UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143 SEC 02

PROBABILITY & STATISTICAL DATA ANALYSIS

-Project 2-

Countries Data

GROUP MEMBERS

NAMES	Matric Number
GOO YE JUI	A20EC0191
LEE MING QI	A20EC0064
KELVIN EE	A20EC0195
LEE JIA XIAN	A20EC0200

LECTURER: ***Mr. Chan Weng Howe***

Table of Content

1.0 INTRODUCTION **3**

2.0 DATA AND ANALYSIS **4**

2.1 HYPOTHESIS TESTING - Two Sample Test 4

2.2 CORRELATION 5

2.3 REGRESSION 6

2.4 CHI-SQUARE TEST OF INDEPENDENCE 8

3.0 DISCUSSION **9**

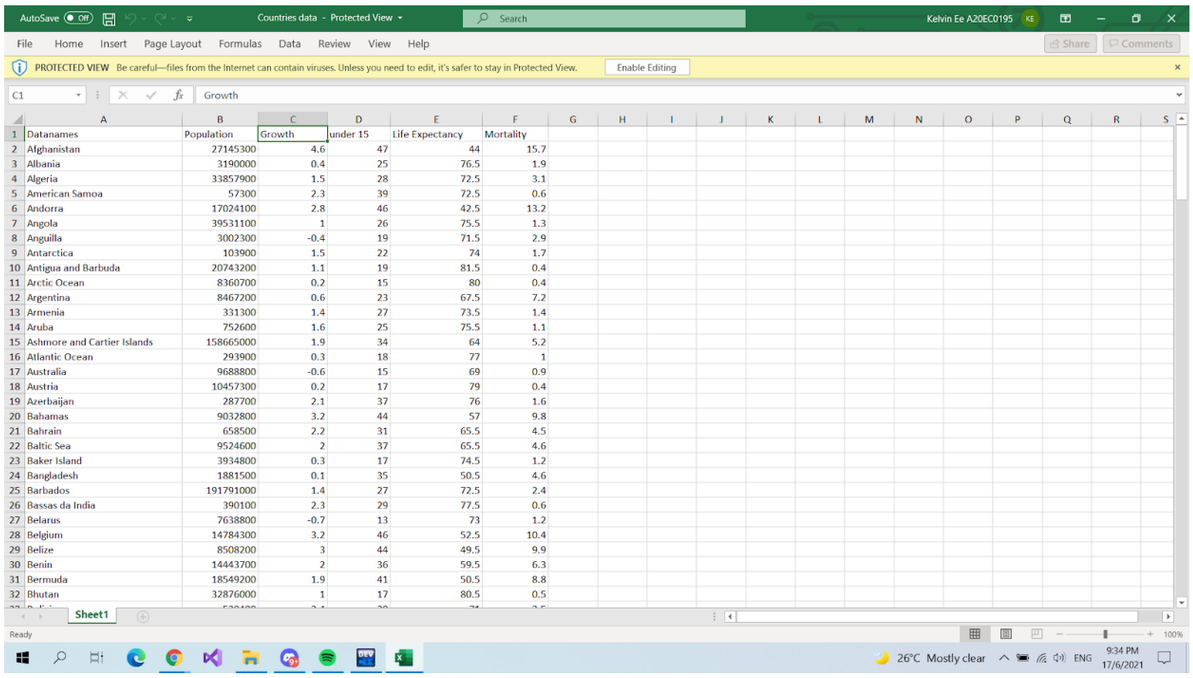
4.0 CONCLUSION **9**

5.0 REFERENCES **10**

6.0 Appendix **11**

1.0 INTRODUCTION

This dataset is provided by Dr. Chan Weng Howe to our group through elearning. The dataset has 198 observations and 6 variables which consists of population, GDP growth, percentage of population under the age of 15, life expectancy and the mortality(per 1000 people) from 198 different countries.



