Explore_bikeshare_data

January 25, 2022

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 [9]: ny <- read.csv('new_york_city.csv')
        wash <- read.csv('washington.csv')
        chi <-  read.csv('chicago.csv')
        us <- rbind(ny,chi)
        library(readr)
        library(tidyr)
        library(plyr)
        library(ggplot2)
        library(dplyr)
        library(lubridate)
        library(tidyverse)
```

```
In [11]: head(ny)
```

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 &

```
In [12]: 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 [13]: 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

```
In [14]: ###### We are looking for missing values
```

```
In [16]: colSums(is.na(ny))
         colSums(is.na(wash))
         colSums(is.na(chi))
         colSums(is.na(us))
```

X 0 **Start.Time** 0 **End.Time** 0 **Trip.Duration** 1 **Start.Station** 0 **End.Station** 0 **User.Type** 0
Gender 0 **Birth.Year** 5218
X 0 **Start.Time** 0 **End.Time** 0 **Trip.Duration** 1 **Start.Station** 0 **End.Station** 0 **User.Type** 0
X 0 **Start.Time** 0 **End.Time** 0 **Trip.Duration** 0 **Start.Station** 0 **End.Station** 0 **User.Type** 0
Gender 0 **Birth.Year** 1747
X 0 **Start.Time** 0 **End.Time** 0 **Trip.Duration** 1 **Start.Station** 0 **End.Station** 0 **User.Type** 0
Gender 0 **Birth.Year** 6965

```
In [40]: ny1 <- drop_na(ny)
         wash1 <- drop_na(wash)
         chi1 <- drop_na(chi)
         us1 <- drop_na(us)
```

```
In [19]: colSums(is.na(ny1))
         colSums(is.na(wash1))
         colSums(is.na(chi1))
         colSums(is.na(us1))
```

X 0 **Start.Time** 0 **End.Time** 0 **Trip.Duration** 0 **Start.Station** 0 **End.Station** 0 **User.Type** 0
Gender 0 **Birth.Year** 0
X 0 **Start.Time** 0 **End.Time** 0 **Trip.Duration** 0 **Start.Station** 0 **End.Station** 0 **User.Type** 0
X 0 **Start.Time** 0 **End.Time** 0 **Trip.Duration** 0 **Start.Station** 0 **End.Station** 0 **User.Type** 0
Gender 0 **Birth.Year** 0
X 0 **Start.Time** 0 **End.Time** 0 **Trip.Duration** 0 **Start.Station** 0 **End.Station** 0 **User.Type** 0
Gender 0 **Birth.Year** 0

