Lab 2

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

```r
library(countrycode)
```

```
## Warning: package 'countrycode' was built under R version 4.3.3
```

```r
library(dplyr)
```

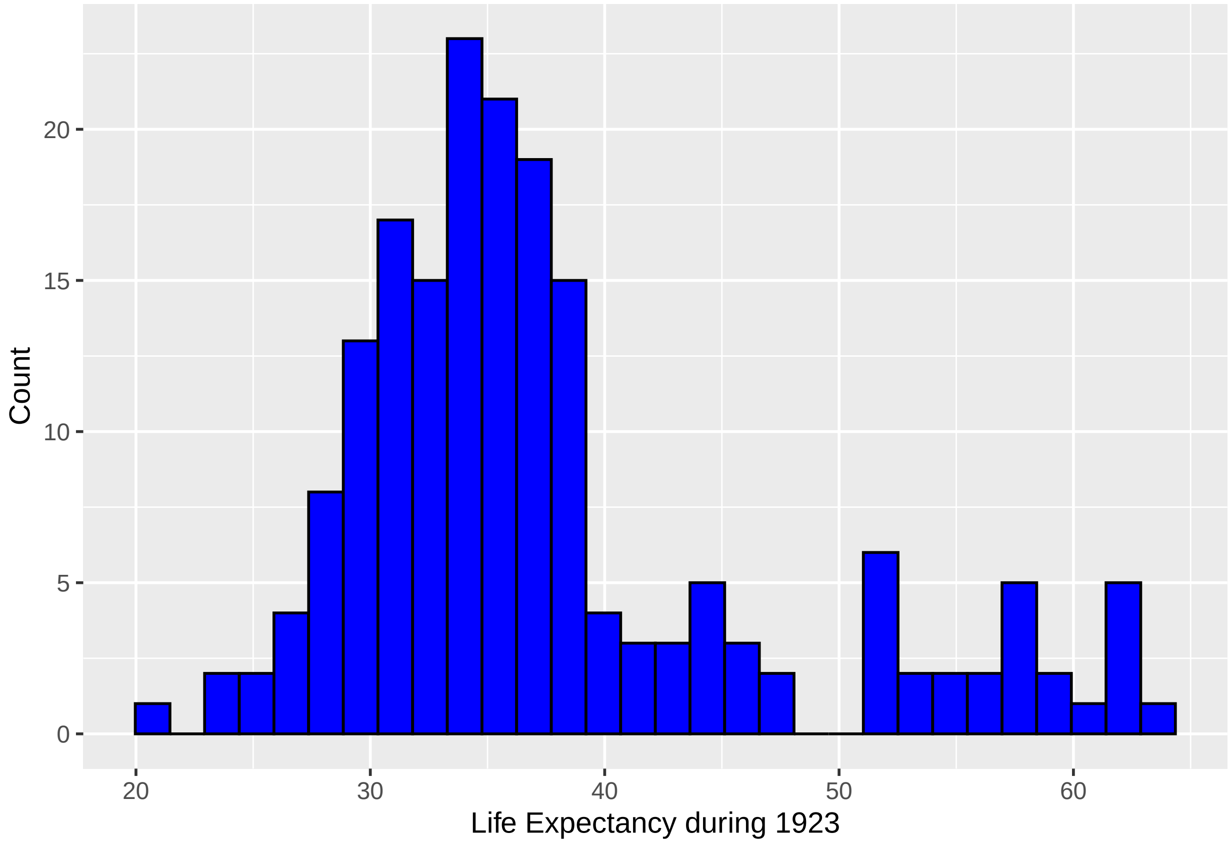
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Below we are going to subset the life expectancy data so we only see the contries and the years 1923
life_exp <- read.csv("lex.csv")
#glimpse(life_exp)
data <- life_exp[c("country", "X1923", "X2023")]
#glimpse(data)
data$continent <- countrycode(sourcevar = data[, "country"],
origin = "country.name",
destination = "continent")

data <- na.omit(data)

glimpse(data)
```

```
## Rows: 186
## Columns: 4
## $ country   <chr> "Afghanistan", "Angola", "Albania", "UAE", "Argentina", "Arm~
## $ X1923     <dbl> 38.6, 36.5, 34.3, 40.0, 52.7, 33.3, 35.3, 62.0, 51.9, 26.3, ~
## $ X2023     <dbl> 64.0, 65.9, 77.3, 74.6, 77.3, 76.5, 77.0, 83.5, 82.8, 71.5, ~
## $ continent <chr> "Asia", "Africa", "Europe", "Asia", "Americas", "Asia", "Ame~
```

