‘Numpy’ and ‘Pandas’ function libraries are used to process, integrate and analysis the data sets. ‘Matplotlib.pyplot’ and ‘seaborn’ are used to draw the graphs in python.

The first thing to do is reading the CSV files. Finding out the data series that we will use for further steps and slice them off. Because the values are read as ‘strings’, therefore, we also need to convert ‘strings’ to ‘float’. Replacing the missing value by average number. Combining those data series in one data table.

Next, a few of boxplots were drawn to look for if there are suspected outliers in the data series. Scatter plots show the correlation between two data sets in a direct-viewing way. The three scatter graphs have ‘Percentage of people buying private health insurance’ as the y-axis and five kinds of ‘Premature death rate’ as the x-axis. In the end of this part, I print out the Pearson correlation coefficients between every two data sets in a table.

The next part is about heatmap. The correlation table was drawn as a heatmap, so it’s easier to see the how do the two data sets correlate each other, strong or weak.

The entropy function is copied from the workshop 7 exercises. Before calculating the entropy, I divide the ‘Percentage of people buying PHI’ and ‘Premature death rate’ into 3 buckets, which are the lowest one-third, middle one-third and highest one-third, name them as ‘Division of PHI’ and ‘Division of death’. Adding the two data series into the initial data table. Using the two data sets to calculate the mutual information about entropy.

The final part is about classification. It includes the decision trees and KNN. The 66% data from the data table are used to be training set, the rest data are the test set. The ‘total death’, ‘lung cancer’, ’circulatory system disease’, ‘colorectal cancer’ and ‘Ischaemic heart disease' are the testing array, the ‘Division of PHI’ is the classlabel. Comparing with 5 nearest neighbours to estimate the result of the test set. In the end, the test accuracy is 0.41, that means the correlation between two data sets is weak. I also plotted a graph with the accuracy by choosing different numbers of nearest neighbours. From the graph, we can see the highest accuracy happen when we choose about 1,29 or 30 nearest neighbours. Whatever the number of nearest neighbour change, the accuracy is below 0.55.