0.0.2 Question 1

Comparison of the mileage rate by gender in New York City

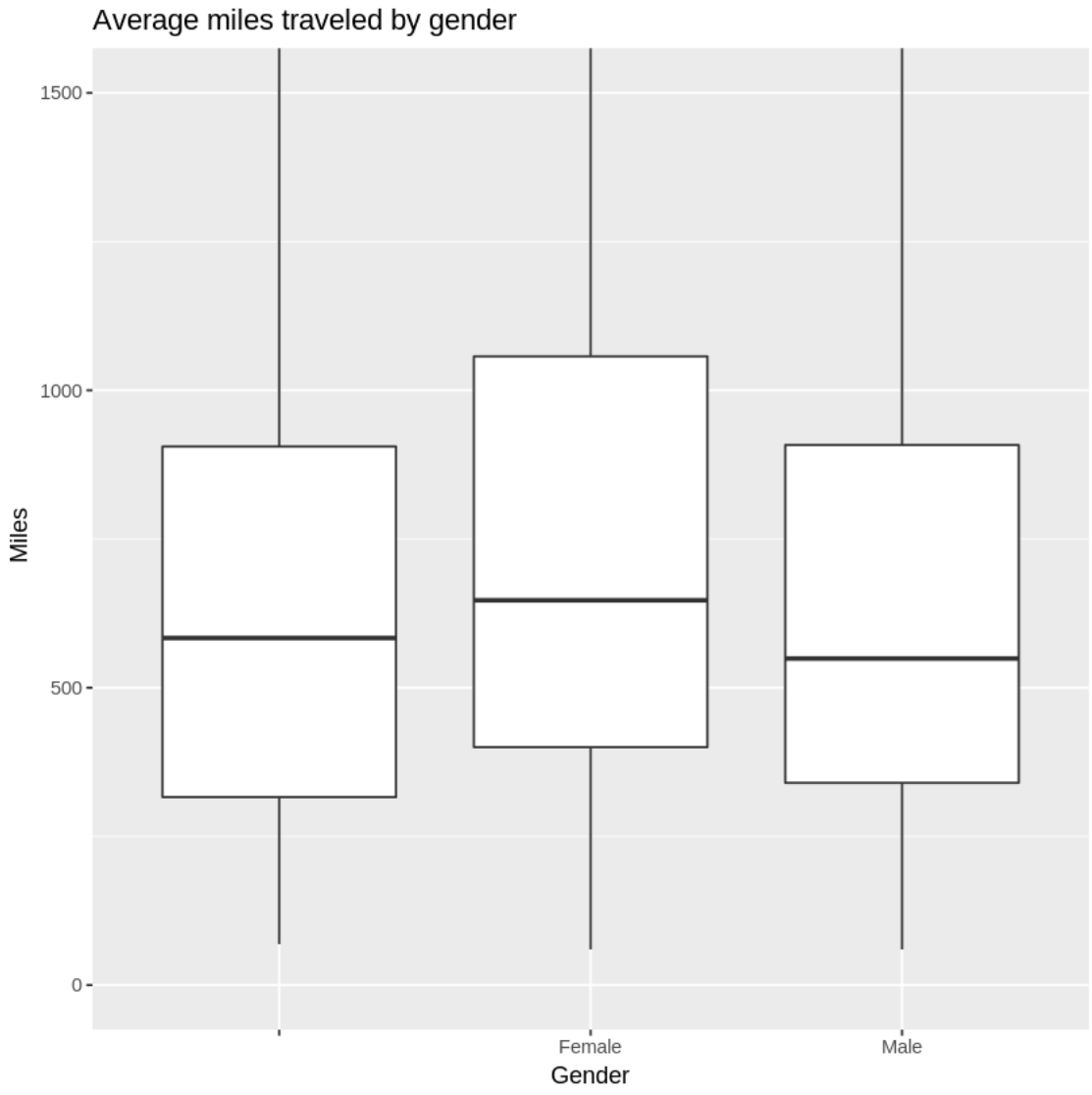
```
In [20]: summary(us1$Trip.Duration[ny1$Gender=="Male"])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     60     341     552     759     915 1088634
```

In [21]:
```
summary(us1$Trip.Duration[ny1$Gender=="Female"])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   395.0   635.0   851.5  1042.0 77918.0
```

In [22]:
```
qplot(x = Gender  , y = Trip.Duration , data = subset(us1,!is.na(Gender)), geom = 'boxp
              main = "Average miles traveled by gender" , xlab = "Gender" , ylab = " Miles") +
```