```
library(ggplot2)
ggplot(data = data, aes(x=X1923))+geom_histogram(fill="blue",color="black")+labs(x="Life Expectancy dur
```

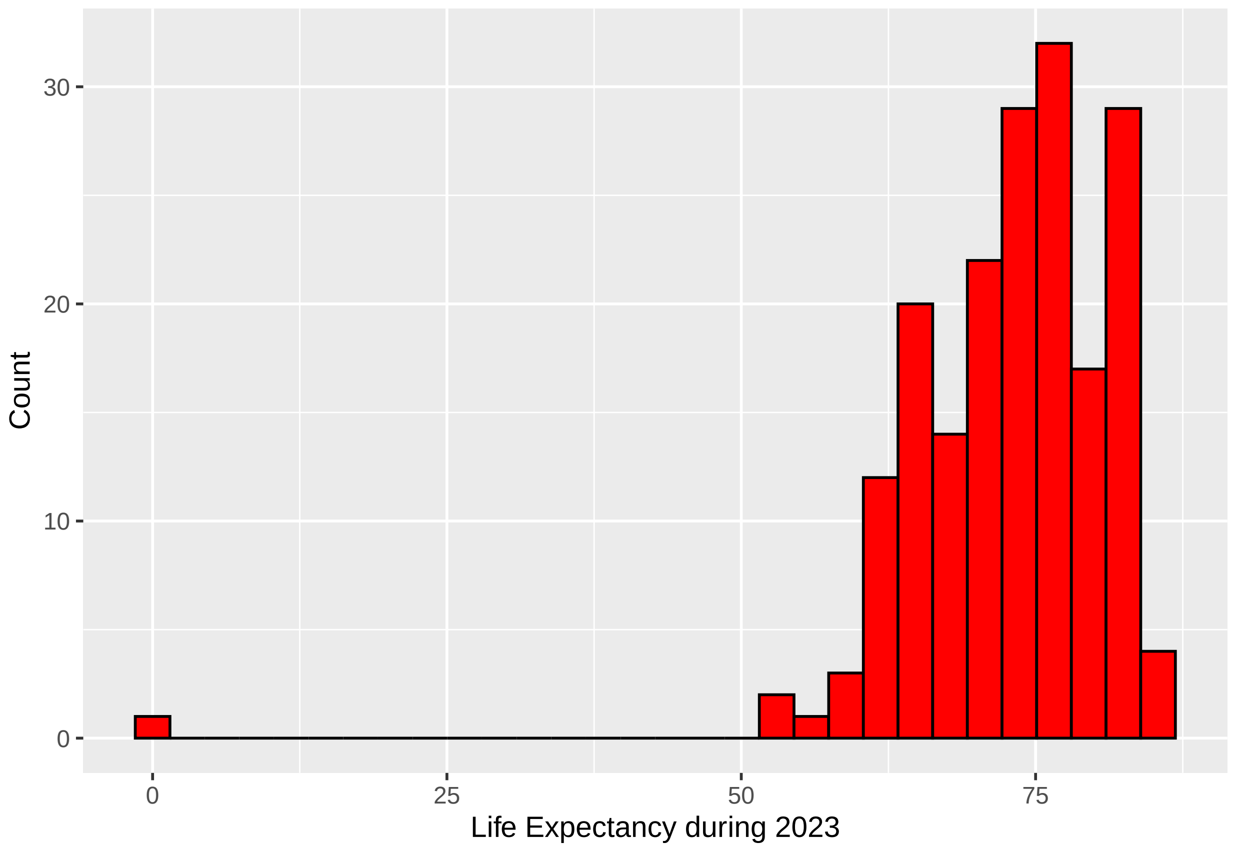
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



The Histogram of life expectanyc during 1923 is skewed to the right meaning the life expectancy is lower, and people lived bewteen 30 and 40 years. The histogram also has a some isolated cases of older life expectancy. At the time during 1923 there were better living conditions in some countries then others. Also with the development of medinces and more inventions in the western world its safe to assume these countries with the higher life expectancy are in Europe.

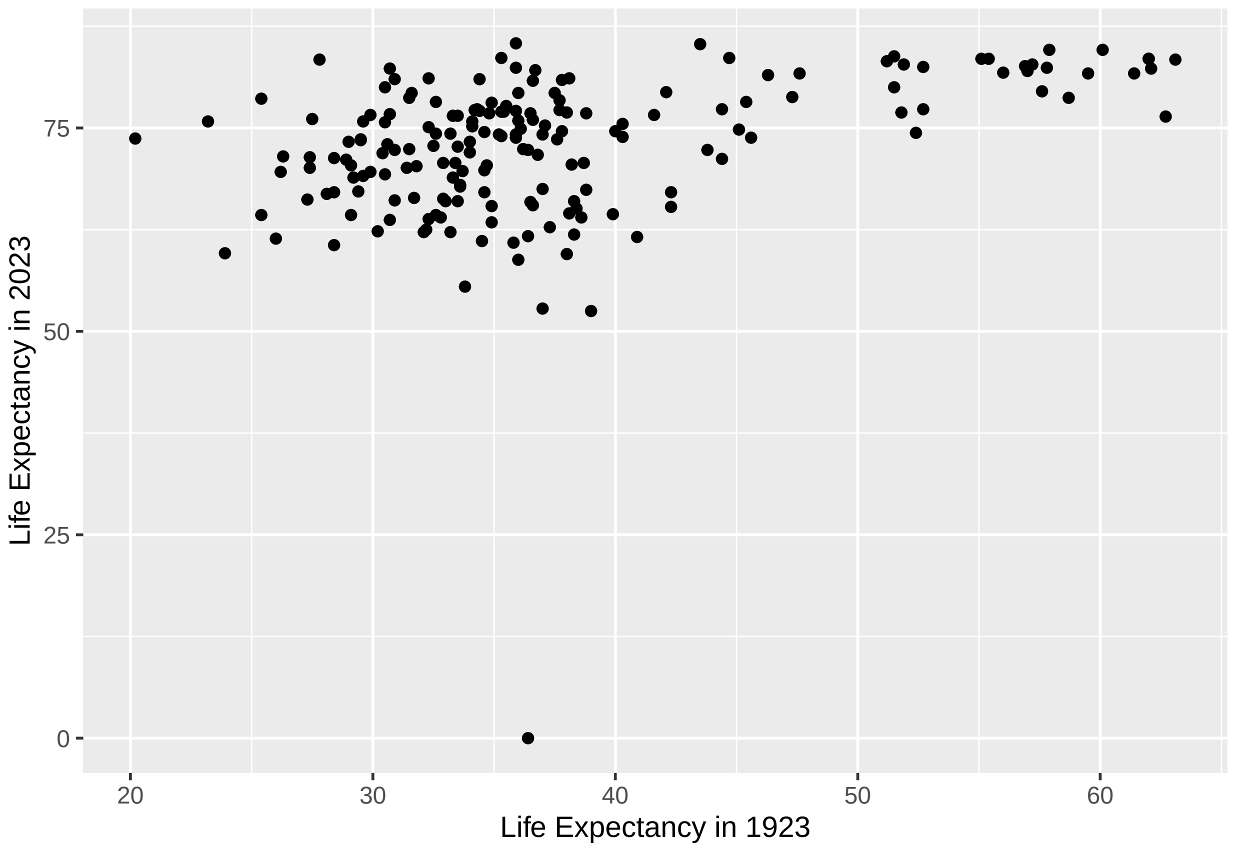
```
ggplot(data = data, aes(x=X2023))+geom_histogram(fill="red",color="black")+labs(x="Life Expectancy durin
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



