HW2 - Akshaj Kammari

Due: 02/09/2024

```
library(readr)
library(ggplot2)
library(dplyr)
```

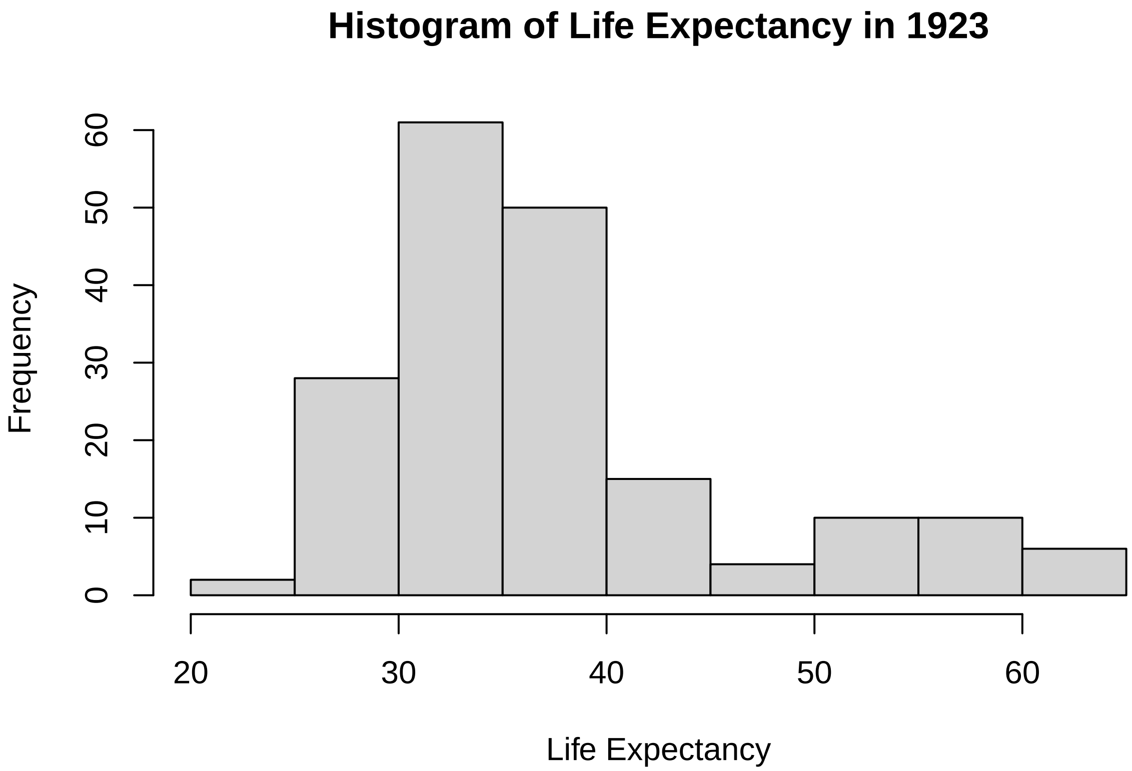
```
##
## Attaching package: 'dplyr'

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

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

