Analysis of the effect of various parameters on UBER Ridership

Nafiur Rahman

Std ID: 34276352

Nafiur.rahman@alumni.ubc.ca

Data 501– Introduction to Data Analytics

April 15, 2020

Abstract

In this project, a dataset on Uber pickups in New York City, dated from January 2015 to June 2015, was analyzed applying exploratory data analysis techniques to explore the factors that affect Uber car's demand in NYC. R and Python were used as primary analysis tools for the whole process, starting from simple univariate plots and analysis to multivariate analysis, regression techniques, and statistical hypothesis tests.

Introduction

Data storytelling is an essential component of data science, through which companies can understand the background of various operations. Uber is the most popular ride-hailing service among general people. At harsh weather or during holidays, the availability of uber varies greatly. With the help of visualization, Uber can avail the benefit of understanding the complex data and gain insights that would help them to craft decisions accordingly.

Methods and Discussion

Data Cleaning

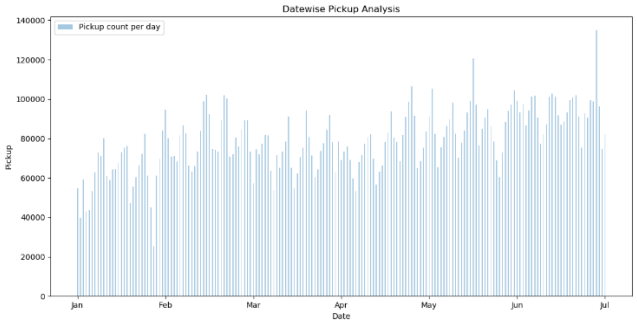
From our dataset, we observe that the Uber pickups of all the boroughs of NYC were segregated. There are some data labeled as 'NA'; as Python reads these values as **NaN**, these were replaced by "NA" after reading the data frame by **Pandas** [1].

In R, no such cleaning is required.

Univariate Plots Section

Pickups

The analysis was started by simply plotting the pickup count of each day in New York City [2]. Our dataset has dates in the format *"2015-01-01 1:00:00 AM"*. For this reason, we created a pivot table with the sum of the pickups of each hour and got each day pickup count.



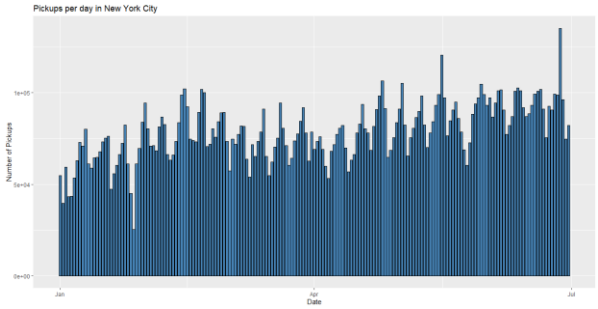
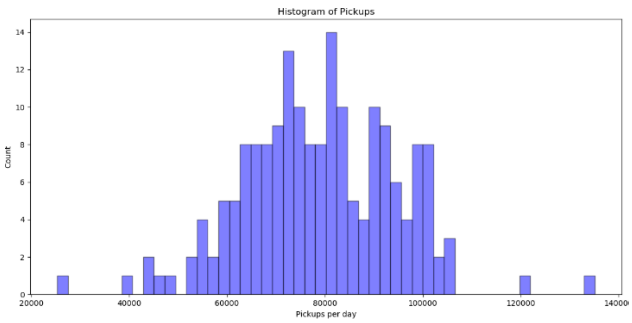


Figure 1(a): Python

Figure 1(b): R

We can notice a gradual rise of pickup counts as time progresses.

This plot gives us information that is difficult to interpret. Let us plot the histogram of these pickups now [3].



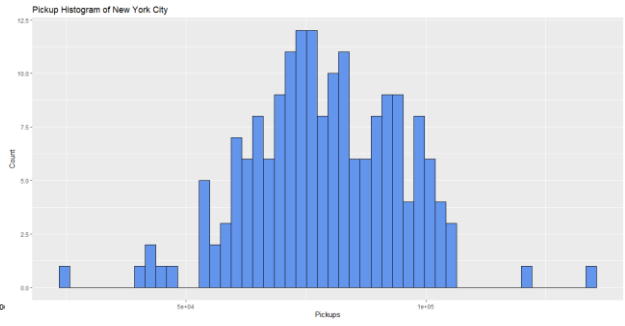
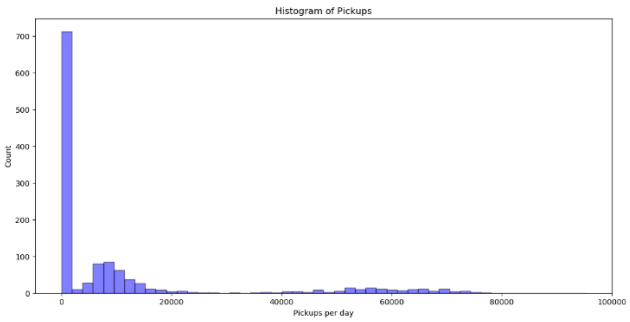


Figure 2(a): Python

Figure 2(b): R

We can see a normal distribution of the pickups in NYC.

Now, to have an insight into the pickups at different boroughs, we add the column boroughs to our pivot table [4].



