

Explore_bikeshare_data

March 3, 2021

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 [1]: library(ggplot2)
library(dplyr)
ny = read.csv('new_york_city.csv')
wash = read.csv('washington.csv')
chi = read.csv('chicago.csv')
head(ny)
summary(ny)
names(ny)
```

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

X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Station
5688089	2017-06-11 14:55:05	2017-06-11 15:08:21	795	Suffolk St & Stanton St	W Broadw
4096714	2017-05-11 15:30:11	2017-05-11 15:41:43	692	Lexington Ave & E 63 St	1 Ave & E 7
2173887	2017-03-29 13:26:26	2017-03-29 13:48:31	1325	1 Pl & Clinton St	Henry St &
3945638	2017-05-08 19:47:18	2017-05-08 19:59:01	703	Barrow St & Hudson St	W 20 St & 8
6208972	2017-06-21 07:49:16	2017-06-21 07:54:46	329	1 Ave & E 44 St	E 53 St & 3
1285652	2017-02-22 18:55:24	2017-02-22 19:12:03	998	State St & Smith St	Bond St &

X	Start.Time	End.Time
Min. : 47	2017-05-11 18:26:10: 3	2017-01-03 08:54:10: 2
1st Qu.:1712425	2017-01-04 13:58:24: 2	2017-01-04 17:21:55: 2
Median :3418634	2017-01-09 09:36:01: 2	2017-01-05 17:25:17: 2
Mean :3415873	2017-01-21 15:36:56: 2	2017-01-12 08:34:01: 2
3rd Qu.:5123382	2017-01-21 17:49:59: 2	2017-01-12 09:41:54: 2
