PROJECT REPORT – GROUP 2

Impact of GDP and Government Health Expenditure on Life Expectancy: An analysis of WHO Data

DEPARTMENT OF HEALTH INFORMATICS

IU LUDDY SCHOOL OF INFORMATICS, COMPUTING & ENGINEERING

IUPUI

Dr. SAPTARSHI PURKAYASTHA

May 04, 2023

TEAM MEMBERS:

MANYA PHILIPS PILLI

RAKESH BATHINI

SANTOSH BOKKA

DIVYA GOTTIMUKKULA

PRUDHVI RAJ TERLI

Introduction

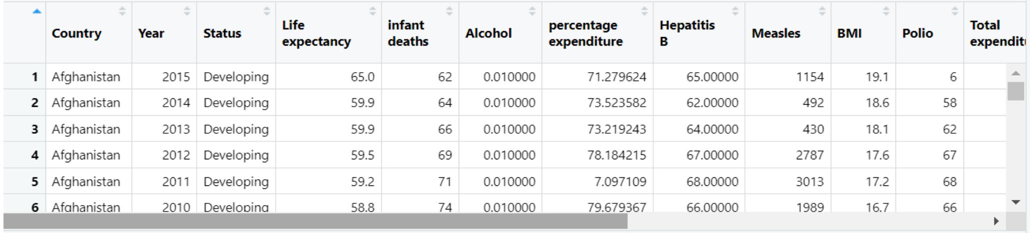
The study by Zaman et al. (2017) suggests that there is a significant relationship between GDP, total government expenditure on health, and life expectancy. This relationship can provide important insights for policymakers and healthcare experts to develop effective strategies for enhancing public health and increasing life expectancy in different countries. The relationship between these variables may vary across different countries due to variations in the level of economic development, government policies, and healthcare systems. Therefore, it is important to conduct statistical analysis using correlation and regression analysis to investigate the relationship between these variables and understand the nature of the relationship.

Problem Statement

The aim of this study is to examine the relationship between GDP, total government expenditure on health, and life expectancy in different countries. This investigation will provide crucial insights for policymakers and healthcare experts, enabling them to develop effective strategies that enhance public health and increase life expectancy. Given the variations in economic development, government policies, and healthcare systems across countries, it is essential to conduct a comprehensive statistical analysis using correlation and regression analysis methods to understand the nature of the relationship between these variables.

Dataset

The dataset contains health related factors that impact the life expectancy of various countries from the years 2000-2015. It has 2938 rows and 22 variables.



	Country	Year	Status	Life expectancy	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	Polio	Total expendit
1	Afghanistan	2015	Developing	65.0	62	0.010000	71.279624	65.00000	1154	19.1	6	
2	Afghanistan	2014	Developing	59.9	64	0.010000	73.523582	62.00000	492	18.6	58	
3	Afghanistan	2013	Developing	59.9	66	0.010000	73.219243	64.00000	430	18.1	62	
4	Afghanistan	2012	Developing	59.5	69	0.010000	78.184215	67.00000	2787	17.6	67	
5	Afghanistan	2011	Developing	59.2	71	0.010000	7.097109	68.00000	3013	17.2	68	
6	Afghanistan	2010	Developing	58.8	74	0.010000	79.679367	66.00000	1989	16.7	66	

1. Country: Categorical variable representing the name of the country.
2. Year: integer variable representing the year of observation (ranging from 2000 to 2015).
3. Status: categorical variable representing whether the country is developed or developing.
4. Life expectancy: continuous variable representing the average number of years a newborn is expected to live, measured in years and the distribution is approximately normal, with a slight positive skewness.

- Summary: min=36.3, max=8.90, mean=69.23, std=9.52.

5. Adult Mortality: continuous variable representing the probability of dying between 15 and 60 years per 1000 population.

- Summary: min=1, max=723, mean=164.8, std=124.1.
- Distribution: the distribution is highly skewed to the right.

6. Infant deaths: integer variable representing the number of infant deaths per 1000 population.

- Summary: min=0, max=180, mean=30.3, std=117.9.
- Distribution: the distribution is highly skewed to the right.

