Explore_bikeshare_data

July 21, 2020

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

```
In [2]: head(ny)
```

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 &

```
In [3]: 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 [4]: 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 [5]:
```
#libraries
library(ggplot2)
```

0.0.2 Question 1

Your question 1 goes here.

- What is the most common month?

In [6]:
```
Na.Func <- function(N) {
    test.logic.null <- any(is.na(N))
    sum.null <- sum(is.na(N))
    colsums.null <- colSums(is.na(N))

    return(c(test.logic.null,sum.null,colsums.null))


}
Na.Func(chi)
Na.Func(ny)
Na.Func(wash)
```

1 1 **2** 1747 **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
1 1 **2** 5219 **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
1 1 **2** 1 **X** 0 **Start.Time** 0 **End.Time** 0 **Trip.Duration** 1 **Start.Station** 0 **End.Station** 0 **User.Type** 0

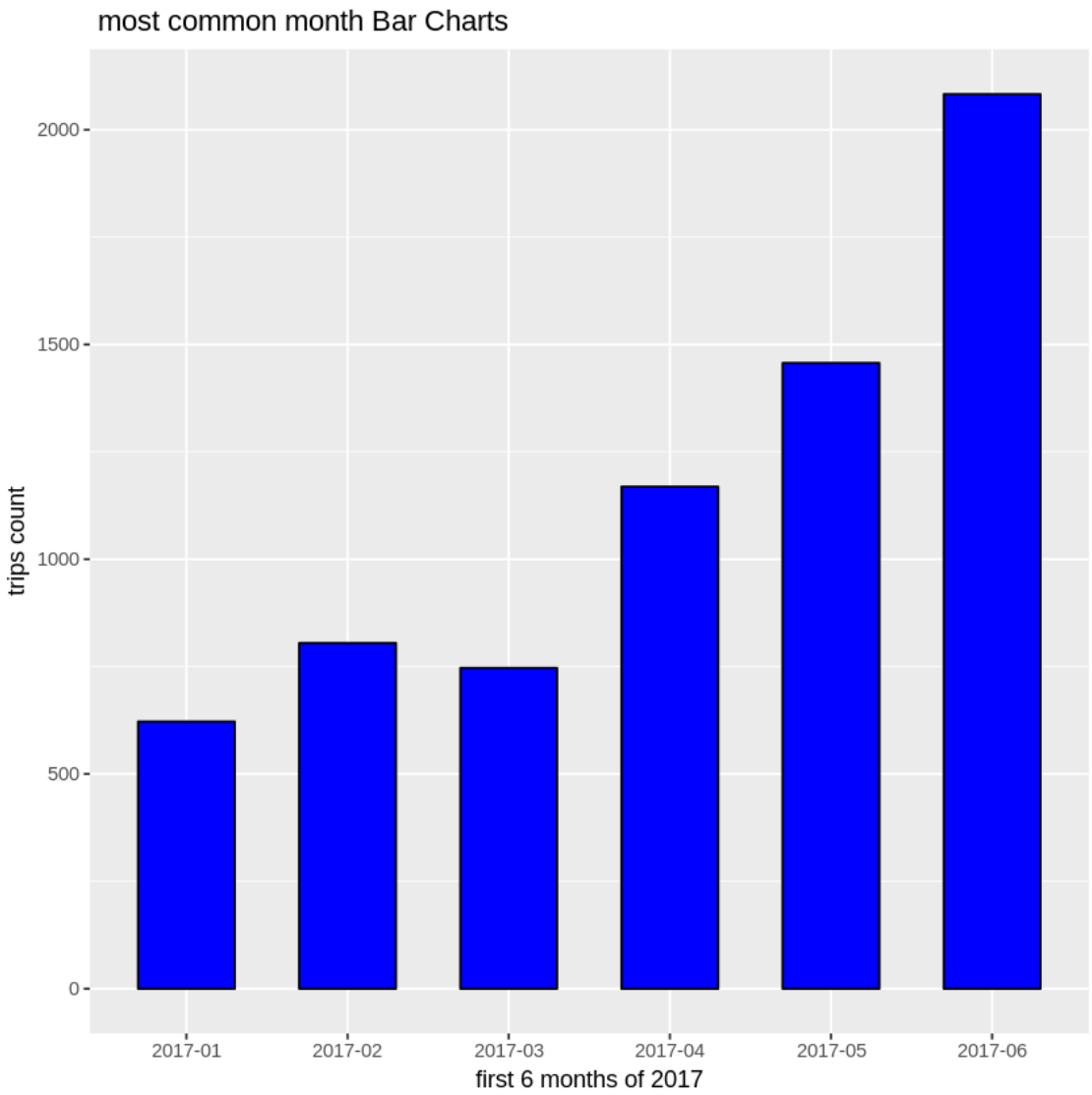
In [7]:
```
#Creating a mode function for the starting months.
Uniq.Func <- function(DS){
  U.F <- unique(DS)
  U.F[which.max(tabulate(match(DS, U.F)))]
}
#Calculating which month appeared the most.
Uniq.Func(chi)
Uniq.Func(ny)
Uniq.Func(wash)
```

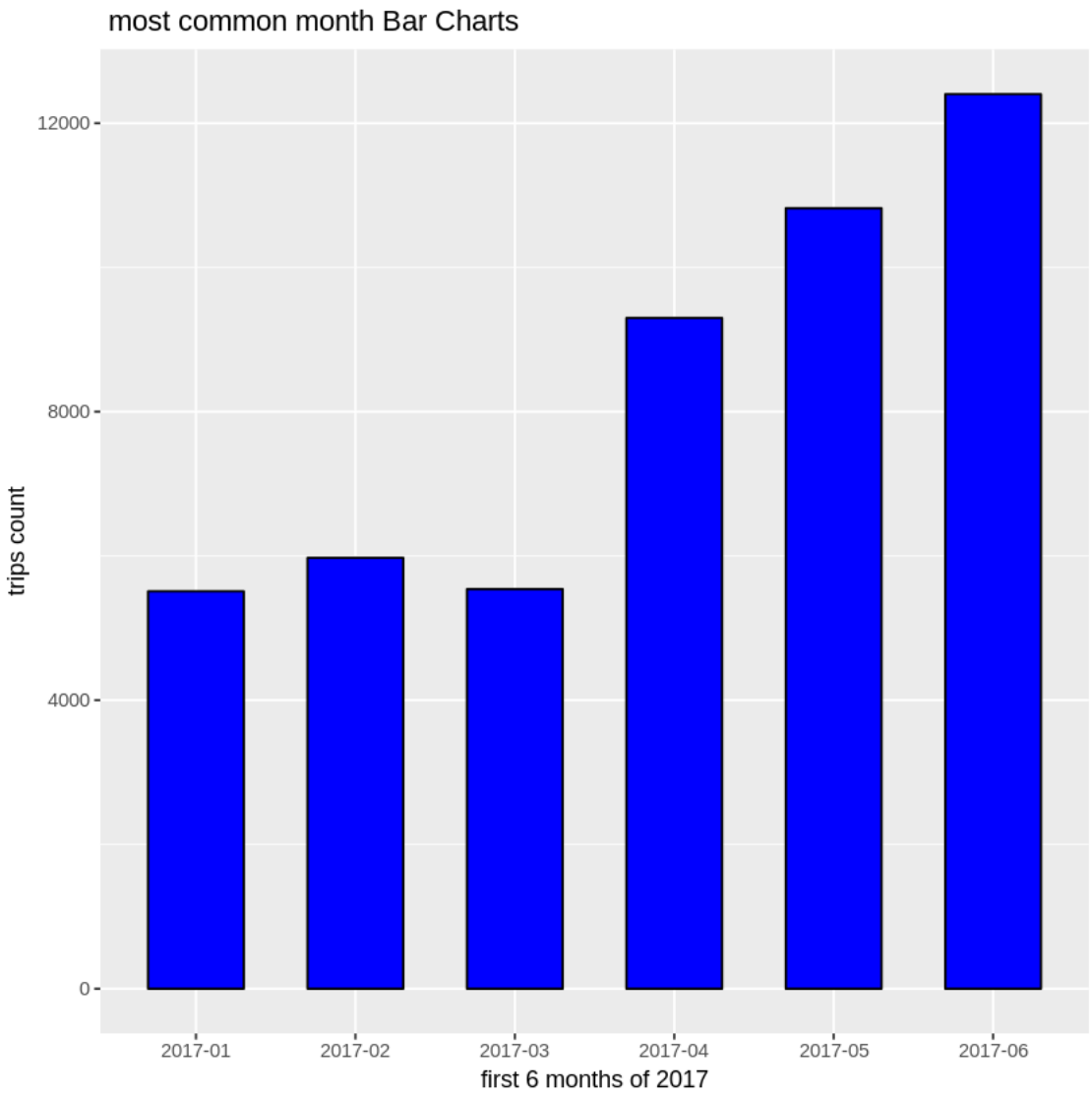
X
1423854
955915
9031
304487
45207
1473887
961916
65924
606841
135470
175805
71678
19061
1023296
611000
958716
718598
931608
1469705
475456
849468
1420915
161454
1413814
717248
1451810
385517
1186035
512692
261757
1171002
1371090
550447
243811
220157
1406209
1019063
1521139
1464943
252627
253570
152475
260811
602912
768050
601593
1256568
53692
1374113
1397375
1035179

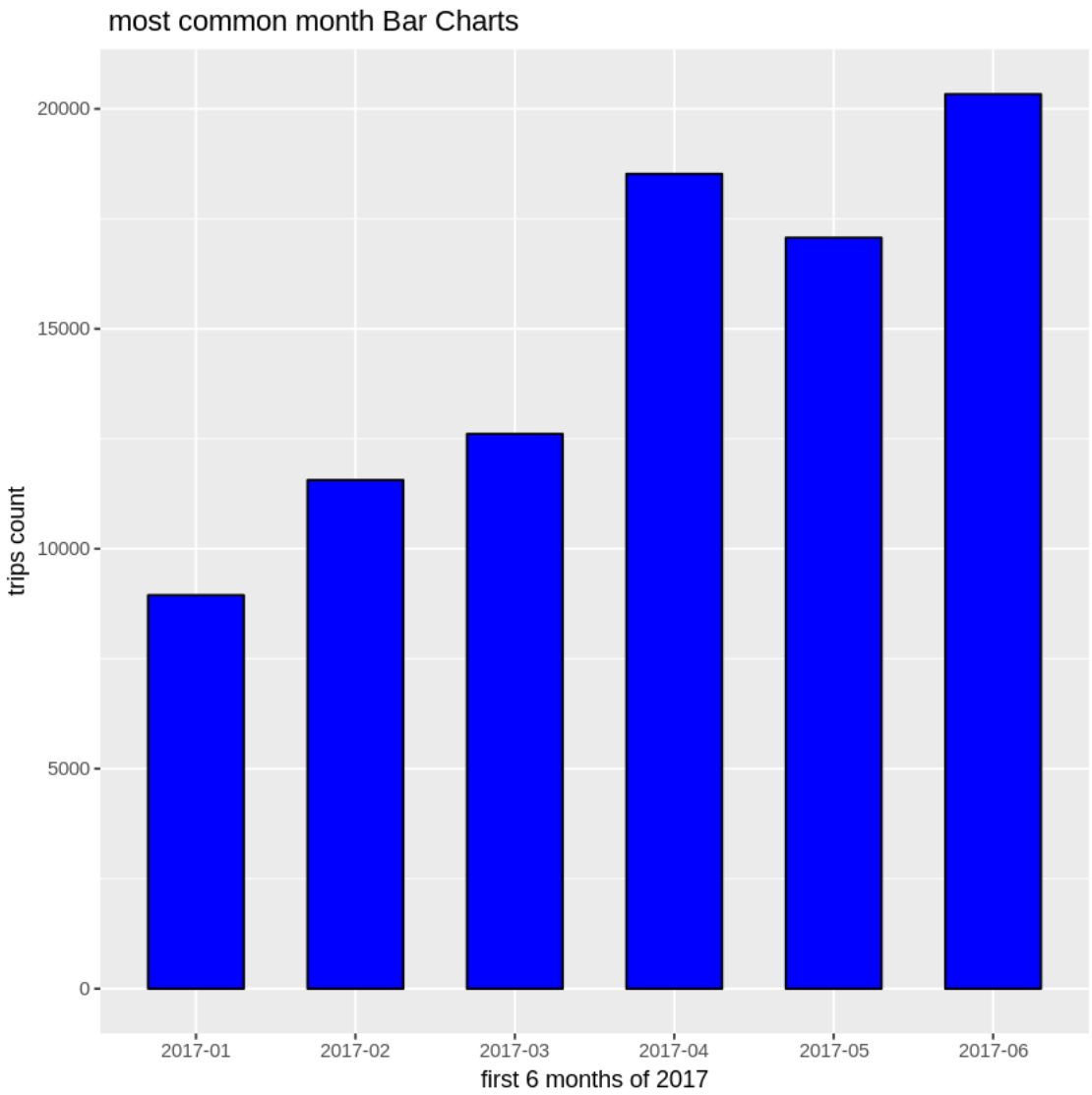
X
5688089
4096714
2173887
3945638
6208972
1285652
1675753
1692245
2271331
1558339
2287178
2744874
3398180
991609
1512596
187466
2195658
6388534
4733837
5857
1132766
3358474
1778858
2497952
2905932
3123311
2959550
2067887
3518426
5383277
4203558
5352804
3532080
2935863
6156169
250989
1275463
3161273
4211314
1717314
1881529
1047928
5623210
3197884
5758780
5082390
2521938
3759176
246004
1951636
6227477

X
1621326
482740
1330037
665458
1481135
1148202
1594275
1601832
574182
327058
854729
721887
143181
26953
1308542
1592744
1139713
1478516
679837
1392498
746605
868881
1161275
305000
1174029
1157984
694527
228536
669432
1036255
676408
551347
905153
625949
1667242
241483
893870
1673660
531841
1143611
760548
585303
774610
1681261
1068808
385427
242173
982818
1742606
1583466
1500573

In [8]:
```
Data.visualization <- function(DS) {