```
world_data <- read.csv("Worldlife.csv")
```

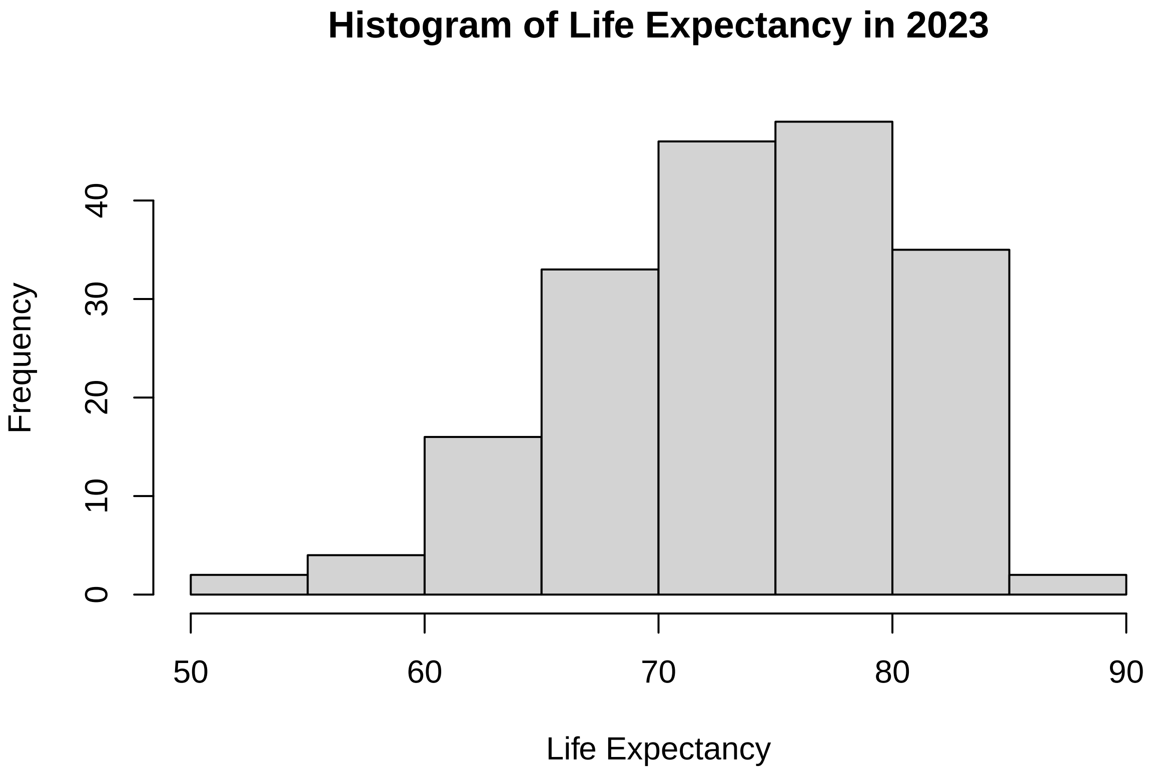
###Regression of life expectancy in 2023 on life expectancy in 1923

###1.Have a histogram of the life expectancy in 1923, and a histogram of life expectancy in 2023, describe the distribution in each year. Are there any observations that are kind of isolated from the rest of the observations in either graph? If so identify them and find possible reasons for that

```
hist(world_data$life1923, main = "Histogram of Life Expectancy in 1923", xlab = "Life Expectancy")
```



