**Supervised Machine Learning - Regression**

#### Ankit Butola(105902492)

Contents

[Group Report 3](#_TOC_250008)

[INTRODUCTION 4](#_TOC_250007)

[AIM 6](#_TOC_250006)

[Independent and dependent variables 7](#_TOC_250005)

[Data Preparation 8](#_TOC_250004)

[Feature Selection 20](#_TOC_250003)

[Model Development and Evaluation 22](#_TOC_250002)

[Model Comparison 25](#_TOC_250001)

[Bibliography 26](#_TOC_250000)

# Group Report

For this project, we needed to identify individuals from our class to establish a group to perform the given tasks, thus we organized a group of three people: 1. Muktesh Sahu, 2. Maya Kumbhar, 3. Ankit Butola, Due to the college hybrid class approach and ongoing pandemic situation, it was extremely impossible to meet physically, however the assignment specifics were shared and communicated using WhatsApp Groups, and Anydesk and managed to set the college meet up occasionally. We then held Zoom meetings on a regular basis. We talked more about the format of tasks and appropriate datasets for these assignments during these discussions. After some productive discussions on dataset research, we've completed the Life Expectancy dataset, which includes all of the desired variables, both dependent and independent, that will be needed to complete the CA1 tasks. Each of the team members was assigned a portion of the assignment to accomplish and eventually everyone input was pulled together to complete this project.

# INTRODUCTION

Our data set is about life expectancy, or the estimated number of years a person will live based on statistics. It is determined by the geographical location of the place. The data for multiple countries on life expectancy and health variables came from the pharmacy data repository. It has been observed that, in comparison to the previous 3 decades, there has been a significant improvement in the health domain which resulted in lower human mortality rates, particularly in various emerging countries. The majority of the missing data was for population, Hepatitis B, and GDP, according to the findings. Data was lacking from lesser-known countries such as Vanuatu, Tonga, Togo, Cabo Verde, and others. We have followed Crisp data modelling methodology followed with extraction, transformation and finding a best suitable model/algorithm to predict the target values.

**h"ariable Descriptions:**

1. countrv (Nominal) - the county in o hich the indicators are from (i.e. jotted States of

.-tozerica or Coztgo)

1. year (Ordixtal) - the calendar '<» a< indicators are from (ranging from 2000 to 2015)
2. status orriinaJ) - whether a counizn' *14* considered to be developing' or Developed"

by IT£O 4t&Ddards

1. life\_expectancv (Ratio) - the life expect of people in years for a particular

coontzn' and year

?. adu1t\_cnorta11h’ (Ratio) - the adult mortaliD’ rate per 1000 population (i.e. number of people ds’inr beni’eexi IN and 60 i’ears per 1000 population). if the rate is 263 then that means 263 people v ilJ die out of 1000 benveen the ages of 1 I and 60. another

2fi.3 °.z

fi. iofaot\_deaM (RaBo) - number of infant deal per 1000 population: similar to above.

but for infants

alcohol (Ratio) - a *cost"4* alcobol consumption rate eiemured W *ñl* T *4* Of pure alcohol conception per capita

1. percen4age\_expenditure (Ratio) - expenditure on hearth as a percentage of Gross

Doxo.estic Product (gdp)

1. hepatitii\_b (R2ftlD) - number of 1 5’ear olds o’ith Hepatitis B immunization over all I

year olds in population

1. ieasle s (Ratio) - number Df reported kleasles cases per 1000 population
2. bmi (Itztm a1 Ordinal) - average Body kMs Index (B AH) of a count ‘s totai

population

IN. under-five\_deaM (Ratio) - number of people under a< aee of fii’e deaths per

1000 population

). OliO QUID) - O42Otber of 1 year olds o ith Polio immunization over &e number of all

1 year olds izi population

1. totaJ\_expenñture (Ratio) - government expenditure on Health as a percentage of total gm’ernment expenditure
2. diphtheria (Rate) - Diphtheria tetanus toxoid and pertussls (DTP3) immunization rate

of 1 year o1d4

lfi. Uz aids (Ratio) - deaths per 1000 live b caund A H\"‘.AIDS for people under 5. number of people under ? v ho die due 4o HR" .AIDS per 1000 b .

1 7. gdp (Ratio) - Gro ss Domestic Product per capi4a 1 8. popul£ftion (Ratio) - popuJatloo Df a count’

1. thintiess\_1-19 5’ears (Ratio) - rate of throne zz among people aged 10-19 (Note: variable should be renamed to thintiess\_lO-IS ’ears to xnore accurateh’ represent ae variable)
2. ess\_5-9\_ ears (Ratio) - ia4e of Princess amoog people aged ?-9
3. income\_coD2QDS1tion\_of\_resources (Ratio) - Humazi De ’elopesent Index in terms of

income composition of remurces (index ranging from 0 to 1)

2?. schooling (Ratio) - ax crate number of i ears of schooling of a populaBon

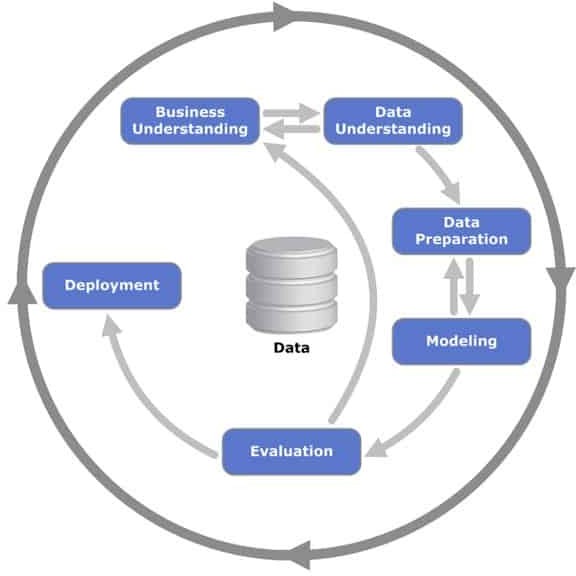
# AIM

Our goal is to learn about the criteria that the World Health Organization utilizes to forecast people's life expectancy in different countries. With the use of simple Linear regression and using optimisation algorithms such as stochastic gradient descent and eventually comparing the models based on the accuracy and other evaluations.

METHODOLOGY and DATASET

**DataSource:** [https://www.kaggle.com/kumarajarshi/life-expectancy-](https://www.kaggle.com/kumarajarshi/life-expectancy-who) [who](https://www.kaggle.com/kumarajarshi/life-expectancy-who)

**Methodology Used:** Crisp-DM



# Independent and dependent variables

The variables in a study of a cause-and-effect relationship are called the independent and dependent variables.

* The independent variable is the cause. Its value is independent of other variables.
* The dependent variable is the effect. Its value depends on changes in the independent variable.

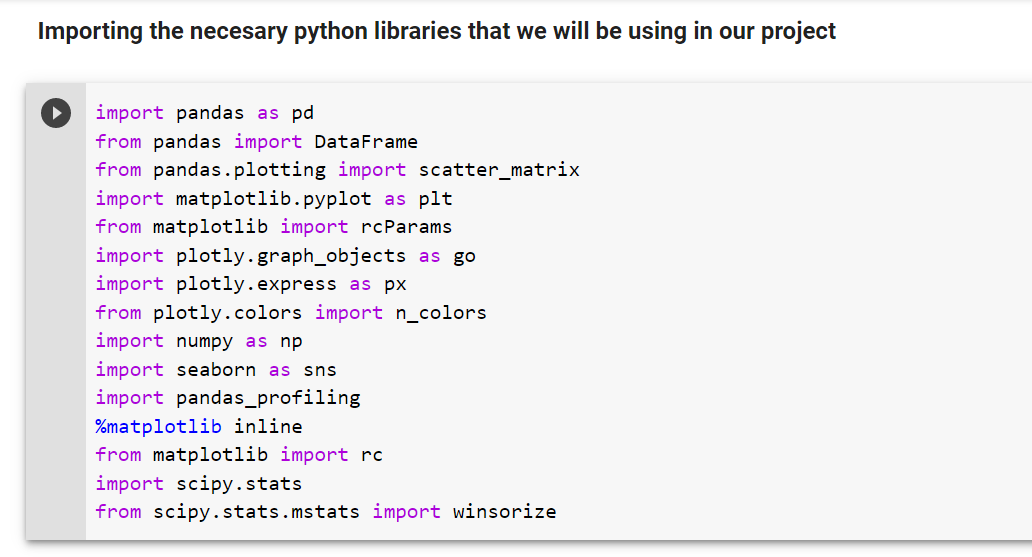
In the Life Expectancy dataset that we have chosen, the variable **life\_expectancy** is the dependent variable as it describes the ratio of the person's life. Whereas, below is the list of independent variables that can cause variations in the **life\_expectancy** figure(the variable country and status has been removed from the list of independent variables as it will not affect the expectancy rate much as compared to the others):

