School of Chemical and Biomedical Engineering

CB0494 Mini-Project

Mini-Report

Done by:

1. Ng Yu Xuan - U1921991J
2. Ng Jun Xiang - U1921209J

TABLE OF CONTENTS

Background

Life expectancy is defined as a statistical measure of the average time an organism is expected to live, depending on the year of its birth. Life expectancy is a volatile variable that can be affected by an assortment of other different variables for example, Adult Mortality, Schooling etc. Hence, there is a need to study, analyse and compare such variables against life expectancy to uncover the crucial characteristics and trends behind such data. Perhaps in the future, governing bodies of such countries can utilise such relevant data to understand which areas to improve on so as to increase their respective country's general life expectancy.

In this report, the life expectancy dataset used was obtained from Kaggle [1]. The life expectancy dataset provides data of different countries from years 2000 to 2015 regarding life expectancy and variables that may be related to life expectancy.

Objectives

It is commonly known that developed countries tend to have higher life expectancy. Life expectancy can be attributed to many other factors.

1. Using this dataset, what are the best predictors of Life Expectancy?
2. How can we use these predictors to affirm what we were told or what is commonly known?

Setup/Techniques used

1. Data Analysis (Non-Interactive Plots and Correlation Heatmap)
2. Supervised Learning Univariate Linear Regression
3. K-means++ Clustering and Plotly visualization

Results

Preparation of Data

To perform analysis on the dataset, cleaning of the data was deemed the priority. It is found that the values for "Measles" are the exact case numbers and the values are not in per 1000 population, for example China in 2015, the number 42 361 in the data is the same number of cases of Measles in China rather than cases per 1000 in China [2]. "Infant Deaths" and "Under-five deaths" are also supposedly per 1000 populations as well. However, only values per 1000 live births were found instead [3][4]. Thus, values for "Infant Deaths" and "Under-five deaths" were assumed to be accurate.

Furthermore, values that exceed the maximum allowable based on the description are removed from the dataset. For example, values for the average "BMI" for some countries are exceedingly high which is not possible and hence, removed from the dataset [5].

Afterwards, the mean values for each variable were found for each country. It was found that "Measles", "BMI", "GDP" and "Population" has a high percentage of missing values, with fractions 48/193, 108/193, 30/193 and 48/193 missing data, respectively. Thus, these 4 variables were dropped. The rest of the variables with little missing values were filled with the mean of the variable (Column) to get a better representation of data. Hence, figure 1 below shows the result of the cleaned data.

	Country	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	thinness 1-19 years	tl
0	Afghanistan	Developing	58.19375	269.0625	78.2500	0.014375	34.960110	64.562500	107.5625	48.3750	8.252500	52.3125	0.10000	16.58125	1
1	Albania	Developing	75.15625	45.0625	0.6875	4.848750	193.259091	98.000000	0.9375	98.1250	5.945625	98.0625	0.10000	1.61875	
2	Algeria	Developing	73.61875	108.1875	20.3125	0.406667	236.185241	78.000000	23.5000	91.7500	4.604000	91.8750	0.10000	6.09375	
3	Angola	Developing	49.01875	328.5625	83.7500	5.740667	102.100268	70.222222	132.6250	46.1250	3.919333	47.6875	2.36875	6.19375	
4	Antigua and Barbuda	Developing	75.05625	127.5000	0.0000	7.949333	1001.585226	98.266667	0.0000	96.9375	4.791333	98.3125	0.12500	3.42500	

Figure 1. Resulting Dataset from Cleaning

Correlation Heatmap

Figure 2 below shows the correlation between life expectancy and the other variables. It is observed that "Adult Mortality", "Income Composition of Resources", "Schooling", "Polio", "Diphtheria" and "HIV/AIDS" have the strongest correlation with "Life Expectancy", with correlation values of -0.9, 0.78, 0.7, 0.63, 0.63 and -0.59 respectively.

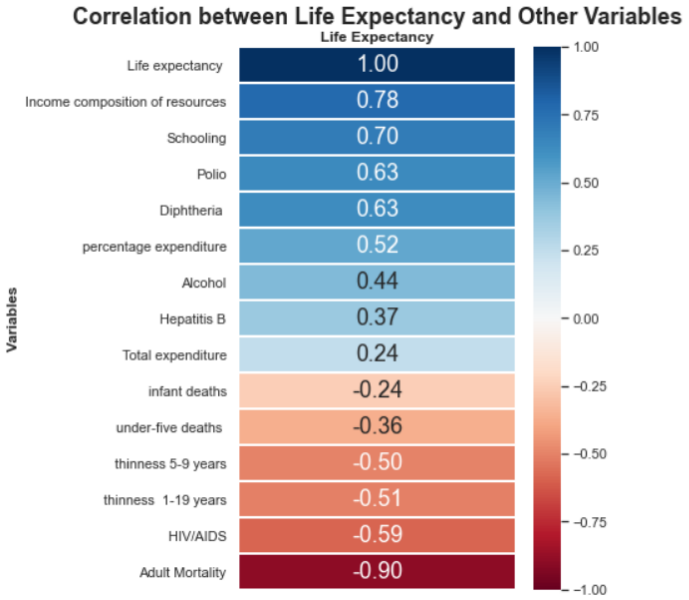


Figure 2. Correlation Matrix for Life Expectancy

The rest of the variables generally do not have high enough correlation with life expectancy, Hence, only the top 6 variables as mentioned above will be used for further analysis.

