```
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
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.3.2
```

```
library(nortest)
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

```
## Loading required package: lattice
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
file <- read.csv("life_data.csv")
```

Distribution of Life Expectancy Over Countries

We consider data which tells the life expectancy in each country of the world over several years. To be more specific, data set provides information from 2000 to 2015. This will give us more precise results and make them less biased.

Let us start with some basic information in order to get known to what we are dealing with.

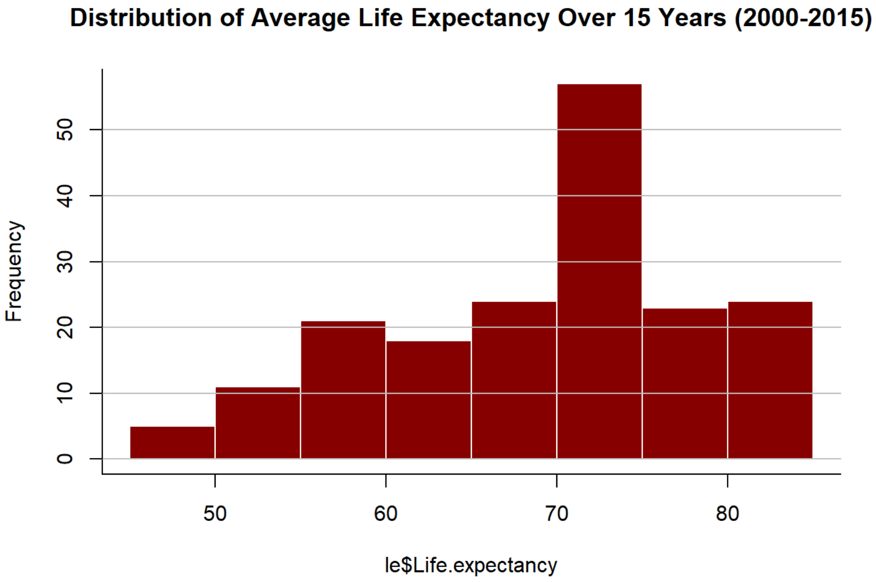
```
# Display the average life expectancy of each country over 15 years
le <- file %>%
  group_by(Country) %>%
  summarise(Life.expectancy = mean(Life.expectancy))

le
```

```
## # A tibble: 193 × 2
##    Country             Life.expectancy
##    <chr>                         <dbl>
##  1 Afghanistan                    58.2
##  2 Albania                        75.2
##  3 Algeria                        73.6
##  4 Angola                         49.0
##  5 Antigua and Barbuda            75.1
##  6 Argentina                      75.2
##  7 Armenia                        73.4
##  8 Australia                      81.8
##  9 Austria                        81.5
## 10 Azerbaijan                     70.7
## # ℹ 183 more rows
```

As we can see, the majority lies between 60-70 years, although there are countries where average life expectancy is ~50-60 years or even 80+ years.

By plotting a histogram of the average life expectancy, we can say for sure what value is the most “popular” among all other.

```
hist(le$Life.expectancy,
     col = "darkred",
     border = "white",
     main = "Distribution of Average Life Expectancy Over 15 Years (2000-2015)",
     breaks = c(45, 50, 55, 60, 65, 70, 75, 80, 85)
)

box(bty = "l")

grid(nx = NA, ny = NULL, lty = 1, lwd = 1, col = "gray")
```



Classification

For further analysis, we will divide data into two categories: “long-living” (life expectancy >= 70) and “short-living” (life expectancy < 70).

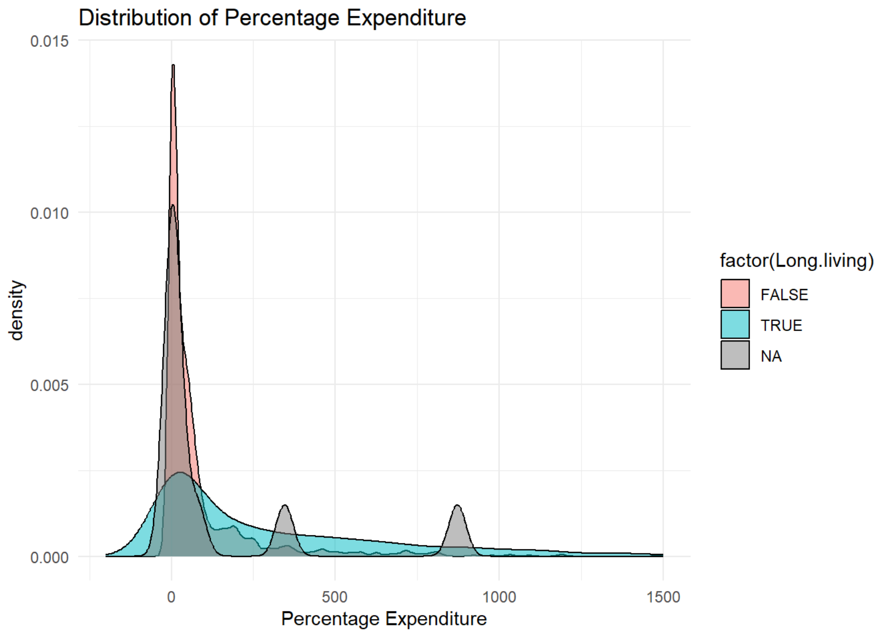
```
le$Long.living <- le$Life.expectancy >= 70
file$Long.living <- file$Life.expectancy >= 70

