DATA 602-L01 – Group 7 – Final Project

Alvin Hui Tsz Chuen David Errington Lalith Nandakumar
Judy Kurupakorn Natcha

October 15, 2024

Contents

1 Introduction

For our project, we chose to look at the global life expectancy in 2019 and compare how it relates to two key factors: Gross Domestic Product (GDP) per capita, and health expenditure.

1.1 Scope & Objectives

The main purpose of this project is to test our statistical knowledge and apply some of the concepts learned throughout our studies in DATA 602. This will involve performing some initial exploratory data analysis on the data sets we selected to gain valuable insights and inform our research questions. From there, we will be moving on to some hypothesis testing and simple linear regression analysis to determine the relationship between each of these three variables.

1.2 Data Sets

1.2.1 Life Expectancy vs. GDP per capita

The first data set that we chose to analyze was global life expectancy vs. GDP per capita. This data set—which is publicly available on OurWorldinData.org, a project by the United Kingdom-based non-profit organization Global Change Data Lab—consists of:

- Eight (8) columns (including `Entity`, `Code`, `Year`, `Period life expectancy at birth`, `GDP per capita`, `900793-annotations`, `Population (historical)`, & `Continent`);
- 250 countries, across 7 different continents; and
- Roughly 65,000 records from 1543 to 2021.

This was obviously a lot of data to process, so to narrow down our research we chose to only focus our analysis on the year 2019, as this was the most recently available year which would still provide us with enough information to gather some valuable insights and avoid any potential anomalies due to the outbreak of the COVID-19 pandemic.

1.2.2 Life Expectancy vs. Health expenditure

The second data set that we chose to analyze was global life expectancy vs. health expenditure. This data set—which is also publicly available on OurWorldinData.org—consists of:

- Seven (7) columns (including country/`Entity`, `Code`, `Year`, `Life expectancy`, `Health expenditure per capita` , `Population (historical)`, & `Continent`);
- 250 countries, across 7 different continents; and
- Roughly 60,000 records from 1950 to 2023.

Once again, we chose to only focus our analysis on 2019, as this was the most recently available year which would still provide us with enough information to gather some insights all while staying consistent with our first data set.

1.2.3 Definitions

According to Our World in Data (2024),

> "**Health expenditure** [expressed in *international dollars* at 2015 prices] includes all health care financing schemes, [such as] government, compulsory contributory insurance, voluntary insurance, household out-of-pocket payments, and rest of the world financing schemes."
>
> "**International dollars** are a hypothetical currency that is used to make meaningful comparisons of monetary indicators of living standards. Figures expressed in international dollars are adjusted for inflation within countries over time, and for differences in the cost of living between countries."

1.3 Research Questions

For our project, we chose to focus our analysis on two main questions:

I. How does GDP per capita and health expenditure affect life expectancy?
   - What is the relationship between them, if any?

II. Which of the two factors (GDP per capita or health expenditure) has the most significant influence/correlation on determining one's life expectancy?

In trying to answer these questions, we ran the data through three different methods of analysis. The first was a high-level, exploratory analysis of the data to try and identify which countries or continents had the highest/lowest values of life expectancy, GDP per capita, and health expenditure. The second involved a hypothesis test of whether there was a statistically significant difference in the average life expectancy at birth across each continent. For the third and last, we chose to conduct a regression analysis to determine which of the two main factors we identified showed the most direct relationship or prominent correlation with life expectancy. The results of our findings, from each of these three analyses, are presented in the following sections below.

2 Exploratory Data Analysis & Visualization

2.1 Data Loading & Preparation

In this exploratory data analysis (EDA), we aim to explore and visualize the relationships among key factors to uncover potential insights and patterns. The analysis is organized into four subsections: an overview visualization by continent, an identification of the top countries with the highest metrics, a clustered bar chart revealing two variables, and box plots to illustrate the distribution and variability across regions. This structured approach will enhance our understanding of the relationships between life expectancy and important socioeconomic and health-related factors.

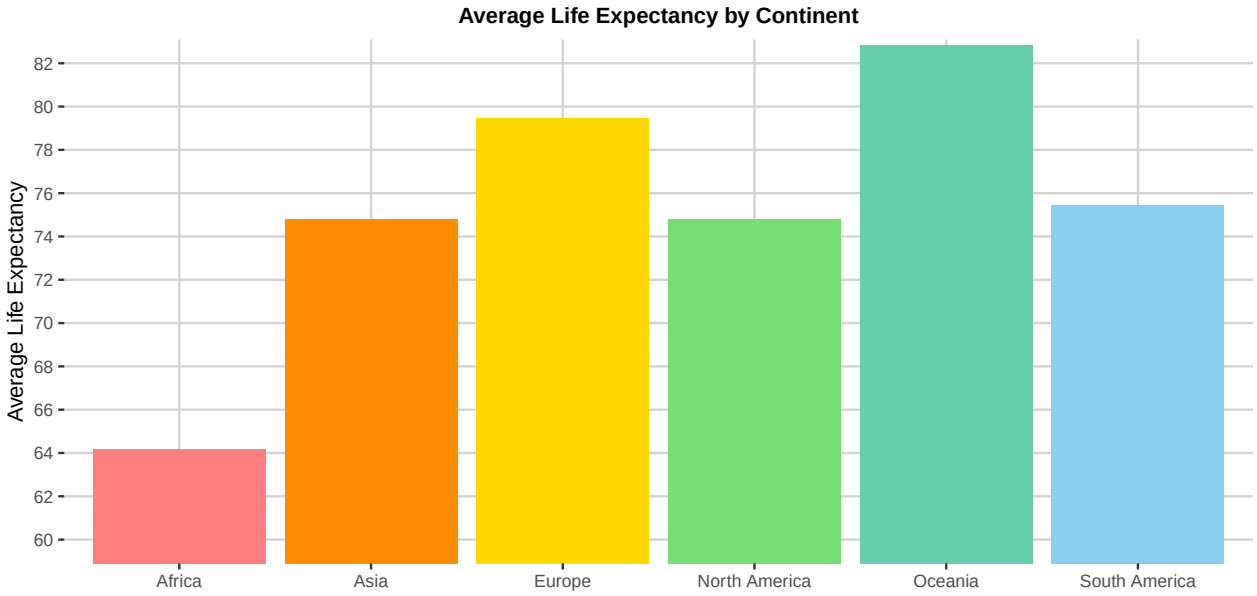
```
library(readr)
library(dplyr)
library(ggplot2)
library(ggpubr)
library(tidyverse)
library(tidyr)
library(patchwork)
# Import data
data <- read_csv("dataset_life_expectancy_2019.csv")
# Standardize column names
names(data)[names(data) == "Entity"] <- "entity"
names(data)[names(data) == "Code"] <- "code"
names(data)[names(data) == "Year"] <- "year"
names(data)[names(data) == "Period life expectancy at birth - Sex: all - Age: 0"] <-
  "life_expectancy"
names(data)[names(data) == "GDP per capita"] <- "gdp_per_capita"
names(data)[names(data) == "Population (historical)"] <- "population"
names(data)[names(data) == "Continent"] <- "continent"
names(data)[names(data) == "Health Expenditure"] <- "health_expense"
names(data)[names(data) == "life_Category"] <- "life_category"
names(data)[names(data) == "health_expenditure_category"] <-
  "Health_Expenditure_Category"
names(data)[names(data) == "GDP_Category"] <- "gdp_category"
names(data)[names(data) == "Population_Category"] <- "population_category"
```

2.2 Overview by Continent

This section explores continent-level data to examine the average life expectancy across different regions. The analysis will cover the total amounts of various variables while comparing average life expectancy with the averages of each variable to uncover relationships and trends. This approach will help illuminate how different factors interact and impact life expectancy across continents.

2.2.1 Average Life Expectancy by Continent

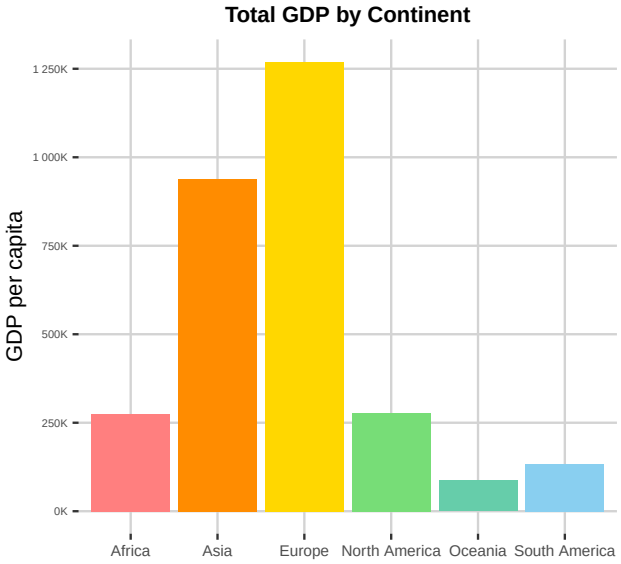
The bar graph below depicting average life expectancy across continents reveals distinct differences among regions. Oceania has the highest average at 82.85 years, despite representing only two countries, while Europe follows closely with an average of 79.48 years across 38 countries. Asia and North America show similar averages of 74.81 and 74.79 years, respectively, with South America at 75.43 years. Africa has the lowest average life expectancy at 64.18 years.

