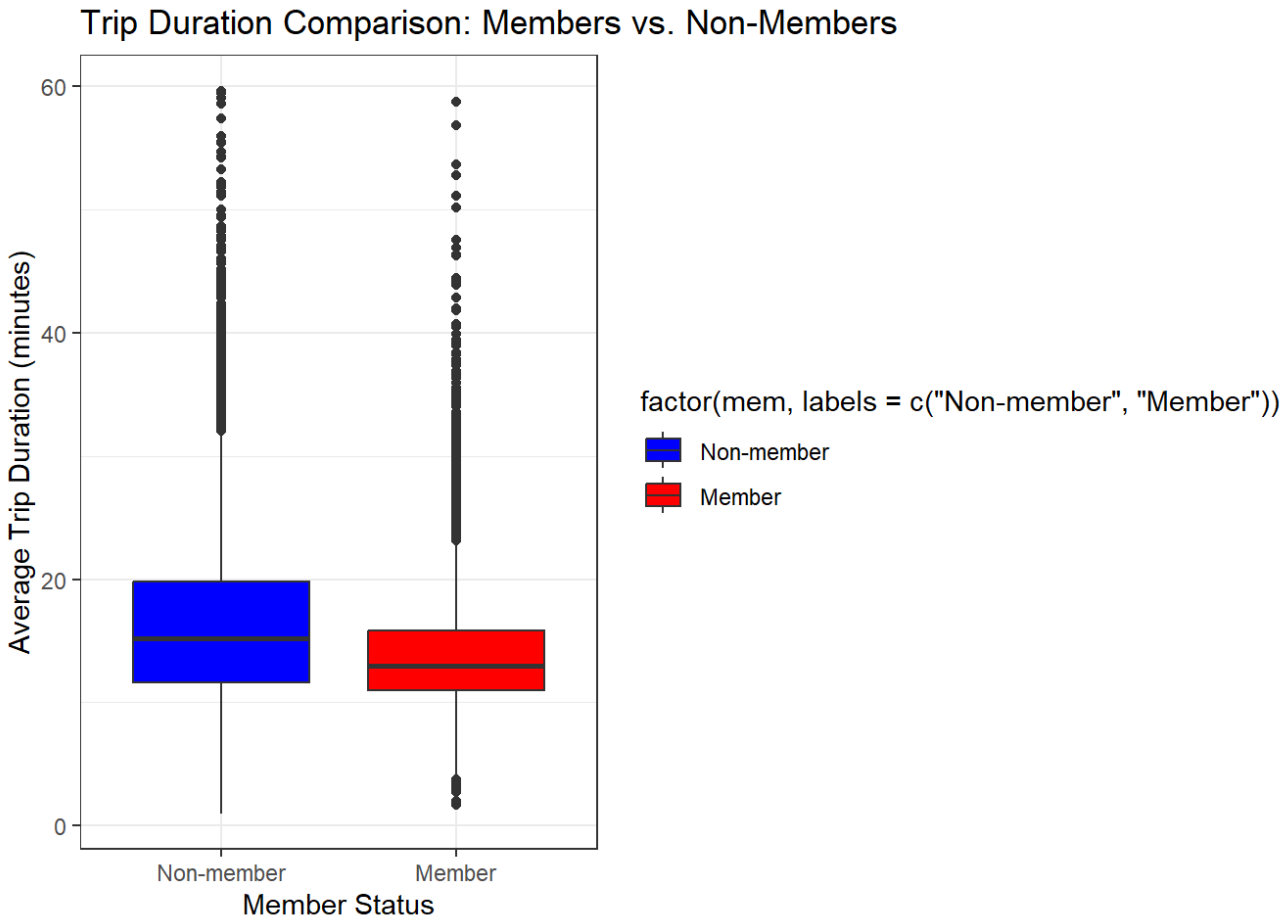
Observations:

We can observe here that the median revenue is noticeably higher during week-ends than during weekdays. This is not surprising because people generally go outside to do some activities during weekend, given the fact that they are working during weekdays. The interesting point here is that, even long week-ends are more profitable than weekdays, they are less profitable than regular week-ends. This need further exploration but it could be explained by the fact that people are generally going on holidays-trips during long week ends and consequently, are not Montreal.

Duration exploration

Let’s see how duration is affected by different factors.

