
UBER
Uber

UBER

Contents

```
#install.packages("knitr") library(knitr)
 #install.packages("tinytex")
 #tinytex::install_tinytex()



#tinytex::install_tinytex(repository = 'http://dante.ctan.org/tex-archive/')
```

Import all Libraries

```
invisible(library(ggplot2))    #Visualization
invisible(library(ggThemeAssist))      #Add ins for Visualization
invisible(library(dplyr))              #data manipulation
invisible(library(lubridate)) #date time
invisible(library(scales))     #graphical scaling
invisible(library(tidyr))              #cleaning data
#install.packages("DT")         #installing DT
invisible(library(DT))                 #table formatted
invisible(library(readr)               #reading
)                                      CSV
```

Loading datasets for six months (April to September)

```
apr_data <- read_csv("C:/Users/Suma/Desktop/R/Project - UBER Dataset/uber-raw-data-apr14.
may_data <- read_csv("C:/Users/Suma/Desktop/R/Project - UBER Dataset/uber-raw-data-may14.
jun_data <- read_csv("C:/Users/Suma/Desktop/R/Project - UBER Dataset/uber-raw-data-jun14.
jul_data <- read_csv("C:/Users/Suma/Desktop/R/Project - UBER Dataset/uber-raw-data-jul14.
aug_data <- read_csv("C:/Users/Suma/Desktop/R/Project - UBER Dataset/uber-raw-data-aug14.
sep_data <- read_csv("C:/Users/Suma/Desktop/R/Project - UBER Dataset/uber-raw-data-sep14.
```

Combining all the datasets by stacking them one over the other

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

Let's take a look at the dataset from a high level

```
head(data_2014)
```

```
## # A tibble: 6 x 4
##   `Date/Time`        Lat   Lon Base
##   <chr>            <dbl> <dbl> <chr>
## 1 4/1/2014 0:11:00  40.8 -74.0 B02512
## 2 4/1/2014 0:17:00  40.7 -74.0 B02512
## 3 4/1/2014 0:21:00  40.7 -74.0 B02512
## 4 4/1/2014 0:28:00  40.8 -74.0 B02512
## 5 4/1/2014 0:33:00  40.8 -74.0 B02512
## 6 4/1/2014 0:33:00  40.7 -74.0 B02512
```

Also let's explore the structure of the data to understand it better

```
str(data_2014)
```

```
## spec_tbl_df [4,534,327 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Date/Time: chr [1:4534327] "4/1/2014 0:11:00" "4/1/2014 0:17:00" "4/1/2014 0:21:00" "4/1/2014 0:28
##  $ Lat      : num [1:4534327] 40.8 40.7 40.7 40.8 40.8 ...
##  $ Lon      : num [1:4534327] -74 -74 -74 -74 -74 ...
##  $ Base     : chr [1:4534327] "B02512" "B02512" "B02512" "B02512" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Date/Time` = col_character(),
##   ..   Lat = col_double(),
##   ..   Lon = col_double(),
##   ..   Base = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Clean the data

```
# Changing the 'date/time' data type to dates

data_2014$`Date/Time` <- as.POSIXct(data_2014$`Date/Time`,format = "%m/%d/%Y %H:%M:%S")

# Extracting the time component

data_2014$Time <- format(as.POSIXct(data_2014$`Date/Time`,format = "%m/%d/%Y %H:%M:%S"),
                         format = "%H:%M:%S")

