Data Analysis of Uber Demand in New York

1. Introduction

The rapid growth of ride-hailing services such as Uber has transformed urban transportation, especially in bustling metropolitan cities like New York. As the city experiences millions of daily commuters, understanding the factors influencing the demand for Uber rides is crucial for both service providers and city planners. This project aims to explore and analyse the demand for Uber pickups in New York over a six-month period, using a comprehensive dataset that includes not only ride-hailing data but also weather conditions, public holidays, and borough-specific information.

The dataset, comprising over 29,000 observations across 13 variables, provides insights into a variety of factors that could impact Uber demand, such as **wind speed**, **visibility, temperature**, **precipitation**, and **snow depth**. In addition, the inclusion of borough information allows us to examine regional patterns in Uber usage across New York City's diverse neighborhoods. Furthermore, the dataset highlights whether or not a given day is a public holiday, which may have a significant impact on the number of people opting for Uber rides.

The primary objective of this project is to explore the relationships between these variables and Uber pickups. This information is not only valuable for Uber as it seeks to optimize its operations but also for city policymakers, as it can contribute to improving traffic management, urban planning, and service accessibility.

Through detailed statistical analysis and visualization techniques, this project will identify key drivers of Uber demand, offering insights that could inform future decision-making for both Uber and New York City's transportation infrastructure.

2. Content

The dataset consists of **29,101 observations** across **13 variables**. These variables capture different aspects of Uber ride demand in New York, including pickup counts, weather conditions, and temporal factors.

pickup_dt:

[The time period when Uber pickups were recorded, in the format "DD-MM-YYYY & HH:MM".]

borough:

[The New York City borough where the pickups occurred. The boroughs include areas like Manhattan, Brooklyn, and others.]

pickups:

[The number of Uber pickups during the observed period, indicating demand.]

spd:

[Wind speed (in miles per hour), which can influence the comfort of traveling.]

vsb:

[Visibility (in miles) during the observed period, important for travel safety and willingness to use Uber.]

temp:

[Temperature (in Fahrenheit), which affects people's likelihood to use transportation.]

dewp:

[Dew point (in Fahrenheit), a measure of moisture in the air that correlates with humidity.]

slp:

[Sea-level pressure, influencing weather conditions but not likely directly related to Uber demand.]

pcp01:

[Liquid precipitation in the last hour (in inches), affecting the ease of travel.]

pcp06:

[Liquid precipitation in the last 6 hours (in inches), impacting traffic and mobility.]

pcp24:

[Liquid precipitation in the last 24 hours (in inches), similarly affecting travel conditions.]

sd:

[Snow depth (in inches), which can heavily impact Uber demand during winter weather.]

hday:

[A categorical variable indicating if the day was a holiday (Y for holiday and N for non-holiday).]

3. Question

The Uber practice dataset has 29,101 observations across 13 variables with information related to:

[Uber pickups in New York for a period of 6 months]

[Weather data]

[Location Id to Borough Mapping]

[Public holidays of New York]

The objective of this practice problem is to explore the attached dataset in R and generate insights about it. On a high level, you should explore what affects the demand of Uber taxis in New York. Your exploration report can look into the following aspects :

Q1. [Summary statistics and structure of the dataset]

Q2. [Univariate analysis and their graphical analysis]

Q3. [Bivariate analysis and their graphical analysis]

Q4. [Final Insights from the dataset]

Data Dictionary :

- ⇨ pickup_dt: [Time period of the observations]
- ⇨ borough: [NYC's borough]
- ⇨ pickups: [Number of pickups for the period]
- ⇨ spd: [Wind speed in miles/hour]
- ⇨ vsb: [Visibility in Miles to nearest tenth]
- ⇨ temp: [temperature in Fahrenheit]
- ⇨ dewp: [Dew point in Fahrenheit]
- ⇨ slp: [Sea level pressure]
- ⇨ pcp01: [1-hour liquid precipitation]
- ⇨ pcp06: [6-hour liquid precipitation]
- ⇨ pcp24: [24-hour liquid precipitation]
- ⇨ sd: [Snow depth in inches]
- ⇨ hday: [Being a holiday (Y) or not (N)]

4. Objective of the Study

The objective of this study is to explore and analyse the factors affecting the demand for Uber pickups in New York City using a dataset that includes weather conditions, borough information, and public holiday data over a 6-month period. The goal is to uncover insights into how various **environmental, temporal**, and **locational factors** influence the number of Uber pickups. By performing a detailed exploratory data analysis (EDA), we aim to understand patterns in Uber demand and identify the key drivers that impact ride requests.

