##### **Uber Insights Initiative**

##### 

UNITEDWORLD SCHOOL OF COMPUTATIONAL INTELLIGENCE (USCI)

Summative Assessment (SA)

Submitted BY

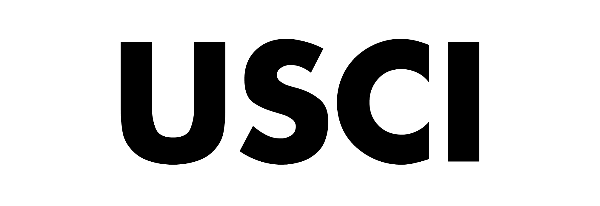
Shubh Sathwara

(Enrl. No.: 202007010158)

**Course Code and Title: 21BSCS23C02- R Programing**

B.Sc. (Hons.) Computer Science / Data Science / AIML

III Semester – July – Nov 2023



Nov/Dec 2023

The goal of the extensive data analysis project known as the Uber Visualization Project is to extract useful information from a sizable dataset of Uber ride records that covers the six-month period from April to September 2014. The analysis, visualization methods, and implications of our findings for Uber's ride-sharing service are all covered in detail in this report.

This project's main goal is to find hidden patterns and trends in the Uber ride data by applying cutting-edge data analysis and visualization techniques. By doing this, we hope to arm Uber with the information it needs to boost user satisfaction, increase operational effectiveness, and make wise strategic choices.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO** | **TITLE** | **PAGE NO** |
|  |  |  |
| **1** | **INTRODUCTION** | 4-5 |
| 1.1 System Description |  |
|  |  |
| **2** | **DEVELOPMENT PROCESS** | 6-14 |
| 2.1 Hardware & Software Requirements |  |
| 2.2 Code Implementation |  |
|  |  |  |
| **3** | **RESULTS AND CONCLUSION** | 15 |
| 3.1 Result |  |
| 3.2 Scope of future work |  |
|  |  |  |
| **4** | **APPENDICES: SAMPLE SOURCE CODE** | 16-20 |
|  |  |  |
| **5** | **REFERENCES** | 21 |
|  |  |  |
|  |  |  |
|  |  |  |

**Chapter-1**

**INTRODUCTION**

This chapter covers the project's overall introduction, the current system, and the proposed system. General provides a comprehensive overview of the project and its implementation tools. The current system's limitations are explained by the existing system, and the proposed system offers a way around them.

**1.1 SYSTEM DESCRIPTION**

With a focus on data import and data visualization, we present a thorough overview of the Uber ride analysis system's data in this section.

**Data Import and Preparation**

A number of crucial R libraries, such as "ggplot2", "lubridate", "dplyr,", "ggthemes", "tidyr", "DT” and “scales”, are the foundation of our system. Each one has a distinct function in the analysis and visualization process. We can produce intelligent visualizations from the Uber ride data thanks to these libraries.

**Data Import**

In order to begin our analysis, we imported CSV files containing Uber ride data for several months in 2014, from April to September. After that, a number of transformation and preprocessing procedures were applied to this raw data in order to prepare it for analysis and visualization.

**Data Transformation**

Important changes were made to the data, with particular attention to the date and time fields. We ensured data uniformity by converting these fields into the proper formats. We also extracted the components for day, month, year, hour, minute, and second, which allowed us to obtain a deeper understanding of temporal patterns.

**Data Visualization**

Our system incorporates a range of visualizations to uncover patterns and trends within the Uber ride data:

**1. Trips by Hour in a Day:** This graphic representation provides insights into the daily ride patterns by displaying the distribution of journeys over the course of a day.

**2. Trips by Hour and Month:** Through the use of hourly and monthly trip counts, we are able to see how riding patterns vary throughout different time periods.

**3. Trips Every Day:** The daily ride counts are shown in this image, which enables us to spot patterns and variances in ride volume.

**4. Trips by Day and Month:** It draws attention to the connection between the day of the month and the month itself, exposing trends that may change from one month to the next.

**5. Number of Trips by Months:** This graphic shows the number of journeys that take place in six month from a yearly viewpoint.

**8. Heatmaps:** A useful tool for visualizing data patterns is a heatmap. In order to depict ride patterns by the hour, day, month, and day of the week, we use heatmaps.

