Name: Ethan Tan
Admin: p2012085
Class: DAAA/FT/2A/03
Subject: Mathematics for A.I. (MAI) Assignment 1

Question 1

Part (a)

Should PCA be carried out on covariance or correlation matrix? Explain.

	Protein_g	Fat_g	Carb_g	Sugar_g	VitA_mcg	Calcium_mg
0	18.52	21.08	0.00	0.00	288	140
1	3.86	0.18	7.50	7.50	2	143
2	24.69	19.40	3.41	0.56	155	1146
3	23.76	32.11	2.57	0.52	264	685
4	1.03	4.38	6.89	6.89	61	32

PCA should be carried out on the correlation matrix. All the variables (nutrients) measure mass (per 100g). Although they currently have different scales (g/mg/mcg), it won't matter when they are standardized (converted to z-scores).

Part (b)

Extract the principal components. Justify your decision and interpret the principal components. You should include the necessary tables, outputs and graphs.

Step 1: Standardize the data

	Protein	Fat	Carbohydrate	Sugar	Vitamin A	Calcium
0	0.474880	0.623704	-0.854253	-0.654585	1.268127	-0.752101
1	-0.884563	-1.064148	-0.259528	0.023174	-0.981604	-0.743525
2	1.047033	0.488030	-0.583852	-0.603979	0.221924	2.123707
3	0.960793	1.514470	-0.650461	-0.607593	1.079339	0.805867
4	-1.146993	-0.724962	-0.307899	-0.031951	-0.517498	-1.060835

The data are now unitless and measure how far (the number of standard deviations) each observation is away from the mean.

Step 2: Perform Eigendecomposition

```
array([[ 0.51534,  0.48046, -0.15449, -0.10489,  0.46399,  0.50342],
       [ 0.17718, -0.1608 ,  0.65792,  0.67293,  0.11191,  0.21105],
       [-0.43594,  0.40172,  0.1047 ,  0.14705,  0.63571, -0.46029],
       [ 0.01062, -0.76116, -0.18808, -0.12491,  0.60314,  0.07594],
       [ 0.30162, -0.02685,  0.64551, -0.6405 ,  0.06444, -0.2779 ],
       [-0.64954,  0.0428 ,  0.2833 , -0.29782,  0.01178,  0.6381 ]])
```

Using sklearn's decomposition module, the eigenvalues (pca.explained_variance_), eigenvectors (pca.components_) and their corresponding explained variances (pca.explained_variance_ratio_) are extracted. The eigenvectors are arranged row-wise (each row is an eigenvector).

Step 3: Summarize the Principal Components

	Eigenvalue	Explained Variance	Cumulative Explained Variance	Protein	Fat	Carbohydrate	Sugar	Vitamin A	Calcium
PC 1	2.99627	0.49596	0.49596	0.51534	0.48046	-0.15449	-0.10489	0.46399	0.50342
PC 2	1.95078	0.32290	0.81886	0.17718	-0.16080	0.65792	0.67293	0.11191	0.21105
PC 3	0.56346	0.09327	0.91213	-0.43594	0.40172	0.10470	0.14705	0.63571	-0.46029
PC 4	0.29968	0.04960	0.96173	0.01062	-0.76116	-0.18808	-0.12491	0.60314	0.07594
PC 5	0.16542	0.02738	0.98911	0.30162	-0.02685	0.64551	-0.64050	0.06444	-0.27790
PC 6	0.06578	0.01089	1.00000	-0.64954	0.04280	0.28330	-0.29782	0.01178	0.63810

All the important information is collated and compiled into a DataFrame.

Step 4: Select the Principal Components

Some Principal Components (PCs) are dropped during dimension reduction. Just enough PCs that can explain most of the variance in the original data are kept.

