HOMEWORK 1
Submitted by: SHARANG BIRADAR N11407631

Part 1: CitiBike Descriptive

Analytical Questions

Q. Compute summary statistics for tripduration

data <- read.csv("~/Desktop/Assignment/CitiBike Data.csv")
View(data)
attach(data)
summary(tripduration)

```
> summary(tripduration)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   60.0   408.0   648.0   822.5  1038.0 33610.0
>
```

Q. Compute summary statistics for age

```
age<-2016-birth.year
summary(age)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
19.00   32.00   38.00   40.43   47.00   96.00
```

Q. Compute summary statistics for tripduration in minutes (Need to transform tripduration from seconds to minutes)

```
tripduration_min<- tripduration/60
summary(tripduration_min)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00    6.80   10.80   13.71   17.30  560.10
```

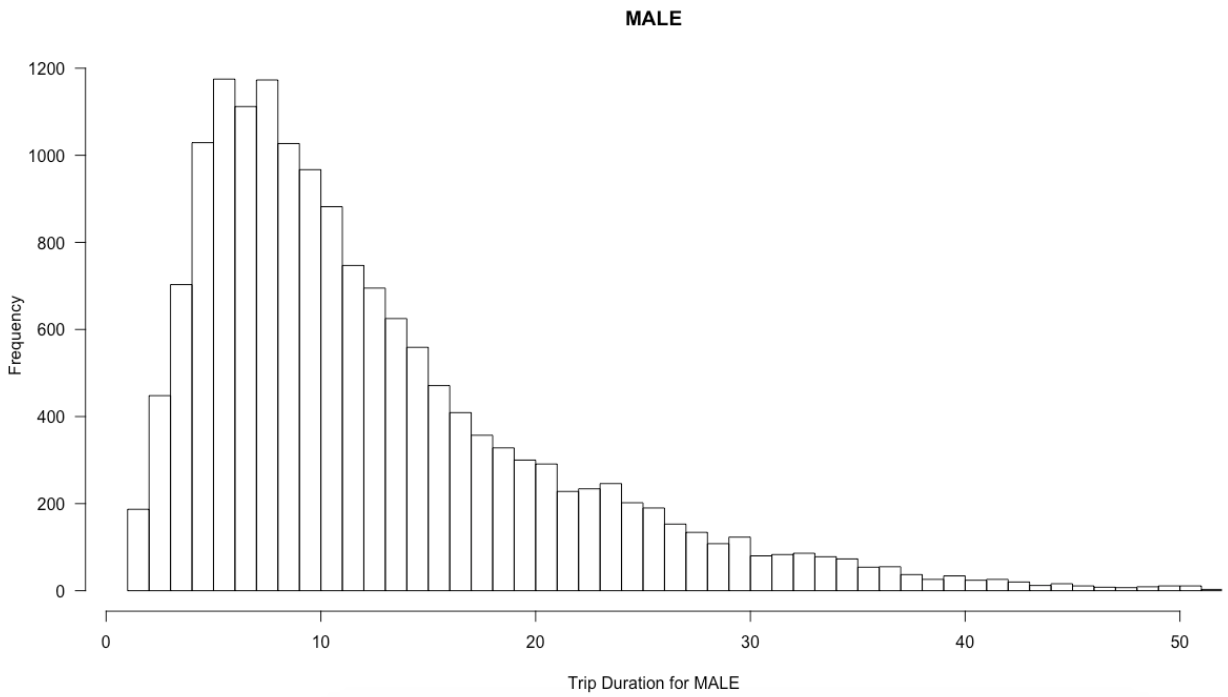
Q. Compute the correlation between age and tripduration

```
> cor(age, tripduration_min)
[1] 0.01140616
```

