# Bike Share Analysis Udacity Project 2

Emmanuel Teikutey

9/29/2021

# Table of Contents

INTRODUCTION - BIKE SHARE DATA .......... 3
Major Questions to Answer .......... 3
THE DATA SET .......... 4
DATA ANALYSIS/ ANSWER TO QUESTIONS .......... 6
QUESTION ONE: The most common month of travel in the cities .......... 6
Visualisation: The most common of month in the cities .......... 8
QUESTION 2: What is the Most Common Start Station? .......... 12
VISUALISATION: Most Common Start Station .......... 13
QUESTION 3: What is the average travel time for users in different cities? .......... 15
VISUALISATION: Average time travel for cities .......... 16
OBESERVATIONS .......... 18
RECOMMENDATIONS .......... 18

## INTRODUCTION - BIKE SHARE DATA

Over the past decade, bicycle-sharing systems have been growing in number and popularity in cities across the world. Bicycle-sharing systems allow users to rent bicycles on a very short-term basis for a price. This allows people to borrow a bike from point A and return it at point B, though they can also return it to the same location if they'd like to just go for a ride. Regardless, each bike can serve several users per day.

Thanks to the rise in information technologies, it is easy for a user of the system to access a dock within the system to unlock or return bicycles. These technologies also provide a wealth of data that can be used to explore how these bike-sharing systems are used.

**#PROJECT OVERVIEW**

In this project, I will make use of R/ R studio to explore and visualise data related to bike share systems for three major cities in the United States—Chicago, New York City, and Washington. I will write code to import the data and answer interesting questions about it by computing descriptive statistics and making visualizations!

## Major Questions to Answer

There are a number of different areas of interest to explore in the data set but for the purpose of this Project I will be focusing on the following three questions:

1. The most common month of travel in the cities

2. The most common Start Station in the cities

3. Total travel time of users in the different cities.

## THE DATA SET

Randomly selected data for the first six months of 2017 are provided for all three cities.

Before we start answering the questions, let’s check the nature of the data set involved.

```
Ny=read.csv('new-york-city.csv')
wash=read.csv('washington.csv')
chi = read.csv('chicago.csv')
```

```
head(ny)
```

```
##         X          Start.Time            End.Time Trip.Duration
## 1 5688089 2017-06-11 14:55:05 2017-06-11 15:08:21           795
## 2 4096714 2017-05-11 15:30:11 2017-05-11 15:41:43           692
## 3 2173887 2017-03-29 13:26:26 2017-03-29 13:48:31          1325
## 4 3945638 2017-05-08 19:47:18 2017-05-08 19:59:01           703
## 5 6208972 2017-06-21 07:49:16 2017-06-21 07:54:46           329
## 6 1285652 2017-02-22 18:55:24 2017-02-22 19:12:03           998
##             Start.Station            End.Station  User.Type Gender
Birth.Year
## 1 Suffolk St & Stanton St W Broadway & Spring St Subscriber   Male
1998
## 2 Lexington Ave & E 63 St        1 Ave & E 78 St Subscriber   Male
1981
## 3       1 Pl & Clinton St   Henry St & Degraw St Subscriber   Male
1987
## 4   Barrow St & Hudson St        W 20 St & 8 Ave Subscriber Female
1986
## 5         1 Ave & E 44 St        E 53 St & 3 Ave Subscriber   Male
1992
## 6     State St & Smith St    Bond St & Fulton St Subscriber   Male
1986
```

```
head(wash)
```

```
##         X          Start.Time            End.Time Trip.Duration
## 1 1621326 2017-06-21 08:36:34 2017-06-21 08:44:43       489.066
## 2  482740 2017-03-11 10:40:00 2017-03-11 10:46:00       402.549
## 3 1330037 2017-05-30 01:02:59 2017-05-30 01:13:37       637.251
## 4  665458 2017-04-02 07:48:35 2017-04-02 08:19:03      1827.341
## 5 1481135 2017-06-10 08:36:28 2017-06-10 09:02:17      1549.427
## 6 1148202 2017-05-14 07:18:18 2017-05-14 07:24:56       398.000
##                                 Start.Station
## 1                        14th & Belmont St NW
## 2                  Yuma St & Tenley Circle NW
## 3              17th St & Massachusetts Ave NW
## 4            Constitution Ave & 2nd St NW/DOL
## 5 Henry Bacon Dr & Lincoln Memorial Circle NW
## 6                                1st & K St SE
##                                              End.Station  User.Type
## 1                                       15th & K St NW Subscriber
## 2                         Connecticut Ave & Yuma St NW Subscriber
## 3                                        5th & K St NW Subscriber
## 4                           M St & Pennsylvania Ave NW   Customer
## 5                                Maine Ave & 7th St SW Subscriber
## 6 Eastern Market Metro / Pennsylvania Ave & 7th St SE Subscriber
```

```
head(chi)
```

```
##         X          Start.Time            End.Time Trip.Duration
## 1 1423854 2017-06-23 15:09:32 2017-06-23 15:14:53           321
## 2  955915 2017-05-25 18:19:03 2017-05-25 18:45:53          1610
## 3    9031 2017-01-04 08:27:49 2017-01-04 08:34:45           416
## 4  304487 2017-03-06 13:49:38 2017-03-06 13:55:28           350
## 5   45207 2017-01-17 14:53:07 2017-01-17 15:02:01           534
## 6 1473887 2017-06-26 09:01:20 2017-06-26 09:11:06           586
##                    Start.Station                  End.Station  User.Type
Gender
## 1          Wood St & Hubbard St      Damen Ave & Chicago Ave Subscriber
Male
## 2           Theater on the Lake Sheffield Ave & Waveland Ave Subscriber
Female
## 3            May St & Taylor St          Wood St & Taylor St Subscriber
Male
## 4 Christiana Ave & Lawrence Ave St. Louis Ave & Balmoral Ave Subscriber
Male
## 5        Clark St & Randolph St Desplaines St & Jackson Blvd Subscriber
Male
## 6  Clinton St & Washington Blvd         Canal St & Taylor St Subscriber
Male
##   Birth.Year
## 1       1992
## 2       1992
## 3       1981
## 4       1986
## 5       1975
## 6       1990
```

```
dim(ny)
```

```
## [1] 300000      10
```

```
dim(wash)
```

```
## [1] 300000      8
```

```
dim(chi)
```

```
## [1] 300000      10
```

