**Assignment 3**

**Author:** Hamna Ashraf, Student # 8826836

**Course:** PROG8431

**Question 1**

1. **Data Transformation (1 point):** 
   1. Standardize all the variables using either of the two normalized functions demonstrated in class. Describe why you chose the method you did.

Solution:

For standardization, I used the min-max normalization technique. It is because we learnt that min-max should be used when the data is tightly clustered and when we’re sure that there are no outliers in our data.

**Before standardizing:**

I plotted the histograms of all variables and checked the data. There were no outliers, and the distribution was within a narrow range.

A graph of a bar graph

Description automatically generated A graph of food and its bar

Description automatically generated with medium confidence

These are two of the graphs, for my data. The rest of them can be viewed in R file, having similar scale and distribution with no outliers.

A group of numbers on a white background

Description automatically generated

**After standardizing data looked like this:**

A screenshot of a computer code

Description automatically generated

**2.Descriptive data analysis (1 point)**

**2.1** Create graphical summaries of the data (boxplots or histograms) and comment on any observations you make.

A group of graph of different sizes

Description automatically generated with medium confidence

Observation:

* The scale changed to 0-1.
* The graph ‘hous’ seems a bit right skewed.
* The graph ‘othr’ seem a bit left skewed.

However, this is not due to the presence of outliers.

**3. Clustering the data (4 points):**

**3.1** Using the K-Means procedure, create clusters with k=3,4,5,6,7. You will be using only two variables as your centroids (**Food and Work**).

A screenshot of a computer

Description automatically generated

The clusters are successfully created without any errors having 2 centroids.

**3.2** **Create the WSS plots as demonstrated in class and select a suitable k value based on the “elbow”. [NOTE – you can create the chart in Excel].**

The library ‘factoextra’ is used in R to create the wss plot and we can determine the optimal cluster based on the elbow shape in graph.

A graph with a line

Description automatically generated

k = 3

There is no drastic change seen in tot\_ss value after cluster 3. Therefore, **3** is the **optimal cluster**.

**4.Evaluation of clusters (7 points)**

**4.1** Based on the “k” chosen above, create scatter plot showing the clusters and colour-coded datapoints for each of “k-1”, “k”, “k+1”. For example, if you think the “elbow” is at k=4 create the charts for k=3, k=4 and k=5. (2 points)

We choose the k = 3, based on the plot we saw in the previous question,

**Scatterplot k = 2:**

**A screen shot of a graph

Description automatically generated**

**c2** means 2 clusters.

**Scatterplot k = 3:**

**A screen shot of a graph

Description automatically generated**

**Scatterplot k = 4:**

A screen shot of a graph

Description automatically generated

**4.2 Based on the WSS plot (3.2) and the charts (4.1), choose one set of clusters that best describes the data. (1 point)**

I think that clusters *k = 4* best separates into different categories among the other plots we have against other clusters. We can see clearly that each cluster is far away from the other and the data is tightly clustered within each cluster.

**4.3 Based on the step 4.2, create summary tables for the segmentation/clustering scheme. (2 points)**

Based on the previous step we selected clusters 4 to be suitable for our data.

The summary table is as follows:

A screenshot of a computer

Description automatically generated

**4.4 Create suitable descriptive names for each cluster.**

A screenshot of a computer

Description automatically generated

Based on the values for each cluster, I have given meaningful values.

**4.5 Suggest possible uses for this clustering scheme**

* Consumer Behavior Analysis

Clustering is very useful to understand the customer behavior

* Personalized Recommendations:

Based on the spending patterns and likes and dislikes of a customer, clustering schemes can be used give personalized recommendations.