Question 1

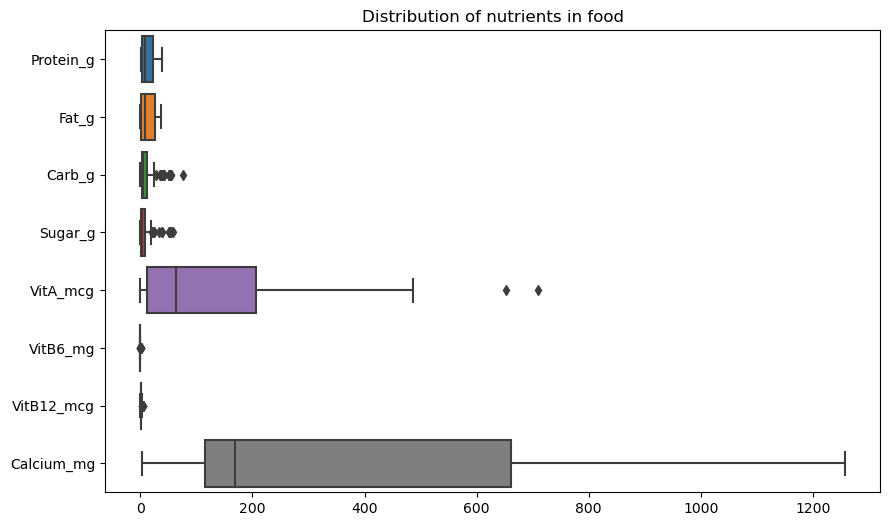
(a) Present your PCA analysis with all the necessary outputs and graphs. Explain all decisions made in the analysis.

Principal components are sets of linear combinations of variables that explain the variance-covariance structure. When doing PCA, the general objective is to perform dimension reduction without much loss of information and data visualization and interpretation. Such visualizations include

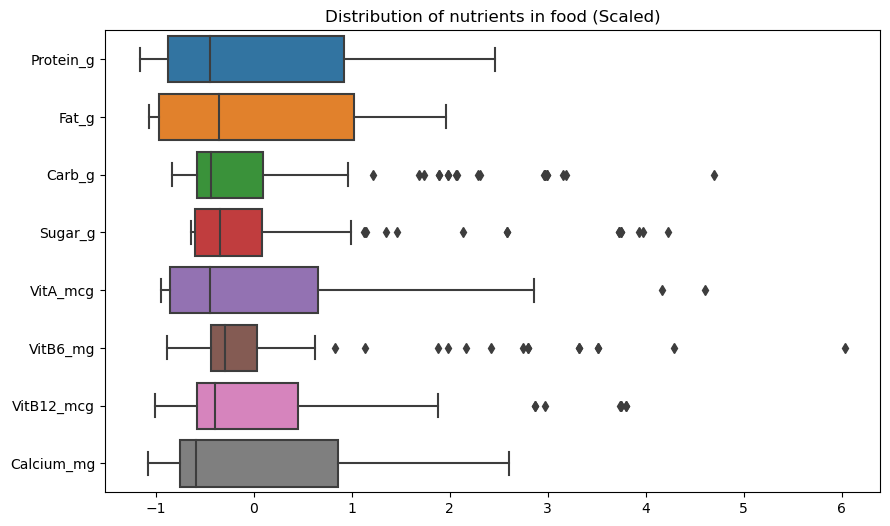
* Score plot (Shows the transformed data from the original uncentered data if no PC is discarded)
* Loading plot (Shows the direction of the original axis, after reflection and rotation such that the principal components are now the new axis)

To determine which PCs we should keep, we can

* Look at the eigenvalues (According to kaiser’s rule, we should retain PCs whose eigenvalues are greater than 1)
* Look at percentage of total variance (It is suggested that PCs that explain less than 5-10% should be discarded)
* Look at a scree plot (It is suggested to retain PCs on the left elbow)



To begin the analysis, we will apply StandardScalar as the given data has values of different units of measurement. For example, protein is measured in grams while calcium is measured in milligrams



To find out which PCs we will be retaining, we will be looking at the different methods of choosing.

