**Data Analysis and Visualization 2 – Project**

**Group Members:** Gurkaranjit, Dhruv, Matthew

**Task 2:**

**1. Research Question and Relevant Attributes**

**Research Question:**

1. Is there a significant difference in life expectancy between developed and developing countries?
2. In addition to testing the difference in life expectancy between developed and developing countries, we also tested the change in **Adult Mortality** between the years 2000 and 2015

**Relevant Attributes:**

**Life\_expectancy**: This is the target variable, as we want to compare the average life expectancy in developed and developing countries.

**Economy\_status\_Developed**:This binary variable (1 for developed, 0 for non-developed) helps categorize countries into "developed" and "developing" categories.

**Economy\_status\_Developing**: Like the previous attribute, but it indicates whether a country is developing (1 for developing, 0 for not).

**Adult Mortality:** Shows deaths of adults per 1000 population

These attributes are important for answering the research question, as they directly allow us to separate the countries into the two groups (developed and developing) and compare their life expectancy. And the adult mortality shows us if the mortality rate has improved over time

**2. Data Processing Techniques and Concepts**

This was handled by **task 1**

**3. Hypothesis Testing Process**

**Null Hypothesis (H₀):** There is no significant difference in life expectancy between developed and developing countries.

**Alternative Hypothesis (H₁):** There is a significant difference in life expectancy between developed and developing countries.

We will use a **two-sample t-test** to compare the life expectancy means between the two groups (developed and developing countries). The following steps were taken:

* Separate the data into two groups: one for developed countries (Economy\_status\_Developed = 1) and one for developing countries (Economy\_status\_Developing = 1).
* Extract the Life\_expectancy values for each group.
* We use the **two-sample t-test** formula, which calculates the difference between the means of the two groups, adjusts for the variability (standard deviations), and compares it to the pooled variance.

The reason we chose the t-test is because two groups are independent, the plots show normality for the most part with minor skewness coming from outliers. However, t-test can deal with minor outliers. The plots also show equal data spread showing similar variance.

A graph of a number of people

AI-generated content may be incorrect.A green graph with white background

AI-generated content may be incorrect.

**Test Results**

A computer screen shot of a test

AI-generated content may be incorrect.

**Results from the Test**:

* **t-statistic**: 53.685
* **p-value**: 2.2 x 10-16
* **Confidence Interval**: [11.71972, 12.60832]
* **Mean of Developed Countries**: 78.50574
* **Mean of Developing Countries**: 66.34173

Since the p-value is extremely small at 1% and 5%, we **reject** the **null** **hypothesis**. This shows a significant difference in life expectancy between developed and developing countries. It also shows developed countries having a higher life expectancy. Therefore with 95% and 99% confidence, we can reject the null hypothesis that there is no difference between developing and developed country’s life expectancy rate.

**Additional Hypothesis Test**

In this additional hypothesis test, we examined the change in Adult Mortality from 2000 to 2015. This was done using a paired t-test, as we are comparing two related samples Adult Mortality in the same countries but in two different years.

**Null Hypothesis** **(H₀)**: There is no significant difference in adult mortality between 2000 and 2015.

**Alternative Hypothesis** **(H₁)**: There is a significant difference in adult mortality between 2000 and 2015.

Steps that were taken to carry out this test was:

* Extract the Adult Mortality data for the years 2000 and 2015 for the same countries.
* We perform a paired t-test because we are comparing the same countries' adult mortality data at two different time points.

**Results from the Test**

t-statistic: 4.4625

p-value: 0.00001434

Confidence Interval: [30.13465,77.81711]

Mean Difference: 54.02588A computer code with black text

AI-generated content may be incorrect.

Since the p-value is smaller than 0.05 and 0.01, we reject the null hypothesis and conclude that there is a significant change in adult mortality between 2000 and 2015. This indicates improvements in healthcare or health conditions over time.

***Task 3:***

**Model Selection:**

We are choosing multiple linear regression to model how the several independent variables are influencing the dependent variable of Life\_expectancy. Why we choose linear because life\_expectancy is a continuous numeric variable. Our goal is to explore how changes in predictors (GDP, schooling, alcohol consumption, etc.) affect the life expectancy. With regression, it allows us to quantify individual contributions of predictors.

To build the model, we will include relevant predictors:

* Schooling
* GDP\_per\_capita
* Alcohol\_consumption
* BMI
* Adult\_mortality
* Infant\_deaths

**Multiple Linear Regression Assumptions**

1. Checking if residual values are normally distributed
2. Checking multicollinearity

The residual values are normally distributed based on the qq plots below which satisfies the first assumption for multiple linear regression. For the multicollinearity, we checked this by using the correlation heatmap from task 1 to make sure the independent variables are not highly correlated with each other.

A graph with a line drawn on it

AI-generated content may be incorrect.

A graph of a graph showing a number of schooling

AI-generated content may be incorrect.

A graph of black dots

AI-generated content may be incorrect.

