**MAI Assignment 1**

*Done By: Darius Khoo*

*Admin No.: 2222259*

*Class: 2A/06*

**Question 1 (a): Present your PCA analysis with all the necessary outputs and graphs. Explain all decisions made in the analysis.**

For this project I decided to use correlation matrix to perform PCA mainly due to the fact that the data set does not have the same scale throughout with some columns measured by g (grams) while some other columns measured by mg (micrograms) with different scales, and different magnitude present in the dataset I felt that correlation matrix would be most suitable for this case.

- For the first step of PCA I standardized that data

A screenshot of a computer

Description automatically generated with medium confidence

After that is done the data is ready to perform PCA which is exactly what I did

A screenshot of a computer screen

Description automatically generated with low confidence

Since Correlation matrix is used, Kaiser’s Rule can be implemented, From the above table we just eliminate any pc that has a eigenvalue of less than 1, leaving us with PC1 and PC2

A picture containing text, diagram, line, plot

Description automatically generated

elbow

Although the scree plot shows the elbow at around PC4 kaiser’s rule tell us that the first 2 PCs are more than enough hence I decided to just keep 2 pc’s

A picture containing text, diagram, screenshot, line

Description automatically generated

Based on the loading plot both our PCs we can infer that:

PC1:

We can see that all nutrients contribute positively to PC1 with all of them contributing roughly the same amount except sugar and carbohydrates which contribute extremely weakly towards PC1

**PC2:**

Vitamin B12, B6, Sugar and carbohydrates contributes positively to PC2 with carbohydrates and sugar contributing strongest compared to the weaker vitamin B6 and B12

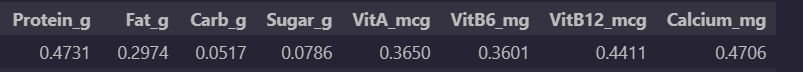
The rest of the features (calcium, protein, Vitamin A, Fat) all contributes negatively. Towards PC2 with all of them contributing around the same amount

**Question 1 (b): Which type(s) of dairy product has/have the following attributes?**

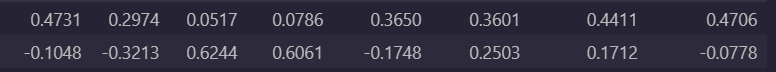
1. **Low carbohydrates and sugar but high in other nutrients**
2. **High carbohydrates and sugar but low in other nutrients.**

We can infer the answer to this question using a combination of the PCA and the score plot

Another being to infer from biplot, but I find biplot confusing and hard to infer so I stuck with what I was familiar with, that being inferring based on PCA and score plot



PC1 and PC2



Based on the the 2 PC’s we can see what are the estimated values that each product will have

1. **Low carbohydrates and sugar but high in other nutrients**

When a product has low carbohydrates and sugar but high in the rest, we can evaluate that on the 2 PCs.

For PC 1, that product should be positive, due to the fact that carbs and sugar having a much weaker contribution to PC1 while the remaining nutrients has a higher contribution.

For PC 2 we can see that it will be negative, since carbohydrates and sugar are the highest contributors to PC2 have a low amount will make the value lesser furthermore with the others being high in amounts, the negatives nutrients will contribute more to the result as, well making it negative.

In Conclusion for the first item, it should have a positive PC1 score while having a negative PC2 score, lets look at the score plot to see the result

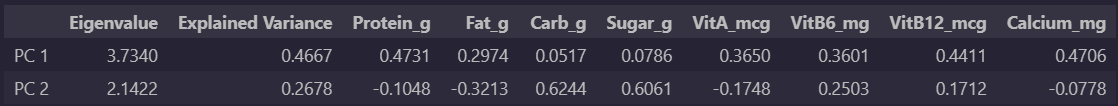
A screen shot of a graph

Description automatically generated with low confidence

Based on the score plot we can infer where the item roughly will be on the score plot, as we can see all the products around that area (negative PC2, Positive PC1) are cheese products. Hence, we can conclude that cheese products are low in carbohydrates and sugar but high in other nutrients

1. **High carbohydrates and sugar but low in other nutrients.**

For this next item we do the same steps first looking at the pc’s



For PC1:

A high carbohydrate and sugar but low in other nutrients would still show a positive score for PC1, since all the features in PC1 has a positive contribution, it should not result in negative values, however since carbohydrates and sugar contribute so little compared to the other features the value of PC1 should be on the lower end of positive.

For PC2:

A high carbohydrate and sugar but low in other nutrients will show a positive value in PC2, since carbohydrate and sugar contributes large amount positively towards PC2 and since other features are low it totals up to a higher positive value for PC2 since (high x high contribution) + (low x low contribution) = high result, we can tell that PC2 should be a larger positive value

Let’s look at the score plot to see where that item would plot on the graph

A screen shot of a graph

Description automatically generated with low confidence

From the score plot we can infer that even though there are not a lot of points around the area we predicted, the possible products would either be milk or cheese as there are 2 of each points in that area, We can conclude that milk and cheese are products that are high in carbohydrates and sugar but low in other nutrients

**Question 1 (c): 2 dairy products have their nutritional values listed below. Which type of dairy product is each of them likely to be? Use a suitable number of principal components to help you with your analysis.**

1. **Product 1**
2. **Product 2**

For this Question the first step in need to do is add the points into the dataset, when new values are considered, it changes the PCA results hence I need to add both products in as new records. And revaluate the new PCA results and infer from then on.

