DATA ANALYSIS

ON

LIFE EXPECTANCY

Country and Year wise Data For 16 Years From 2000-2015

Report submitted to the

SDM COLLEGE(Autonomous)

In partial fulfilment of the degree of

MASTER OF SCIENCE

IN

STATISTICS

By

SPOORTHI K

Under the supervision of

Mr. Pradeep K

Department of Postgraduate Studies

in Statistics

SRI DHARMASTHALA MANJUNATHESHWARA

COLLEGE(Autonomous)

UJIRE-57424

March 2023

Contents

1 Chapter 1

1.1Introduction:

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that effect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single data-set. On initial visual inspection of the data showed some missing values. As the data-sets were from WHO, we found no evident errors. Missing data was handled in R software by using Miss map command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model data-set. The final merged file (final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and social factors.

1.2 Objectives and Scope of Study:

- Does various predicting factors affect the Life expectancy? What are the predicting variables actually affecting the life expectancy?
- What is the impact of schooling on the lifespan of humans?
- Does Life Expectancy have positive or negative relationship with income composition?
- Is there a significant difference between life expectancy in developing countries and developed?
- Which factors contribute highly to life expectancy?
- Predicting life expectancy for future data of India and its neighbouring countries.
- Predicting life expectancy for future data of top 10 highly populated countries.

Scope of the study

- People can become more aware of their general health and improvise on it. Insurance Companies can use these predictions to provide individualized services.
- Governments can use predictions to allocate limited resources efficiently. Social welfare, Health-care funding to individuals and in areas of greater needs can be assigned effectively.
- It can benefit for policy making, and help optimize an individual's health, or the services they receive.

1.3 Literature Review:

V.M Shkolnikov et al, proposed that Predicting life span for human being is a vital step. It is an emerging research area that is gaining interest but involved lot of challenges due to the limited number of resources (i.e., datasets) available. By obtaining the Date of birth, Environmental factors, Food habits, Diseases and Medical history, a lot of investigations will be conducted to predict the sustainability of human.

D.F.Andrews et al, proposed that when there is change in small fraction the data techniques will be resistant. Otherwise, when the efficiency of statistics held high then the techniques will be robust. If the accuracy score is excellent then the result of the predicted one is accurate.

D.M.J Naimark et al, proposed that the expectancy of the life can be grasped to equal to area under a certain region He proposed it is necessary to understand the baseline risk under the control group. By the help of different models, we can predict the life expectancy.

A.A. Bhosale et al, proposed that expectancy of the life mainly target on predicting models using trends. He proposed life expectancy rely on weight, adult mortality, heart rate, respiration rate for human beings. The inspection provides the standard life expectancy is forecasted by variables that can be easily calculated.

M.K.Z. Sormin et al, proposed to rough calculate the life expectancy of the population across the world so that it will be helpful to the particular country to increase their health of the human beings. The Cyclic Order Weight neural network method is used for the appraise.

K.J.Preacher et al, proposed that slopes, significance and bands of confidence are used to test the steps in multiple linear regression but this trend has been outdated and extended to multilevel linear regression. When one dependent value is depended on more independent values then we use multiple linear regression.

2 Chapter 2:

2.1 Materials and Methods:

A Life Expectancy data has been collected from the website of “https://www.kaggle.com” this data was collected during the year 2000-2015.

2.2 About the Data:

country (Nominal) - the country name

year (Ordinal) - the calendar year the indicators are from (ranging from 2000 to 2015)

status (Nominal) - whether a country is considered to be 'Developing' or 'Developed' by WHO standards

life_expectancy (Ratio) - the life expectancy of people in years for a particular country and year

adult_mortality (Ratio) - the adult mortality rate per 1000 population (i.e. number of people dying between 15 and 60 years per 1000 population); if the rate is 263 then that means 263 people will die out of 1000 between the ages of 15 and 60; another way to think of this is that the chance an individual will die between 15 and 60 is 26.3%

infant_deaths (Ratio) - number of infant deaths per 1000 population; similar to above, but for infants

alcohol (Ratio) - a country's alcohol consumption rate measured as liters of pure alcohol consumption per capita

percentage_expenditure (Ratio) - expenditure on health as a percentage of Gross Domestic Product (gdp)

hepatitis_b (Ratio) - number of 1 year olds with Hepatitis B immunization over all 1 year olds in population

