**A MINI PROJECT on**

**“Uber Analysis”**

Submitted to

Mr. Bheemisetty Venkata Sivaiah Asst. Prof., CSE

**BACHELOR OF TECHNOLOGY**

**IN**

**INFORMATION TECHNOLOGY**

|  |  |
| --- | --- |
| **B Pallavi** | **21121A3530** |
| **S** **Bindu** | **21121A3544** |
| **T Kushal** | **21121A3545** |
| **R Raju**  **K Mallikarjuna** | **21121A3539**  **21121A3516** |
|  |  |
|  |  |
|  |  |
|  |  |

**Submitted by**



Department of Information Technology

## SREEVIDYANIKETHANENGINEERINGCOLLEGE

(AUTONOMOUS)

(Affiliated to JNTUA, Anantapur, Approved by AICTE, Accredited by NBA &

NAAC) Sree Sainath Nagar, Tirupati – 517 102, A.P., INDIA

2023-2024

## 

## TABLE OF CONTENTS

1. ABSTRACT ....................................................................... 3
2. KEYWORDS ...................................................................... 3
3. INTRODUCTION................................................................ 3
4. FLOWCHART………………………………………………………………… 4
5. DATASET………………………………………………………………………. 5
6. ABOUT R STUDIO…………………………………………………………. 5
7. ABOUT TABLEAU SOFTWARE……………………………………….. 5
8. LIBRARIES USED…………………………………………………………… 6
9. ALGORITHM…………………………………………………………………. 6
10. RESULTS…………………………………………………………….. 7
11. CONCLUSION…………………………………………………….. 12
12. REFERENCES…………………………………………............... 12

**Abstract**— Data analytics has helped companies optimize and grow their performance for decades. Data analytics and visualization have aided us with several benefits, a few of them being identifying emerging trends, studying relationships and patterns in data, analysis in-depth, and cherry on top are the insights we draw from these patterns. It is a requirement of time that we study this concept thoroughly for all the benefits it provides. This project is all about understanding one such data set of Uber from New York City and is very component to understanding the use of data analytics and visualization. It is generated with the help of ‘R’ programming language using libraries such as ggplot2, lubridate, dplyr and tidyr. Through projects like this, we can gain knowledge of various complex operations performed in data visualization. It will enable us to recognize the patterns in the data of this huge organization and provide critical insights into untapped information. Also, guides us in understanding the operations of ggplot2 library.

**Keywords**— Uber, Data analytics, Data visualization, R programming, ggplot, lubridate, dplyr.

1. **INTRODUCTION**

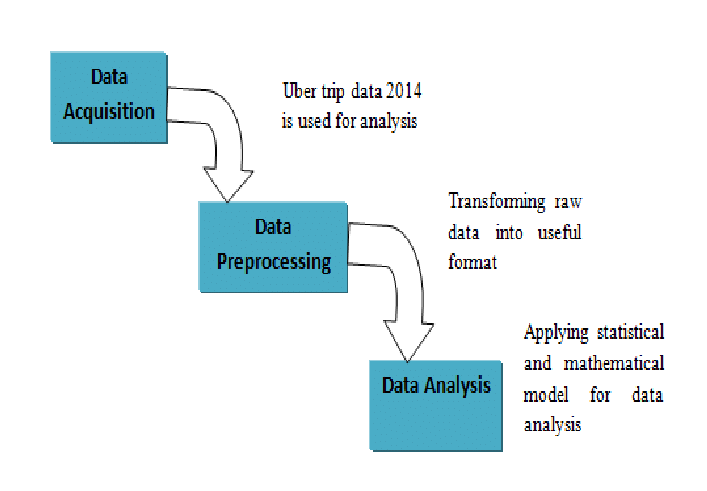
Uber has emerged as the leading company in the provision of new transportation options within the contemporary world. Uber, then, is primarily in the business of networking, and all the company's emerging operations can be conceptualized in terms of simply providing a medium through which the relevant supply can meet up with the relevant demand. Analytics is a tremendously growing niche that people apply in their businesses to give it a boost. This is more of a data visualization project that will enhance our knowledge towards using the ggplot2 library for understanding the data and for developing an intuition for understanding the customers who avail the trips Solution to this issue is understanding what Customer segmentation (aka Market Segmentation). Customer segmentation can be explained as a game where a kid separates balls and cubes based on their shape or colors. In simple language, customer segmentation is segregating customers, and markets on different criteria and dividing them based on various characteristics. The Uber data is not as detailed as the taxi data, in particular, Uber provides time and location for pickups only, not drop-offs, but I wanted to provide an amalgamated dataset including all available taxi and Uber data. Uber analyzes historical data for say, the last 3 or 4 weeks and identifies pockets within the city that witness extremely high demand.

