Ain Shams University
Faculty of Computer & Information Sciences
Computer Science Department

Data Science Project Documentation

Project Idea:

"Household Income Analysis"

Team Members:

1st Team Member Name:

Kareem Saeed Ragab

1st Team Member ID:

2018170282

1st Team Member Department:

Computer Science

2nd Team Member Name:

Nada El Sayed Anies

2nd Team Member ID:

2018170430

2nd Team Member Department:

Computer Science

Part #1: "Review of Big Data Analytic Methods"

1- Step # 1: Retrieve and Clean Up Data using R

1.1- Analyze the data:

- The screenshot from the R code.

```
      zcta           sex                meanage        meaneducation    meanemployment
 Min.   :  601   Length:64076       Min.   :  0.00   Min.   : 0.00   Min.   :0.000
 1st Qu.:27305   Class :character   1st Qu.: 36.65   1st Qu.:11.91   1st Qu.:1.542
 Median :49909   Mode  :character   Median : 39.30   Median :12.46   Median :1.813
 Mean   :49801                      Mean   : 39.68   Mean   :12.53   Mean   :1.787
 3rd Qu.:72007                      3rd Qu.: 42.28   3rd Qu.:13.11   3rd Qu.:2.077
 Max.   :99950                      Max.   :137.08   Max.   :19.00   Max.   :3.000
 meanhouseholdincome
 Min.   :     0
 1st Qu.: 37642
 Median : 44163
 Mean   : 48245
 3rd Qu.: 54373
 Max.   :250000
```

- Columns names: zcta , sex , meanage , meaneducation , meanemployment

1.2- Number of rows in the zeta table:

- 64076

1.3- Are there any duplicated rows of data in the zeta table?

- There is no duplicated rows of data.

1.4- According to 1.3, no duplicated rows.

1.5- Saved the table, Included in the R code.

2- Step # 2: Data Analysis in R

2.1- Loaded the data, Included in the R code.

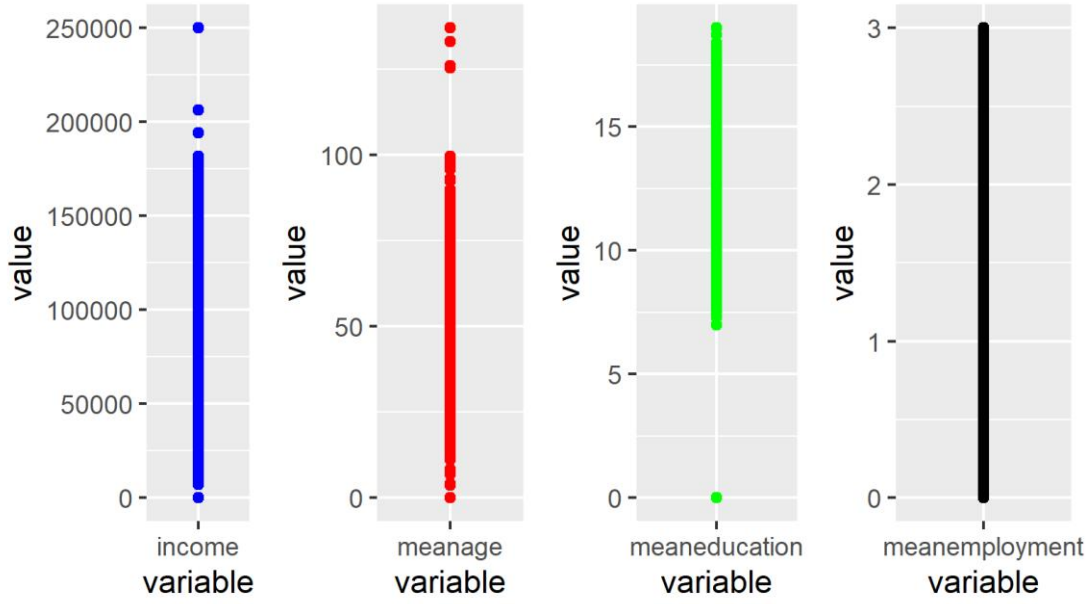
2.2- Changed the column names, Included in the R code.

