**DATA ANALYTICS (CSE4027)**

FINAL REPORT

**UBER DATA ANALYSIS**

**PREDICTION**

**ABSTRACT:**

Data analytics help organizations and businesses gain insight into the enormous amount of knowledge they need for further production and growth. The world as we know it will continue to be increasingly driven by data. whether a business is moving in the right direction, data analytics is fundamentally the key to effective decision making. In our project we will analyse the uber pickups in city. 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 it also guide us in understanding the operations of ggplot2 library. For understanding the data and for developing an intuition for understanding the customers who will avail the trips.

**INTRODUCTION:**

Uber was founded 12 years ago, and it was already one of the fastest-growing companies in the world. Uber is a ridesharing company that hires independent contractors as drivers. It's one of many services today that contribute to sharing economy, supplying a means of connecting existing resources instead of providing the physical resources themselves. Uber is currently seen to take over 69% of the United States’ market share of consumer transport, and 25% of food delivery with its app uber eats. As of 2018, Uber is available in 700 cities and 63 countries, and enlists an average of 50,000 new drivers monthly. As of 2019, there are approximately 4 million registered Uber drivers. Almost half of all registered Uber drivers have multiple jobs.

Uber annual revenue



By looking at the above table we can observe that uber annual revenue increases yearly and uber is now available more than 10,000 cities globally.

**PROBLEM STATEMENT:**

Uber has been highly praised for giving independent contractors an opportunity to earn money as long as they have a car while also offering convenient ways for consumers to get around at lower costs. The biggest issues Uber faces include legal action because drivers are not licensed, rider and driver safety, protection and security of customer and driver information, and a lack of adequate insurance coverage. The aim of this project is to study and analyse uber data and making predictions based on uber dataset taking years 2018, 2019 and 2020.

**OBJECTIVE:**

India is Uber’s second largest market after the United States. Uber has a bright future and expansion opportunities are great. It is therefore important for Uber to ensure the safety of their riders and the drivers. They should also adopt controls to ensure that independent contractors using their app obeys relevant country laws. Increasing the demand for rides-for-services is a continuous or future challenge that requires attention primarily to safety improvements and rates that have a cost/benefit to both passengers and drivers. Unpredictable demand is a future risk that could be met with product diversification.

The objective of our project is to predict, analyse and estimate the value of the city market for uber. Our main objectives include visualization of Uber’s growth in city. Characterize the demand based on identified patterns in the time series. And the estimation of market for uber. And prediction of uber demand growth.

**PLAN OF ACTION:**

**Step-1: Understanding the dataset**

Data understanding is the knowledge that you have about the data, the needs that the data will satisfy, its content and location.

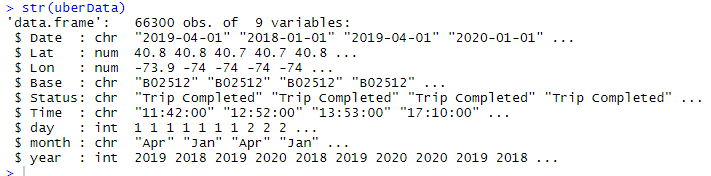
**Dataset Description:**

**Dataset name:** uberdata

**Dataset file: **

**Dataset size:** [66300, 9]

**Data types:**

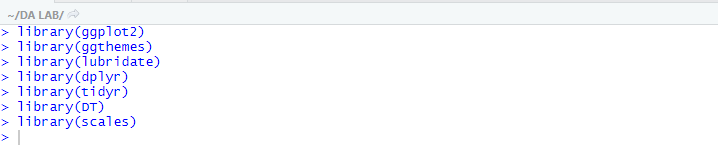


**Dataset Attributes:**

The dataset contains information about the Datetime, Latitude, Longitude and Base of each uber ride that happened from the month of January to December at a city.

* **Date/Time:** The date and time of uber pickup
* **Lat:** The latitude of the uber pickup
* **Lon:** The longitude of the uber pickup
* **Base:** The base code for uber pickup
* **Status:** The status is for whether uber ride completed or cancelled
* **Time:** The time of uber pickup
* **Day:** The day of uber pickup
* **Month:** The respective month that upper pickup took place
* **Year:** The respective year that upper pickup took place

**Loading necessary packages that are required:**



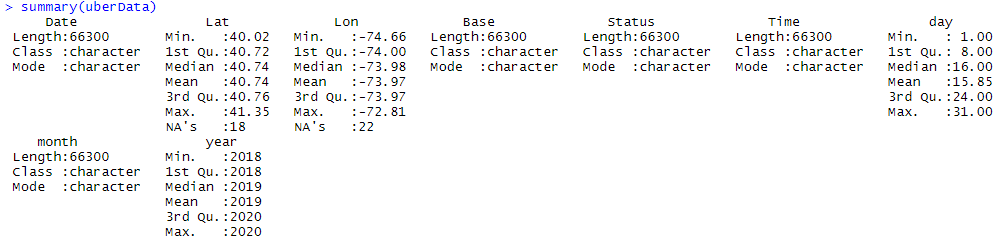
* **ggplot2()**- This is the backbone of this project. ggplot2 is the most popular data visualization library that is most widely used for creating aesthetic visualization plots.
* **ggthemes()**- This is more of an add-on to our main ggplot2 library. With this, we can create better create extra themes and scales with the mainstream ggplot2 package. provide some extra geoms, scales, themes for ggplot
* **lubridate()**- This package makes easier to work with dates and times. Our dataset involves various time-frames. In order to understand our data in separate time categories, we will make use of the lubridate package.
* **dplyr()**- Helps to resolve with most frequent data. This package is the lingua franca of data manipulation in R.
* **tidyr()**- provides important functions that can be used for data cleaning. This package will help you to tidy your data. The basic principle of tidyr is to tidy the columns where each variable is present in a column, each observation is represented by a row and each value depicts a cell.
* **DT()**- provides an R interface to the javascript library data tables. With the help of this package, we will be able to interface with the JavaScript Library called – Datatables.
* **Scales()**- With the help of graphical scales, we can automatically map the data to the correct scales with well-placed axes and legends.To customize to control the appearance of axis and legend lables.

**Loading or Reading the Dataset:**

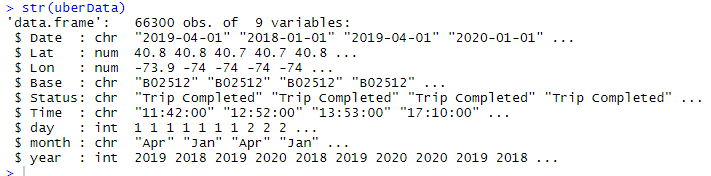


**Understanding the Dataset:**

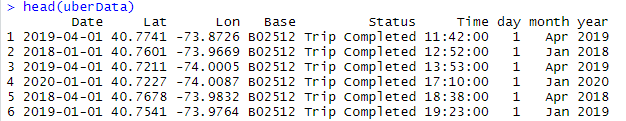
**Summary of data:**



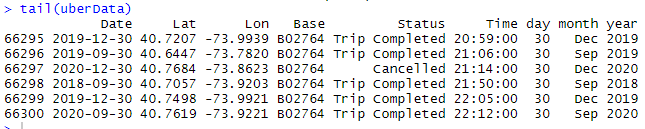
**Structure of data:**



**First n rows present in input data frame:**

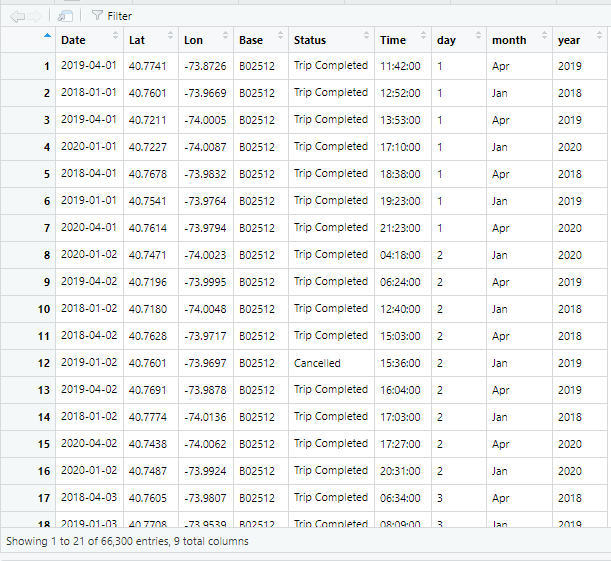


**Last n rows present in input data frame:**



**View contents in input data frame:**





**Dimension of data:** to see number of rows and columns present in data



**Unique level from Dataset:** To see how many levels present in Status column



**Step-2: Preprocessing dataset**

Data preprocessing is technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. The following preprocessing techniques are applied to the dataset:

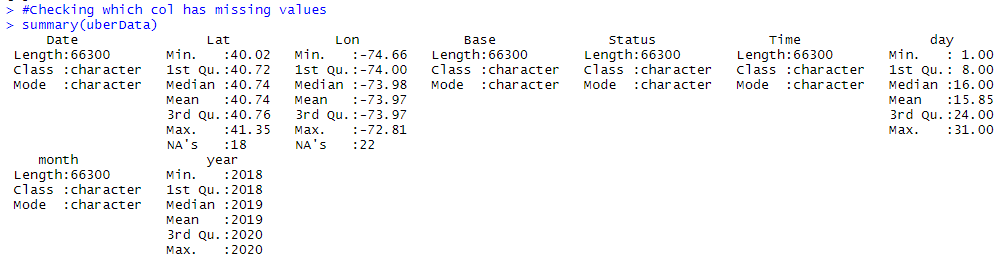
* **Checking missing values**: Missing values were found



* **Checking for how many missing values are present in data:**

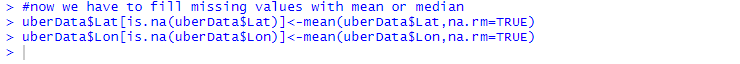


* **Checking for which column has missing values:**



From the above summary we can see Lat has 18 missing values and Lon has 22 missing values.

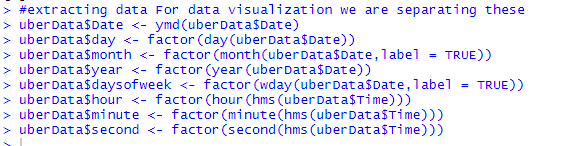
* **Filling the missing values with mean or medium:**



* **After filling with mean checking for missing values:**

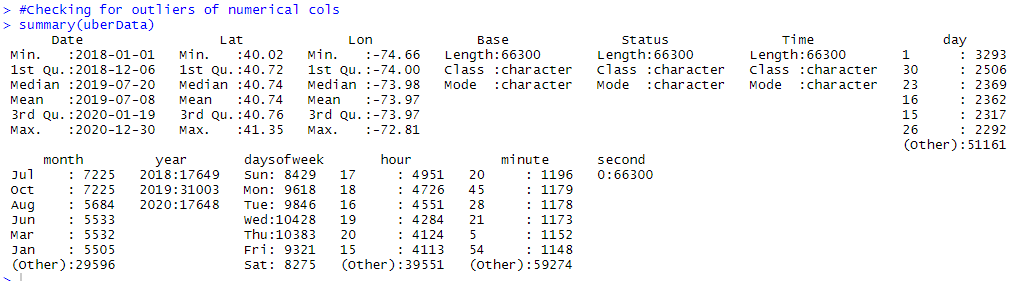


* **Extracting data:**

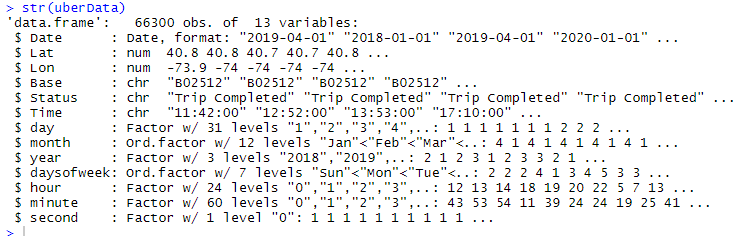


* **Dimension and summary of uber data after adding few columns:**





* **First 6 rows present in input data frame:** 
* **Structure of data after adding few columns:** To see datatype of each col



From above structure of data we can observe days of week, hour, minute and second columns are added to check number of uber rides are taking

place in a month, year, week, hour, minute and second.

* **Unique level in dataset from year:** To check how many years are there

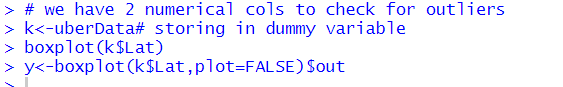


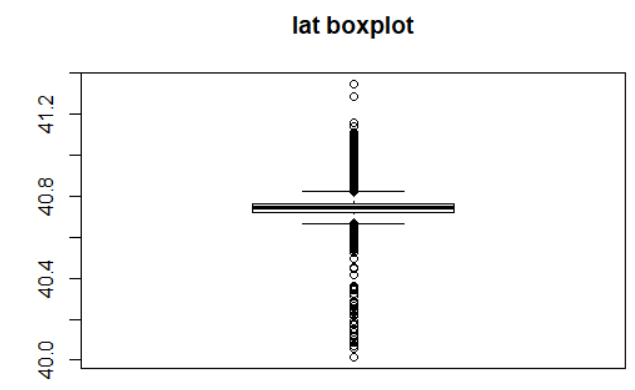
We can observe that we have 3 levels in dataset from column named year so from this we are retrieving uber pickups in year 2018, 2019, 2020.

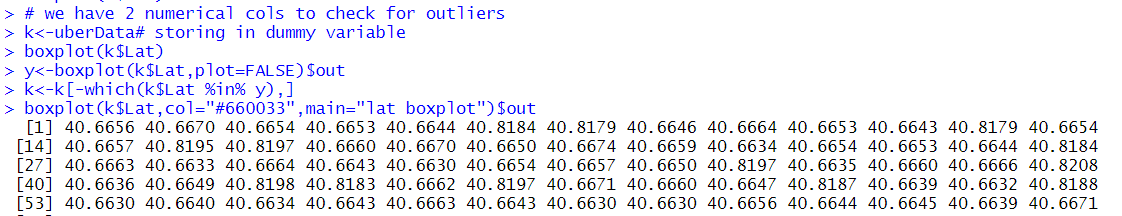
* **Checking for outliers:**

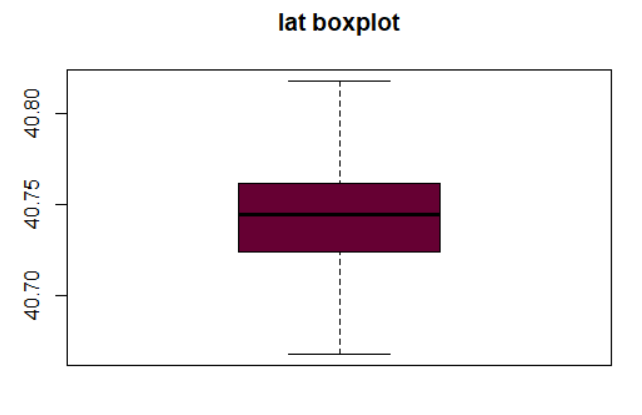
We have 2 numerical columns to check for outliers

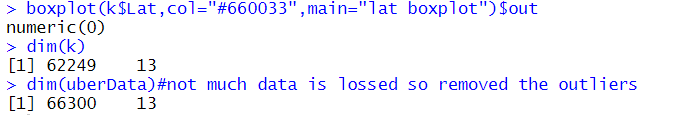
Lat column:







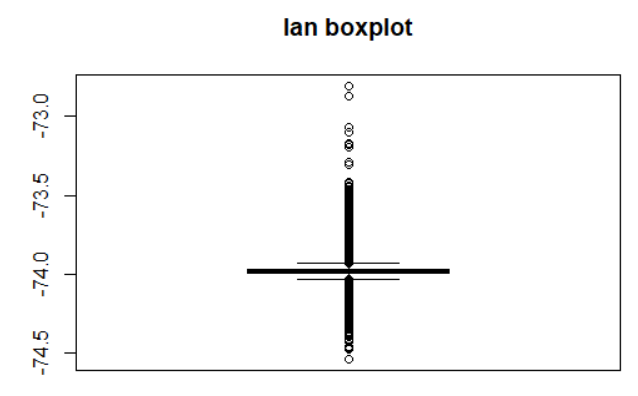




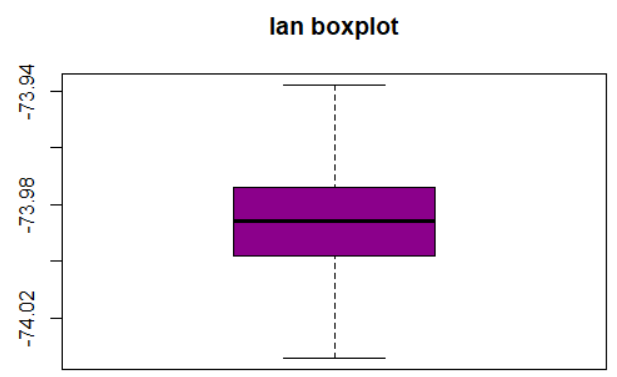
As we can see much data is not lost so we can remove outliers

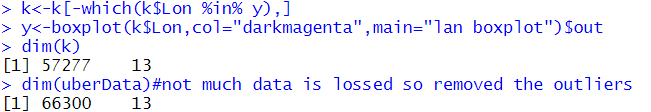
Longitude Column:



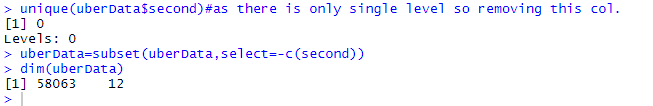








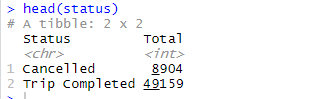
**Removing unnecessary columns:** As there is only single level and it does not help us neither in prediction nor improve accuracy so we are dropping this second column.



* **Storing data in dummy variable**

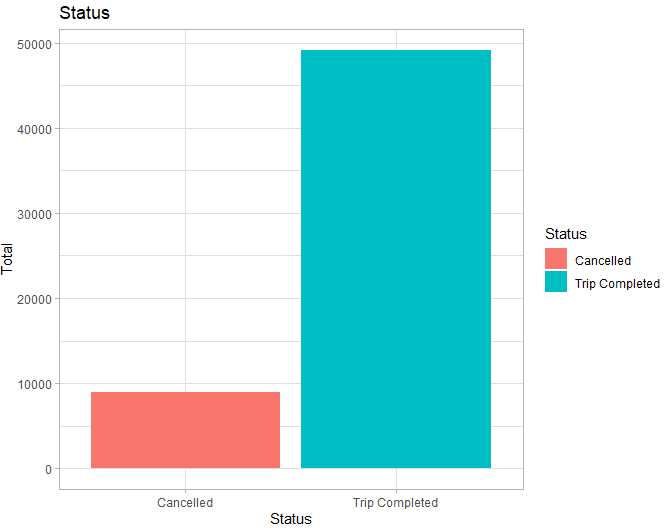


**Total count of levels in status for uber pickups:**



**Graph plot for status of uber pickup:** we can see very less cancelled trips are there

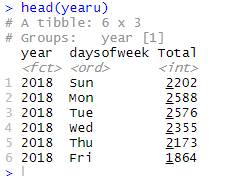


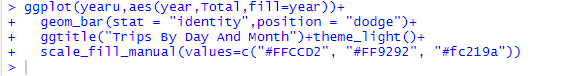


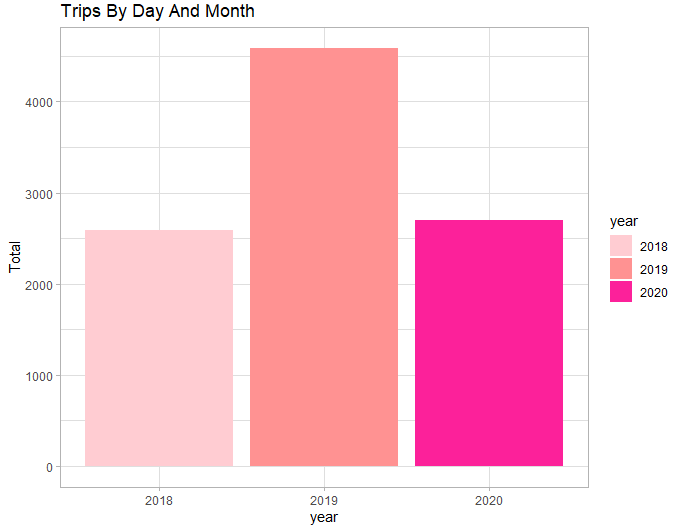
**Observation**: Trips completed is high compared to trips cancelled.

**Graph plot for total trips by Day and Month:**



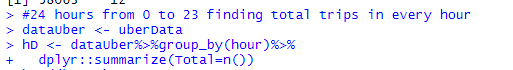




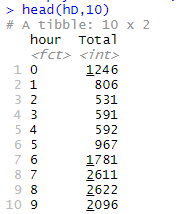


**Observation**: we can observe from above graph plot that total uber pickups are more in the year 2019 compared to years 2018 and 2020.

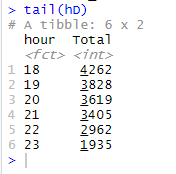
* **Finding total trips in every hour (24 hours -> from 0 to 23)**

we will use the **ggplot** function to plot the number of trips that the passengers had made in a day. We will also use **dplyr** to aggregate our data. In the resulting visualizations, we can understand how the number of passengers fares throughout the day. 

**Total number of trips in every hour (From first 10 rows of data)**

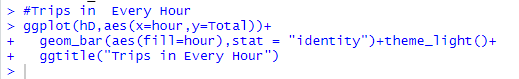


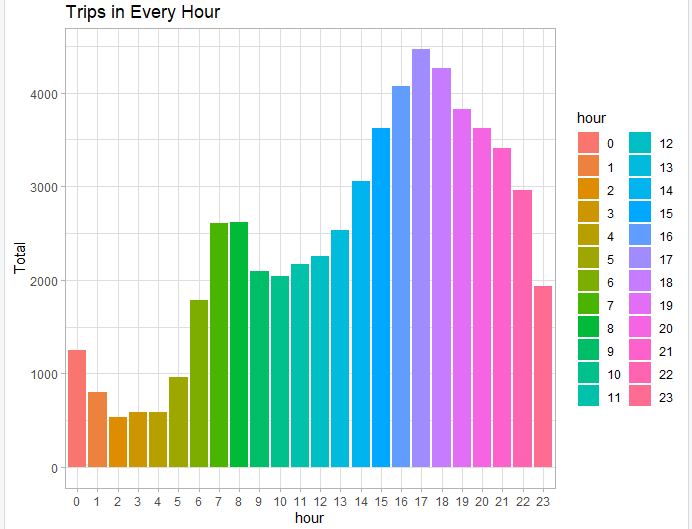
**Total number of trips in every hour (From last 10 rows of data)**



**Step-3: Plotting graphs for better understanding of relation between attributes:**

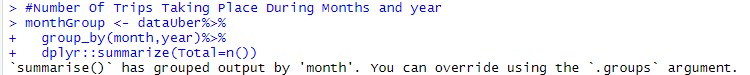
* **Trips every hour:**



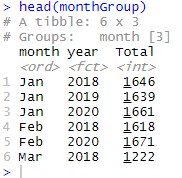


**Observation:** we can observe from the above graph the journey peak between 6-8 am and then peak again even more from 3pm rises steadily until its on its peak at around 5pm before the pickups steadily drop till late night.