Method 1 : Looking at eigenvalues

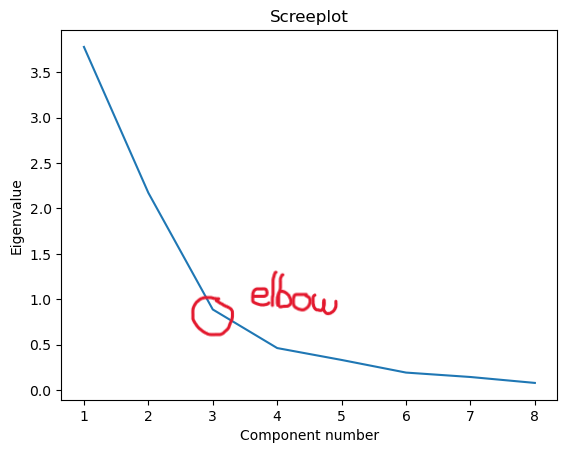


Based on Kaiser’s Rule, we should only select eigenvalues that are more than 1. Therefore, by Kaiser’s Rule, we extract the first 2 PCs whose eigenvalues are more than 1 (3.7798 and 2.1723 respectively)

Method 2 : Cumulative explained variance



The cumulative explained variance of the fist 3 PCs are to be extracted as they explain more that 80% of the total explained variance. Furthermore, they each explain more than 5-10% of variability.



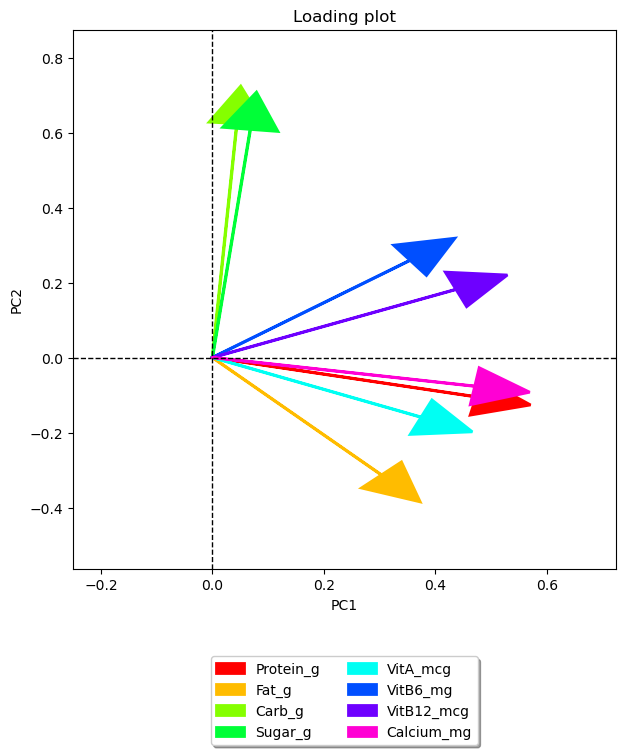
Method 3 : Scree plot

This scree plot suggests remaining the first 2 PCs only with an elbow at PC 3.

With no given requirement on how I need to select my PCs, I will use a confluence between the methods to find the appropriate number of components to keep. With Kaiser’s rule and the scree plot showing us to keep the first 2 PCs, I have decided to drop the rest.

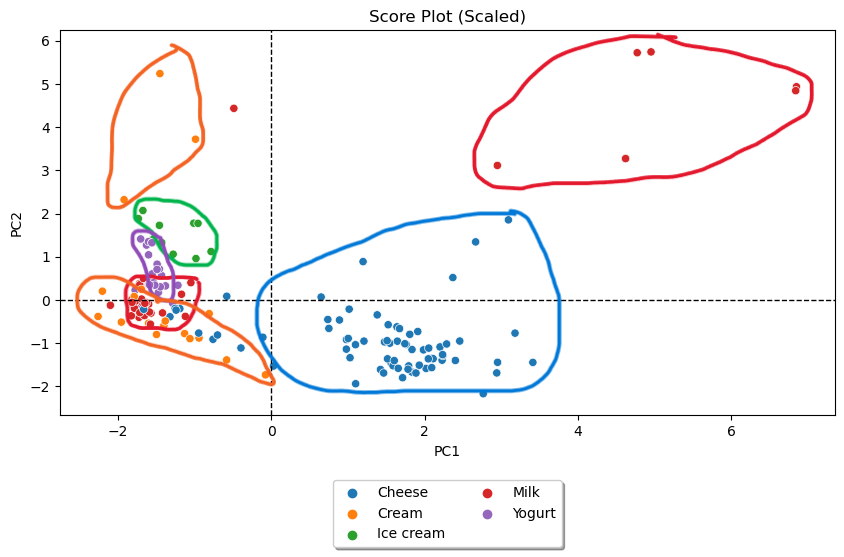
Now, we will look at the visualizations of the PCs