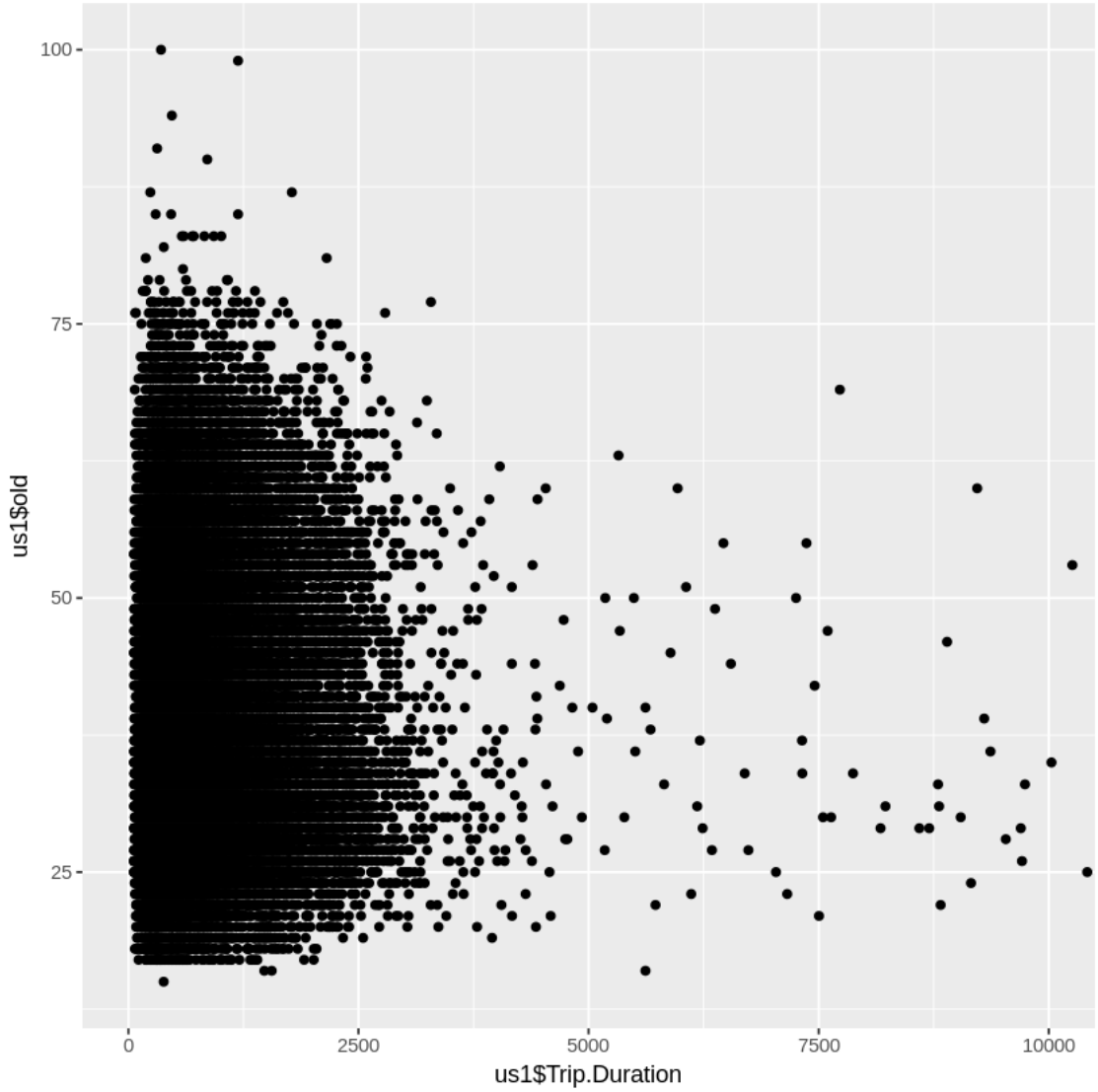
Comparing the average miles traveled, we find that women travel on average more miles than men in New York City

0.0.3 Question 2

Do younger customers drive more miles than older customers

```
In [23]: ny1["old"] <- 2017-ny1$Birth.Year
         chi1["old"] <- 2017-chi1$Birth.Year
         us1["old"] <- 2017-us1$Birth.Year
```

```
In [24]: ggplot(data= us1, aes(x=us1$Trip.Duration, y=us1$old)) + geom_point(aes())  + coord_car
```



We find that the older a person gets, the fewer miles traveled, based on the data of bike renters in New York and Washington.