The histogram of the life expectancy during 2023 on the other hand is skewed to the left showing that in 2023, people lived longer lives wiht majority of the data being around 77 and 80. Where as in 1923 it was closer to 30-40.

```
ggplot(data, aes(x=X1923, y=X2023))+geom_point()+labs(x="Life Expectancy in 1923", y= "Life Expectancy i
```



The points show a positive correlation as they tend to stay higher and increase. Countries that had a higher life expectancy in 1923 also had a higher life expectancy in 2023 shown by the upword trend of the dots otherwise known as a postive corrilation.

```
#simple linear regression
#install.packages("moderndive")
library(moderndive)
```

```
## Warning: package 'moderndive' was built under R version 4.3.3
```

```
model1 <- lm(data = data, X1923~X2023)
get_regression_table(model1)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     8.48      5.1       1.66   0.098   -1.58    18.5
## 2 X2023         0.4       0.07      5.76   0         0.263    0.538
```

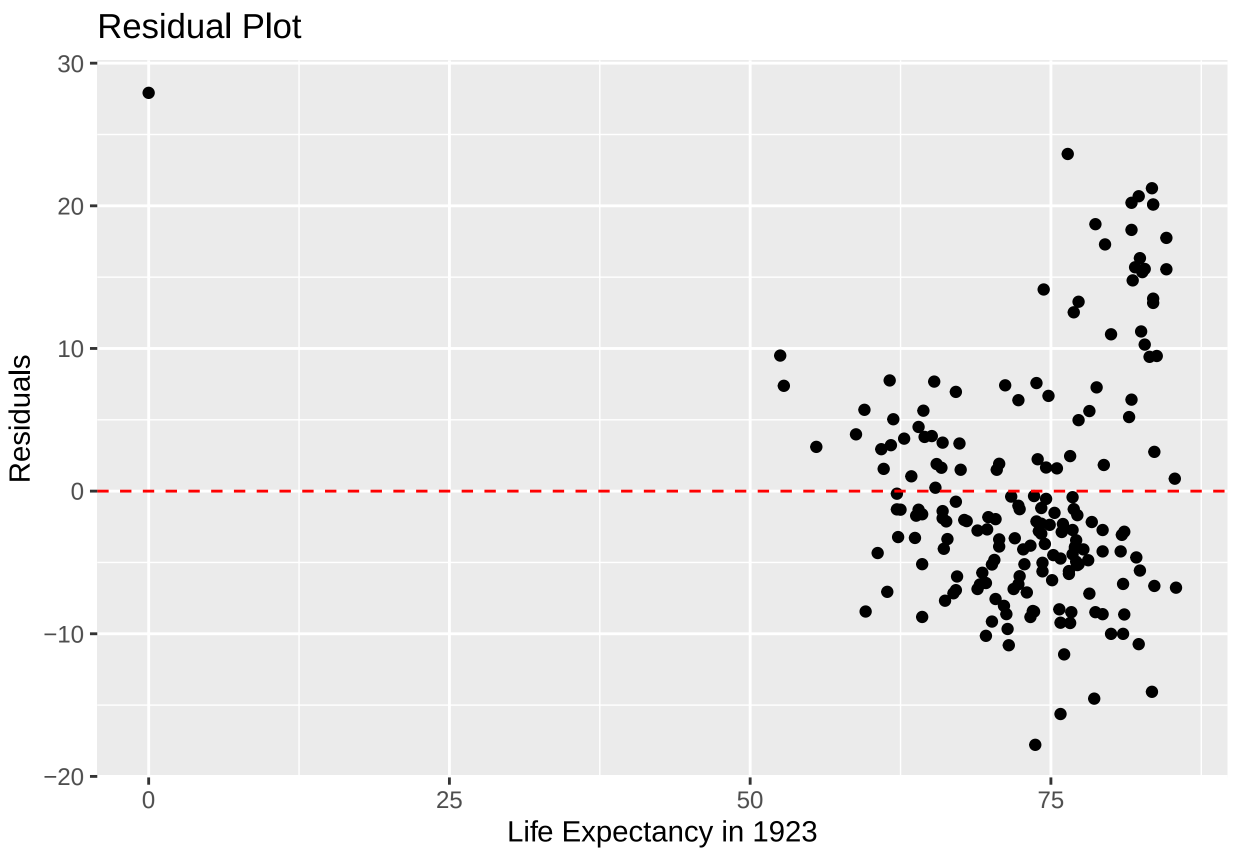
```
#print(model1)
```

The relationship between 1923 and 2023 is somewhat significant. The R squared values is 15.26% so there is definatly some relationshoip between the two. The P value though is also greater than 0.05 so we also can not reject the null hypothesis so there may be no relationship too.