# Format the date to ymd hms

data_2014$`Date/Time` <- ymd_hms(data_2014$`Date/Time`)
```

Manipulate the data

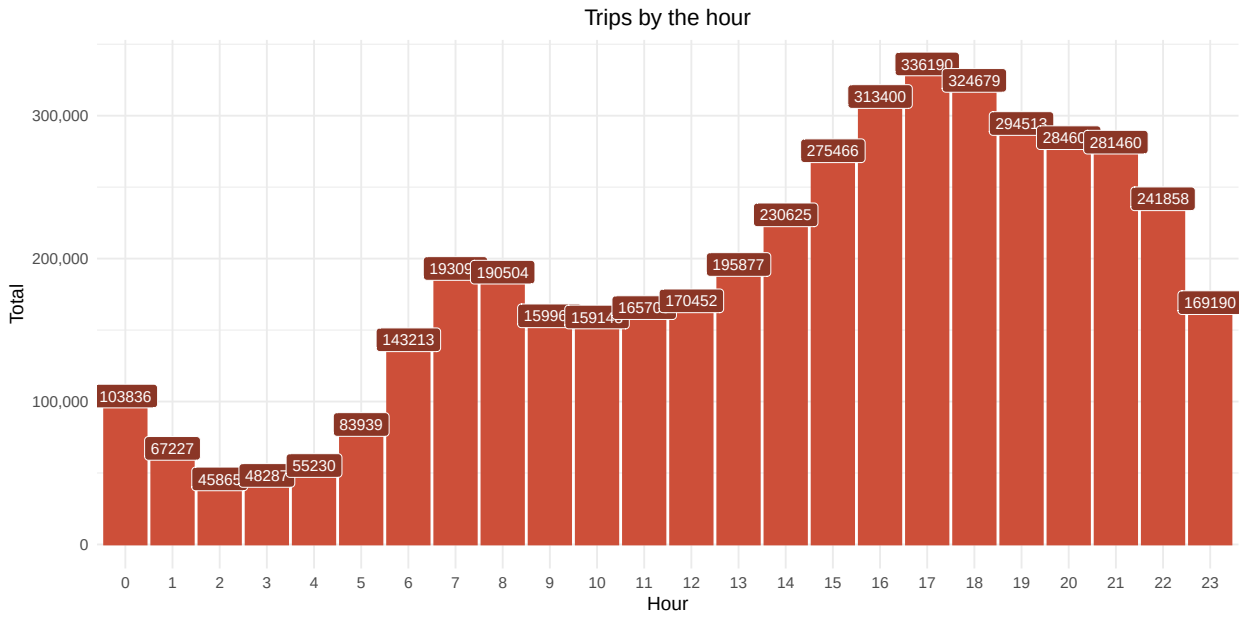
```
# Extracting the day number, month name and day name from the date time

data_2014$Day <- format(day(data_2014$`Date/Time`))
data_2014$Month <- format(month(data_2014$`Date/Time`, label = TRUE))
data_2014$dayoftheweek <- format(wday(data_2014$`Date/Time`, label = TRUE))

# Extracting the hour, minute and second from the date time

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`)))
```

Visualizations

What time of the day do most of the bookings happen?

```
## Step 1 : Creating a new table with no of trips per hour
hour_data <- data_2014 %>%
  group_by(Hour) %>%
  summarise(Total=n())


## Step 2 : Visualising in a tabular format

datatable(hour_data)
```

```
ggplot(hour_data, aes(Hour,Total)) +
  geom_bar(stat = "identity", fill = "tomato3", color = "tomato3") +
  ggtitle("Trips by the hour") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5)) +
  geom_label(mapping = aes(label = Total),
             fill = "tomato4",
             color = "white",
             size = 3) +
  scale_y_continuous(labels = comma)
```

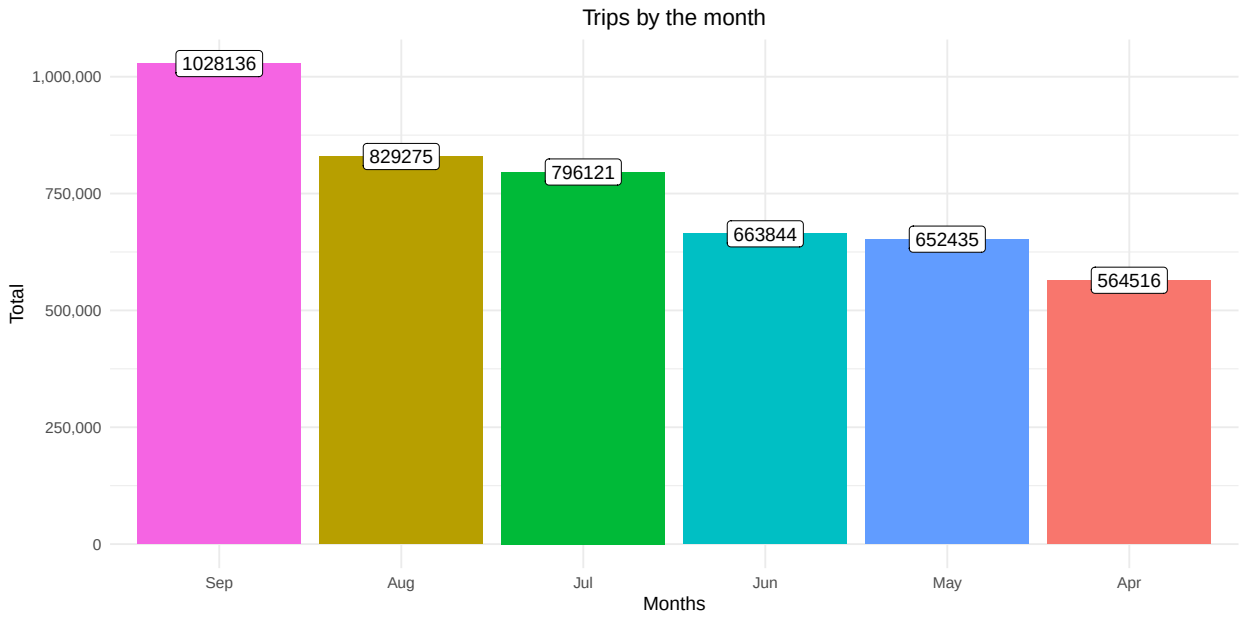


As seen in the viz above,most of the trips via Uber appear to happen between 3pm and 9pm, with a small peak from 6am - 8am.This might be directly related to working professionals, who may book a cab for their commutes to and from their office. There is a huge spike in rides around 5pm, which is the end of workday for most offices on an average.

How many rides were booked from April to September?