    # Convert DS Variable to d.t
    d.t = DS

    # The as.Date methods accept character strings, factors, logical NA and objects
     # Character strings are processed as far as necessary for the format specified:
      # any trailing characters are ignored.

    d.t$Start.Time <- as.Date(d.t$Start.Time)

    ggplot(aes(format(Start.Time, "%Y-%m")), data = na.omit(d.t)) +
      geom_bar(width = 0.6, color= 'black', fill = 'blue') +
      ggtitle(' most common month Bar Charts ') +
      labs(x = 'first 6 months of 2017', y = 'trips count')

}

Data.visualization(chi)
Data.visualization(ny)
Data.visualization(wash)
```


most common month Bar Charts
trips count
2000
1500
1000
500
0
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
first 6 months of 2017


most common month Bar Charts
12000
8000
4000
0
trips count
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
first 6 months of 2017



Summary of your question 1 results goes here.

In [9]:
```
" From the resulting plots we can see that the month with the most number
of trips is June (06/2017) and its the case for all the three cities
Which makes sense, because people drive bikes more in the summer than the
other seasons"

#Summary For All Datasets
summary(chi['Start.Time'])
print('-------------------------------------------------------------------------------')
summary(ny['Start.Time'])
print('-------------------------------------------------------------------------------')
summary(wash['Start.Time'])
```

' From the resulting plots we can see that the month with the most number\nof trips is June (06/2017) and its the case for all the three cities\nWhich makes sense, because people drive bikes more in the summer than the\nother seasons'

