Dataset Name: LIFE EXPECTANCY

Group Name: Zzz-02

On Campus/cloud: On Campus

Student ID	Student name	Individual Contribution
218047883	Amber Jain	5
218302635	Isaac Gleeson	5
218188385	Miles Danswan	5
218043304	Cedric Quenette	5

*** 5 – Contributed significantly, attended all meetings**
4 – Partial contribution, attended all meetings
3 – Partial contribution, attended few meetings
1 – No contribution, attended few meetings
0 – No contribution, did not attended any meetings

NOTE: IF ANY OF THE CELLS IN INDIVIDUAL CONTRIBUTION MARK IS EMPTY ALL STUDENTS WOULD GET 3 MARK BY DEFAULT

Section 1: Introduction and getting to know your data (max 2 pages)

Why and how was the data collected?

The data in this dataset comes from two sources: The first source is the Global Health Observatory (GHO) data repository under the World Health Organization (WHO) which was collected to keep track of the health status of countries. The dataset is related to life expectancy. The second source, used to fill in the economic data, is the United Nations (UN) website. The goal of the consolidated data set is to help countries make informed decisions on which areas to invest in to improve the life expectancy of their population. Previous studies did not take into consideration immunization and human development index (HDI). This new data opens new areas of exploration.

Feature Analysis

Variable	Description and Unit	Data type	Number of missing values
Country	Country name	Nominal	0
Year	Year	Ordinal	0
Status	Developed or Developing status	Nominal	0
Life Expectancy	Life Expectancy in age	Discrete	10
Adult Mortality	Probability of dying between 15 and 60 years per 1000 population	Discrete	10
Infant Deaths	Number of Infant Deaths per 1000 population	Discrete	0
Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)	Continuous	194
Percentage Expenditure	Expenditure on health as a percentage of GDP per capita(%)	Discrete	0
Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)	Discrete	553
Measles	Number of reported cases per 1000 population	Continuous	0
BMI	Average Body Mass Index of entire population	Continuous	34

Under-five Deaths	Number of under-five deaths per 1000 population	Continuous	0
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)	Discrete	19
Total Expenditure	General government expenditure on health as a percentage of total government expenditure (%)	Discrete	226
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)	Discrete	19
HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)	Continuous	0
GDP	Gross Domestic Product per capita (in USD)	Continuous	448
Population	Population of the country	Continuous	652
Thinness 10-19 Years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)	Discrete	34
Thinness 5-9 Years	Prevalence of thinness among children for Age 5 to 9(%)	Discrete	34
Income Composition of Resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)	Discrete	167
Schooling	Number of years of Schooling(years)	Discrete	163

Initial Observations and Plans

More developed countries tend to have generally better outcomes on their health variables. Meaning, countries with the attribute 'Developed' have better scores on average across all health metrics - higher health expectancy, lower adult mortality rate, lower infant mortality rate, lower diseases (Hepatitis B, Measles etc.) There is a clear correlation between how developed a country is and how positive their health markers are. An interesting observation is that developed countries do not generally spend a higher percentage of their total government expenditure on healthcare than developing countries. It may be the case that developed countries simply have, on an absolute scale, more money to spend on healthcare per capita. This therefore leads to better health outcomes.

Overall, this dataset is very versatile and beyond the obvious observations, there are several more profound observations which can be made. This is because there is a lot of data representing a variety of factors. We plan to consider all the data available and extract findings from it.

Section 2: Exploratory Data Analysis and Results (max 7 pages)

Data Cleaning

Detecting & Replacing missing values

The first anomaly observed in the data was the presence of whitespace in column names which had to be removed. This was done by the 'rename' function in Python.

