# Homework 2

Tania Ommer

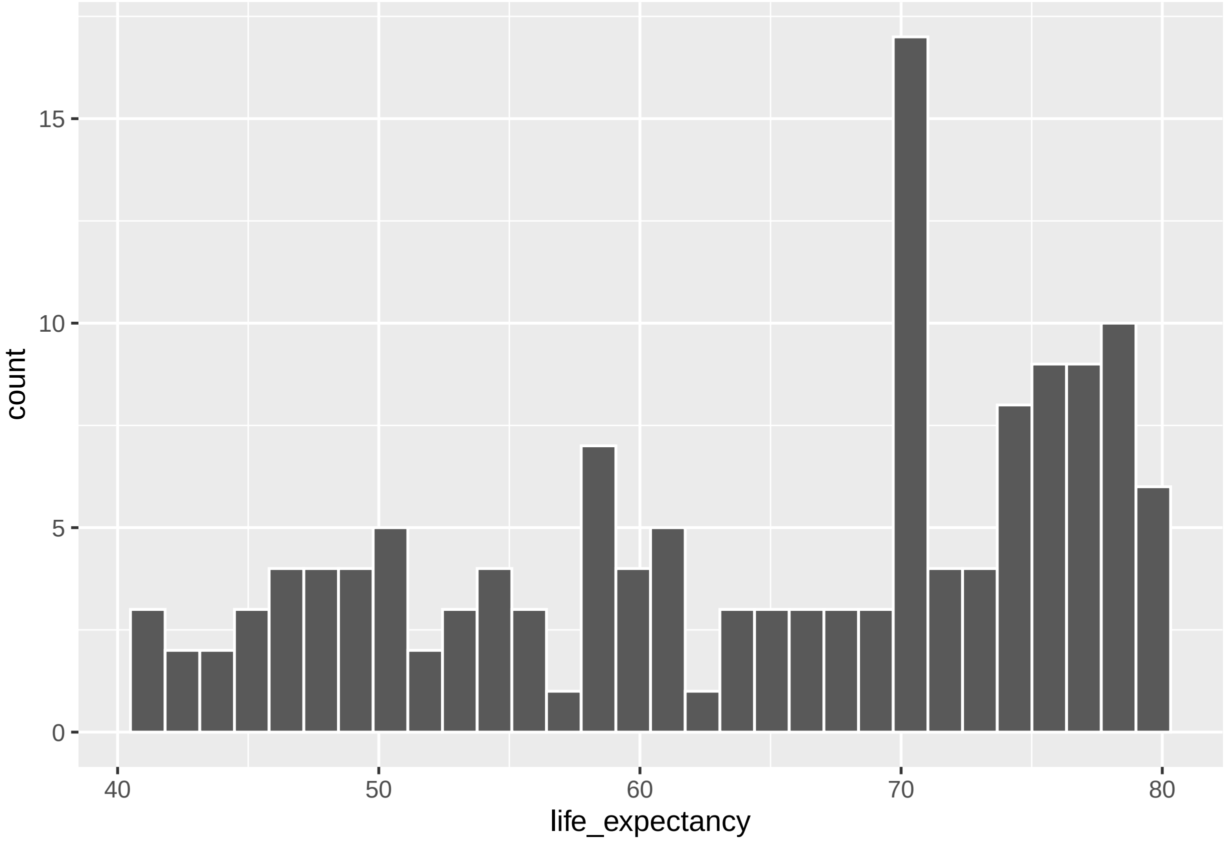
2023-09-26

### Life expectancy in US

1. Have a histogram of the life expectancy, describe the distribution of it

```
ggplot(data = US, mapping = aes(x = life_expectancy)) +
  geom_histogram(color = "white")
```

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



```
summary(US)
```

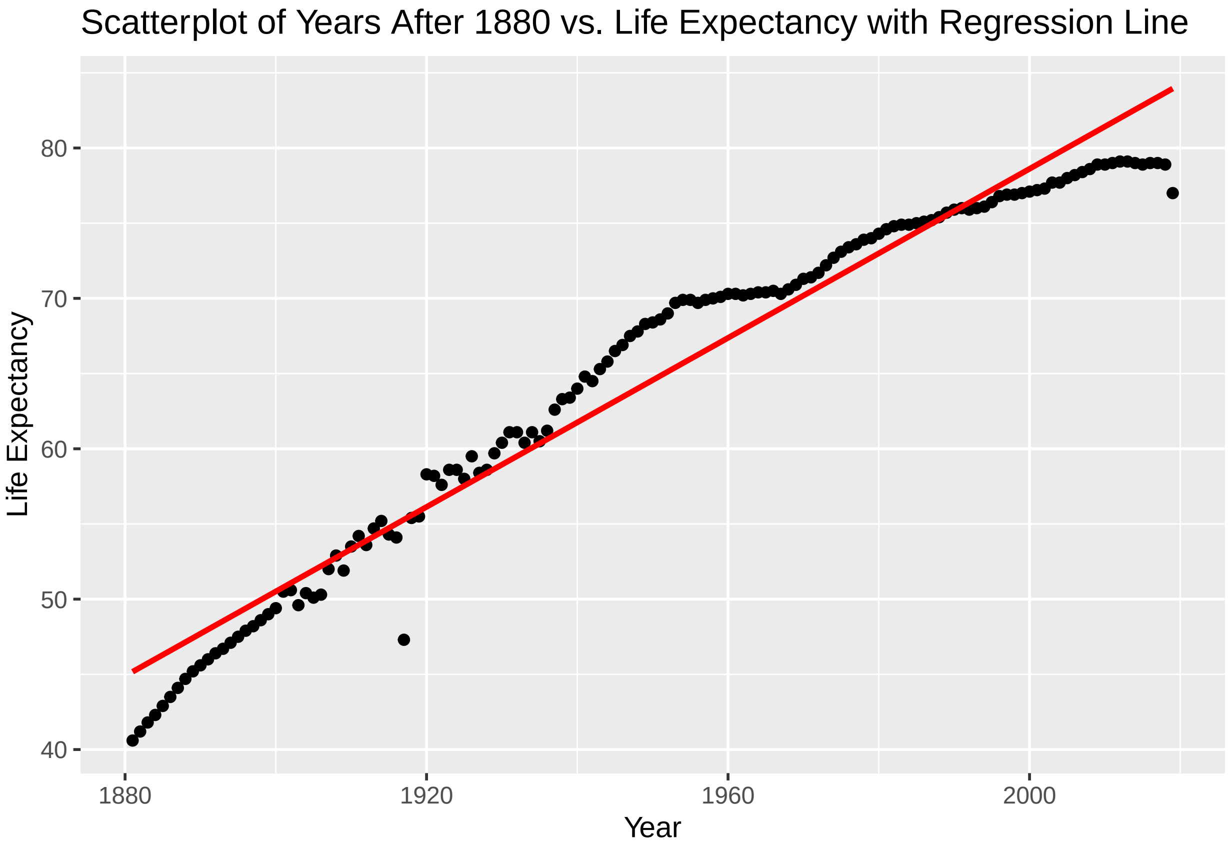
```
##       year      life_expectancy
##  Min.   :1881   Min.   :40.60  
##  1st Qu.:1916   1st Qu.:54.50  
##  Median :1950   Median :68.40  
##  Mean   :1950   Mean   :64.56  
##  3rd Qu.:1984   3rd Qu.:74.95  
##  Max.   :2019   Max.   :79.10
```

The distribution of the histogram of life expectancy is skewed left with the higest peak around the life expectancy value of 70 and several smaller peaks. The life expectancy values range from approximately 40 to 80. The median life expectancy is 68.40 and the mean life expecctancy is 64.56.