```
               Start.Time
 2017-01-24 07:40:32:   2
 2017-04-22 13:16:25:   2
 2017-05-27 15:17:50:   2
 2017-06-10 13:29:41:   2
 2017-06-20 17:05:11:   2
 2017-06-21 13:18:52:   2
 (Other)            :8618


[1] "--------------------------------------------------------------------------------"


               Start.Time
 2017-05-11 18:26:10:    3
 2017-01-04 13:58:24:    2
 2017-01-09 09:36:01:    2
 2017-01-21 15:36:56:    2
 2017-01-21 17:49:59:    2
 2017-01-21 20:08:29:    2
 (Other)            :54757


[1] "--------------------------------------------------------------------------------"


               Start.Time
 2017-02-19 12:19:00:    6
 2017-02-20 11:35:00:    6
 2017-02-24 17:46:00:    6
 2017-03-01 08:20:00:    6
 2017-03-02 08:39:00:    6
 2017-03-09 17:31:00:    6
 (Other)            :89015
```

0.0.3 Question 2

Your question 2 goes here.

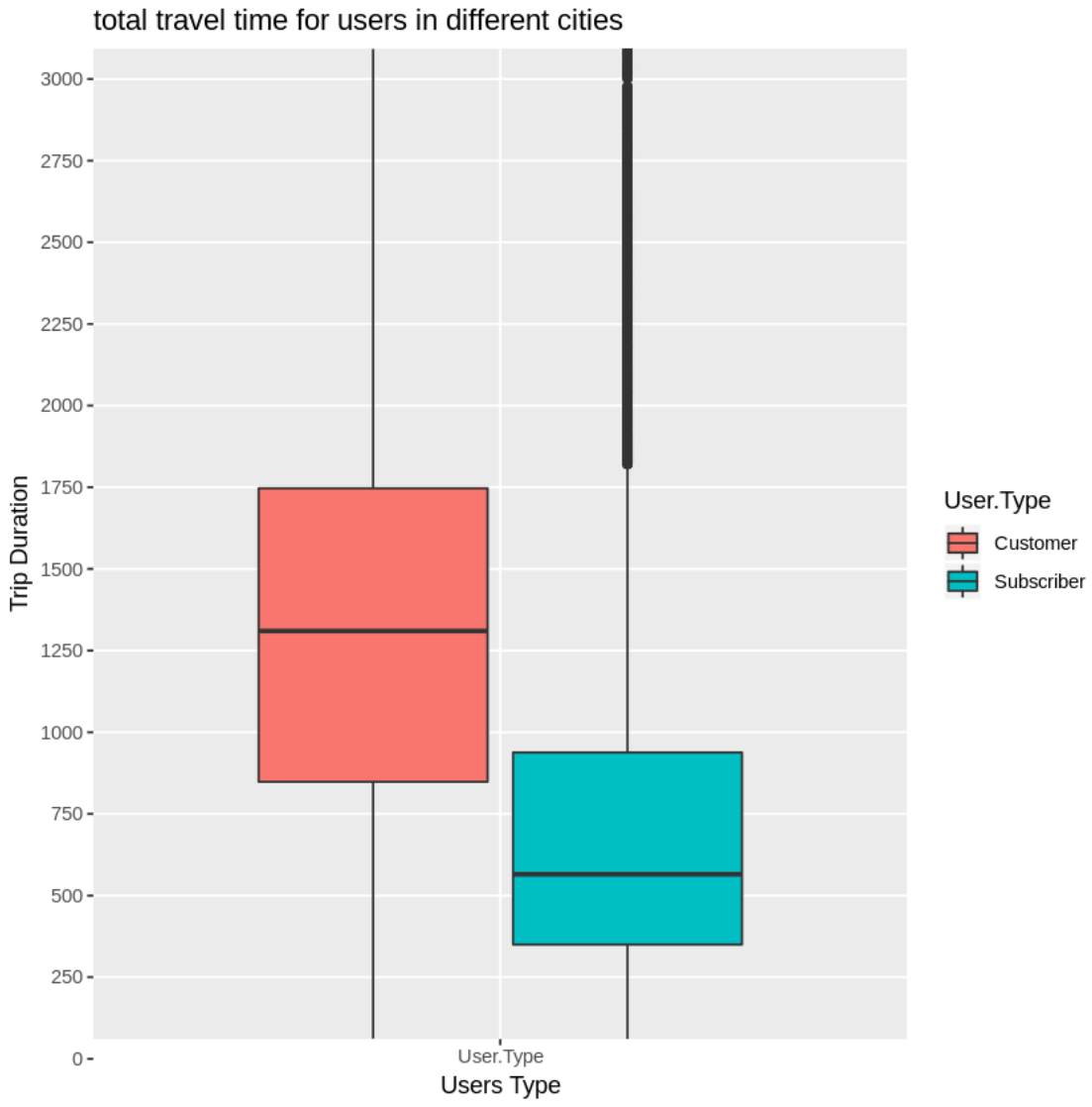
- What is the total travel time for users in different cities?
- What is the average travel time for users in different cities?

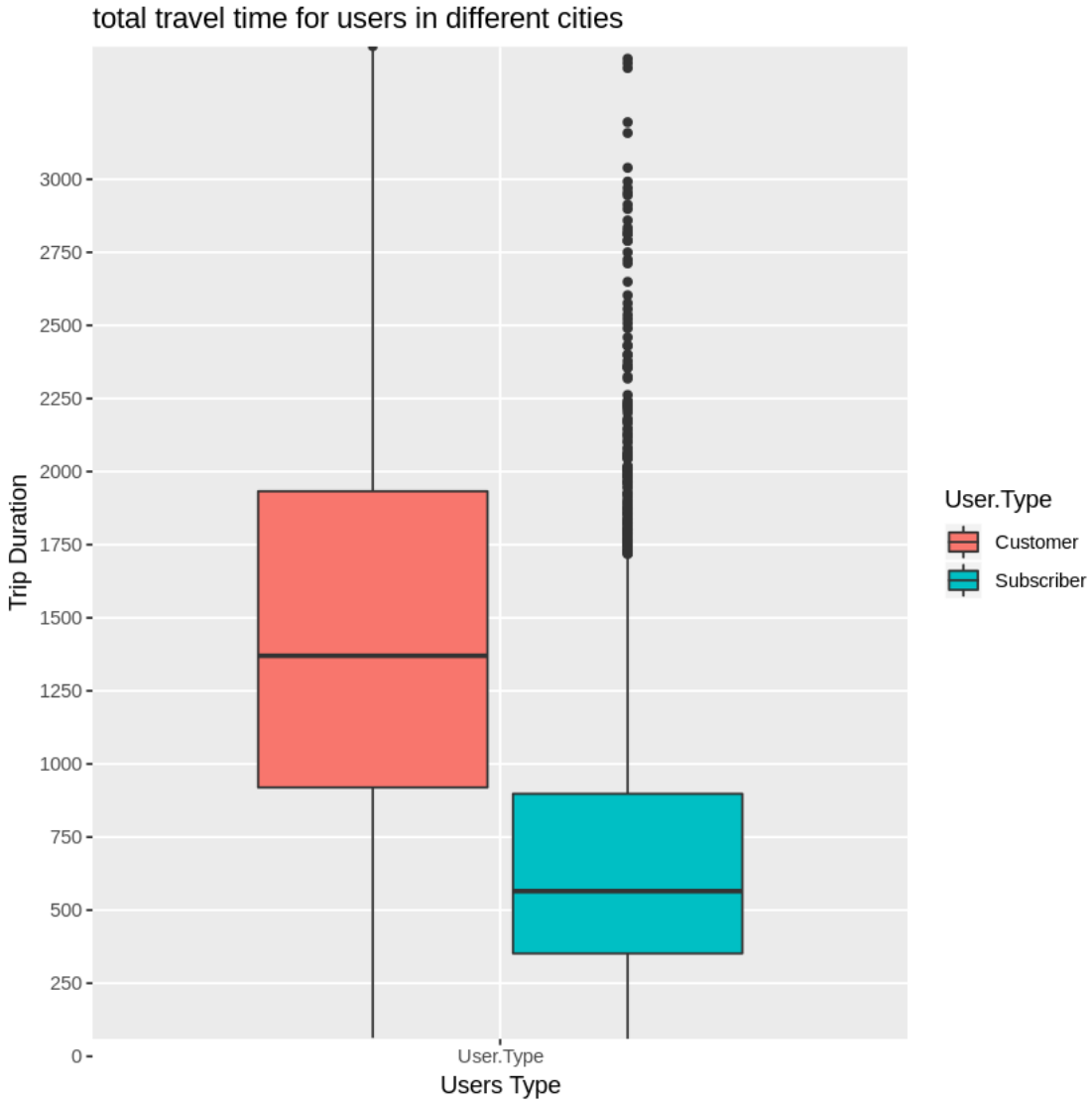
```
In [10]: head(ny, 3)
         dim(ny['Trip.Duration'])
```

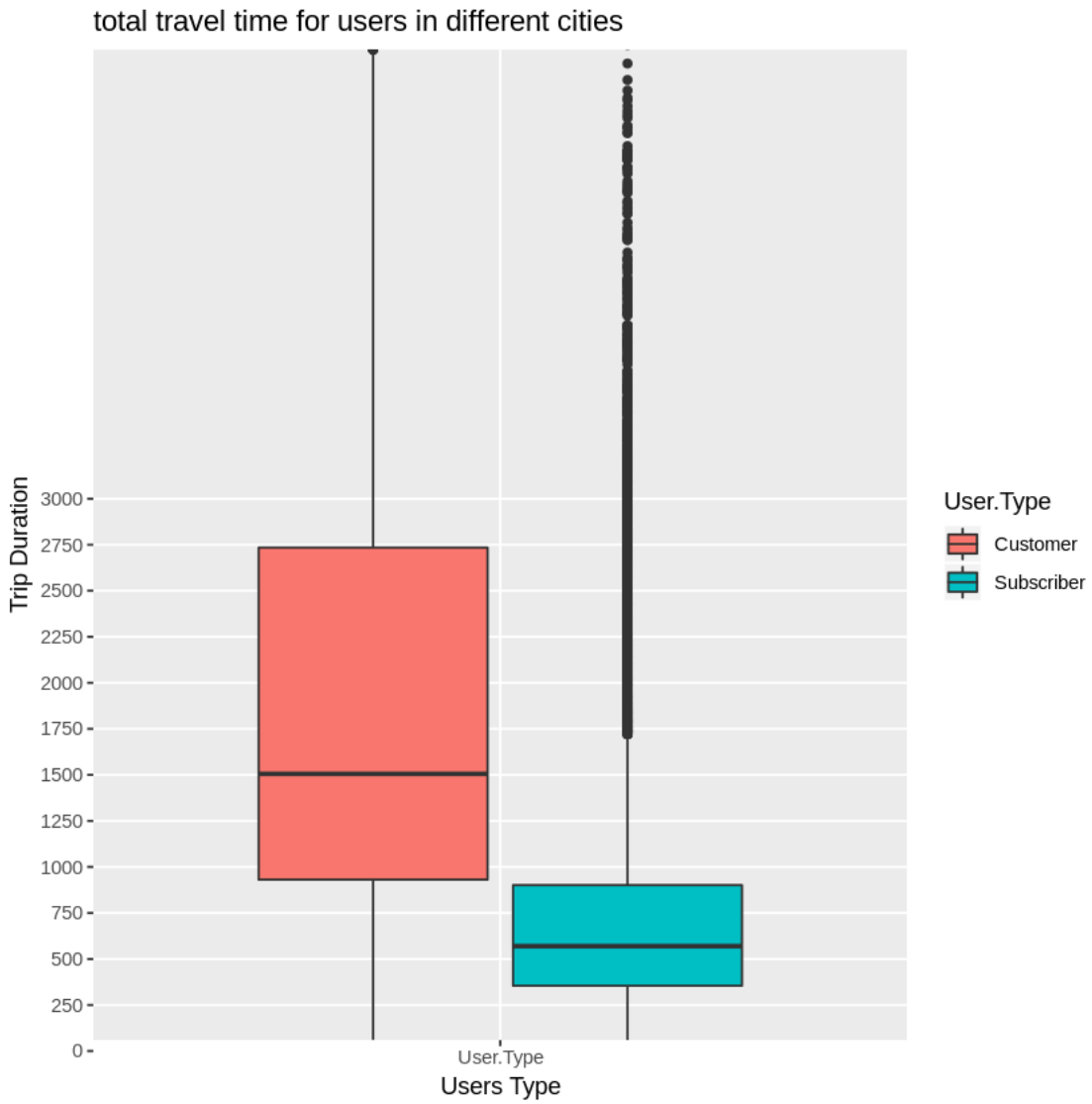
X	Start.Time	End.Time	Trip.Duration	Start.Station	End.Station
5688089	2017-06-11 14:55:05	2017-06-11 15:08:21	795	Suffolk St & Stanton St	W Broadw
4096714	2017-05-11 15:30:11	2017-05-11 15:41:43	692	Lexington Ave & E 63 St	1 Ave & E 7
2173887	2017-03-29 13:26:26	2017-03-29 13:48:31	1325	1 Pl & Clinton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

1. 54770 2. 1