summary(le)
```

```
##    Country          Life.expectancy Long.living    
##  Length:193         Min.   :46.11   Mode :logical  
##  Class :character   1st Qu.:62.48   FALSE:79       
##  Mode  :character   Median :72.49   TRUE :104      
##                     Mean   :69.22   NA's :10       
##                     3rd Qu.:75.11                  
##                     Max.   :82.54                  
##                     NA's   :10
```

Influence of Expenditures on Life Expectancy

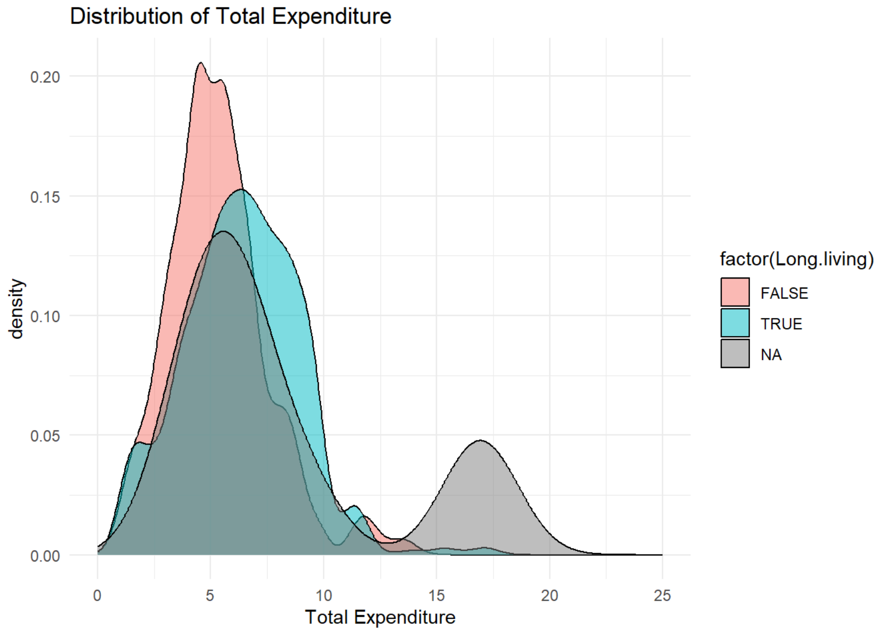
```
ggplot(
  file,
  aes(
    x = percentage.expenditure,
    fill = factor(Long.living)
  )
) + geom_density(alpha=0.5) +
      xlab(label = "Percentage Expenditure") +
      ggtitle("Distribution of Percentage Expenditure") +
      theme_minimal() + xlim(-200, 1500)
```

```
## Warning: Removed 328 rows containing non-finite values (`stat_density()`).
```



```
ggplot(
  file,
  aes(
    x = Total.expenditure,
    fill = factor(Long.living)
  )
) + geom_density(alpha=0.5) +
      xlab(label = "Total Expenditure") +
      ggtitle("Distribution of Total Expenditure") +
      theme_minimal() + xlim(0, 25)
```

```
## Warning: Removed 226 rows containing non-finite values (`stat_density()`).
```



TASK 1

Firstly, we want to see if our life expectancy is improved over years.

We took two random variables: weighted average life expectancy - calculated by formula as sum of life expectancy of every country multiplied by its country population and divided by total population in the world in exact year.

We will test hypothetis:

H_0: correlation between weighted average life expectancy and years is not significant.

H_1: correlation between weighted average life expectancy and years is significant.

To this step we are using Spearman correlation test instead of Pearson correlation test because there are a lot of thoughts whether it is crucial to have the normality of the data assumption of Pearson correlation itself.

$\rho(rho) = \frac{\sum(x'-\bar{x'})\cdot(y'-\bar{y'})}{\sqrt{\sum(x'-\bar{x'})^2\cdot\sum(y'-\bar{y'})^2}}$ where x' and y' are the ranks of our r.v. (x = years, y = weighted average life expectancy) and $\bar{x}$ and $\bar{y}$ are the means of our variables (weighted average life expectancy and years).

We can use it because we have three assumptions.

1. Our two variables are measured on the interval.

2. Our two variables can represent paired observations.

3. There is a monotonic relationship between them.

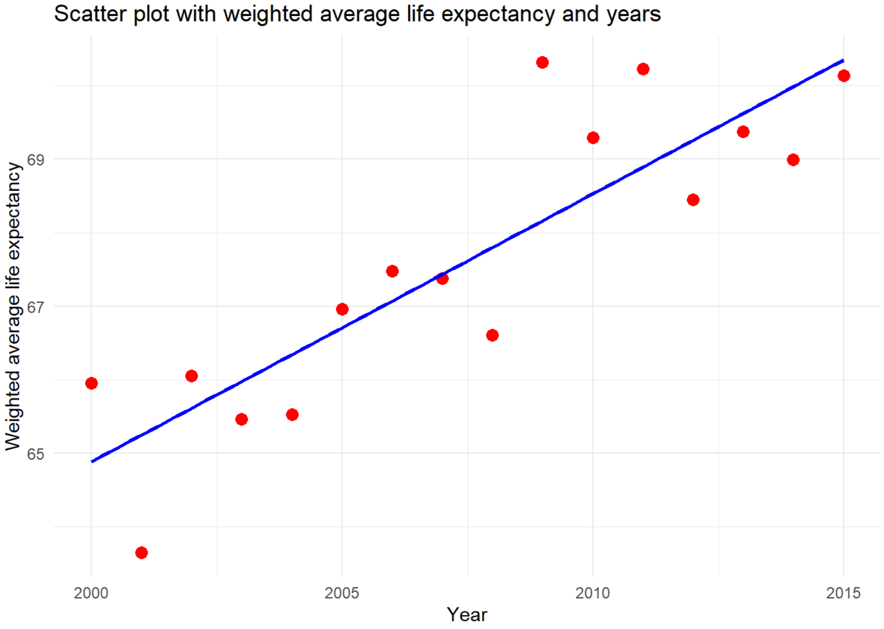
We will use in-built function cor.test().