**9. Map Visualization:** We can gain spatial insights and understand where most rides take place in New York by looking at a geographical visualization of ride locations.

We can learn a great deal about Uber ride patterns and trends thanks to this extensive system of data analysis and visualization, which we will go into more detail about in this report.

**Chapter-2**

**DEVELOPMENT PROCESS**

One of the steps in the software development life cycle is the development process and documentation. Analysis of the project's requirements, design, and implementation are all part of the development process.

**2.1 Hardware & Software Requirements**

The models listed below represent the minimal requirements needed to complete the project.

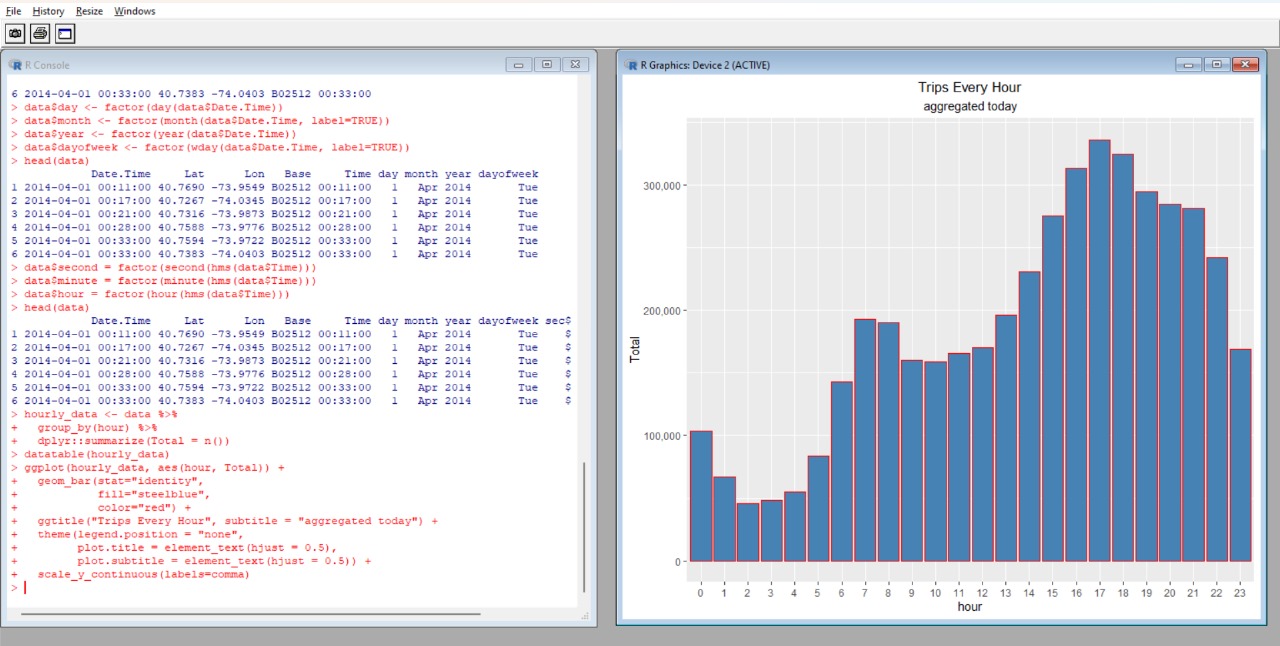
**Software Requirements**: -

* Operating System: Windows 7
* Technology: R Programming
* Web Technologies: R Studio 4.3.1

**Hardware Requirements: -**

* Processor: Intel i3
* Hard Disk: 500 GB
* RAM: 2 GB
* Cache Memory: 512KB
* Operating System: Windows XP/ Windows7