5. If the life expectancy of 1923 were to increase by one year would lead to a 0.4 increase in 2023 as shown by the estamate value of 2023 becuase it is 0.4 greater than 1923. Further 1/0.4 =. about 2.5 so for every 1 year increase in 1923, we see a 2.5 year increase for the life expectancy in 2023.

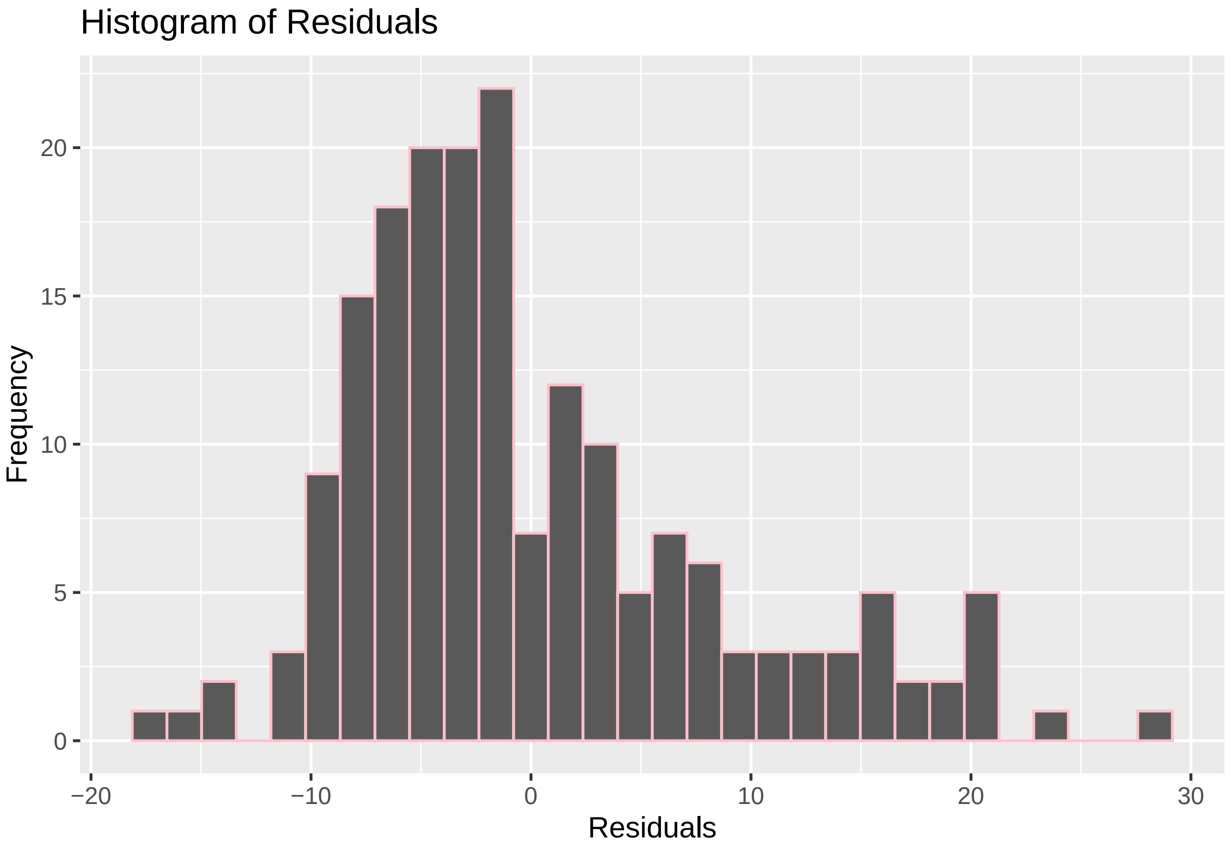
6. based on the model and the equation of lienar regresion , we use the formula 34.3-8.4 = 25.8 then we divide this by 0.4 to get 64.5 years would be the individuals life expectancy in 2023

```r
library(ggplot2)
model1 <- lm(data = data, X1923~X2023)
points <-  get_regression_points(model1)
ggplot(data = points, aes(x =X2023 , y = residual))+ geom_point()+geom_hline(yintercept = 0, linetype =
  labs(title = "Residual Plot", x = "Life Expectancy in 1923", y = "Residuals")
```



```r
ggplot(data = points, aes(x=residual))+geom_histogram(color='pink') +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



7. The values seem to be clustering around 75 but the data seems random to me becuase there is no explicit trend or pattern that can be seen. In the histogram the residuals are slighly skewed to the right but there is no other trend that is apparent to us from these 2 plots. So I'd assume the data is somewhat random.

