Name : Silviana

Class : DIT/FT/2A/14

Student ID : P1939213

Module : MS0240 - MATHEMATICS FOR AI

# Question 1

A car magazine compiled the attributes of 387 cars that were produced in year 2004.

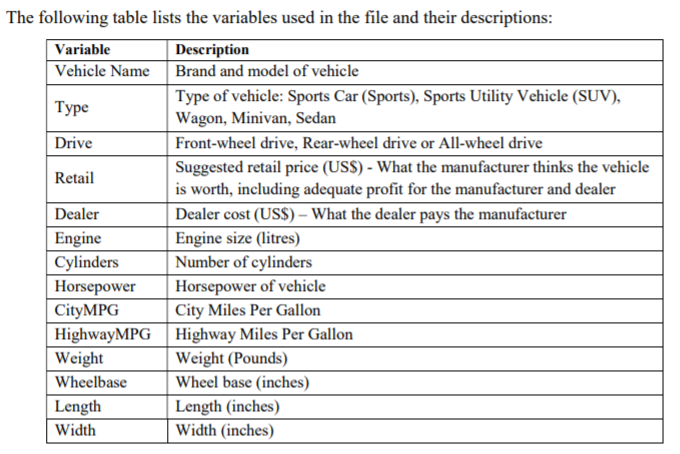


Figure 1‑1 legends

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

For this dataset, PCA should be carried out on correlation matrix because the variables are on different scales. If we want to use the covariance matrix, the data should be scaled using the sklearn.preprocessing.StandardScaler before applying PCA.

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

To choose the principal components, we are going to look at 3 criterias: the eigenvalues, cumulative explained variance, and screeplot.

A screen shot of a computer

Description automatically generated

Figure 1‑2 PCA result for all features

#1 – Kaiser’s Rule (retain PCs with eigenvalues > 1)  
From *Figure 1-2,*  we can see that only PC1 and PC2 have eigenvalues of above 1. Using this rule, we are going to consider keeping just PC1 and PC2.

#2 – Cumulative Explained Variance  
We are processing this data for descriptive purposes. Therefore, we can retain PCs whose cumulative explained variance is at least 70% to 80%. Considering this rule, we are to keep PC1 and PC2.

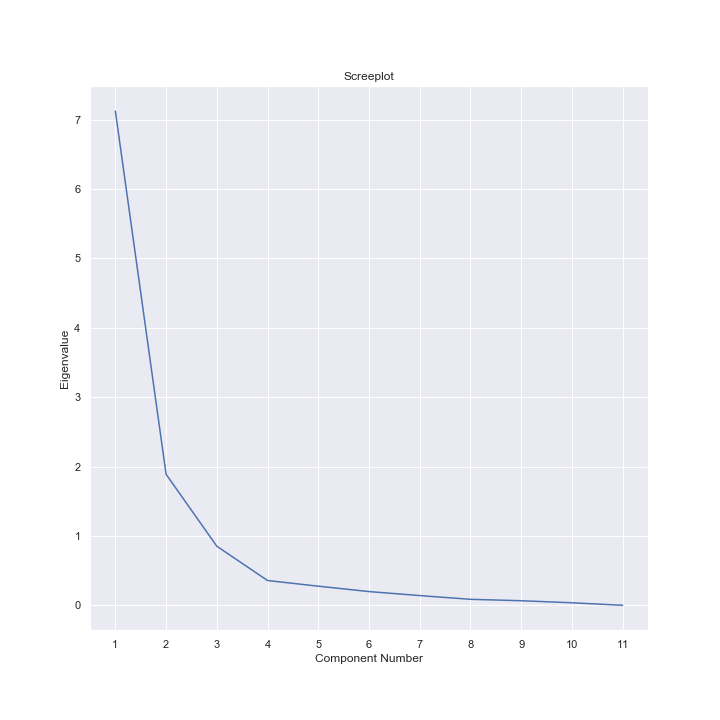


Figure 1‑3 Screeplot for all features

#3 – Screeplot  
The inflection point is located at PC4. Therefore, we are going to keep PC1, PC2, and PC3.

By looking at the combination of the 3 guidelines above, we are going to keep PC1 and PC2.