To grow business in this competitive environment data analysis is necessary. Data analysis reports, and other kinds of analysis and report documents must be developed by businesses so that they can have references for peculiar activities and undertakings especially when making decisions for the future operations of the company. The Excel files with the weather data and Uber pick-up data should be joined together for the analysis. A data analysis can be developed accordingly if you can arrange all the information based on the activity that you will undergo. In the Uber data analysis R project, we observed how to create data visualizations. Uber is the only mobility company to assess and publish real-world sustainability data.

In this R project, our objective is to analyze the Uber Pickups in New York City dataset. This is more of a data visualization project that will guide us towards using the ggplot2 library for understanding the data and developing an intuition to understand the customers who avail the trips. Our main objectives are:

* Visualize Uber's growth in NYC
* Characterize the demand based on identified patterns in the time series
* Estimate the value of the NYC market for Uber
* Other insights about the usage of the service
* Attempt to predict the demand's growth

1. **Flowchart**



**The diagram above represents the method flow of this project.**

1. **DATASET**

The dataset contains information about Uber pickups in New York City from April 2014 to September 2014. It has over 500k pickups (rows) and the following 4 columns:

* Date/Time - The date and time of pickup
* Lat - Latitude of pickup
* Long - Longitude of pickup Base

1. **ABOUT R STUDIO**

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax highlighting editor that supports direct code execution, as well as tools for plotting, history, and debugging, and workspace management.

1. **ABOUT TABLEAU SOFTWARE**

Tableau is a platform used to visualize analytical data, transforming the whole narrative of using data for gaining insights. It has empowered data enthusiasts and organizations to make the most of the analytical data available.

1. **LIBRARIES USED**

* Ggplot2 - it is the main part of the project, and it is used widely to create aesthetic visualization plots.
* Ggthemes – it is a library for many themes from which the user can get the desired scale for their database.
* Lubridate – it consists of time frames, and it should be in separate time categories.
* Tidyr – This function will classify the huge data into many columns and rows which will make it easier to manipulate it.

1. **ALGORITHM**

1. Importing the Essential Packages

2. Creating vector of colors to be implemented in our plots

3. Reading the Data into their designated variables

4. Plotting the trips by the hours of a day

5. Plotting data by trips during every day of the month

6. Finding the number of Trips taking place during months in a year

7. Finding out the number of Trips by bases

8. Creating a Heatmap visualization of day, hour, and month

9. Creating a map visualization of rides in New York

10. Collecting insights from all visualizations

1. **PROPOSED SYSTEM**

We proposed that we will build a data visualization project with ggplot2 using R and its libraries. Analyze various parameters like (a) Trips by the hours in a day and (b) Trips during months in a year. At the end create visualizations for different timeframes of the year. Explain how time affects customer trips.