```
get_regression_summaries(model1)
```

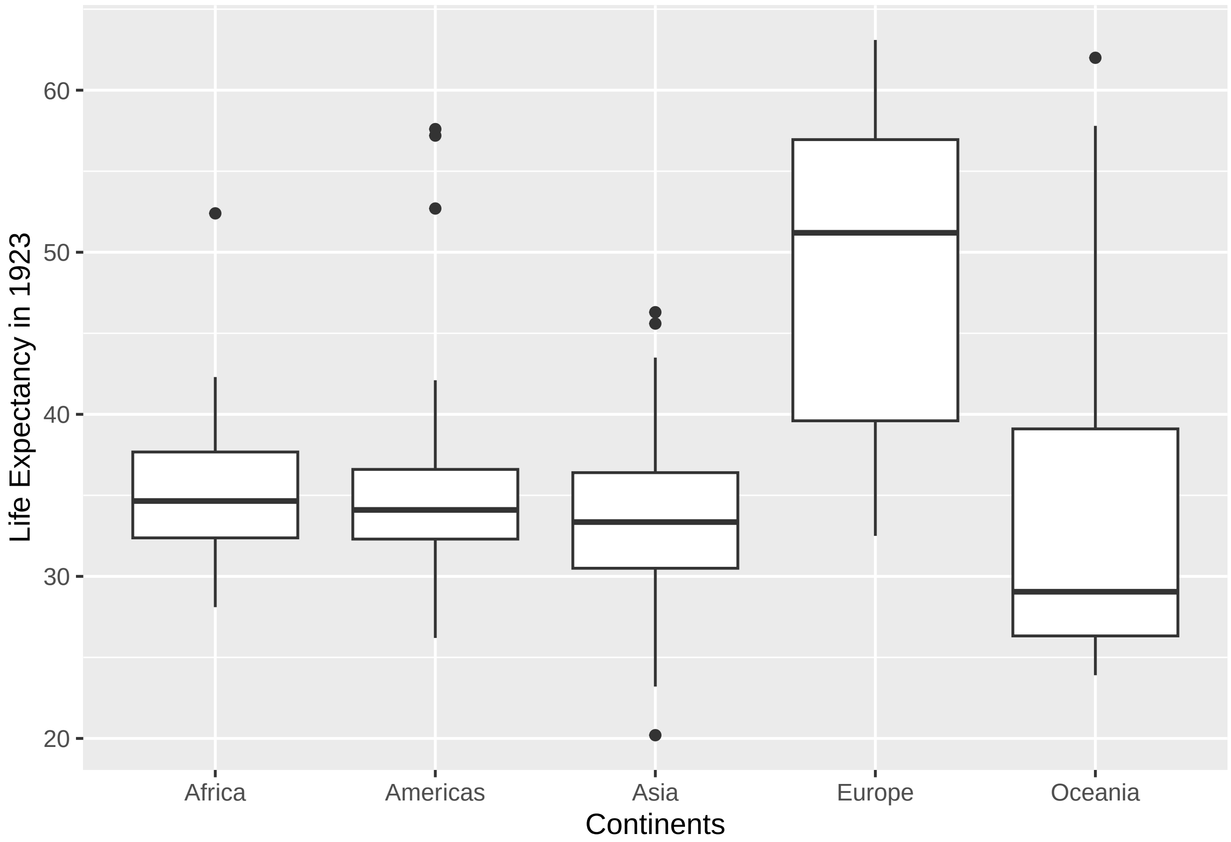
```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.153         0.148  70.5  8.39  8.44      33.1       0     1   186
```

8. Our R^2 value is 0.153 so 15% of the time the variation can be described by the life expectancy in 2023. So it is somewhat relevant or 15% relevant.