7. Alcohol: continuous variable representing the alcohol consumption rate, measured in liters per capita.

- Summary: min=0.01, max=17.87, mean=4.6, std=3.94.
- Distribution: the distribution is highly skewed to the right.

8. Percentage expenditure: continuous variable representing the expenditure on health as a percentage of GDP.

- Summary: min=0.0, max=20,000.0, mean=5.94, std=17.52.
- Distribution: the distribution is highly skewed to the right.

9. Hepatitis B: Continuous variable representing immunization coverage among 1-year-olds against Hepatitis B.

- Summary: min=1.0, max=99.0, mean=80.94, std=25.07.
- Distribution: the distribution is slightly skewed to the left.

10. Measles: integer variable representing the number of reported cases of measles per 1000 population.

- Summary: min=0, max=212183, mean=2419.59, std=11467.27.
- Distribution: the distribution is highly skewed to the right.

11. BMI: continuous variable representing the average Body Mass Index (BMI) of the total population.

- Summary: min=1.0, max=87.3, mean=38.32, std=20.04.
- Distribution: the distribution is highly skewed to the right.

12. Under-five deaths: integer variable representing the number of deaths of children under five years of age per 1000 population.

- Summary: min=0, max=250, mean=42.04, std=160.45.
- Distribution: the distribution is highly skewed to the right.

13. Polio: continuous variable representing the immunization coverage among 1-year-olds against Polio.

- Summary: min=3.0, max=99.0, mean=82.55, std=23.36.
- Distribution: the distribution is slightly skewed to the left.

14. Total expenditure: continuous variable representing the total health expenditure as a percentage of GDP.

- Summary: min=0.37, max=17.6, mean=5.94, std=2.38.
- Distribution: the distribution is highly skewed to the right.

15. Diphtheria: continuous variable representing the percentage of the population who received Diphtheria Tetanus Pertussis (DTP) immunization among 1-year-olds.

- Summary: min=2, max=99, mean=82.32, std=23.74.
- Distribution: the distribution is slightly skewed to the left. Most of the values fall between 80 and 100 with some values towards the lower end of the distribution.

16. HIV/AIDS: continuous variable representing the deaths per 1,000 live births due to HIV/AIDS.

- Summary: min=0.1, max=50.6, mean=1.74, std=4.5.
- Distribution: the distribution is highly skewed to the right.

17. GDP: continuous variable representing the Gross Domestic Product per capita.

- Summary: min=1.68, max=119172.7, mean=7483.16, std=14270.17.
- Distribution: the distribution is highly skewed to the right.

18. Population: continuous variable representing the population of the country.

- Summary: min=34, max=1378.68M, mean=1292854.46, std=4233914.38.
- Distribution: the distribution is highly skewed to the right.

19. Thinness 1-19 years: continuous variable representing the rate of thinness among children and adolescents aged 10 to 19 years old.

- Summary: min=0.1, max=27.7, mean=5.04, std=4.59.
- Distribution: the distribution is highly skewed to the right.

20. Thinness 5-9 years: continuous variable representing rate of thinness among children aged 5 to 9 years old.

- Summary: min=0.1, max=28.6, mean=4.94, std=4.6.
- Distribution: the distribution is highly skewed to the right.

21. Income composition of resources: continuous variable representing the Human Development Index (HDI) in terms of income composition of resources.

- Summary: min=0.0, max=0.94, mean=0.63, std=0.21.
- Distribution: the distribution is slightly skewed to the left.

22. Schooling: continuous variable representing the average number of years of schooling for the population aged 15 years and older.

- Summary: min=0.0, max=20.7, mean=11.99, std=3.36.
- Distribution: the distribution is slightly skewed to the left.

Stratified sampling was done based on the ‘Status’ of variables, as developed and developing countries.

