ANALYSIS OF UBER PICKUPS AND WEATHER IN NEW YORK CITY

Final Project Report

MSCI 3250 Data Wrangling – Fall19

Dec. 13th, 2019

Team 10:

Alice Jiang

Kaidi Meng

Fangxing Wei

Xin Shu

Content

1. Motivation

Uber is a useful App which connects drivers to customer. By using this app, drivers can seek out more customer and find them more precisely, because Uber will automatically send customer's location and provide navigation to a specific driver. Moreover, customer know that it is so hard to find a taxi when there are tons of people who need a ride, plus sometimes they do not want to wait outside for a taxi when it is raining or in a bad weather condition. Uber offers such a good service for customer so that they can wait inside a room for the driver to pick them up. However, where should drivers go when they want an order, and under what weather condition will they receive more orders? Those specific locations, specific weather conditions and temperature information in New York City is very valuable for each Uber driver because they can improve their profit and lower their cost based on that.

In this project, we used data about Uber pick up information from Kaggle to analyze the pickup rate in New York City in order to give Uber drivers some suggestions when they are providing the service.

2. Data

2.1 dataset

In this project, there are three primary sources of data, dataset named "Uber Pickups in New York City"[1] on Kaggle website, average temperature, and rainfall for each date in NYC that matching with the dataset.

The dataset contains 4,534,368 examples in 6 separated csv files, one for each month (through April 2014 to September 2014). In order to integrate the raw data with web scarping data more effectively, we decided to randomly choose 5,000 records in each month in Excel and manually combine data from April to September into one file.

2.2 Scrapping Temperature

We scraped every day's average temperature from April 2014 to September 2014 from Almanac.com[2], and then built a data frame with the average temperature data. Then we added one column to contains the date information. The *all_temp* data frame we got contains 183 observations of 2 variables.

2.3 Scrapping Rainfall

We also scraped every day's rainfall amount from April 2014 to September 2014 from cnyweather.com[3]. We saved these data as a data frame and added one column to show the date information. This *all_rain* data frame we got also contains 183 observations of 2 variables.

2.4 Integration Temperature and Rainfall into Dataset

[1] https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city
[2] https://www.almanac.com/weather/history/NY/New%20York/2014-04-01
[3] http://www.cnyweather.com/wxraindetail.php?year=2014

Based on these 2 data frames we mentioned above, we made a horizontal merge of these 2 data frames with our original dataset based on date column. This made it very clear and easy for us to check and analyze the impact that the weather factors have on Uber pickup counts. And we dropped the *Base* column as it does not mean something to our research. After these changes, we got a data frame with 30,000 observations of 8 variables.

Since the web scarping information was originally factors, for our later calculations to be carried out accurately, we transformed the Temp and Rainfall columns to numeric data type.

Table 1 data dictionary

Column	Type	Description
Date	date	Date of the ride (e.g. 2014-04-01)
Lat	numeric	Lattitude of the pickup location (e.g. 40.7575)
Lon	numeric	Longitude of the pickup location (e.g. -73.9818)
Time	numeric	Extracted hour of the pickup time (e.g. 17)
Day	character	Day of the week (e.g. Monday)
Temp	numeric	Temperature of the pickup date (e.g. 45.1)
Rainfall	numeric	Rainfall of the pickup date (e.g. 0.00)
TimePeriod	factor	Time interval of the pickup (e.g. Morning)

3. Analysis

The goal of this project is to examine how time and weather (including temperature and rainfall) affect Uber orders amount.