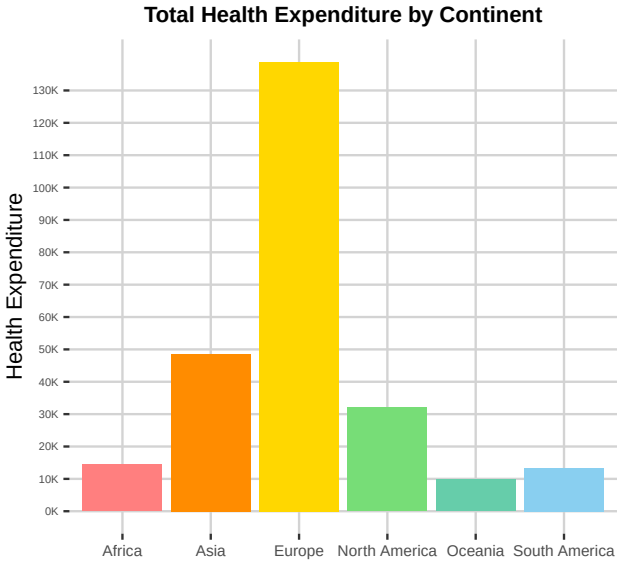


2.2.2 Summary of the Variables by Continent

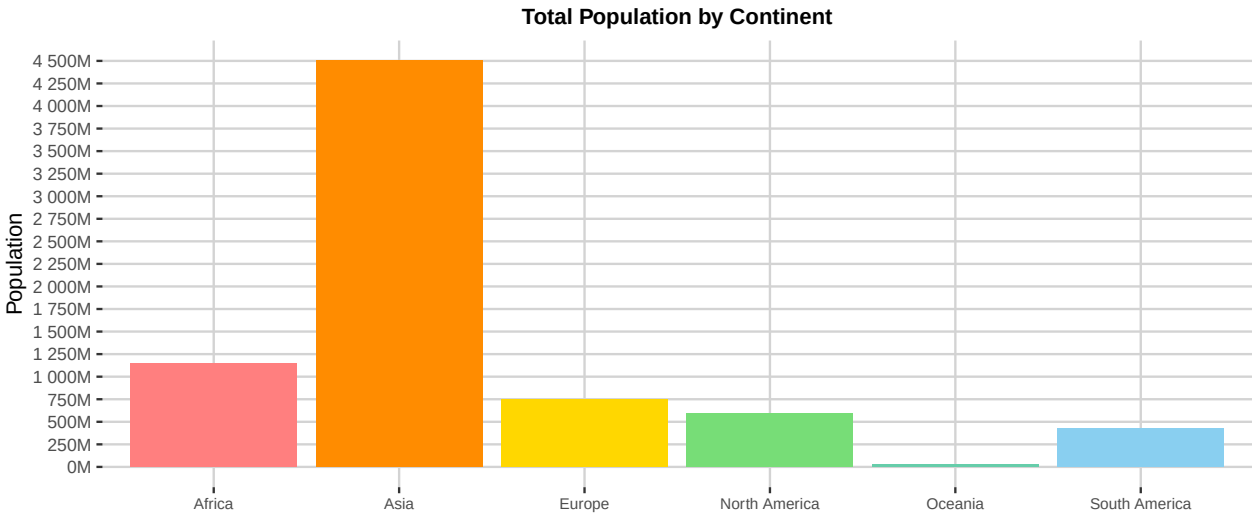
The total GDP graph indicates that Asia and Europe lead in economic output, significantly outpacing North America, Africa, South America, and Oceania. This concentration suggests potential economic disparities and implications for resource availability.

Health expenditure is highest in Europe, followed by Asia and North America, while Africa, South America, and Oceania show much lower spending. These differences highlight varying priorities in health investment, with Europe's higher expenditure potentially linked to better health outcomes.





For the Total Population Analysis, Asia has the largest population, followed by Africa, Europe, and North America, with Oceania and South America having smaller populations. High populations in Asia and Africa may drive demand for resources, affecting GDP and health spending.



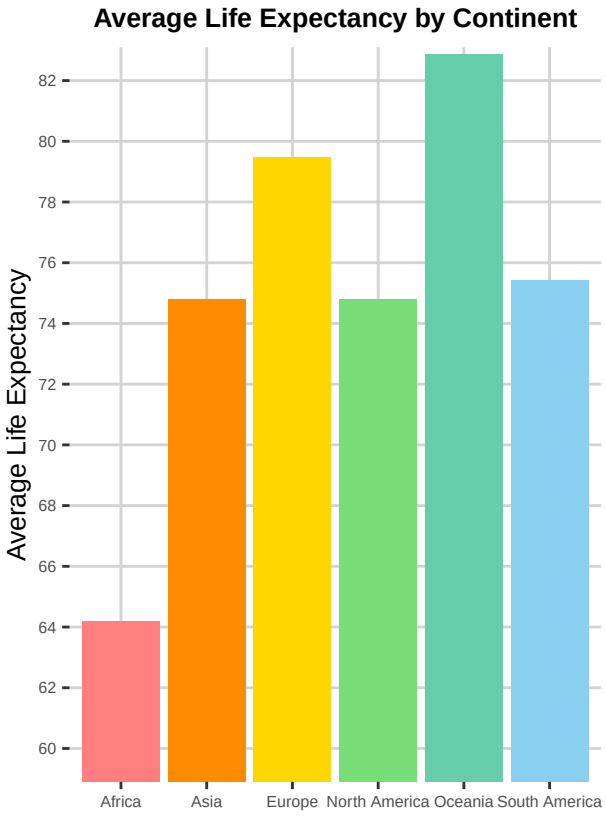
Overall, the analysis reveals significant economic and health disparities among continents, with higher GDP and health expenditures correlating with larger populations in Asia and Europe. These relationships suggest that economic resources and health investments are crucial for improving life expectancy across regions.

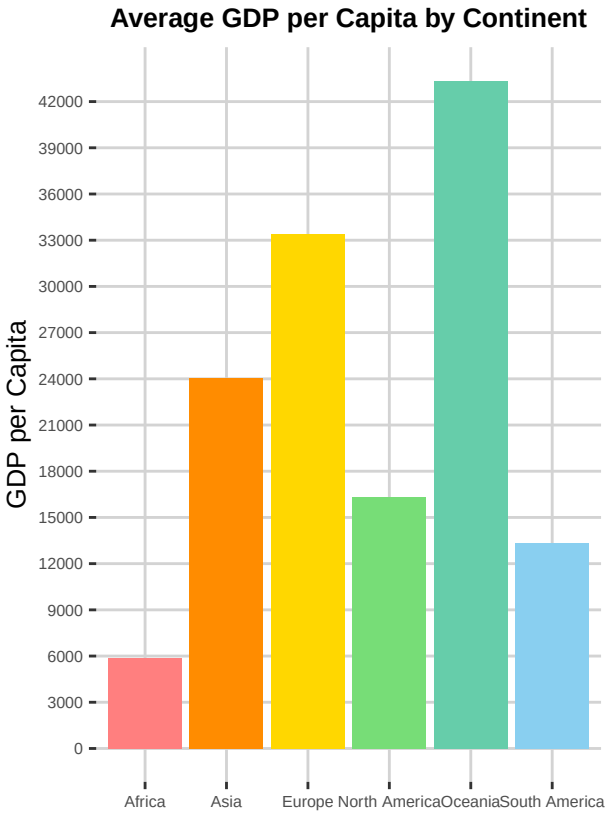
2.2.3 Average of the Variables by Continent

Average Life Expectancy with Average GDP per capita by Continent

The relationship between average life expectancy and GDP per capita varies significantly across continents. Oceania exhibits the highest average GDP per capita, which correlates with its leading life expectancy of 82.85 years. Similarly, Europe, with an average GDP per capita of about $33K, shows a high average life expectancy of 79.48 years.

In contrast, Africa, with the lowest GDP per capita, has the lowest average life expectancy at 64.18 years. This suggests a strong inverse relationship between economic wealth and life expectancy in this region. The data from Asia, North America, and South America reflect intermediate values, where higher GDP per capita generally aligns with increased life expectancy.



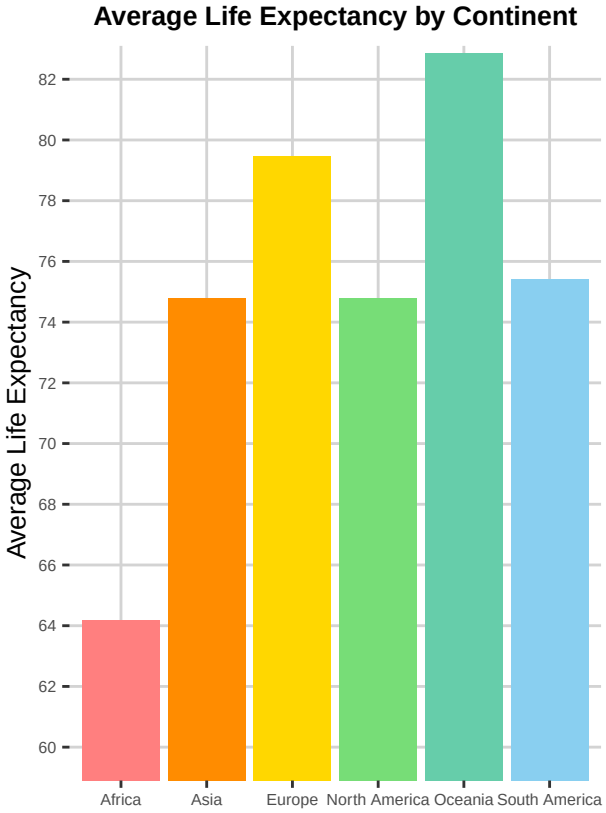


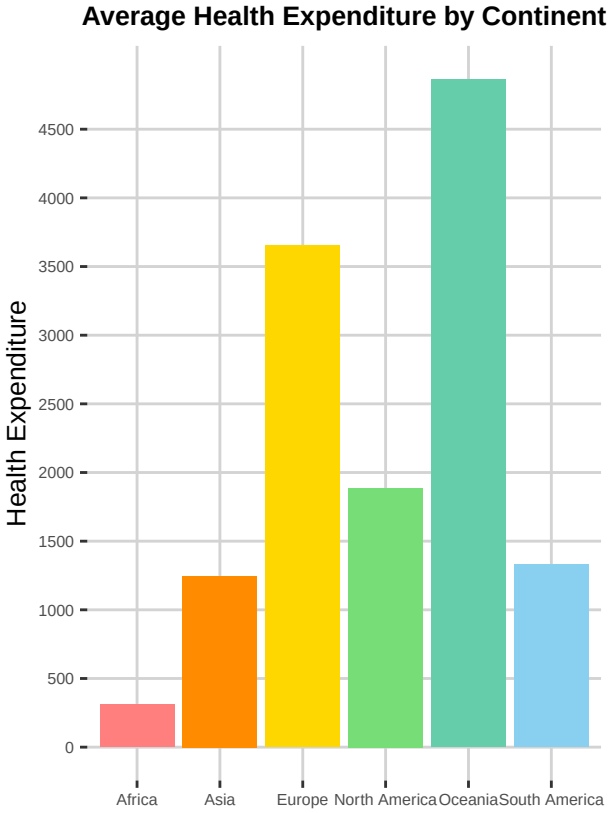
Overall, this analysis indicates that economic resources play a significant role in shaping life expectancy across different continents.