```
library(dplyr)
counts <- data %>%
  group_by(continent) %>%
  summarise(counts = n_distinct(country))
print(counts)
```

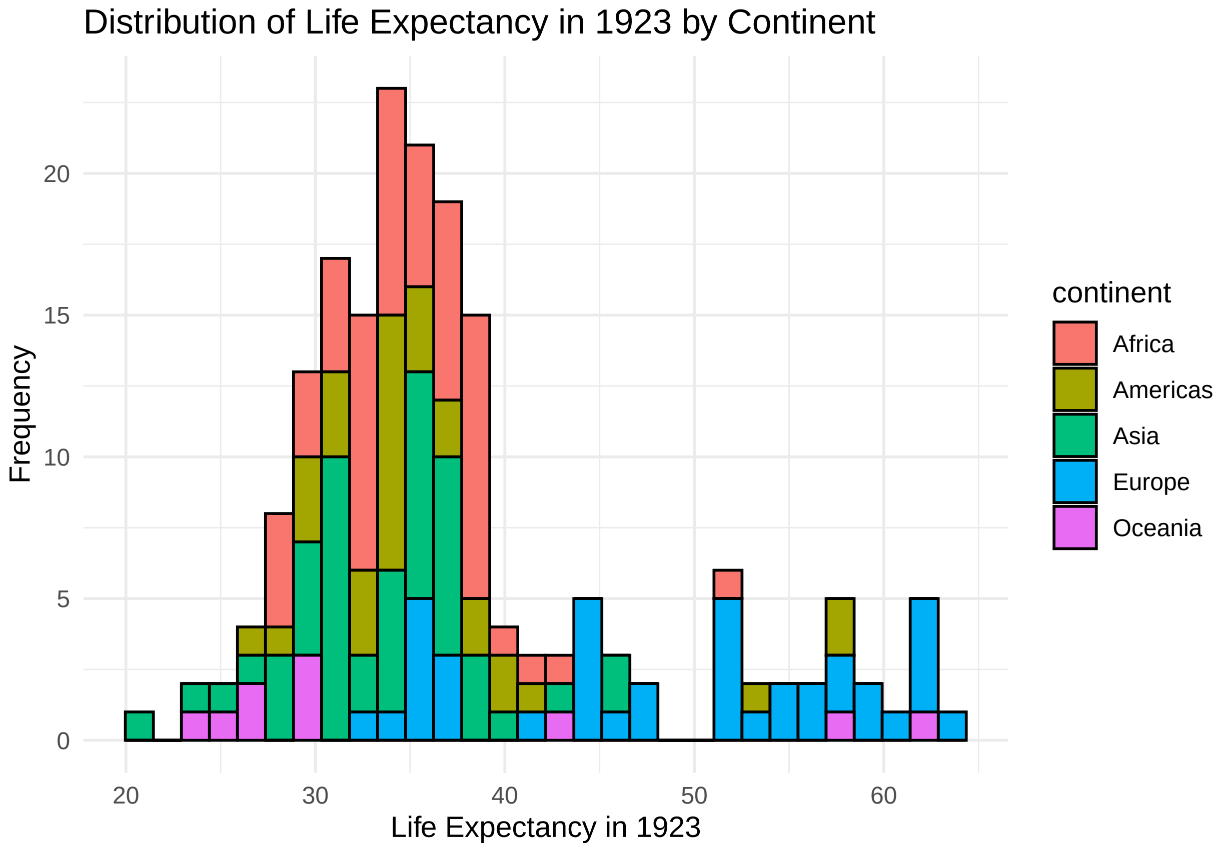
```
## # A tibble: 5 x 2
##   continent counts
##   <chr>      <int>
## 1 Africa        54
## 2 Americas      33
## 3 Asia          50
## 4 Europe        39
## 5 Oceania       10
```

```
#Box plots by contnenet
ggplot(data, aes(x=continent, y = X1923))+ geom_boxplot()+ labs(x= " Continents ", y = "Life Expectancy
```



```
# histogram
ggplot(data, aes(x = X1923, fill = continent)) +
  geom_histogram( position = "stack", color="black") +
  labs(title = "Distribution of Life Expectancy in 1923 by Continent",
       x = "Life Expectancy in 1923",
       y = "Frequency") +
  theme_minimal()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



This histogram is skewwed to the right.But there are some outliers which are mostly European Countries that had higher life expectancy in 1923.

```
summary_table <- data %>%
  group_by(continent) %>%
  summarise(
    mean_life_exp = mean(X1923, na.rm = TRUE),
    median_life_exp = median(X1923, na.rm = TRUE)
  )

# Print the summary table
print(summary_table)
```

```
## # A tibble: 5 x 3
##   continent mean_life_exp median_life_exp
##   <chr>             <dbl>           <dbl>
## 1 Africa             34.9            34.7
## 2 Americas           35.8            34.1
## 3 Asia               33.5            33.3
## 4 Europe             48.7            51.2
## 5 Oceania            35.2            29.0
```

```
model2 <- lm(data=data, X1923~continent)
get_regression_table(model2)
```

```
## # A tibble: 5 x 7
##   term                 estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept              34.9       0.972    35.9     0        33.0     36.9
## 2 continent: Americas     0.879     1.58      0.557   0.578    -2.24     4.00
## 3 continent: Asia        -1.45      1.40     -1.03    0.303    -4.22     1.32
## 4 continent: Europe      13.8       1.50      9.18    0        10.8     16.8
## 5 continent: Oceania      0.282     2.46      0.115   0.909    -4.57     5.14
```