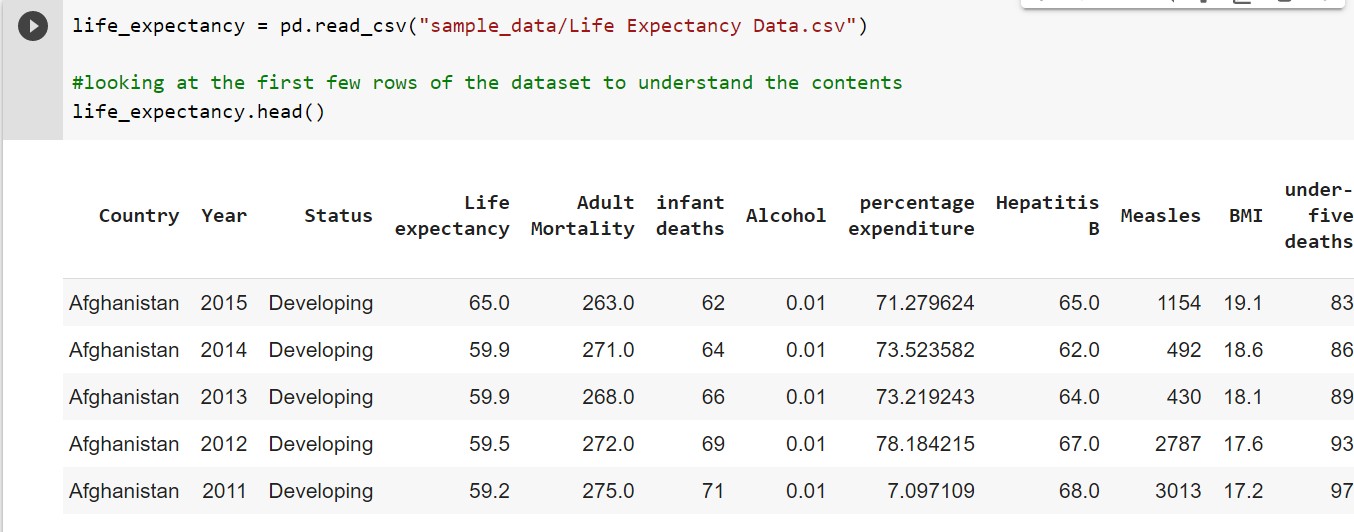


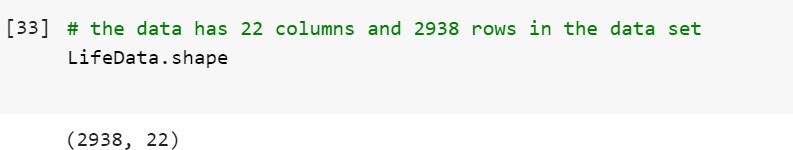
**Linear regression model** can be used in this scenario as the dependent variable is continuous in nature and will predict the value of a life\_expectancy based on the value of dependent variables.

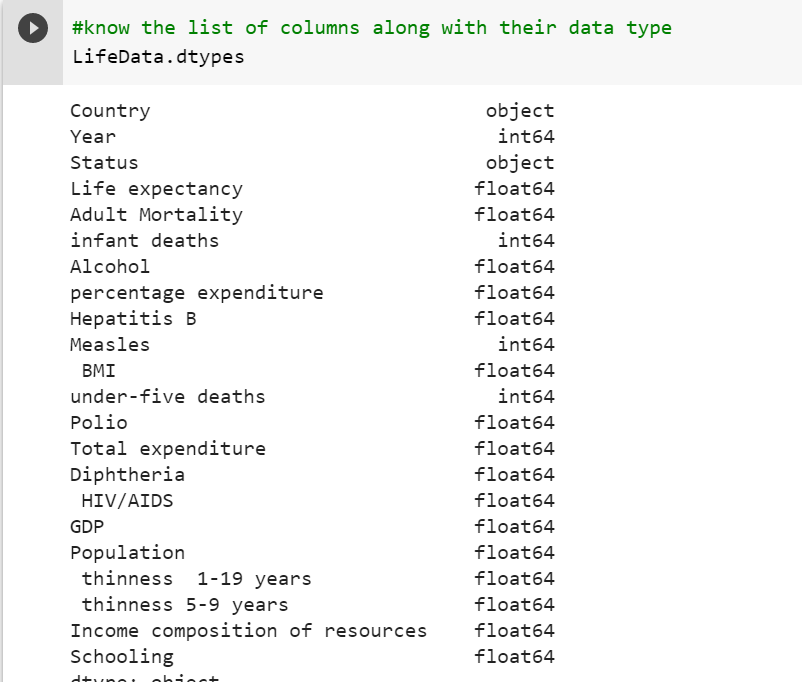
# Data Preparation

Before starting with the data preprocessing let us first understand the data by reading it, importing necessary python libraries and the dataset

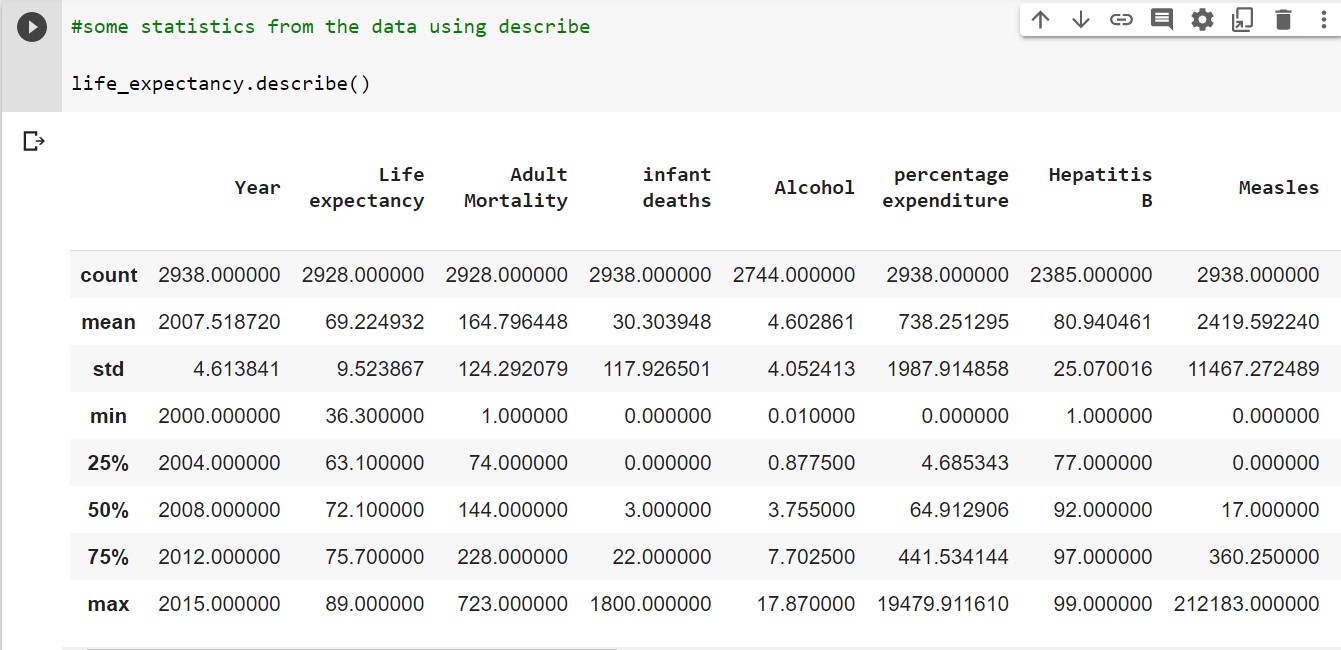






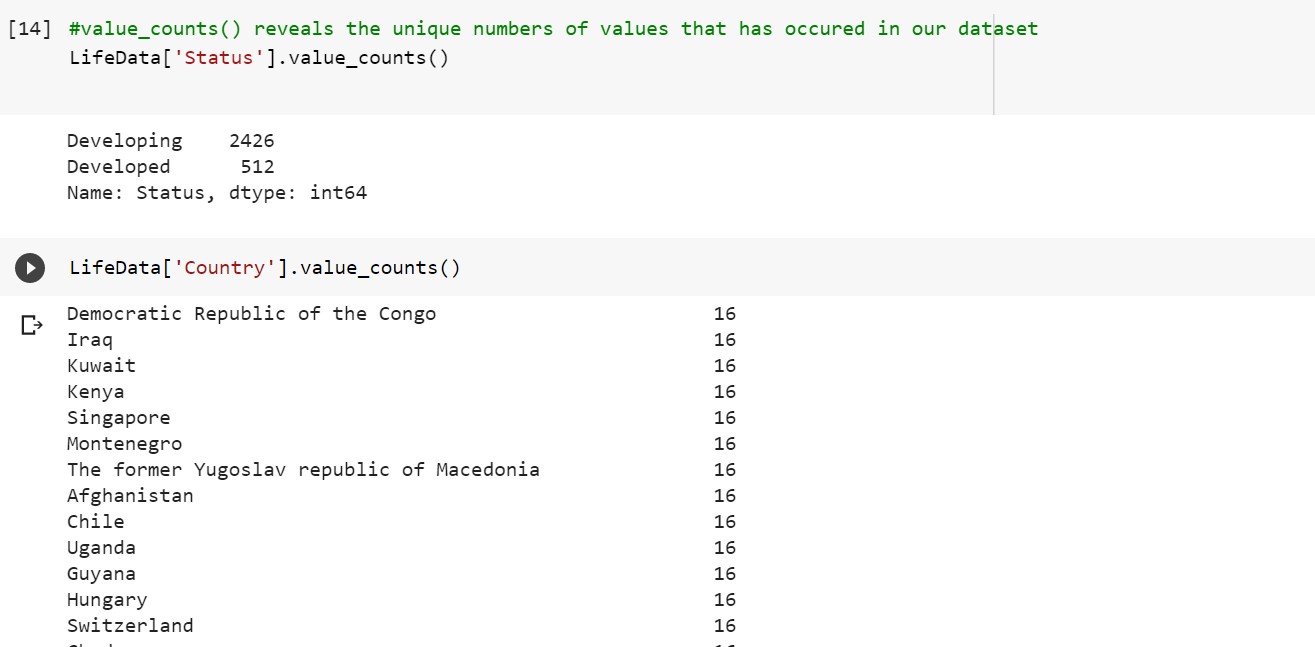


Now let's look at some statistics from the data using Pandas' describe function, which is used to display some basic statistical information of a data frame or a series of numeric values such as percentile, mean, std, and so on.