Max. :6816152	2017-01-21 20:08:29: 2	2017-01-12 20:34:42: 2
	(Other) :54757	(Other) :54758

Trip.Duration	Start.Station
Min. : 61.0	Pershing Square North: 592
1st Qu.: 368.0	W 21 St & 6 Ave : 385
Median : 610.0	Broadway & E 22 St : 383
Mean : 903.6	E 17 St & Broadway : 380
3rd Qu.: 1051.0	West St & Chambers St: 364
Max. :1088634.0	W 20 St & 11 Ave : 329
NA's :1	(Other) :52337

End.Station	User.Type	Gender	Birth.Year
Pershing Square North: 556	: 119	: 5410	Min. :1885
E 17 St & Broadway : 445	Customer : 5558	Female:12159	1st Qu.:1970
Broadway & E 22 St : 427	Subscriber:49093	Male :37201	Median :1981
W 21 St & 6 Ave : 365			Mean :1978
W 20 St & 11 Ave : 344			3rd Qu.:1988
W 38 St & 8 Ave : 338			Max. :2001
(Other) :52295			NA's :5218

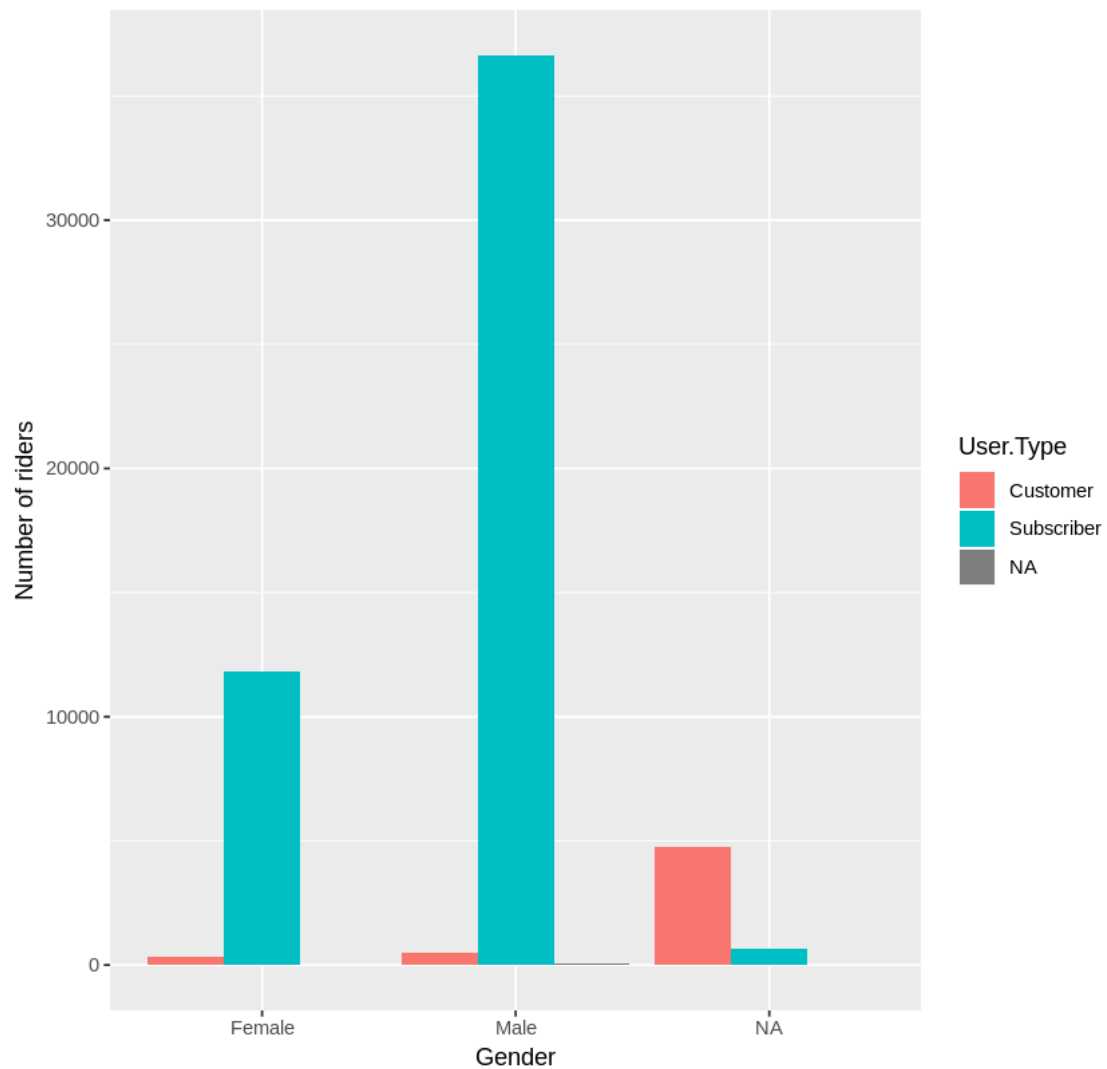
1. 'X' 2. 'Start.Time' 3. 'End.Time' 4. 'Trip.Duration' 5. 'Start.Station' 6. 'End.Station'
7. 'User.Type' 8. 'Gender' 9. 'Birth.Year'

0.0.2 Question 1

Does Gender influence the type of rider in New York?

```
In [3]: ggplot(ny, aes(x= Gender, fill = User.Type))+
  geom_bar(position = "dodge") +
  ylab('Number of riders')+
  ggtitle('Viz 1: Side by Side bar chart showing User type by Gender in NY')