2. Does it appear to be some linear relationship between life expectancy and the number of years since 1880 (using the scatterplot)? Is it a positive or negative trend?

```r
ggplot(data = US, mapping = aes(x = year, y = life_expectancy)) +
  geom_point() + geom_smooth(method = "lm", col = "red", se = FALSE) +
  labs(x = "Year", y = "Life Expectancy") +
  ggtitle("Scatterplot of Years After 1880 vs. Life Expectancy with Regression Line")
```

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



Based on the scatterplot above, there appears to be a linear relationship between life expectancy and the number of years since 1880. Both the regression line and the general pattern of points on the scatterplot show the positive relationship the two variables.

3. Are there any unusual points in that trend? What could be the possible reason for that?

There are two outlier points that stray from the positive linear trend between life expectancy and number of years since 1880. One unusual point is right before the year 1920 and the other unusual point is around the year 2020. This could be due to historical events such as pandemics occurring during both 1920 and 2020. The 1920 Spanish flu pandemic as well as the 2020 COVID-19 pandemic both took significant tolls on US life expectancy, causing them to drop, as mortality rates increased in the nation.

4. What is the correlation between life expectancy and number of years since 1880?

```r
correlation <- cor(US$life_expectancy, US$year)
cat("Correlation coefficient:", correlation)
```

```
## Correlation coefficient: 0.9789403
```

The correlation between life expectancy and number of years since 1880 is approximately 0.98. Since the coefficient is very close to 1, there is a strong positive correlation between life expectancy and the number of years since 1880.

5. Run a simple regression. Is the model significant?

```r
regression <- lm(life_expectancy ~ year, data = US)
get_regression_table(regression)
```

```
## # 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 -483.        9.76      -49.5       0 -503.    -464.
## 2 year         0.281     0.005      56.1       0    0.271    0.291
```

```r
summary(regression)
```

```
##
## Call:
## lm(formula = life_expectancy ~ year, data = US)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9900 -1.7413  0.2189  1.5626  4.2937
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.834e+02  9.765e+00  -49.50   <2e-16 ***
## year         2.810e-01  5.007e-03   56.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.368 on 137 degrees of freedom
## Multiple R-squared:  0.9583, Adjusted R-squared:  0.958
## F-statistic:  3150 on 1 and 137 DF,  p-value: < 2.2e-16
```

In the table above, "life expectancy" is the dependent variable and "the number of years since 1880" is the dependent variable. Yes, the model is highly statistically significant because the p-value found in the summary is less than the significance level of 0.5, meaning the null hypothesis was rejected.

6. On average, what is the increase in life expectancy per year?

The average increase in life expectancy per year is approximately 2.810e-01 (0.2810108), as found in the regression model summary above, under the coefficient estimate for "year". The coefficient of 2.810e-01 means that, on average, life expectancy increases by approximately 0.281 years for each additional year after 1880.

7. Predict the life expectancy in year 2021

```r
predict_year <- 2021
predicted_life <- predict(regression, newdata = data.frame(year = predict_year))
cat("Predicted life expectancy in 2021:", predicted_life, "years\n")
```

```
## Predicted life expectancy in 2021: 84.51507 years
```

The predicted life expectancy in 2021 is about 85.51 years. This was predicted using the life expectancy data for the number of years after 1880.

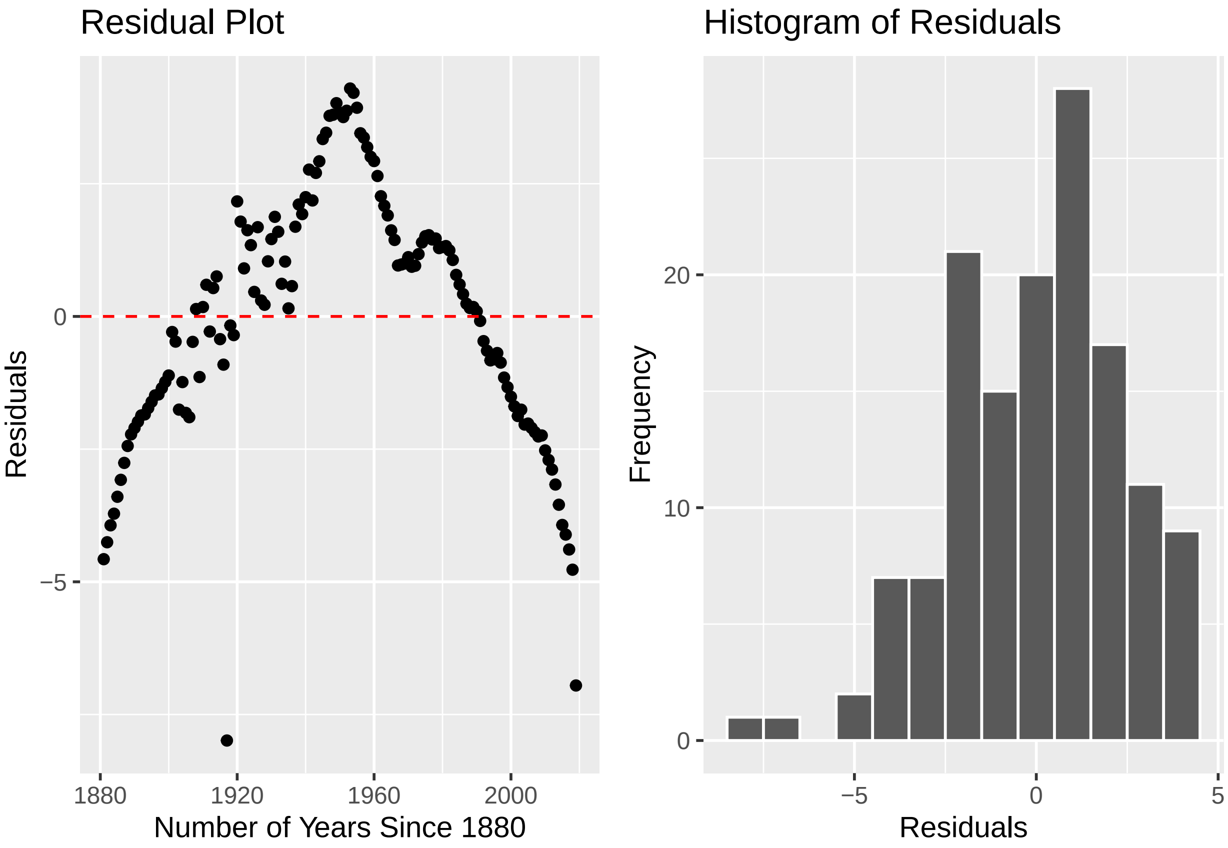
8. Have a residual plot of residual against number of years since 1880, and a histogram of the residual. Describe whether the residual seems to be random, explain why.

```r
residuals <- residuals(regression)

residual_plot <- ggplot(data = data.frame(Year = US$year, Residual = residuals), aes(x = Year, y = Resi
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Residual Plot",
    x = "Number of Years Since 1880",
    y = "Residuals"
  )

histogram_plot <- ggplot(data = data.frame(Residual = residuals), aes(x = Residual)) +
  geom_histogram(binwidth = 1, color = "white") +
  labs(
    title = "Histogram of Residuals",
    x = "Residuals",
    y = "Frequency"
  )

grid.arrange(residual_plot, histogram_plot, ncol = 2)
```



