**Objective:**

This is more of a data visualization project that will guide you towards using the ggplot2 library for understanding the data and for developing an intuition for understanding the customers who avail the trips.

**Activity outcomes:**

In this R project, we will analyze the **Uber Pickups in New York City dataset**.

**Essential Libraries:**

* *library(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.

* *library(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.

* *library(lubridate)*

Our dataset involves various time-frames. In order to understand our data in separate time categories, we will make use of the lubridate package.

* *library(dplyr)*

This package is the lingua franca of [data manipulation in R](https://data-flair.training/blogs/manipulating-and-processing-data-in-r/).

* *library(tidyr)*

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.

* *library(DT)*

With the help of this package, we will be able to interface with the [JavaScript](https://data-flair.training/blogs/javascript-tutorials-home/)Library called – Datatables

.

* *library(scales)*

With the help of graphical scales, we can automatically map the data to the correct scales with well-placed axes and legends.

**Setting the path for assessing dataset:**

*setwd('C:/Users/nazir/Downloads/Uber-dataset')*

The purpose of this line is to set the path directory of the dataset that we will use in our analysis.

**Making the vector of colors**

In this step of data science project, we will create a vector of our colors that will be included in our plotting functions. You can also select your own set of colors.

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

 We will read several csv files that contain the data from April 2014 to September 2014. We will store these in corresponding data frames like apr\_data, may\_data, etc.

*apr\_data <- read.csv("uber-raw-data-apr14.csv")*

*may\_data <- read.csv("uber-raw-data-may14.csv")*

*jun\_data <- read.csv("uber-raw-data-jun14.csv")*

*jul\_data <- read.csv("uber-raw-data-jul14.csv")*

*aug\_data <- read.csv("uber-raw-data-aug14.csv")*

*sep\_data <- read.csv("uber-raw-data-sep14.csv")*

We will combine all of this data into a single dataframe called ‘data\_2014’.

*data\_2014 <- rbind(apr\_data,may\_data, jun\_data, jul\_data, aug\_data, sep\_data)*

**Formatting time**

In the next step, we will perform the appropriate formatting of Date.Time column. Then, we will proceed to create factors of time objects like day, month, year etc.

*data\_2014$Date.Time <- as.POSIXct(data\_2014$Date.Time, format = "%m/%d/%Y %H:%M:%S")*

*data\_2014$Time <- format(as.POSIXct(data\_2014$Date.Time, format = "%m/%d/%Y %H:%M:%S"), format="%H:%M:%S")*

*data\_2014$Date.Time <- ymd\_hms(data\_2014$Date.Time)*

*data\_2014$day <- factor(day(data\_2014$Date.Time))*

*data\_2014$month <- factor(month(data\_2014$Date.Time, label = TRUE))*

*data\_2014$year <- factor(year(data\_2014$Date.Time))*

*data\_2014$dayofweek <- factor(wday(data\_2014$Date.Time, label = TRUE))*

*data\_2014$hour <- factor(hour(hms(data\_2014$Time)))*

*data\_2014$minute <- factor(minute(hms(data\_2014$Time)))*

*data\_2014$second <- factor(second(hms(data\_2014$Time)))*

**Computimg number of trips made per hour**

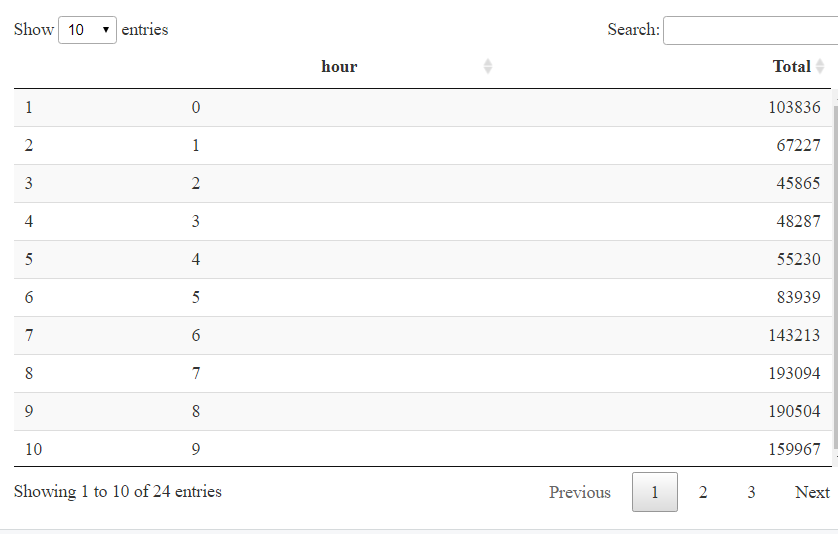
Next we will use the ggplot function to plot the number of trips that the passengers had made in a day. We will also use dplyr to aggregate our data. In the resulting visualizations, we can understand how the number of passengers fares throughout the day. We observe that the number of trips are higher in the evening around 5:00 and 6:00 PM.