```
In [11]: Total.T.D.Users <- function(DS){
             ggplot(aes("User.Type",`Trip.Duration`, fill = User.Type), data = subset(DS, User.T
               geom_boxplot()+
               scale_y_discrete(limits = seq(0,3000,250))+
               ggtitle('total travel time for users in different cities')+
               labs(x = 'Users Type', y = 'Trip Duration')

         }
         Total.T.D.Users(ny)
         Total.T.D.Users(chi)
         Total.T.D.Users(wash)
```




total travel time for users in different cities
Trip Duration
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
User.Type
Users Type
User.Type
Customer
Subscriber




```
In [12]: "We notice that the customers numbers in trip duration in the 3 states
          are much more than the numbers of subscribers
          the customers seems to be very enthusiastic!!!"

          by(ny$Trip.Duration, ny$User.Type, summary)
          print("------------------------------------------------------------")
          by(chi$Trip.Duration, chi$User.Type, summary)
          print("------------------------------------------------------------")
          by(wash$Trip.Duration, wash$User.Type, summary)
```

'We notice that the customers numbers in trip duration in the 3 states \nare much more than the numbers of subscribers \nthe customers seems to be very enthusiastic!!!'

ny$User.Type:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  201.0   762.5  1112.0  1838.5  1519.0 51595.0       1
------------------------------------------------------------
ny$User.Type: Customer
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
     62.0     848.2    1310.0    2193.1    1747.0 1088634.0
------------------------------------------------------------
ny$User.Type: Subscriber
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    61.0    350.0    565.0    755.4    938.0 110648.0


[1] "------------------------------------------------------------"


chi$User.Type:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3020    3020    3020    3020    3020    3020
------------------------------------------------------------
chi$User.Type: Customer
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   63.0   919.8  1370.0  1930.0  1932.8 85408.0
------------------------------------------------------------
chi$User.Type: Subscriber
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     60     352     565     685     898   21634


[1] "------------------------------------------------------------"


wash$User.Type:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
     NA      NA      NA     NaN      NA      NA       1
------------------------------------------------------------
wash$User.Type: Customer
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    61.3    930.9   1505.1   2634.4   2734.3 904591.4
------------------------------------------------------------
wash$User.Type: Subscriber
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
    60.27    354.55    569.69    733.33    901.17 170032.91
```