All three of the data files contain 300000 rows and 10 columns for Chicago and NewYork whiles Washington has 8 columns. But all data files have same core six (6) columns:

- Start Time (e.g., 2017-01-01 00:07:57)
- End Time (e.g., 2017-01-01 00:20:53)
- Trip Duration (in seconds - e.g., 776)
- Start Station (e.g., Broadway & Barry Ave)
- End Station (e.g., Sedgwick St & North Ave)
- User Type (Subscriber or Customer)

The Chicago and New York City files also have the following two columns:

- Gender
- Birth Year

# DATA ANALYSIS/ ANSWER TO QUESTIONS

## QUESTION ONE: The most common month of travel in the cities

The date and time is grouped in the datetime. The montths are also numerically labelled and I have relabel them characteristically. I need to extract the month from the date time and to do this I need to use the lubridate function which provides tools that make it easier to parse and manipulate dates.

```
library(lubridate)

##
## Attaching package: 'lubridate'

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

month <- function(city) {
  start_date <-
    sapply(strsplit(as.character(city$Start.Time), " "), "[", 1)
  # extract month from the datatime
  month <- substr(x = start_date, 6, 7)
}
# Call the month function to Chicago
month_chi <- month(chi)
# set a new column with the month data
chi['month'] <- month_chi
```

```r
# list the month column to check when needed
#chi['month']

# Call the month function to New York when needed
month_ny <- month(ny)
ny['month'] <- month_ny
#ny['month']

# Call the month function to Washington when needed
month_wash <- month(wash)
wash['month'] <- month_wash
#wash['month']


# replacing '01' with 'january' and '02' with 'February' and so on ...
old <- c('01', '02', '03', '04', '05', '06')
new <- c('January', 'Febraury', 'March', 'April', 'May', 'June')
chi$month[chi$month %in% old] <-
  new[match(chi$month, old, nomatch = 0)]
ny$month[ny$month %in% old] <-
  new[match(ny$month, old, nomatch = 0)]
wash$month[wash$month %in% old] <-
  new[match(wash$month, old, nomatch = 0)]

# find the unique values of months
# check the list when needed
#chi['month']
uniqv_chi <- unique(chi$month)
#ny['month']
uniqv_ny <- unique(ny$month)
#wash['month']
uniqv_wash <- unique(wash$month)

# check the list when needed
#uniqv_chi
#uniqv_ny
#uniqv_wash

# find the mode of the month
common_month <- function(data_column,uniqv) {
  uniqv[which.max(tabulate(match(data_column, uniqv)))]
}

# call the most common month function to Chicago
common_m_chi <- common_month(chi$month,uniqv_chi)
cat('The most common month for Chicago is:', common_m_chi,'\n')

## The most common month for Chicago is: June
```

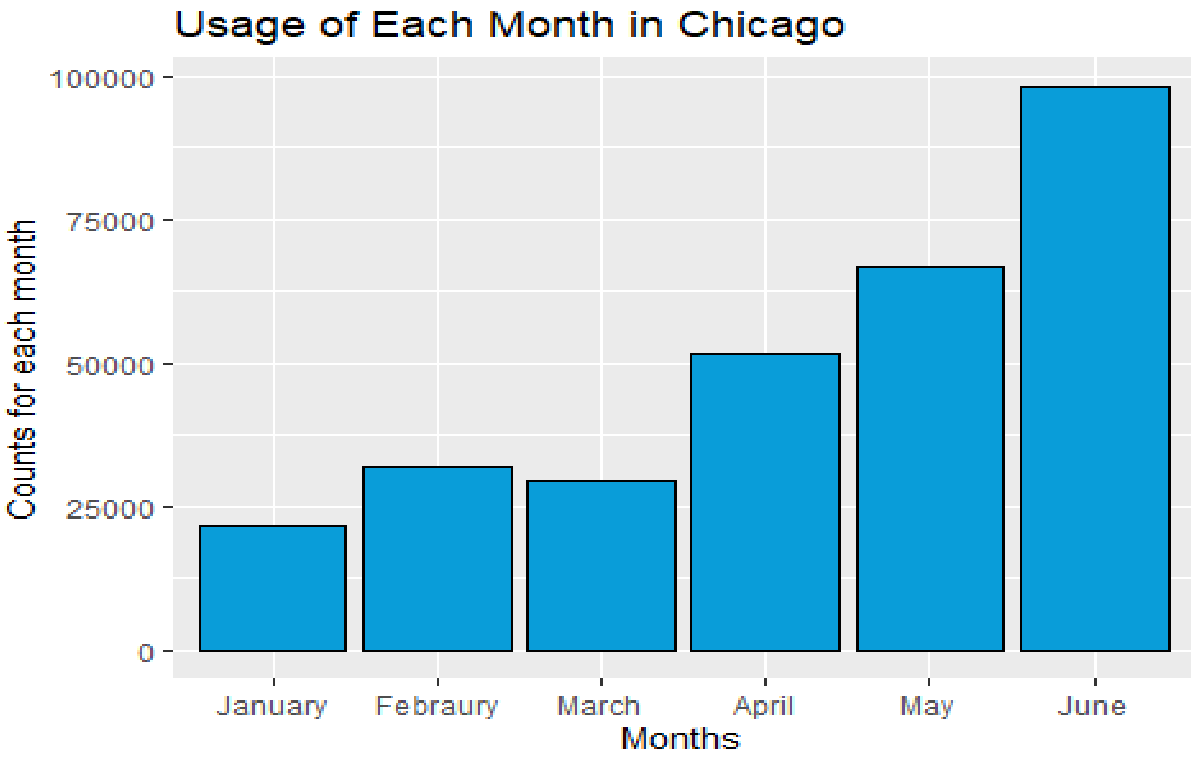
```
# call the most common month function to New York
common_m_ny <- common_month(ny$month,uniqv_ny)
cat('The most common month for New York is:', common_m_ny,'\n')

## The most common month for New York is: June

# call the most common month function to Washington
common_m_wash <- common_month(wash$month,uniqv_wash)
cat('The most common month for Washington is:', common_m_wash)

## The most common month for Washington is: June
```

