Global Life Expectancy

Polina Ovchinnikova

Yale School of Public Health
BIS 634: Computational Methods for Informatics
Fall 2022

INTRODUCTION

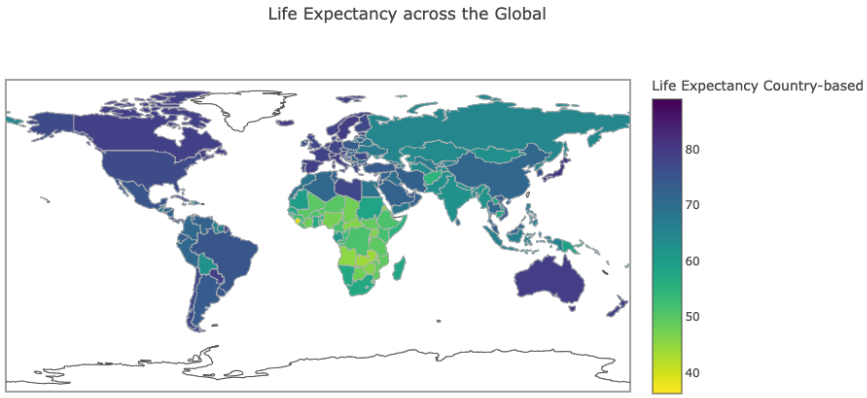
1.1 Background / Motivation

The online data set is the Life Expectancy (WHO), which consists of data from 2000 to 2015 for 193 countries. The World Health Organization (WHO) keeps track of all countries' health status and other related factors. The dataset is made available to the public for health data analysis. It is taken from Kaggle, where it is available and licensed on Kaggle. Furthermore, it follows the FAIRness principles as it is

- Findable - dataset is public and can be easily found through an internet search
- Accessible - people can copy, modify, distribute and perform work without asking permission (via Kaggle API)
- Interoperable - the dataset is stirred in .cvs format and has a formal, accessible, shared, and broadly accessible application
- Reusable - the dataset is published with clear and accessible data usage license

They are mainly used for research data. Still, the principles apply to any open-access digital resource related to scientific activity. The dataset related to life expectancy, and health factors for 193 countries during 2000-2015, consists of 22 columns and 2938 rows collected from the same WHO data repository website. Furthermore, we found no evident errors as the datasets were from WHO. Missing data were handled by filling null values with the means. The result indicated that most of the missing data were for population, Hepatitis B, and GDP.

The motivation behind this project was to find some factors that play a role in life expectancy. The focus was to see if countries' "Developing" and "Developed" status plays a significant role in life expectancy and other variables. Developing countries have a low gross domestic product (GDP) per person and tend to rely on agriculture as their prime industry. So, we wanted to see how it might describe other factors that play a role in life expectancy. Looking at the graph below, we can note that counties primarily located in Africa and parts of Asia have a lower life expectancy.



Acknowledgments: The data was collected from WHO and United Nations websites with the help of Deeksha Russell and Duan Wang.

1.2 Description of Dataset

The dataset related to life expectancy, and health factors for 193 countries during 2000-2015, consists of 22 columns and 2938 rows collected from the same WHO data repository website.

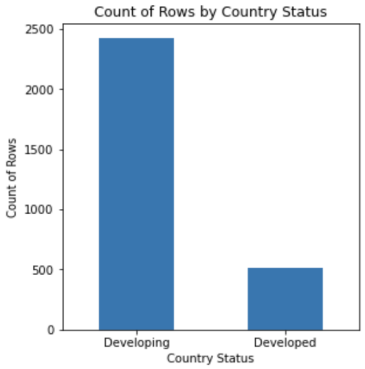
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Country                          2938 non-null   object
 1   Year                             2938 non-null   int64
 2   Status                           2938 non-null   object
 3   Life expectancy                  2928 non-null   float64
 4   Adult Mortality                  2928 non-null   float64
 5   infant deaths                    2938 non-null   int64
 6   Alcohol                          2744 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2385 non-null   float64
 9   Measles                          2938 non-null   int64
 10   BMI                             2904 non-null   float64
 11  under-five deaths                2938 non-null   int64
 12  Polio                            2919 non-null   float64
 13  Total expenditure                2712 non-null   float64
 14  Diphtheria                       2919 non-null   float64
 15   HIV/AIDS                        2938 non-null   float64
 16  GDP                              2490 non-null   float64
 17  Population                       2286 non-null   float64
 18   thinness  1-19 years            2904 non-null   float64
 19   thinness 5-9 years             2904 non-null   float64
 20  Income composition of resources  2771 non-null   float64
 21  Schooling                        2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