Loading plot :

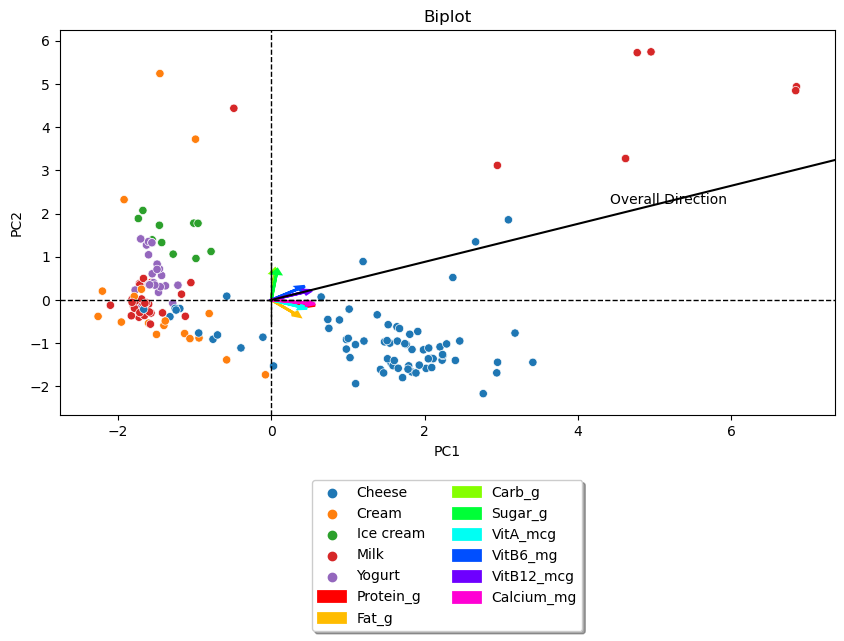


From the loading plot, we can tell that Carb\_g and Sugar\_g are highly correlated, Calcium\_mg and Protein\_g are highly correlated as well. Shorter arrows like VitA\_mcg are relatively less important than longer arrows like Carb\_g and Sugar\_g.

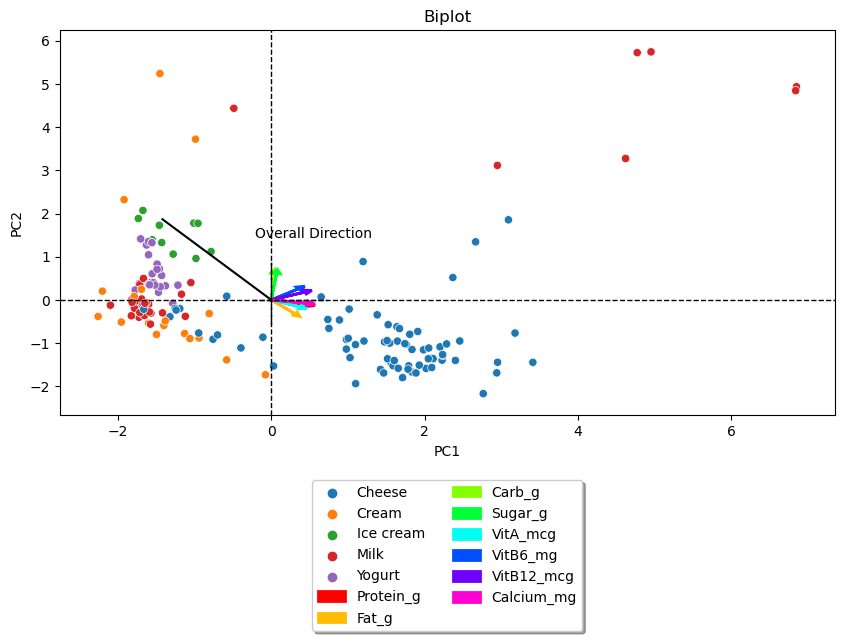
Score plot :

We can see that the different products do have some overlapping. There are types of products - Yoghurt, Milk and Cream that are very close on the score plot. There are also outliers like Milk and Cream which are very far from the rest of their product cluster. Cheese is the only product that has a clear cluster with minimal outliers.

