ASSIGNMENT 1 – Exploration of R

In this assignment I will be performing an exploratory analysis of the **countyComplete** dataset in R language. I will be selecting 3 variables from the dataset and performing some basic descriptive statistics on them. We first import and summarize the structure of the dataset into R. We do this with the following command.

county = read.csv("countyComplete.csv")
str(county)

```
data.frame':   3143 obs. of  53 variables:
 $ state                              : Factor w/ 51 levels "Alabama","
Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ name                               : Factor w/ 1877 levels "Abbevill
e County",..: 83 90 101 151 166 227 237 250 298 320 ...
 $ FIPS                               : int  1001 1003 1005 1007 1009 1
011 1013 1015 1017 1019 ...
 $ pop2010                            : int  54571 182265 27457 22915 5
7322 10914 20947 118572 34215 25989 ...
 $ pop2000                            : int  43671 140415 29038 20826 5
1024 11714 21399 112249 36583 23988 ...
 $ age_under_5                        : num  6.6 6.1 6.2 6 6.3 6.8 6.5
6.1 5.7 5.3 ...
```

summary(county)

```
state                     name            FIPS          pop2010
 Texas   : 254   Washington County:  30   Min.   : 1001   Min.   :     82
 Georgia : 159   Jefferson County :  25   1st Qu.:18178   1st Qu.:  11104
 Virginia: 134   Franklin County  :  24   Median :29177   Median :  25857
 Kentucky: 120   Jackson County   :  23   Mean   :30390   Mean   :  98233
 Missouri: 115   Lincoln County   :  23   3rd Qu.:45082   3rd Qu.:  66699
 Kansas  : 105   Madison County   :  19   Max.   :56045   Max.   :9818605
 (Other) :2256   (Other)          :2999
```

The above shows only the partial output of the str() and summary() function which are used to briefly describe the structure of the dataset. Here we can see that the dataset contains 3143 observations (rows) across 53 variables (columns). It also shows the type of value each column holds such as categorical value, numeric value, integer value. From the summary() function we can get vital information of the variables according to their datatypes.

Next, we will be choosing any three variables from the dataset and provide an exploratory analysis by calculating the measures of central tendency and dispersion. We choose **per_capita_income**, **bachelors** and **mean_work_travel** as the three variables.

Part 1) We start with summarizing the variable **per_capita_income** using the R function **summary()**.

pci = county$per_capita_income
summary(pci)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7772   19030   21773   22505   24814   64381
```

We then calculate the measures of central tendencies for the variable **per_capital_income**.

(1) Mean

mean(pci)

```
[1] 22504.7
```

(2) Median

median(pci)

```
[1] 21773
```

Now, we calculate the measures of dispersion of the variable.

(1) Standard deviation

sd(pci)

```
[1] 5408.668
```

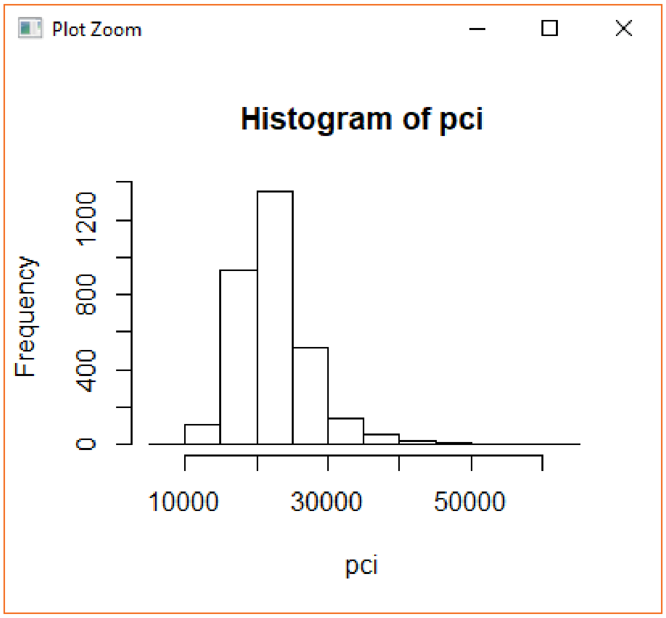
(2) Variance

var(pci)