After carefully reviewing the data and grouping them by 'Countries', it is evident that the data contains a lot of null values. The number of missing values and their percentage can be seen in Figure 1.

```
Life expectancy                   0.340368
Adult Mortality                   0.340368
Alcohol                           6.603131
Hepatitis B                      18.822328
BMI                               1.157250
Polio                             0.646698
Total expenditure                 7.692308
Diphtheria                        0.646698
GDP                              15.248468
Population                       22.191967
thinness  1-19 years              1.157250
thinness 5-9 years                1.157250
Income composition of resources   5.684139
Schooling                         5.547992
dtype: float64
```

Figure 1: Missing value in %

Most of the columns except GDP, Population and Hepatitis B have less than 15% missing values which were filled by a mixture of methods such as Linear Interpolation, mean, median and the Python's 'ffill'. A brief outline about these methods is presented below:

- Linear Interpolation: This is a technique where the value is calculated and filled on the basis of a linear function which is calculated using the preceding and succeeding value [1].
- Mean: The value is filled with the mean of the data which is grouped by either 'Country', 'Year' or 'Status'. For Example, to fill the missing data for 'Alcohol', each the mean of each country is calculated and the missing values are filled based on that.
- Median: The middle value when the entire series is sorted in ascending order. This is particularly used for 'Life expectancy' as most of the missing data is spread out rather than grouped by a single attribute.
- 'ffill' : This special method in Python is used here the most as it copies the preceding value of the missing data to fill it.

As a result, the majority of the data was filled with appropriate values using the above method. As GDP, Population and Hepatitis B had less than 15% missing values, they were ignored to avoid misinterpretation of data. Figure 2 displays the result.

```
Country                           0.000000
Year                              0.000000
Status                            0.000000
Life expectancy                   0.000000
Adult Mortality                   0.000000
infant deaths                     0.000000
Alcohol                           0.000000
percentage expenditure            0.000000
Hepatitis B                       4.901293
Measles                           0.000000
BMI                               0.000000
under-five deaths                 0.000000
Polio                             0.000000
Total expenditure                 0.000000
Diphtheria                        0.000000
HIV/AIDS                          0.000000
GDP                              14.363513
Population                       22.191967
thinness  1-19 years              0.000000
thinness 5-9 years                0.000000
Income composition of resources   0.000000
Schooling                         0.000000
dtype: float64
```

Figure 1: Result after filling null values

Detecting and Replacing Outliers

Box Plots were drawn to detect Outliers, anything lying outside the 1.5 times Interquartile range were considered as outliers. An example BoxPlot for 'Adult Mortality' is shown in Figure 3.

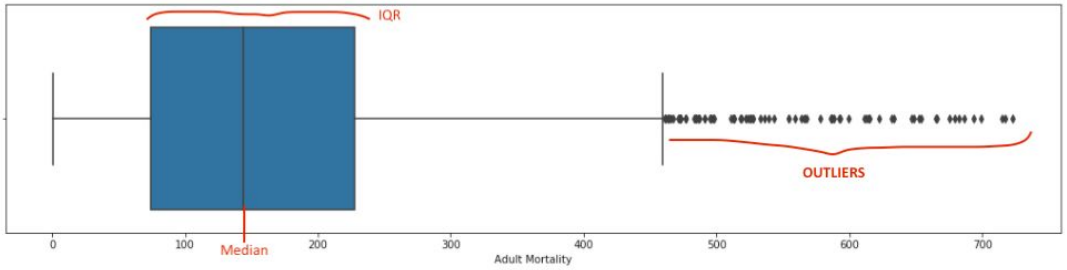


Figure 3: Outliers detection using Box Plot

The box plot shows the outliers as marked in red. Similarly, Box Plots were observed for all the columns and were replaced with **Winsorization** using SciPy library. Winsorization is a technique where extreme values are replaced by lesser extreme values [2].

This enables replacement of outliers with the boundary values. After winsorization, all the outliers were successfully removed and the resultant Box Plots had no outliers. For Example, figure 4.

