Explore_bikeshare_data

April 1, 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 [87]: library(ggplot2)
```

```
In [88]: ny = read.csv('new_york_city.csv')
         wash = read.csv('washington.csv')
         chi = read.csv('chicago.csv')
```

```
In [89]: 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 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 [90]: 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	Ca

```
In [91]: 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

0.1 Clean and Organize The Data

0.2 Add CITY Column and city information to each data set.

```
In [92]: ny$City <- c('NY')
         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 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 [93]: wash$City <- c('Wash')
         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 [94]: chi$City <- c('Chi')
         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	Cai

0.3 Add Gender and Birth.Year columns to Wash Dataset

```
In [95]: wash$Gender <- c(NA)
         wash$Birth.Year <- c(NA)
         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

0.4 Let's merge the datasets together

```
In [96]: bsd <- rbind(ny, wash, chi)
```

```
In [97]: # replace any blank spaces with NA
         bsd[bsd == ""] <- NA
```

```
In [98]: head(bsd, 100000)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station
5688089	2017-06-11 14:55:05	2017-06-11 15:08:21	795	Suffolk St & Stanton St
4096714	2017-05-11 15:30:11	2017-05-11 15:41:43	692	Lexington Ave & E 63 St
2173887	2017-03-29 13:26:26	2017-03-29 13:48:31	1325	1 Pl & Clinton St
3945638	2017-05-08 19:47:18	2017-05-08 19:59:01	703	Barrow St & Hudson St
6208972	2017-06-21 07:49:16	2017-06-21 07:54:46	329	1 Ave & E 44 St
1285652	2017-02-22 18:55:24	2017-02-22 19:12:03	998	State St & Smith St
1675753	2017-03-06 16:22:53	2017-03-06 16:30:51	478	Front St & Gold St
1692245	2017-03-07 07:42:24	2017-03-07 08:49:42	4038	E 89 St & York Ave
2271331	2017-04-02 08:02:36	2017-04-02 09:28:08	5132	Central Park S & 6 Ave
1558339	2017-03-01 23:01:31	2017-03-01 23:06:41	309	E 3 St & 1 Ave
2287178	2017-04-02 14:37:20	2017-04-02 14:56:12	1131	Bank St & Washington St
2744874	2017-04-13 13:40:39	2017-04-13 13:45:59	319	Front St & Maiden Ln
3398180	2017-04-27 23:27:31	2017-04-28 00:05:53	2301	E 10 St & 5 Ave
991609	2017-02-13 15:40:53	2017-02-13 16:00:26	1172	1 Ave & E 68 St
1512596	2017-02-28 19:26:43	2017-02-28 19:35:21	518	N 11 St & Wythe Ave
187466	2017-01-11 11:30:30	2017-01-11 11:35:15	285	E 17 St & Broadway
2195658	2017-03-29 20:19:44	2017-03-29 20:24:07	263	State St & Smith St