*hour\_data <- data\_2014 %>%*

*group\_by(hour) %>%*

*dplyr::summarize(Total = n())*

*datatable(hour\_data) #creates table*

**

visualize data on ggplot

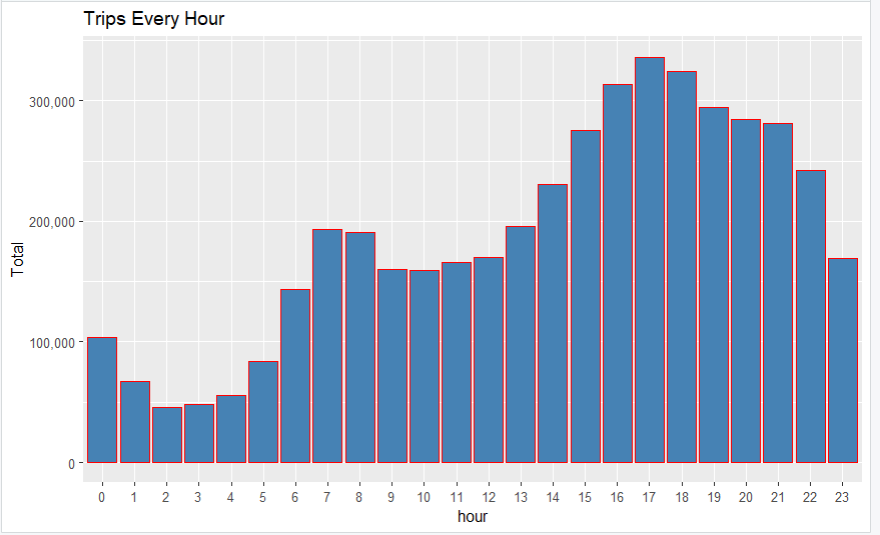
*ggplot(hour\_data, aes(hour, Total)) +*

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

*ggtitle("Trips Every Hour") +*

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

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



compute trips per hour of every month

*month\_hour <- data\_2014 %>%*

*group\_by(month, hour) %>%*

*dplyr::summarize(Total = n())*

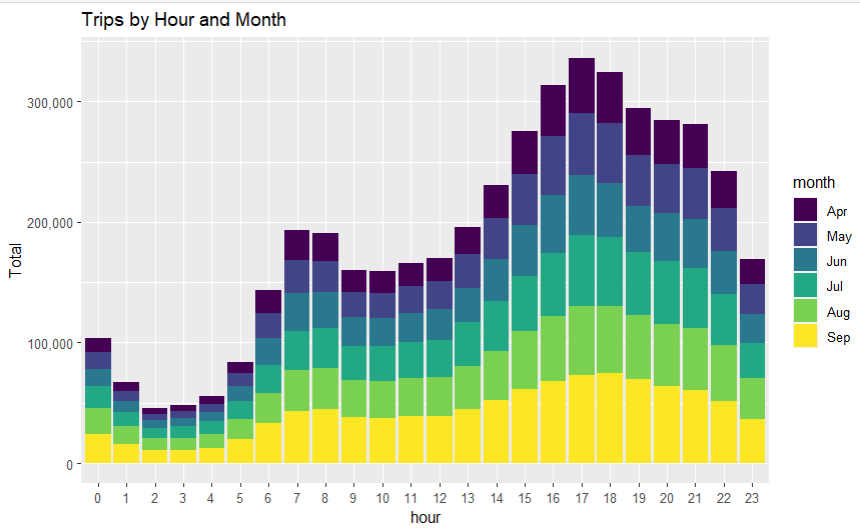
visualize it on ggplot

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

*geom\_bar( stat = "identity") +*

*ggtitle("Trips by Hour and Month") +*

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



computing no. of on daily base

*day\_group <- data\_2014 %>%*

*group\_by(day) %>%*

*dplyr::summarize(Total = n()) #n() use for current group size*

*datatable(day\_group)*

**

 We will learn how to plot our data based on every day of the month. We observe from the resulting visualization that 30th of the month had the highest trips in the year which is mostly contributed by the month of April.