	Datanames	Population	Growth	under 15	Life Expectancy	Mortality
2	Afghanistan	27145300	4.6	47	44	15.7
3	Albania	3190000	0.4	25	76.5	1.9
4	Algeria	33857900	1.5	28	72.5	3.1
5	American Samoa	57300	2.3	39	72.5	0.6
6	Andorra	17024100	2.8	46	42.5	13.2
7	Angola	39531100	1	26	75.5	1.3
8	Anguilla	3002300	-0.4	19	71.5	2.9
9	Antarctica	103900	1.5	22	74	1.7
10	Antigua and Barbuda	20743200	1.1	19	81.5	0.4
11	Arctic Ocean	8360700	0.2	15	80	0.4
12	Argentina	8467200	0.6	23	67.5	7.2
13	Armenia	331300	1.4	27	73.5	1.4
14	Aruba	752600	1.6	25	75.5	1.1
15	Ashmore and Cartier Islands	158665000	1.9	34	64	5.2
16	Atlantic Ocean	293900	0.3	18	77	1
17	Australia	9688800	-0.6	15	69	0.9
18	Austria	10457300	0.2	17	79	0.4
19	Azerbaijan	287700	2.1	37	76	1.6
20	Bahamas	9032800	3.2	44	57	9.8
21	Bahrain	658500	2.2	31	65.5	4.5
22	Baltic Sea	9524600	2	37	65.5	4.6
23	Baker Island	3934800	0.3	17	74.5	1.2
24	Bangladesh	1881500	0.1	35	50.5	4.6
25	Barbados	191791000	1.4	27	72.5	2.4
26	Bassas da India	390100	2.3	29	77.5	0.6
27	Belarus	7638800	-0.7	13	73	1.2
28	Belgium	14784300	3.2	46	52.5	10.4
29	Belize	8508200	3	44	49.5	9.9
30	Benin	14443700	2	36	59.5	6.3
31	Bermuda	18549200	1.9	41	50.5	8.8
32	Bhutan	32876000	1	17	80.5	0.5

Figure 1.0 Countries Data

Early studies had found out that there is a strong relation between economic growth and population health status. Richer countries with better standards of living, health systems and investment in determinants of health tend to have higher average life expectancy (Schultz, T. P. ,2010 ; Jetter, M., Laudage, S., & Stadelmann, D. ,2019 ; Hague, S., Gottschalk, R., & Martins, P. ,2008). However, some of the countries have failed to achieve the life expectancy that their income would predict (Preston, S. H. ,1975). Apart from that, child mortality can be considered as one of the best measures of the health status of a country (Wang, L. ,2002). Besides, the age group of 0-14 represents the fertile age. The larger the fraction of the population who are in the fertile age range represents higher fertility rate, and this will influence the youth dependency ratio of the country. A high youth dependency ratio indicates that a greater investment needs to be made in schooling and other services for children.

From the background studies above, we can know that there are a lot of factors affecting life expectancy including social factors, demographic variables and mortality. We know that the life expectancy reflects the Human Development Index (HDI) of the country. Therefore, we chose to study on this topic and we wish to investigate whether these factors really affect life expectancy. We aimed to conduct an inferential statistical analysis on life expectancy.

2.0 DATA AND ANALYSIS

The aim of this study is to discuss a country from a dataset provided which includes population, GDP growth, percentage of population under the age of 15, life expectancy and the mortality(per 1000 people) from 198 different countries. We will carry out 4 different tests which are Hypothesis testing-two samples test, correlation test, regression test, and chi-square test of independence.

2.1 HYPOTHESIS TESTING - Two Sample Test

A healthy gross domestic product (GDP) growth rate sustains the economy in the expansion phase of the business cycle for as long as possible. The GDP growth rate represents how much more the economy produced compared to the previous quarter. According to Jones(2016), an ideal GDP rate is between 2% to 3% so that a country's economy can safely maintain itself without causing negative side effects.

In this test, we assume the healthy GDP growth rate to be 2%. The purpose of this test is to test whether those countries with healthy GDP growth rate have different mean of life expectancy than those with unhealthy GDP growth rate. These two samples contain 64 healthy GDP growth rate countries and 134 unhealthy GDP growth rate countries. The population variance is unknown and assumed to be unequal in this case. The test is being conducted at a=0.05, where a is the significant level of confidence.

For the significance test, our hypothesis statements were as below :

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

where μ_1 is mean of life expectancy for countries with healthy GDP growth rate and

μ_2 is the mean of life expectancy for countries with unhealthy GDP growth rate.

From the result, the mean of GDP-healthy countries based on life expectancy = 61.40625 while GDP-unhealthy countries mean = 71.57463. The degree of freedom =109.76 floor to 109. The test statistics, t_0 = -6.5574. The p-value of the test statistic is 0.000000001853.At α=0.025, critical value of $t_{0.025,109}$ = -1.981967 and 1.981967.