6388534	2017-06-23 21:21:59	2017-06-23 21:30:45	525	E 2 St & Avenue C
4733837	2017-05-24 08:53:32	2017-05-24 09:04:30	658	Central Park West & W 76 St
5857	2017-01-01 13:32:39	2017-01-01 13:49:57	1038	W 22 St & 8 Ave
1132766	2017-02-18 13:29:08	2017-02-18 13:30:31	82	E 71 St & 1 Ave
3358474	2017-04-27 09:44:35	2017-04-27 09:48:00	204	University Pl & E 14 St
1778858	2017-03-09 11:15:39	2017-03-09 11:29:03	803	E 25 St & 2 Ave
2497952	2017-04-08 13:39:48	2017-04-08 14:04:24	1476	Dean St & Hoyt St
2905932	2017-04-16 17:36:06	2017-04-16 18:02:52	1605	Allen St & Stanton St
3123311	2017-04-21 09:41:14	2017-04-21 09:48:36	441	Lexington Ave & E 63 St
2959550	2017-04-17 18:27:23	2017-04-17 18:56:33	1750	NYCBS Depot - SSP
2067887	2017-03-25 12:02:11	2017-03-25 12:08:44	393	W 26 St & 8 Ave
3518426	2017-04-29 23:58:44	2017-04-30 00:02:19	215	Great Jones St
5383277	2017-06-06 11:23:30	2017-06-06 11:26:56	205	W 43 St & 10 Ave
1057950	2017-05-04 08:13:09	2017-05-04 08:19:13	364.314	Convention Center / 7th & M St NW
1383239	2017-06-03 03:23:05	2017-06-03 03:38:57	952.490	Jefferson Dr & 14th St SW
641304	2017-03-30 19:13:00	2017-03-30 19:21:00	438.568	Park Rd & Holmead Pl NW
596371	2017-03-26 17:19:00	2017-03-26 17:58:00	2349.762	Iwo Jima Memorial/N Meade & 14th
500280	2017-03-15 09:11:00	2017-03-15 09:16:00	273.657	3rd & H St NW
348469	2017-02-23 09:03:00	2017-02-23 09:29:00	1540.182	1st & K St SE
684766	2017-04-04 05:13:27	2017-04-04 05:38:43	1515.387	Jefferson Dr & 14th St SW
1356630	2017-06-01 05:39:02	2017-06-01 05:46:57	474.116	17th & G St NW
1206321	2017-05-19 02:40:05	2017-05-19 03:06:59	1614.570	1st & H St NW
1220990	2017-05-20 02:52:34	2017-05-20 03:25:45	1990.920	Jefferson Memorial
497757	2017-03-13 22:34:00	2017-03-13 22:49:00	885.052	8th & D St NW
177776	2017-02-01 11:14:00	2017-02-01 11:21:00	426.896	15th & K St NW
444814	2017-03-07 07:49:00	2017-03-07 07:51:00	162.470	13th St & Eastern Ave
1269132	2017-05-24 07:07:27	2017-05-24 07:12:48	321.456	N Quincy St & Glebe Rd
526329	2017-03-19 15:17:00	2017-03-19 16:31:00	4474.716	Jefferson Dr & 14th St SW
768506	2017-04-11 06:00:08	2017-04-11 06:05:30	321.834	15th & L St NW
908107	2017-04-21 06:38:17	2017-04-21 06:47:31	553.851	22nd & I St NW / Foggy Bottom
1677607	2017-06-25 11:39:42	2017-06-25 11:46:12	389.562	8th & O St NW
1177653	2017-05-16 09:36:18	2017-05-16 09:52:15	956.722	M St & New Jersey Ave SE
1736751	2017-06-29 11:36:27	2017-06-29 11:56:08	1181.215	Hains Point/Buckeye & Ohio Dr SW
885206	2017-04-19 08:00:28	2017-04-19 08:14:46	857.324	11th & Kenyon St NW