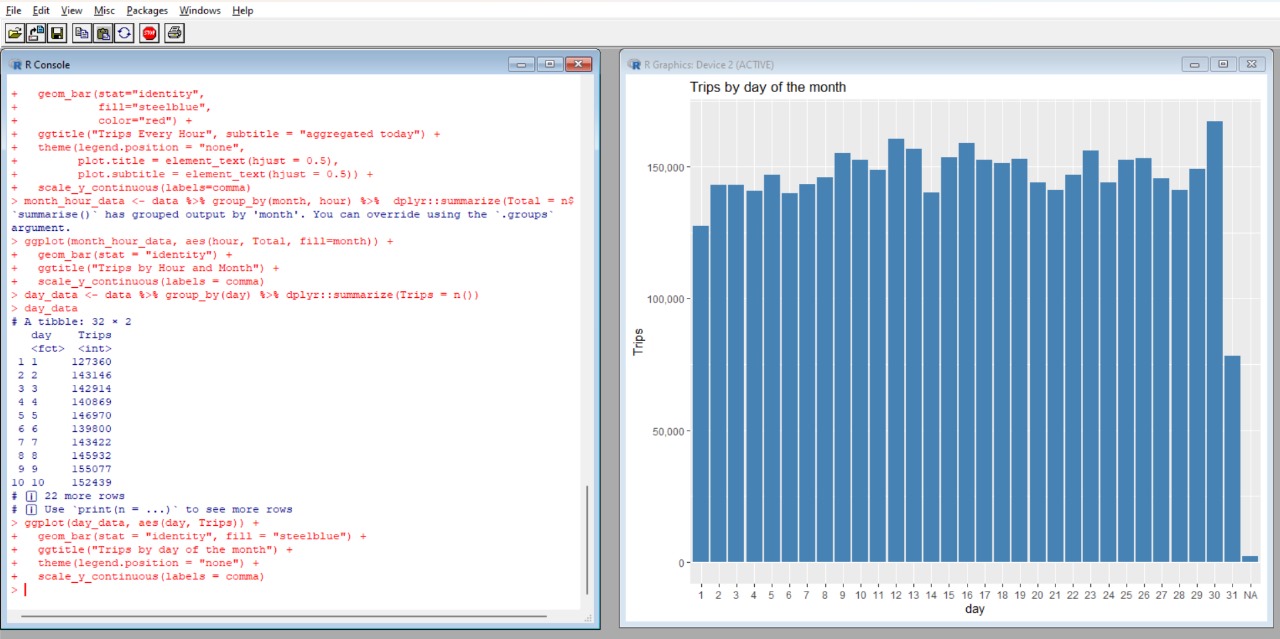
**2.2 Code Implementation**



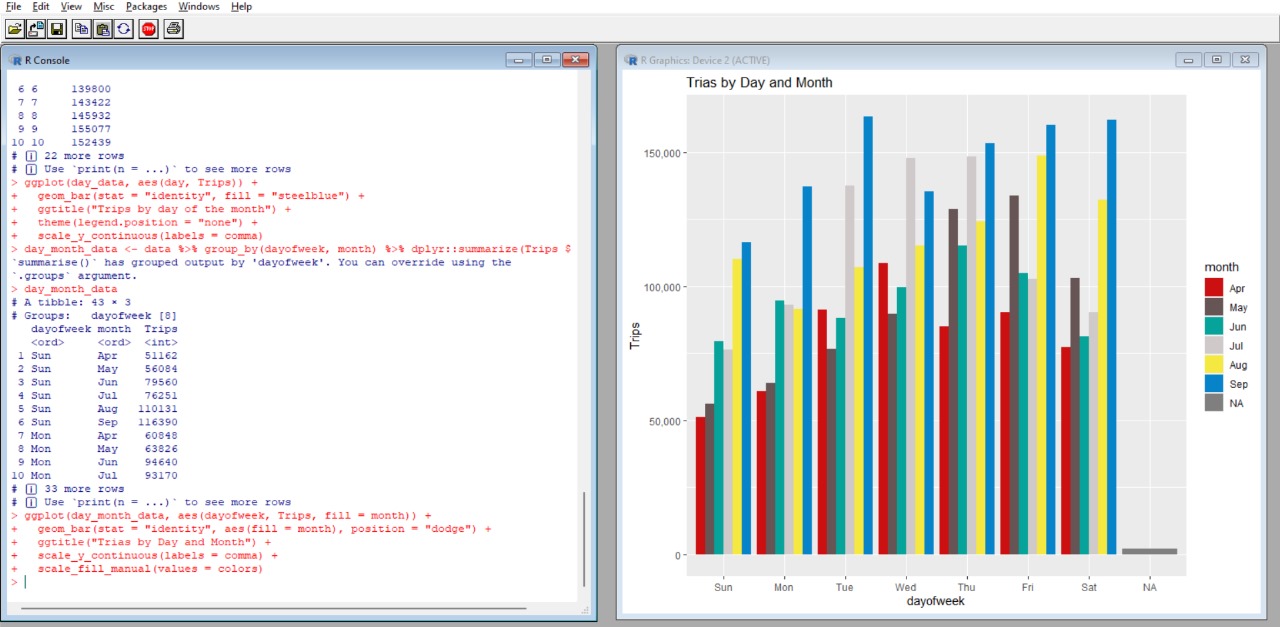
**Figure 1: Bar Graph Representing Trips made in an Hour in a Day**



