Lab2

Raj Shah

2025-02-24

```r
# Load necessary libraries
library(tidyverse)
```

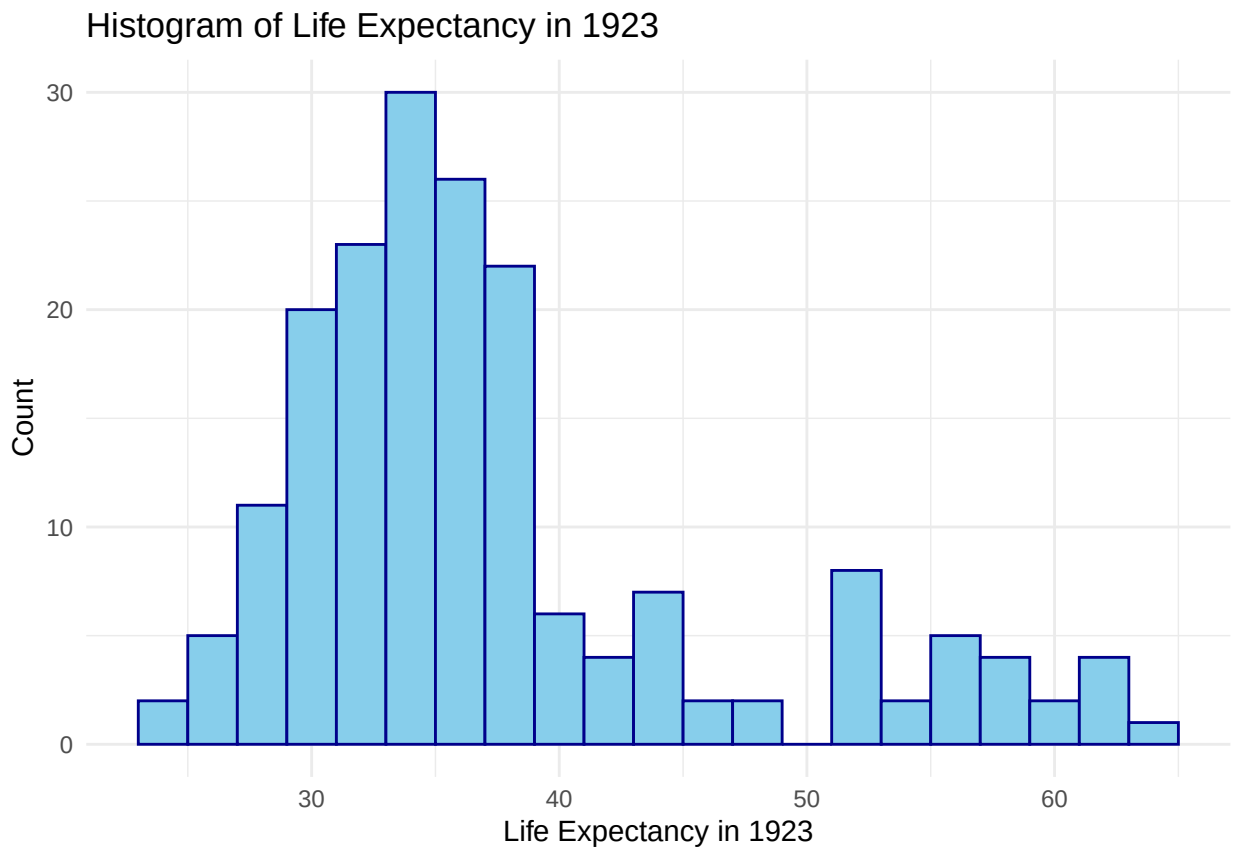
```
## Warning: package 'tidyverse' was built under R version 4.4.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

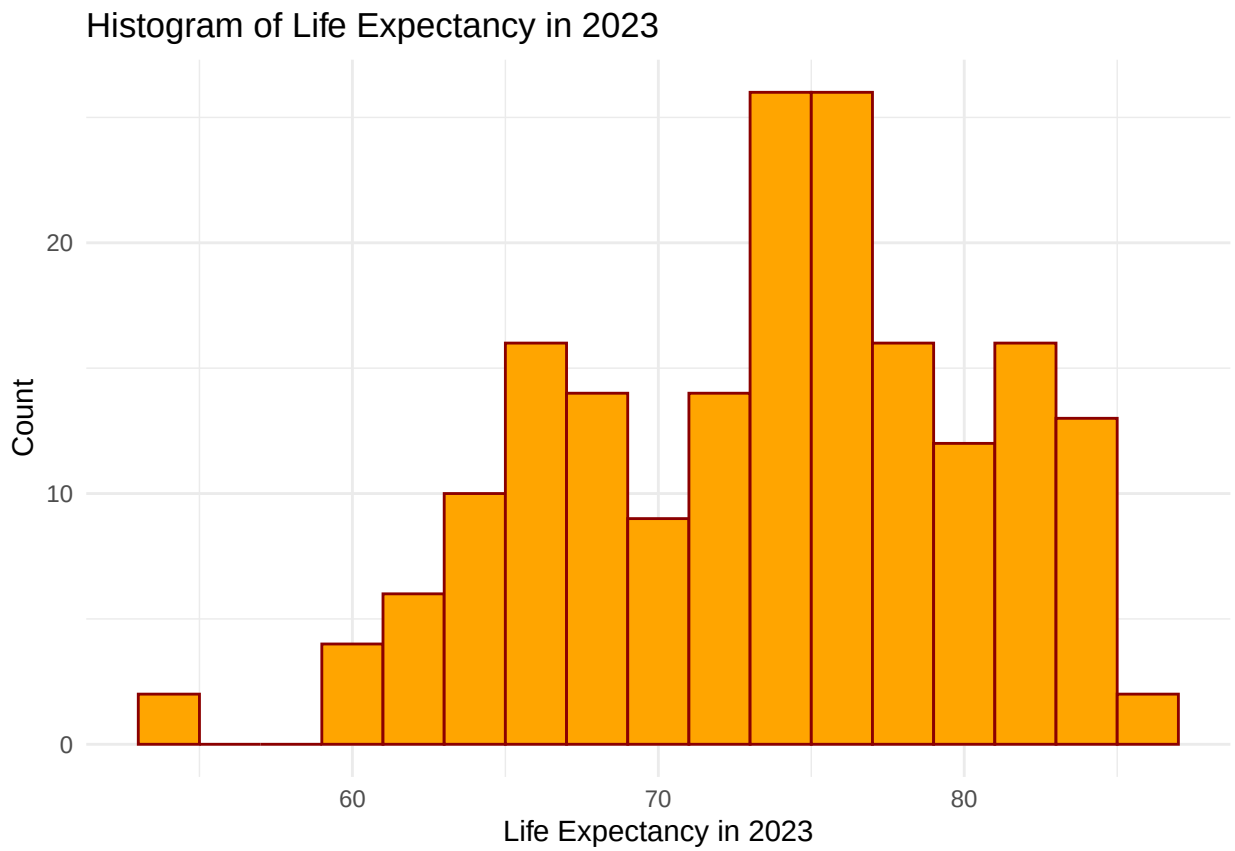
```r
library(ggplot2)
library(countrycode)
```

