Artificial Intelligence Project:

Uber Car Traffic in New York City Dataset Analysis

Yu Wu (193080180)

Huiqiao Yang (186803710)

Wilfrid Laurier University

CP 468: Artificial Intelligence

Dr. Sukhjit Singh Sehra

Edition: July 19, 2022

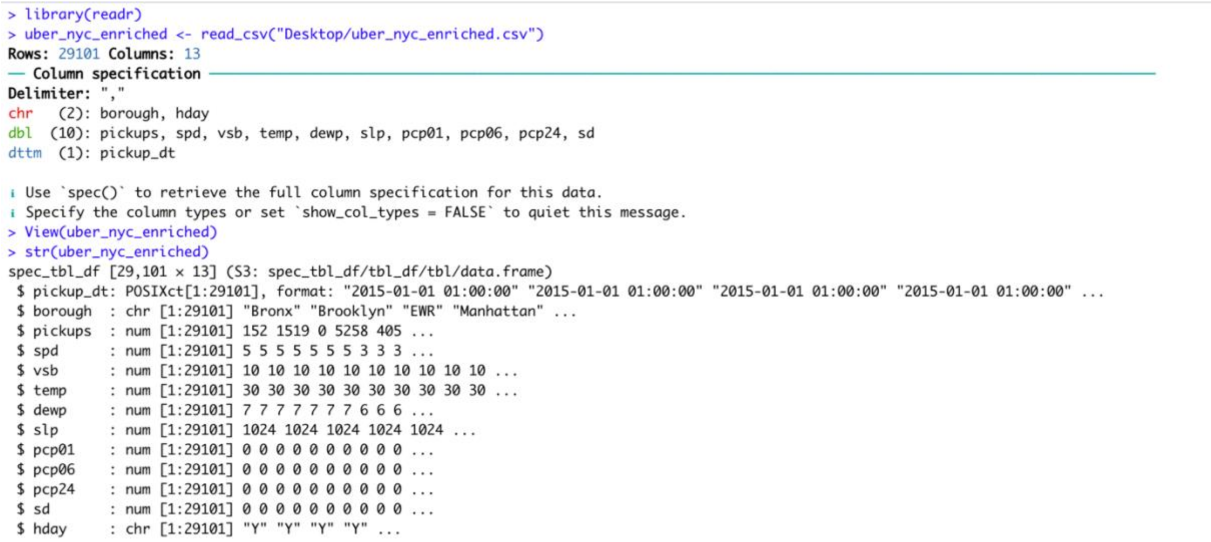
I. Introduction

Uber is an essential tool for most people to get around daily, and this project aims to predict when people prefer to go out by analyzing a subset of Uber pickup data from different weather and weekends in New York City, USA.

II. Project Description

Data Collection

The uber car traffic dataset from Kaggle contains a subset of Uber pickup data for weather, boroughs, and holidays. By importing database into the R studio and view the basic 13 variables in the *uber_nyc_enriched.csv* databse,



```
> library(readr)
> uber_nyc_enriched <- read_csv("Desktop/uber_nyc_enriched.csv")
Rows: 29101 Columns: 13
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr   (2): borough, hday
dbl  (10): pickups, spd, vsb, temp, dewp, slp, pcp01, pcp06, pcp24, sd
dttm  (1): pickup_dt

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> View(uber_nyc_enriched)
> str(uber_nyc_enriched)
spec_tbl_df [29,101 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ pickup_dt: POSIXct[1:29101], format: "2015-01-01 01:00:00" "2015-01-01 01:00:00" "2015-01-01 01:00:00" "2015-01-01 01:00:00" ...
 $ borough  : chr [1:29101] "Bronx" "Brooklyn" "EWR" "Manhattan" ...
 $ pickups  : num [1:29101] 152 1519 0 5258 405 ...
 $ spd      : num [1:29101] 5 5 5 5 5 5 5 3 3 3 ...
 $ vsb      : num [1:29101] 10 10 10 10 10 10 10 10 10 10 ...
 $ temp     : num [1:29101] 30 30 30 30 30 30 30 30 30 30 ...
 $ dewp     : num [1:29101] 7 7 7 7 7 7 7 6 6 6 ...
 $ slp      : num [1:29101] 1024 1024 1024 1024 1024 ...
 $ pcp01    : num [1:29101] 0 0 0 0 0 0 0 0 0 0 ...
 $ pcp06    : num [1:29101] 0 0 0 0 0 0 0 0 0 0 ...
 $ pcp24    : num [1:29101] 0 0 0 0 0 0 0 0 0 0 ...
 $ sd       : num [1:29101] 0 0 0 0 0 0 0 0 0 0 ...
 $ hday     : chr [1:29101] "Y" "Y" "Y" "Y" ...
```

