UTM

UNIVERSITI TEKNOLOGI MALAYSIA

PROBABILITY & STATISTICAL DATA ANALYSIS
SECTION-10

SECI1143
20232024/2

Project 2

LECTURER'S NAME:

DR MOHAMAD SHUKOR BIN TALIB

GROUP MEMBERS :

NAME	MATIRC NO.
CHANG WEN XUEN	A23CS5012
LIM CHEN XI	A23CS0103
FOO MING KUANG	A23CS5026
KAREN VOON XIU WEN	A23CS0229

Table of content

1.0 Introduction of background

Understanding the factors that influence life expectancy is crucial for improving public health outcomes and guiding policy decisions. Although numerous studies have been conducted to explore these factors, many have overlooked critical variables such as immunization rates and the Human Development Index (HDI). This study aims to address these gaps by examining life expectancy from a more comprehensive perspective, incorporating data from 2000 to 2015 across 193 countries.

1.1 Purpose of the Study

The primary objective of this study is to identify and analyze the key determinants of life expectancy using a robust dataset from the World Health Organization (WHO) and supplementary economic data from the United Nations. By integrating various health, economic, social, and immunization-related factors, we aim to develop a predictive model that can help countries understand and improve their life expectancy rates. This comprehensive approach allows for a deeper insight into how different variables interact and contribute to life expectancy across diverse global contexts.

1.2 Motivation and Interest

The motivation for this study stems from the observation that significant variables like immunization coverage and HDI have been underrepresented in previous research. As global health initiatives continue to evolve, understanding the impact of these factors becomes increasingly important. Moreover, by analyzing data over a 15-year period, we can observe the effects of health sector developments and other changes in a way that single-year studies cannot.

The inclusion of immunization factors such as Hepatitis B, Polio, and Diphtheria vaccination rates is particularly significant. Immunization is a well-established method for preventing disease and improving population health, yet its direct impact on life expectancy has not been thoroughly examined in past studies. By including these variables, this study aims to provide a more holistic view of the determinants of life expectancy.

1.3 Expectations from the Data

Impact of Immunization: Higher immunization rates for Hepatitis B, Polio, and Diphtheria are expected to positively correlate with increased life expectancy.

Economic Factors: Higher GDP and health expenditure are anticipated to be positively associated with life expectancy, indicating better overall health outcomes in wealthier countries.

Social Factors: Higher levels of education and better income distribution are expected to positively impact life expectancy, enhancing public health and longevity.

2.0 Dataset

Dataset URL: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

2.1 Data description

Variables (Description)	Type of Variable	Measurement Level
Country	Qualitative	Nominal
Year	Quantitative	Interval
Status	Qualitative	Nominal
Life expectancy	Quantitative	Ratio
Adult Mortality	Quantitative	Ratio
Infant Deaths	Quantitative	Ratio
Alcohol	Quantitative	Ratio
Percentage expenditure	Quantitative	Ratio
Hepatitis B	Quantitative	Ratio
Measles	Quantitative	Ratio
BMI	Quantitative	Ratio
Under-Five Deaths	Quantitative	Ratio
Polio	Quantitative	Ratio
Total expenditure	Quantitative	Ratio
Diphtheria	Quantitative	Ratio
HIV/AIDS	Quantitative	Ratio
GDP	Quantitative	Ratio
Population	Quantitative	Ratio

Thinness 1-19 Years	Quantitative	Ratio
Thinness 5-9 Years	Quantitative	Ratio
Income Composition of Resources	Quantitative	Ratio
Schooling	Quantitative	Ratio

2.2 Statistical Test Analysis

Selected Variables	**Objectives**	**Test Analysis and Expected Outcome**
Life expectancy	To test whether the average life expectancy in the dataset is significantly different from a global benchmark	Test Analysis: Hypothesis Test: One Sample t-Test Expected Outcome: If the p-value is less than 0.05, we will reject the null hypothesis and conclude that the average life expectancy in the dataset is significantly different from 70 years. Otherwise, we will fail to reject the null hypothesis.
Life expectancy, GDP	To determine the strength and direction of the relationship between life expectancy and GDP.	Test Analysis: Correlation Test: Pearson Correlation Expected Outcome: A positive correlation is expected, indicating that higher GDP is associated with higher life expectancy. The scatter plot should show an upward trend.
Life expectancy (dependent variable), Adult Mortality, BMI, GDP, HIV/AIDS (independent variables)	To predict life expectancy based on adult mortality, BMI, GDP, and HIV/AIDS.	Test Analysis: Regression Test: Multiple Linear Regression Expected Outcome: The regression model is expected to show that lower adult mortality, higher BMI, higher GDP, and lower HIV/AIDS rates are significant predictors of higher life expectancy. The scatter plot should show how the predicted values align with actual life expectancy values.

