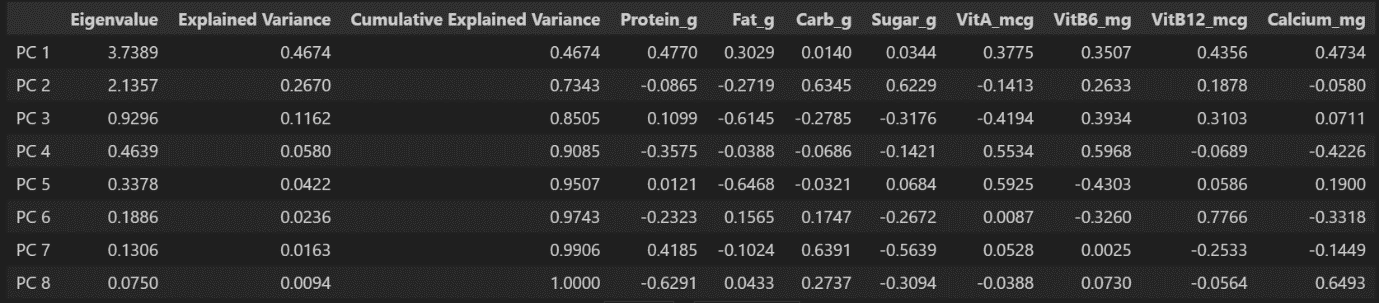
**Question 1**

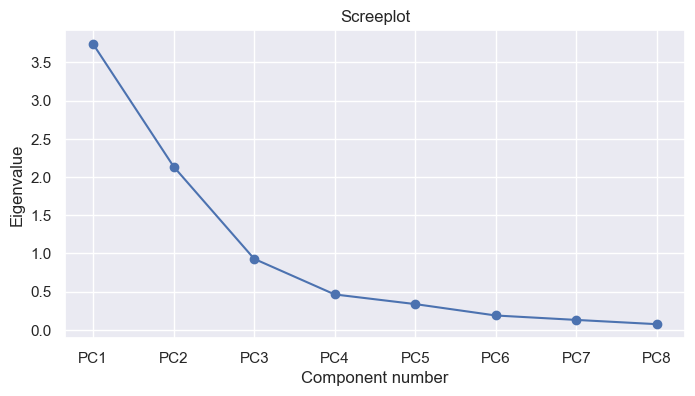
**Part a**

Since the variables have different units of measurements (grams, micrograms and milligrams), we will be carrying out PCA on the correlation matrix instead of on the covariance matrix. To do so, we must standardize the data.

PCA Results



Scree plot

****

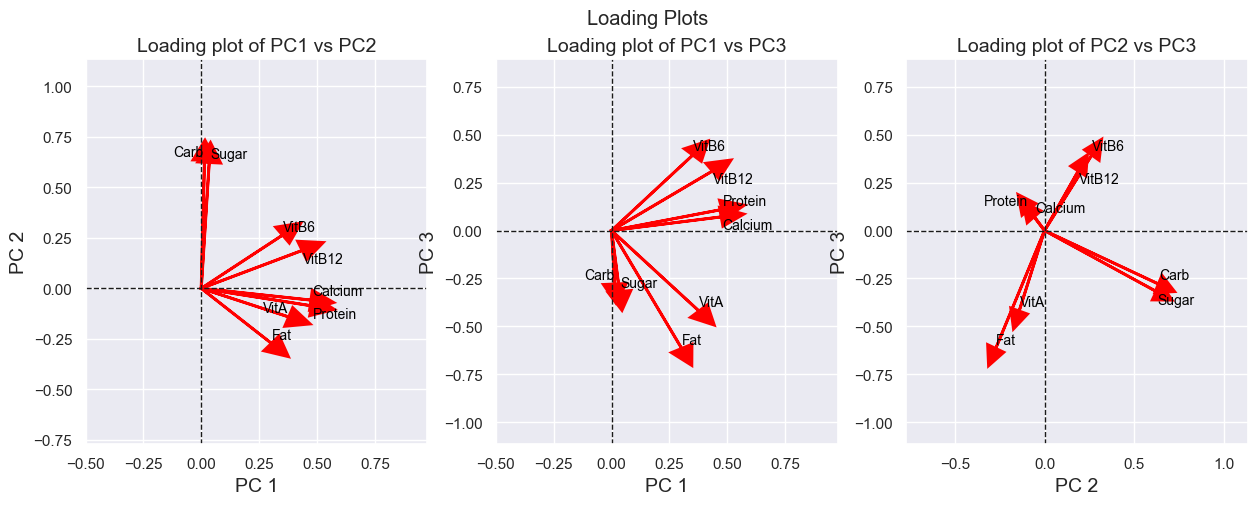
**Extraction of PCs**

Number of PCs to extract:

* By Kaiser’s rule, extract the first 2 PCs where eigenvalues (3.74 and 2.14) are > 1
* 1st 3 PCs already account for 85% of the total variance
* Scree Plot shows elbow at PC4, suggesting 3 PCs to extract

Let’s extract the first three PCs.