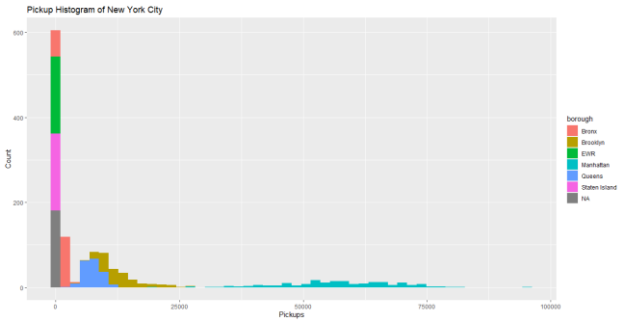
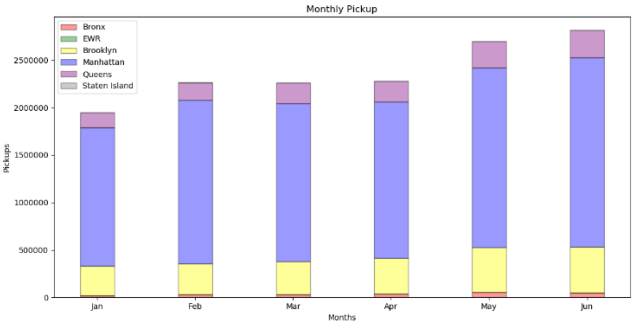


Figure 3(a): Python *Figure 3(b): R*

The distribution is a strange one. It looks like a union of normal distributions. We can suspect that the different boroughs have very discrete distributions. Moreover, we can see that the majority of 0 pickups are created solely by EWR, Staten Island, and from pickup data that we are missing the borough.

To see which borough dominates most, we plot a bar graph for each month [5].



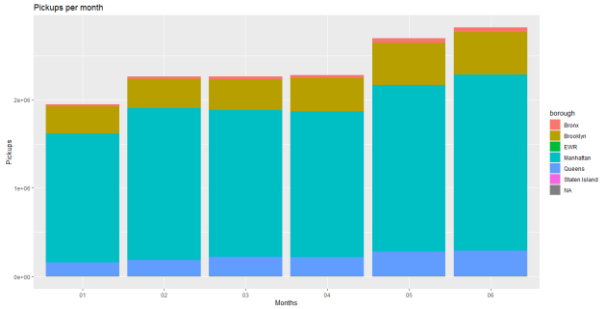


Figure 4(a): Python *Figure 4(b): R*

So, Manhattan is the borough with most pickups in New York City.

Now let us investigate the histogram of pickups of different boroughs [6].

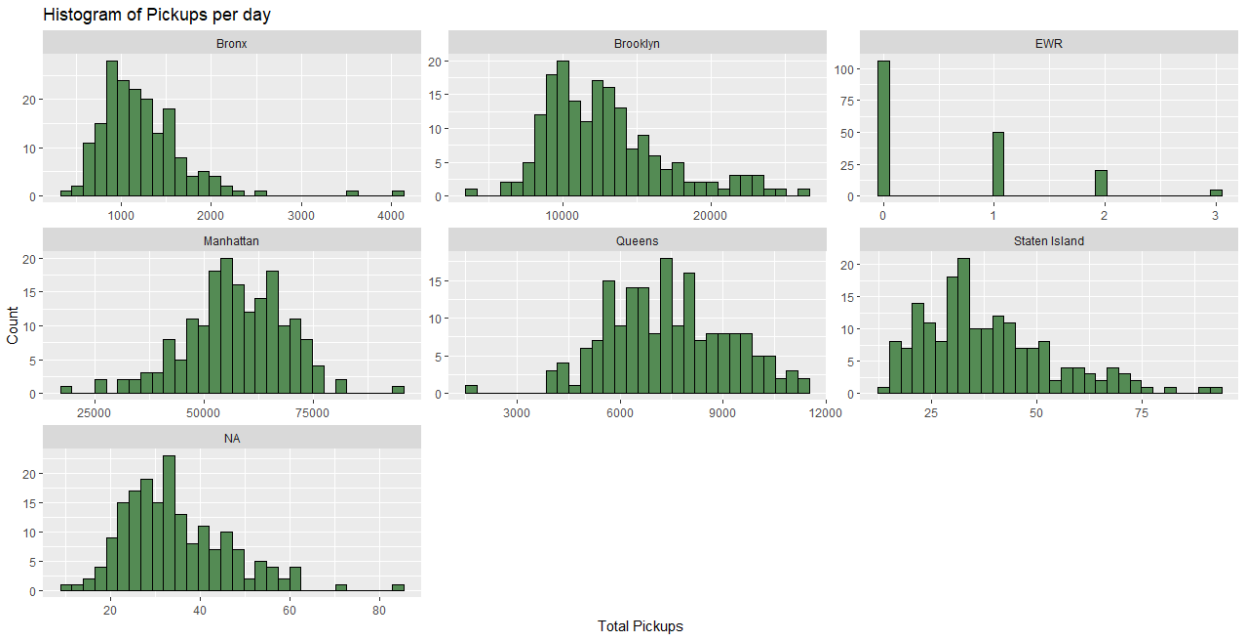


Figure 5

Some observations from these graphs:

- There is a clear difference in ridership between the different boroughs. Manhattan has by far the highest demand, followed by Brooklyn, Queens Bronx.
- EWR and Staten Island have very few pickups.
- All four significant boroughs' pickups follow normal distributions.

Weather

The univariate plots of several weather parameters are provided below [7].