In [13]: "We note that the descending order of the clients is in the 3 states respectively
1) Washonton
2) New York
3) Chicago
And in terms of subscribers

```
1) New York
2) Washonton
3) Chicago.
"

by(ny$Trip.Duration , ny$User.Type, mean )
print("_____________________________________________________________")
by(chi$Trip.Duration , chi$User.Type, mean )
print("_____________________________________________________________")
by(wash$Trip.Duration , wash$User.Type, mean )
```

'We note that the descending order of the clients is in the 3 states respectively\n1) Washonton\n2) New York\n3) Chicago\nAnd in terms of subscribers\n1) New York\n2) Washonton\n3) Chicago.\n'

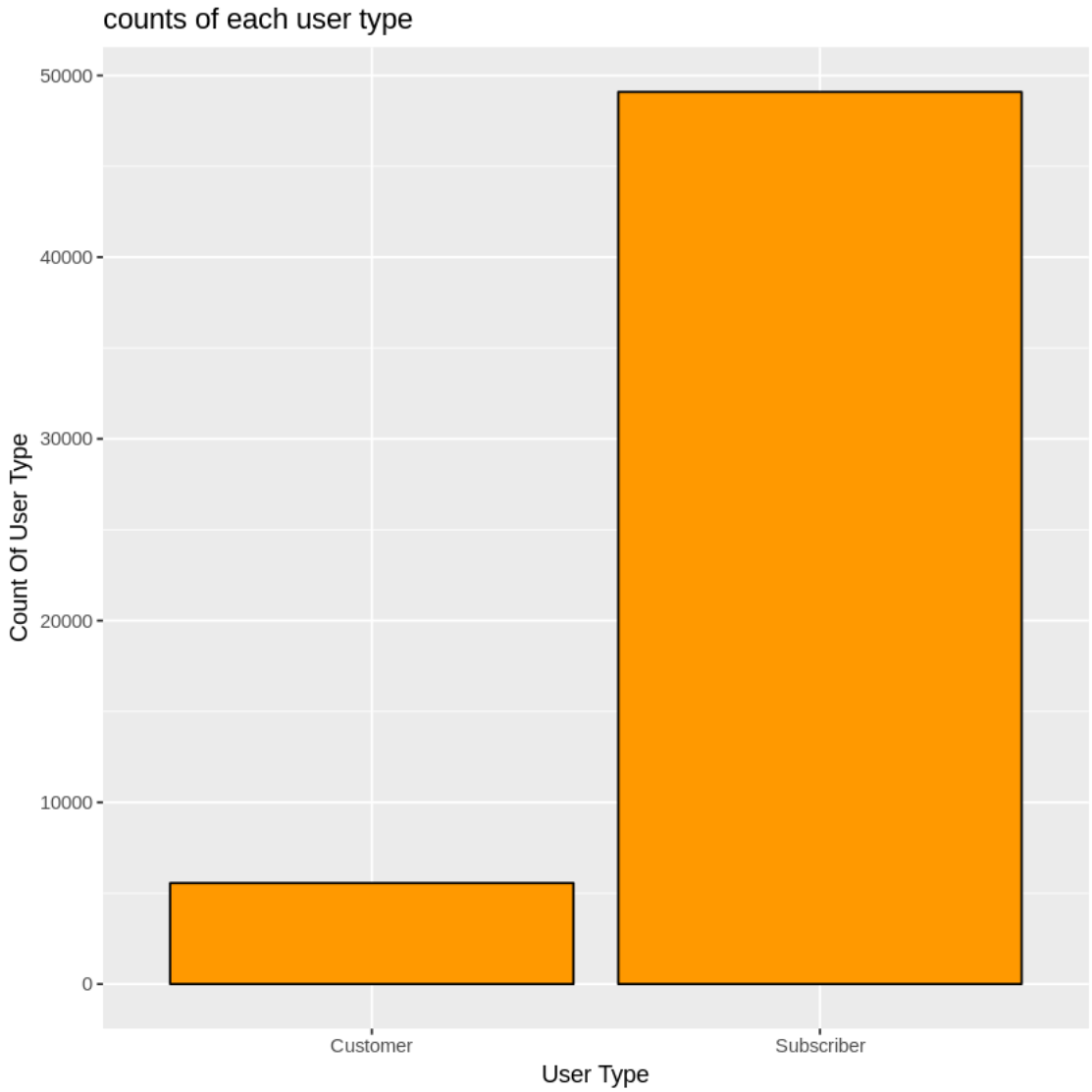
```
ny$User.Type:
[1] NA
------------------------------------------------------------
ny$User.Type: Customer
[1] 2193.076
------------------------------------------------------------
ny$User.Type: Subscriber
[1] 755.3829


[1] "_____________________________________________________________"


chi$User.Type:
[1] 3020
------------------------------------------------------------
chi$User.Type: Customer
[1] 1929.977
------------------------------------------------------------
chi$User.Type: Subscriber
[1] 685.027


[1] "_____________________________________________________________"


wash$User.Type:
[1] NA
------------------------------------------------------------
wash$User.Type: Customer
[1] 2634.429
------------------------------------------------------------
wash$User.Type: Subscriber
[1] 733.326
```

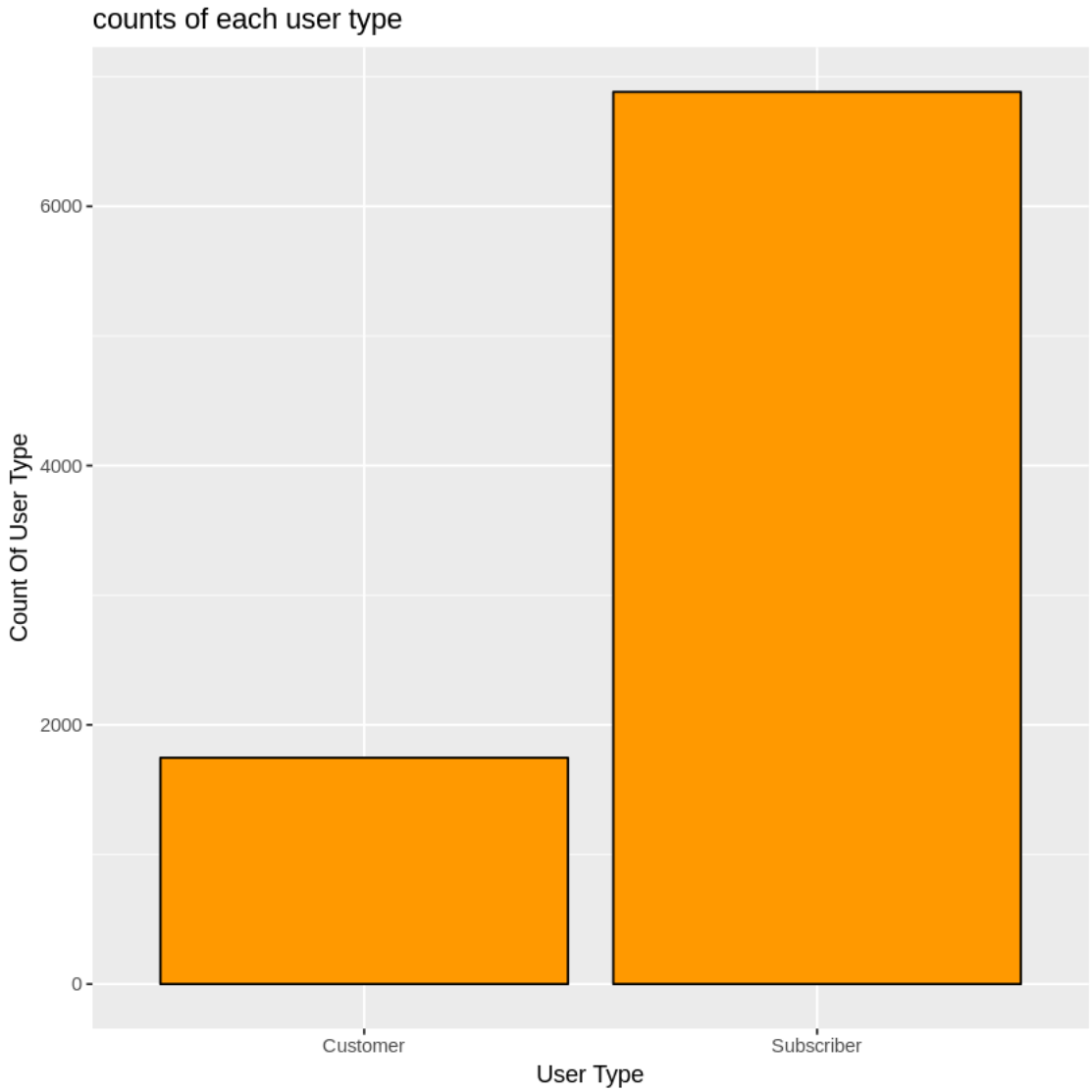
Summary of your question 2 results goes here.