Method 1 - Kaiser's Rule

	Eigenvalue	Explained Variance	Cumulative Explained Variance	Protein	Fat	Carbohydrate	Sugar	Vitamin A	Calcium
PC 1	2.99627	0.49596	0.49596	0.51534	0.48046	-0.15449	-0.10489	0.46399	0.50342
PC 2	1.95078	0.32290	0.81886	0.17718	-0.16080	0.65792	0.67293	0.11191	0.21105
PC 3	0.56346	0.09327	0.91213	-0.43594	0.40172	0.10470	0.14705	0.63571	-0.46029
PC 4	0.29968	0.04960	0.96173	0.01062	-0.76116	-0.18808	-0.12491	0.60314	0.07594
PC 5	0.16542	0.02738	0.98911	0.30162	-0.02685	0.64551	-0.64050	0.06444	-0.27790
PC 6	0.06578	0.01089	1.00000	-0.64954	0.04280	0.28330	-0.29782	0.01178	0.63810

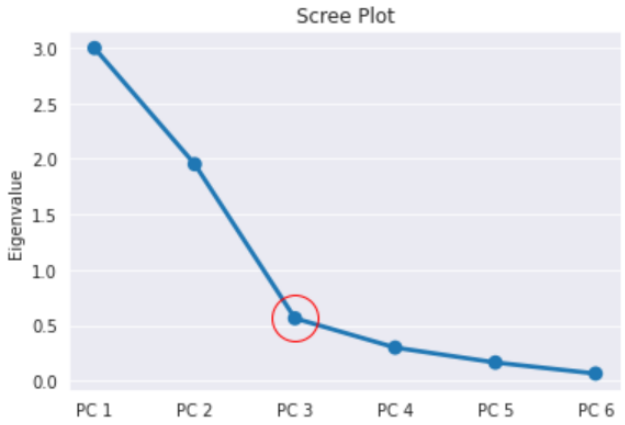
PCA was carried out on the correlation matrix, so Kaiser's Rule can be used in this case. Kaiser's Rule states that only PCs that have eigenvalues greater than 1. By Kaiser's Rule, only the top 2 PCs should be kept.

Method 2 - Cumulative Explained Variance

	Eigenvalue	Explained Variance	Cumulative Explained Variance	Protein	Fat	Carbohydrate	Sugar	Vitamin A	Calcium
PC 1	2.99627	0.49596	0.49596	0.51534	0.48046	-0.15449	-0.10489	0.46399	0.50342
PC 2	1.95078	0.32290	0.81886	0.17718	-0.16080	0.65792	0.67293	0.11191	0.21105
PC 3	0.56346	0.09327	0.91213	-0.43594	0.40172	0.10470	0.14705	0.63571	-0.46029
PC 4	0.29968	0.04960	0.96173	0.01062	-0.76116	-0.18808	-0.12491	0.60314	0.07594
PC 5	0.16542	0.02738	0.98911	0.30162	-0.02685	0.64551	-0.64050	0.06444	-0.27790
PC 6	0.06578	0.01089	1.00000	-0.64954	0.04280	0.28330	-0.29782	0.01178	0.63810

For this analysis, 80% is the benchmark for sufficient explained variance. The top 2 PCs already explain more than 80% (82%) of the total variance. Hence, only the top 2 PCs should be retained.

Method 3 - Scree Plot



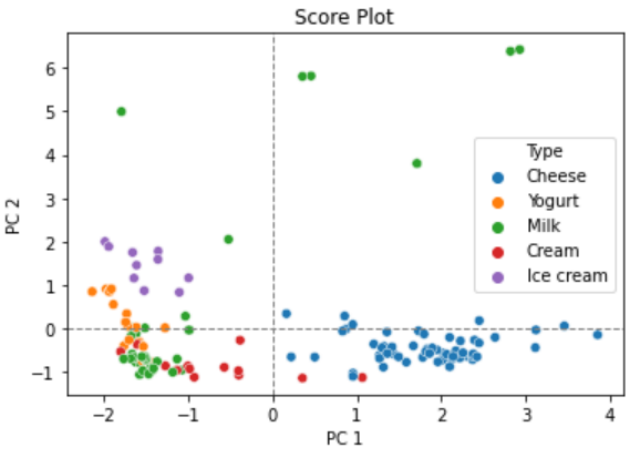
By the scree plot, there is an elbow at the 3rd PC. Therefore, only the top 2 PCs should be kept.