```r
hist(world_data$life2023, main = "Histogram of Life Expectancy in 2023", xlab = "Life Expectancy")
```



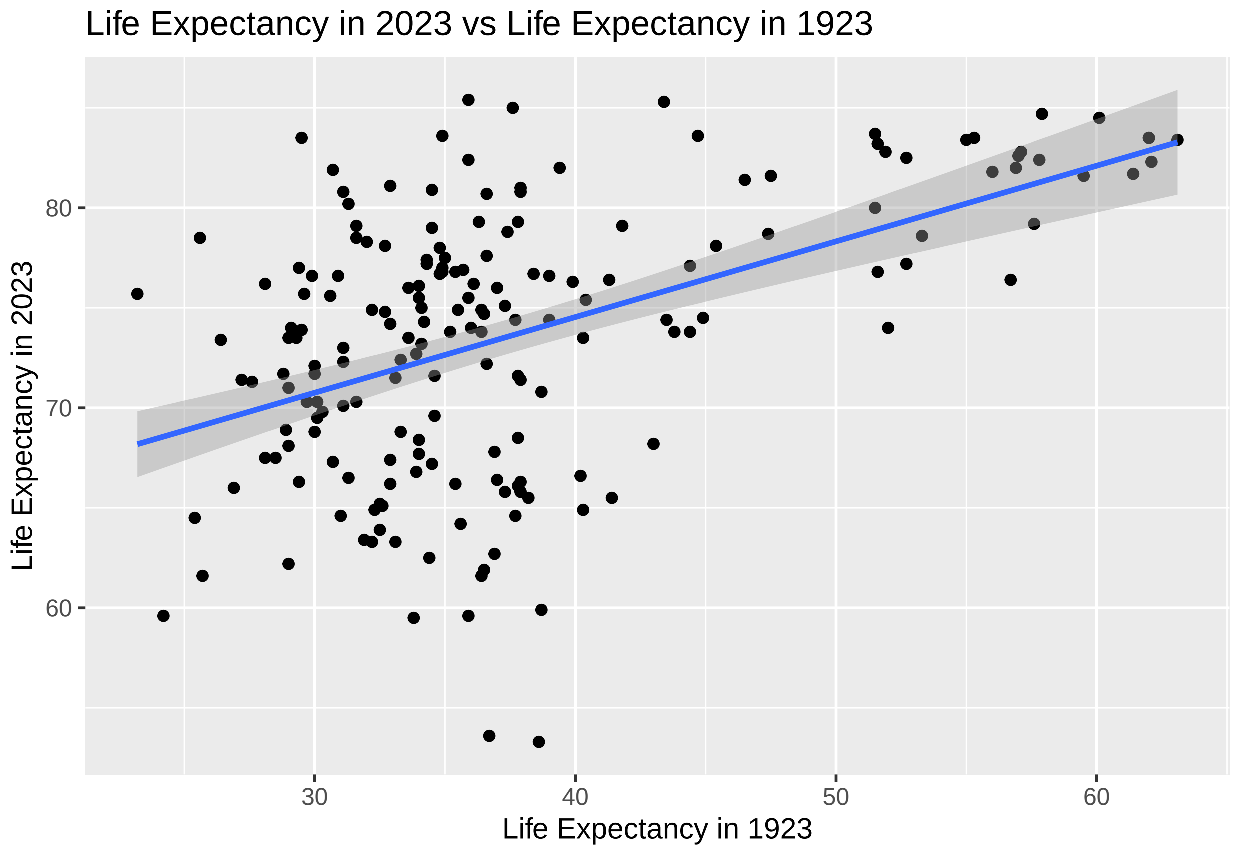
#In 1923, the plot is skewed right, while in 2023, the plot is skewed left. This is due to the fact that life expectancy has increased as time went on.

###2.Does it appear to be some linear relationship between life expectancy in 2023 vs. life expectancy in

1923 (using the scatterplot)? Is it a positive or negative trend?

```
ggplot(world_data, aes(x = life1923, y = life2023)) + geom_point() + geom_smooth(method = "lm") + xlab(
```

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



#Positive trend

###3.What is the correlation between life expectancy in 2023 and in 1923?

```
cor(world_data$life1923, world_data$life2023)
```

```
## [1] 0.4929469
```

###4.Run a simple regression. Is the model significant?

```
model <- lm(life2023 ~ life1923, data = world_data)
summary(model)
```

```
##
## Call:
## lm(formula = life2023 ~ life1923, data = world_data)
##
## 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
```

#Yes, the model is significant.

###5.If life expectancy in 1923 could increase by 1 year, what would be the expected increase in life expectancy in 2023?

