Journal of Engineering Sciences
Vol 12, Issue 7, July/2021
ISSN NO:0377-9254

UBER DATA ANALYSIS USING GGPLOT

Mrunal Patil[1], Vidya Kumari[2], Adarsh Patil[3], Laxmikant Ahire[4] and
Asst.Prof. Umakant Mandawkar[5]

[1,2,3,4] *B.Tech, Computer Science and Engineering, Sandip University, Nashik, India*

[5] *Asst Prof., Computer Science and Engineering, Sandip University, Nashik, India*

***Abstract*— Data analytics has helped companies optimize and grow their performance for decades. Data analytics and visualization has aided us with several benefits, 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 requirement of time that we study this concepts in thoroughly for all this benefits it provides. This project is all about understanding one such data set of uber from New York City and is very component to understand 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 data of this huge organization and provides critical insights of untapped information. Also guide us in understanding the operations of ggplot2 library.**

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

1. INTRODUCTION

Uber has emerged as 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, cubes based on their shape or colors.

In simple language customer segmentation is segregating customers, market on different criteria and dividing them on the basis of various characteristics. The Uber data is not as detailed as the taxi data, in peculiar Uber provides time and location for pickups only, not drop-offs, but I wanted to provide an amalgamate dataset including all available taxi and Uber data. Uber analyze historical data for say, last 3 or 4 weeks and identifies pockets within the city that witness extremely high demand.

To grow business with 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. 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

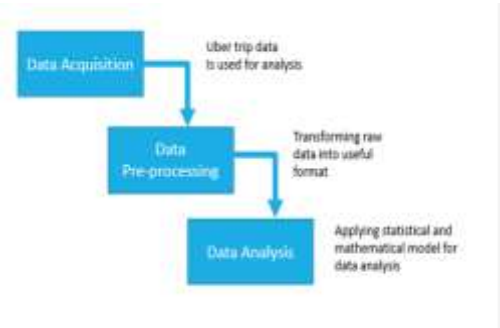
2. LITERATURE REVIEW

Two papers of related works are of interest to same projects – those related to case studies of uber data in different cities and ones related to urban transport development.

Ggplot2 is now over 10 years old and is used by 100's of 1000's of people to make millions of plots. It is an R package dedicated to data visualization. It can greatly enhance the quality and aesthetics of your graphics, and will make you much more efficient in creating them. Ggplot2 allows building almost any type of chart. It is a system for declarative creating graphics, based on the grammar of graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

3. METHOD AND MATERIAL

3.1. FLOWCHART



The diagram above represents the method flow of this project.

3.2. 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

3.3. 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.

3.4. 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 most of the analytical data available.

3.5. 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.

3.6. ALGORITHM

1. Importing the Essential Packages
2. Creating vector of colours to be implemented in our plots
3. Reading the Data into their designated variables
4. Plotting the trips by the hours in 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

4. PROPOSED SYSTEM

We proposed that we will build a data visualization project with ggplot2 using R and its libraries.
Analyse various parameters like (a) Trips by the hours in a day (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 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.

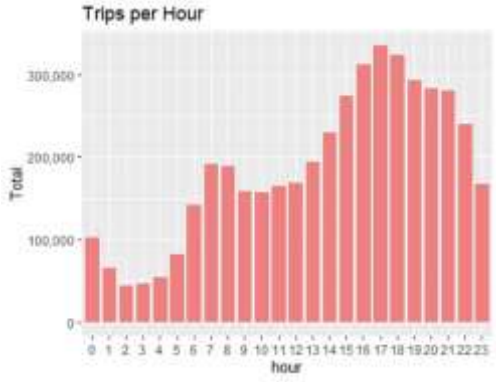
Based on the data, we will find the destination people travel the most that generate high airline revenues for travel, formed on booked trip count.

As the diagram suggest the proposed system will involve a beginning with extraction of data. Here we used the data set of New York City from month of April 2014 to September 2014. Cleaning the data and classification has 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. Final stage consists of Visualization of obtained results, which we did using Tableau Software. It helps us clearly understand the information in form of graphs. Giving us desired insights.