In all cases/ cities, it is evidently clear that June is the most common month.

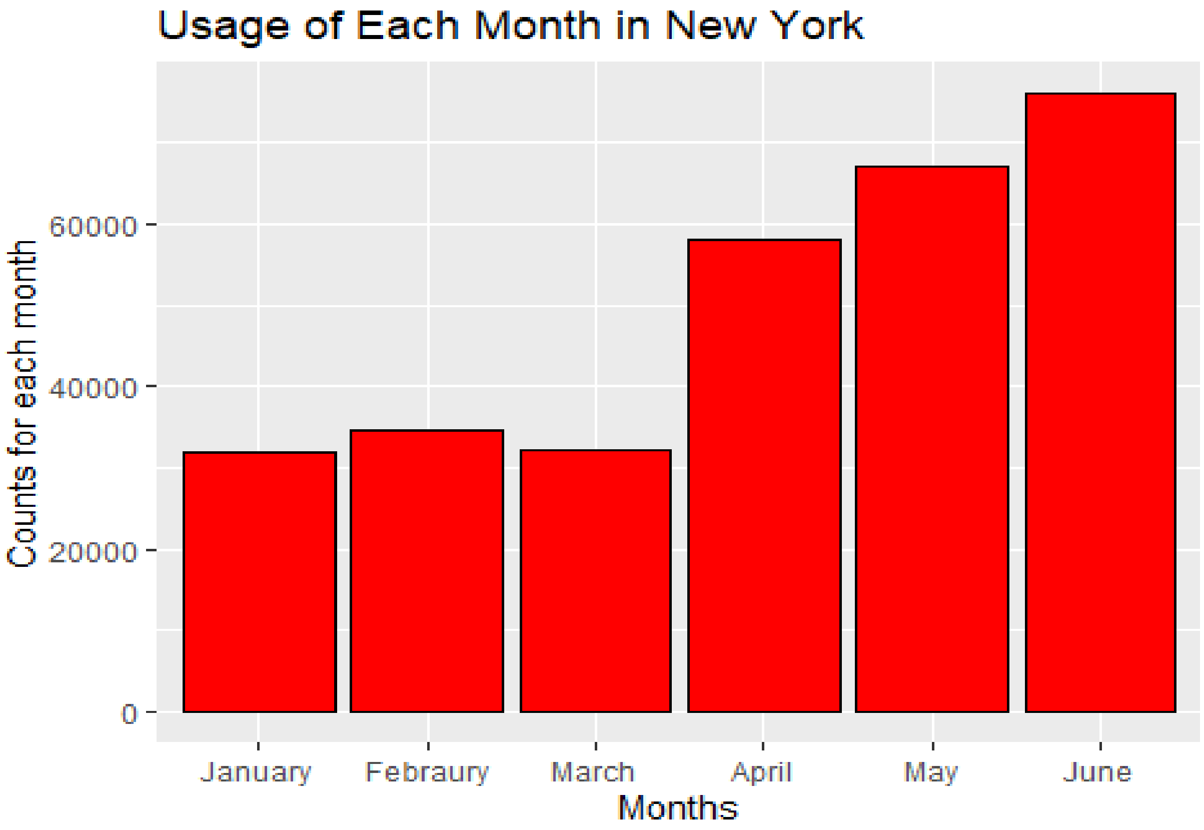
## Visualisation: The most common of month in the cities

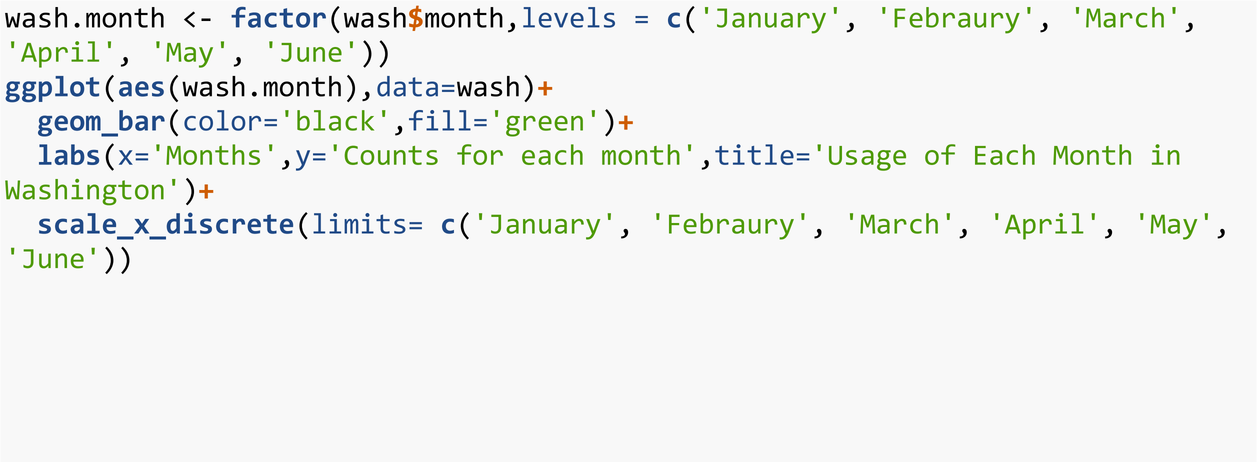
For visualization I chose to use ggplot2 function to create my graphs.