measles (Ratio) - number of reported Measles cases per 1000 population

bmi (Interval/Ordinal) - average Body Mass Index (BMI) of a country's total population

under-five_deaths (Ratio) - number of people under the age of five deaths per 1000 population

polio (Ratio) - number of 1 year olds with Polio immunization over the number of all 1 year olds in population

total_expenditure (Ratio) - government expenditure on health as a percentage of total government expenditure

diphtheria (Ratio) - Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1 year olds

hiv/aids (Ratio) - deaths per 1000 live births caused by HIV/AIDS for people under 5; number of people under 5 who die due to HIV/AIDS per 1000 births

gdp (Ratio) - Gross Domestic Product per capita

population (Ratio) - population of a country

thinness_1-19_years (Ratio) - rate of thinness among people aged 10-19 (Note: variable should be renamed to thinness_10-19_years to more accurately represent the variable)

thinness_5-9_years (Ratio) - rate of thinness among people aged 5-9

income_composition_of_resources (Ratio) - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

schooling (Ratio) - average number of years of schooling of a population

2.3 Statistical techniques:

Histogram: A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

Pair Plot: Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our data-set.

Heat Map: A heatmap is a graphical representation of data that uses a system of color-coding to represent different values. Heatmaps are used in various forms of analytics but are most commonly used to show user behaviour on specific webpages or webpage templates.

Box Plot: In descriptive statistics, a box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.

Pearson's rank correlation test: The Pearson's rank correlation coefficient is a method of testing the strength and direction (positive or negative) of the correlation (relationship or connection) between two variables.

t test for Independence: The Independent Samples t Test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Independent Samples t Test is a parametric test. This test is also known as: Independent t Test.

Multiple linear regression: Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.

Python libraries:

Pandas: Pandas mainly used for machine learning in from of dataframes. Pandas allows various data manipulation operations such as groupby, join ,merge, data cleaning.

NumPy: NumPy for numerical analysis.

Matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension Numpy. Matplotlib is a data visualization library

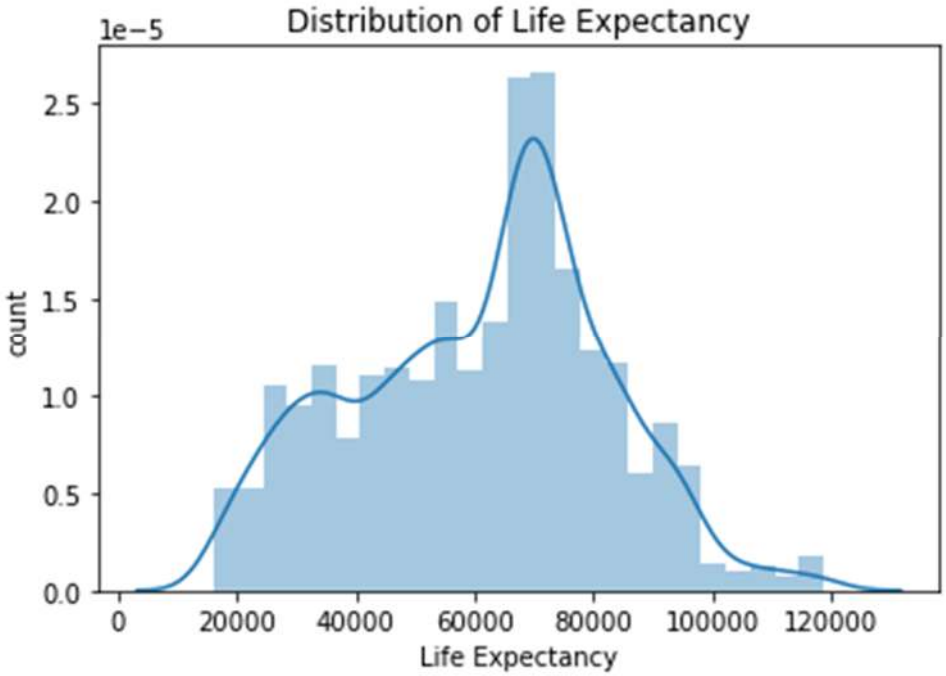
Seaborn: Seaborn is Python data visualization library based on matplotlib.