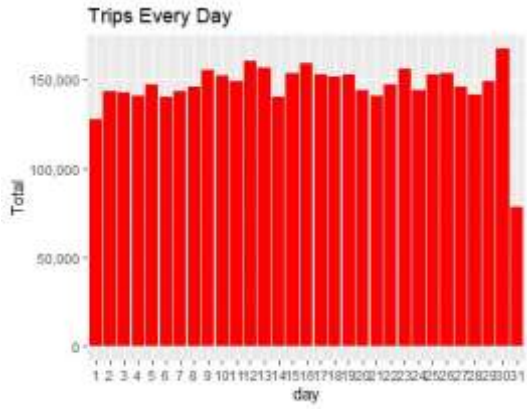
5. V. RESULTS

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

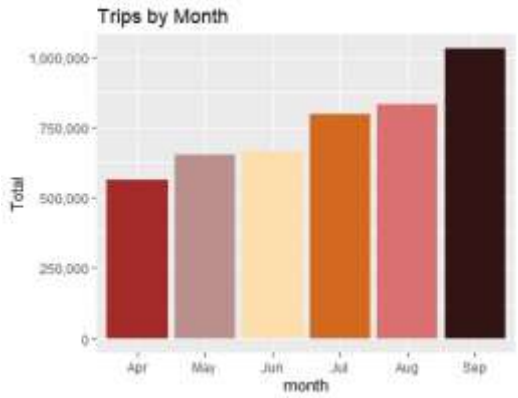
1.Trips by the hours in a day:



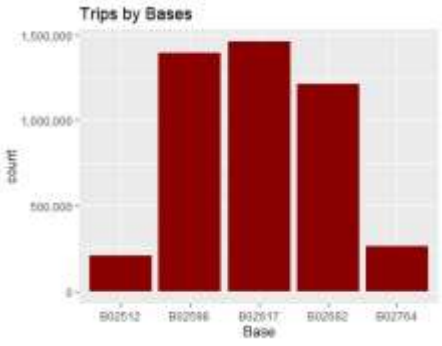
2. Trips during every day of the month:



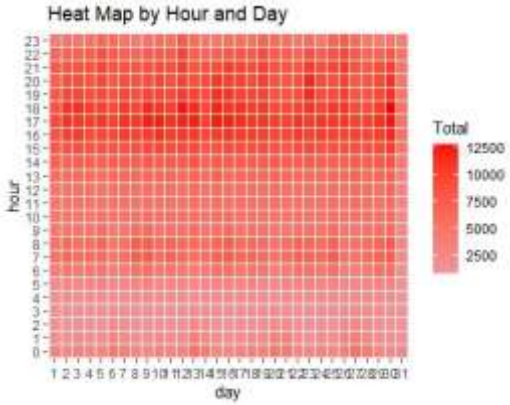
3. Trips taking place during months in a year:



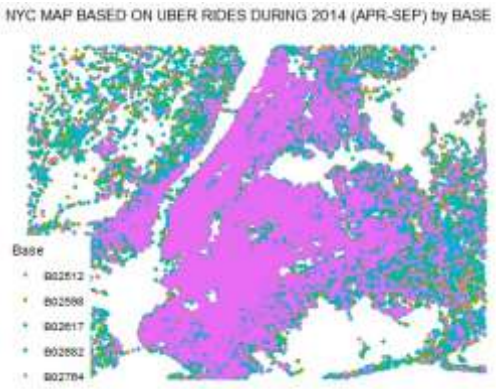
4. Number of Trips by bases:



5. Heatmap visualization of day, hour and month:



6. Map visualization of rides in New York:



6. CONCLUSION

At the end of this Uber data analysis R project, we studied how to create data visualizations. We used package ggplot2 that 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.

7. FUTURE SCOPE

We can use this data for training a model using ML and building a smart AI based predictive system.

Model can automatically send the insights to the authorities or drivers related to areas having most trips and passenger count in certain areas.

This big data can be used to study passenger's behavior.

REFERENCES

[1] https://ggplot2.tidyverse.org/
[2]https://github.com/geoninja/Uber-Data-Analysis
[3]https://www.researchgate.net/publication/333667985_A_Preliminary_Exploration_of_Uber_Data_as_an_Indicator_of_Urban_Liveability
[4]https://rpubs.com/Unsa/582359
[5]https://www.skyfilabs.com/project-ideas/uber-data-analysis
[6]https://www.uber.com/us/en/careers/teams/data-science/
[7]https://iedu.us/tag/project-in-r-uber-data-analysis-project/
[8]https://growvation.com/paritosh-sankhla/project/uber-data-analysis/5e95ee80-9455-4473-acaf-b670fe2abc8b
[9] Aguinaldo Bezerra, Gisliany Alves, Ivanovitch Silva, Pierangelo Rosati (2019). "A Preliminary Exploration of Uber Data as an Indicator of Urban Liveability". Research Gate. DOI: 10.1109/CyberSA.2019.8899714.
[10] Widdoes, Kaylene, "Case Study of Uber Data in the Central London Area" (2016). Honors Research Projects. 411.
[11] Uber Technologies, Inc., "Facts and figures," 2018.
[12] A. Ley and P. Newton, "Creating and sustaining liveable cities," in developing living cities: From analysis to action. World Scientific, 2010, pp. 191–229.
[13] R. Cervero, Transit-oriented development in the United States: Experiences, challenges, and prospects. Transportation Research Board, 2004, vol. 102.