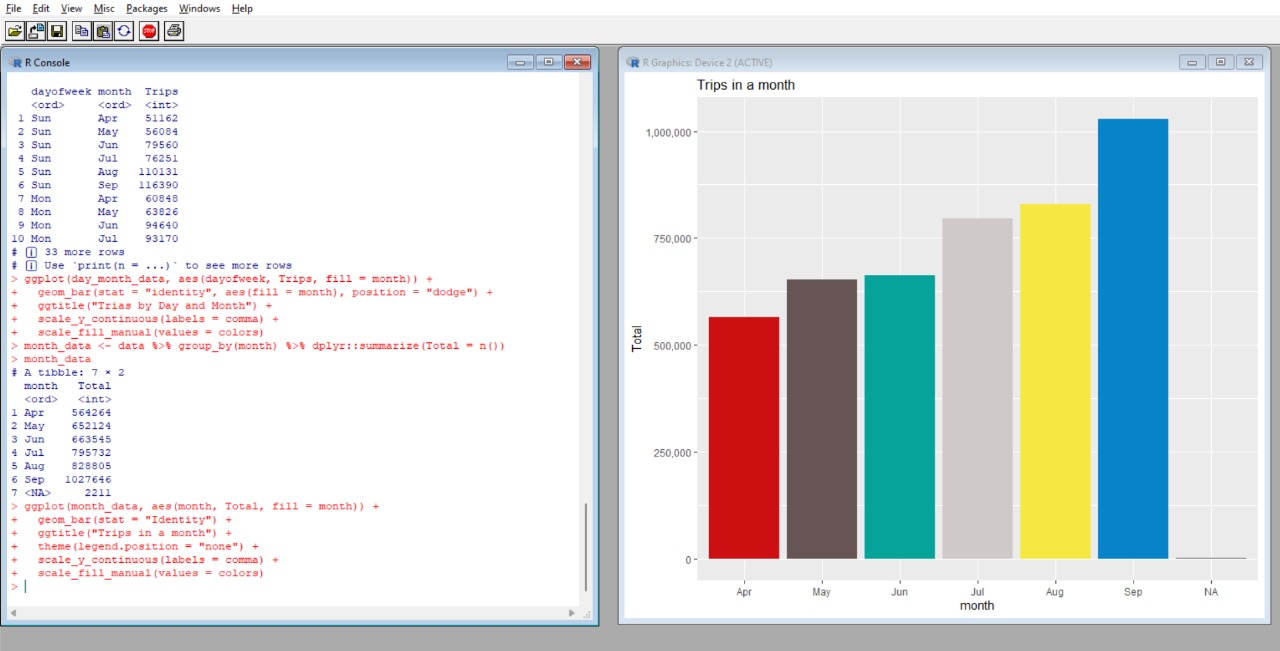
**Figure 2: Bar Graph Representing Trips made per Hour in a Day filtered by Months**