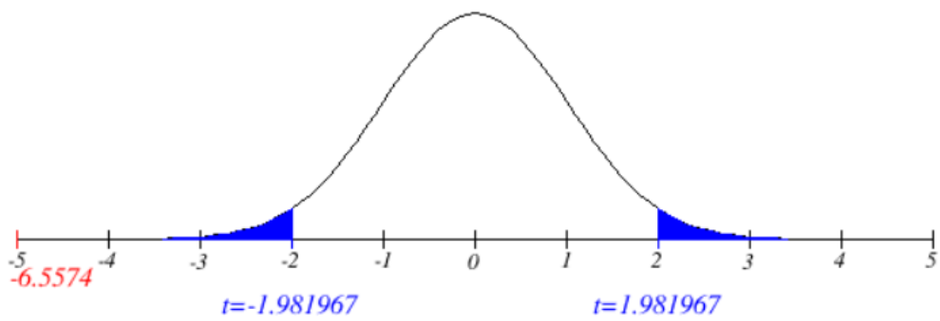


Figure 2.0 Critical Region

Analysis: Since t_0=-6.5574 < -1.981967 and P-value = 0.000000001853<0.025, H_0 is rejected.

Conclusion: There is enough evidence to prove that the GDP-healthy countries have different mean of life expectancy to the GDP-unhealthy countries.

2.2 CORRELATION

In this correlation test, we will investigate the strength of the linear relationship between the percentage of population under the age of 15 and the life expectancy. We calculate the correlation coefficient, r to know the strength of the linear relationship between them. A significance test for correlation is also conducted to show that whether there is enough evidence of a linear relationship between them at the significance level,$\alpha = 0.05$.

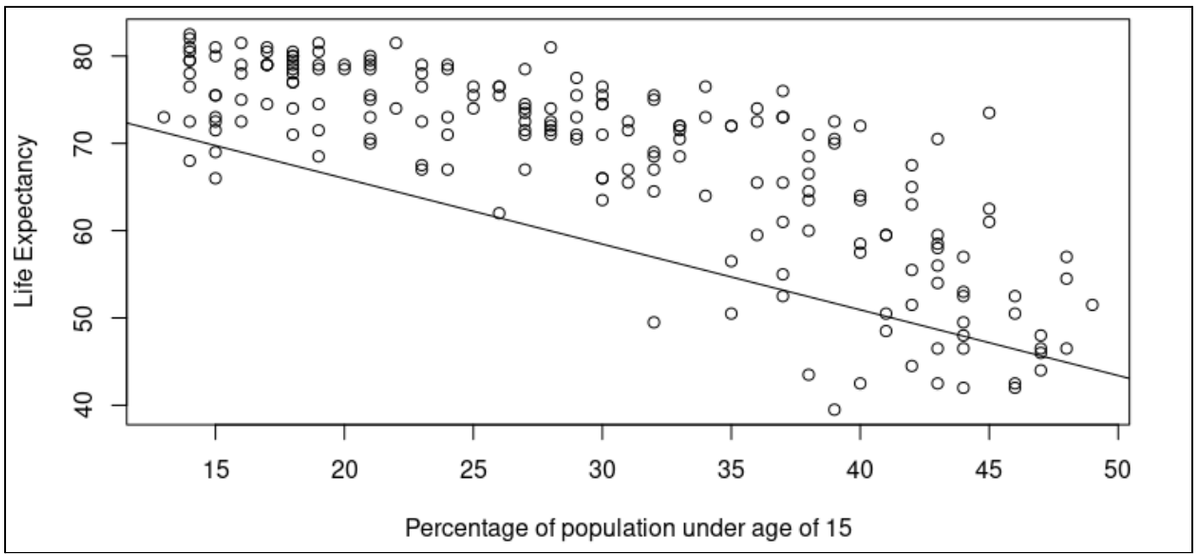


Figure 3.0: Relationship between Life Expectancy & Percentage of population under age 15

From the correlation test conducted, the correlation coefficient, r obtained = -0.7796239 which indicates that there is a moderate negative linear correlation between life expectancy and percentage of population under age of 15.

