Uber Data

Eva Shah

1/14/2021

```
webshot::install_phantomjs()

## It seems that the version of `phantomjs` installed is greater than
or equal to the requested version.To install the requested version or
downgrade to another version, use `force = TRUE`.

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.2

library(ggthemes)
library(lubridate)

## Warning: package 'lubridate' was built under R version 3.6.2

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.6.2

library(readr)
aprdata = read.csv("~/Desktop/data sets/Uber-dataset/apr.csv")
maydata = read.csv("~/Desktop/data sets/Uber-dataset/may.csv")
jundata = read.csv("~/Desktop/data sets/Uber-dataset/jun.csv")
```

```r
juldata = read.csv("~/Desktop/data sets/Uber-dataset/jul.csv")
augdata = read.csv("~/Desktop/data sets/Uber-dataset/aug.csv")
sepdata = read.csv("~/Desktop/data sets/Uber-dataset/sep.csv")
#Combining all the data
data = rbind(aprdata, maydata, jundata, juldata, augdata, sepdata)
str(data)
```

```
## 'data.frame':    4534327 obs. of  4 variables:
##  $ Date.Time: Factor w/ 260093 levels "4/1/2014 0:00:00",..: 11 17
21 28 33 33 38 44 54 58 ...
##  $ Lat      : num  40.8 40.7 40.7 40.8 40.8 ...
##  $ Lon      : num  -74 -74 -74 -74 -74 ...
##  $ Base     : Factor w/ 5 levels "B02512","B02598",..: 1 1 1 1 1 1 1
1 1 1 ...
```

```r
#now formatting date and time
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)

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))

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

data$hour <- factor(hour(hms(data$Time)))
data$minute <- factor(minute(hms(data$Time)))
data$second <- factor(minute(hms(data$Time)))

library(scales)
```

```
## Warning: package 'scales' was built under R version 3.6.2
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
library(dplyr)
library(DT)
```

```
## Warning: package 'DT' was built under R version 3.6.2
```

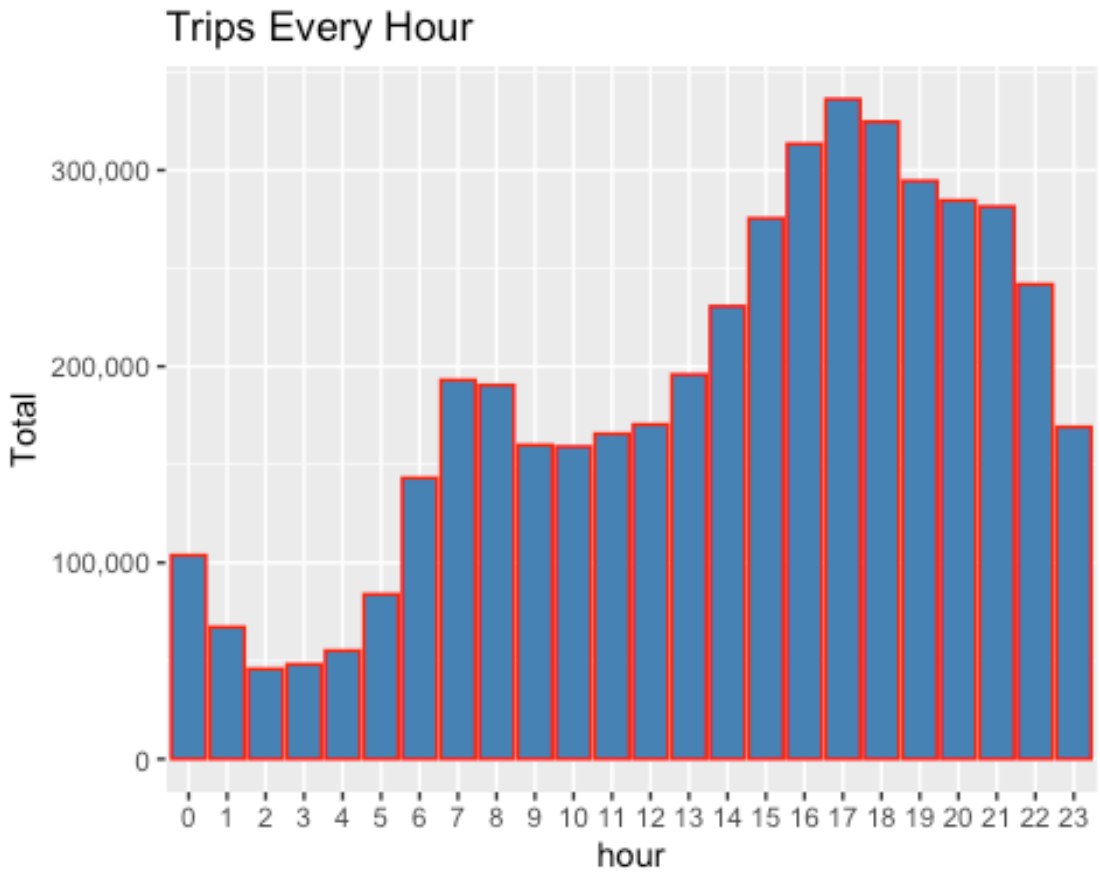
```r
#Data exploration

hour_data <- data %>%
           group_by(hour) %>%
               dplyr::summarize(Total = n())

## `summarise()` ungrouping output (override with `.groups` argument)

datatable(hour_data)
```