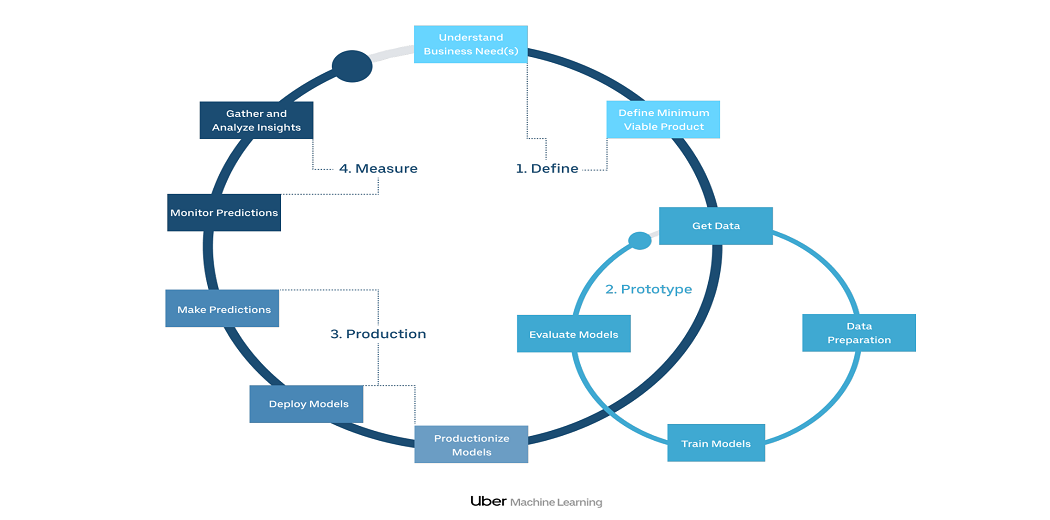
• Customers are often dissatisfied with traditional cab companies because of their high prices and long stand-by time and hence can exploit new and big markets.

• Find the days on which each basement has a greater number of active vehicles.

• Can tap growing markets in suburban areas where taxi services are not available.

• Estimated Time of Arrival can be reduced with an increase in the number of Uber drivers which successively will make Uber more liked by the customers and hence, the company will get more revenue and drivers will also be profited.

Based on the data, we will find the destination people travel the most that generate high airline revenues for travel, based on booked trip count.



**System of working Uber Analysis**

As the diagram suggests the proposed system will involve a beginning with extraction of data. Here we used the data set of New York City from April 2014 to September 2014. Cleaning the data and classification has a high level of importance for more appropriate results. Later on, we enter the stage of clustering and analysis wherein we will use ggplot2 library in R studio for reading the data in related variables and plotting trips across different parameters such as week, month, etc. The final stage consists of the Visualization of obtained results, which we did using Tableau Software. It helps us clearly understand the information in the form of graphs. Giving us desired insights.

1. **RESULTS**

At the end of all procedure, we get to see different graphs giving us unbelievable insights. We have plotted different graphs as mentioned below:

* 1. **DATA VISUALISATION**

1. **plotting trips every hour**

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

geom\_bar(stat="identity",

fill="steelblue",

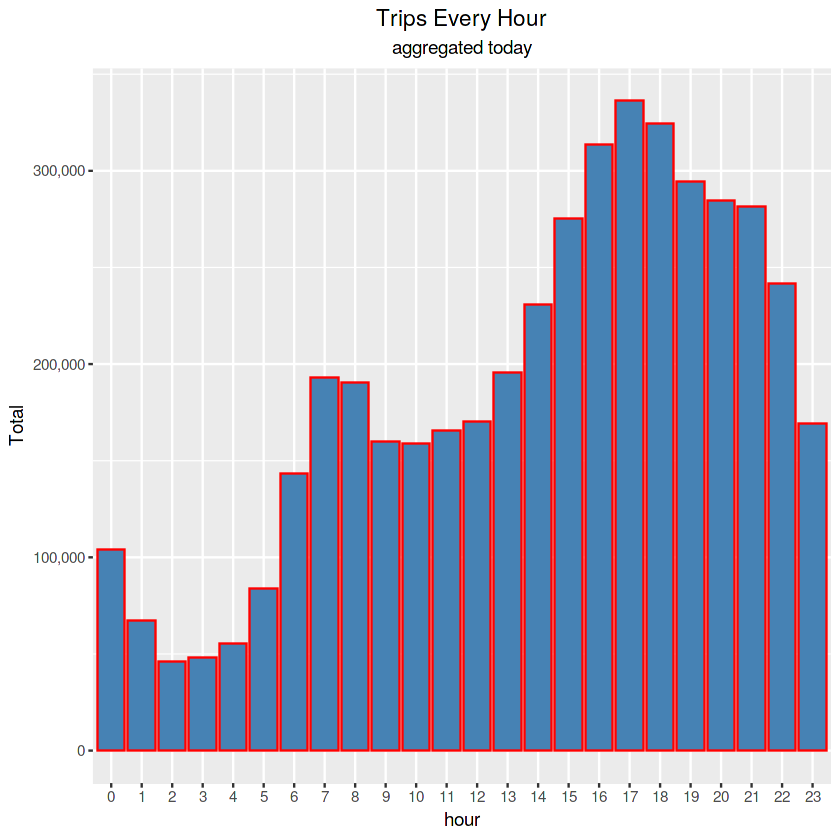
color="red") +

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)



1. **Plotting trips by hour and month**

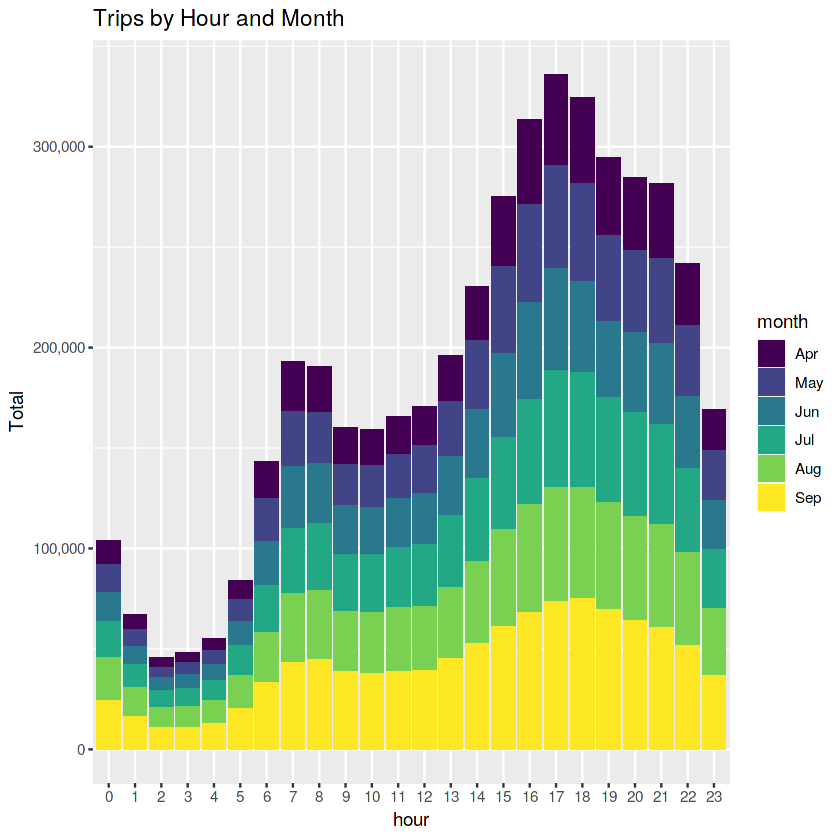
month\_hour\_data <- data %>% group\_by(month, hour) %>% dplyr::summarize(Total=n())