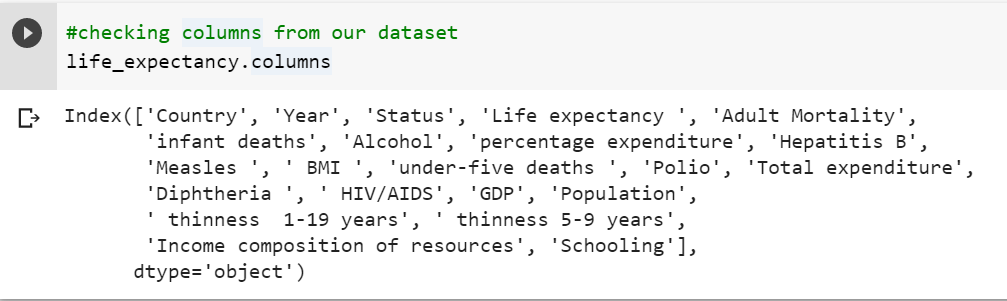


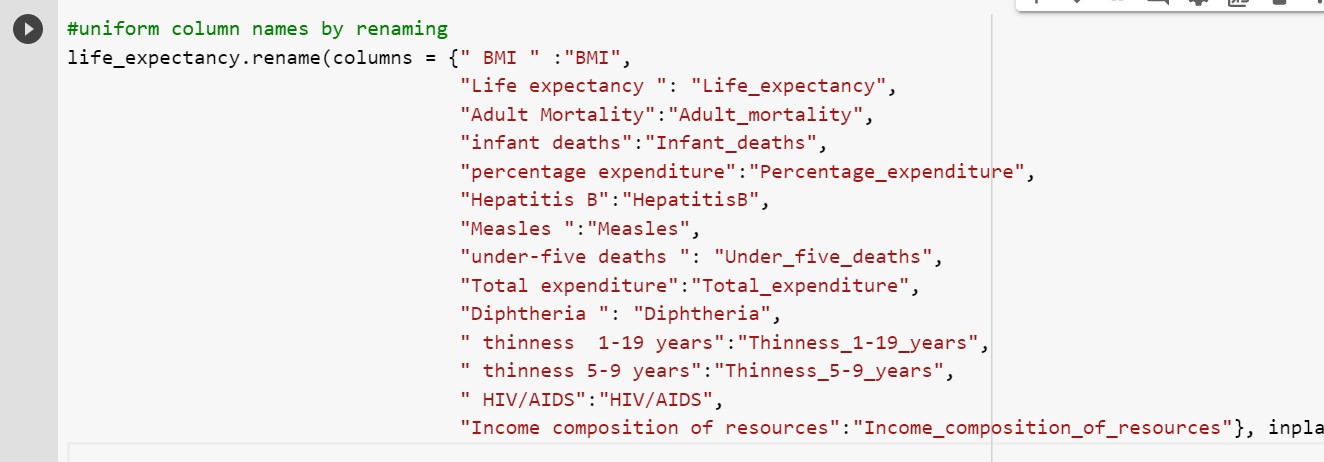
We can also check the unique number of values that has occurred in our dataset,

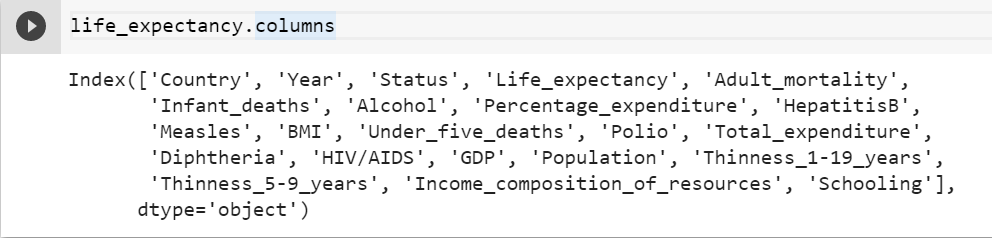
for e.g.



Let us check the attributes/variables from our set and to identify the categorical variables



So, in our dataset we have only two categorical variables which are country and status. Also, from the above script we can also see that the column names are not well uniformed. Now let us uniform the columns,

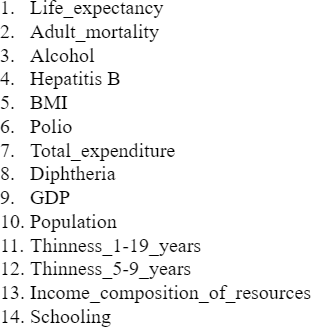


Data Cleaning

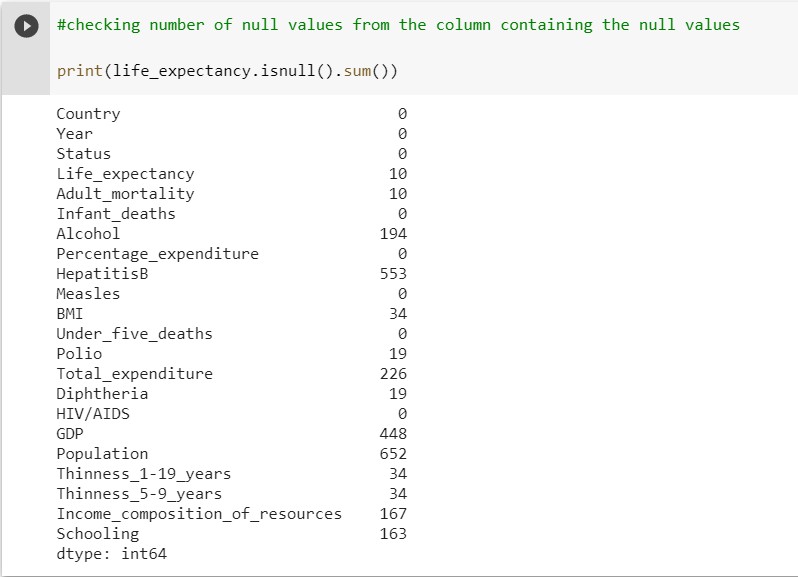
Let us firstly start with eliminating the null values form the dataset:



As per the above output, we can say that below columns contains the null values:

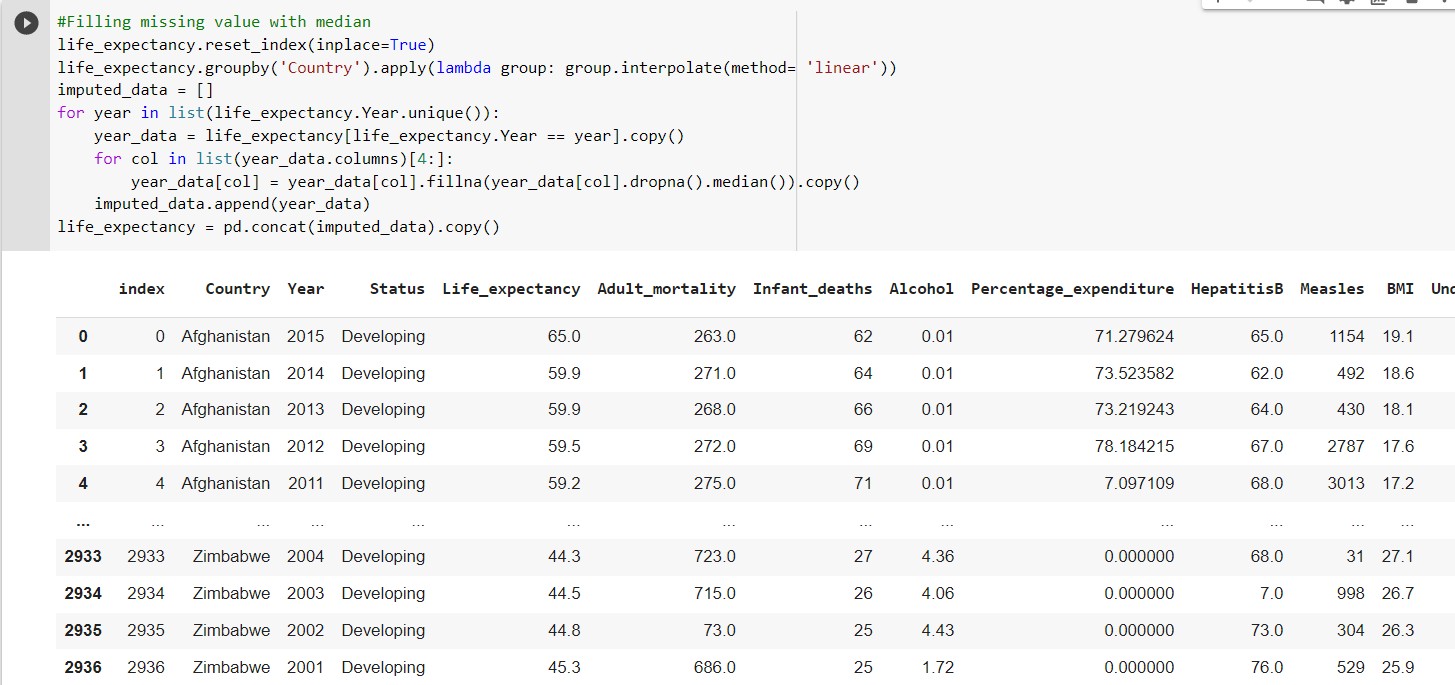


Almost all the columns contain null values in our dataset, we can also check the frequency of null values in particular variable.



