MAI Assignment 1 Answers

Name: Lim Jun Jie

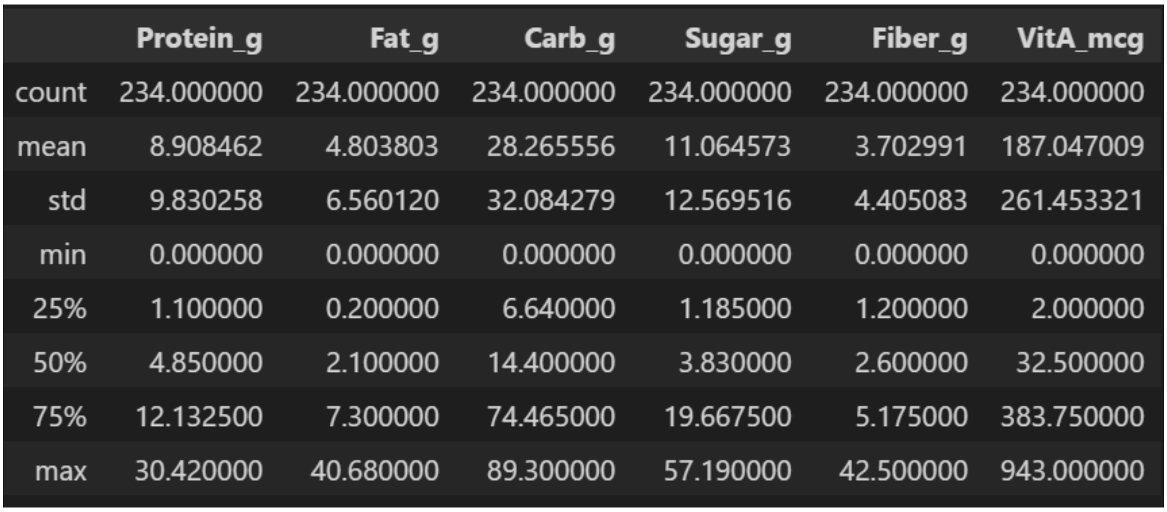
Admission Number: 2100788

Class: DAAA/FT/2A/02

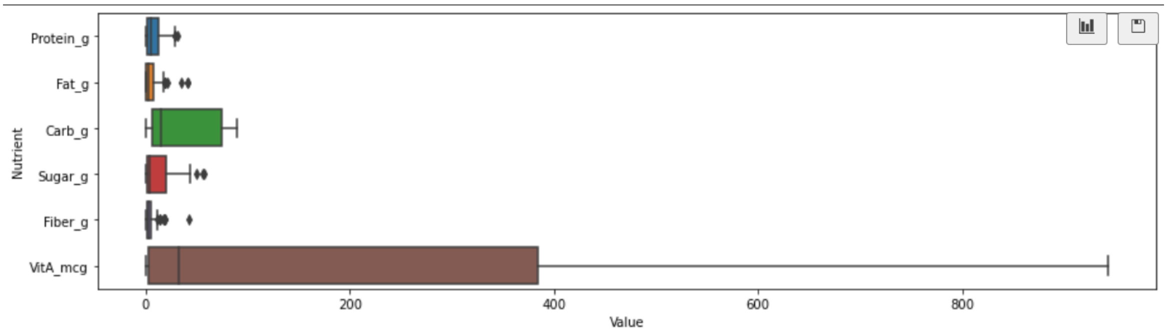
Question 1(a)

PCA should be carried out on correlation matrix.

1. The mean for vitA variable is 187 (3s.f). However, the mean for the 5 other variables ranges from 3.70 to 28.3 (3s.f). This indicates that the variables are of different magnitudes.
2. The variables are also measured in different scales since vitA variable is measured in micrograms while the 5 other variables are measured in grams
3. Since the variables have different scales and largely differing magnitudes, PCA on the correlation matrix is preferred to prevent the variable with the largest magnitude from dominating the first PC



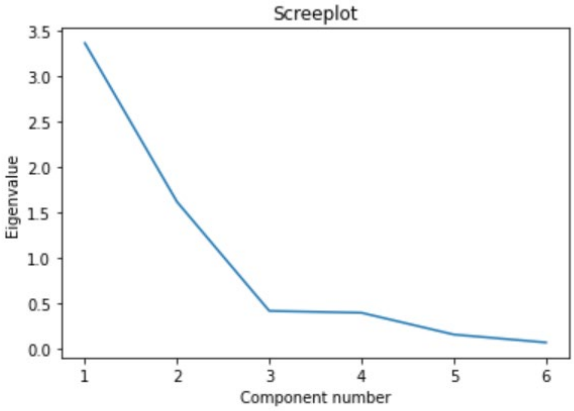
	Protein_g	Fat_g	Carb_g	Sugar_g	Fiber_g	VitA_mcg
count	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000
mean	8.908462	4.803803	28.265556	11.064573	3.702991	187.047009
std	9.830258	6.560120	32.084279	12.569516	4.405083	261.453321
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.100000	0.200000	6.640000	1.185000	1.200000	2.000000
50%	4.850000	2.100000	14.400000	3.830000	2.600000	32.500000
75%	12.132500	7.300000	74.465000	19.667500	5.175000	383.750000
max	30.420000	40.680000	89.300000	57.190000	42.500000	943.000000

