Uber Rides Data Analysis

Karthika Chiluka

15/06/2022

## Summary

For many individuals, Uber has been a crucial business both for the companies and customers. People can request cabs and move from one place to the other at a fee without even having to own a car. Through the utilization of monitored systems, the operations of Uber business are made effective and convenient. Recently, there has been an increase in the number of uber rides requests, and this has called for an effective system that is reliable in providing predictions on the periods, seasons, and times when there is an increased demand for uber rides requests to ensure there are sufficient cabs available at the moment to avoid losses and customer inconveniences. Therefore, this project implements a data analysis approach that provides insights into factors influencing Uber ride demand. The project analyzes the days of the week and weather conditions with a higher and lower count of uber trips that customers requested. With the project, it is possible to analyze when to have a higher or lower number of cabs available and have proper planning in an automated functionality.

## Introduction

Uber is a ride-sharing company that offers its customers services for connecting local drivers and riders for transport services at a fee (Rogers, 2015). Uber services have been expanding and globally have become very popular. They include an app that helps individuals request rides at any given time from available local Uber drivers. For efficient and reliable services, there needs to have Uber cabs available for any request made by customers. Uber Company has been trying to ensure that they satisfy their customers by providing reliable services at all times. However, at times it happens that there are few Uber cabs than the demand at a given season, time of the day, and weather conditions and vice versa, which affects the Uber business largely, and they have been making losses (Geitung, 2017). With this project's multivariate data analysis model, it is possible to analyze the various factors like day of the week, time of the day, and weather conditions and how they affect Uber rides. Machine learning methods are also applied in this project to perform data analysis on the Uber rides demands dataset. The data mining and analysis methods involve utilizing computational methods that help determine the essential patterns of a large dataset (Radhika & Masood, 2021). For completion of this project includes a literature review, theory, data, methodology, results, implications, and conclusion sections.

## Literature Review

In a study by Brodeur & Nield (2018), the authors illustrated a significant correlation between whether it rained and the number of Uber rides. In their study, the authors were able to identify that when it is raining, the number of Uber rides per hour is approximately 19% higher (Brodeur & Nield, 2018). According to the authors, this illustrated a higher demand for Uber rides during rainy hours. The weather has had a significant effect on the demand for Uber rides. In a study by Cohen et al. (2016), the authors were able to identify that the day of the week affected the demand for Uber rides. During the weekends, it was identified that the demand for Uber rides was a bit higher compared to weekdays. This might have resulted from people's engagement in different activities where they have to travel to several places for refreshment and personal reasons. However, little research has been done regarding the main factors affecting the demand for Uber rides for different pickups. There has only been research on factors that affect a surge in the prices of Uber rides. Therefore, this project focuses on environmental factors such as weather conditions and time and how they affect the demand for Uber rides for different pickups or boroughs for proper planning and management in terms of the availability of Uber cabs or taxis. Therefore, this project needs to answer the research question of "Which weather conditions mainly affect the demand for Uber rides for different pickups?" Also, the research intends to examine whether temperature and time of the day are associated with the demand for Uber rides.

## Data

The Uber dataset was downloaded from <https://www.kaggle.com/datasets/yannisp/uber-pickups-enriched/download>. This dataset involves:

## 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.

## # A tibble: 6 x 13

## pickup\_dt borough pickups spd vsb temp dewp slp pcp01 pcp06

## <dttm> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

## 1 2015-01-01 01:00:00 Bronx 152 5 10 30 7 1024. 0 0

## 2 2015-01-01 01:00:00 Brooklyn 1519 5 10 30 7 1024. 0 0

## 3 2015-01-01 01:00:00 EWR 0 5 10 30 7 1024. 0 0

## 4 2015-01-01 01:00:00 Manhatt~ 5258 5 10 30 7 1024. 0 0

## 5 2015-01-01 01:00:00 Queens 405 5 10 30 7 1024. 0 0

## 6 2015-01-01 01:00:00 Staten ~ 6 5 10 30 7 1024. 0 0

## # ... with 3 more variables: pcp24 <dbl>, sd <dbl>, hday <chr>

First, the pickup\_dt column is converted to the Date type variable. The POSIXct function is used.

uber$pickup\_dt <- as.POSIXct(uber$pickup\_dt, format="%m/%d/%Y %H:%M:%S")

uber$Time <- format(as.POSIXct(uber$pickup\_dt, format = "%m/%d/%Y %H:%M:%S"), format="%H:%M:%S")

uber$pickup\_dt <- ymd\_hms(uber$pickup\_dt)

The Date column is split into four columns day,month,year and weekday. This is because they will be essential in analysis for this project.

