A Data Analytics Project on Factors Affecting Life Expectancy in Various Countries from 2000-2015

Gunnika Singh Sobti (100328439)
Henrique de Almeida Pires Pontes (100328913)
Karan Patel (100330009)
Venkata Sai Shouri Gupta Thallam (100331046)

4/4/19

DANA-4800-001 - Data Analysis and Stat Inference

I. OBJECTIVE & INTRODUCTION

<u>Objective</u>: To study the factors that are affecting the Life Expectancy in the countries all over the world from 2005 – 2015.

<u>Introduction</u>:

"HEALTH IS WEALTH" – In search, of this we wanted to find out how wealthier is each country in terms of their people's health. We started finding a data sets of all the countries in the world from the year 2005 – 2015 which contains many factors that effect the people's health in every country. Few examples are the nation status (developed, developing, undeveloped), GDP of that nation, BMI, Total Expenditure of the government on their people's health etc. We have a data set which contains of all these factors which effect the life expectancy in each country from 2005 – 2015.

You might be thinking what will be result of studying this and how is it helpful. It is very important to know the life expectancy of every country for their importance and the factors affecting the life expectancy because everyone in this world wants to live longer, healthier and happier in this point of view the countries can correct those factors which are mostly affecting the life expectancy. It not only helps the countries but also the United nations to make the life expectancy better all over the world.

This will help every nation to know about their present status in health wise comparing to other countries and take necessary actions to achieve more life expectancy than global average by helping each other which will create a great co-operation between all the countries among the world to make the peoples life healthier and better.

There will be many factors which will be affecting the life expectancy, but the main aim of this analysis is to find the factors which are affecting the most.

II. SAMPLING DESIGN

The Sampling Design we have considered for the data is Stratified Random Sampling. We have considered this because of the consideration of grouping together of “n” which is 150. n is the number of countries in the world. We have divided the groups on two steps one is country i.e. each country is a group, in which it has the data of 2005 – 2015. The stratification is mainly done on the nations status like developing/ developed/ under developed. In each status countries will be grouped and under each country we have the data of 2005 – 2015.

We can’t consider the Simple Random Sample as it we needed to select the subset of the population. In this case as most of the data we have is genuine we can’t consider the subset as it is year wise and year wise the values might change and even the status will also the countries numbers. So, in this case considering Simple Radom Sample won’t result in the optimised and accurate analysis.

We can’t consider Cluster Sampling too, even though we use the grouping logic but tge disadvantage is that the variation between clusters is great relative to the variation within clusters, cluster sampling can result in inaccurate estimates.

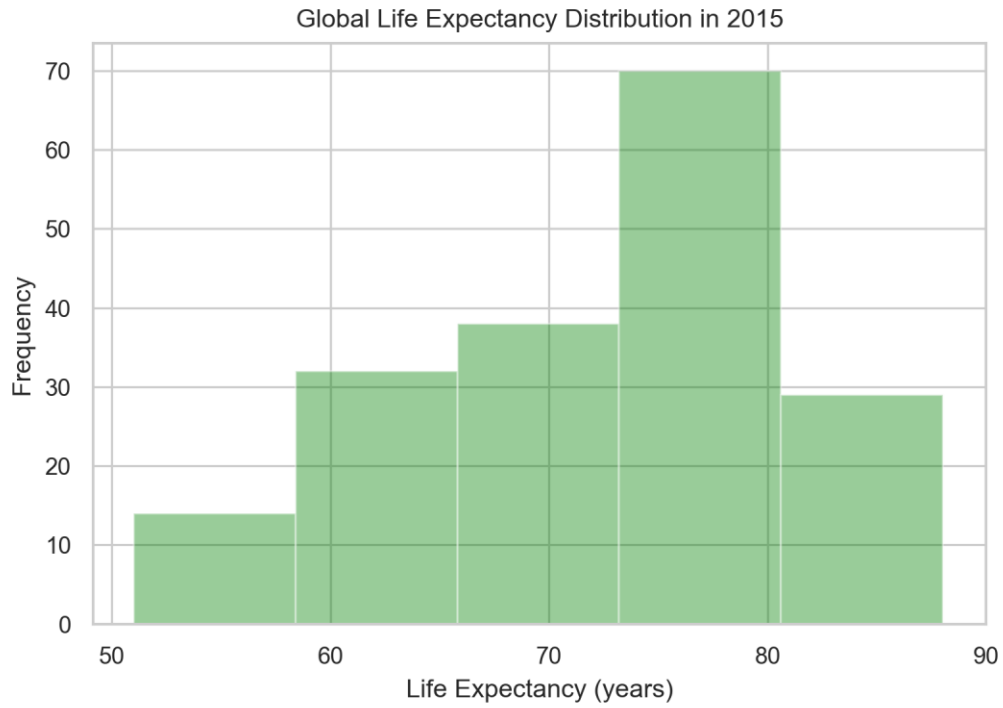
BIAS:

