# UBER DATA ANALYSIS PROJECT

**AYUSH SRIVASTAVA**

In this R project, I will analyze the ***Uber Pickups in New York City dataset***. This is more of a data visualization project that will guide towards using the ggplot2 library for understanding the data and for developing an intuition for understanding the customers who avail the trips.

Talking about our Uber data analysis project, data storytelling is an important component of Machine Learning through which companies are able to understand the background of various operations. With the help of visualization, companies can avail the benefit of understanding the complex data and gain insights that would help them to craft decisions

# 1. Importing the Essential Packages

In the first step of our R project, we will import the essential packages that I will use in this uber data analysis project. Some of the ***important libraries of R*** that we will use are –

- 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.

- 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.

- dplyr

This package is the lingua franca of ***data manipulation in R***.

- tidyr

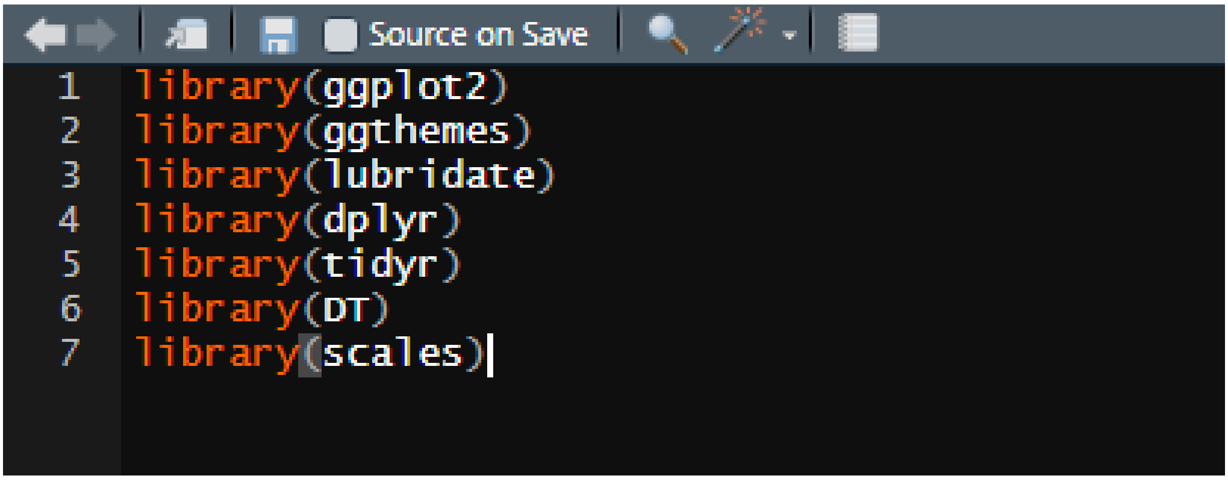
This package will help you to tidy your data. The basic principle of tidyr is to tidy the columns where each variable is present in a column, each observation is represented by a row and each value depicts a cell.

- DT

With the help of this package, we will be able to interface with the ***Java Script*** Library called – Datatables.

- scales

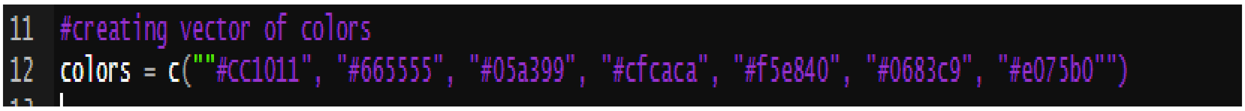
With the help of graphical scales, we can automatically map the data to the correct scales with well-placed axes and legends.



```
library(ggplot2)
library(ggthemes)
library(lubridate)
library(dplyr)
library(tidyr)
library(DT)
library(scales)
```

## 2. Creating vector of colours to be implemented in our plots

In this step of data science project, we will create a vector of our colours that will be included in our plotting functions.

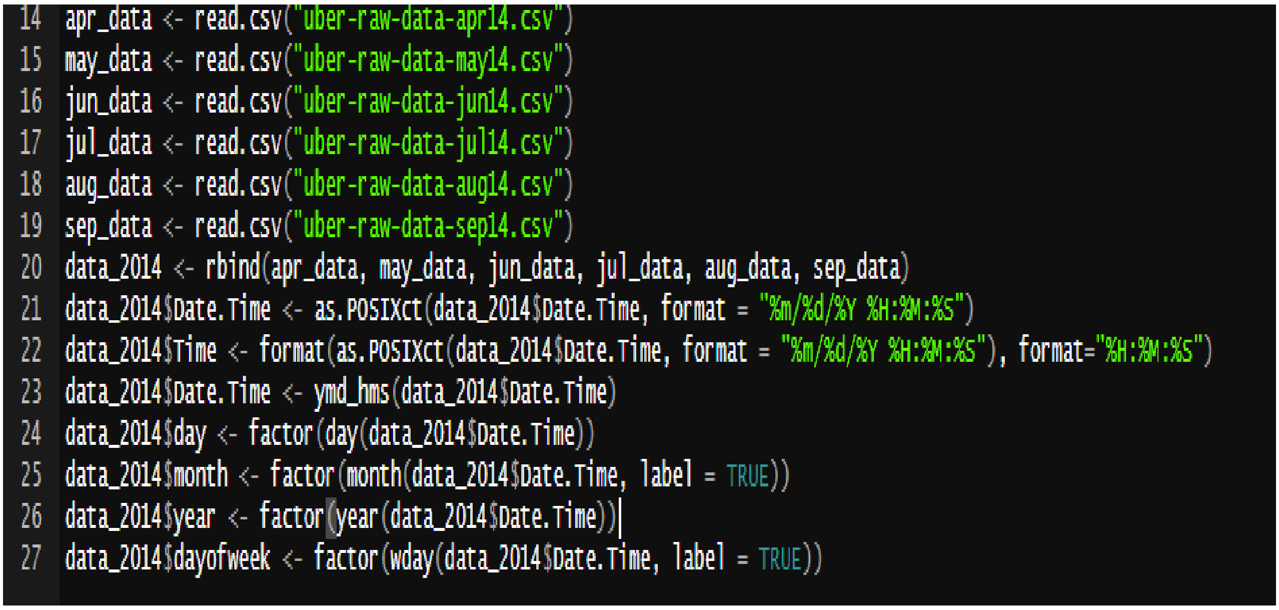


```
#creating vector of colors
colors = c(""#CC1011", "#665555", "#05a399", "#cfcaca", "#f5e840", "#0683c9", "#e075b0"")
```

## 3. Reading the Data into their designated variables

Now, I 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. After we have read the files, we will combine all of this data into a single dataframe called 'data_2014'.

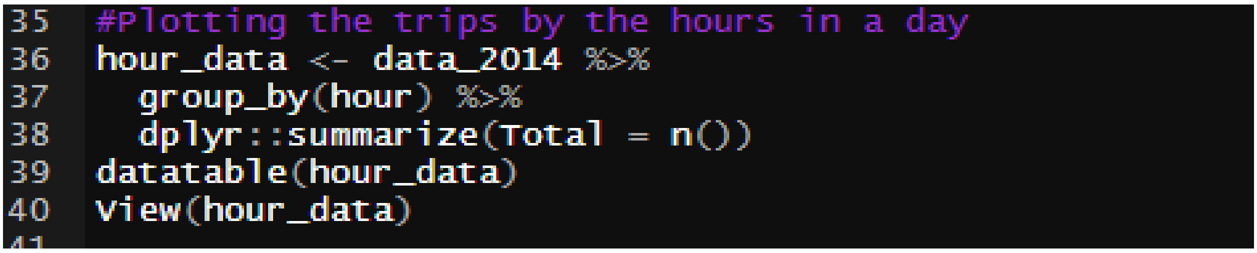
Then, in the next step, we will perform the appropriate formatting of Date.Time column. Then, we will proceed to create factors of time objects like day, month, year etc.



```
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")
data_2014 <- rbind(apr_data, may_data, jun_data, jul_data, aug_data, sep_data)
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))
```

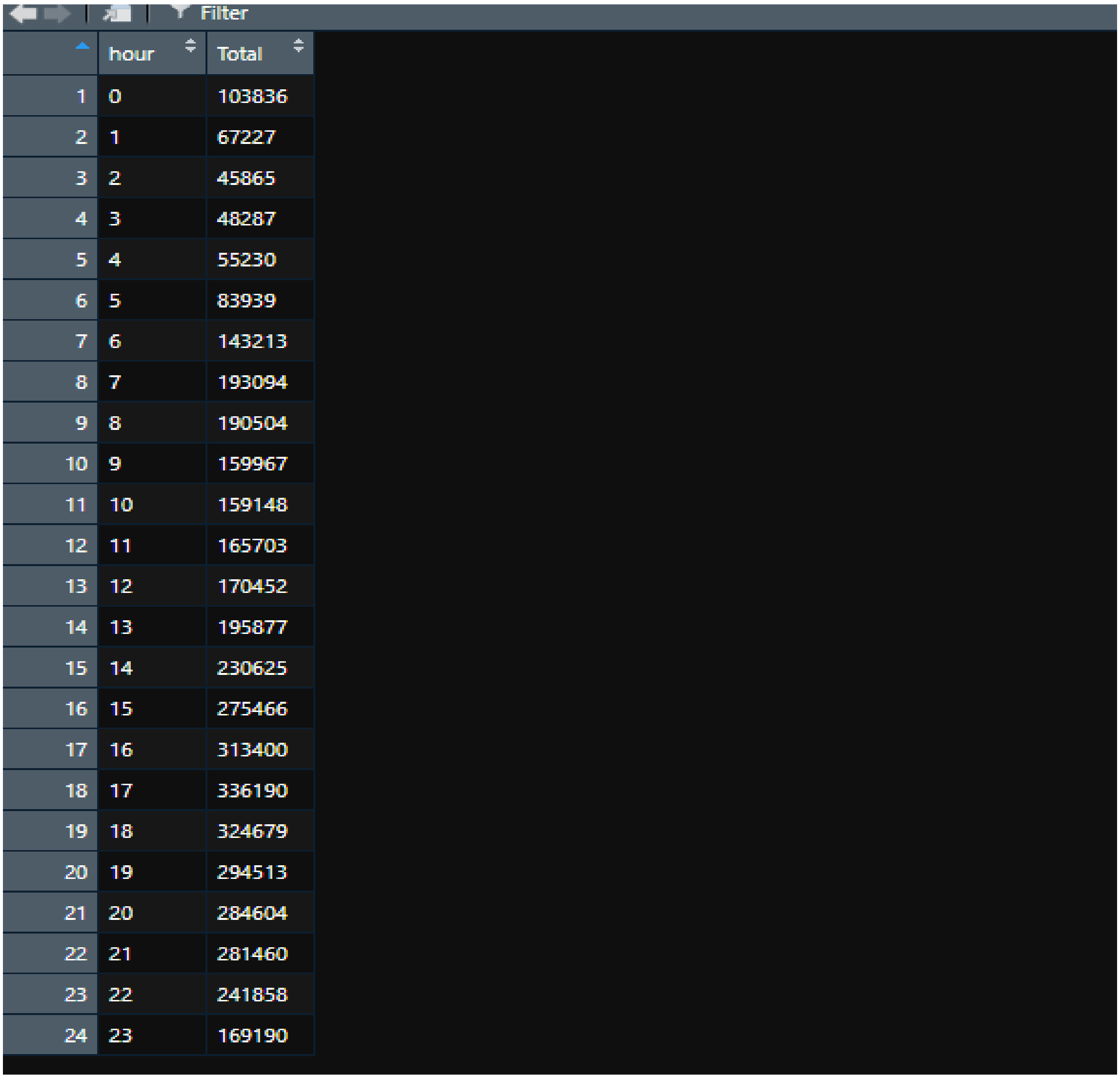
# Plotting the trips by the hours in a day