```
## Warning: package 'countrycode' was built under R version 4.4.2
```

```r
# Read the dataset
df <- read.csv("C:/Users/rajsh/OneDrive/Desktop/Inference Data Science 291/LAB2/Worldlife_cleaned.csv")

# Assign unique colors for each continent
continent_colors <- c("Africa" = "red", "Americas" = "blue", "Asia" = "purple", "Europe" = "green", "Oc

#### Part 1: Regression of Life Expectancy in 2023 on Life Expectancy in 1923 ####

# Question 1: Histograms of Life Expectancy in 1923 and 2023
ggplot(df, aes(x = life1923)) +
  geom_histogram(binwidth = 2, fill = "skyblue", color = "darkblue", alpha = 1.0) +
  ggtitle("Histogram of Life Expectancy in 1923") +
  xlab("Life Expectancy in 1923") +
  ylab("Count") +
  theme_minimal()
```

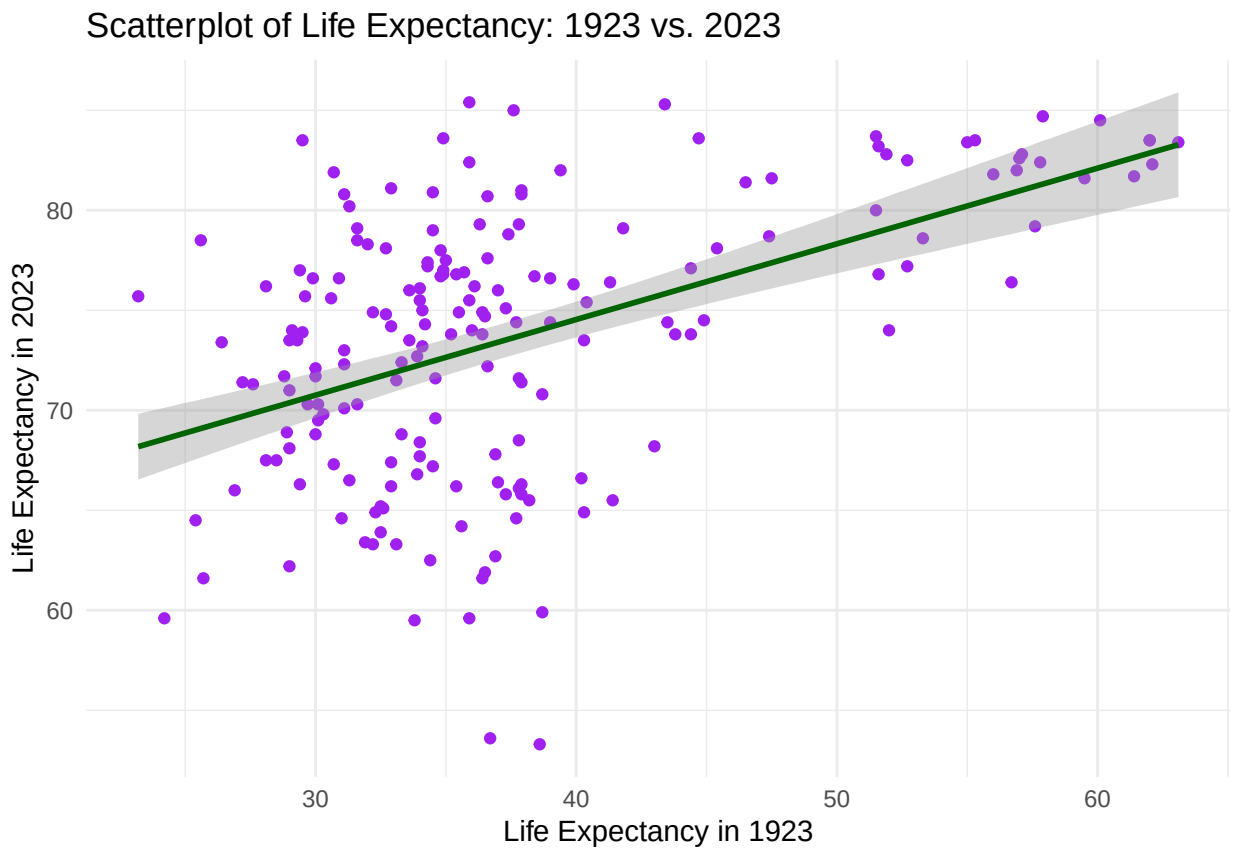