Wind Speed

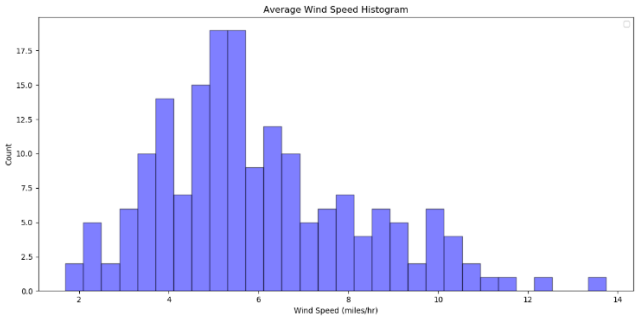


Figure 6(a): Python

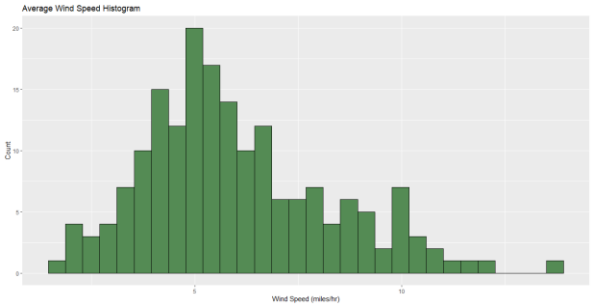


Figure 6(b): R

The histogram is positively skewed with a Mode of 7 miles/hour, which means that most of the time, there was a light breeze. The speed tops at 14 miles/hour, which is not even a strong breeze.

Visibility

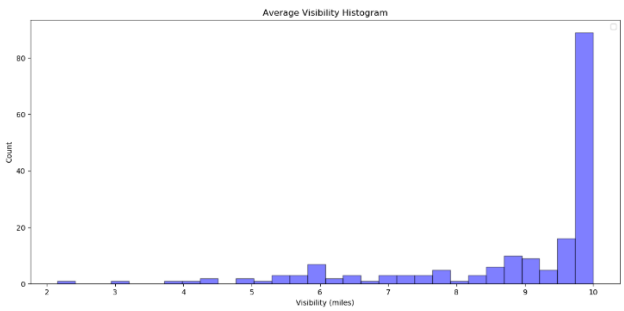


Figure 7(a): Python

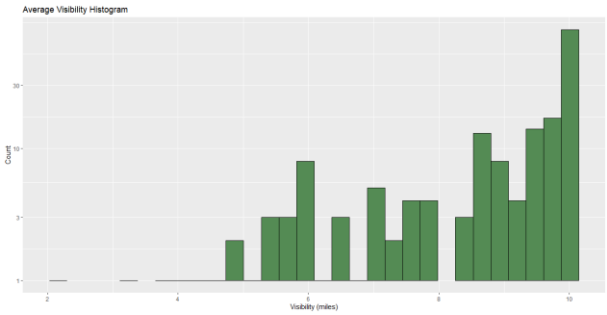
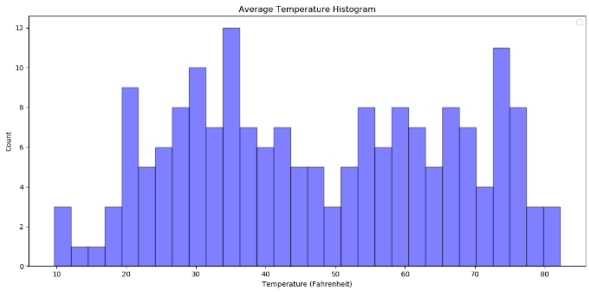


Figure 7(b): R

In R (green graph), the Y scale is scaled logarithmically for a better understanding of the plot. Analyzing the plots, it is certain that most of the time, there was a clear sky.

Temperature



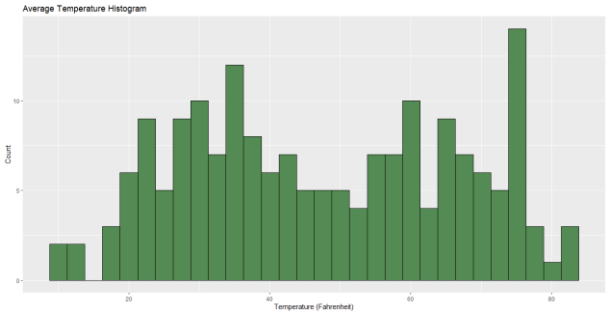
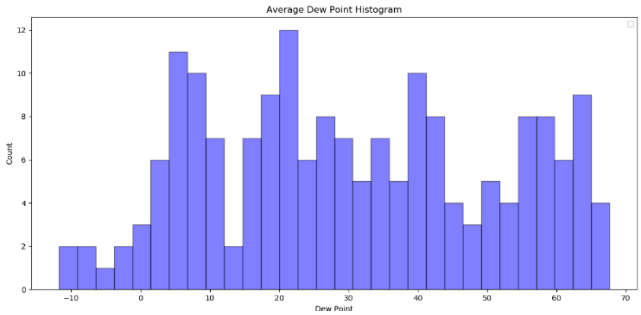


Figure 8(a): Python *Figure 8(b): R*

The distribution of the temperature is bi-modal with one peak around 35 degrees and the other near 75.

Dew Point



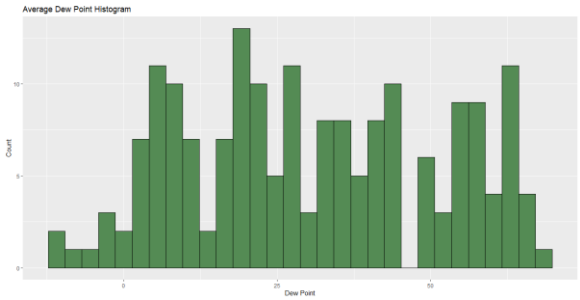
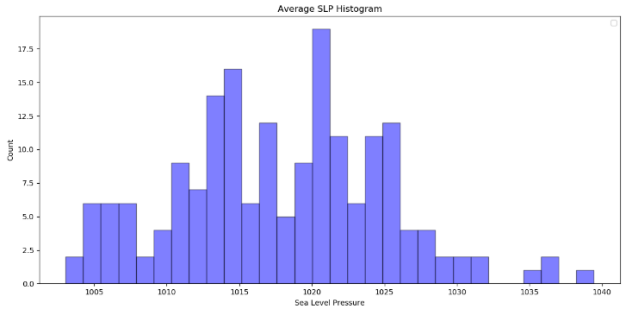


Figure 9(a): Python *Figure 9(b): R*

Dew point is the temperature at which airborne water vapor will condense to form liquid dew. A higher dew point means there will be more moisture in the air. Since the dew point is correlated with temperature (by definition), their distributions appear similar.