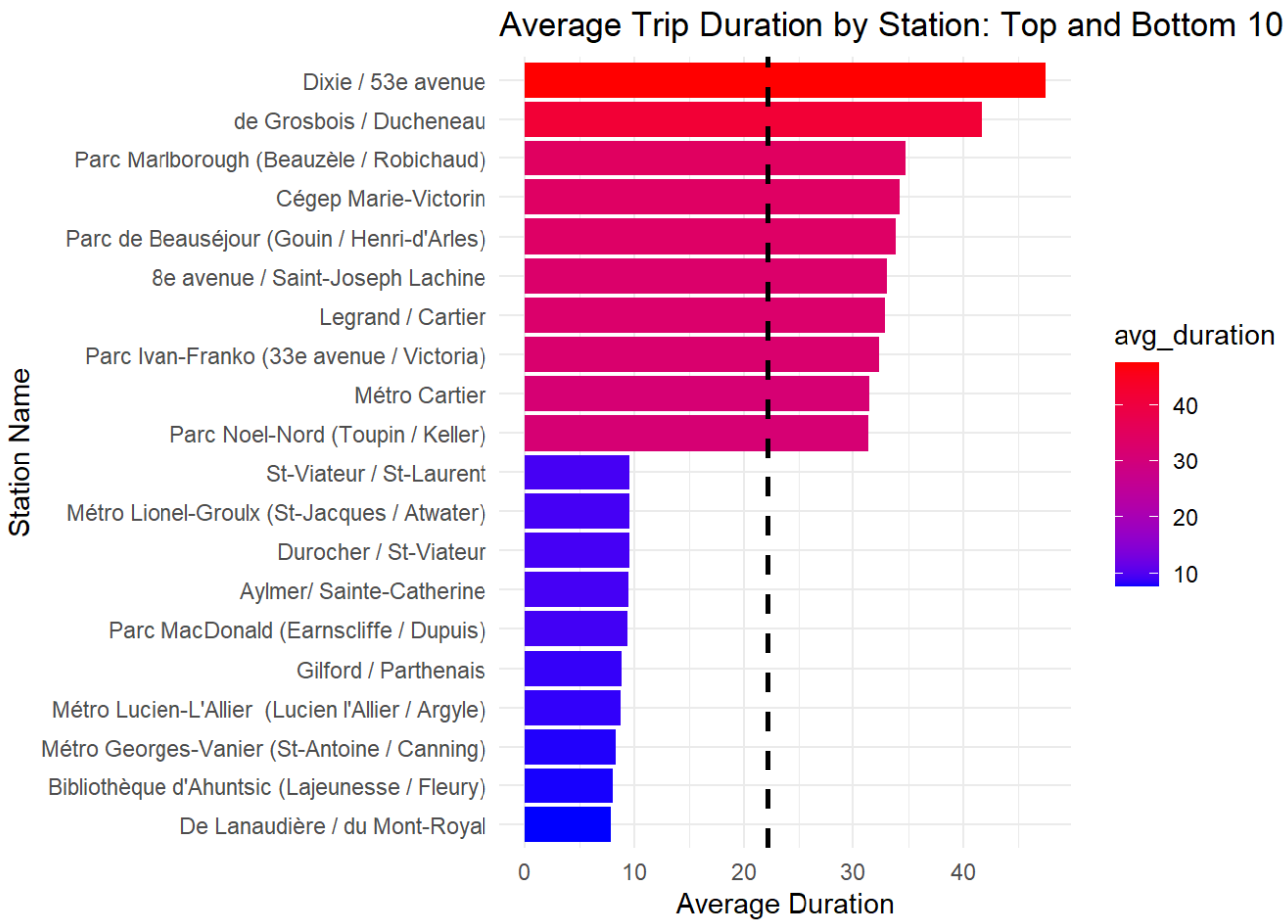
Duration in function of membership



Observations:

This graph reveals disparities in trip duration between members and non-members. The data indicate that non-members generally engage in longer trips compared to our membership base. Moreover, the frequency of outliers (exceptionally long trips) is notably higher among non-members.