**Loading Plot**

****

**Interpretation of PCs**

**PC1: 0.4770z1 + 0.3029z2 + 0.0140z3 + 0.0344z4 + 0.3775z5 + 0.3507z6 + 0.4356z7 + 0.4734z8**

All the loadings are in the same direction. This PC seems to measure general nutrition.

**PC2: - 0.0865z1 - 0.2719z2 + 0.6345z3 + 0.6229z4 - 0.1413z5 + 0.2633z6 + 0.1878z7 - 0.0580z8**

The loadings on Calcium, Protein, Vitamin A and Fat are opposite in sign to the other loadings. The magnitude of the loadings on Carbohydrate and Sugar are also larger than the loadings on Vitamin B6 and Vitamin B12. This PC seems to measure the level of carbohydrates and sugar against other types of nutrients in dairy products.

**PC3: 0.1099z1 - 0.6145z2 - 0.2785z3 - 0.3176z4 - 0.4194z5 + 0.3934z6 + 0.3103z7 + 0.0711z8**

The loadings on Fat, Carbohydrate, Sugar and Vitamin A are opposite in sign to the other loadings. This PC seems to measure a contrast of more beneficial nutrients in optimal amounts such as protein, calcium, vitamin B6 and B12 against other nutrients that can have negative health implications when consumed in excessive amounts such as Fats, Carbohydrates and Sugar.

**Score Plot**

**A picture containing text, screenshot, diagram, line

Description automatically generated**

From the score plot, certain types of milk and cheese tend to score high on PC1.

Yogurt, Ice Cream alongside certain types of milk and certain types of cream tend to score high on PC2 while cheese scores lower on PC2.

Cheese, Milk, Yogurt, and certain types of milk tend to score higher on PC3.

**Part b**

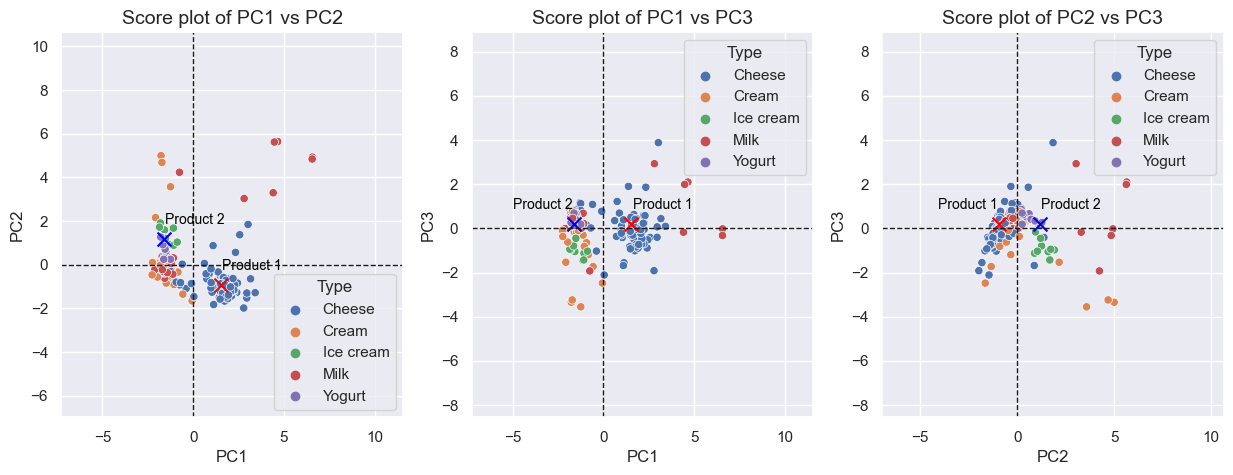
Since PC2 measures the level of carbohydrates and sugar against other types of nutrients in dairy products, we will be looking at the score plots with PC2 in one of the axis to answer this part.

