Explore_bikeshare_data-Copy1

October 19, 2019

0.0.1 Explore Bike Share Data

For this project, your goal is to ask and answer three questions about the available bikeshare data from Washington, Chicago, and New York. This notebook can be submitted directly through the workspace when you are confident in your results.

You will be graded against the project Rubric by a mentor after you have submitted. To get you started, you can use the template below, but feel free to be creative in your solutions!

```
In [35]: ny = read.csv('new_york_city.csv')
         wash = read.csv('washington.csv')
         chi = read.csv('chicago.csv')
```

```
In [4]: head(wash)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station
1621326	2017-06-21 08:36:34	2017-06-21 08:44:43	489.066	14th & Belmont St NW
482740	2017-03-11 10:40:00	2017-03-11 10:46:00	402.549	Yuma St & Tenley Circle NW
1330037	2017-05-30 01:02:59	2017-05-30 01:13:37	637.251	17th St & Massachusetts Ave NW
665458	2017-04-02 07:48:35	2017-04-02 08:19:03	1827.341	Constitution Ave & 2nd St NW/DOL
1481135	2017-06-10 08:36:28	2017-06-10 09:02:17	1549.427	Henry Bacon Dr & Lincoln Memorial
1148202	2017-05-14 07:18:18	2017-05-14 07:24:56	398.000	1st & K St SE

```
In [5]: head(chi)
```

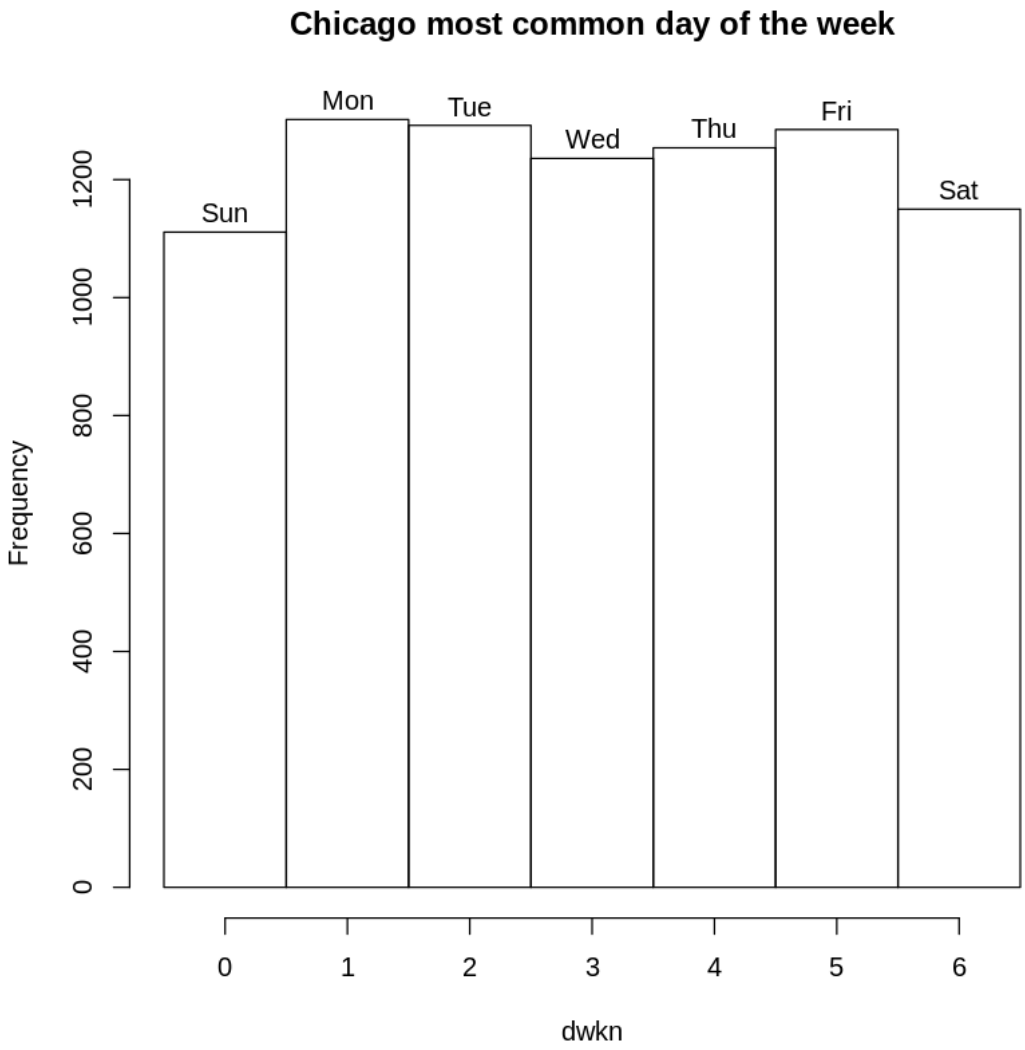
X	Start.Time	End.Time	Trip.Duration	Start.Station	End
1423854	2017-06-23 15:09:32	2017-06-23 15:14:53	321	Wood St & Hubbard St	Da
955915	2017-05-25 18:19:03	2017-05-25 18:45:53	1610	Theater on the Lake	She
9031	2017-01-04 08:27:49	2017-01-04 08:34:45	416	May St & Taylor St	Wo
304487	2017-03-06 13:49:38	2017-03-06 13:55:28	350	Christiana Ave & Lawrence Ave	St.
45207	2017-01-17 14:53:07	2017-01-17 15:02:01	534	Clark St & Randolph St	Des
1473887	2017-06-26 09:01:20	2017-06-26 09:11:06	586	Clinton St & Washington Blvd	Car

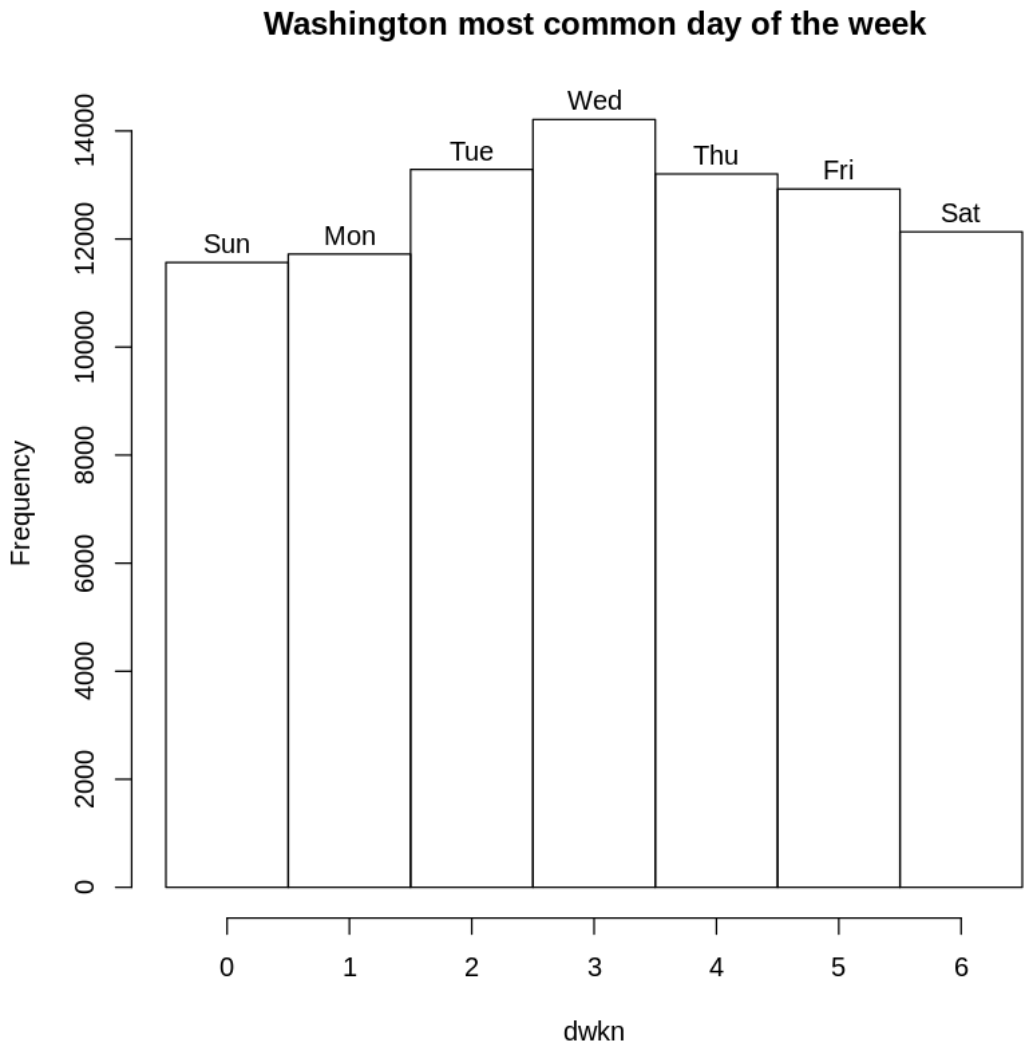
0.0.2 Question 1

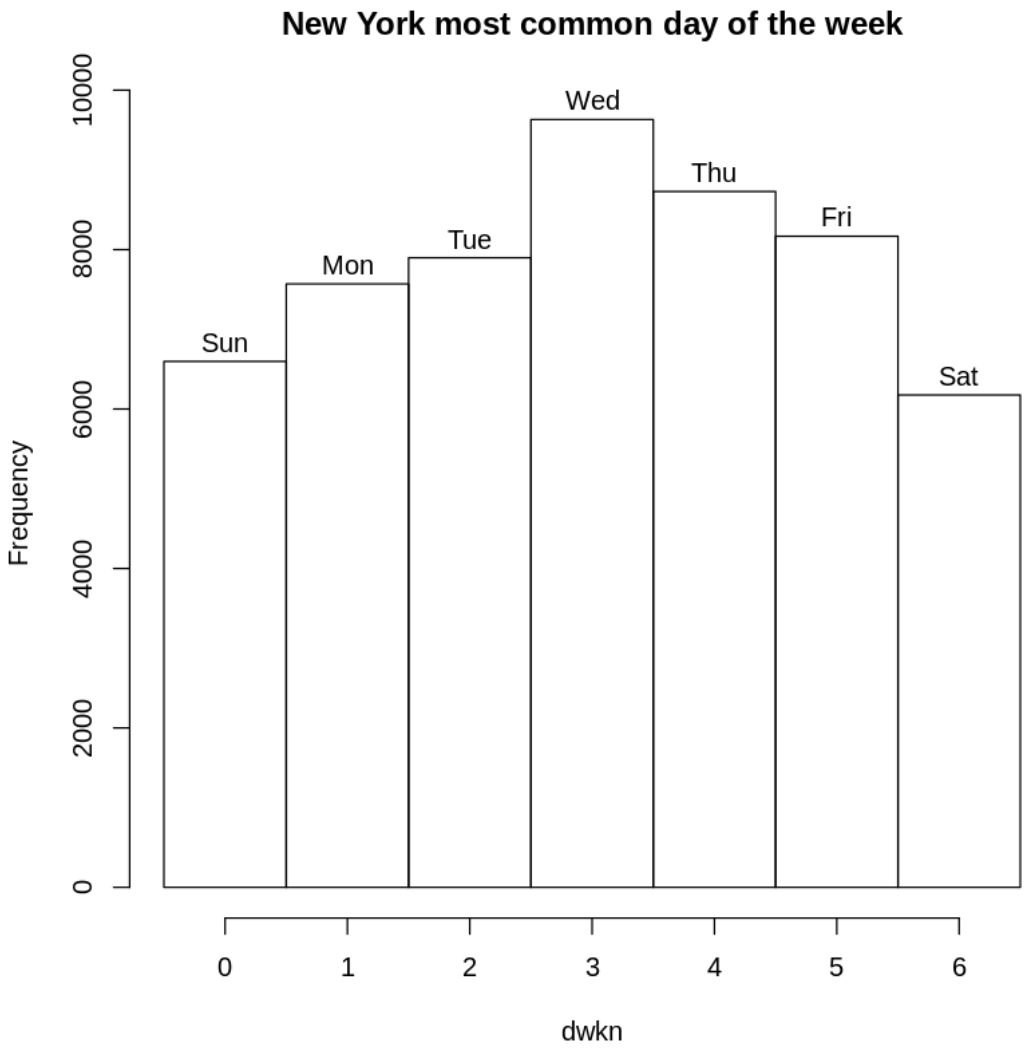
Your question 1 goes here.