```
month_data <- data_2014 %>%
  group_by(Month) %>%
  summarise(Total=n())


datatable(month_data)
```

```
ggplot(data = month_data, mapping = aes(x = reorder(Month,-Total),y = Total)) +
  geom_bar(stat = "identity", mapping = aes(fill= Month)) +
  ggtitle("Trips by the month") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5)) +
  xlab("Months") +
  scale_y_continuous(labels = comma) +
geom_label(mapping = aes(label = Total),
           size = 4,
           color = "black",
           fill = "white")
```

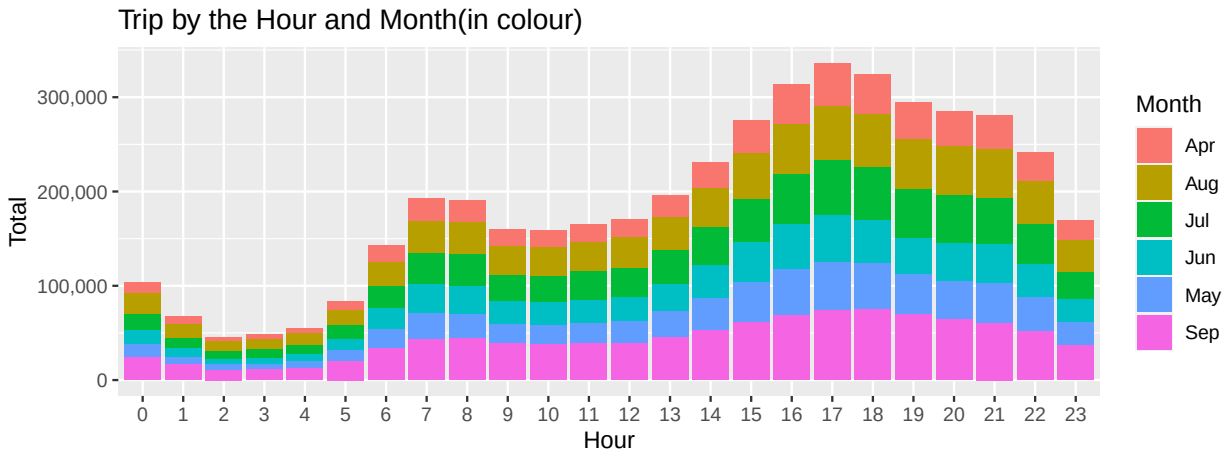


UBER rides have almost doubled from April (564,516) to September (1,028,136) as per the dataset at hand.

How many rides were booked at each time of the day, split by all 6 month?

```
month_hour_data <- data_2014 %>%
  group_by(Month,Hour) %>%
  summarise(Total=n())


datatable(month_hour_data)
```

```
ggplot(month_hour_data, aes(Hour,Total,fill=Month)) +
  geom_bar(stat = "identity") +
  ggtitle("Trip by the Hour and Month(in colour)") +
  scale_y_continuous(label = comma)
```

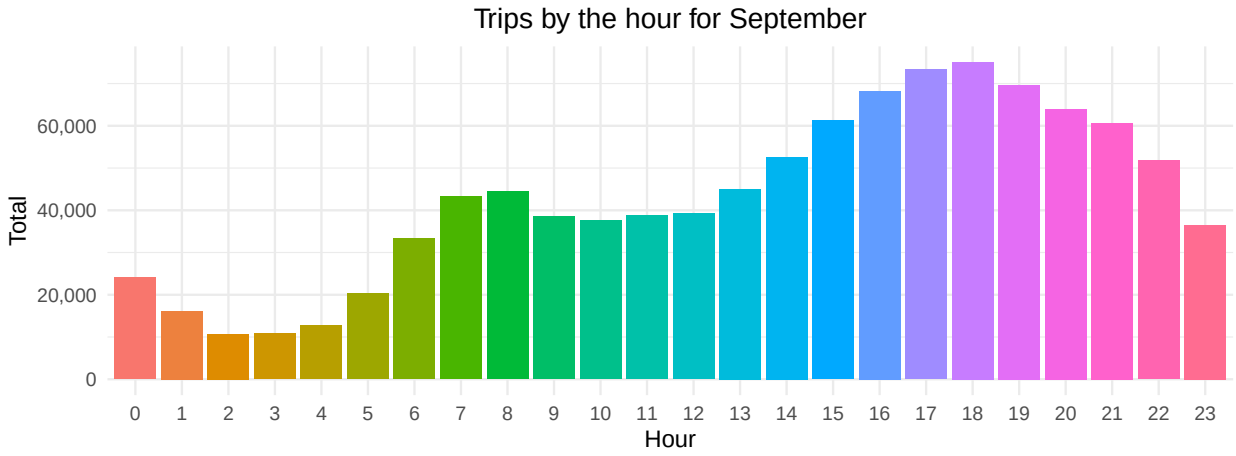


This graph confirms our previous conclusion that most rides happened between 3pm and 5pm with a slight peak between 5 am - 8am, along with September being the highest booked month.

How many rides were booked in the highest booked month - September, split by time of the day?