2.3- What are the mean and median average incomes?

- Incomes column mean: 48245
- Incomes column median: 44163

2.4- Plot a scatter plot of the data?

- The scatter plot from the R code.



- **Do you have any outlier values?**
  YES
- **What are these outlier values?**
  In incomes and meanage columns in range more than 200,000 (>200,000) and less than 7,000 (<7,000) as shown in the plot.

2.5- Deleting the outlier values, Included in the R code.

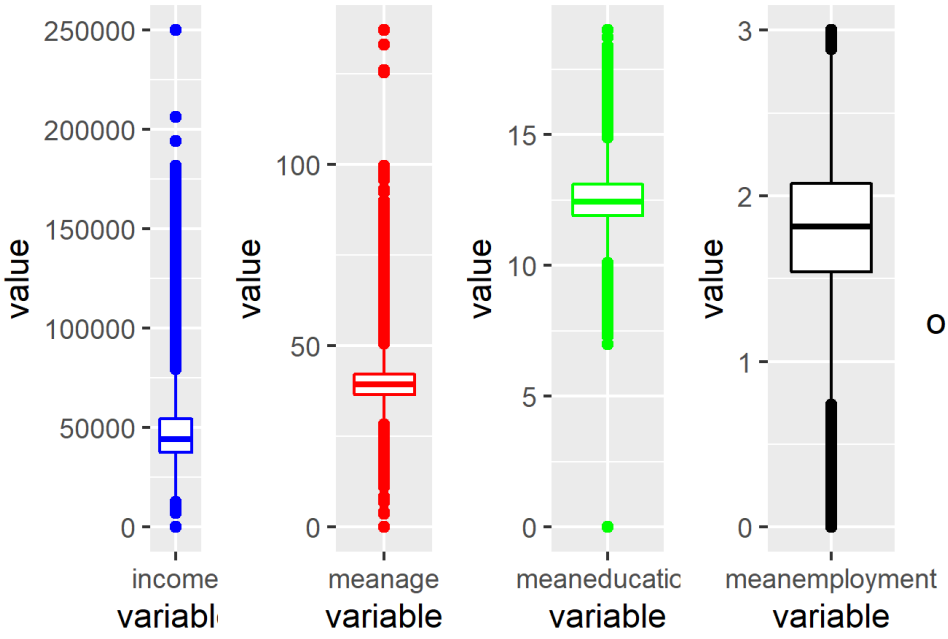
2.6- The mean of the data:

- The mean before deleting the outlier values: 48245.24
- The mean after deleting the outlier values: 48464.95

3. Step # 3: Visualize your data

3.1- Create a box plot of the data?