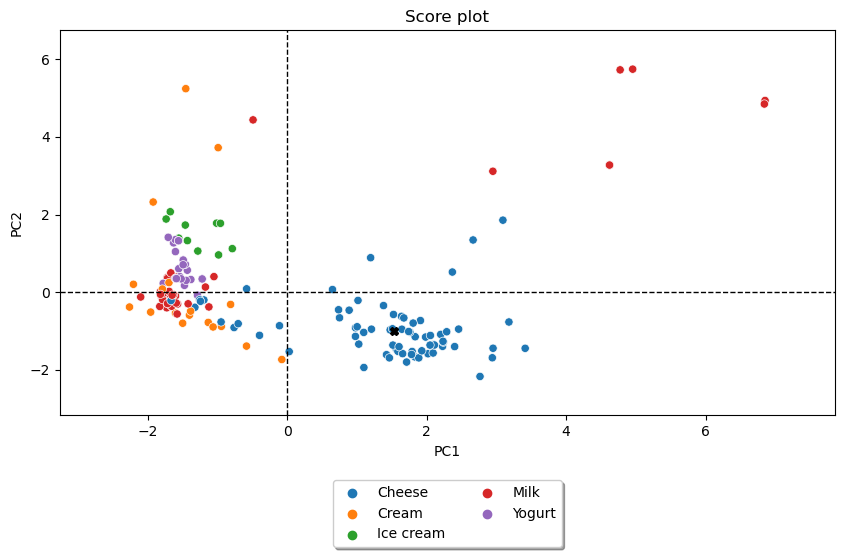
(b) Which type(s) of dairy product has/have the following attributes?

(i) Low carbohydrates and sugar but high in other nutrients. 

With taking the lowest values for Carb\_g and Sugar\_g and the highest values for the rest of the products, we have a line that generally shows the direction of what the product might be on the score plot. The product is likely to be either Milk or Cheese

(ii) High carbohydrates and sugar but low in other nutrients. 

With taking the highest values for Carb\_g and Sugar\_g and the lowest values for the rest of the products, we have a line that generally shows the direction of what the product might be on the score plot. The product is likely to be Ice Cream.

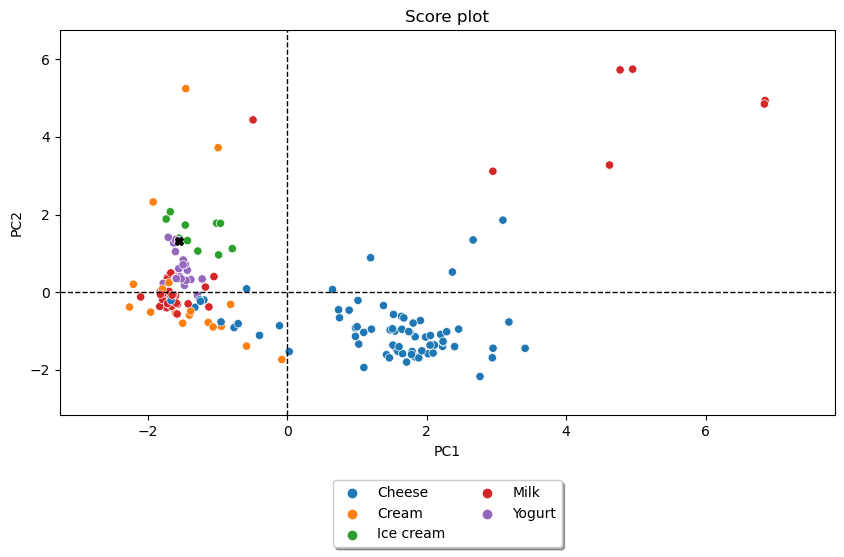


(c) 2 dairy products have their nutritional values listed below. Which type of dairy product is each of them likely to be? Use a suitable number of principal components to help you with your analysis.