The bias we have is the non – response bias. As we can see in the data set that most of the parts in the data set are incomplete, which makes the analysis inappropriate. So, to get out of this inappropriate analysis we don’t forecast/ predict the values where it’ll lead to the non – realistic analysis. So, we just clean the data so that the non – response bias is taken out of the data set.

III. ANALYSIS & DISCUSSION

1. UNIVARIATE ANALYSIS

a) Life expectancy (Numerical variable)



The histogram above shows the life expectancy distribution from the studied countries in 2015. It is important to highlight that most countries had a life expectancy between 70 and 80 years old in 2015.

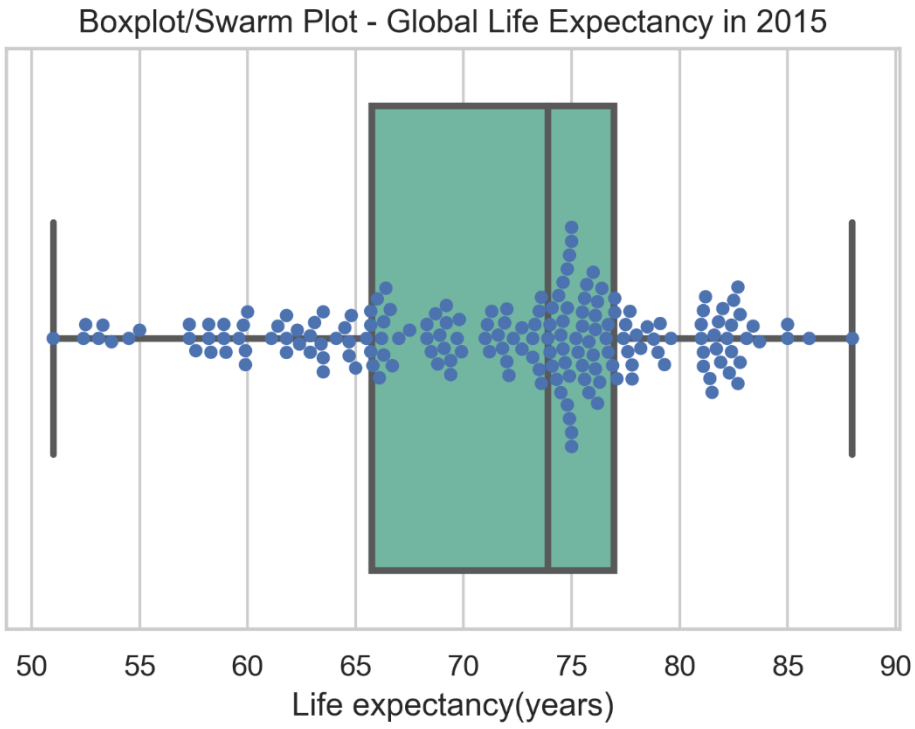
The five-number below shows the lowest global life expectancy at 51 years and the highest at 88 years. The median global life expectancy in 2015 was 73.9 years. As of 2015, the country with the lowest life expectancy was Sierra Leone while the highest was Slovenia.

```
     Five number summary
Global Life expectancy in 2015
Min: 51.000
Q1: 65.750
Median: 73.900
Q3: 76.950
Max: 88.000
```

Interval Estimate of μ at 95% Confidence Interval for Life Expectancy in 2015

(70.3168,72.91700)

The graph bellow presents a boxplot of the global life expectancy distribution in 2015 with a swarm plot displaying the data points. There is no presence of outliers, also it is possible to see the concentration of countries around the median of 73.9 years. Sierra Leone and Slovenia are isolated in the lower and higher end of the boxplot but still are not considered to be an outlier. Based on the Inter Quartile Range (IQR) rule for outliers, the lower limit is 48.95 and the upper limit is 93.75.