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, I understand how the number of passengers fares throughout the day. I observe that the number of trips are higher in the evening around 5:00 and 6:00 PM.

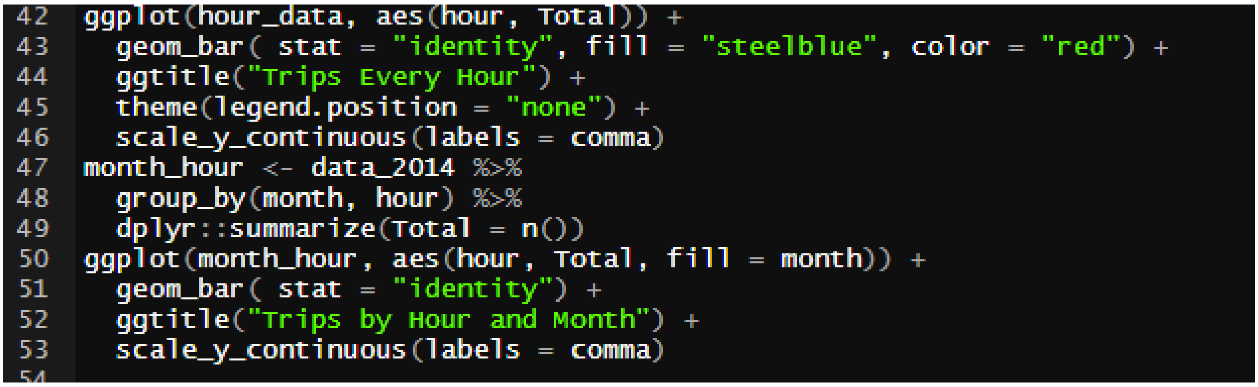


```
#Plotting the trips by the hours in a day
hour_data <- data_2014 %>%
  group_by(hour) %>%
  dplyr::summarize(Total = n())
datatable(hour_data)
view(hour_data)
```

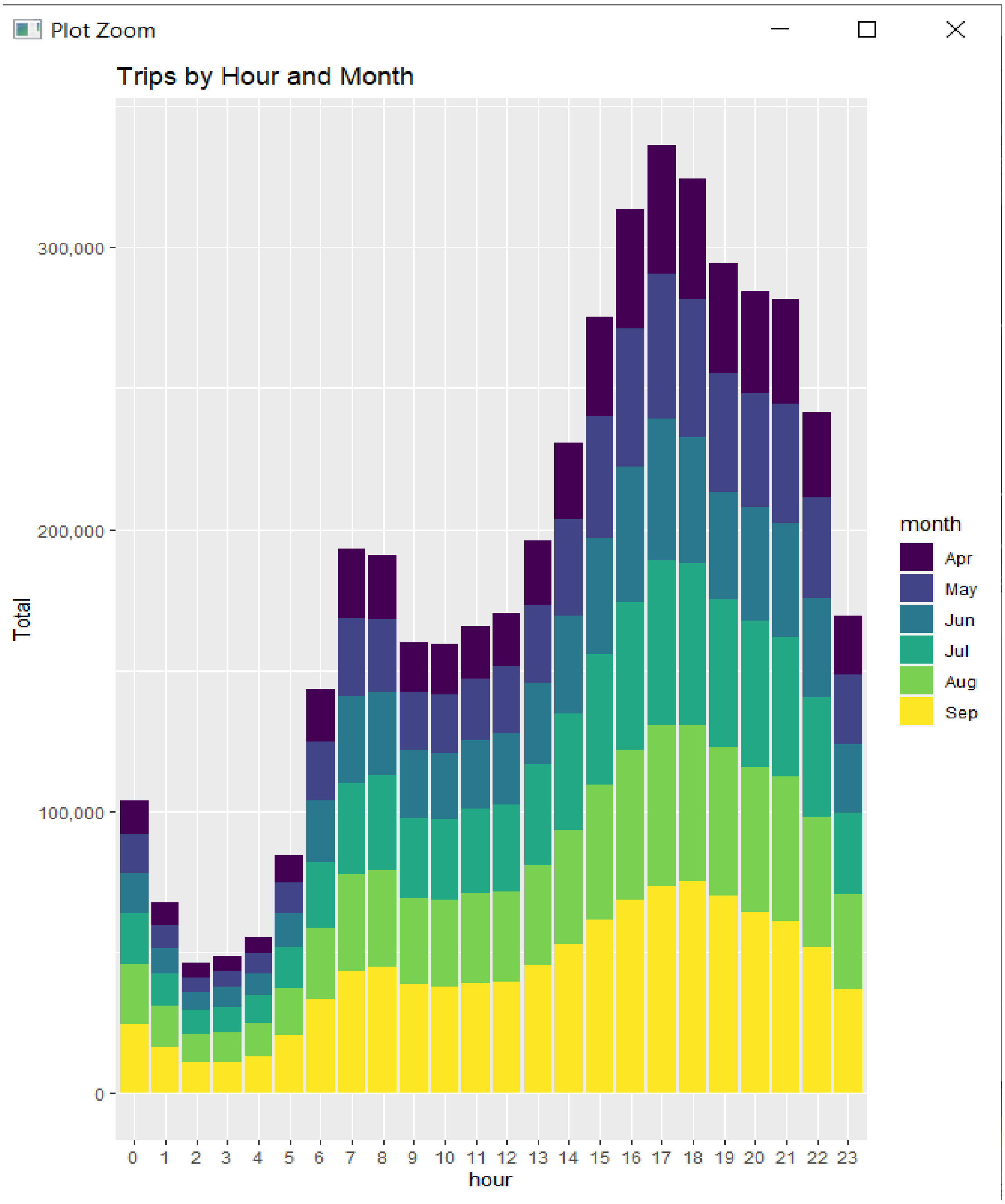
OUTPUT:



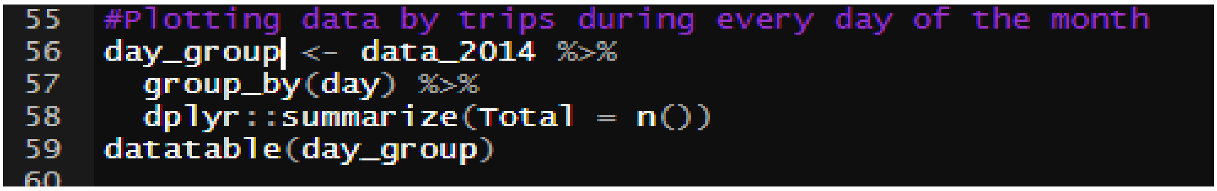
| | hour | Total |
|---|---|---|
| 1 | 0 | 103836 |
| 2 | 1 | 67227 |
| 3 | 2 | 45865 |
| 4 | 3 | 48287 |
| 5 | 4 | 55230 |
| 6 | 5 | 83939 |
| 7 | 6 | 143213 |
| 8 | 7 | 193094 |
| 9 | 8 | 190504 |
| 10 | 9 | 159967 |
| 11 | 10 | 159148 |
| 12 | 11 | 165703 |
| 13 | 12 | 170452 |
| 14 | 13 | 195877 |
| 15 | 14 | 230625 |
| 16 | 15 | 275466 |
| 17 | 16 | 313400 |
| 18 | 17 | 336190 |
| 19 | 18 | 324679 |
| 20 | 19 | 294513 |
| 21 | 20 | 284604 |
| 22 | 21 | 281460 |
| 23 | 22 | 241858 |
| 24 | 23 | 169190 |