```
In [36]: #Most common day of the week
         # Chicago
         library(ggplot2)
```

```r
library(lubridate)
date_Start.Time <- as.Date(chi$Start.Time)
date <- c(date_Start.Time)
x <- ymd(date)
dwka <- format(x , "%a")
dwkn <- as.numeric( format(x , "%w") )
hist( dwkn , breaks= -.5+0:7, labels= unique(dwka[order(dwkn)]), main = "Chicago most c

# Wash
library(ggplot2)
library(lubridate)
date_Start.Time <- as.Date(wash$Start.Time)
date <- c(date_Start.Time)
x <- ymd(date)
dwka <- format(x , "%a")
dwkn <- as.numeric( format(x , "%w") )
hist( dwkn , breaks= -.5+0:7, labels= unique(dwka[order(dwkn)]), main = "Washington mos

# NY
library(ggplot2)
library(lubridate)
date_Start.Time <- as.Date(ny$Start.Time)
date <- c(date_Start.Time)
x <- ymd(date)
dwka <- format(x , "%a")
dwkn <- as.numeric( format(x , "%w") )
hist( dwkn , breaks= -.5+0:7, labels= unique(dwka[order(dwkn)]), main = "New York most
```


Chicago most common day of the week
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Frequency
0
200
400
600
800
1000
1200
0
1
2
3
4
5
6
dwkn


Washington most common day of the week
Frequency
0
2000
4000
6000
8000
10000
12000
14000
Sun
Mon
Tue
Wed
Thu
Fri
Sat
0
1
2
3
4
5
6
dwkn



In [ ]:

The most common day of the week in Chicago is Monday, Wednesday in Washington and New York.

0.0.3 Question 2

Your question 2 goes here.

In [41]:
```
#What are the counts of each user type?

table(chi$User.Type)
table(wash$User.Type)
```

```r
table(ny$User.Type)

#What are the counts of each gender (only available for NYC and Chicago)?

table(chi$Gender)
table(wash$Gender)
table(ny$Gender)


#What are the earliest, most recent, most common year of birth (only available for NYC

summary(ny$Birth.Year)
summary(chi$Birth.Year)

library(ggplot2)
ggplot(data = ny, aes(x = ny$Birth.Year)) +
  geom_histogram(binwidth = 1)+
  scale_x_continuous(limits = c(1920, 2001), breaks = seq(1920, 2001, 10)) + labs(title


ggplot( data = chi, aes(x = chi$Birth.Year)) +
  geom_histogram(binwidth = 1)+
  scale_x_continuous(limits = c(1920, 2001), breaks = seq(1920, 2001, 10)) + labs(title
```