Product 1:

* Protein: 22.17 g
* Vitamin A: 181 mcg
* Fat: 22.35 g
* Vitamin B6: 0.034 mg
* Carbohydrate: 2.22 g
* Vitamin B12: 2.28 mg
* Sugar: 1.01 g
* Calcium: 505 mg

The product is marked by a X on the score plot and it is likely to be Cheese

Product 2: 

* Protein: 4.32 g
* Vitamin A: 13 mcg
* Fat: 1.42 g
* Vitamin B6: 0.047 mg
* Carbohydrate: 23.0 g
* Vitamin B12: 0.53 mg
* Sugar: 14.58 g
* Calcium: 114 mg

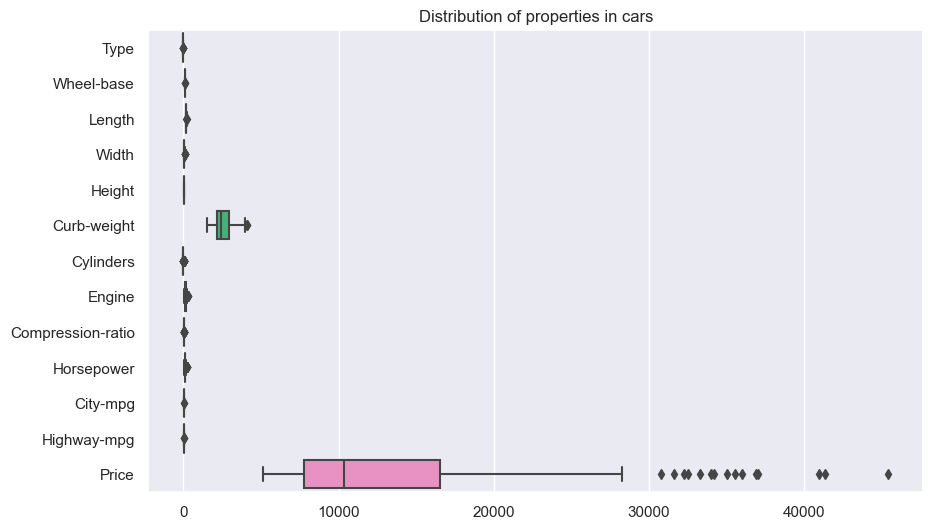
The product is marked by a X on the score plot and it is likely to be either Ice Cream or Yoghurt

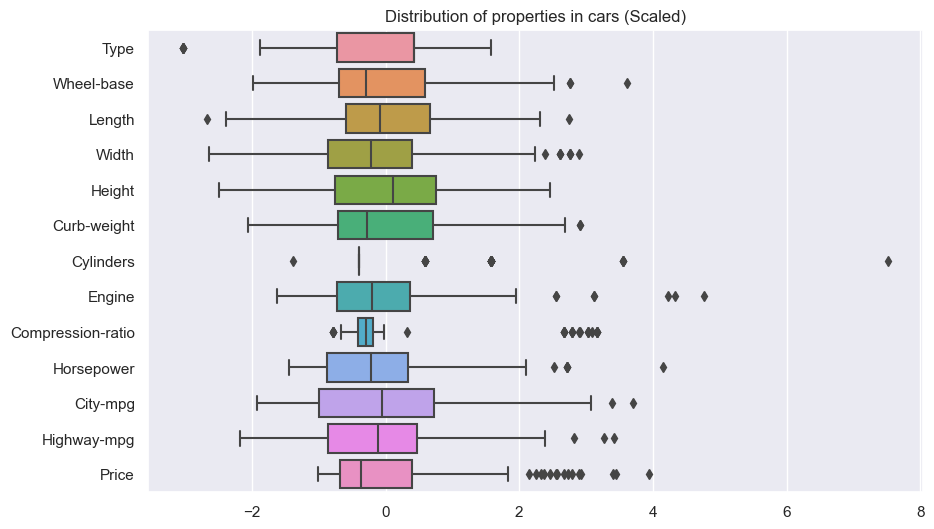
(d) Describe your observations so far, comparing what you have done in part (c) and the decision(s) you have made in the earlier part of PCA in (a). Compare how you may have expected the principal components to perform and how they have actually performed.