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

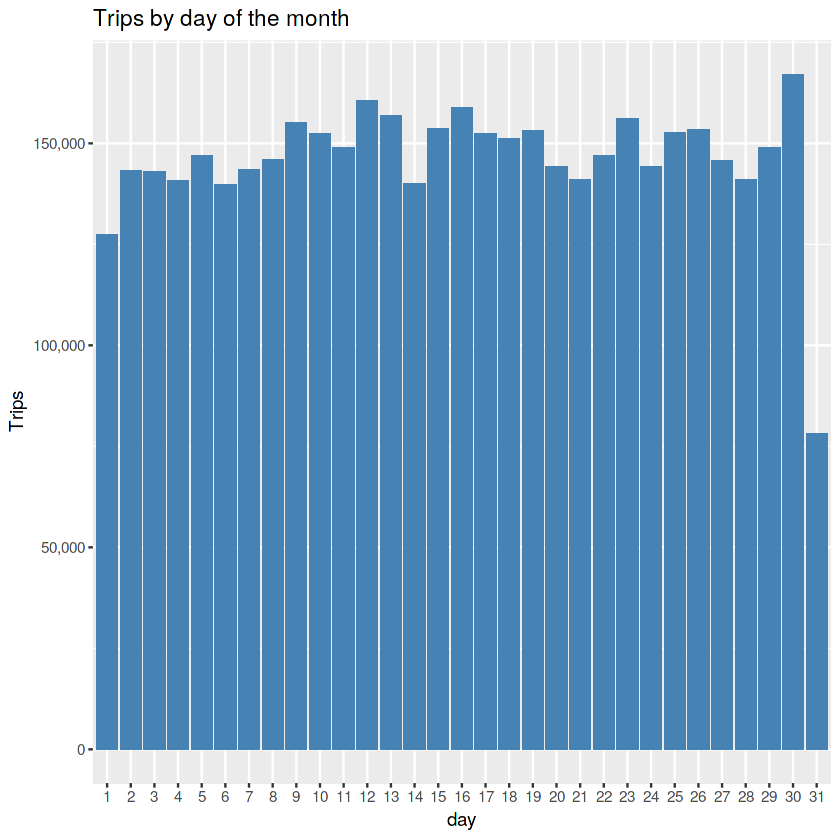
## ggplot(day\_data, aes(day, Trips)) +

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

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

## theme(legend.position = "none") +

## scale\_y\_continuous(labels = comma)



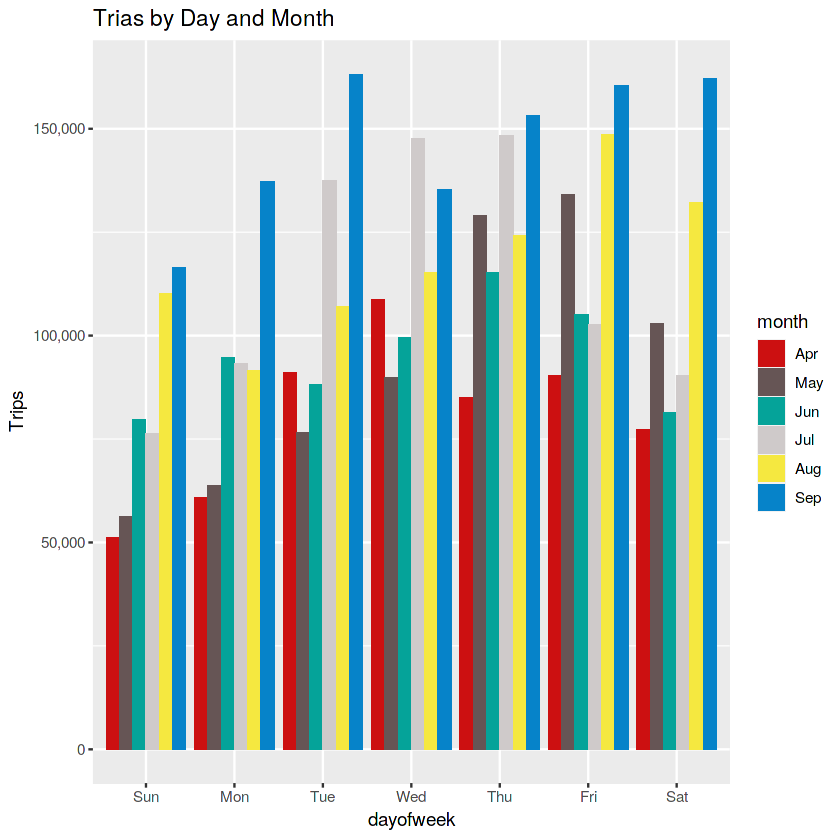
ggplot(day\_month\_data, aes(dayofweek, Trips, fill = month)) +