Average Life Expectancy with Average Health Expenditure by Continent

The relationship between average life expectancy and health expenditure across continents reveals clear patterns. Oceania leads with the highest average health expenditure, correlating with a life expectancy of 82.85 years. Europe follows with an average health expenditure of $3,600 and a life expectancy of 79.48 years. Conversely, Africa has the lowest health expenditure, aligning with its lowest life expectancy of 64.18 years.

This trend indicates that increased health spending is linked to longer life expectancy, as seen in Asia and North America, where higher expenditures ($1,200 and $1,800, respectively) correspond with life expectancies of 74.81 and 74.79 years. Overall, the data highlights the positive relationship between health expenditure and life expectancy, highlighting the importance of investing in health services to improve population health outcomes.



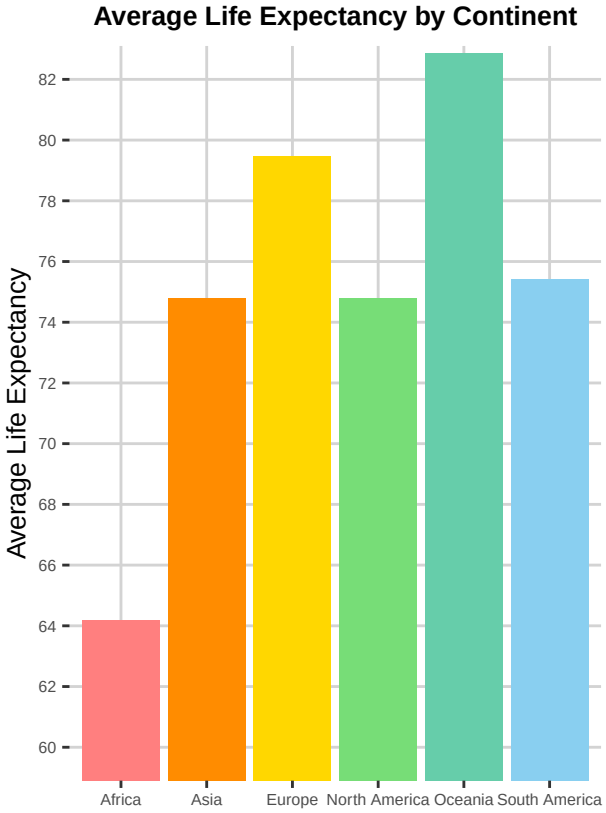


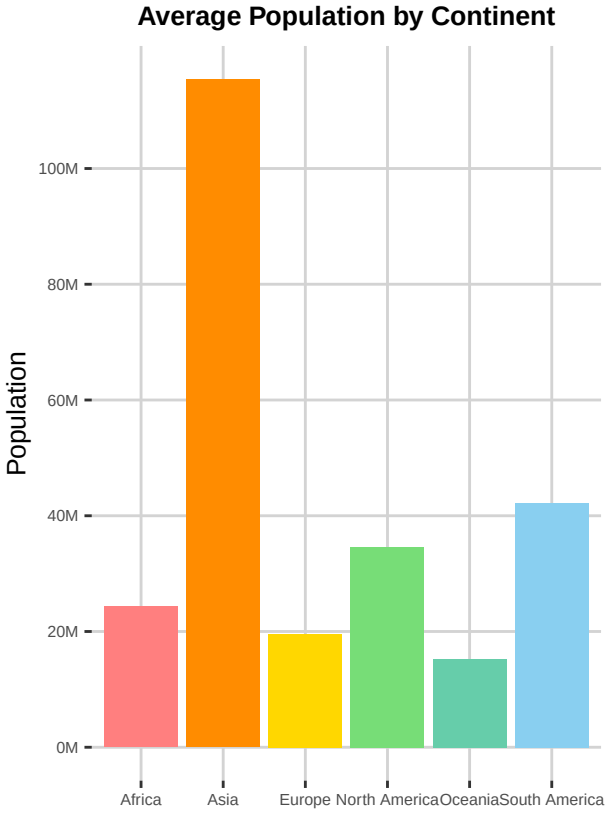
Average of Life Expectancy with Population by Continent

The relationship between average life expectancy and population size across continents doesn't show a direct, consistent correlation. Oceania, with the smallest average population of 15 million, has the highest average life expectancy of 82.85 years, while Africa, with an average population of 24 million, has the lowest life expectancy at 64.18 years.

Asia, the most populous continent with an average population of 115 million, shows a mid-range life expectancy of 74.81 years. Similarly, South America and North America, with average populations of 42 million and 34 million, respectively, each have an average life expectancy in the mid-70s.

In general, the data indicates that population size alone may not be a strong indicator of life expectancy, as other factors such as healthcare access, economic conditions, and health expenditure also play critical roles.