```
life_expectancy <- file[, c("Country", "Year", "Life.expectancy", "Alcohol", "percentage.expenditure", "Total.expenditure",
"Population")]
countries_year_life_expectancy <- file[, c("Country", "Year", "Life.expectancy", "Population")]

cleaned_data <- countries_year_life_expectancy %>%
  filter(!is.na(Life.expectancy) & !is.na(Population) & Population != 0)

weighted_avg_life_expectancy <- cleaned_data %>%
  filter(Year >= 2000 & Year <= 2015) %>%
  group_by(Year) %>%
  summarise(WeightedAvgLifeExpectancy = sum(Life.expectancy * Population) / sum(Population))

total_population <- cleaned_data %>%
  group_by(Year) %>%
  summarise(total_population = sum(Population, na.rm = TRUE))

weighted_avg_and_total_population <- merge(weighted_avg_life_expectancy, total_population, by = "Year")

correlation_test_spearman <- cor.test(
  weighted_avg_and_total_population$Year,
  weighted_avg_and_total_population$WeightedAvgLifeExpectancy,
  method = "spearman"
)

# Print the results of the correlation test
print(correlation_test_spearman)
```

```
##
##  Spearman's rank correlation rho
##
## data:  weighted_avg_and_total_population$Year and weighted_avg_and_total_population$WeightedAvgLifeExpectancy
## S = 112, p-value = 5.244e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8352941
```

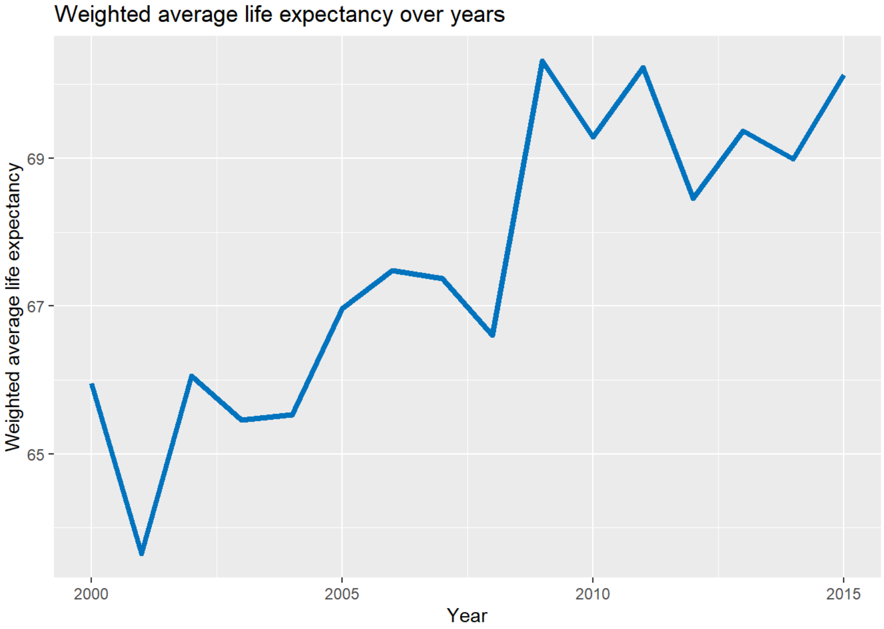
```
ggplot(weighted_avg_and_total_population, aes(x = Year, y = WeightedAvgLifeExpectancy)) +
  geom_point(color = "red", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatter plot with weighted average life expectancy and years",
       x = "Year",
       y = "Weighted average life expectancy") +
  theme_minimal()
```

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

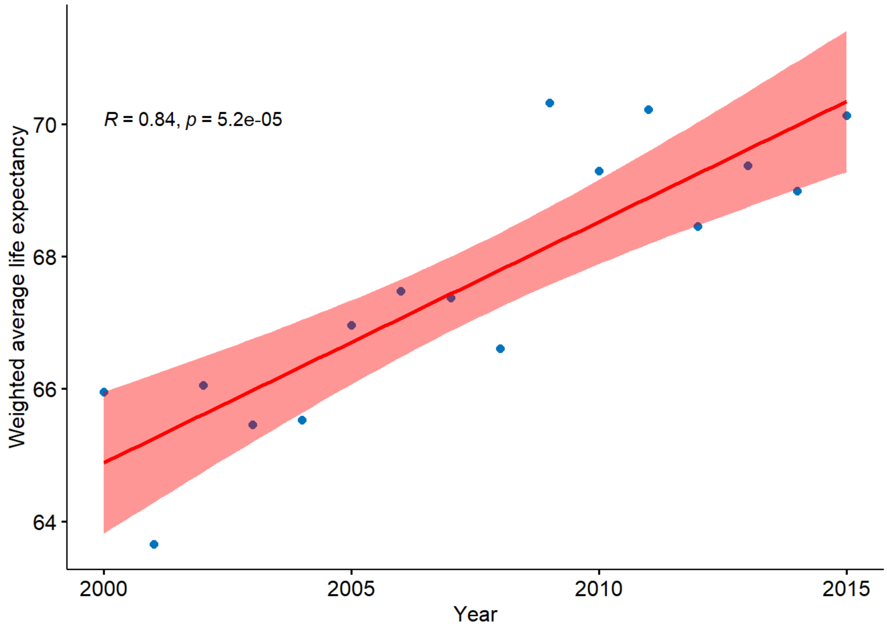