Data Cleaning

The first step in the data cleaning process was to strip the column names to remove unnecessary spaces or characters. This was done using the strip () function in Python. The second step in the data cleaning process involved checking the dataset for any duplicate rows, which can occur when multiple entries for the same observation are present. To identify duplicate rows, we used the duplicated () function in Python, which returns a Boolean value for each row in the dataset indicating whether it is a duplicate of a previous row. After applying this function, we found that there were no duplicate rows in the dataset, indicating that each observation was unique. Therefore, we did not need to remove any duplicate rows using the drop_duplicates() function in Python. The next step involved checking for null values in the dataset. This was done using the is null() function in Python. It was found that there were null values in some of the variables. To handle the null values, they were replaced with the mean value of the respective column. This was done using the fillna() function in Python.

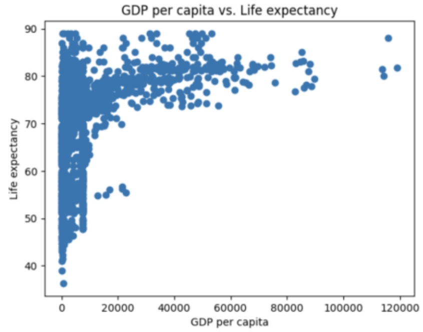
Exploratory Data Analysis

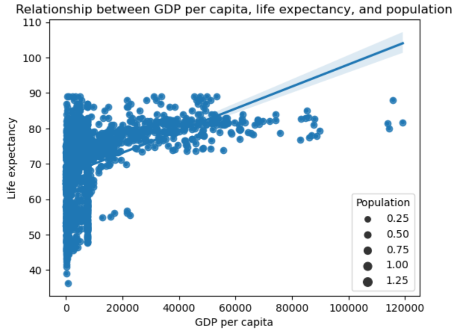
Box plots

Box plots are used to show how data is distributed and to identify potential outliers. The Interquartile Range (IQR) approach was used to find and eliminate outliers in the variables (Sharma, 2021). A box plot showed the changes in the data's distribution after locating and eliminating outliers.

Scatter plots

Scatter plots between Life expectancy and GDP show positive association which shows that as GDP value increases, the life expectancy value also increases. This indicates a relationship between the two variables.





Barchart

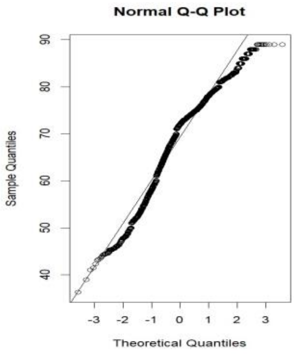
Barchart was used to see predictor variables and display their importance. This helped us identify any patterns or trends in the data like the strong association between Life expectancy and Schooling as we all Income composition of resources and its moderate association GDP and percentage expenditure.

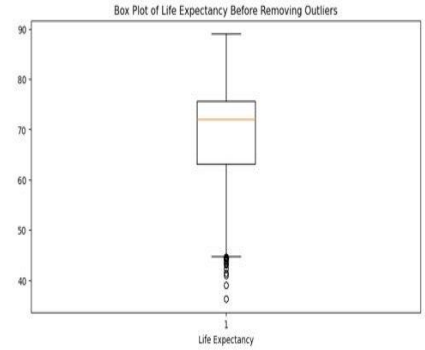
Summary Statistics

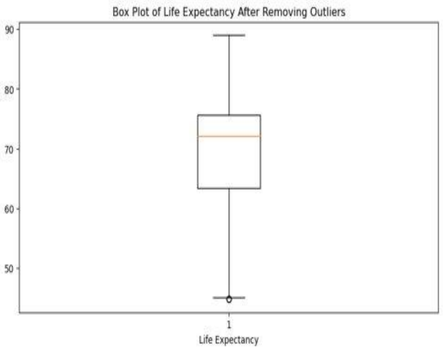
The describe() function is used to get the relevant summary statistics of the data including mean, median, standard deviation, min and max values, and quartiles. This provided an initial understanding of the data and helped in identifying any potential issues.

Data distribution using Q-Q plot

The qqnorm() function is used to check the normality of the data distribution and the resulting Q-Q plot appears to have an approximately normal distribution. Therefore, t-test and ANOVA can be used for analysis.







Subsets

We stratified the countries in the dataset into two groups - developed and developing countries based on their "Status" variable, and then performed analysis tests on each group separately.