0.0.4 Question 3

What is the average age for renting bikes in washenton and New York City ?

```
In [32]: us1<- drop_na(us)
         us1 <- rbind(ny1,chi1)
         str(us1)
```

```
'data.frame':        56435 obs. of  11 variables:
 $ X            : int  5688089 4096714 2173887 3945638 6208972 1285652 1675753 1692245 1558339 2
 $ Start.Time   : Factor w/ 63133 levels "2017-01-01 00:17:01",..: 45448 32799 17316 31589 49688
 $ End.Time     : Factor w/ 63130 levels "201","2017-01-01 00:30:56",..: 45432 32783 17295 31567
 $ Trip.Duration: int  795 692 1325 703 329 998 478 4038 309 319 ...
 $ Start.Station: Factor w/ 1108 levels "","1 Ave & E 16 St",..: 522 406 10 93 5 521 325 309 245
 $ End.Station  : Factor w/ 1108 levels "","1 Ave & E 16 St",..: 613 8 362 558 269 107 389 110 2
 $ User.Type    : Factor w/ 3 levels "","Customer",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Gender       : Factor w/ 3 levels "","Female","Male": 3 3 3 2 3 3 3 3 3 3 ...
 $ Birth.Year   : num  1998 1981 1987 1986 1992 ...
 $ old          : num  19 36 30 31 25 31 35 33 25 62 ...
 $ old.group    : Factor w/ 5 levels "15-24","25-34",..: 1 3 2 2 1 2 2 2 1 5 ...
```

```
In [33]: us1["old"] <- 2017-us1$Birth.Year
         us1["old.group"] <- cut(us1$old,c(15,25,35,45,55,100),c("15-24","25-34","35-44","45-54"
```

```
In [34]: str(us1)
```

```
'data.frame':        56435 obs. of  11 variables:
 $ X            : int  5688089 4096714 2173887 3945638 6208972 1285652 1675753 1692245 1558339 2
 $ Start.Time   : Factor w/ 63133 levels "2017-01-01 00:17:01",..: 45448 32799 17316 31589 49688
 $ End.Time     : Factor w/ 63130 levels "201","2017-01-01 00:30:56",..: 45432 32783 17295 31567
 $ Trip.Duration: int  795 692 1325 703 329 998 478 4038 309 319 ...
 $ Start.Station: Factor w/ 1108 levels "","1 Ave & E 16 St",..: 522 406 10 93 5 521 325 309 245
 $ End.Station  : Factor w/ 1108 levels "","1 Ave & E 16 St",..: 613 8 362 558 269 107 389 110 2
 $ User.Type    : Factor w/ 3 levels "","Customer",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Gender       : Factor w/ 3 levels "","Female","Male": 3 3 3 2 3 3 3 3 3 3 ...
 $ Birth.Year   : num  1998 1981 1987 1986 1992 ...
 $ old          : num  19 36 30 31 25 31 35 33 25 62 ...
 $ old.group    : Factor w/ 5 levels "15-24","25-34",..: 1 3 2 2 1 2 2 2 1 5 ...
```

```
In [35]: head(us1)
```