From the observations in part (c), we can see that the 2 PCs that were used did well in classifying the two products. I expected some overlapping between the products as the total explained variance we captured using the 2 PCs was less than 80%. This meant that there was another 20% that we could use for the data. However, using the 2 PCs to classify the product shows that using the 2 PCs could be just as good as using all the information, making PCA meaningful and useful in dimension reduction and at the same time retain most of the data.

Question 2

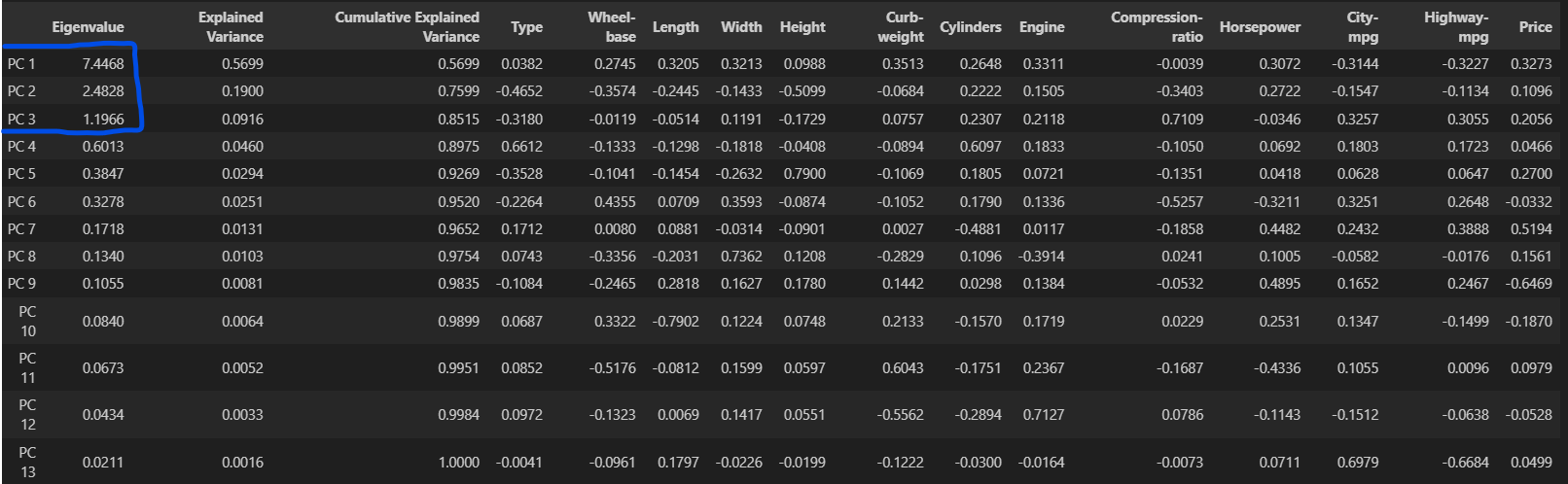
(a) Present your PCA analysis with all the necessary outputs and graphs. Explain all decisions made in the analysis.

To begin the analysis, we will apply StandardScalar as the given data has values of different units of measurement and apply LabelEncoder for categorical data. The different features have very different numerical values, for example, Price is in the 10000 range while the others are within the 100 range. Type also consists of many different categorical values like Convertible, Hatchback and Sedan