Summary of Principal Component Selection

All 3 methods agree that the top 2 Principal Components should be kept.

Step 5: Plot the Transformed Data

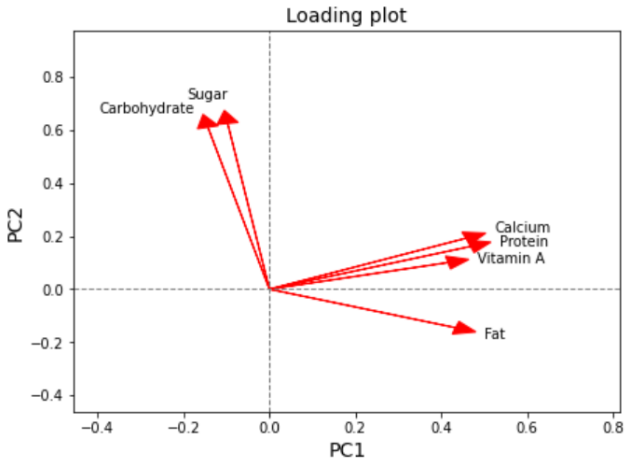
Score Plot



From the score plot, it can be observed that the cheese, yogurt and ice cream clusters are quite distinct in terms of PC1 and PC2. The cream data points have more variability across the PC1 axis, and the milk data points are scattered across both axes (perhaps due to outliers).

Step 6: Interpret the Principal Components

Loading Plot



From the loading plot, the interpretations of the PCs are as follows.

PC 1:

- PC 1 seems to measure the contrast between the concentrations of sugar and non-sugar nutrients
- Carbohydrate and Sugar are sugar nutrients
- Carbohydrate and Sugar have negative loading values for PC 1, while
- Calcium, Protein, Vitamin A and Fat have positive loading values for PC 1
- A higher score for PC 1 means that the dairy product has a greater concentration of non-sugar nutrients as compared to sugar nutrients
- A lower score for PC 1 means that the dairy product has a greater concentration of sugar nutrients as compared to non-sugar nutrients

PC 2:

- PC 2 seems to measure the contrast between the concentrations of fats and the other nutrients
- Fat is the only variable with a negative loading value
- A higher score for PC 2 means that the dairy product has a lower concentration of fats as compared to other nutrients
- A lower score for PC 2 means that the dairy product has a greater concentration of fats as compared to other nutrients

Part (c)

Which type(s) of dairy product has/have the following attributes? Explain your answer with the aid of a suitable graph with colour or marker to display "Type" information.

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

   Cheese. The cheese (blue) products have generally high PC 1 scores. Recalling the interpretations of the PCs, "a higher score for PC 1 means that the food has a greater concentration of non-sugar nutrients as compared to sugar nutrients". Therefore, the dairy product(s) with low carbohydrates and sugar but high in other nutrients should have a high PC 1 score.