A diagram of a graph

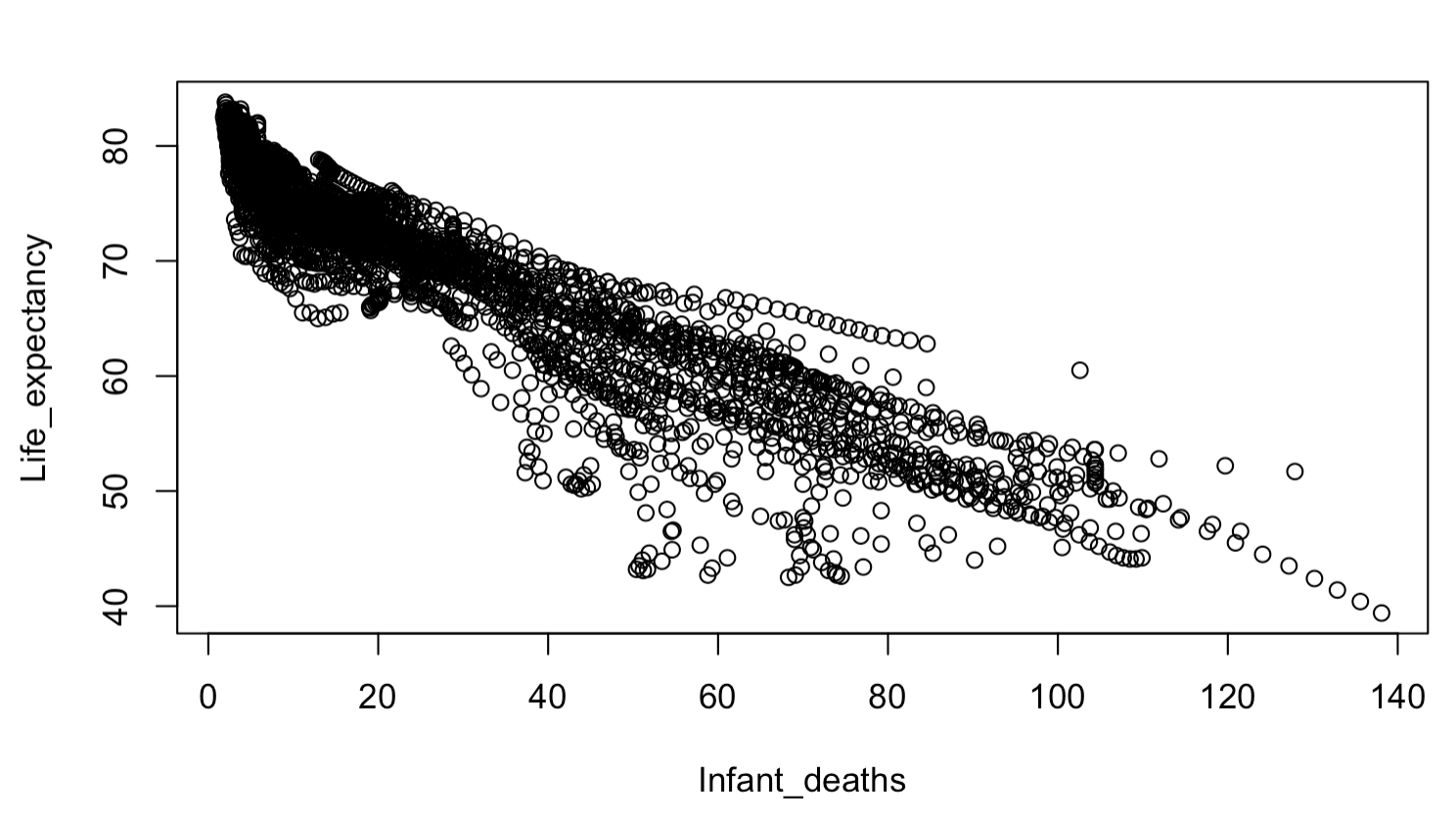
AI-generated content may be incorrect.

A black and white diagram

AI-generated content may be incorrect.

A graph of a graph showing a number of dots

AI-generated content may be incorrect.



A screenshot of a computer code

AI-generated content may be incorrect.

*Interpretation of Coefficient Table:*

|  |  |  |
| --- | --- | --- |
| Predictor | Estimate | Meaning |
| Intercept | ~82.94 | This is the base life expectancy when all predictors are 0 |
| Schooling | +0.1560 | For every extra year of schooling, life expectancy increases by 0.16 years |
| GDP\_per\_capita | +0.0000268 | An $1000 increase in GDP adds ~0.027 years to life expectancy |
| Alcohol\_consumption | +0.0738 | The consumption of alcohol correlates positively. Our reasoning behind this is it reflects wealthier countries |
| BMI | -0.107 | A higher BMI correlates negatively here. We believe it’s due to obesity-related health issues. |
| Adult\_mortality | -0.0478 | Each increase in adult mortality per 1000 reduces life expectancy by ~0.048 |
| Infant\_deaths | -0.1339 | Each increase in infant deaths reduces life expectancy by 0.13 years |

*Model Fit and Evaluation:*

|  |  |  |
| --- | --- | --- |
| Metric | Value | Meaning |
| Residual Std. Error | 1.427 | On average, the predictions were off by about 1.43 years which is good for life expectancy data |
| Multiple R-squared | 0.977 | Our model explains 97.7% of the variation in life expectancy which is very high |
| Adjusted R-squared | 0.977 | Very high confirms the fit even with multiple predictors |
| F-statistic | 20250+ | Extremely high value which indicates a strong overall model significance |