5. Solution

Q1. Summary Statistics and Structure of the Dataset :-

```
# [  Install Necessary Packages & Load Library's  ] :

install.packages("dplyr")
install.packages("ggplot2")
install.packages("lubridate")

library(dplyr)
library(ggplot2)
library(lubridate)

# [ Load the Dataset ] :

getwd()
uber_data <- read.csv("D:\\Project ( UBER )\\uber_nyc_enriched.csv")
uber_data

# [ Convert Pickup Date to Date-Time Format ] :

uber_data$pickup_dt <- ymd_hms(uber_data$pickup_dt)

# [ Get Structure ] :
```

str(uber_data)

OutPut =>

```
data.frame':   29101 obs. of  13 variables:
$ pickup_dt: POSIXct, format: "2001-01-20 15:01:00" "2001-01-20 15:01:00" "2001-01-20 15:01:00"
...
$ borough  : chr  "Bronx" "Brooklyn" "EWR" "Manhattan" ...
$ pickups  : int  152 1519 0 5258 405 6 4 120 1229 0 ...
$ spd      : num  5 5 5 5 5 5 3 3 3 ...
$ vsb      : num  10 10 10 10 10 10 10 10 10 10 ...
$ temp     : num  30 30 30 30 30 30 30 30 30 30 ...
$ dewp     : num  7 7 7 7 7 7 6 6 6 ...
$ slp      : num  1024 1024 1024 1024 1024 ...
$ pcp01    : num  0 0 0 0 0 0 0 0 0 0 ...
$ pcp06    : num  0 0 0 0 0 0 0 0 0 0 ...
$ pcp24    : num  0 0 0 0 0 0 0 0 0 0 ...
$ sd       : num  0 0 0 0 0 0 0 0 0 0 ...
$ hday     : chr  "Y" "Y" "Y" "Y" ...
```

[Summary of Uber Data] :

summary(uber_data)

OutPut =>

```
  pickup_dt                      borough            pickups            spd        
Min.   :2001-01-20 15:01:00.00   Length:29101       Min.   :   0.0   Min.   : 0.000  
1st Qu.:2008-04-20 15:11:00.00   Class :character   1st Qu.:   1.0   1st Qu.: 3.000  
Median :2016-01-20 15:17:00.00   Mode  :character   Median :  54.0   Median : 6.000  
Mean   :2015-11-20 11:28:16.89                      Mean   : 490.2   Mean   : 5.985  
3rd Qu.:2023-04-20 15:20:00.00                      3rd Qu.: 449.0   3rd Qu.: 8.000  
Max.   :2031-05-20 15:23:00.00                      Max.   :7883.0   Max.   :21.000  
      vsb              temp            dewp             slp            pcp01        
Min.   : 0.000   Min.   : 2.00   Min.   :-16.00   Min.   : 991.4   Min.   :0.00000  
1st Qu.: 9.100   1st Qu.:32.00   1st Qu.: 14.00   1st Qu.:1012.5   1st Qu.:0.00000  
Median :10.000   Median :46.00   Median : 30.00   Median :1018.2   Median :0.00000  
Mean   : 8.818   Mean   :47.67   Mean   : 30.82   Mean   :1017.8   Mean   :0.00383  
3rd Qu.:10.000   3rd Qu.:64.50   3rd Qu.: 50.00   3rd Qu.:1022.9   3rd Qu.:0.00000  
Max.   :10.000   Max.   :89.00   Max.   : 73.00   Max.   :1043.4   Max.   :0.28000  
     pcp06             pcp24               sd              hday          
Min.   :0.00000   Min.   :0.00000   Min.   : 0.000   Length:29101      
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.: 0.000   Class :character  
Median :0.00000   Median :0.00000   Median : 0.000   Mode  :character  
Mean   :0.02613   Mean   :0.09046   Mean   : 2.529                     
3rd Qu.:0.00000   3rd Qu.:0.05000   3rd Qu.: 2.958                     
Max.   :1.24000   Max.   :2.10000   Max.   :19.000                     
```

This will give :