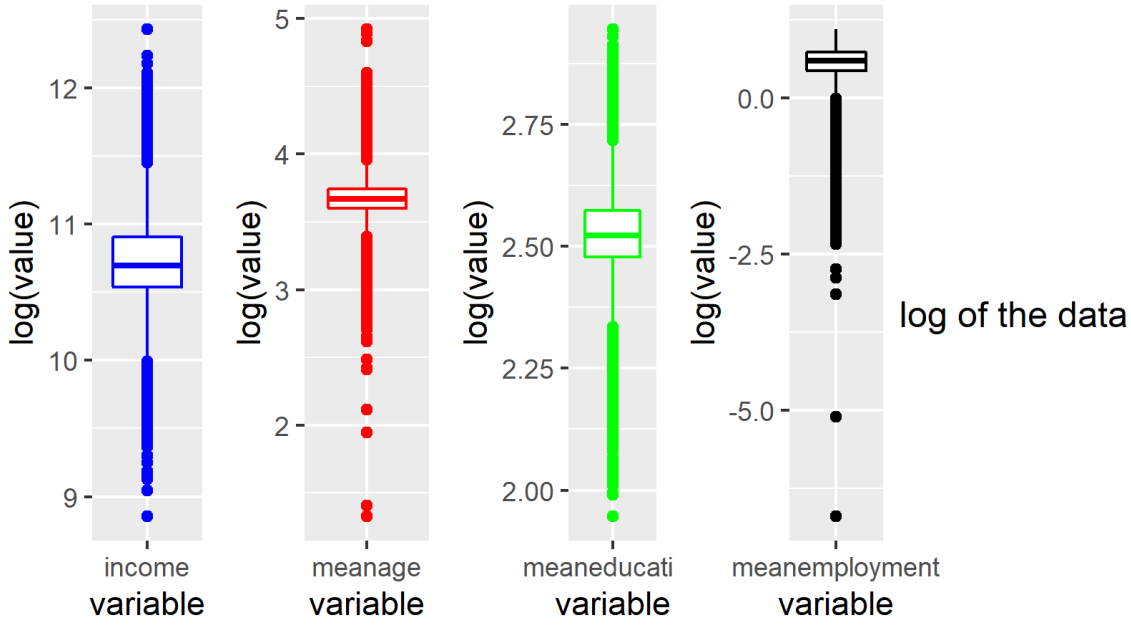
- The screenshot from the R code.



original data

3.2- Create a box plot of the data with the (log scale)?

- The screenshot from the R code.



log of the data

3.3- What can you conclude from this data analysis/visualization?

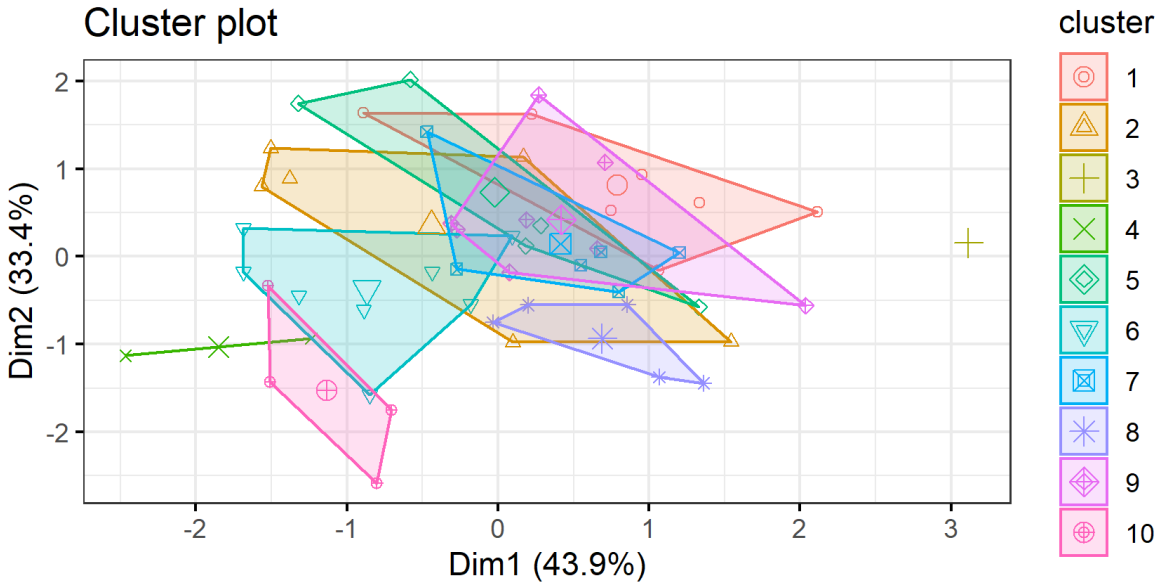
- In the end of the step #1 of data analysis/visualization, it's important to do preprocessing for the data to study and clean it from any duplicated data and outlier values to help you in the next process such as (K-means clustering).