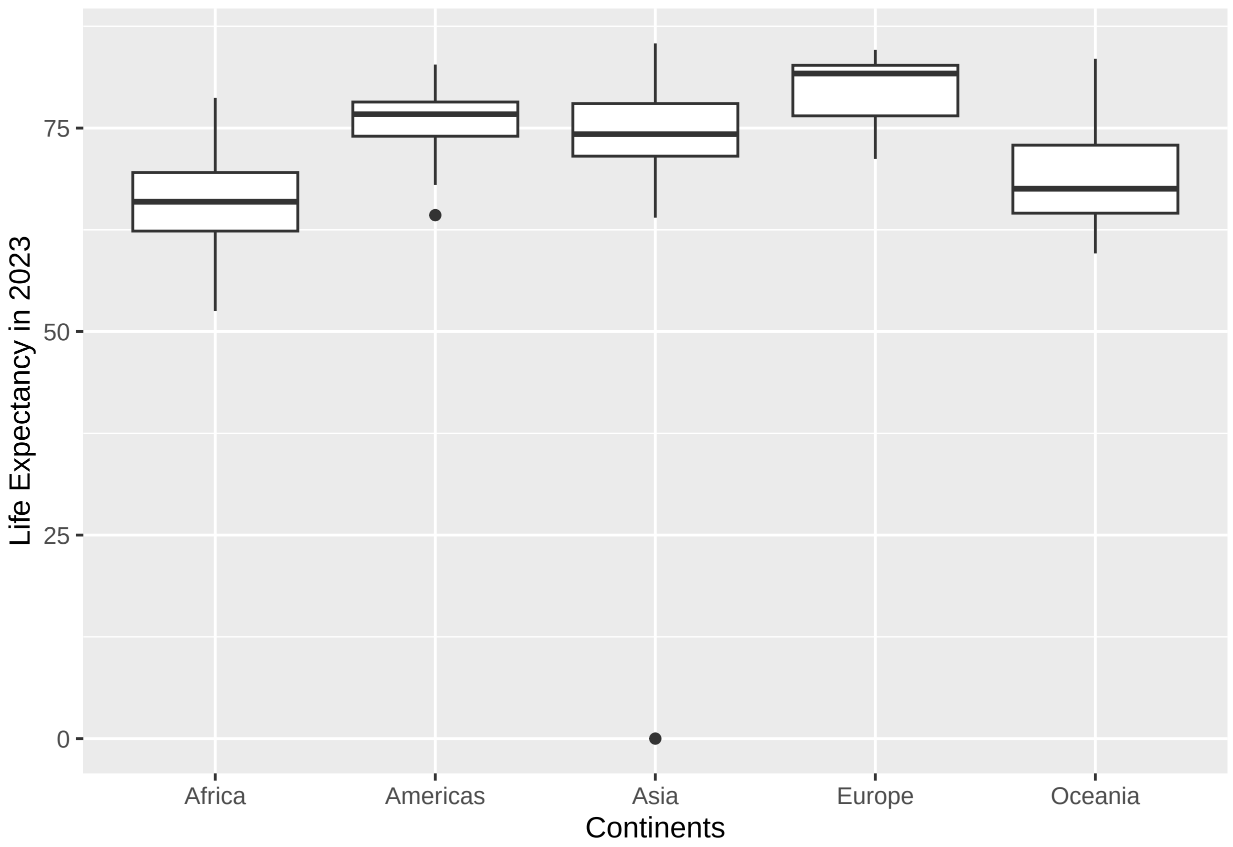
Based on the regresion table the estimate life expectanyc Africa is calcuated to have an average life expectancy of 35, America is 35, Asia is 33 Europe is 47 and Oceania is 35. These are all very close to the values calculated above. These estimates from the regresion model are correctly calcualted and match the means. To check the significance of these levels we need to look at the P-value next. If we set our reference P-value to 0.05 the only stastically significant continent is Europe. We can not reject the null hypothesis for the rest of the continents becuase there P-vakue is to high so they are insignificant.

```
data$continent <- factor(data$continent)
new_model <- lm(data=data, X1923~relevel(continent, ref = "Europe"))
get_regression_table(new_model)
```

```
## # A tibble: 5 x 7
##   term                         estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 "intercept"                      48.7      1.14      42.6        0     46.5     51.0
## 2 "relevel(continent, re~       -13.8      1.50      -9.18       0    -16.8    -10.8
## 3 "relevel(continent, re~       -12.9      1.69      -7.64       0    -16.2     -9.58
## 4 "relevel(continent, re~       -15.2      1.53      -9.98       0    -18.3    -12.2
## 5 "relevel(continent, re~       -13.5      2.53      -5.33       0    -18.5     -8.51
```

As shown the reference level effects the p value, so when we use Eruope as our reference all continents are statstically significant becuase they are different from the intercept which is Europe.

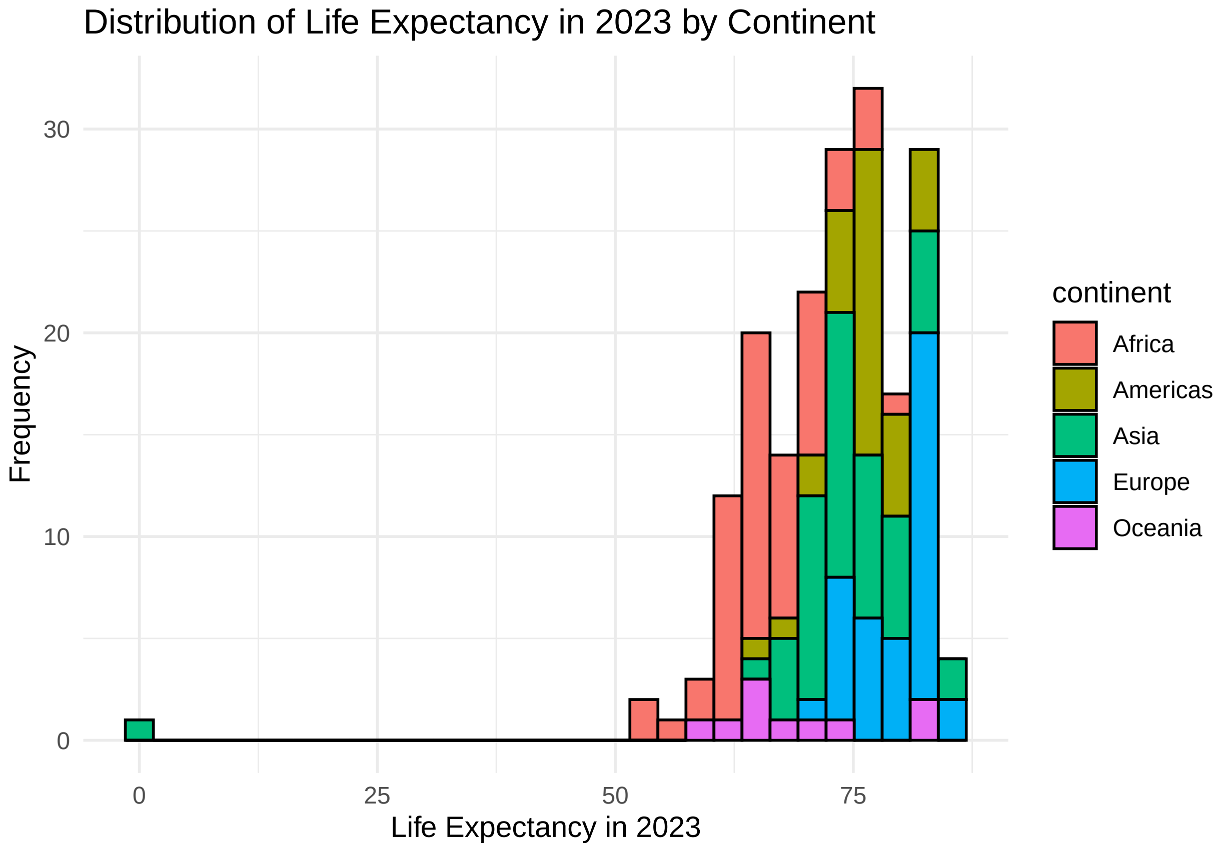
```
ggplot(data, aes(x=continent, y = X2023))+ geom_boxplot()+ labs(x= " Continents ", y = "Life Expectancy
```