* **Graph plot for total trips taking place during Months and year:**



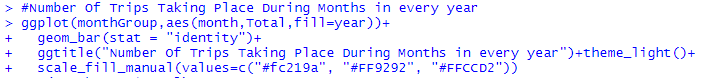
**Total number of trips (From first 10 rows of data)**

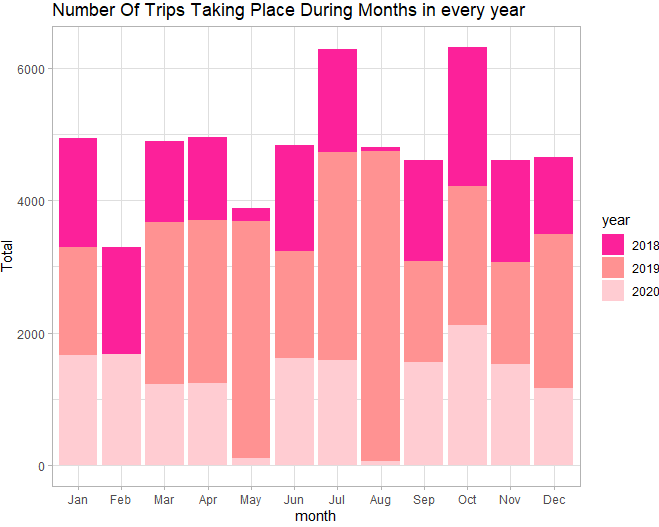


**Maximum number of trips that occurred (Month versus Year)**



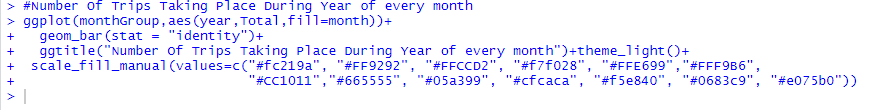
**Plotting graphs for trips by Months and Year**



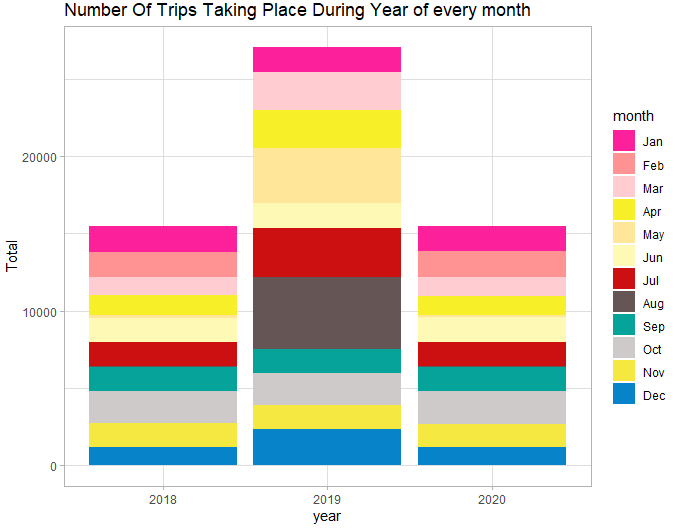


**Observation:** From above graph plot we can observe the number of trips that are taking place in each month of the year, we observe that most trips were made in months of July and October.

* **Plotting Trips by Year of every Month**

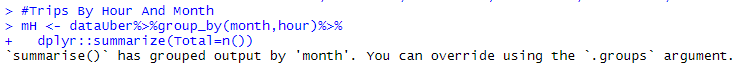


**Plotting graphs for number of trips taking place during every year of every Month:**

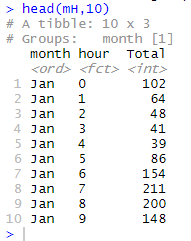


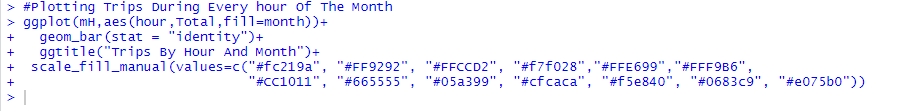
**Observation:** From the above graph plot we visualized the number of trips that are taking place each month of the year and in the output visualization, we observed that most trips were made during the month of August in the year 2019.

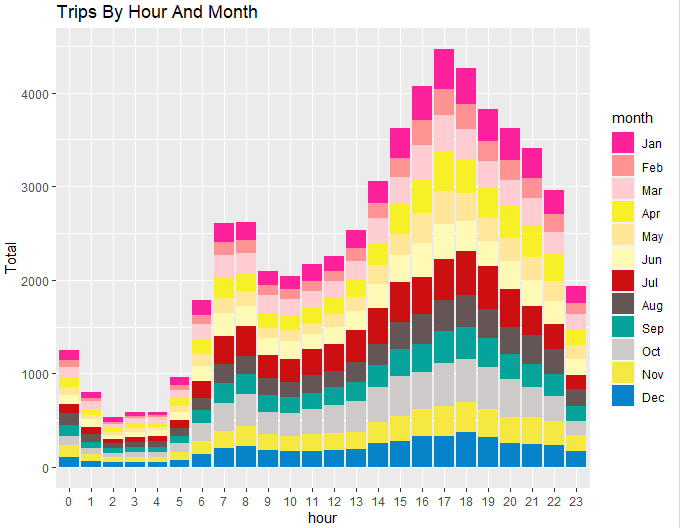
* **Plotting Trips By Hour and Month**



**Total trips count that takes place between months versus Hour:**

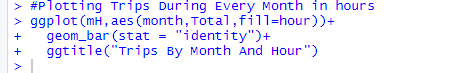


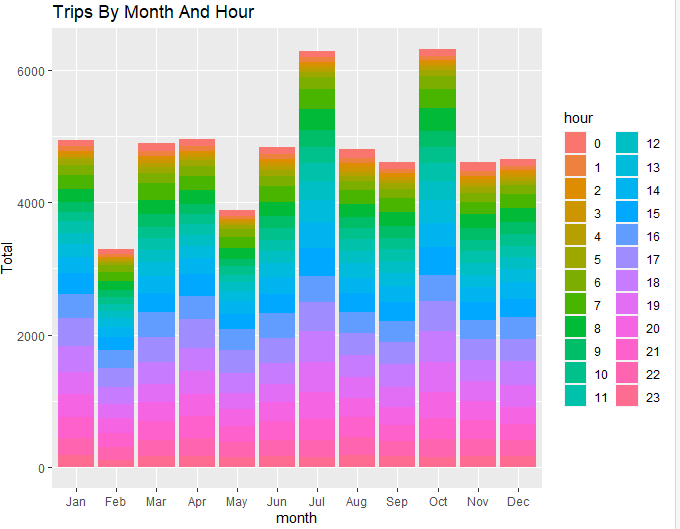




**Observation:** From the above graph, the graph plot taken by total number of trips by hour and month and we can observe that highest number of trips were made during the months

* **Plotting Trips by Month and Hour**

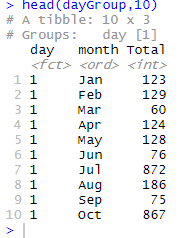


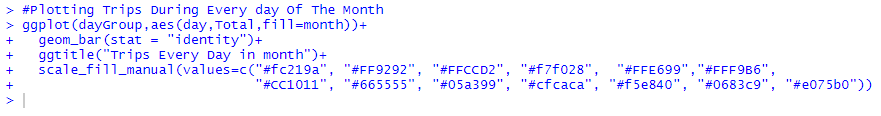


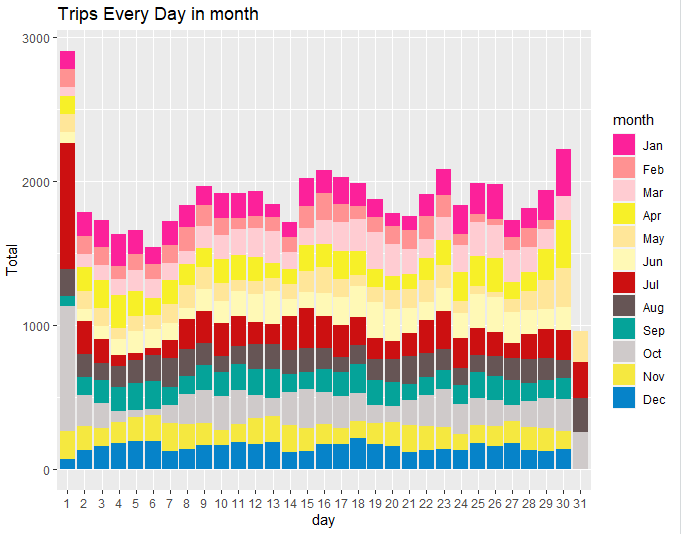
**Observation:** From the above graph plot visualization we can observe that number of trips taken by month and hour and highest number of trips were made during the months of July and October.