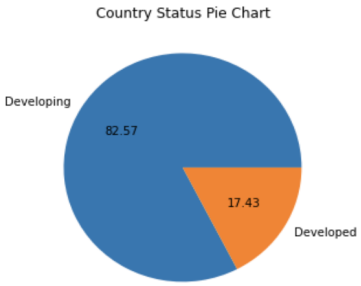
Columns Meaning:

- Country: Country
- Year: Year
- Status: Country Developed or Developing status
- Life expectancy: Life expectancy in age
- Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- Infant deaths: Number of Infant Deaths per 1000 population

- Alcohol: Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol) -percentage expenditure: Expenditure on health as a percentage of - Gross Domestic Product per capita(%)
- Hepatitis B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- Measles: Measles - number of reported cases per 1000 population
- BMI: Average Body Mass Index of the entire population
- Under-five deaths: Number of under-five deaths per 1000 population
- Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)
- Total expenditure: General government expenditure on health as a percentage of total government expenditure (%)
- Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years)
- GDP: Gross Domestic Product per capita (in USD)
- Population: Population of the country
- Thinness 1-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- Thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9(%)
- Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Schooling: Number of years of Schooling(years)

The graph below shows that the majority of the data comes from countries listed as 'Developing' - 82.57%. And given that the majority of the data comes from developing countries rather than developed countries, it is likely that any model used will reflect findings for developing countries more correctly than developed countries.

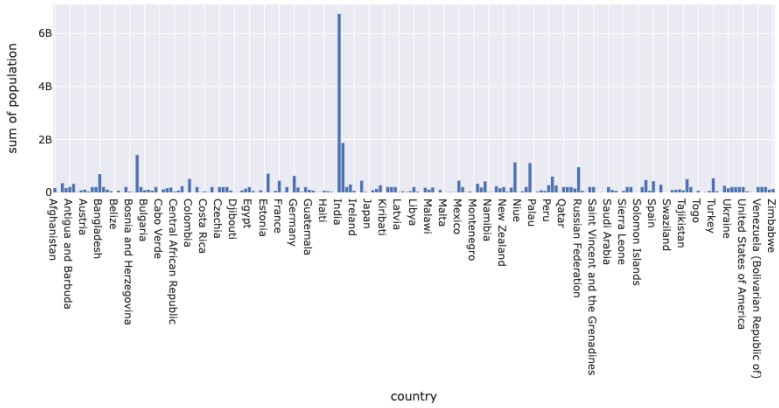




The table below shows that "Afghanistan" occurs the most frequently in the provided data set. As a result, there may be some correlation with some of the indirect analyses that may be done.

	Country	Status
count	2938	2938
unique	193	2
top	Afghanistan	Developing
freq	16	2426

In addition, if we look at the graph below that shows the sum of the population it can be noted that India has the highest sum.



Here we can observe the top 10 best and worst life expectancy counties. On the right, we can observe that Japan, followed by Sweden, has the highest life expectancy. And on the left, counties like Sierra Leone and the Central African Republic have one of the lowest life expectancies.

	country	life_expectancy
84	Japan	82.53750
165	Sweden	82.51875
75	Iceland	82.44375
166	Switzerland	82.33125
60	France	82.21875
82	Italy	82.18750
160	Spain	82.06875
7	Australia	81.81250
125	Norway	81.79375
30	Canada	81.68750

	country	life_expectancy
152	Sierra Leone	46.11250
31	Central African Republic	48.51250
94	Lesotho	48.78125
3	Angola	49.01875
100	Malawi	49.89375
32	Chad	50.38750
44	Côte d'Ivoire	50.38750
192	Zimbabwe	50.48750
164	Swaziland	51.32500
123	Nigeria	51.35625

DATA CLEANING

2.1 Data Cleaning

The data consisted of some null values that have been dealt with by filling them in with the data's mean values. The majority of missing values in the data came from: Population, Hepatitis B, and GDP. Plus, any extra empty space was removed and the variable of the dataset was changed to lower case to make it easier to use it.

```
# Looking for null value in the data
df.isnull().sum()
```

```
Country                             0
Year                                0
Status                              0
Life expectancy                    10
Adult Mortality                    10
infant deaths                       0
Alcohol                           194
percentage expenditure              0
Hepatitis B                       553
Measles                             0
 BMI                               34
under-five deaths                   0
Polio                              19
Total expenditure                 226
Diphtheria                         19
 HIV/AIDS                           0
GDP                               448
Population                        652
 thinness  1-19 years              34
 thinness 5-9 years                34
Income composition of resources   167
Schooling                         163
dtype: int64
```