```
coef(model)["life1923"]
```

```
##  life1923
## 0.3783683
```

#37.83%

###6.One country had life expectancy 34.3 in 1923 and no record for its life expectancy in 2023. Predict its life expectancy in 2023.

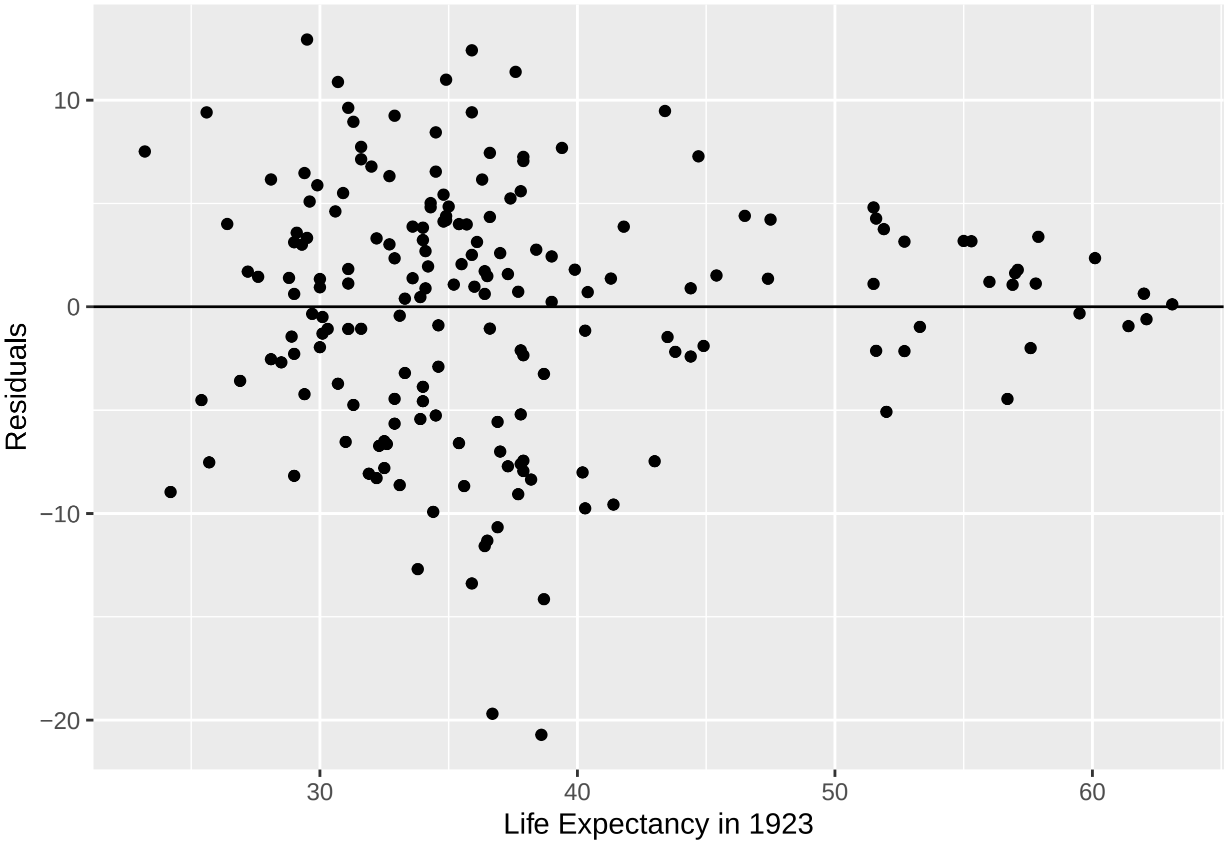
```
predict(model, newdata = data.frame(life1923 = 34.3))
```