Show 10 entries Search:

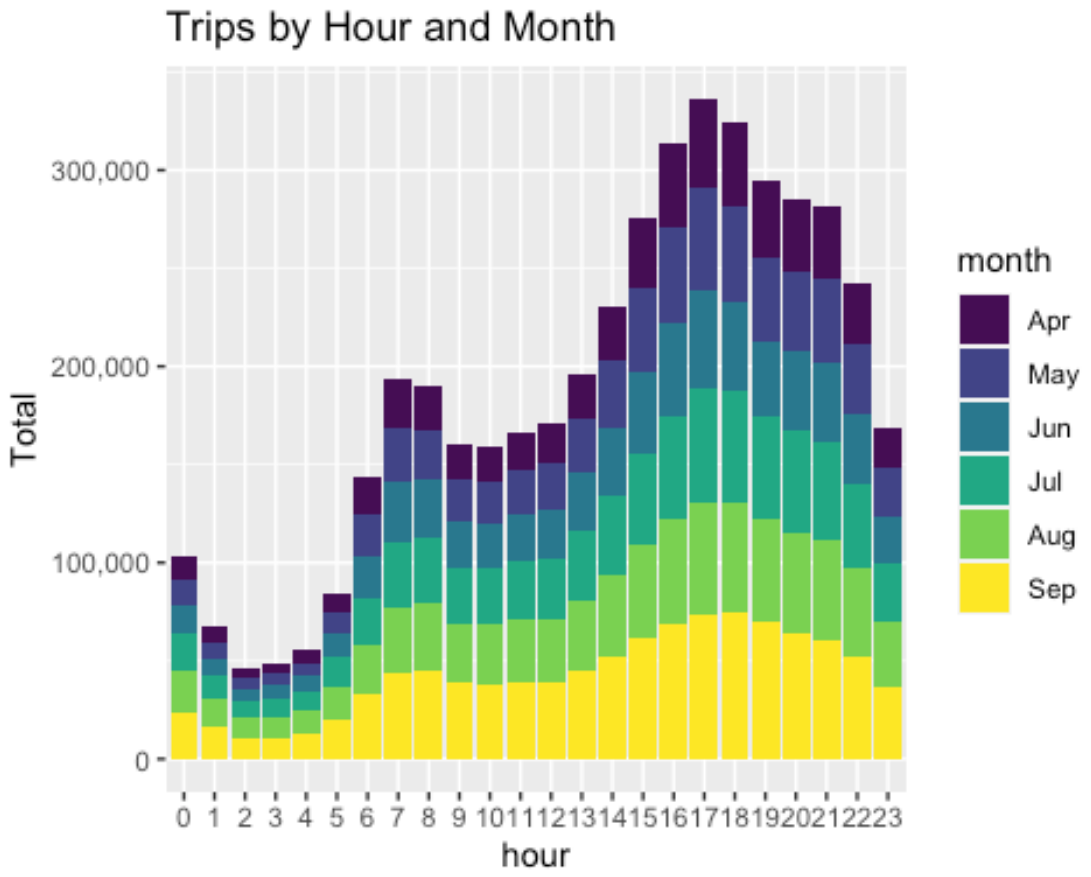
	hour	Total
1	0	103836
2	1	67227
3	2	45865
4	3	48287
5	4	55230
6	5	83939
7	6	143213
8	7	193094
9	8	190504
10	9	159967