```
ggplot(weighted_avg_and_total_population, aes(x = Year, y = WeightedAvgLifeExpectancy)) +
  geom_line(color = "#0073C2FF", size = 1.5) +
  labs(title = "Weighted average life expectancy over years",
   x = "Year",
   y = "Weighted average life expectancy",
  theme_minimal() +
  theme(text = element_text(size = 12),
        plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5)))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
ggscatter(weighted_avg_and_total_population,
          x = "Year",
          y = "WeightedAvgLifeExpectancy",
          cor.coef = TRUE,
          conf.int = TRUE,
          add = "reg.line",
          col = "red",
          color = "#0073C2FF",
          cor.method = "spearman",
          xlab = "Year",
          ylab = "Weighted average life expectancy")
```



As we can see from results and from plots our p-value is 5.244e-05 and correlation coefficient (R) = 0.8352941.

This p-value s less than the significance level $\alpha = 0.05$ so we reject H_0.

We can conclude that weighted average life expectancy and years are significantly correlated with a correlation coefficient of 0.8352941 and p-value of 5.244e-05.

And yet there is a positive correlation coefficient = 0.8352941.

It means as one increases the other tends to increase as well.

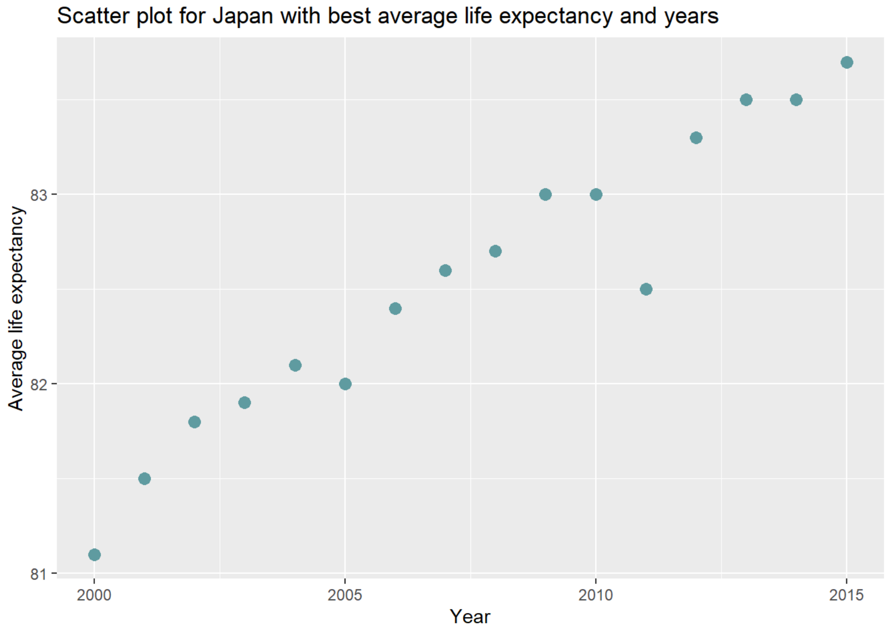
We can see that some dots are not in our confidence interval, so we can check more.

Let’s check whether we can make such a conclusion that weighted average life expectancy during 2000 - 2015 is increasing?

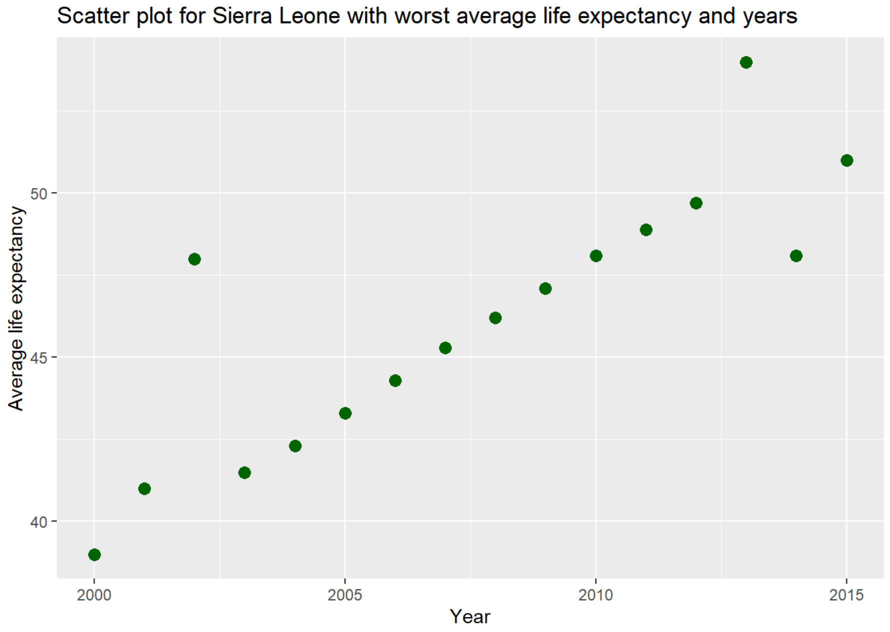
We decided to take the country with the worst average life expectancy and the best average life expectancy and do the same steps and see what we will get.

