1a) PCA should be carried out on correlation matrix as some numerical variables are of different magnitudes. For example, the magnitude of VitA\_mcg is larger than other variables. Moreover, we can see a difference in the units of measurements, (micrograms for VitA\_mcg and other numerical variables in grams). The problem of using covariance matrix is that variable (VitA\_mcg) with the highest variance will dominate the first principal component.

Graphical user interface, text

Description automatically generated

1b) Since PCA is performed on correlation matrix, using several guidelines, Kaiser’s rule states that we should retain PCs whose eigenvalues are greater than 1. Given the output below, extract first 2 PC, where eigenvalues are more than 1.

Retain two PC, which already has a cumulative explained variance of 82.6%. It is sufficient as we do not

Graphical user interface

Description automatically generated

Looking at the scree plot, shows an ‘’elbow” at PC3, hence retain PCs on the left of elbow, which is first 2 PCs.

Chart, line chart

Description automatically generated

Finally, extract first 2 PCs.

PC1 measures a contrast in Carb & Sugar & fiber & vitamin A **versus** protein & fat. If the product has be**low average or low** amounts protein and fats, and **above average or high** amounts in carb, sugar, fiber and VitA, it will have a high value for PC1.

Likewise, if it has above average amounts of protein and fats, and below average amounts in other nutrients, then PC1 will have a negative/small value.

**Looking at the loading plot, the loadings for protein, fat on PC1, -0.2884, and -0.2594 respectively, correspond to a negative loading, whereas the loadings for other nutrients on PC1 are positive, which highlights a contrast.**

Loadings for Carb, VitA, Fibre and sugar are quite large while the loadings of Fat and Protein are smaller.

|  |
| --- |
|  |

PC2 measures a weightage average of all nutrients, with more ‘weights’ on protein and fats. If the product has above average or high amounts protein and fats, and below average or low amounts of other nutrients, then PC2 will have a high value. And more nutritious food

* An increase in proteins and fat would cause a larger significance/increase in PC2 as compared to the rest of the nutrients
* Food with more protein, and fats have higher value of PC2.

For PC 2, the magnitude of protein and fat is higher than magnitudes of other nutrients, and all loadings are in similar direction.

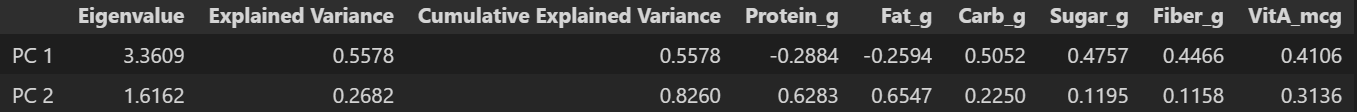
Chart, line chart

Description automatically generated

1c)

Text, letter

Description automatically generated



i) Using the score plot, we can see that **Breakfast Cereals** type scored high for PC1, which suggest it has low or below average amounts of fats and protein, and high or above average amounts of other nutrients. Similarly, it scored lower for PC2.

ii) Next, we can see that the **poultry products** type scored low for PC1, which means have high/above average amounts of protein and fats, and low/below average amounts of other nutrients. Similarly, it scored higher for PC2.

Chart, scatter chart

Description automatically generated

1d)

Normalization: Text

Description automatically generated with medium confidence

Std: standard deviation Text

Description automatically generated with medium confidence

(3sf calculations)

**Nutritional value of a particular food product:**

Z-Score for Protein : (0.3 – 8.91)/9.83 = -0.876

Z-Score for Fat : (0.1- 4.8)/6.56 = -0.716

Z-Score for Carbohydrate : (19.6 – 28.3)/32.1 = -0.271

Z-Score for Sugar : (16.7-11.1)/12.6 = 0.444

Z-Score for fiber : (2.9 -3.7)/4.41 = -0.181

Z-Score for VitA: (2- 187)/261 = -0.709

**PC1 and PC2:**

Text, letter

Description automatically generated

Sub in values into PC1 and PC2:

Text, letter

Description automatically generated

Looking at the score plot, we find the product that scores 0.14 on PC1, and -1.27 on PC2.

Chart, scatter chart

Description automatically generated

Hence, the product type would likely be fruits and fruit juices, given that the point are in the group of fruits and fruit juices

**Q2)**

1. PCA should be carried out on correlation matrix, given that the constituents/variables have different units of measurement and are of different magnitudes.

Malic acid is measured in g/L, magnesium is measured mg/L, while alcohol is measured in %v/v at 20 oC .

Values of proline also has a much larger magnitude than other variables like colour intensity.