Q. Plot the histograms and box plots for tripduration by gender

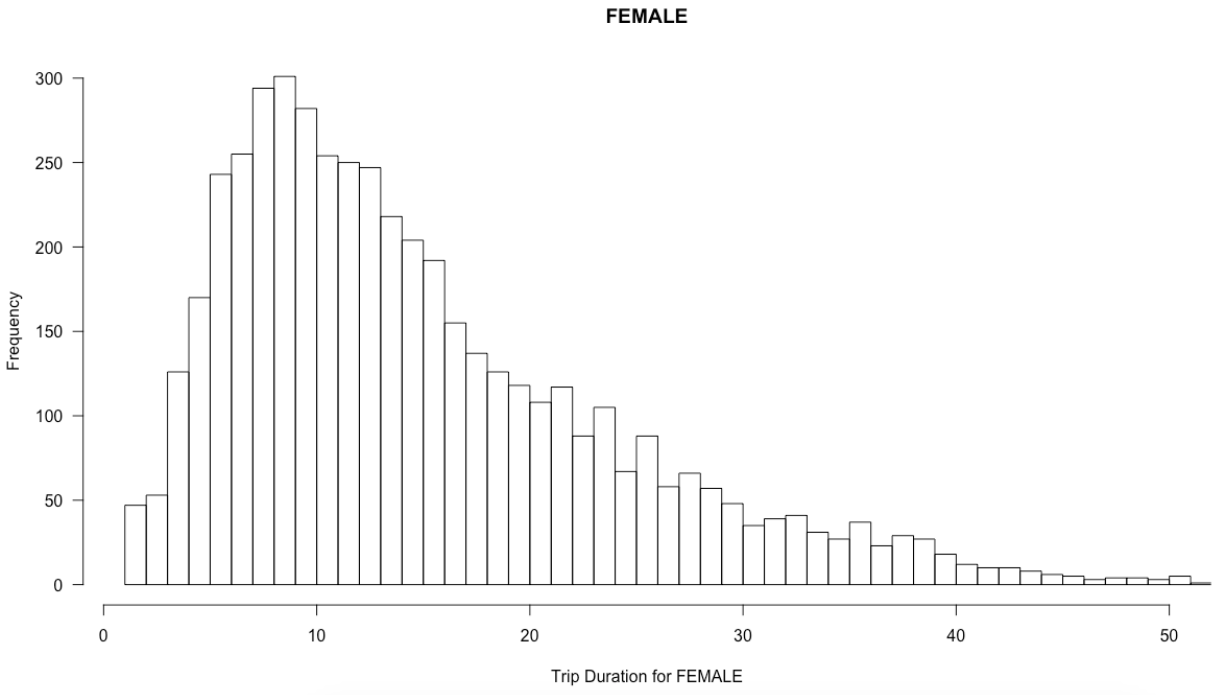
a. Histogram for Male

```
hist(tripduration_min[gender==1],xlim=c(1,50),breaks=500,
     main="MALE", las=1, xlab="Trip Duration for MALE")
```



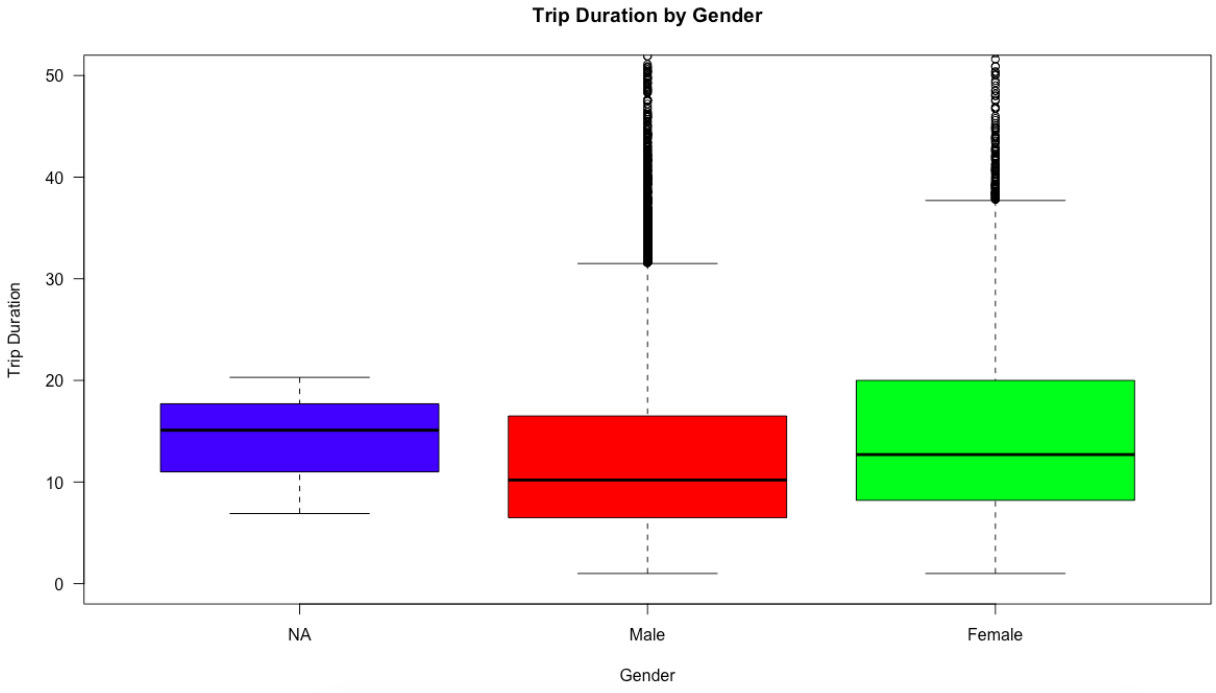
b. Histogram for Female

```
hist(tripduration_min[gender==2],xlim=c(1,50),breaks=500,
     main="FEMALE", las=1, xlab="Trip Duration for FEMALE")
```