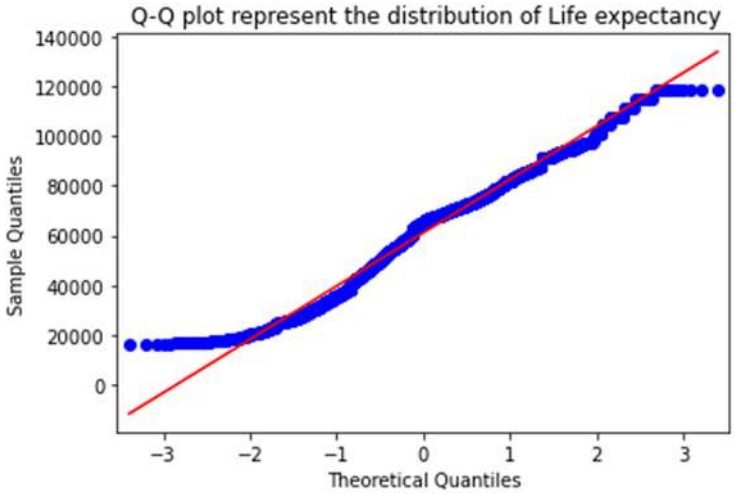
3 Chapter 3

Results and Discussion

3.1 Exploratory Data Analysis:

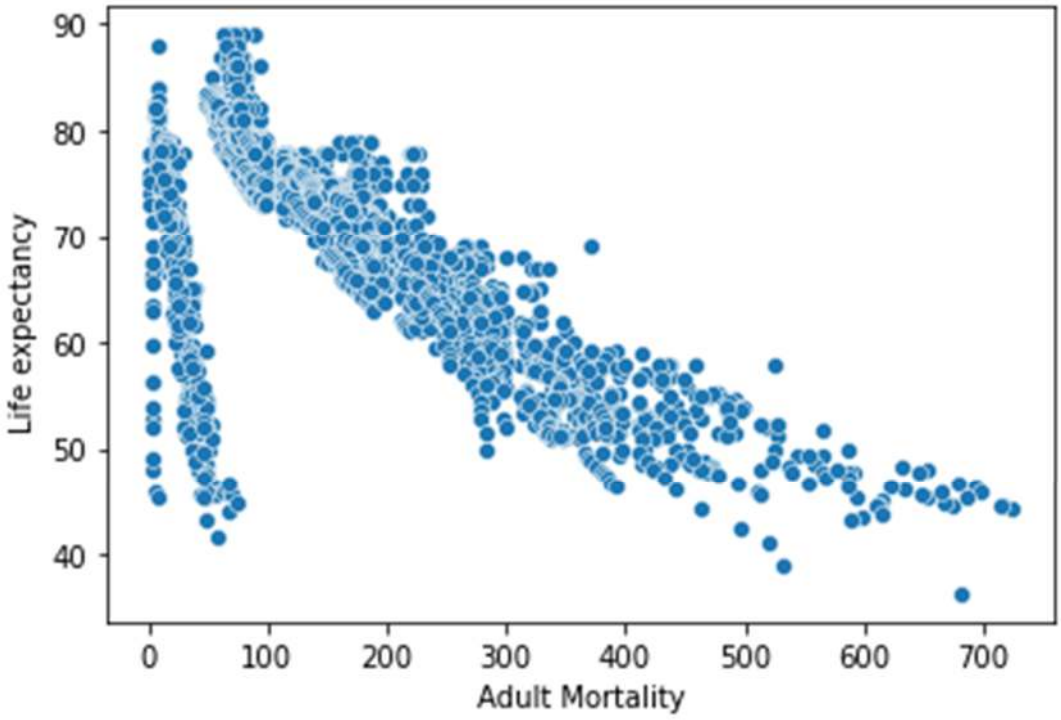
3.1.1 Distribution of the target variable:





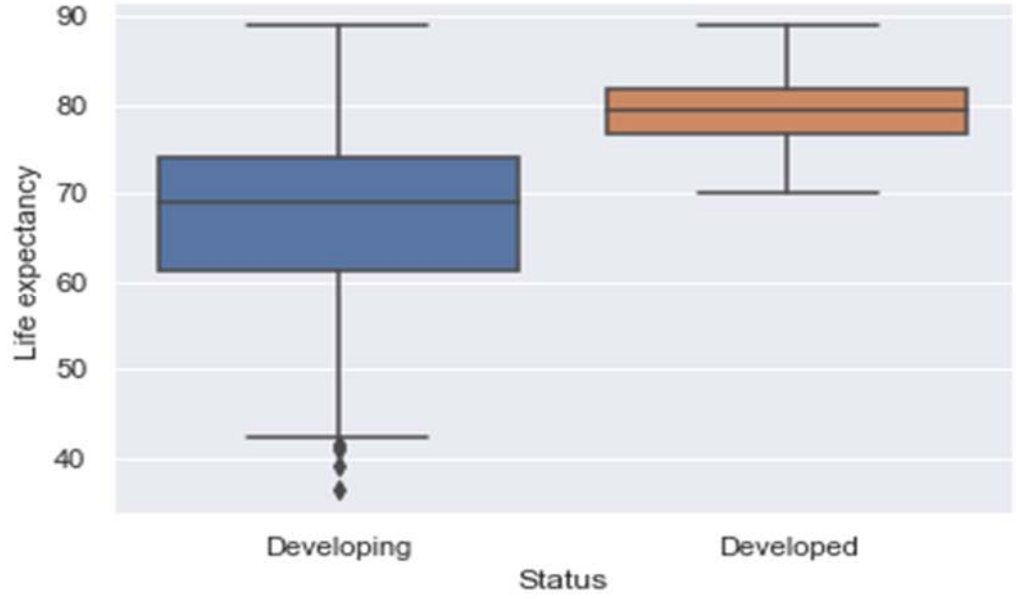
Conclusion: By the distribution plot and Q-Q plot, it is observed that Life Expectancy is normally distributed.

3.1.2 Relationship between Life expectancy and Adult mortality:



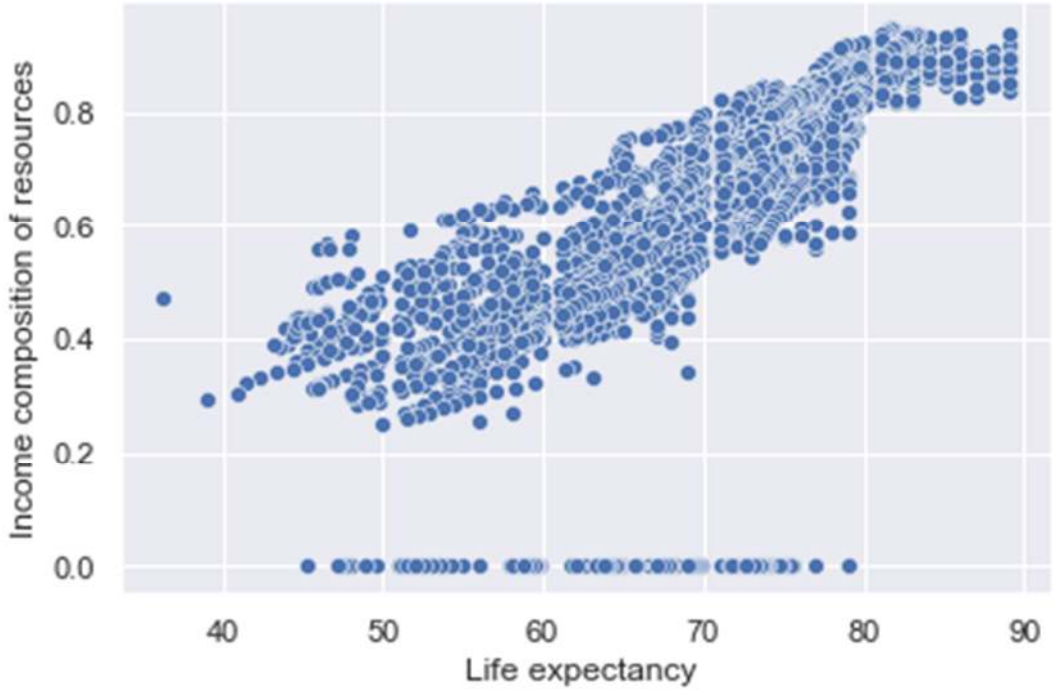
Conclusion: The visualization shows a strong negative correlation between Life Expectancy and Adult Mortality. Which means chance of an individual will die between 15 and 60 is decreases with increase in Life expectancy.