Statistical Methods

T-Test

First, a t-test was done between the life expectancy of developed and developing countries. The results indicate that sample mean of life expectancy in developed countries is 79.19785, while the sample mean of life expectancy in developing countries is 67.12018. The 95% confidence interval for the difference in means is (11.28313, 12.87222), indicating that we can be 95% confident that the true difference in means lies between these two values.

Then, a t-test was done between the GDP of developed and developing countries. The sample mean of GDP for developed countries is 20232.108 and for developing countries is 4792.531. The 95 percent confidence interval for the difference in means is (14318.08, 16561.07), meaning that we can be 95 percent confident that the true difference in means lies within this range. The t-statistic is 26.994, indicating a significant difference between the means of the two groups. The degrees of freedom are 2936, and the p-value is less than 2.2e-16, which suggests that the true difference in means is not equal to 0.

ANOVA

After Levene's Homogeneity test was done to check for equal variances, the ANOVA test was used to analyze the impact of the "Status" variable on the "Life.expectancy" variable. The table shows the sum of squares, mean squares, F value, and p-value for the model. The F value of 888.3 and the p-value of <2e-16 suggest that there is a significant effect of the "Status" variable on the "Life.expectancy" variable. The "Status" variable is a significant predictor of the "Life.expectancy" variable, meaning that there is a significant difference in the mean life expectancy between developed and developing countries. The residual row shows the sum of squares and mean square errors of the model, which represent the variation in the "Life.expectancy" variable that cannot be explained by the "Status" variable.

Pearson's Correlation Coefficient

Since we are considering continuous variables in our study, we have chosen the Pearson's correlation coefficient. First, we checked the correlation between Life expectancy and GDP along with percentage expenditure after which we added Schooling variable as well. The result was almost the same around 0.42 for both with minor difference. However, GDP and Schooling showed slightly higher correlation comparatively.

Multiple Linear Regression

The data was subset including the relevant columns such as Life expectancy, GDP, and percentage expenditure to run the regression model. Using the summary() function, we obtained various statistics including estimated coefficients, along with standard errors, t-values, and p-values. The goodness of the fit of the model can be assessed using the R-squared value, along with the F-statistic and its corresponding p-value.

The coefficient for GDP is statistically significant with p-values < 0.05, indicating that it has a significant linear relationship with life expectancy. The coefficient for percentage expenditure is not statistically significant (p-value > 0.05), indicating that there is no significant linear relationship between percentage expenditure and life expectancy.

The "Multiple R-squared" value (0. 1853) indicates that the model explains 18.53% of the variability in life expectancy. The "Adjusted R-squared" value (0.1853) adjusts for the number of predictor variables in the model. The "F-statistic" value (333.8) and its associated p-value (< 2.2e-16) indicate that the overall model is statistically significant, meaning that at least one of the predictor variables is significantly associated with life expectancy.

A multiple linear regression model was used to examine the relationship between Life expectancy, GDP, percentage expenditure, and Schooling. The results showed that the coefficients for GDP and Schooling were statistically significant with a p-value < 0.05, while percentage expenditure was not. Specifically, Schooling had a strong and significant impact on life expectancy, as evidenced by a high t-value of 46.397, indicating that this association is not due to chance and is statistically meaningful.

Limitations

- Causation vs. correlation: These statistical methods can only identify correlations between variables but cannot prove causation. For example, while GDP, percentage expenditure, and schooling may be correlated with life expectancy, they may not necessarily cause an increase in life expectancy.
- Confounding variables: The effects of confounding variables, which are variables that are correlated with both the independent and dependent variables, can distort the results of these statistical methods.
- Biased data: The accuracy of these statistical methods depends on the quality of the data used. If the data is biased or incomplete, the results obtained may not accurately reflect the true relationship between the variables.

Conclusion

The two-sample t-test was conducted to compare the mean life expectancy between two groups, and it suggested that, on average, people in developed countries have a significantly higher life expectancy compared to those in developing countries. The two-sample t-test which was conducted to compare the mean percentage expenditure between two groups and the results suggest that, on average, developed countries have significantly higher percentage expenditure compared to developing countries. This establishes the impact of a country’s economy on life expectancy.