#plot it

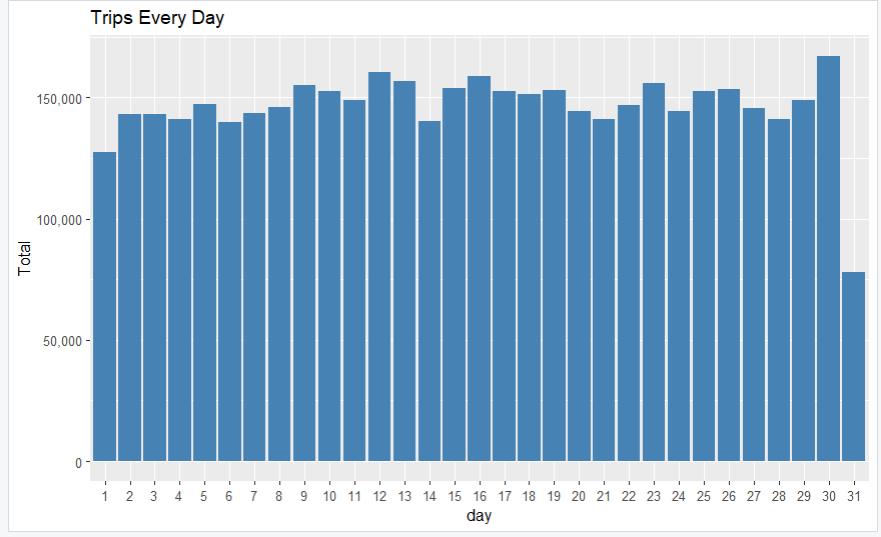
*ggplot(day\_group, aes(day, Total)) +*

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

*ggtitle("Trips Every Day") +*

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

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



#compute daily bases trips in each month

*day\_month\_group <- data\_2014 %>%*

*group\_by(month, day) %>%*

*dplyr::summarize(Total = n())*

#plot it

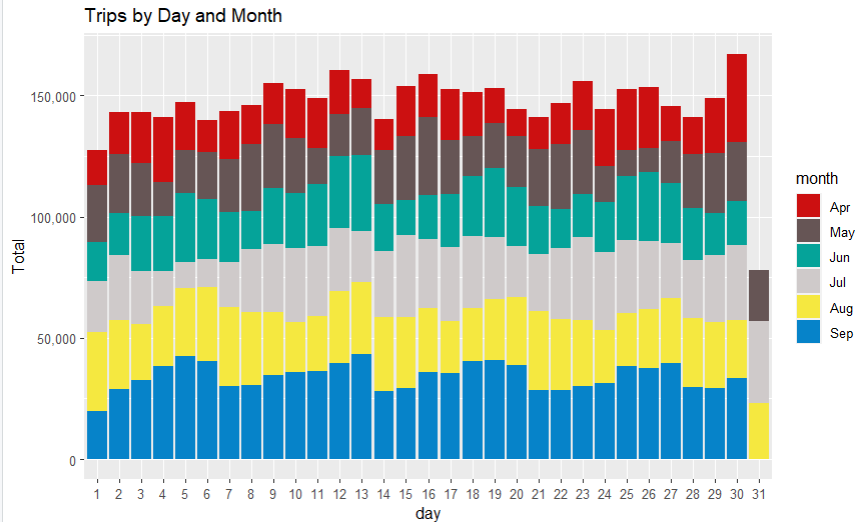
*ggplot(day\_month\_group, aes(day, Total, fill = month)) +*

*geom\_bar( stat = "identity") +*

*ggtitle("Trips by Day and Month") +*

*scale\_y\_continuous(labels = comma) +*

*scale\_fill\_manual(values = colors)*



#group data by month

*month\_group <- data\_2014 %>%*

*group\_by(month) %>%*

*dplyr::summarize(Total = n())*

*datatable(month\_group)*

**

We will visualize the number of trips that are taking place each month of the year. In the output visualization, we observe that most trips were made during the month of September. Furthermore, we also obtain visual reports of the number of trips that were made on every day of the week.