A screenshot of a computer

Description automatically generated with medium confidence

Product 2

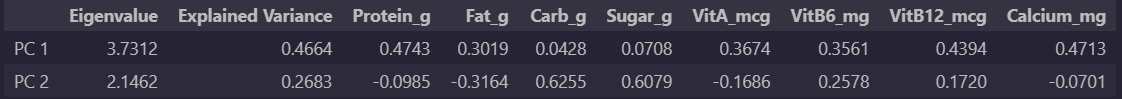
Product 1

After adding those 2 records into the dataset, we can evaluate a new PCA based on the new data and evaluate based on that.

A screenshot of a computer screen

Description automatically generated with low confidence

To be consistent I used kaiser method to evaluate the PCA as well dropping PCs with eigenvalues less than 1 leaving the same 2 PCs behind but this time the equation of each PC is slightly different



Now I don’t need to evaluate the position of the items using guessing as I have the exact values of each nutrients (feature) so I can just plot the score plot and see where it is at

A screen shot of a graph

Description automatically generated with low confidence

Product 1

Product 2

As we can see from the score plot, I plotted.

1. Product 1

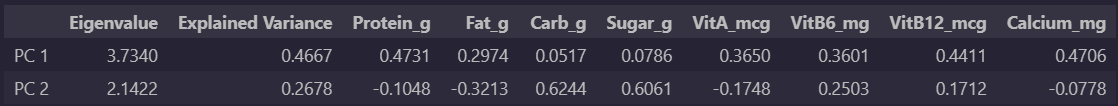
All the products surrounding product 1 in the score plot is cheese showing that there is a high chance that product 1 is a cheese product as well.

1. Product 2

For product 2 it is not as clear cut as product 1 where all the products that are around it, is 1 food type, around product 2 there seem to be both ice cream products and yogurt products in its near vicinity, however after carefully looking at the point I notice way more yogurt products around it compared to ice cream, hence there is a high chance that product 2 is a yogurt product

Question 1 (d): Describe your observations so far, comparing what you have done in part (c) and the decision(s) you have made in the earlier part of PCA in (a). Compare how you may have expected the principal components to perform and how they have actually performed.

Based on the PCA calculated in part 1



If I did not add the data points in and just calculated the values for product 1 and 2

I would get these results.

Product 1:

Product 2:

PC1: -1.47

PC2: 1.30

PC1: 1.60

PC2: -1.09

While I got these results

Product 1:

Product 2:

PC1: -1.47

PC2: 1.31

PC1: 1.60

PC2: -1.10

As we can see even though the difference is not a lot, note that the values were rounded up, so the difference is not exactly what you see,

How I got the first result was taking the standardized values of both product 1 and 2 and substituted them in the PC1 and PC2 created in part (a), while the second one was the values calculated in part (c) where I added the values into the dataset and calculated a new PCA, Which is why, the actual results were not that far off from what I expected, as even though the values were added in to the dataset there were over hundreds of records in the original dataset which made me expect a change in result but nothing major due to the size of the original dataset. Which was exactly what was shown in the output.

Question 2 (a): Present your PCA analysis with all the necessary outputs and graphs. Explain all decisions made in the analysis.

For Question 2 since the magnitude and the scale of the dataset being different, with some data being measured in cm while some are in dollars, we use correlation matrix as well to perform this PCA.

First, we perform standardization to the dataset.

A screenshot of a computer

Description automatically generated with medium confidence

With that being done the data now can be used in our PCA calculations, I put it pca\_result function and got an output of

A screenshot of a computer screen

Description automatically generated with low confidence

I decided to choose the number PCs to extract based on the cumulative variance explained, as long as the cumulative variance explained reaches 80%, we can discard the rest

From the table though it seems that we only need till PC3 to hit that 80% cumulative variance with the total being 88.03%, hence discarding the rest.

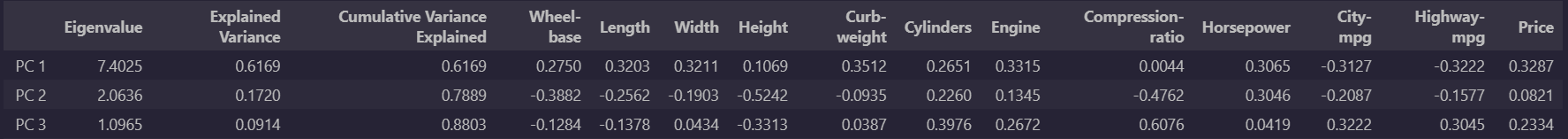
Let’s also take a look at the scree plot:

A picture containing text, diagram, plot, line

Description automatically generated

elbow

Based on the scree plot we can see that the elbow is at PC4 but since the explained variance of the first 3 PCs already exceed 80% I feel like less is better and concluded that I will only keep the first 3 PC’s, resulting in.