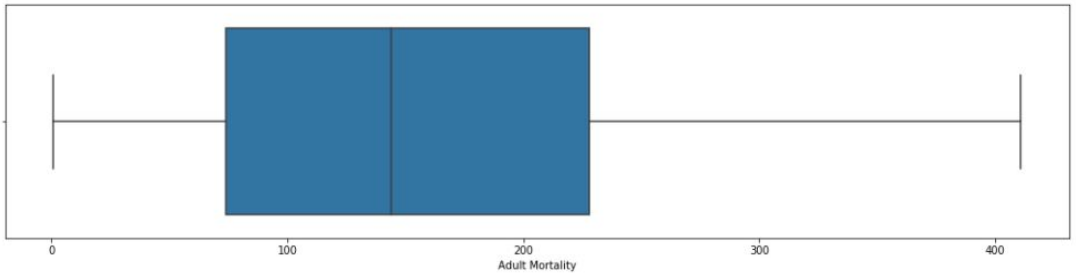


Figure 4: Outliers replaced by Winsorization

General Exploration of Data

Life Expectancy (Target variable) vs all other attributes

This dataset has a target variable (Life Expectancy), against which all other variable's impacts can be measured and evaluated. To do this, Pearson's Correlation Coefficient can be calculated which can help us to narrow down the exploration.

Pearson's Correlation helps us to study the correlation between any two elements. Using the 'Seaborn' library a heatmap is plotted which shows the Pearson Correlation values between each column in the data set.

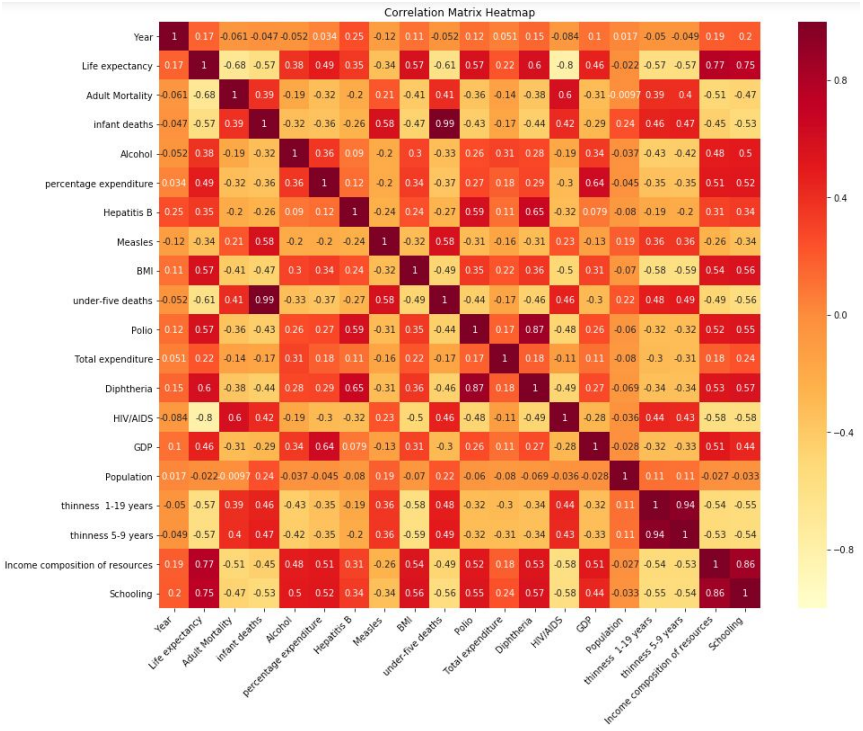


Figure 5: Correlation Heatmap

From the above heatmap, these are the following stark observations can be made based on Correlation values and colour of Heatmap:

Highly correlated variables (corr greater than 0.6):

- 'Income Composition of resources' and 'Schooling', coefficient of +0.86.
- 'Income Composition of resources' and 'Life Expectancy', coefficient of +0.77
- 'Life Expectancy' and 'Schooling', coefficient of +0.75
- 'Under five deaths' and 'infant mortality', coefficient of +0.99
- 'GDP' and 'Percentage Expenditure', coefficient of +0.64
- 'Polio' and 'Diphtheria', coefficient of +0.87.
- 'Thinness 5-9 years' and 'Thinness 10-19 years', coefficient of 0.86.

Negatively correlated variables (corr less than -0.6):

- 'Under 5 deaths' and 'Life Expectancy', coefficient of -0.61.
- 'Life Expectancy' and 'Adult Mortality', coefficient of -0.68.
- 'Life Expectancy' and 'HIV/AIDS', coefficient of -0.80.