* **Plotting Trips By Month and Hour**



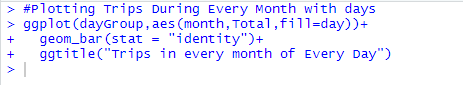


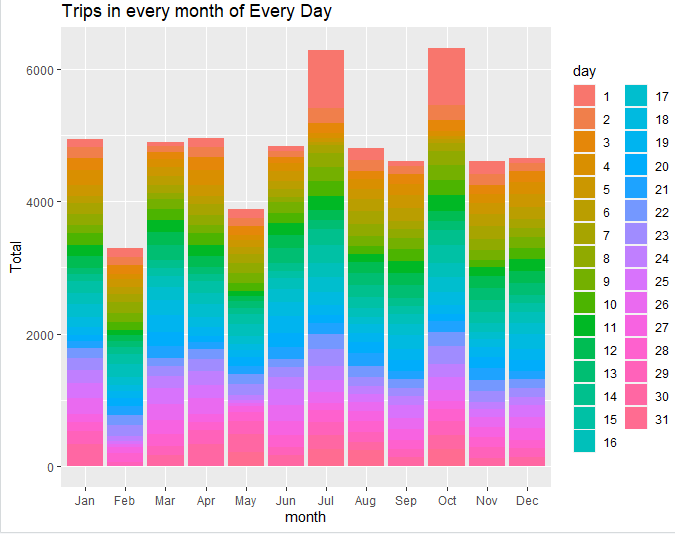




**Observation:** From the above graph we can observe that we have plotted the graph with number of trips every day in month, we observed that higher number of pickups were made during the month of July.

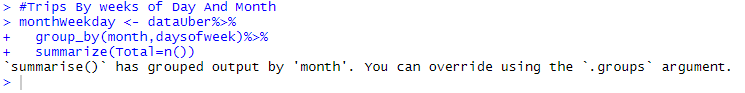
* **Plotting Trips During every Month with Days**



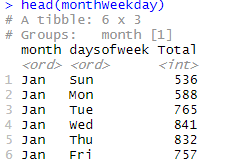


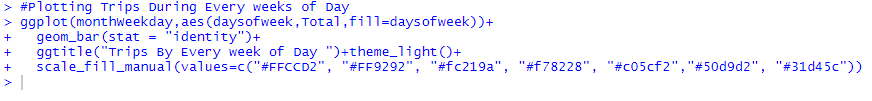
**Observation:** From the above graph we can observe that higher number of trips in every month of every day were made during in the months of July and October.

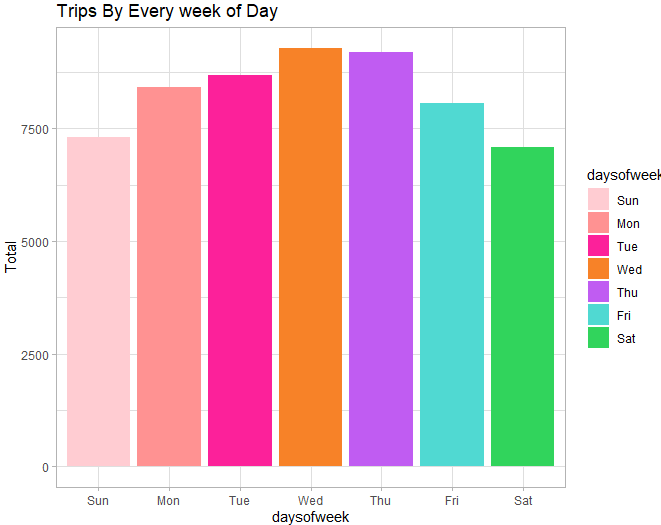
* **Plotting Trips During every Month with Days**



**Total count of uber pickups that were made with month and days of week (first 6 rows of data):**

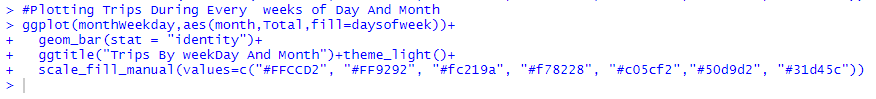






**Observation:** From the above visualization we can conclude that higher number of rides were made during Wednesday in a week.

* **Plotting Trips During every day of week and Month:**

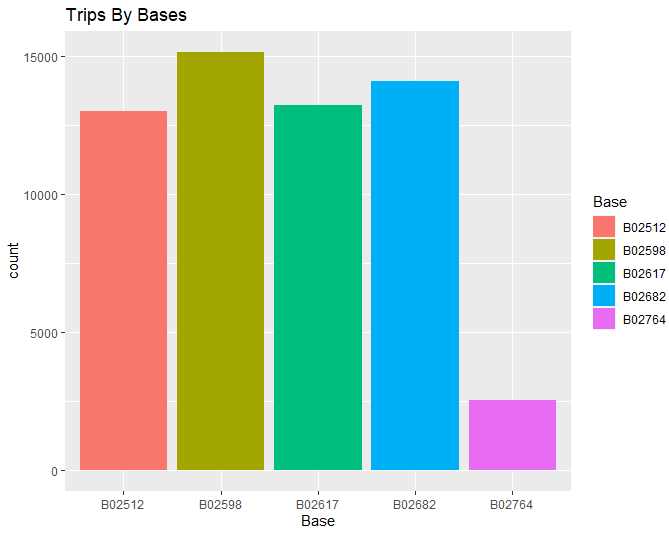




**Observation:** From the above graph plot visualization we can conclude that higher number of uber rides were made during Monday, Tuesday, Wednesday and Thursday in the months of July and October.

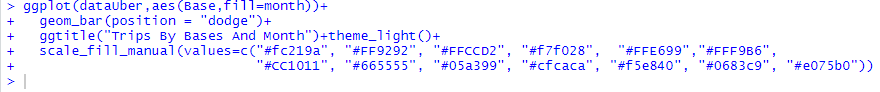
* **Plotting Graph plot Trips By Bases:**

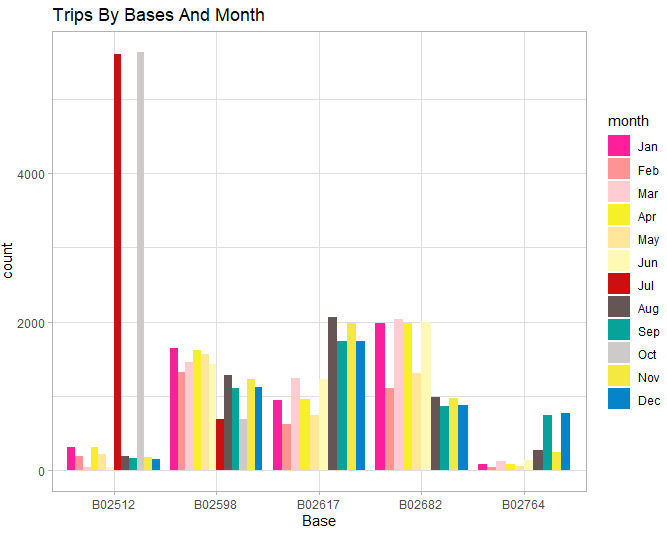




**Observation:** From the above visualization we plot the number of trips that have been taken by passengers from each of the bases. There are five bases in all out of which, we observe that B02598 had the highest number of trips.