Linear Regression

Supervised learning univariate linear regression was then performed for the top 6 variables to find out the best predictors for life expectancy. The dataset was split into 75:25 ratio for train and test set, respectively. Fixed and random split linear regression were both performed and the R^2 values found were then compared in figure 3 below.

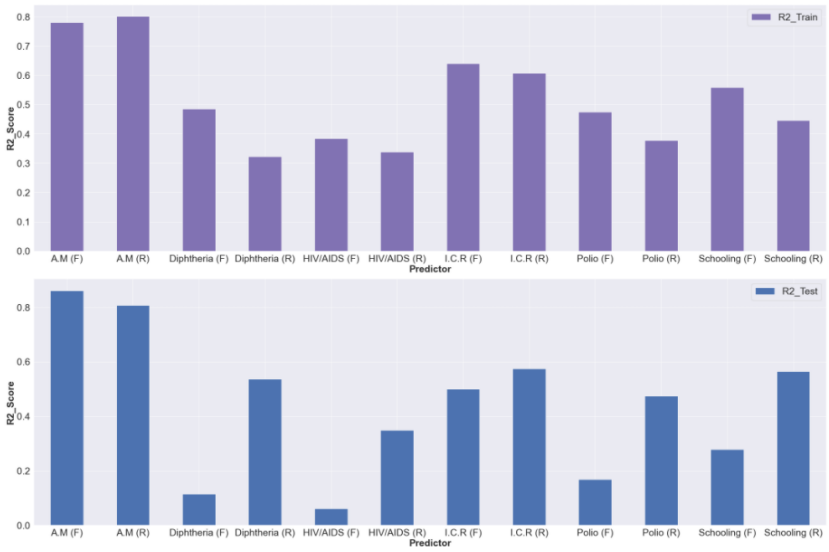


Figure 3. Bar plot of R^2 Values for both Fixed and Random Split Train and Test Set

For the train set, R^2 values were quite consistent between fixed and random split. However, differences in R^2 values were more obvious in the test set as seen above. This may be due to the ordering of countries in alphabetical order or the small number of datapoints used that gives rise to biasness. Thus, random split would be more ideal to use to analyse linear regression.

Figure 4. Bar plot of R^2 Values for Random Split Train and Test Set

Figure 4 above displays the average order of R^2 values after multiple runs. “Adult Mortality”, “Income Composition of Resources” R^2 values are the most stable and do not vary as much as the R^2 values of “Schooling”, “Polio”, “Diphtheria” and “HIV/AIDS”. Thus, “Adult Mortality” and “Income Composition of Resources” can be considered as the most reliable and accurate in predicting life expectancy values.

With the best predictors of life expectancy uncovered, further analysis will be performed with "Adult Mortality", "Income Composition of Resources" and "Schooling" (For Interest) to determine whether developed countries tend to have high life expectancy values.

Category Plot

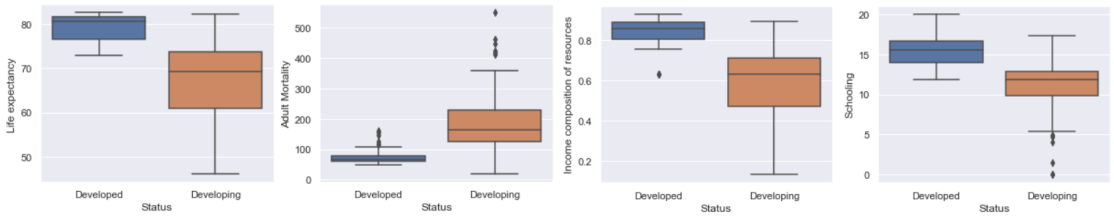


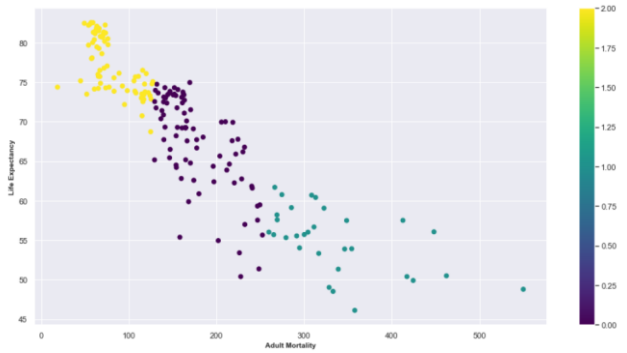
Figure 5. Category plots of Life expectancy, Adult Mortality, Income Composition of Resources and Schooling against Status.

From figure 5 above, category plot was used to observe how the country status relates to life expectancy and other predictors. It is observed that developed countries have higher life expectancy, lower adult mortality, higher income composition of resources and higher schooling. Thus, this affirms that developed countries have higher life expectancy.