Sea Level Pressure



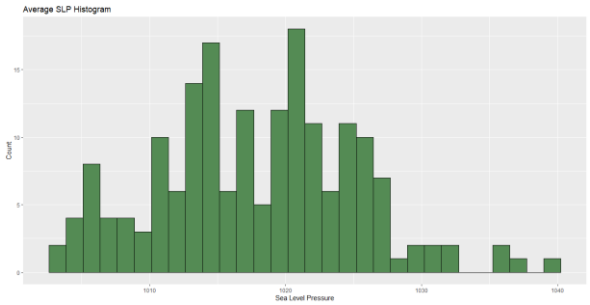
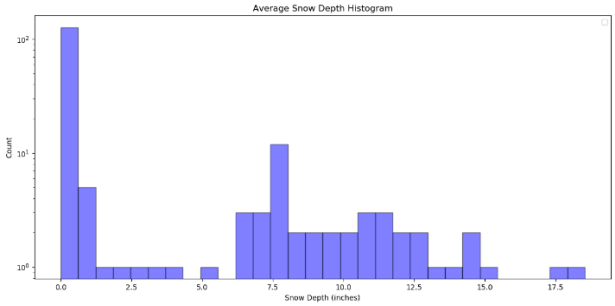


Figure 10(a): Python *Figure 10(b): R*

Air pressure affects the weather by influencing the movement of air around the planet; areas of low pressure generally develop clouds and precipitation, while areas of high pressure tend to bring clear, sunny weather conditions. Air pressure affects the weather at a later time; thus, there might be a delayed effect in the ridership. The plot shows a normal distribution.

Snow Depth



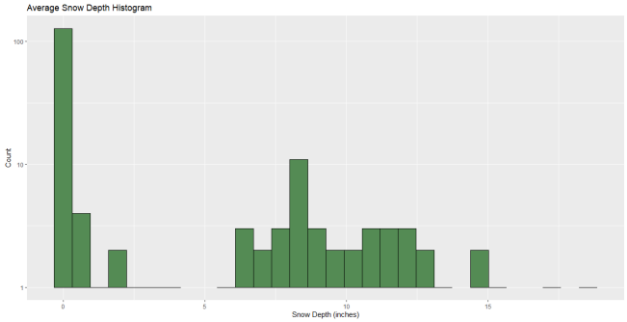


Figure 11(a): Python *Figure 11(b): R*

As the Y-axis is scaled logarithmically, we can decide that most of the time, there is no snowfall at all.

Univariate Analysis

Most of the variables have normal distributions with or without scaling the X-axis. There is a couple of bi-modal distribution denoting a probable rapid change in their value on a time scale.

Finally, there are some variables representing weather variables with default/expected values at the edge of the scale (like visibility = 10), creating geometric distributions.

Bivariate Plot Section

Pickups Vs. Date

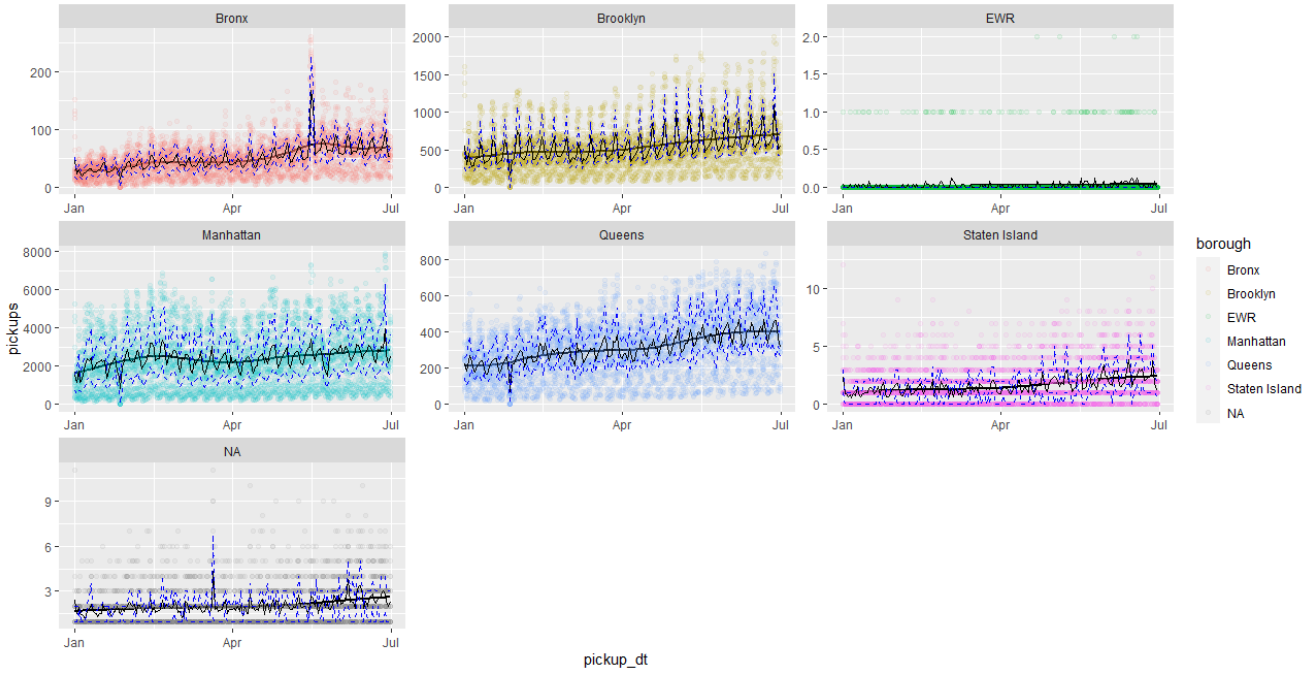


Figure 12

The figure [8] above illustrates the number of pickups at different boroughs from January 2015 to June 2015. As Manhattan is the borough with the most pickups, we will further analyze for Manhattan only.