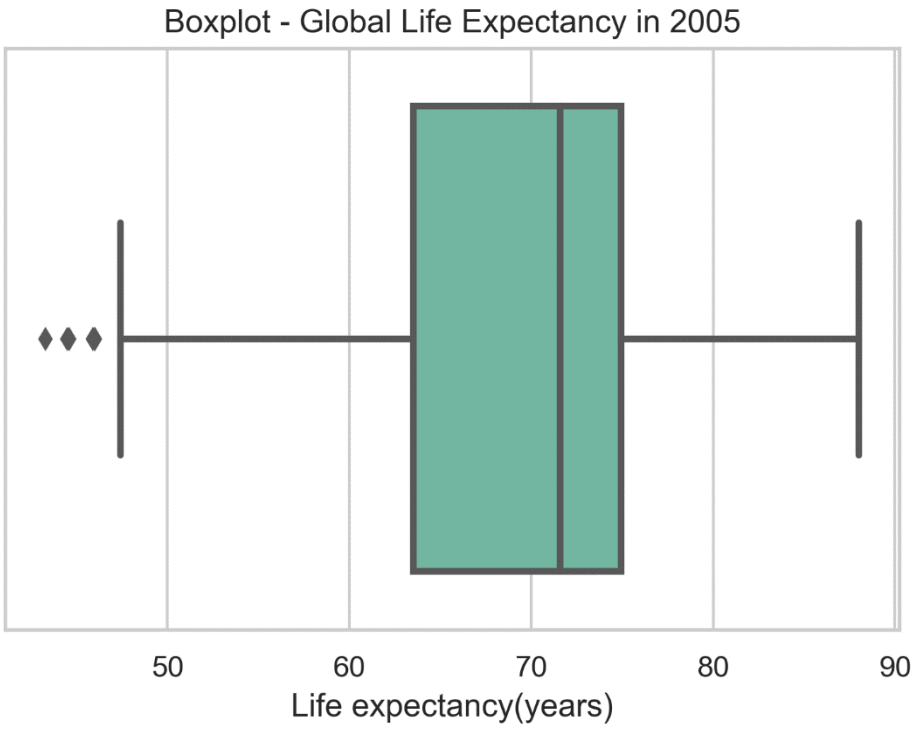


In comparison, the year of 2005 presented 6 outliers in the lower limit. Based on the IQR rule for outliers, the lower limit is 46.325 and these countries were below that.

```
Life Expectancy Outliers in 2005

country
Sierra Leone                43.3
Lesotho                     44.5
Zimbabwe                    44.6
Central African Republic    45.9
Malawi                      46.0
Swaziland                   46.0
Name: life_expectancy, dtype: float64
```

The boxplot bellow displays the distributional characteristics of the life expectancy data in 2005, highlighting the outliers.



As presented in the five-number summary, the median age of life expectancy in 2005 was 71.60 (2.3 years lower than 2015). The minimum value was 43.3 (7.7 years lower than in 2015) and the maximum of 88 (same as 2015).

```
     Five number summary
Global Life expectancy in 2005
Min: 43.300
Q1: 63.500
Median: 71.600
Q3: 74.950
Max: 88.000
```