K-Means Clustering

Clustering was performed to cluster the countries with similar life expectancy levels and said variable levels to visualise the trend better. K-means ++ was first utilized to obtain the optimum K value of 3. With the optimum K value known, clustering was then performed for "Adult Mortality", "Income Composition of Resources" and "Schooling".

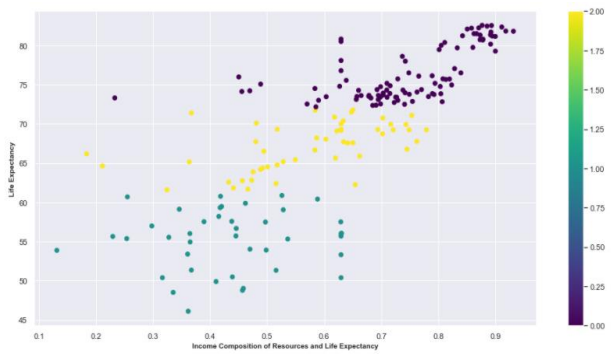
Next, each cluster was then tabularised and compared with one another in regard to the number of both developing and developed countries, respectively.



Status	Cluster 1	Cluster 0	Cluster 2
Developed	0	3	29
Developing	30	87	44

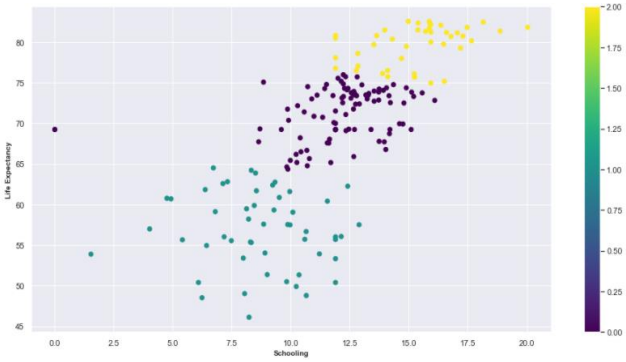
Figure 6. Clustering for Life expectancy and Adult Mortality (Left) and Table for Number of Countries in each Cluster for Different Status (Right)

This essentially displays the disparity of developed and developing countries among the 3 clusters as well as proving that developed countries tend to have lower life expectancy and "Adult Mortality" as compared to developing countries. The same was done with "Income Composition of Resources" and "Schooling".



Status	Cluster 1	Cluster 2	Cluster 0
Developed	0	0	32
Developing	41	57	63

Figure 7. Clustering for Life expectancy and Income Composition of Resources (Left) and Table for Number of Countries in each Cluster for Different Status (Right)



Status	Cluster 1	Cluster 0	Cluster 2
Developed	0	6	26
Developing	52	93	16

Figure 8. Clustering for Life expectancy and Schooling (Left) and Table for Number of Countries in each Cluster for Different Status (Right)

From figures 6 to 8, it can be observed that countries with similar traits ("Life Expectancy" and "Adult Mortality"/ "Income Composition of Resources"/ "Schooling") are clustered together and that developed countries tend to be in clusters with higher life expectancy. Thus, this further reinforces that developed countries possess higher levels of life expectancy. Additional plotly visualisation has been done in the code to better visualise the data on the world map and observe where countries with similar traits are located.

Conclusion

Although the data is not very reliable, the data has been thoroughly cleaned so as to provide relevant information and trends, ensuring that key data in the dataset can still be used for analysis.

Through analysis of the given dataset with the necessary tools like linear regression, K-means clustering coupled with K++ Means, and plotly. As shown above, Adult Mortality and Income composition of Resources are the best predictors for life expectancy. Using these predictors, we have then affirmed that developed countries generally possessed higher life expectancy levels as compared to developing countries.

Appendix A. References

1. K. Rajarshi, “Life Expectancy (WHO),” *Kaggle*, 10-Feb-2018. [Online]. Available: https://www.kaggle.com/kumarajarshi/life-expectancy-who. [Accessed: 01-Apr-2021].
2. “Immunization Country Profile,” *World Health Organization*. [Online]. Available: https://apps.who.int/immunization_monitoring/globalsummary/countries?countrycriteria%5Bcountry%5D%5B%5D=CHN. [Accessed: 03-Apr-2021].
3. “Mortality rate, infant (per 1,000 live births),” *Data*. [Online]. Available: https://data.worldbank.org/indicator/SP.DYN.IMRT.IN. [Accessed: 03-Apr-2021].
4. “Mortality rate, under-5 (per 1,000 live births),” *Data*. [Online]. Available: https://data.worldbank.org/indicator/SH.DYN.MORT. [Accessed: 03-Apr-2021].
5. “Rankings - Mean BMI > BMI > Data Visualizations > NCD,” *RisC*. [Online]. Available: https://ncdrisc.org/bmi-mean-ranking.html. [Accessed: 03-Apr-2021].

Appendix B. Contribution Table

C23 Team 9	Coding (%)	Report (%)	Presentation (%)
Ng Yu Xuan	50	50	50
Ng Jun Xiang	50	50	50