For the significance test, our hypothesis statements were as below :

H_0: p=0 (No linear correlation)

H_1: p≠0 (Linear correlation exist)

In the significance test for correlation, the test statistics, t_0= -17.429. The degree of freedom, df = 196. Since it is a two-tailed test, at α=0.025, critical value of $t_{0.025,196}$= -1.972141 and 1.972141.

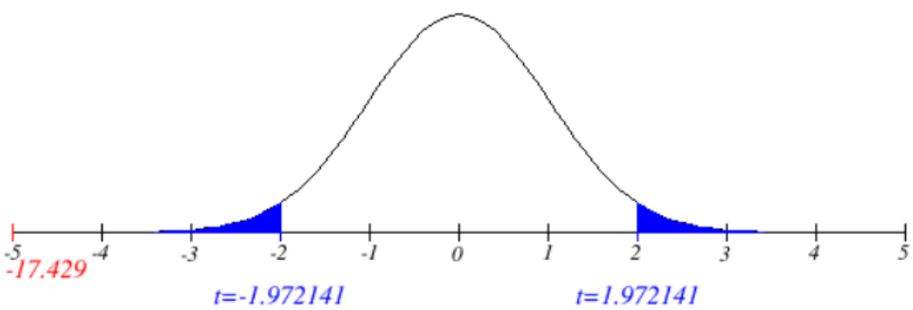


Figure 4.0 Critical Region

Analysis: Hence, we will reject H_0 since t_0 = -17.429 < -1.972141 and p-value = 2.2e-16 < 0.025 ,H_0 is rejected.

Conclusion: There is sufficient evidence of a linear correlation between life expectancy and percentage of population under age of 15.

2.3 REGRESSION

In this regression test, we will build an estimated regression model for mortality rate and life expectancy to test the relationship between them. This regression analysis is conducted to predict the value of life expectancy based on the mortality rate of the country and explain the impact of changes in mortality rate on life expectancy. The independent variable,x is mortality rate and the dependent variable, y is life expectancy.

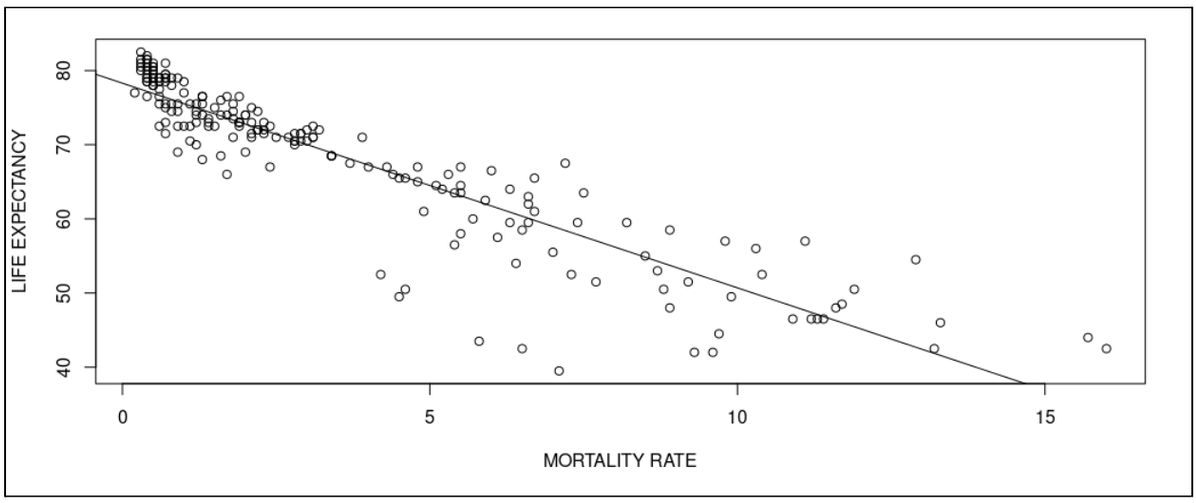


Figure 5.0 : Relationship between mortality rate vs life expectancy

From the linear regression model we built based on mortality rate and life expectancy, we obtained that the estimated regression equation as below :

$\hat{y}$ = 78.29593 - 2.76x

Where,

$\hat{y}$ = Life expectancy

x = Mortality rate

b_0 = 78.29593

b_1 = - 2.76

Based on the regression equation, we conclude that the average value of life expectancy is 78.29593 when the mortality rate is 0. Also, the life expectancy decreases by 2.76 for each additional 1 of mortality rate.