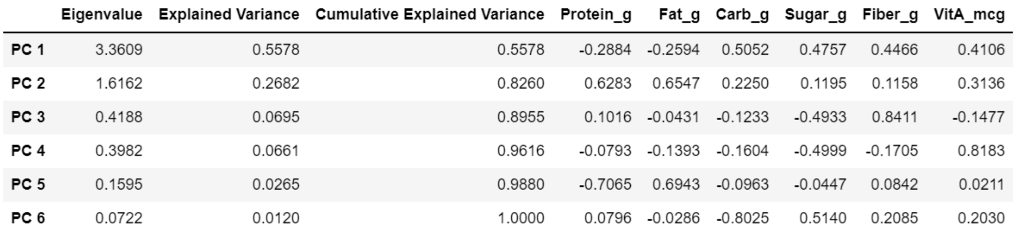


Question (b)

There are 3 criteria that can be used to determine the number of Principal Components to extract:

1. Use Kaiser's Rule – Kaiser's Rule states that we should retain PCs whose eigenvalues are greater than 1 as an eigenvalue that is less than 1 explains less variance than a variable would. (Only when correlation matrix is used in the PCA)
2. Look at cumulative explained variance – Retain PCs with cumulative percentage of at least 80% to 90%
3. Retain PCs left to the elbow before the inflection point shows up and the graph starts to flatten

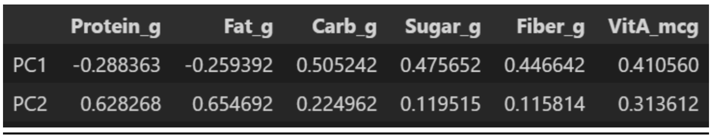




	Eigenvalue	Explained Variance	Cumulative Explained Variance	Protein_g	Fat_g	Carb_g	Sugar_g	Fiber_g	VitA_mcg
PC 1	3.3609	0.5578	0.5578	-0.2884	-0.2594	0.5052	0.4757	0.4466	0.4106
PC 2	1.6162	0.2682	0.8260	0.6283	0.6547	0.2250	0.1195	0.1158	0.3136
PC 3	0.4188	0.0695	0.8955	0.1016	-0.0431	-0.1233	-0.4933	0.8411	-0.1477
PC 4	0.3982	0.0661	0.9616	-0.0793	-0.1393	-0.1604	-0.4999	-0.1705	0.8183
PC 5	0.1595	0.0265	0.9880	-0.7065	0.6943	-0.0963	-0.0447	0.0842	0.0211
PC 6	0.0722	0.0120	1.0000	0.0796	-0.0286	-0.8025	0.5140	0.2085	0.2030

1. Using Kaiser's Rule, only the first 2 PCs will be selected as the eigenvalues of PC 1 and 2 are greater than 1.
2. The cumulative explained variance indicates that only the first 2 PCs should be extracted as the first 2 PCs have a cumulative explained variance of 82.6%, which is more than 80% of total variance.
3. The scree plot shows that the elbow is located at PC3, starting from PC3, graph starts to become flat, and the eigenvalues differ a lot from PC 1 and 2, suggesting that only the first 2 PCs on the left of the elbow should be extracted.

   Since the 3 criteria all suggest extracting PC 1 and 2, I will only extract the first 2 Principal Components.



	Protein_g	Fat_g	Carb_g	Sugar_g	Fiber_g	VitA_mcg
PC1	-0.288363	-0.259392	0.505242	0.475652	0.446642	0.410560
PC2	0.628268	0.654692	0.224962	0.119515	0.115814	0.313612

Interpreting PC1

PC1 = -0.288363c_1 -0.259392c_2 +0.505242c_3 +0.475652c_4 +0.446642c_5 +0.410560c_6 (where c_1 to c_6 are values of each nutrient, in their respective units of measurement)

1. For PC1, the loadings on Protein and Fat are negative while other loadings are positive. PC1 measures a weighted average of nutrients contrasting against Protein and Fat.

2. For PC1 to be positive, the values of Protein and Fat should be below average while the values of Carbohydrate, Sugar, Fiber and Vitamin A should be above average.
3. For PC1 to be negative, the values of Protein and Fat should be above average while the values of Carbohydrate, Sugar, Fiber and Vitamin A should be below average