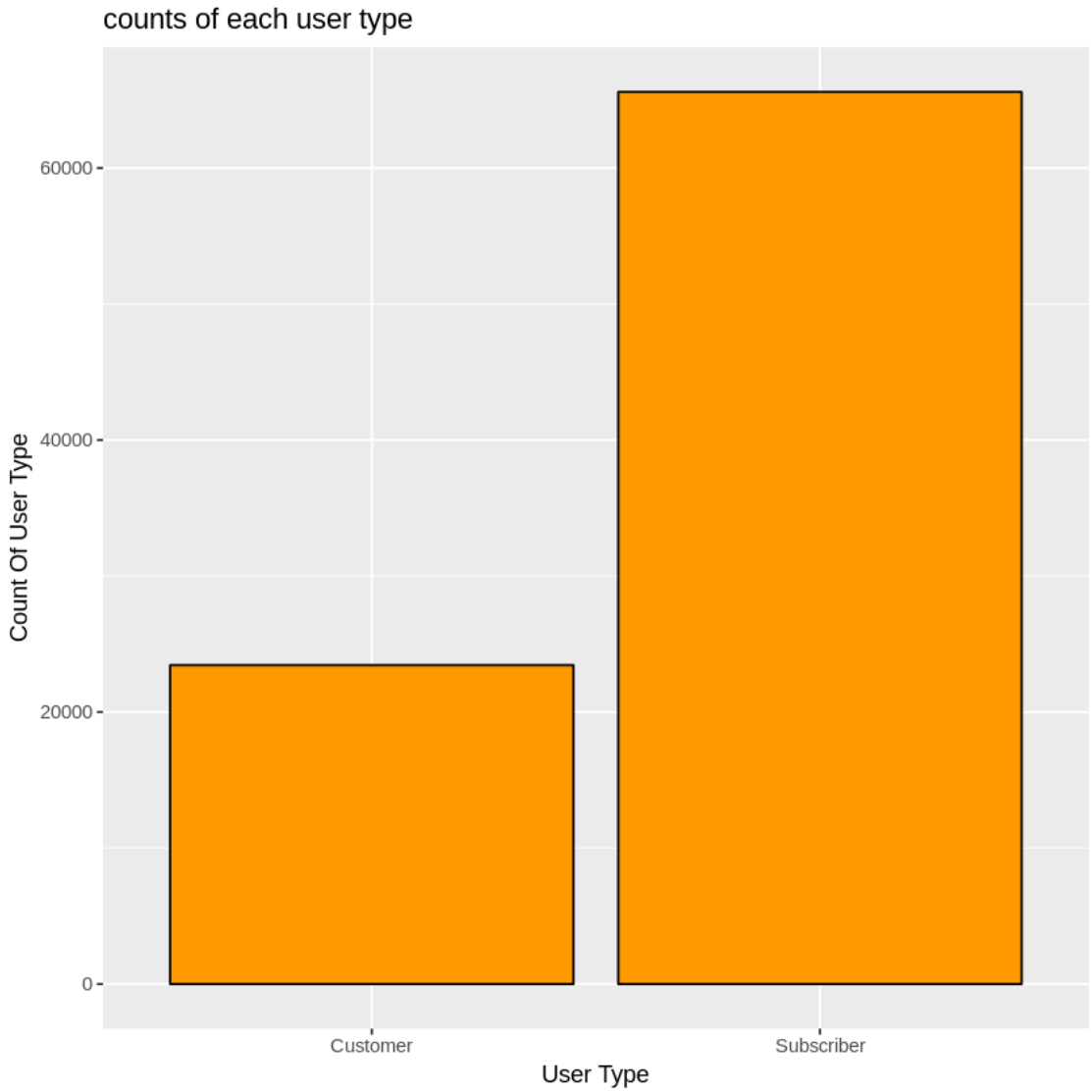
0.0.4 Question 3

Your question 3 goes here. - What are the counts of each user type? - What are the counts of each gender (only available for NYC and Chicago)? - What are the earliest, most recent, most common year of birth (only available for NYC and Chicago)?

```
In [14]: User.Type <- function(DS) {
             ggplot(aes(x = User.Type), data=subset(DS, User.Type != "")) +
             geom_bar(color ='black', fill = '#ff9900') +
             ggtitle('counts of each user type')+
             labs(x = 'User Type', y = 'Count Of User Type')

         }
         User.Type(ny)
         User.Type(chi)
         User.Type(wash)
```


counts of each user type
50000
40000
30000
20000
10000
0
Count Of User Type
Customer
Subscriber
User Type


counts of each user type
Count Of User Type
6000
4000
2000
0
Customer
Subscriber
User Type



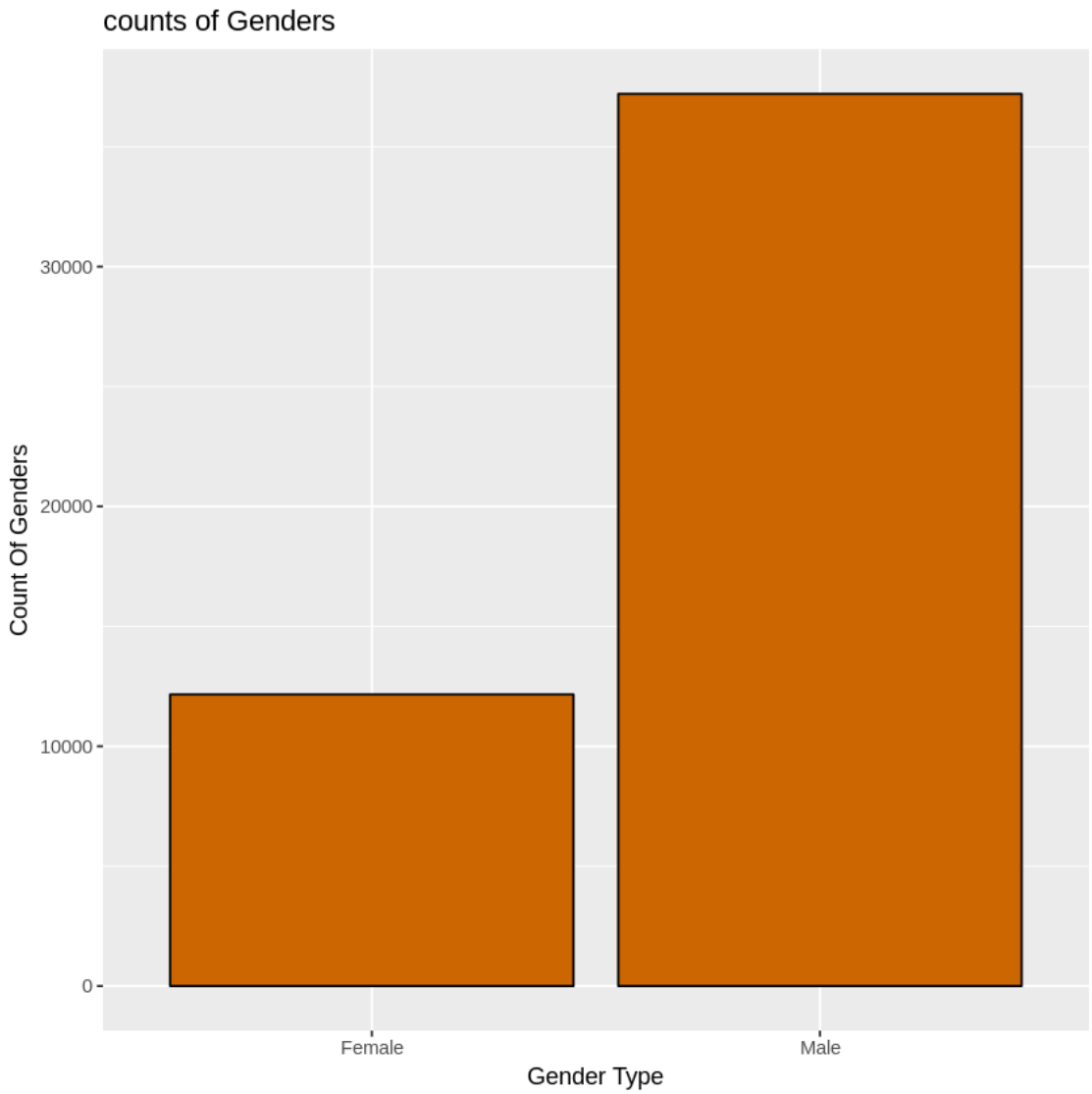
In [15]: *# We notice here that subscribers in the three states have more numbers than the custom*