Interval Estimate of μ at 95% Confidence Interval for life expectancy is 2015

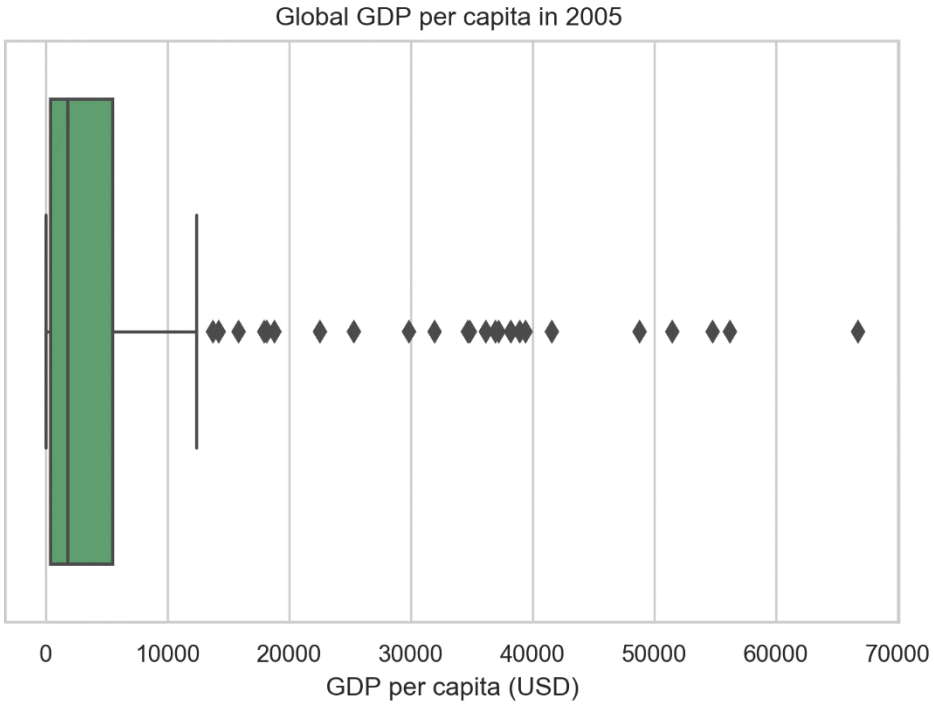
(67.2409,70.3774)

b) Gross Domestic Product (GDP) per capita in USD (Numerical Variable)

The five-number summary for the GDP per capita of the studied countries in 2005 presents a minimum value of 15.57 dollars and a maximum value of 66775.394 dollars. The median was US$ 1791.07.

```
     Five number summary
Global Gross Domestic Product per capita in 2005
Min: 15.570
Q1: 371.420
Median: 1791.070
Q3: 5459.431
Max: 66775.394
```

Based on the IQR criteria to find outliers. The lower limit is -7260.595 and the upper limit of 13091.446. Given that, it does not exist a negative GDP per capita, we do not have any lower outliers but there are 24 outliers above the lower limit as it is possible to see in the boxplot.



The boxplot shows that 75% of the studied countries GDP per capita were less than US$ 5459.431(Q3) in 2005 and half of them were lower than US$ 1791.07(Median).

Confidence Interval of μ at 95% Confidence Interval of GDP in 2005

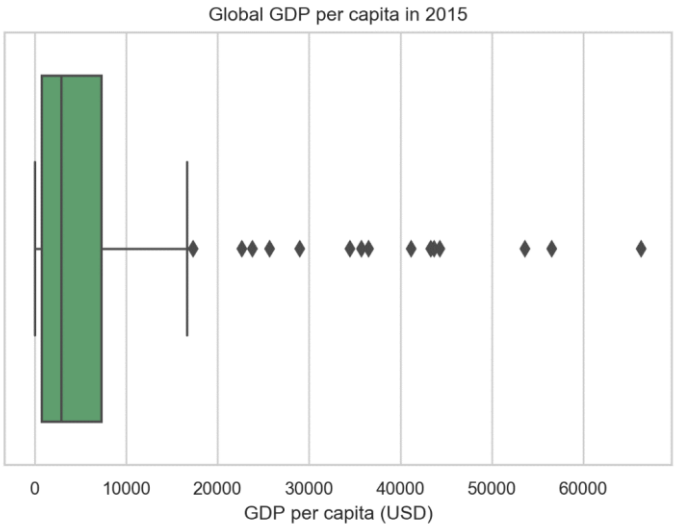
(5152.194,9347.3801)

```
GDP per capita outliers in 2005

country
Norway                  66775.39440
Iceland                 56249.75550
Switzerland             54797.54663
Qatar                   51488.49529
Denmark                 48799.82370
Netherlands             41577.16900
United Arab Emirates    39439.81970
Finland                 38969.17163
Austria                 38242.42520
Japan                   37217.64873
Belgium                 36967.28292
Canada                  36189.58838
France                  34879.72633
Germany                 34696.62920
Italy                   31959.26215
Singapore               29869.85398
Cyprus                  25324.48666
Greece                  22551.73574
Portugal                18784.94850
Slovenia                18169.18910
Bahrain                 17959.17854
Malta                   15835.34667
Barbados                14223.86576
Saudi Arabia            13739.82945
Name: gdp, dtype: float64
```