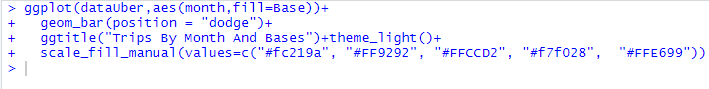
* **Plotting Graph plot Trips By Bases and Month:**

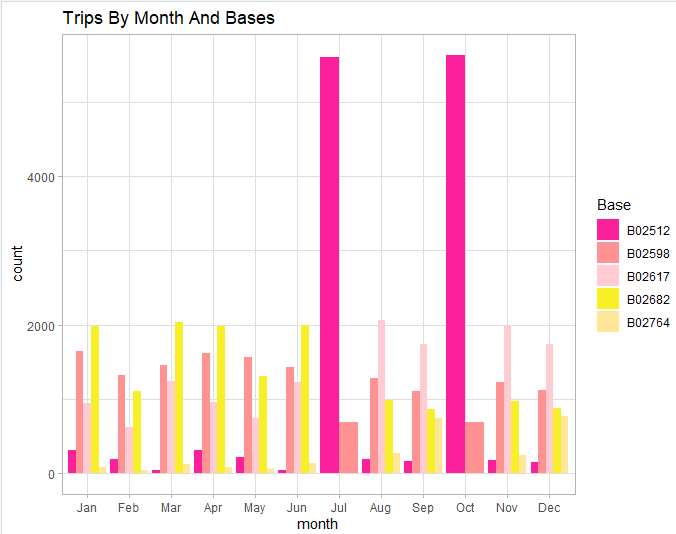




**Observation:** From the above visualization we plot the number of trips that have been taken by passengers from each of base and month. Higher number of trips were made with base B02512 in the month of July and October.

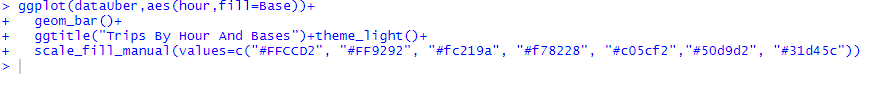
* **Plotting Graph plot with Trips By Month and Bases:**

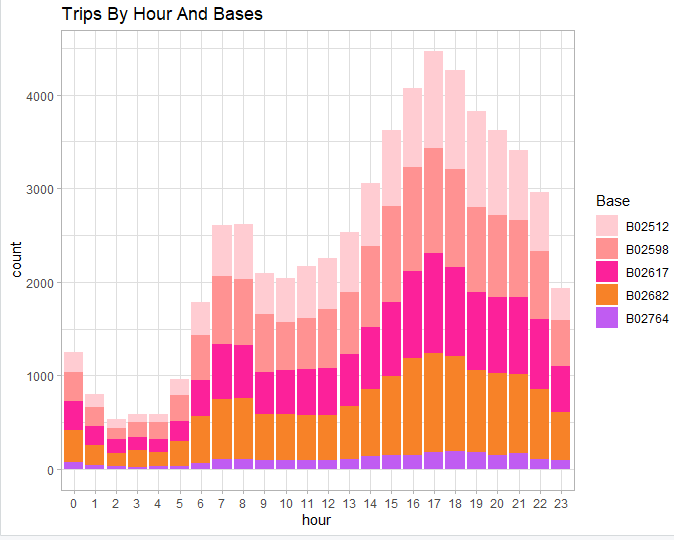




**Observation:** From the above visualization we plot the number of trips that have been taken by passengers from each of month and base. Higher number of trips were made with base B02512 in the month of July and October.

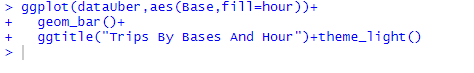
* **Plotting Graph plot with Trips By Hour and Bases:**

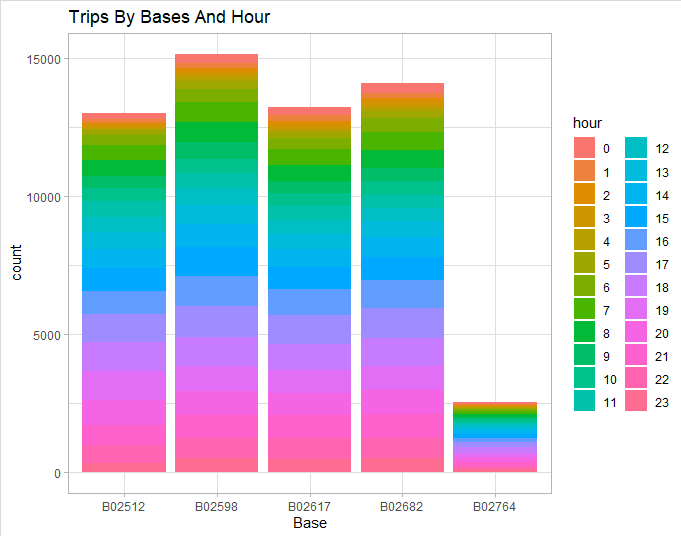




**Observation:** From the above visualization we plot the number of trips that have been taken by passengers from each of hour and bases. Higher number of trips were made with base B02512 at 5pm.

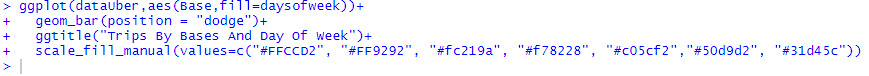
* **Plotting Graph plot with Trips By Base and Hour:**

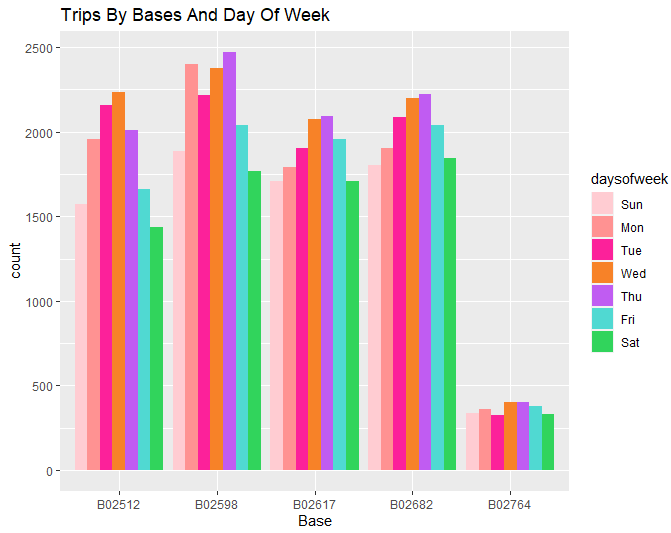




**Observation:** From the above visualization we plot the number of trips that have been taken by passengers from each of bases and hour. Higher number of trips were made with base B02598.

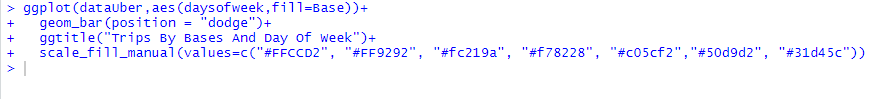
* **Plotting Graph plot with Trips By Base and Day Of Week:**

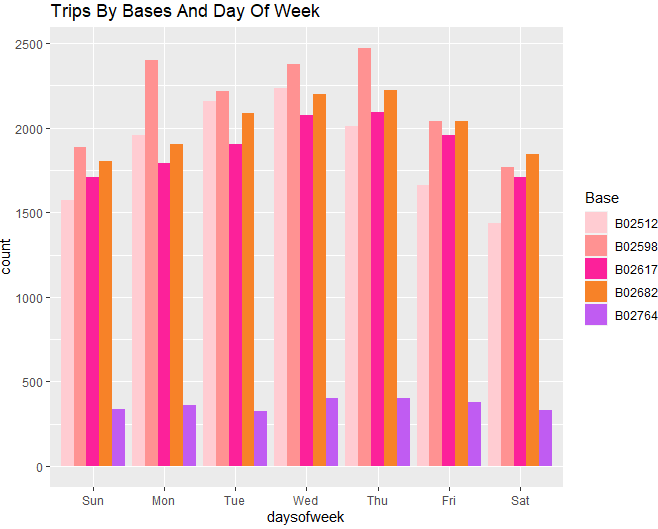




**Observation:** From the above visualization we plot the number of trips that have been taken by passengers from each of bases and Day of week. Higher number of trips were made with base B02598 during Thursday.