Some key observations related to the target variable 'Life Expectancy':

1.) HIV/AIDS had a larger impact on Life expectancy than all other diseases as proved by the lowest coefficient.
2.) Schooling and Income composition by resources have by far the largest positive impact on Life Expectancy.

Developed vs Developing Countries: Year progression

An year by year analysis was done as shown in Figure 6 for all the variables grouped by 'Status' of the Country. These line plots are helpful in understanding how Developing and Developed Countries have evolved over the years with respect to each feature.

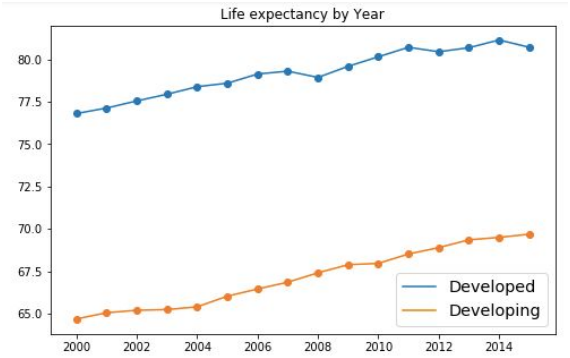


Figure 6: Status and Year Analysis of Life expectancy

Similarly it was done for all the other columns and following were the initial observations:

- Life Expectancy has increased for both developed and developing nations, but developing nations have had a more rapid increase in it.
- Infant deaths have decreased for both over the years.
- BMI has increased for both, with Developed nations showing sharper growth.
- Income and Schooling has also increased for both.
- Alcohol Consumption has increased in Developing nations over the last 4 years, and has decreased in Developed nations.

Investigating the relationship between Schooling and Life expectancy

Initial investigation

With a primary interest in this investigation being between the relationships that existed with Schooling, our Hypothesis before starting our initial investigation was that a greater Schooling period would lead to a higher Life expectancy.

Steps taken in our early investigation into the dataset included the basic steps. Including printing the shape of the dataset, the head and tail, min and max and what started our exploration of Schooling and life expectancy was the *figure 5:* heatmap showing the range of Pearson correlation strength between each relationship. One of the strongest relationships in this dataset was found to be a strong positive relationship between Schooling and Life expectancy.
The next step in the investigation was to narrow in further on the Schooling and Life expectancy relationship and did so with the dot plot below.

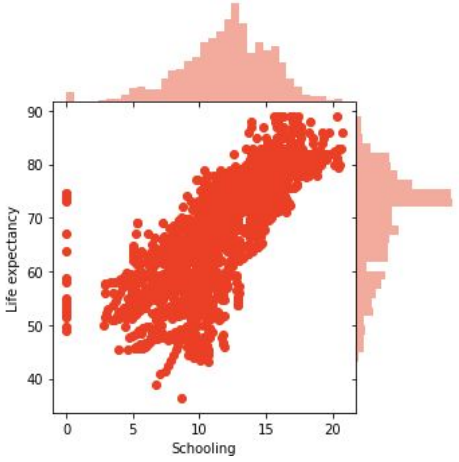
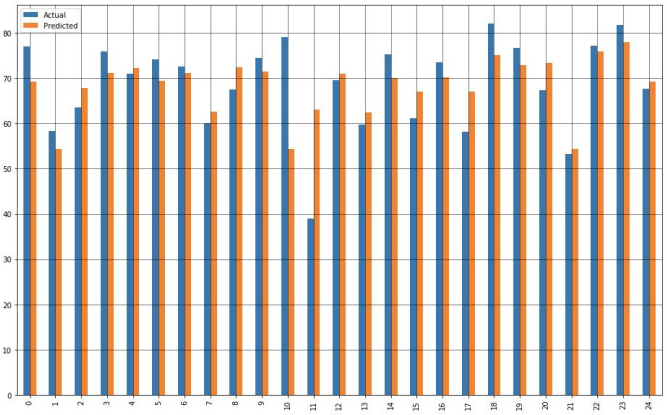


Figure 7: correlation between Schooling and Life expectancy

This visualization shows the strong positive relationship that the Pearson calculation gave us; a strength of 0.751975.