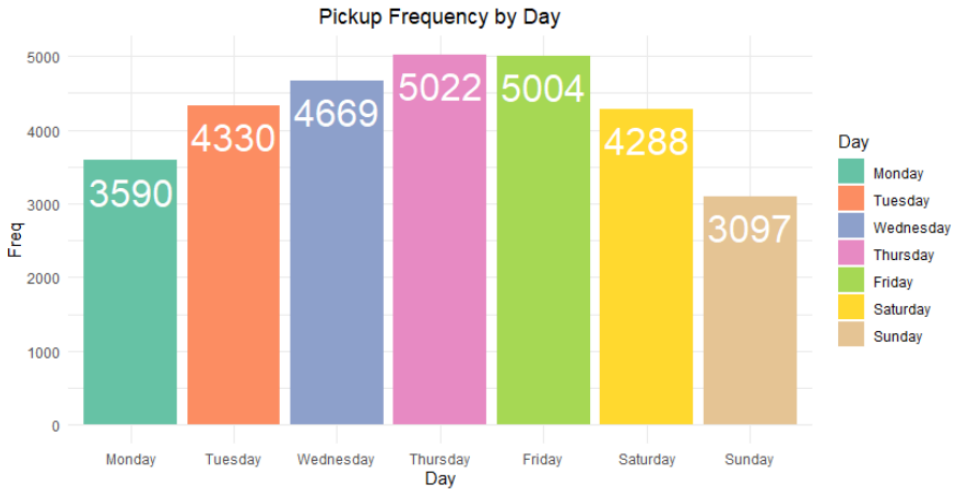
3.1 Highest Frequency by Day

Which day of a week has highest frequency for Uber orders?

We created a table to measure the order amount in each day. In order to interpret easily, the table is shown as a bar plot as Figure 1. The amount of Uber pickups in each day was labeled in the top of each bar.

According to Figure 1, Thursday and Friday have the highest amount of Uber orders in our dataset. This draws another interpretation, since Thursday and Friday are close to the weekends, more people are likely to go out frequently before weekends by using Uber.

Figure 1 Pickup Frequency by Day



3.2 Busiest Period

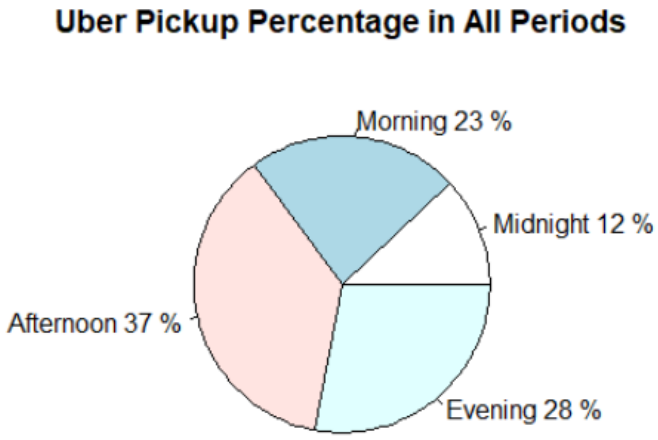
When is the busiest time for Uber driver?

To better understand when the busiest hour for Uber pickups is, we investigate it by cutting a day into four sections. Midnight is from 0:00-6:00, Morning is from 7:00-12:00, Afternoon is from 13:00-18:00, and Evening is from 19:00-23:00. We created a table to count amount of Uber pickups in each period and proportion shown in Table 2.

Table 2 Number of Pickup & Percentage in Each Period

Period	**Midnight**	**Morning**	**Afternoon**	**Evening**
Amount	3581	6955	11103	8361
Percentage (%)	11.94	23.18	37.01	27.87

