Question 1

1a)

Results of PCA:

A picture containing text, screenshot, font, keyboard

Description automatically generated

A picture containing text, screenshot, font

Description automatically generated

A graph with a line

Description automatically generated with low confidence

* Using the 3 rules given to us when attempting to extract principal components (PC), we shall take the first two. This is as the elbow of the scree plot is at the 3rd PC, only the first 2 PCs have a eigenvalue >1, and they explain 74% of the total variance.

Loading Plot:

A picture containing line, diagram, text, screenshot

Description automatically generated

Score Plot:

A picture containing text, screenshot, diagram, line

Description automatically generated

* From the graphs, we can see that PC1 measures when products have high sugar and carbohydrate content, while PC2 measures when products have high vitamin content.

1bi) The product is likely to be cheese. As the product is said to have the attributes “Low carbohydrates and sugar but high in other nutrients”, this would mean that is has a low PC2 load, but a high PC1 load. Hence, we place the product in the right quadrant. Next, when we refer to the scoreplot, we can see that there are many Cheese datapoints there. Hence, we can infer that the product is likely to be cheese.

1bii) The product is likely to be Cream or Milk. As the product is said to have the attributes “High carbohydrates and sugar but low in other nutrients”, this would allow me to infer that it has a high PC2 load, and a low PC1 load. This would place it in the upper quadrants, meaning that is likely to be Milk or Cream, especially after we refer to the scoreplot.

1c) Product 1: The product is likely to be Cheese.

1cii) Product 2: The product is likely to be Yogurt.

1d) At first, I expected the first 2 Principal Components to be able to explain most of the variance, and allow me to acutely identify the products passed on that alone. However, in actuality, I had to use one more PC (PC3) to allow me to identify certain products, namely, product 2.

Question 2

2a)

Results of PCA:

A picture containing text, screenshot, keyboard

Description automatically generated

A picture containing text, screenshot, font

Description automatically generated

A graph with a line

Description automatically generated with low confidence

* Using the screeplot, we can see that the elbow is at the second PC, hence we should only take the first PC.
* However, as not enough variability is shown in the first PC (62%), we shall also take the second PC, as they explain 79.2% of the variability combined. In addition, if we were to follow Kaiser’s rule, it would advise that we take both PC1 and PC2, as they both have a eigenvalue > 1.

A picture containing text, line, diagram, screenshot

Description automatically generated

A picture containing text, screenshot, diagram

Description automatically generated

* From the graphs, we can see that PC1 measures positively against all variables, except both “Highway-mpg” and “City-mpg”, which is measures negatively against. I
* We can also see that PC2 measures negatively against most variables, and measures positively with “Horsepower”, “Cylinders”, “Engine”, “Price”

2b) Compared to the PCA analysis done in Question 1, I can see that PCA is a lot more useful for clustering as compared to classification. One reason why I believe so, is due to the scoreplot generated for Question 2. It is particularly difficult to infer where a certain point is based on the second scoreplot, while it is much easier to using the clearly defined clusters in Question 1’s scoreplot.