After obtaining the reduced PCA, we can look at the loading plot and infer more from there.

A picture containing text, diagram, line, plot

Description automatically generated

Although we extracted 3 PC’s why is it in the graph we can only see 2?

Well in a graph it can only show up to 2 dimensions hence it just shows the first 2 which are PC1 and PC2, although not shown on the graph don’t worry since PC3 is still used in evaluating results in a 3d space it is just not shown in a 2d graph.

Back to this loading plot, we can infer for

PC1: We can see that highway mpg and city mpg contributes negatively to PC1,

While compression ratio, height, wheel-base, length, width, curb-weight, price, engine, cylinders, horsepower all contributes positively to PC1, even though compression ratio looks like it doesn’t if you look closely it slightly point to the right side of PC1 it just means it’s contribution towards PC1 is extremely weak.

PC2: we can infer that highway mpg, city mpg, compression ratio, height, wheel-base, length, width and curb weight all contribute negatively to PC2

While price, engine, cylinder, horsepower all contributes positively to PC2.

After all these the last analysis we can do is to take a look at the score plot

A screen shot of a computer screen

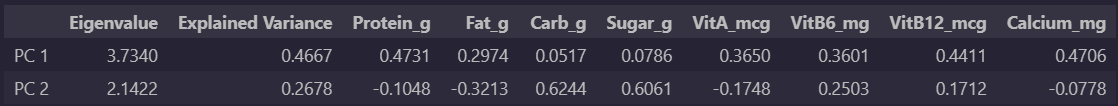
Description automatically generated with low confidence

Question 2 (b): Explain the difference between the PCA results of this dataset and the dairy nutrition dataset in Question 1, and thus comment on the usefulness of PCA for classification or clustering purposes.

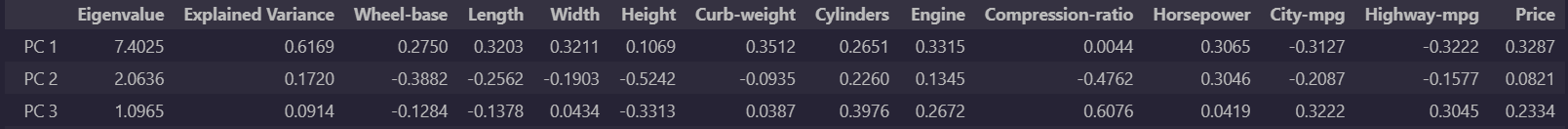
The difference between the PCA results can be observed from 2 parts, PCA output, score plot let’s look at both right now.

First let’s compare the PC’s we extracted.

Question 1:



Question 2:



Even though the criteria I used to determine the amount of PC’s to extract slightly differs, we can see quite a big difference in results as in question 1 due to Kaisers rule we choose to extract 2 PCs while in question 2 we decided based on the cumulative variance explained which led me to extract 3 PCs, another change is that there are way more features/ categories in question 2 will elaborate more on this later.

I would say that PCA for both datasets are extremely useful.

Since PCA’s general objectives are:

1. Dimension reduction without much loss of information
2. Data visualization and interpretation

Based on the first point of dimension reduction some people would say PCA in question 1 was much more useful since we reduced to 2 PCs compared to 3 PCs in question 2, however keep in mind that the features/categories in the original dataset was different as mention for question 2 there were originally 12 features which means 12 PCs having reduced to 3 after performing PCA, comparatively question 1 had originally 8 features reducing to 2 features after performing PCA,

based on this I would say that questions 2’s dimension reduction from PCA was much better as it managed to drop much more PCs compared to question 1

Next let’s look at the difference in score plot.

Question 1:

A screen shot of a graph

Description automatically generated with low confidence

Question 2:

A screen shot of a computer screen

Description automatically generated with low confidence

From the above 2 scores plot let’s infer what we can see in terms of classification/clustering

Question 1:

The score plot for question 1 is really clear with clusters of each object although some items pass over to the clusters of other objects, we can still clearly identify the clusters of each object making for easy inference in the future. The obvious clusters shows that PCA is useful for clustering/classification for this data/question.

Question 2:

Unlike the score plot in question 1 the data points for question 2 are all around the place almost seems random with no clear groups or clusters which makes it hard to infer anything from the score plot. Making PCA quite bad for clustering/classification for this data/question

In conclusion for Question 1 PCA is useful for dimension reduction and also extremely useful for classification/clustering.

While for question 2 PCA is slightly more useful in terms of dimension reduction however a lot worse and not useful in terms of clustering and classification