```
##        1
## 72.38312
```

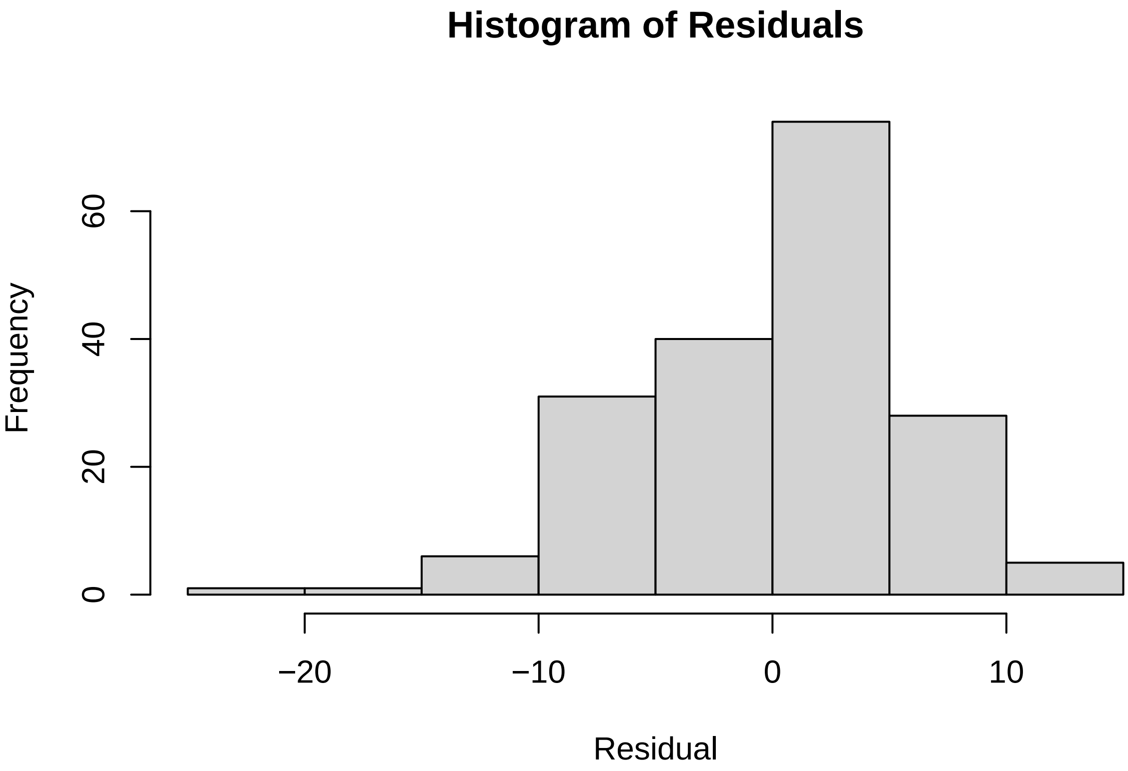
#72.38 years

###7.Have a residual plot of residual against life expectancy in 1923 (with a horizontal line at y=0), and a histogram of the residual. Describe whether the residual seems to be random, explain why.

```
world_data$residuals <- resid(model)
ggplot(world_data, aes(x = life1923, y = residuals)) + geom_point() + geom_hline(yintercept = 0) + xlab
```



```
hist(world_data$residuals, main = "Histogram of Residuals", xlab = "Residual")
```



#Yes, a little, because there are points located throughout the plot, but the majority of those points lie between the ages 30 and 40.

###8.What is the percentage of total variability in life expectancy in 2023 that can be explained through the linear model using life expectancy in 1923?