The five-number summary for the GDP per capita of the studied countries in 2015 presents a minimum value of 33.68 dollars (US$ 18.11 increase from 2005) and a maximum value of 66346.523 dollars (US$ -428.87 lower than 2005). The median was US$ 2916.23 (US$ 1125.16 increase from 2005).

```
     Five number summary
Gross Domestic Product in 2015
Min: 33.680
Q1: 766.010
Median: 2916.230
Q3: 7290.107
Max: 66346.523
```



The boxplot above shows the presence of 15 outliers. Based on the IQR criteria to spot outliers, countries that had a Global GDP per capita higher than US$ 17076.25 in 2015 are considered outliers. Qatar had the highest GDP per capita in 2015 (US$ 66346.52) while Burundi had the lowest (US$ 33.68). 75% of the studied countries in 2015, had a GDP per capita lower than US$ 7290.11.

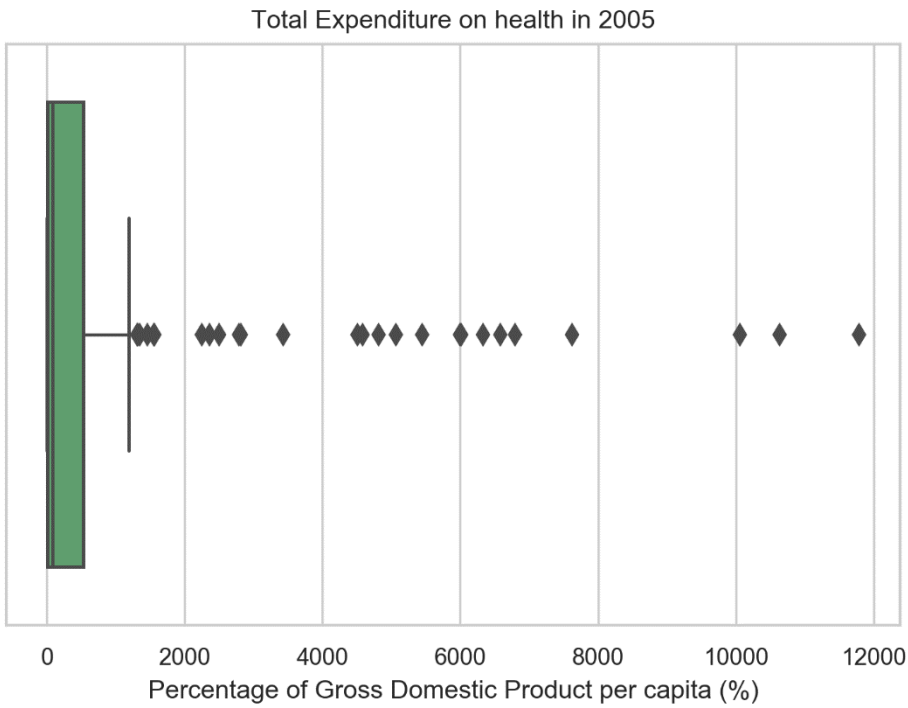
Confidence Interval of u at 95% Confidence Interval for GDP at 2015