#plot it

*ggplot(month\_group, aes(month, Total, fill = month)) +*

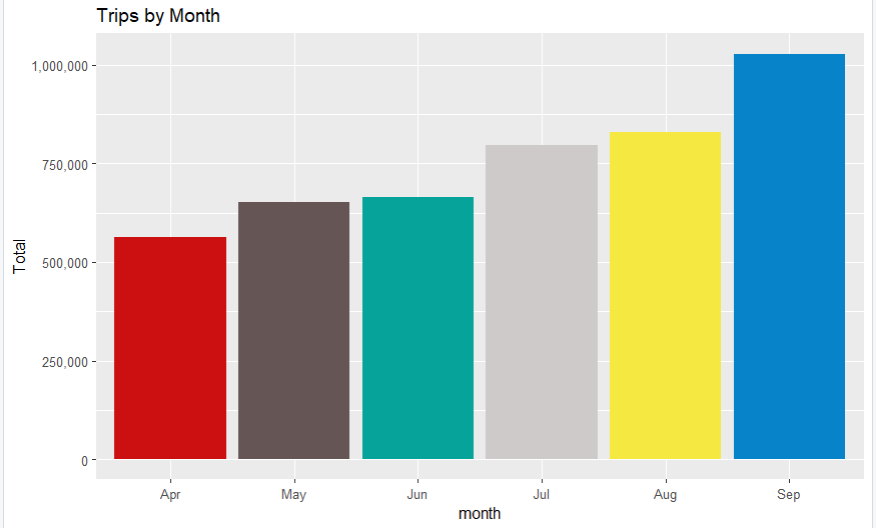
*geom\_bar( stat = "identity") +*

*ggtitle("Trips by Month") +*

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

*scale\_y\_continuous(labels = comma) +*

*scale\_fill\_manual(values = colors)*



#compute trips on weekly bases of each month

*month\_weekday <- data\_2014 %>%*

*group\_by(month, dayofweek) %>%*

*dplyr::summarize(Total = n())*

#plot it

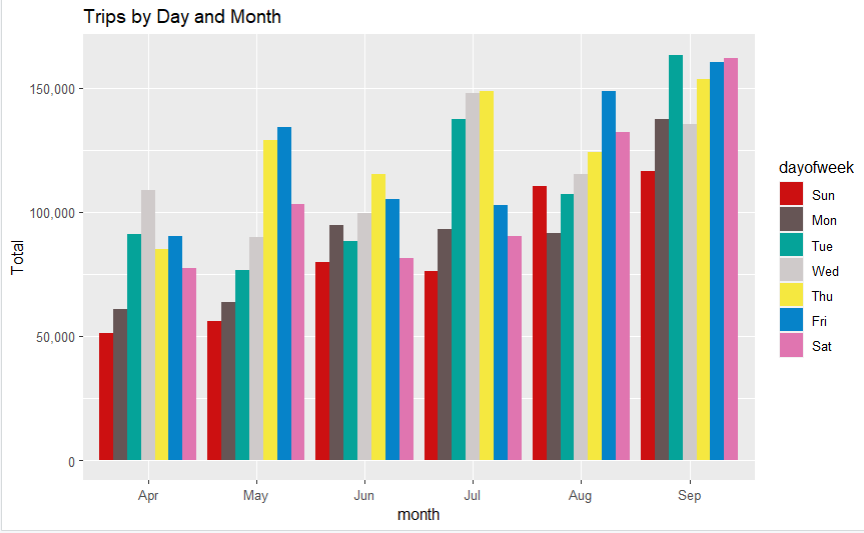
*ggplot(month\_weekday, aes(month, Total, fill = dayofweek)) +*

*geom\_bar( stat = "identity", position = "dodge") +*

*ggtitle("Trips by Day and Month") +*

*scale\_y\_continuous(labels = comma) +*

*scale\_fill\_manual(values = colors)*

**

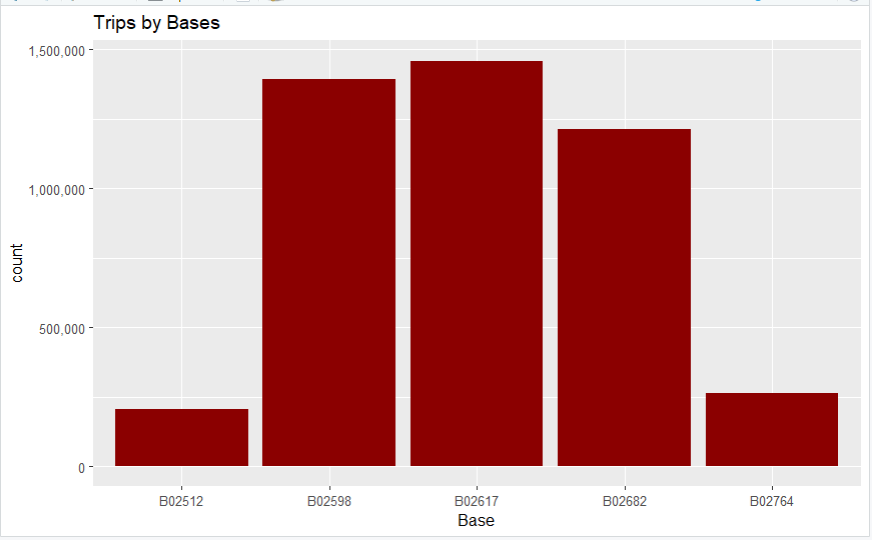
#number of trips taken by the passengers from each of the bases