Many columns have null values, but the total number of missing values is insufficient to eliminate the columns. As a result, substituting missing values is a smart idea. All columns with missing values are numeric continuous variables, we also know.

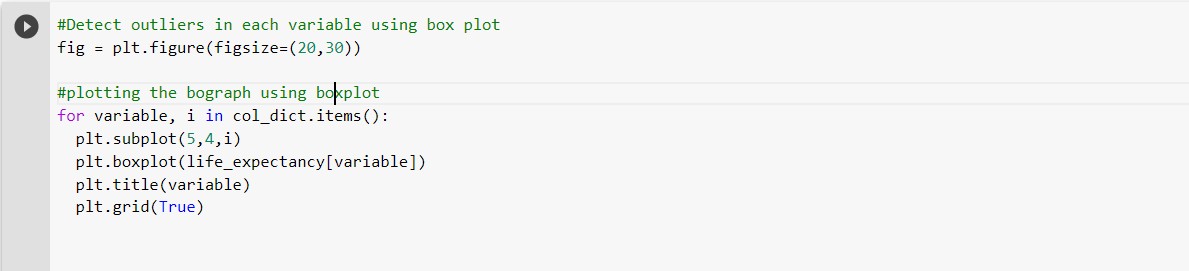
Because of the outliers, using a central tendency average to fill in the missing numbers is not a good idea. We may also use the median to fill it:

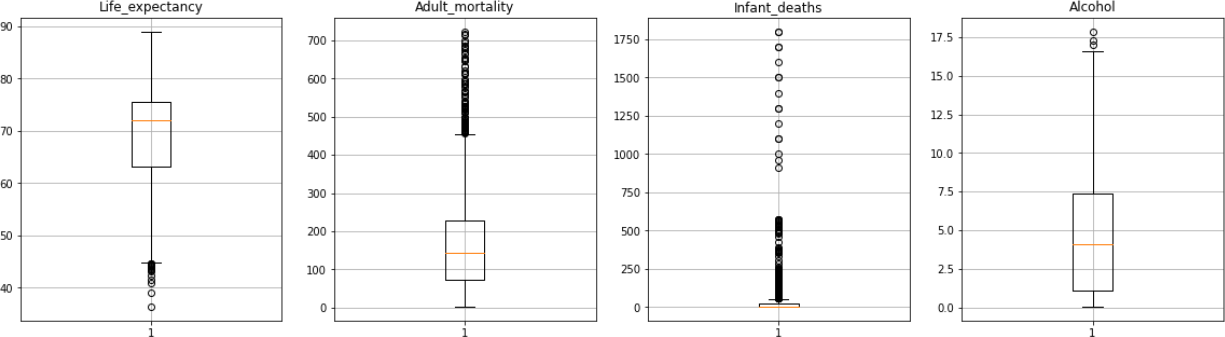


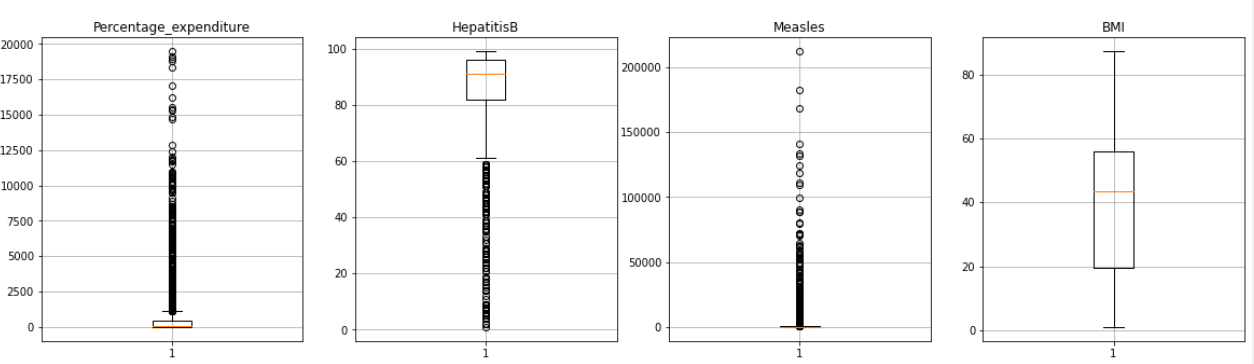
Removing Outliers:

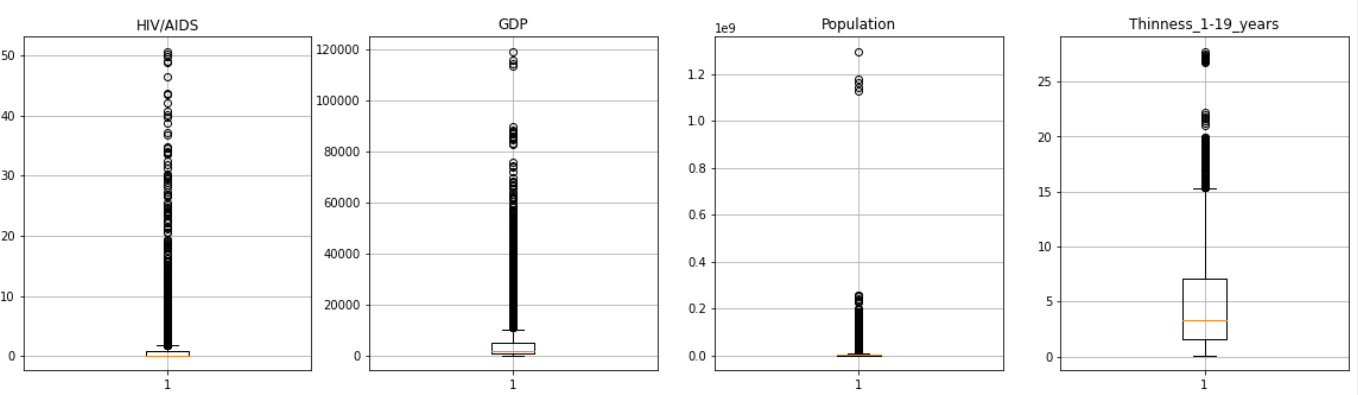
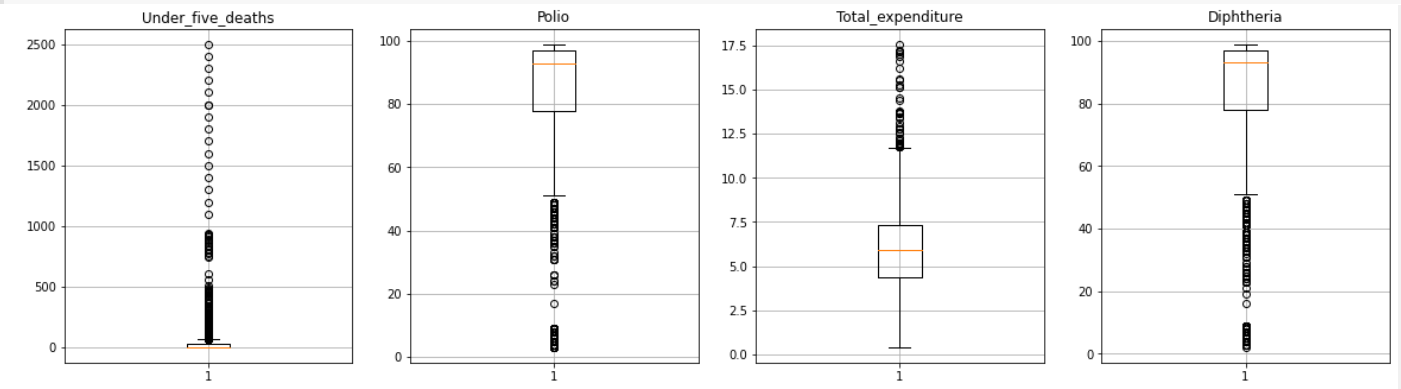
Outliers are extreme values that fall a long way outside of the other observations. For example, in a normal distribution, outliers may be values on the tails of the distribution.