Duration in function of station



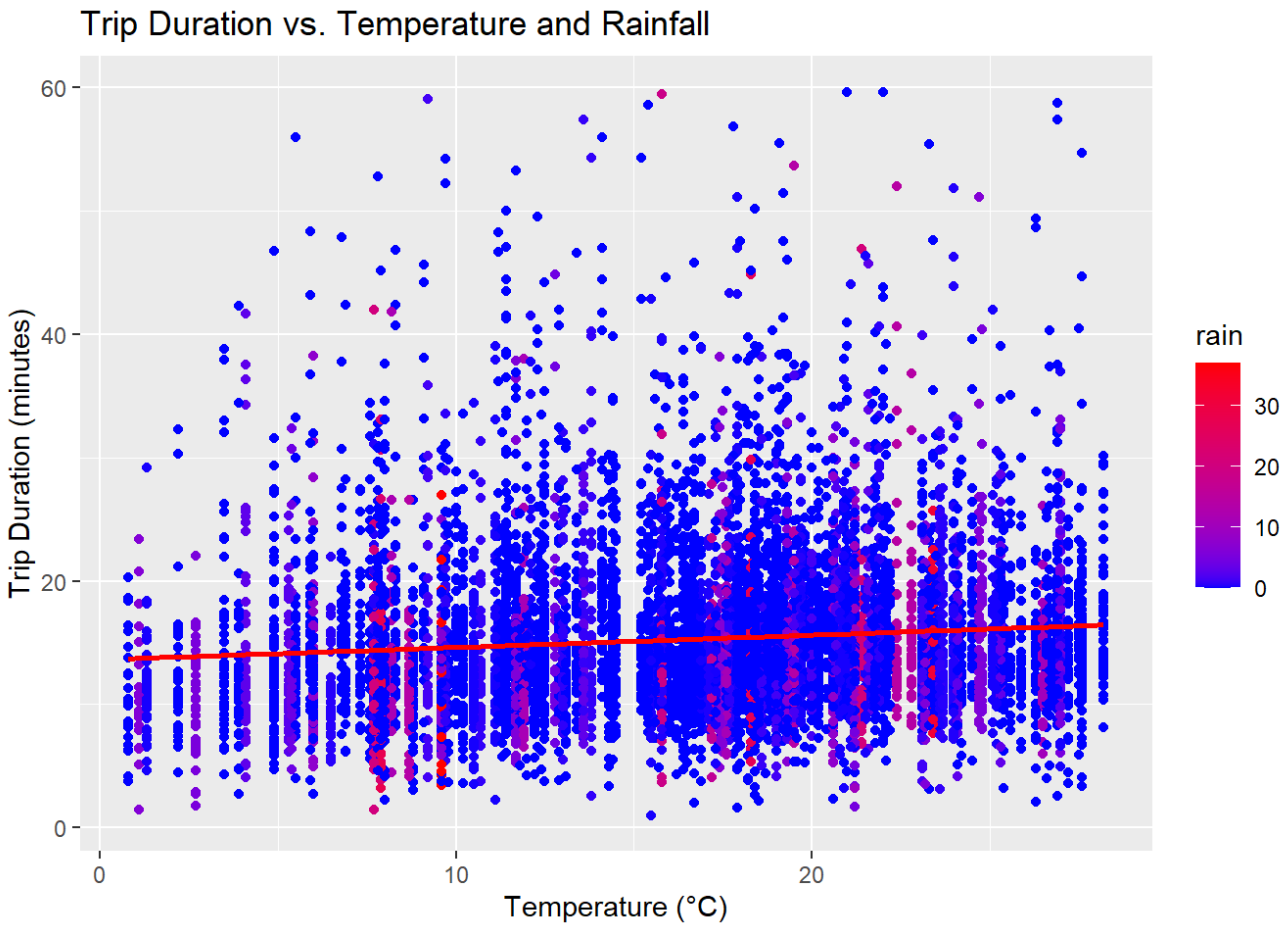
Observations:

Given the fact that there are more than 600 stations, we did only observed extreme values and the mean of all stations. The graph shows that the average duration is around 22 min, with the duration of the lowest stations being slightly under 10 min and the duration of the highest stations being around 30 min. This need deeper exploration to see if the lowest stations are in highly concentrated district and if the highest stations are in more distant places.

Duration in function of weather

Duration in function of temperature/Rain

```
## `geom_smooth()` using formula = 'y ~ x'
```