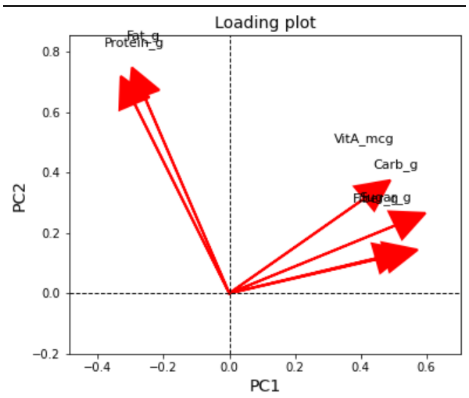
Interpreting PC2

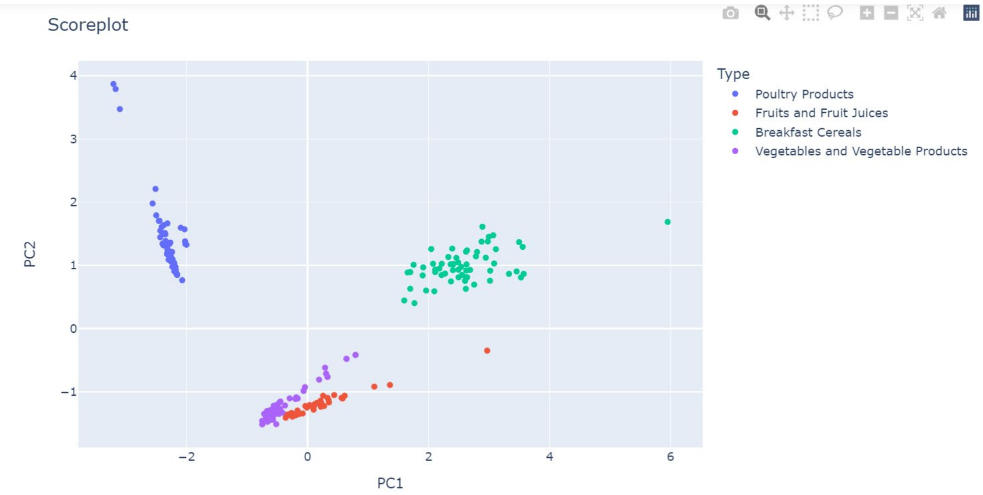
PC2 = $0.628268c_1 + 0.654692c_2 + 0.224962c_3 + 0.119515c_4 + 0.115814c_5 + 0.313612c_6$ (where c_1 to c_6 are values of each nutrient, in their respective units of measurement)

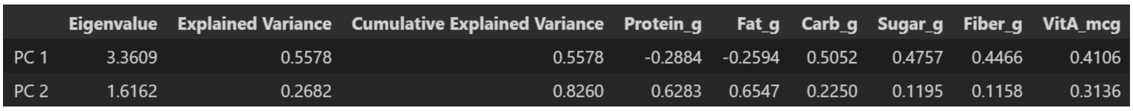
1. The loading for every nutrient in PC2 is positive.
2. For PC2 to be positive, all nutrients should be above average (as this is a scaled dataset, values below the mean are negative)
3. For PC2 to be negative, values of all the nutrients should be below average.

Question 1(c)

Loading Plot, Score Plot and Linear System of PC1 and PC2 used to answer part (c).



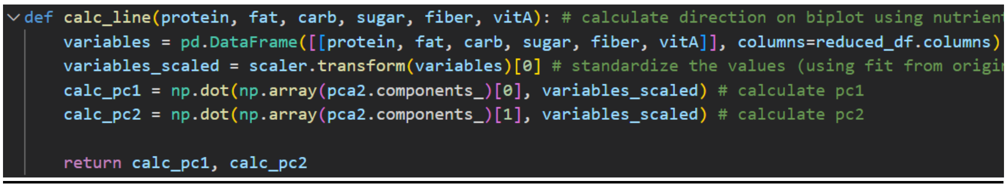




	Eigenvalue	Explained Variance	Cumulative Explained Variance	Protein_g	Fat_g	Carb_g	Sugar_g	Fiber_g	VitA_mcg
PC 1	3.3609	0.5578	0.5578	-0.2884	-0.2594	0.5052	0.4757	0.4466	0.4106
PC 2	1.6162	0.2682	0.8260	0.6283	0.6547	0.2250	0.1195	0.1158	0.3136

This is the function used to calculate the direction of line based on nutrient values. I will be using this function to determine the Type of food.

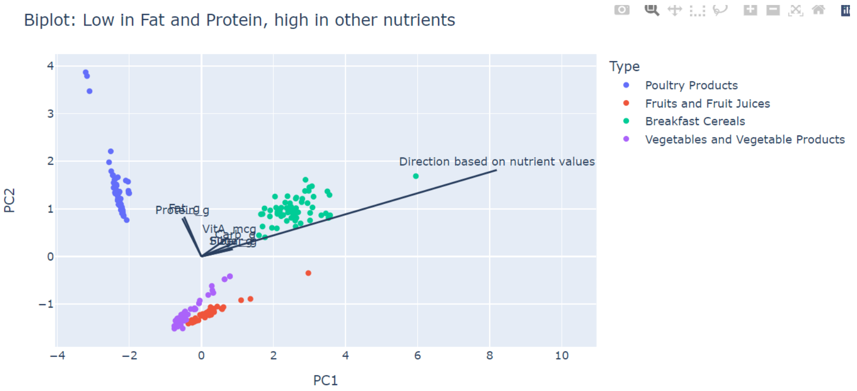
1. The nutrient values passed into this function are standardized using StandardScaler which was fit on the original dataset.
2. PC1 and PC2 for the nutrient values entered is calculated using the linear equations of PC1 and PC2.
3. A line is drawn from the origin to the point of the calculated PC1 and PC2



```
def calc_line(protein, fat, carb, sugar, fiber, vitA): # calculate direction on biplot using nutrien
    variables = pd.DataFrame([[protein, fat, carb, sugar, fiber, vitA]], columns=reduced_df.columns)
    variables_scaled = scaler.transform(variables)[0] # standardize the values (using fit from origi
    calc_pc1 = np.dot(np.array(pca2.components_)[0], variables_scaled) # calculate pc1
    calc_pc2 = np.dot(np.array(pca2.components_)[1], variables_scaled) # calculate pc2

    return calc_pc1, calc_pc2
```

