**Assignment 3**

**Author:** Vidhya Venugopal, 8902970

**Course:** PROG8430

1. Data Transformation

1.1 # Standardize all of the variables using either of the two functions demonstrated in class. Describe why you chose the method you did:

I have used Min-max normalization. This method scales the variables to a specific range, typically between 0 and 1. It is useful when you want to preserve the original distribution and maintain the relationships between variables. Min-max normalization is commonly used when working with algorithms that rely on similarities between variables, such as clustering or nearest-neighbour methods.

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, font, screenshot, electric blue

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated with low confidence

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, screenshot, font, number

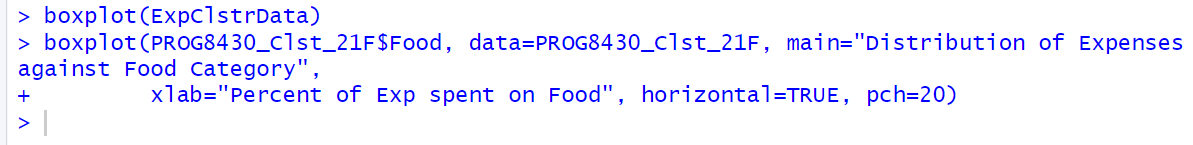
Description automatically generated

A picture containing text, screenshot, font, number

Description automatically generated

1. Descriptive Data Analysis

2.1 # Create graphical summaries of the data (as demonstrated in class: boxplots or histograms) and comment on any observations you make



A picture containing text, screenshot, rectangle, line

Description automatically generated

A picture containing text, screenshot, display, line

Description automatically generated

A picture containing text, screenshot, display, diagram

Description automatically generated

A picture containing text, screenshot, display, line

Description automatically generated

A picture containing text, screenshot, line, rectangle

Description automatically generated

A picture containing text, screenshot, display, diagram

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

Based on the above box plot graph, it is been observed that the data which is the expenses spent in the Food and Work category is normally distributed whereas the data for other categories are Negative skew and not normally distributed.

1. Clustering

3.1 # Create segmentation/cluster schemes for k=3,4,5,6,7

#K = 3

A picture containing text, font, screenshot, white

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, pattern, symmetry

Description automatically generated

A picture containing text, font, screenshot, paper

Description automatically generated

A picture containing text, screenshot, font, line

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence



#K = 4

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, paper, document, pattern

Description automatically generated

A picture containing text, font, paper, black and white

Description automatically generated

A picture containing text, screenshot, font, line

Description automatically generated

A screenshot of a computer screen

Description automatically generated with low confidence



#K = 5

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, pattern, black and white

Description automatically generated

A picture containing text, font, paper, black and white

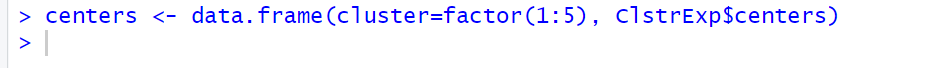
Description automatically generated

A picture containing text, screenshot, font, algebra

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence



#K = 6

A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, pattern, screenshot

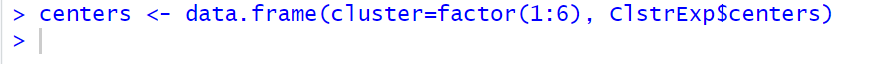
Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence



#K = 7

A screenshot of a computer

Description automatically generated with medium confidence

A grid of numbers on a white background

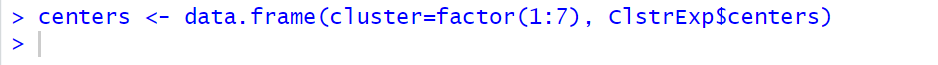
Description automatically generated with low confidence

A screenshot of a computer

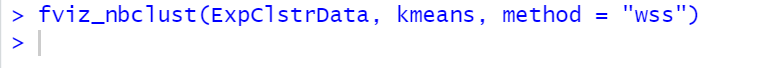
Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence



3.2 # Create the WSS plots as demonstrated in class and select a suitable k value based on the “elbow”.