```
sept_hour <- data_2014 %>%
  group_by(Hour,Month) %>%
  filter(Month == "Sep") %>%
  summarise(Total=n())


datatable(sept_hour)
```

```
ggplot(sept_hour, aes(Hour,Total,fill=Hour)) +
  geom_bar(stat = "identity") +
  ggtitle("Trips by the hour for September") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(labels = comma)
```



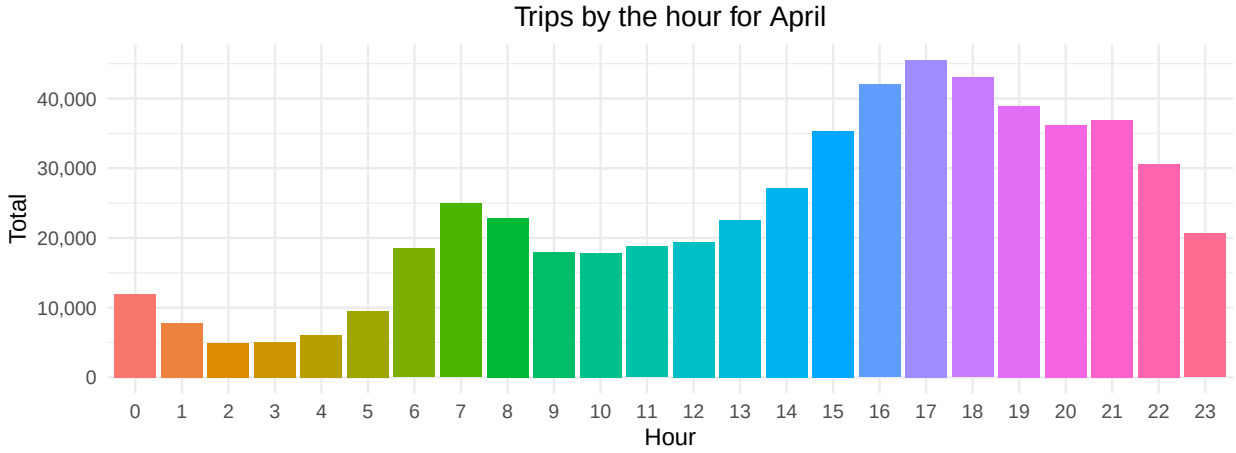
<br

We can see that the number of rides hit their peak of 70,000 bookings at 6pm alone. Since its already assumed that 6pm is a high demand time slot for office goers, we can find out if it was a result of any campaign that the rides were so high in september, and re run those again.

How many rides were booked in the lowest booked month - April, split by time of the day?

```
  april_hour <- data_2014 %>%
  group_by(Month,Hour) %>%
  filter(Month == "Apr") %>%
  summarise(Total = n())