Life expectancy, Status	To test if there are significant differences in life expectancy between developing and developed countries.	Test Analysis: ANOVA (Analysis of Variance) test Expected Outcome: A significant ANOVA result (p-value < 0.05) would indicate that there are significant differences in life expectancy between developing and developed countries.
Status (Developing/Developed), Hepatitis B immunization coverage (categorized: low, medium, high)	To determine if there is an association between a country's development status and its Hepatitis B immunization coverage.	Test Analysis: Chi-Square Test of Independence Expected Outcome: A significant chi-square result (p-value < 0.05) would indicate an association between a country's development status and its Hepatitis B immunization coverage, suggesting that immunization coverage varies between developing and developed countries.

3.0 Data Analysis

3.1 Hypothesis Testing – 1 Sample

Based on the test we want to determine whether the average life expectancy in the dataset is significantly different from a global benchmark.

Let p = average life expectancy in dataset

Hypothesis statement:

H_0: p = 70

H_1: p ≠ 70

Significance level, α = 0.05

```
# Perform the t-test
t_test_result <- t.test(data$'Life expectancy' , mu = 70)
print(t_test_result)
```

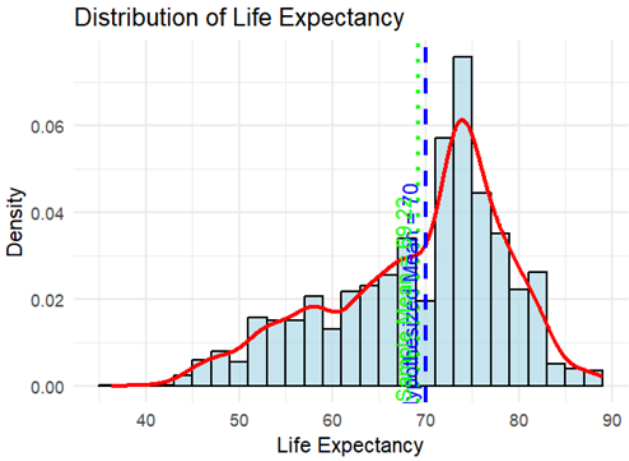
Using RStudio, we can get the result that test statistic t_0= -4.4036, the average life expectancy in your sample data is approximately 69.22 years and the p-value=1.1×10^{-5}

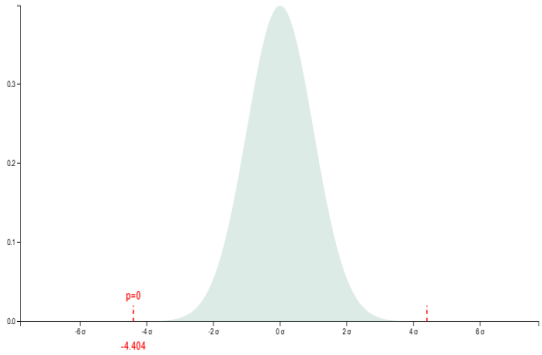
```
> print(t_test_result)

        One Sample t-test

data:  data$"Life expectancy"
t = -4.4036, df = 2927, p-value = 1.103e-05
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
 68.87982 69.57004
sample estimates:
mean of x
 69.22493

> print(paste("p-value:", p_value))
[1] "p-value: 1.10263601464004e-05"
```





Distribution Graph when test statistic t_0= -4.4036 and P-value=1.1×10^{-5}

Conclusion:

Given that we are dealing with life expectancy data and likely do not have the population standard deviation, we applied t-test. Since P-value $1.1\times10^{-5}< 0.05$, we reject the null hypothesis, there is sufficient evidence to conclude that the average life expectancy is significantly different from 70 years. The p-value obtained from the t-test is approximately 1.1×10^{-5}, which is a very small number, indicating strong evidence against the null hypothesis. This means that the average life expectancy in the dataset is significantly different from 70 years.