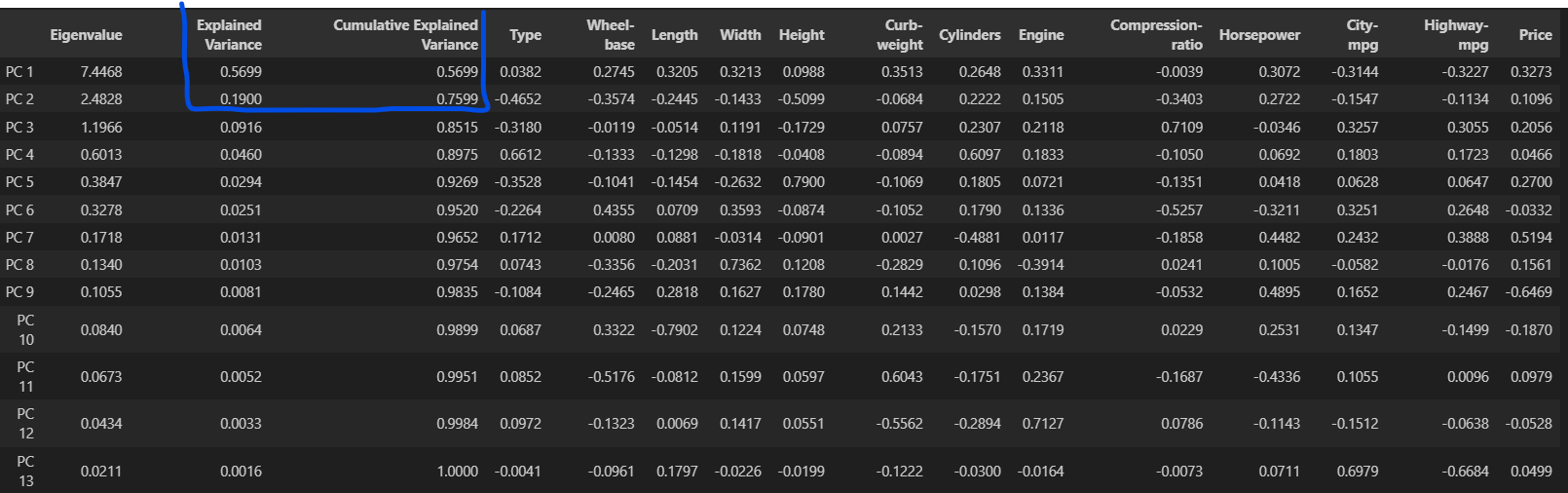
In this analysis, the goal would be to be able to use the properties of each car to find each brand.

To find out which PCs we will be retaining, we will be looking at the different methods of choosing.

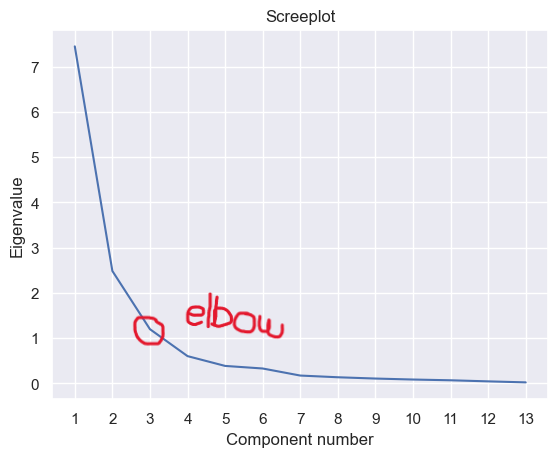
Method 1 : Looking at eigenvalues



Based on Kaiser’s Rule, we should only select eigenvalues that are more than 1. Therefore, by Kaiser’s Rule, we extract the first 3 PCs whose eigenvalues are more than 1 (7.4468, 2.4828 and 1.1966 respectively)

Method 2 : Cumulative explained variance

The explained variance of the first 2 PCs each explain more than 5-10% of variability. However, they only have 75.99% of the total variances. Although the 3rd PC would add the total variance to 85.15%, it has a explained variance of less than 10%.

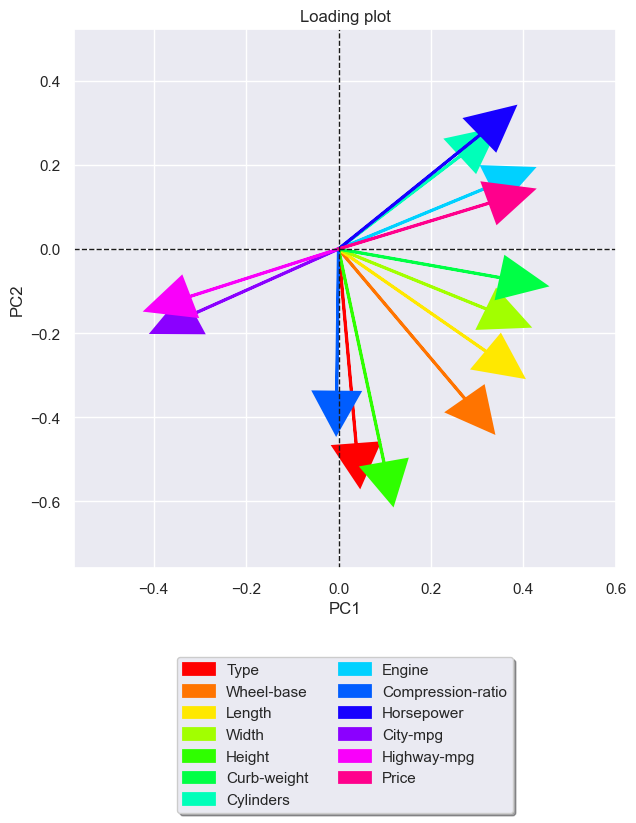
Method 3 : Scree plot 

This scree plot suggests remaining the first 2 PCs only with an elbow at PC 3.

