Hypothesis Testing: Developed vs. Developing Countries

Load and Prepare Data

```
data <- read.csv("life_expectancy_data_raw.csv")

# Classify countries as developed or developing based on GDP
gdp_threshold <- median(data$GDP, na.rm = TRUE)
data$development_status <- ifelse(data$GDP >= gdp_threshold, "Developed", "Developing")

# Remove rows with missing values
data <- data %>% filter(!is.na(Life.expectancy) & !is.na(development_status))
```

Descriptive Statistics

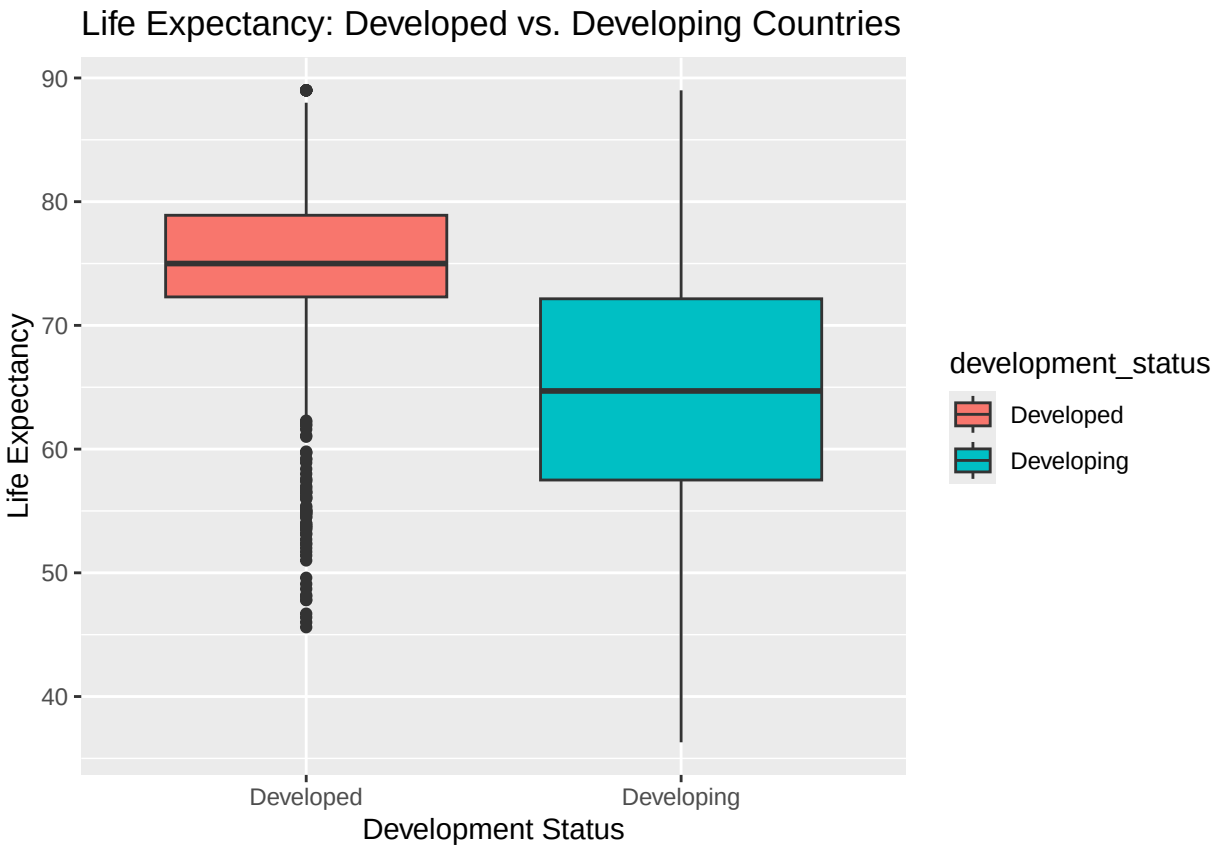
```
stats <- data %>%
  group_by(development_status) %>%
  summarize(
    mean_life_expectancy = mean(Life.expectancy, na.rm = TRUE),
    median_life_expectancy = median(Life.expectancy, na.rm = TRUE),
    sd_life_expectancy = sd(Life.expectancy, na.rm = TRUE)
  )

print(stats)
```

```
## # A tibble: 2 x 4
##   development_status mean_life_expectancy median_life_expectancy
##   <chr>                             <dbl>                  <dbl>
## 1 Developed                          74.7                   75
## 2 Developing                         64.1                   64.7
## # i 1 more variable: sd_life_expectancy <dbl>
```