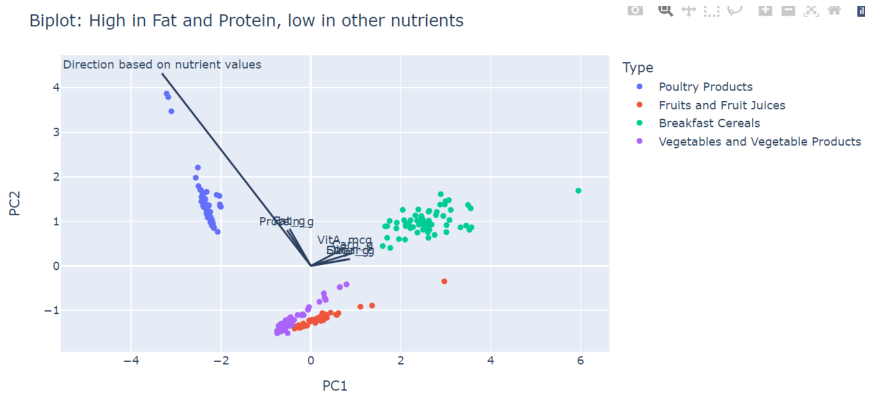
Question 1(c)(i): Low Protein and Fat, high in all other nutrients



1. The maximum values of other nutrients are used, while lower quartile values of Protein and Fat is used to mimic the scenario of low Protein and Fat, but high in all other nutrients
2. When Protein and Fat values are low, and values of other nutrients are high, the direction seems to point towards the Breakfast Cereals cluster.
3. The biplot also seems to imply that when all other nutrients except Protein and Fat are high, the possible type of food product would be Breakfast Cereals, since the loadings of all other nutrients are in the direction of Breakfast Cereals cluster.

Thus, the possible type of food product that is low in fat and protein but high in other nutrients is Breakfast Cereals.

Question 1(c)(ii): High Protein and Fat, low in all other nutrients



1. The maximum values of Protein and Fat are used, while lower quartile values of all other nutrients are used to mimic the scenario of high Protein and Fat, but low in all other nutrients
2. When protein and fat values are high and other values are low, the direction based on calculated PC1 and PC2 points towards Poultry Products cluster
3. The biplot also seem to suggest that high Protein and Fat food product would be the Poultry Products, since the direction of the loadings of Protein and Fat point towards the Poultry Products Cluster.

Hence, the possible type of food product that has high protein and fat and low in other nutrients is Poultry Products.

Question 1(d)

Working:

Values = (0.3, 0.1, 19.6, 16.7, 2.9 ,2)

Scaled Values = (-0.87758779, -0.71856717, -0.27066624, 0.44930193, -0.18267824, -0.70928026)

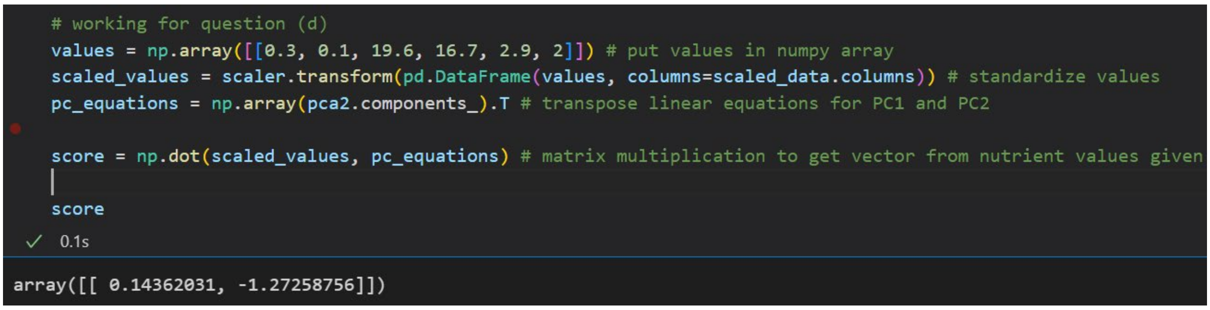
Score = Scaled Values x $\begin{pmatrix}\text{PC1 Components}\\ \text{PC2 Components}\end{pmatrix}$.Transposed

Score =

(-0.87758779, -0.71856717, -0.27066624, 0.44930193, -0.18267824, -0.70928026) x

$\begin{pmatrix}(-0.28836335, -0.25939213, 0.50524221, 0.47565165, 0.4466416\ , 0.41055986)\\ (0.62826837, 0.65469229, 0.22496195, 0.11951497, 0.11581441, 0.3136115)\end{pmatrix}$.Transposed

Score = (0.14362031, -1.27258756)