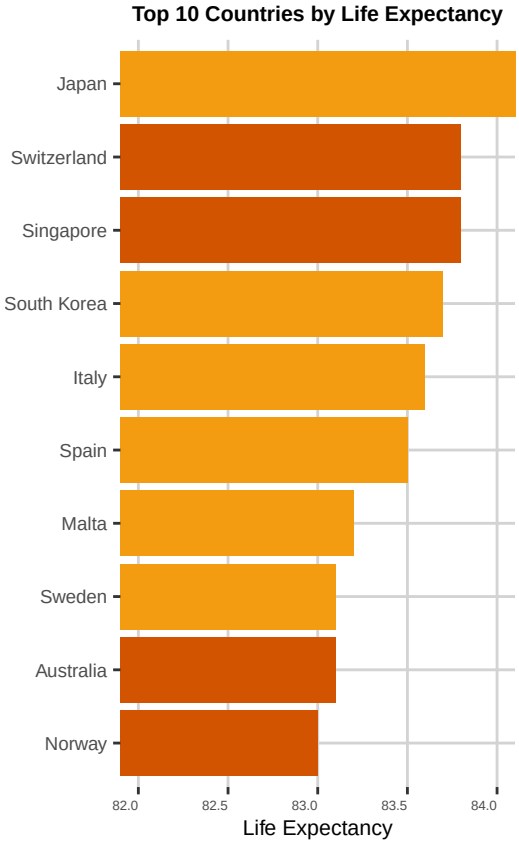


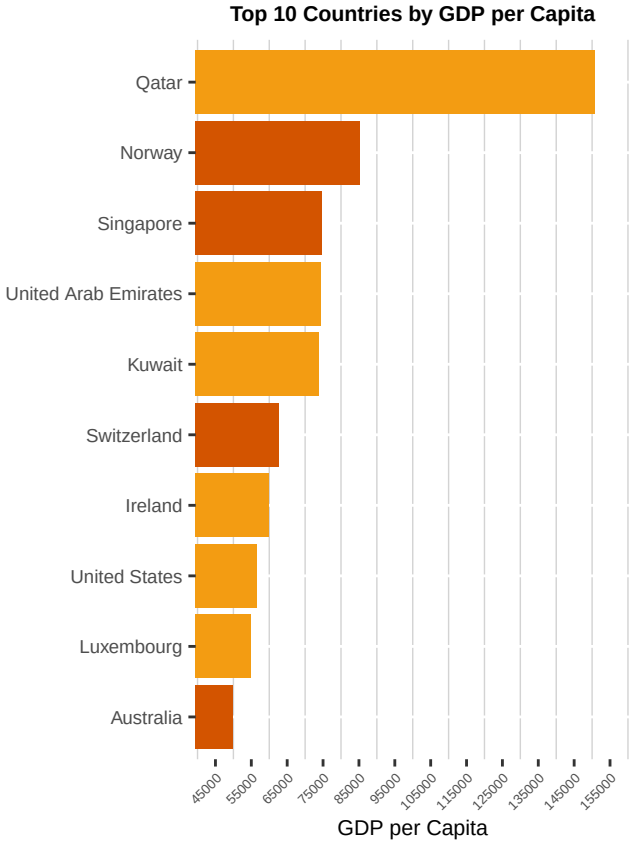


2.3 Top 10 Countries for Analyzation with Life Expectancy

2.3.1 Top 10 Countries for Life Expectancy & GDP per capita

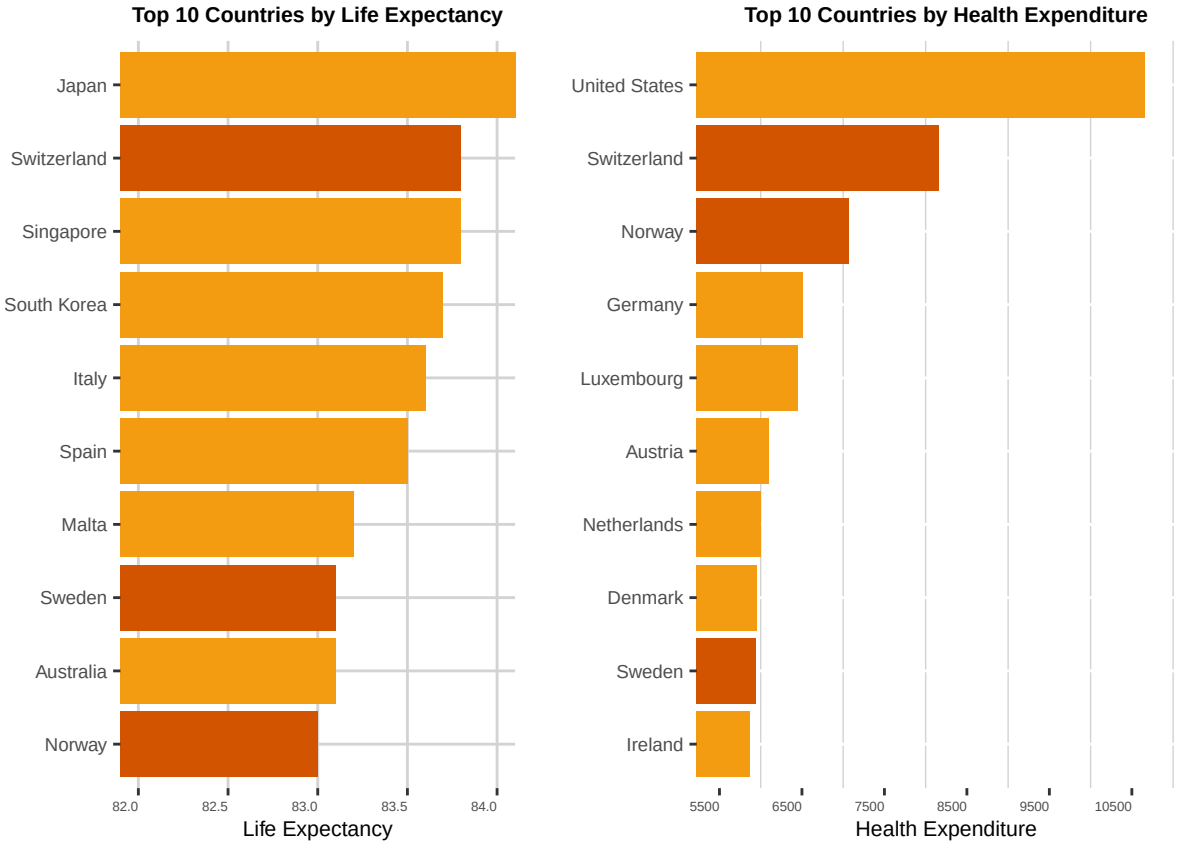
In analyzing the top 10 countries by life expectancy and GDP per capita, four countries: Norway, Switzerland, Singapore, and Australia, appear in both lists. Out of the total 153 countries in the dataset, these four common countries represent a small yet significant subset. These countries not only have high life expectancies but also rank among the wealthiest in terms of GDP per capita, suggesting a possible connection between economic prosperity and longer life spans.





2.3.2 Top 10 Countries for Life Expectancy & Health Expenditure

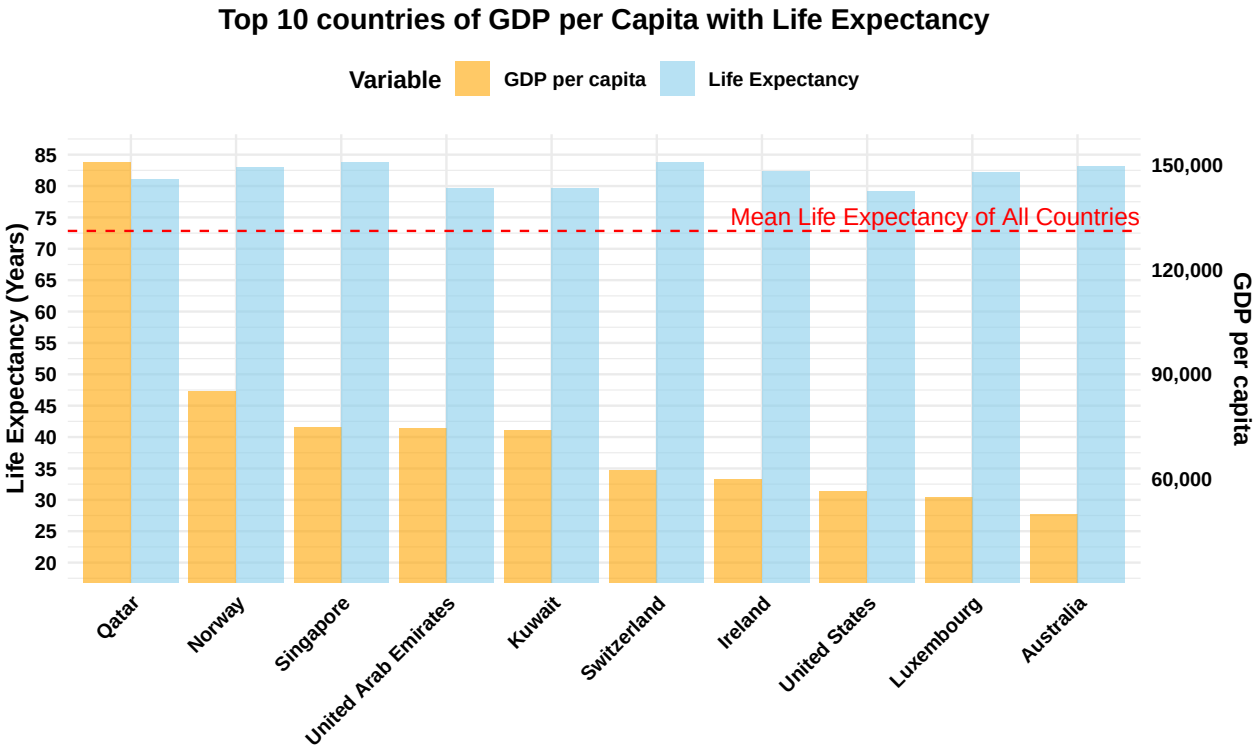
Among the 153 countries, Norway, Switzerland, and Sweden are present in both the top 10 for life expectancy and health expenditure, making up about 2% of all nations. These countries excel in both areas, suggesting a connection between higher healthcare spending and longer life spans. This pattern emphasizes the potential role of significant health investments in promoting longevity in these nations.



2.4 Clustered Bar Charts

2.4.1 GDP per capita & Life Expectancy

The analysis focuses on the top 10 countries with the highest GDP per capita and their corresponding life expectancy values, visualized in a dual Y-axis bar graph. All top 10 countries exceed the overall mean life expectancy, indicating that higher GDP per capita is associated with longer life spans. For instance, Qatar (around 82 years) and Norway (approximately 84 years). The mean life expectancy of the top 10 countries by GDP per capita is 81.76 years, significantly higher than the overall mean life expectancy of 72.85 years across all countries. This difference highlights the potential impact of economic factors on health outcomes.



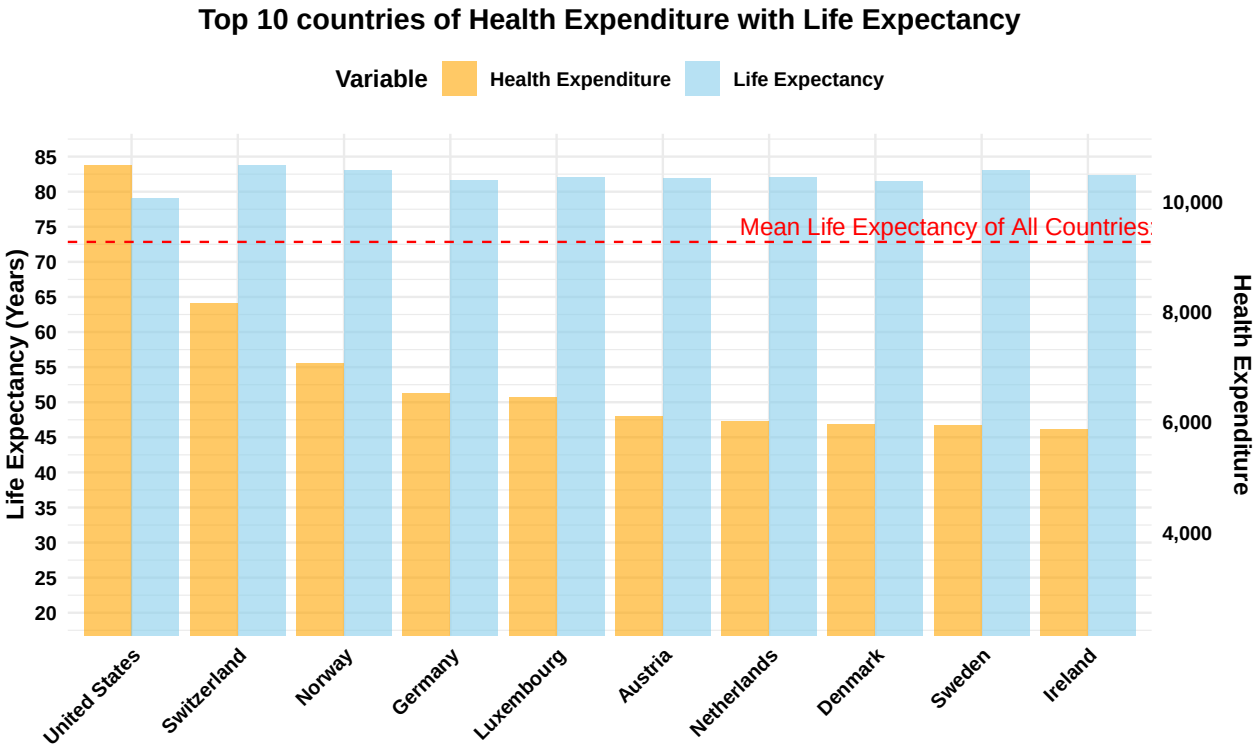

```
## [1] "Mean Life Expectancy by Top 10 Countries of GDP per Capita: 81.76"
```

```
## [1] "Mean Life Expectancy of All Countries: 72.85"
```

2.4.2 Health Expenditure & Life Expectancy

This analysis focuses on the relationship between health expenditure and life expectancy in the top 10 countries by health expenditure. Notably, the United States, despite its high spending, reflects potential inefficiencies in healthcare delivery that may negatively impact health outcomes. In contrast, countries like Switzerland, Norway, and several others illustrate the benefits of significant health investments, resulting in high life expectancy rates.