ny$Gender[ny$Gender==""] <-NA
ny$User.Type[ny$User.Type == ""] <-NA
ny %>%
  filter(!is.na(Gender) & !is.na(User.Type)) %>%
  ggplot(aes(x = Gender, fill = User.Type))+
  geom_bar(position = "dodge") +
  ylab('Number of riders') +
  ggtitle("Viz 2: Distribution of User type by Gender in NY without missing values")
ny2 <-ny %>%
  filter(!is.na(Gender) & !is.na(User.Type))
table(ny2$Gender, ny2$User.Type)
chisq.test(ny2$Gender, ny2$User.Type, correct =FALSE)
```

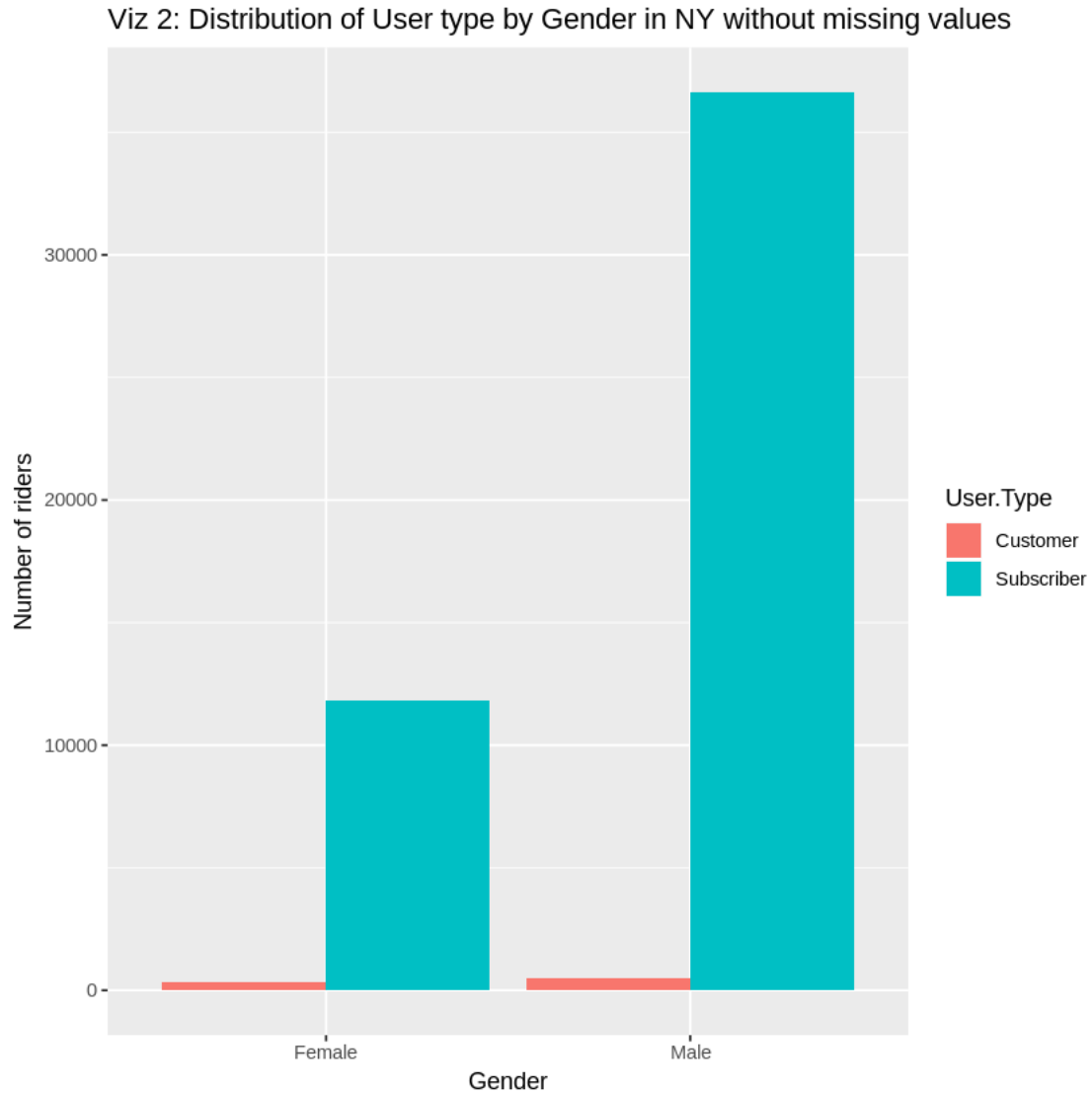
Viz 1: Side by Side bar chart showing User type by Gender in NY



	Customer	Subscriber
Female	324	11804
Male	491	36625

Pearson's Chi-squared test

data: ny2\$Gender and ny2\$User.Type
X-squared = 102.15, df = 1, p-value < 2.2e-16



by referring to first chart that show difference by gender between riders it seem that there is a big difference between male riders and female riders as male riders above 35,000 and females near to 10,000 so the answer is. maybe yes with a large percent it influence the type of rider in New York

0.0.3 Question 2

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

```
In [4]: duration_chi = mean(chi$Trip.Duration)
duration_ny = mean(ny$Trip.Duration)