This relationship has shown to be visually and numerically strong, strong enough to be put through a regression prediction algorithm.



	Actual	Predicted
0	77.0	69.287185
1	58.3	54.271199
2	63.5	67.720299
3	75.8	71.115218
4	71.0	72.290382
...	...	...
581	54.0	66.414561
582	65.6	68.112020
583	65.1	67.720299
584	74.5	69.156611
585	75.7	73.987841

Figure 8,9 (respectively): Actual to Regression predicted

Showing our regression prediction algorithm with a close relationship to our original data, this is one prediction technique that we can use to help us perform extrapolation analysis from our dataset. We can identify two columns (10,11) where the range between 'Actual' and 'Predicted' is to be considered inaccurate and requires further scrutiny.
Testing these results against other prediction methods would help validate the reliability of any extrapolation, although the only way to be certain of the models accuracy is to wait and record the true data in later years.

Investigating Factors Affecting Health Outcomes in Children

We have also investigated which factors affect the health of children.

Preventing Redundant Analysis

In the analysis, we have determined that thinness 5-9 years and thinness 1-19 years as well as under-five deaths and infant deaths is sufficiently similar to avoid analysing all of them. We have omitted one in each pair, keeping under-five deaths and thinness 1-19. To prove they are indeed similar enough, we have visualised the data and calculated their Pearson's correlation[5].

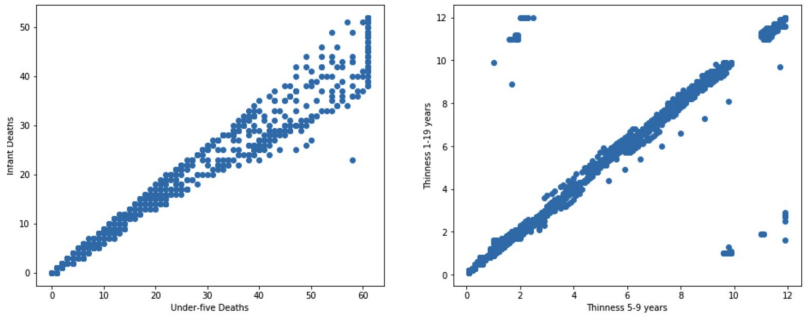


Figure 10: under-five deaths vs. infant deaths & thinness 1-19 vs. thinness 5-9

The correlation coefficient of under-five deaths and infant deaths is: 0.992 and of thinness 5-9 years and thinness 10-19 years is: 0.942. The combination of the visualisation and the correlation *confirms* that both under-five deaths is very similar to infant deaths and thinness 1-19 years is very similar to thinness 5-9 years.

Selected Metrics and Factors

The metrics related to child health are: **thinness 10-19 years (%), HIV/AIDS (deaths per 1000 live births 0-4 years), under-five deaths (per 1000 population)**. We have isolated the factors affecting the aforementioned metrics the most by creating a correlation heatmap (fig. 5).
The selected factors with the strongest impact on child health metrics are: **Immunisation (Diphtheria, Hepatitis B, Polio) (immunisation coverage among 1-year-olds (%)), Schooling (number of years of schooling), BMI (Average Body Mass Index of the Population).**

Difference Between Developing and Developed Countries

To determine whether there is a statistically significant difference between developing and developed countries, we have performed a **Student's t-test**[5] of factors affecting child health. The resulting t-values for under-five deaths is 17.2, for thinness 25, for HIV/AIDS is 16.9. We do not need to specify a p-value because the t-values are so high, meaning that there is a huge difference between developing and developed countries. Thus, we will separate the analysis.