Let's explore our dataset with some built-it functions.

```
In [99]: names(bsd)
```

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' 10. 'City'

```
In [100]: dim(bsd)
```

1. 152451 2. 10

```
In [101]: summary(bsd)
```

```
       X                          Start.Time                    End.Time
 Min.   :      7   2017-02-19 12:19:00:     6   2017-03-09 17:54:00:     7
 1st Qu.: 589310   2017-02-20 11:35:00:     6   2017-03-28 18:11:00:     7
 Median :1184899   2017-02-24 17:46:00:     6   2017-01-13 17:48:00:     6
 Mean   :1781625   2017-03-01 08:20:00:     6   2017-01-31 08:49:00:     6
 3rd Qu.:2085970   2017-03-02 08:39:00:     6   2017-02-13 18:09:00:     6
 Max.   :6816152   (Other)            :152420   (Other)            :152418
                   NA's               :     1   NA's               :     1
 Trip.Duration                                  Start.Station
 Min.   :     60   Columbus Circle / Union Station     :  1700
 1st Qu.:    392   Lincoln Memorial                    :  1546
 Median :    667   Jefferson Dr & 14th St SW           :  1488
 Mean   :   1098   Massachusetts Ave & Dupont Circle NW:  1219
 3rd Qu.:   1159   Jefferson Memorial                  :  1068
 Max.   :1088634   (Other)                             :145428
 NA's   :2         NA's                                :     2
                                                End.Station
 Columbus Circle / Union Station                      :  1767
 Jefferson Dr & 14th St SW                            :  1603
 Lincoln Memorial                                     :  1514
 Massachusetts Ave & Dupont Circle NW                 :  1344
 Smithsonian-National Mall / Jefferson Dr & 12th St SW:  1103
 (Other)                                              :145117
 NA's                                                 :     3
     User.Type         Gender        Birth.Year        City
          :     0          :    0   Min.   :1885    Length:152451
 Customer  : 30754   Female:13882   1st Qu.:1970    Class :character
 Subscriber:121576   Male  :42360   Median :1981    Mode  :character
 NA's      :   121   NA's  :96209   Mean   :1979
                                    3rd Qu.:1988
                                    Max.   :2002
                                    NA's   :96016
```