```
filtered_data <- countries_year_life_expectancy %>%
  filter(Year >= 2000 & Year <= 2015)

worst_country <- filtered_data %>%
  group_by(Country) %>%
  summarise(AvgLifeExpectancy = mean(Life.expectancy, na.rm = TRUE)) %>%
  arrange(AvgLifeExpectancy) %>%
  slice(1)

best_country <- filtered_data %>%
  group_by(Country) %>%
  summarise(AvgLifeExpectancy = mean(Life.expectancy, na.rm = TRUE)) %>%
  arrange(desc(AvgLifeExpectancy)) %>%
  slice(1)

worst_country_name <- worst_country$Country
best_country_name <- best_country$Country

worst_best_df <- filtered_data %>%
  filter(Country %in% c(worst_country_name, best_country_name))

worst_life_expectancy <- worst_best_df %>%
  filter(Country == worst_country_name) %>%
  arrange(Year)

best_life_expectancy <- worst_best_df %>%
  filter(Country == best_country_name) %>%
  arrange(Year)
```

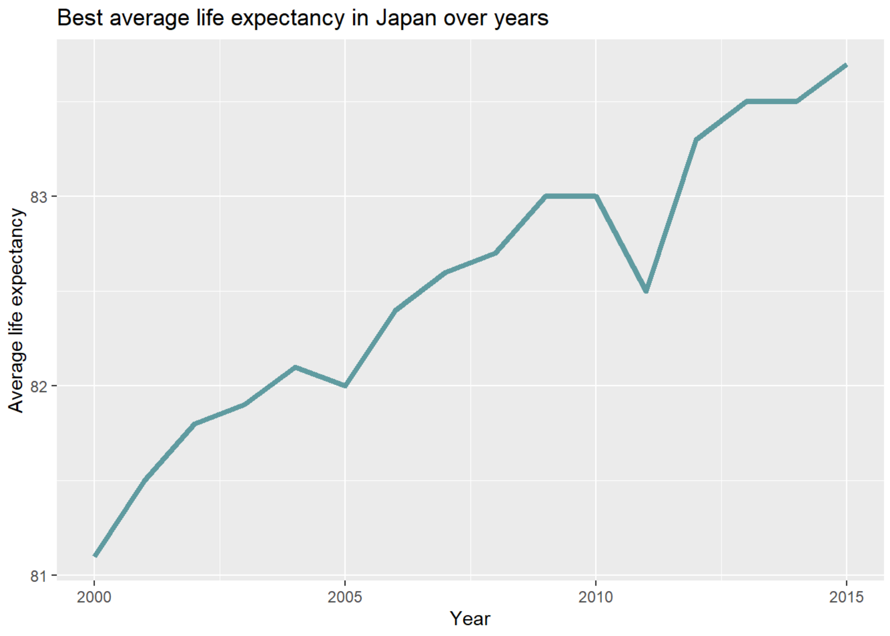
```
ggplot(best_life_expectancy, aes(x = Year, y = Life.expectancy)) +
  geom_point(color = "cadetblue", size = 3) +
  labs(title = paste("Scatter plot for", best_life_expectancy$Country, "with best average life expectancy and years"),
   x = "Year",
   y = "Average life expectancy",
   theme_minimal() +
   theme(text = element_text(size = 12),
  plot.title = element_text(hjust = 0.5),
  plot.caption = element_text(hjust = 0.5)))
```


Scatter plot for Japan with best average life expectancy and years

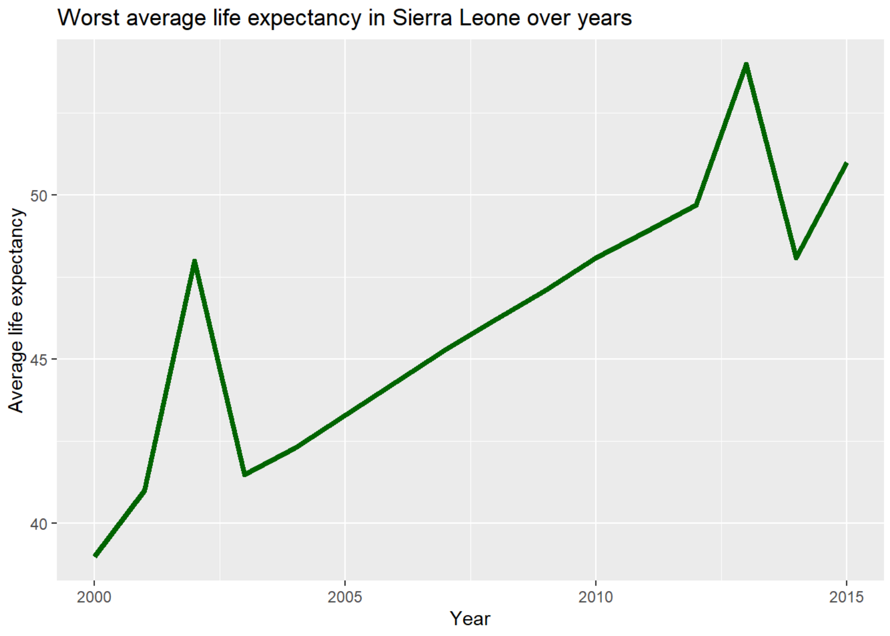
```
ggplot(worst_life_expectancy, aes(x = Year, y = Life.expectancy)) +
  geom_point(color = "darkgreen", size = 3) +
  labs(title = paste("Scatter plot for", worst_life_expectancy$Country, "with worst average life expectancy and years"),
   x = "Year",
   y = "Average life expectancy",
   theme_minimal() +
   theme(text = element_text(size = 12),
  plot.title = element_text(hjust = 0.5),
  plot.caption = element_text(hjust = 0.5)))
```


Scatter plot for Sierra Leone with worst average life expectancy and years

```
ggplot(best_life_expectancy, aes(x = Year, y = Life.expectancy)) +
  geom_line(color = "cadetblue", size = 1.5) +
  labs(title = paste("Best average life expectancy in", best_life_expectancy$Country, "over years"),
   x = "Year",
   y = "Average life expectancy",
  theme_minimal() +
  theme(text = element_text(size = 12),
        plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5)))
```