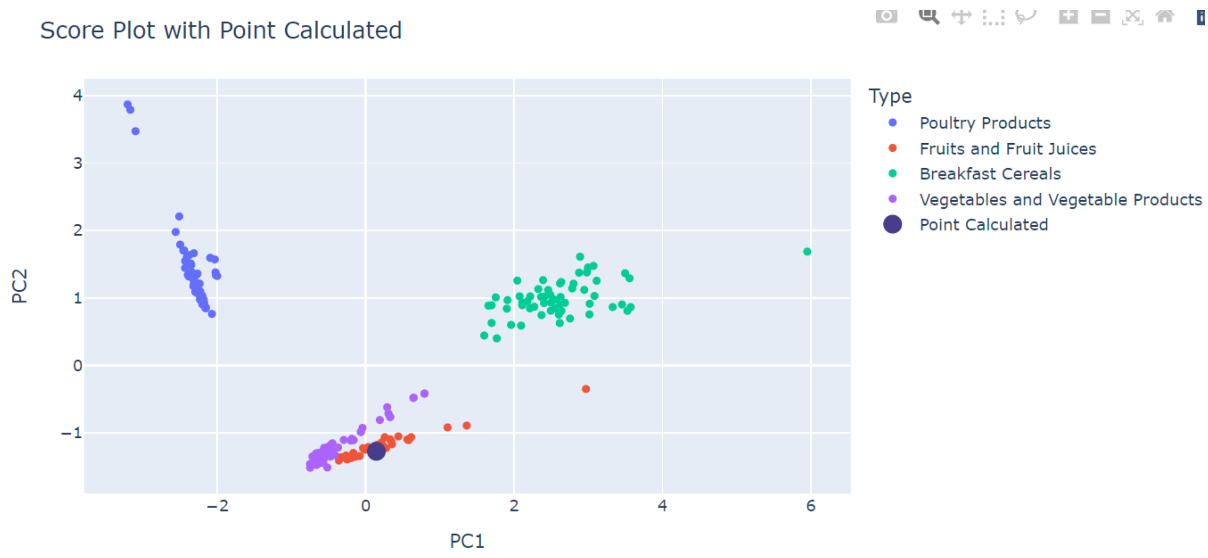
```
# working for question (d)
values = np.array([[0.3, 0.1, 19.6, 16.7, 2.9, 2]]) # put values in numpy array
scaled_values = scaler.transform(pd.DataFrame(values, columns=scaled_data.columns)) # standardize values
pc_equations = np.array(pca2.components_).T # transpose linear equations for PC1 and PC2

score = np.dot(scaled_values, pc_equations) # matrix multiplication to get vector from nutrient values given

score
```

✓ 0.1s

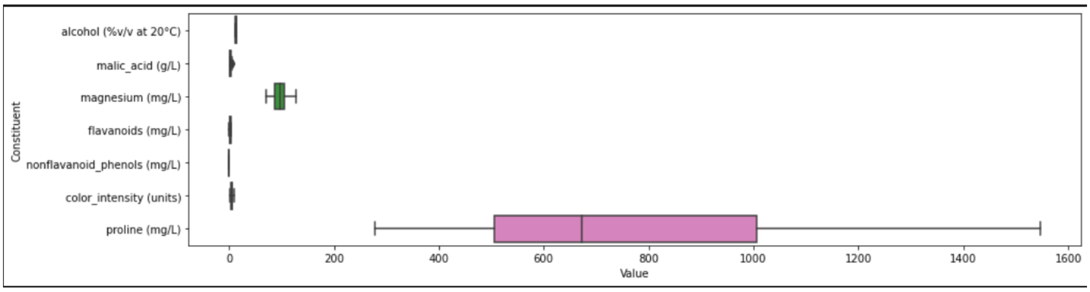
```
array([[ 0.14362031, -1.27258756]])
```

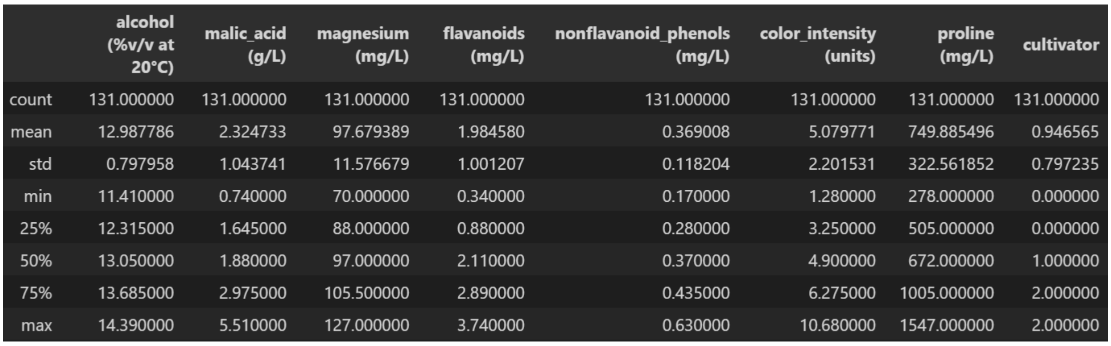


When the calculated score is plotted in the score plot, the point lies within the Fruits and Fruit Juices cluster, and the nearby datapoints surrounding it are all from the Fruits and Fruit Juices type. Thus, the type of the food product is likely to be Fruits and Fruit Juices.