ANOVA test was conducted, and it indicates that the "Status" variable has a significant impact on the outcome being measured and this test indicates a highly significant difference between the groups, with the variable "Status" having a strong impact on the outcome being measured. This can be used as evidence to reject the null hypothesis since the economic status of a country is impactful on Life expectancy.

The multiple linear regression model was fitted to predict "Life expectancy" using the predictors "GDP," "percentage expenditure," and "Schooling.". This regression model suggests that GDP and Schooling have significant relationships with life expectancy, while "percentage expenditure" does not have a significant relationship. A higher R-squared value indicates that the model explains a greater proportion of the variance in the response variable, which is desirable. Initially, the model's performance was decent, but after adding schooling as a predictor, it improved significantly. This suggests that higher literacy rates may contribute to higher GDP and a stronger economy, which in turn may lead to higher life expectancy.

So, finally with higher F-statistic and p-value of less than 0.05 indicates that we can reject the null hypothesis with 95% confidence interval, meaning that there is a significant relationship between at least one of the predictor variables (GDP, percentage expenditure, and Schooling) and the response variable (life expectancy).

References

Life Expectancy (WHO). (n.d.). Life Expectancy (WHO) | Kaggle. https:///datasets/kumarajarshi/life-expectancy-who

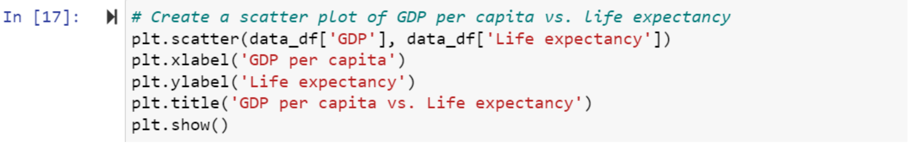
pandas.DataFrame.describe — pandas 2.0.1 documentation. (n.d.). Pandas.DataFrame.Describe — Pandas 2.0.1 Documentation. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html

Sharma, N. (2021, August 3). Ways to Detect and Remove the Outliers. Medium. https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

Zaman, S. B., Hossain, N., Mehta, V., Sharmin, S., & Mahmood, S. A. I. (2017). An association of total health expenditure with GDP and life expectancy. Journal of Medical Research and Innovation, 1(2), AU7-AU12. https://doi.org/10.15419/jmri.72

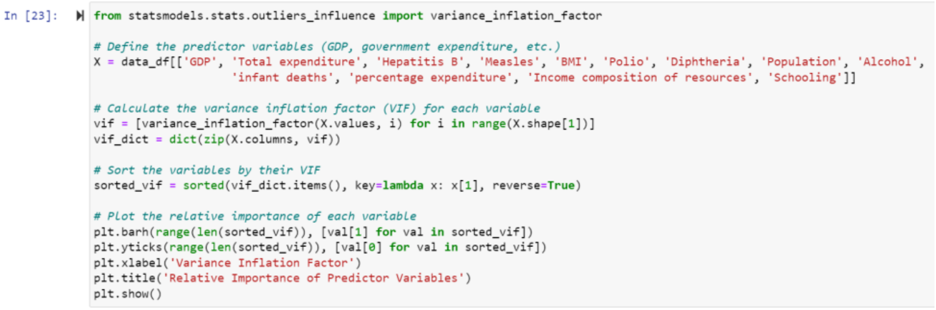
Appendix

These are some of the following codes we used in our project.