```
ggplot(worst_life_expectancy, aes(x = Year, y = Life.expectancy)) +
  geom_line(color = "darkgreen", size = 1.5) +
  labs(title = paste("Worst average life expectancy in", worst_life_expectancy$Country, "over years"),
   x = "Year",
   y = "Average life expectancy",
  theme_minimal() +
  theme(text = element_text(size = 12),
        plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5)))
```



```
cor.test(best_life_expectancy$Year, best_life_expectancy$Life.expectancy, method = "spearman")
```

```
## Warning in cor.test.default(best_life_expectancy$Year,
## best_life_expectancy$Life.expectancy, : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  best_life_expectancy$Year and best_life_expectancy$Life.expectancy
## S = 23.033, p-value = 1.254e-09
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9661277
```

```
cor.test(worst_life_expectancy$Year, worst_life_expectancy$Life.expectancy, method = "spearman")
```

```
## Warning in cor.test.default(worst_life_expectancy$Year,
## worst_life_expectancy$Life.expectancy, : Cannot compute exact p-value with ties
```

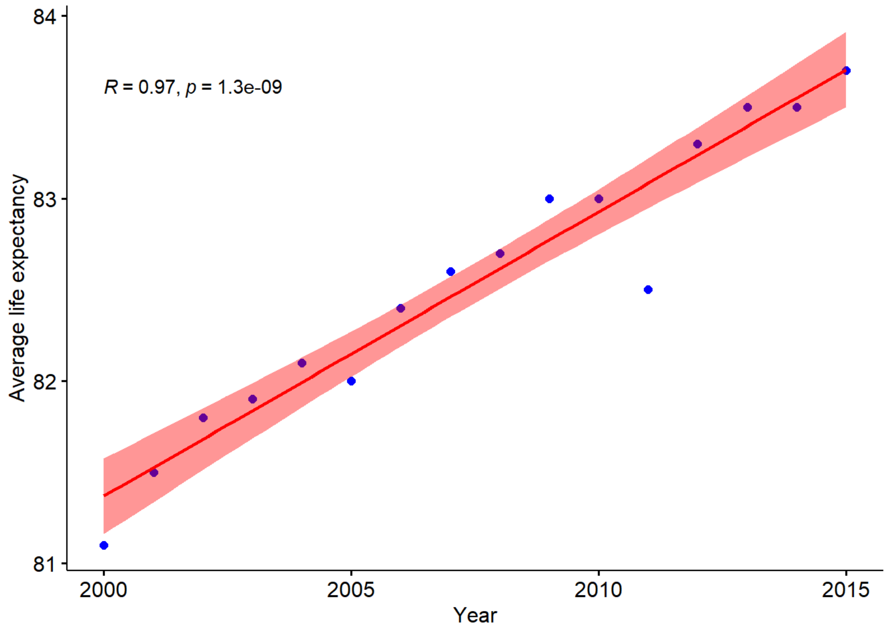
```
## 
##  Spearman's rank correlation rho
## 
## data:  worst_life_expectancy$Year and worst_life_expectancy$Life.expectancy
## S = 75.555, p-value = 4.16e-06
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho 
## 0.8888891
```

As we can see for best country p-value = 1.254e-09 and correlation coefficient is 0.9661277.

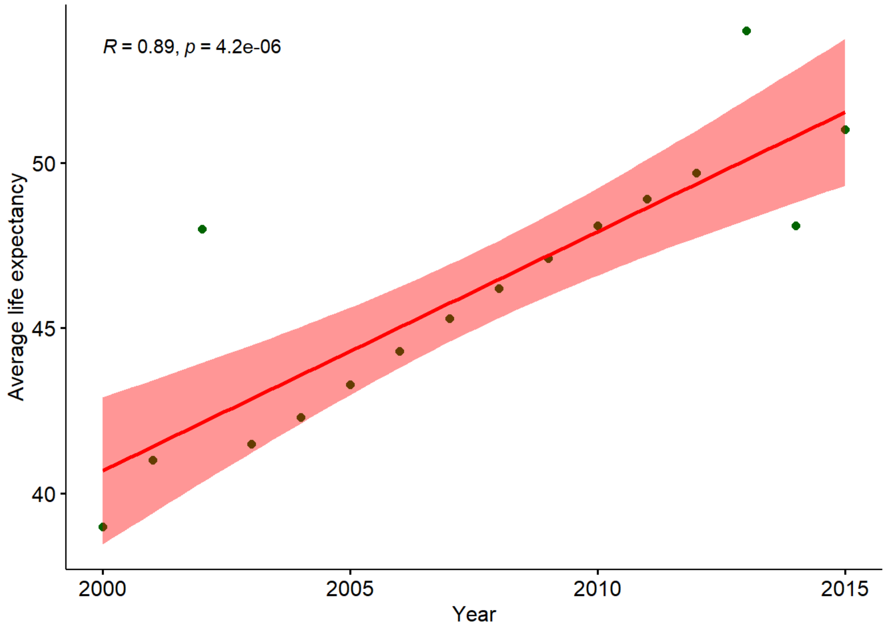
And for worst country p-value = 4.16e-06 correlation coefficient = 0.8888891.

Both of them have strong and positive correlation coefficients and both of them reject the absence of significance between years and average life expectation.