The mean life expectancy for the top 10 countries with the highest health expenditure is 82.03 years, which is markedly higher than the overall average of 72.85 years for all countries. This analysis reveals a clear trend: countries with higher health expenditure generally tend to have higher life expectancy, highlighting the importance of effective healthcare funding in improving population health.




```
## [1] "Mean Life Expectancy by Top 10 Countries of Health Expenditure:  82.03"
```

```
## [1] "Mean Life Expectancy of All Countries: 72.85"
```

2.5 Box Plot Analysis

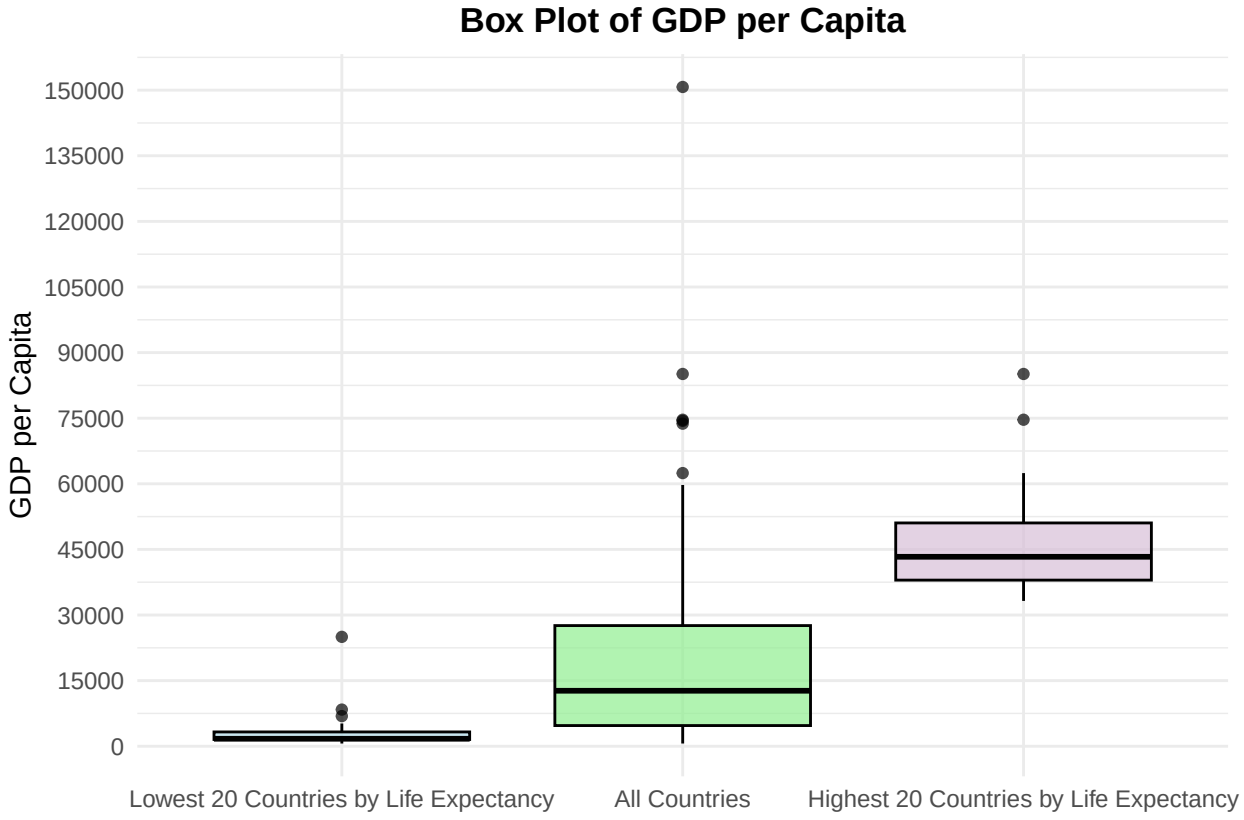
These visualizations explore how GDP per capita and health expenditure vary among three groups based on life expectancy: the Top 20 countries, the all countries, and the lowest 20 countries. By examining these distributions, we aim to gain insights into the relationship between life expectancy, economic prosperity (measured by GDP), and health spending. This analysis seeks to uncover possible connections between higher life expectancy and greater investments in both the economy and healthcare.

2.5.1 Box Plot of GDP per capita

The box plot below reveals significant disparities in GDP per capita among the three groups. The Top 20 Life Expectancy countries exhibit a notably higher median GDP per capita compared to the Overall and 20 lowest Life Expectancy groups. This indicates a strong association between economic prosperity and longevity.

In contrast, the 20 lowest Life Expectancy countries display a very low median GDP per capita, with many outliers suggesting a wider variability in this group. The Overall group sits in between, indicating a mix of countries with diverse economic conditions.

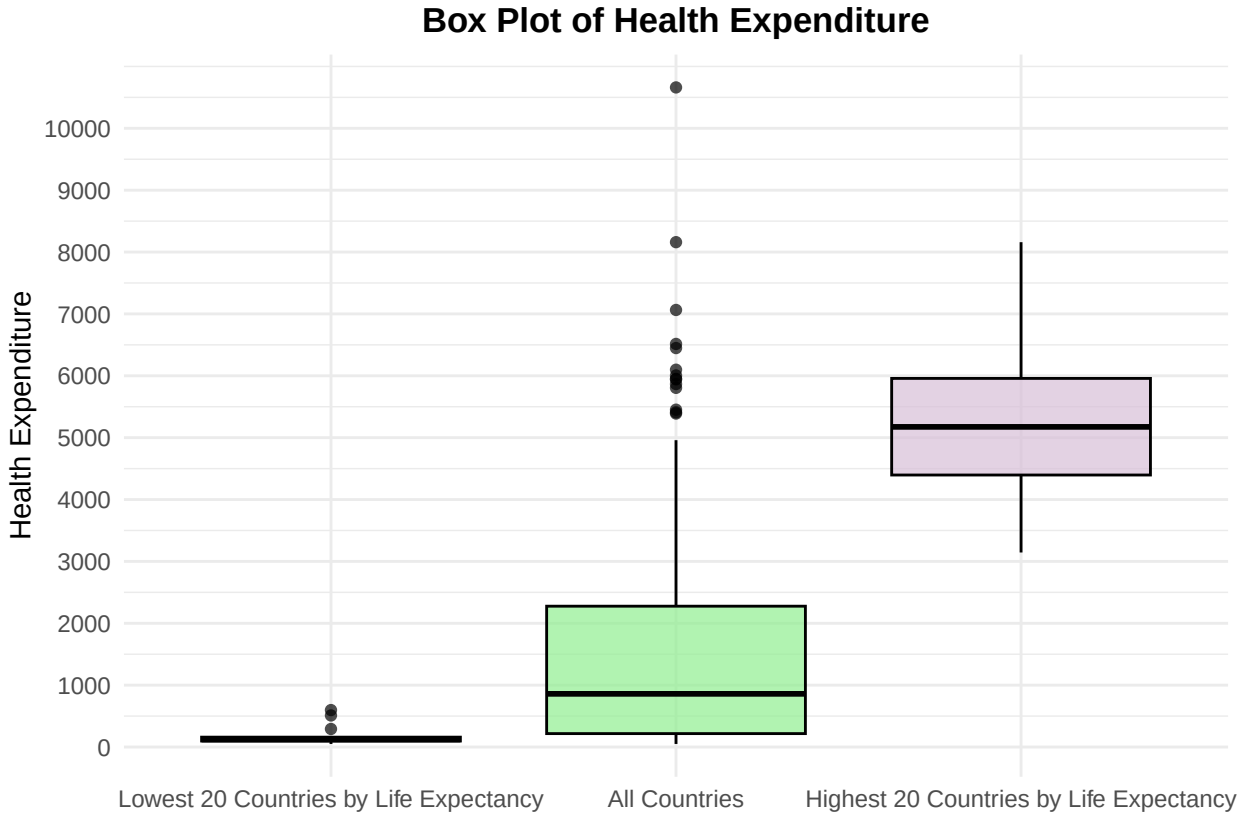
This visualization suggests that GDP per capita may be a key indicator of life expectancy. Countries with lower economic indicators often experience reduced life expectancies, highlighting the impact of economic factors on health outcomes.



2.5.2 Box Plot of Health Expenditure

The box plot of health expenditure across different life expectancy groups reveals notable differences in spending. The 20 lowest Life Expectancy group shows critically low health expenditure, often below $1,000 per capita.

In contrast, the Top 20 Life Expectancy countries group demonstrates significantly higher median health expenditure, typically exceeding $5,000 per capita, with a minimum of $3,000 and a maximum of $8,000. This indicates that countries with better health outcomes prioritize health financing. Given that the minimum health expenditure in the Top 20 Life Expectancy group surpasses the maximum in the Bottom 20 Life Expectancy group, it highlights the extent to which higher life expectancy correlates with substantial investment in health care.



3 Hypothesis Testing

3.1 Focus Questions

1. What is the effect of GDP per capita on average life expectancy, and does this effect vary by continent?
2. What is the effect of health expenditure on average life expectancy, and does this effect vary by continent?

3.2 Why Analysis of Covariance (ANCOVA)?

ANCOVA allows comparison across multiple groups, reducing the risk of Type I errors that occur with multiple t-tests. It explains the variance in data, can include covariates, and is ideal for continuous outcomes like life expectancy. Additionally, post-hoc tests (e.g., Tukey's honestly significant difference test) offer detailed pairwise comparisons when significant differences are found.

3.3 Selection of Variables

The first model utilizes GDP per capita as the continuous independent variable and continent as the categorical independent variable, predicting the dependent variable, average life expectancy at birth.

In the second model, health expenditure serves as the continuous independent variable, alongside continent as the categorical independent variable, with the dependent variable being average life expectancy at birth.

3.4 Initial Hypothesis

3.4.1 Model I

- Null Hypothesis (H_0): There is no relationship between continent and life expectancy, controlling for GDP per capita.
- Alternative Hypothesis (H_1): There is a significant relationship between continent and life expectancy, controlling for GDP per capita.