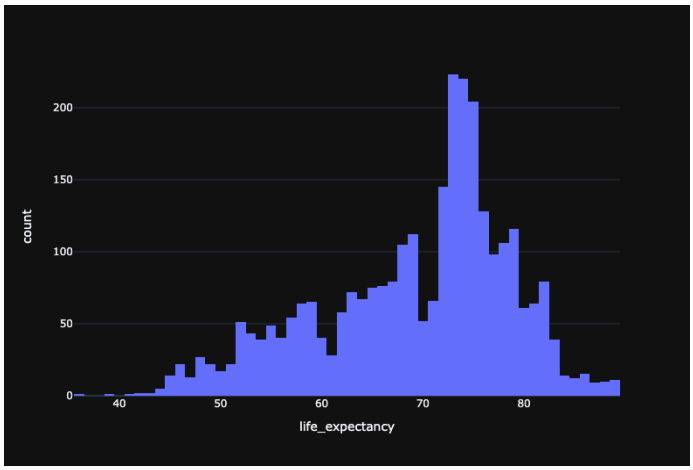
```
# Looking for null value in the data after fitting
df.isnull().sum()
```

```
Country                           0
Year                              0
Status                            0
Life expectancy                   0
Adult Mortality                   0
infant deaths                     0
Alcohol                           0
percentage expenditure            0
Hepatitis B                       0
Measles                           0
 BMI                              0
under-five deaths                 0
Polio                             0
Total expenditure                 0
Diphtheria                        0
 HIV/AIDS                         0
GDP                               0
Population                        0
 thinness  1-19 years             0
 thinness 5-9 years               0
Income composition of resources   0
Schooling                         0
dtype: int64
```

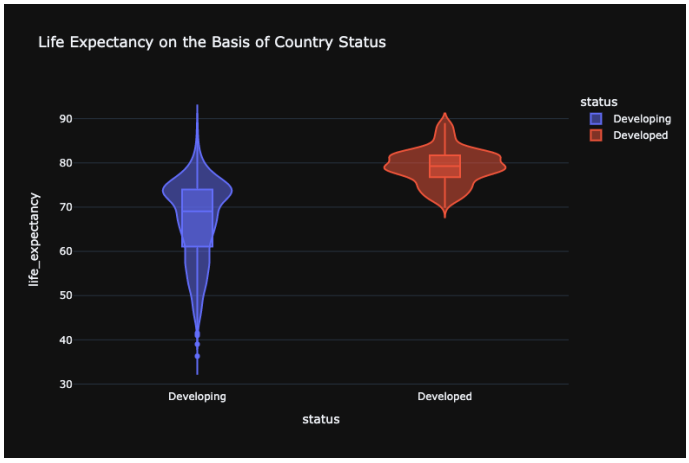
2.2 Distribution

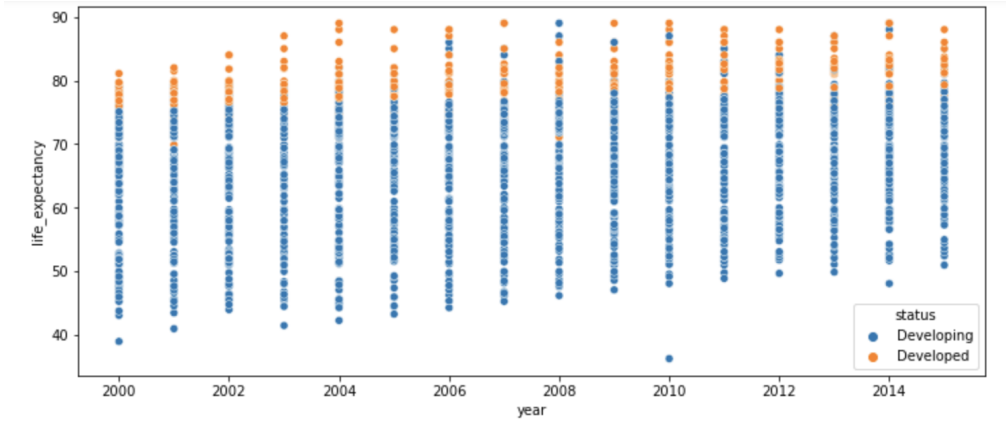
Visually, it is plain to see several outliers for all of these variables - including the target variable, life expectancy.

- Furthermore, the highest distribution of life expectancy is between 75 years of age and 80.