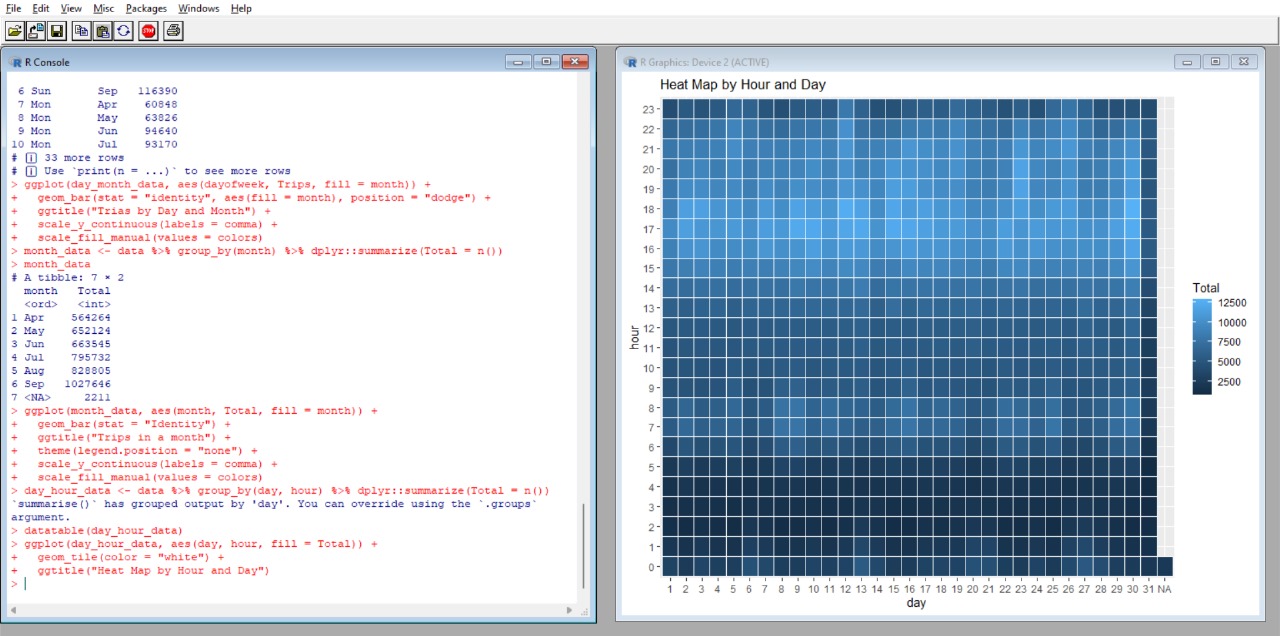
**Figure 3: Bar Graph Representing Trips made per Day of a Month**



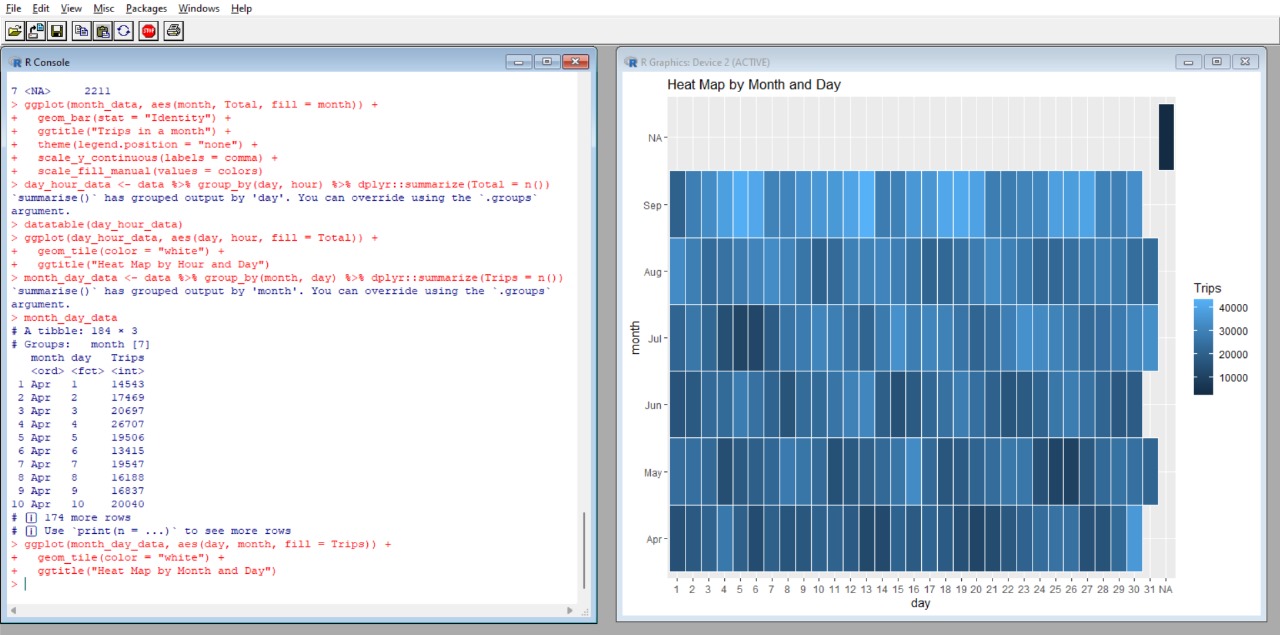
**Figure 4: Bar Graph Representing Trips made per Day of Week filtered by Months**