The residual plot does not seem to be random, as the points on the scatterplot appear to follow a trend, rather than being randomly scattered around the dashed line. The histogram of residuals has a distribution that is skewed left with gaps in between the residual values. The plot also appears to have a pattern. Thus, the residuals do not appear to be random.

9. What is the percentage of total variability in life expectancy that can be explained through the linear model using number of years since 1880?

```r
summary(regression)
```

```
##
## Call:
## lm(formula = life_expectancy ~ year, data = US)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9900 -1.7413  0.2189  1.5626  4.2937
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.834e+02  9.765e+00  -49.50   <2e-16 ***
## year         2.810e-01  5.007e-03   56.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.368 on 137 degrees of freedom
## Multiple R-squared:  0.9583, Adjusted R-squared:  0.958
## F-statistic:  3150 on 1 and 137 DF,  p-value: < 2.2e-16
```

```r
r_squared <- summary(regression)$r.squared
percentage <- r_squared * 100

percentage
```

```
## [1] 95.83241
```

If you take the R-squared value from the linear regression summary and multiply it by 100, you will get the percentage of total variability. Therefore, approximately 95.83% of the total variability in life expectancy is explained by the linear model, and the remaining 4.17% represents unexplained variability or random variation.

## Life expectancy in the world - 1919

1. How many countries are there in each continent?

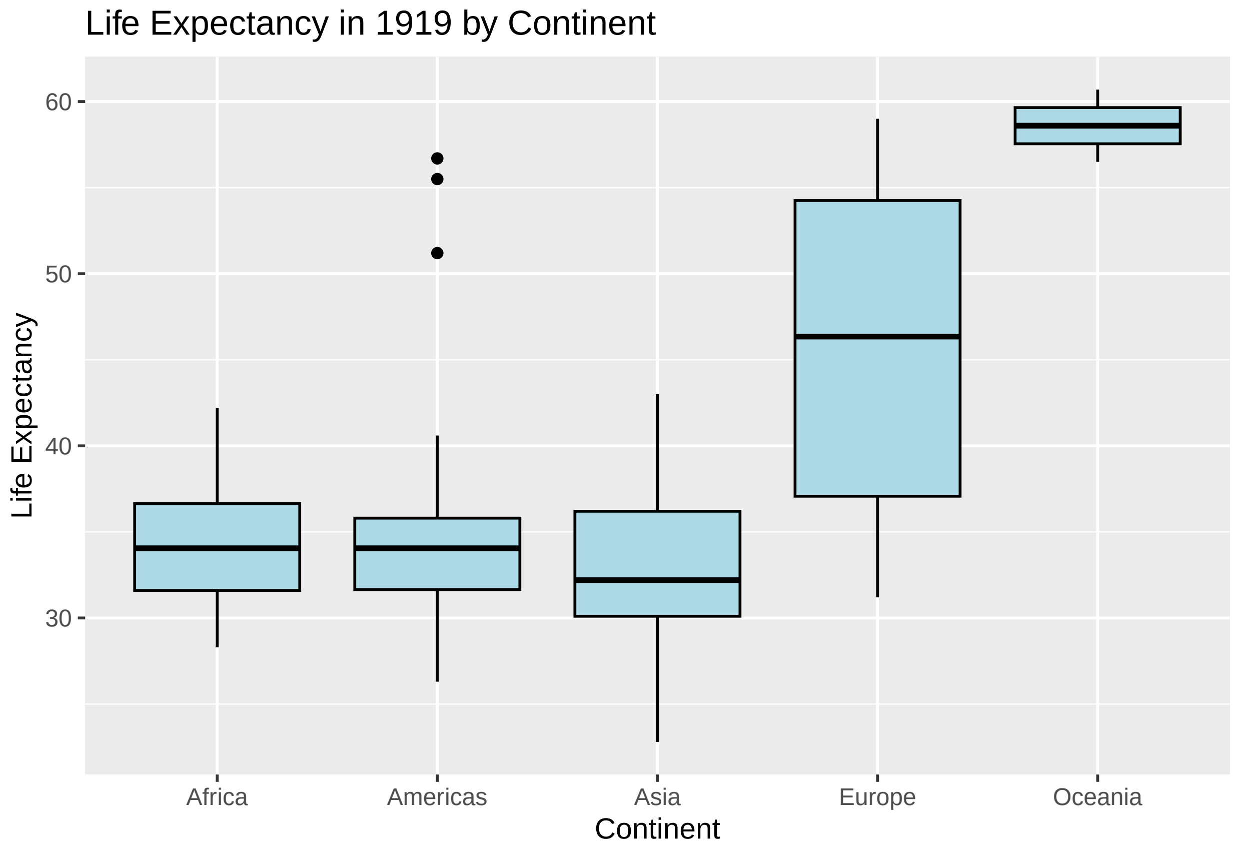
```r
table(World$continent)
```

```
##
##   Africa Americas     Asia   Europe  Oceania
##       50       24       29       30        2
```

The table above shows there are 50 countries in Africa, 24 countries in America, 29 countries in Asia, 30 countries in Europe, and 2 countries in Oceania.