0.4.1 Q1: Which gender has the longest duration time?

I'll add the following code so that the graphs do not display numeric values in scietific notation.

```
In [102]: options(scipen=999)
```

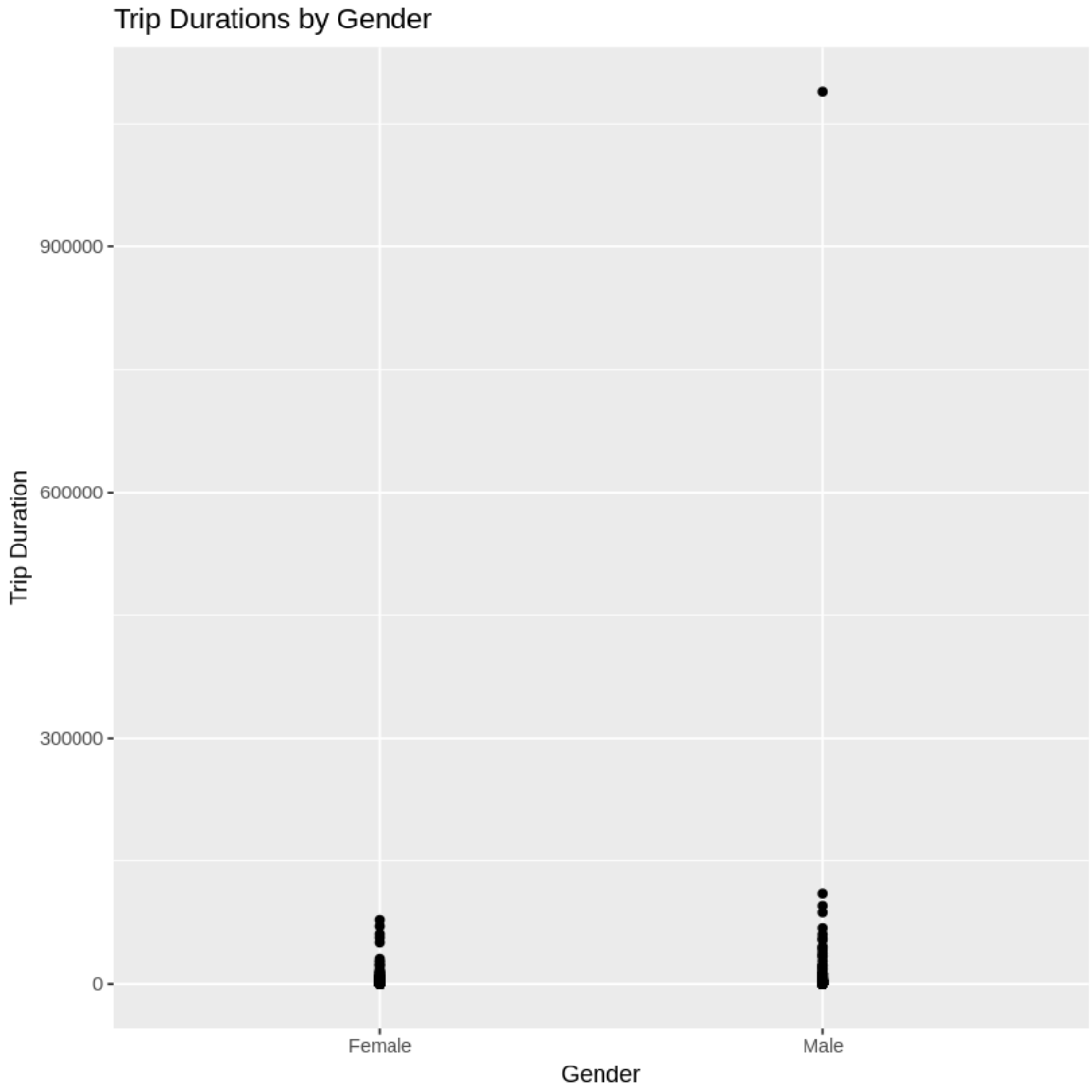
I will omit the NA rows where gender information is missing and save it to a new dataset.

```
In [103]: bsd_omit <- na.omit(bsd)
          head(bsd_omit)
```