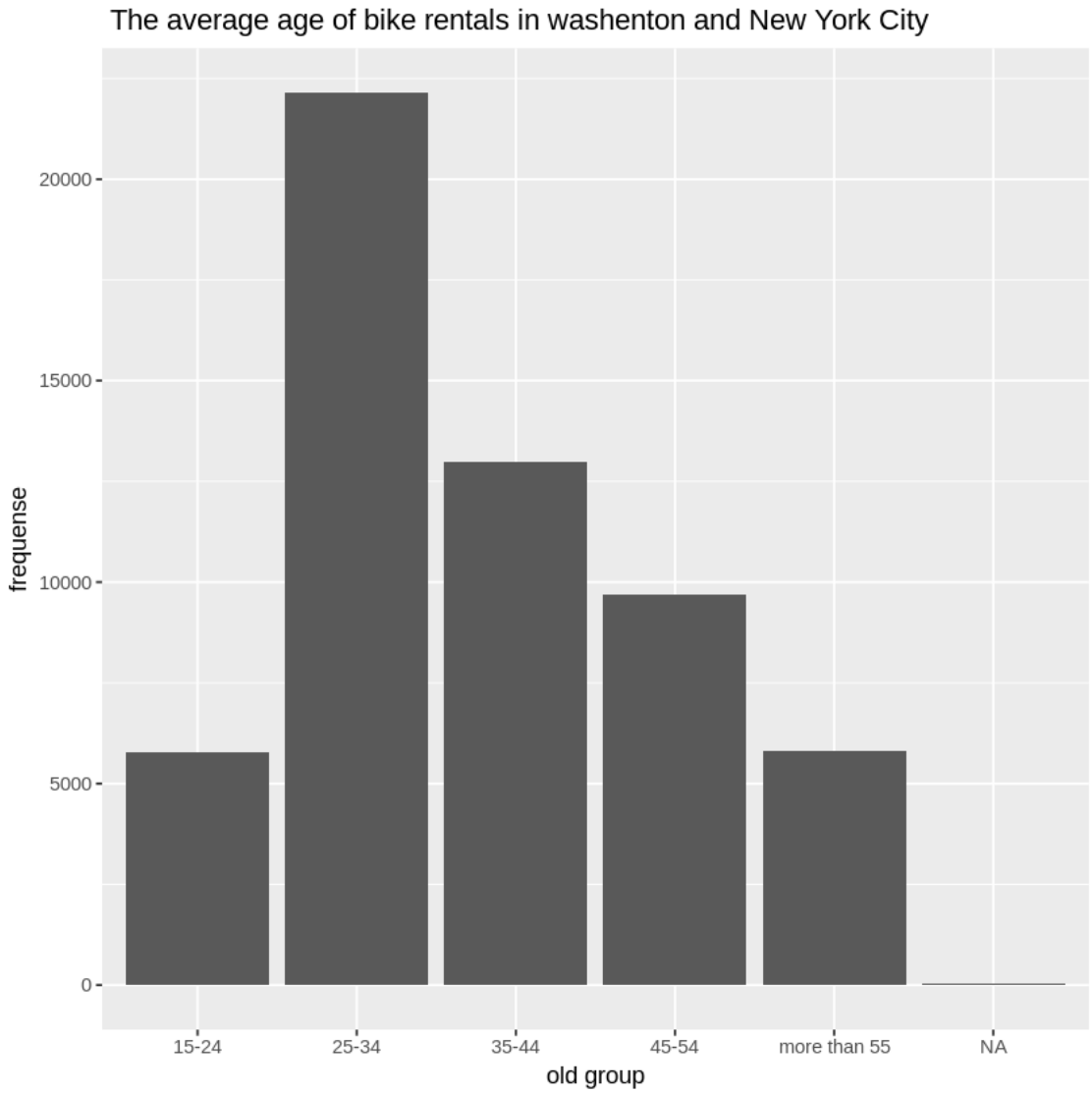
X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Station
5688089	2017-06-11 14:55:05	2017-06-11 15:08:21	795	Suffolk St & Stanton St	W Broadwa
4096714	2017-05-11 15:30:11	2017-05-11 15:41:43	692	Lexington Ave & E 63 St	1 Ave & E 7
2173887	2017-03-29 13:26:26	2017-03-29 13:48:31	1325	1 Pl & Clinton 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 &

```
In [36]: summary(us1$old.group)
```

15-24 5781 **25-34** 22141 **35-44** 12992 **45-54** 9675 **more than 55** 5821 **NA's** 25

```
In [38]: qplot( x = old.group, data = us1 , gemo = "line", main = " The average age of bike rent
```

```
Warning message:
Ignoring unknown parameters: gemo
```



We find that the highest average age in the number of bicycle rental is the 25-34 age group, and the reason for this may be the availability of sufficient time to make such trips

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