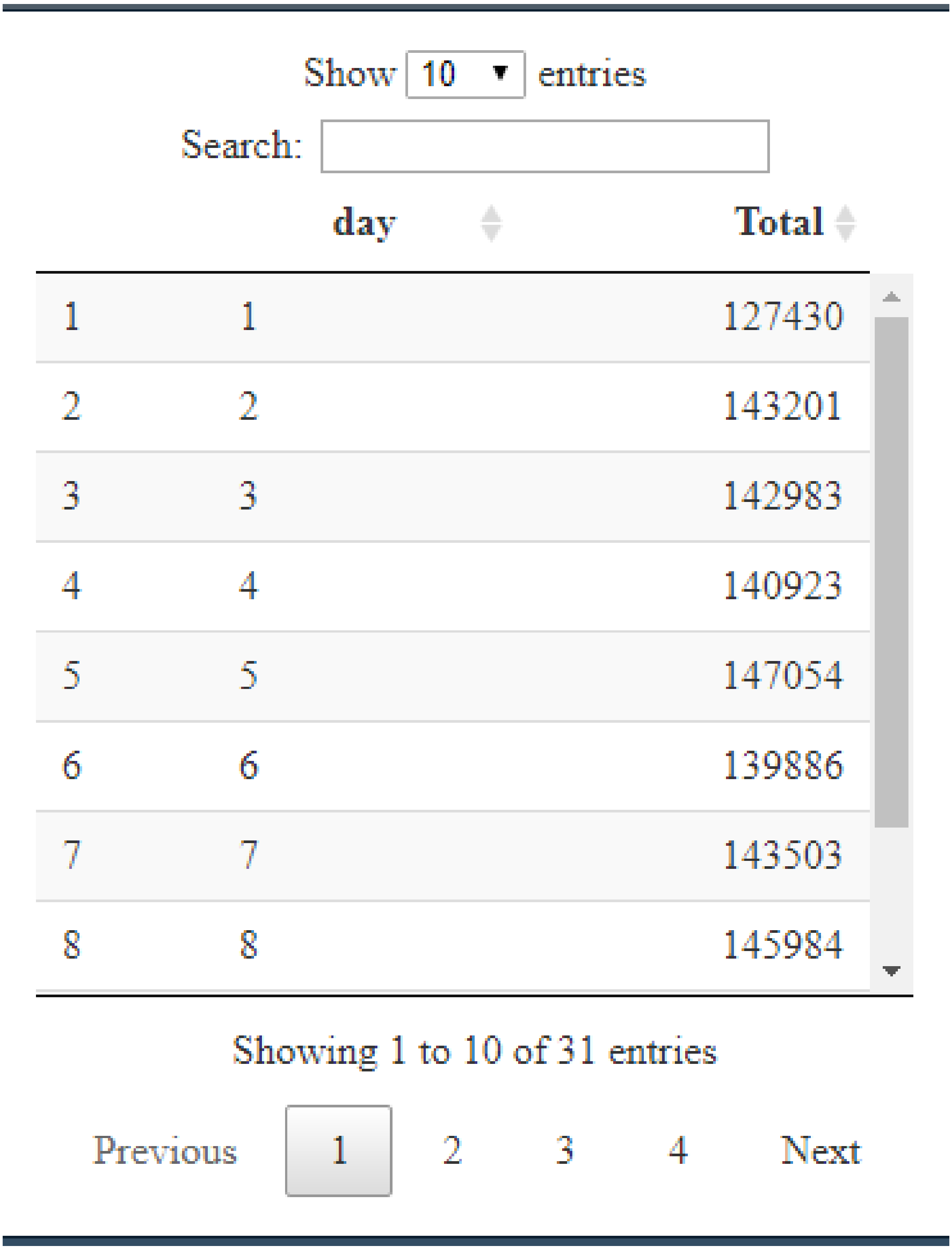
```
ggplot(hour_data, aes(hour, Total)) +
  geom_bar( stat = "identity", fill = "steelblue", color = "red") +
  ggtitle("Trips Every Hour") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = comma)
month_hour <- data_2014 %>%
  group_by(month, hour) %>%
  dplyr::summarize(Total = n())
ggplot(month_hour, 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



```
#Plotting data by trips during every day of the month
day_group <- data_2014 %>%
  group_by(day) %>%
  dplyr::summarize(Total = n())
datatable(day_group)
```

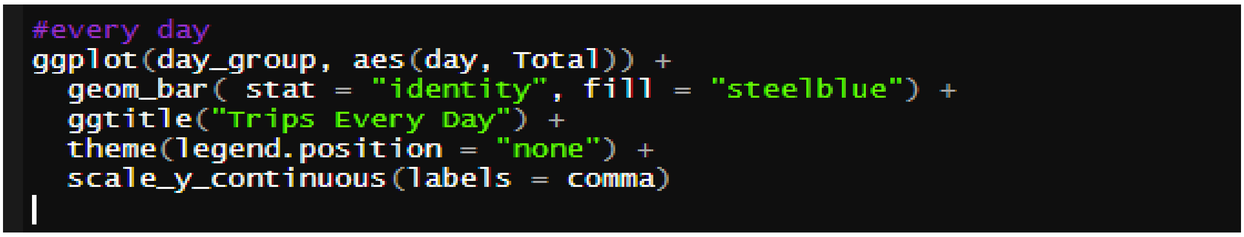


Show 10 entries

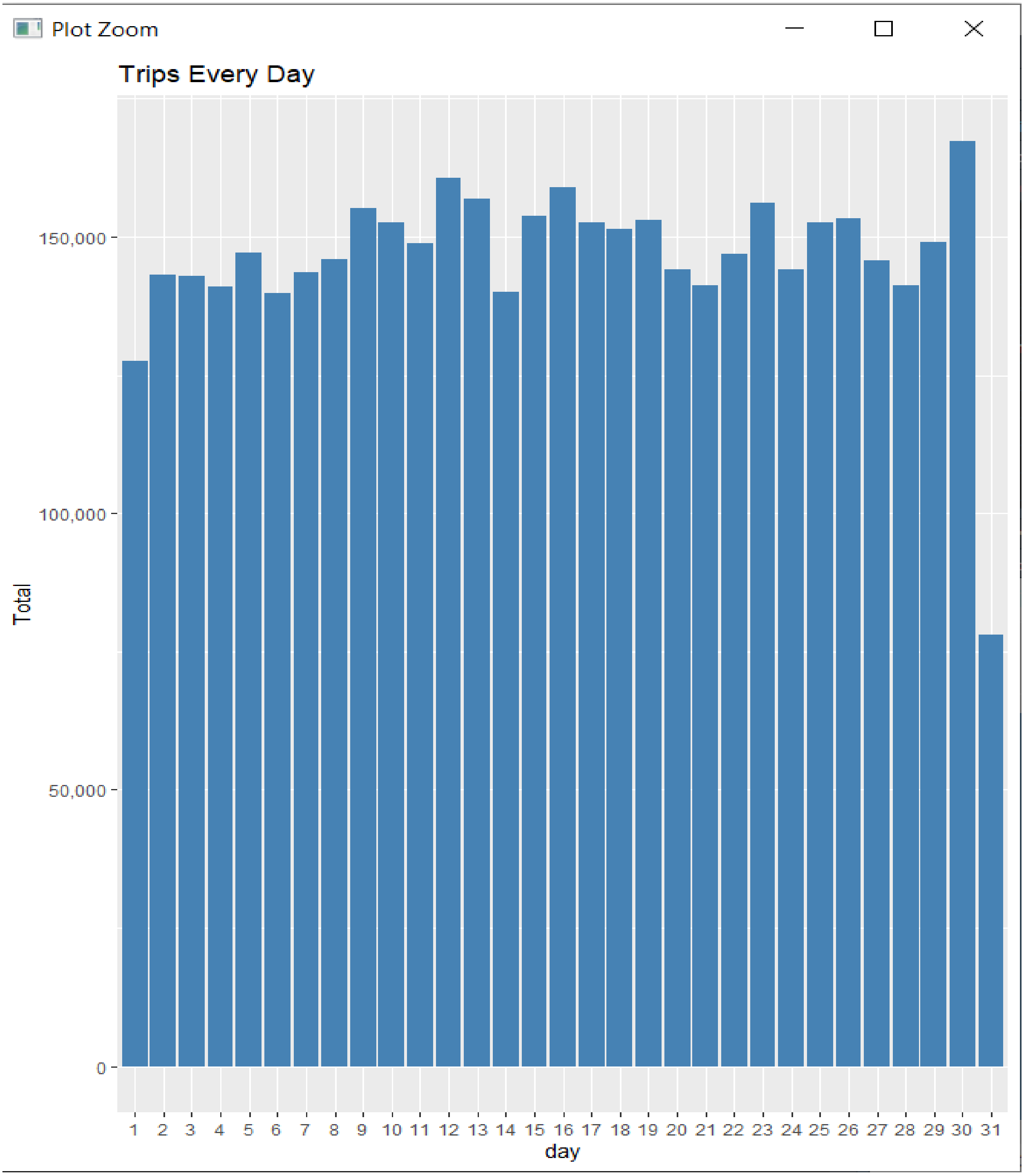
Search:

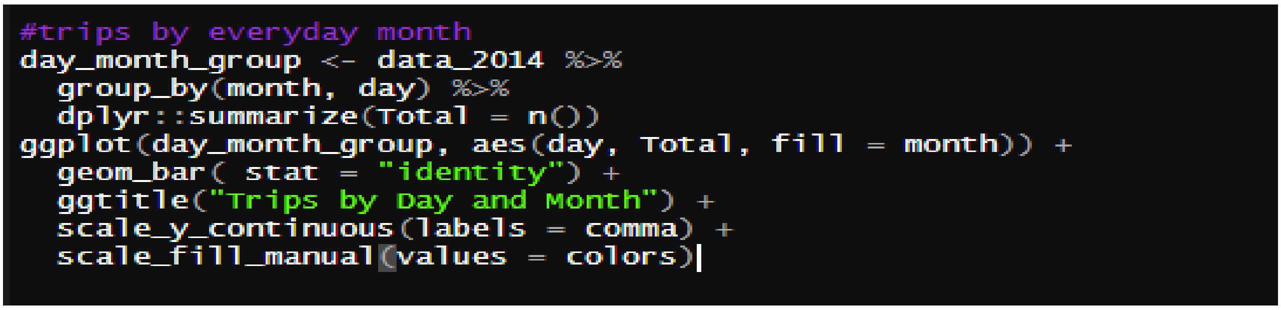
| | day | Total |
|---|---|---|
| 1 | 1 | 127430 |
| 2 | 2 | 143201 |
| 3 | 3 | 142983 |
| 4 | 4 | 140923 |
| 5 | 5 | 147054 |
| 6 | 6 | 139886 |
| 7 | 7 | 143503 |
| 8 | 8 | 145984 |

Showing 1 to 10 of 31 entries

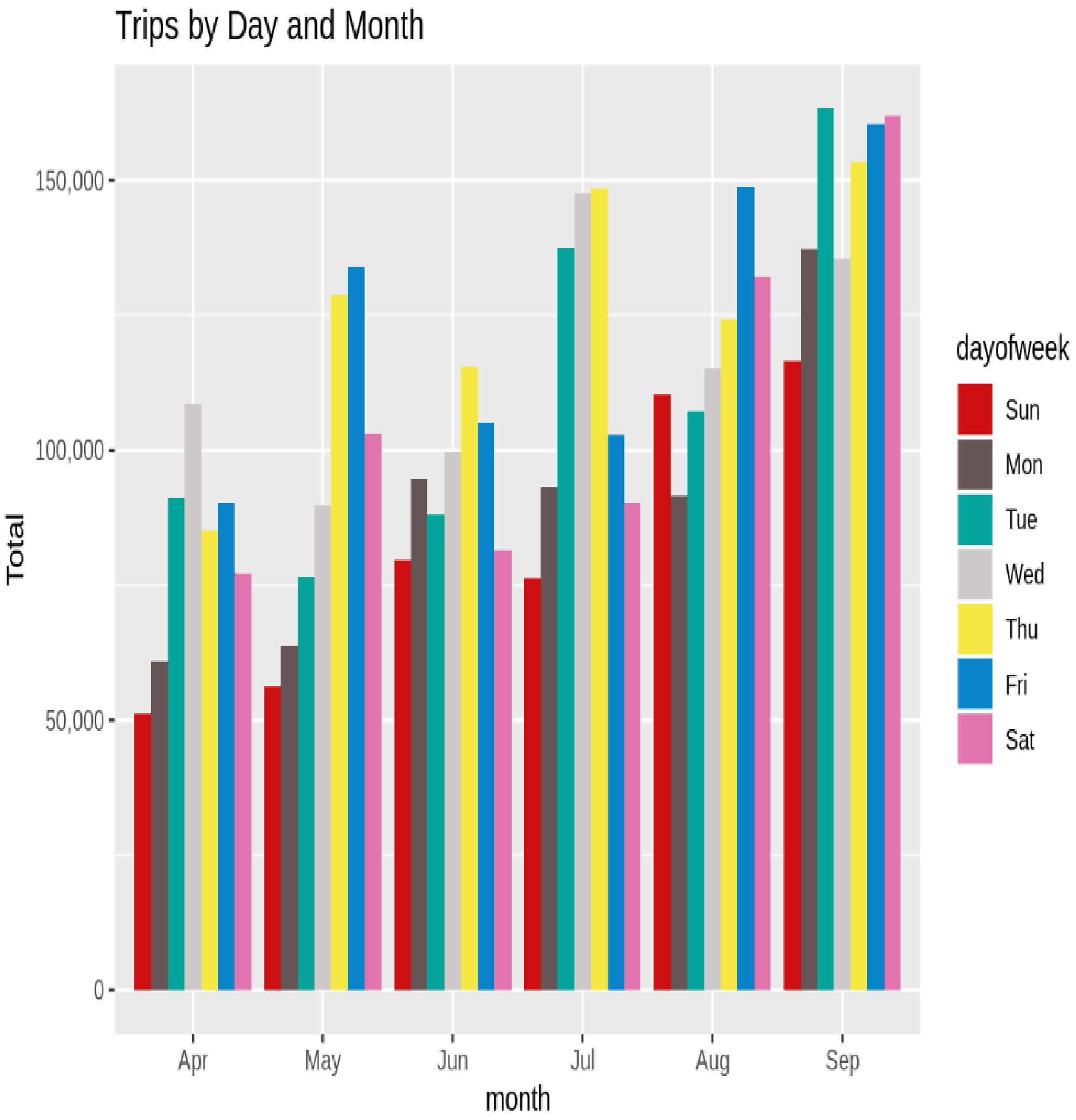