3.1.3 Life Expectancy of developed and developing countries.



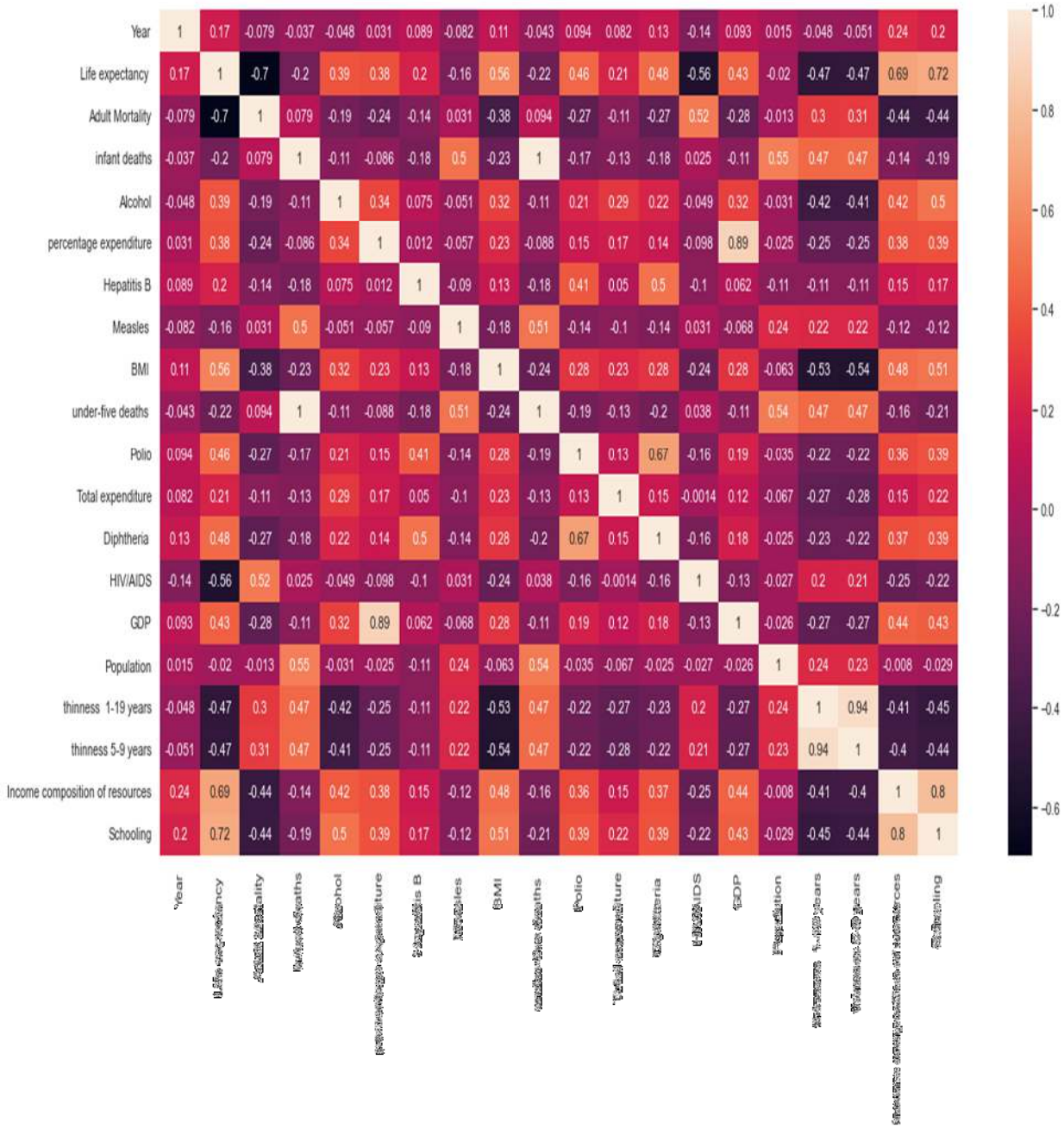
Conclusion: Life expectancy of developed countries is higher than the life expectancy of developing countries.

3.1.4 Relationship between Life expectancy and Income Composition of resources:



Conclusion: Life Expectancy and Income composition of resources are highly positively correlated. - Human Development Index in terms of income composition of resources (index ranging from 0 to 1) is increases with increase in Life expectancy.

3.1.5 Relationship between different variables:



Conclusion: With heat map we find out all the important variable that have high impact in predicting our final output. So we can see that mortality of the adults, alcohol, percentage cost, hepatitis, measles, bmi, under five deaths, polio, total expenditure, diptheria, hiv, population, schooling have a strong impact in predicting our final resultant variable.