# Create individual columns for month day and year

uber$day <- factor(day(uber$pickup\_dt))

uber$month <- factor(month(uber$pickup\_dt, label=TRUE))

uber$year <- factor(year(uber$pickup\_dt))

uber$dayofweek <- factor(wday(uber$pickup\_dt, label=TRUE))

The next step involve creation three new columns for storing Hour, Minute ,and Second of the trip.

# Add Time variables as well

uber$second = factor(second(hms(uber$Time)))

uber$minute = factor(minute(hms(uber$Time)))

uber$hour = factor(hour(hms(uber$Time)))

## Methodology

Having the data ready for analysis, the next step involved data analysis depending on time analysis and weather condition analysis basis. First, an illustration of the distribution of total trips in different hours in a day is done over six months by using the group by the Hour column.

#trips by hours in a day

data\_per\_hour <- uber %>%

group\_by(hour) %>%

dplyr::summarize(Total = n())

Then a bar chart showing the distribution of trips in a day is created.

# Plot the data by hour

ggplot(data\_per\_hour, aes(hour, Total)) +

geom\_bar(stat="identity",

fill="blue",

color="red") +

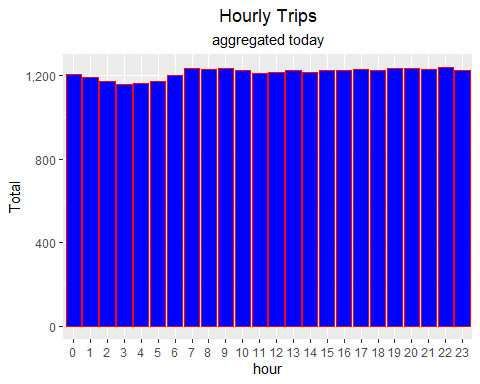
ggtitle("Hourly Trips", subtitle = "aggregated today") +

theme(legend.position = "none",

plot.title = element\_text(hjust = 0.5),

plot.subtitle = element\_text(hjust = 0.5)) +

scale\_y\_continuous(labels=comma)



Once the distribution of trips in a day is analyzed, the next is the analysis of the distribution of trips in each month. This is performed using the group by month column.

#trips by month

Monthly\_Data <- uber %>% group\_by(month) %>% dplyr::summarize(Total = n())

Monthly\_Data

## # A tibble: 6 x 2

## month Total

## <ord> <int>

## 1 Jan 4897

## 2 Feb 4492

## 3 Mar 4957

## 4 Apr 4798

## 5 May 5058

## 6 Jun 4899

The distribution hs been plotted in a barchart

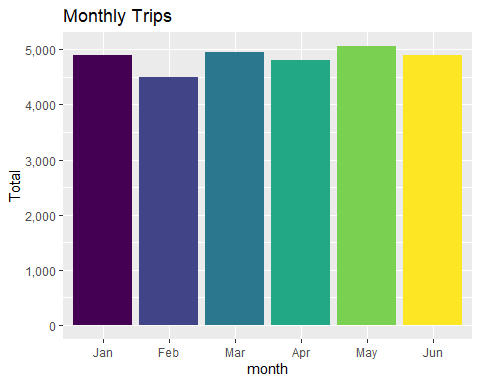
ggplot(Monthly\_Data, aes(month, Total, fill = month)) +

geom\_bar(stat = "Identity") +

ggtitle("Monthly Trips") +

theme(legend.position = "none") +

scale\_y\_continuous(labels = comma)



The next stage involved the illustration of how weather conditions affect the Uber rides, by concentrating on the pickup areas labelled pickups.

histogram <- function(varname, bs = NULL, bw = NULL){

h <- ggplot(uber.spread, aes\_string(varname)) + geom\_histogram(bins = bs, binwidth = bw)

return(h)

}

uber.spread <- uber %>% spread(borough, pickups, fill = 0)

d <- melt(uber.spread %>% dplyr::select(spd:sd)) #spd:sd = all the weather variables

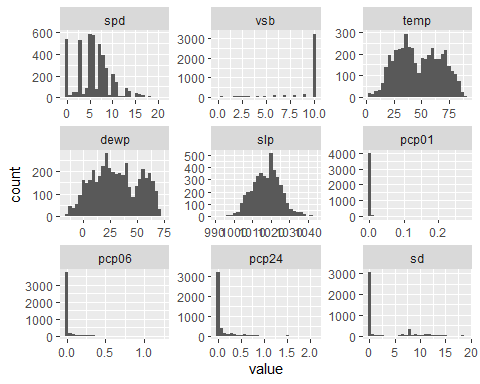
## No id variables; using all as measure variables

ggplot(d, aes(value)) +

geom\_histogram() +

facet\_wrap(~variable , scales = 'free')

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Next, involved the creation of the borough column as a one variable, named pickups and divide the days of the week as weekdays named wday.