```r
#every day
ggplot(day_group, aes(day, Total)) +
  geom_bar( stat = "identity", fill = "steelblue") +
  ggtitle("Trips Every Day") +
  theme(legend.position = "none") +
  scale_y_continuous(labels = comma)
```



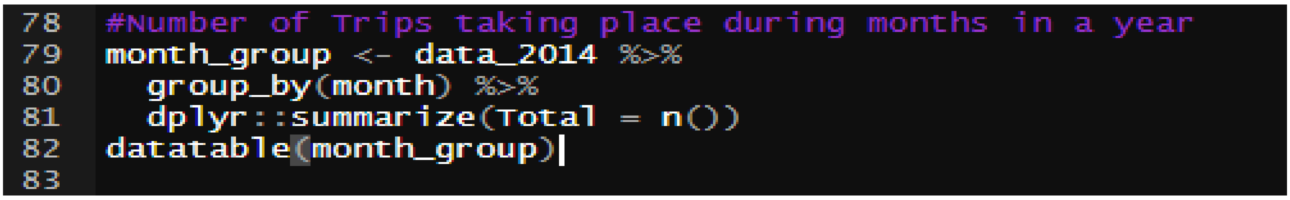


```
#trips by everyday month
day_month_group <- data_2014 %>%
  group_by(month, day) %>%
  dplyr::summarize(Total = n())
ggplot(day_month_group, aes(day, Total, fill = month)) +
  geom_bar( stat = "identity") +
  ggtitle("Trips by Day and Month") +
  scale_y_continuous(labels = comma) +
  scale_fill_manual(values = colors)
```

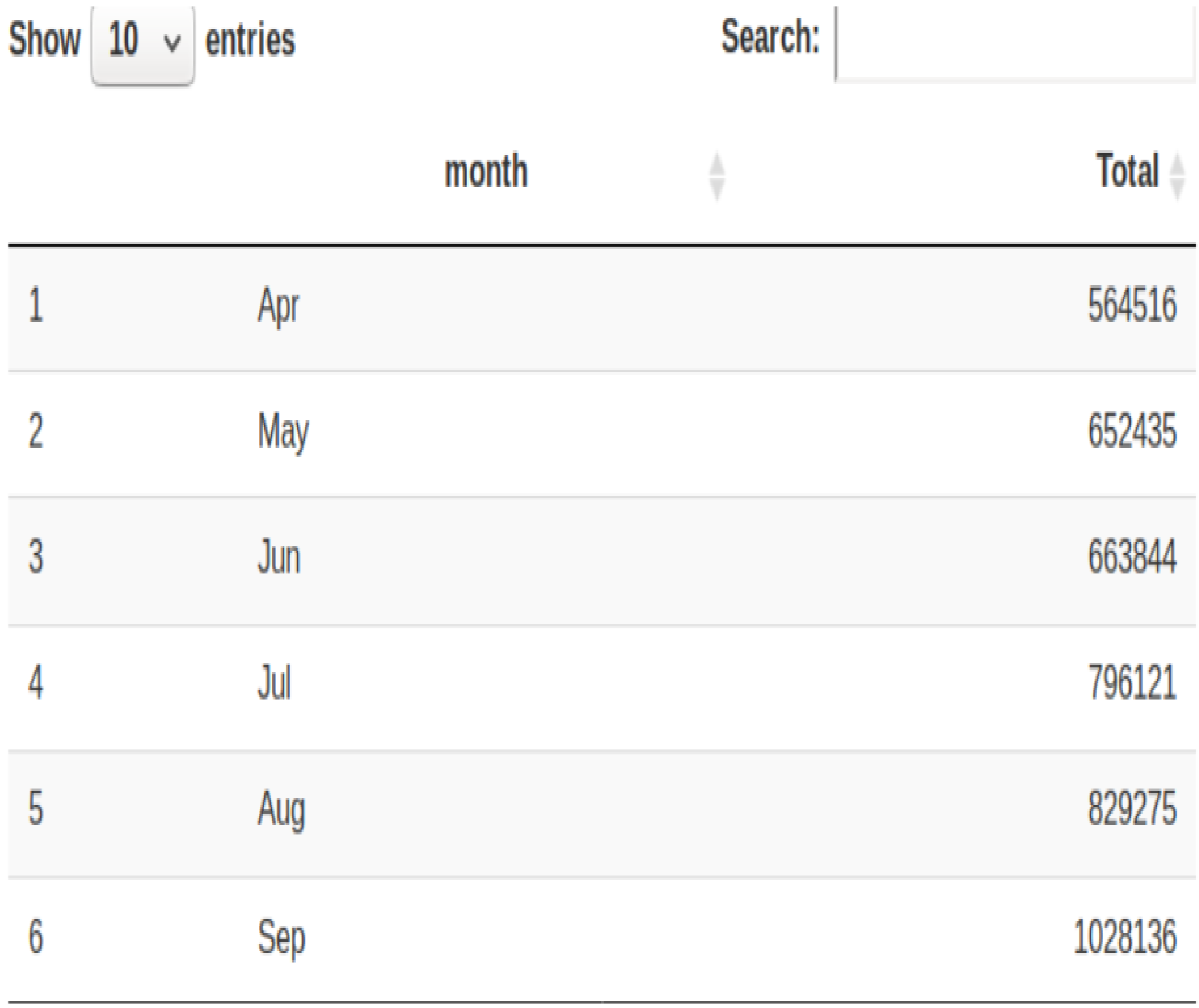


# Number of Trips taking place during months in a year

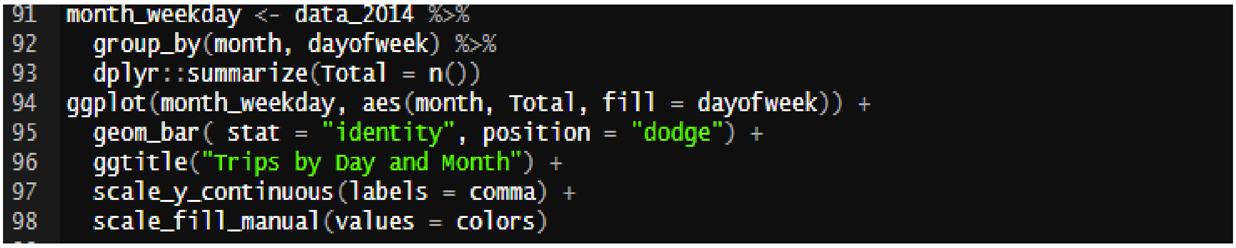
visualize the number of trips that are taking place each month of the year. In the output visualization, I observe that most trips were made during the month of September. Furthermore, I also obtain visual reports of the number of trips that were made on every day of the week.