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

## Warning in as.POSIXlt.POSIXct(Sys.time()): unable to identify current
timezone 'C':
## please set environment variable 'TZ'
```

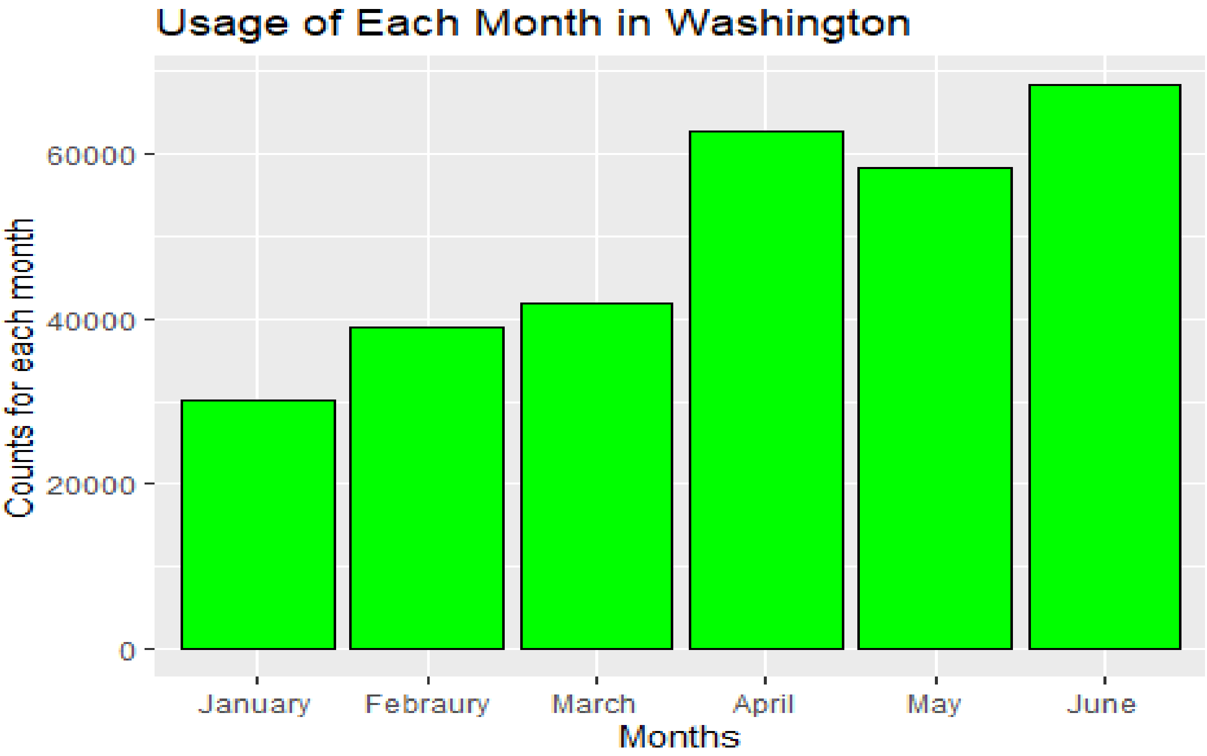


```
ny.month <- factor(ny$month,levels = c('January', 'Febraury', 'March',
'April', 'May', 'June'))
ggplot(aes(ny.month),data=ny)+
  geom_bar(color='black',fill= 'red')+
  labs(x='Months',y='Counts for each month',title='Usage of Each Month in New
York')
```





```
wash.month <- factor(wash$month,levels = c('January', 'Febraury', 'March',
'April', 'May', 'June'))
ggplot(aes(wash.month),data=wash)+
  geom_bar(color='black',fill='green')+
  labs(x='Months',y='Counts for each month',title='Usage of Each Month in
Washington')+
  scale_x_discrete(limits= c('January', 'Febraury', 'March', 'April', 'May',
'June'))
```



**Numeric Summary**

Chicago

```
summary(chi$month)
```

```
January February    March    April      May     June 

   21809    32057    29639    51659    66755    98081 
```

```
New York
summary(ny$month)
```

```
January February    March    April      May     June 

   31882    34741    32164    58176    67015    76022 
```

```


Washington
summary(wash$month)
```

```
January Febraury    March    April      May     June 

   30053    38932    41863    62620    58193    68339 
```

It is evidently clear from the numerical summaries that June was the most common month in all three cities.

## QUESTION 2: What is the Most Common Start Station?

To answer this question, I need to assign unique values and find the mode for the various start stations in the cites.