duration_wash = mean(wash$Trip.Duration)
cat('The average travel time for Chicago is:', duration_chi, '\n')
cat('The average travel time for New York is:', duration_ny, '\n')
cat('The average travel time for Washington is:', duration_wash, '\n')
```

The average travel time for Chicago is: 937.1728
The average travel time for New York is: NA
The average travel time for Washington is: NA

```
In [100]: chart <- function(city) {  
  var1 <- ggplot(aes(x=Trip.Duration),data=city)+  
    ggtitle("The Bar Plot of Average Travel Time .")+  
    geom_histogram(binwidth = 100)+  
    scale_x_continuous(limits = c(0,5000))+  
    labs(x='Usages',y='Travel Time')+  
    geom_hline(aes(yintercept = mean(Trip.Duration)),col='red',size=1)  
  return(var1)  
}  
chart(ny)  
chart(chi)  
chart(wash)
```

Warning message:

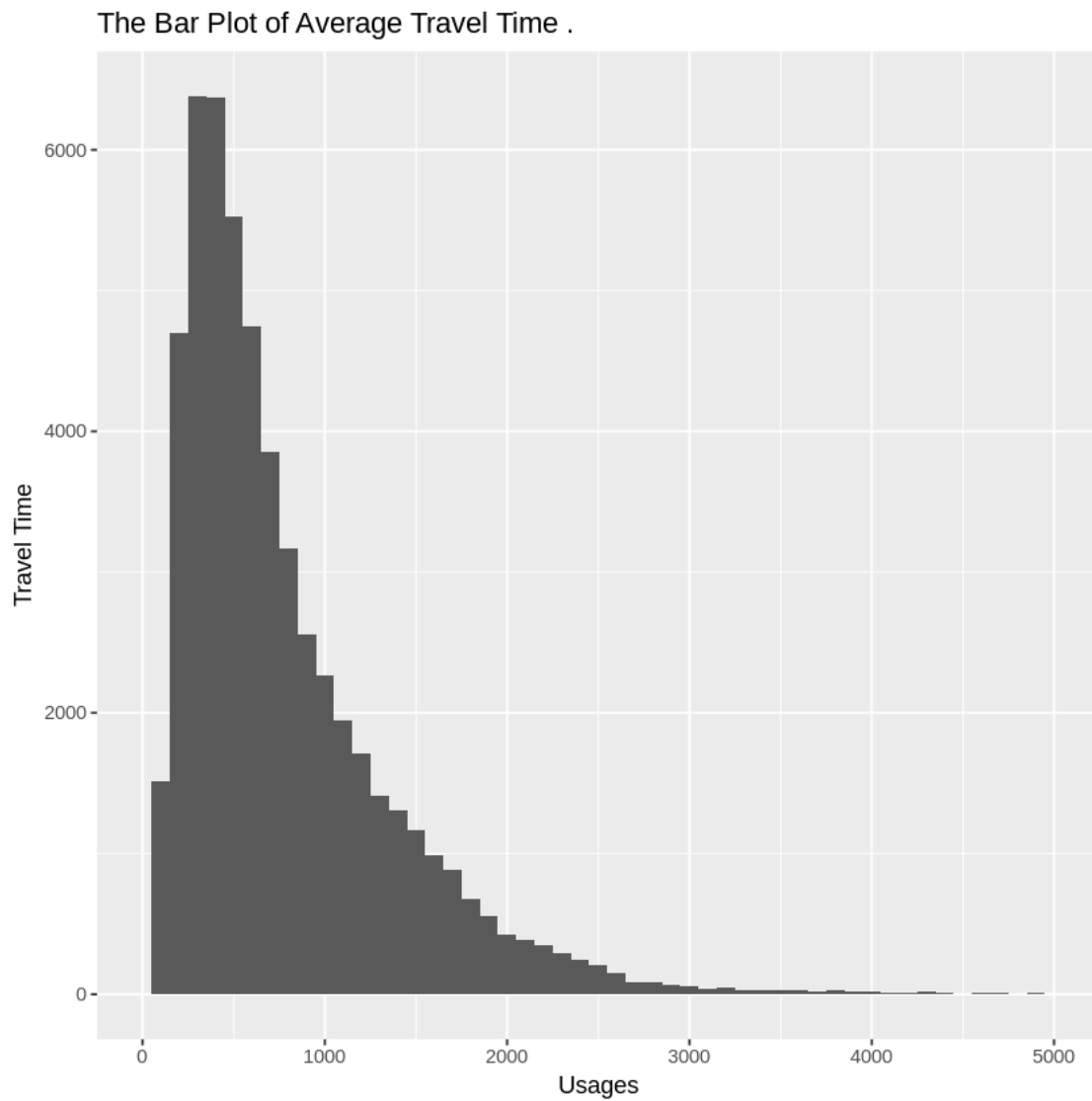
Removed 299 rows containing non-finite values (stat_bin).Warning message:

Removed 2 rows containing missing values (geom_bar).Warning message:

Removed 54770 rows containing missing values (geom_hline).Warning message:

Removed 93 rows containing non-finite values (stat_bin).Warning message:

Removed 2 rows containing missing values (geom_bar).