(513.324,1133.3118)

c) Total Expenditure on health as a percentage of Gross Domestic Product per capita (%) (Numerical variable)

The five-number summary shows that the minimum spend on health was 0% of the GDP per capita and the maximum was 11792.535%. The country that most spend on health as a percentage of their GDP per capita was Norway and a total of 27 countries have spent the least in 2005.

```
     Five number summary
Total Expenditure on health
percentage of Gross Domestic Product per capita (%) in 2005
Min: 0.000
Q1: 9.190
Median: 79.420
Q3: 529.644
Max: 11792.535
```



The boxplot above shows the presence of 26 outliers. Based on the IQR criteria, countries that have spend more than 1310.32% of their GDP per capita in 2005 on health, are considered outliers.

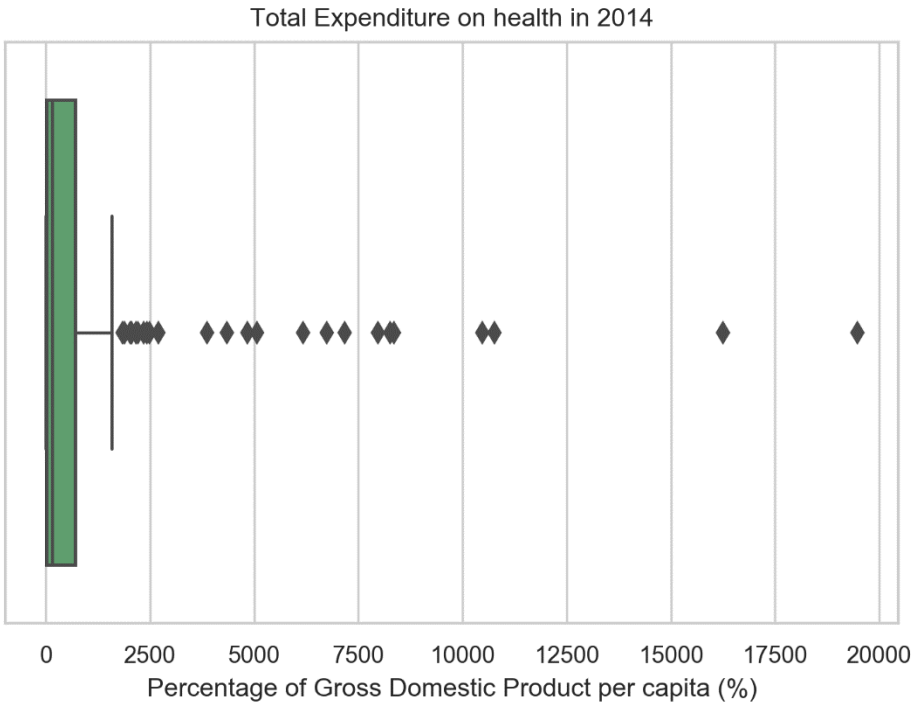
The five-number summary bellow for the 2014 data, still shows a minimum of 0% and maximum of 19479.91% of total expenditure on health as a percentage of GDP per capita (%). The maximum increased 7687.375 % from 2005 to 2014. Considering the median value, now 50% of the countries spend more than 151.10%, that was a 71.68 % increase from 2005. As of 2014, 25% of the countries spend less than 11.06%, a 1.87% increase from 2005.

```
     Five number summary
Total Expenditure on health
percentage of Gross Domestic Product per capita (%) in 2014
Min: 0.000
Q1: 11.060
Median: 151.100
Q3: 703.208
Max: 19479.912
```

The boxplot bellow highlights the presence of 24 outliers. Based on the IQR criteria, a country that has spend more than 1741.43% of their GDP per capita on health in 2014 are considered an outlier.

Interval Estimate of μ at 95% Confidence Interval for Total Expenditure on Health in 2005

(513.324,1133.3118)



Switzerland spent the most on health compared to their GDP per capita in 2014 and 29 countries have spent the least.