2. Have a side-by-side boxplot of life expectancy in 1919 by continent and describe it.

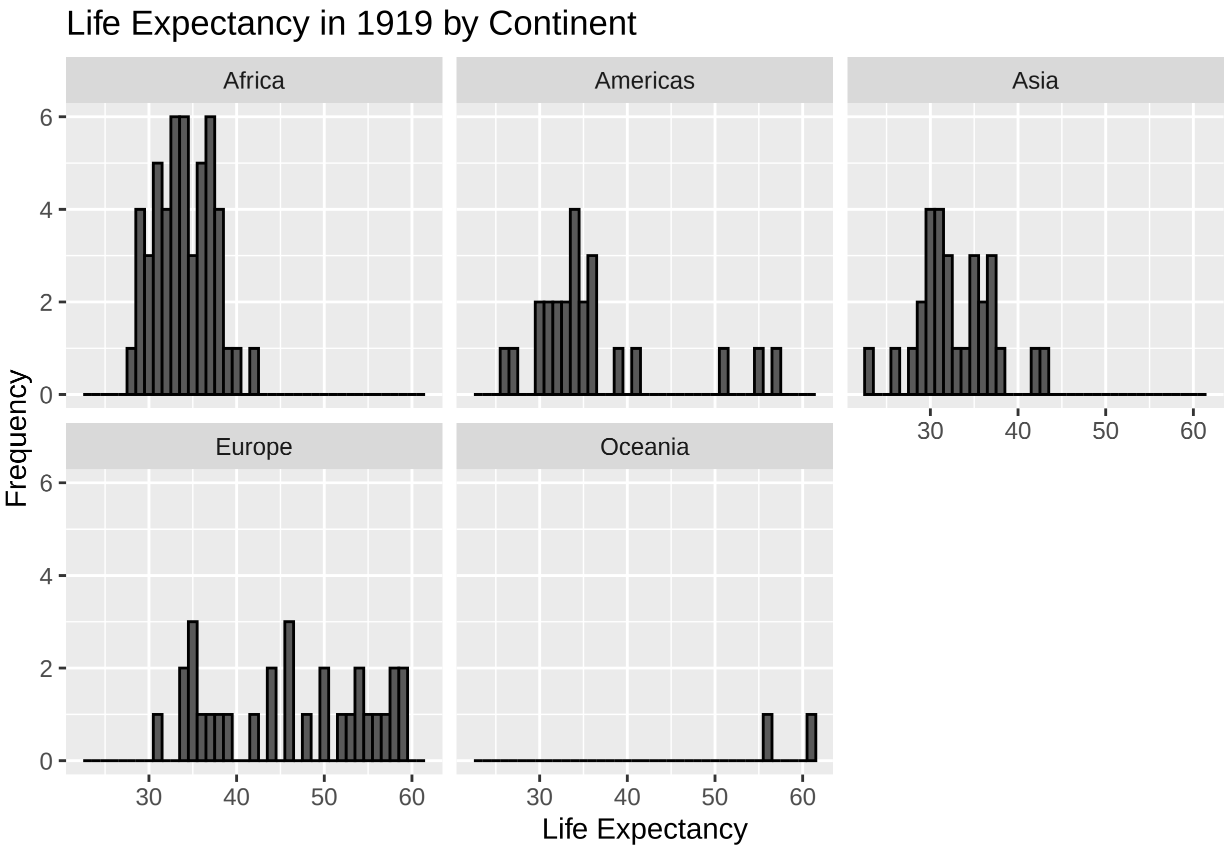
```r
ggplot(data = World, aes(x = continent, y = life1919)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(
    title = "Life Expectancy in 1919 by Continent",
    x = "Continent",
    y = "Life Expectancy"
  )
```



In the side-by-side boxplot above, the distributions vary for each continent. The Americas is the only continent that has outlier points in between the life expectancy value of 50 and 60 years. Africa, Asia, and the Americas all have similar median values at around life expectancy values of 30 to 35 years, while Europe's median life expectancy is around 45 years and Oceania's around 60 years. Oceania has the lowest variability in its data while Europe has the highest variability, as seen by analyzing the whiskers of the plot.

3. Have a histogram of life expectancy in 1919 by continent

```
# Create the histogram
ggplot(data = World, aes(x = life1919)) +
  geom_histogram(binwidth = 1, color = "black") +
  facet_wrap(~ continent, nrow = 2) +
  labs(
    title = "Life Expectancy in 1919 by Continent",
    x = "Life Expectancy",
    y = "Frequency"
  )
```



4. Have a table summarizing the mean and median of life expectancy in 1919 in each continent

```
summary_table <- World %>%
  group_by(continent) %>%
  summarize(mean_life_expectancy = mean(life1919),
            median_life_expectancy = median(life1919))
kable(summary_table)
```

| continent | mean_life_expectancy | median_life_expectancy |
|---|---:|---:|
| Africa | 34.10600 | 34.05 |
| Americas | 35.91667 | 34.05 |
| Asia | 33.01379 | 32.20 |
| Europe | 46.18667 | 46.35 |
| Oceania | 58.60000 | 58.60 |

5. Fit a regression model of life expectancy in 1919 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?

```
regression_life <- lm(life1919 ~ continent, data = World)
summary(regression_life)
```

```
##
## Call:
## lm(formula = life1919 ~ continent, data = World)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -14.9867  -3.2113  -0.8138   3.2940  20.7833 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        34.1060     0.8649  39.432  < 2e-16 ***
## continentAmericas   1.8107     1.5188   1.192    0.235    
## continentAsia      -1.0922     1.4276  -0.765    0.446    
## continentEurope    12.0807     1.4124   8.553 2.87e-14 ***
## continentOceania   24.4940     4.4103   5.554 1.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.116 on 130 degrees of freedom
## Multiple R-squared:  0.473,  Adjusted R-squared:  0.4568 
## F-statistic: 29.17 on 4 and 130 DF,  p-value: < 2.2e-16
```

```
get_regression_table(regression_life)
```

```
## # 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              34.1      0.865    39.4     0        32.4     35.8 
## 2 continent: Americas     1.81     1.52      1.19    0.235    -1.19     4.82
## 3 continent: Asia        -1.09     1.43     -0.765   0.446    -3.92     1.73
## 4 continent: Europe      12.1      1.41      8.55    0         9.29    14.9 
## 5 continent: Oceania     24.5      4.41      5.55    0        15.8     33.2
```

```
# create new data frame with continents
continents <- data.frame(continent = unique(World$continent))

# predict the average life expectancy for each continent
continents$Estimated_Avg_Life_Expectancy <- predict(regression_life, newdata = continents)
print(continents)
```

```
##   continent Estimated_Avg_Life_Expectancy
## 1      Asia                      33.01379
## 2    Europe                      46.18667
## 3    Africa                      34.10600
## 4  Americas                      35.91667
## 5   Oceania                      58.60000
```

Since the regression model above was fit using the default reference level, the coefficient for "Africa" is serving as the reference level for the "continent" variable. The coefficients for the rest of the continents in the model represent the differences in average life expectancy compared to the reference level, Africa. The table created above shows the estimated average life expectancies in each continent.