Outputs:

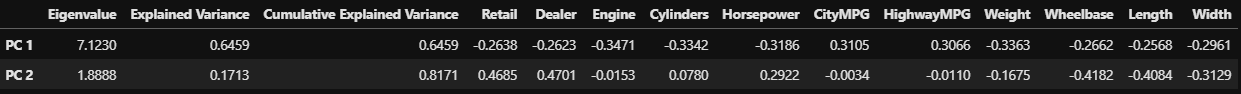


Figure 1‑4 Reduced dimensionality to 2 dimensions

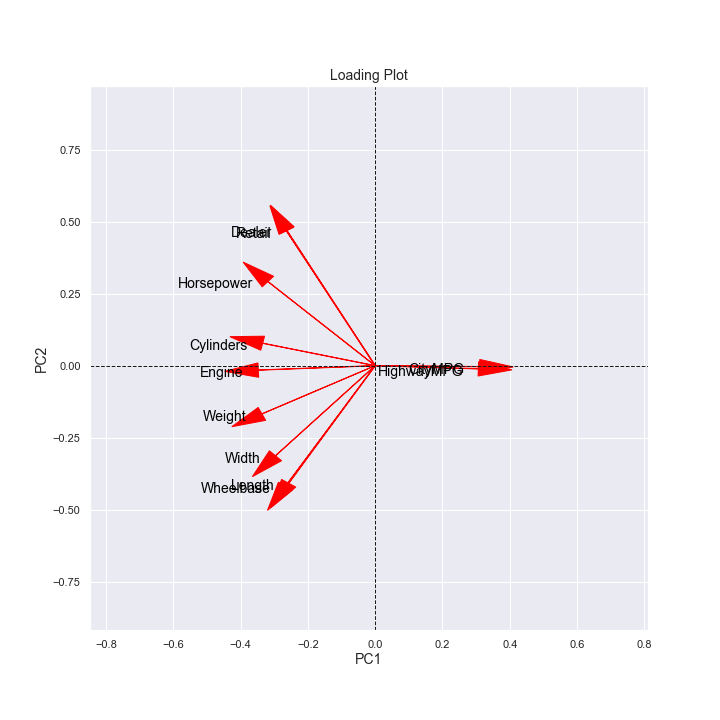


Figure 1‑5 Loading plot of PC1 and PC2

Interpretation:  
PC1 :

HighwayMPG and CityMPG are opposite in sign to the other loadings.

A car that has above average highwayMPG and cityMPG with below average retail price, dealer price, engine, cylinders, horsepower, weight, wheelbase, length and width will have a positive PC1 score.

This PC seems to measure the contrast between high efficiency and low efficiency cars.

The characterictics of cars that are represented by this principal components would be cars that are cheaper, more efficient (high mpg), less powerful (low cylinder, horsepower), and smaller in size.

PC2:

Retail, dealer, cylinders, and horsepower are opposite in sign on engine size, cityMPG, highwayMPG, weight, wheelbase, length, and width.   
The weighting on engine size, the number of cylinder, cityMPG, and highwayMPG are quite small compared to other loadings.

A car with above average retail price, dealer price, number of cylinder, and horsepower with below average engine size, cityMPG, highwayMPG, weight, wheelbase, length, and width will have a positive PC2 score.

This PC seems to measure the contrast between luxury and powerful cars against non-luxury cars. The car characteristics that are represented by this principal component are are expensive, powerful, small and does not save fuel.

* 1. Which type(s) of vehicles has/have the following attributes? Explain your answer with the aid of a suitable graph with colour or marker to display ‘Type’ information.

A picture containing man

Description automatically generated

Figure 1‑6 Score Plot

* + 1. Big size, not so expensive, but not so much horsepower.   
       left/bottom, bottom, left  
       The type of cars with these attributes would be on the bottom left section of the principal components intersection. As we can see from *Figure 1-6* many SUVs and Minivans would fall into this category.
    2. Small size, expensive, but a lot of horsepower.

Right/up, left/up, left/up

The type of cars with these attributes would be on the upper left section of the scoreplot. As we can see from *Figure 1-6*, many sports cars and sedans fall under this category.

* 1. A vehicle has the attributes listed below.   
     What type of vehicle is it likely to be? Show your working and explain.

Retail: US$46,300   
Dealer: US$41,200   
Engine: 5.7 litres   
Cylinders: 8   
Horsepower: 312   
CityMPG: 12   
HighwayMPG: 15   
Weight: 5826 pounds   
Wheelbase: 127 inches   
Length: 209 inches   
Width: 79 inches



Figure 1‑7 Mean of all numeric columns

Using the numpy.mean() method, we can get the mean of all numerical columns on the dataset. Now, we will compare the given values to the mean.

Retail > mean,   
dealer > mean,   
engine > mean,   
cylinders > mean,   
horsepower > mean,   
cityMPG < mean,   
highwayMPG < mean,   
weight > mean,   
wheelbase > mean,   
length > mean,   
width > mean

From the comparison above, we can look back at the score plot and predict that this car could be on the upper left side of the score plot because the retail and dealer prices are above average and the cityMPG and highwayMPG is lower than the mean. This car could be a sports car or a sedan or a rare SUV (because there are only a few SUV data points on the scoreplot).

# Question 2

Fisher’s Iris data set, named after British statistician and biologist Ronald Fisher, is one of the most commonly used in machine learning. The data set consists of 50 samples from each of three species or ‘class’ of Iris – Iris Setosa, Iris Virginica and Iris Versicolor. Four variables were measured from each sample – the length and the width of the sepals and petals, in centimetres. The aim is to use data to distinguish the species from each other (i.e. classification).