Confidence Interval of u at 95% Confidence Interval for Total Expenditure on health in 2014

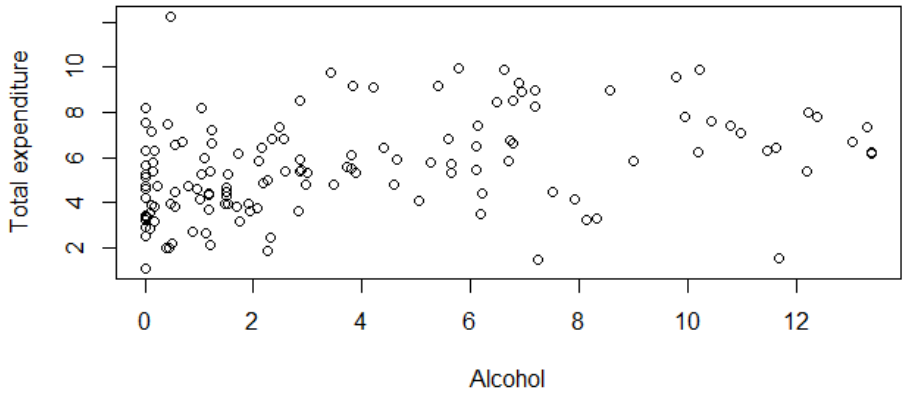
(593.300,1410.52)

2. BIVARIATE ANALYSIS

Bivariate analysis is the simultaneous analysis of 2 variables (attributes). It explores the concept of relationship between 2 variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.

1) Total Expenditure & Alcohol

Scatterplot: It shows low positive correlation



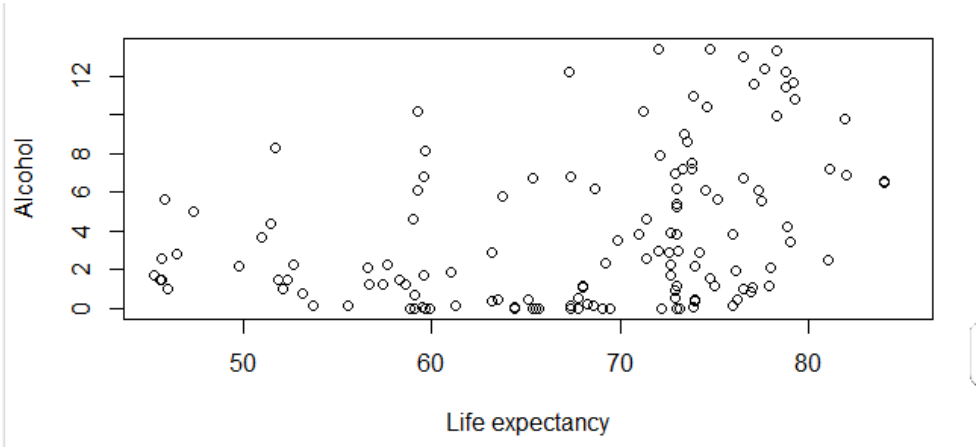
Correlation Coefficient: **0.3720862**

Conclusion: We can say that as Alcohol consumption is higher in countries where total expenditure is large but we cannot say for sure that increase in total expenditure is due to excessive alcohol consumption.

Hence, there is a low positive correlation between Total Expenditure and Alcohol.