Showing 1 to 10 of 24 entries Previous 1 2 3 Next

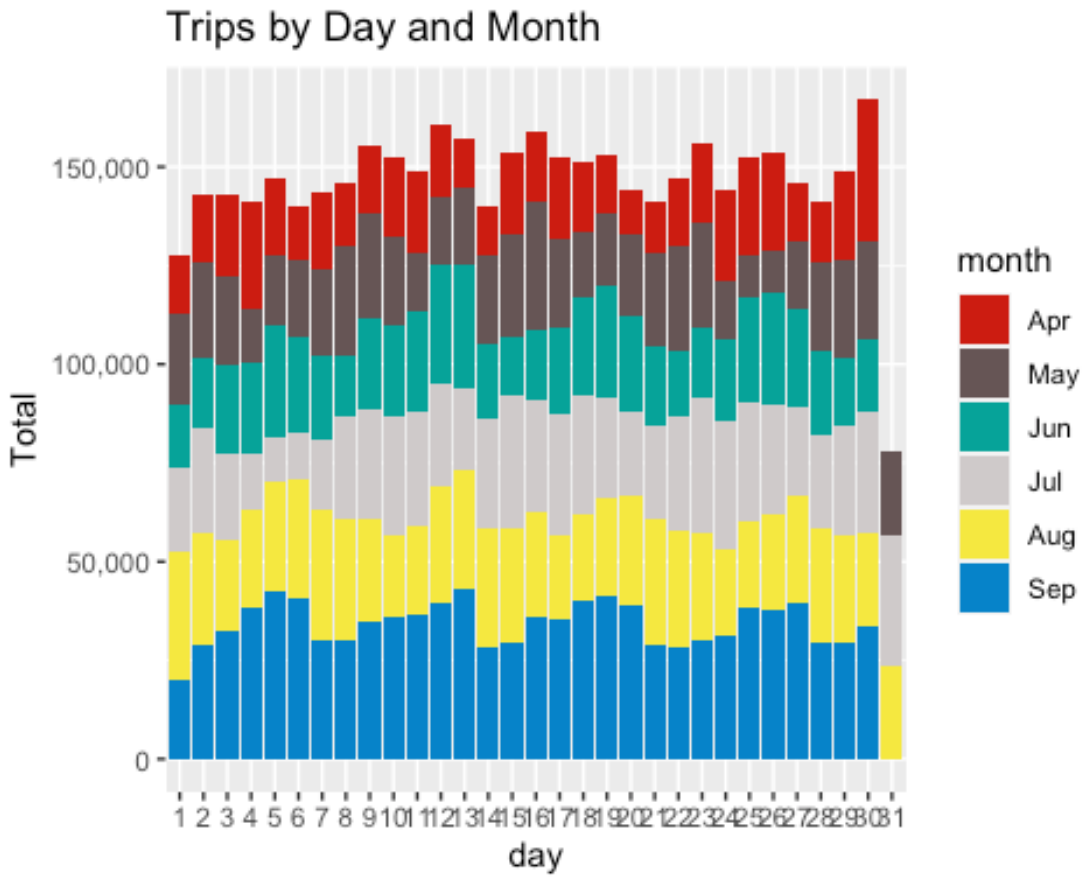
```r
##BAR GRAPHS

##Trips by Hour and Month
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 %>%
          group_by(month, hour) %>%
             dplyr::summarize(Total = n())

## `summarise()` regrouping output by 'month' (override with `.groups`
argument)

ggplot(month_hour, aes(hour, Total, fill = month)) +
       geom_bar( stat = "identity") +
          ggtitle("Trips by Hour and Month") +
           scale_y_continuous(labels = comma)
```




```
## Trips by Day and Month

day_month_group <- data %>%
         group_by(month, day) %>%
             dplyr::summarize(Total = n())

## `summarise()` regrouping output by 'month' (override with `.groups`
argument)

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)
```