Question 2(a)

PCA should be carried out on correlation matrix.





	alcohol (%v/v at 20°C)	malic_acid (g/L)	magnesium (mg/L)	flavanoids (mg/L)	nonflavanoid_phenols (mg/L)	color_intensity (units)	proline (mg/L)	cultivator
count	131.000000	131.000000	131.000000	131.000000	131.000000	131.000000	131.000000	131.000000
mean	12.987786	2.324733	97.679389	1.984580	0.369008	5.079771	749.885496	0.946565
std	0.797958	1.043741	11.576679	1.001207	0.118204	2.201531	322.561852	0.797235
min	11.410000	0.740000	70.000000	0.340000	0.170000	1.280000	278.000000	0.000000
25%	12.315000	1.645000	88.000000	0.880000	0.280000	3.250000	505.000000	0.000000
50%	13.050000	1.880000	97.000000	2.110000	0.370000	4.900000	672.000000	1.000000
75%	13.685000	2.975000	105.500000	2.890000	0.435000	6.275000	1005.000000	2.000000
max	14.390000	5.510000	127.000000	3.740000	0.630000	10.680000	1547.000000	2.000000

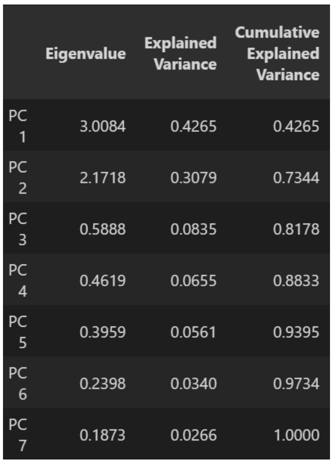
Reasons:

1. The 7 variables have different units of measurement such as grams/milligrams per litre, colour intensity units and alcohol percentage of total volume.
2. From the boxplot and descriptive summary, it is seen that the variables have different magnitudes and standard deviation. Proline has the highest median of 672mg/L and a large standard deviation of 322.56mg/L. Meanwhile, nonflavonoid phenols has the lowest median of 0.37mg/L and a very low standard deviation of 0.118mg/L

Thus, as the variables have different units of measurement and differing magnitudes, PCA should be done on the correlation matrix to avoid an inaccurate analysis (e.g., The variable with the largest magnitude dominating PC1.)

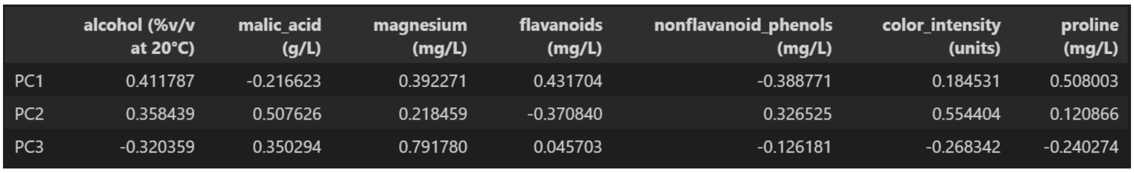
Question 2(b)

Since we should only retain the PCs with cumulative explained variance of at least 80% of total variance, the decision is only needed to be based on this factor.



	Eigenvalue	Explained Variance	Cumulative Explained Variance
PC 1	3.0084	0.4265	0.4265
PC 2	2.1718	0.3079	0.7344
PC 3	0.5888	0.0835	0.8178
PC 4	0.4619	0.0655	0.8833
PC 5	0.3959	0.0561	0.9395
PC 6	0.2398	0.0340	0.9734
PC 7	0.1873	0.0266	1.0000

From the PCA results table above, PC3 has a cumulative explained variance of 81.78%, thus only the first 3 PCs are to be extracted.



	alcohol (%v/v at 20°C)	malic_acid (g/L)	magnesium (mg/L)	flavanoids (mg/L)	nonflavanoid_phenols (mg/L)	color_intensity (units)	proline (mg/L)
PC1	0.411787	-0.216623	0.392271	0.431704	-0.388771	0.184531	0.508003
PC2	0.358439	0.507626	0.218459	-0.370840	0.326525	0.554404	0.120866
PC3	-0.320359	0.350294	0.791780	0.045703	-0.126181	-0.268342	-0.240274

Interpreting PC1

$PC1 = 0.411787c_1 -0.216623c_2 +0.392271c_3 +0.431704c_4 -0.388771c_5 +0.184531c_6 +0.508003c_7$

1. For PC1, the loadings on malic acid and nonflavonoid phenols is negative. This PC seems to measure a weighted average of constituents in wine, contrasting against malic acid and nonflavonoid phenols.
2. For a positive PC1, the values of malic acid and nonflavonoid phenols should be below average and alcohol, magnesium, flavonoids, colour intensity and proline should be above average
3. For a negative PC1, the values of malic acid and nonflavonoid phenols should be above average, while the other constituents should be below average

Interpreting PC2

$PC2 = 0.358439c_1 +0.507626c_2 +0.218459c_3 -0.370840c_4 +0.326525c_5 +0.554404c_6 +0.120866c_7$

The loading on flavonoids is negative. This PC seems to measure a weighted average of constituents in wine contrasting against flavonoids.

