Question 1

1a)

Steps

1. Carry out PCA on correlation matrix as variables have different magnitude.
2. Drop ‘Type’ and ‘Description’

A picture containing text, screenshot, font, number

Description automatically generated

1. Standardized the data.

A picture containing text, screenshot, font, white

Description automatically generated

1. PCA Samples using pca.transform()

A picture containing text, font, screenshot, white

Description automatically generated

1. PCA Results

A picture containing text, screenshot, font, white

Description automatically generated

1. Scree plot

A graph with a line

Description automatically generated with low confidence

1. Extract the 2 PCs

A close-up of numbers

Description automatically generated with low confidence

1. Loading plot

A picture containing text, screenshot, diagram, line

Description automatically generated

1. Obtain scores from original data that are not centred.

A screenshot of a calculator

Description automatically generated with medium confidence

1. Plot the Score Plot

A picture containing text, screenshot, diagram, plot

Description automatically generated

PCA on correlation matrix (data is standardized):

1. By Kaiser’s rule, extract the first 2 PCs where eigenvalues (3.78,2.17) are > 1.
2. First 2 PCs accounted for 74.4% of the total variance.
3. Scree plot shows elbow at PC4, suggesting that the 1st 3 PCs to extract.

Let’s extract the first 2 PCs only.

PC1 = 0.4672x1 + 0.2910x2 + 0.0722x3 + 0.0973x4 + 0.3685x5 + 0.3653x6 + 0.4428x7 + 0.4658x8

The loadings are all positive in sign. However, the loading on carbohydrates and sugar is quite small compared to the other loadings.

This PC seems to measure the overall nutritional information of the food products.

PC2 = -0.1304x1 -0.3270x2 + 0.6181x3 + 0.6013x4 -0.1793x5 + 0.2452x6 + 0.1731x7 – 0.1013x8

The loading on protein, fat, vitamin A and calcium is opposite in sign to the other loading.

This PC seems to measure a contrast of protein, fat, vitamin A and calcium against the remaining variables.

From the Score plot, Milk tends to score high on PC1 and Cream tend to score low on PC1.

Milk and Cream tend to score high on PC2, on the other hand, Cheese tend to score low on PC2.

1bi) Low carbohydrates and sugar but high in other nutrients: Cheese

Looking at score plot, a high PC1 score will indicate a product that is high in other nutrients and on the other hand, a low PC2 score will indicate a product that is low in carbohydrates and sugar.

1bii) High carbohydrates and sugar but low in other nutrients: Cream and Ice Cream

Looking at score plot, a high PC2 score will indicate a product that is high in carbohydrates and sugar, on the other hand, a low PC1 score will indicate a product that is low in other nutrients.

1c) Product 1

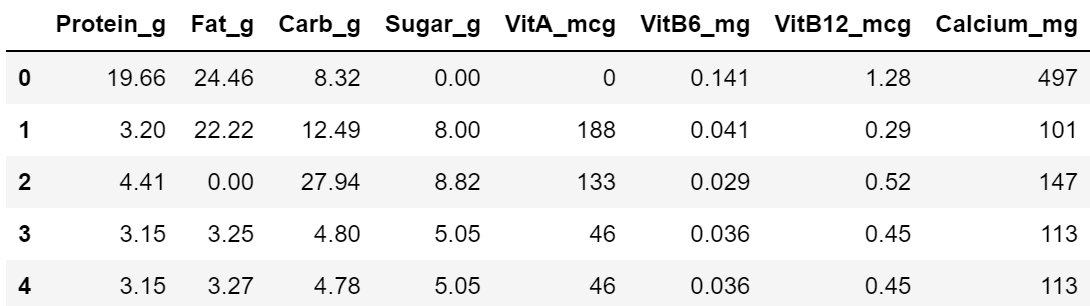
Steps

1. Insert Product 1 and drop ‘Description’.

A screenshot of a computer program

Description automatically generated with low confidence

1. Drop ‘Type’



1. Standardized the data.

A screenshot of a calculator

Description automatically generated with low confidence

1. Transform the recentred data.

A picture containing text, font, white, screenshot

Description automatically generated

1. Obtain PCA Results

A picture containing text, screenshot, font, number

Description automatically generated

1. Plot the Scree Plot

A graph with a line

Description automatically generated with low confidence

PCA on correlation matrix:

1. By Kaiser’s rule, extract the first 2 PCs where eigenvalues (3.77,2.17) are > 1.
2. First 3 PCs accounted for 85.2% of the total variance.
3. Scree plot shows elbow at PC4, suggesting that the 1st 3 PCs to extract.

Let’s extract first 3 PCs.

1. Extract the 3 PCs
2. Obtain scores from original data that are not centred.
3. Plot the Scree Plots to determine the type of product.

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, plot, diagram

Description automatically generated

A picture containing text, screenshot, line, plot

Description automatically generated

From the score plot, product 1 is ‘Unknown’, the data points around product 1 is mostly all Cheese. Hence, Product 1 is Cheese.

Product 2

A screenshot of a computer program

Description automatically generated with medium confidence

A picture containing text, screenshot, font, number

Description automatically generated

A graph with a line

Description automatically generated with low confidence

A picture containing screenshot, text, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing screenshot, text, line, diagram

Description automatically generated

PCA on correlation matrix:

1. By Kaiser’s rule, extract the first 2 PCs where eigenvalues (3.78,2.18) are > 1.
2. First 3 PCs accounted for 85.4% of the total variance.
3. Scree plot shows elbow at PC4, suggesting that the 1st 3 PCs to extract.