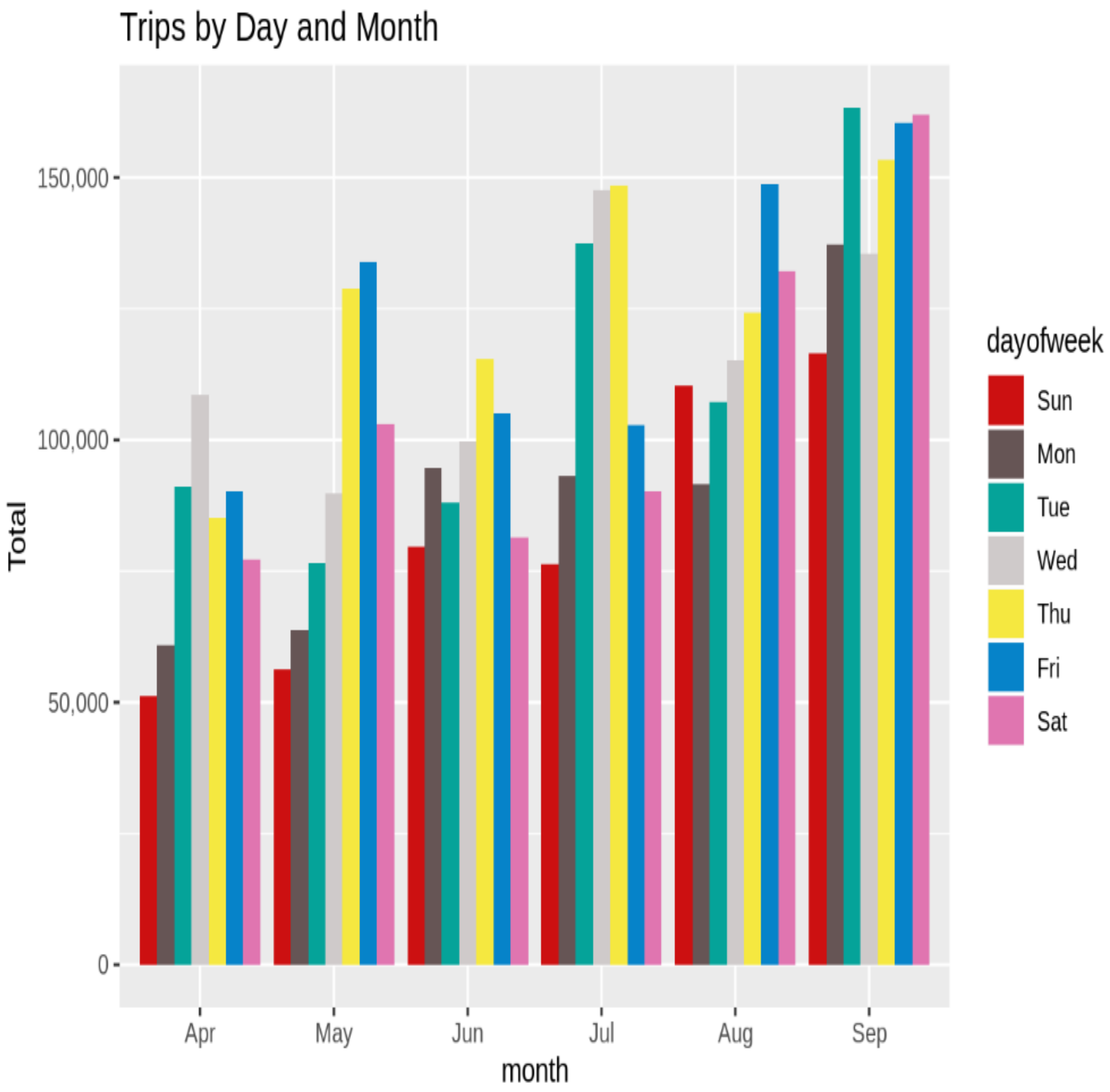
```
#Number of Trips taking place during months in a year
month_group <- data_2014 %>%
  group_by(month) %>%
  dplyr::summarize(Total = n())
datatable(month_group)
```



| | month | Total |
|---|---|---|
| 1 | Apr | 564516 |
| 2 | May | 652435 |
| 3 | Jun | 663844 |
| 4 | Jul | 796121 |
| 5 | Aug | 829275 |
| 6 | Sep | 1028136 |

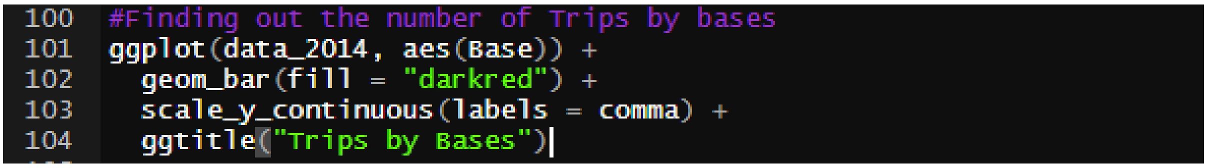


```
month_weekday <- data_2014 %>%
  group_by(month, dayofweek) %>%
  dplyr::summarize(Total = n())
ggplot(month_weekday, aes(month, Total, fill = dayofweek)) +
  geom_bar( stat = "identity", position = "dodge") +
  ggtitle("Trips by Day and Month") +
  scale_y_continuous(labels = comma) +
  scale_fill_manual(values = colors)
```

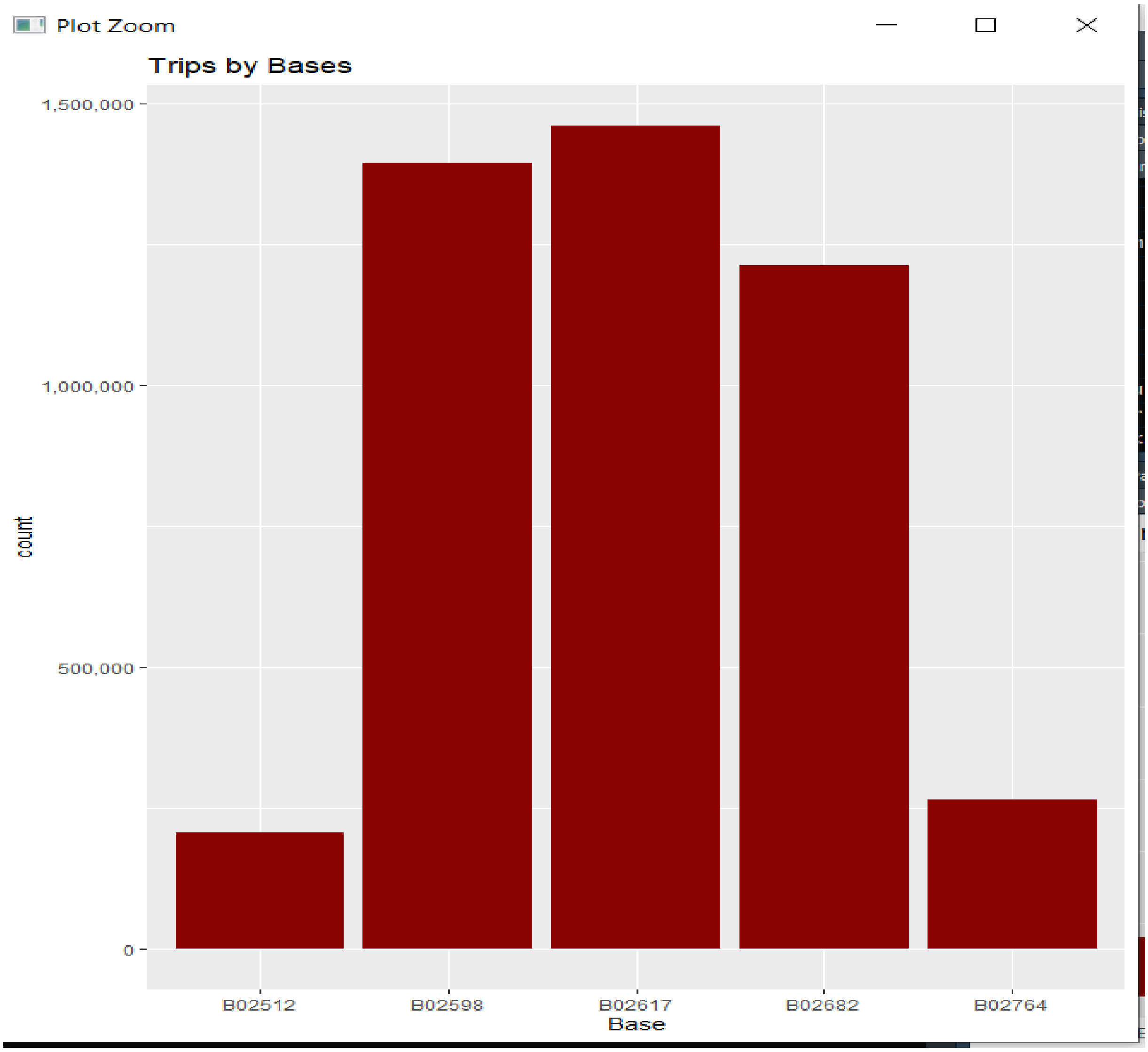


# Finding out the number of Trips by bases

In the following visualization, I 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.

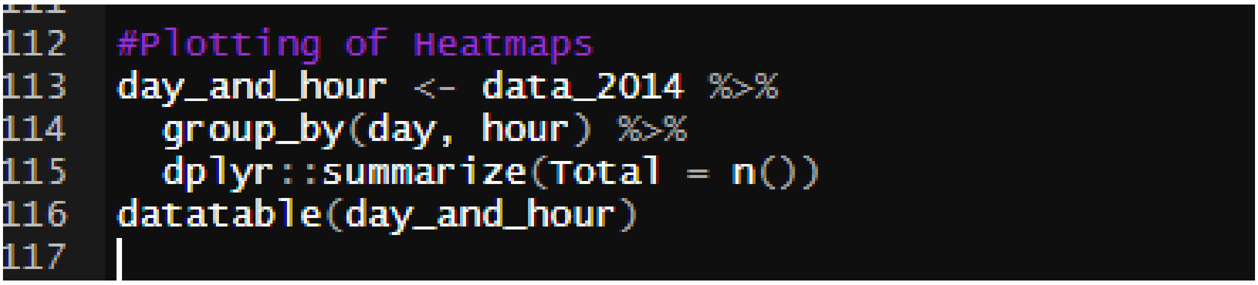


