library(ggplot2)

library(lubridate)

library(dplyr)

library(tidyr)

library(tidyverse) # metapackage of all tidyverse packages

library(scales)

# Read the data for each month separately

apr <- read.csv("/content/uber-raw-data-apr14.csv")

may <- read.csv("/content/uber-raw-data-may14.csv")

june <- read.csv("/content/uber-raw-data-jun14.csv")

july <- read.csv("/content/uber-raw-data-jul14.csv")

aug <- read.csv("/content/uber-raw-data-aug14.csv")

sept <- read.csv("/content/uber-raw-data-sep14.csv")

# Combine the data together

data <- rbind(apr, may, june, july, aug, sept)

cat("The dimensions of the data are:", dim(data))

o/p: The dimensions of the data are: 1115394 4

# Print the first 6 rows of the data

head(data)

| A data.frame: 6 × 4 | | | | |
| --- | --- | --- | --- | --- |
|  | **Date.Time** | **Lat** | **Lon** | **Base** |
|  | **<chr>** | **<chr>** | **<chr>** | **<chr>** |
| **1** | 4/1/2014 0:11:00 | 40.769 | -73.9549 | B02512 |
| **2** | 4/1/2014 0:17:00 | 40.7267 | -74.0345 | B02512 |
| **3** | 4/1/2014 0:21:00 | 40.7316 | -73.9873 | B02512 |
| **4** | 4/1/2014 0:28:00 | 40.7588 | -73.9776 | B02512 |
| **5** | 4/1/2014 0:33:00 | 40.7594 | -73.9722 | B02512 |
| **6** | 4/1/2014 0:33:00 | 40.7383 | -74.0403 | B02512 |

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

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

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

# Create individual columns for month day and year

data$day <- factor(day(data$Date.Time))

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

data$year <- factor(year(data$Date.Time))

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

# Add Time variables as well

data$second = factor(second(hms(data$Time)))

data$minute = factor(minute(hms(data$Time)))

data$hour = factor(hour(hms(data$Time)))

# Look at the data

head(data)

#Data Visualisation

#Plotting the trips by hours in a day

hourly\_data <- data %>%

group\_by(hour) %>%

dplyr::summarize(Total = n())

# Plot the data by hour

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

geom\_bar(stat="identity",

fill="steelblue",

color="red") +

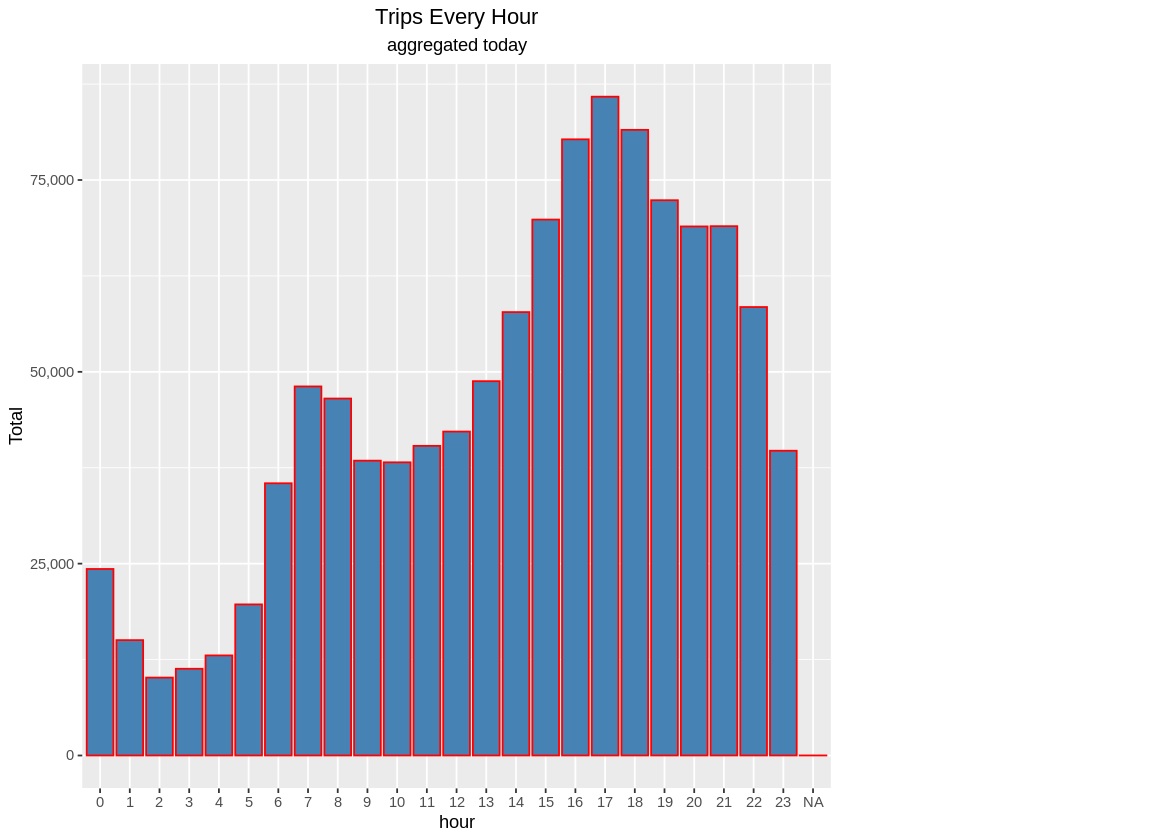
ggtitle("Trips Every Hour", subtitle = "aggregated today") +