**Figure 5: Bar Graph Representing Trips made per Months**



**Figure 6: Heat Map by Hour and Day**



**Figure 7: Heat Map by Months and Day**



**Figure 8: Heat Map by Month and Day**

**Chapter-3**

**RESULTS & CONCLUSION**

**3.1 Result**

These are the final points that we have deduced from this whole project.

1. **Hourly Ride Patterns:** An important finding of our analysis was the hourly ride patterns. Peaks are displayed in the visualizations at specific times of the day, which can be related to things like rush hours and recreational pursuits.
2. **Monthly Variations:** Monthly fluctuations in ride patterns were noted, with certain months recording higher ride volumes than others. Uber may find this information useful for planning and allocating resources.
3. **Daily Ride Trends:** Some days of the month seem to draw more riders than others, according to our analysis of daily ride trends. These insights can help Uber plan its marketing campaigns or allocate resources more efficiently on certain days.
4. **Heatmaps:** Heatmaps made it easier to identify trends and variations by visually representing the intensity of ride patterns by hour, month, and day of the week.

**3.2 Scope of Future Work**

These are the key points to be noted, this was the main purpose of the whole project.

1. **Predictive Modelling:** Create models to predict ride demand so that resources can be distributed.
2. **Real-time Analytics:** Use real-time data analysis to react quickly to shifts in demand.
3. **Customer Segmentation:** Riders can be categorized for individualized services and promotions.
4. **External Data Integration:** Add outside data sources to your analysis, such as the weather.
5. **Advanced Visualization:** Examine interactive dashboards to investigate data in real time.
6. **Machine Learning:** Utilize machine learning for demand forecasting, route optimization, and anomaly detection.
7. **Feedback Analysis:** Examine customer feedback sentiment to enhance services.
8. **Cost Optimization:** Find the sources of costs to improve operational effectiveness.
9. **Sustainability Initiatives:** Encourage eco-friendly modes of transportation.
10. **Market Expansion:** Make wise decisions about expansion using data insights.

**Chapter-4**

**APPENDICES**

**Sample Coding:**

library(ggplot2)

library(ggthemes)

library(lubridate)

library(dplyr)

library(tidyr)

library(DT)

library(scales)

**# Creating vector of colors for the plots**

colors = c("#CC1011", "#665555", "#05a399", "#cfcaca", "#f5e840", "#0683c9", "#e075b0")

getwd()

**# [1] "C:\Users\DELL\Desktop\Uber data analysis"**

**# Read the data for each month separately**

apr <- read.csv("C:/Users/DELL/Desktop/Uber data analysis/Uber dataset/uber-raw-data-apr14.csv")

may <- read.csv("C:/Users/DELL/Desktop/Uber data analysis/Uber dataset/uber-raw-data-may14.csv")

june <- read.csv("C:/Users/DELL/Desktop/Uber data analysis/Uber dataset/uber-raw-data-jun14.csv")

july <- read.csv("C:/Users/DELL/Desktop/Uber data analysis/Uber dataset/uber-raw-data-jul14.csv")

aug <- read.csv("C:/Users/DELL/Desktop/Uber data analysis/Uber dataset/uber-raw-data-aug14.csv")

sept <- read.csv("C:/Users/DELL/Desktop/Uber data analysis/Uber dataset/uber-raw-data-sep14.csv")

data <- rbind(apr, may, june, july, aug, sept)

cat("The dimensions of the data are:", dim(4534327 4))

head(data)