```
ggscatter(best_life_expectancy,
          x = "Year",
          y = "Life.expectancy",
          color = "blue",
          cor.coef = TRUE,
          conf.int = TRUE,
          add = "reg.line",
          col = "red",
          cor.method = "spearman",
          xlab = "Year",
          ylab = "Average life expectancy")
```



```
ggscatter(worst_life_expectancy,
          x = "Year",
          y = "Life.expectancy",
          color = "darkgreen",
          cor.coef = TRUE,
          conf.int = TRUE,
          add = "reg.line",
          col = "red",
          cor.method = "spearman",
          xlab = "Year",
          ylab = "Average life expectancy")
```



As we can see from scatter plots - there is a tendency that when years increasing, life expectancy also increases.

From correlation plot we can see that our R is strong because there dots are close to our regression line.

These additional tests for worst and best countries helps us to confidently say that as year are increasing the weighted average life expectancy has a tendency to increase also. But it doesn’t mean that are dependent, this is important because there are a tons of factors.

TASK 2

Next testing will be normality among total average life expectancy distributed by countries.

So out hypothesis will be next:

H_0: life expectancy normally distributed.

H_1: life expectancy is not normally distributed.

We will use Anderson-Darling test for normality (ad.test()).

The Anderson-Darling test is a statistical test used to assess whether a sample comes from a particular distribution, in this case, whether the life expectancy follows a normal distribution. The test is an extension of the Kolmogorov-Smirnov test and is particularly useful when dealing with ties in the data. (In our case we have repeated values).

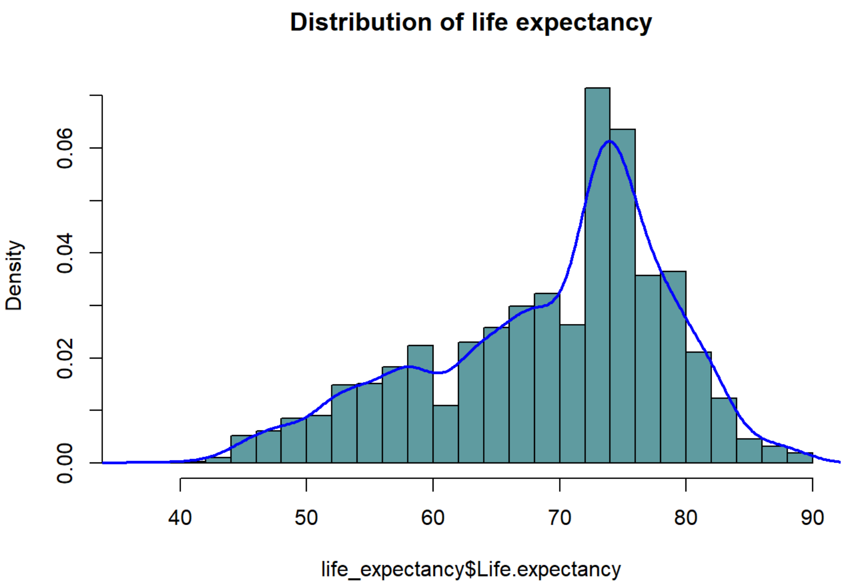
The test makes use of the cumulative distribution function.

The Anderson-Darling test is an EDF omnibus test for the composite hypothesis of normality.

The test statistic is$A = -n - \frac{1}{n}\sum_{i=1}^{n}[2i-1][\ln(p(i)) + \ln(1-p(n-i+1))]$ where $p(i) = \Phi\left(\frac{x(i)-\bar{x}}{s}\right)$.

Here, Φ is the cumulative distribution function of the standard normal distribution, and $\bar{x}$ and s are the mean and standard deviation of the data values. The p-value is computed from the modified statistic $Z = A(1.0 + \frac{0.75}{n} + \frac{2.25}{n^2})$ according to Table 4.9 in Stephens (1986).

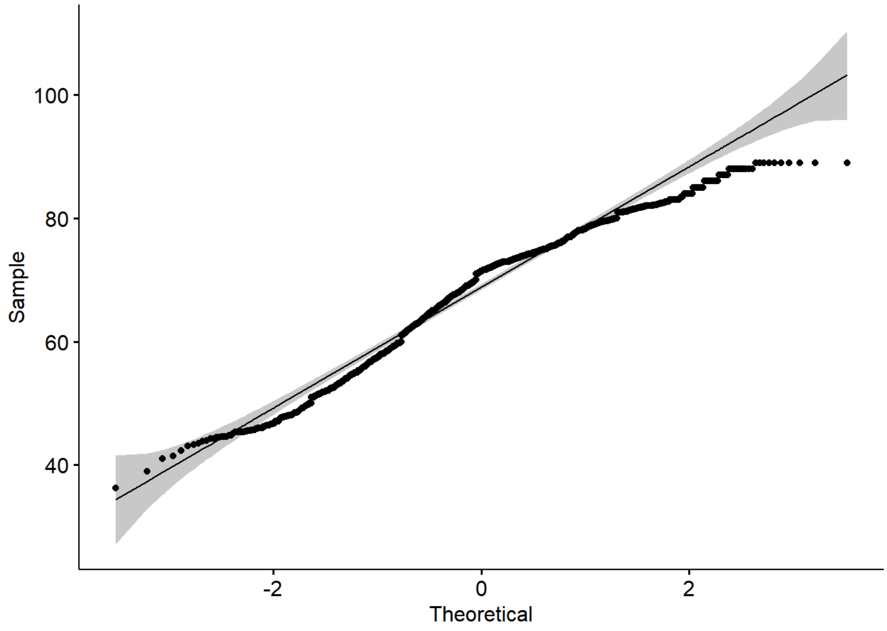
```
hist(life_expectancy$Life.expectancy,
     col = "cadetblue",
     main = "Distribution of life expectancy",
     breaks = 20,
     probability = TRUE)

lines(density(life_expectancy$Life.expectancy, na.rm = TRUE), col = "blue", lwd = 2)
```



```
ad.test(life_expectancy$Life.expectancy)
```

```
##
##  Anderson-Darling normality test
##
## data:  life_expectancy$Life.expectancy
## A = 49.53, p-value < 2.2e-16
```