3.2.1 Testing the relationship between Life expectancy and Schooling.

Hypothesis

- H_0: Life expectancy and Schooling are uncorrelated.
- H_1: there is a correlation between Life expectancy and Schooling.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

We reject the null hypothesis if p value<0.05

```
r= 0.7150663398620065
p value= 3.1353648e-30
```

Hence we reject the null hypothesis.

Conclusion: There is a correlation between Life expectancy and Schooling.

3.2.2 Testing the significant difference of average life expectancy between developed and developing countries.

Hypothesis

- H_0: There is no significant difference of average life expectancy between developed and developing countries.
- H_1: There is significant difference of average life expectancy between developed and developing countries.

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

We reject the null hypothesis if p value<0.05

```
t= 12.845221420127023
p value= 2.893075763217165e-37
```

Hence we reject the null hypothesis.

Conclusion: There is significant difference of average life expectancy between developed and developing countries.

3.3 Multiple Linear Regression

In Multiple Linear Regression, the target variable(Y) is a linear combination of multiple predictor variables x_1, x_2, x_3, ...,x_n.

$$Y=b_o+b_1x_1+b_2x_2\ldots+b_nx_n$$

Y= Output/Response variable

b_0, b_1, b_2, b_3 , b_n....= Coefficients of the model.

x_1, x_2, x_3, x_4,...= Various Independent/feature variable.

Root mean square error(RMSE):

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. Formally it is defined as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values
$y_1, y_2, \ldots, y_n$ are observed values
n is the number of observations

R2 score:

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively.

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2=coefficient of determination

RSS=sum of squares of residuals

TSS=total sum of squares

3.3.1 Predicting life expectancy of India and its neighbouring countries.

The countries include India, Bangladesh, Bhutan, Afghanistan, China, Pakistan, Myanmar, Nepal,By using multiple linear regression,

$Y=b_o+b_1x_1+b_2x_2\ldots+b_nx_n$

Y=Life expectancy of India and its border sharing countries.

X=All the regressor of India and its border sharing countries.

	Actual	predicted
0	71.8	69.3
1	66.4	64.6
2	66.6	65.4
3	73.9	73.3
4	65.2	62.5
5	66.0	66.6
6	68.2	69.6
7	59.9	61.1
8	65.6	64.9
9	68.0	69.4
10	65.8	66.1
11	66.0	66.5
12	69.8	67.4
13	64.2	63.0
14	57.3	59.7
15	75.0	76.3
16	64.2	63.7
17	66.2	67.0
18	65.3	65.4
19	69.9	68.6
20	69.4	70.8
21	71.7	69.0
22	63.3	63.3
23	58.8	59.8
24	56.2	58.9
25	65.0	60.0
26	62.5	64.8
27	63.9	62.8
28	58.1	59.6
29	63.1	63.9
30	68.4	68.3
31	65.0	65.3
32	65.1	67.0
33	67.5	65.8
34	66.0	68.1
35	64.3	66.0
36	62.8	64.2
37	62.9	61.7
38	62.9	64.7

Root mean square error: 1.7149230770024328
R2 score: 0.8247540507130977

Conclusion: The model accuracy is 82.5%.

3.3.2 Predicting life expectancy of top 10 most populated countries.

The countries include India, China, United States of America, Indonesia, Pakistan, Nigeria, Brazil, Bangladesh, Russian Federation, Mexico
By using multiple linear regression,

$Y=b_o+b_1X_1+b_2X_2\ldots+b_nX_n$

Y=Life expectancy of top 10 most populated countries.

X=All the regressor of top 10 most populated countries.