3.2 Correlation Test

In this correlation analysis, the variables that we used are life expectancy and GDP of the dataset. To determine the strength and direction of the relationship (linear relationship) between life expectancy and GDP, we applied Pearson's Product-Moment Correlation test in this analysis.

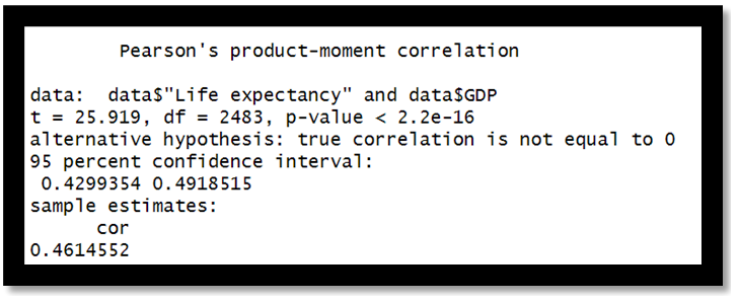
Hypothesis statement:

H_0: $\rho = 0$ (no linear correlation)

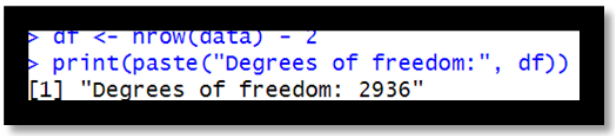
H_1: $\rho \neq 0$ (linear correlation exists)

Test statistics:

By using RStudio, we can get the result test statistic = 25.919 and $t_{0.025,2483}$ =1.9608.



When n is large, the difference between n and n−2n becomes negligible. Pearson correlation tests might slightly adjust degrees of freedom, often due to computational or methodological conventions. The degree of freedom should be df = 2938-2 = 2936.



From t-table, since this is a two-tailed test, there are two critical values:

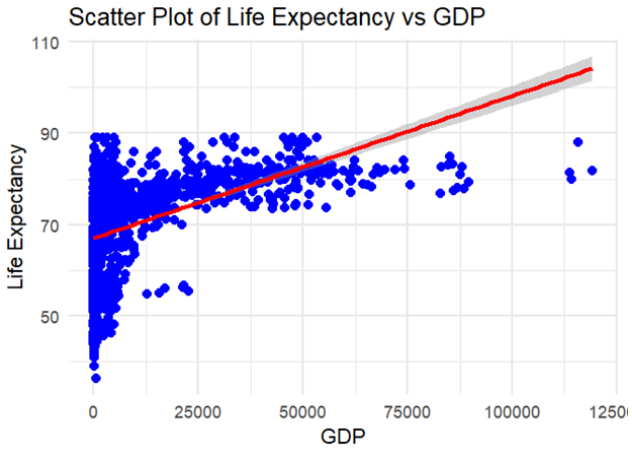
Lower tail critical value $-t_{0.025,2936} = -1.9608$

Upper tail critical value $t_{0.025,2936} = 1.9608$

From RStudio, we also get p-value = 2.2e - 16.

Thus, the H0 will be rejected if t < -1.9608 or t > 1.9608. Otherwise, fail to reject H_0

```
[1] "Critical value: 1.9607723063617"
```



From the scatter plot above, we can see that the points slope slightly upward, it indicates that there is a positive correlation between life expectancy and GDP, that is the higher the life expectancy, the higher the GDP.

```
"Correlation coefficient (r): 0.461455192620738"
```

By using RStudio, we also get a sample correlation coefficient, r = 0.461455, which indicates that there is a moderate positive linear correlation between life expectancy and GDP.

Conclusion:

Through the result of the scatter plot graph and the correlation test, it is clearly shown that the life expectancy is correlated with the GDP. They have a positive linear relationship,

So when the GDP increases, life expectancy increases. Since the test statistic t=25.919 > upper tail critical value $t_{0.025,2936} = 1.9608$, we reject the null hypothesis. There is sufficient evidence to conclude that there is a linear relationship between life expectancy and GDP.

3.3 Regression Test - Multiple Linear Regression

We used the life expectancy data of a random sample of countries for the regression test. In this test, we aimed to find out whether there is a linear relationship between life expectancy and the variables: adult mortality, BMI, GDP, and HIV/AIDS. The dependent variable, denoted as y, is life expectancy, while the independent variables are adult mortality, BMI, GDP, and HIV/AIDS.