```
    Customer Subscriber
1       1746       6883
```

```
    Customer Subscriber
1      23450      65600
```

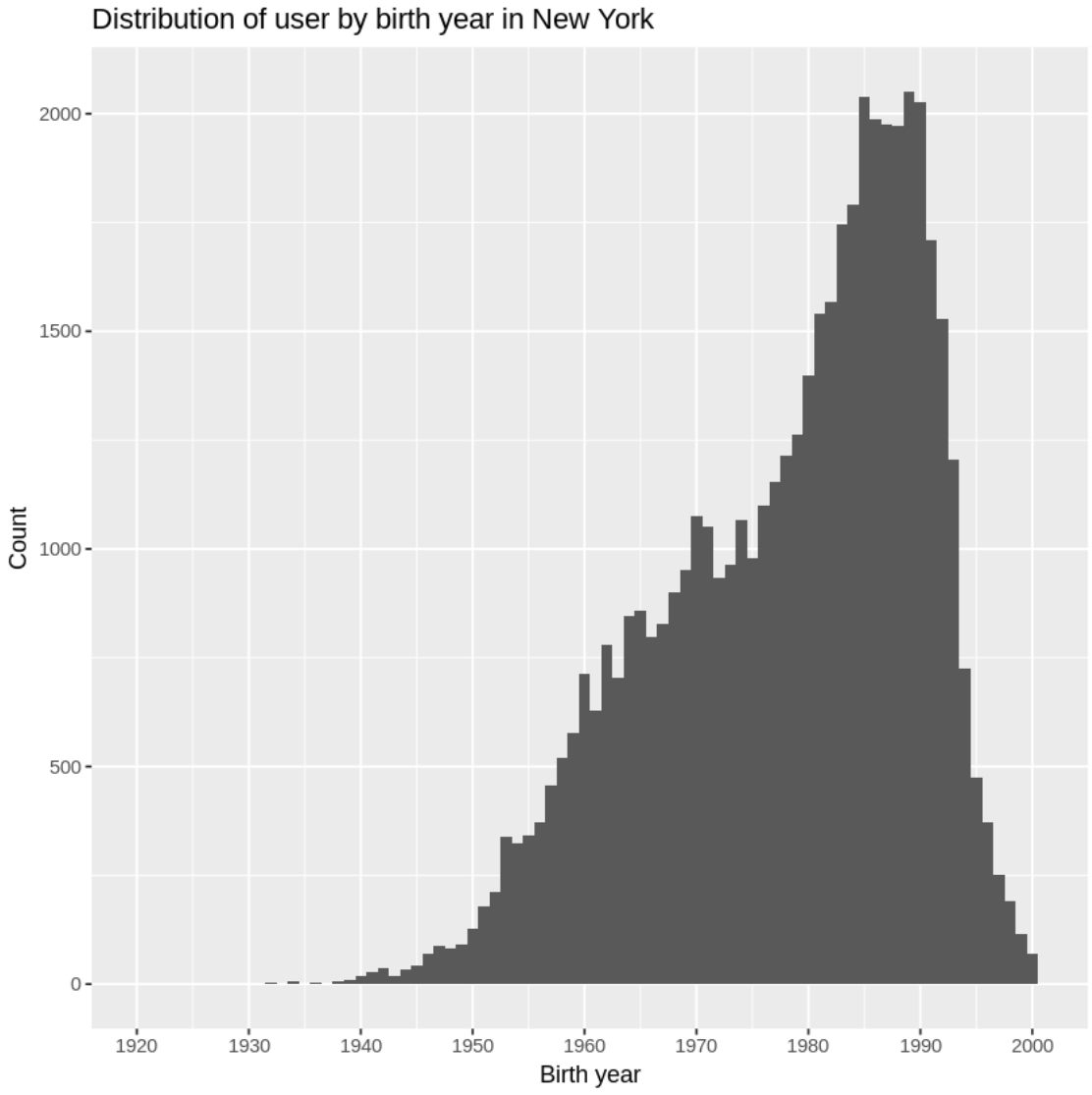
```
    Customer Subscriber
119      5558      49093
```

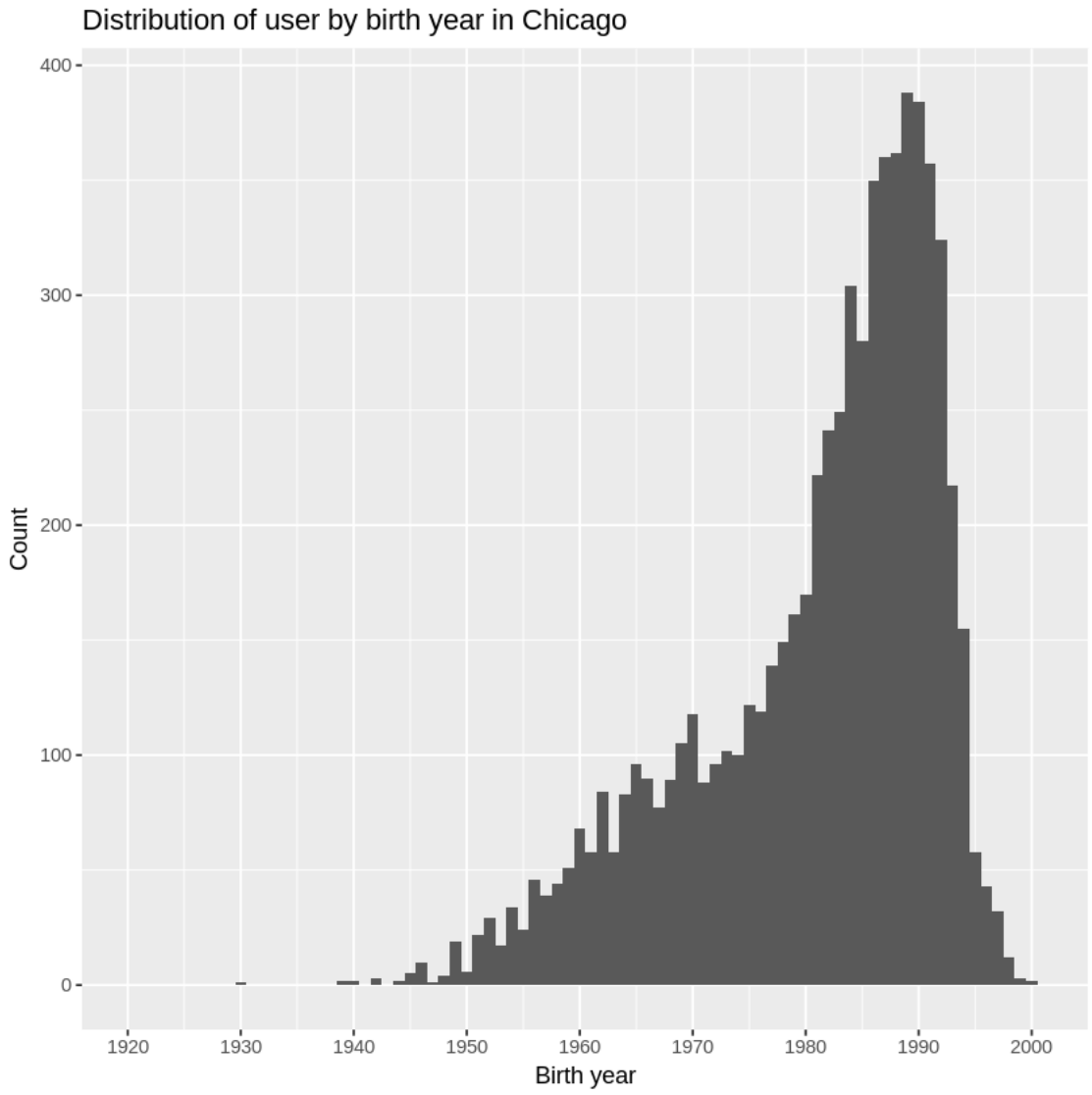
```
     Female   Male
1748   1723   5159
```

```
< table of extent 0 >
```

```
     Female   Male
```

```
  5410   12159   37201

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1885    1970    1981    1978    1988    2001    5218

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1899    1975    1984    1981    1989    2002    1747

Warning message:
Removed 5238 rows containing non-finite values (stat_bin).Warning message:
Removed 2 rows containing missing values (geom_bar).Warning message:
Removed 1754 rows containing non-finite values (stat_bin).Warning message:
Removed 2 rows containing missing values (geom_bar).
```





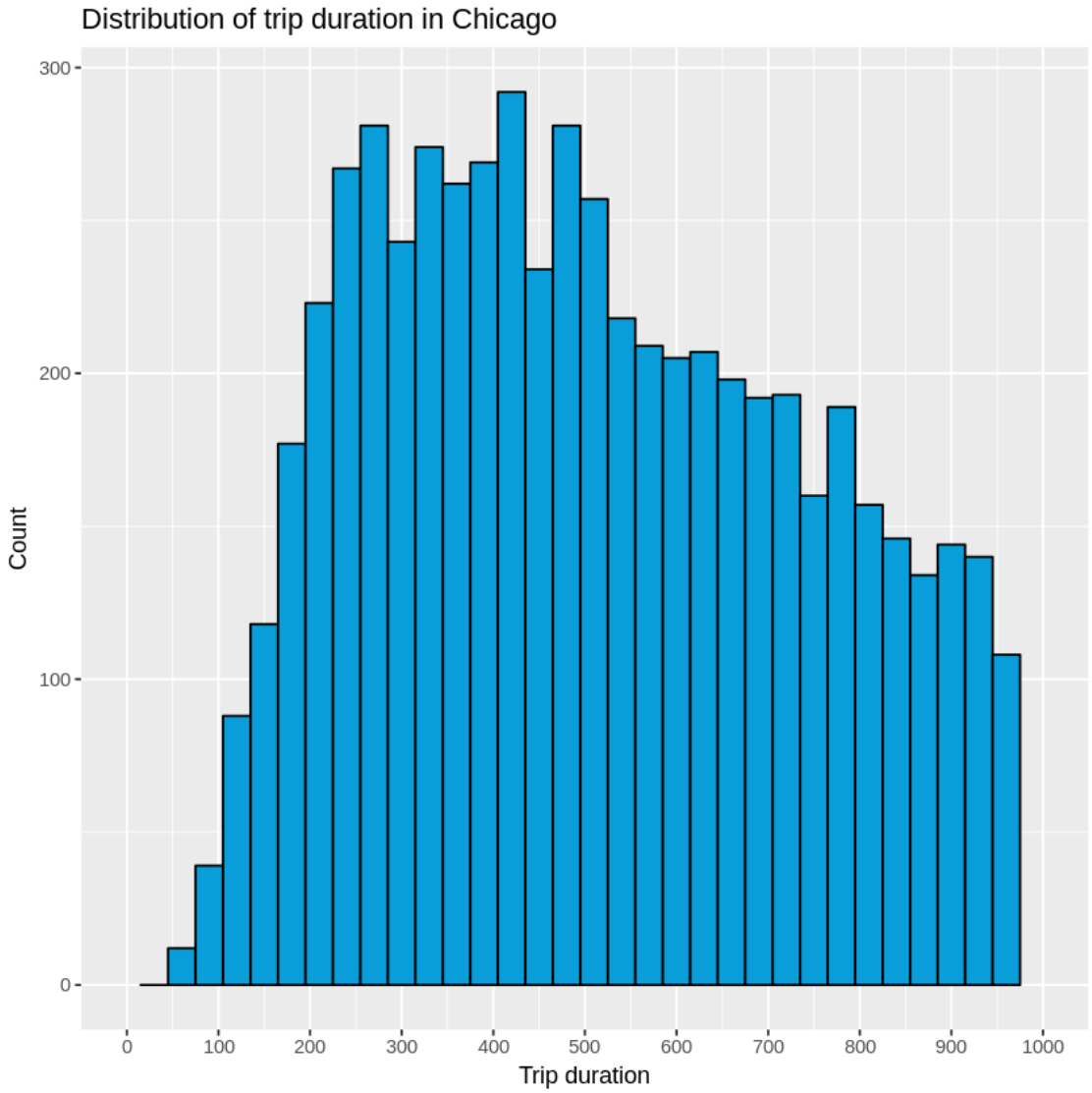
In [30]:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   394.2   670.0   937.2  1119.0 85408.0
```