*ggplot(data\_2014, aes(Base)) +*

*geom\_bar(fill = "darkred") +*

*scale\_y\_continuous(labels = comma) +*

*ggtitle("Trips by Bases")*

**

In the following visualization, we plot the number of trips that have been taken by the passengers from each of the bases. There are five bases in all out of which, we observe that B02617 had the highest number of trips. Furthermore, this base had the highest number of trips in the month B02617. Thursday observed highest trips in the three bases – B02598, B02617, B02682.

#number of trips taken by the passengers from each of the bases of each month

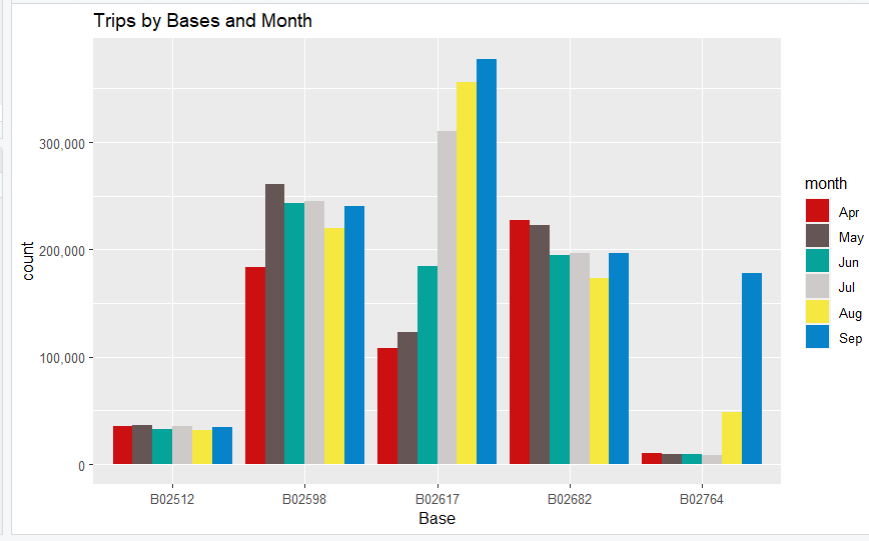
*ggplot(data\_2014, aes(Base, fill = month)) +*

*geom\_bar(position = "dodge") +*

*scale\_y\_continuous(labels = comma) +*

*ggtitle("Trips by Bases and Month") +*

*scale\_fill\_manual(values = colors)*



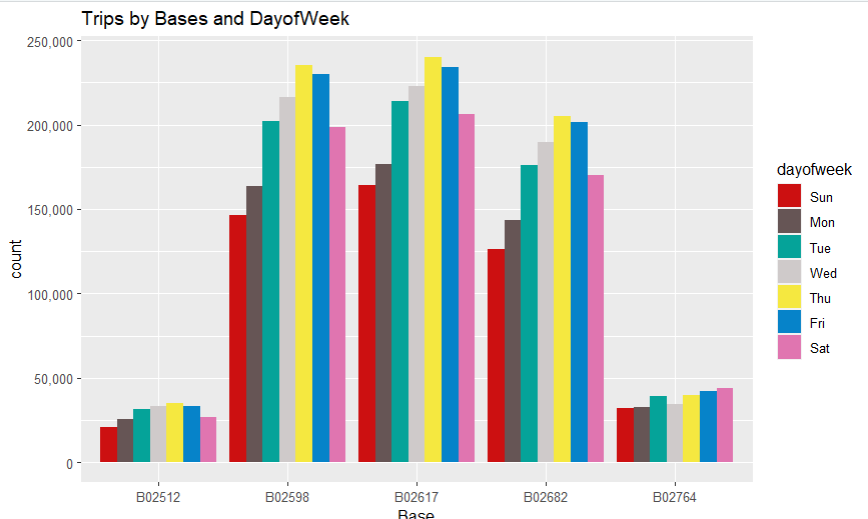
#number of trips taken by the passengers from each of the bases of each week *ggplot(data\_2014, aes(Base, fill = dayofweek)) +*