```
uniqv_ss_chi <- unique(chi$Start.Station)
uniqv_ss_ny <- unique(ny$Start.Station)
uniqv_ss_wash <- unique(wash$Start.Station)

common_ss <- function(data_column,uniqv_ss) {
  uniqv_ss[which.max(tabulate(match(data_column, uniqv_ss)))]
}
```

Call the most common month function to Chicago

```
common_ss_chi <- common_ss(chi$Start.Station,uniqv_ss_chi)
print('The most common start station for Chicago is:')

## [1] "The most common start station for Chicago is:"

common_ss_chi

## [1] "Streeter Dr & Grand Ave"
```

Call the most common month function to New York

```
common_ss_ny <- common_ss(ny$Start.Station,uniqv_ss_ny)
print('The most common start station for New York is:')

## [1] "The most common start station for New York is:"

common_ss_ny

## [1] "Pershing Square North"
```

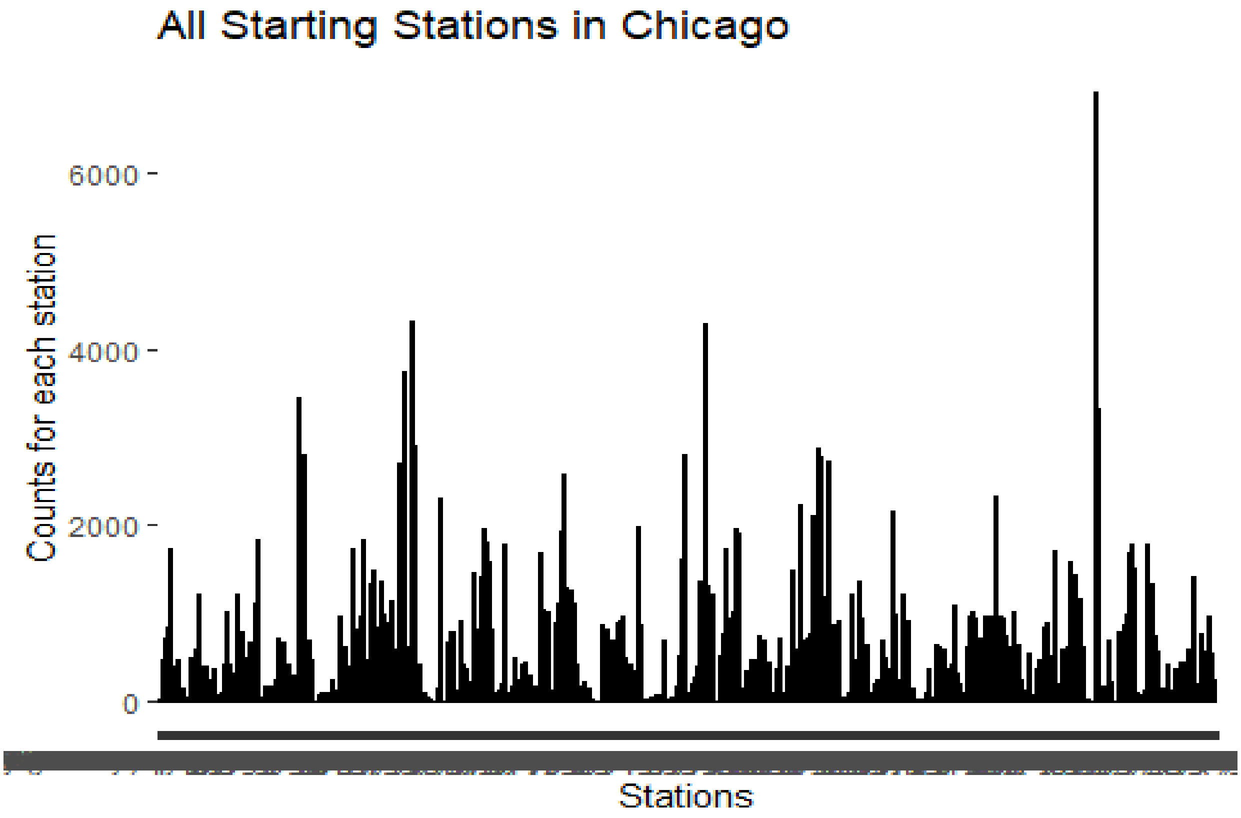
Call the most common month function to Washington

```
common_ss_wash <- common_ss(wash$Start.Station,uniqv_ss_wash)
print('The most common start station for Washington is:')

## [1] "The most common start station for Washington is:"

common_ss_wash

## [1] "Columbus Circle / Union Station"
```

## VISUALISATION: Most Common Start Station

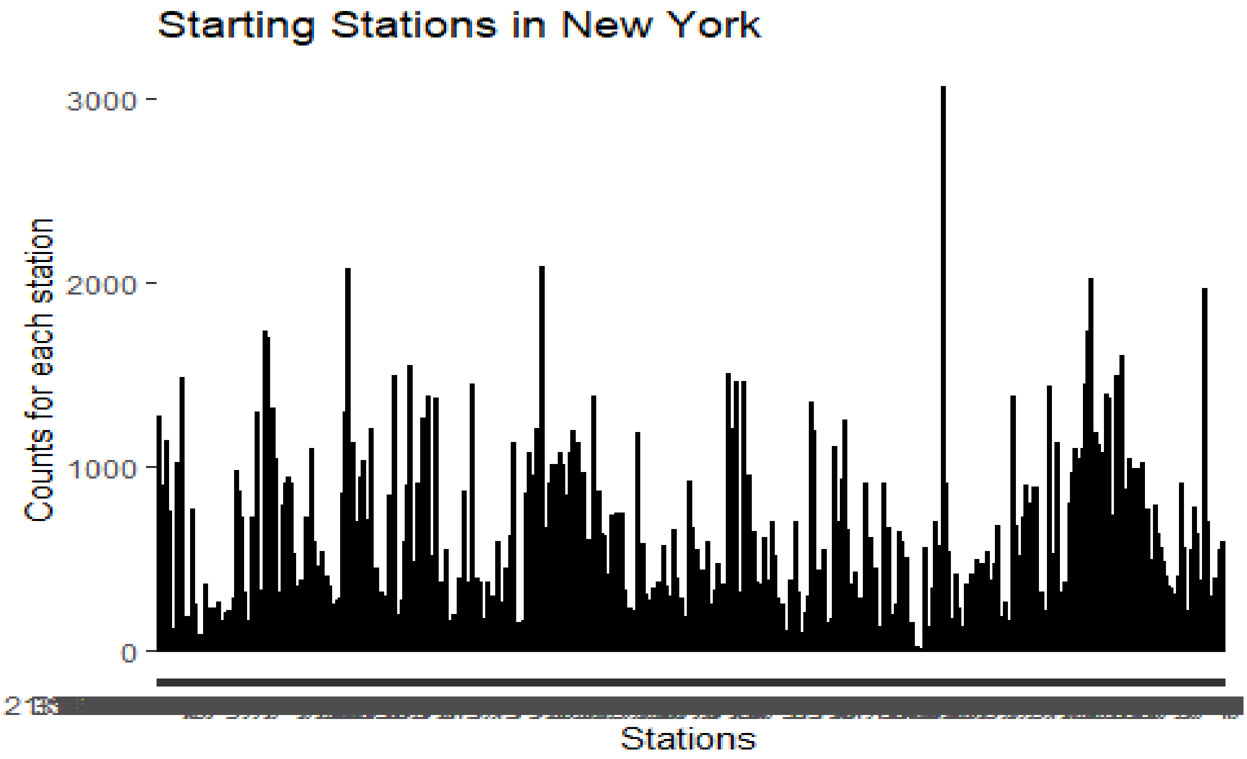
### Plot for Chicago

