EE2211 Tutorial 2 (Python coding)

(Data Reading and Visualization, simple data structure)

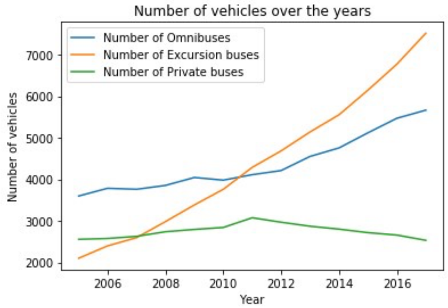
Question 1:

A Comma Separated Values (CSV) file is a plain text file that contains a list of data. These files are often used for exchanging data between different applications. Download the file "government-expenditure-on-education.csv" from https://data.gov.sg/dataset/government-expenditure-on-education. Plot the educational expenditure over the years. (Hint: you might need "`import pandas as pd`" and "`import matplotlib.pyplot as plt`".)

(Data Reading and Visualization, slightly more complicated data structure)

Question 2:

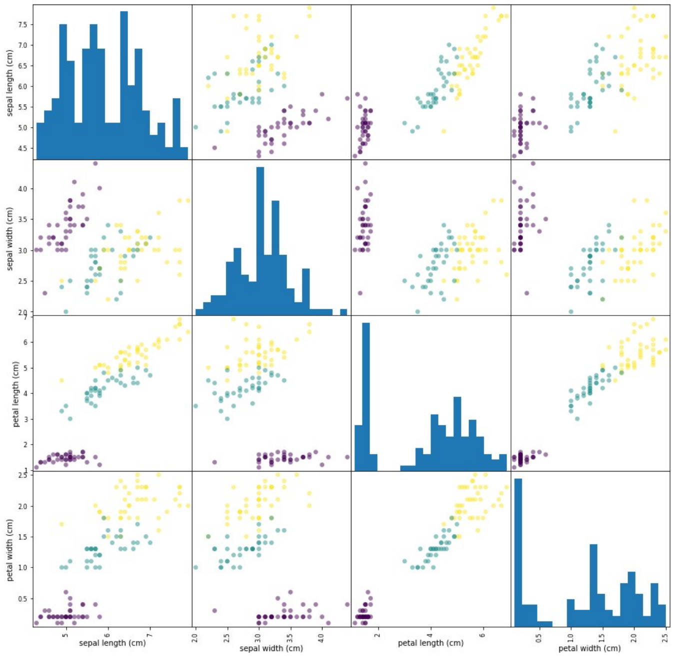
Download the CSV file from https://data.gov.sg/dataset/annual-motor-vehicle-population-by-vehicle-type. Extract and plot the number of Omnibuses, Excursion buses and Private buses over the years as shown below. (Hint: you might need "`import pandas as pd`" and "`import matplotlib.pyplot as plt`".)



(Data Reading and Visualization, distribution)

Question 3:

The "iris" flower data set consists of measurements such as the length, width of the petals, and the length, width of the sepals, all measured in centimeters, associated with each iris flower. Get the data set "from sklearn.datasets import load_iris" and do a scatter plot as shown below. (Hint: you might need "`from pandas.plotting import scatter_matrix`")



(Data Wrangling/Normalization)

Question 4:

You are given a set of data for supervised learning. A sample block of data looks like this:

```
“       1.2234, 0.3302, 123.50, 0.0081, 30033.81, 1
        1.3456, 0.3208, 113.24, 0.0067, 29283.18, -1
        0.9988, 0.2326, 133.45, 0.0093, 36034.33, 1
        1.1858, 0.4301, 128.55, 0.0077, 34037.35, 1
        1.1533, 0.3853, 116.70, 0.0066, 22033.58, -13
        1.2755, 0.3102, 118.30, 0.0098, 30183.65, 1
        1.0045, 0.2901, 123.52, 0.0065, 31093.98, -1
        1.1131, 0.3912, 113.15, 0.0088, 29033.23, -1          ”
```

Each row corresponds to a sample data measurement with 5 input features and 1 response.

(a) What kind of undesired effect can you anticipate if this set of raw data is used for learning?

(b) How can the data be preprocessed to handle this issue?

(Missing Data)

Question 5:

The Pima Indians Diabetes Dataset involves predicting the onset of diabetes within 5 years in Pima Indians given medical details. Download the Pima-Indians-Diabetes data from https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv.

It is a binary (2-class) classification problem. The number of observations for each class is not balanced. There are 768 observations with 8 input variables and 1 output variable. The variable names are as follows:

0. Number of times pregnant.
1. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
2. Diastolic blood pressure (mm Hg).
3. Triceps skinfold thickness (mm).
4. 2-Hour serum insulin (mu U/ml).
5. Body mass index (weight in kg/(height in m)^2).
6. Diabetes pedigree function.
7. Age (years).
8. Class variable (0 or 1).

(a) Print the summary statistics of this data set.

(b) Count the number of “0” entries in columns [1,2,3,4,5].

(c) Replace these “0” values by “NaN”.

(Hint: you might need the “`.describe()`” and “`.replace(0, numpy.NaN)`” functions “`from pandas import read_csv`”.)

(In Quiz and Exam format)

Question 6:

Disease Outbreak Response System Condition (DORSCON) in Singapore is a colour-coded framework that shows the current disease situation. The framework provides us with general guidelines on what needs to be done to prevent and reduce the impact of infections. There are 4 statuses – Green, Yellow, Orange and Red, depending on the severity and spread of the disease. Which type of data does DORSCON belong to ?

(1) Categorical; (2) Ordinal; (3) Continuous; (4) Interval

(In Quiz and Exam format)

Question 7:

A boxplot is a standardized way of displaying the dataset based on a five-number summary: the minimum, the maximum, _*BLANK1*_, and the first and third quartiles, where the number of data points that fall between the first and third quartiles amounts to _*BLANK2*_ percent of the total number of data on display.