Part #2: "Advanced Analytics/Methods (K-means)"

1- The table is created, Included in the R code

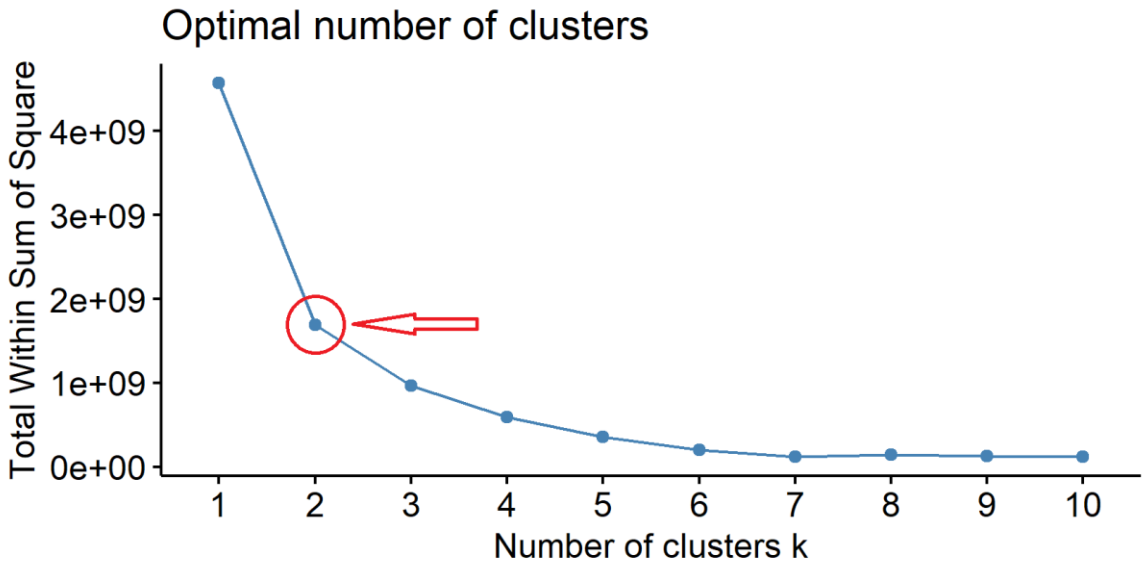
2- Cluster the data using K-means and plot the result?

- The plot screenshot from the R code, **Note** K = 10 :



3- Determine the reasonable value of k using (elbow)?