X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Statio
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 &

I'll plot the duration data, separated by gender.

```
In [105]: ggplot(bsd_omit, aes(x=Gender, y=Trip.Duration)) +
              geom_point() +
              xlab("Gender") +
              ylab("Trip Duration") +
              ggtitle("Trip Durations by Gender")
```



It looks like we have a major outlier on the male side of the data. I will update the dataset to drop the outlier.

In [106]:
```
#find the row index of the outlier
which(bsd_omit$Trip.Duration > 900000)
```

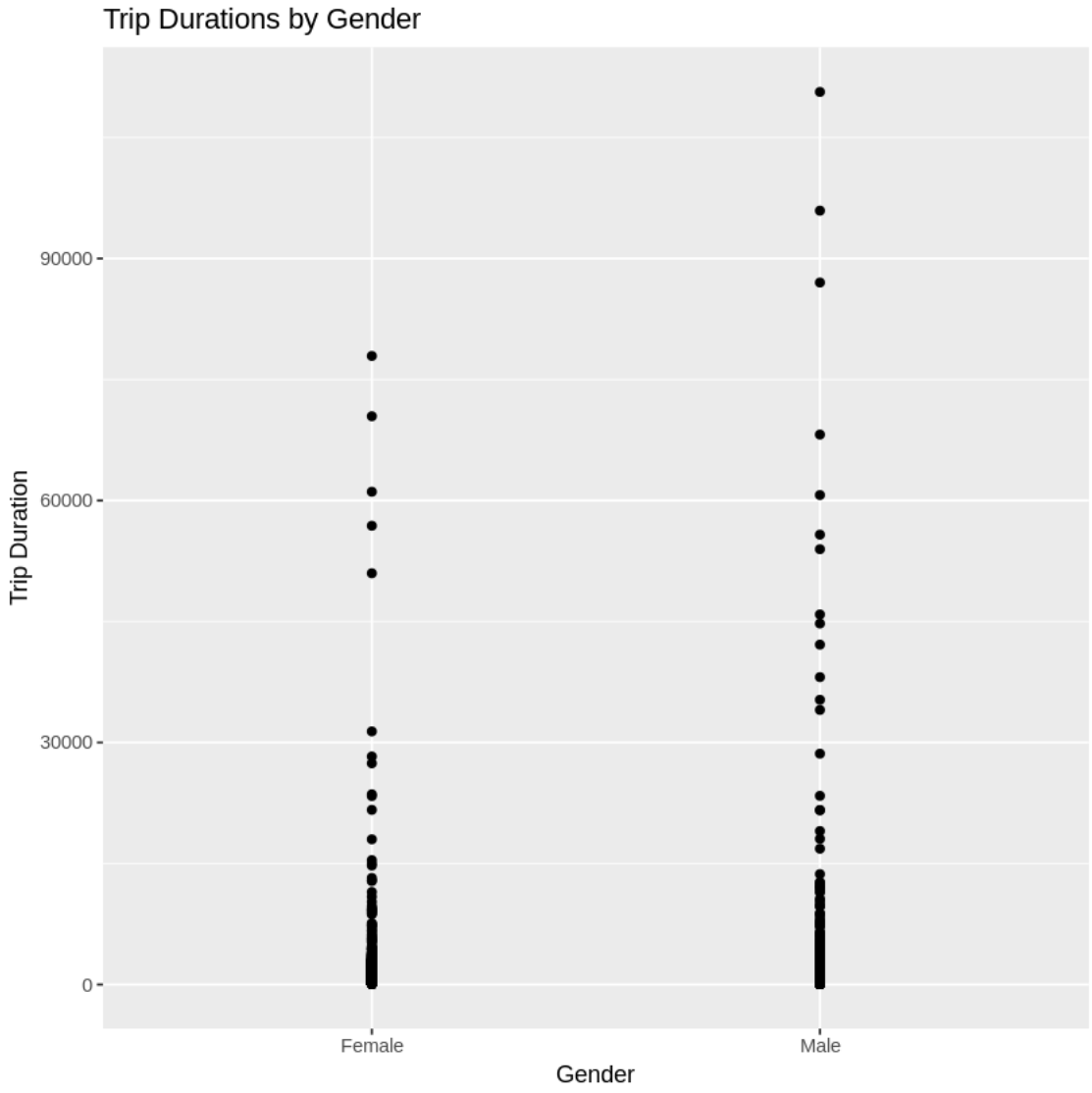
11425

In [107]:
```
# confirm
bsd_omit[11425,]
```

	X	Start.Time	End.Time	Trip.Duration	Start.Station	End.S
12719	6080502	2017-06-18 16:48:13	2017-07-01 07:12:08	1088634	Hope St & Union Ave	Halse

In [108]:
```
#drop the row
bsd_omit = bsd_omit[-c(11425),]
```

```
In [110]: #plot the data again
          ggplot(bsd_omit, aes(x=Gender, y=Trip.Duration)) +
              geom_point() +
              xlab("Gender") +
              ylab("Trip Duration") +
              ggtitle("Trip Durations by Gender")
```



It appears that the male customers have the longest trip duration in this dataset.

0.4.2 Question 2

What do the trip duration lengths look like by birth year?

Let's identify any outliers like the ones in the previous question.

```
In [111]: #find the index of any trip duration over 900000 minutes
          which(bsd$Trip.Duration > 900000)
```