Less spread on these box plots then in 1923. The data is more focused around the mean except for a few outliers.

```
ggplot(data, aes(x = X2023, fill = continent)) +
  geom_histogram( position = "stack", color="black") +
  labs(title = "Distribution of Life Expectancy in 2023 by Continent",
       x = "Life Expectancy in 2023",
       y = "Frequency") +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This Plot is skewed to the left, most of the data points and life expectancies are for higher values unlike 1923 in which the data was skewed to the right.

```
summary_table <- data %>%
  group_by(continent) %>%
  summarise(
    mean_life_exp = mean(X2023, na.rm = TRUE),
    median_life_exp = median(X2023, na.rm = TRUE)
  )

# Print the summary table
print(summary_table)
```

```
## # A tibble: 5 x 3
##   continent mean_life_exp median_life_exp
##   <fct>             <dbl>           <dbl>
## 1 Africa             66.0            66.0
## 2 Americas           76.1            76.7
## 3 Asia               73.4            74.2
## 4 Europe             79.4            81.7
## 5 Oceania            69.6            67.6
```

```
model3 <- lm(data=data, X2023~continent)
get_regression_table(model3)
```

```
## # A tibble: 5 x 7
##   term                 estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept               66.0       1.01     65.2    0         64.0     68.0
## 2 continent: Americas     10.1       1.64      6.15   0          6.86    13.4
## 3 continent: Asia          7.42      1.46      5.08   0          4.54    10.3
## 4 continent: Europe       13.5       1.56      8.60   0         10.4     16.5
## 5 continent: Oceania       3.63      2.56      1.42   0.158     -1.42     8.69
```

Based on regreison model the average life expectanyc in Africa is 66 years, America it is 76 years, Asia it is 73 years, Europe it is 79 years, and in Ociainia it is 69 years. Again these values match up to the means of the life expectancy for each country showing the model is accurate. In terms of signifcance all the values are singificant except Ociania. For Ociania the P-value is to hgih and we fail to reject the null hypothesis.

```
# new reference
new_model2 <- lm(data=data, X2023~relevel(continent, ref = "Europe"))
get_regression_table(new_model2)
```

```
## # A tibble: 5 x 7
##   term                         estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 "intercept"                     79.4       1.19     66.7    0         77.1    81.8
## 2 "relevel(continent, re~   -13.5       1.56     -8.60   0        -16.5   -10.4
## 3 "relevel(continent, re~    -3.34      1.76     -1.9    0.059     -6.82    0.129
## 4 "relevel(continent, re~    -6.04      1.59     -3.80   0         -9.17   -2.90
## 5 "relevel(continent, re~    -9.82      2.64     -3.72   0        -15.0    -4.62
```

When we switch the reference point again we get new p- values. These new p valeus are all significant except for the Americas. Our value we comapre is 0.05 and 0.059 is just barely bigger so we fail to reject the null hypothesis fro the Americas. Therefore making it not statistically significant.

Last Question :

As shown by the histograms by continent, in 1923, the data was right-skewed, meaning most countries had lower life expectancies, but a few had much higher values. The spread was large, indicating a significant variation in life expectancy between countries. By 2023, the histogram is left-skewed, with most values clustered around 75 years of age. This shift indicates that global life expectancy has risen by approximately 30–35 years, with fewer countries having extremely low life expectancy. The boxplots further confirm this trend. In 1923, most continents had median life expectancies between 30 and 40 years, with Europe being the exception, having a significantly higher average and a much larger spread. In 2023, life expectancy averages across continents became more similar, clustering around 60–70 years, and the spread of data decreased, meaning differences within continents have become smaller over time. Similarly, from the regression table in 1923, when Africa was the reference, we see that aside from Europe, the average life expectancy across continents was within one year of each other, and the p-values were not statistically significant. However, in 2023, the regression model shows that life expectancies are more spread out, with differences of around seven or more years between continents. Additionally, p-values are lower, indicating more statistical significance. This suggests that in 2023, the country you live in has a greater statistical effect on life expectancy for three out of the five continents when Africa was the reference, whereas in 1923, only Europe had a significant effect on life expectancy compared to Africa.