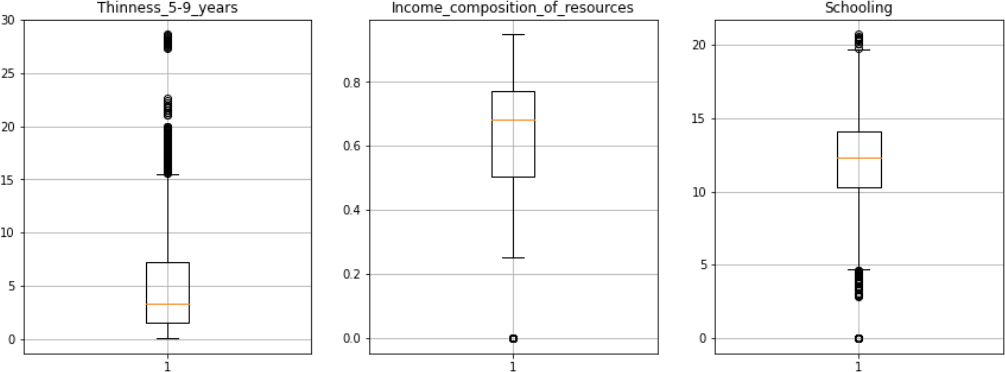
Let us look at the outliers in our dataset.



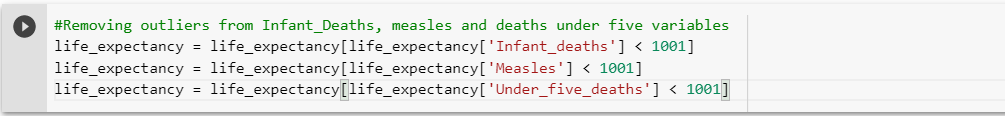




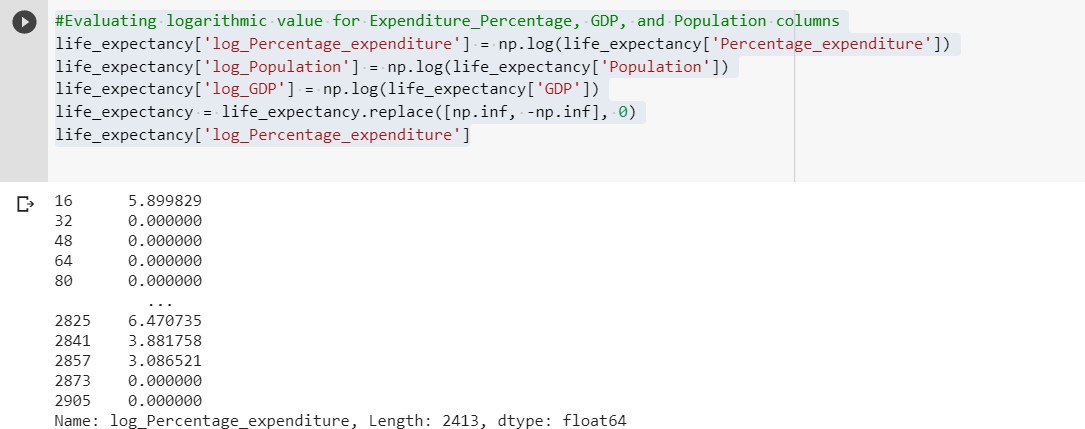




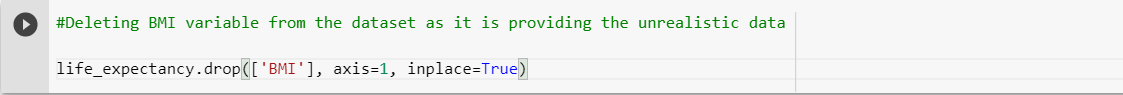
Since some of the variables are beyond the expected value we will therefore eliminate them as outlier.



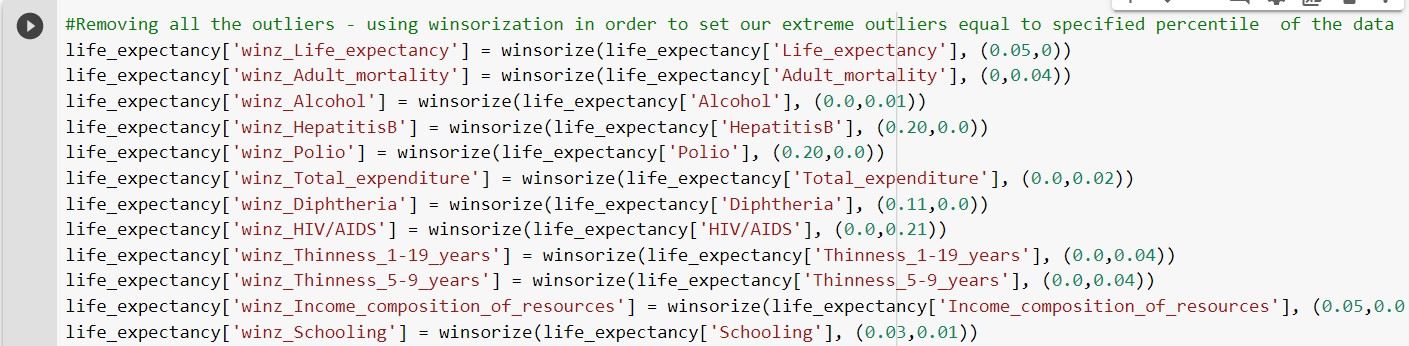
Furthermore, as almost all our variables contains outliers we will be therefore using the winsorization technique through which we can specify specific data for the outliers.



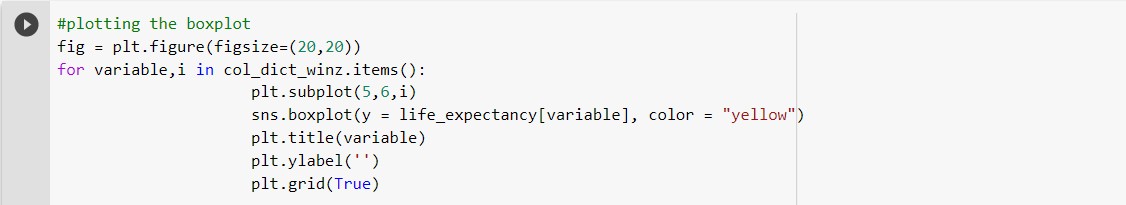
We will be deleting the BMI column from out dataset as it depicts the unrealistic values so rather than setting and extreme value we will simply refer to drop it.

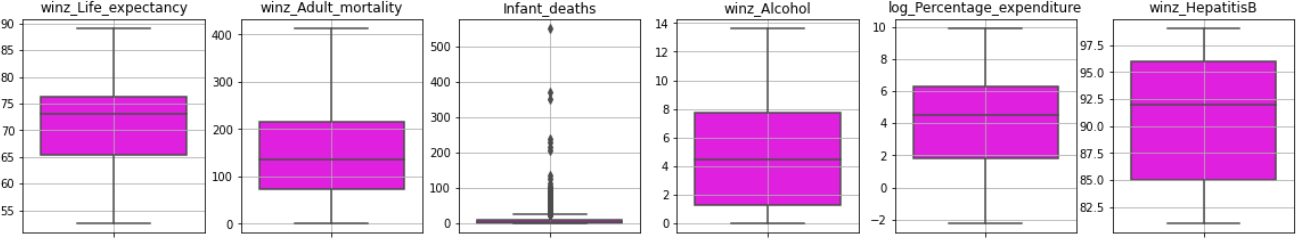


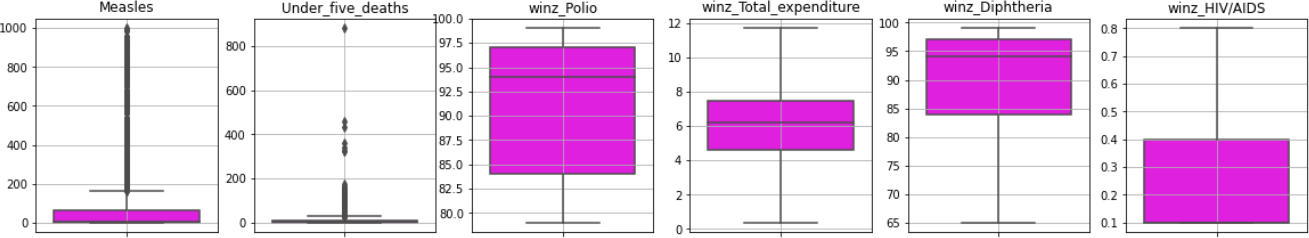
Now, we will be applying winsorization to get rid of outliers from Life Expectancy dataset.