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)



## **Creating a map visualization of rides in NYC:**

## min\_lat <- 40

## max\_lat <- 40.91

## min\_long <- -74.15

## max\_long <- -73.7004

## ggplot(data, aes(x=Lon, y=Lat)) +

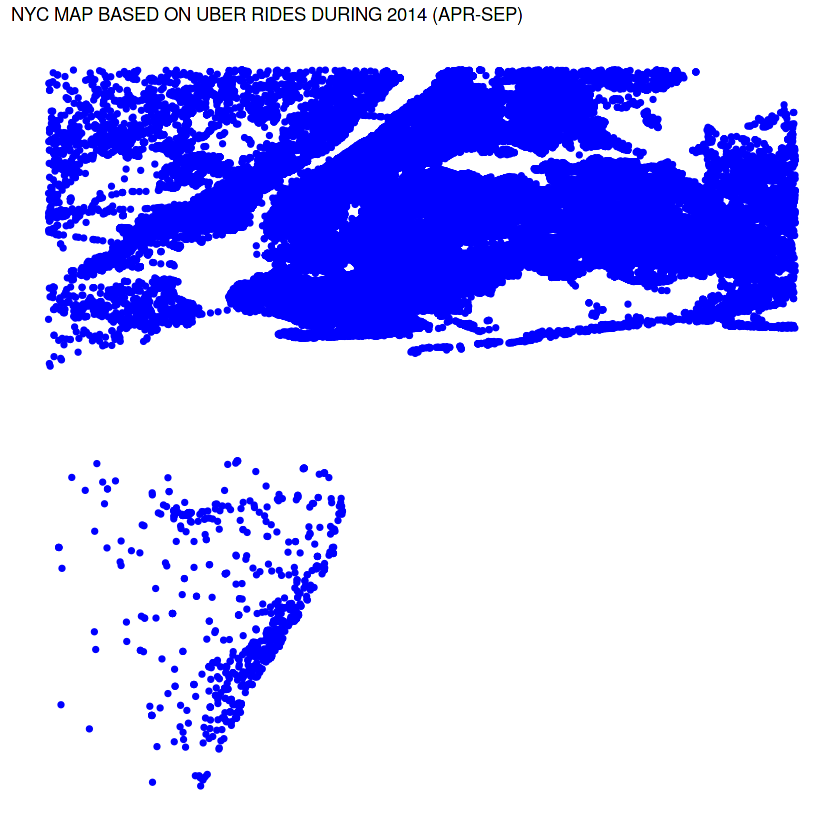
## geom\_point(size=1, color = "blue") +

## scale\_x\_continuous(limits=c(min\_long, max\_long)) +

## scale\_y\_continuous(limits=c(min\_lat, max\_lat)) +

## theme\_map() +

## ggtitle("NYC MAP BASED ON UBER RIDES DURING 2014 (APR-SEP)")



1. **CONCLUSION**

At the end of this Uber data analysis R project, we studied how to create data visualizations. We used package ggplot2 which helped us to plot various types of visualizations that pertained to several time-frames of the year. With this, we conclude how time and place affected customer trips.

Finally, we made visualization a Geo plot of New York that provided us with the details of how various users made trips from different bases.

1. **FUTURE SCOPE**

We can use this data for training a model using ML and building a smart AI-based predictive system. The model can automatically send insights to the authorities or drivers related to areas having the most trips and passenger count in certain areas. This big data can be used to study passenger’s behavior.

1. **REFERENCES**

[1] https://ggplot2.tidyverse.org/

[2]https://rpubs.com/Unsa/582359

[3]https://www.skyfilabs.com/project-ideas/uber data-analysis [4][https://www.uber.com/us/en/careers/teams/data science/](https://www.uber.com/us/en/careers/teams/data%20science/)

[5]https://iedu.us/tag/project-in-r-uber-data analysis-project/ [6]https://growvation.com/paritosh sankhla/project/uber.in