```
ggplot(df, aes(x = life2023)) +
  geom_histogram(binwidth = 2, fill = "orange", color = "darkred", alpha = 1.0) +
  ggtitle("Histogram of Life Expectancy in 2023") +
  xlab("Life Expectancy in 2023") +
  ylab("Count") +
  theme_minimal()
```




```
# Explanation:
# - The histogram shows that life expectancy in 1923 was mostly between 30-40 years,
#   while in 2023 it shifted to 60-80 years.
# - Advances in healthcare and economic development contributed to this change.

# Question 2: Scatterplot of Life Expectancy in 1923 vs. 2023
ggplot(df, aes(x = life1923, y = life2023)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", color = "darkgreen") +
  ggtitle("Scatterplot of Life Expectancy: 1923 vs. 2023") +
  xlab("Life Expectancy in 1923") +
  ylab("Life Expectancy in 2023") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Explanation:
# - The scatterplot suggests a positive relationship between life expectancy in 1923 and 2023.
# - However, it is not perfectly linear, indicating the influence of other factors.

# Question 3: Correlation between Life Expectancy in 1923 and 2023
correlation <- cor(df$life1923, df$life2023, use = "complete.obs")
print(paste("Correlation between life expectancy in 1923 and 2023:", round(correlation, 3)))
```

```
## [1] "Correlation between life expectancy in 1923 and 2023: 0.493"
```

```
# Explanation:
# - The correlation is 0.493, suggesting a moderate positive relationship.