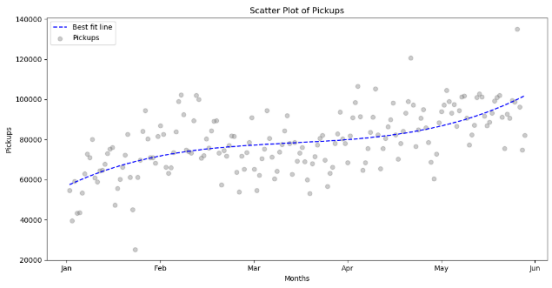


Figure 13(a): Python

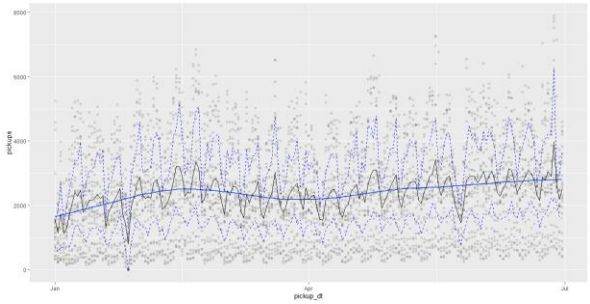
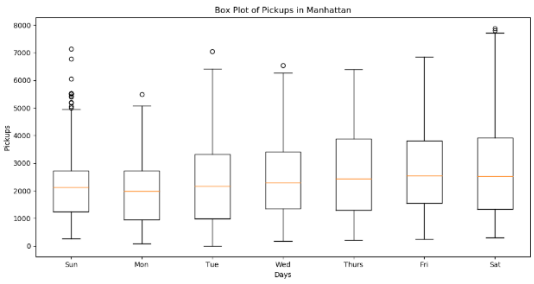


Figure 13(b): R

The graph [9] also shows the mean, 25th, and 75th quantile. Plotting the pickups VS datetime, we can see that there is a clear pattern. There are 26 peaks, as many as the number of weeks in the investigated period. Also, there is a general rising in the number of pickups over time, which is aligned with the findings of the pair plots.

Distribution of Pickups per day

As Manhattan is the busiest borough of all, we plot the pickups per week [10].



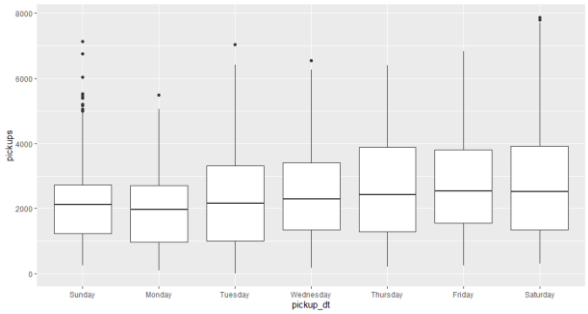
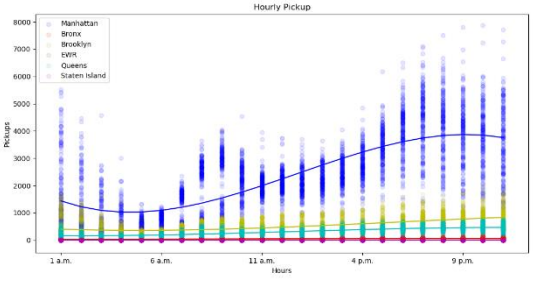


Figure 14(a): Python *Figure 14(b): R*

We notice that there is a pattern during the week. The demand starts low on Monday and then rises until Saturday when it peaks. On Sunday, the demand falls, and then we go back to Monday.

Pickups Vs. Time of the day

Below is the hourly pickup plot of the NYC [11]