Based off the regression model summary and table, Asia and the Americas are both not statistically significant since they both have p-values greater than the default p-value of 0.05. To represent this in the model, there are no significance codes asterisks present next to the p-values for Asia and America. Eurpose, Oceania, and the reference level Africa all have low p-values and are statistically significant, as marked by the asterisks.

6. Rerun the regression by using different reference levels

```
# converted to unordered levels
World$continent <- factor(World$continent)
```

```
World$continent <- relevel(World$continent, ref = "Asia")
new_model1 <- lm(life1919 ~ continent, data = World)
get_regression_table(new_model1)
```

```
## # 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              33.0       1.14    29.1     0        30.8      35.3
## 2 continent: Africa       1.09      1.43     0.765   0.446   -1.73       3.92
## 3 continent: Americas     2.90      1.69     1.72    0.088   -0.436      6.24
## 4 continent: Europe      13.2       1.59     8.27    0        10.0      16.3
## 5 continent: Oceania     25.6       4.47     5.72    0        16.7      34.4
```

```
World$continent <- relevel(World$continent, ref = "Europe")
new_model2 <- lm(life1919 ~ continent, data = World)
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               46.2      1.12     41.4    0         44.0     48.4
## 2 continent: Asia        -13.2      1.59     -8.27   0        -16.3    -10.0
## 3 continent: Africa      -12.1      1.41     -8.55   0        -14.9     -9.29
## 4 continent: Americas    -10.3      1.68     -6.13   0        -13.6     -6.96
## 5 continent: Oceania      12.4      4.47      2.78   0.006     3.58     21.2
```

```
World$continent <- relevel(World$continent, ref = "Americas")
new_model3 <- lm(life1919 ~ continent, data = World)
get_regression_table(new_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             35.9       1.25     28.8    0        33.4    38.4
## 2 continent: Europe     10.3       1.68      6.13   0         6.96   13.6
## 3 continent: Asia       -2.90      1.69     -1.72   0.088    -6.24    0.436
## 4 continent: Africa     -1.81      1.52     -1.19   0.235    -4.82    1.19
## 5 continent: Oceania    22.7       4.50      5.04   0        13.8    31.6
```

```
World$continent <- relevel(World$continent, ref = "Oceania")
new_model4 <- lm(life1919 ~ continent, data = World)
get_regression_table(new_model4)
```

```
## # 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               58.6      4.32     13.6    0          50.0    67.2
## 2 continent: Americas    -22.7      4.50     -5.04   0         -31.6   -13.8
## 3 continent: Europe      -12.4      4.47     -2.78   0.006     -21.2    -3.58
## 4 continent: Asia        -25.6      4.47     -5.72   0         -34.4   -16.7
## 5 continent: Africa      -24.5      4.41     -5.55   0         -33.2   -15.8
```

Above are regression model tables with varying reference level. Reference levels used above are Asia, Europe, Americas, and Oceania. A regression model table with Africa as the reference level was not created, as it was already made in the previous question, where a model was constructed using the default reference level (Africa).

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?

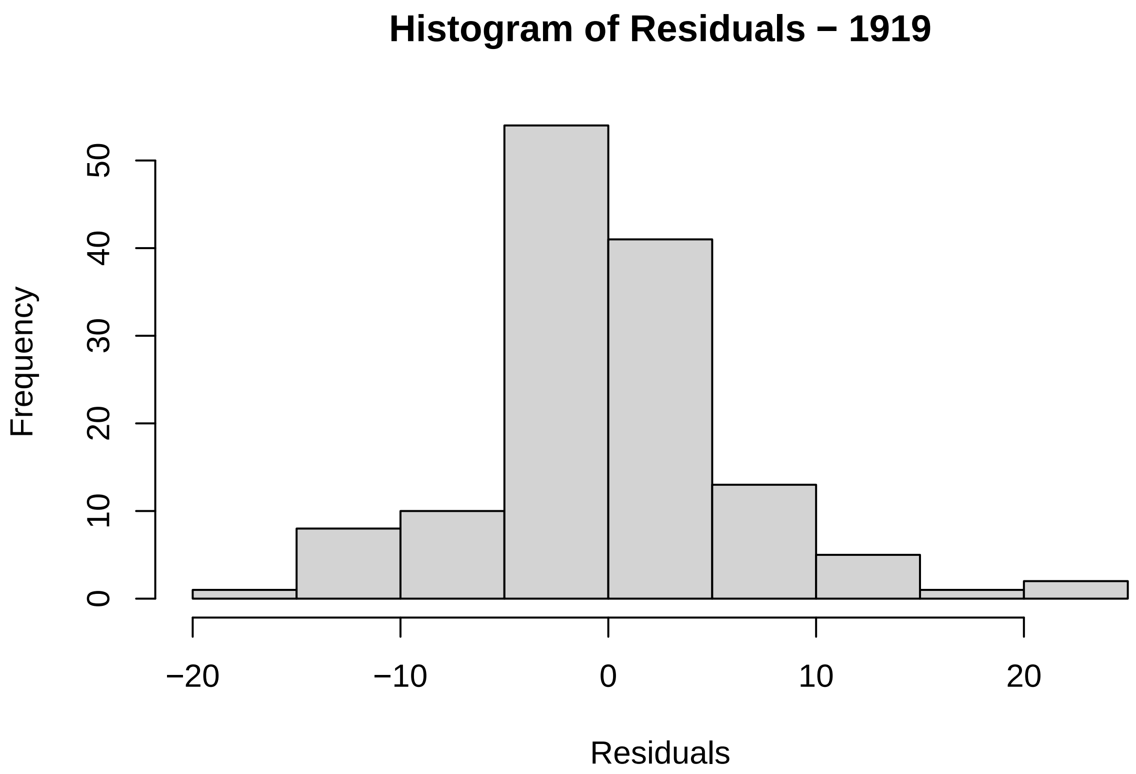
I would regroup the continents based off their coefficient values. Since Europe and Oceania have the highest coefficient values, I would group them together as “High Life Expectancy”. The rest of the continents (Asia, Africa, and the Americas) have much lower coefficient values and would be grouped as “Low Life Expectancy”.