datatable(april_hour)
```

```
ggplot(april_hour, aes(Hour,Total,fill=Hour)) +
  geom_bar(stat = "identity") +
  ggtitle("Trips by the hour for April") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(labels = comma)
```

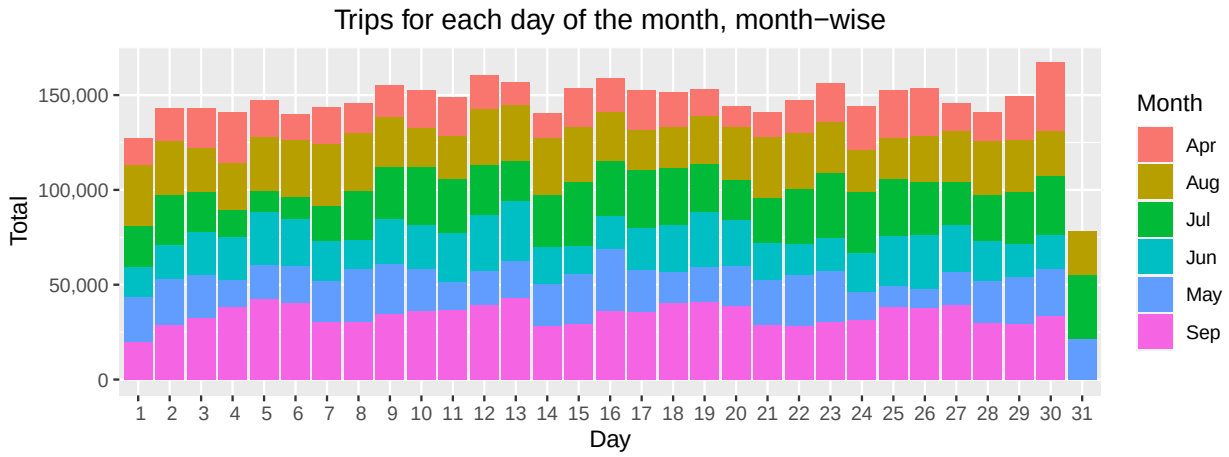


April saw a high of 50,000 rides at peak hours, and a low of close to 5000 rides from 2 am - 4am.

Are there any trends day wise for each month?

```
day_data <- data_2014 %>%
  group_by(Month,Day) %>%
  summarise(Total=n())

datatable(day_data)
```

```
ggplot(day_data, aes(Day,Total,fill=Month)) +
  geom_bar(stat="identity") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Trips for each day of the month, month-wise") +
  scale_y_continuous(labels = comma)
```



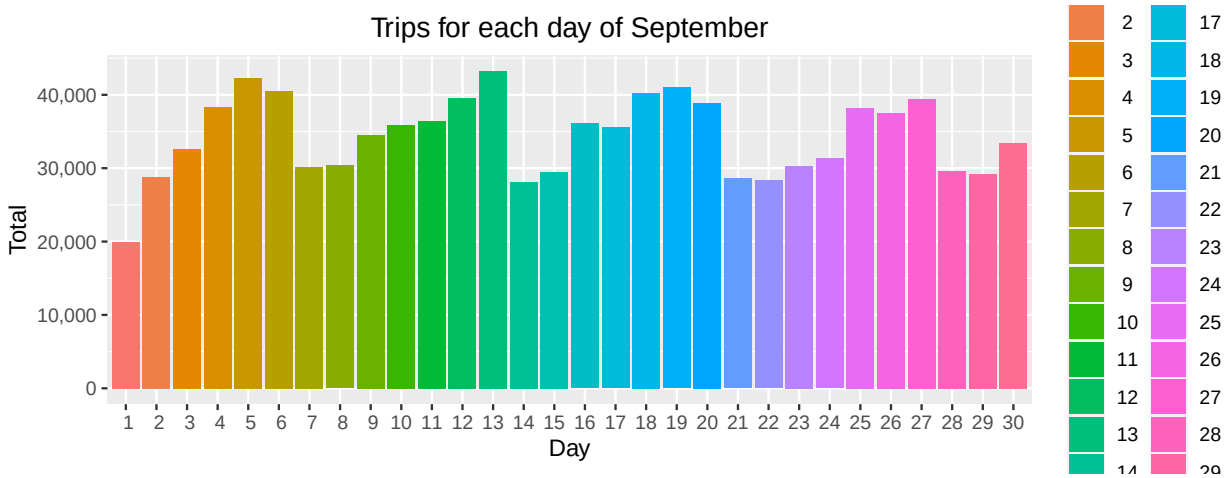
More or less uniform distribution, when analysed for all months, can dig deeper into each month to see any trends.

Note : 31st appears in only half the months, therefore its count is skewed.

Is there any trend for all days of September?

```
sept_day <- data_2014 %>%
  group_by(Month,Day) %>%
  filter(Month == "Sep") %>%
  summarise(Total=n())