```
library(ggplot2)
library(scales)

ggplot(aes(chi$Start.Station),data=chi)+
  geom_bar(color='black',fill='#099DD9')+
  labs(x='Stations',y='Counts for each station',title='All Starting Stations
in Chicago')
```



### Plot for New York

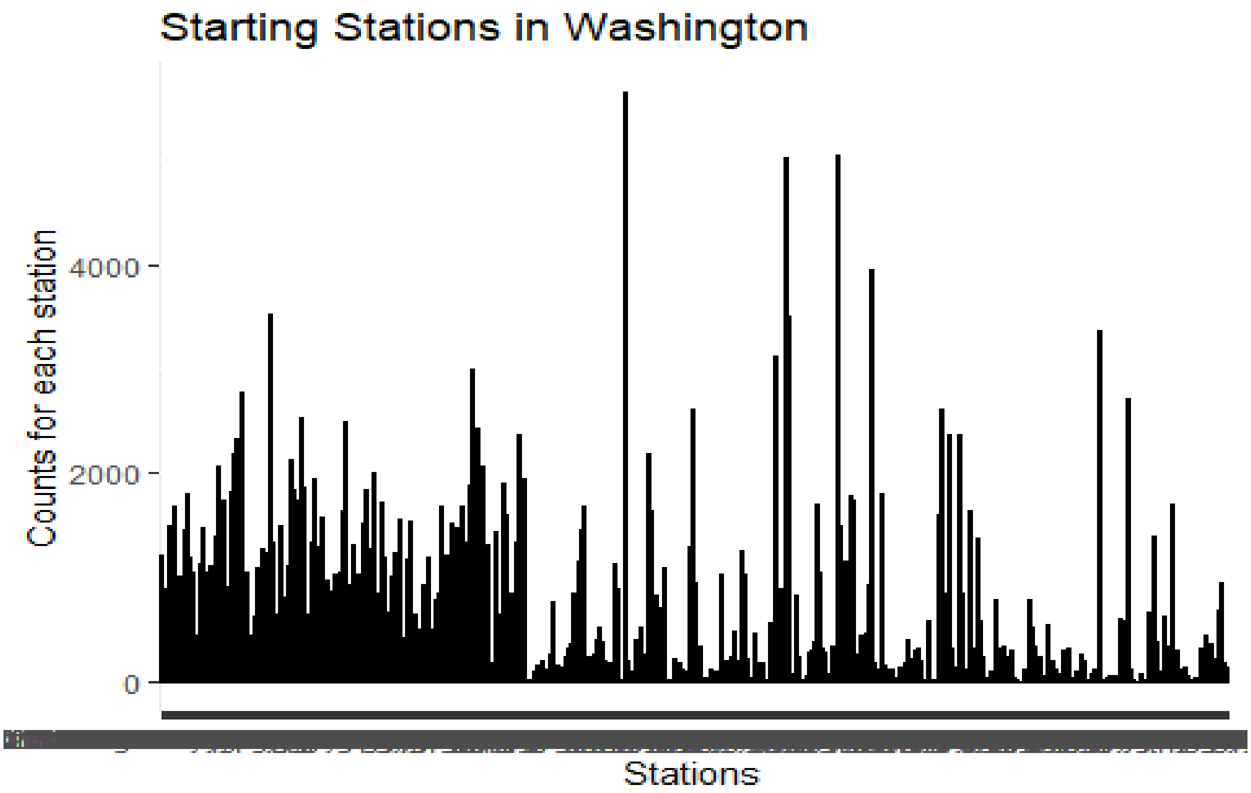
```
ggplot(aes(ny$Start.Station),data=ny)+
  geom_bar(color='black',fill='#099DD9')+
  labs(x='Stations',y='Counts for each station',title='Starting Stations in
New York')
```



### Plot for Washington

```
ggplot(aes(wash$Start.Station),data=wash)+
  geom_bar(color='black',fill='#099DD9')+
  labs(x='Stations',y='Counts for each station',title='Starting Stations in
Washington')
```

```
## Warning: Use of `wash$Start.Station` is discouraged. Use `Start.Station`
## instead.
```



## QUESTION 3: What is the average travel time for users in different cities?

To get the average travel time for users, I run the mean function on the Trip duration column for the different cities.