*geom\_bar(position = "dodge") +*

*scale\_y\_continuous(labels = comma) +*

*ggtitle("Trips by Bases and DayofWeek") +*

*scale\_fill\_manual(values = colors)*



**Creating a Heatmap visualization of day, hour and month**

#no. of trips each hour of all the days of each month

*day\_and\_hour <- data\_2014 %>%*

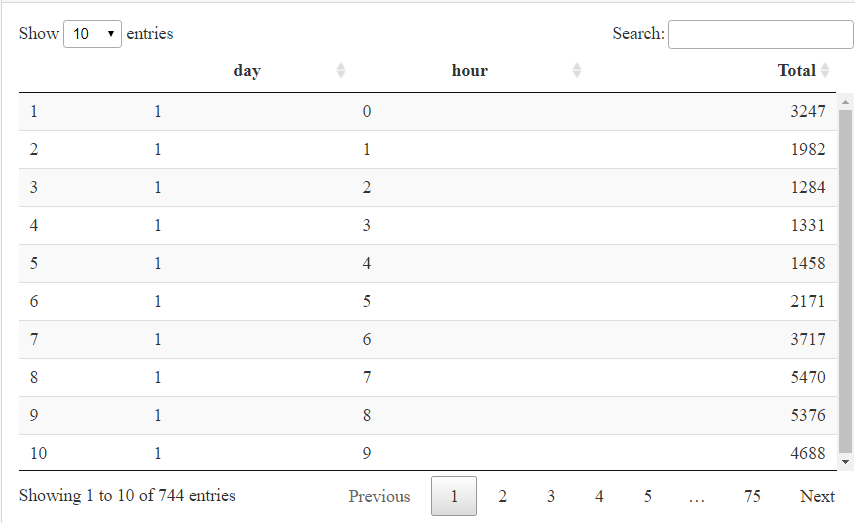
*group\_by(day, hour) %>%*

*dplyr::summarize(Total = n())*

We will plot [Heatmap](https://en.wikipedia.org/wiki/Heat_map) by Hour and Day.