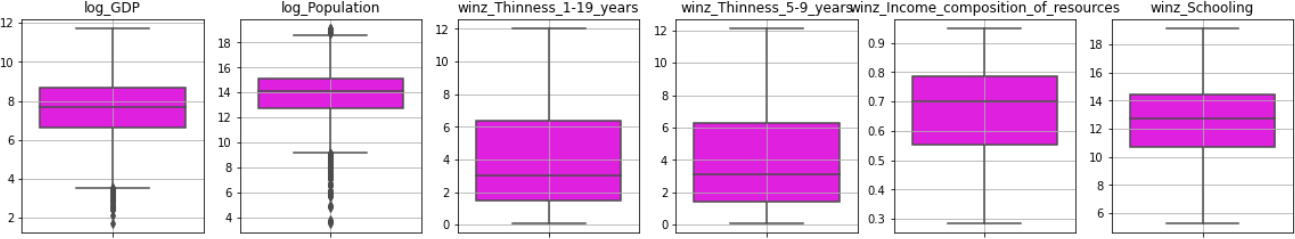






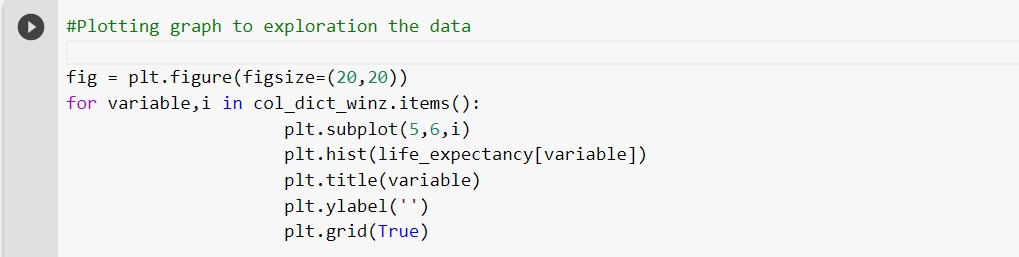


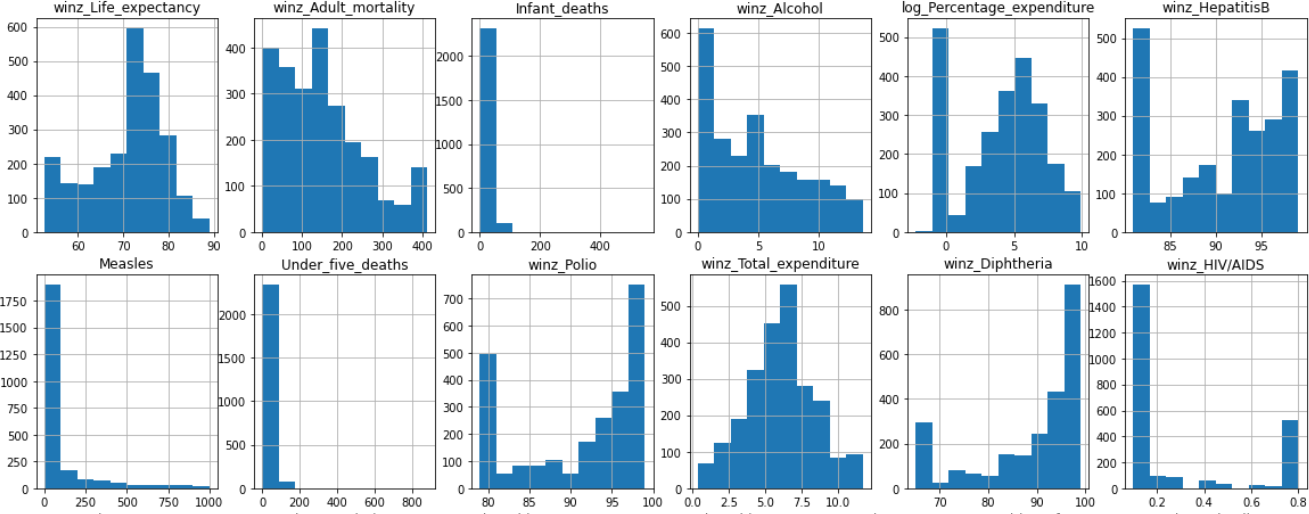


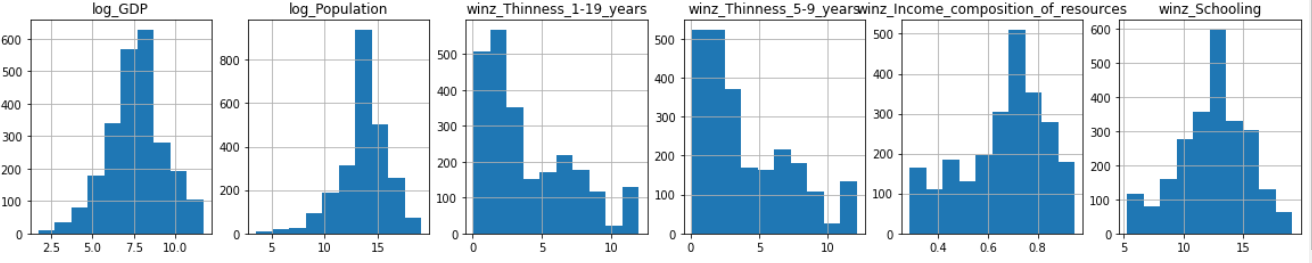


At this stage we have cleaned our dataset by renaming the column to uniform the column names, we have eliminated the unwanted column(s) by dropping it and have removed all the outliers from our dataset. At this point our dataset is clean and now we can use this

transformed dataset to explore our data and to understand the correlation between the variables.







#PIotting a heatmap to check the correlation between the variables in our datase life exp = life expectancy[

[’Year’,

’Country’, ’Status’,

’winz Life expectancy’, ’winz Adult mortality’, ’Infant deaths’,

’winz Alcohol’,

’lo Percentage expenditure’, ’winz\_HepatitisB’,

’ Measles,’

Under f1ve deaths, ’

’winz Polio’,

’winz Total expenditure’, ’winz Diphtheria’,

’winz HIV/AIDS’,

’lo GDP’,

’lo Population’,

’winz Thinness 1-19pears’, ’winz Thinness 5-9pears’,

’winz Income composition of resources’, ’winz Schooling’j

plt.figure(figsize=(15,10))

sns.heatmap(life exp.corr(), annot =True, linewidths = 4)

- MMMMMMMWMMMMMMMMMM \*"

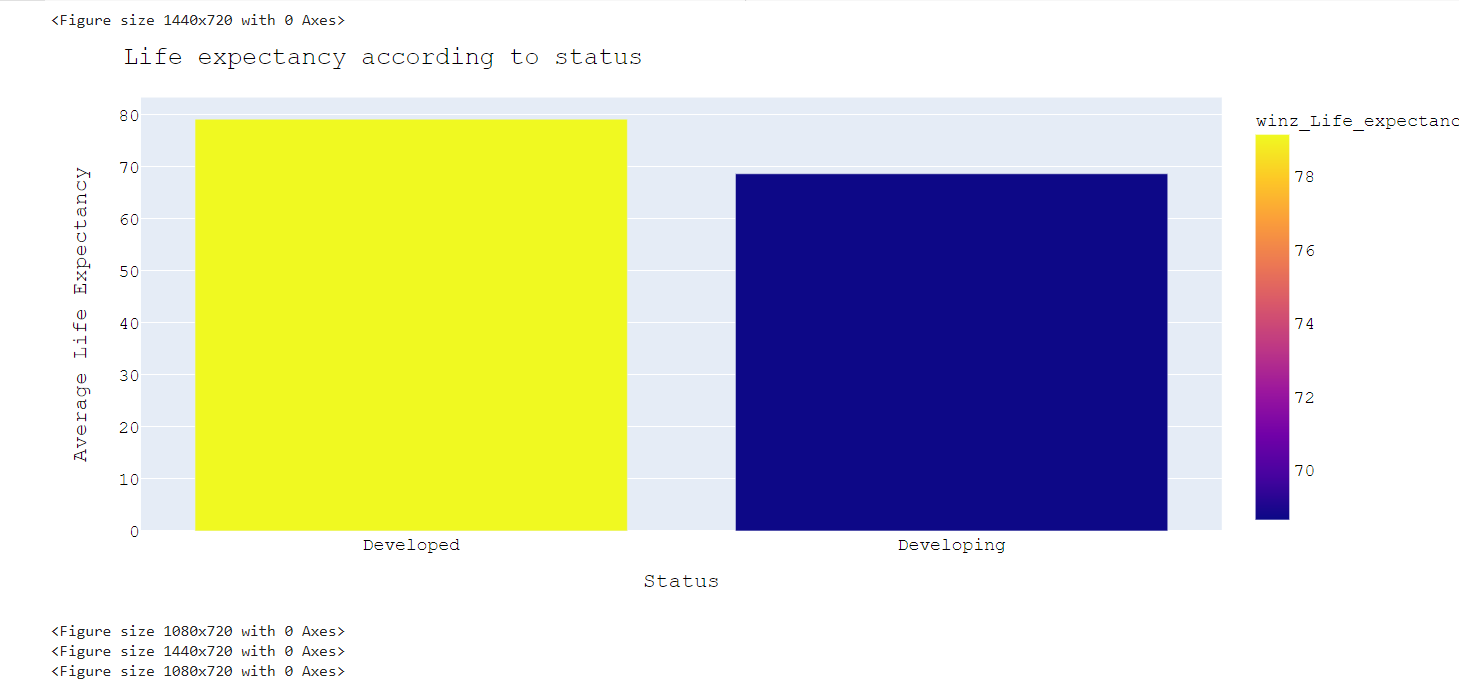


So below are the following observations:

1. Adult mortality has a negative association with education, a positive relationship with HIV and AIDS, and a negative relationship with the composition of resource income.
2. There is a strong positive link between infant deaths and under five deaths.
3. There is a beneficial association between education and alcohol.
4. Education, the mix of resource income, GDP, and life expectancy all have a positive association with percentage expenditure.
5. Polio and diphtheria have a high positive connection with hepatitis B.
6. Diphtheria, hepatitis B, and life expectancy all have a substantial favorable connection with polio.
7. Polio and life expectancy have a strong beneficial association with diphtheria.