3.4.2 Model II

- Null Hypothesis (H_0): There is no relationship between continent and life expectancy, controlling for health expenditure.
- Alternative Hypothesis (H_1): There is a significant relationship between continent and life expectancy, controlling for health expenditure.

3.5 Testing Both Models in R

3.5.1 Model I

```
library(dplyr)
dat <- read.csv("dataset_life_expectancy_2019.csv")
anova_gdp_continent <- aov(Period.life.expectancy.at.birth...Sex..all...Age..0 ~
                           GDP.per.capita * Continent, data = dat)
summary(anova_gdp_continent)
```

```
##                            Df Sum Sq Mean Sq F value  Pr(>F)
## GDP.per.capita              1   4124    4124 279.183 < 2e-16 ***
## Continent                   5   2539     508  34.372 < 2e-16 ***
## GDP.per.capita:Continent    5    261      52   3.536 0.00483 **
## Residuals                 141   2083      15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
tukey_result <- TukeyHSD(anova_gdp_continent, "Continent")
print(tukey_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Period.life.expectancy.at.birth...Sex..all...Age..0 ~ GDP.per.capita * Contir
##
## $Continent
##                                     diff        lwr       upr     p adj
## Asia-Africa                    6.0311657  3.6257733  8.436558 0.0000000
## Europe-Africa                  8.3518822  5.9292545 10.774510 0.0000000
## North America-Africa           7.9759814  4.8330527 11.118910 0.0000000
## Oceania-Africa                 9.2111805  1.1934389 17.228922 0.0143262
## South America-Africa           9.3693123  5.5020250 13.236600 0.0000000
## Europe-Asia                    2.3207164 -0.2105578  4.851991 0.0924938
## North America-Asia             1.9448157 -1.2826021  5.172234 0.5072061
## Oceania-Asia                   3.1800147 -4.8712215 11.231251 0.8632761
## South America-Asia             3.3381466 -0.5981125  7.274406 0.1464540
## North America-Europe          -0.3759008 -3.6161844  2.864383 0.9994297
## Oceania-Europe                 0.8592983 -7.1971040  8.915701 0.9996216
## South America-Europe           1.0174302 -2.9293848  4.964245 0.9758798
## Oceania-North America          1.2351990 -7.0662763  9.536674 0.9981003
## South America-North America    1.3933309 -3.0323017  5.818964 0.9434935
## South America-Oceania          0.1581319 -8.4437518  8.760016 0.9999999
```

3.5.2 Model II

```r
anova_HE_continent <- aov(Period.life.expectancy.at.birth...Sex..all...Age..0 ~
                            Health.Expenditure * Continent, data = dat)
summary(anova_HE_continent)
```

```
##                                Df Sum Sq Mean Sq F value   Pr(>F)
## Health.Expenditure              1   4710    4710 394.662  < 2e-16 ***
## Continent                       5   2038     408  34.152  < 2e-16 ***
## Health.Expenditure:Continent    5    576     115   9.652 6.07e-08 ***
## Residuals                     141   1683      12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
tukey_result <- TukeyHSD(anova_HE_continent, "Continent")
print(tukey_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Period.life.expectancy.at.birth...Sex..all...Age..0 ~ Health.Expenditure * Co
##
## $Continent
##                                     diff       lwr       upr     p adj
## Asia-Africa                   8.00117085  5.839095 10.163247 0.0000000
## Europe-Africa                 5.91878923  3.741221  8.096357 0.0000000
## North America-Africa          6.18527577  3.360268  9.010283 0.0000000
## Oceania-Africa                5.89406669 -1.312645 13.100778 0.1765867
## South America-Africa          8.39015669  4.914063 11.866251 0.0000000
## Europe-Asia                  -2.08238162 -4.357606  0.192843 0.0935201
```

```
## North America-Asia          -1.81589508 -4.716845  1.085055 0.4635900
## Oceania-Asia                -2.10710416 -9.343922  5.129714 0.9592162
## South America-Asia           0.38898584 -3.149103  3.927075 0.9995611
## North America-Europe         0.26648654 -2.646028  3.179001 0.9998217
## Oceania-Europe              -0.02472254 -7.266184  7.216739 1.0000000
## South America-Europe         2.47136746 -1.076210  6.018944 0.3403423
## Oceania-North America       -0.29120908 -7.752953  7.170535 0.9999974
## South America-North America  2.20488092 -1.773079  6.182841 0.5989728
## South America-Oceania        2.49609000 -5.235675 10.227855 0.9373843
```

3.6 Results & Interpretation

3.6.1 Model I

Analysis of Covariance (ANCOVA):

The ANCOVA results show that there is a relationship between continent and life expectancy, controlling for GDP per capita. The mean life expectancy in some continents is significantly different than others. For instance, countries with higher GDP per capita tend to have higher life expectancy (as we saw in our exploratory data analysis), but the effect of GDP per capita on life expectancy varies across continents, indicating that the relationship between GDP per capita and life expectancy is not the same for all continents.

Tukey's Honestly Significant Difference (HSD) Test:

The Tukey HSD results indicate significant differences in life expectancy between several continent pairs, particularly between Africa and Asia, Europe, North America, Oceania, and South America, all with p-values less than 0.05.

3.6.2 Model II

Analysis of Covariance (ANCOVA):

The ANCOVA results indicate that there is a relationship between continent and life expectancy, controlling for health expenditure. The mean life expectancy varies significantly between continents. For instance, countries with higher health expenditure generally have higher life expectancy (as we saw in our exploratory data analysis), but the effect of health expenditure on life expectancy is not uniform across all continents. This suggests that the relationship between health expenditure and life expectancy differs depending on the continent.

Tukey's Honestly Significant Difference (HSD) Test:

The Tukey HSD results show significant differences in life expectancy between several continent pairs, particularly between Africa and Asia, Europe, North America, and South America, all with p-values less than 0.05.

4 Regression Analysis

4.1 Selection of Variables

From the analysis in the EDA section, we identified GDP per capita and health expenditure as potential influence factors of life expectancy. In this section, we are going the examine this further, deciding which factor shows a more direct and prominent correlation with life expectancy.

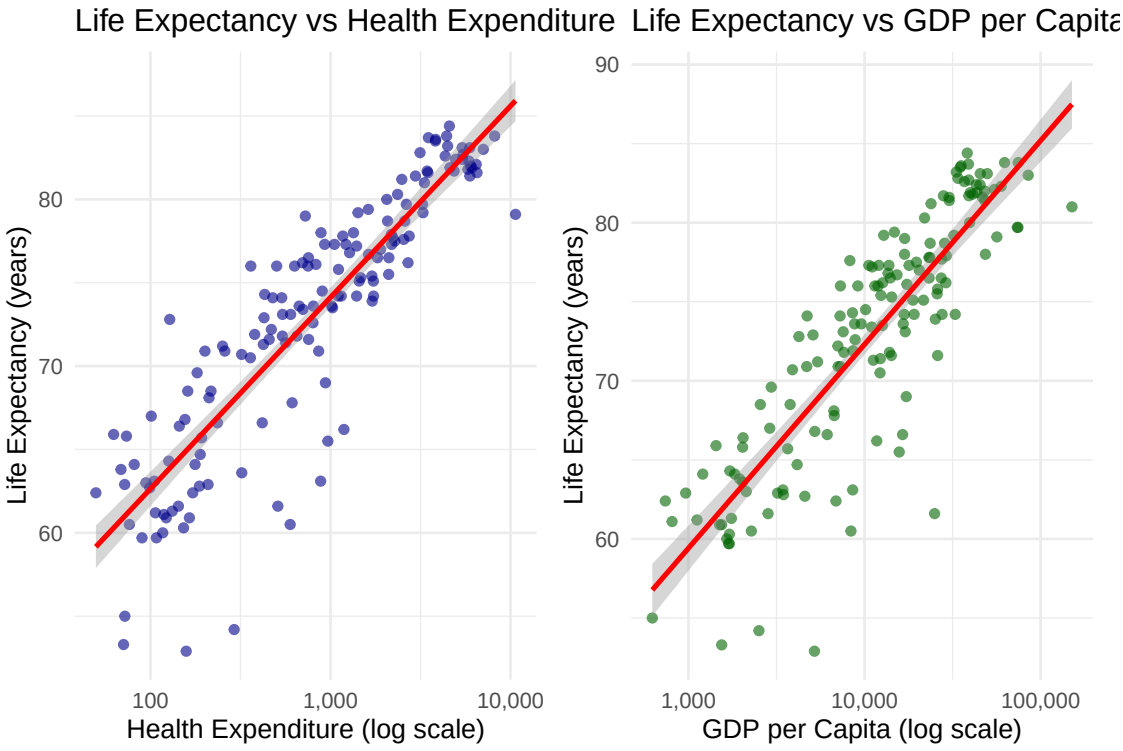
It is worth noting that, as health expenditure and GDP span a wide range of values, a log transformation has been applied to these values. Without a log transformation, the larger values would dominate the plot, and the smaller values would be compressed, making it difficult to discover meaningful patterns, especially for countries with lower health expenditure and/or GDP. By applying a logarithmic scale, we compress the larger values and spread out the smaller ones, providing a clearer view of the relationship across all data points.

4.2 Scatter Plot Analysis

The two scatter plots below compare the relationship between life expectancy and the two factors/predictors: Health expenditure and GDP, both on a logarithmic scale.