#create data table

*datatable(day\_and\_hour)*

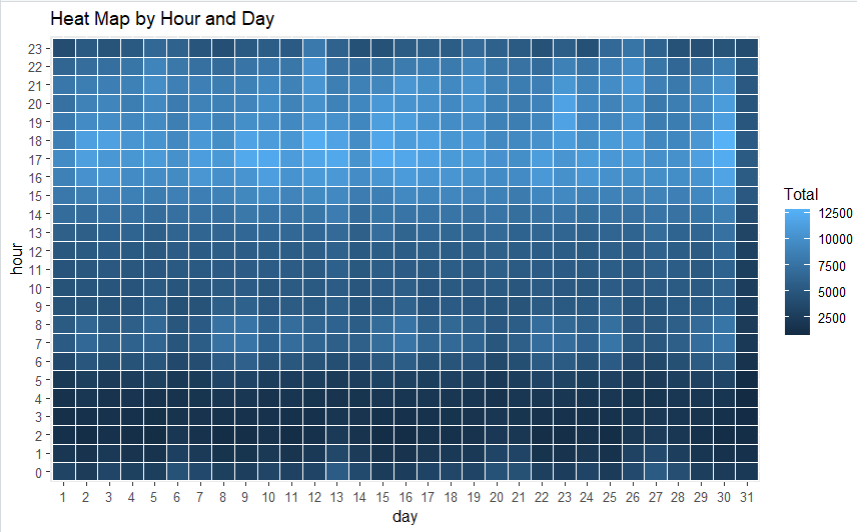


[**Heatmap**](https://en.wikipedia.org/wiki/Heat_map)**by Hour and Day.**

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

*geom\_tile(color = "white") +*

*ggtitle("Heat Map by Hour and Day")*

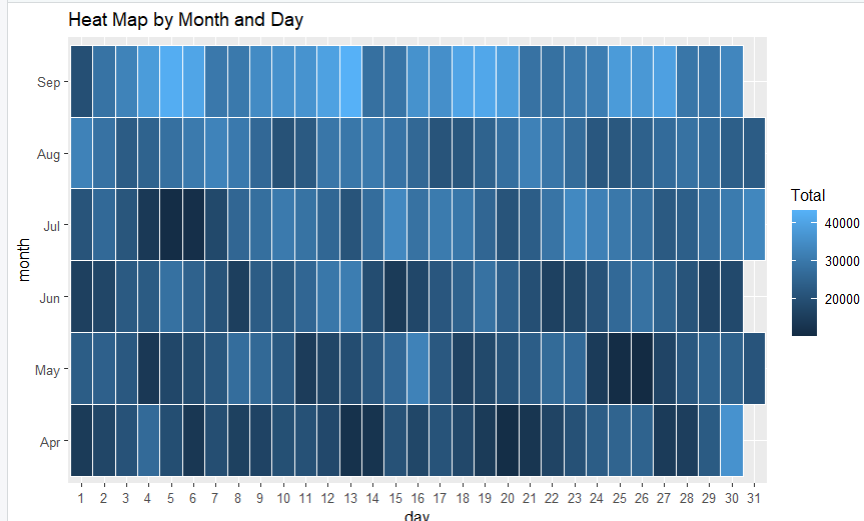


**Heatmap by Month and Day**

*ggplot(day\_month\_group, aes(day, month, fill = Total)) +*

*geom\_tile(color = "white") +*

*ggtitle("Heat Map by Month and Day")*

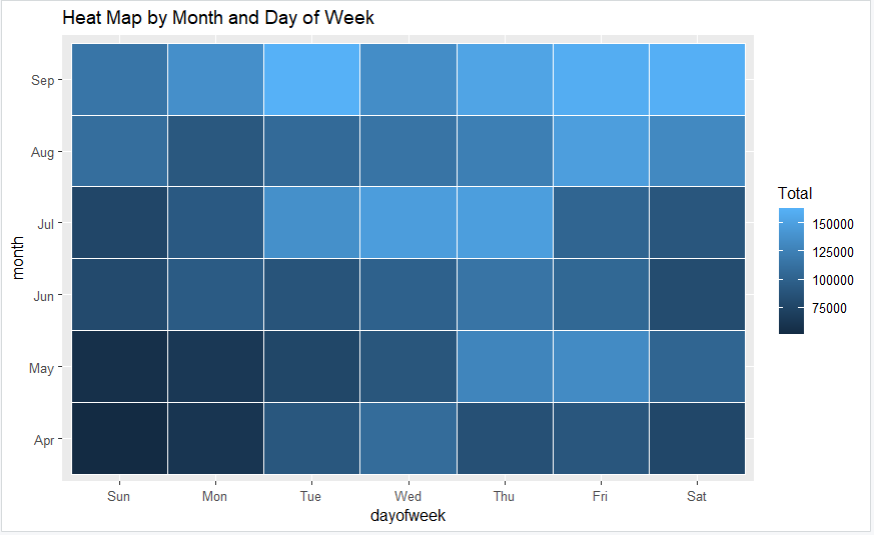


**Heatmap by Month and Day of the Week.**

*ggplot(month\_weekday, aes(dayofweek, month, fill = Total)) +*

*geom\_tile(color = "white") +*

*ggtitle("Heat Map by Month and Day of Week")*

**

**Heatmap that delineates Month and Bases.**

*month\_base <- data\_2014 %>%*

*group\_by(Base, month) %>%*

*dplyr::summarize(Total = n())*

**Heatmap, by bases and day of the week.**

*day0fweek\_bases <- data\_2014 %>%*

*group\_by(Base, dayofweek) %>%*

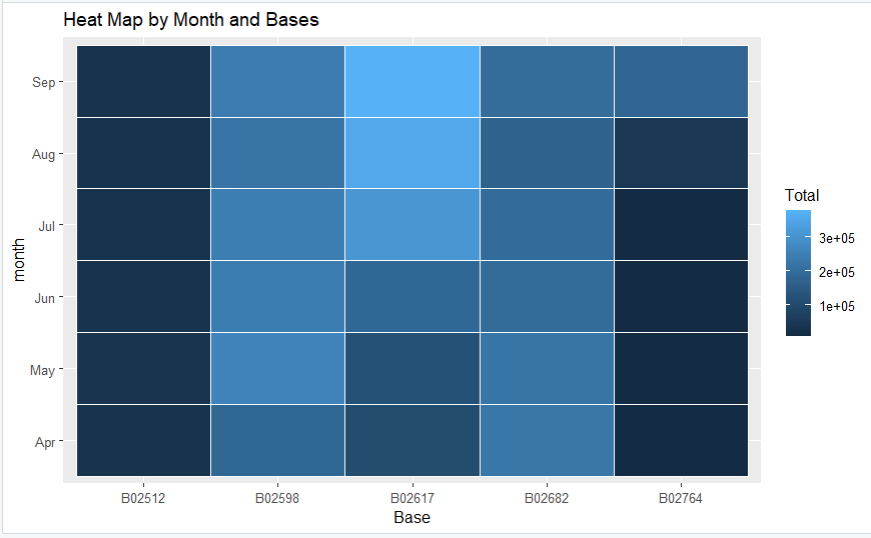