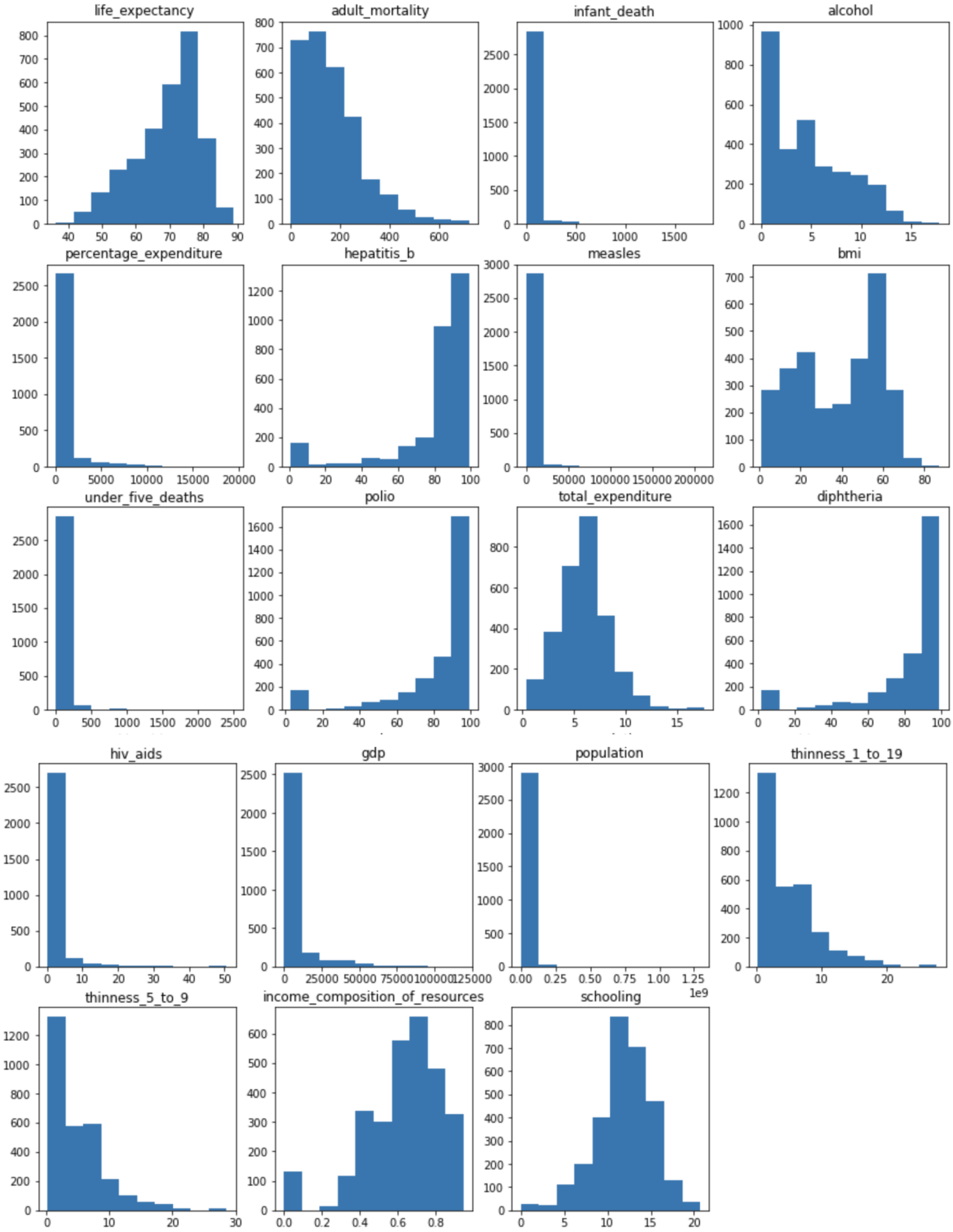
- Looking at the graph below, it can be seen that “Developed” countries have a far less life expectancy range, only around 65 and a bit over 90. In “Developing” countries, we can note that the range is really wide compared to “Developed” countries and ranges from 30 years of age to a bit over the 90s. Furthermore, by comparing the means (the wide part of the violin graph), it could be observed that “Developing” countries have a mean of around 73 and “Developed” around 80.





- This graph also allows us to see how the range varies depending on the state of each country. Throughout the past 15 years, each status's range has remained constant.

The distribution of each variable is shown in the graph below, highlighting any outliers and demonstrating that even the targeted variable, life expectancy, does not follow a normal distribution. Therefore, the majority of the variables do not follow a normal distribution.


life_expectancy
adult_mortality
infant_death
alcohol
percentage_expenditure
hepatitis_b
measles
bmi
under_five_deaths
polio
total_expenditure
diphtheria
hiv_aids
gdp
population
thinness_1_to_19
thinness_5_to_9
income_composition_of_resources
schooling
1e9

2.3 Outliers

When applying Tukey's approach, an outlier is defined as more than 1.5 times the IQR. Additionally, it enabled us to determine the percentage of data that is an outlier as well as how many outliers each variable has:

----------life_expectancy----------
Number of outliers: 17
Percent of data that is outlier: 0.58%
----------adult_mortality----------
Number of outliers: 86
Percent of data that is outlier: 2.93%
----------infant_death----------
Number of outliers: 315
Percent of data that is outlier: 10.72%
----------alcohol----------
Number of outliers: 3
Percent of data that is outlier: 0.1%
----------percentage_expenditure----------
Number of outliers: 389
Percent of data that is outlier: 13.24%
----------hepatitis_b----------
Number of outliers: 316
Percent of data that is outlier: 10.76%
----------measles----------
Number of outliers: 542
Percent of data that is outlier: 18.45%
----------bmi----------
Number of outliers: 0
Percent of data that is outlier: 0.0%
----------under_five_deaths----------
Number of outliers: 394
Percent of data that is outlier: 13.41%
----------polio----------
Number of outliers: 279
Percent of data that is outlier: 9.5%
----------total_expenditure----------
Number of outliers: 51
Percent of data that is outlier: 1.74%
----------diphtheria----------
Number of outliers: 298
Percent of data that is outlier: 10.14%
----------hiv_aids----------
Number of outliers: 542
Percent of data that is outlier: 18.45%
----------gdp----------
Number of outliers: 300

Percent of data that is outlier: 10.21%
----------population----------
Number of outliers: 194
Percent of data that is outlier: 6.6%
----------thinness_1_to_19----------
Number of outliers: 100
Percent of data that is outlier: 3.4%
----------thinness_5_to_9----------
Number of outliers: 99
Percent of data that is outlier: 3.37%
----------income_composition_of_resources----------
Number of outliers: 130
Percent of data that is outlier: 4.42%
----------schooling----------
Number of outliers: 77
Percent of data that is outlier: 2.62%

With 542, or 18.45% of the data, the "measles" and "HIV/AID" variable has the highest percentage of outliers. Then comes "Under the five deaths," which has 394 values and a 13.41% outlier rate in the data.

Since none of the outlier percentages in the data reached 25%, it was decided to keep the outlier because it would only slightly affect the analysis's results.

DATA ANALYSIS
3.1 Linear Regression

OLS Regression Results

Dep. Variable:	life_expectancy	**R-squared:**	0.820
Model:	OLS	**Adj. R-squared:**	0.819
Method:	Least Squares	**F-statistic:**	663.3
Date:	Mon, 12 Dec 2022	**Prob (F-statistic):**	0.00
Time:	17:58:51	**Log-Likelihood:**	-8268.0
No. Observations:	2938	**AIC:**	1.658e+04
Df Residuals:	2917	**BIC:**	1.670e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	73.4394	34.723	2.115	0.035	5.356	141.523
year	-0.0092	0.017	-0.533	0.594	-0.043	0.025
status	1.5897	0.270	5.886	0.000	1.060	2.119
adult_mortality	-0.0198	0.001	-24.926	0.000	-0.021	-0.018
infant_death	0.0998	0.008	11.839	0.000	0.083	0.116
alcohol	0.0620	0.026	2.381	0.017	0.011	0.113
percentage_expenditure	8.534e-05	8.47e-05	1.008	0.314	-8.07e-05	0.000
hepatitis_b	-0.0147	0.004	-3.752	0.000	-0.022	-0.007
measles	-1.96e-05	7.66e-06	-2.558	0.011	-3.46e-05	-4.58e-06
bmi	0.0444	0.005	8.998	0.000	0.035	0.054
under_five_deaths	-0.0747	0.006	-12.094	0.000	-0.087	-0.063
polio	0.0285	0.004	6.385	0.000	0.020	0.037
total_expenditure	0.0661	0.034	1.930	0.054	-0.001	0.133
diphtheria	0.0402	0.005	8.544	0.000	0.031	0.049
hiv_aids	-0.4708	0.018	-26.667	0.000	-0.505	-0.436
gdp	3.347e-05	1.3e-05	2.571	0.010	7.94e-06	5.9e-05
population	2.751e-10	1.69e-09	0.163	0.871	-3.04e-09	3.59e-09
thinness_1_to_19	-0.0818	0.050	-1.624	0.105	-0.181	0.017
thinness_5_to_9	0.0073	0.050	0.147	0.883	-0.090	0.105
income_composition_of_resources	5.7738	0.641	9.003	0.000	4.516	7.031
schooling	0.6574	0.042	15.693	0.000	0.575	0.740