Figure 2 Uber Pickup Percentage in All Periods

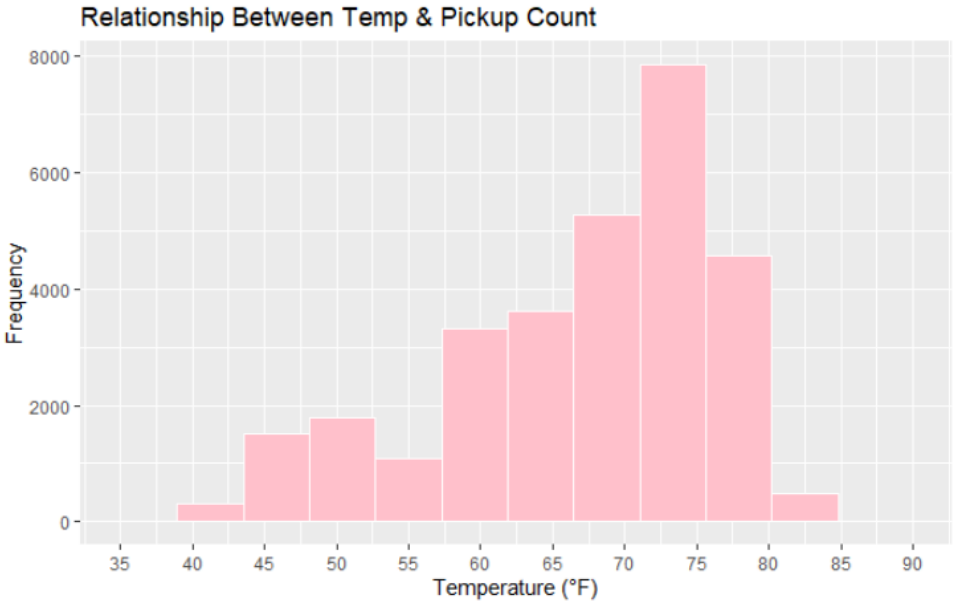


By creating a pie chart based on the percentage values from Table 2, we can easily observe that Uber order in afternoon occupies the largest portion in whole day. We can tell that people in our data set are likely to using Uber around the latter half of a day. Which means both Uber users and Uber drivers are very active in the afternoon.

3.3 Relationship between Uber pickup and Temperature

Besides the relationship between times/days and Uber pickup rate, we took more elements in consideration. We constructed the frequency of Uber orders for each day's temperature. Since the temperature only appears when there's an Uber order after we merged all the tables, the number of the frequency that a certain temperature appears is equal to the Uber pickup orders in that specific day. Therefore, by counting the frequency of one temperature, we can get the number of pickup orders as well. Figure 3 shows the relationship between temperature and Uber pickups. It reflects that pickup counts reach the highest when the temperature is between 70°F and 75° F.

Figure 3 Relationship Between Temperature and Uber Pickup



Furthermore, based on result of kruskal test, the p-value is 0.4836, which is greater than 0.05, suggesting that there is NOT a significant difference in number of Uber pickups in different temperature.

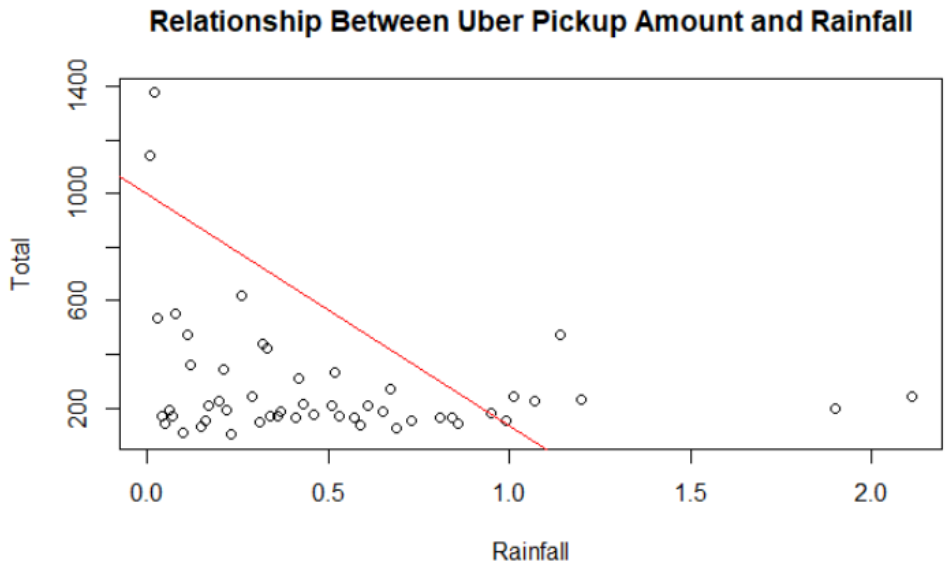
Figure 4 Kruskal Test on Pickup and Temperature

```
> kruskal.test(temp_summary$n~temp_summary$Temp)

        Kruskal-Wallis rank sum test

data:  temp_summary$n by temp_summary$Temp
Kruskal-Wallis chi-squared = 132, df = 132, p-value = 0.4836
```

3.4 Relationship between Uber Pickup and Rainfall

To take our analysis to a wider level, we explored the relationship between Uber pickup and rainfall using linear regression model. The correlation between rainfall (in inch) and number of Uber pickups is –0. 2536957. We can tell the relationship between Uber pickups and rainfall is negative, which means in our situation, the lower the rainfall, the more the pickup orders.