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 that can be explained through the linear model using this new continent indicator?

```
# new variable 'new_continent' based on coefficients
World$new_continent <- factor(ifelse(World$continent %in% c("Europe", "Oceania"), "High Life Expectancy
model_new_continent <- lm(life1919 ~ new_continent, data = World)
summary(model_new_continent)
```

```
##
## Call:
## lm(formula = life1919 ~ new_continent, data = World)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7625  -3.4704  -0.3204   2.9086  22.4796
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        46.962      1.112   42.23   <2e-16 ***
## new_continentLow Life Expectancy  -12.742      1.273  -10.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.291 on 133 degrees of freedom
## Multiple R-squared:  0.4296, Adjusted R-squared:  0.4253
## F-statistic: 100.2 on 1 and 133 DF,  p-value: < 2.2e-16
```

```
residuals <- residuals(model_new_continent)
hist(residuals, main = "Histogram of Residuals - 1919", xlab = "Residuals")
```



```r
r_squared <- summary(model_new_continent)$r.squared
percentage <- r_squared * 100
cat("Percentage of Total Variability:", percentage)
```

```
## Percentage of Total Variability: 42.95822
```

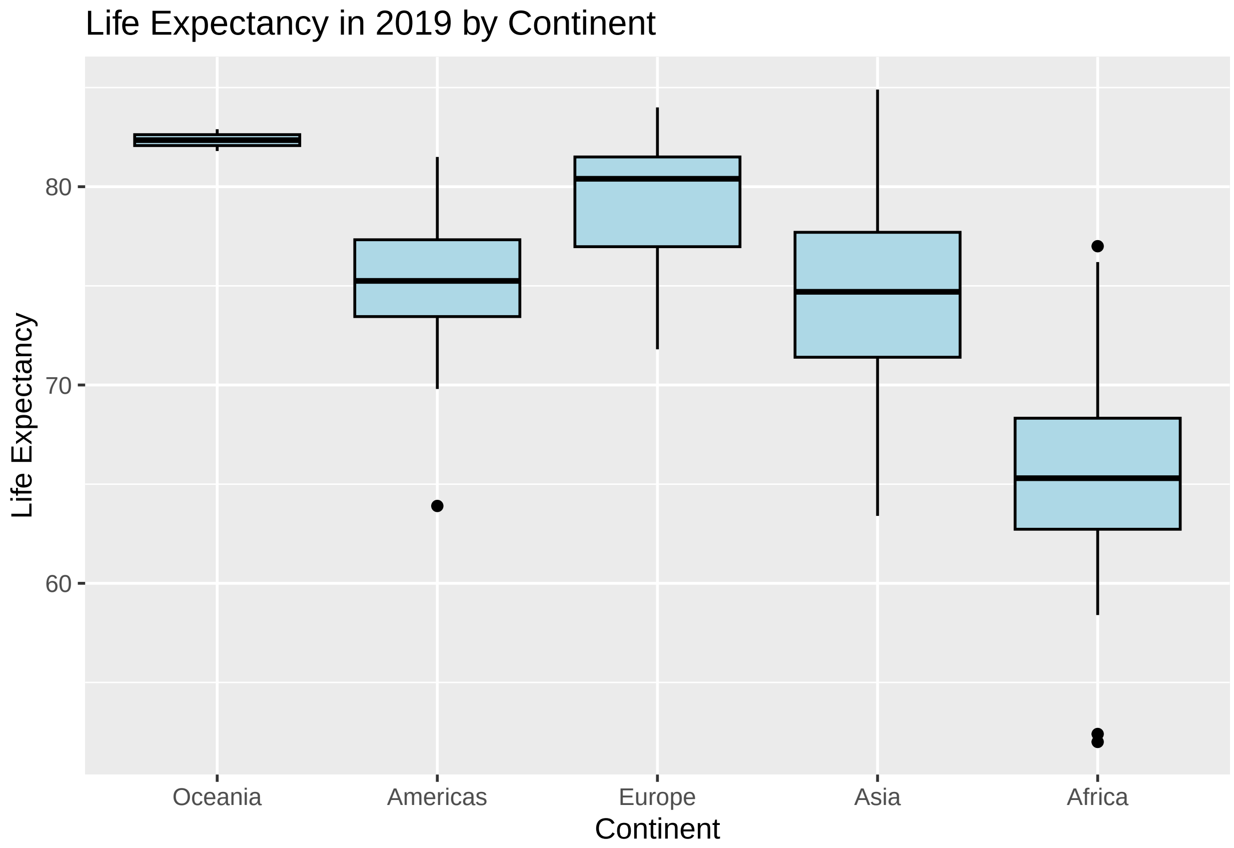
The histogram of the Residuals follows a normal distribution. The distribution is symmetric, as the tallest peak is at 0 with equivalent sized smaller peaks on both sides. The mean of the Residuals is very close to 0, indicating that the model is an accurate predictor.

The percentage of total variability in life expectancy that can be explained through the linear model using this new continent indicator is about 42.96%.

### Life expectancy in the world - 2019

2. Have a side-by-side boxplot of life expectancy in 2019 by continent and describe it.

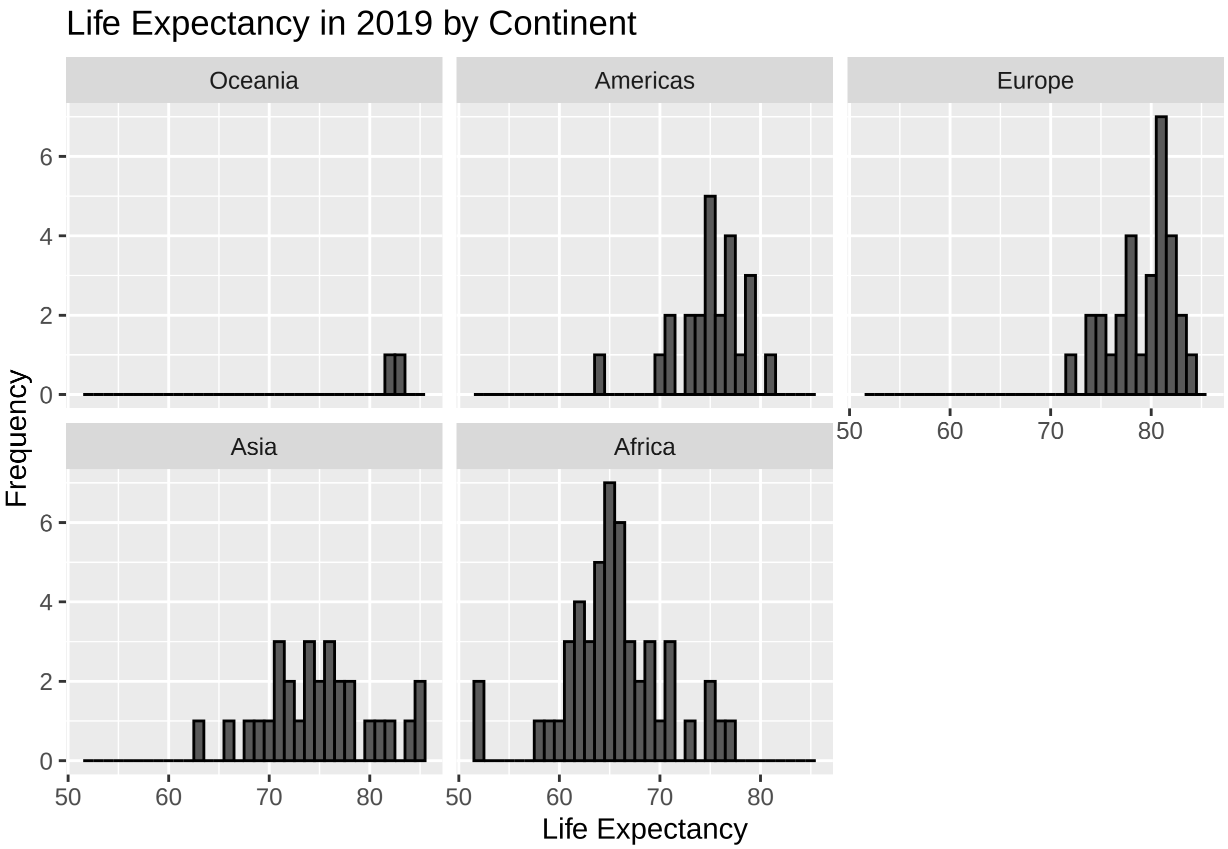
```r
ggplot(data = World, aes(x = continent, y = life2019)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(
    title = "Life Expectancy in 2019 by Continent",
    x = "Continent",
    y = "Life Expectancy"
  )
```



