**Full Data Science Pipeline**

*Foundations in Data Science B7IS136*

Justin Lazarev

STUDENT NO. 10567499

Submission Date:

Word Count:

# Initial impressions of the Data and Preliminary Steps Taken

Upon opening the csv file, initial impressions of data dictated to several data cleaning methods in need of implementation. I would then create target variables that would be used to predict Life Expectancy.

I began this pipeline project by importing the data set and all libraries that I would need to fully clean, plot and regress the data. These libraries include pandas, matplotlib, seaborn and numpy. For the regression I utilised the sklearn library. I used a few of the simple data frame viewing commands initially to view specifics about the data such as the columns and the data type. I renamed the columns as some of the names contained blank space characters in strange places that caused commands later to not function as intended. I also used the describe command to get an initial impression of the simple statistics of each column such as mean and standard deviation.

# Data Cleaning

For the data cleaning section of the project I employed a command that listed the columns with null values in descending order. For columns that contained more than 200 null values I removed the whole column. This was true for 4 columns. For other null values, I removed the null rows.

After this, I plotted a boxplot to view outliers of the dataset. With the low correlation of measles to life expectancy, I can remove the measles column among others to further improve my accuracy with my prediction model within the regression cell.

# Feature Selection

Feature selection involved creating a correlation matrix to show all correlations with the remaining columns. Using this matrix, I could view which of the data was interesting to plot and perform linear regression upon. I decided to plot Life Expectancy. I also decided to remove Infant deaths, Measles and Under-Five Deaths as their correlation was extremely low and wouldn’t influence the analysis in a positive way.

# Linear Regression

With the linear regression I removed the columns with low correlation in the variable assignment cell so that if the removed columns were to be used, then the user would just have to not run the cell. The x and y variables are assigned with y being the life expectancy column and the x being every other column in the dataset.

In the next cell, I printed the variables and split the data into training sets and testing sets. Training sets is a subset used to train a model. A test set is a subset to test the trained model. I then made a variable that was the Linear Regression method. This variable(method) was then used for the training sets.

I then found the mean absolute error, mean squared error and r2 score. These statistics were found using the predicted value for y and the test value for y.

* Mean Absolute Error – The average of the errors between predicted and real data.
* Mean Squared Error – Measures the average of the squares of the errors.
* R2 score – the ratio of the variance in the predicted data versus the observed data.

I then plot both the test y values and the predicted y values against each other and inserted a best fit line.

# Visualisations

The visualisations here have a strong correlation. According to the data, Life Expectancy and Schooling relate heavily. Having more schooling indicates longer life. Similarly having more schooling indicates higher alcohol usage.