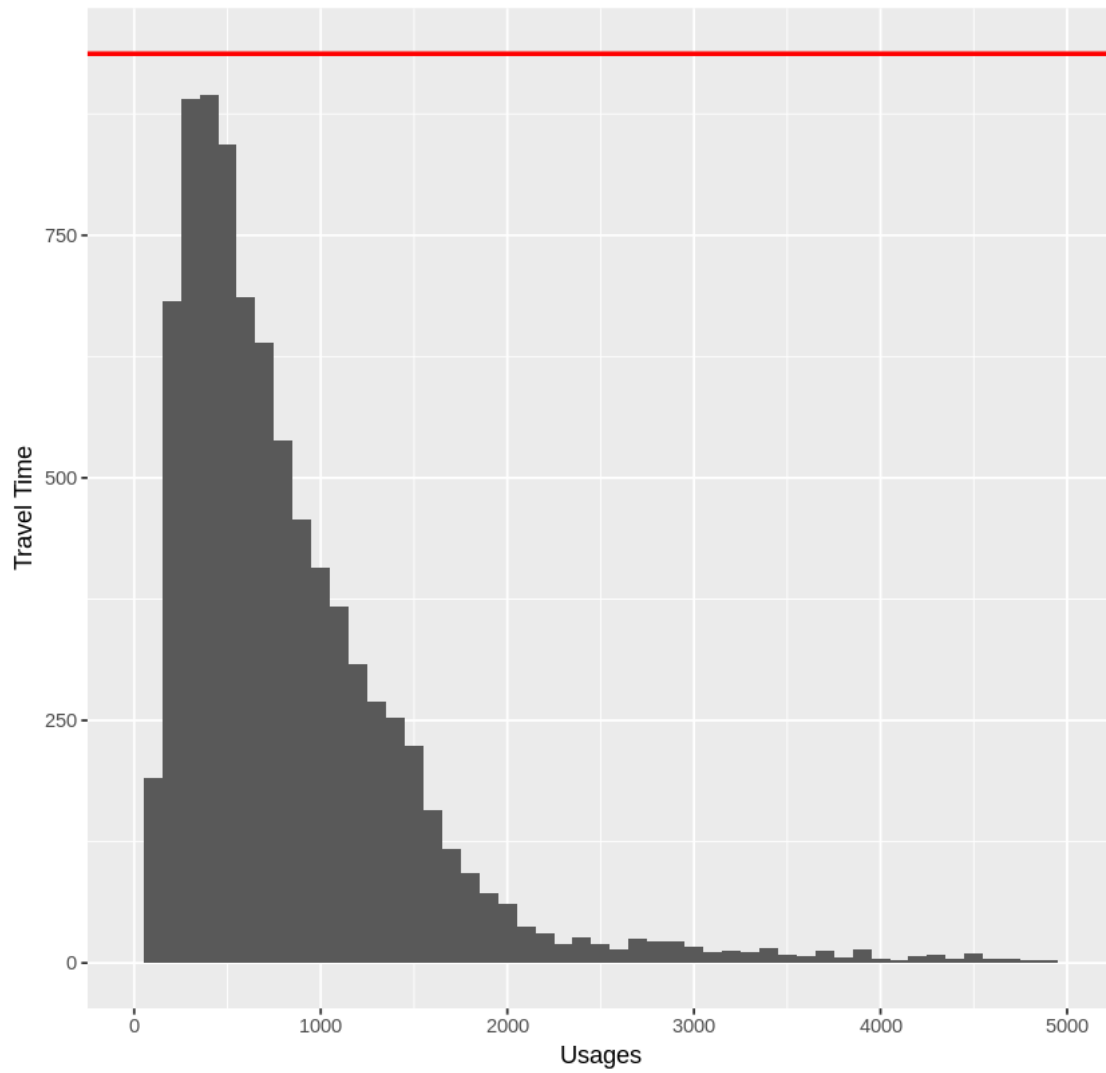
Warning message:

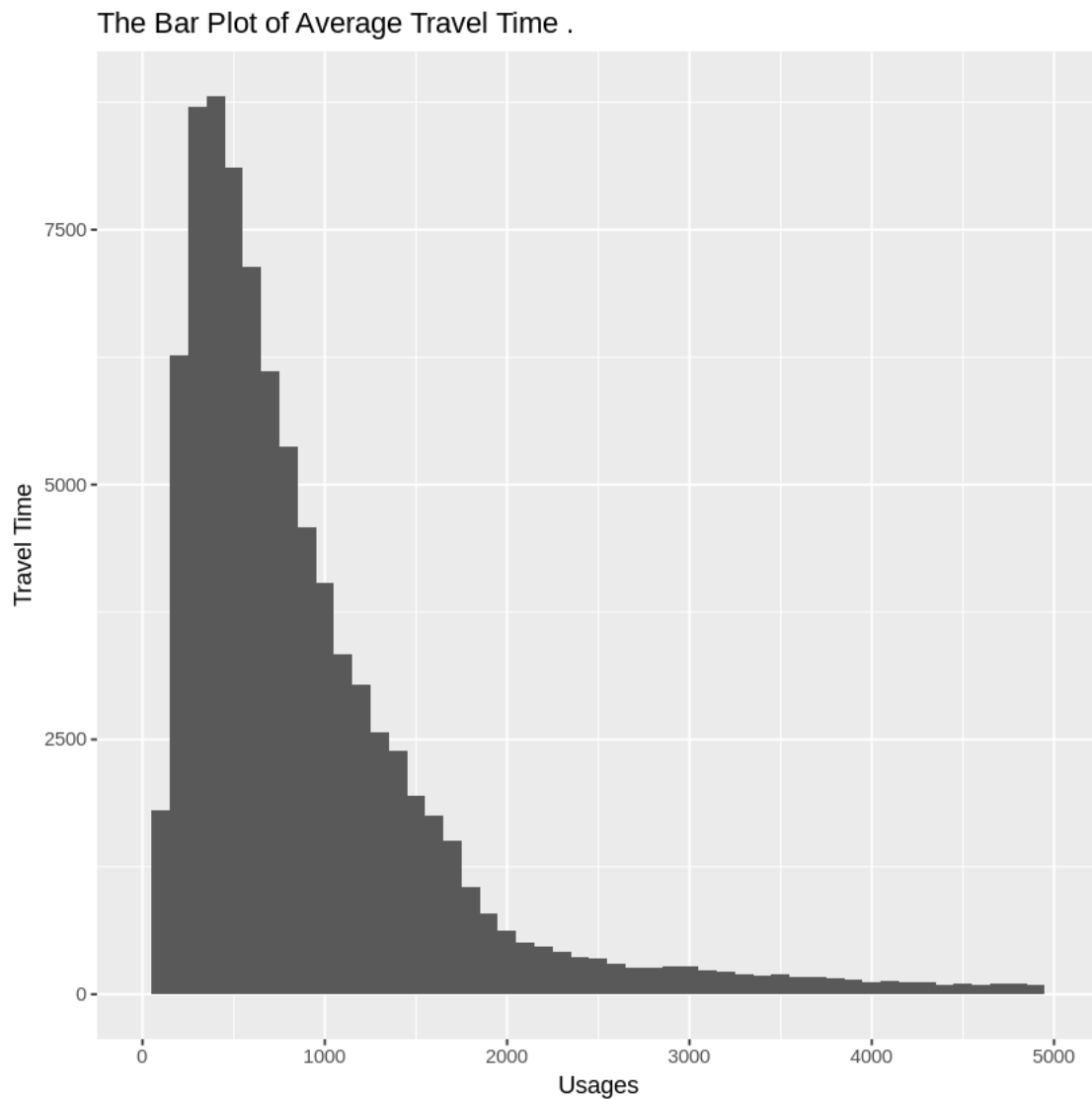
Removed 2830 rows containing non-finite values (stat_bin).Warning message:

Removed 2 rows containing missing values (geom_bar).Warning message:

Removed 89051 rows containing missing values (geom_hline).

The Bar Plot of Average Travel Time .





the charts shows that the top users avg travel time for the different cities goes to washington then to newyork and chicago

0.04 Question 3

Are male riders younger or older than female in New York in general?

