MCS - 7103 - ASSIGNMENT ONE
OKELLO ANDREW PETERS
2024/HD05/21944U 2400721944

RIGOROUS EXPLORATORY DATA ANALYSIS OF DATA CONCERNING THE VARIABLES INFLUENCING LIFE EXPECTANCY IN COUNTRIES UTILIZING PYTHON LIBRARIES INCLUDING PANDAS, MATPLOTLIB, AND SEABORN

Brief Background and Motivation:

In a recent post about **THE STATE OF UGANDA'S ECONOMY AT A GLANCE,** the Ministry of Planning, Finance, and Economic Development (MOFPED) mentioned that life expectancy of Ugandans has increased from 63 years in 2011 to 68 years in 2022. This is the ninth of the 10-fold growth points listed. These days, people are concerned about how long they will live because of new variables that are posing a threat to human health, like processed foods, industrialization, illness, and many others.

Considering the background above, I chose a raw dataset called life expectancy to better explore the relationship between life expectancy and a number of putative contributing factors. These results can assist national governments and policy makers in making the best choices and implementation of programs that will increase the average life expectancy of their citizens.

Table of Contents

Research Questions before the Dataset

1. What is Life Expectancy?
2. How do we measure life expectancy?
3. What are the control variables?
4. How can we increase a person's life expectancy?

Research Questions after the Dataset

1. What are the main control variables that determine life expectancy?
2. What variable should be maximized to elevate life expectancy?
3. What variable should be minimized to boost life expectancy?

Disclaimer

To prevent this report from being bulky, I only inserted graphs or visuals for multivariate analysis since that's where my solid findings came from. However, the rest of the graphs for univariate and bivariate analysis can be clearly seen in the python file. I also included two graphs for Bivariate analysis

Data Wrangling

During this phase, I thoroughly cleaned the data.
I began by removing any extra white space from the feature names (column names) that appeared at the beginning or end of the string. I took out the white spaces.
I must encode the categorical data to display it on a heatmap because categorical data cannot be displayed on one.

To prevent squandering my work on things that don't directly affect life expectancy. To gain a broad idea of how each attribute and the target variable were correlated, I utilized an annotated heatmap. All columns with correlations less than 0.5 were deliberately omitted.

I also checked for duplicates and null values.

EXPLORATORY DATA ANAYSIS (EDA)

UNIVARIATE ANALYSIS

Status Column

Using the histogram, we can observe that most of the countries used in this dataset where developing countries since there frequency is way too higher than the frequency of the developed countries. Developed countries are represented by zero and developing countries are represented by one.

Adult Mortality Column

The distribution of this feature shows that the data in this column is skewed to the right, which means that we have more observations at the left, however we have some extreme values that are pulling the data to the right. This means that the mean is greater than the median. It also demonstrates the presence of some outliers in this feature.

Life Expectancy Column

According to the feature distribution, it is skewed to the left which means that most values are at the right however, we have extreme values which are pulling the data to the left. This means that the mean is less than the median value. This demonstrates the presence of outliers which I must deal with.

Infant Deaths Column

The distribution is skewed to of the data makes sense since some countries may have zero deaths of infants and others may have more.

BMI Column

According to the histogram, we can observe that the BMI column has two distinct peaks or modes areas where values are concentrated. This means that the data has two different groups or patterns. There are two main values where values are concentrated.

Under-five deaths

This means that the higher the population density of a particular country, the lower the under-five deaths in that country.

Polio

The higher the population density of a particular country, the higher the reported polio cases in that country.

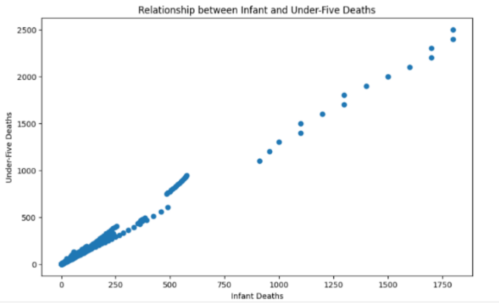
Diphtheria

The data is skewed to the left which means most values are at the right. However, there are outliers that are skewing the data towards the left. According to the distribution, the higher the population density of a specific country, the higher the diphtheria cases reported by that country.