datatable(sept_day)
```

```
ggplot(sept_day, aes(Day,Total,fill=Day)) +
  geom_bar(stat="identity") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Trips for each day of September") +
  scale_y_continuous(labels = comma)
```

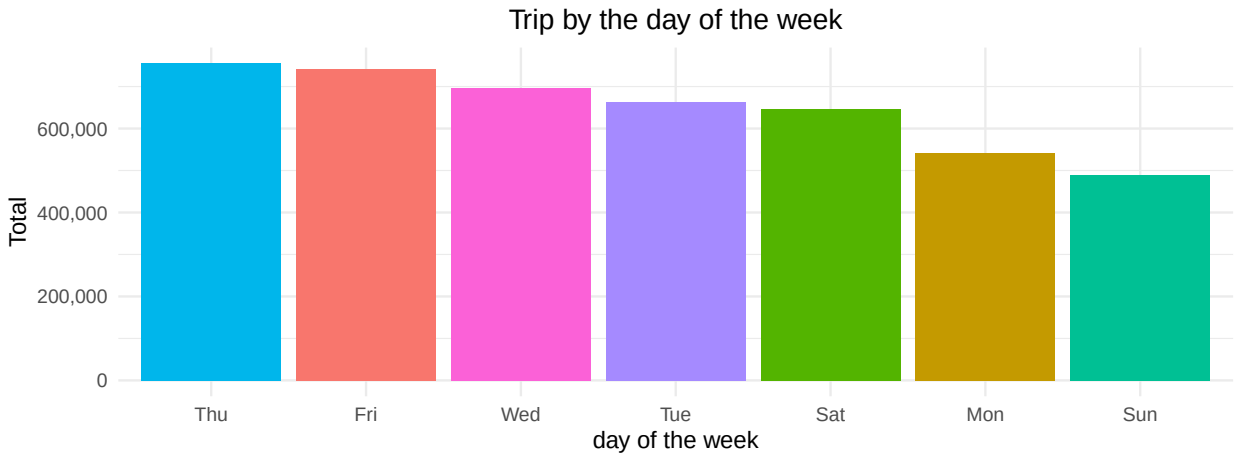


The above chart is a deeper look into the rides booked by the day in the month of September. There appears to be a cycle after five days, which is most definitely the weekdays, and a fall over the weekends. But lets dig deeper into that too.

Is there a pattern in UBER rides with respect to the day of the week?

```
# Plotting the number of trips grouped by the day of the week

day_of_the_week_data <- data_2014 %>%
  group_by(dayoftheweek) %>%
  summarise(Total = n())

datatable(day_of_the_week_data)
```

```
ggplot(day_of_the_week_data,aes(reorder(dayoftheweek,-Total),Total,fill=dayoftheweek)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  ggtitle("Trip by the day of the week") +
  xlab("day of the week") +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(label = comma)
```

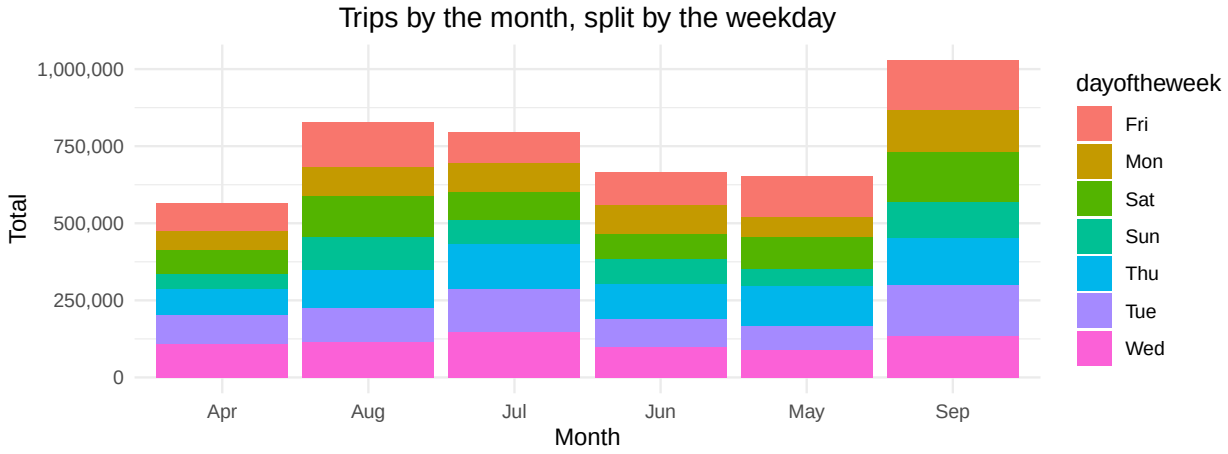


Thursday and Friday experience the most UBER rides, while the lowest number of bookings happen on Sunday, given its a holiday at offices in general.

How many trips occur in each month, split by the day of the week?