For a positive PC2, the values of the other constituents should be above average while value of flavonoids should be below average

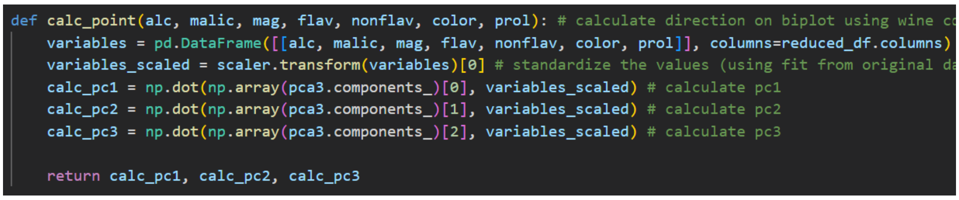
For a negative PC2, the value of flavonoids should be above average, and the values other constituents should be below average.

Interpreting PC3

$PC3 = -0.320359c_1 +0.350294c_2 +0.791780c_3 +0.045703c_4 -0.126181c_5 -0.268342c_6 -0.240274c_7$

1. The loadings on alcohol, nonflavonoid phenols, colour intensity and proline are negative. This PC seems to measure a weighted average of constituents in wine contrasting against alcohol, nonflavonoid phenols, colour intensity and proline.
2. For PC3 to be positive, the values of alcohol, nonflavonoid phenols, colour intensity and proline should be below average and other constituents should be above average.
3. For PC3 to be negative, the values of alcohol, nonflavonoid phenols, colour intensity and proline should be above average while the other values should be below average.

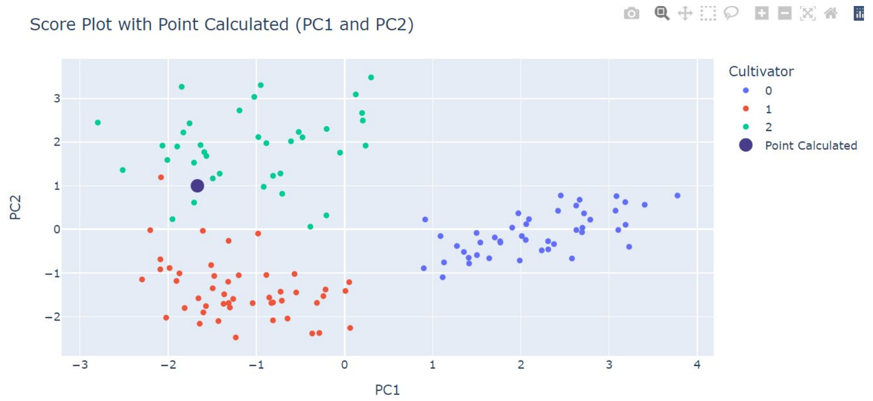
Question 2(c)



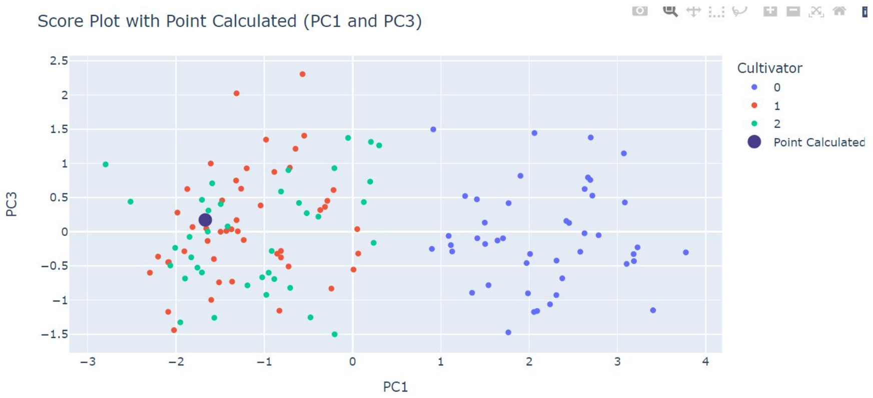
```
def calc_point(alc, malic, mag, flav, nonflav, color, prol): # calculate direction on biplot using wine co
    variables = pd.DataFrame([[alc, malic, mag, flav, nonflav, color, prol]], columns=reduced_df.columns)
    variables_scaled = scaler.transform(variables)[0] # standardize the values (using fit from original da
    calc_pc1 = np.dot(np.array(pca3.components_)[0], variables_scaled) # calculate pc1
    calc_pc2 = np.dot(np.array(pca3.components_)[1], variables_scaled) # calculate pc2
    calc_pc3 = np.dot(np.array(pca3.components_)[2], variables_scaled) # calculate pc3

    return calc_pc1, calc_pc2, calc_pc3
```