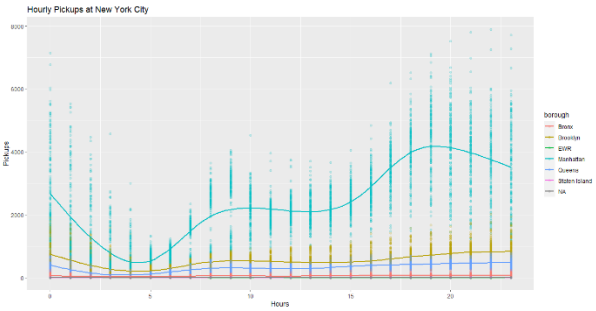


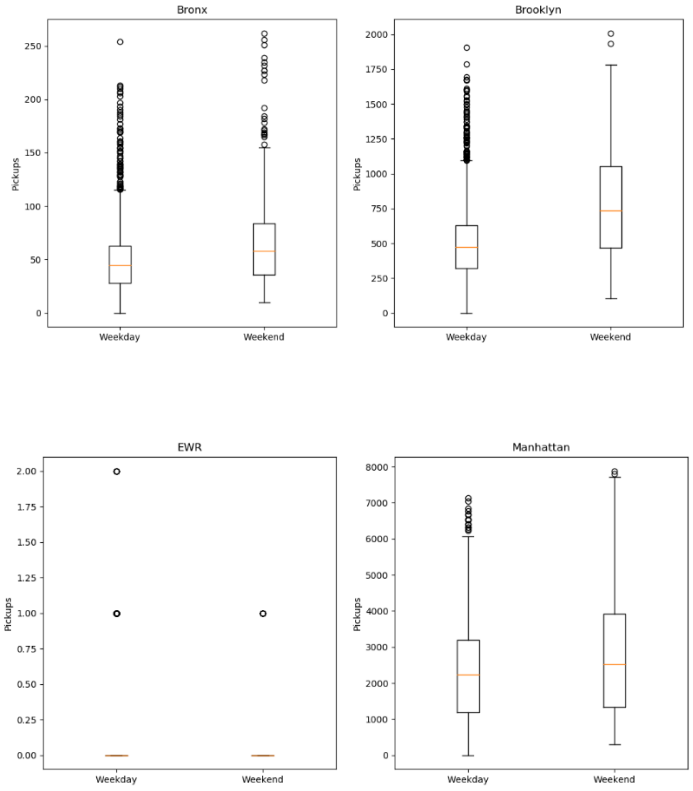
Figure 15(a): Python *Figure 15(b): R*

For Manhattan, there is a clear pattern of the ridership on a daily level. The traffic starts low at 5 a.m., starts rising until 9-10 a.m. when it hits a plateau. At around 2 p.m., it starts growing again until 8 p.m. when it hits the daily maximum. Even without the regression line, the pattern is clear.

We can see a kind of split at around 7:00 until 10:00, and the spread is getting higher during evening and night. Since 7:00-10:00 is the period when most of the people commute to their offices; I assume it depicts different ridership patterns between working and non-working days. I will explore it further in the multivariate plots section.

Working Vs. Non-Working Day

This plot shows the effect of a working day on the number of pickups [12].



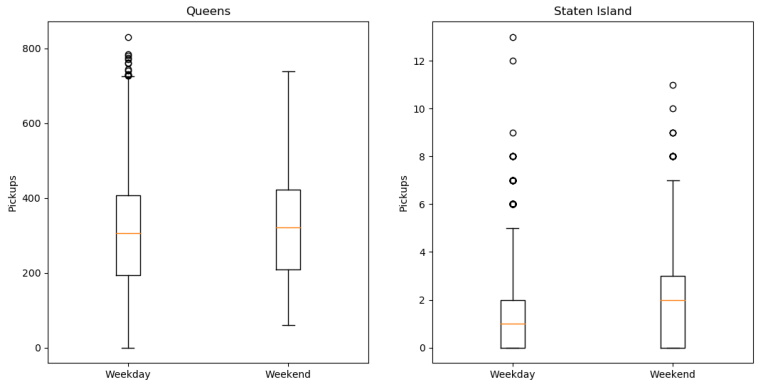


Figure 16

So, we can conclude that people usually take more uber rides on weekends rather than on weekdays. This probably indicates that a significant amount of people takes public transports (Subway, Bus, etc.) while they are going for their work.

Holiday Vs. Pickup

Now we plot a boxplot of pickups at Manhattan depending on holidays [13].

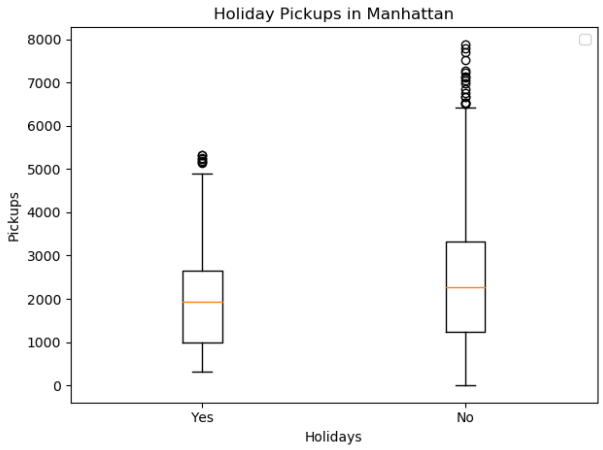


Figure 17

We can see that people take more Uber rides on holidays. As the means are close to each other, we would like to perform a T-Test [14] on the data to decide if holiday has an effect on the number of pickups at a borough. The result of the T-Test is:

```
> t.test(query_table$pickups~query_table$hday, mu=0, alt="two.sided", conf=0.95,
 paired=FALSE)

        Welch Two Sample t-test

data:  query_table$pickups by query_table$hday
t = 3.5766, df = 182.95, p-value = 0.0004455
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 163.8149 566.9347
sample estimates:
mean in group N mean in group Y
       2401.303        2035.928
```

As the p-value << 0.05, we can reject the null hypothesis (The means are different). As a result, we can conclude that holidays influence Uber ridership.