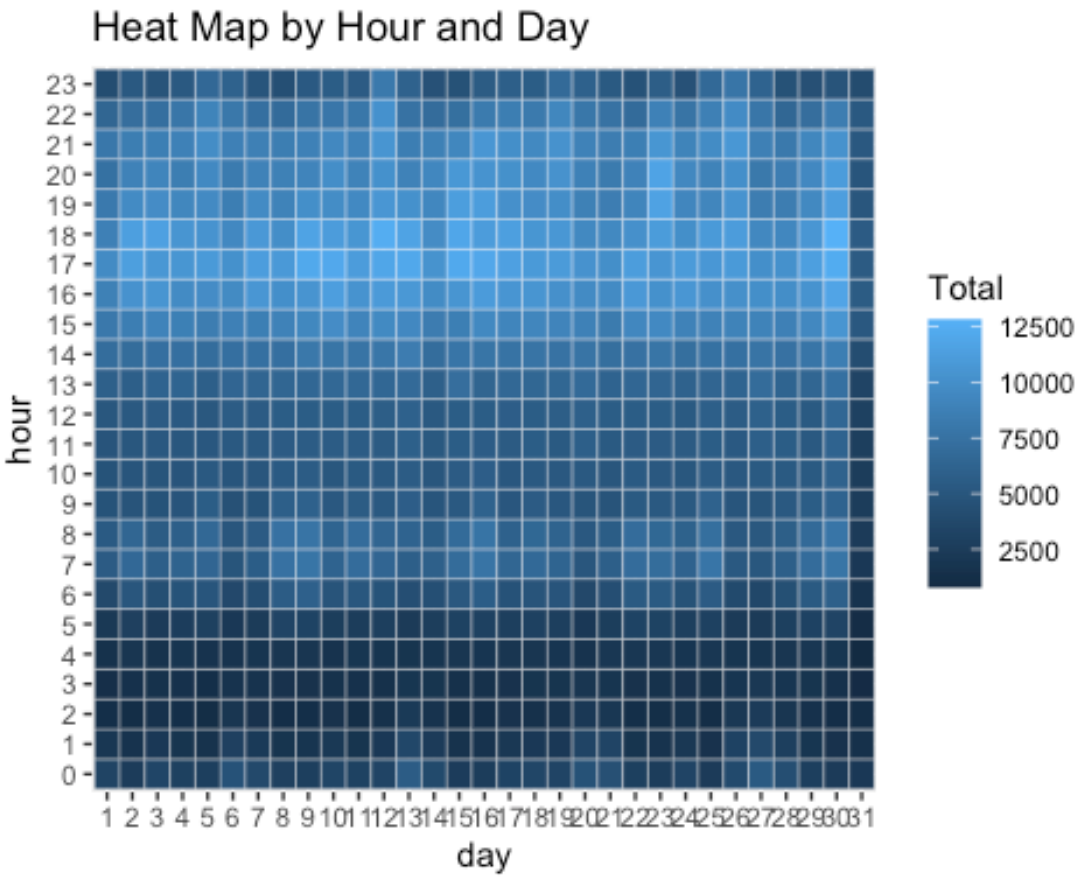

```
##HEAT MAPS

day_and_hour <- data %>%
         group_by(day, hour) %>%
            dplyr::summarize(Total = n())

## `summarise()` regrouping output by 'day' (override with `.groups`
argument)

datatable(day_and_hour)
```

Show 10 entries Search:

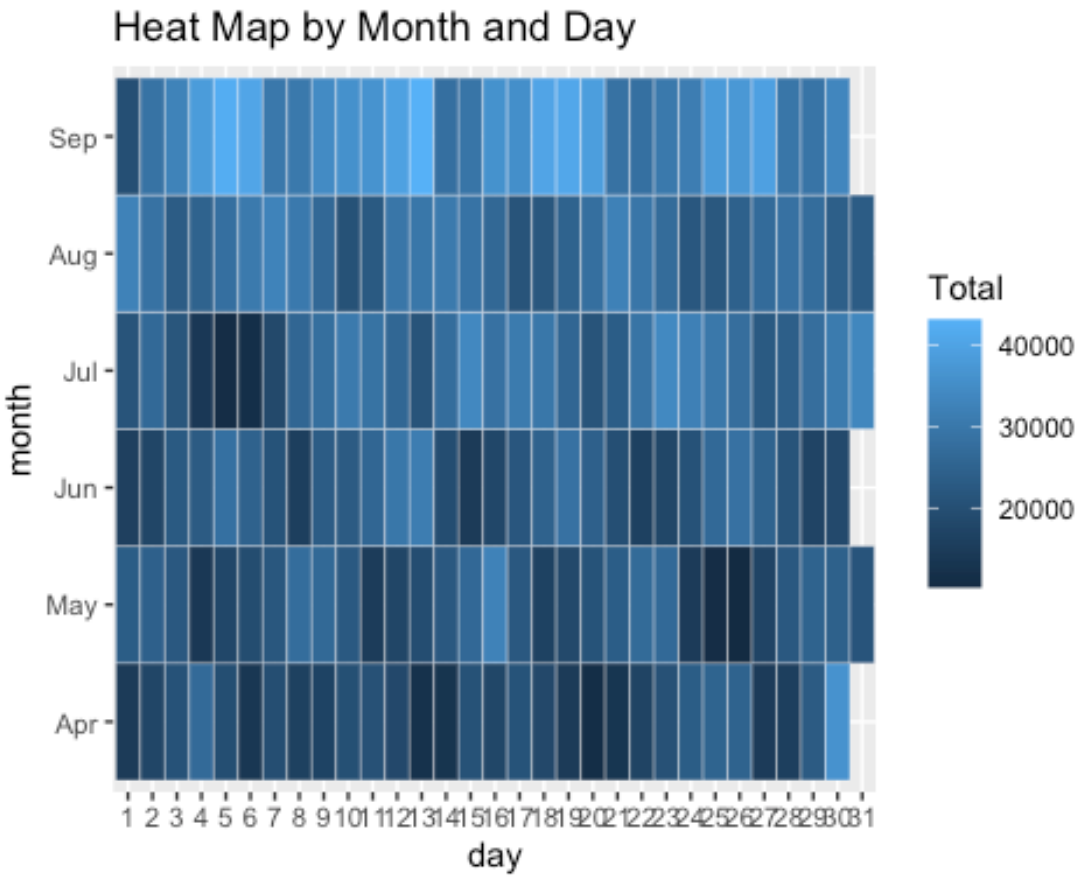
	day	hour	Total
1	1	0	3247
2	1	1	1982
3	1	2	1284
4	1	3	1331
5	1	4	1458
6	1	5	2171
7	1	6	3717
8	1	7	5470
9	1	8	5376
10	1	9	4688