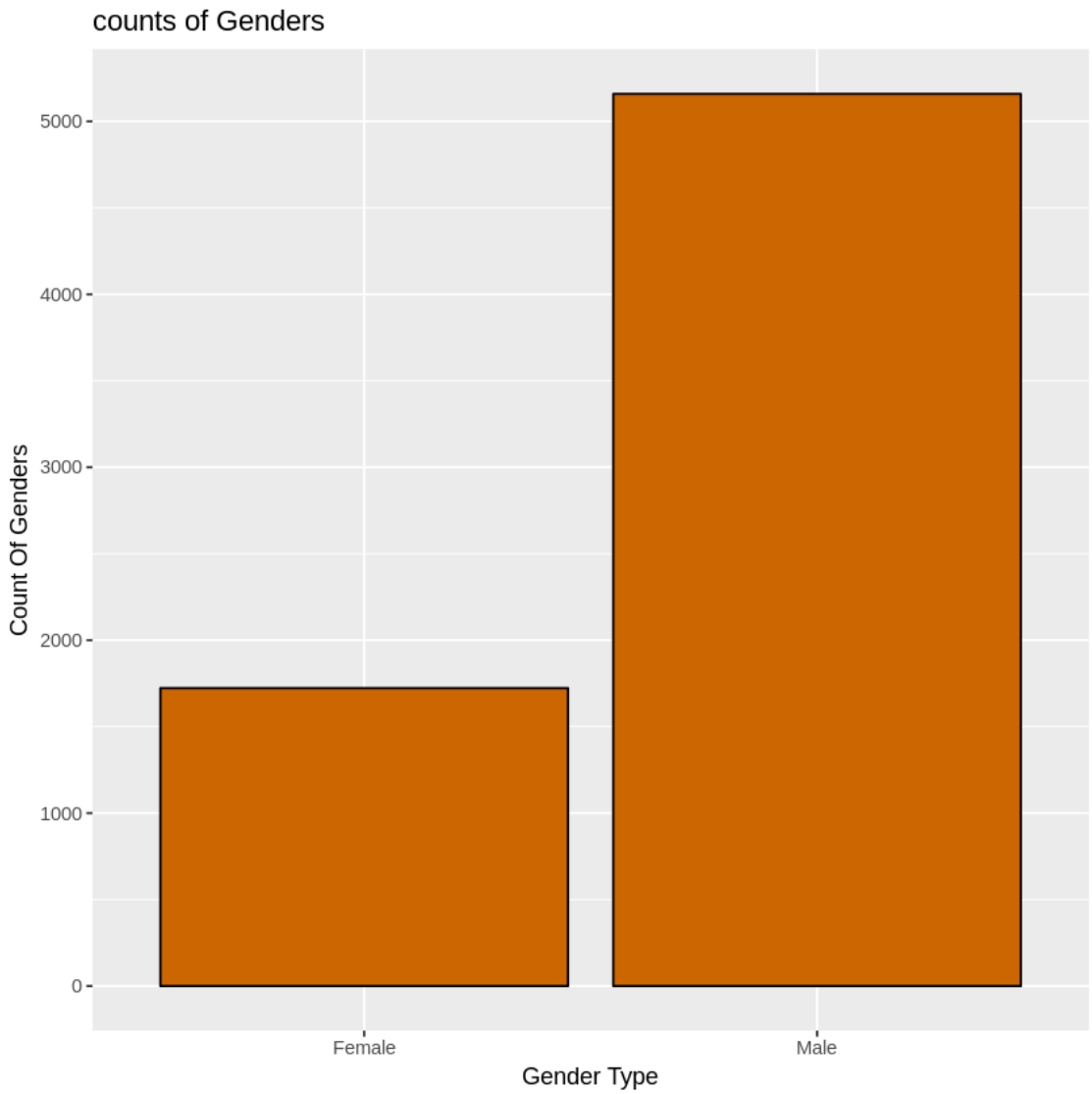
```
summary(ny['User.Type'])
print('--------------')
summary(chi['User.Type'])
print('--------------')
summary(wash['User.Type'])
```

```
     User.Type
          :  119
 Customer  : 5558
 Subscriber:49093
```

```
[1] "--------------"


      User.Type
           :   1
 Customer  :1746
 Subscriber:6883


[1] "--------------"


      User.Type
           :    1
 Customer  :23450
 Subscriber:65600
```

```
In [16]: Count.Gender <- function(DS) {
             ggplot(aes(x = Gender), data = subset(DS, Gender != "")) +
             geom_bar(color ='black', fill = '#cc6600') +
             ggtitle('counts of Genders')+
             labs(x = 'Gender Type', y = 'Count Of Genders')

         }
         Count.Gender(ny)
         Count.Gender(chi)
```


counts of Genders
Count Of Genders
30000
20000
10000
0
Female
Male
Gender Type



In [17]: *# We notice here that MAle in the New york And Chicago have more numbers than the Femal*