* **Plotting Graph plot with Trips By Day of week and Base:**



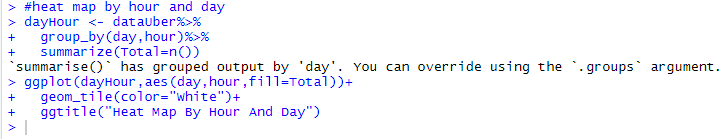


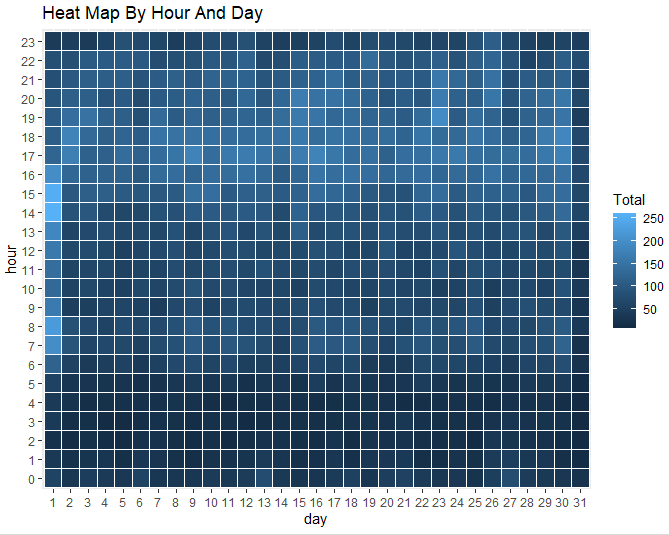
**Observation:** From the above visualization we plot the number of trips that have been taken by passengers from each of Day of week and base. Higher number of trips were made during Thursday with base B02598.

**Step-4: Plotting Heat Map for better understanding of relation between attributes:**

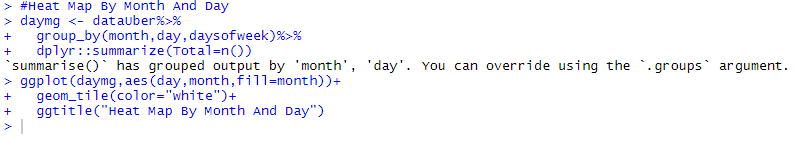
**A heatmap is a graphical representation of data that uses a system of color-coding to represent different values. Heatmaps are used in various forms of analytics**