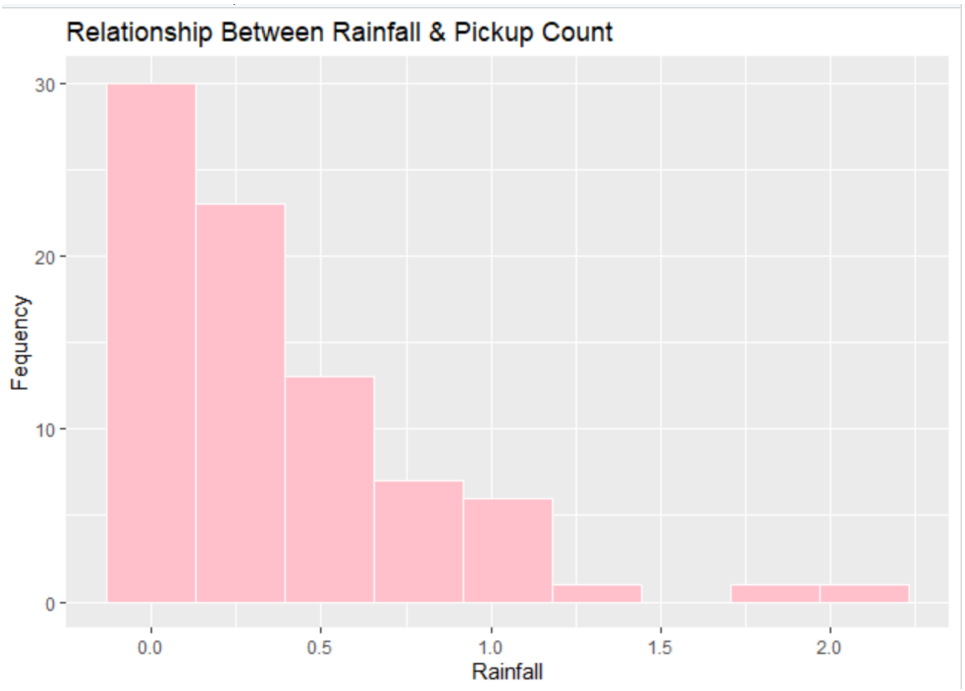
Figure 5 Relationship Between Uber Pickup Amount and Rainfall



As shown in Figure 5, the Uber pickup orders are likely to increase when there's less rain. This result is opposite with our hypothesis before we build the model. In normal logic, the heavier the rainfall, the more people prefer to call Uber. Our result may lead by limited information compared with the real world, and the timespan is from April to September only.

From the plot shown below (Figure 6), it tells the number of people in our dataset using Uber under different rainfall levels. We can also conclude that, under our data environment, people tend to use Uber more when there's less rainfall.

Figure 6 Relationship between Rainfall & Pickup



3.5 Hottest Pickups Locations in NYC

Which location in New York City is the hottest place for Uber pickups?

Our dataset contains longitude and latitude for each Uber picks, which can help us locate the specific place. In order to investigate this question, we grouped by longitude and latitude and sorted in descending order to check which location is the hottest. We extracted and visualize the top 100 locations as shown in Figure 7. The dots are focus on four main zones – mid and lower Manhattan, LaGuardia Airport (LGA), John F. Kennedy Airport (JFK), and Newark Airport (EWR). There has huge number of pickups around Manhattan. This result is under our expectations, because Manhattan has densely population in New York City. Based on the results, we can conclude that people in our dataset have lots of activities around these four places.

Does the result change if we only extract the top 10 locations?

In order to observe the result better, we visualized the top 10 places shown in Figure 8. The occurrences around the airports (especially JKF and LGA) occupy relatively high proportion, but Manhattan zone reduced a lot. The later 90% people of our dataset may prefer call Uber around Manhattan. All in all, airports (LGA and JFK) are the hottest pickup locations.