# Question 4: Simple Linear Regression
model <- lm(life2023 ~ life1923, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = life2023 ~ life1923, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.710  -3.687   1.020   3.961  12.933
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.40509    1.90355  31.208  < 2e-16 ***
## life1923     0.37837    0.04923   7.685 8.83e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.949 on 184 degrees of freedom
## Multiple R-squared:  0.243,  Adjusted R-squared:  0.2389
## F-statistic: 59.06 on 1 and 184 DF,  p-value: 8.834e-13
```

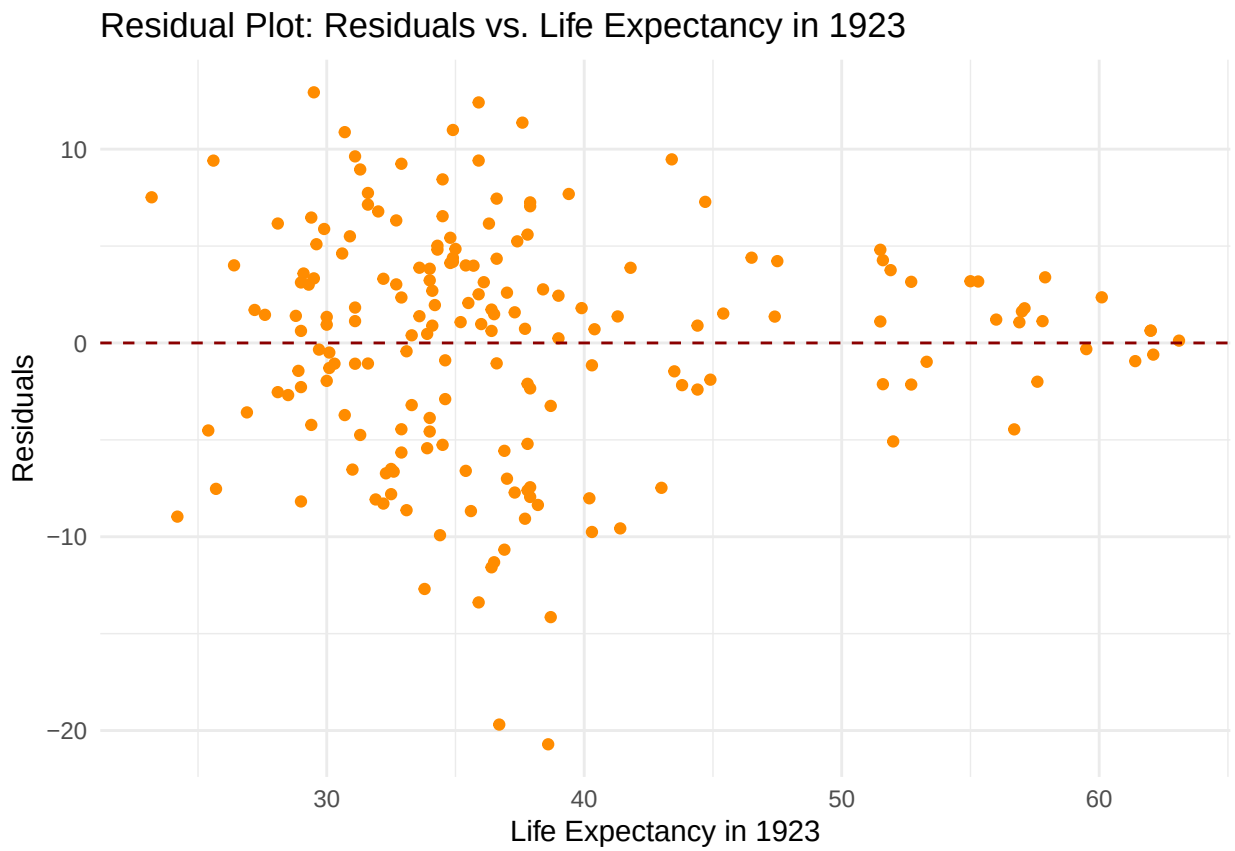
```
# Explanation:
# - The regression equation is: Life Expectancy in 2023 = 59.405 + 0.378 * Life Expectancy in 1923
# - The model is statistically significant, with an R-squared of 24.3%.

# Question 5: Expected Increase in Life Expectancy
expected_increase <- model$coefficients["life1923"]
print(paste("Expected increase in life expectancy for 1-year increase in 1923:", round(expected_increas
```

```
## [1] "Expected increase in life expectancy for 1-year increase in 1923: 0.378"
```

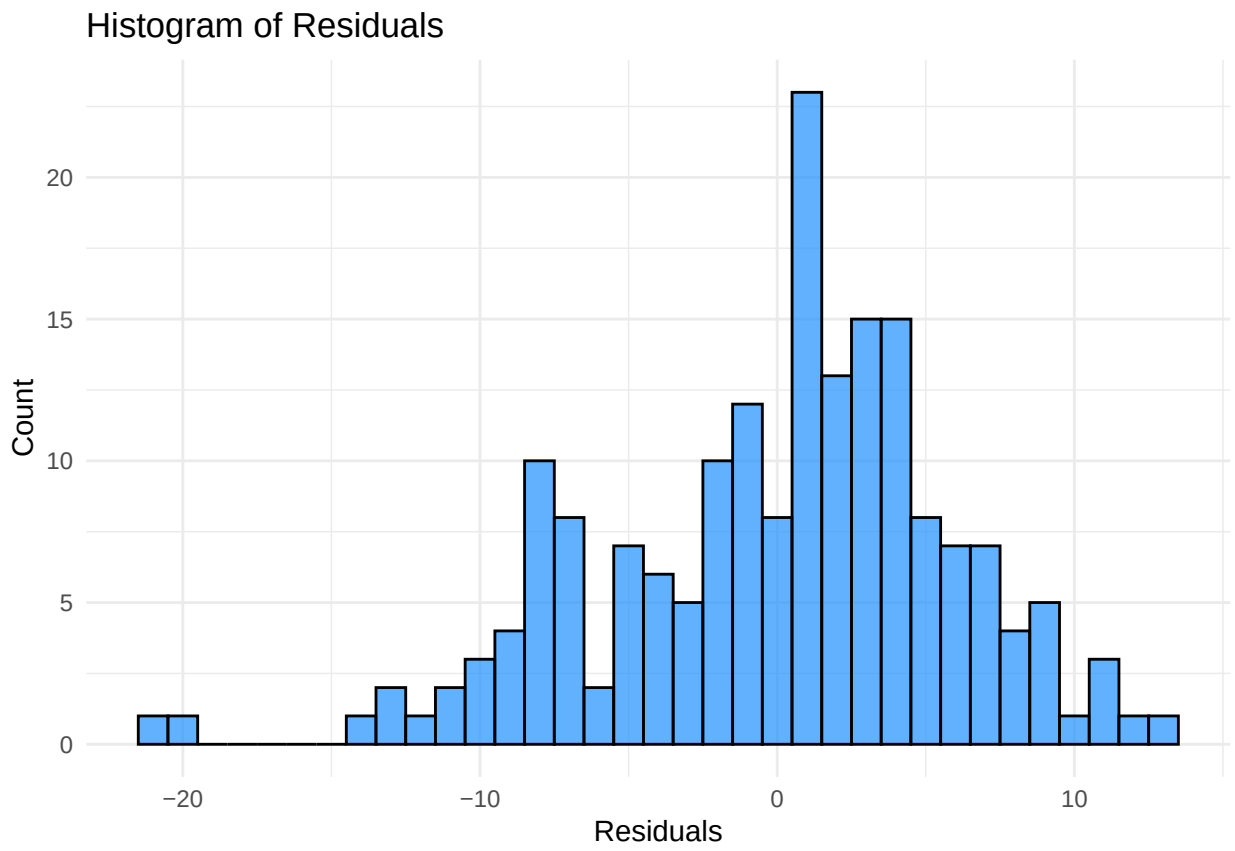
```
# Explanation:
# - A 1-year increase in 1923 is associated with a 0.378-year increase in 2023.