there are 29,101 hourly aggregated observations with 13 variables in the dataset.

<u>Variables Description</u>

pickup_dt: pickup date

borough: place to pickup

pickups: number of pickups

spd: wind speed

vsb: visibility

temp: temperature

dewp: dew point (humidity)

slp: sea level pressure

pcp01: precipitation for last 1 hour

pcp06: precipitation for last 6 hours

pcp24: precipitation for last 24 hours

sd: depth of snow

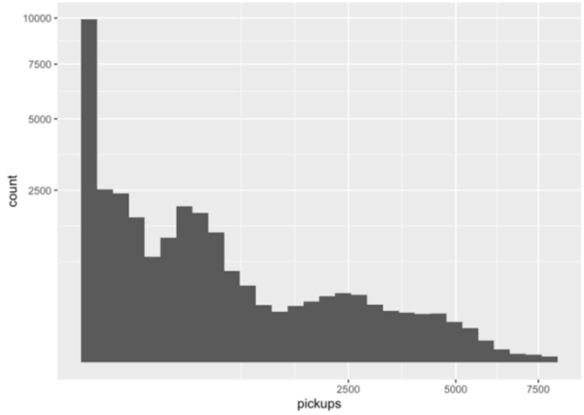
hday: holiday for ‘Y’(1), ‘N’(0) otherwise (dummy variable)

Exploratory Data Analysis

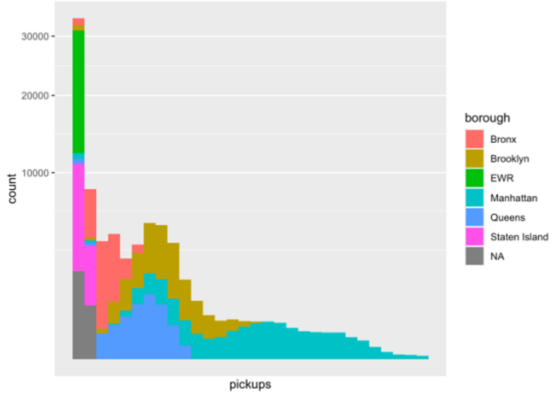
This project will apply R and exploratory data analysis techniques to explore factors such as weather/humidity/holidays that influence Uber car demand in New York City.

The detailed analysis process is as follows:

- Univariate Plots & Analysis



Distribution of Uber pickups in different boroughs

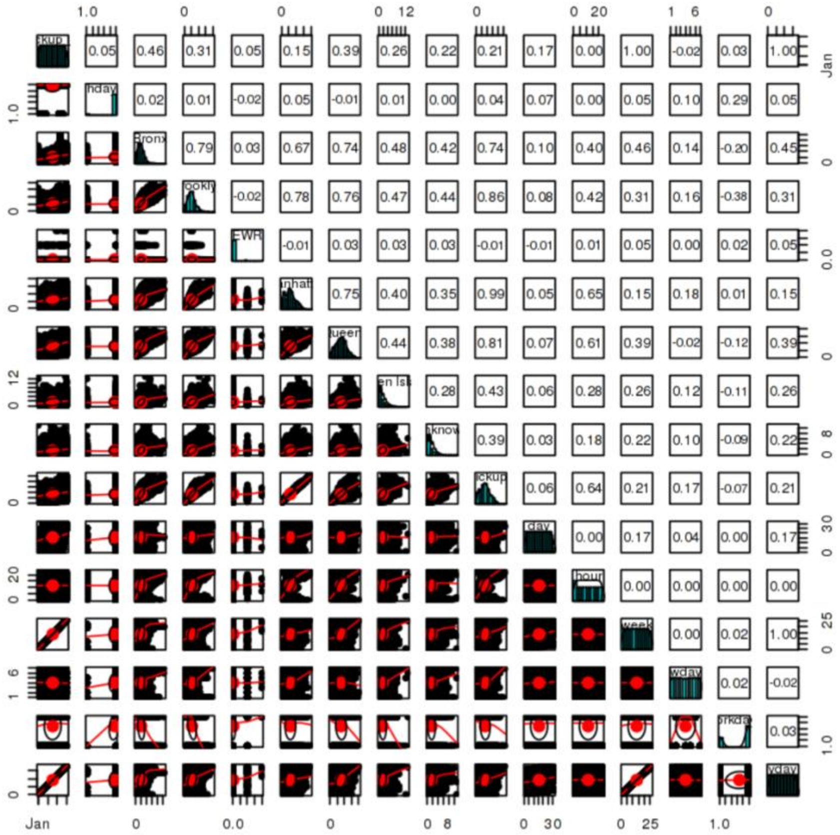


Through the graphical analysis above,

There is a clear difference in patronage between the different boroughs. By far, Manhattan has the highest demand, followed by Brooklyn, Queens and the Bronx. Meanwhile, there are few pickups in EWR and Staten Island.

Although pickups in the major boroughs of Manhattan, Brooklyn, Queens and the Bronx follow a normal distribution on a square root scale, there appears to be a difference of about 1,500 in Manhattan among them. It can therefore be assumed that there must be a pattern in demand where demand rises rapidly from about 1,000 to close to 2,500.