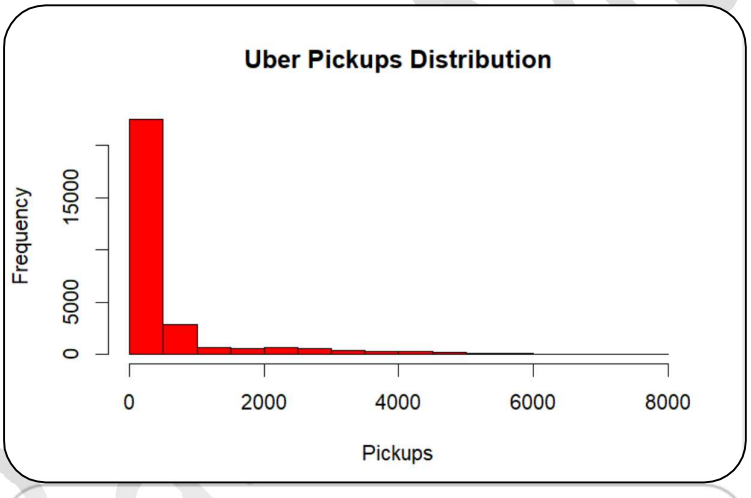
- The data types of each column (e.g., numeric, factor, character).
- Summary statistics (like mean, min, max, and NA values) for numerical columns (e.g., pickups, temp, etc.).

Q2. Univariate Analysis :-

[Plot Histogram for Uber Pickups] :

```
hist(uber_data$pickups, main="Uber Pickups Distribution",
    xlab="Pickups", col="red")
```

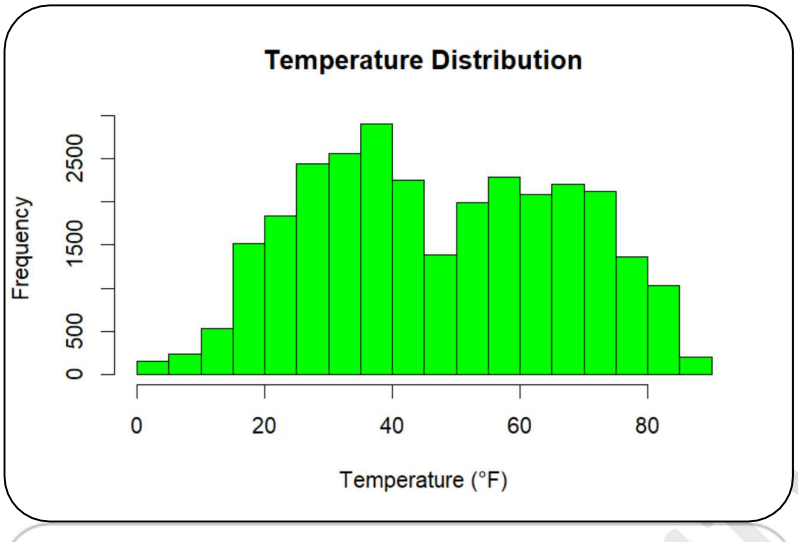
OutPut =>



[Plot histogram for temperature] :

```
hist(uber_data$temp, main="Temperature Distribution",
    xlab="Temperature (°F)", col="green")
```

OutPut =>



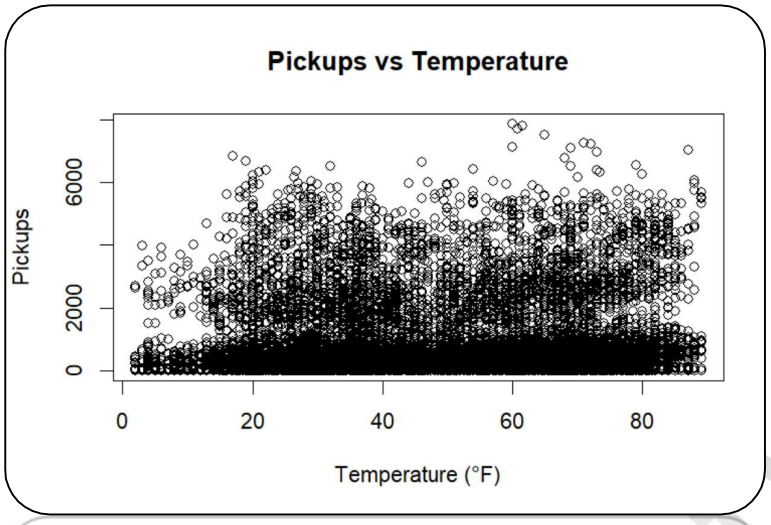
(This will help visualize the distribution of **pickups** and **temperature**.)