(b)(i) Low carbohydrates and sugar but high in other nutrients suggest a low PC2 value. The type of dairy product would be Cheese.

(b)(ii) High carbohydrates and sugar but low in other nutrients suggest a high PC2 value. The type of dairy product would be Ice Cream and Yogurt.

**Part c**

First, we must standardise the 2 dairy products based on the sample mean and sample standard deviation from the initial dataset. After that, we can calculate the scores for the 2 dairy products by multiplying the scaled products with the loadings. Finally, we can add the 2 new dairy products to the score plot and see where it lies.

****

From the new score plot, it seems that Product 1 is likely Cheese while Product 2 is likely Yogurt.

**Part d**

So far, it seems that each type of dairy product has a unique combination of different PC values.

In part C where two 2 products have to be identified with PCA, I had to first standardize the data using the mean and standard deviation of the original dataset just like how the original dataset was standardized back in part a. This is to ensure that the same pre-processing steps are being used.

I expected one of the principal components to measure total vitamins by grouping up Vitamin A, Vitamin B6 and Vitamin B12 but in the end, there were no PCs where Vitamin A had the same sign as Vitamin B6 and B12 other than in PC1 where all loadings had the same signs.

**Question 2**

Since the variables have different units of measurements, we will be carrying out PCA on the correlation matrix instead of on the covariance matrix. To do so, we must standardize the data.

**Part a**

PCA Results

****

Scree plot

A picture containing text, line, plot, screenshot

Description automatically generated

**Extraction of PCs**

Number of PCs to extract:

* By Kaiser’s rule, extract the first 2 PCs where eigenvalues (7.52, 2.01, 1.09) are > 1
* 1st 3 PCs already account for 85% of the total variance
* Scree Plot shows elbow at PC4, suggesting 3 PCs to extract

Let’s extract the first three PCs.

**Loading Plot**

**A picture containing text, diagram, line, plot

Description automatically generated**

***PC1: 0.2772z1 + 0.3212z2 + 0.3212z3 + 0.1009z4 + 0.3494z5 + 0.2655z6 + 0.3317z7 - 0.0024z8 + 0.3093z9 - 0.3125z10 - 0.3211z11 + 0.3280z12***

The loadings on Compression-ratio, City-mpg, Highway-mpg are opposite in sign to the other loadings***.*** PC1 seems to capture a combination of length, width, curb-weight, engine, horsepower. This indicates that PC1 is likely a measure of the physical dimensions and performance of the car.

***PC2: -0.3777z1 - 0.2461z2 - 0.1798z3 - 0.5322z4 - 0.0941z5 + 0.2142z6 + 0.1237z7 - 0.5003z8 + 0.2945z9 - 0.2088z10 - 0.1637z11 + 0.0614z12***

The loadings on Wheel-base, length, width, height, curb-weight, compression-ratio, city-mpg, highway-mpg are opposite in sign on the other loadings. The coefficients of PC2 are negative for most variables indicating the PC2 is likely capturing features that have an inverse relationship with PC2. This indicates that PC2 is likely a measure of more efficient cars.

***PC3: -0.1292z1 - 0.1377z2 + 0.0461z3 - 0.3602z4 + 0.0397z5 + 0.4055z6 + 0.2723z7 + 0.5898z8 + 0.0505z9 + 0.3217z10 + 0.2998z11 + 0.2203z12***

The loadings on Wheel-base, length, height are opposite in sign on the other loadings. The coefficients that are positive include engine specifications (cylinders, engine) and factors related to fuel efficiency (city-mpg, highway-mpg). This indicates that PC3 is likely a measure of more powerful and higher performance cars.

**Score Plot**

**A picture containing text, diagram, screenshot, line

Description automatically generated**

From the score plot, Sedans and Hardtops generally score higher on PC1 while Hatchbacks generally score lower on PC1

Hatchbacks, Hardtops and Convertibles tend to score higher on PC2 while Wagons tend to score lower on PC2

Hardtops and Convertibles tend to score higher on PC3 while wagons tend to score lower on PC3.

**Part b**

The PCA results of this dataset are not as obvious as the results from the dairy nutrition dataset in Question 1. From the score plot, the clusters are not very clear and the difference for the different car types are not very large. As such, this makes PCA not as useful for classification or clustering purposes since information needed to distinguish the different clusters may have been lost during PCA.