uber.spread <- uber.spread %>%

mutate(pickups = Bronx +Brooklyn + EWR + Manhattan + Queens + `Staten Island`

+ !is.na(NA)) %>%

mutate(day = day(pickup\_dt)) %>%

mutate(hour = hour(pickup\_dt)) %>%

mutate(week = week(pickup\_dt)) %>%

mutate(wday = wday(pickup\_dt, label = TRUE)) %>%

mutate(workday = ifelse(wday == 'Sat' | wday == 'Sun' |

hday == 'Y', 'N', 'Y')) %>%

mutate(yday = yday(pickup\_dt))

uber <- uber %>%

mutate(day = day(pickup\_dt)) %>%

mutate(hour = hour(pickup\_dt)) %>%

mutate(week = week(pickup\_dt)) %>%

mutate(wday = wday(pickup\_dt, label = TRUE)) %>%

mutate(workday = ifelse(wday == 'Sat' | wday == 'Sun' |

hday == 'Y', 'N', 'Y')) %>%

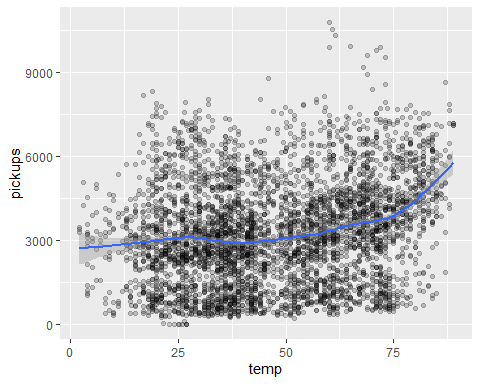
mutate(yday = yday(pickup\_dt))

ggplot(uber.spread, aes(temp, pickups)) +

geom\_jitter(alpha = 0.2) +

geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



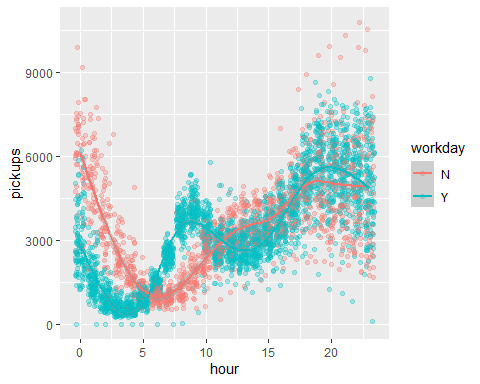
Then next will involve an illustration of the distribution of pickups between the working days and non-working days.

ggplot(uber.spread, aes(hour, pickups)) +

geom\_jitter(alpha = 0.3, aes(colour = workday)) +

geom\_smooth(aes(color = workday))

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



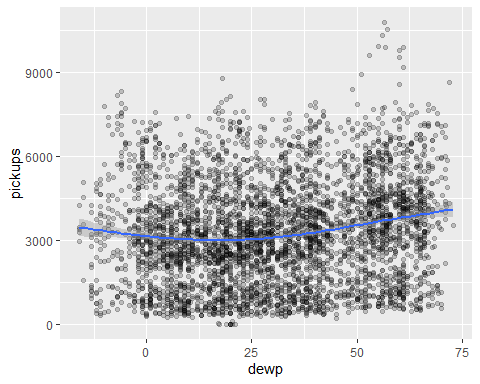
The next stage involved showing how Dew point named Dewp correlates with the pickups and how it affects it.

ggplot(uber.spread, aes(dewp, pickups)) +

geom\_jitter(alpha = 0.2) +

geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



Next, was to show the distribution between the precipitation named pcp01 and pickups.

ggplot(uber.spread, aes(pcp01, pickups)) +

xlim(0,quantile(uber.spread$pcp01, 0.95)) +

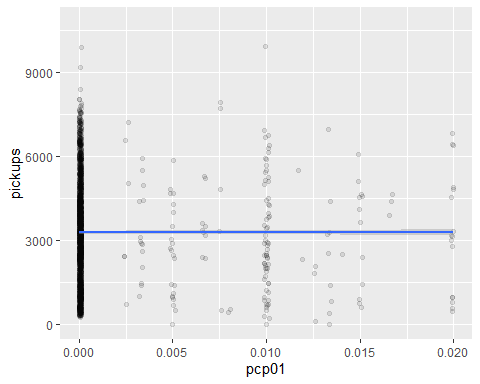
geom\_jitter(alpha = 0.1) +

geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 214 rows containing non-finite values (stat\_smooth).