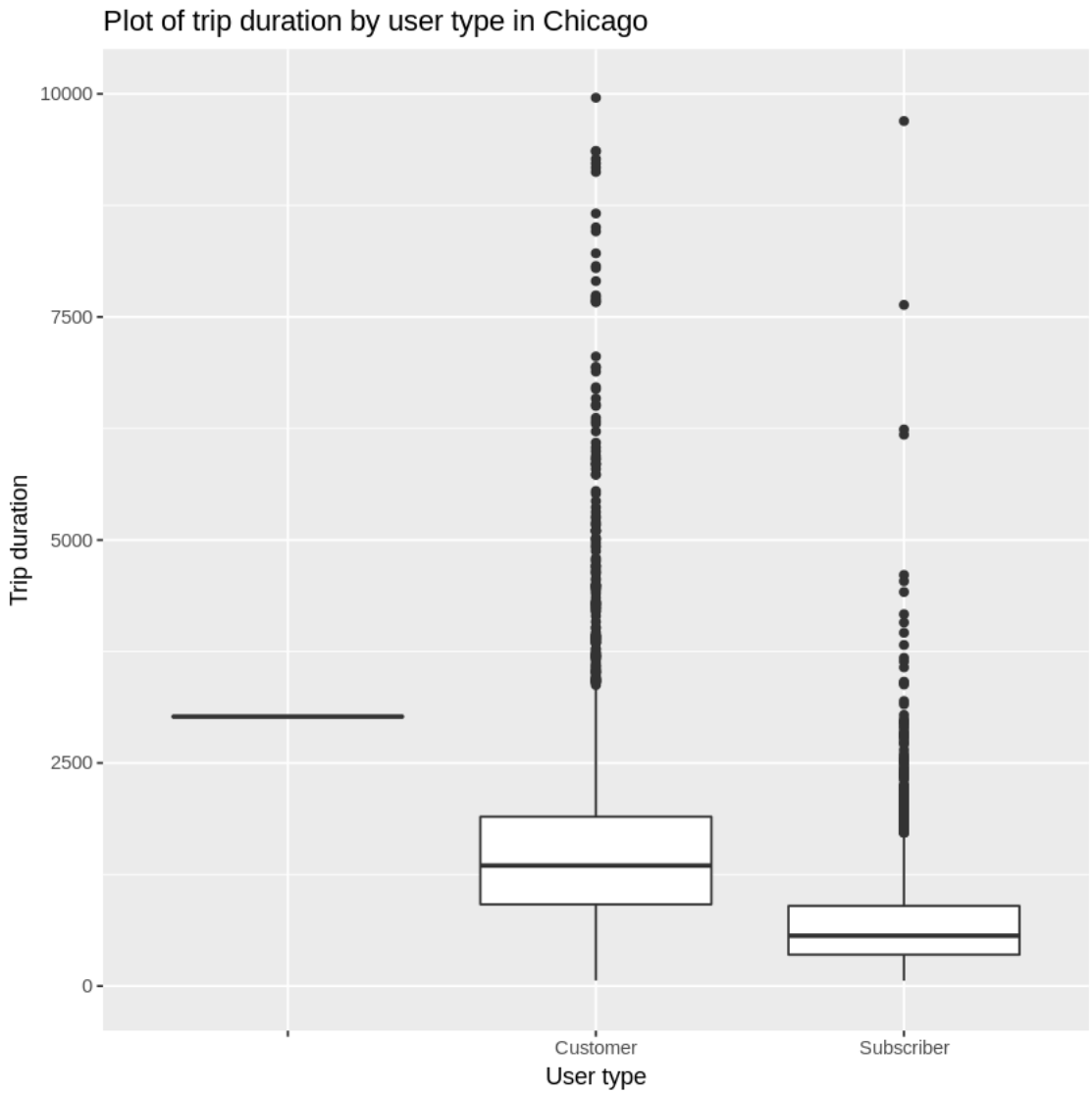
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
    60.3    410.9    707.0   1234.0   1233.2 904591.4        1
```

```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
   61.0     368.0     610.0     903.6    1051.0 1088634.0         1
```

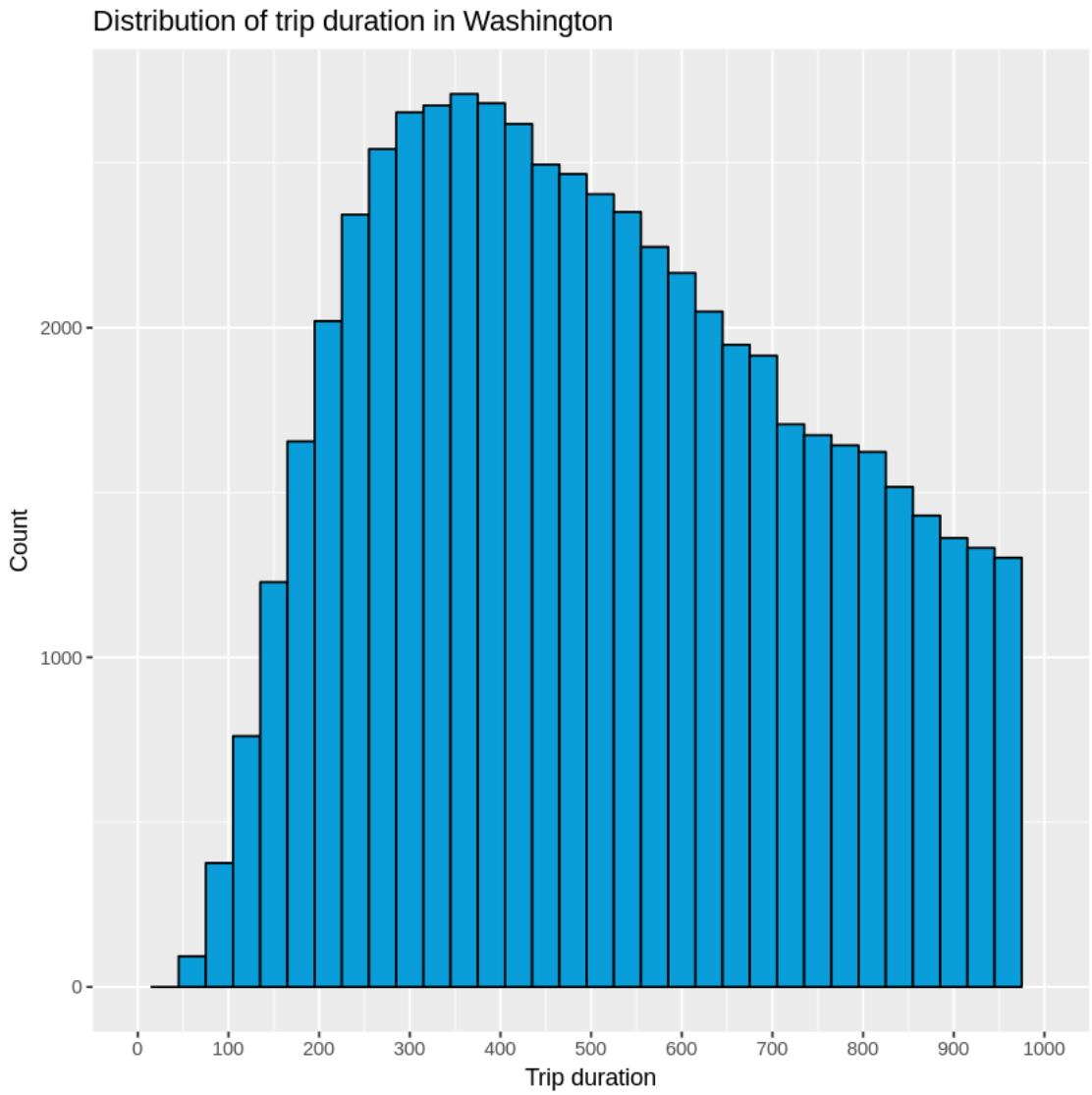
```
Warning message:
Removed 2597 rows containing non-finite values (stat_bin).Warning message:
Removed 2 rows containing missing values (geom_bar).Warning message:
Removed 27 rows containing non-finite values (stat_boxplot).
```



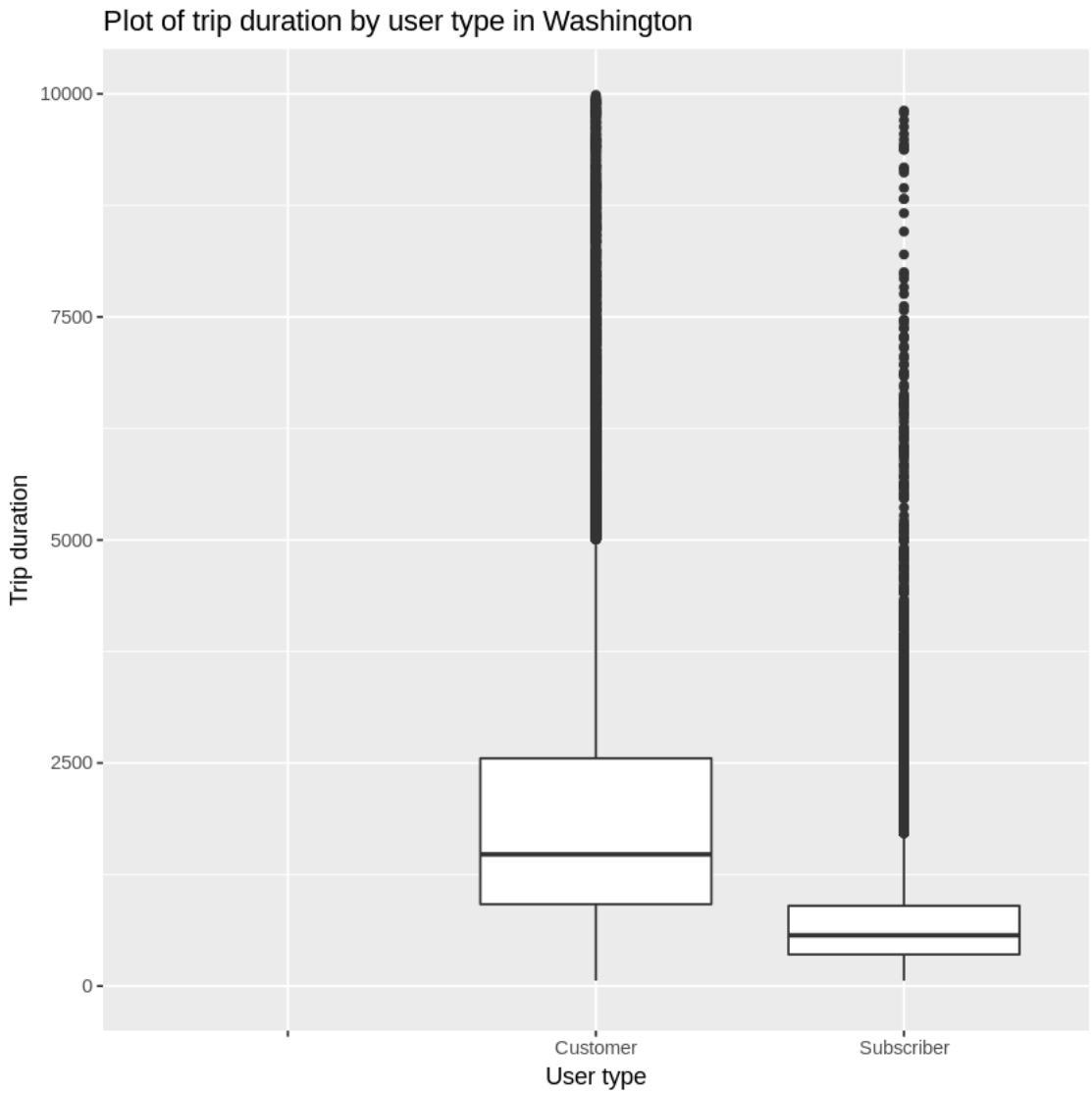