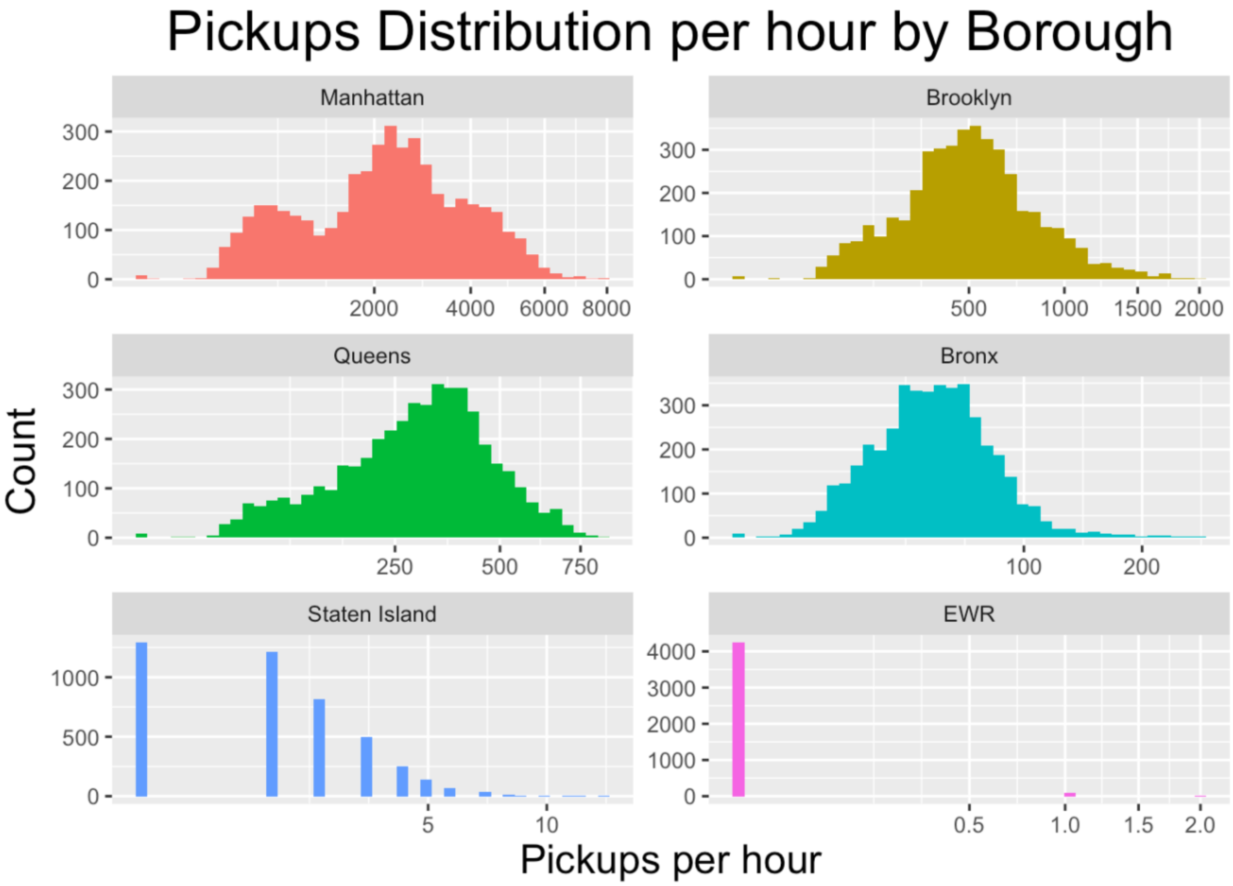
- Bivariate Plots & Analysis



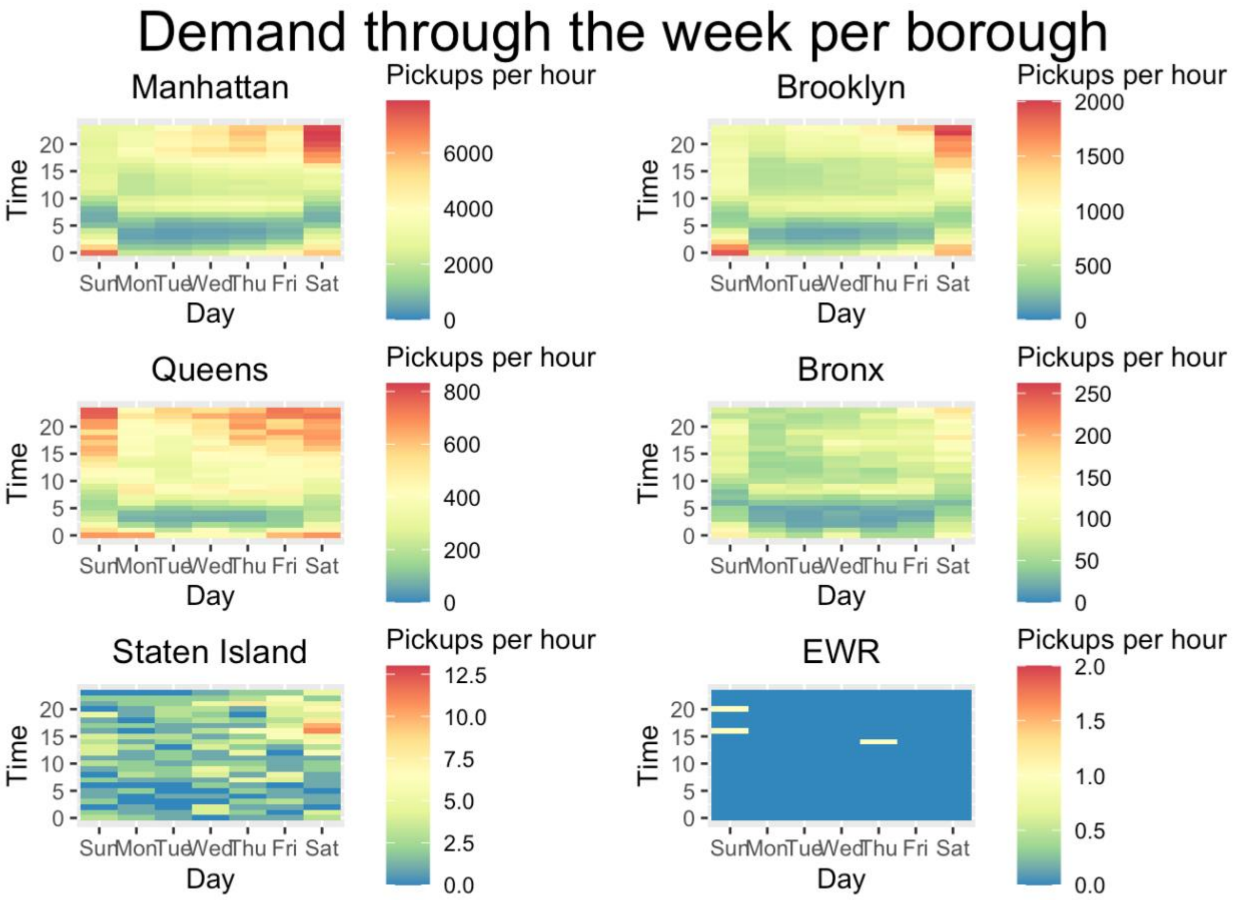
From the above pairs we can see that，

this is a time variable matrix which shows the effect of date time on each borough. This matrix shows a strong correlation between the Bronx, Queens, and Brooklyn and date time, which means that ridership is higher in these locations compared to others. Secondly, this variable actually has no effect on EWR and has very little effect on Manhattan. This suggests that demand for uber has increased significantly in some areas, while demand for uber has increased slowly and consistently in others. In addition, the matrix shows that the traffic in all boroughs except Staten Island and EWR is time dependent.

III. Result



The distribution of the four boroughs is mainly normal bimodal because of the rapid rise in demand in the morning. The pickups on Staten Island follow a geometric distribution due to very little demand in the area. On the EWR, demand is almost zero, with only a very few pickups that we might consider to be outliers.



The demand pattern for each borough is presented in the graph above, which shows that all four major boroughs follow roughly the same pattern during the day and during the week. During times that are not holidays, pickup demand drops after midnight and then begins to rise rapidly during the morning peak around 6am. It plateaus in the afternoon and begins to rise again during the evening peak. During the week, demand starts to drop on Monday and then rises. It peaks on Saturday and then drops again on Sunday. This pattern is more pronounced in Manhattan and Brooklyn.