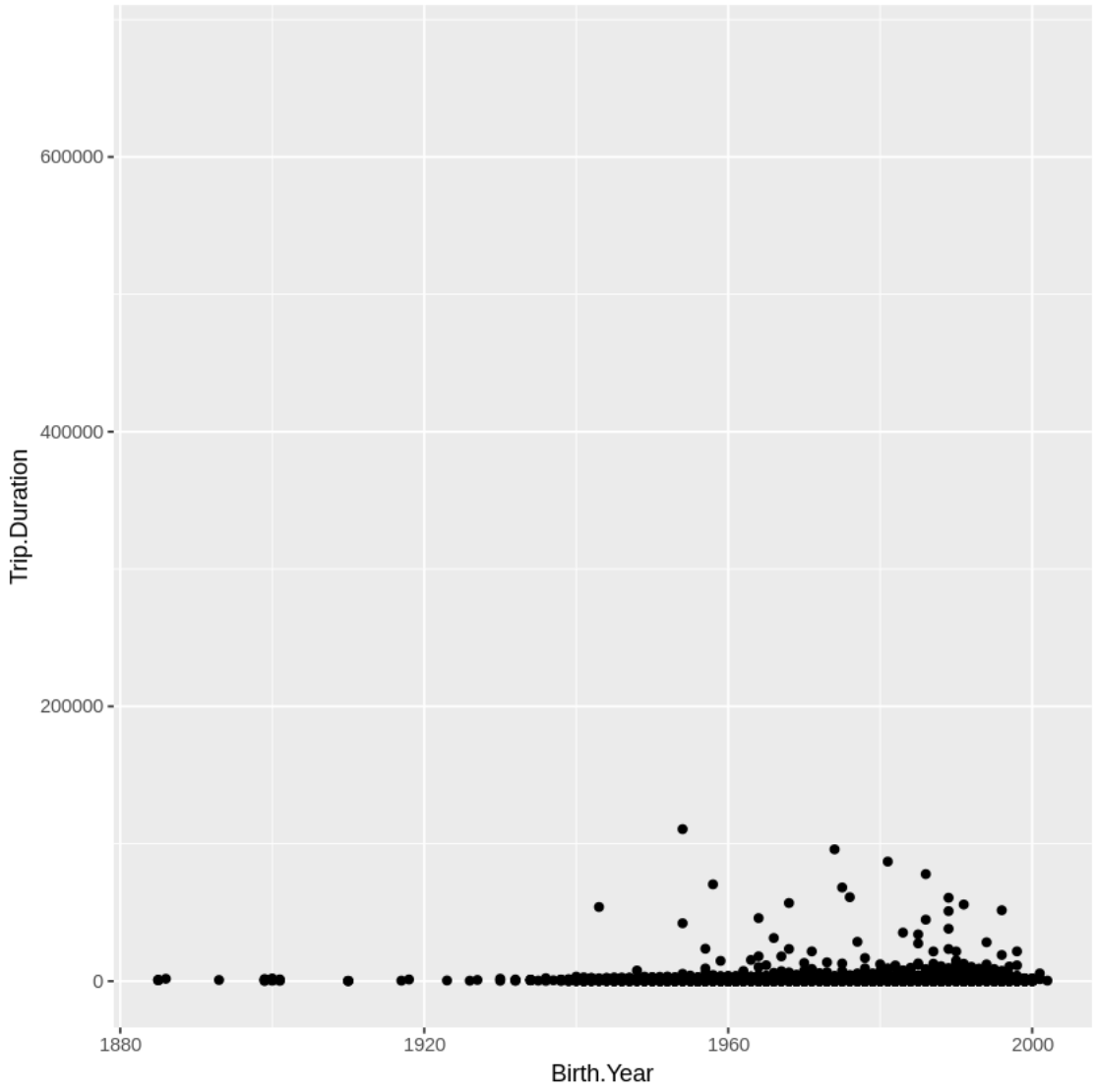
1. 12719 2. 82409

```
In [112]: #drop the first index outlier
          bsd = bsd[-c(12719),]
```

```
In [113]: #drop the second index outlier
          bsd = bsd[-c(82408),]
```

```
In [114]: ggplot(aes(x = Birth.Year, y = Trip.Duration), data = bsd) + geom_point()
```