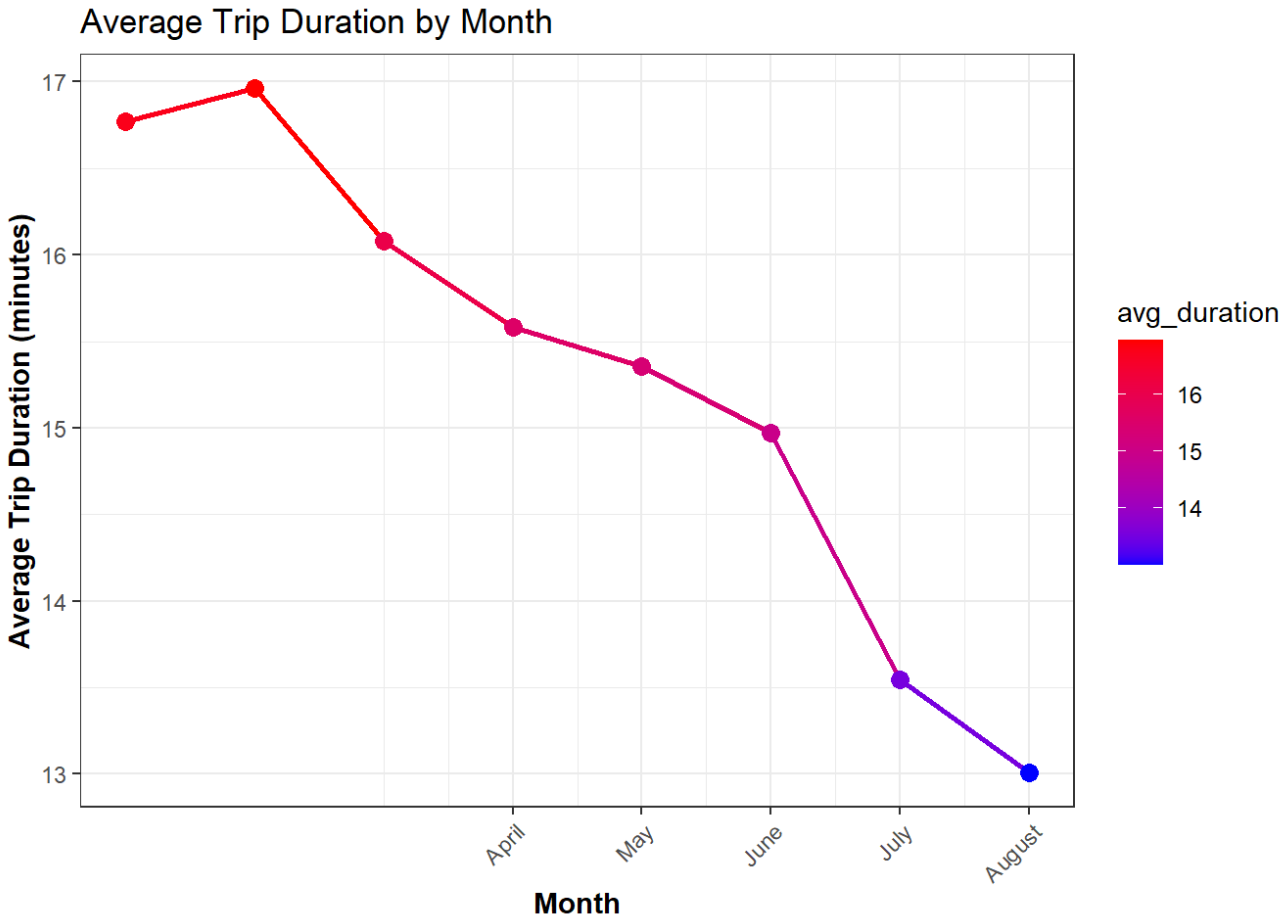


Observations:

Our analysis reveals a positive correlation between trip duration and temperature. While the relationship appears to be somewhat weak, it is nonetheless evident. Additionally, we observed that rainy days tend to have less extreme trip durations compared to sunny days—an outcome that aligns with general expectations.

Duration in function of time

Duration in function of months



Observations:

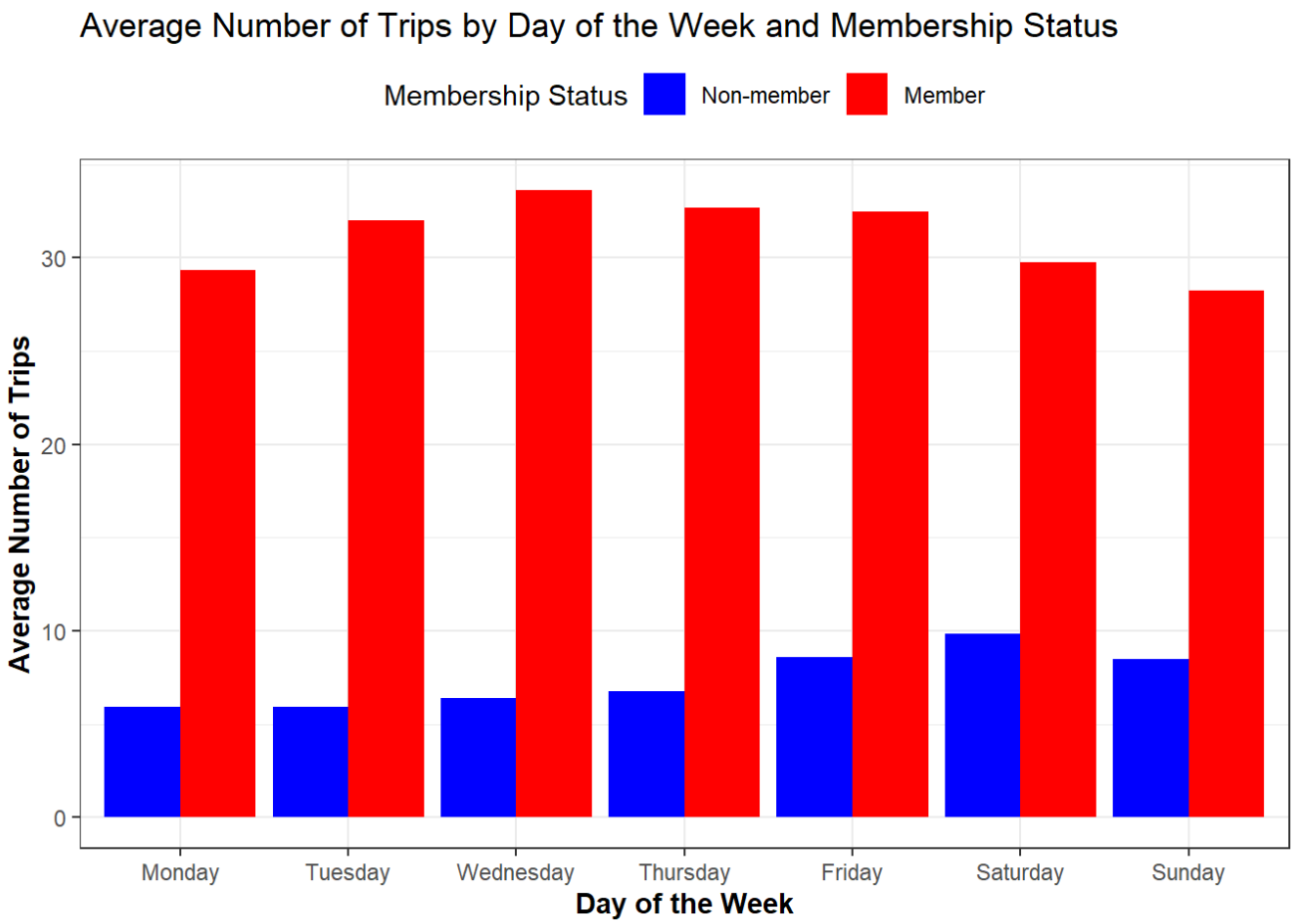
The analysis demonstrates a noticeable yet subtle declining trend in average trip duration from April through November. While the trend is consistent, it's crucial to note that the variations are confined within a range of approximately 13 to 17 minutes—a span that may not be significant from an operational or customer experience standpoint.