The multiple linear regression model was constructed as follows:

$$LifeExpectancy = \beta0 + \beta1(AdultMortality) + \beta2(BMI) + \beta3(GDP) + \beta4(HIV\ AIDS) + \epsilon$$

The summary of the regression model is as follows:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.789e+01  3.429e-01  198.00   <2e-16 ***
AdultMortality -2.702e-02  1.110e-03  -24.34   <2e-16 ***
BMI             1.487e-01  6.140e-03   24.21   <2e-16 ***
GDP             1.513e-04  8.228e-06   18.38   <2e-16 ***
HIVAIDS        -4.836e-01  2.396e-02  -20.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the plot, we can see the line is nearly at a 45-degree angle, which indicates a strong linear relationship between the predicted and actual life expectancy values. This suggests that the model is a good fit for the data.

From the analysis, we obtain the following regression coefficients:

- **Intercept(**β0**)**: 67.89, which represents the estimated life expectancy when all predictor

variables are zero.

- **Coefficient for Adult Mortality(β1)** : -0.027, indicating that for every one-unit increase in adult mortality, life expectancy decreases by 0.027 years, holding other factors constant.
- **Coefficient for BMI (β2)**: 0.15, suggesting that for every one-unit increase in BMI, life expectancy increases by 0.15 years, holding other factors constant.
- **Coefficient for GDP (β3)**: 0.00015, showing that for every one-unit increase in GDP, life expectancy increases by 0.00015 years, holding other factors constant.
- **Coefficient for HIV/AIDS (β4)**: -4.84, meaning that for every one-unit increase in HIV/AIDS prevalence, life expectancy decreases by 4.84 years, holding other factors constant.

The estimated regression model is as below:

$$\hat{y} = 67.89 - 0.027\,(AdultMortality) + 0.15(BMI) + 0.00015(GDP) - 4.84(HIV\ AIDS)$$

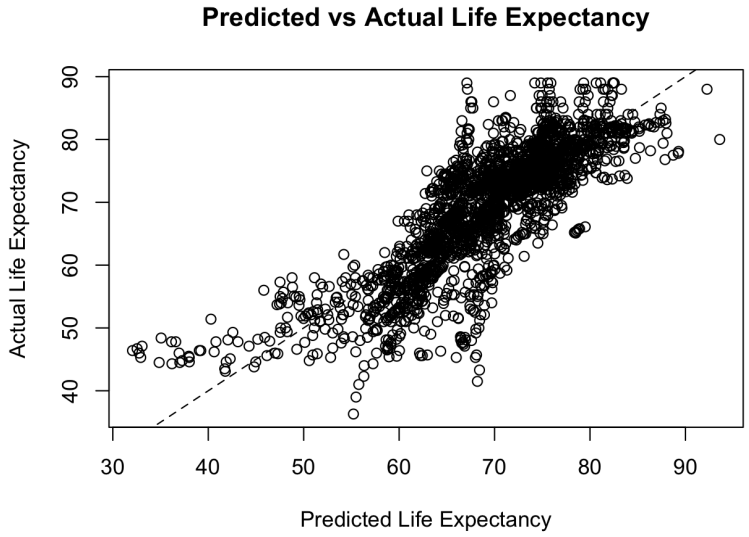
This equation indicates that the life expectancy can be predicted based on the values of adult mortality, BMI, GDP, and HIV/AIDS prevalence. The intercept value of 67.89 shows the baseline life expectancy when all predictor variables are zero.

To measure the goodness-of-fit of the model, we calculate the coefficient of determination (R^2):

$$R^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} = 0.877$$

This R^2 value of 0.877 indicates that approximately 87.7% of the variance in life expectancy is explained by the independent variables in the model. This high R^2 value suggests that the model provides a good fit for the data.

The following is the scatter plot of the predicted life expectancy values against the actual life expectancy values:



Graph 3: predicted life expectancy values against the actual life expectancy values.

Based on the plot, we can see the line is nearly at a 45-degree angle, which indicates a strong linear relationship between the predicted and actual life expectancy values. This suggests that the model is a good fit for the data.