# Question 6: Predict Life Expectancy for Missing Data
life1923_value <- 34.3
predicted_life2023 <- predict(model, newdata = data.frame(life1923 = life1923_value))
print(paste("Predicted life expectancy in 2023:", round(predicted_life2023, 2)))
```

```
## [1] "Predicted life expectancy in 2023: 72.38"
```

```
# Explanation:
# - If a country had 34.3 years of life expectancy in 1923, it is predicted to have 72.38 years in 2023

# Question 7: Residual Plot and Histogram
df$residuals <- model$residuals
ggplot(df, aes(x = life1923, y = residuals)) +
  geom_point(color = "darkorange") +
  geom_hline(yintercept = 0, color = "darkred", linetype = "dashed") +
  ggtitle("Residual Plot: Residuals vs. Life Expectancy in 1923") +
  xlab("Life Expectancy in 1923") +
  ylab("Residuals") +
  theme_minimal()
```




```
ggplot(df, aes(x = residuals)) +
  geom_histogram(binwidth = 1, fill = "dodgerblue", color = "black", alpha = 0.7) +
  ggtitle("Histogram of Residuals") +
  xlab("Residuals") +
  ylab("Count") +
  theme_minimal()
```




```
# Explanation:
# - The residuals appear randomly scattered, suggesting a valid model.
# - The histogram shows an approximately normal distribution.

# Question 8: Variability Explained
r_squared <- summary(model)$r.squared
print(paste("Percentage of total variability explained:", round(r_squared * 100, 2), "%"))
```

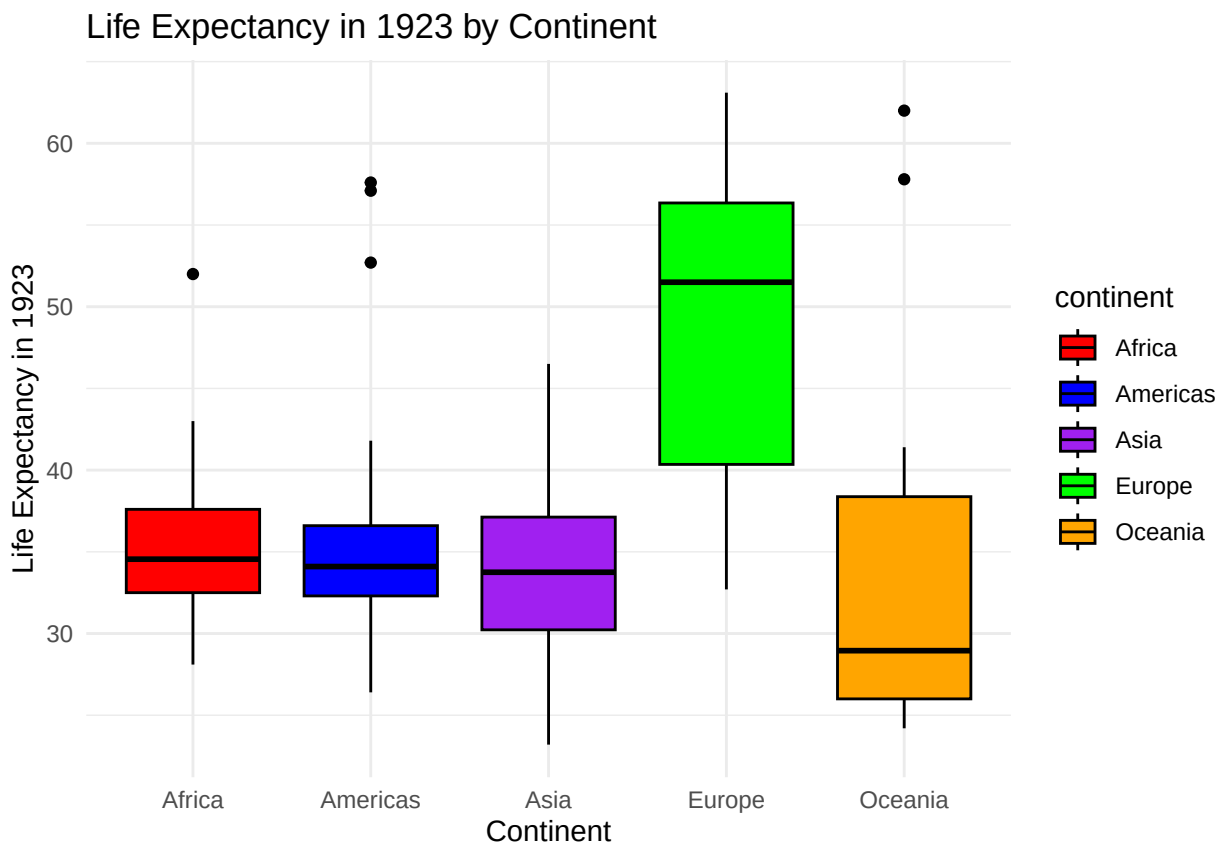
```
## [1] "Percentage of total variability explained: 24.3 %"
```