	Actual	predicted
0	73.3	74.9
1	77.5	77.2
2	67.7	68.0
3	73.6	74.4
4	79.1	78.9
5	71.8	72.0
6	68.6	68.6
7	76.7	74.5
8	76.1	75.9
9	74.1	75.0
10	67.3	66.3
11	74.7	76.5
12	78.2	78.6
13	66.4	68.2
14	68.1	68.8
15	52.0	52.1
16	75.0	72.7
17	66.0	66.8
18	66.8	65.2
19	68.0	68.3
20	79.3	74.5
21	51.6	55.8
22	77.8	77.2
23	68.7	71.3
24	53.2	52.0
25	71.7	68.0
26	66.3	64.8
27	72.7	73.3
28	69.6	71.1
29	75.0	75.7
30	65.0	68.0

31	47.4	46.3
32	54.5	58.8
33	69.1	70.8
34	75.0	73.9
35	71.0	69.3
36	63.2	62.3
37	65.5	65.9
38	74.2	71.8
39	71.4	73.2
40	76.6	73.9
41	75.7	76.4
42	70.0	73.5
43	55.0	50.3
44	69.9	70.4
45	74.8	71.5
46	48.5	54.2
47	77.5	77.1

Root mean square error: `2.165398142244227`
R2 score: `0.9310762589263261`

Conclusion: The model accuracy is 93.1%

4. Chapter 4:

4.1 Conclusion:

The following conclusions are obtained:

From the distribution plot and Q-Q plot, I concluded that the Life Expectancy is normally distributed.

There is a strong negative correlation between Life Expectancy and Adult Mortality and also Life Expectancy and Income composition of resources are highly positively correlated.

Life expectancy of developed countries is higher than the life expectancy of developing countries.

mortality of the adults, alcohol, percentage cost, hepatitis, measles, bmi, under five deaths, polio, total expenditure, diptheria, hiv, population, schooling have a strong impact in predicting our final resultant variable.

There is a correlation between Life expectancy and Schooling.

There is significant difference of average life expectancy between developed and developing countries.

I used multiple linear regression for predicting Life expectancy of India and its neighbouring countries. Its model accuracy is 82.5%

I also used multiple linear regression for predicting Life expectancy of top 10 populated countries. Its model accuracy is 93.1%

5 Chapter 5:

5.1 Summary:

Life expectancy is a statistical measure of the average time a human being is expected to live This problem statement provides a way to predict average life expectancy of people living in a country when various factors such as year, GDP, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related to deaths that happened in the country are given.

The dataset started with 22 variables with Life Expectancy as the target variable and 2938 rows. The first step was to clean the data, this included detecting and dealing with both missing values and outliers. The data contain lot of missing values. I imputed the missing values with the mean of the data. Once the missing values were sorted, the next step was detecting and dealing with outliers. Detection of outliers is done using the boxplots and by using IQR method I eliminate the outliers.

Now the cleaned data is analysed. I performed Exploratory data analysis. I have come to the conclusion that the Life Expectancy is normally distributed, there is a strong negative correlation between Life Expectancy and Adult Mortality and also Life Expectancy and Income composition of resources are highly positively correlated, Life expectancy of developed countries is higher than the life expectancy of developing countries, mortality of the adults, alcohol, percentage cost, hepatitis, measles, bmi, under five deaths, polio, total expenditure, diptheria, hiv, population, schooling have a strong impact in predicting our final resultant variable.

After EDA, I performed statistical tests. Here I have used Pearson's rank correlation test to check the relationship between Life expectancy and Schooling and used t test for independence to check whether there is significant difference of average life expectancy between developed and developing countries. I came to the conclusion that, there is a correlation between Life expectancy and Schooling and there is significant difference of average life expectancy between developed and developing countries.

I used multiple linear regression for predicting Life expectancy of India and its neighbouring countries. Its model accuracy is 82.5%. I also used multiple linear regression for predicting Life expectancy of top 10 populated countries. Its model accuracy is 93.1%.

6 Chapter 6:

6.1 Bibliography

1. https://www.geeksforgeeks.org/python-programming-language/

2.https://towardsdatascience.com/anova-t-test-and-other-statistical-tests-with-python-e7a36a2fdc0c

3. https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/

4. https://towardsdatascience.com/what-really-drives-higher-life-expectancy-e1c1ec22f6e1