From the multiple linear regression analysis, it is evident that:

- Higher adult mortality rates are associated with lower life expectancy.
- Higher BMI is associated with higher life expectancy.
- Higher GDP per capita is associated with higher life expectancy, although with a relatively small effect.
- Higher HIV/AIDS prevalence is associated with significantly lower life expectancy.

The significant coefficients and the high R^2 value suggest that the regression model provides a reasonable prediction of life expectancy based on the selected variables. Policymakers and health organizations can use these findings to focus on interventions that target these significant predictors to improve the overall life expectancy of populations.

3.4 ANOVA (Analysis of Variance) test

In this data analysis, the ANOVA test is used to assess the equality of life expectancy means across different country statuses. The significance level used to test the null hypothesis is $\alpha = 0.05$.

Hypothesis Statement:

H_0: $\mu_{Developed} = \mu_{Developing}$

H_1: $\mu_{Developed} \neq \mu_{Developing}$

The number of samples (n), mean of samples ($\bar{x}$) and standard deviation of samples (s) for each sample are calculated as below:

Status	n	$\bar{x}$	s
Developed	2426	79.19785	3.930942
Developing	512	67.11147	9.006092

The mean between samples is $\bar{\bar{x}}$= 73.15466, the standard deviation between samples is s$\bar{x}$= 5.184567, variance between samples is ns$\bar{x}$2 = 61759.17 and variance within samples is sp 2 = 26.879.

```
> # Print the calculated values and F-critical value
> cat("Mean Life Expectancy for Developed countries: ", mean_developed, "\n")
Mean Life Expectancy for Developed countries:  79.19785
> cat("Standard Deviation for Developed countries: ", std_developed, "\n")
Standard Deviation for Developed countries:  3.930942
> cat("Sample size for Developed countries: ", n_developed, "\n\n")
Sample size for Developed countries:  512

> cat("Mean Life Expectancy for Developing countries: ", mean_developing, "\n")
Mean Life Expectancy for Developing countries:  67.11147
> cat("Standard Deviation for Developing countries: ", std_developing, "\n")
Standard Deviation for Developing countries:  9.006092
> cat("Sample size for Developing countries: ", n_developing, "\n\n")
Sample size for Developing countries:  2416
```

Diagram 3.4.1: Finding values of n, $\bar{x}$ and s using R studio.

$$MS_{between} = \frac{SS_{between}}{df_{between}} \quad MS_{within} = \frac{SS_{within}}{df_{within}}$$

Formula for MSB and MSW.

$$F = \frac{MS_{between}}{MS_{within}}$$

Formula for F-test statistic

```
> cat("Sum of Squares Between Groups (SSB): ", ss_between, "\n")
Sum of Squares Between Groups (SSB):  61714.72
> cat("Degrees of Freedom Between Groups (df_between): ", df_between, "\n")
Degrees of Freedom Between Groups (df_between):  1
> cat("Mean Square Between Groups (MSB): ", ms_between, "\n\n")
Mean Square Between Groups (MSB):  61714.72

> cat("Sum of Squares Within Groups (SSW): ", ss_within, "\n")
Sum of Squares Within Groups (SSW):  203776
> cat("Degrees of Freedom Within Groups (df_within): ", df_within, "\n")
Degrees of Freedom Within Groups (df_within):  2926
> cat("Mean Square Within Groups (MSW): ", ms_within, "\n\n")
Mean Square Within Groups (MSW):  69.64321
```

Diagram 3.4.2: calculated results for MSB And MSW using R studio.

Using the formula for the F-test statistic and applying it in R-programming, we find the value of F-test statistic =886.1556. The critical value of F at $\alpha = 0.05$ is obtained from the F-distribution table, which is 3.844639.

```
> cat("F-test statistic: ", f_statistic, "\n")
F-test statistic:  886.1556
> cat("F-critical value at alpha =", alpha, "is:", f_critical, "\n")
F-critical value at alpha = 0.05 is: 3.844639
>
```

Diagram 3.4.3 : finding values of F-test statistic and F-critical value using R studio.

From the results, since F-test statistic > F-critical value (886.1556 > 3.844639), we reject the null hypothesis, H_0, as there is sufficient evidence to claim that there is a significant difference in the means of life expectancy between Developed and Developing countries.