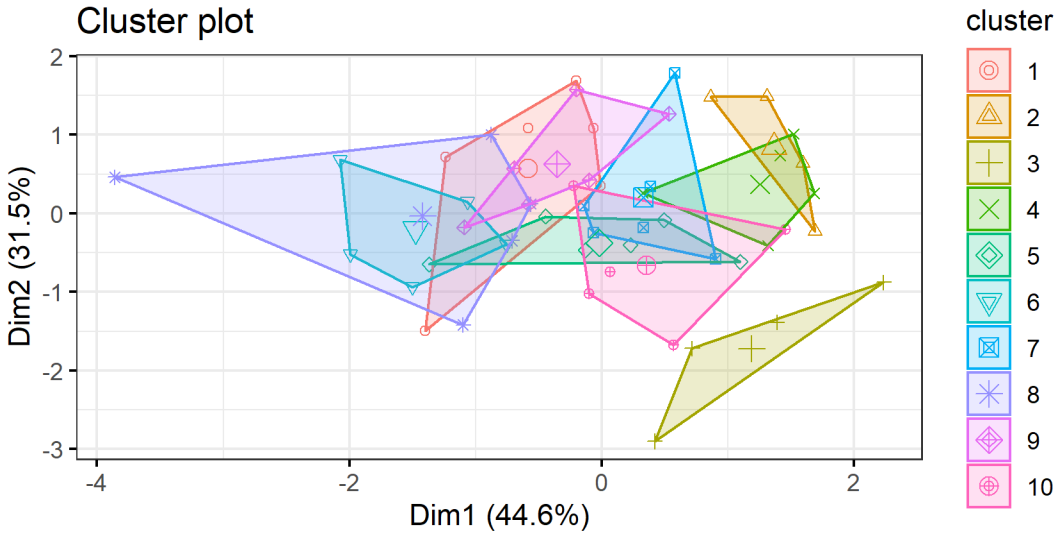
- Within cluster sum of squares plot



- As shown in the elbow plot above the best K value in the clustering using K-means equal 2.

4- Cluster the transformed data to (log10 scale) using K-means and plot the result?

- The plot screenshot from the R code, **Note** <u>K = 10</u> :

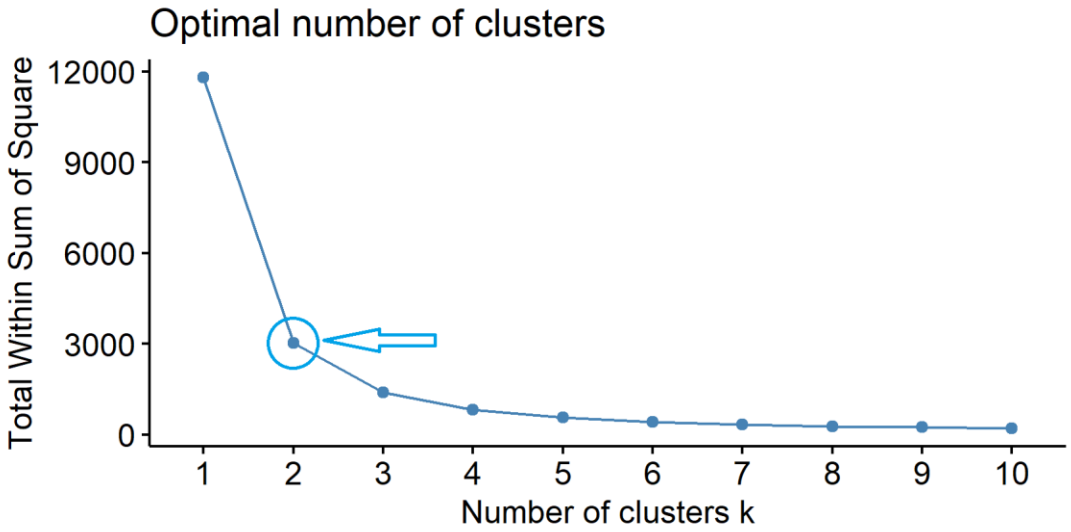


- The clustering grouping changed with all the points grouped to a cluster. Before scaling the data, there are points some of them take one cluster with its own, so it will lead to overfitting.

- **Why?** Because the scaling the data with log10 distributes the data with each other and distance between them became less, so the K-means algorithm work again with the updated data.