In the side-by-side boxplot above, the distributions vary for each continent. The Americas and Africa both have outlier points. The Americas' outlier point is in between the life expectancy value of 60 and 70 years and Africa's outlier points are between 50 and 60 years, and 70 and 80 years. Europe, Asia, and the Americas all have similar median values at around life expectancy values of 70 to 80 years, while Africa's median life expectancy is around 65 years and Oceania's around 80 years. Oceania has the lowest variability in its data while Asia has the highest variability, as seen by analyzing the whiskers of the plot.

3. Have a histogram of life expectancy in 2019 by continent

```
# Create the histogram
ggplot(data = World, aes(x = life2019)) +
  geom_histogram(binwidth = 1, color = "black") +
  facet_wrap(~ continent, nrow = 2) +
  labs(
    title = "Life Expectancy in 2019 by Continent",
    x = "Life Expectancy",
    y = "Frequency"
  )
```



4. Have a table summarizing the mean and median of life expectancy in 2019 in each continent

```
summary_table <- World %>%
  group_by(continent) %>%
  summarize(mean_life_expectancy = mean(life2019),
            median_life_expectancy = median(life2019))
kable(summary_table)
```

| continent | mean_life_expectancy | median_life_expectancy |
|---|---:|---:|
| Oceania | 82.35000 | 82.35 |
| Americas | 75.02917 | 75.25 |
| Europe | 79.31000 | 80.40 |
| Asia | 75.05517 | 74.70 |
| Africa | 65.58800 | 65.30 |

5. Fit a regression model of life expectancy in 2019 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?

```
regression_life <- lm(life2019 ~ continent, data = World)
summary(regression_life)
```

```
##
## Call:
## lm(formula = life2019 ~ continent, data = World)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -13.5880  -2.6100   0.0708   2.3208  11.4120 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         82.350      3.226  25.527  < 2e-16 ***
## continentAmericas   -7.321      3.358  -2.180   0.0310 *  
## continentEurope     -3.040      3.332  -0.912   0.3632    
## continentAsia       -7.295      3.335  -2.187   0.0305 *  
## continentAfrica    -16.762      3.290  -5.095  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.562 on 130 degrees of freedom
## Multiple R-squared:  0.6146, Adjusted R-squared:  0.6028 
## F-statistic: 51.84 on 4 and 130 DF,  p-value: < 2.2e-16
```

```r
get_regression_table(regression_life)
```

```
## # 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              82.4       3.23    25.5     0         76.0    88.7  
## 2 continent: Americas    -7.32      3.36    -2.18    0.031    -14.0    -0.678
## 3 continent: Europe      -3.04      3.33    -0.912   0.363     -9.63    3.55 
## 4 continent: Asia        -7.30      3.34    -2.19    0.031    -13.9    -0.696
## 5 continent: Africa     -16.8       3.29    -5.10    0        -23.3   -10.3  
```

```r
# create new data frame with continents
continents <- data.frame(continent = unique(World$continent))

# predict the average life expectancy for each continent
continents$Estimated_Avg_Life_Expectancy <- predict(regression_life, newdata = continents)
print(continents)
```

```
##   continent Estimated_Avg_Life_Expectancy
## 1      Asia                      75.05517
## 2    Europe                      79.31000
## 3    Africa                      65.58800
## 4  Americas                      75.02917
## 5   Oceania                      82.35000
```

Since the regression model above was fit using the default reference level, the coefficient for “Oceania” is serving as the reference level for the “continent” variable. The coefficients for the rest of the continents in the model represent the differences in average life expectancy compared to the reference level, Oceania The table created above shows the estimated average life expectancies in each continent.

Based off the regression model summary and table, Europe not statistically significant since it has a p-value greater than the default p-value of 0.05. To represent this in the model, there is no significance codes asterisks present next to the p-value for Europe. Asia, Africa, the Americas and the reference level Oceania all have low p-values and are statistically significant, as marked by the asterisks.

6. Rerun the regression by using different reference levels

```r
# converted to unordered levels
World$continent <- factor(World$continent)
```

```r
World$continent <- relevel(World$continent, ref = "Asia")
new_model1 <- lm(life2019 ~ continent, data = World)
get_regression_table(new_model1)
```

```
## # 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             75.1       0.847    88.6     0        73.4     76.7
## 2 continent: Oceania     7.30      3.34      2.19    0.031    0.696    13.9
## 3 continent: Americas   -0.026     1.26     -0.021   0.984   -2.52      2.46
## 4 continent: Europe      4.26      1.19      3.58    0        1.90      6.60
## 5 continent: Africa     -9.47      1.06     -8.89    0      -11.6      -7.36
```

```r
World$continent <- relevel(World$continent, ref = "Europe")
new_model2 <- lm(life2019 ~ continent, data = World)
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              79.3      0.833    95.2     0         77.7    81.0
## 2 continent: Asia        -4.26     1.19     -3.58    0         -6.60   -1.90
## 3 continent: Oceania      3.04     3.33      0.912   0.363     -3.55    9.63
## 4 continent: Americas    -4.28     1.25     -3.43    0.001     -6.75   -1.81
## 5 continent: Africa     -13.7      1.05    -13.0     0        -15.8   -11.6
```

```r
World$continent <- relevel(World$continent, ref = "Americas")
new_model3 <- lm(life2019 ~ continent, data = World)
get_regression_table(new_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            75.0       0.931    80.6     0        73.2      76.9
## 2 continent: Europe     4.28      1.25      3.43    0.001     1.81      6.75
## 3 continent: Asia       0.026     1.26      0.021   0.984   -2.46      2.52
## 4 continent: Oceania    7.32      3.36      2.18    0.031     0.678    14.0
## 5 continent: Africa    -9.44      1.13     -8.33    0      -11.7      -7.2
```

```r
World$continent <- relevel(World$continent, ref = "Africa")
new_model4 <- lm(life2019 ~ continent, data = World)
get_regression_table(new_model4)
```

```
## # 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              65.6      0.645    102.         0    64.3      66.9
## 2 continent: Americas     9.44     1.13       8.33       0     7.2      11.7
## 3 continent: Europe      13.7      1.05      13.0        0    11.6      15.8
## 4 continent: Asia         9.47     1.06       8.89       0     7.36     11.6
## 5 continent: Oceania     16.8      3.29       5.10       0    10.3      23.3
```