c. Boxplot for Trip Duration by Gender

```
boxplot(tripduration_min~ gender, ylim=c(0,50),
    names=c("NA","Male","Female"), col=c("Blue", "Red", "Green"),
    las=1, ylab="Trip Duration", xlab= "Gender",
    main="Trip Duration by Gender")
```



Business Question

Q. What is the total revenue assuming all users riding bikes from 0 to 45 minutes pay $3 per ride and user exceeding 45 minutes pay an additional $2 per ride.

```
 rev_045<-sum(tripduration_min<45)*3
 rev_045
1] 61911
 rev_45<-sum(tripduration_min>45)*5
 rev_45
1] 1080
 totalrev<-rev_045+rev_45
 totalrev
1] 62991
```

Q. Looking at tripduration in minutes, what can you say about the variance in the data.

```
var(tripduration_min)
1] 196.2727
```

a. What does this mean for the pricing strategy?

```
summary (tripduration_min)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    6.80   10.80   13.71   17.30  560.10
```

```
sd(tripduration_min)
] 14.00974
```

The mean of the data is 13.71 minutes and the Standard Deviation is about 14 minutes which is between the trip duration of 0- 45 minutes. Therefore more focus of price strategy should be in this price range.

b. What does this mean for inventory availability?

Like I mentioned in previous question the mean is about 13.71 minutes and inventory mean is the Standard Deviation is about 14 minutes, so there is availability of a bike every 14 minutes on an average. Thus inventory is available about every 14minutes on an average.

Q. A business manager wants to reallocate the $5M marketing budget using a gender segmentation strategy. Specifically, the manager is asking you to create two models:

a. A model that use % of male vs females in the dataset

```
male<- sum(gender==1)
male
1] 15961
female<- sum(gender==2)
female
1] 4889
total_population<-male+female
total_population
1] 20850
male_percentage<- (male/total_population)*100
male_percentage
1] 76.55156
female_percentage<- (female/total_population)*100
female_percentage
1] 23.44844
```

MODEL A: This model suggests that of the $5 million marketing budget 77.55% is allocated to male population and 23.44% is allocated to female population.

b. A model based on average trip duration by gender

i. For Male Trip Average in Minutes

```
maletrip_avg<- mean(tripduration_min[gender==1])
maletrip_avg
1] 13.11829
```

ii. For Female Trip Average in Minutes

```
femaletrip_avg<- mean(tripduration_min[gender==2])
femaletrip_avg
1] 15.63737
```

iii. Total Trip Average and Percentage for both Gender

```
totaltrips_avg<- maletrip_avg + femaletrip_avg
totaltrips_avg
1] 28.75566
malepercent_avg<-(maletrip_avg/totaltrips_avg)*100
malepercent_avg
1] 45.61985
femalepercent_avg<-(femaletrip_avg/totaltrips_avg)*100
femalepercent_avg
1] 54.38015
```

MODEL B: This model suggests that of the $5 million marketing budget 45.61% is allocated to male population and 54.38% is allocated to female population.