2. High carbohydrates and sugar but low in other nutrients.

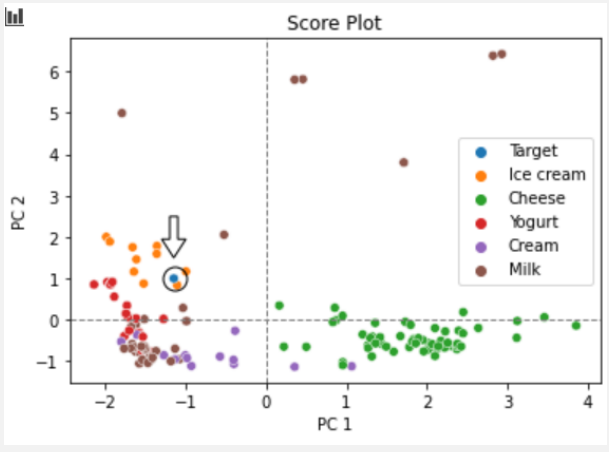
   Yogurt and ice cream. The yogurt (orange) and ice cream (purple) products have generally low PC 1 scores. Recalling the interpretations of the PCs, "a lower score for PC 1 means that the food has a greater concentration of sugar nutrients as compared to non-sugar nutrients". Therefore, the dairy product(s) with high carbohydrates and sugar but low in other nutrients should have low PC 1 scores.

   For milk (green) and cream (red) products, some data points have positive PC 1 scores while others have negative PC 1 scores. Due to this inconsistency, they cannot be definitively classified to have high carbohydrates and sugar but low in other nutrients.

Part (d)

A dairy product has its nutritional value listed below. Which type of dairy product is it likely to be? Show your working and explain.

Protein	4.8 g	Sugar	19.1 g
Fat	23 g	Vitamin A	17 mcg
Carbohydrate	26.2 g	Calcium	121 mg



The dairy product is likely to be ice cream. The given (above) data was scaled and transformed by PCs 1 and 2, and plotted on a labelled score plot, together with the rest of the data points. As the target datum point (blue) was situated in the Ice cream (orange) cluster, it is likely that the dairy product is an ice cream.

Note: the colour scheme used here is different from that of previous score plot

Question 2

Part (a)

Should PCA be carried out on covariance or correlation matrix? Explain.

	RI	Na	Mg	Al	Si	K	Ca
0	1.52369	13.44	0.00	1.58	72.22	0.32	12.24
1	1.51915	12.73	1.85	1.86	72.69	0.60	10.09
2	1.51508	15.15	0.00	2.25	73.50	0.00	8.34
3	1.52171	11.56	1.88	1.56	72.86	0.47	11.41
4	1.51556	13.87	0.00	2.54	73.23	0.14	9.41

PCA should be carried out on the covariance matrix. RI measures the ratio of the velocity of light in a vacuum to its velocity in the glass objects. On the other hand, Na, Mg, Al, Si, K and Ca measure the proportions of their corresponding oxides in the objects by weight. As RI and the 6 other quantitative variables have different scales and measure different things, it doesn't make sense to carry out PCA on the correlation matrix. Hence, PCA should be carried out on the covariance matrix.

Part (b)

Extract the principal components. Justify your decision and interpret the principal components. You should include the necessary tables, outputs and graphs.

Step 1: Perform Eigendecomposition

```
array([[ 0.002, -0.19 , -0.082, -0.168,  0.123, -0.457,  0.84 ],
       [ 0.001, -0.328, -0.538,  0.187, -0.322,  0.615,  0.292],
       [-0.001,  0.345, -0.759,  0.154,  0.463, -0.213, -0.149],
       [-0.001, -0.651,  0.098, -0.059,  0.709,  0.197, -0.146],
       [-0.   ,  0.465,  0.059, -0.62 ,  0.27 ,  0.52 ,  0.23 ],
       [ 0.002, -0.312, -0.339, -0.725, -0.302, -0.241, -0.336],
       [ 1.   ,  0.001,  0.001,  0.002,  0.002,  0.001, -0.001]])
```