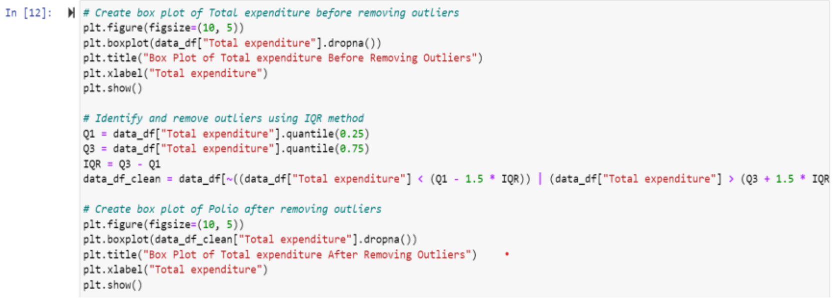
```
In [17]:  # Create a scatter plot of GDP per capita vs. life expectancy
          plt.scatter(data_df['GDP'], data_df['Life expectancy'])
          plt.xlabel('GDP per capita')
          plt.ylabel('Life expectancy')
          plt.title('GDP per capita vs. Life expectancy')
          plt.show()
```

To create a scatter plot.



```
In [23]:  from statsmodels.stats.outliers_influence import variance_inflation_factor

          # Define the predictor variables (GDP, government expenditure, etc.)
          X = data_df[['GDP', 'Total expenditure', 'Hepatitis B', 'Measles', 'BMI', 'Polio', 'Diphtheria', 'Population', 'Alcohol',
                       'infant deaths', 'percentage expenditure', 'Income composition of resources', 'Schooling']]

          # Calculate the variance inflation factor (VIF) for each variable
          vif = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
          vif_dict = dict(zip(X.columns, vif))

          # Sort the variables by their VIF
          sorted_vif = sorted(vif_dict.items(), key=lambda x: x[1], reverse=True)

          # Plot the relative importance of each variable
          plt.barh(range(len(sorted_vif)), [val[1] for val in sorted_vif])
          plt.yticks(range(len(sorted_vif)), [val[0] for val in sorted_vif])
          plt.xlabel('Variance Inflation Factor')
          plt.title('Relative Importance of Predictor Variables')
          plt.show()
```

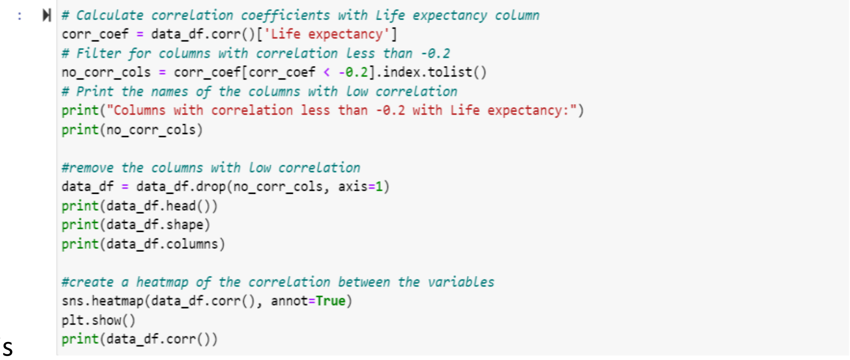
To create bar graph



```
In [12]:  # Create box plot of Total expenditure before removing outliers
          plt.figure(figsize=(10, 5))
          plt.boxplot(data_df["Total expenditure"].dropna())
          plt.title("Box Plot of Total expenditure Before Removing Outliers")
          plt.xlabel("Total expenditure")
          plt.show()

          # Identify and remove outliers using IQR method
          Q1 = data_df["Total expenditure"].quantile(0.25)
          Q3 = data_df["Total expenditure"].quantile(0.75)
          IQR = Q3 - Q1
          data_df_clean = data_df[~((data_df["Total expenditure"] < (Q1 - 1.5 * IQR)) | (data_df["Total expenditure"] > (Q3 + 1.5 * IQR

          # Create box plot of Polio after removing outliers
          plt.figure(figsize=(10, 5))
          plt.boxplot(data_df_clean["Total expenditure"].dropna())
          plt.title("Box Plot of Total expenditure After Removing Outliers")
          plt.xlabel("Total expenditure")
          plt.show()
```