Above are regression model tables with varying reference level. Reference levels used above are Asia, Europe, Americas, and Africa. A regression model table with Oceania as the reference level was not created, as it was already made in the previous question, where a model was constructed using the default reference level (Oceania).

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?

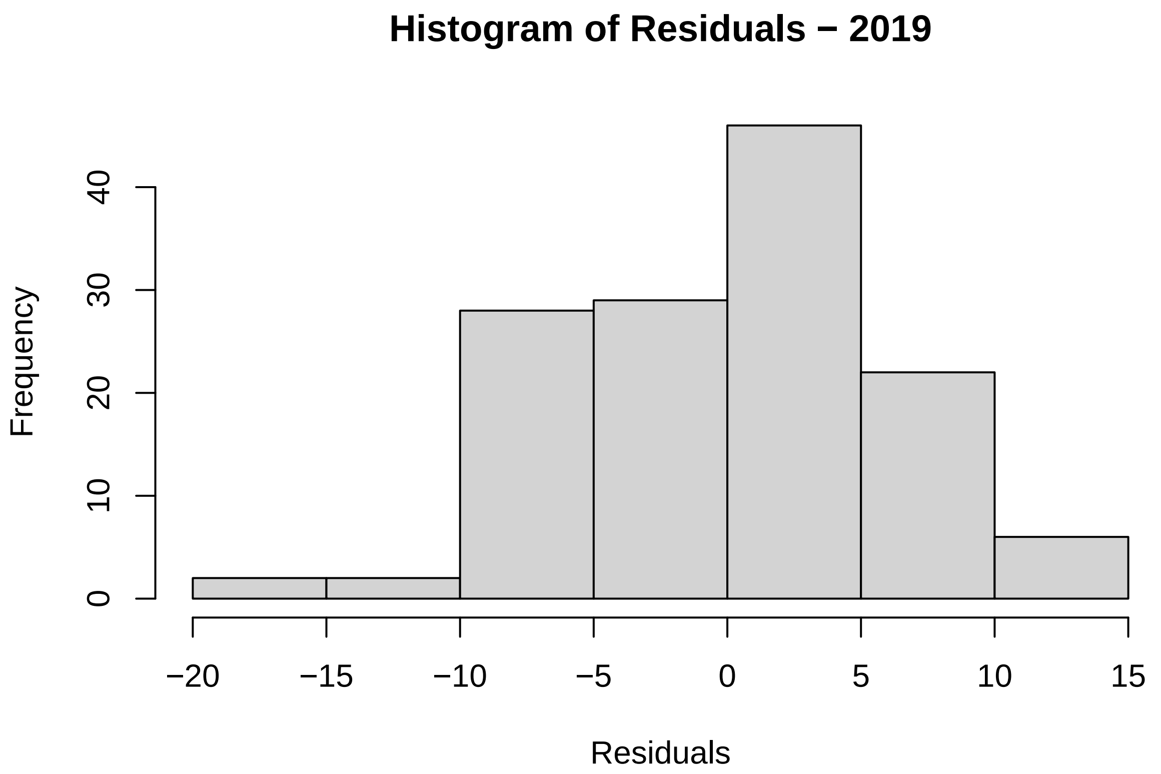
I would regroup the continents based off their coefficient values. Since Europe and Oceania have the highest coefficient values, I would group them together as “High Life Expectancy”. The rest of the continents (Asia, Africa, and the Americas) have much lower coefficient values and would be grouped as “Low Life Expectancy”.

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 that can be explained through the linear model using this new continent indicator?

```
#life2019 as dependent variable
model_2019 <- lm(life2019 ~ new_continent, data = World)
summary(model_2019)
```

```
##
## Call:
## lm(formula = life2019 ~ new_continent, data = World)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4534  -4.4534   0.6466   4.2966  14.4466
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         79.500      1.086  73.176  < 2e-16 ***
## new_continentLow Life Expectancy    -9.047      1.244  -7.273 2.71e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.146 on 133 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.2792
## F-statistic:  52.9 on 1 and 133 DF,  p-value: 2.71e-11
```

```
residual_2019 <- residuals(model_2019)
hist(residual_2019, main = "Histogram of Residuals - 2019", xlab = "Residuals")
```



```
r_squared_2019 <- summary(model_2019)$r.squared
percentage_2019 <- r_squared_2019 * 100
cat("Percentage of Total Variability:", percentage_2019)
```

```
## Percentage of Total Variability: 28.45715
```

The histogram of the Residuals follows a normal distribution. The distribution is slightly skewed left with the highest peak being in between the residual values 0 and 5. The mean of the Residuals is very close to 0, indicating that the model is an accurate predictor.

The percentage of total variability in life expectancy that can be explained through the linear model using this new continent indicator is about 28.46%.

10. Describe whether you see any difference happened in these 100 years.

The coefficients for the "Low Life Expectancy" category are negative in 1919 and 2019, indicating that the continents grouped under "Low Life Expectancy" did tend to have significantly lower life expectancy relative to the reference group ("High Life Expectancy"). Since the R-squared value in 1919 (43.96%) is higher compared to 2019 (28.46%), the linear regression model that uses "new_continent" grouping variable as a predictor explains a larger proportion of the variability in life expectancy in 1919 compared to 2019. This explains why although the "Low Life Expectancy" group had consistently low expectancies in both 1919 and 2019, the R-squared percentage value is higher in 1919, as the continent grouping was a stronger predictor for this year. When looking at the histograms for 1919 and 2019, it is clear that many of the countries around the world have significantly higher life expectancies in 2019. This could be due to new vaccinations, health precautions, and overall boosted economic landscapes in several countries over the years.