theme(legend.position = "none",

plot.title = element\_text(hjust = 0.5),

plot.subtitle = element\_text(hjust = 0.5)) +

scale\_y\_continuous(labels=comma)



#Plotting trips by hour and month

# Aggregate the data by month and hour

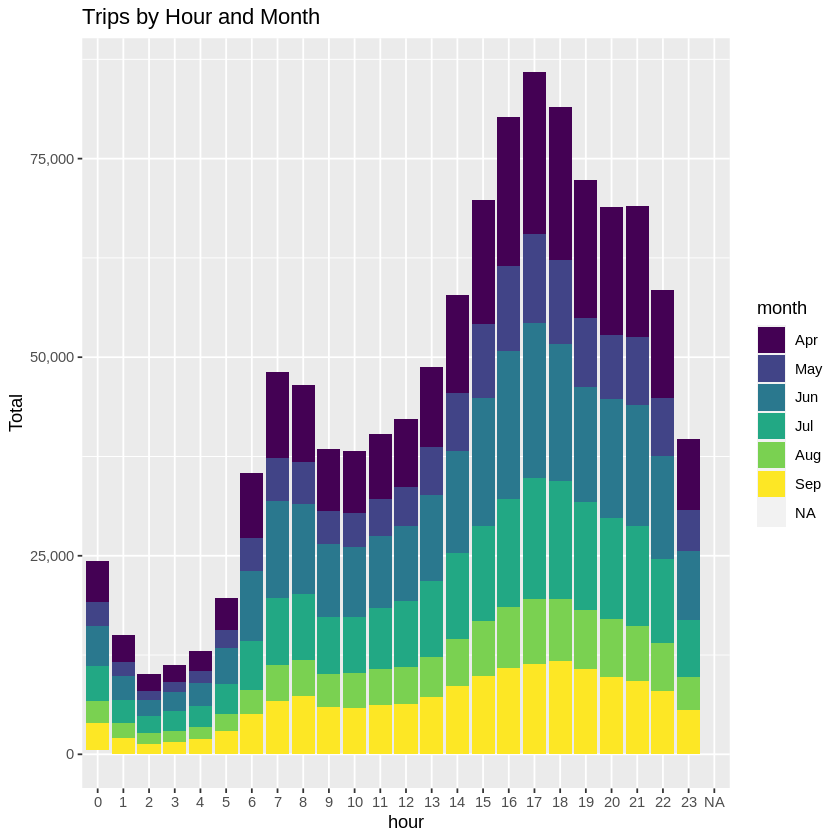
month\_hour\_data <- data %>% group\_by(month, hour) %>% dplyr::summarize(Total = n())

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

# Aggregate data by day of the month

day\_data <- data %>% group\_by(day) %>% dplyr::summarize(Trips = n())