```
summary(model)$r.squared
```

```
## [1] 0.2429967
```

#24.3%

##PART 2: Regression of life expectancy on continent

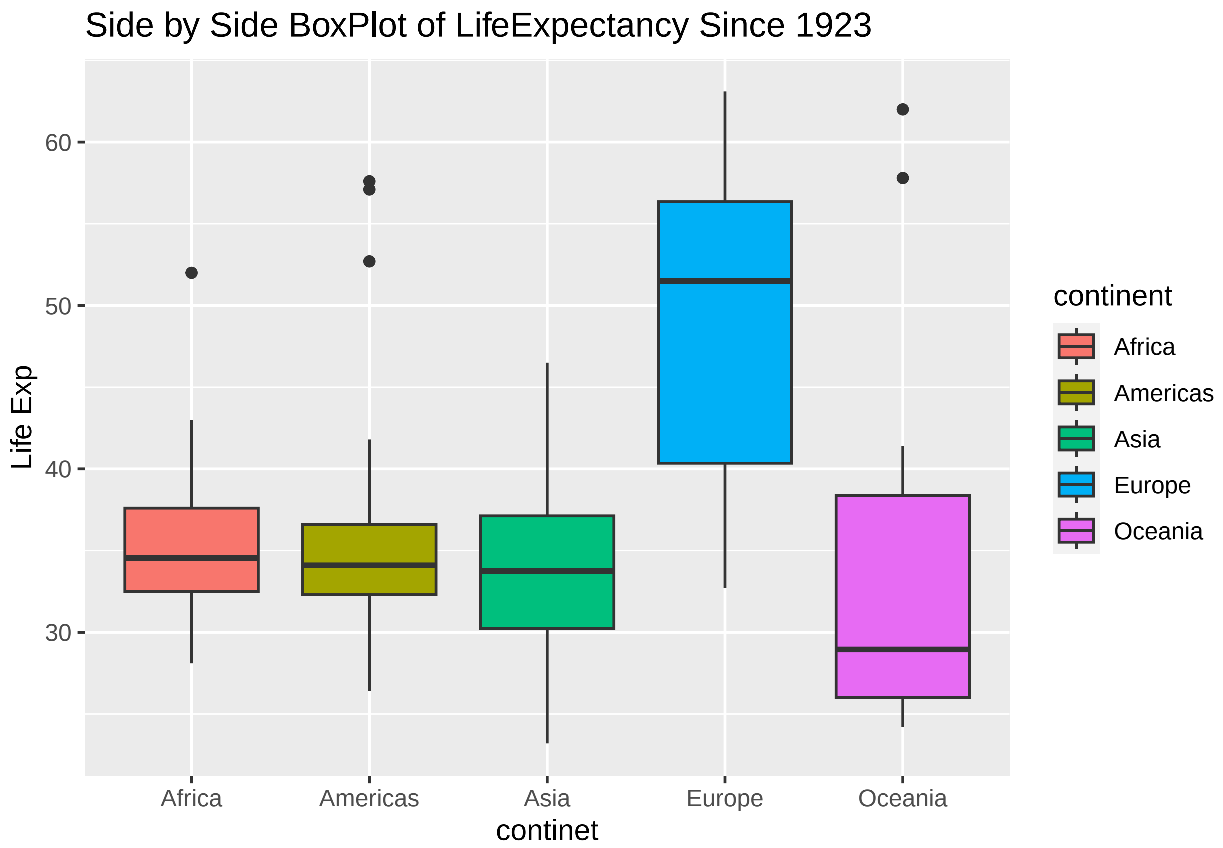
###1.How many countries are there in each continent?

```
world_data = read.csv("Worldlife.csv")
table(world_data$continent)
```

```
##
##   Africa Americas     Asia   Europe  Oceania
##       54       33       50       39       10
```

###2.Have a side-by-side boxplot of life expectancy in 1923 by continent and describe it.

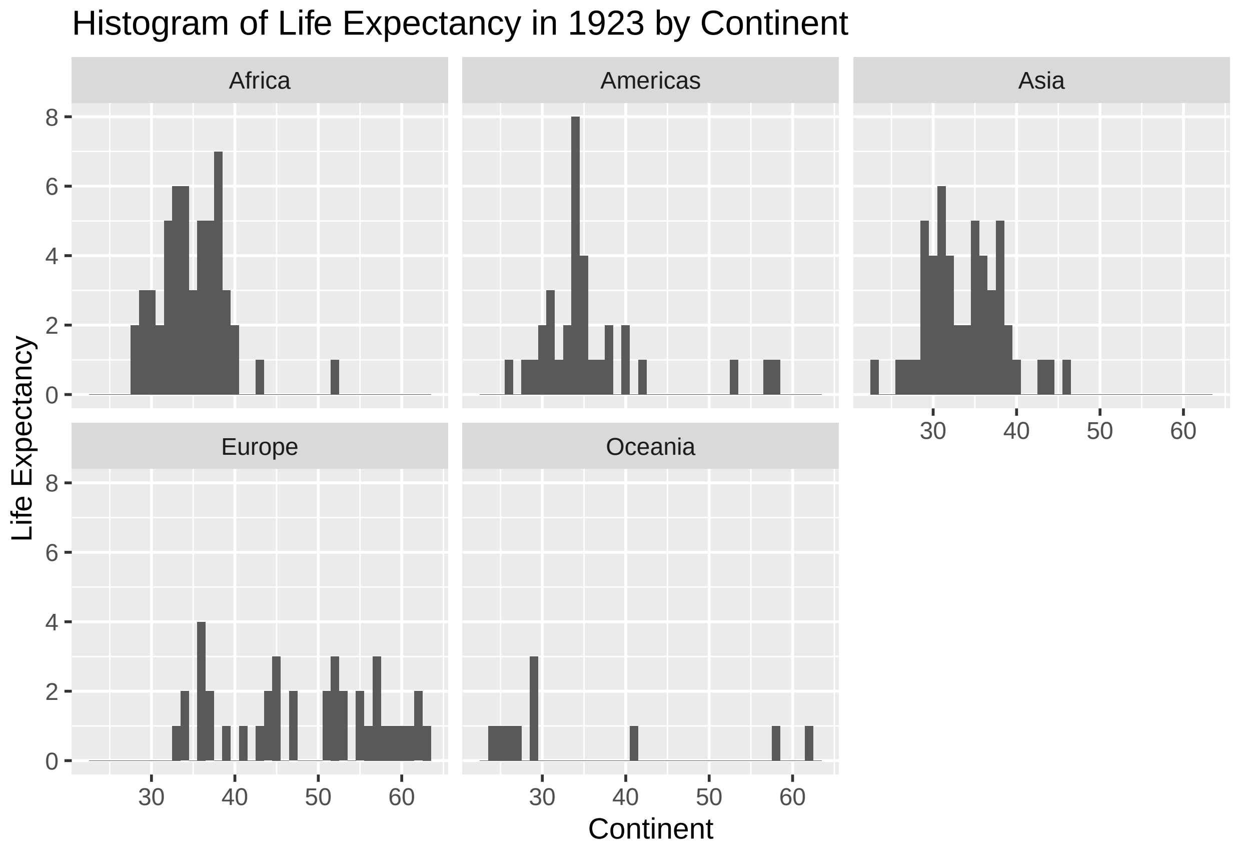
```
ggplot(data = world_data, aes(x=continent, y = life1923, fill = continent)) + geom_boxplot() + xlab("co
```



#All of the continents except Europe lie between 30-40s. Europe lies between 40-55.

###3.Have a histogram of life expectancy in 1923 by continent.