Q3. Bivariate Analysis :-

[Scatter plot for pickups vs temperature] :

```
plot(uber_data$temp, uber_data$pickups, main="Pickups vs Temperature",
    xlab="Temperature (°F)", ylab="Pickups")
```

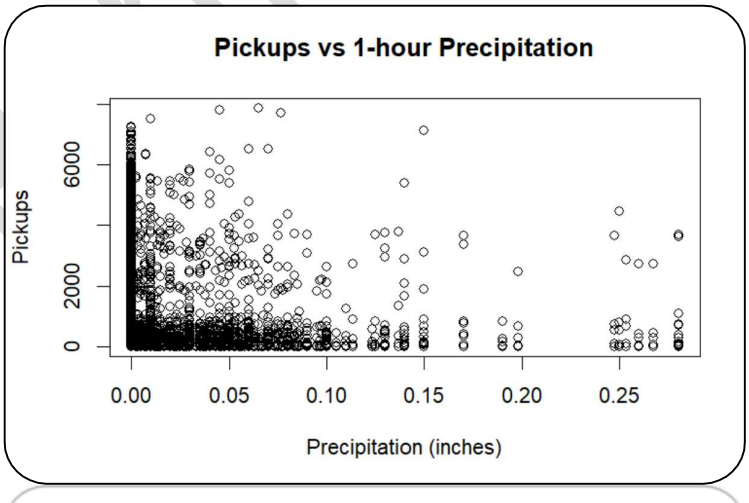
OutPut =>



[Scatter plot for pickups vs 1-hour precipitation] :

```
plot(uber_data$pcp01, uber_data$pickups, main="Pickups vs 1-hour Precipitation",
    xlab="Precipitation (inches)", ylab="Pickups")
```

OutPut =>



These scatter plots help you see if there's a relationship between **temperature** or **precipitation** and **Uber pickups**.

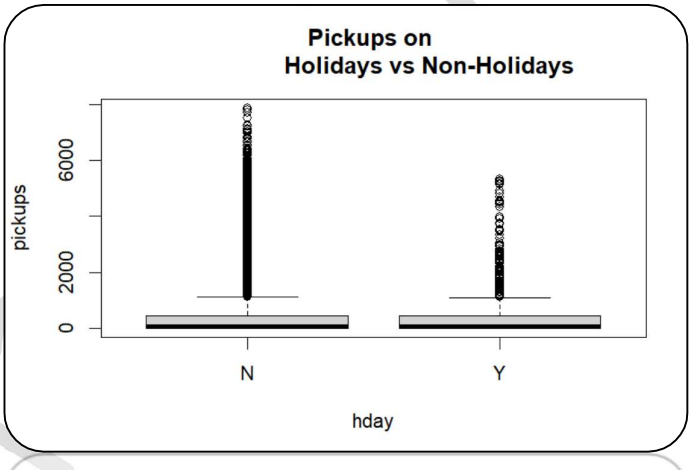
Q4. Final Insights from the Dataset :-

[Box plot to compare pickups on holidays vs non-holidays] :

```
uber_data$hday <- factor(uber_data$hday, levels = c("N", "Y"))
        boxplot(pickups ~ hday, data = uber_data, main = "Pickups on
        Holidays vs Non-Holidays")
```

Note : (N = No & Y = Yes)

OutPut =>



[Bar plot for pickups by borough] :

```
 library(ggplot2)
ggplot(uber_data, aes(x=borough, y=pickups)) + geom_bar(stat="identity",
        fill="lightblue") + theme_minimal() + labs(title="Uber
        Pickups by Borough", x="Borough", y="Total Pickups")
```

OutPut =>

Q5. Uber Pickups for a 6-Month Period :-

```
# [ Filter data for January to June ] :

uber_data$pickup_month <- format(uber_data$pickup_dt, "%b")


# [ Filter for first 6 months ] :

six_months_data <- uber_data %>% filter(pickup_month %in% c("Jan", "Feb", "Mar", "Apr",
"May", "Jun"))


# [ Calculate total pickups for the first 6 months ] :

total_pickups_6_months <- sum(six_months_data$pickups, na.rm = TRUE)


print(paste("Total Uber pickups for the 6-month period : ", total_pickups_6_months))
```

OutPut =>

```
[1] "Total Uber pickups for the 6-month period :  14265773"
```

6. Acknowledgement

I would like to express my sincere gratitude to all of my batch mates who contributed to the successful completion of this project.

First and foremost, I would like to thank my Instructor **Debjit Saha** for his valuable guidance, support, and encouragement throughout the project. Their expertise and feedback were instrumental in shaping the direction of this analysis.

I also want to acknowledge the creators and maintainers of the Uber dataset, whose efforts in collecting and making this data publicly available have provided invaluable resources for this study. This dataset served as the foundation for exploring the patterns in Uber demand and the factors influencing it.

7. Conclusion

In conclusion, this study has highlighted the importance of understanding the various factors that influence Uber ride demand in New York City. By leveraging insights on weather patterns, temporal trends, and public holidays, Uber can enhance its operational strategies, optimize resource allocation, and improve customer satisfaction. Ultimately, this study underscores the dynamic nature of Uber demand and the need for adaptive, data-driven approaches in the ride-sharing industry.