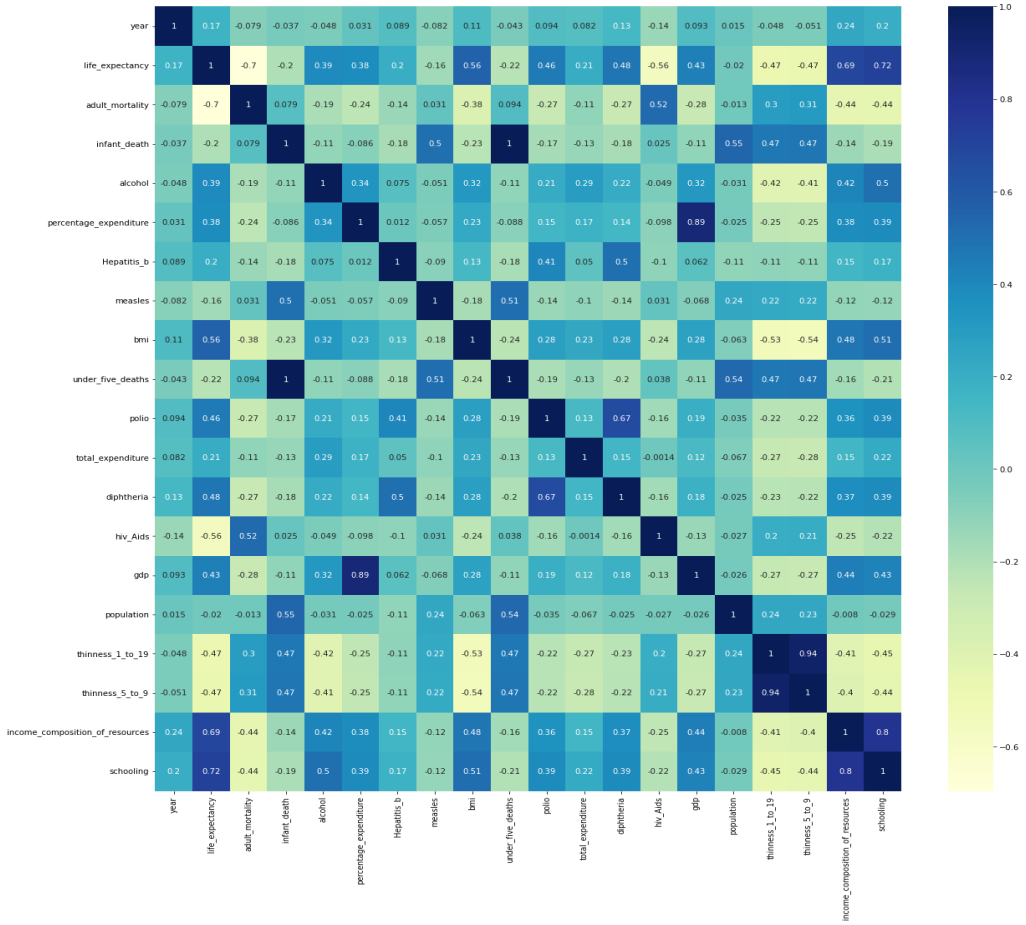
By analyzing the linear regression, the R-squared of 0.820 shows us that there is a strong correlation between the variables and the dependent value of Life Expectancy. Also, it should be

noted that the variables that have shown statistical significance looking at the p-value are “Alcohol,” “ Measles,” and “GDP.”

Omnibus:	135.918	**Durbin-Watson:**	0.701
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	398.080
Skew:	-0.175	**Prob(JB):**	3.62e-87
Kurtosis:	4.769	**Cond. No.**	2.57e+10

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.57e+10. This might indicate that there are strong multicollinearity or other numerical problems.

3.2 Correlation



```
schooling                          0.715066
adult_mortality                    0.696359
income_composition_of_resources    0.692483
bmi                                0.559255
hiv_Aids                           0.556457
diphtheria                         0.475418
thinness_1_to_19                   0.472162
thinness_5_to_9                    0.466629
polio                              0.461574
gdp                                0.430493
alcohol                            0.391598
percentage_expenditure             0.381791
under_five_deaths                  0.222503
total_expenditure                  0.207981
Hepatitis_b                        0.203771
infant_death                       0.196535
year                               0.169623
measles                            0.157574
population                         0.019638
Name: life_expectancy, dtype: float64
```

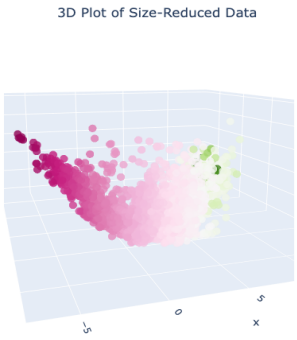
Based on the correlation, we can note that there are a few more variables that also might play a role in life expectancy, as the there correlation is close to 1, showing a strong correlation and impact on Life Expectancy:

- Life expectancy + Schooling = 0.72
- Life expectancy + Adult Mortality = 0.69
- Life expectancy + Income of Resources = 0.69
- Life expectancy + BMI = 0.56
- Life expectancy + Diphtheria = 0.48
- Life expectancy + Polio = 0.46
- Life expectancy + Population = 0.43

Thus, not only the status of the country but also factors such as GDP, Schooling, Adult mortality (expected), Income of Resources, Alcohol, and Measles play a significant role in life expectancy.