BIVARIATE ANALYSIS

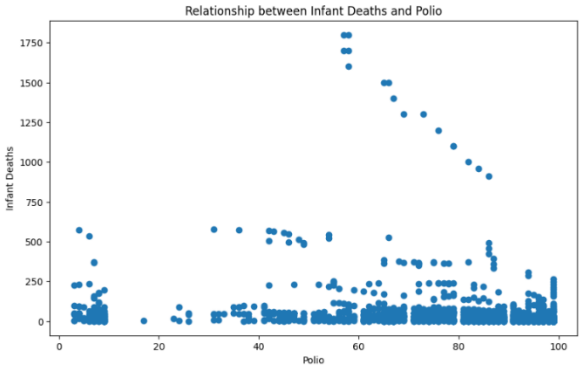
I'm attempting to understand how each feature and the goal feature relate to one another in this analysis. I'm attempting to ascertain which variables have a greater influence on the target variable and whether there is a direct or inverse link between the two.

Infant deaths vs Under five deaths



From the scatter plot of these two variables, it observed that these two features are directly proportional to each other. Which means that as infant deaths increase, under-five deaths also increase, and the vice visor is true.

Infant Deaths vs Polio



According to the scatter plot, as more people of a given country are vaccinated with the polio vaccine, there is a strong decline in the number of infant deaths.

Under-five deaths and Life Expectancy

Life Expectancy vs. Adult Mortality

From the scatter plot of these two variables, it is observed that as the number of people who die between 15 and 16 years increases (Adult Mortality) for a given country, the life expectancy decreases, and the reverse is also true.

Status vs Life Expectancy

From the scatter plot of these two features, it is plain as day that for developed countries, the life expectancy is high since a person is expected to die when he is at least 70 years old from the scatter plot. But for developing countries which are represented by 1, the life expectancy is high and low at the same time, a person can possibly die even before they are 40 years old, but they can also make it to 90 years.

Adult Mortality vs Infant deaths

According to the scatter plot of these two features, it seems that as adult mortality increases, infant deaths increase slightly reaching a point beyond which they remain constant. This means that infants deaths are less likely to be affected by the variation in Adult Mortality rate for a given country.

Life Expectancy vs GDP

From the scatter plot, the Gross Domestic Product (GDP) positively affects life expectancy. When the GDP of a given country is low, the life expectancy is for that country is down but as the GDP increases, the life expectancy increases. If the GDP of the country increases, the people of that country will live longer since quality goods and services are produced and readily available to them.

Infant Deaths vs BMI

As the number of infant-deaths declines, BMI increases since children are not dying, they're growing into adults. But as infant deaths increase, BMI numbers drop.

Polio vs Life Expectancy

I was puzzled about if the polio column represents number of polio patients in a country or the number of people vaccinated with the polio vaccine. So, I plotted this scatter plot to help me understand the relationship between polio and life expectancy. From the plot, as the number of people vaccinated with the polio vaccine increases in a particular country, the life expectancy of people in that country increases which means the polio column represents the number of people vaccinated with the polio vaccine in given country each year.

Diphtheria vs Life Expectancy

From the scatter plot, as more people in a particular country are administered with the polio vaccine, the life expectancy of people in that country increases and the vice visor is also true.