2) Alcohol & Life Expectancy

Scatterplot: It shows low negative correlation

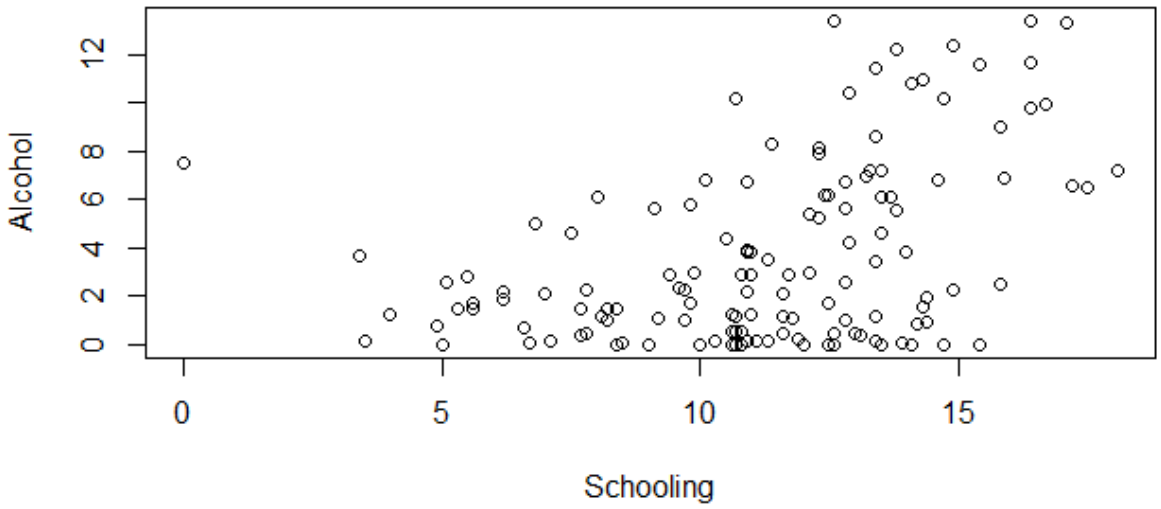


Correlation Coefficient: **0.3498**

Conclusion: We can observe that countries which have higher life expectancy tend to have a higher alcohol consumption rate but we cannot say for sure it is a cause and effect relationship.
Hence, there is a low negative correlation between Alcohol and Life Expectancy.

3) Schooling & Alcohol

Scatterplot: It shows a low positive relationship.



Correlation Coefficient: **0.424593**

Conclusion: We can observe from the graph that countries which have greater number of years of school have higher alcohol consumption rate but we cannot say for sure if number of years taken to complete school is higher because of large alcohol consumption

Hence, there is a low positive correlation between Schooling and Alcohol

Categorical & Categorical

1) Year & Status

H_0: The variable year and the variable Status are independent to each other in the population.

H_1: The variable year and the variable Status are not independent to each other in the population.

```
        Pearson's Chi-squared test

data:  life$Year and life$Status
X-squared = 0.097282, df = 15, p-value = 1

Warning message:
In chisq.test(life$Year, life$Status) :
  Chi-squared approximation may be incorrect
>
```

Since the p value is greater than 0.05 we fail to reject H_0, i.e the variable year and the variable status are independent to each other in the population.

IV. LIMITATIONS

As the data is collected from the open source platform, we can't predict the genuinely of the data.

The data is not collected totally i.e few countries have the missing data in few years. There is no continuity of the data.

The missing values can not be forecasted as it might the result in outliers and inappropriate output.

The data is limited and takes for the years 2005 - 2015 and can not be forecasted to the current year.

V. CONCLUSION

UNIVARIATE ANALYSIS:

1) It is important to highlight that most countries had a life expectancy between 70 and 80 years old in 2015. The median age of life expectancy in 2005 was 71.60 (2.3 years lower than 2015). The minimum value was 43.3 (7.7 years lower than in 2015) and the maximum of 88 (same as 2015).

2) Qatar had the highest GDP per capita in 2015 (US$ 66346.52) while Burundi had the lowest (US$ 33.68). 75% of the studied countries in 2015, had a GDP per capita lower than US$ 7290.11. In 2005 75% of the studied countries GDP per capita were less than US$ 5459.431(Q3) and half of them were lower than US$ 1791.07(Median).

3) Switzerland spent the most on health compared to their GDP per capita in 2014 and 29 countries have spent the least. Norway is the country that most spend on health as a percentage of their GDP per capita.

BIVARIATE ANALYSIS:

1) If the alcohol consumption increases the total expenditure of people in country can be seen higher than normal.

2) As the life expectancy increases for people in different country the alcohol consumption rate also increases.

3) Increase in alcohol consumption can be seen when the student's age increases.

VI. APPENDIX

DATASET:

We have considered the readily available data set of Life Expectancy (WHO)

Statistical Analysis on factors influencing Life Expectancy from KAGGLE and the link is:

https://www.kaggle.com/kumarajarshi/life-expectancy-who/version/1