In both plots, a positive linear relationship is observed, indicating when either health expenditure or GDP increase, life expectancy also tends to increase.

```
# Scatter plot: Life Expectancy vs Health Expenditure
plot_health <- ggplot(data, aes(x = health_expense, y = life_expectancy)) +
  geom_point(alpha = 0.6, color = "darkblue") +
  geom_smooth(method = "lm", color = "red") +
  scale_x_log10(labels = scales::comma) +  # Log scale to handle skewness
  labs(title = "Life Expectancy vs Health Expenditure",
       x = "Health Expenditure (log scale)",
       y = "Life Expectancy (years)") +
  theme_minimal()
# Scatter plot: Life Expectancy vs GDP
plot_gdp <- ggplot(data, aes(x = gdp_per_capita, y = life_expectancy)) +
  geom_point(alpha = 0.6, color = "darkgreen") +
  geom_smooth(method = "lm", color = "red") +
  scale_x_log10(labels = scales::comma) +  # Log scale to handle skewness
  labs(title = "Life Expectancy vs GDP per Capita",
       x = "GDP per Capita (log scale)",
       y = "Life Expectancy (years)") +
  theme_minimal()
# Combine the two plots side by side
plot_health + plot_gdp
```



Comparing the two plots above, we can observe that the blue dots in the health expenditure plot are scattered slightly closer to the regression line. This indicates that the model based on health expenditure is more accurate in predicting life expectancy compared to the model based on GDP.

4.3 Correlation Analysis

The correlation analysis examines the relationship between life expectancy and two key factors/predictors: Health expenditure and GDP. The results are as follows:

- Correlation between Life Expectancy and `log10(Health Expenditure)`: 0.887
- Correlation between Life Expectancy and `log10(GDP)`: 0.861

Both predictors show a strong positive relationship with life expectancy. Specifically, the correlation between life expectancy and health expenditure is 0.887, while the correlation with GDP is 0.861. This suggests that, although both variables are closely related with life expectancy, health expenditure shows a slightly stronger correlation, making it a more accurate predictor of life expectancy compared to GDP.

```
# Correlation between log10(Health Expenditure) and Life Expectancy
cor_log_health_life <- cor(log10(data$health_expense), data$life_expectancy,
                           use = "complete.obs")
cat("Correlation between log10(Health Expenditure) and Life Expectancy:",
    round(cor_log_health_life, 3), "\n")
```

```
## Correlation between log10(Health Expenditure) and Life Expectancy: 0.887
```

```
# Correlation between log10(GDP) and Life Expectancy
cor_log_gdp_life <- cor(log10(data$gdp_per_capita), data$life_expectancy,
                        use = "complete.obs")
cat("Correlation between log10(GDP) and Life Expectancy:",
    round(cor_log_gdp_life, 3),"\n")
```

```
## Correlation between log10(GDP) and Life Expectancy: 0.861
```

4.4 Variance Analysis (R-squared)

With regard to the variance analysis, we have obtained the results as follows:

- R-squared for Life Expectancy vs. Health Expenditure: 0.7875
- R-squared for Life Expectancy vs. GDP: 0.7405

The R-squared measures the proportion of variance of the dependency variable (i.e., Life expectancy) that can be explained by the independent variable(s) (i.e., Health expenditure or GDP). The results above show that health expenditure has a better prediction accuracy.

```
# R-squared value for Life Expectancy vs Health Expenditure:
model_health <- lm(life_expectancy ~ log10(health_expense), data = data)
r_squared_health <- round(summary(model_health)$r.squared, 4)
cat("R-squared for Life Expectancy vs Health Expenditure:", r_squared_health, "\n")
```

```
## R-squared for Life Expectancy vs Health Expenditure: 0.7875
```

```
# R-squared value for Life Expectancy vs GDP:
model_gdp <- lm(life_expectancy ~ log10(gdp_per_capita), data = data)
r_squared_gdp <- round(summary(model_gdp)$r.squared, 4)
cat("R-squared for Life Expectancy vs GDP:", r_squared_gdp, "\n")
```

```
## R-squared for Life Expectancy vs GDP: 0.7405
```

4.5 Practical Consideration

In addition to the statistical measures above, it is also important to consider the practical relevance of the variables. When analyzing life expectancy, the focus of our project, health expenditure directly affects the resources allocated to the healthcare system, which intuitively impacts life expectancy. Better access to high-quality healthcare services logically leads to better health outcomes and, therefore, longer life expectancy.

Thus, health expenditure is the more direct variable, compared to the GDP variable, in predicting life expectancy.

4.6 Conclusion

With the statistical measures and practical considerations established above, we can conclude that there is sufficient evidence to support the conclusion that there is a strong positive relationship between life expectancy and health expenditure.

The remaining sections will focus on building the linear regression model, model prediction, and model evaluation.

4.7 Regression Model

4.7.1 Regression Model Development

Using the simple linear regression approach, the relationship between life expectancy and health expense/expenditure can be expressed as follows:

$$Y = \beta_0 + \beta_1 X + e$$

where,

- Y: Dependent (response) variable (i.e., `life_expectancy`);
- X: Independent (predictor) variable (i.e., `log10(health_expense)`);
- β_0: Intercept (the value of Y when $X = 0$);
- β_1: Slope of the regression line; and
- e: Error (i.e, the difference between the observed values and predicted values).

The relationship can be further rewritten in the following *deterministic model*:

$$Y = \beta_0 + \beta_1 X$$

The `lm` function in R was then used to calculate the values of β_0 and β_1 (see below for detailed coding implementation):

- $\hat{\beta}_0 = 39.69$
- $\hat{\beta}_1 = 11.48$

Therefore, the linear regression model of life expectancy and health expenditure is as follows:

$$\hat{y} = 39.69 + 11.48X$$

where $\hat{y}$ = Life expectancy, and X = log(Health expenditure).

```
# Build a linear regression model using log-transformed health expenditure
model <- lm(life_expectancy ~ log10(health_expense), data = data)
# Extract model coefficients (intercept and slope)
intercept <- coef(model)[1]
slope <- coef(model)[2]
# Print the regression model equation in a readable format
cat("Regression Model Equation:\n", "Life Expectancy =", round(intercept, 2), "+",
    round(slope, 2), "* log(Health Expenditure)")
```

```
## Regression Model Equation:
##  Life Expectancy = 39.69 + 11.48 * log(Health Expenditure)
```

4.7.2 Regression Model Interpretation

Interpretation of $\hat{\beta}_0 = 39.69$:

- This coefficient represents the theoretical life expectancy when the log of health expenditure is zero. However, this value has little practical interpretation because it is outside the range of our sample data.

Interpretation of $\hat{\beta}_1 = 11.48$:

- This coefficient represents the change in life expectancy affected by the change in the log of health expenditure. More precisely, a 1% increase in health expenditure is associated with an increase in life expectancy of approximately 0.0499 years (i.e., $4.99 \times \log(1.01) = 0.0499$), or about 18 days.

4.7.3 Regression Model Predictions

With the linear regression model developed, we can make predictions about life expectancy based on a given health expenditure.

For example, consider the following scenarios:

- Country A: health expenditure = 1,000
- Country B: health expenditure = 5,000
- Country C: health expenditure = 10,000

The life expectancy of newborns in these countries can be estimated using the model developed in the previous section (refer to the code below for calculation details):

- In Country A, with a health expenditure of 1,000, the estimated life expectancy of a newborn is 74.12 years.
- In Country B, with a health expenditure of 5,000, the estimated life expectancy of a newborn is 82.15 years.
- In Country C, with a health expenditure of 10,000, the estimated life expectancy of a newborn is 85.60 years.

It is also important to note that our regression model was developed based on a specific data set, which has a range of health expenditure between $49.79 and $10,661.03. Therefore, it should only be used to predict the life expectancy of countries with health expenditures within this range.

```
# Predict life expectancy for specific values of health expenditure
new_data <- data.frame(health_expense = c(1000, 5000, 10000))
# Predict life expectancy with confidence and prediction intervals
predictions <- predict(model, newdata = new_data, interval = "prediction")
# Print prediction results
cat("Predicted Life Expectancy (with Prediction Intervals):\n")
```

```
## Predicted Life Expectancy (with Prediction Intervals):
```

```
predictions
```

```
##        fit      lwr      upr
## 1 74.12189 67.06371 81.18007
## 2 82.14502 75.04503 89.24501
## 3 85.60040 78.46306 92.73773
```

4.7.4 Regression Model Evaluation

Model evaluation is an important step to ensure the reliability and accuracy of our regression model. Since the coefficients ($\hat{\beta}_0$ and $\hat{\beta}_1$) are estimated based on sample data, they are subject to variability.

By conducting thorough evaluations, we can verify the significance of our coefficient estimations and confirm whether our model meets crucial assumptions about the residuals, including their independence, normality, and homoscedasticity (equal variance).

Evaluation 1: Is there a linear relationship between Life Expectancy and Health Expenditure?

From the scatter plotted in section 4.2, we can confirmed that a strong positive linear relationship can be observed between life expectancy and health expenditure.

Evaluation 2: What is the strength of linear relationship?

From the correlation calculated in section 4.3, the correlation coefficient is 0.887, indicating a very strong positive correlation between life expectancy and health expenditure.

Evaluation 3: Are the coefficient estimates significant?

Our developed regression model is based on a specific set of data. If we were to obtain different samples from the same population, we might get different coefficients and thus a different regression model.

To determine whether the relationship we observed in the regression model is likely to exist in the population, we would need to conduct a hypothesis test:

- Null hypothesis (H_0): $\beta_1 = 0$
- Alternative hypothesis (H_A): $\beta_1 \neq 0$

Based on the calculations below (refer to the coding section below for details), the p-value is nearly zero, indicating strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude that the coefficient is statistically significant.

In summary, the result of the hypothesis test supports that there is a relationship between life expectancy and health expenditure.

```
model_summary <- summary(model)
p_value_health <- model_summary$coefficients[2, 4]
cat("P-value for log(Health Expenditure):",
    format(p_value_health, scientific = TRUE))
```

```
## P-value for log(Health Expenditure): 1.199119e-52
```

Evaluation 4: Are the assumptions of Independence and Equality of the variance of the residuals met?