```
Warning message:
Removed 30046 rows containing non-finite values (stat_bin).Warning message:
Removed 2 rows containing missing values (geom_bar).
```



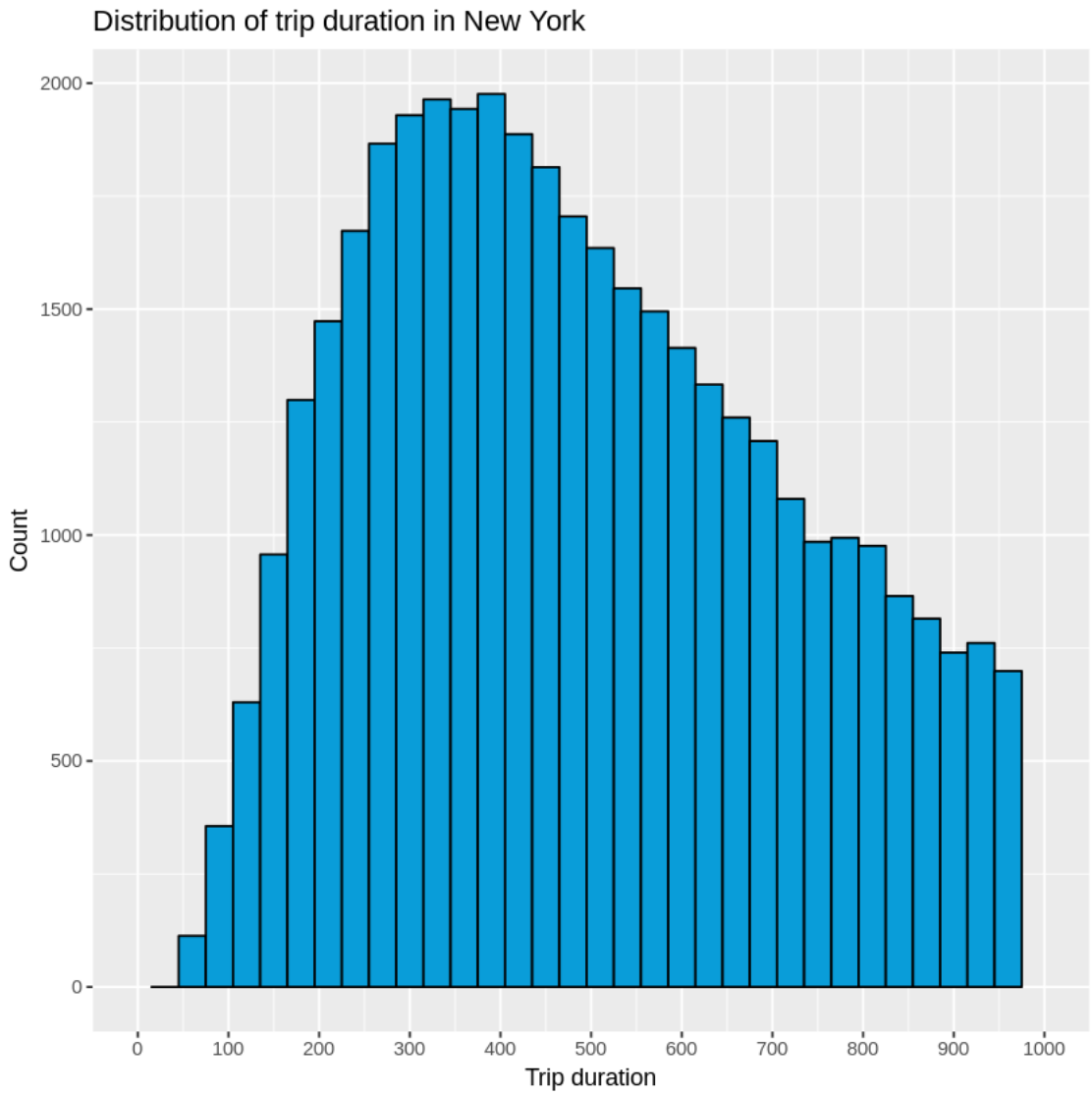

```
Warning message:
Removed 613 rows containing non-finite values (stat_boxplot).
```

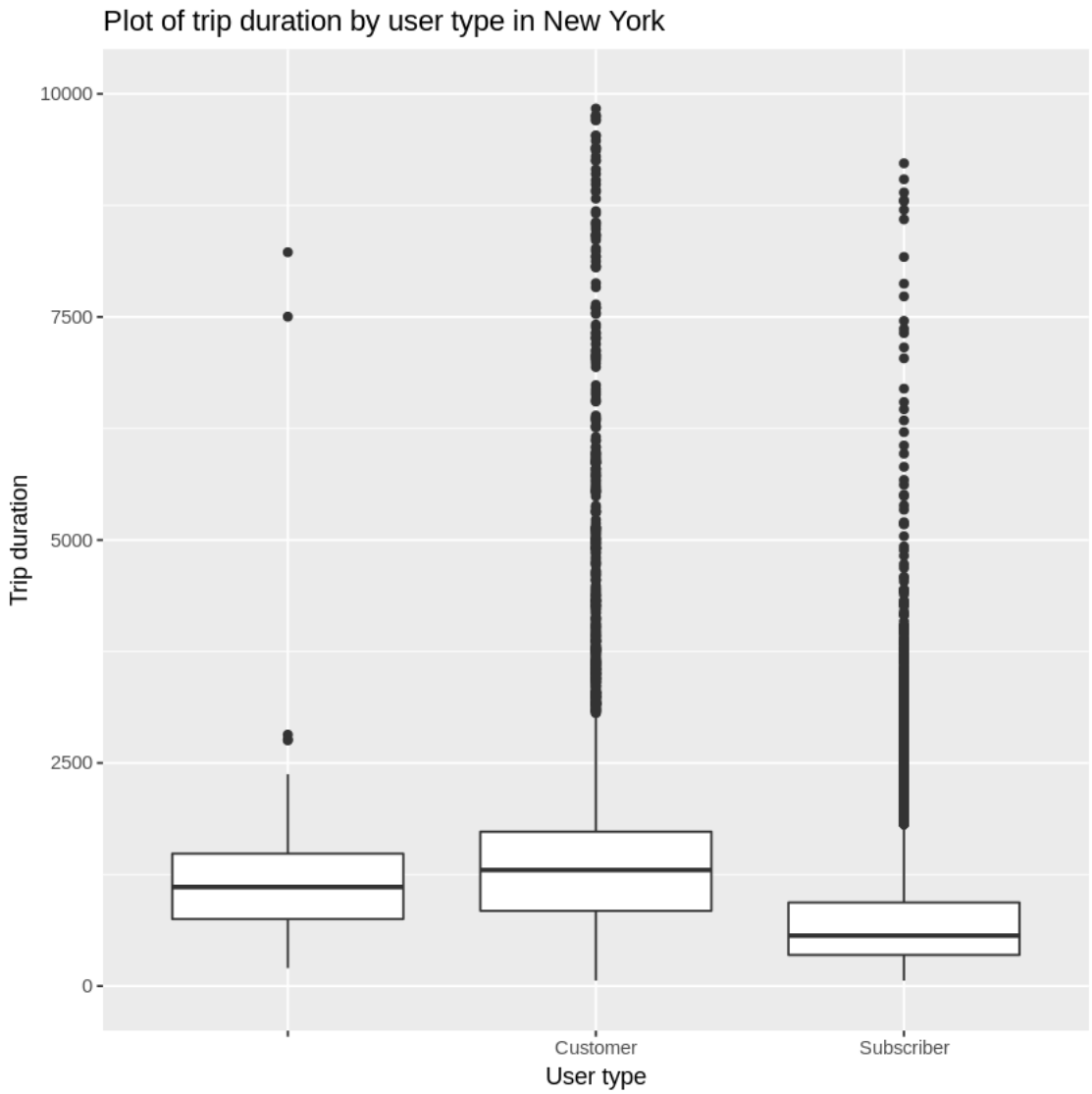