HIV/AIDS vs Life Expectancy

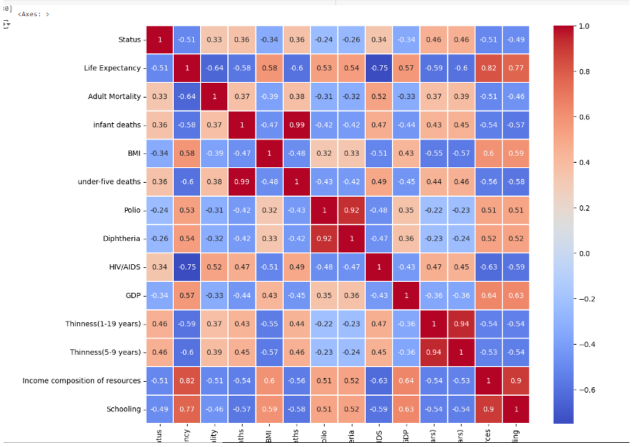
From the scatter plot, we observe that as the number of HIV/AIDS cases increase in a particular country, the life expectancy drastically drops. The two variables are inversely proportional to each other.

Schooling vs Life Expectancy

As more people in a particular country access education (schooling), the life expectancy of those people in that country also increases. The vice visor is also true.

MULTIVARIATE ANALYSIS

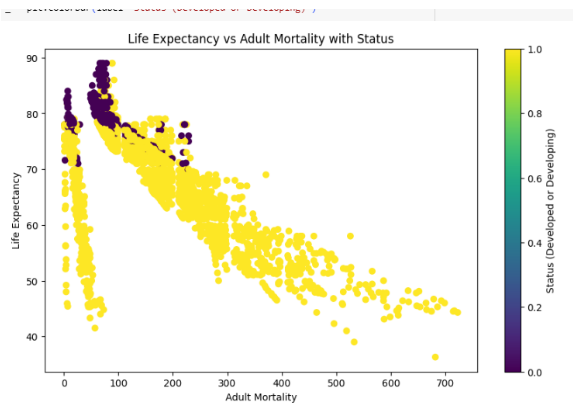
I constructed a correlation matrix using a heatmap to help me understand the correlation between the target variable and each of the other control variables. Variable or feature with a correlation greater than 0.6 were considered to have a strong impact on the output variable.



```
<Axes: >
```

	atus	ncy	ility	aths	BMI	aths	olio	eria	IDS	GDP	ars)	ars)	rces	ling
Status	1	-0.51	0.33	0.36	-0.34	0.36	-0.24	-0.26	0.34	-0.34	0.46	0.46	-0.51	-0.49
Life Expectancy	-0.51	1	-0.64	-0.58	0.58	-0.6	0.53	0.54	-0.75	0.57	-0.59	-0.6	0.82	0.77
Adult Mortality	0.33	-0.64	1	0.37	-0.39	0.38	-0.31	-0.32	0.52	-0.33	0.37	0.39	-0.51	-0.46
infant deaths	0.36	-0.58	0.37	1	-0.47	0.99	-0.42	-0.42	0.47	-0.44	0.43	0.45	-0.54	-0.57
BMI	-0.34	0.58	-0.39	-0.47	1	-0.48	0.32	0.33	-0.51	0.43	-0.55	-0.57	0.6	0.59
under-five deaths	0.36	-0.6	0.38	0.99	-0.48	1	-0.43	-0.42	0.49	-0.45	0.44	0.46	-0.56	-0.58
Polio	-0.24	0.53	-0.31	-0.42	0.32	-0.43	1	0.92	-0.48	0.35	-0.22	-0.23	0.51	0.51
Diphtheria	-0.26	0.54	-0.32	-0.42	0.33	-0.42	0.92	1	-0.47	0.36	-0.23	-0.24	0.52	0.52
HIV/AIDS	0.34	-0.75	0.52	0.47	-0.51	0.49	-0.48	-0.47	1	-0.43	0.47	0.45	-0.63	-0.59
GDP	-0.34	0.57	-0.33	-0.44	0.43	-0.45	0.35	0.36	-0.43	1	-0.36	-0.36	0.64	0.63
Thinness(1-19 years)	0.46	-0.59	0.37	0.43	-0.55	0.44	-0.22	-0.23	0.47	-0.36	1	0.94	-0.54	-0.54
Thinness(5-9 years)	0.46	-0.6	0.39	0.45	-0.57	0.46	-0.23	-0.24	0.45	-0.36	0.94	1	-0.53	-0.54
Income composition of resources	-0.51	0.82	-0.51	-0.54	0.6	-0.56	0.51	0.52	-0.63	0.64	-0.54	-0.53	1	0.9
Schooling	-0.49	0.77	-0.46	-0.57	0.59	-0.58	0.51	0.52	-0.59	0.63	-0.54	-0.54	0.9	1