```
ggplot(data = world_data, aes(x = life1923)) + geom_histogram(binwidth = 1) + labs(x = "Continent", y =
```



###4.Have a table summarizing the mean and median of life expectancy in 1923 in each continent.

```
MeanMedian = world_data %>% group_by(continent) %>% summarise(meanData = mean(life1923), medianData = me
MeanMedian
```

```
## # A tibble: 5 x 3
##   continent meanData medianData
##   <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
```

###5.Fit a regression model of life expectancy in 1923 on continent using default reference level. What is the estimated average life expectancy in each continent? Compare the results with the previous summary table. Are there any levels that are insignificant?

```
world_model_1923_default <- lm(life1923 ~ continent, data = world_data)
summary(world_model_1923_default)
```

```
##
## Call:
## lm(formula = life1923 ~ continent, data = world_data)
##
## 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
```

###6.Rerun the regression by using different reference levels.

```
world_data$continent <- as.factor(world_data$continent)
world_data$continent_relevel <- relevel(world_data$continent, ref = "Asia")
model_1923_relevel <- lm(life1923 ~ continent_relevel, data = world_data)
summary(model_1923_relevel)
```

```
## 
## Call:
## lm(formula = life1923 ~ continent_relevel, data = world_data)
## 
## 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)                 33.7940     0.9831  34.376   <2e-16 ***
## continent_relevelAfrica      1.1097     1.3643   0.813    0.417    
## continent_relevelAmericas    2.1454     1.5591   1.376    0.171    
## continent_relevelEurope     14.6445     1.4851   9.861   <2e-16 ***
## continent_relevelOceania     1.2660     2.4080   0.526    0.600    
## ---
## 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
```

###7.If we want to regroup the 5 levels in continent to have a new continent indicator, how will you regroup based on the output previously?