```
duration_chi = mean(chi$Trip.Duration)
duration_ny = mean(ny$Trip.Duration)
duration_wash = mean(wash$Trip.Duration)
```

Average time travel for Chicago

```
cat('The average travel time for Chicago is:', duration_chi,'\n')

## The average travel time for Chicago is: 936.2393
```

Average time travel for New York

```
cat('The average travel time for New York is:', duration_ny,'\n')
```

```
## The average travel time for New York is: 899.6842
```

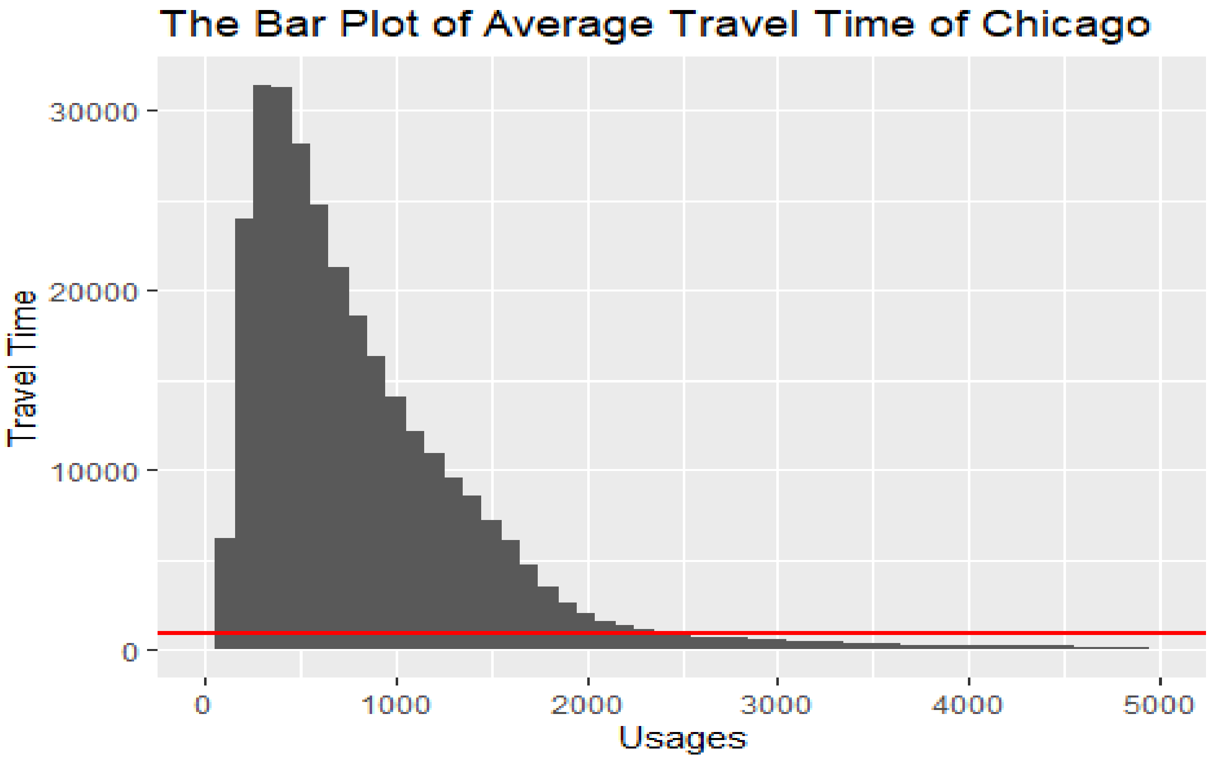
Average time travel for New York

```
cat('The average travel time for Washington is:', duration_wash,'\n')

## The average travel time for Washington is: 1237.28
```

## VISUALISATION: Average time travel for cities

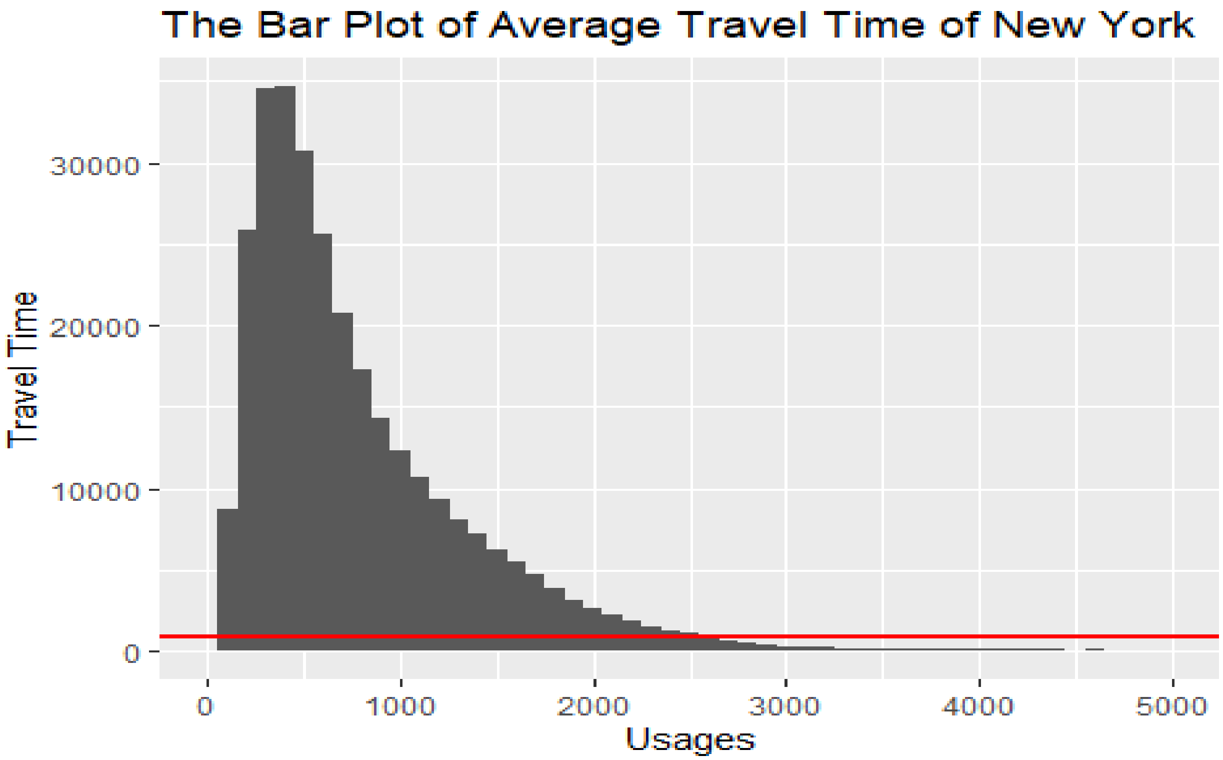
### Plot for Chicago