```
month_weekday_data <- data_2014 %>%
  group_by(Month,dayoftheweek) %>%
  summarise(Total=n())

datatable(month_weekday_data)
```

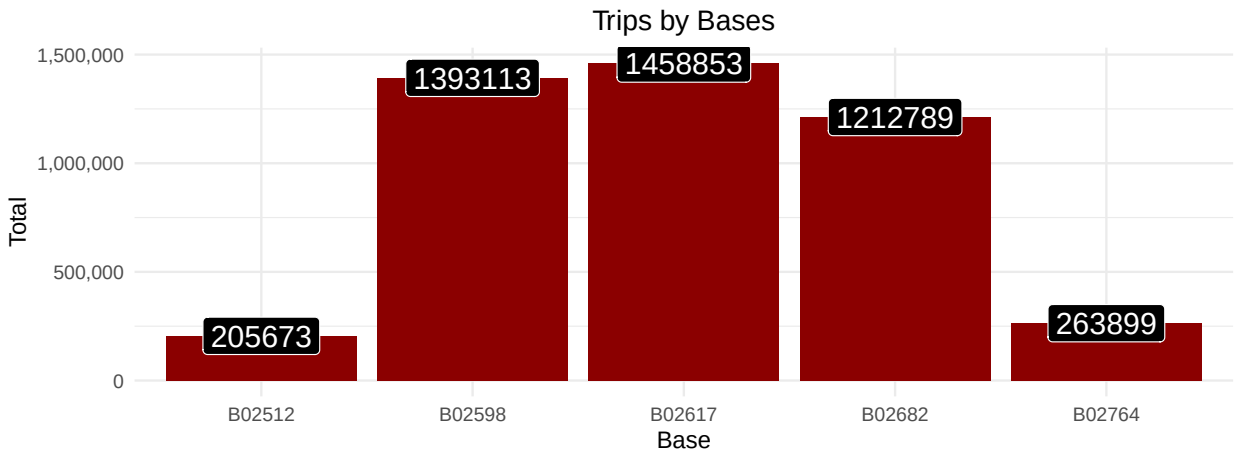
```
ggplot(month_weekday_data,aes(Month,Total,fill=dayoftheweek)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  ggtitle("Trips by the month, split by the weekday") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(labels = comma)
```



As we had seen before, September has the highest number of trips with Friday, Thursday and Tuesday being the highest booked days.

How many rides are associated with each of the five UBER bases?

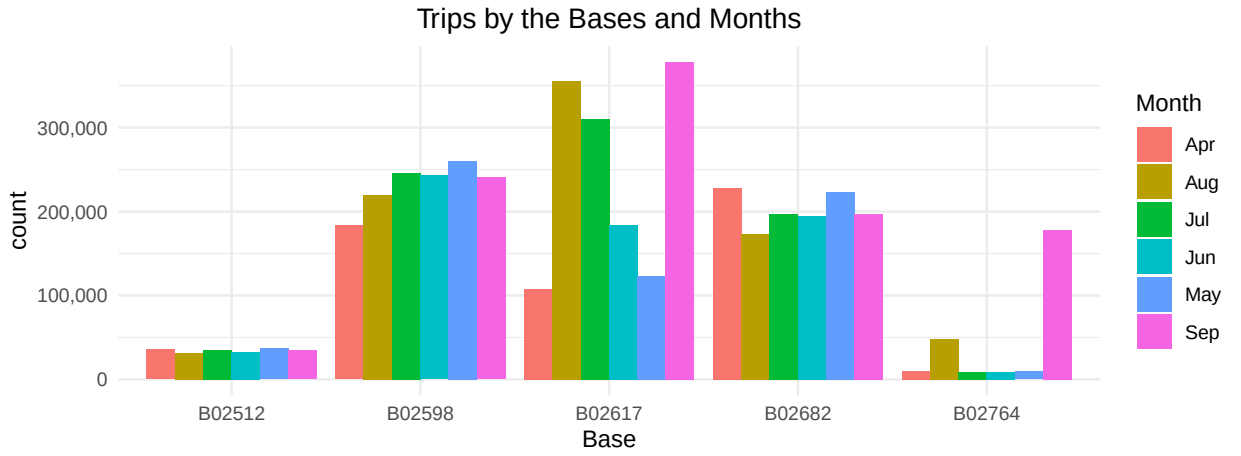
```
A <- data_2014 %>%
  group_by(Base) %>%
  summarise(Total = n())