```
world2 <- world_data %>% mutate(ifAsia = ifelse(continent == "Asia", 0, 1))
```

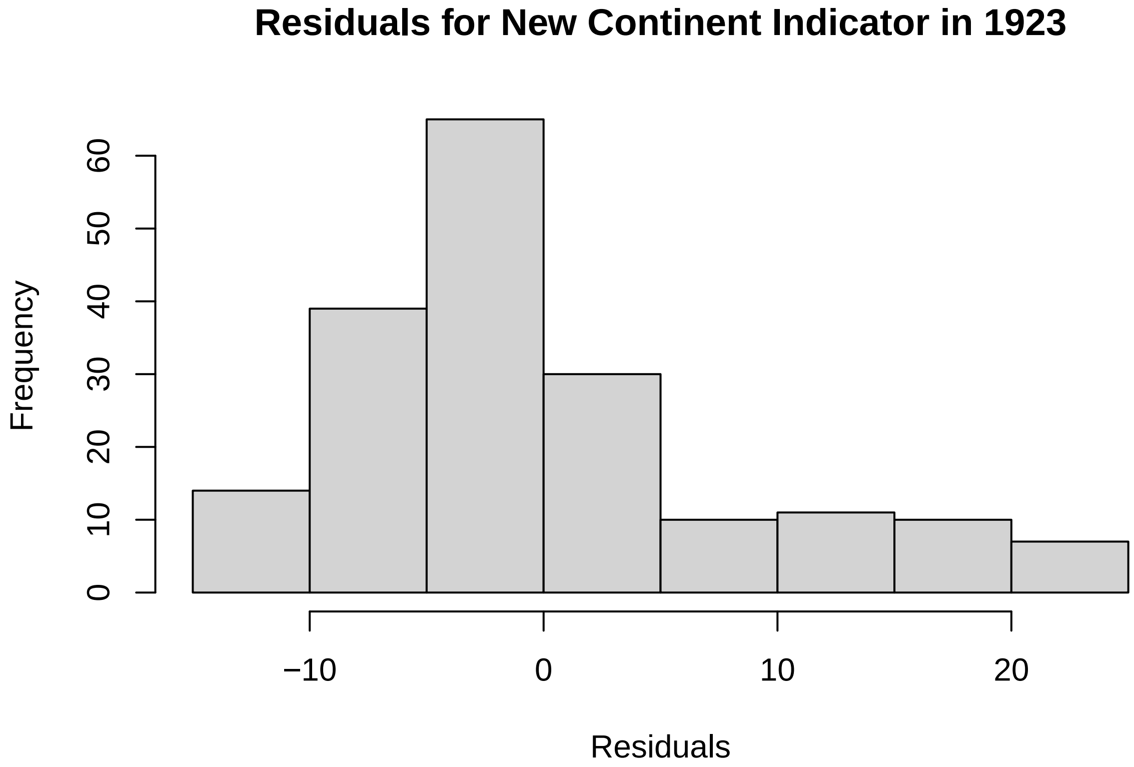
###8.Run the model using your new continent indicator and get the histogram of residual. Describe the residual. What is the percentage of total variability in life expectancy in 1923 that can be explained through the linear model using this new continent indicator?

```
world_data$new_continent_indicator <- ifelse(world_data$continent == "Asia", "Asia", "Others")
model_new_indicator_1923 <- lm(life1923 ~ new_continent_indicator, data = world_data)
summary(model_new_indicator_1923)
```

```
## 
```

```
## Call:
## lm(formula = life1923 ~ new_continent_indicator, data = world_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.848  -5.223  -2.448   3.581  24.052
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      33.794      1.215  27.804  < 2e-16 ***
## new_continent_indicatorOthers     5.254      1.421   3.696 0.000289 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.594 on 184 degrees of freedom
## Multiple R-squared:  0.06912,    Adjusted R-squared:  0.06406
## F-statistic: 13.66 on 1 and 184 DF,  p-value: 0.0002886
```

```r
world_data$residuals_new_indicator_1923 <- resid(model_new_indicator_1923)
hist(world_data$residuals_new_indicator_1923, main = "Residuals for New Continent Indicator in 1923", xl
```



###9.Repeat the previous steps (Q2-Q8) using life expectancy in 2023 as the dependent variable.

###10.Describe whether you see any different patterns happened in these 100 years. #The average life expectancy has went up. This may be due to finding vaccines to illnesses and having a better understanding of life which helps increase life expectancy.