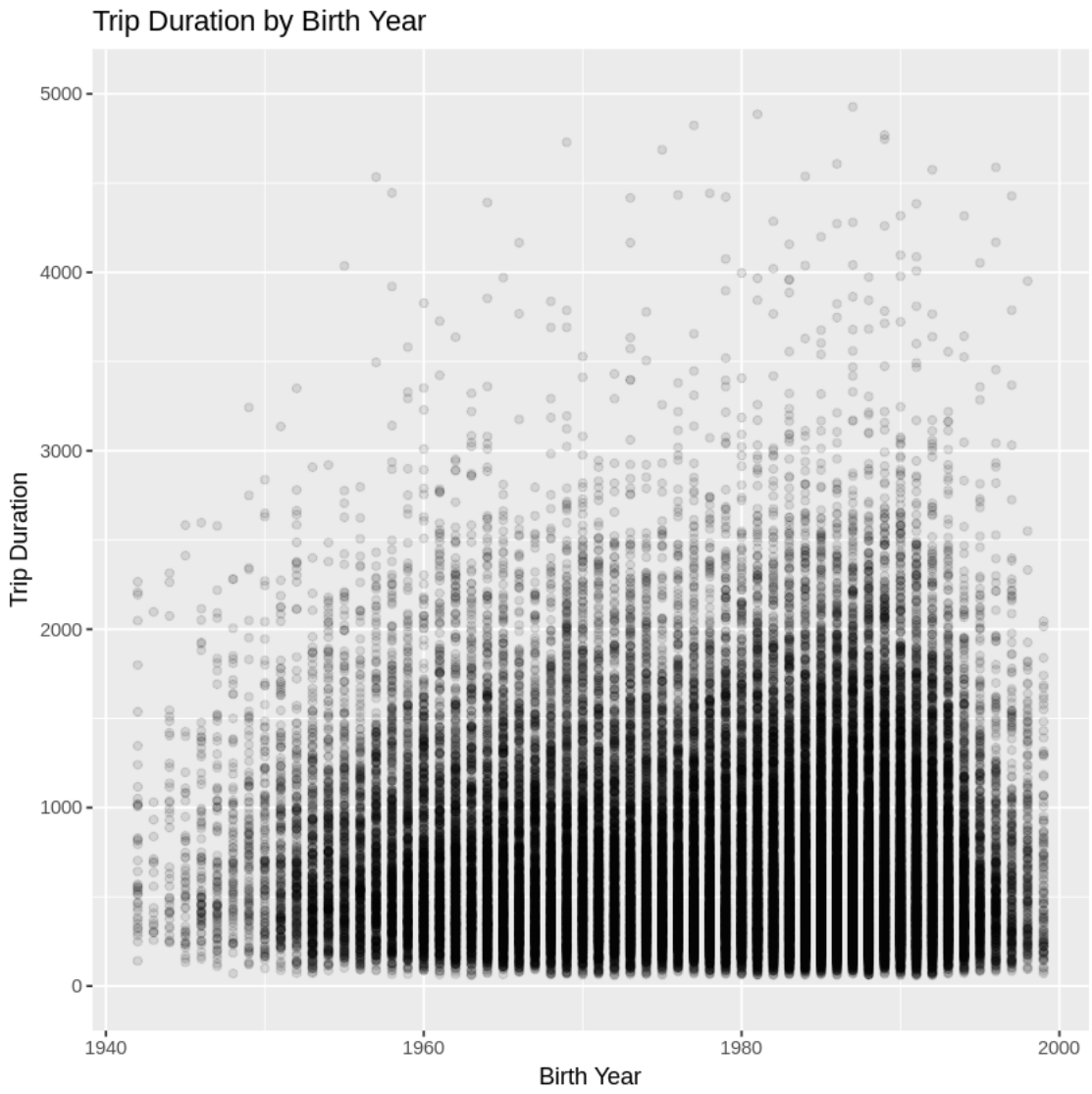
```
Warning message:
Removed 96015 rows containing missing values (geom_point).
```



Let's try to clean up this dataset a bit more: - The data provided is from 2017 - According to Motivate's website, no one under the age of 18 is allowed to rent, so we'll cap the max Birth Year off at 1999, - Since its likely some of the birth year data is in correct, we'll also shoot for an age range from 18 - 75, meaning the easliest year we'll keep in the dataset is 1942

```
In [115]: #update the plot with limits on the birth year and also apply a gradient to see a conc
          ggplot(aes(x = Birth.Year, y = Trip.Duration), data = bsd) +
              geom_point(alpha = 1/10) +
              xlim (1942, 1999) +
              ylim (0,5000) +
              xlab("Birth Year") +
              ylab("Trip Duration") +
              ggtitle("Trip Duration by Birth Year")
```

```
Warning message:
Removed 96311 rows containing missing values (geom_point).
```