## Warning: Removed 2280 rows containing missing values (geom\_point).



Next involves the illustration of the association between the pickups and wind speed named spd.

ggplot(uber, aes(spd, pickups)) +

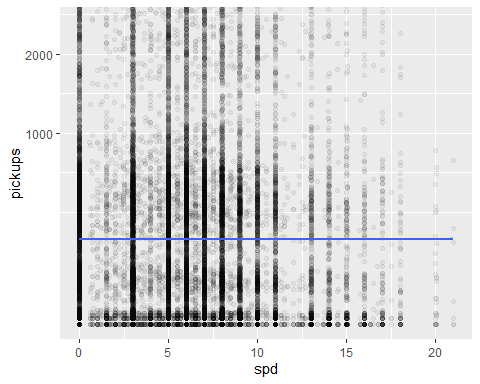
geom\_jitter(alpha = 0.05) +

geom\_smooth() +

scale\_y\_sqrt() +

coord\_cartesian(ylim = c(0, 2500))

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

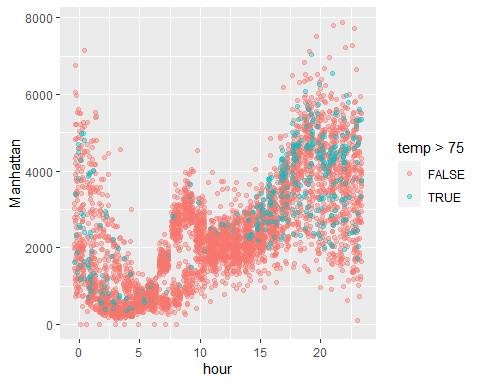


## Results

The above figures illustrated the various relationships between the main variables, that is, weather conditions, time of the day, day of the week, and month, and the count of Uber rides and pickups. To compare these variables, there was the creation of different distribution plots. Temperature and rain

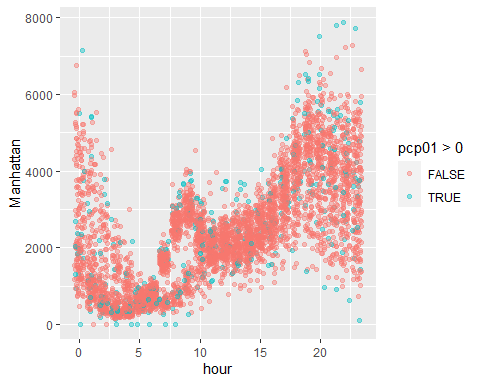
ggplot(uber.spread, aes(hour, Manhattan)) +

geom\_jitter(alpha = 0.4, aes(color = temp > 75))



ggplot(uber.spread, aes(hour, Manhattan)) +

geom\_jitter( alpha = 0.4, aes(color = pcp01 > 0))

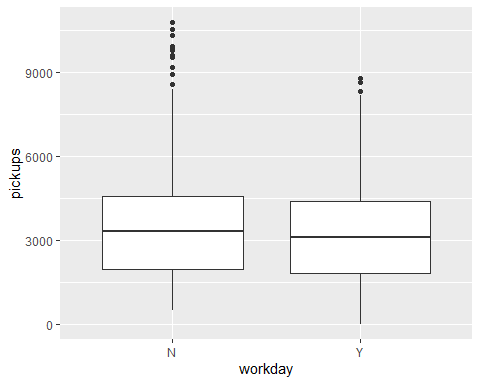


From the two graphs, it is clear that neither rain nor temperature affects Uber rides. However, there is a positive correlation between temperature and demand for Uber rides. It is clear that there are higher temperatures after 1500hrs, and this is when there is a higher demand for Uber rides.

The results are illustrated in the box-and-whisker plot in terms of working and non-working days and the demand for Uber rides.

ggplot(uber.spread, aes(workday, pickups)) +

geom\_boxplot()



The box-and-whisker plot illustrates that the Uber rides pattern is changed by the Non-working days throughout the day, but they do not significantly affect the total demand of the uber rides on a given day.

## Implication

From the results of this project, it would be recommended that further analysis should be conducted through the implementation of advanced machine learning models to provide predictions on the Uber rides dataset.

## Conclusion

Temperature and time of the day are the main contributing factors to the demand for Uber rides for different pickups. There are more Uber rides requests when there are higher temperatures during the day, especially at 1500hrs. The project has also identified that non-working days had a slightly higher demand for Uber rides than working days.

# References

Brodeur, A., & Nield, K. (2018). An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC. Journal of Economic Behavior & Organization, 152, 1-16.

Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (2016). Using big data to estimate consumer surplus: The case of uber (No. w22627). National Bureau of Economic Research.

Geitung, I. (2017). Uber drivers in Cape Town: Working conditions and worker agency in the sharing economy (Master’s thesis).

Radhika, A., & Masood, M. S. (2021). Effective dimensionality reduction by using soft computing method in data mining techniques. Soft Computing, 25(6), 4643-4651.

Rogers, B. (2015). The social costs of Uber. U. Chi. L. Rev. Dialogue, 82, 85.