Assignment 1

Instructions:

- There are 2 questions in this assignment, complete both.
- Type your answers for both questions in a Word document. Indicate clearly the question number and part. Your submission should not exceed 10 pages.
- Present your Python code for questions 1 and 2 in separate Jupyter notebooks, i.e. you should submit two Jupyter notebooks, one for each question. Indicate clearly which question number and part the code is for.
- Submit the *Declaration of Academic Integrity* before submitting your assignment.

Question 1 (60 marks)

A car magazine compiled the attributes of 428 cars that were produced in year 2004. The data can be found in the file *cars.csv*.

The following table lists the variables used in the file and their descriptions:

Variable	Description
Vehicle Name	Brand and model of vehicle
Type	Type of vehicle: Sports Car (Sports), Sports Utility Vehicle (SUV), Wagon, Minivan, Sedan
Drive	Front-wheel drive, Rear-wheel drive or All-wheel drive
Retail	Suggested retail price (US$) - What the manufacturer thinks the vehicle is worth, including adequate profit for the manufacturer and dealer
Dealer	Dealer cost (US$) – What the dealer pays the manufacturer
Engine	Engine size (litres)
Cylinders	Number of cylinders
Horsepower	Horsepower of vehicle
CityMPG	City Miles Per Gallon
HighwayMPG	Highway Miles Per Gallon
Weight	Weight (Pounds)
Wheelbase	Wheel base (inches)
Length	Length (inches)
Width	Width (inches)

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

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

(c) 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.

(i) Big size, not so expensive, but not so much horsepower.

(ii) Small size, expensive, but a lot of horsepower.

(d) 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	

Question 2 (40 marks)

Fisher's Iris data set, named after British statistician and biologist Ronald Fisher, is one of the most commonly used in machine learning. Refer to the data file *iris.csv*, 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).

(a) Produce scatterplot for each pair of the variables. You may produce a scatterplot matrix using *pairplot* from *Seaborn* library, or simply individual scatterplots using *scatter* from *Matplotlib*. Either way, use colour or marker to display '*class'* information. Comment on how the irises can be distinguished and identified from your plot(s).

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

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

(d) Comment on how the irises can be distinguished and identified after PCA.