For the coefficient of determination, R^2= 0.8299, where $0<R^2<1$. This indicates that there is a strong linear relationship between mortality rate and life expectancy. From the R^2 value, we can conclude that 82.99% of the variation in life expectancy can be explained by the variation in mortality rate.

To test the regression, at the level of confidence, α=0.05 is being used.

H_0:β_1=0 (No linear relationship between mortality rate and life expectancy)

H_1: $\beta_1 \neq 0$ (Linear relationship exist between mortality rate and life expectancy)

In the significance test for regression model, the test statistics, t_0= -30.92. The degree of freedom, df = 196. Since it is a two-tailed test, at α=0.025, critical value of $t_{0.025,196}$= -1.972141 and 1.972141.

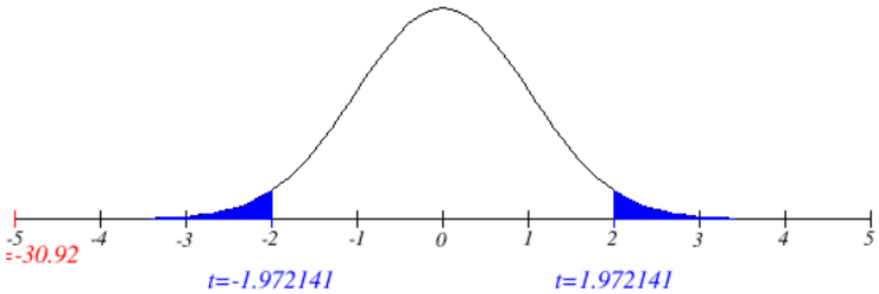


Figure 6.0 : Critical Region

Analysis: Hence, we will reject H_0 since t_0= -30.92 < -1.972141 and p-value = 2.2e-16 < 0.025 ,$H_0$is rejected.

Conclusion: There is sufficient evidence of a strong linear relationship between mortality rate and life expectancy.

2.4 CHI-SQUARE TEST OF INDEPENDENCE

In this chi-square test, we want to find out whether there is a relationship between the two variables that is population size and GDP growth at the significance level of 0.05. The population size is grouped into three different groups which is below 1M, between 1M and 10M and more than 10M while the GDP growth is grouped into two different groups which are healthy(GDP >=2) and unhealthy(GDP<2).

The two-way contingency table that is obtained from the result which includes the observed count and the expected count is shown below:

	GDP growth			
	Healthy		Unhealthy	
Population size	Obs.	Exp.	Obs.	Exp.
Below 1M	15	15.19192	32	31.80808
between 1M and 10M	24	23.59596	49	49.40404
More than 10M	25	25.21212	53	52.78788

Table 1.0 : Frequency Table of Population Size and GDP Growth

The hypothesis statement is as below:

H_0: Population size and GDP growth are independent.

H_1: Population size and GDP growth are dependent.

Based on the chi-square test, we obtained that the test statistic value, $\chi^2 = 0.016442$. The degree of freedom, df = 2. At $\alpha = 0.05$, df = 2, the critical value for the test, χ^2 is 5.991465.

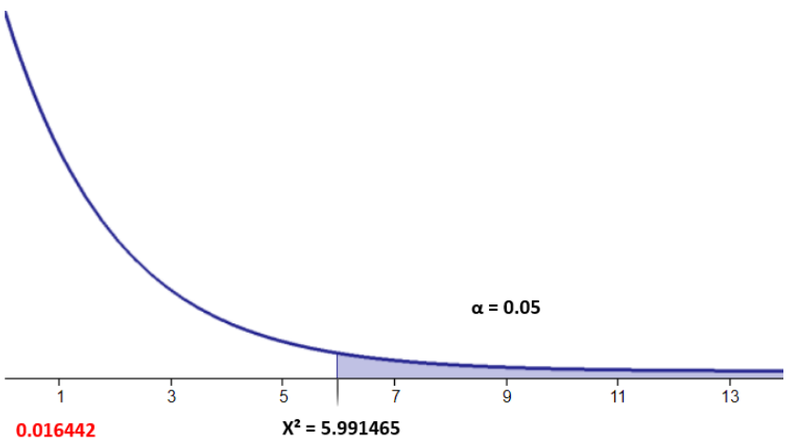


Figure 7.0 : Critical Region

Analysis: Hence, $\chi^2 = 0.016442 < 5.991465$, means the test statistics will not fall within the critical region, thus we fail to reject the null hypothesis, H_0 at significance level, $\alpha = 0.05$.