```
# Explanation:
# - The model explains 24.3% of the variation in life expectancy in 2023.

#### Part 2: Regression of Life Expectancy on Continent ####

# Question 1: Count of Countries per Continent
continent_counts <- df %>%
  group_by(continent) %>%
  summarise(num_countries = n())

print(continent_counts)
```

```
## # A tibble: 5 x 2
##   continent num_countries
##   <chr>             <int>
```

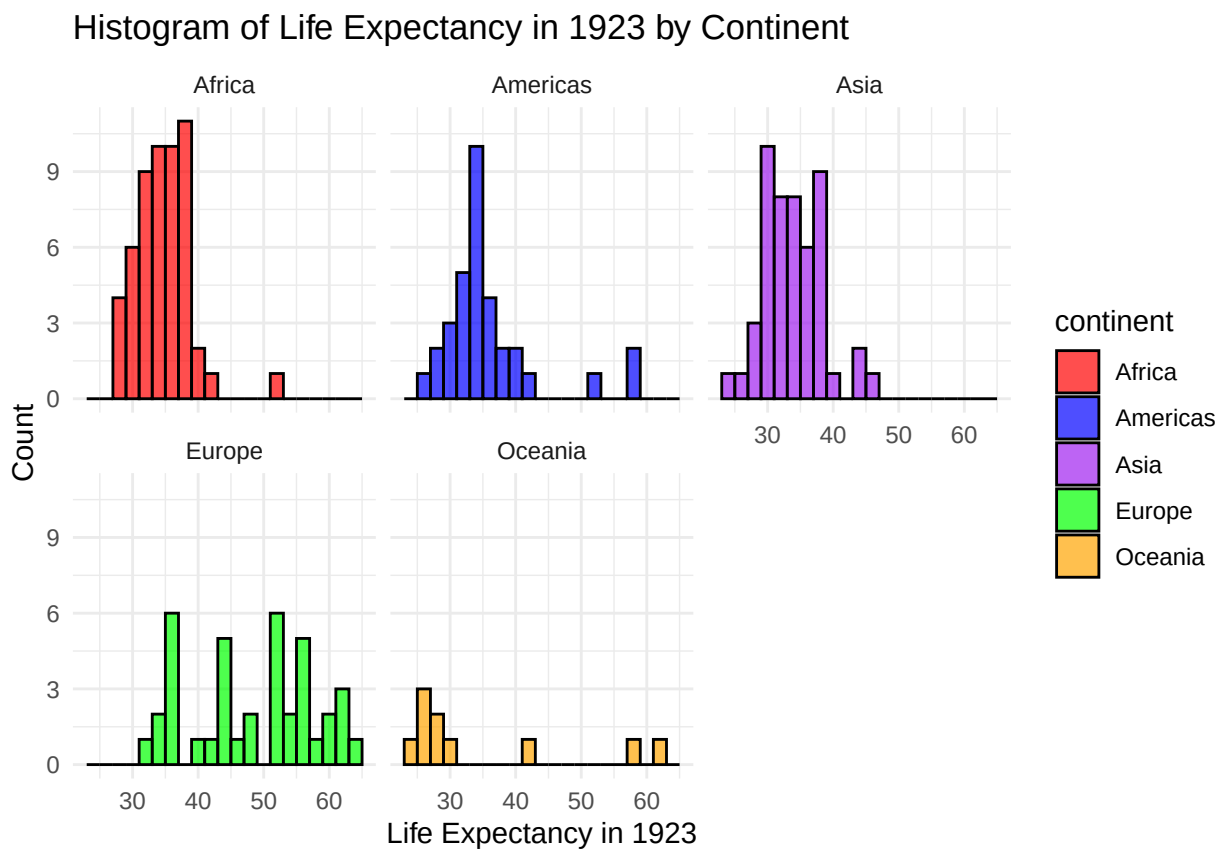
```
## 1 Africa                54
## 2 Americas              33
## 3 Asia                  50
## 4 Europe                39
## 5 Oceania               10
```

```r
# Explanation:
# - This shows the number of countries per continent in the dataset.