Showing 1 to 10 of 744 entries Previous 1 2 3 4 5 … 75 Next

```
##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")
```

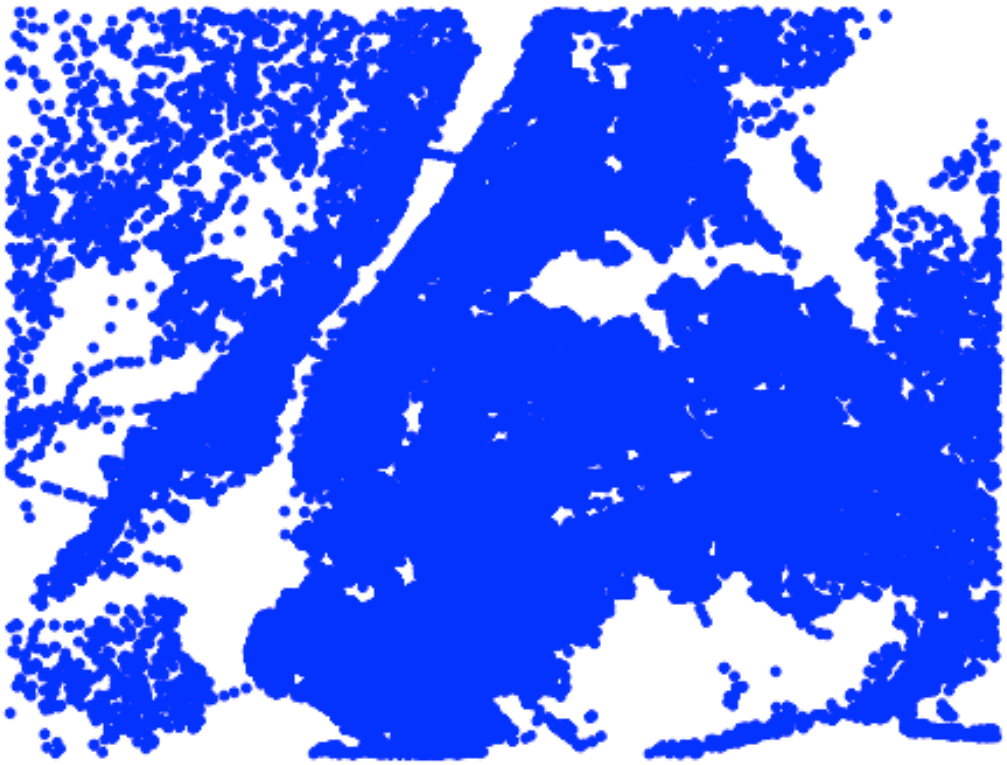



```
## HEAT MAP BY MONTH AND DAY

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




```
##GEO PLOT

min_lat <- 40.5774
max_lat <- 40.9176
min_long <- -74.15
max_long <- -73.7004

ggplot(data, 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)")

## Warning: Removed 71701 rows containing missing values (geom_point).
```

NYC MAP BASED ON UBER RIDES DURING 2014 (APR-SEP)



```
#Now geo data by base
ggplot(data, 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")

## Warning: Removed 71701 rows containing missing values (geom_point).
```

NYC MAP BASED ON UBER RIDES DURING 2014 (APR-SEP) by BASE

In this project, we performed different visualizations by utilizing ggplot to understand the number of trips by day, month and through every base. The bar graph shows the most number of rides were in the month of september. The heat map shows that along with highest number of rides in September, the most rides were used around 5 pm to 6 pm.The geo-mao visualizations shows rides throughout NY and from each base.