Name: Justin Wong Juin Hng

Admin Number: p2112646

Question 1)

a)

PCA should be carried out on a covariance matrix as we can convert micrograms to grams.

Graphical user interface, application

Description automatically generated

A picture containing polygon

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

When we look at the screen plot, we can see that the elbow of the curve is located at PC2.

This means that all Principal Components equal to or below 2 are irrelevant.

This is further supported by the explained variance of PC1 itself, accounting for 87.28% of the variance of the dataset

c)

**A picture containing chart

Description automatically generated**

When we analyse the direction and magnitude of the arrows, we can see that Protein and Fat influence

PC2 the most, and Carbohydrates and Sugar have a sizeable impact on the influencing of PC1.

ci)

Graphical user interface

Description automatically generated with low confidence

Knowing that, when we want to find a product with low protein and fat, PC2 value must be low and PC1 value must be high. This means that breakfast cereals are the food product.

cii)

Knowing that, when we want to find a product with high protein and fat, PC2 value must be high and PC1 value must be low. This means that poultry are the food product.

d)

Chart

Description automatically generated

When we plot the score plot, we can see that the unknown variable is low on both PC1 and PC2 and is in the cluster of Fruits and Juices. This suggests that the unknown product is Fruits and Juices.

2a) We should carry it out on a correlation matrix as the units are not the same and cannot be converted to the same.

2b)

Table

Description automatically generated

A picture containing chart

Description automatically generatedTable

Description automatically generated

We extract the first 3 PCs as it explains the minimum amount of variance required, at 81.78% variance.

Additionally, the scree plot shows that the elbow of the plot is at pc=3, but since PC1 and PC2 only explain for 73% of the variance of the dataset, we keep PC3.

c)

A picture containing chart

Description automatically generated

We can see that there is a distinct clustering of grouped data with PC1 and PC2. We can roughly identify that the unknown cultivator is most likely cultivator number 2, where the unknown cultivator (number 3) is most surrounded by.

A picture containing scatter chart

Description automatically generated

We can observe that even though there is a very clear distinction for Cultivator 0, the clustering of data around the left of the y axis is incredibly cluttered and makes it very hard to distinguish which cultivator the unknown data point belongs to.

A picture containing scatter chart

Description automatically generated

We can see some very slight clustering and grouping data distributed horizontally with PCA 2 and PCA 3,

Still, we can see that the previous trend of grouped data surrounding the unknown is the pink square, which corresponds to Cultivator 2.

d)

I think that PC3 is not useful. Even though it managed to produce a plot where we could distinguish between the different cultivators, it was not as effective as the PC1 vs PC2 plot at distinctively grouping and clustering the data points in such a way that was immediately obvious from a first look. On top of that, it provided a useless score plot (PC1 vs PC2), where an analysis of the score plot showed that it was ineffective at determining the group accurately.

Adding on to that, PC3 only accounts for 8% of the explained variance.

After adding all these factors together, it is safe to conclude that PC3 is not useful in helping me identify the cultivator the wine in part (c) is likely to originate from.