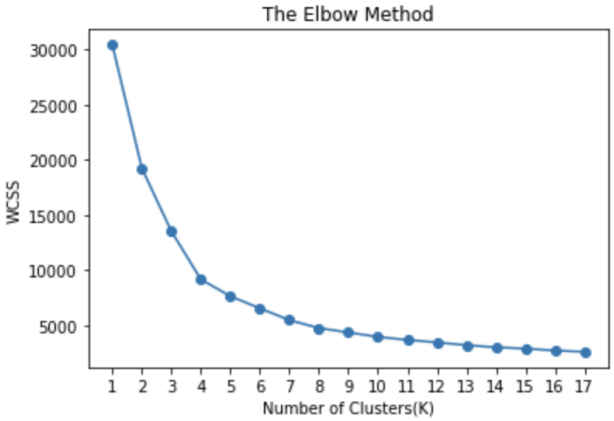
3.3 PCA & Elbow Method

High-dimensional data processing will require a lot of computing power and cost. Therefore, the dimensional reduction must be made before K-Means Clustering can be performed. We can therefore use PCA to preprocess the data before K-Means Clustering is performed.



Thus, based on the PCA plot, there are 3 clusters (pink, light pink, and green); therefore, we can now implement the Elbow Method to find the optimal number of clusters.

For the Elbow Method, each cluster is formed by calculating and comparing the data point distance within a cluster to its center. We employ the Within-Cluster-Sum-of-Squares (WCSS) method to determine the appropriate number of clusters. The objective of WCSS is to minimize the sum of the squares representing the distances from each data point in each cluster to its associated center. The WCSS tends to be 0 because the data points become the centers, and the distance will be zero. This will result in a perfect cluster, but this is nearly impossible because there are many clusters as there are observations. Assume that there are n observations in a dataset and that we specify n number of clusters, which means k = n. As a result, by fitting the model over a range of K, we utilize the Elbow graph to determine the best value for K. For a variety of K values, we initialize the K-Means method at random before plotting it against the WCSS.

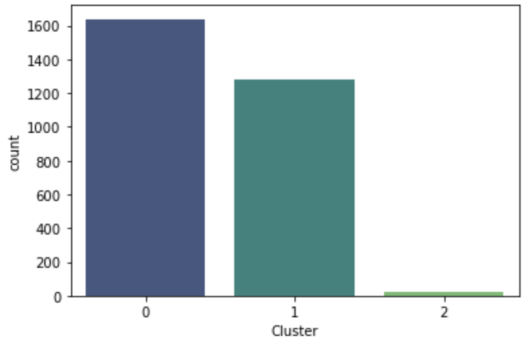


This means that the ideal value for K is 3, which we can note. Additionally, it should be noticed that the WCSS value falls as the number of clusters increases. The "elbow," depending on the rate of reduction, indicates the model fits best at that moment. Therefore, choose K accordingly. The graph shows a dramatic decrease in WCSS from clusters 1 to 2 to 3. The decline is minor after 3. So, we settled on 3 as K's ideal value. The Elbow Method lets us determine that three clusters are the optimal balance.

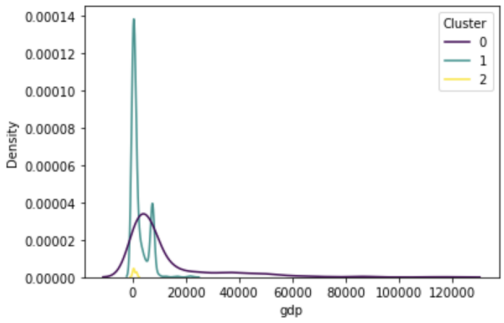
3.3 K-means Clustering

For K-Means Clustering, to find the life expectancy, we set the k=3, as the PCA and Elbow Method showed that the optimal number of clusters would be 3. Thus, we divided the cluster into 3 categories:

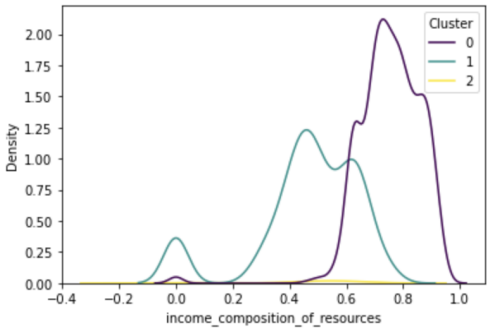
Cluster 0 = High life expectancy,
Cluster 1 = Moderate life expectancy,
Cluster 2 = Low life expectancy



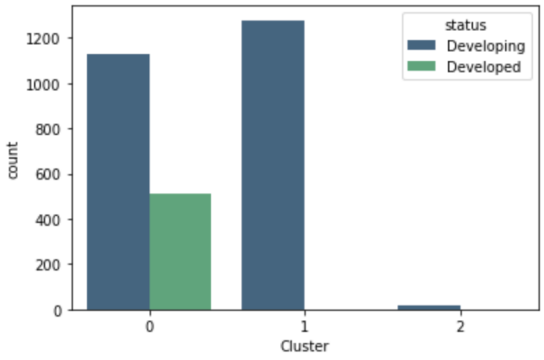
Here, low life expectancy has a minor count compared to high and moderate life expectancy, indicating that most of the countries in this data have mode high to moderate life expectancy.