*dplyr::summarize(Total = n())*

#plot it

*ggplot(month\_base, aes(Base, month, fill = Total)) +*

*geom\_tile(color = "white") +*

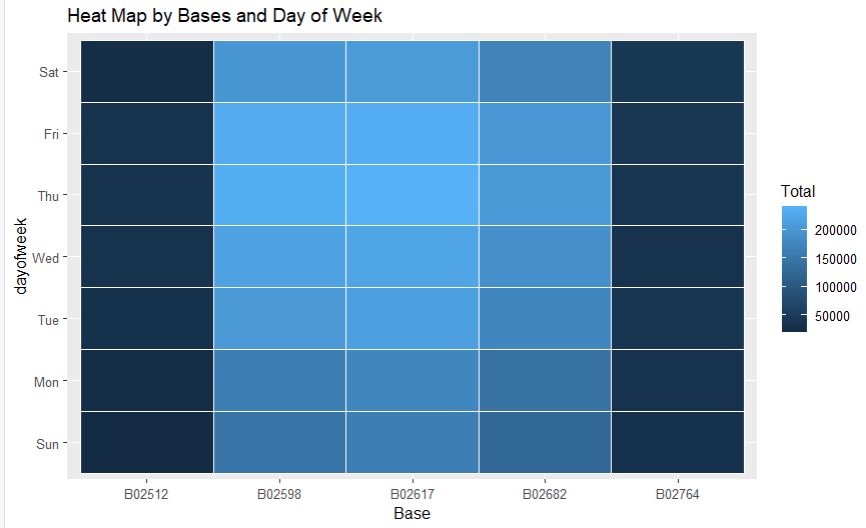
*ggtitle("Heat Map by Month and Bases")*



*ggplot(day0fweek\_bases, aes(Base, dayofweek, fill = Total)) +*

*geom\_tile(color = "white") +*

*ggtitle("Heat Map by Bases and Day of Week")*

**

**Creating a map visualization of rides in New York**

In the final section, we will visualize the rides in New York city by creating a geo-plot that will help us to visualize the rides during 2014 (Apr – Sep) and by the bases in the same period.

*min\_lat <- 40.5774*

*max\_lat <- 40.9176*

*min\_long <- -74.15*

*max\_long <- -73.7004*

*#geo-plot based on uber rides in 2014*

*ggplot(data\_2014, 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)")*

*ggplot(data\_2014, aes(x=Lon, y=Lat, color = Base)) +*

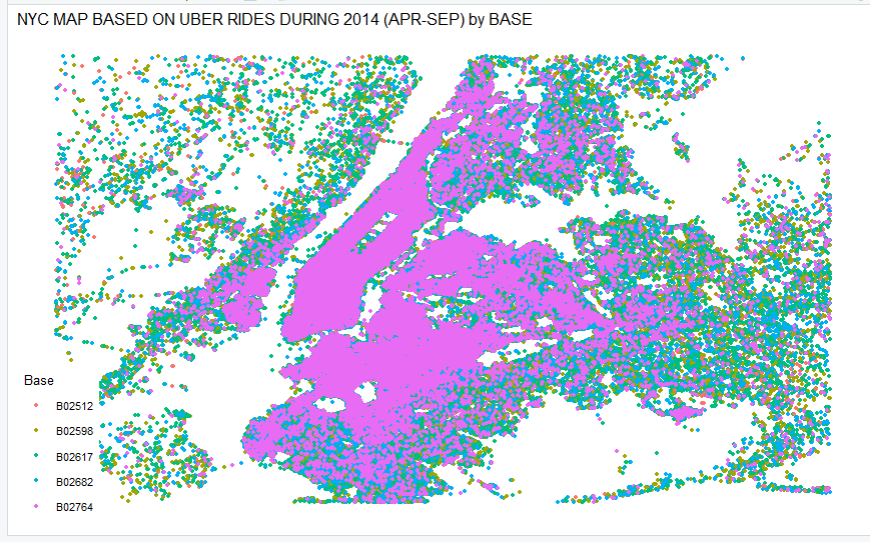
*geom\_point(size=1) +*

*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) by BASE")*

**