ggplot(A,mapping = aes(x = Base, y = Total)) +
  geom_bar(stat = "identity", fill = "darkred") +
  theme_minimal() +
  scale_y_continuous(label=comma) +
  ggtitle("Trips by Bases")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_label(mapping = aes(label = Total),
             size = 5,
             color = "white",
             fill= "black")
```



Bases B02512 and B02764 are not generating many rides and we should dig deeper to find out the cost to benefit ratio for these bases and conclude if they are even profitable.

Is there any change over the months for rides associated with each base?

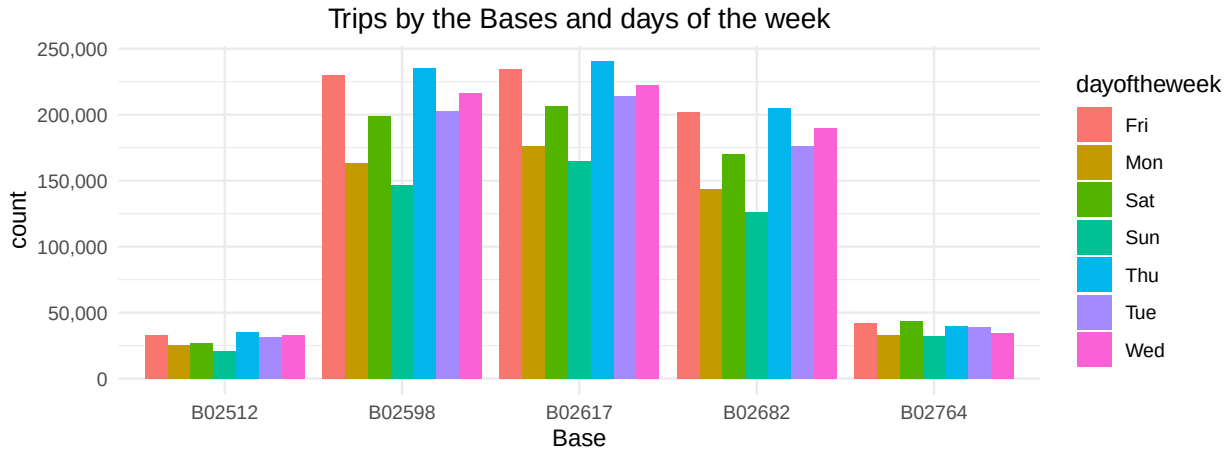
```
ggplot(data_2014,aes(Base,fill=Month)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  scale_y_continuous(labels = comma) +
  ggtitle("Trips by the Bases and Months") +
  theme(plot.title = element_text(hjust = 0.5))
```



September saw an unusually high no of trips for Base B02764 as well as B02617. Also, April and May were lower for Base B02617 compared to the other months.

Is there any change over the days of the week, for rides associated with each base?

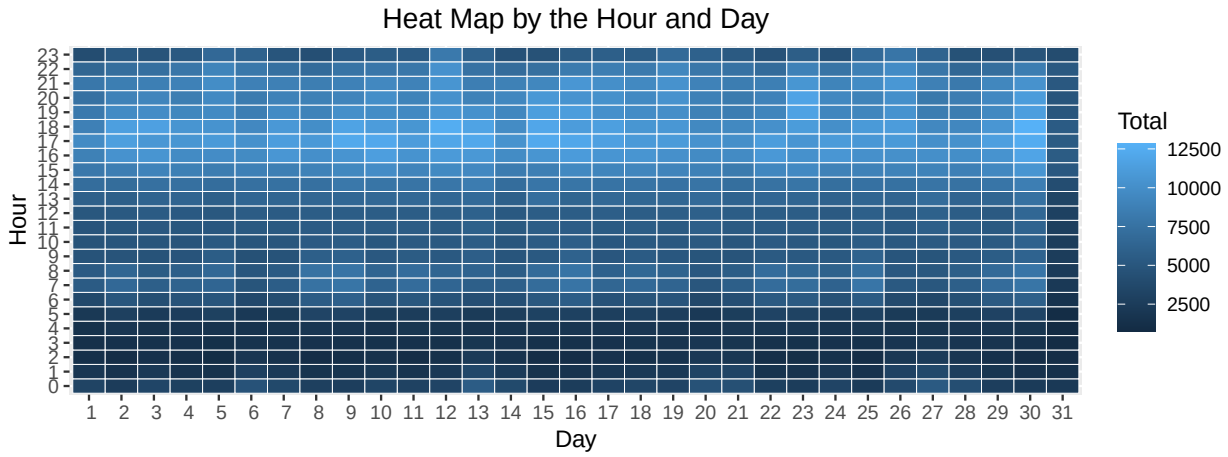
```
ggplot(data_2014,aes(Base,fill=dayoftheweek)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(labels = comma) +
  ggtitle("Trips by the Bases and days of the week")
```



Each base experiences the same spike during Fridays and Thursdays, along with a dip during Sundays.

Heat Map - Uber Rides by day and hour

```
invisible(day_and_hour <- data_2014 %>%
  group_by(Day,Hour) %>%
  dplyr::summarise(Total=n()))



ggplot(day_and_hour, aes(Day,Hour,fill=Total)) +
  geom_tile(color="white") +

  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Heat Map by the Hour and Day")
```



As seen earlier, the rides from 3pm to 9pm are higher, and lower from midnight to 5am.