With no given requirement on how I need to select my PCs, I will use a confluence between the methods to find the appropriate number of components to keep. With Kaiser’s rule and the scree plot showing us to keep the first 2 PCs, I have decided to drop the rest.

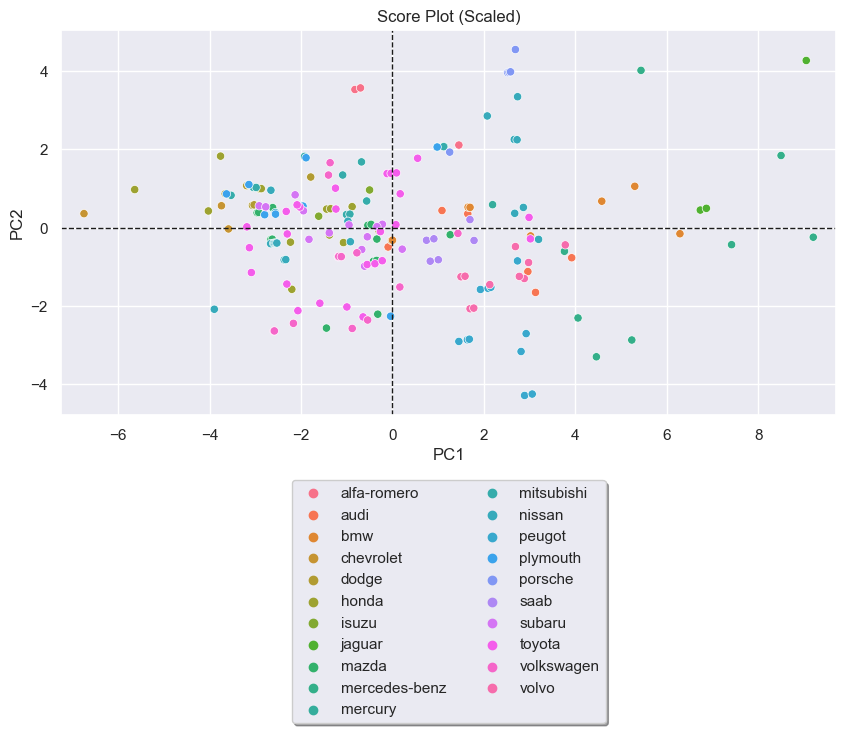
Now, we will look at the visualizations of the PCs

Loading plot :

From the loading plot, we can tell that there are many features that are highly correlated (close to 0 degrees between the two arrows) as well as features that are almost not correlated (close to 180 degrees) . Shorter arrows like cylinders are relatively less important than longer arrows like horsepower and wheel\_base.

Score plot :

From the score plot, there are no outstanding clusters between all the different brands of cars. Furthermore, there are any outliers at the parameter of the plot. They deviate significantly and are further from the rest of their car brand. There are also no visible relationships or trends in the plot.



(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 score plot in part (a) suggests that there is no distinct grouping of the car brands based on the analyzed features. The cars from different brands have relatively similar properties and it is challenging to differentiate solely based on those parameters.

With a data set like the Cars data set, where there is a difficulty in presenting the different cars, PCA does not provide any help in the classification or clustering of cars. Performing PCA could actually be doing us more harm than good as there could be more data points to show the differences in the cars.

As compared to the score plot in question 1, it suggests that there is clear segregations between the different product types, with some outliers, based on the analyzed features. The products have relatively different properties and it is not as difficult to differentiate based on the parameters.

Furthermore, with a data set like the dairy nutrition data set, where there is a clear difference between most of the products, PCA is especially useful in the classification and clustering. PCA can be used on the data set to perform dimension reduction without losing much of the information.