```
ggplot(aes(x=Trip.Duration),data=chi)+
  geom_histogram(binwidth = 100)+
  ggtitle('The Bar Plot of Average Travel Time of Chicago')+
  scale_x_continuous(limits = c(0,5000))+
  labs(x='Usages',y='Travel Time')+
  geom_hline(aes(yintercept = mean(Trip.Duration)),col='red',size=1)
```



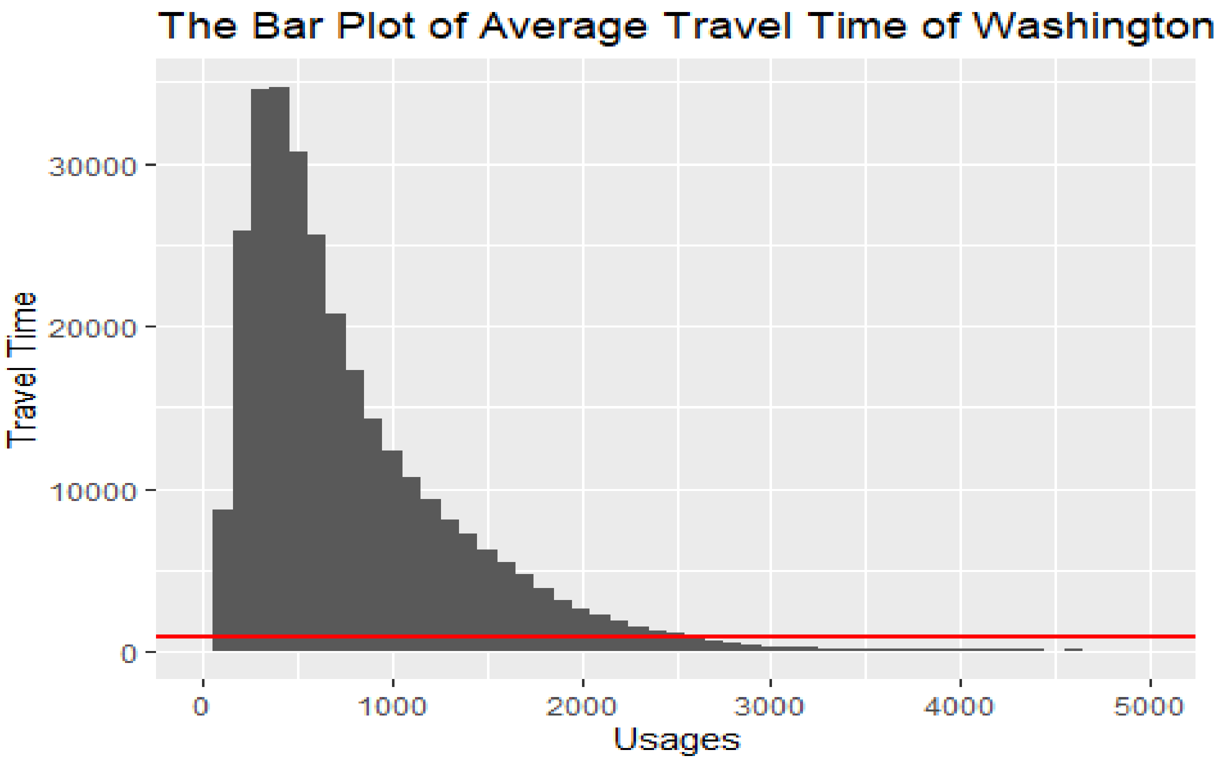
### Plot for New York

```r
ggplot(aes(x=Trip.Duration),data=ny)+
  geom_histogram(binwidth = 100)+
  ggtitle('The Bar Plot of Average Travel Time of New York')+
  scale_x_continuous(limits = c(0,5000))+
  labs(x='Usages',y='Travel Time')+
  geom_hline(aes(yintercept = mean(Trip.Duration)),col='red',size=1)
```



### Plot for Washington

```r
ggplot(aes(x=Trip.Duration),data=ny)+
  geom_histogram(binwidth = 100)+
  ggtitle('The Bar Plot of Average Travel Time of Washington')+
  scale_x_continuous(limits = c(0,5000))+
  labs(x='Usages',y='Travel Time')+
  geom_hline(aes(yintercept = mean(Trip.Duration)),col='red',size=1)
```



## OBESERVATIONS

- June is the most common months amongst the months under consideration
- The city with the highest travel time is New york
- The most common start station for Chicago is Street Dr & Grand Av, the one for New York is Perching Square North and Washington is Columbus Circle/ Union Station

## RECOMMENDATIONS

- More promotional activities should be carried out in other months other than June
- Offer discounted pricing in areas with low patronage