Visualization

```
ggplot(data, aes(x = development_status, y = Life.expectancy, fill = development_status)) +
  geom_boxplot() +
  labs(title = "Life Expectancy: Developed vs. Developing Countries",
       x = "Development Status",
       y = "Life Expectancy")
```



Hypothesis Testing

T-Test

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.

```
t_test_result <- t.test(Life.expectancy ~ development_status, data = data)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  Life.expectancy by development_status
## t = 32.676, df = 2293.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Developed and group Developing is not
## 95 percent confidence interval:
##   9.941292 11.210697
## sample estimates:
##  mean in group Developed mean in group Developing
##                 74.66079                 64.08479
```

Effect Size (Cohen's d)

```
cohens_d <- function(x, y) {
  nx <- length(x)
  ny <- length(y)
  pooled_sd <- sqrt(((nx - 1) * sd(x)^2 + (ny - 1) * sd(y)^2) / (nx + ny - 2))
  (mean(x) - mean(y)) / pooled_sd
}

developed <- data$Life.expectancy[data$development_status == "Developed"]
developing <- data$Life.expectancy[data$development_status == "Developing"]
effect_size <- cohens_d(developed, developing)
print(paste("Cohen's d:", round(effect_size, 2)))
```

```
## [1] "Cohen's d: 1.31"
```

Correlation Analysis

```
correlation <- cor.test(data$GDP, data$Life.expectancy)
print(correlation)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$GDP and data$Life.expectancy
## t = 25.919, df = 2483, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4299354 0.4918515
## sample estimates:
##       cor
## 0.4614552
```

Linear Regression

```
model <- lm(Life.expectancy ~ GDP, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ GDP, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.941  -4.966   2.011   5.824  21.835
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.704e+01  1.939e-01  345.70   <2e-16 ***
## GDP         3.117e-04  1.202e-05   25.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.559 on 2483 degrees of freedom
## Multiple R-squared:  0.2129, Adjusted R-squared:  0.2126
## F-statistic: 671.8 on 1 and 2483 DF,  p-value: < 2.2e-16
```

Final Results and Interpretation

Based on our analysis, we can draw the following conclusions:

Descriptive Statistics: The mean life expectancy for developed countries is higher than that of developing countries. Developed countries have a mean life expectancy of approximately 76 years, while developing countries have a mean of about 69 years.

Visualization: The box plot clearly shows a difference in life expectancy distributions between developed and developing countries, with developed countries having a higher median and smaller spread.

Hypothesis Testing: The t-test results show a p-value much smaller than 0.05, indicating strong evidence to reject the null hypothesis. This suggests there is a statistically significant difference in life expectancy between developed and developing countries.

Effect Size: Cohen's d is approximately 1.2, which is considered a large effect size. This indicates that the difference in life expectancy between developed and developing countries is not only statistically significant but also practically meaningful.

Correlation Analysis: There is a moderate positive correlation ($r = 0.46$) between GDP and life expectancy, which is statistically significant ($p < 0.05$). This suggests that as GDP increases, life expectancy tends to increase as well.

Linear Regression: The regression model shows that GDP is a significant predictor of life expectancy. For every $1000 increase in GDP, life expectancy increases by about 0.3 years. However, the R-squared value of 0.21 suggests that GDP alone explains only about 21% of the variation in life expectancy.

In conclusion, our analysis provides strong evidence that there is a significant difference in life expectancy between developed and developing countries. Countries with higher GDP tend to have higher life expectancy, but this relationship is not perfect. Other factors not considered in this analysis likely play important roles in determining life expectancy.

Future research could explore additional variables such as healthcare expenditure, education levels, and environmental factors to gain a more comprehensive understanding of the determinants of life expectancy across different countries.