```
#Finding out the number of Trips by bases
ggplot(data_2014, aes(Base)) +
  geom_bar(fill = "darkred") +
  scale_y_continuous(labels = comma) +
  ggtitle("Trips by Bases")
```

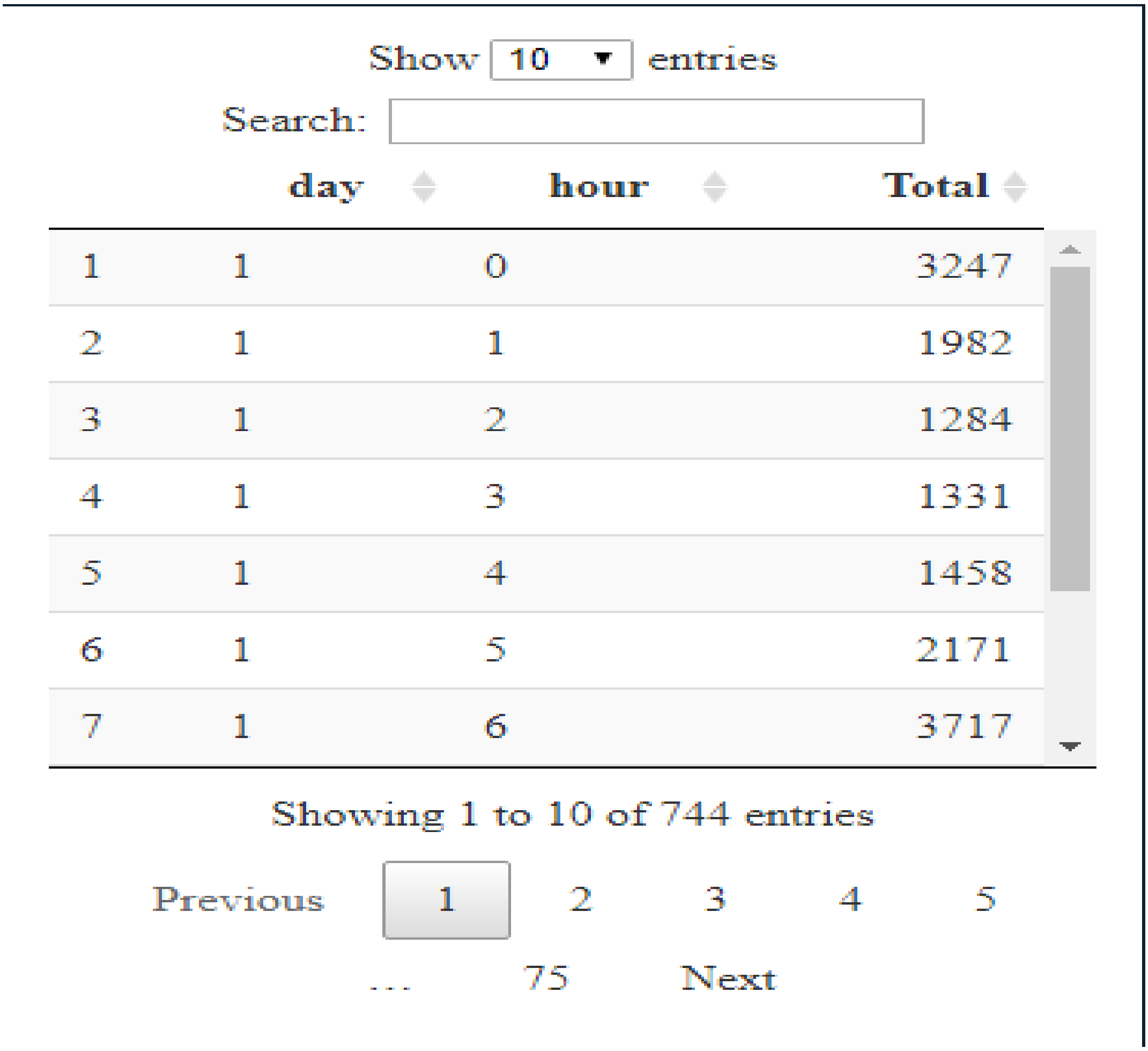


# Creating a Heatmap visualization of day, hour and month

In this section, will plot heatmaps using ggplot().