```
In [101]: qplot(x = Gender, y = Birth.Year,
               data = subset(ny, !is.na(Gender)),
               geom = 'boxplot') +
           ggtitle('Viz 7:Initial Distribution of Birth Year of riders by Gender in NY')
```

```
qplot(x = Gender, y = Birth.Year,
```



```

data = subset(ny, !is.na(Gender)),
geom = 'boxplot') +
coord_cartesian(ylim = c(1962,1990))+
ggtitle('Viz 8:Distribution of Birth Year of riders by Gender in NY (Birth year betw

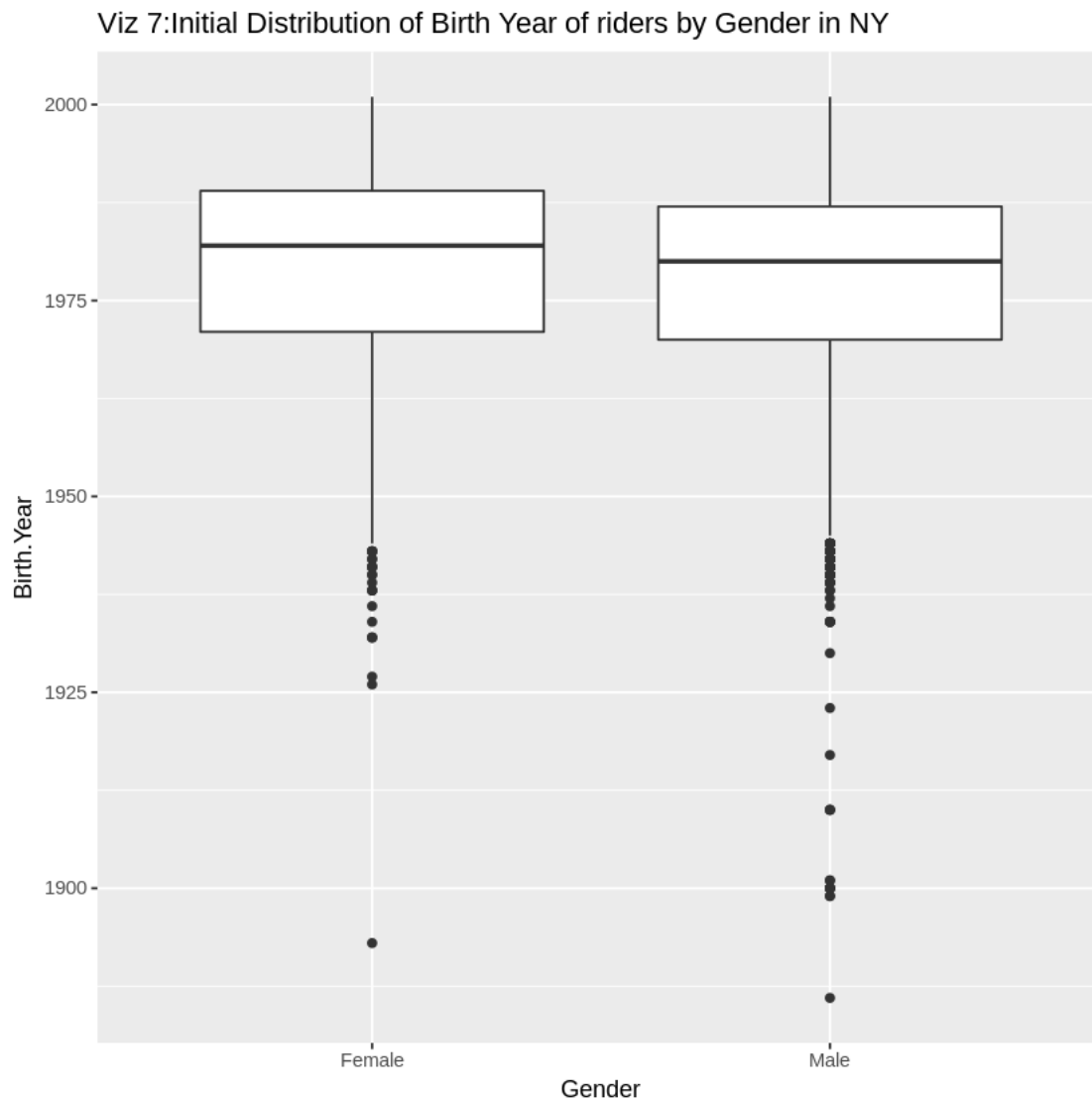
by(ny$Birth.Year, ny$Gender, summary)

```

Warning message:

Removed 1 rows containing non-finite values (stat_boxplot).Warning message:

Removed 1 rows containing non-finite values (stat_boxplot).