day\_data

# Plot the data for the day

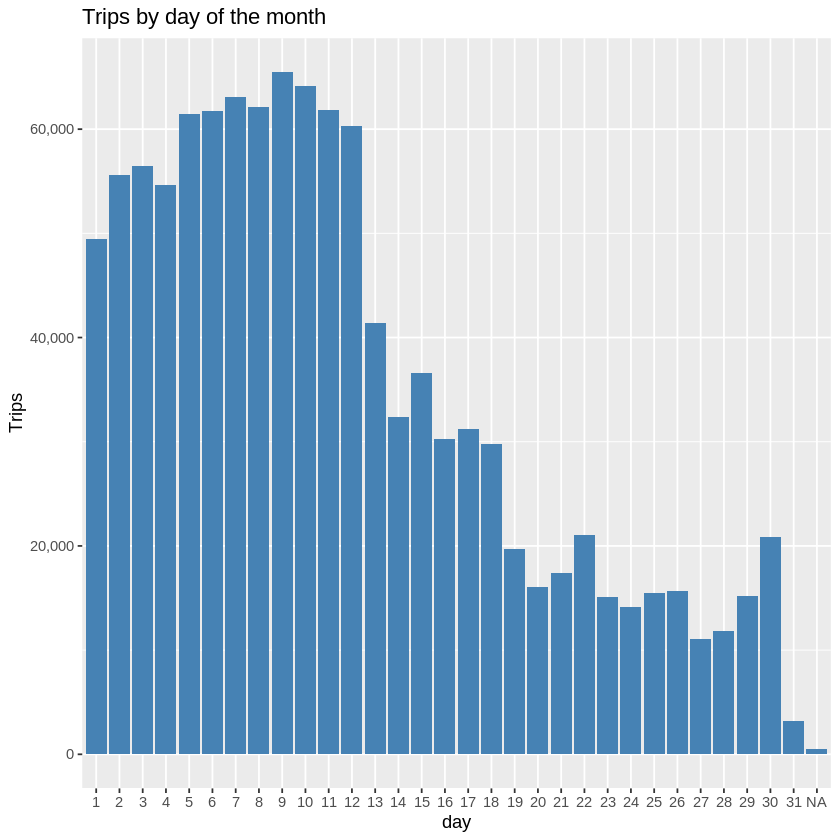
ggplot(day\_data, aes(day, Trips)) +

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

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

theme(legend.position = "none")

scale\_y\_continuous(labels = comma)



# Collect data by day of the week and month

day\_month\_data <- data %>% group\_by(dayofweek, month) %>% dplyr::summarize(Trips = n())

day\_month\_data

# Plot the above data

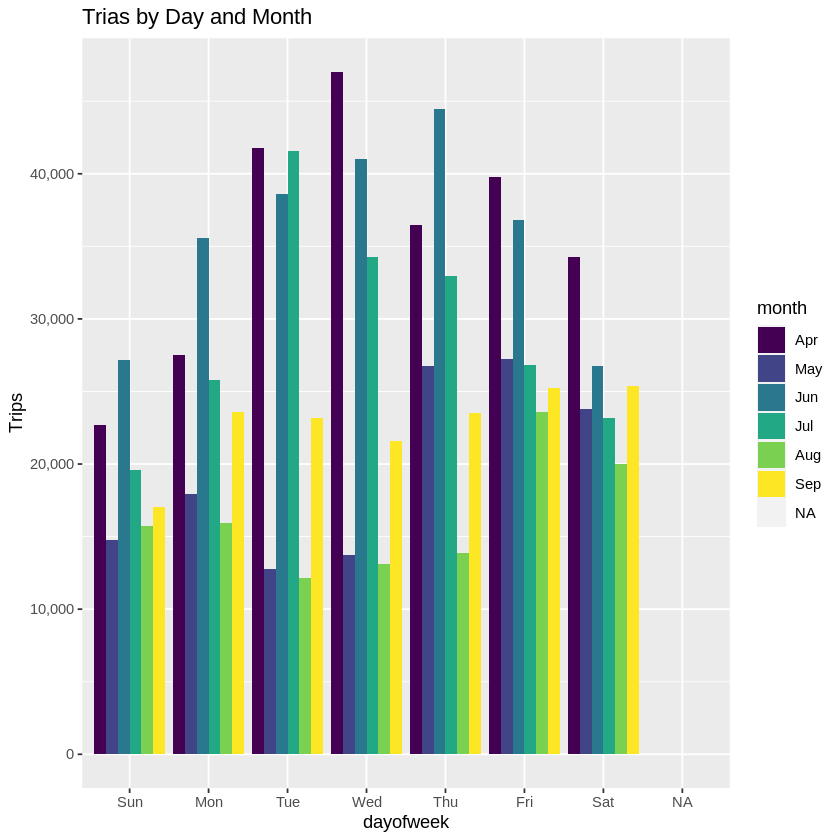
ggplot(day\_month\_data, aes(dayofweek, Trips, fill = month)) +

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)