```
#Plotting of Heatmaps
day_and_hour <- data_2014 %>%
  group_by(day, hour) %>%
  dplyr::summarize(Total = n())
datatable(day_and_hour)
```

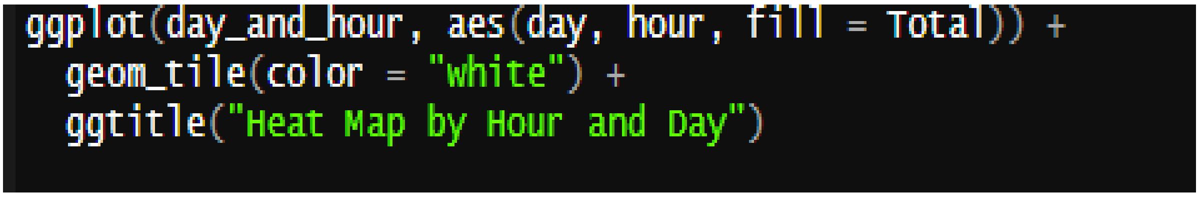


Show 10 ▾ entries

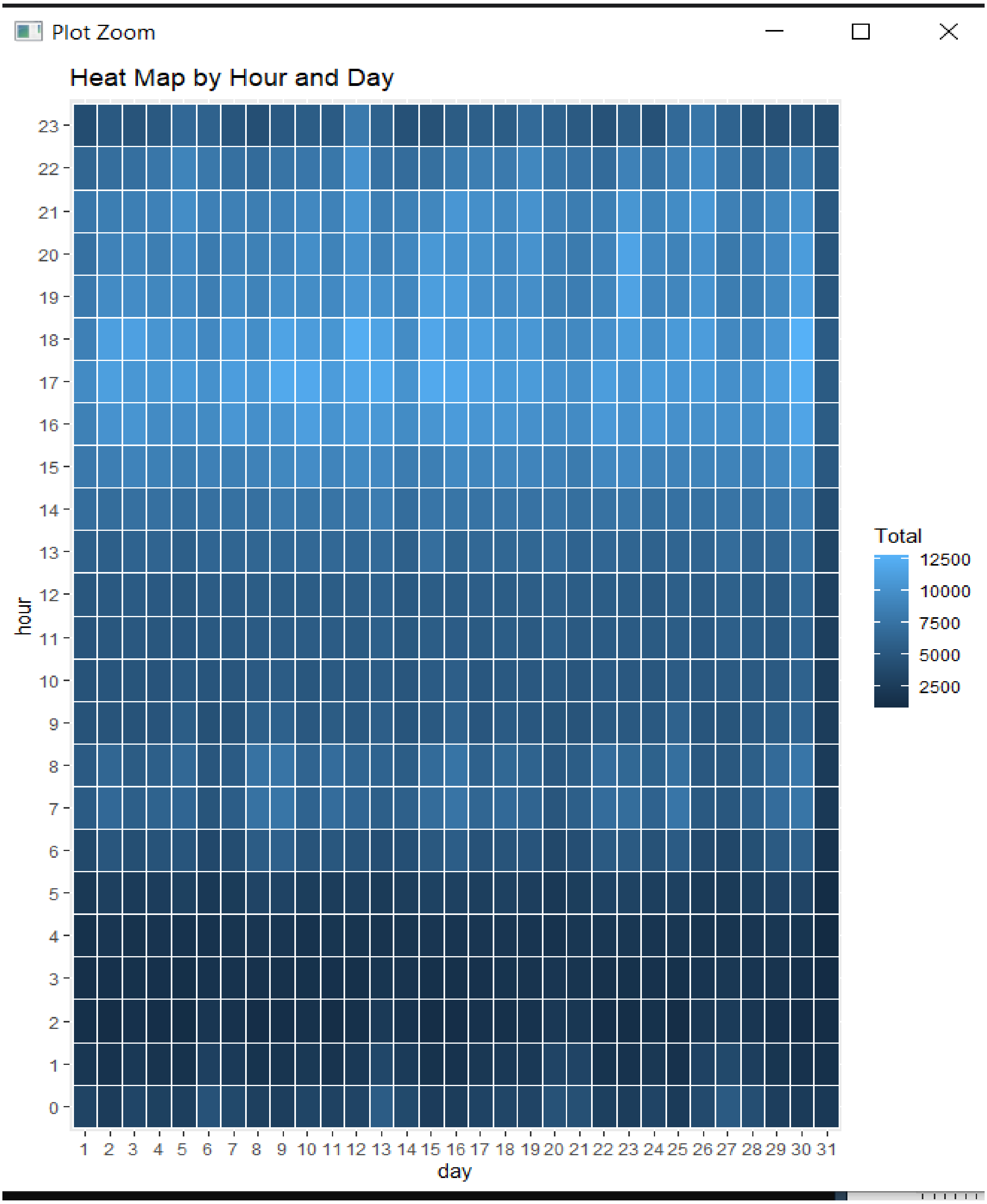
Search:

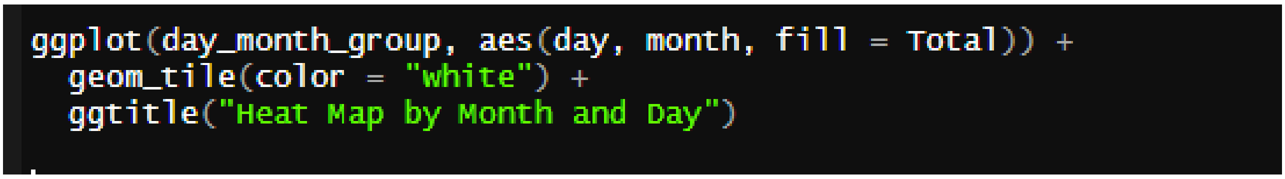
|  | day | hour | Total |
|---|---|---|---|
| 1 | 1 | 0 | 3247 |
| 2 | 1 | 1 | 1982 |
| 3 | 1 | 2 | 1284 |
| 4 | 1 | 3 | 1331 |
| 5 | 1 | 4 | 1458 |
| 6 | 1 | 5 | 2171 |
| 7 | 1 | 6 | 3717 |

Showing 1 to 10 of 744 entries

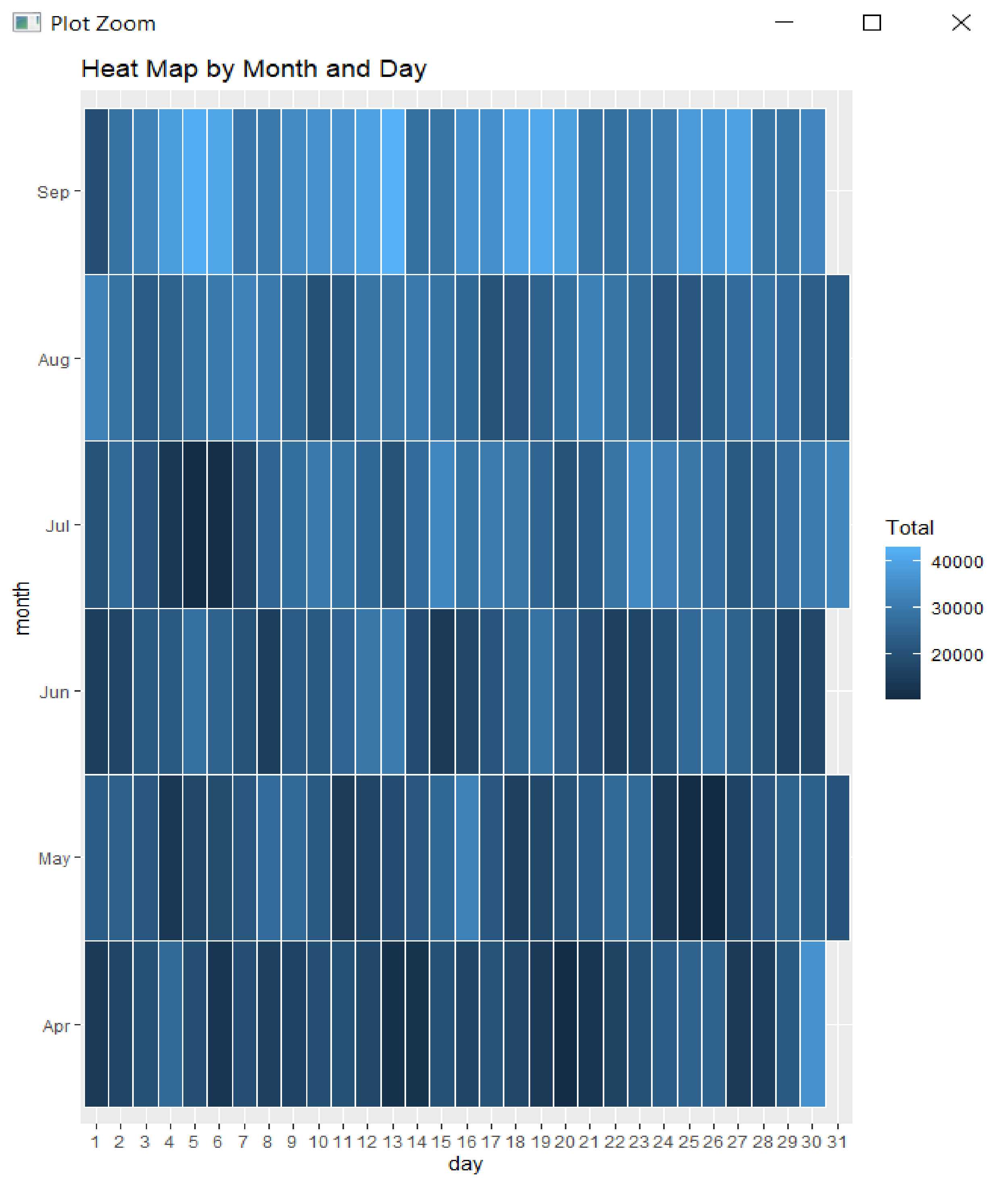


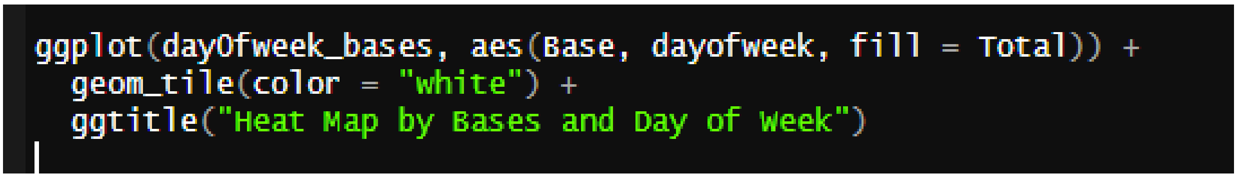
```
ggplot(day_and_hour, aes(day, hour, fill = Total)) +
  geom_tile(color = "white") +
  ggtitle("Heat Map by Hour and Day")
```



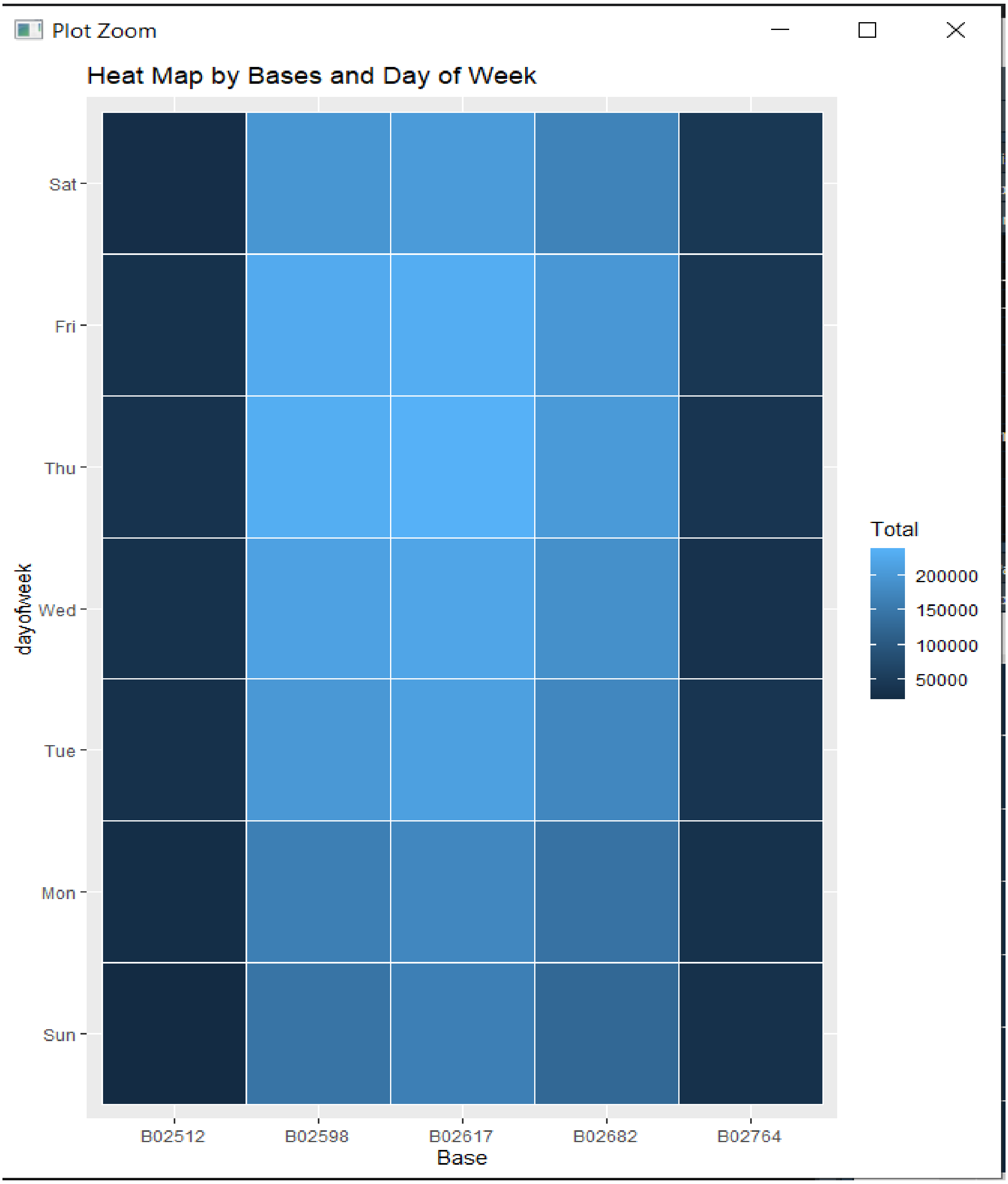