Furthermore, looking more into how GDP plays a role in life expectancy, we can note that moderate life expectancy has the most density compared to low life expectancy (that is barely visible) and high life expectancy.



Income comparison of resources shows that high life expectancy has the most resources, which does not come as a surprise. However, it was interesting to note that low life expectancy, even though barely seen, does have some of the high-income resources at some point (line near 1).

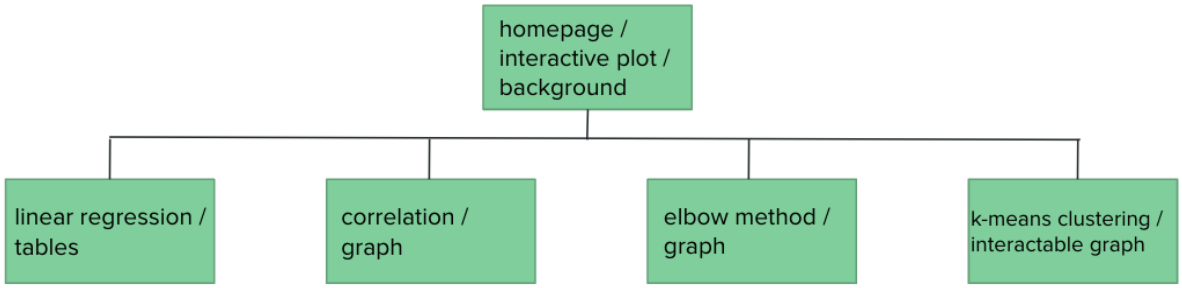


As it should be recalled having “Developing” countries take up to 83% of the data, we should note that high life expectancy has the most “Developed” and moderate life expectancy has the most “Developing” countries. However, I was surprised to see that not many countries have a low life expectancy, as I was expecting more based on the number of “Developing” countries in the dataset.

API and WEB FRONT-end

4.1 Server API

Using Flask, the below API server framework was created:



4.2 Web front-end.

The web contains a home page that gives an overview of the data, a bit of background, and an interactive world plot that would allow us to see life expectancy in different parts of the world. Then 4 tabs would allow users to view the analyzes that have been performed:

- Linear Regression - tables to see the correlation and p-values of the variables mapped against the dependent variable of life expectancy
- Correlation - plot allows seeing the relationship between variables
- Elbow Method - intractable plot to see how the K and the WCSS as you look at different x values along the line
- K-Means Clustering - an interactive plot that allows changing of the number of clusters to see how it affects the position of countries when it comes to status

DISCUSSION

4.1 Surprising Analysis

Some of the surprising analyzes that I encountered during the project were:

1. The percentage of "Developing" countries in the dataset and how it plays a role in the K-means clustering analysis. Furthermore, I found it interesting that "Afghanistan" has the most significant frequency.
2. GDP, although playing a significant role in life expectancy, only correlated with 0.43. However, it did show statistical significance when p-values were conducted using linear regression
3. I was surprised that the "measles" & "HIV/AIDS" variables have the most significant outliers, with 542, making it 18.45% of the data.

Thus, I believe it is critical to continue the study of life expectancy, especially in "Developing" countries, as it can showcase the significant facts that might play a vital role in human life.

4.2 Difficulties

The biggest challenge of this project was creating a webpage using Flask, as it was my first time using Python to create an integrative webpage that would include an indescribable graph. So, integrating plot graphs and the K-mean clustering page would have been the most challenging part regarding the webpage construction. Furthermore, I found it challenging to develop needed graphics that would describe the data well, as 22 variables could have been shown and analyzed in this project.

4.3 Conclusion

It is important to see what factors affect life expectancy as it would help to understand how life expectancy could be made longer in countries where the life expectancy mean is only around 50 years.

By observing the analyses, we could conclude that by looking at the "Developing" and "Developed" life expectancy, schooling might be one of the significant values that play a role in life expectancy. But we should note that some limitations of the data consist in the number of outliers some variables have and the skewness of the dataset towards the "Developing" counties. Furthermore, for future work, it would be interesting to see if the analyzes would change if outliers were removed. As well as performing more predictable algorithms/ machine learning analysis to see what variables they would choose that have the most impact on life expectancy.

REFERENCES

- Kumar R.(2017). Life Expectancy(WHO).from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
- https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/
- https://en.wikipedia.org/wiki/Elbow_method_(clustering)

- https://www.datasklr.com/principal-component-analysis-and-factor-analysis/principal-component-analysis