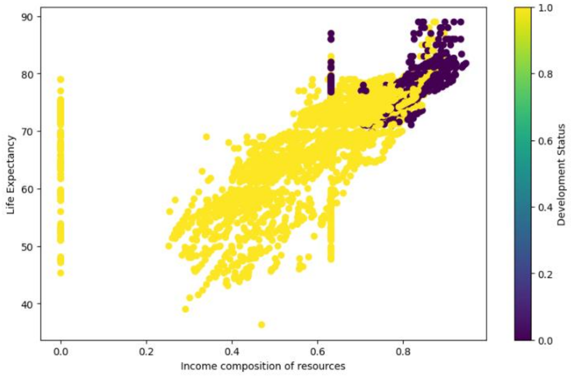
Life Expectancy vs Adult Mortality with Status of a Country



```
plt.colorbar(label='Status (Developed or Developing)')
```

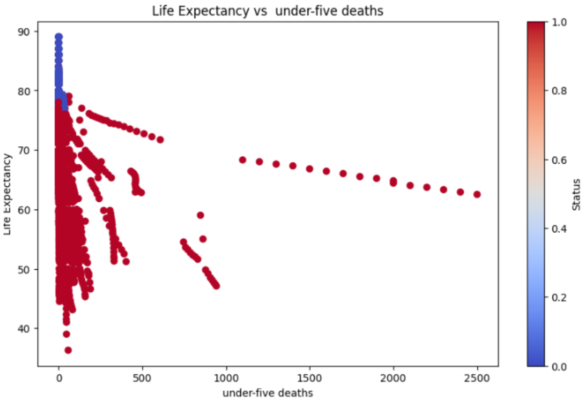
According to the scatter plot above, as the adult mortality rate increases, especially for developing countries, the life expectancy of that country declines. Developed countries have a low Adult Mortality rate which makes them to have a high life expectancy. However, there is also a slight decline in the life expectancy of developed countries whose mortality rate is slightly higher than the other countries.

Income Composition of Resources vs. Life Expectancy by Development Status



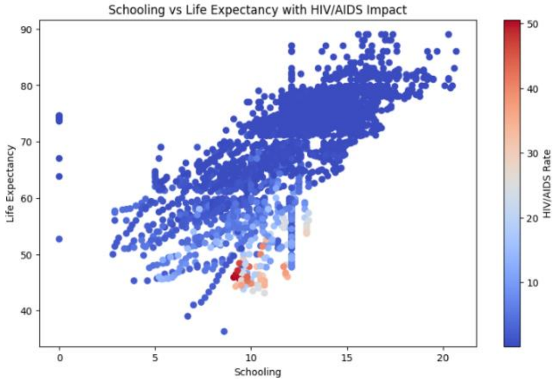
According to the scatter plot, developed countries which are represented by 0 have a larger number of income composition of resources compared to developing countries represented by 1, this has resulted into developed countries having a higher life expectancy compared to developing countries. Developed countries are represented by dark purple dots and developing countries are represented by yellow dots.

Life expectancy vs under-five deaths with status



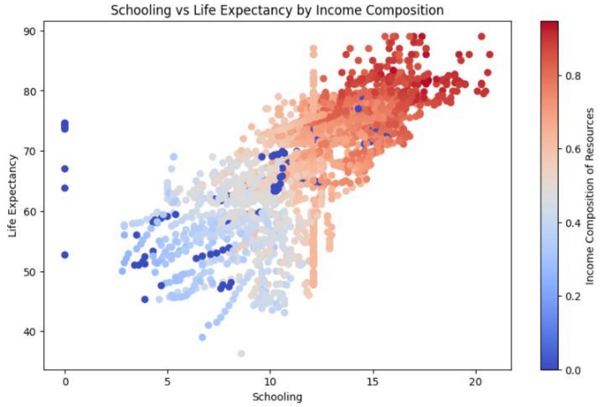
According to the scatter plot, all countries with a low number of infant deaths have a high life expectancy but still developed countries have a higher life expectancy than developing countries.