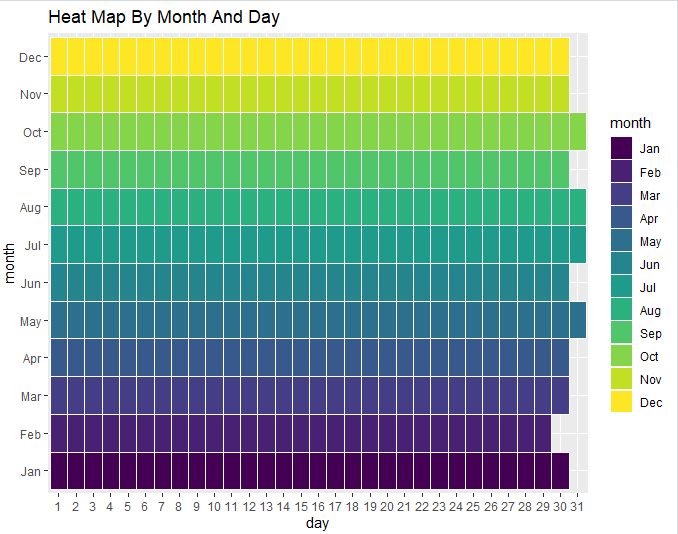
* **Heat Map By Hour and Day:**



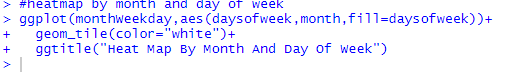


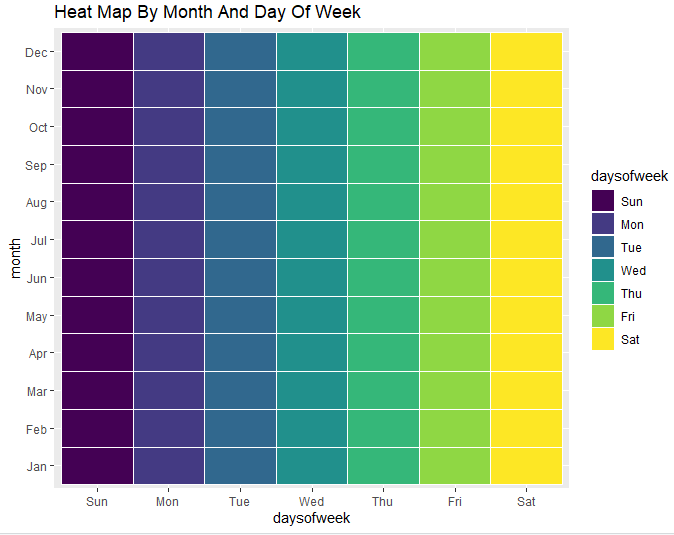
* **Heat Map By Month and Day:**



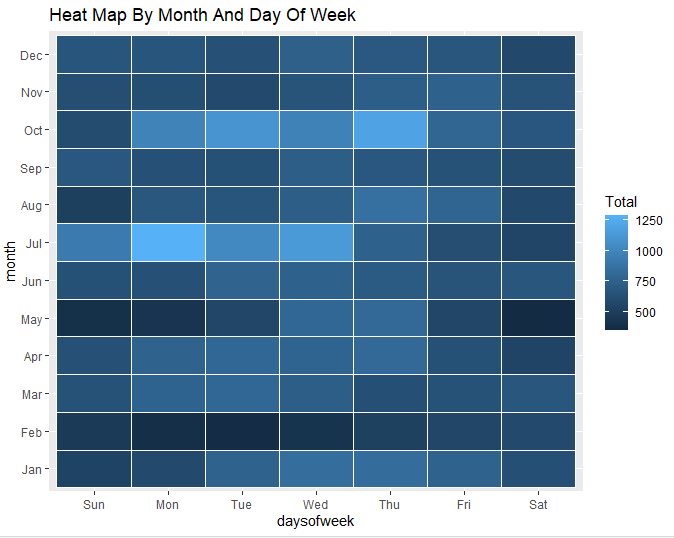


* **Heat Map By Month and Day of week**

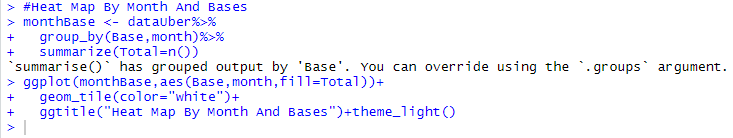


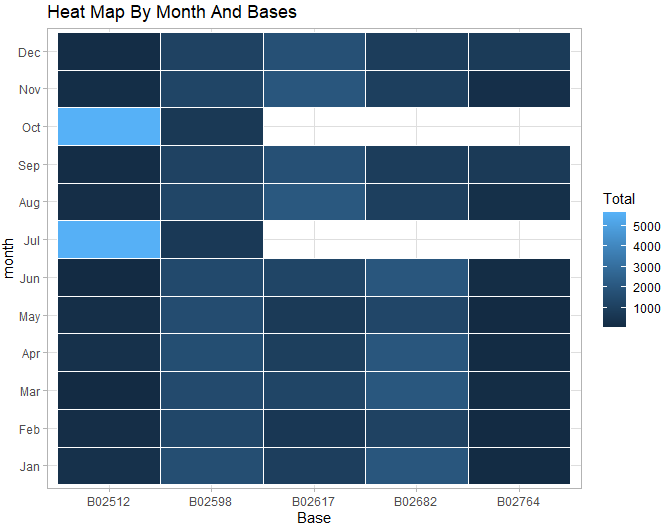




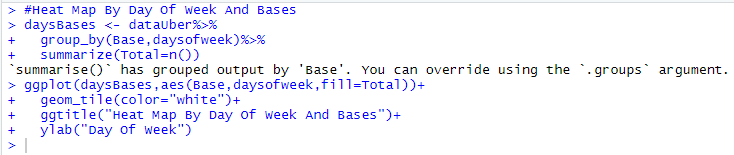


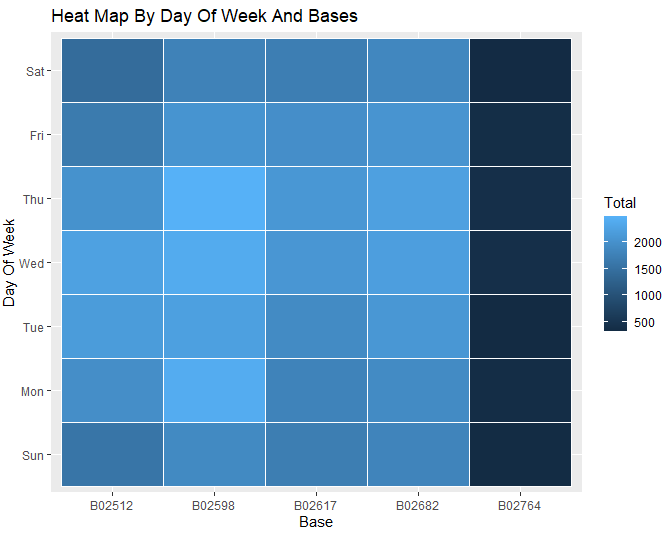
* **Heat Map By Month and Day of week**



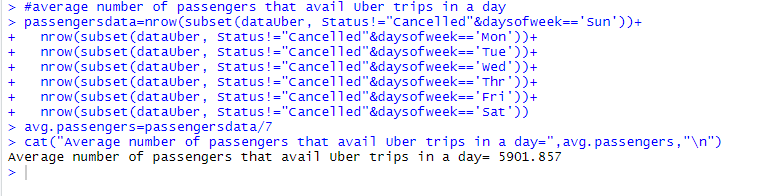


* **Heat Map By Day of week and Bases**





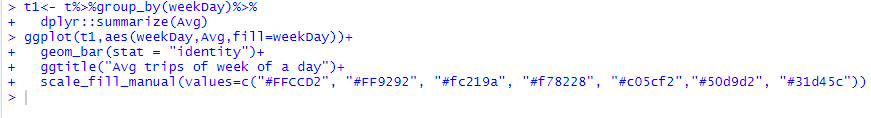
* **Average number of passengers that avail uber trips in a day**



* **Average number of passengers that avail uber trips in days of Sunday, Monday, Tuesday, Wednesday, Thursday, Friday and Saturday**



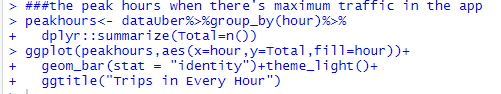


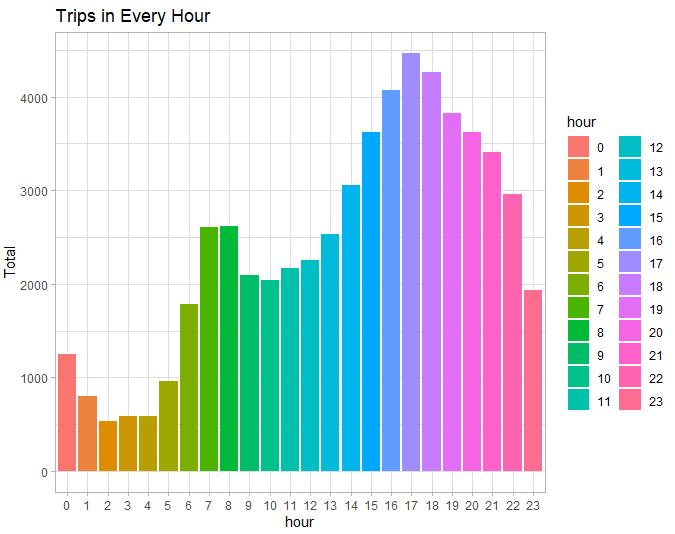




**Observation:** From the above visualization we plot average trips of week of a day and we can say Wednesday and Thursday is the average trips avail by passengers.

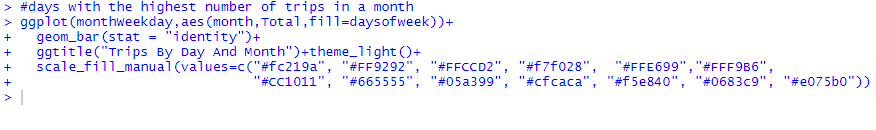
* **The peak hours where there is maximum traffic in the app**

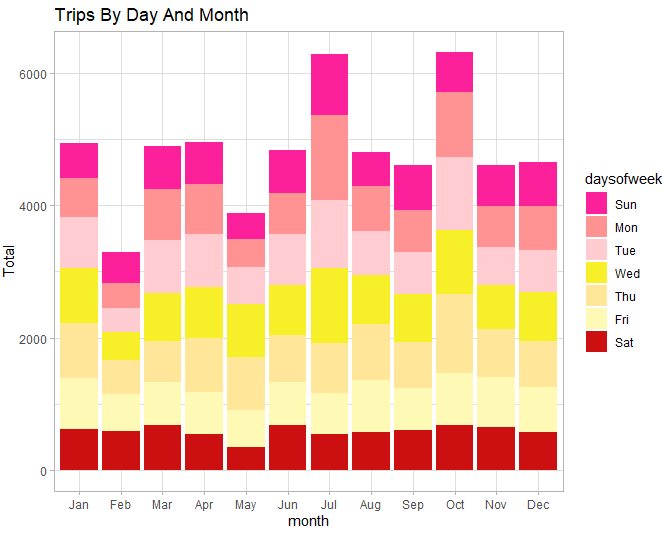




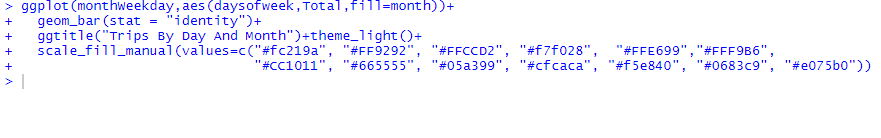
**Observation:** From the above graph plot visualization we plot trips in every hour we can observe that higher number of trips were made around 5-6 pm.

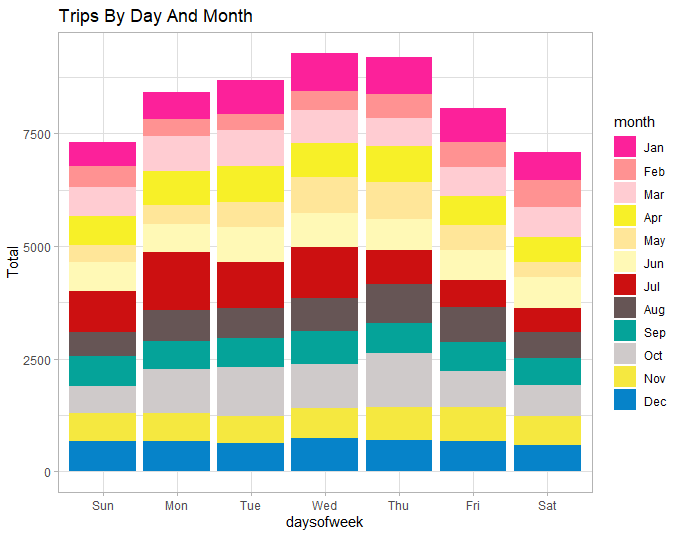
* **Days with highest number of trips in a month**





**Observation:** From the above graph plot visualization we can observe highest number of trips were made in the months of July and October.





**Observation:** From the above graph plot visualization we can observe that higher number of trips were made in the month of January and during Wednesdays.

**CONCLUSION:**

At the end of the Uber data analysis R project, we observed how to create data visualizations. We made use of packages like ggplot2 that allowed us to plot various types of visualizations that pertained to several time-frames of the year. With this, we could conclude how time affected customer trips. Finally, we made a geo plot of city that provided us with the details of how various users made trips from different bases.