To create Box plot

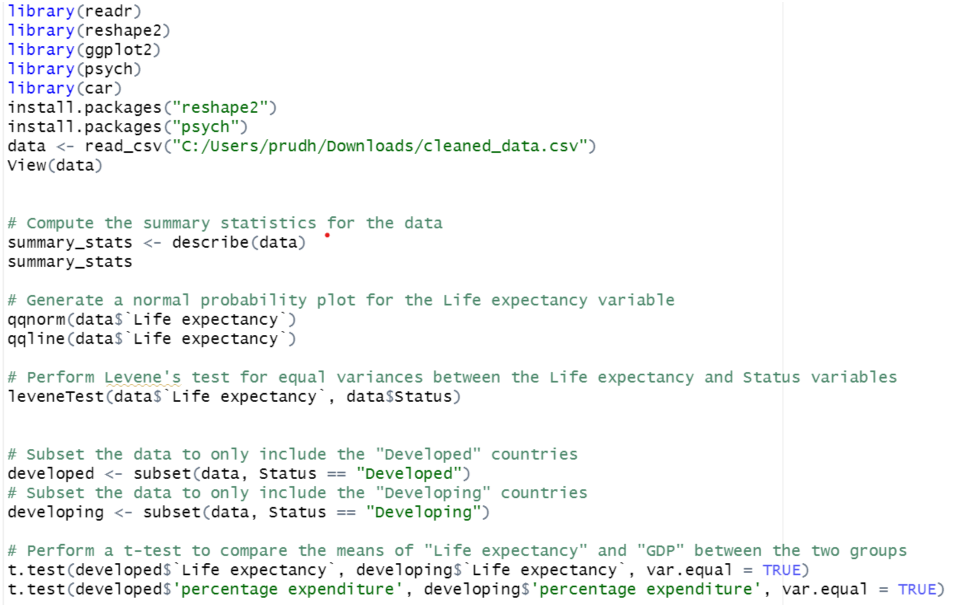
To



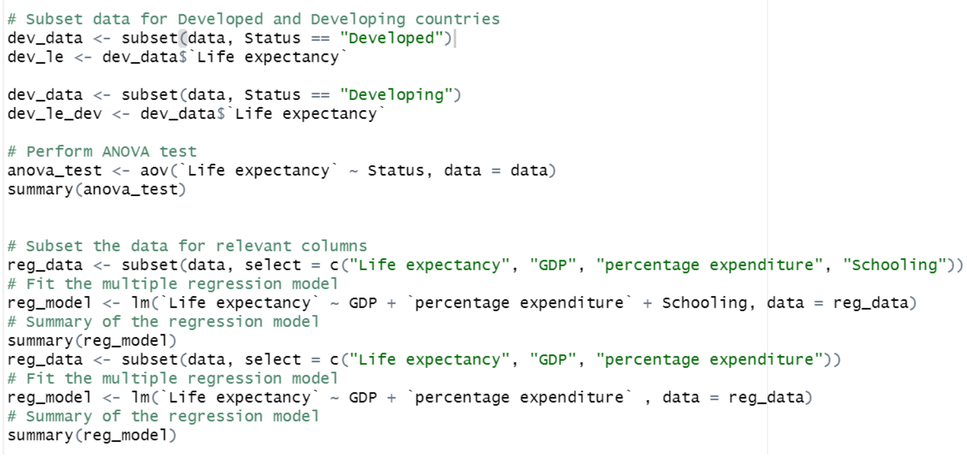
```
:  # Calculate correlation coefficients with Life expectancy column
   corr_coef = data_df.corr()['Life expectancy']
   # Filter for columns with correlation less than -0.2
   no_corr_cols = corr_coef[corr_coef < -0.2].index.tolist()
   # Print the names of the columns with low correlation
   print("Columns with correlation less than -0.2 with Life expectancy:")
   print(no_corr_cols)

   #remove the columns with low correlation
   data_df = data_df.drop(no_corr_cols, axis=1)
   print(data_df.head())
   print(data_df.shape)
   print(data_df.columns)

   #create a heatmap of the correlation between the variables
   sns.heatmap(data_df.corr(), annot=True)
   plt.show()
   print(data_df.corr())
```