Weather

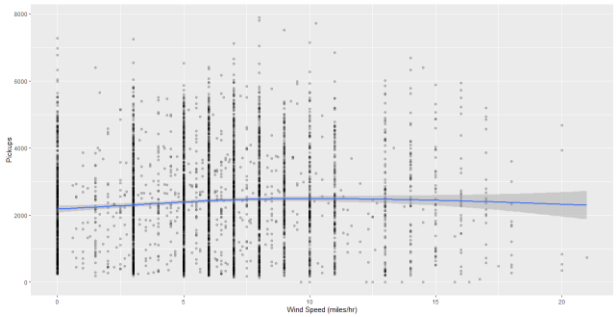


Figure 18: Wind Speed

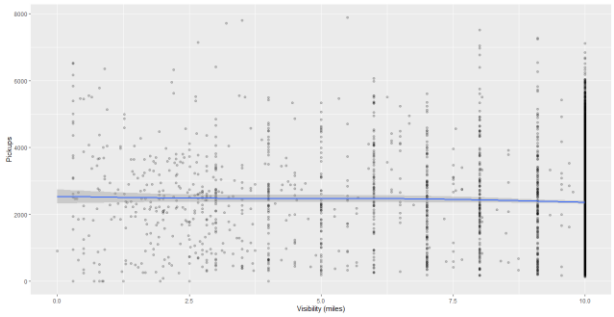


Figure 19: Visibility

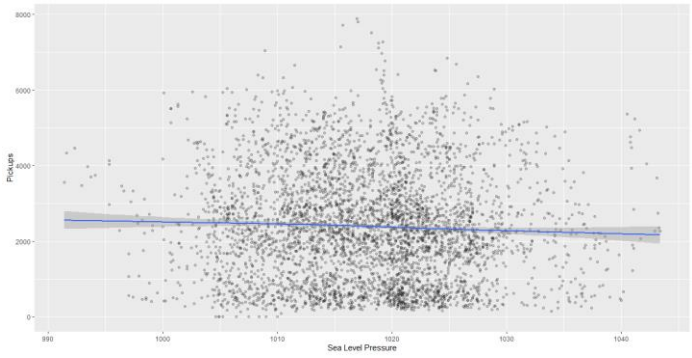
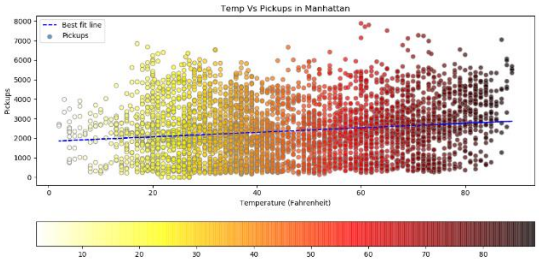


Figure 18: Wind Speed

There is a slight correlation between these parameters and ridership, but I don't think it is strong enough to affect the ridership.

Temperature



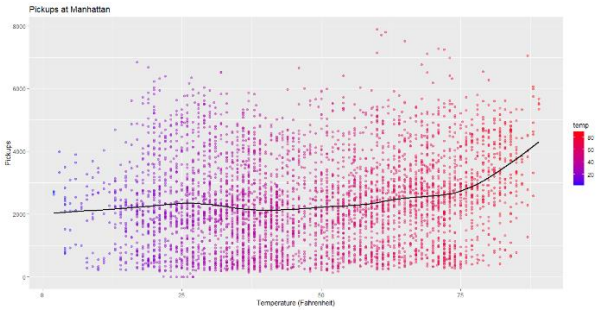
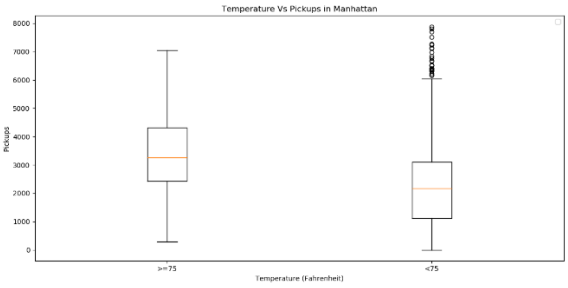


Figure 19(a): Python *Figure 19(b): R*

Although all the other weather parameters don't have a strong correlation with the number of pickups, the temperature seems to have a strong positive correlation on Uber Ridership [15]. To investigate more details, let us plot a boxplot of pickups with temperature as a factor [16].



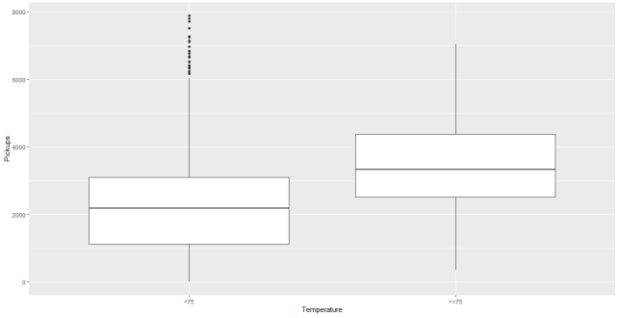


Figure 20(a): Python *Figure 20(b): R*

This box plot is self-explanatory. It clearly shows that at higher temperatures, people are prone to take more Uber rides.

Bivariate Analysis

It seems that time variables have a much stronger effect than weather variables on ridership. We noticed a powerful influence of time of the day with the demand, being able to explain 61% of the variance by itself. There is also a pattern on week level with the demand starting low on Monday and rising until it tops on Saturday then begins again to decrease. On a more macroscopic level,

there is a rise in demand on the evaluated period starting at the beginning of the year at around 2,000 pickups per hour and reaching 3,500 pickups by the end of June.

In the case of the weather, the analysis does not provide any strong indication that any weather variable affects the ridership. The only exception might be the temperature on its highest values.

Multivariate Plot Section

Borough & Time of the Day

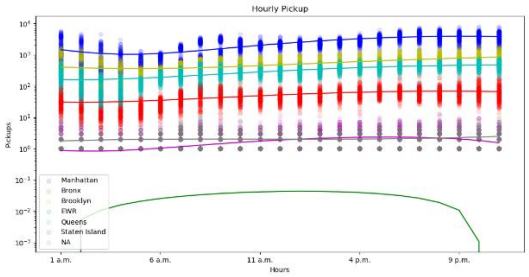


Figure 21(a): Python

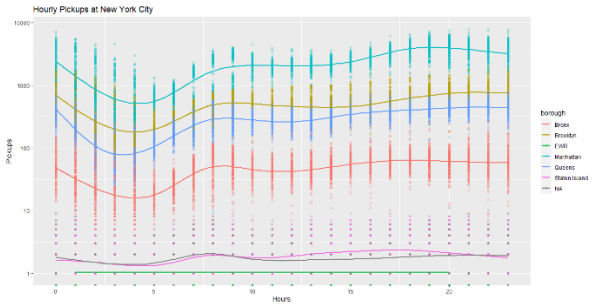


Figure 21(b): R

We have used the same plot of [11] but in the logarithmic scale for more in-depth analysis. The time of the day and the borough are two of the most significant variables in predicting the ridership. I am evident that the four considerable boroughs follow the same pattern. The same applies to Staten Island, but the values are much more disperse. Finally, EWR seems to have a random demand, with most of the values being zero with a few 1s and 2s.