**# Readable format for the DateTime**

data$Date.Time <- as.POSIXct(data$Date.Time, format="%m/%d/%Y %H:%M:%S")

data$Time <- format(as.POSIXct(data$Date.Time, format = "%m/%d/%Y %H:%M:%S"), format="%H:%M:%S")

data$Date.Time <- ymd\_hms(data$Date.Time)

head(data)

**# Create individual columns for month day and year**

data$day <- factor(day(data$Date.Time))

data$month <- factor(month(data$Date.Time, label=TRUE))

data$year <- factor(year(data$Date.Time))

data$dayofweek <- factor(wday(data$Date.Time, label=TRUE))

head(data)

**# Add Time variables**

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

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

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

head(data)

**## Data Visualisation**

**# Plotting the trips by hours in a day**

hourly\_data <- data %>%

group\_by(hour) %>%

dplyr::summarize(Total = n())

datatable(hourly\_data)

**# Plot the data by hour**

ggplot(hourly\_data, aes(hour, Total)) +

geom\_bar(stat="identity",

fill="steelblue",

color="red") +

ggtitle("Trips Every Hour", 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)

**# Plotting trips by hour and month**

**# Aggregate the data by month and hour**

month\_hour\_data <- data %>% group\_by(month, hour) %>% dplyr::summarize(Total = n())

ggplot(month\_hour\_data, aes(hour, Total, fill=month)) +

geom\_bar(stat = "identity") +

ggtitle("Trips by Hour and Month") +

scale\_y\_continuous(labels = comma)

**# Plotting data by trips during every day of the month**

day\_data <- data %>% group\_by(day) %>% dplyr::summarize(Trips = n())

day\_data

**# Plot the data for the day**

ggplot(day\_data, aes(day, Trips)) +

geom\_bar(stat = "identity", fill = "steelblue") +

ggtitle("Trips by day of the month") +

theme(legend.position = "none") +

scale\_y\_continuous(labels = comma)

**# Collect data by day of the week and month**

day\_month\_data <- data %>% group\_by(dayofweek, month) %>% dplyr::summarize(Trips = n())

day\_month\_data

**# Plot the Collect data by day of the week and month**

ggplot(day\_month\_data, aes(dayofweek, Trips, fill = month)) +

geom\_bar(stat = "identity", aes(fill = month), position = "dodge") +

ggtitle("Trias by Day and Month") +

scale\_y\_continuous(labels = comma) +

scale\_fill\_manual(values = colors)

**# Number of Trips place during months in a year**

month\_data <- data %>% group\_by(month) %>% dplyr::summarize(Total = n())

month\_data

**# Plot the Number of Trips place during months in a year**

ggplot(month\_data, aes(month, Total, fill = month)) +

geom\_bar(stat = "Identity") +

ggtitle("Trips in a month") +

theme(legend.position = "none") +

scale\_y\_continuous(labels = comma) +

scale\_fill\_manual(values = colors)

**# Heatmap by Hour and Day**

day\_hour\_data <- data %>% group\_by(day, hour) %>% dplyr::summarize(Total = n())

datatable(day\_hour\_data)

**# Plot a Heatmap by Hour and Day**

ggplot(day\_hour\_data, aes(day, hour, fill = Total)) +

geom\_tile(color = "white") +

ggtitle("Heat Map by Hour and Day")

**# Plot Heatmap by Day and Month**

month\_day\_data <- data %>% group\_by(month, day) %>% dplyr::summarize(Trips = n())

month\_day\_data

**# Plot a heatmap by Collect data by month and day**

ggplot(month\_day\_data, aes(day, month, fill = Trips)) +

geom\_tile(color = "white") +

ggtitle("Heat Map by Month and Day")

**# Plot a heatmap by day of the week and month**

ggplot(day\_month\_data, aes(dayofweek, month, fill = Trips)) +

geom\_tile(color = "white") +

ggtitle("Heat Map by Month and Day")

**Chapter-5**

**REFERENCES**

1. **Data Set:** https://www.kaggle.com/code/amirmotefaker/uber-data-analysis-using-r
2. **Software:** R Studios
3. **Language Used:** R programming
4. **AI Used:** [https://chat.openai.com](https://chat.openai.com/) & https://quillbot.com
5. Guidance of Ms. Deepti Amita