5- The elbow plot after scaling data with log10

- Within cluster sum of squares plot



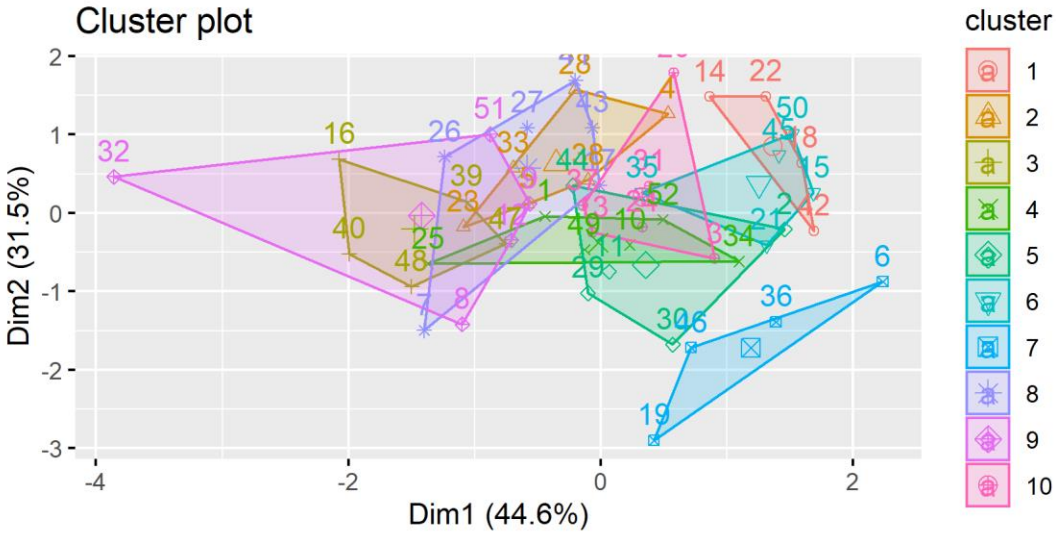
- Still the best choice of K equal 2. According to the elbow plot above.

6- Have you observed an outlier in the data?

YES

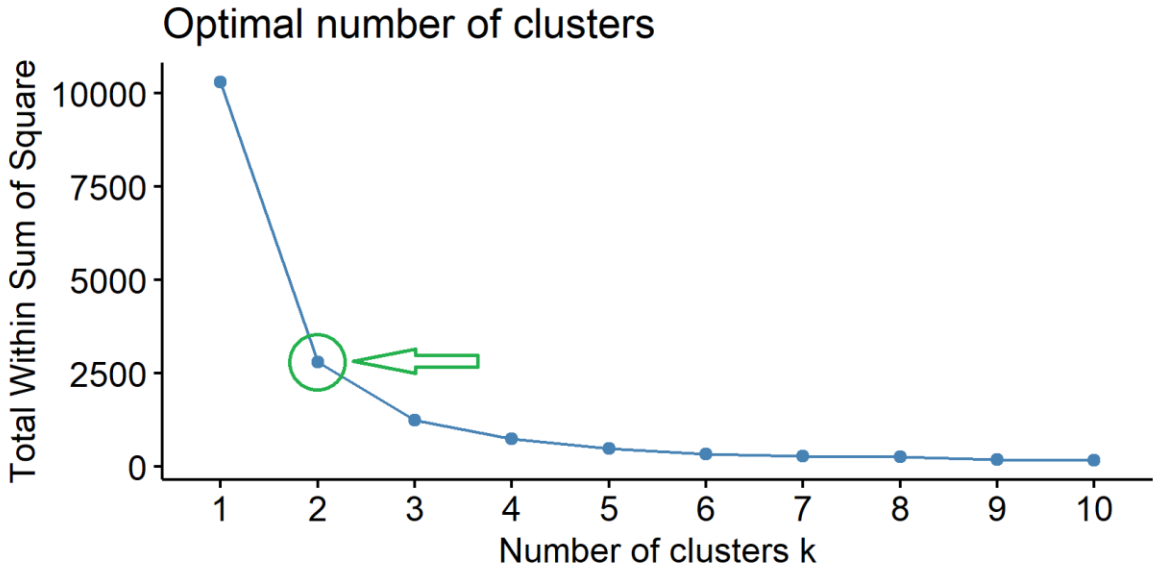
6.1- Calculated the outlier values by two ways:

- **Theoretically:** to get the rows of outlier values from the dataset.
  The outlier values exists in the rows: 32, 20, 41, 7, 17
- **Graphically**



6.2- The elbow plot after removing the outlier values

- Within cluster sum of squares plot



- Still the best choice of K equal 2. According to the elbow plot above.

Addition

- **Cluster the data with the best choice of <u>K = 2</u> and plot the result?**