```
Warning message:
Removed 14808 rows containing non-finite values (stat_bin).Warning message:
Removed 2 rows containing missing values (geom_bar).
```




```
Warning message:
Removed 114 rows containing non-finite values (stat_boxplot).
```


Distribution of trip duration in New York
Count
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
900
1000
Trip duration



In [ ]: The first 3 tables show the break-down user by type in Chicago, Washington and NY. The n

0.0.4 Question 3

Your question 3 goes here.

In [ ]: library(ggplot2)

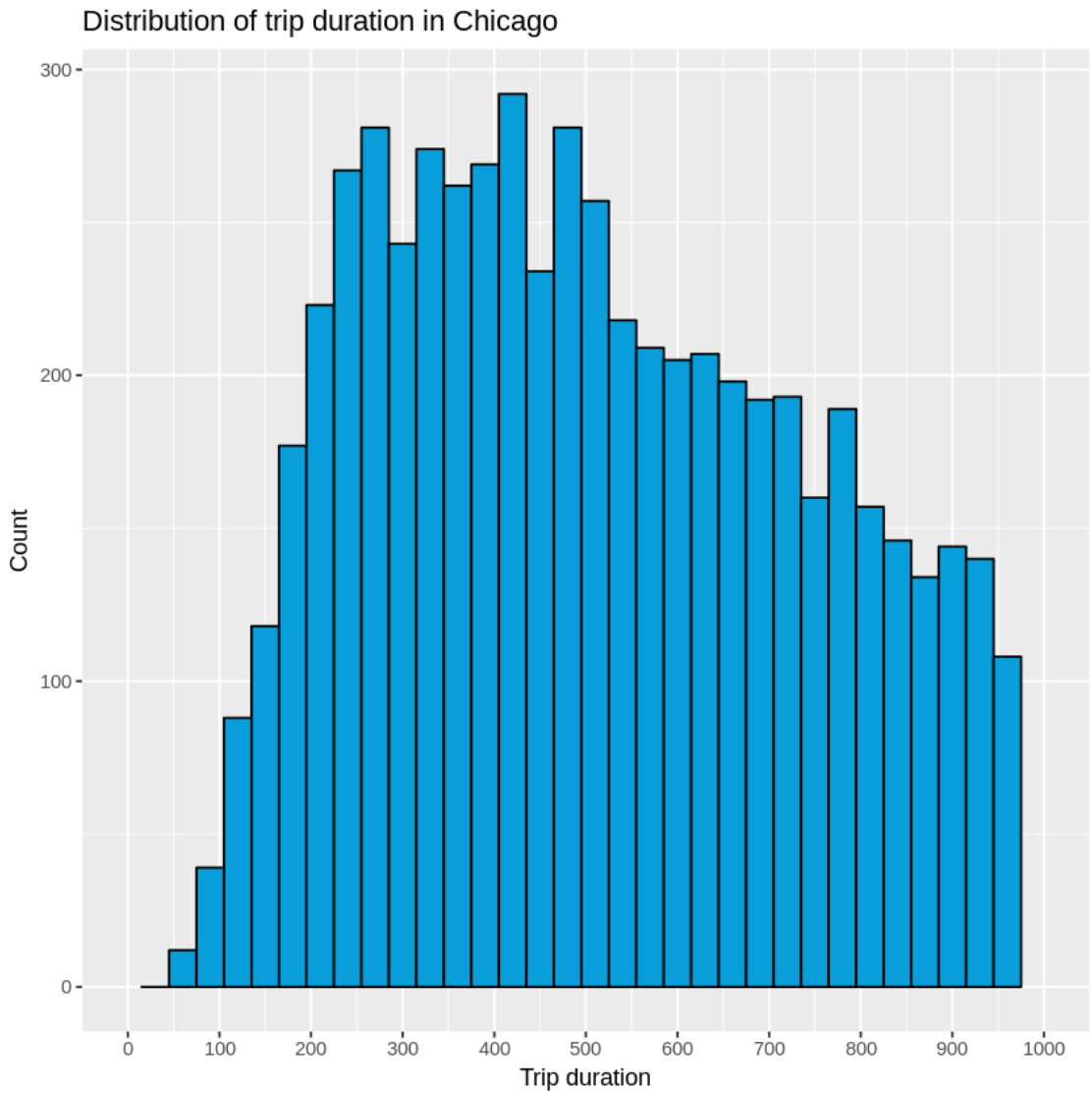
In [38]: *#3 Trip duration*

#What is the total travel time for users in different cities?

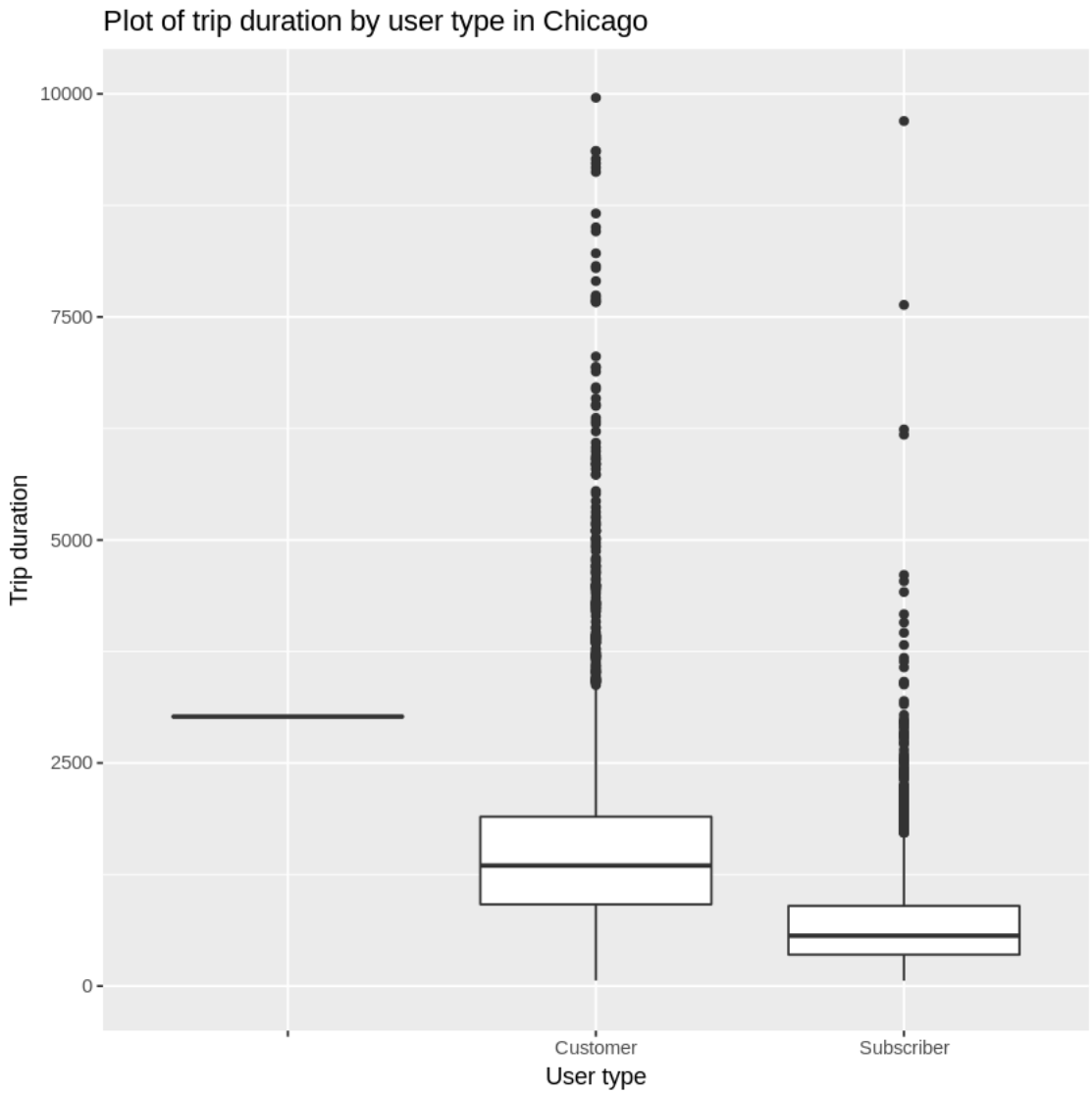
#What is the average travel time for users in different cities?