create Heat map after analysis

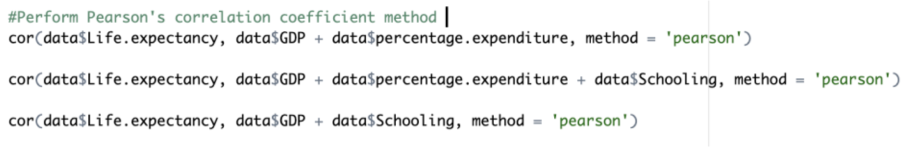
For Normality testing and hypothesis testing, the following codes were used in R.



```
library(readr)
library(reshape2)
library(ggplot2)
library(psych)
library(car)
install.packages("reshape2")
install.packages("psych")
data <- read_csv("C:/Users/prudh/Downloads/cleaned_data.csv")
View(data)


# Compute the summary statistics for the data
summary_stats <- describe(data)
summary_stats

# Generate a normal probability plot for the Life expectancy variable
qqnorm(data$`Life expectancy`)
qqline(data$`Life expectancy`)

# Perform Levene's test for equal variances between the Life expectancy and Status variables
leveneTest(data$`Life expectancy`, data$Status)


# Subset the data to only include the "Developed" countries
developed <- subset(data, Status == "Developed")
# Subset the data to only include the "Developing" countries
developing <- subset(data, Status == "Developing")

# Perform a t-test to compare the means of "Life expectancy" and "GDP" between the two groups
t.test(developed$`Life expectancy`, developing$`Life expectancy`, var.equal = TRUE)
t.test(developed$'percentage expenditure', developing$'percentage expenditure', var.equal = TRUE)
```

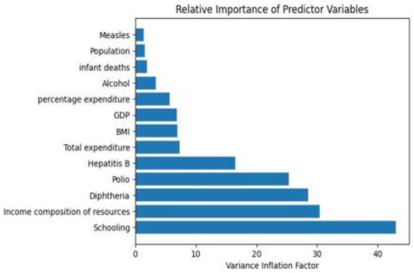


```
# Subset data for Developed and Developing countries
dev_data <- subset(data, Status == "Developed")
dev_le <- dev_data$`Life expectancy`

dev_data <- subset(data, Status == "Developing")
dev_le_dev <- dev_data$`Life expectancy`

# Perform ANOVA test
anova_test <- aov(`Life expectancy` ~ Status, data = data)
summary(anova_test)


# Subset the data for relevant columns
reg_data <- subset(data, select = c("Life expectancy", "GDP", "percentage expenditure", "Schooling"))
# Fit the multiple regression model
reg_model <- lm(`Life expectancy` ~ GDP + `percentage expenditure` + Schooling, data = reg_data)
# Summary of the regression model
summary(reg_model)
reg_data <- subset(data, select = c("Life expectancy", "GDP", "percentage expenditure"))
# Fit the multiple regression model
reg_model <- lm(`Life expectancy` ~ GDP + `percentage expenditure` , data = reg_data)
# Summary of the regression model
summary(reg_model)
```



```
#Perform Pearson's correlation coefficient method
cor(data$Life.expectancy, data$GDP + data$percentage.expenditure, method = 'pearson')

cor(data$Life.expectancy, data$GDP + data$percentage.expenditure + data$Schooling, method = 'pearson')

cor(data$Life.expectancy, data$GDP + data$Schooling, method = 'pearson')
```

```
Call:
lm(formula = `Life expectancy` ~ GDP + `percentage expenditure` + 
    Schooling, data = reg_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-25.7290  -3.0024   0.7817   4.1059  28.7741 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               4.566e+01  4.793e-01  95.268  < 2e-16 ***
GDP                       1.243e-04  2.034e-05   6.109 1.13e-09 ***
`percentage expenditure` -1.104e-04  1.317e-04  -0.838    0.402    
Schooling                 1.894e+00  4.082e-02  46.397  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.521 on 2934 degrees of freedom
Multiple R-squared:  0.5301,	Adjusted R-squared:  0.5296 
F-statistic:  1103 on 3 and 2934 DF,  p-value: < 2.2e-16
```


Relative Importance of Predictor Variables
Measles
Population
infant deaths
Alcohol
percentage expenditure
GDP
BMI
Total expenditure
Hepatitis B
Polio
Diphtheria
Income composition of resources
Schooling
Variance Inflation Factor