Day & Time of the Day

This shows a tile plot of pickups of the city of Manhattan [17].

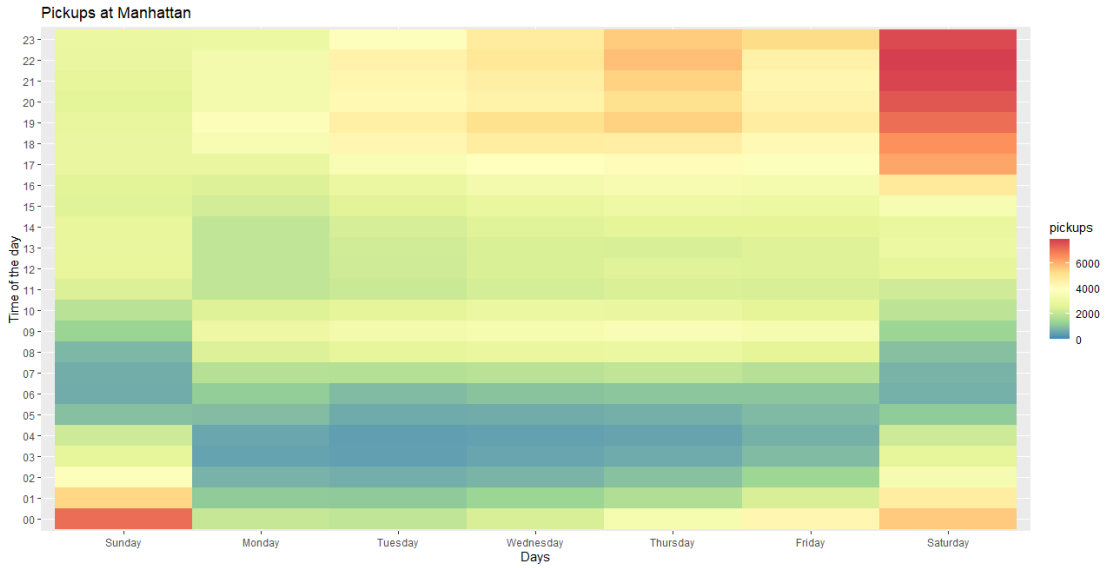


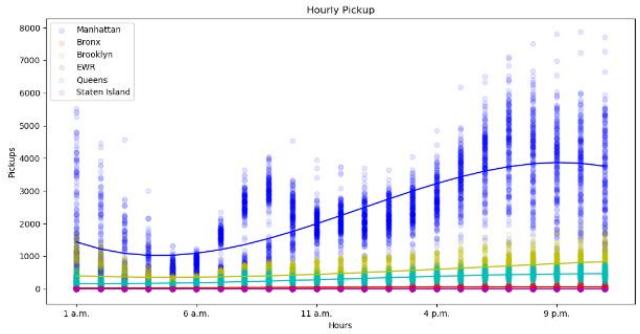
Figure 22

In the above heat map, we can see the ridership throughout the week. We can see the same pattern throughout the week, with the demand rising from Monday onward, especially in the afternoon/evening hours and peaking on Saturday. We can also notice a transposition of the demand during Saturday and Sunday for 3-4 hours comparing to working days.

Working Vs. Non-Working Day

Figure 23

This plot distinguishes between the effect of weekend or weekday on pickups [18]. Previously we saw a split in the hourly pickup from 7-10 a.m. Figure 23 is the explanation of why there was a split before.



Multivariate Analysis

Some indications became clear facts during this section.

- There is a clear pattern of ridership, both during the day and during the week, followed by the four major boroughs.
- Holidays and weekends change the ridership through the day, but they do not have any significant effect on the total daily ridership.
- Surprisingly, there is not a single weather variable affecting the ridership. I was expecting that rainy day or freezing days to have a positive impact on the ridership, but there are no pieces of evidence to support my intuition.

Code References

[1] *data_cleaning.py*

[2] *pickups_per_day.py; pickups_per_day.R*

[3] *pickups_per_day_2.py; pickups_per_day_2.R*

[4] *pickups_per_day_borough.py; pickups_per_day_borough.R*

[6] *pickups_each_month.py; pickups_each_month.R*

[5] *pickups_each_borough.R*

[7] *weather.py; weather.R*

[8] *scatter_pickup_all.R*

[10] *scatter_pickup.py; scatter_pickup.R*

[9] *weekly_pickup.py; weekly_pickup.R*

[11] *hourly_pickup.py; hourly_pickup.R*

[12] *working_day.py*

[13] *holiday.py*

[14] *holiday_ttest.R*

[15] *temperature.py; temperature.R*

[16] *temp_box.py; temp_box.R*

[17] *pickup_tile.R*

[18] *hourly_workday.R*

Reference

1. Uber pickups in New York City, from 01/01/2015 to 30/06/2015 (uber-raw-data-janjune-15.csv) from kaggle
2. Weather data from National Centers for Environmental Information.
3. LocationID to Borough mapping from FiveThirtyEight