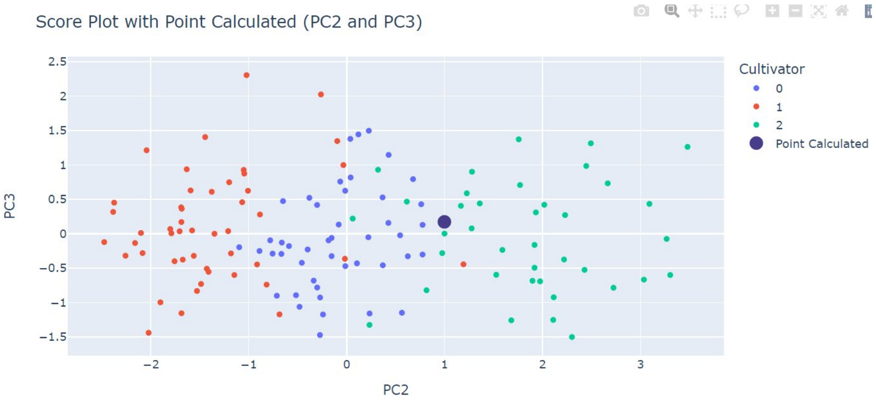
This function will be used to calculate PC1, PC2 and PC3 based on the attributes of the wine



When the point calculated is plotted in PC1 against PC2 graph, the point lies within the Cultivator 2 cluster. Its immediate neighbours are also of Cultivator 2 type



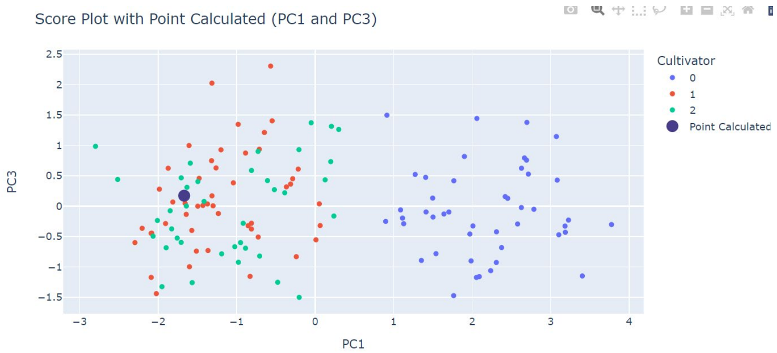
When the point is plotted in the score plot of PC1 and PC3, the calculated point is located within the mixed cluster of Cultivator 1 and 2, and a decision cannot be made as it is a mixed cluster, where there is no clear indication of Cultivator 1 and Cultivator 2 clusters



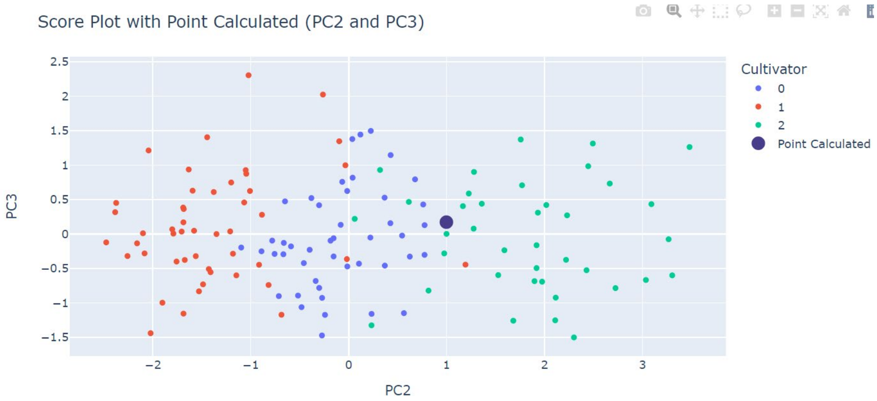
When the point is plotted in the score plot of PC2 and PC3, the point lies in between the cluster of Cultivator 0 and 2. However in this plot we can see that some points of Cultivator 2 are mixed within the cluster of Cultivator 0.

Thus, based on the insights the 3 plots gave, since all 3 plots suggest Cultivator 2 as one of the possible Cultivators, this wine is likely to have come from Cultivator 2.

Question 2(d)



Using the score plot of PC3 and PC1, plotting the point calculated using the wine's attributes in part (c). The point calculated lies in the mixed cluster of Cultivator 1 and Cultivator 2. Hence, this score plot of PC3 and PC1 is not useful in helping to identify the cultivator of wine in part (c) as a conclusion cannot be drawn between Cultivator 1 and 2.



When using the score plot of PC3 and PC2, when plotting the calculated point using the same wine attributes from part (c), the point lies in the middle of Cultivator 0 cluster and Cultivator 2 cluster. From this plot, it cannot be decided that whether the point calculated belongs to Cultivator 0 or Cultivator 2. Thus, this plot itself is not useful in helping to identify whether the point belongs to Cultivator 0 or 2.

Both plots which are using PC3 suggested Cultivator 2 as a possible cultivator. Even though PC3 plotted with PC1 and PC2 respectively cannot give as a solid conclusion on identifying the cultivator, it still was not completely useless and provided useful information as it constantly suggested Cultivator 2 as one of the possible cultivators when plotted against PC1 and PC2.

Thus, my conclusion is that PC3 was useful in helping to identify the cultivator of the wine in part (c) since it provided us some useful information.