Analysis 1: Immunisation vs Child Mortality

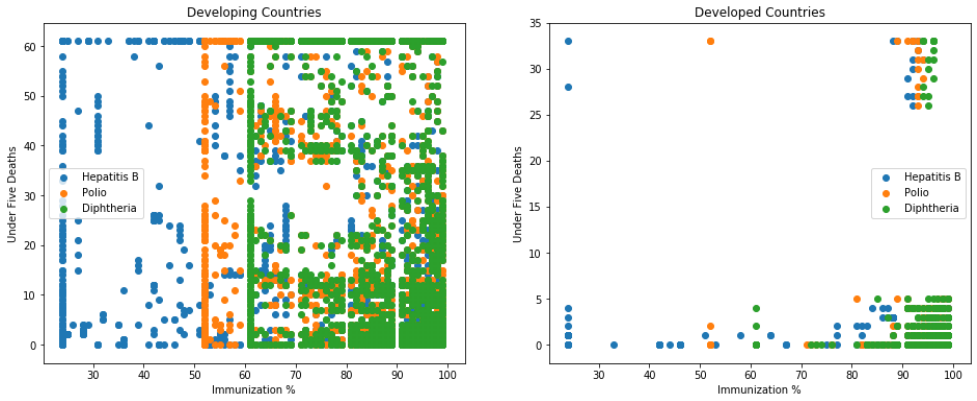


Figure 10: under-five deaths vs. immunisation (developed and developing)

We will investigate the significance of immunisation on reducing child mortality. Thus, we will find the effect of Hepatitis B, Polio and Diphtheria immunisation on under-five deaths. The hypothesis is that the higher the rate of immunisation, the lower the child mortality. First, we visualise how under-five deaths is impacted by immunisation in developing countries and developed countries.

The graphs do not show any obvious correlation in the data. using Pearson's correlation we expect a negative correlation between immunisation and death. The rationale is that a higher rate of immunity from diseases translates to a lower death rate from those diseases. Additionally, countries which adopt a higher rate of immunisation should also have better quality of healthcare.

For **developing countries**, the correlation between **Hepatitis-B** and under-five deaths is **-0.264**, for **Polio** it is **-0.402**, for **Diphtheria** it is **-0.421**.
For **developed countries**, the correlation between **Hepatitis B** and under-five deaths is **0.003**, for **Polio** it is **-0.179** and for **Diphtheria** it is **0.035**.

For developing countries, increased immunisation clearly decreases child mortality, however, for developed countries this is not the case. It is possible that developed countries have a low prevalence of these diseases due to herd immunity, thus, making immunisation less necessary.

Furthermore, higher immunisation also correlates with a general higher standard of healthcare in a country, meaning that they have access to better healthcare in general.

Analysis 2: Schooling vs Child Health

We have also investigated the significance of the number of years spent in school with child health metrics. The *hypothesis* is that the *more years of schooling, the better the health outcomes* of children. The graphs did not show a conclusive correlation, however, Pearson's correlation proved conclusive. For **developing countries**, the correlation with schooling and **under-five deaths** is **-0.504**, for **thinness** it is **-0.446**, for **HIV** it is **-0.532**. For **developed countries**, for **under-five deaths** it is **-0.441**, for **thinness -0.206**, and there is **no correlation for HIV/AIDS** because there are no HIV related child deaths in developed countries.
We can see a clear impact on schooling in improving the health of children.

Analysis 3: BMI

We will analyse the effect of BMI on child health metrics. We expect BMI to be normally distributed with most of the population being around the mean BMI. We also expect negative child health outcomes to increase as BMI becomes too high or too low. BMI in the dataset is the average across each population but still serves as a good overall indicator for children.

The visual analysis **does not** show that BMI is normally distributed, however, to **confirm**, we will perform a statistical analysis, a normality test. The one we have selected is the Shapiro-Wilk test which will identify whether BMI was taken from a Gaussian distribution (normal distribution). From the literature, the Shapiro-Wilk test[6] is suitable for the size of our dataset as well (thousands of data points or under). The resulting p-values for both developing and developed countries was 0.00. There is no need to specify an alpha value because it will be greater than zero. We can see clearly that **BMI was not taken from a Gaussian distribution**.

We then move into finding a correlation between BMI and children health metrics.
For **developing countries**, the correlation coefficient between BMI and **under-five deaths** is **-0.474**, for thinness **-0.561**, for **HIV/AIDS -0.478**. For **developed countries**, **under-five deaths -0.044**, **thinness -0.131**, **HIV/AIDS has no correlation** because of no child HIV/AIDS related deaths in developed countries. We can see a negative correlation between BMI and health factors in developing countries and almost no correlation in developed countries. This means that for developing countries, as BMI increases, negative health factors in children decrease.