Proline ranges from (278 - 1547), colour intensity ranges from (1.28 – 10.68)

The problem of using covariance matrix is that variable (proline (mg/L) with the highest variance will dominate the first principal component.

Timeline

Description automatically generated with medium confidence

1. From output below, to retain PCs with cumulative percentage of at least 80% of total variance. By kaiser’s rule, it states that we should retain PCs whose eigenvalues are greater than 1. Given the output below, it shows 2 PCs to extract, where eigenvalues are more than 1.

Moreover, 3 PCs are enough to account for 81.8% of total variance. Looking at the scree plot, shows an ‘’elbow” at PC3, hence retain PCs on the left of elbow, which is first 2 PCs. Ultimately, we want to retain PCs with cumulative percentage of at least 80% of total variance, we will keep first 3 PCs.

Text, letter

Description automatically generated

A picture containing text, road, monitor, black

Description automatically generated

Chart, line chart

Description automatically generated

***Interpretation of PCs.***

Chart

Description automatically generated

* PC1 measures a contrast in alcohol, magnesium, flavonoids, color intensity, proline versus malic acid, and nonflavanoid phenols. If the wine has below average or low amounts of malic acid and nonflavanoid phenols, and above average or high amounts of other constituents, PC1 will have a high value. Likewise, if it has above average amounts of malic acid and nonflavanoid phenols, and below average amounts in other constituents, then PC1 will have a negative/small value.
* Looking at the loadings, the loadings for malic acid and nonflavanoid phenols on PC1, -0.2166, -0.3888 respectively, correspond to negative loadings, whereas the loadings for other constituents on PC1 are positive, which highlights a contrast.

Chart, radar chart

Description automatically generated

* PC2 measures a contrast in flavonoids versus other constituents. If the wine has below average or low amounts of flavonoids, and above average or high amounts of other constituents, then PC2 will have a high value. Similarly, if it has above average flavonoids, and below average or very low amounts of other constituents, then PC2 will have a small value.
* Loadings for flavonoids is -0.3708 on PC2, corresponds to negative loading, whereas loadings for other constituents on PC2 are positive, which highlights a contrast.

Chart, line chart

Description automatically generated

PC3: measure contrast in flavonoids, malic acid and magnesium vs other constituents. If wine has below average or low amounts of flavonoids, malic acid and magnesium, and above average/high amounts of other constituents, then PC3 will score small/negative value. Likewise, if it has above average/high amount of flavonoids, malic acid and magnesium, and low/below average amounts of other constituents, then PC3 will score high. The loadings of flavonoids, malic acid and magnesium are 0.0457, 0.3505, 0.7918 respectively on PC3, correspond to positive loadings, whereas loadings for other constituents are negative, highlighting a contrast.

- Loadings of flavonoids are very small, as more variance is captured in PC1, PC2

C)

A picture containing text, meter

Description automatically generated

Std = standard deviation

A black screen with white text

Description automatically generated with low confidence

**For attributes of a particular wine:**

Z-score for alcohol: (12.75- 13.0)/0.798 = -0.313

Z-score for malic\_acid: (3.11-2.32)/1.04 = 0.76

Z-score for magnesium: (96-97.7)/11.6 = -0.147

Z-score for flavonoids: (0.61 – 1.98)/1 = -1.37

Z-score for nonflavanoid\_phenols: (0.52 – 0.369)/0.118 = 1.28

Z-score for color\_intensity: (4.56- 5.08)/2.2 = -0.236

Z-score for proline : (632 – 750)/323 = -0.365

Chart, radar chart

Description automatically generated

Through a score plot, we can display the cultivator information.

Sub in values into PC1 PC2 and PC3:

Text, letter

Description automatically generated

PC1 : y1 = -1.67, PC2 : y2 = 0.992, PC3 : y3= 0.177

We can locate the point.

Chart, radar chart, scatter chart

Description automatically generated

Looking at the score plot, we can find the cultivator through the annotation above. And Likely the wine comes from **cultivator 2** as the point can be seen in that group.

d)

Chart, scatter chart

Description automatically generated

Using a score plot, and locating the values for PC1 = -1.67, and PC2 = 0.992, we can see that that particular wine comes from cultivator 2 group. Hence using PC3 does not really help in classifying the cultivator of the particular wine, as seen in score plot, using PC1 and PC2 indicates that the particular wine is from cultivator 2.

Furthermore, looking at the strip plot below, we can see that in general, data points of 3 cultivators have similar PC3 scores, hence making it not useful to determine which cultivator the wine came from.Graphical user interface, chart, scatter chart

Description automatically generated

Hence, PC3 is not useful in helping me identify which cultivator the wine in part c is likely to originate from