1. Produce scatterplot for each pair of the variables and use colour or marker to display ‘class’ information. Comment on how the irises can be distinguished and identified from your plot(s).

A close up of text on a white background

Description automatically generated

Figure 2‑1 Pairplot of Iris Variables

From figure 2-1, we can see that petal length and width are separating setosa from versicolor and virginica. However, the distinction between versicolor and virginica is not as clear.

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

For this dataset, PCA can be carried out on covariance matrix or correlation matrix because all of the features are of the same unit of measurement.

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

To extract the principal components, we are going to evaluate this dataset based on the 3 aforementioned criterias: the eigenvalues, cumulative explained variance, and screeplot.

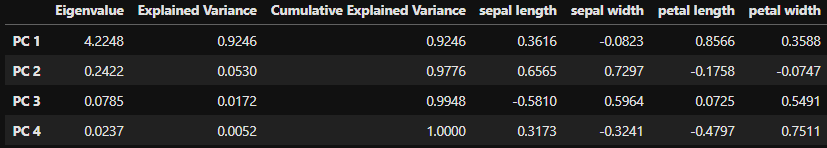


Figure 2‑2 PCA result for all features

#1 – Kaiser’s Rule (retain PCs with eigenvalues > 1)  
From *Figure 2-2* above*,*  we can see that only PC1 has eigenvalues of above 1. Using this rule, we are going to consider keeping only PC1.

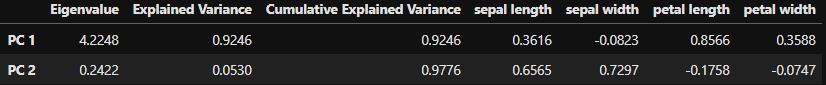
#2 – Cumulative Explained Variance  
We are processing this data for descriptive purposes. Therefore, we can retain PCs whose cumulative explained variance is at least 70% to 80%. Considering this rule, we can keep PC1 as it accounts for 92.46% of the total variance.

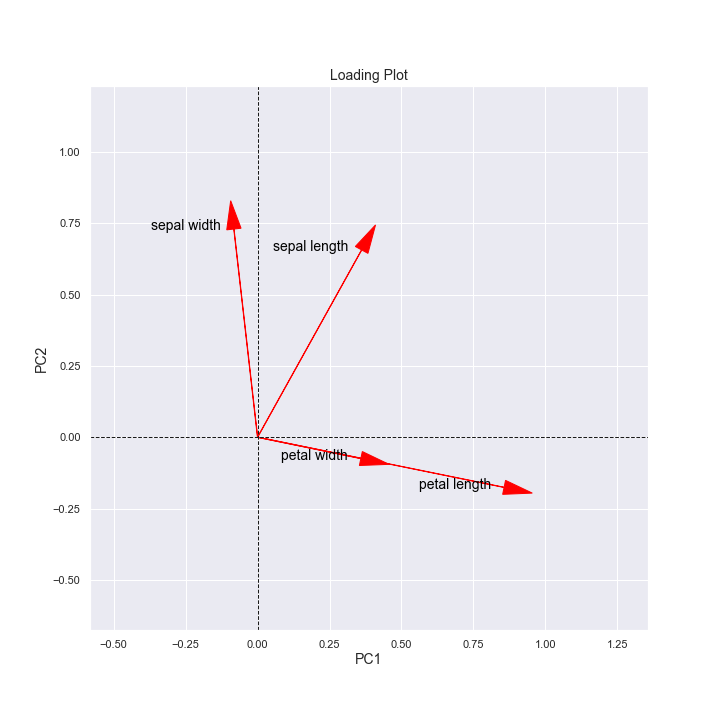
A picture containing table, white

Description automatically generated

#3 – Screeplot  
The inflection point is located at PC3. Therefore, based on this plot, we are to keep PC1 and PC2.

By looking at the combination of the 3 guidelines above, we are going to keep PC1.

Outputs:



Interpretation:

PC1:

For the first component, the loadings on sepal width and petal length are quite big. The direction of sepal width is opposite of the other measurements. This PC seems to tell us that bigger sepal length, petal width, petal length, and small sepal width would result in large positive value. This PC seems to measure the contrast between the other species and setosa because Setosa has the setosa generally has smaller sepal length, petal width, and petal length compared to the other

1. Comment on how the irises can be distinguished and identified after PCA.

A picture containing white

Description automatically generated

Iris-setosa can be identified by its small petal width, petal length, and sepal length, with big sepal width.

Iris virginica generally has larger petal width, petal length, and sepal length than iris-versicolor although they both collide sometimes.

Iris versicolor is generally everage because it’s located in the middle.