Conclusion: On comparing Model A distinguishes between male and female population. And our focus should be on male population on who account for 77.55% of the total $5 million marketing budget.

Part 2: Teach Me Something.

Q. Write a couple of sentences about what your dataset contains (column names, types) and why you chose the dataset.

Q. Teach me one thing about your dataset

I chose the dataset of University Rankings which explains about the ranking of University according to quality of education, quality of faculty, alumni employment, publications etc.
Also I was curious to find out NYU's overall rank (23 is not bad).
Source: https://www.kaggle.com/mylesoneill/world-university-rankings

```
> dim(data)
[1] 2200   14
```

We can see there are 2200 observations and 14 variables.

```
> getwd()
[1] "/Users/SHARANG/Desktop/DATA"
> setwd("/Users/SHARANG/Desktop/DATA")
> data <- read.csv("~/Desktop/DATA/cwurData.csv")
>   View(data)
> summary(data)
   world_rank                                institution            country    national_rank
 Min.   :   1.0   Arizona State University          :   4   USA           :573   Min.   :  1.00
 1st Qu.: 175.8   Boston University                 :   4   China         :167   1st Qu.:  6.00
 Median : 450.5   Brown University                 :   4   Japan         :159   Median : 21.00
 Mean   : 459.6   California Institute of Technology:   4   United Kingdom:144   Mean   : 40.28
 3rd Qu.: 725.2   Carnegie Mellon University       :   4   Germany       :115   3rd Qu.: 49.00
 Max.   :1000.0   Columbia University              :   4   France        :109   Max.   :229.00
                  (Other)                         :2176   (Other)       :933
 quality_of_education alumni_employment quality_of_faculty  publications      influence       citations
 Min.   :  1.0        Min.   :  1.0     Min.   :  1.0      Min.   :   1.0   Min.   :  1.0   Min.   :  1.0
 1st Qu.:175.8        1st Qu.:175.8     1st Qu.:175.8      1st Qu.: 175.8   1st Qu.:175.8   1st Qu.:161.0
 Median :355.0        Median :450.5     Median :210.0      Median : 450.5   Median :450.5   Median :406.0
 Mean   :275.1        Mean   :357.1     Mean   :178.9      Mean   : 459.9   Mean   :459.8   Mean   :413.4
 3rd Qu.:367.0        3rd Qu.:478.0     3rd Qu.:218.0      3rd Qu.: 725.0   3rd Qu.:725.2   3rd Qu.:645.0
 Max.   :367.0        Max.   :567.0     Max.   :218.0      Max.   :1000.0   Max.   :991.0   Max.   :812.0

  broad_impact       patents          score             year
 Min.   :   1.0   Min.   :  1.0   Min.   : 43.36   Min.   :2012
 1st Qu.: 250.5   1st Qu.:170.8   1st Qu.: 44.46   1st Qu.:2014
 Median : 496.0   Median :426.0   Median : 45.10   Median :2014
 Mean   : 496.7   Mean   :433.3   Mean   : 47.80   Mean   :2014
 3rd Qu.: 741.0   3rd Qu.:714.2   3rd Qu.: 47.55   3rd Qu.:2015
 Max.   :1000.0   Max.   :871.0   Max.   :100.00   Max.   :2015
 NA's   :200
```

We can see the mean, median and max value for all the entities in the dataset.

Q. Finally, what is the business application of the findings and dataset. What possibilities do you have now as a business manager?

This data set can help students to search a University which is a good fit for them. Students can make more informed decision about which university to attend.

As the world is progressing at an exceptional rate with new technologies developing on day to day basis. People have realized the importance of education. Even for menial jobs companies are expecting employers to have understanding of software and understand data over a computer. University education helps the employers to understand the values, education and environment the student has been exposed to. This data set helps the employers to differentiate between potential employees from various universities.