**Assignment 3 – Clustering**

Name: Mohammed Adeen Shaik

Student ID: 8969152

**TASK 1: Data Transformation and Descriptive Analysis**

All the numbers in the data frame show the annual spending of each product / variable

* 1. Renaming all variables with initials of First **(M)** and Last**(S)** names. **(MS)**
  2. Create graphical summaries of the data

**Boxplot:**

A screenshot of a computer

Description automatically generated

The data shows some variability, with the lower whisker extending to about 4000 and the upper whisker reaching up to 16000.

A screenshot of a computer

Description automatically generated

Multiple data points (outliers) detected on spending of grocery products that are too high. Some values might be 0 as lower whisker extends to 0

A screenshot of a computer

Description automatically generated

The lower whisker extends to about 2500 and the upper whisker reaches up to 12500, This indicates the spread in data is wider.

A screenshot of a computer

Description automatically generated

Some values might be 0 since the lower whisker extends to 0

A screenshot of a computer

Description automatically generated

Some values might be 0 since the lower whisker extends to 0

A screenshot of a computer

Description automatically generated

Multiple outliers were detected on spending of cleaning products that are too high. Some values might be 0 since the lower whisker extends to 0

**Histograms:**

**A graph of a bar chart

Description automatically generated with medium confidence**

**A graph on a screen

Description automatically generated**

**A graph on a computer screen

Description automatically generated**

**A graph of a graph

Description automatically generated with medium confidence**

**A screenshot of a graph

Description automatically generated**

**A graph on a screen

Description automatically generated**

* 1. Standardize all the variables:

A screenshot of a computer program

Description automatically generated

I used the **scale()** function to standardize all the variables to transform the data such that each variable has a mean of zero and a standard deviation of one. After looking at the data frame, I realized each variable in the data frame have the same value that is the annual spending of the product that variable represents. So, in order to for ensure that each variable contributes equally to the analysis, especially in clustering algorithms like k-means, I have used this function.

**Task 2: Clustering ( Milk and Frozen products )**

3.1: **A screen shot of a graph

Description automatically generated**

The optimal value I have taken is k = 4

**Task 3: Evaluation of Clusters**

**4.1:**

K =4

A screen shot of a graph

Description automatically generated

k-1 = 3

A screen shot of a graph

Description automatically generated

K+1 = 5

A screen shot of a graph

Description automatically generated

4.2:

The set of clusters that best describes the data is k =4, It is the most suitable set that shows the spending on Milk products and Frozen products on a high and low basis as a pair. You will understand when you see the final output.

4.3:

Create summary tables for the clustering scheme selected in 4.2

A screenshot of a computer

Description automatically generated

4.4: Naming the clusters

A screen shot of a graph

Description automatically generated

As I said, I chose this set of clusters ( k =4 ) because it describes the High/Low spending of the products. I have named them as such:

1. **Low Milk/Low frozen:** data points plotted negatively on both axes(Milk and Frozen)
2. **High Milk/Low frozen:** data points plotted positively on the x(Milk) axis and negatively on the y(frozen) axis
3. **High Frozen/Low Milk:** data points plotted positively on the y(frozen) axis and negatively on the x(Milk) axis
4. **High Frozen/high Milk**: data points plotted positively on the y(frozen) axis and positively on the x( Milk) axis
5. Suggestions on uses of the clustering scheme.

With this clustering scheme, we can suggest many changes to the **wholesale food distributor to improvise their selling strategy** based on the information from the chart and the summary table.

such suggestions can be:

 **Targeted Marketing**: Customize campaigns based on cluster behavior, such as offering frozen product discounts to the "Low Milk / High Frozen" cluster.

 **Stock Optimization**: Adjust inventory levels based on cluster demand patterns to prevent overstock and stockouts.

 **Personalized Recommendations**: Provide product suggestions tailored to each cluster’s preferences, enhancing customer satisfaction.

** Dynamic Pricing**: Implement flexible pricing strategies that reflect the demand patterns of different clusters.