```r
summary(chi$Trip.Duration)
summary(wash$Trip.Duration)
summary(ny$Trip.Duration)

ggplot(data = chi, aes(x = chi$Trip.Duration)) +
geom_histogram(binwidth = 30, color = 'black', fill = '#099DD9') +
scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 100)) + labs(title="Distr

qplot(x = User.Type, y = Trip.Duration, data = subset(chi, !is.na(User.Type)), geom = '


ggplot(data = wash, aes(x = wash$Trip.Duration)) +
  geom_histogram(binwidth = 30, color = 'black', fill = '#099DD9') +
scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 100)) + labs(title="Distr


qplot(x = User.Type, y = Trip.Duration, data = subset(wash, !is.na(User.Type)), geom =


ggplot(data = ny, aes(x = ny$Trip.Duration)) +
  geom_histogram(binwidth = 30, color = 'black', fill = '#099DD9') +
scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 100)) + labs(title="Distr


qplot(x = User.Type, y = Trip.Duration, data = subset(ny, !is.na(User.Type)), geom = 'b
```

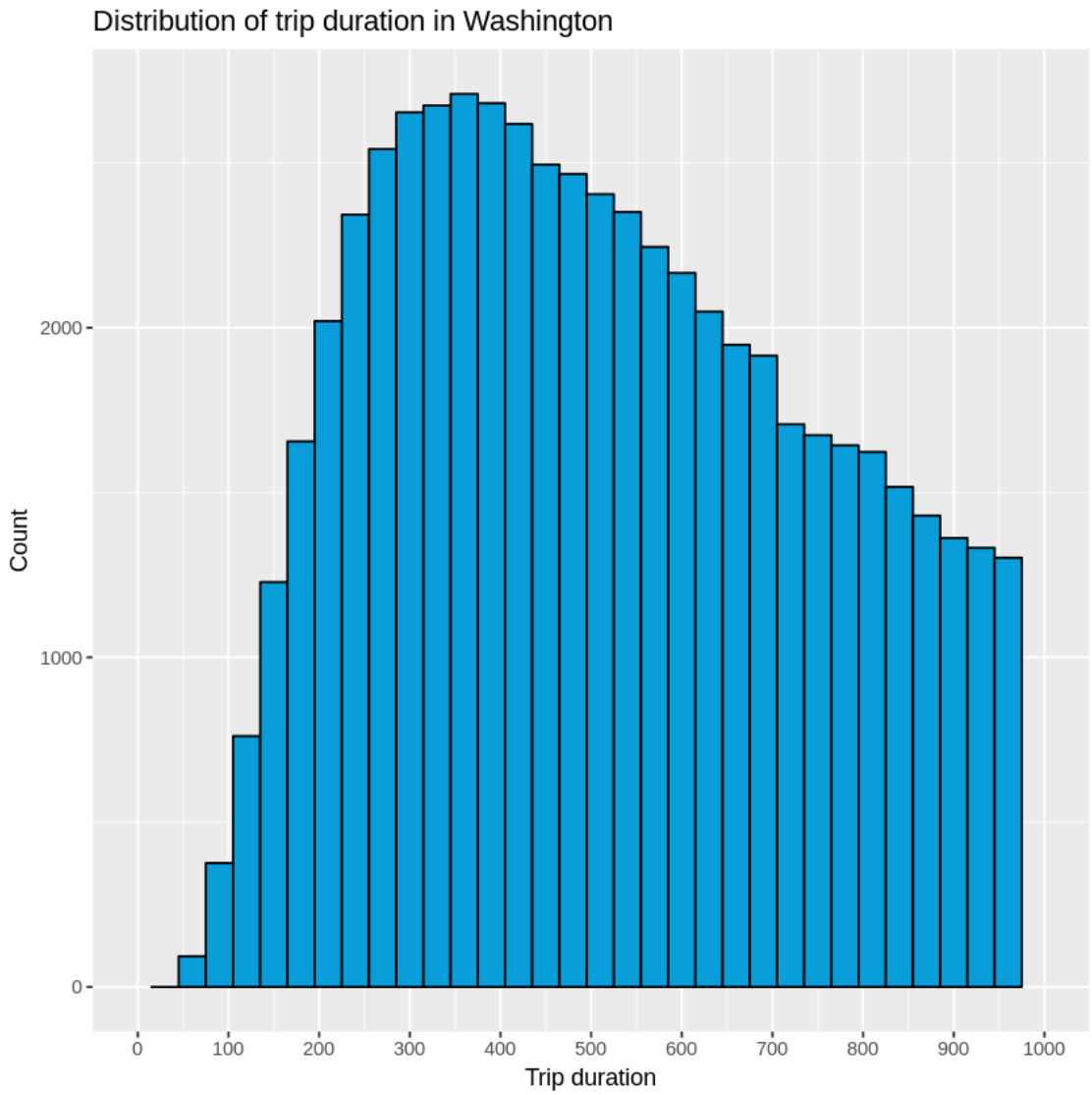
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   394.2   670.0   937.2  1119.0 85408.0


    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
    60.3    410.9    707.0   1234.0   1233.2 904591.4        1


     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
     61.0     368.0     610.0     903.6    1051.0 1088634.0         1


Warning message:
Removed 2597 rows containing non-finite values (stat_bin).Warning message:
Removed 2 rows containing missing values (geom_bar).Warning message:
Removed 27 rows containing non-finite values (stat_boxplot).
```



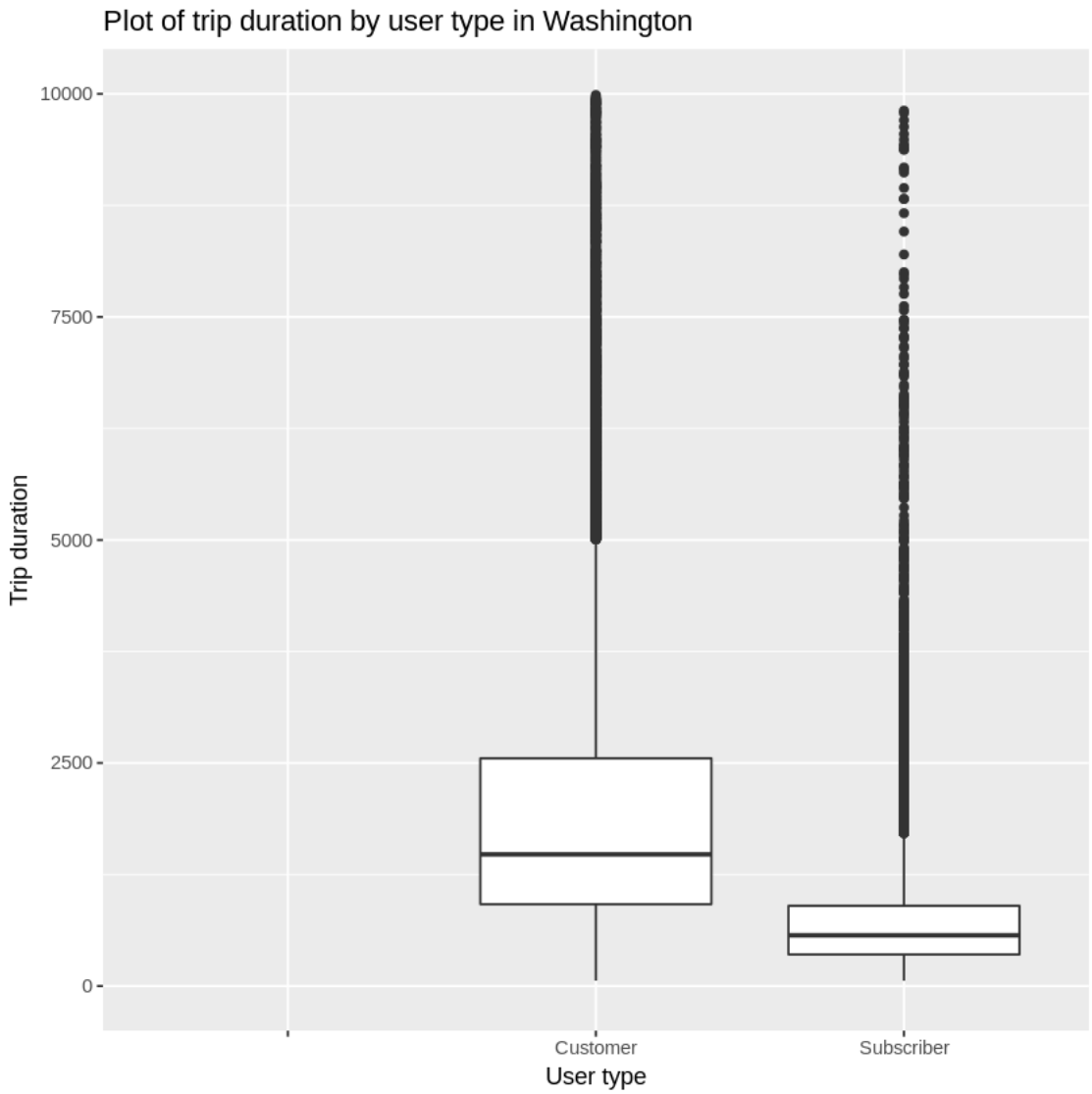

```
Warning message:
Removed 30046 rows containing non-finite values (stat_bin).Warning message:
Removed 2 rows containing missing values (geom_bar).
```