3.5 Chi-Square Test of Independence

Chi-square test of independence is used to test the relationship between two nominal variables. In this analysis, we use the variables Status (Developing/Developed) and Hepatitis B immunization coverage (categorized: low, medium, high) to assess whether there is an association between a country's development status and its Hepatitis B immunization coverage. We'll use the following rules for categorization:

- Low: Hepatitis B < 60
- Medium: 60 ≤ Hepatitis B < 80
- High: Hepatitis B ≥ 80

The formula for the test statistic is:

$$\chi^2 = \sum \frac{(Oij - Eij)^2}{Eij}$$

Where:

- Oij = Observed count at row i column j
- Eij = Expected count at row i column j

Hypothesis Statement:

H_0: Hepatitis B immunization coverage is independent of a country's development status.

H_1: Hepatitis B immunization coverage is dependent on a country's development status.

Test Statistic:

A significant level of 0.05 ($\alpha = 0.05$) is used to test the claim that Hepatitis B immunization

coverage is dependent on a country's development status.

$$df = (2 - 1)(3 - 1)$$

```
> alpha <- 0.05
> df <- (2 - 1) * (3 - 1)
> critical_value <- qchisq(1 - alpha, df)
> critical_value
[1] 5.991465
```

Diagram 3.5.1: Value of critical value using RStudio

Observed Frequencies for Variables Status and Hepatitis B immunization coverage:

Status	Low	Medium	High
Developed	24	9	306
Developing	298	335	1413

```
[1] "Observed Frequencies:"
> print(observed)

             High  Low Medium
Developed     306   24      9
Developing   1413  298    335
```

Diagram 3.5.2: Observed Frequencies for Variables Status and Hepatitis B immunization coverage using RSudio

Expected Frequencies:

Status	Low	Medium	High
Developed	45.76855	48.8956	244.3358
Developing	276.23145	295.1044	1474.6642

```
[1] "Expected Frequencies:"
> print(expected)

                  High       Low   Medium
Developed    244.3358  45.76855  48.8956
Developing 1474.6642 276.23145 295.1044
```

Diagram 3.5.2: Expected Frequencies for Variables Status and Hepatitis B immunization coverage using RStudio

Calculate Test Statistic value:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

```
> # Print chi-square test results
> print(chi_square_test)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 68.156, df = 2, p-value = 1.585e-15
```

Diagram 3.5.3: Test Statistic value using RStudio χ^2 = 68.156.

Since the test statistic value $\chi^2 = 68.156$ is greater than the critical value $\chi^2(2, 0.05) = 5.991$, it falls within the critical region. Thus, we reject $\mathbf{H}_0$. In Conclusion, there is sufficient evidence to conclude that there is an association between a country's development status and its Hepatitis B immunization coverage, at $\alpha = 0.05$. This suggests that Hepatitis B immunization coverage varies between developing and developed countries.

4.0 Conclusion

From the 1 sample hypothesis testing, we found out that the average life expectancy in the dataset is significantly different from the global benchmark of 70 years, as the p-value of 1.1×10^-5 is much less than the significance level of 0.05, leading to the rejection of the null hypothesis.

Next, for the correlation analysis, we found out that there is a moderate positive linear relationship between life expectancy and GDP, evidenced by a Pearson correlation coefficient of 0.461455 and a test statistic of 25.919. This led to the rejection of the null hypothesis, indicating a significant correlation between these variables.

For the regression analysis, we found out that the regression model showed strong predictive power with an R^2 value of 0.877, indicating that about 87.7% of the variance in life expectancy is explained by the model. Variables such as adult mortality, BMI, GDP, and HIV/AIDS rates were significant predictors of life expectancy.

From the ANOVA test, we observe that our F-test statistic (886.1556) is significantly greater than the F-critical value (3.844639), resulting in rejecting the null hypothesis H_0. There is sufficient evidence to conclude that there are significant differences in the means of life expectancy between developed and developing countries. Hence, the mean life expectancy varies significantly based on the country's development status

Lastly, for the chi-square test of independence, given a test statistic value of 68.156, which is greater than the critical value of 5.991, we reject the null hypothesis. This provides sufficient evidence to conclude that there is a significant association between a country's development status and its Hepatitis B immunization coverage. This results in a dependency on the country's development status.

In conclusion, our group gets to perform test analysis such as 1 sample hypothesis testing, correlation analysis, regression analysis, ANOVA test and chi-square test of independence using Rstudio. We believe that this project is indeed useful for our future as this project has developed our data analysis skills.