In the two smaller boroughs, Staten Island and EWR, demand in Staten Island appears to be random during the day, but rises slightly as the week progresses. EWR, on the other hand, has virtually no demand.

IV. Conclusion

The dataset used for this project included data on the number of passengers in New York City for the first six months of 2015. During these six months, there was a general upward trend in demand for Uber cars in New York City, with total demand increasing from 2,000 to 3,500 pickups per hour. Before the survey, I believed that weather changes and the factors that affect them would significantly impact Uber traffic in New York City. However, based on the data analysis above, it was shown that the weather variable did not have any or a fragile effect on passenger traffic.

With the above findings, I was able to forecast the range and model based on the customer as well as driver demand. The weekly real-time forecast model can be used to get a general idea of the demand for the next week, and the real-time system can compare the forecast for each district with the location of the Uber car and highlight areas based on the driver's application to help them find the customer's location more efficiently. However, in very irregular conditions yet still prone to misestimation. At the same time, since current observations can affect future forecasts, specific demands may at some point lead to incorrect estimates of later forecasts.

V. Reference

Y. (2017, February 15). *EDA on Uber's ridership.* Kaggle. https://www.kaggle.com/code/yannisp/eda-on-uber-s-ridership/data

VI. Github Link

https://github.com/Otterlen-w/CP468_Project