```
ggplot(day_month_group, aes(day, month, fill = Total)) +
  geom_tile(color = "white") +
  ggtitle("Heat Map by Month and Day")
```



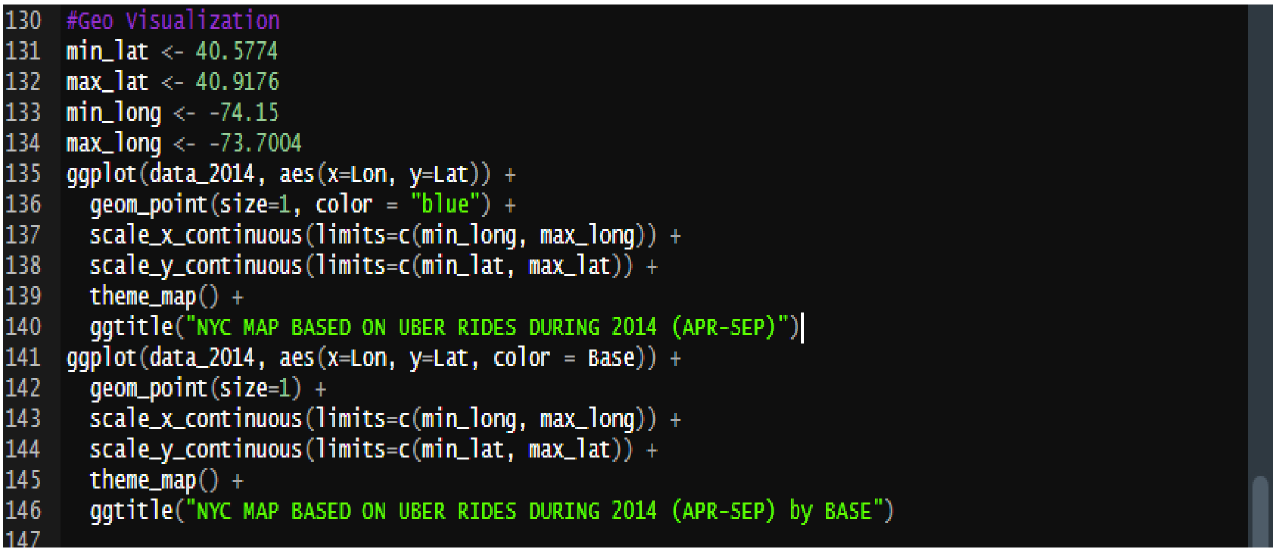


```
ggplot(dayOfweek_bases, aes(Base, dayofweek, fill = Total)) +
  geom_tile(color = "white") +
  ggtitle("Heat Map by Bases and Day of Week")
```

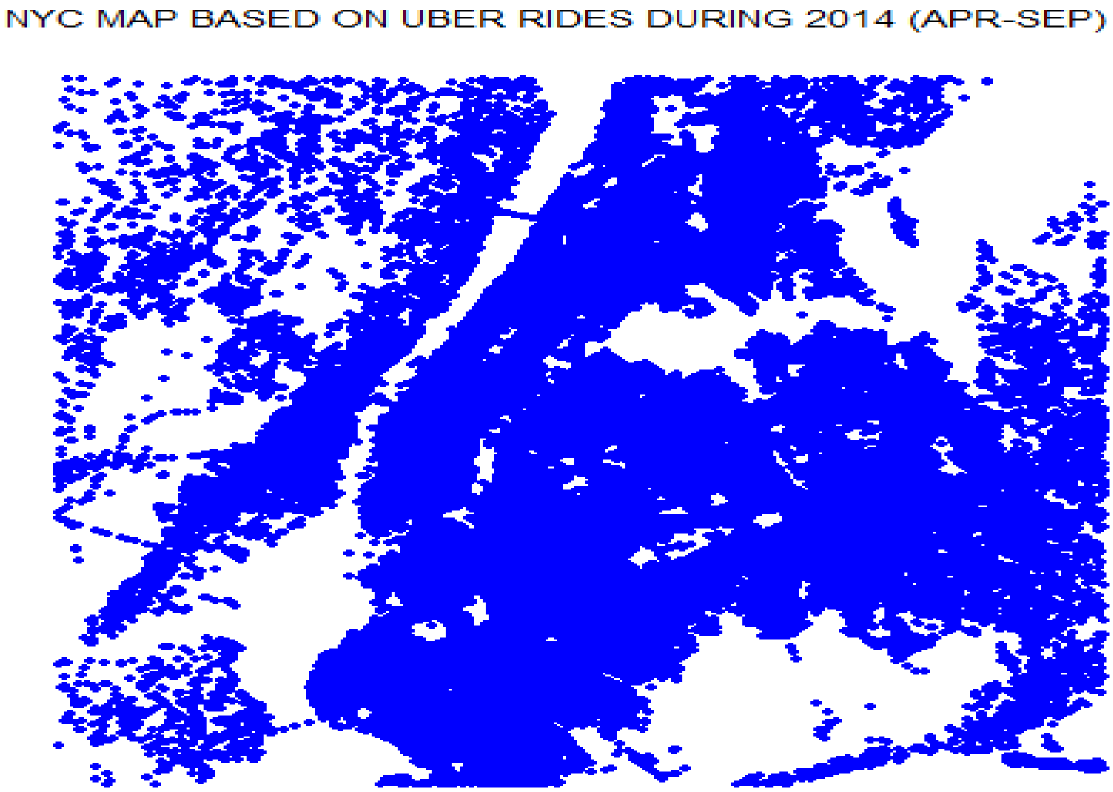


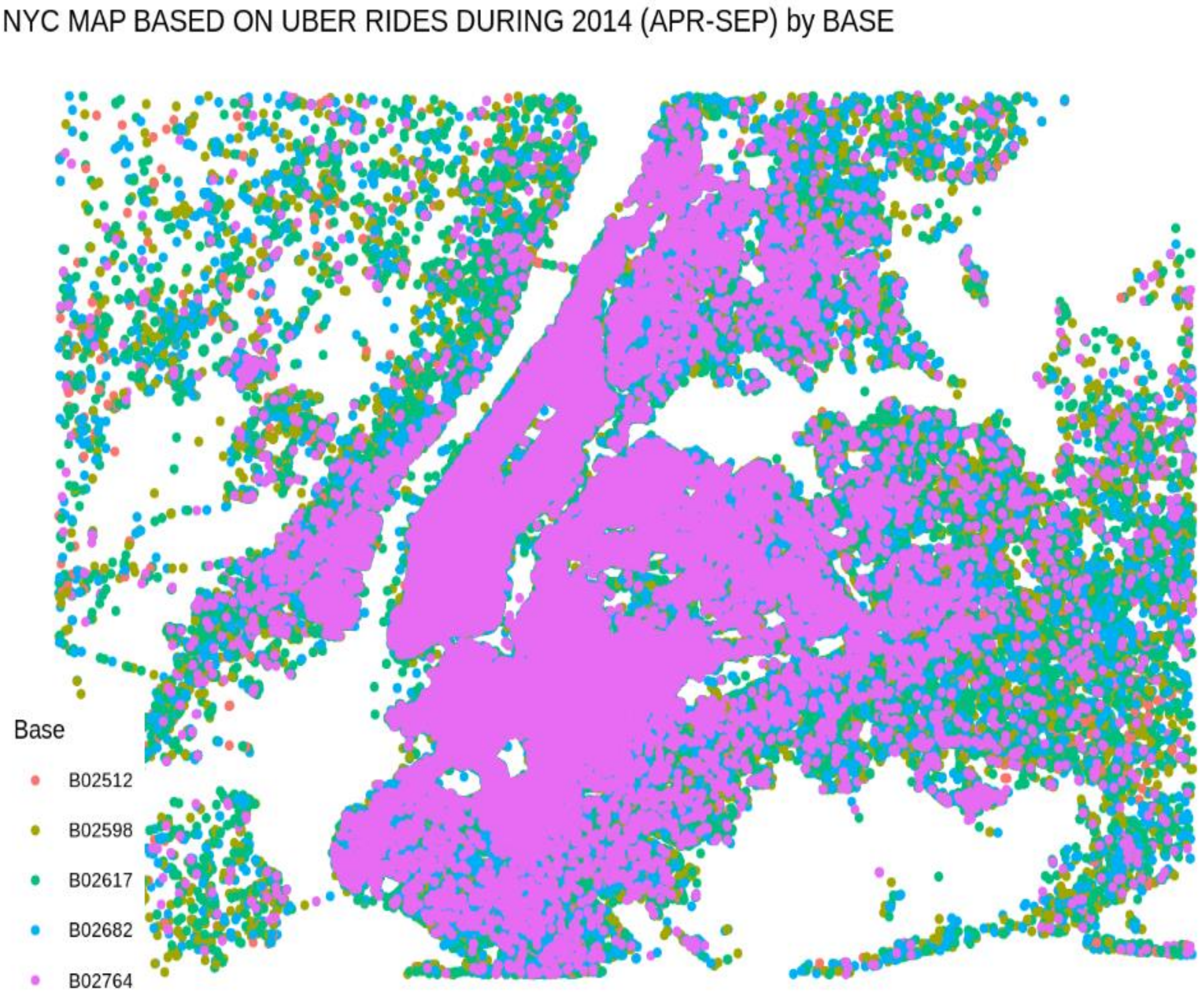
# Creating a map visualization of rides in New York

In the final section, I 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.



```
#Geo Visualization
min_lat <- 40.5774
max_lat <- 40.9176
min_long <- -74.15
max_long <- -73.7004
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")
```




NYC MAP BASED ON UBER RIDES DURING 2014 (APR-SEP) by BASE
Base
B02512
B02598
B02617
B02682
B02764

# Summary

At the end of the Uber data analysis R project, I observed how to create data visualizations. I 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 New York that provided us with the details of how various users made trips from different bases.