Conclusion: We conclude that there is insufficient evidence to prove that population size and GDP growth are dependent.

3.0 DISCUSSION

Based on the result of the hypothesis testing of 2 sample mean, we found out that there is sufficient evidence to prove that the GDP-healthy countries have different mean of life expectancy to the GDP-unhealthy countries.We can conclude that countries with healthy GDP growth not necessarily to have a higher life expectancy than unhealthy-GDP-growth countries. It might be affected by other factors as well like the percentage of expenditure on health on the GDP per capita.

Based on the correlation test, a Pearson's product-moment correlation coefficient is used because both variables are ratio data. We found out that there exists a relationship between the two variables, life expectancy and percentage of population under age of 15. They have a moderate negative linear correlation between each other which means that when the percentage of population under the age of 15 decreased, life expectancy will increase and vice versa.

Based on the regression test, we have obtained the estimated regression equation which is $\hat{y} = 78.29593 - 2.76x$. It helps us to predict the value of the dependent variable life expectancy based on the value of mortality rate. We found out that there exists a strong linear relationship between mortality rate and life expectancy. The higher the mortality rate, the lower the life expectancy based on the graph produced.

Based on the Chi-Square test of independence, we know that the population size and GDP growth are independent of each other. This means that a country with a low population can have a healthy GDP growth rate, which is above 2.0. On the other hand, a country with a high population can also have an unhealthy GDP growth rate, which is below 2.0. Hence, we can conclude that there is no relationship between population size and GDP growth rate.

4.0 CONCLUSION

In this project, we have conducted a series of activities like choosing dataset, pre-processing and analysis process. We have learnt how to choose a dataset by choosing a dataset with complete datas and variables. The dataset should have a reasonable mix of both continuous and categorical variables. Besides, during the pre-processing, we filtered the data that was incomplete and deleted it. We have also learnt how to conduct statistical analysis, which are hypothesis testing for 2 samples, correlation test, regression test and Chi Square test of independence. By using these statistical analysis, we are able to prove the hypothesis made in the early stages of the project.

According to these tests in our project, we can make some conclusions:

- GDP growth is not a factor that affects the life expectancy of a country. (Hypothesis Testing)
- The life expectancy will increase, when the percentage of population under the age of 15 and vice versa. (correlation)
- The mortality rate is considered to reflect the differences in life expectancy quite well. The higher the mortality rate, the lower the life expectancy. (Regression)

- There is no relationship between the population size and GDP growth of the countries. (Chi-Square Test)

5.0 REFERENCES

Hague, S., Gottschalk, R., & Martins, P. (2008). Pro-poor growth: the evidence beyond income.

Jetter, M., Laudage, S., & Stadelmann, D. (2019). The intimate link between income levels and life expectancy: global evidence from 213 years. Social Science Quarterly, 100(4), 1387-1403.

Jones, C. (2016). The Facts of Economic Growth. Retrieved 30 June 2021, from https://web.stanford.edu/~chadj/facts.pdf