# Question 2: Boxplot of Life Expectancy in 1923 by Continent
ggplot(df, aes(x = continent, y = life1923, fill = continent)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = continent_colors) +
  ggtitle("Life Expectancy in 1923 by Continent") +
  xlab("Continent") +
  ylab("Life Expectancy in 1923") +
  theme_minimal()
```




```r
# Explanation:
# - The boxplot highlights regional disparities in 1923.

# Question 3: Histogram of Life Expectancy in 1923 by Continent
ggplot(df, aes(x = life1923, fill = continent)) +
  geom_histogram(binwidth = 2, color = "black", alpha = 0.7) +
  facet_wrap(~continent) +
```

```
  scale_fill_manual(values = continent_colors) +
  ggtitle("Histogram of Life Expectancy in 1923 by Continent") +
  xlab("Life Expectancy in 1923") +
  ylab("Count") +
  theme_minimal()
```




```
# Explanation:
# - The histogram shows variations in life expectancy across continents.

# Question 4: Summary Statistics by Continent
summary_table <- df %>%
  group_by(continent) %>%
  summarise(
    Mean_Life_Expectancy = mean(life1923, na.rm = TRUE),
    Median_Life_Expectancy = median(life1923, na.rm = TRUE)
  )

print(summary_table)
```

```
## # A tibble: 5 x 3
##   continent Mean_Life_Expectancy Median_Life_Expectancy
##   <chr>                    <dbl>                  <dbl>
## 1 Africa                    34.9                   34.6
## 2 Americas                  35.9                   34.1
## 3 Asia                      33.8                   33.8
```

```
## 4 Europe                    48.4                  51.5
## 5 Oceania                   35.1                  29.0
```

```r
# Explanation:
# - This provides mean and median life expectancy for each continent.

# Question 5: Regression Analysis by Continent
df$continent <- as.factor(df$continent)
model_continent <- lm(life1923 ~ continent, data = df)
summary(model_continent)
```

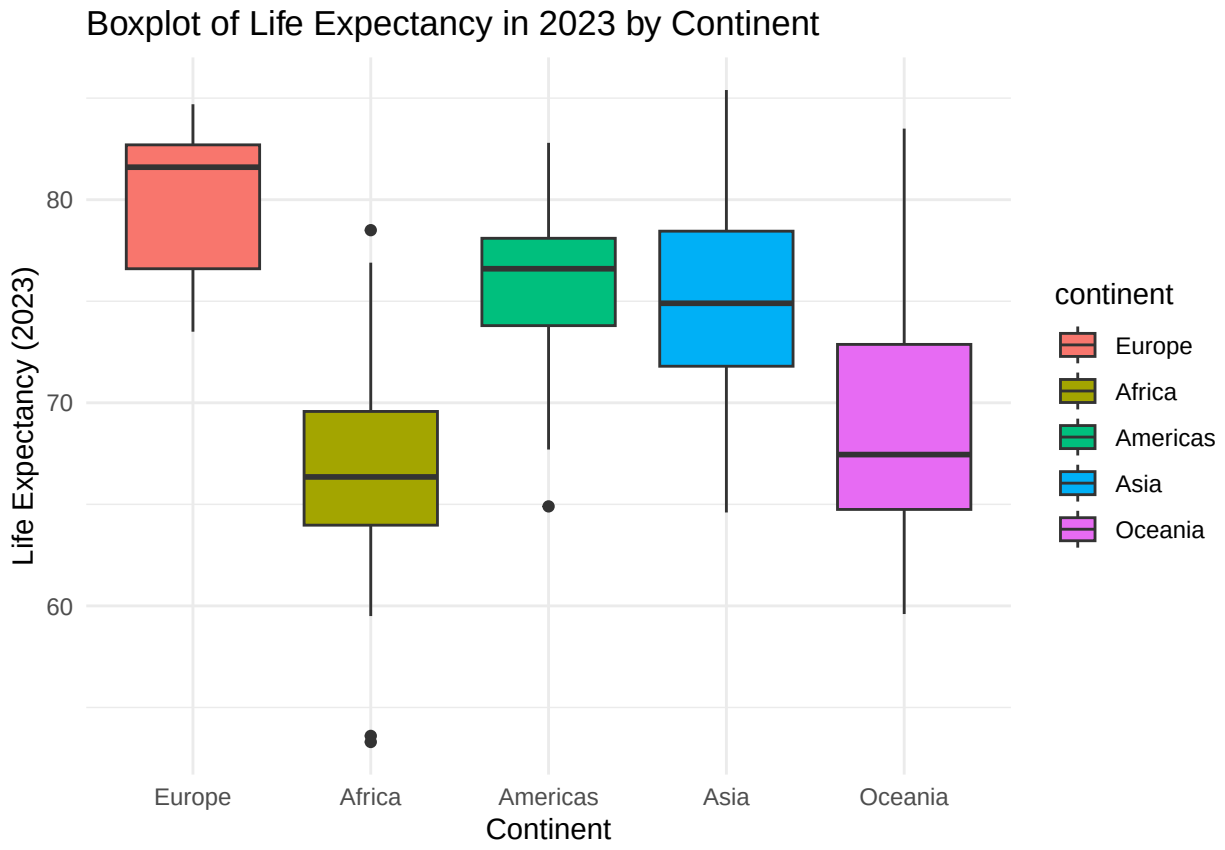
```
##
## Call:
## lm(formula = life1923 ~ continent, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7385  -4.0048  -0.9211   3.1615  26.9400
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        34.9037     0.9460  36.897   <2e-16 ***
## continentAmericas   1.0357     1.5360   0.674    0.501
## continentAsia      -1.1097     1.3643  -0.813    0.417
## continentEurope    13.5348     1.4608   9.265   <2e-16 ***
## continentOceania    0.1563     2.3931   0.065    0.948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.951 on 181 degrees of freedom
## Multiple R-squared:  0.4009, Adjusted R-squared:  0.3877
## F-statistic: 30.28 on 4 and 181 DF,  p-value: < 2.2e-16
```