5. https://ieeexplore.ieee.org/document/9596123

6. https://www.tutorialspoint.com/python/index.htm

7. https://scikit-learn.org/

7 Chapter 7:

Appendix

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt


data=pd.read_csv(r"D:\SPOO\Documents\Msc\Project\Life Expectancy Data.csv")

data.shape

data.info()

data.isnull().sum()


plt.figure(figsize=(20,8))

ax = sns.distplot(data['Life expectancy '])

plt.title('Count of life expectancy')

plt.xticks(rotation=45)

plt.ylabel('count')

import statsmodels.graphics.gofplots as sm

sm.qqplot(data["Life expectancy "],line="s")


plt.title(" Q-Q plot represent the distribution of Life expectancy")


sns.scatterplot(x='Adult Mortality',y='Life expectancy ',data=data)
```

```
sns.scatterplot(x='Schooling',y='Life expectancy ',data=data)
sns.scatterplot(x='under-five deaths ',y='infant deaths',data=data)
sns.boxplot(x='Status',y='Life expectancy ',data=data)
sns.scatterplot(x='Life expectancy ',y='Income composition of resources',data=data)
cor=data.corr()
cor
from scipy.stats import pearsonr
data1=data['Life expectancy ']
data2=data['Schooling']
stat,p=pearsonr(data1,data2)
print(stat,p)
if p>0.05:
    print("independent")
else:
    print("dependent")

data1=data['Life expectancy '][data['Status']=='Developing']
data2=data['Life expectancy '][data['Status']=='Developed']
from scipy.stats import ttest_ind
stat,p=ttest_ind(data1,data2)
if p>0.05:
    print("independent")
else:
    print("dependent")
```

```
data.drop(columns=["under-five deaths ",'infant deaths'],inplace=True)

for column in data[['Life expectancy ', 'Adult Mortality', 'Alcohol', ' BMI ', 'Polio', 'Total expenditure',

      'Diphtheria ', 'Hepatitis B','GDP', 'Population',

      ' thinness  1-19 years', ' thinness 5-9 years',

      'Income composition of resources', 'Schooling']]:

   data[column]=data[column].fillna(value=data[column].mean())

cat=data.select_dtypes(include="O")

num=data.select_dtypes(exclude="O")

for feature in num.columns:

   sns.boxplot(x=num[feature])

   plt.show()

new=data[['Life expectancy ', 'Adult Mortality',

      'Alcohol', 'percentage expenditure', 'Hepatitis B', 'Measles ', ' BMI ',

       'Polio', 'Total expenditure', 'Diphtheria ',

      ' HIV/AIDS', 'GDP', 'Population', ' thinness  1-19 years',

      ' thinness 5-9 years', 'Income composition of resources', 'Schooling']]

for i in new:

   le=pd.DataFrame(data[i])

   med=le.median()

   q3=le.quantile(q=0.75)

   q1=le.quantile(q=0.25)

   iqr=q3-q1
```

```
    iqrll=int(q1-1.5*iqr)
    iqrul=int(q3+1.5*iqr)
    data.loc[data[i]>iqrul,i]=int(le.quantile(q=0.99))
    data.loc[data[i]<iqrll,i]=int(le.quantile(q=0.01))
    sns.boxplot(x=data[i])
    plt.show()
#india and its neighbouring countries
c1=data[data['Country'] == 'India']
c2=data[data['Country'] == 'Bangladesh']
c3=data[data['Country'] == 'Bhutan']
c4=data[data['Country'] == 'Afghanistan']
c5=data[data['Country'] == 'China']
c6=data[data['Country'] == 'Pakistan']
c7=data[data['Country'] == 'Myanmar']
c8=data[data['Country'] == 'Nepal']
da=pd.concat([c1,c2,c3,c4,c5,c6,c7,c8])
da.shape
x=da
x=da.drop('Life expectancy ',axis=1)
x.drop(['Country','Status'],inplace=True,axis=1)
a=data['Life expectancy '][data['Country']=='India']
b=data['Life expectancy '][data['Country'] == 'Bangladesh']
c=data['Life expectancy '][data['Country'] == 'Bhutan']
d=data['Life expectancy '][data['Country'] == 'Afghanistan']
```

```
e=data['Life expectancy '][data['Country'] == 'China']
f=data['Life expectancy '][data['Country'] == 'Pakistan']
g=data['Life expectancy '][data['Country'] == 'Myanmar']
h=data['Life expectancy '][data['Country'] == 'Nepal']
y=pd.concat([a,b,c,d,e,f,g,h])
y
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.30, random_state=101)
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler().fit(x_train)
x_scaled_train=scaler.transform(x_train)
from sklearn.linear_model import LinearRegression
Linear_model= LinearRegression()
Linear_model.fit(x_train,y_train)
predictions1=Linear_model.predict(x_test)
from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_test,predictions1)**(0.5))
from sklearn.metrics import r2_score
r2_score(y_test,predictions1)
```