Let us investigate why an increase in BMI results in better health in children. Our hypothesis is that BMI is correlated with wealth which is in turn correlated with better health outcomes for children.

	Variable	N	Mean	SD	SE	95% Conf.	Interval
0	BMI	2426.0	35.111129	19.546760	0.396853	34.332924	35.889335
1	BMI	512.0	51.803906	17.196829	0.760000	50.310798	53.297015
2	combined	2938.0	38.020150	20.175077	0.372211	37.290329	38.749971

Figure 11: Descriptive analysis of developing vs. developed BMI

From a descriptive analysis of BMI between developing countries (index 0) and developed countries (index 1), we can see that **developing countries** have a **lower mean BMI of 35** than **developed countries** which have a mean **BMI of 52**. Pearson's correlation for **developing countries** between GDP and **under-five deaths** is **-0.250**, **thinness -0.235, HIV/AIDS 0.317**. Thus, the higher the GDP, the better the health metrics for children and the higher the BMI which confirms our hypothesis.

Section 3: Conclusions (max 1 page)

Conclusion

The above exploration of the Life Expectancy dataset has provided conclusive answers to questions regarding several attributes of life.

Education provides the foundations for developing the future, hence, we must ensure that those who are learning, go on to apply their learnings. Our analysis of the relationship between Schooling and Life Expectancy has revealed a positive response to this question, with a strong relationship binding the two attributes. Those who Study for longer periods, consistently have a higher Life Expectancy, and vice-versa.

Living conditions are a common issue in both developing and developed countries. For countries to develop their standard of living and support of health care, it's imperative to understand the factors which are primarily affecting current living conditions. Our analysis of *Immunisation*, *Schooling* and *Body Mass Index* shows that Immunising children at an early age will (in most cases) decrease the mortality rate of children under 5 years of age, keeping children in School can improve the overall Health of children, and as BMI increases, the negative effects on young children will decrease. These results can help countries improve their support for HealthCare in specific areas, with the intention of increasing Life Expectancy and overall living conditions.

The Life Expectancy of Countries will vary for a significant time to come, however, with further clarity to the problem of 'how to efficiently affect living conditions' and 'what areas of Health require primary focus', Countries can implement Health standards which improve the Living conditions and raise the Life Expectancy.

For child health, *schooling* is the most effective way of reducing mortality and improving health outcomes. This is because *schooling* has the highest negative correlation with negative health effects in children. Furthermore, *schooling* increases the wealth of a country (GDP and income composition of resources) which further improves the health outcomes of children. The one drawback is that developed countries have a BMI which is above the healthy range which may lead to health issues.

Outstanding Problems for Machine Learning Execution

- Predict the Life Expectancy for the following **2-5 years** based on Adult Mortality Rates
- Predict in what year a 'Developing Country' will be categorised as 'Developed', based on their development between 2000 and 2015

Section 4: References (max 1 page)

[1] Encyclopediaofmath.org. 2020. Linear Interpolation - Encyclopedia Of Mathematics. [online] Available at: <https://www.encyclopediaofmath.org/index.php/Linear_interpolation> [Accessed 20 April 2020].

[2] Encyclopediaofmath.org. 2020. Linear Interpolation - Encyclopedia Of Mathematics. [online] Available at: <https://www.encyclopediaofmath.org/index.php/Linear_interpolation> [Accessed 20 April 2020].

[3] Chauhan, N., 2020. *A Beginner'S Guide To Linear Regression In Python With Scikit-Learn*. [online] Medium. Available at: <https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f> [Accessed 30 April 2020].

[4] Brownlee, J., 2020. *How To Calculate Correlation Between Variables In Python*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/> [Accessed 30 April 2020].

[5] Docs.scipy.org. 2020. Scipy.Stats.Ttest_Ind — Scipy V1.4.1 Reference Guide. [online] Available at: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html> [Accessed 30 April 2020].

[6] Brownlee, J., 2020. A Gentle Introduction To Normality Tests In Python. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/> [Accessed 30 April 2020].