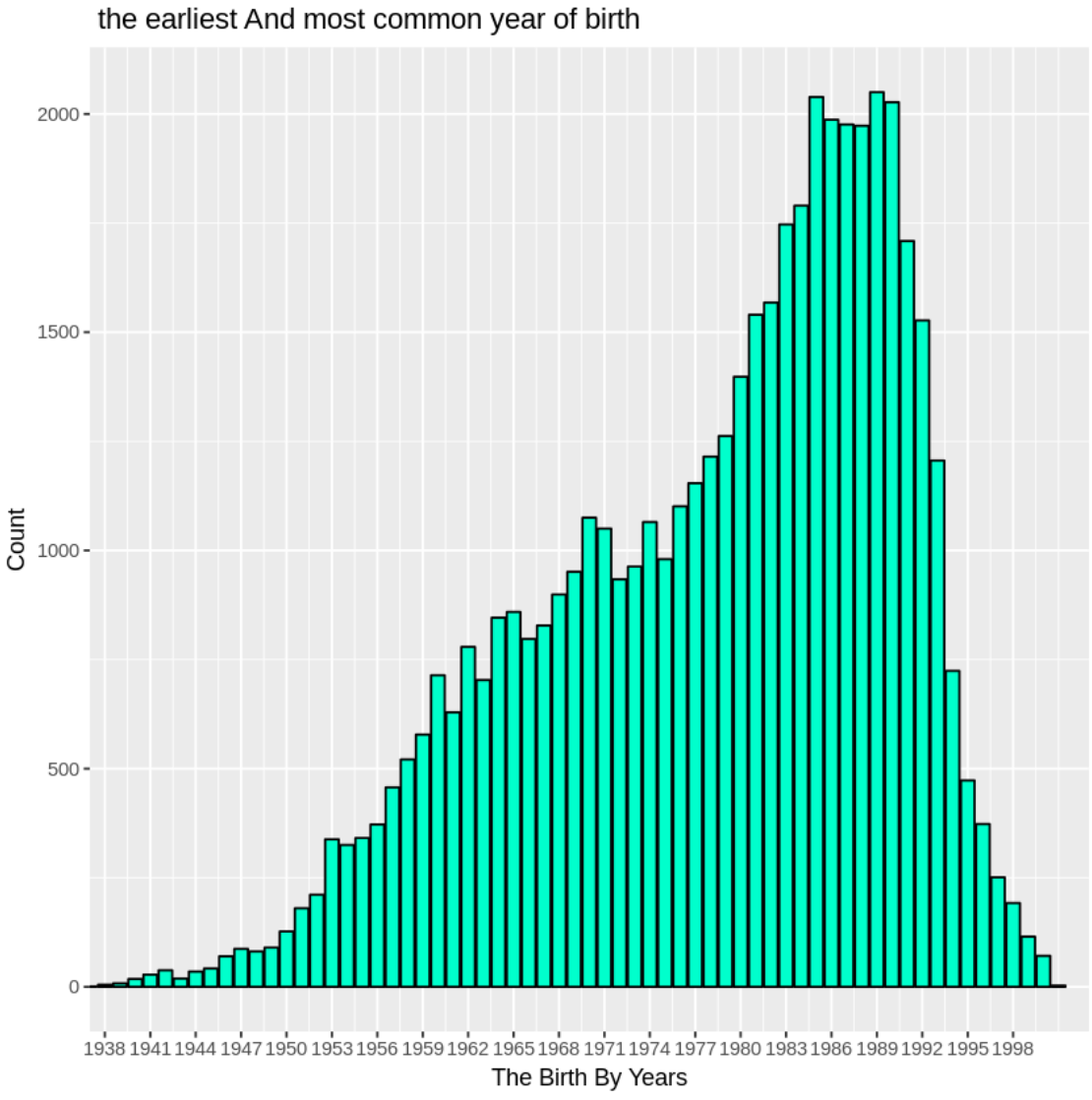
```
summary(ny['Gender'])
print('--------------')
summary(chi['Gender'])
```

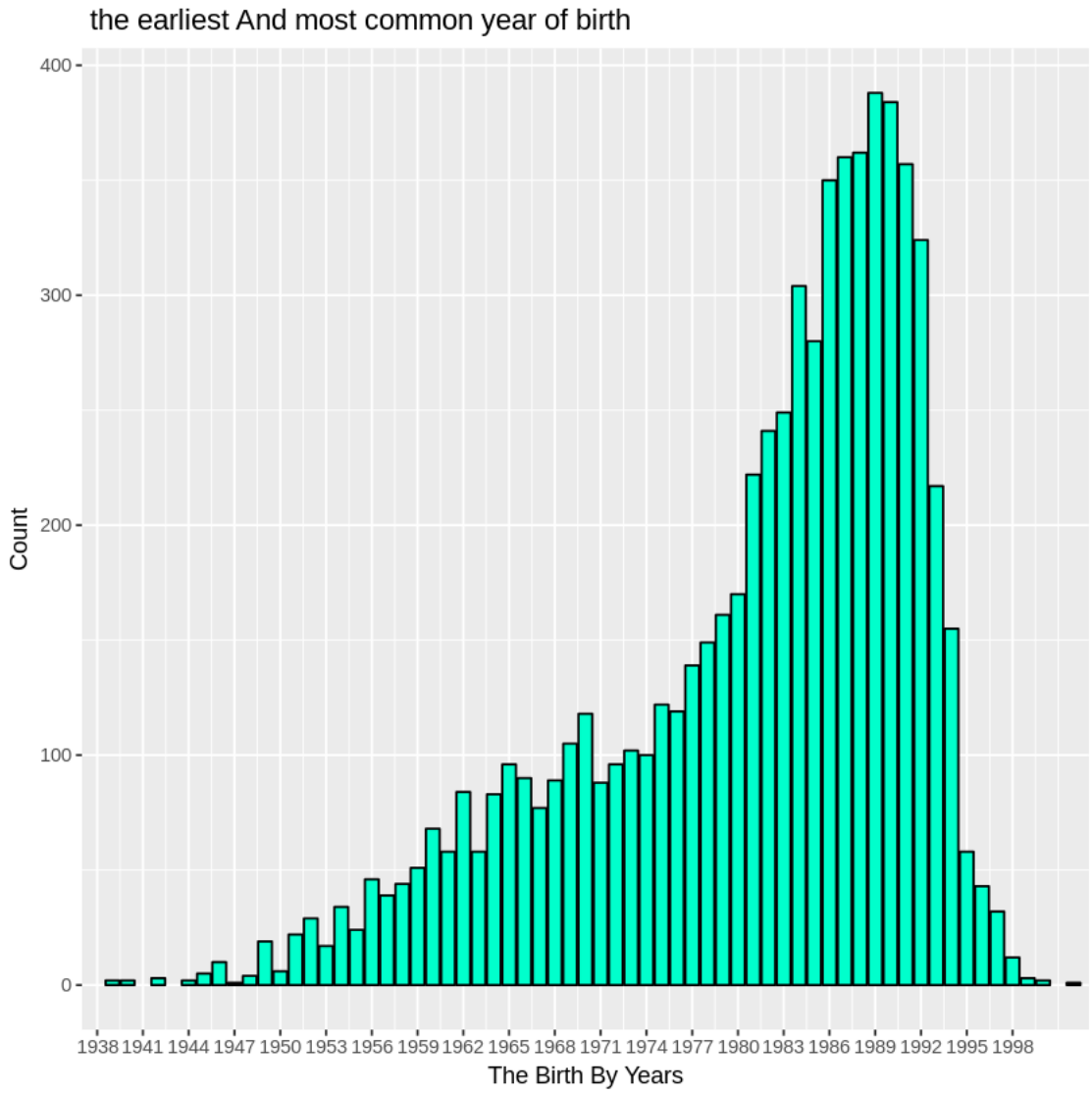
```
   Gender
      : 5410
 Female:12159
 Male  :37201
```

```
[1] "--------------"
```

```
    Gender
       :1748
 Female:1723
 Male  :5159
```

```
In [18]: E.M.Year.Of.Birth <- function(DS) {
             ggplot(aes(x = Birth.Year), data = na.omit(DS))+
             geom_bar(color = 'black', fill = '#00ffcc')+
             scale_x_continuous(breaks = seq(0,2000,3))+
             coord_cartesian(xlim=c(1940,2000)) +
             ggtitle(' the earliest And most common year of birth') +
             labs(x = 'The Birth By Years', y = 'Count')
         }
         E.M.Year.Of.Birth(ny)
         E.M.Year.Of.Birth(chi)
```






```
In [19]: "From the resulting plots we can see that, the earliest, most recent, most common year
         New York State (Between 1985 And 1990), But the biggest of them is (1989),
         And in the state of Chicago most common year of birth (Between 1986 And 1992), But the
         is the same age in New York (1989)
         "
         #Summary For All Datasets
         summary(chi['Birth.Year'])
         print('------------')
         summary(ny['Birth.Year'])
```

'From the resulting plots we can see that, the earliest, most recent, most common year of birth Of\nNew York State (Between 1985 And 1990), But the biggest of them is (1989),\nAnd in the state of Chicago most common year of birth (Between 1986 And 1992), But the biggest of them \nis the same age in New York (1989) \n'

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

```
[1] "------------"
```

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

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