#Number of Trips place during months in a year

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

month\_data

ggplot(month\_data, aes(month, Total, fill = month)) +

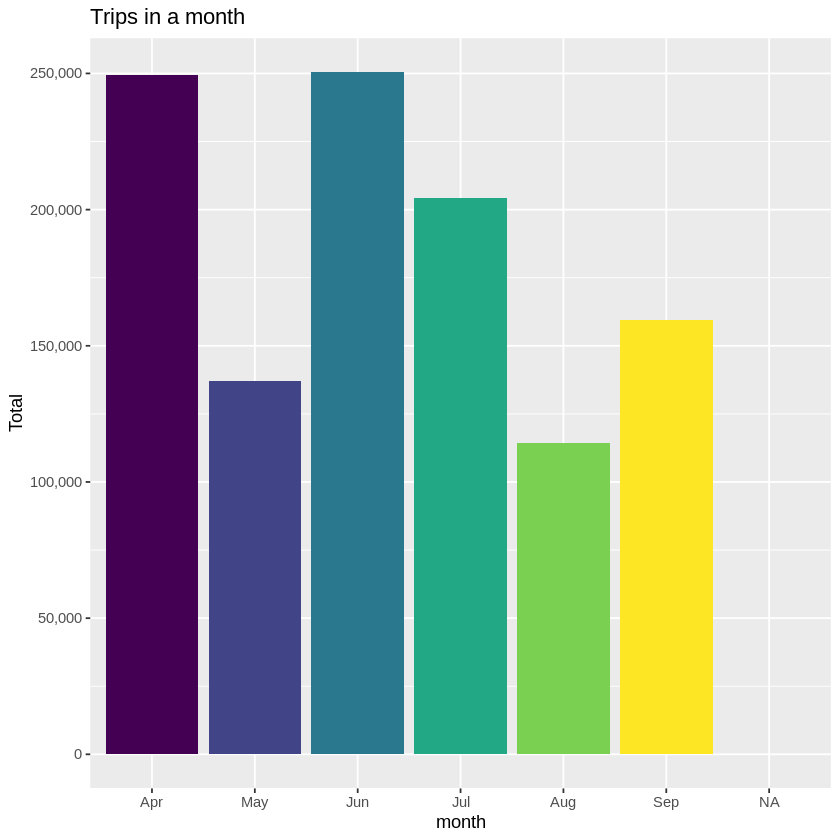
geom\_bar(stat = "Identity") +

ggtitle("Trips in a month") +

theme(legend.position = "none") +

scale\_y\_continuous(labels = comma) +

scale\_fill\_manual(values = colors)



# Collect data by month and day

month\_day\_data <- data %>% group\_by(month, day) %>% dplyr::summarize(Trips = n())

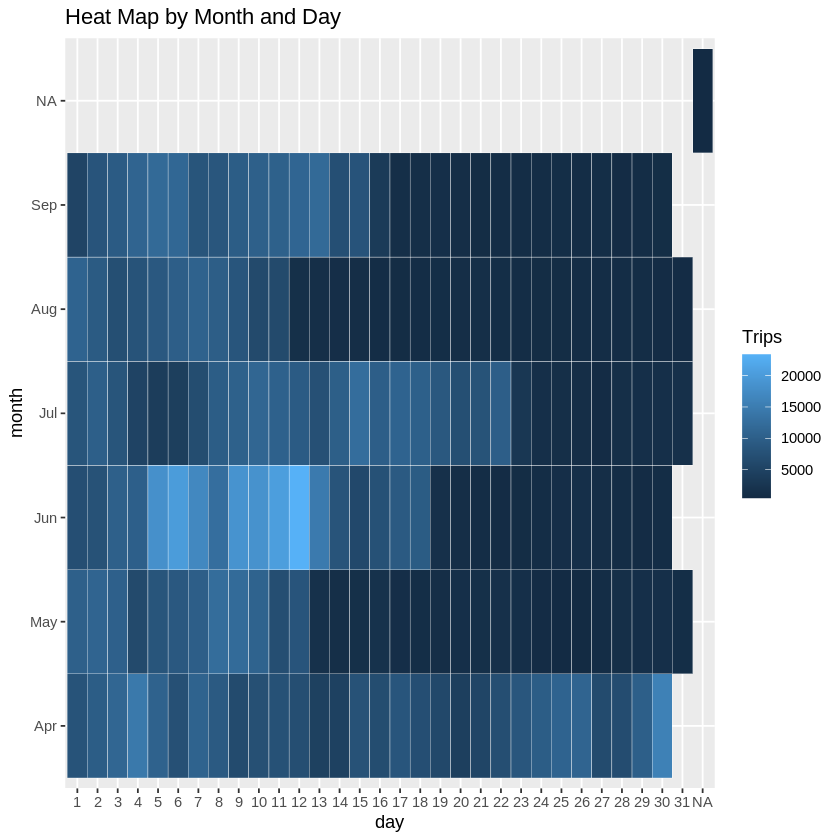
month\_day\_data

# Plot a heatmap

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

geom\_tile(color = "white") +

ggtitle("Heat Map by Month and Day")



# Plot a heatmap by day of the week and month

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

geom\_tile(color = "white") +

ggtitle("Heat Map by Month and Day")