A picture containing text, line, plot, screenshot

Description automatically generated

1. Evaluation of Clusters

2.1 # Based on the “k” chosen above, create a scatter plot showing the clusters and colour-coded datapoints for each of “k-1”, “k”, “k+1”.

Based on the above chart, the elbow value is 3. Therefore, K value = 2:

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, black and white, paper, pattern

Description automatically generated

A picture containing black and white, pattern, text

Description automatically generated

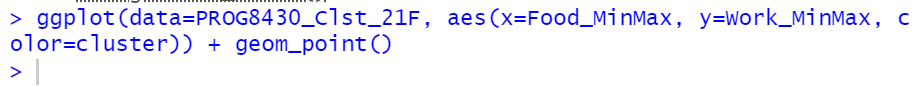
A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence





A picture containing screenshot, text, diagram, plot

Description automatically generated

K value = 3:

A screenshot of a computer

Description automatically generated with medium confidence

A grid of numbers

Description automatically generated with low confidence

A picture containing text, pattern, black and white

Description automatically generated

A screenshot of a computer code

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A picture containing text, font, screenshot, electric blue

Description automatically generated

A picture containing screenshot, text, diagram, plot

Description automatically generated

K value = 4:

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, pattern, black and white

Description automatically generated

A picture containing text, black and white, pattern, paper

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, screenshot, plot, diagram

Description automatically generated

When K = 4 the value with in cluster sum of square by cluster is 85.4 % which is higher than the remaining K =3 is 76.3% & K = 2 is 57.1% clusters

4.3 # Create summary tables for the segmentation/clustering scheme.

A picture containing text, screenshot, font

Description automatically generated

4.4 # Create suitable descriptive names for each cluster

The sizes of the clusters as cluster 1 to 4 = 146, 195, 102, 157. These are the means.

Food - The means for the Food column are different 0.133, 0.150, 0.101 & 0.200 for different clusters and cluster 4 Food mean rate is higher than the remaining clusters. Therefore, the patter of expenses spent on Food is identified.

Entr - The means for the Entr column are different 0.0990, 0.0202, 0.0976 & 0.032 for different clusters and cluster 1 Entr mean rate is higher than the remaining clusters. Therefore, the patter of expenses spent on Entr is identified.

Educ - The means for the Educ column are different 0.124, 0.209, 0.0262 & 0.150 for different clusters and cluster 2 Educ mean rate is higher than the remaining clusters. Therefore, the patter of expenses spent on Educ is identified.

Educ

Tran - The means for the Tran column are different 0.167, 0.0213, 0.223 & 0.0520 for different clusters and cluster 3 Tran mean rate is higher than the remaining clusters. Therefore, the patter of expenses spent on Tran is identified.

Work - The means for the Work column are different 0.0361, 0.0808, 0.0807 & 0.0403 for different clusters and cluster 2 Work mean rate is higher than the remaining clusters. Therefore, the patter of expenses spent on Work is identified.

Hous - The means for the Hous column are different 0.390, 0.498, 0.403 & 0.495 for different clusters and cluster 2 Hous mean rate is higher than the remaining clusters. Therefore, the patter of expenses spent on Hous is identified.

Othr - The means for the Othr column are different 0.0510, 0.0203, 0.0686 & 0.0306 for different clusters and cluster 3 Othr mean rate is higher than the remaining clusters. Therefore, the patter of expenses spent on Othr is identified.

4.5 # Suggest possible uses for this clustering scheme.

These clustering schemes can be used to identify the outliers in datasets. By clustering normal data points together, any data points that do not fit well within a cluster can be flagged as potential anomalies. Therefore, considering our expense data based on the outliers of data, it can be identified that the corresponding person is not doing the transaction and immediately alert the corresponding person as there is difference in the expense pattern.