```r
# Explanation:
# - Europe had significantly higher life expectancy than other continents.

# Question 6: Change Reference Level & Re-run Model
df$continent <- relevel(df$continent, ref = "Europe")
model_continent_europe <- lm(life1923 ~ continent, data = df)
summary(model_continent_europe)
```

```
##
## Call:
## lm(formula = life1923 ~ continent, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7385  -4.0048  -0.9211   3.1615  26.9400
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         48.438      1.113  43.516  < 2e-16 ***
## continentAfrica    -13.535      1.461  -9.265  < 2e-16 ***
```

```
## continentAmericas  -12.499      1.644  -7.602 1.52e-12 ***
## continentAsia      -14.644      1.485  -9.861  < 2e-16 ***
## continentOceania   -13.378      2.464  -5.430 1.80e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.951 on 181 degrees of freedom
## Multiple R-squared:  0.4009, Adjusted R-squared:  0.3877
## F-statistic: 30.28 on 4 and 181 DF,  p-value: < 2.2e-16
```

```r
# Explanation:
# - Setting Europe as the reference confirms it had the highest life expectancy.

# Question 7: Repeat for Life Expectancy in 2023
ggplot(df, aes(x = continent, y = life2023, fill = continent)) +
  geom_boxplot() +
  ggtitle("Boxplot of Life Expectancy in 2023 by Continent") +
  xlab("Continent") +
  ylab("Life Expectancy (2023)") +
  theme_minimal()
```




```r
model_continent_2023 <- lm(life2023 ~ continent, data = df)
summary(model_continent_2023)
```

```
##
```

```
## Call:
## lm(formula = life2023 ~ continent, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5296  -3.0978  -0.2468   2.9618  13.8500
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        79.8179     0.7621 104.738  < 2e-16 ***
## continentAfrica   -12.9883     1.0001 -12.987  < 2e-16 ***
## continentAmericas  -3.7089     1.1257  -3.295  0.00118 **
## continentAsia      -4.4539     1.0167  -4.381 2.00e-05 ***
## continentOceania  -10.1679     1.6869  -6.028 9.12e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.759 on 181 degrees of freedom
## Multiple R-squared:  0.5234, Adjusted R-squared:  0.5129
## F-statistic: 49.69 on 4 and 181 DF,  p-value: < 2.2e-16
```

```
# Explanation:
# - In 2023, Europe still has the highest life expectancy, and Africa the lowest.

# Question 8: Compare Changes Over 100 Years
df %>% group_by(continent) %>%
  summarise(Mean_Life1923 = mean(life1923, na.rm = TRUE),
            Mean_Life2023 = mean(life2023, na.rm = TRUE),
            Change = Mean_Life2023 - Mean_Life1923)
```

```
## # A tibble: 5 x 4
##   continent Mean_Life1923 Mean_Life2023 Change
##   <fct>             <dbl>         <dbl>  <dbl>
## 1 Europe             48.4          79.8   31.4
## 2 Africa             34.9          66.8   31.9
## 3 Americas           35.9          76.1   40.2
## 4 Asia               33.8          75.4   41.6
## 5 Oceania            35.1          69.6   34.6
```

```
# Explanation:
# - Life expectancy increased globally, but regional gaps remain.
```