Let’s extract 3 PCs.

From the score plot, product 2 is ‘Unknown’, the data points around product 1 is mostly all Yogurt. Hence, Product 2 is Yogurt.

1d) In part (c), I used the score plots to determine what type a product is through observing where the data points clusters at. In part (a), I standardized the data, determine how many PCs to extract using scree plot, Kaiser’s rule, and the cumulative explained variance.

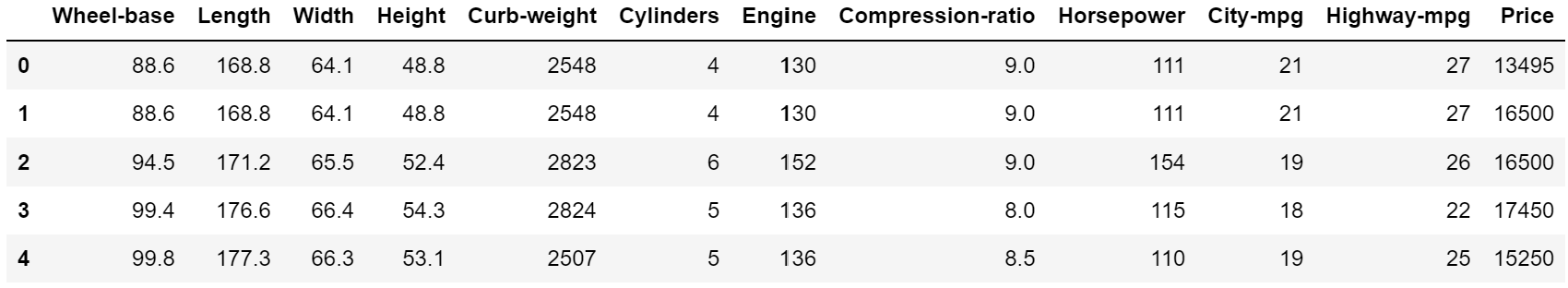
I expected principal components to tell me about the relationships between the different type of products. However, the more principal component is extracted, the more variability is captured which also means that determining the type of product will also be more accurate.

Question 2

2a)

Steps

1. Carry out PCA on correlation matrix as variables have different magnitude.
2. Drop ‘Brand’ and ‘Type’



1. Standardized the data.

A screenshot of a computer

Description automatically generated with low confidence

1. Transform the recentred data with pca.transform()

A picture containing text, font, white, receipt

Description automatically generated

1. Obtain the PCA Results

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

Description automatically generated

1. Plot the Scree Plot

A graph with a line

Description automatically generated with low confidence

PCA on correlation matrix:

1. By Kaiser’s rule, extract the first 3 PCs where eigenvalues (7.54,1.99,1.08) are > 1.
2. First 3 PCs accounted for 88.4% of the total variance.
3. Scree plot shows elbow at PC4, suggesting that the 1st 3 PCs to extract.

Let’s extract the first 3 PCs only.

1. Extract the first 3 PCs
2. Plot the Loading Plot

A picture containing text, screenshot, diagram, line

Description automatically generated

1. Obtain scores from original data that are not centred.

A picture containing text, screenshot, font, number

Description automatically generated

1. Plot the Score Plots

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

PC1 = 0.2780x1 + 0.3217x2 + 0.3217x3 + 0.1031x4 + 0.3492x5 + 0.2662x6 + 0.3307x7 + 0.0019x8 + 0.3077x9 – 0.3122x10 – 0.3221x11 + 0.3273x12

The loading on City Miles Per Gallon, Horsepower of Vehicle is opposite in sign to the other loading.

PC1 seems to measure a contrast of City Miles Per Gallon and Horsepower of Vehicle against the remaining variables.

PC2 = -0.3786x1 – 0.2438x2 – 0.1767x3 – 0.5350x4 – 0.0932x5 + 0.2186x6 + 0.1303x7 – 0.4954x8 + 0.2996x9 – 0.2065x10 – 0.1566x11 + 0.0634x12

The loading on Compression Ratio, City Miles Per Gallon and Horsepower of Vehicle is opposite in sign to the other loading.

PC2 seems to measure a contrast of every variable except Compression Ratio, City Miles Per Gallon and Horsepower of Vehicle against the Compression Ratio, City Miles Per Gallon and Horsepower of Vehicle

PC3 = -0.1227x1 - 0.1356x2 + 0.0469x3 – 0.3656x4 + 0.0432x5 + 0.3895x6 + 0.2701x7 + 0.6048x8 + 0.0507x9 + 0.3212x10 + 0.2939x11 + 0.2152x12]

The loading on Wheel-base and length is opposite in sign to the other loading.

PC3 seems to measure a contrast of Wheel-base and length against the remaining variables.

From the Score plot, Sedan scores high on PC1, PC2 and PC3

Wagon tends to score low on PC2 and PC3.

Hatchback tends to score low on PC1.

2b) It is easier to differentiate the type of product in question 1. However, As seen from the score plot at question 2, it is not easy to differentiate which brand a car is. PCA is useful in classification when identifying a product but not as useful in clustering when identifying a product. PCA is useful in dimension reduction.