```
[1] 29253692
```

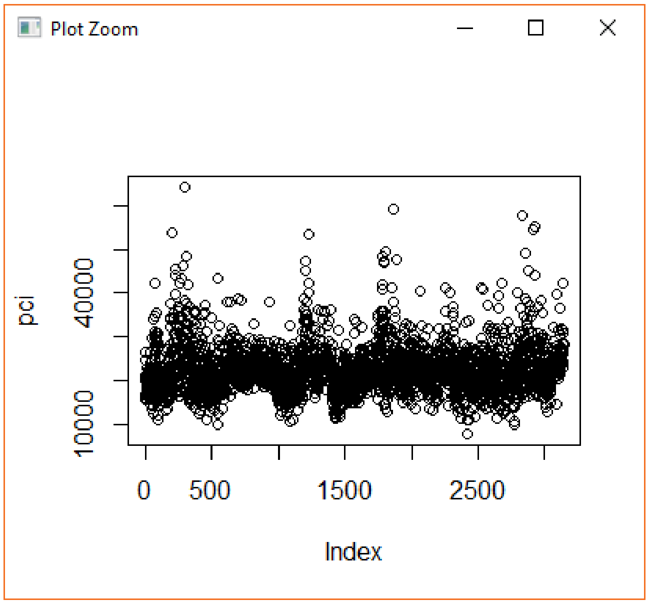
We can generate a histogram to determine the frequency of each values of the variable **per_capital_income** using the code below.

hist(pci)



We now generate the graphical representation of the entire values for the variable **per_capital_income.**

plot(pci)



Part 2) We start with summarizing the variable **bachelors** using the R function **summary()**.

bach = county$bachelors
summary(bach)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   3.70   13.10   16.90   19.03   22.60   71.00 
```

We then calculate the measures of central tendencies for the variable **bachelors**.

(1) Mean

mean(bach)

```
[1] 19.03376
```

(2) Median

median(bach)

```
[1] 16.9
```

Now, we calculate the measures of dispersion for the values of the variable.

(1) Standard deviation

sd(bach)

```
[1] 8.663063
```

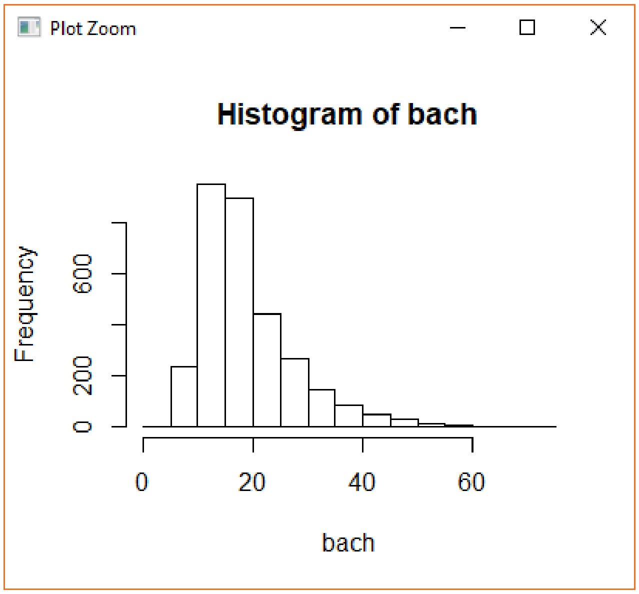
(2) Variance

```
var(bach)
```

```
[1] 75.04865
```