Let us analyse the data for the life expectancy according to the status using a bar chart.

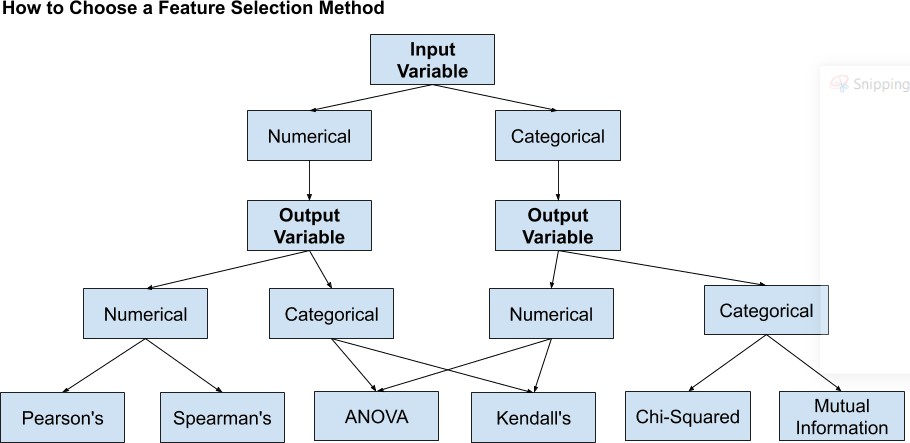




As per the above graph it can be said that the life expectancy for the developed country is more than compared to the developing country.

# Feature Selection

When creating a predictive model, the technique of feature selection is used to reduce the number of input variables. Reduce the amount of input variables to save modelling costs and, in some situations, increase model performance. As a result, choosing a suitable statistical measure for a dataset during filter-based feature selection might be problematic for a machine learning practitioner. Statistics are used to analyze the relationship between each input variable and the goal variable, and the input variables having the strongest association to the target variable are chosen. Since the type of data in both the input and output variables influences the statistical measures used, these procedures can be quick and effective.



Few of the techniques are:

Num I/P, Num O/P

Num I/P, Categorical O/P Categorical I/P, Num O/P Categorical I/P, Categorical O/P

**Numerical O/P**: Regression predictive modelling problem.

**Categorical O/P**: Classification predictive modelling problem.

Since our dataset contains only two categorical values and those also, we have decided to pop out from the independent variables so now we have all the numerical input and output variables as independent and dependent variables. We have made use of correlation techniques on the variables.

As we are going for a linear regression so finding and comparing the correlation can be used to check the correlation between the variables.

If the correlation is or near to 0 than there is no relation,

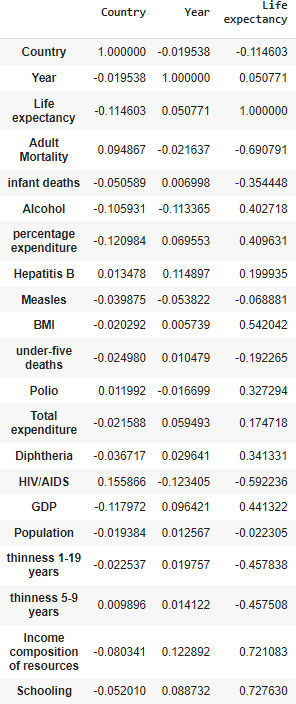
If the correlation value is negative then there is a negative relation which means that if the value of the feature variable decreases then the value of the target value increases (inversely proportional)

And If the value of the correlation is positive then there is a positive relation between target and feature variable which means that if the value of the feature increases then the value of the target values also increases and vice-a-versa(directly proportional)

code:

life\_expectancy= life\_expectancy.dropna() life\_expectancy.corr()

Output:

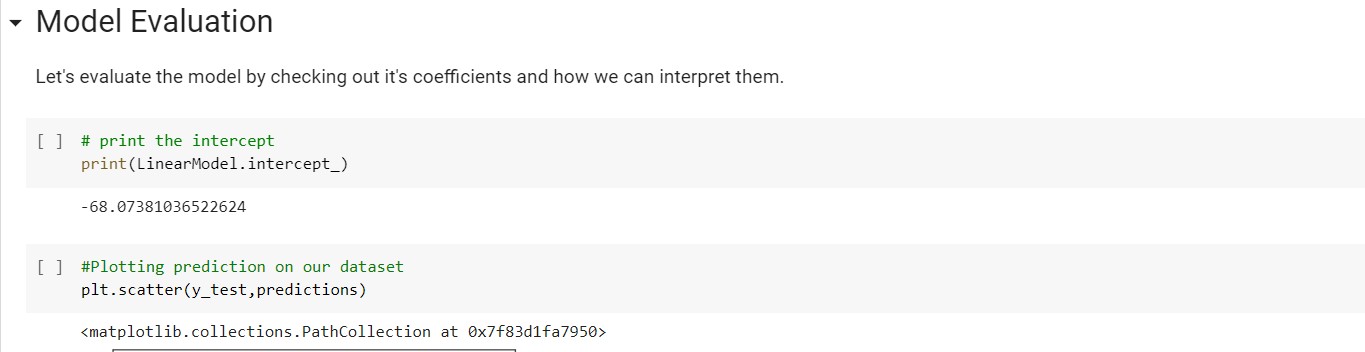
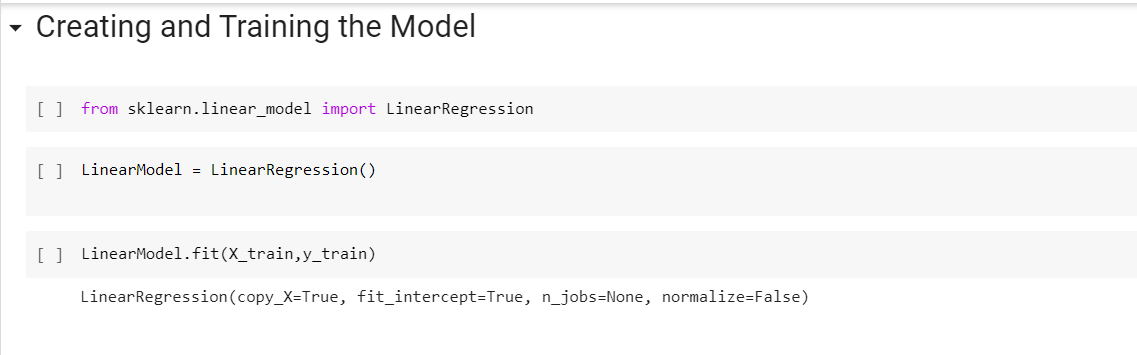
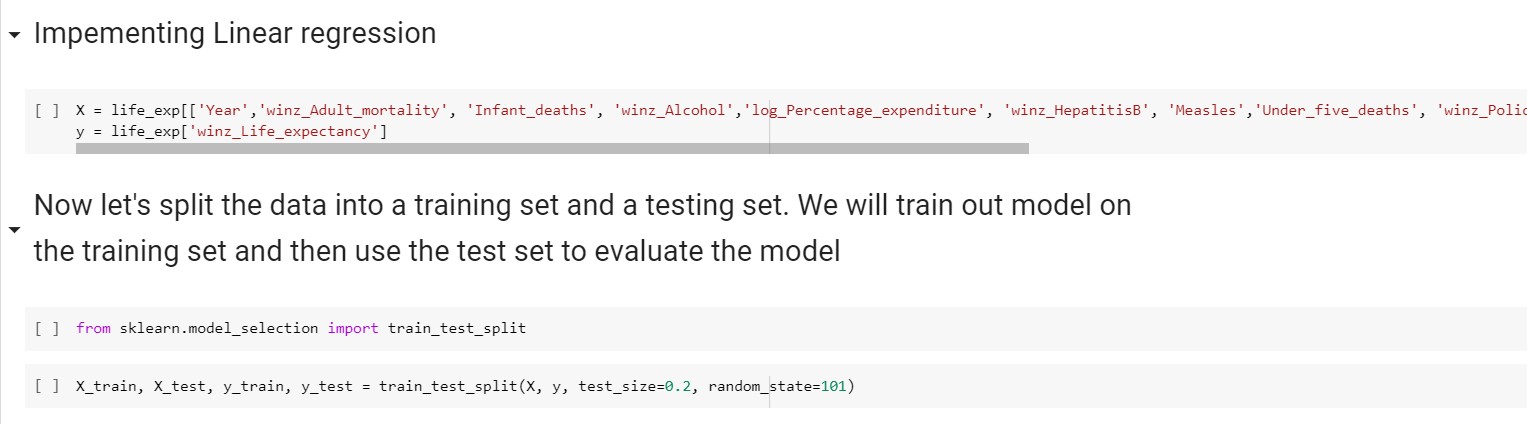