Step 2: Summarize the Principal Components

	Eigenvalue	Explained Variance	Cumulative Explained Variance	RI	Na	Mg	Al	Si	K	Ca
PC 1	4.68843	0.41703	0.41703	0.00155	-0.18998	-0.08208	-0.16838	0.12268	-0.45668	0.83977
PC 2	3.18202	0.28304	0.70007	0.00069	-0.32782	-0.53804	0.18671	-0.32227	0.61546	0.29247
PC 3	2.28842	0.20355	0.90362	-0.00100	0.34499	-0.75864	0.15403	0.46276	-0.21318	-0.14875
PC 4	0.71961	0.06401	0.96763	-0.00131	-0.65129	0.09770	-0.05867	0.70902	0.19696	-0.14602
PC 5	0.26536	0.02360	0.99123	-0.00017	0.46485	0.05937	-0.61976	0.27047	0.51988	0.22991
PC 6	0.09860	0.00877	1.00000	0.00201	-0.31171	-0.33937	-0.72492	-0.30178	-0.24110	-0.33607
PC 7	0.00000	0.00000	1.00000	1.00000	0.00072	0.00056	0.00156	0.00208	0.00090	-0.00113

Step 3: Select the Principal Components

Method 1 - Kaiser's Rule

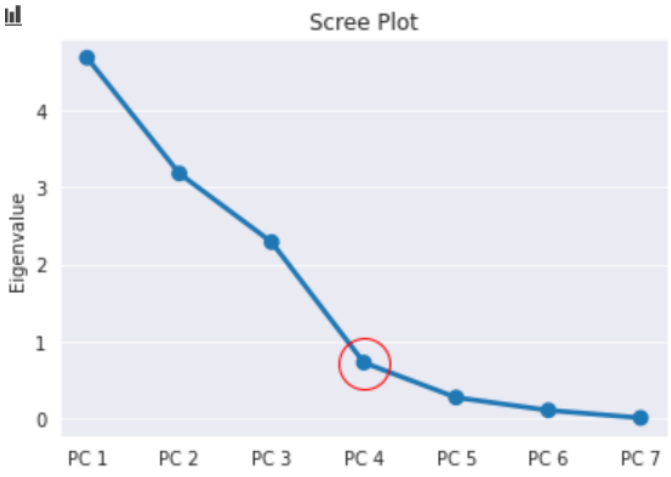
PCA was not carried out on a covariance matrix, so Kaiser's Rule is not applicable in this case.

Method 2 - Cumulative Explained Variance

	Eigenvalue	Explained Variance	Cumulative Explained Variance	RI	Na	Mg	Al	Si	K	Ca
PC 1	4.68843	0.41703	0.41703	0.00155	-0.18998	-0.08208	-0.16838	0.12268	-0.45668	0.83977
PC 2	3.18202	0.28304	0.70007	0.00069	-0.32782	-0.53804	0.18671	-0.32227	0.61546	0.29247
PC 3	2.28842	0.20355	0.90362	-0.00100	0.34499	-0.75864	0.15403	0.46276	-0.21318	-0.14875
PC 4	0.71961	0.06401	0.96763	-0.00131	-0.65129	0.09770	-0.05867	0.70902	0.19696	-0.14602
PC 5	0.26536	0.02360	0.99123	-0.00017	0.46485	0.05937	-0.61976	0.27047	0.51988	0.22991
PC 6	0.09860	0.00877	1.00000	0.00201	-0.31171	-0.33937	-0.72492	-0.30178	-0.24110	-0.33607
PC 7	0.00000	0.00000	1.00000	1.00000	0.00072	0.00056	0.00156	0.00208	0.00090	-0.00113

Once again, 80% is the benchmark in this analysis. The top 3 PCs explain more than 80% (90%) of the total variance. Hence, only the top 3 PCs should be retained.

Method 3 - Scree Plot



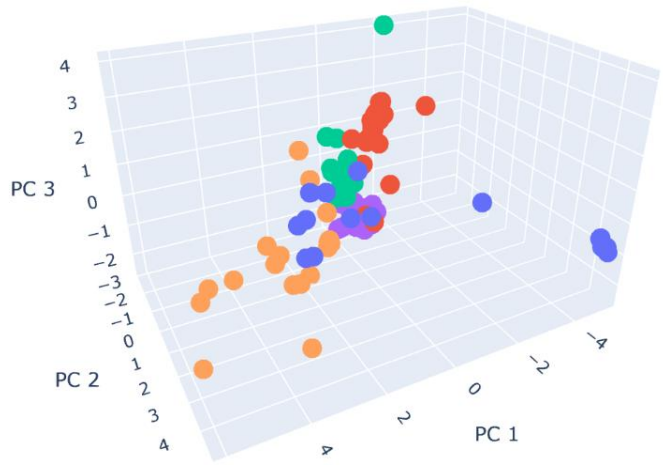
By the scree plot, there is an elbow at the 4th PC. Therefore, only the top 3 PCs should be kept.