Figure 7 Top 100 Hottest Pickups Location

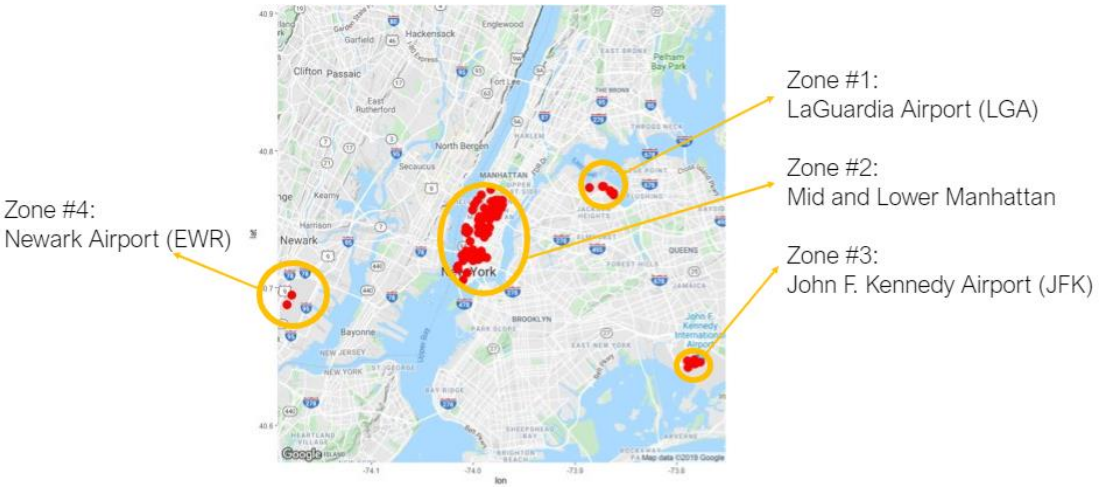
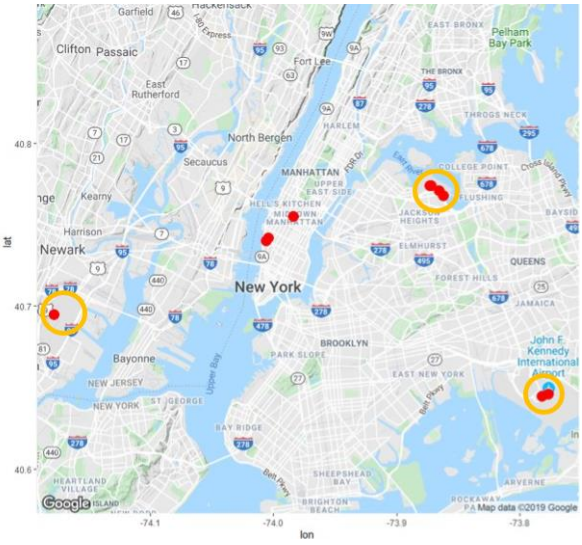


Figure 8 Top 10 Hottest Pickups Location



4. Conclusions

This project incorporates temperature and rainfall data on Uber pickups in New York City from April to September in 2014 to investigate how weather affect Uber orders. Specifically, we performed 5 analyses, considering day, time, temperature, rainfall, and location perspectives. By using summary tables, plots, kruskal test, linear regression, and maps, we can conclude that weather condition can affect Uber pickups. When temperature between 70°F and 75° F with almost no rainfall, people in JFK and LGA have high frequency to place Uber orders on Thursday afternoon.

Moreover, we recommendation that people who interested in being an Uber driver in New York City can spend time around airports on Thursday afternoon. As for Uber rider, they should expect to wait longer time if they must call Uber around airports, or on Thursday afternoon, or when temperature around 70°F to 75° F with little or no rainfall.

5. Limitations

Although we can draw some conclusions based on the data we collected above, there are still some limitations. Since our records are from April to September in 2014, it is not very practical for recent years. Furthermore, we only used about 7% (30,000 out of 4,534,368) of the whole data in Kaggle, it is hard to say our analysis are close to accurate. To improve the whole analysis, we believe that if we add more features to the main table, it would be closer to the actual world, something like when its holiday, or when Uber company sends coupons to users in order to let them call Uber frequently.

The temperature we scrapped from website are average temperatures of the day. This information may not be accurate since temperature changes every hour. For example, if the Uber was pickup at 5:00 in the morning, the temperature at that time may lower than the average. If we can get temperature records for each hour, which means the result might be more accurate.

Another limitation is rounding in time. In order to analyze easily, we extract only hour from time column in raw data. This can mislead part of result, for example, 13:59:00 turned to be 13 in our new data set. If we round the time to 14:00 would be more reasonable. However, our current skills do not allow us to apply such detail analysis.