Schooling vs Life Expectancy with HIV/AIDS Impact



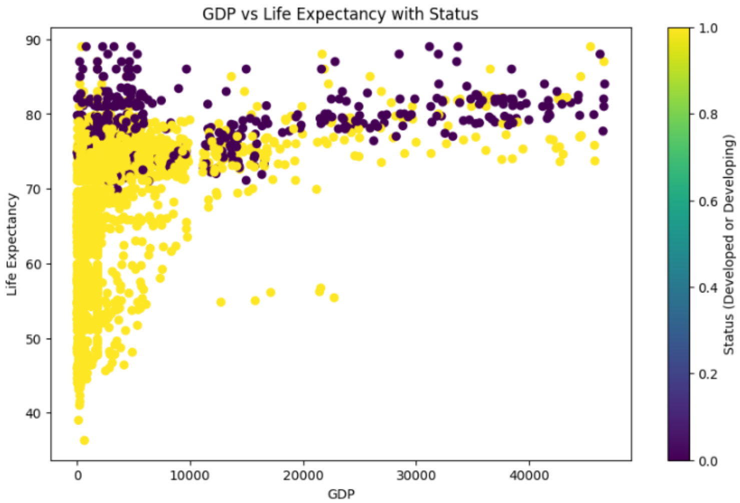
According to the distribution of data on the scatter plot, countries with more educated people who have HIV/AIDs have life expectancy between 50 to 60 years. However, in countries with a larger number of HIV/AIDS cases, regardless of the education level, the life expectancy is too low (below 50 years). Countries with zero reported cases of HIV/AIDS with high schooling have an extremely high life expectancy.

Schooling vs Life Expectancy with Income composition of resources



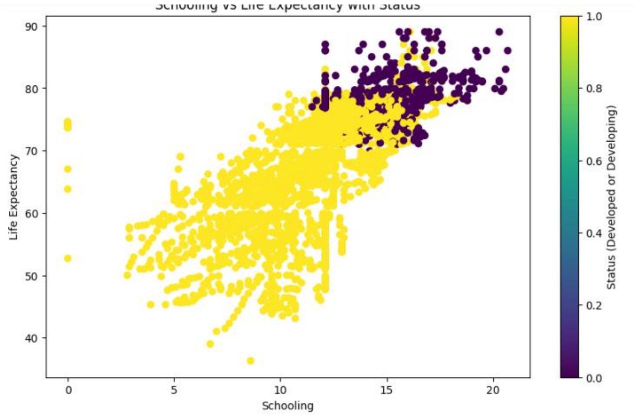
Generally, as the number of people accessing school increases for a country with a large figure of income resources, there is a drastic increase in the life expectancy of people in that country according to the scatter plot of the three variables.

GDP vs Life Expectancy with Status of a Country



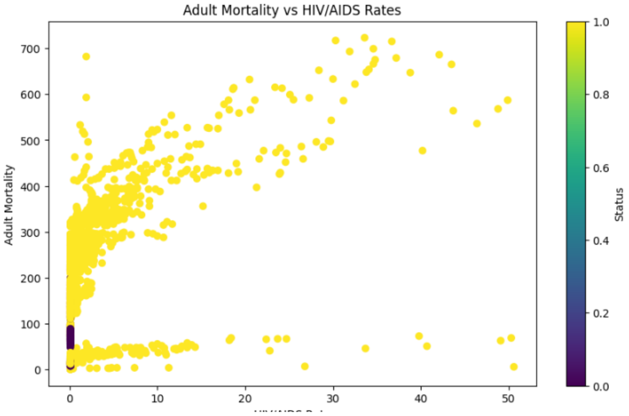
From the scatter plot of these three variables, we observe that developed countries have high life expectancy regardless of the size of their GDP. However, for developing countries, the size of their GDP matters, those with a small GDP have a lower life expectancy compared to those with a high GDP.