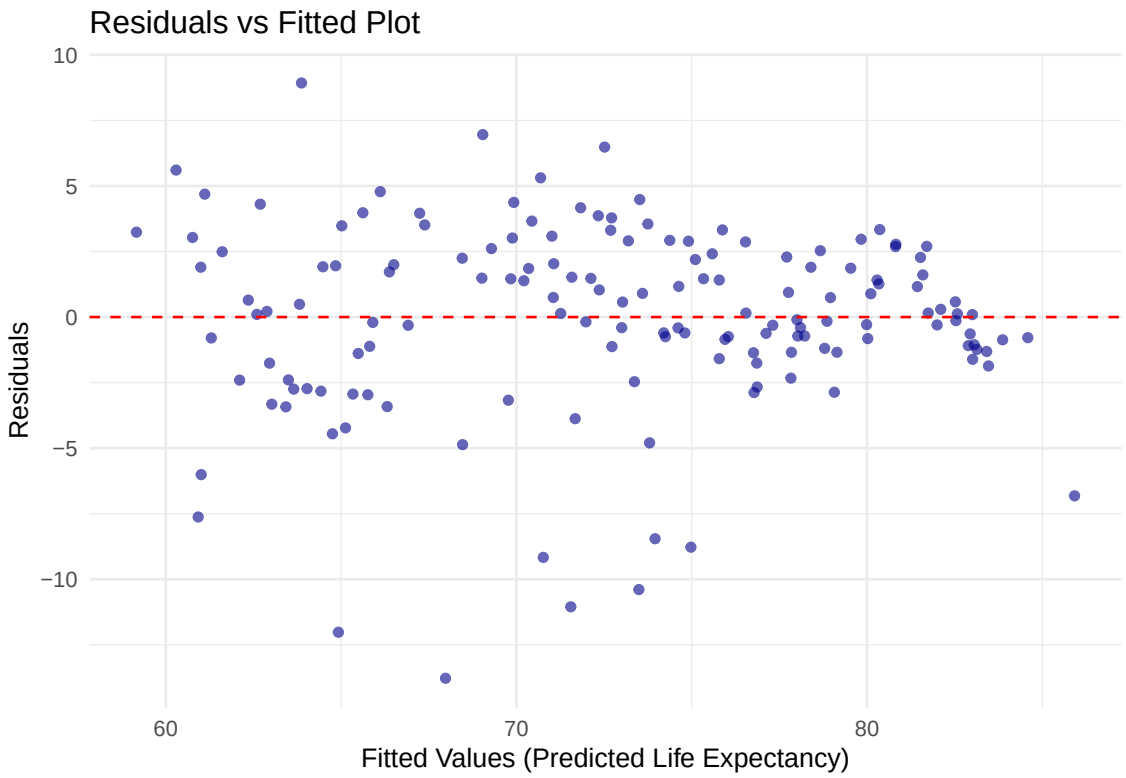
To ensure the residuals are independent and have equal variance, we would need to analyze its *Residuals vs Fitted plot*.

Referring to the plot below, we have the following observations:

- Independence:
  - The residuals are scattered evenly without showing a clear shape or clusters, suggesting that residuals are independent of each other.
- Homoscedasticity (equal variance):
  - The distribution of the residuals spreads evenly and randomly across the fitted values range, indicating a good sign of equal variance in residuals. It is worth mentioning that a slight tendency for the spread of residuals to decrease as the fitted value increase, indicating the our model might be less reliable for prediction with higher life expectancy.

```
# Residuals vs Fitted plot for testing independence and homoscedasticity (equal variance)
fitted_values_health <- fitted(model)
residuals_health <- resid(model)
ggplot(data.frame(fitted_values = fitted_values_health,
                  residuals = residuals_health),
       aes(x = fitted_values, y = residuals)) +
```

```
geom_point(alpha = 0.6, color = "darkblue") +
geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
labs(title = "Residuals vs Fitted Plot",
     x = "Fitted Values (Predicted Life Expectancy)",
     y = "Residuals") +
theme_minimal()
```



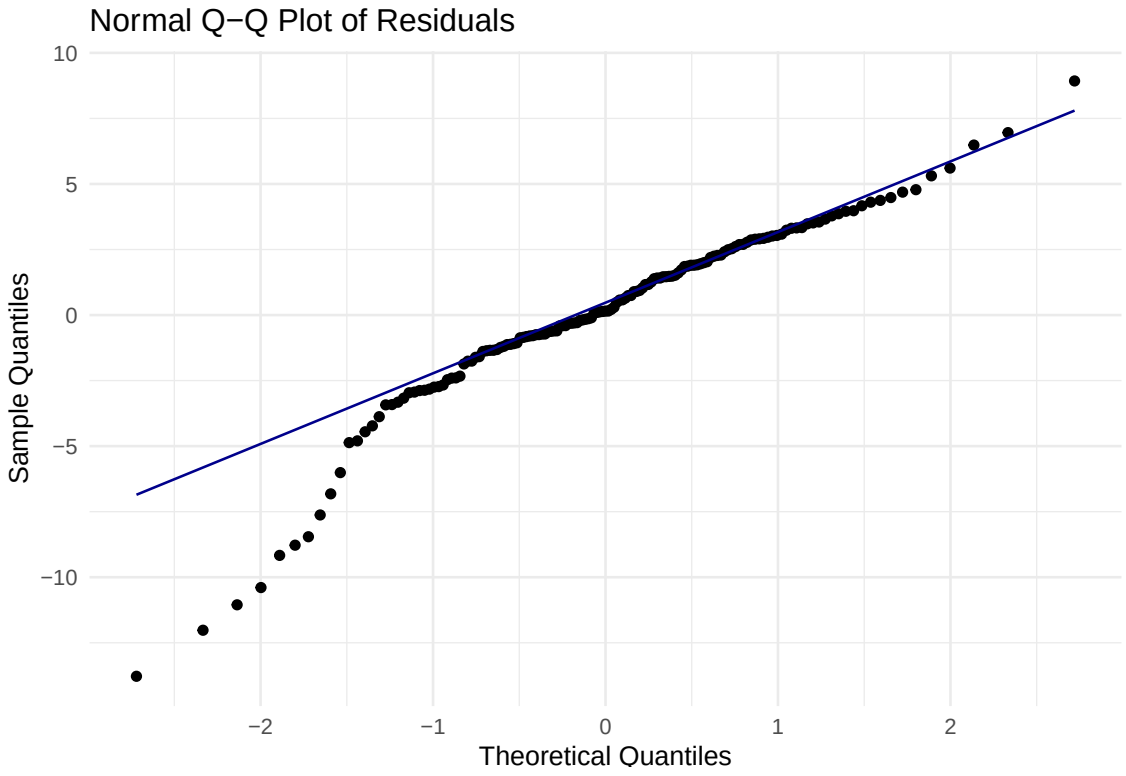
Evaluation 5: Is the assumption of Normality of the residuals met?

To ensure the residuals are normally distributed, we need to analyze its corresponding Q-Q plot.

Referring to the plot below, we have the following observations:

- Overall, the middle part of the plot aligns well with the reference line, indicating that residuals in that area are normally distributed. Therefore, our regression model should perform well on countries with typical medical expenditure.
- The lower tail at the left side shows significant deviation, indicating that some countries have much lower Life Expectancy than our model predicts based on their medical expenditure.
- The upper tail at the right side also shows certain degree of deviation, meaning that some countries have higher Life Expectancy that our model predicts.

```
# Q-Q Plot for testing normality of the residuals
ggplot(data.frame(residuals = residuals_health), aes(sample = residuals)) +
  stat_qq() + stat_qq_line(color = "darkblue") +
  labs(title = "Normal Q-Q Plot of Residuals",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_minimal()
```



Evaluation 6: How much of the variance in Life Expectancy can be explained by the Health Expenditure in our model?

To quantify the relationship between health expenditure and life expectancy, we would need to analyze the coefficient of determination (R-squared) from our regression model.

Based on the result calculated below, the R-squared value is 0.7875, indicating that approximately 78.75% of the variability in life expectancy can be explained by our model using log-transformed health expenditure. This relatively high R-squared value, suggests a strong relationship between the two variables. This implies that increases in health expenditure is strongly associated with increases in life expectancy.

```
model_summary <- summary(model)
r_squared <- model_summary$r.squared
cat("R-squared:", round(r_squared, 4), "\n")
```

```
## R-squared: 0.7875
```

5 Conclusion & Future Steps/Recommendations

5.1 Conclusion

In this project, we set out to explore the relationship between life expectancy and two key socioeconomic factors: Gross Domestic Product (GDP) per capita and health expenditure. Through exploratory data analysis, hypothesis testing, and regression analysis, we aimed to statistically analyze whether these factors have a significant impact on life expectancy across different countries and continents.

Our findings indicated a clear positive relationship between both GDP per capita and health expenditure with life expectancy. However, the strength of this relationship varied between the two factors. Health expenditure demonstrated a slightly stronger correlation with life expectancy than GDP, indicating that countries that invested more in healthcare tended to have better health outcomes, leading to longer life expectancy. This was further supported by the regression analysis, where health expenditure captured approximately 78.75% of the variance in life expectancy, compared to 74.05% for GDP.

5.2 Future Steps/Recommendations

While our analysis provided valuable insights, there are several that could be further explored to improve the accuracy and applicability of our findings:

- **Multivariate Regression Model**: In our project, a simple linear regression was used to analyze the relationship between life expectancy and health expenditure. To further improve the prediction accuracy, a multivariate regression model could provide a more comprehensive view considering additional variables such as environmental factors or eating habits.
- **Time-series Analysis**: Our analysis focused on data from a single year (2019), limiting our ability to observe trends over time. A time-series analysis of life expectancy, GDP, and health expenditure across multiple years would provide deeper insights into how these relationships evolve and whether specific events (e.g., pandemics) have long-term effects on life expectancy.

In conclusion, this project has demonstrated the significant role that economic and healthcare factors, such as GDP and health expenditure, play in influencing life expectancy. While our findings are based on a limited data set and scope, they provide valuable insights into these relationships. Future research could explore these factors more deeply, with the added potential of raising public awareness about the importance of healthcare investment and its direct impact on life expectancy.

References

Our World in Data. (2023). *Life expectancy vs. GDP per capita.* https://ourworldindata.org/grapher/life-expectancy-vs-gdp-per-capita

Our World in Data. (2024). *Life expectancy vs. health expenditure.* https://ourworldindata.org/grapher/life-expectancy-vs-health-expenditure