Exploration of the number of trips

Number of trips in function of time

Average number of trips in function of day of the week and

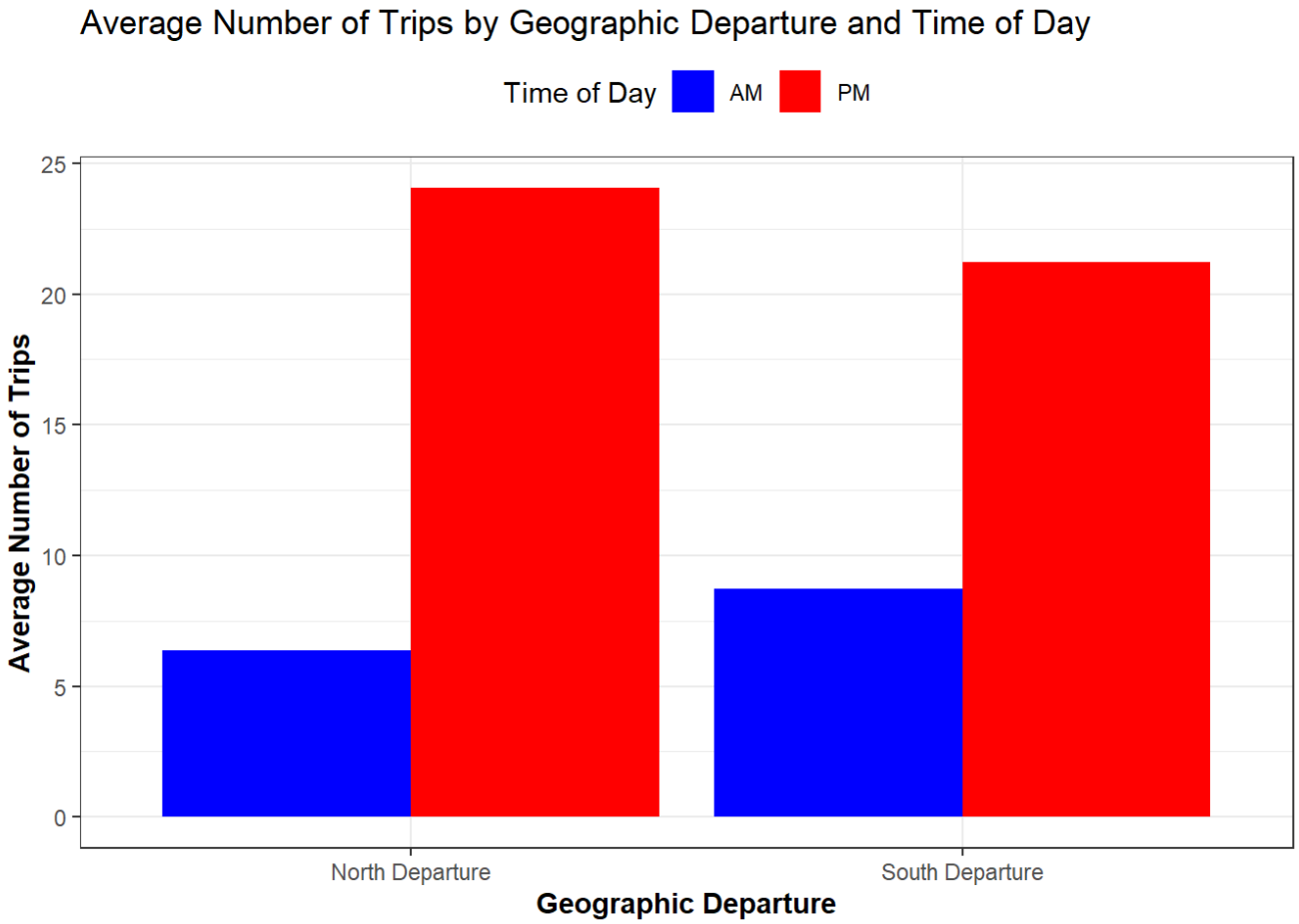
membership



Observations:

This is a fascinating observation. The graph indicates that members and non-members have different usage patterns. Firstly, members account for a significantly larger proportion of Bixi usage than non-members. Secondly, members tend to use Bixi more frequently during weekdays for commuting, while non-members use it more during weekends for leisure activities. It is possible that non-members use Bixi for weekend outings or leisure, while members use it for frequent commuting.

Number of trips in function of time of the day and location



Observations:

Our analysis reveals fascinating insights. On average, Bixi bikes are utilized significantly more in the afternoon compared to the morning. In the morning, there appears to be a higher number of departures from the south side of the city, while in the afternoon, the trend is reversed. This could suggest that Bixi user are more located on the south side of the city.

Conclusion:

Here is a summary of the findings:

- **Weather**: The analysis suggests that weather does not have a significant impact on Bixi usage.
- **Duration**: The duration of Bixi trips varies depending on the station.
- **Revenue**: Bixi revenue fluctuates over the course of the month.
- **Membership**: Most Bixi usage is attributed to members.
- **Usage Patterns**: Members tend to use Bixi more during weekdays, while non-members use it more during weekends.
- **Flow**: There appears to be a higher flow of Bixi trips from the south of the city to the north in the morning, and the opposite pattern in the afternoon.

These insights provide valuable information about Bixi usage patterns and can help inform decision-making processes.

Contribution

Charles Julien : Did univariate analysis as well as number of trip analysis and text structure.

Gabriel : Did part of bivariate exploration and page layout.

Chike Odenigbo: Did summary statistics, missing variable analysis, feature engineering and correlation plots (excluded due to page limit)

Atul Sharma: Did part of Bivariate analysis.