# Model Development and Evaluation

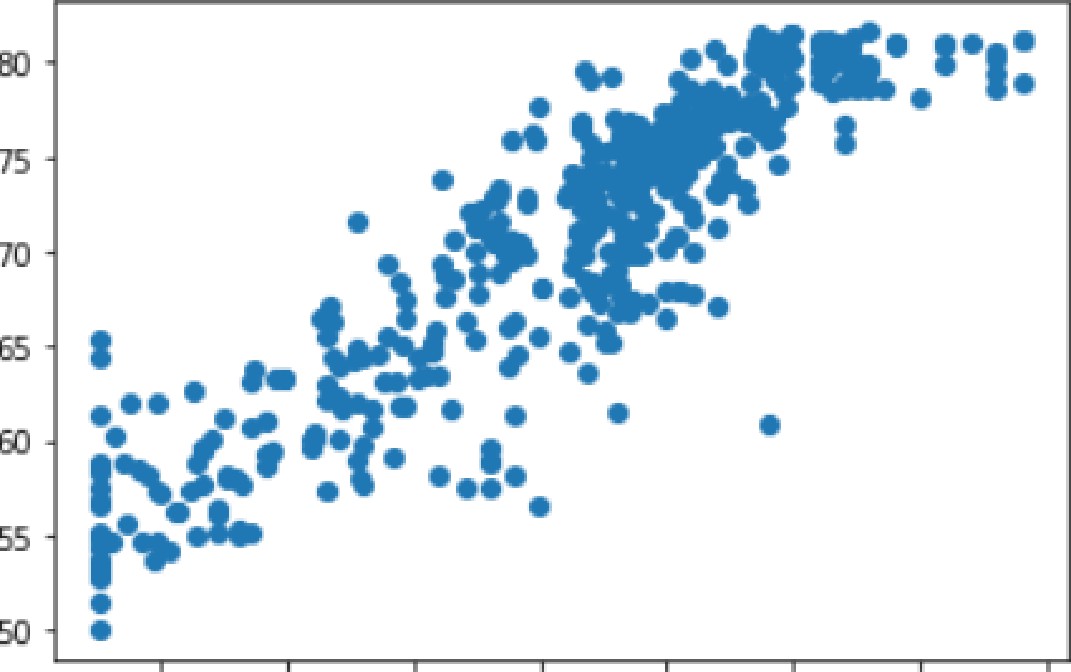
##### Linear Regression

We have implemented the Linear Regression on the life\_expectancy dataset with 84.28 % of accuracy to predict the life\_expectancy with test data.

The colab file with Linear regression is implemented and uploaded on Moodle.



[ ] #Plotting prediction on our dataset pit.sEatter(y test,predictions)

<matplotlib.collections.PathCollection at 0x7f83d1fa7950>

## - Performace of the model

[ ] LinearModel.score(X train, y train)

e. a4zg2s77s gs4tgs

###### Evaluating the model

Lower the MSE, the closer is forecast to actual

MAE and RMSEThese metrics tell us how accurate our predictions are and, what is the amount of deviation from the actual values.

[38] 1 from sklearn imoort metrics

1. print( ' NAE : ' , metrics. frean absolute error(y test, predictions) )
2. print( ' MSE : ' , metrics. frean squared error(y test, p° ed1ctions) )
3. print( ' ROSE: ’, np. sort(metrics. nean squa°eo error(y test, predict ions) ) )

NAE : 2.7229115538906155

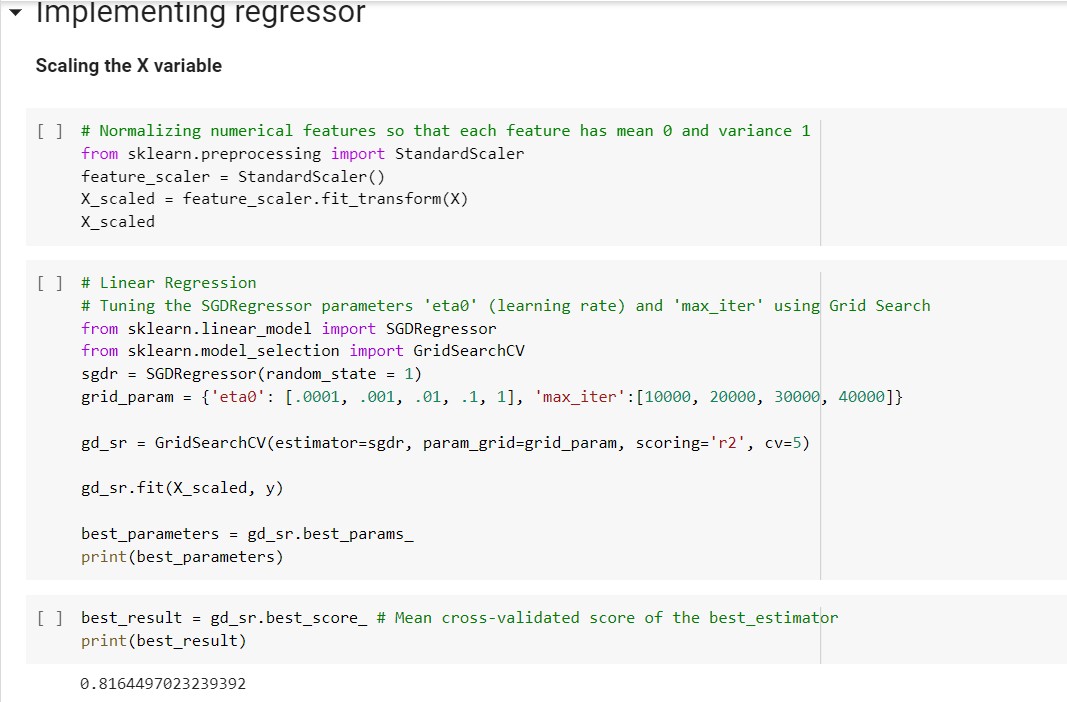
MSE : 13. 850010969276253

MSE : 3.721560286932922

##### Linear Regression with optimization technique: Stochastic Gradient Descent

SGDR is a ML optimization method for finding model params that correspond to the best fit between expected and actual outputs. It's an effective method.

[Stochastic gradient descent](https://en.wikipedia.org/wiki/Stochastic_gradient_descent) is a type of gradient descent algorithm where weights of the model are learned (or updated) based on every training example such that the next prediction could be accurate.

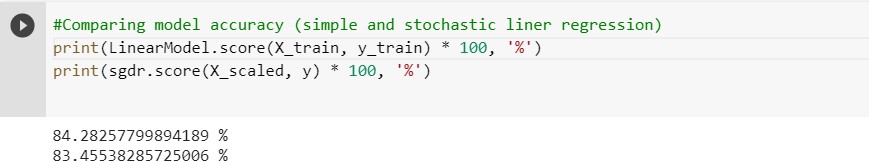


# Model Comparison

The accuracy score of the simple linear model is **83.28 %** and after scaling the X data and applying stochastic gradient descent and fitting the x\_scaled and y with it the accuracy score slightly increases to **83.45%.** This scaling is applied on the original x and y features. After training the dataset with simple linear regression we have tried scaling the x\_train and x\_test also and found that the accuracy is either equal to the simple linear model or increased by

* 1. % respectively.

If we keep on iterating the updated weights of the model there is a chance of acquiring the best accuracy of the trained model.



# Bibliography

Data Preparation:

[https://www.datarobot.com/wiki/data-preparation/#:~:text=What%20is%20Data%20Preparati](https://www.datarobot.com/wiki/data-preparation/#%3A~%3Atext%3DWhat%20is%20Data%20Preparation%20for%2Cuncover%20insights%20or%20make%20predictions) [on%20for,uncover%20insights%20or%20make%20predictions.](https://www.datarobot.com/wiki/data-preparation/#%3A~%3Atext%3DWhat%20is%20Data%20Preparation%20for%2Cuncover%20insights%20or%20make%20predictions)

Stochastic gradient descent:

* + - [https://realpython.com/gradient-descent-algorithm-python/#application-of-the-gradien t-descent-algorithm](https://realpython.com/gradient-descent-algorithm-python/#application-of-the-gradient-descent-algorithm).
    - <https://elearning.dbs.ie/pluginfile.php/1469671/mod_resource/content/1/PP.html>.

Feature selection:

<https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>.

Correlation:

[https://www.codegrepper.com/code-examples/python/python+correlation+between+features+](https://www.codegrepper.com/code-examples/python/python%2Bcorrelation%2Bbetween%2Bfeatures%2Band%2Btarget) [and+target](https://www.codegrepper.com/code-examples/python/python%2Bcorrelation%2Bbetween%2Bfeatures%2Band%2Btarget).

Kaggle:

[https://www.kaggle.com/kaustubhjoshi1910/basic-eda-with-model-training-and-deployment/](https://www.kaggle.com/kaustubhjoshi1910/basic-eda-with-model-training-and-deployment/notebook) [notebook](https://www.kaggle.com/kaustubhjoshi1910/basic-eda-with-model-training-and-deployment/notebook).

Moodle: https://elearning.dbs.ie/pluginfile.php/1469671/mod\_resource/content/1/PP.html.

NOTE: We have prepared the individual files about our contribution and reflection of learning separately and will submit the file altogether on the machine learning moodle detail page.

**Thank You**