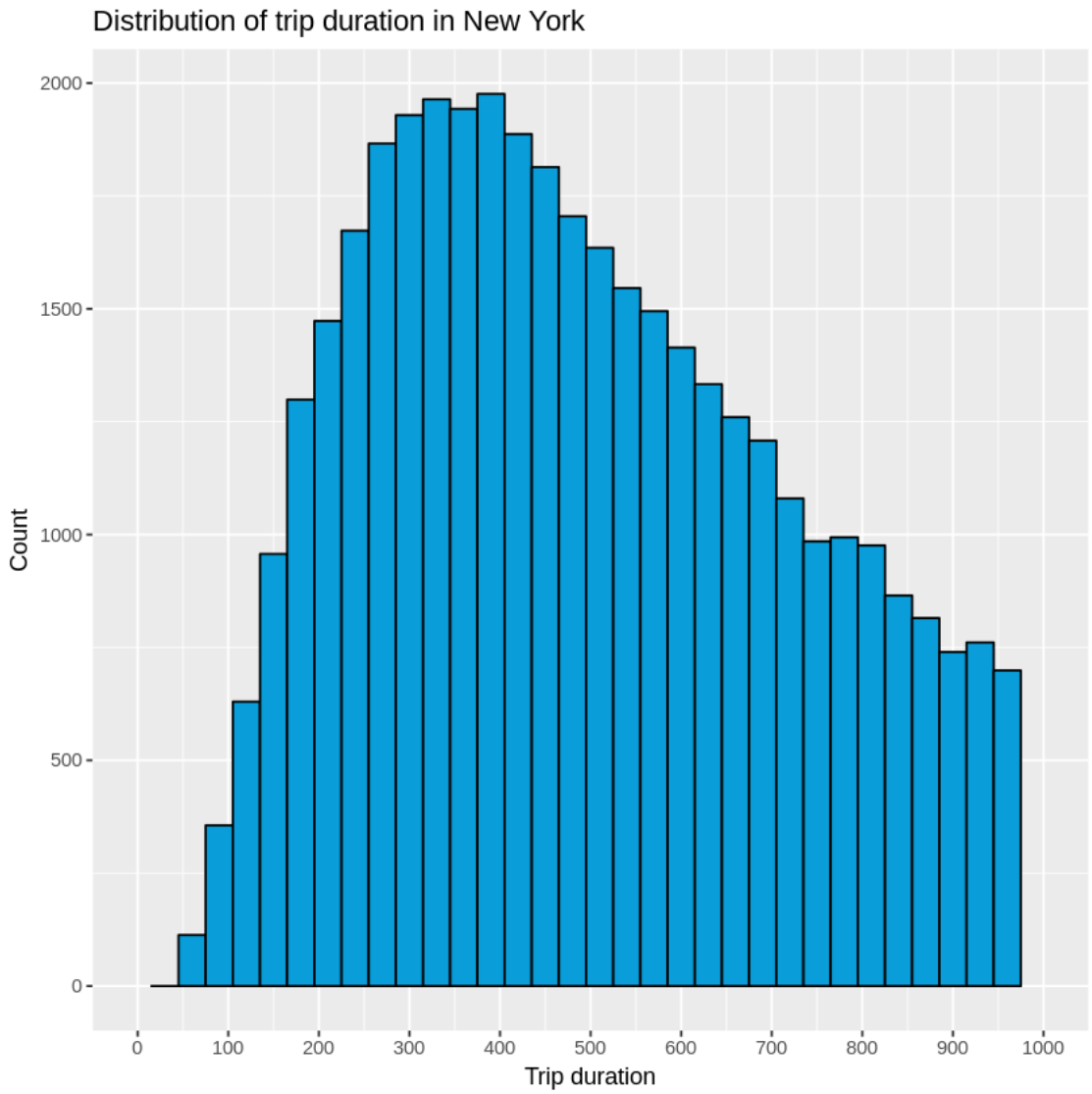

```
Warning message:
Removed 613 rows containing non-finite values (stat_boxplot).
```

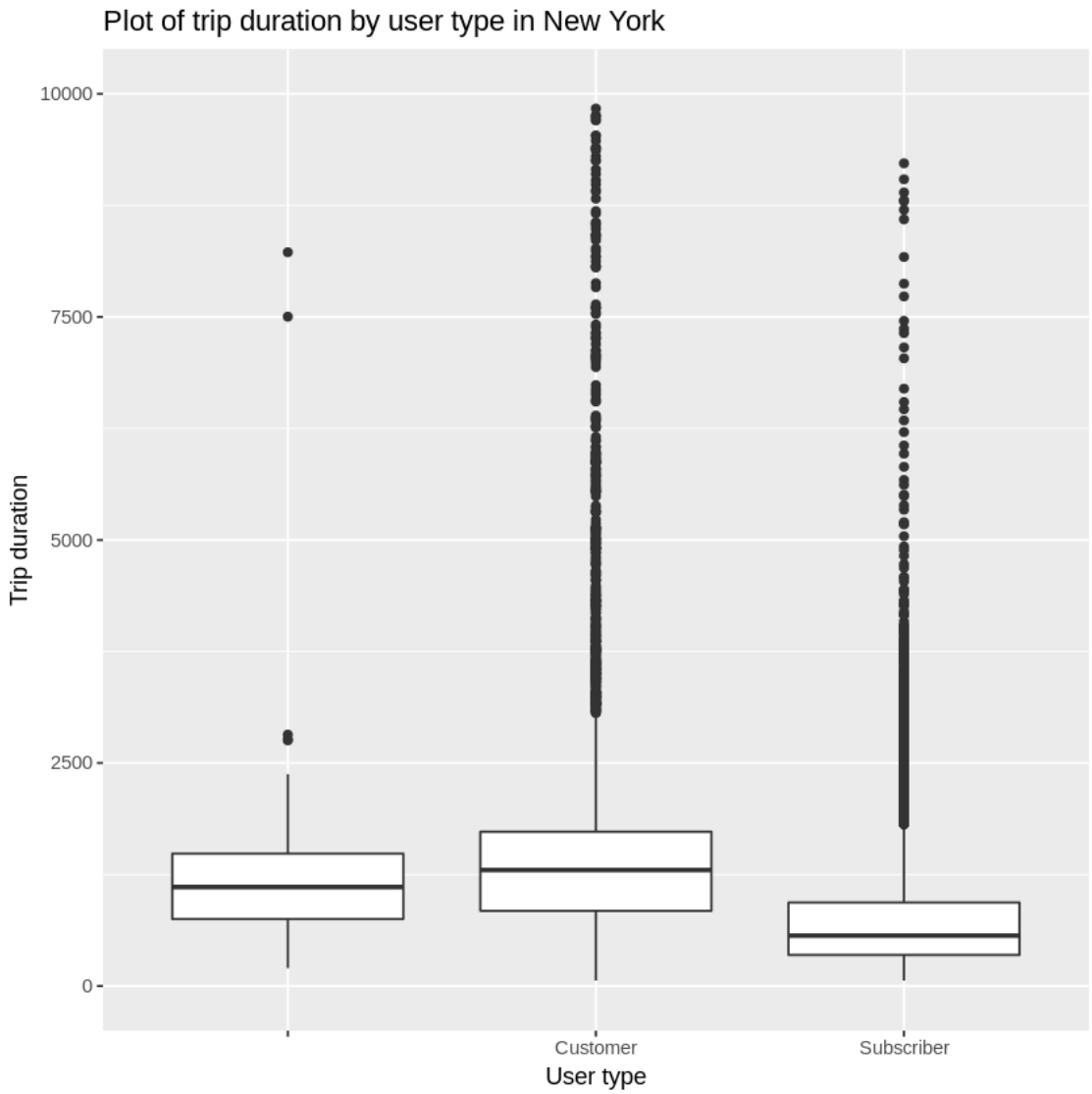



```
Warning message:
Removed 14808 rows containing non-finite values (stat_bin).Warning message:
Removed 2 rows containing missing values (geom_bar).
```




```
Warning message:
Removed 114 rows containing non-finite values (stat_boxplot).
```


Distribution of trip duration in New York
Count
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
900
1000
Trip duration



In the 3 cities the median trip duration is around 600s-700s, the middle 50% of all trips fall within the 368s-1233s range with a number of outliers falling outside the inter-quartile range.

0.1 Finishing Up

Congratulations! You have reached the end of the Explore Bikeshare Data Project. You should be very proud of all you have accomplished!

Tip: Once you are satisfied with your work here, check over your report to make sure that it is satisfies all the areas of the rubric.

0.2 Directions to Submit

Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly,

you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).

Alternatively, you can download this report as .html via the **File** > **Download as** submenu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.

Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

In [ ]:
```
system('python -m nbconvert Explore_bikeshare_data.ipynb')
```