Summary of Principal Component Selection

Both methods 2 and 3 agree that the top 3 Principal Components should be kept.

Step 4: Plot the Transformed Data

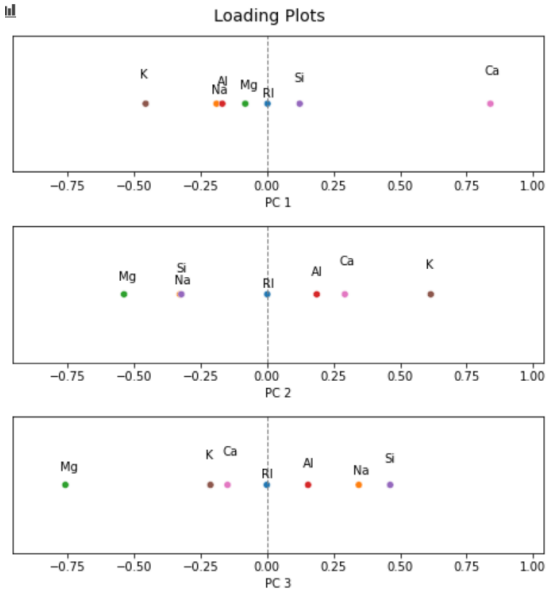
Score Plot



From the 3D score plot, it is quite obvious that the different classes of glass objects form relatively distinct clusters. This is useful in predicting the class of an unknown glass object, like in Part (c).

Step 5: Interpret the Principal Components

Loading Plot



Three separate 1D loading plots are used instead of one 3D graph as it is easier to visualize and compare the loading values.

From the loading plots, the interpretations of the PCs are as follows.

PC 1:

- PC 1 seems to measure the contrast between the concentrations of calcium oxide (CaO) and silicon oxide (SiO_2) against the concentrations of potassium oxide (K_2O), sodium oxide (Na_2O), aluminium oxide (Al_2O_3) and magnesium oxide (MgO)
- Ca and Si have positive loading values while
- K, Na, Al and Mg have negative loading values
- A higher PC 1 score means that the glass object has a higher concentration of silicon oxide and calcium oxide as compared to the rest of the oxides involved
- A lower PC 1 score means that the glass object has a lower concentration of silicon oxide and calcium oxide as compared to the rest of the oxides involved

PC 2:

- PC 2 seems to measure the contrast between the concentrations of potassium oxide, calcium oxide and aluminium oxide against the concentrations of magnesium oxide, sodium oxide and silicon oxide
- K, Ca and Al have positive loading values while
- Mg, Na and Si have negative loading values
- A higher PC 2 score means that the glass object has a higher concentration of potassium oxide, calcium oxide and aluminium oxide as compared to the rest of the oxides involved
- A lower PC 2 score means that the glass object has a lower concentration of potassium oxide, calcium oxide and aluminium oxide as compared to the rest of the oxides involved

PC 3:

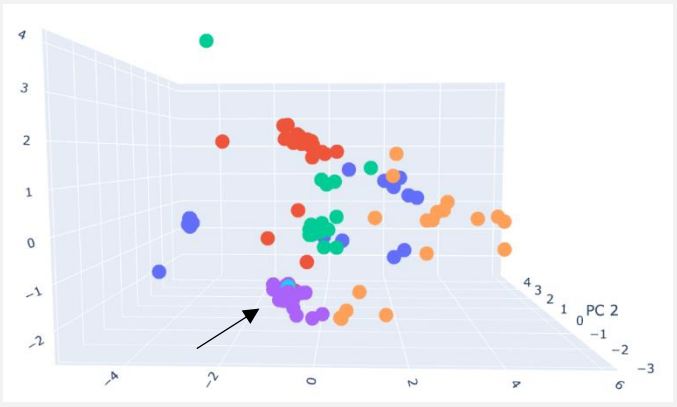
- PC 3 seems to measure the contrast between the concentrations of silicon oxide, sodium oxide and aluminium oxide against the concentrations of magnesium oxide, potassium oxide and calcium oxide
- Si, Na and Al have positive loading values while
- Mg, K and Ca have negative loading values
- A higher PC 3 score means that the glass object has a higher concentration of silicon oxide, sodium oxide and aluminium oxide as compared to the rest of the oxides involved
- A lower PC 3 score means that the glass object has a lower concentration of silicon oxide, sodium oxide and aluminium oxide as compared to the rest of the oxides involved

RI has a very small absolute loading value for all 3 PCs. This means that RI adds little to no information. Thus, it can be considered a redundant variable in this case study.

Part (c)

The following shows the attributes of a glass object. Which class does it likely belong to? Explain your answer with the aid of a suitable graph with colour or marker to display "Class" information.

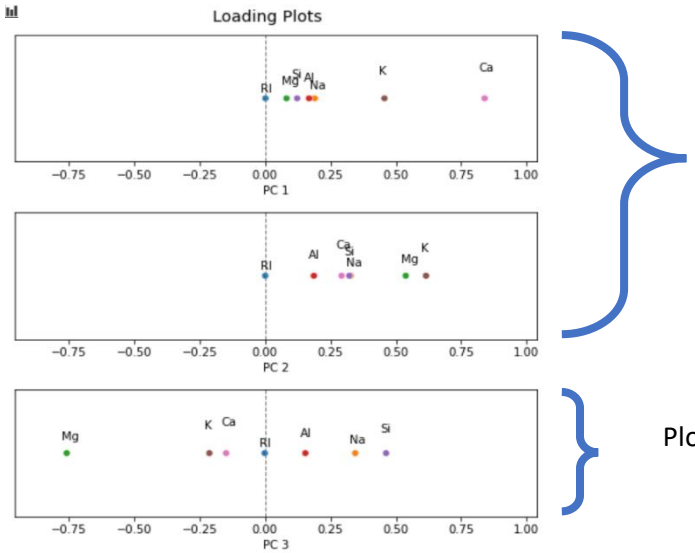
RI	1.51641	Si	73.05
Na	13.04	K	0.53
Mg	3.5	Ca	8.6
Al	1.28	color Containers headlamps Tableware VehWin BuildWin Target	



The glass object likely belongs to the VehWin (Vehicle Window) class. The given (above) data was transformed by PCs 1 and 2, and plotted on a 3D labelled score plot, together with the rest of the data points. As the target datum point (light blue) was situated in the VehWin (purple) cluster, it is likely that the glass object belongs to the VehWin class.

Part (d)

Explain how PC3 is advantageous over the first two principal components.



Plot of absolute loading values
(PC 1 and PC 2)

Plot of unaltered loading values
(PC 3)

PC3 captures the contrast between the more important variables and less important variables of both PC1 and PC2. For PC1, Ca and K have the 2 highest absolute loading values. In other words, Ca and K are the most important variables for PC 1. For PC2, K and Mg are the most important variables. These 3 variables (Mg, K and Ca) have negative loading values for PC3. The rest of the variables (Si, Al and Na) have positive loading values for PC3. PC3 tells the difference between the positive and negative values of the individual variables during transformation by PC1 or PC2. For example, if PC1 was close to 0, and PC3 was large, that means that during the transformation of PC1, the individual transformed variables were quite even.