ny\$Gender:

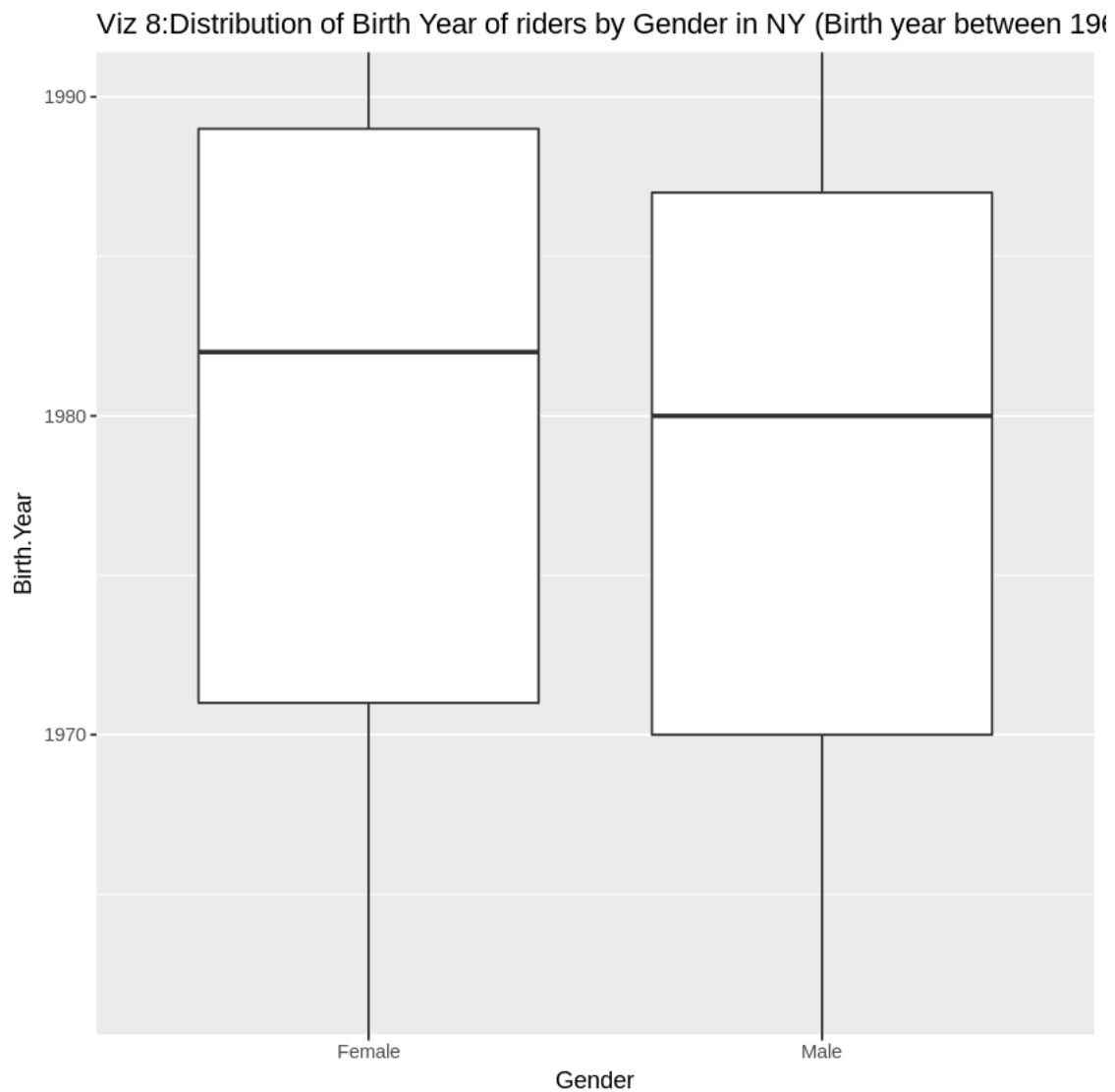
NULL

ny\$Gender: Female

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1893	1971	1982	1979	1989	2001	1

ny\$Gender: Male

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1886	1970	1980	1978	1987	2001



The charts explain that males riders are older than females rider as the median no of female riders is 1982 and for male 1980 and the minnum male riders age are 1886 and females are 1893

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 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** sub-menu, 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')
```