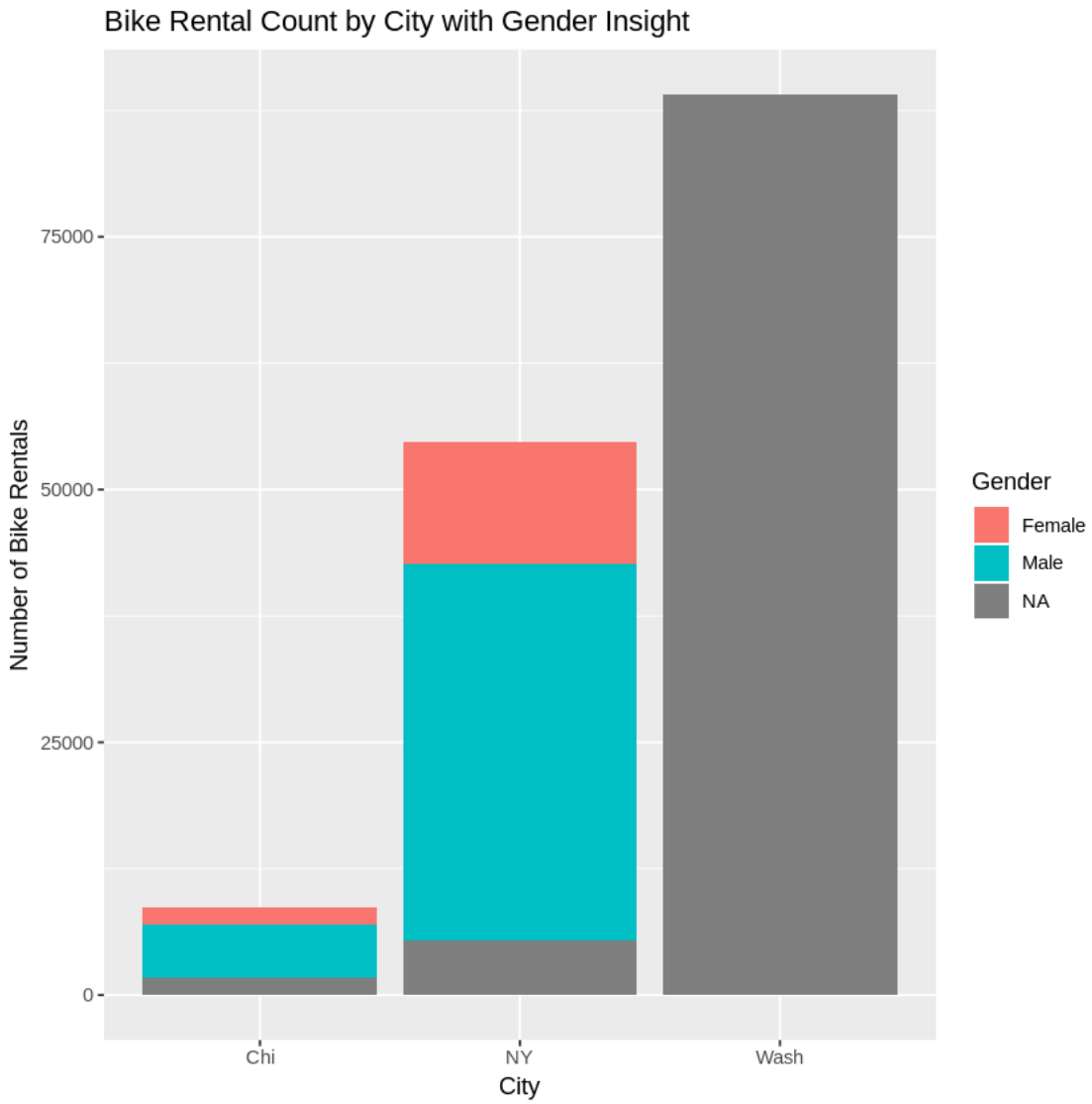


We can see a concentration of customers' trip durations that last between 0 and 1000 minutes, and we can see a trend that younger customers tend to have longer trips than older customers

0.4.3 Question 3

How many bike rentals happen in each city? And can we get a feel for the gender breakdown for those city's where we have gender data?.

```
In [116]: #create a barplot with a gender layer on the bar
          ggplot(data = bsd, aes(x=City)) +
              geom_bar(stat = "count", aes(fill=Gender)) +
              xlab("City") +
              ylab("Number of Bike Rentals") +
              ggtitle("Bike Rental Count by City with Gender Insight")
```



We can see that Washington has the biggest number of bike rentals in the dataset. As for gender, we can see in the citys where gender is reported, that men out number the women in regards to the number of bike rentals in each city.

0.5 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.

0.6 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 [118]: system('python -m nbconvert Explore_bikeshare_data.ipynb')
```