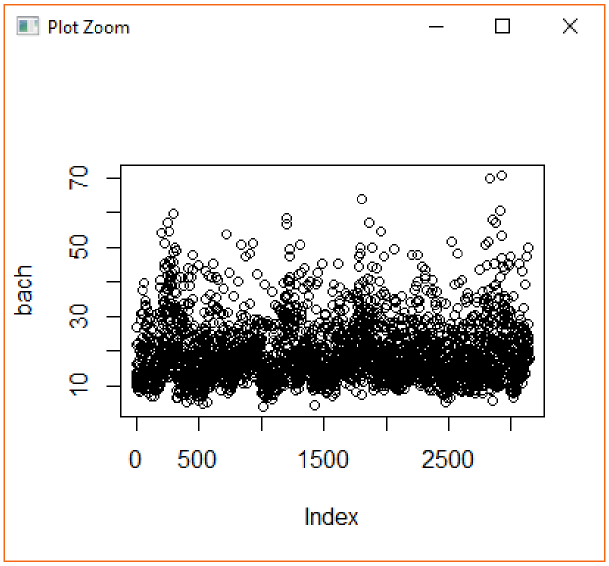
We can generate a histogram to determine the frequency of each values of the variable **bachelors** using the code below.

```
hist(bach)
```



We now generate the graphical representation of the entire values for the variable **bachelors.**

```
plot(bach)
```



Part 3) We start with summarizing the variable **mean_work_travel** using the R function **summary()**.

```
mwt = county$mean_work_travel
summary(mwt)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   4.30   19.00   22.40   22.73   26.10   44.20
```

We then calculate the measures of central tendencies for the variable **mean_work_travel**.

(1) Mean

```
mean(mwt)
```

```
[1] 22.72558
```

(2) Median

```
median(mwt)
```

```
[1] 22.4
```

Now, we calculate the measures of dispersion for the values of the variable.

(1) Standard deviation

```
sd(mwt)
```

```
[1] 5.514159
```

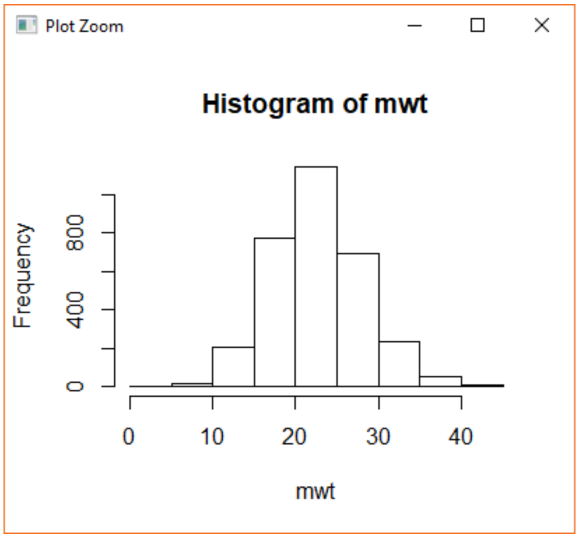
(2) Variance

```
var(mwt)
```

```
[1] 30.40595
```

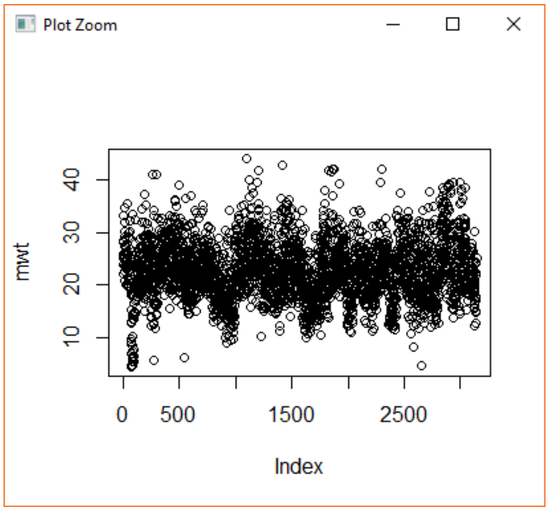
We can generate a histogram to determine the frequency of each values of the variable **mean_work_travel** using the code below.

```
hist(mwt)
```



We now generate the graphical representation of the entire values for the variable **mean_work_travel.**

plot(mwt)



Lastly, we find the relationship between the variables **bachelors** and **per_capita_income**. We do so by using the built-in **cor()** function. We use pearson's product moment correlation. We find the correlation using the code below.

cor(pci, bach, method = "pearson")

```
[1] 0.7924464
```

As the correlation coefficient of bachelors and per capita income is 0.792 which is close to 1, we can conclude that the variables are positively linearly related.