Schooling vs Life Expectancy with Status of a Country



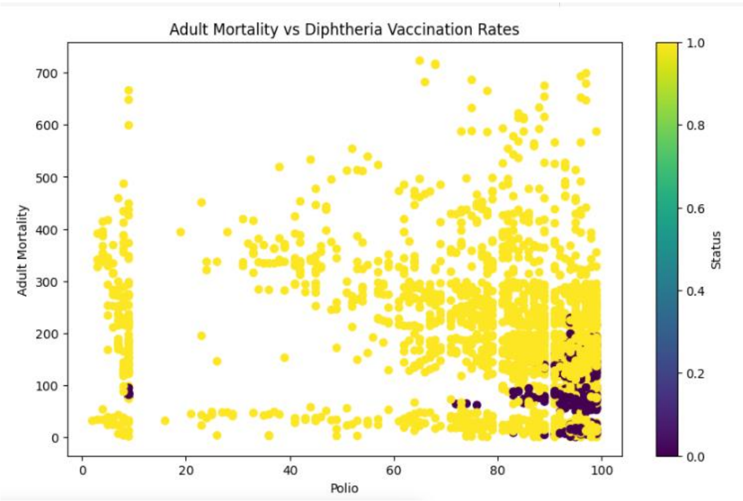
Countries with a status of developed have a high number of educated people which makes their life expectancy too high compared to developing countries. But as the number of people who access education in developing countries increases, the life expectancy equally increases.

Adult Mortality vs HIV/AIDs with Status



From the scatter plot, developing countries have high mortality rates regardless of the number of people with HIV/AIDS in their countries. It means that HIV/AIDS is not a major contributing factor to adult mortality in developing countries.

Adult Mortality vs Diphtheria with Status



As more people are vaccinated with the Diphtheria vaccine in developed countries, the Adult Mortality rate decreases. However, for developing countries, the adult mortality rate remains high even with an increase in diphtheria vaccination. This means diphtheria vaccination alone is not adequate to minimize Adult Mortality.

Conclusions

According to the Exploratory Data Analysis, the main variables to consider while measuring the life expectancy of any country are HIV/AIDS, Schooling, income composition of resources, Adulty Mortality, and under five deaths. I observed that countries with many HIV/AIDS victims had a significantly low life expectancy. I also visualized that as the Adult Mortality especially for developing countries increases, the life expectancy declines. Schooling happens to have the second highest correlation according to the correlation matrix, which means that as the given population of a country accesses education, there is drastic increase in the life expectancy of that country. Income composition of resources is a variable with strongest impact on the Life expectancy of any country, countries with an abundance of resources have a high life expectancy.

Another strong factor was under-five deaths, countries with many under-five deaths had low life expectancy. According to bivariate analysis under-five deaths are caused by infant deaths and infant deaths are caused by low rates of diphtheria and polio vaccinations.

Recommendations

1. Based on the findings of this analysis, governments and policy makers should think about optimizing the use of domestic resources to spur development, as the methods by which these resources are used have the greatest influence on a nation's life expectancy.

2. Since education has the second-highest influence on a nation's life expectancy, underdeveloped nations should think about improving the quality of their educational system by investing more in the construction of schools, institutes, and colleges as well as updating their curricula. Life expectancy rises sharply with higher levels of education.

3. Governments should also start programs where people are sensitized about how to prevent HIV/AIDS. People should be told about prevention of mother to child HIV/AIDS transmission during birth and the young adult should educated about the barrier protection methods and other methods of avoiding HIV. This lowers the number of people with HIV/AIDS thereby boosting the life expectancy of the country.

4. Governments should invest in immunization programs to minimize the number of under-five deaths.