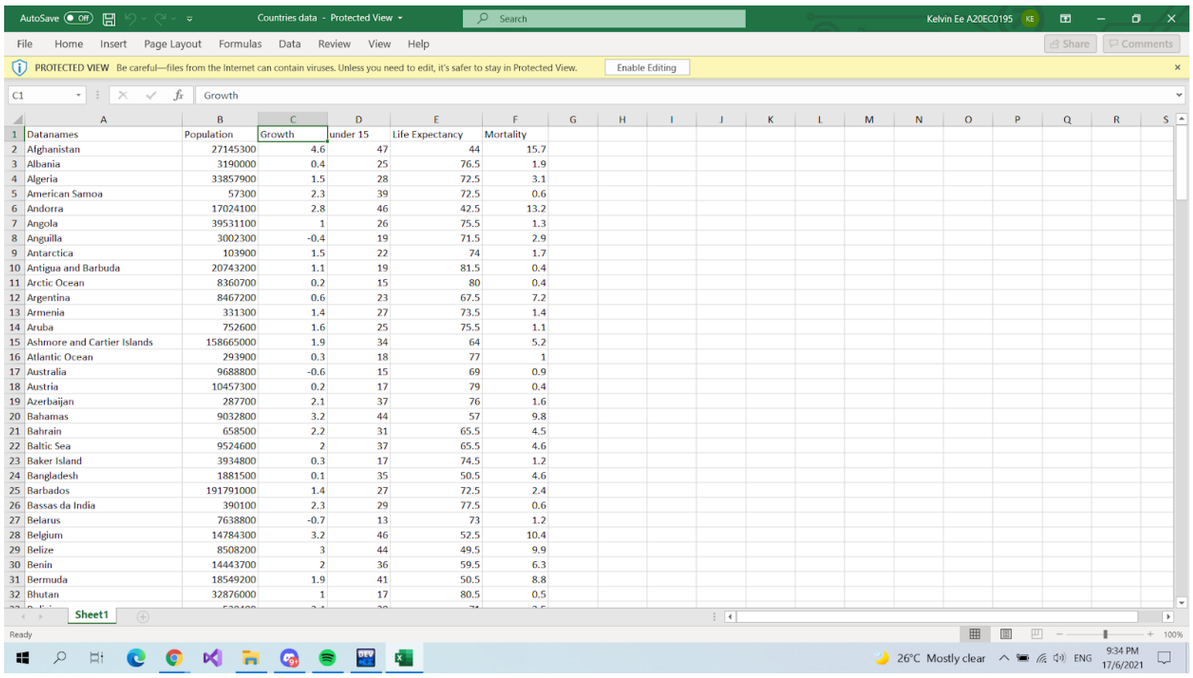
Preston, S. H. (1975). The changing relation between mortality and level of economic development. Population studies, 29(2), 231-248.

Schultz, T. P. (2010). Health human capital and economic development. Journal of African Economies, 19(suppl_3), iii12-iii80.

Wang, L. (2002). Health outcomes in low-income countries and policy implications: Empirical findings from demographic and health surveys (Vol. 2831). World Bank, Environment Department.

World Health Organization. (2019). "Global Health Estimates: Life expectancy and leading causes of death and disability" Retrieved from https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates

6.0 Appendix



	A	B	C	D	E	F
1	Datanames	Population	Growth	under 15	Life Expectancy	Mortality
2	Afghanistan	27145300	4.6	47	44	15.7
3	Albania	3190000	0.4	25	76.5	1.9
4	Algeria	33857900	1.5	28	72.5	3.1
5	American Samoa	57300	2.3	39	72.5	0.6
6	Andorra	17024100	2.8	46	42.5	13.2
7	Angola	39531100	1	26	75.5	1.3
8	Anguilla	3002300	-0.4	19	71.5	2.9
9	Antarctica	103900	1.5	22	74	1.7
10	Antigua and Barbuda	20743200	1.1	19	81.5	0.4
11	Arctic Ocean	8360700	0.2	15	80	0.4
12	Argentina	8467200	0.6	23	67.5	7.2
13	Armenia	331300	1.4	27	73.5	1.4
14	Aruba	752600	1.6	25	75.5	1.1
15	Ashmore and Cartier Islands	158665000	1.9	34	64	5.2
16	Atlantic Ocean	293900	0.3	18	77	1
17	Australia	9688800	-0.6	15	69	0.9
18	Austria	10457300	0.2	17	79	0.4
19	Azerbaijan	287700	2.1	37	76	1.6
20	Bahamas	9032800	3.2	44	57	9.8
21	Bahrain	658500	2.2	31	65.5	4.5
22	Baltic Sea	9524600	2	37	65.5	4.6
23	Baker Island	3934800	0.3	17	74.5	1.2
24	Bangladesh	1881500	0.1	35	50.5	4.6
25	Barbados	191791000	1.4	27	72.5	2.4
26	Bassas da India	390100	2.3	29	77.5	0.6
27	Belarus	7638800	-0.7	13	73	1.2
28	Belgium	14784300	3.2	46	52.5	10.4
29	Belize	8508200	3	44	49.5	9.9
30	Benin	14443700	2	36	59.5	6.3
31	Bermuda	18549200	1.9	41	50.5	8.8
32	Bhutan	32876000	1	17	80.5	0.5

Sample of original data set