```
ggqqplot(cleaned_data$Life.expectancy)
```



From A-D test we get that our p-value < 2.2e-16 and is less than $\alpha = 0.05$ so we reject the null hypothesis that there is no difference between our and normal distribution (normality).

From hist&density plot we see it has significant left skew and it doesn't fit normal distribution.

Inspecting Q-Q plot we can say that out empirical line(dark line) does not follow the theoretical(thin) because empirical should converge to linear function.

So overall conclusion will be that average life expectancy does not follow normal distribution.

TASK 3

```
life_expectancy <- file[, c("Country", "Year", "Life.expectancy", "Alcohol", "Status", "infant.deaths")]
```

H_0: Life expectancy is independent of alcohol consumption.

H_1: Life expectancy is dependent on alcohol consumption.

$TEST$: The Pearson Correlation Test

As discovered later, The Chi-Squared Test of Independence works only for categorical variables, not the continuous as we got in our case. That's why the decision of changing the test was made.

Firstly, we need to know what is the Pearson correlation test? It is a statistic that determines how closely two variables are related. Its value ranges from -1 to +1, with 0 denoting no linear correlation, -1 denoting a perfect negative linear correlation, and +1 denoting a perfect positive linear correlation. A correlation between variables means that as one variable's value changes, the other tends to change in the same way.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively.

- *r = 0 - there is no relation between the variable.*
- *r = 1 - perfectly positively correlated.*
- *r = -1 - perfectly negatively correlated.*
- *r = (0, 0.30) - negligible correlation.*
- *r = (0.30; 0.50) - moderate correlation.*
- *r = (0.50; 1) - highly correlated.*

```
alcohol_consumption <- file$Alcohol
life_expectancy <- file$Life.expectancy

# Drop missing values, if any
data <- data.frame(Alcohol = alcohol_consumption, Life_Expectancy = life_expectancy)
data <- na.omit(data)

# Perform Pearson correlation
correlation <- cor(data$Alcohol, data$Life_Expectancy, method = 'pearson')

# Calculate p-value
p_value <- cor.test(data$Alcohol, data$Life_Expectancy)$p.value

# Print results
cat('Pearson Correlation Coefficient:', correlation, '\n')
```

```
## Pearson Correlation Coefficient: 0.4048768
```

```
cat('P-value:', p_value, '\n')
```

```
## P-value: 2.105875e-108
```

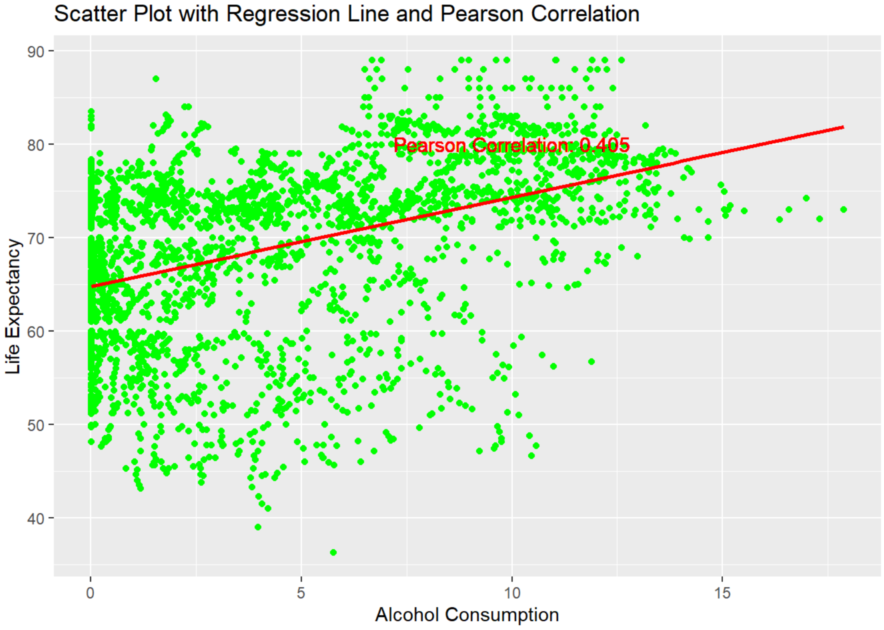
Given that we have already performed the Pearson Correlation Test and obtained a p-value less than 0.05, we can conclude that the correlation between alcohol consumption and life expectancy exists in our data.

In addition to the **p-value**, we would also give some attention to **correlation** results (because it was the initial goal of the research). As our Pearson Correlation Cofficient r is 0.4, we can assume that our correlation is moderate.

To visualize this correlation, we can create a scatter plot with alcohol consumption on one axis and life expectancy on the other. If there is a positive correlation, you would expect to see points tending to move upward from left to right on the plot. If the points on the scatter plot are not tending to move upward from left to right, it suggests that there may not be a simple linear relationship between alcohol consumption and life expectancy. In this case, the correlation coefficient alone might not fully capture the association between these variables.

```
# Scatter plot with regression line and Pearson correlation coefficient
gg <- ggplot(data, aes(Alcohol, Life_Expectancy))
gg <- gg + geom_point(color = 'green')  # Set points color to blue
gg <- gg + geom_smooth(method="lm", se=FALSE, color="red")
gg <- gg + geom_text(x=10, y=80, label=sprintf('Pearson Correlation: %.3f', correlation), size=4, color='red')
gg <- gg + labs(title='Scatter Plot with Regression Line and Pearson Correlation',
               x='Alcohol Consumption', y='Life Expectancy')
print(gg)
```

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



```
Okay, if points are tending to move upward from left to right on the plot, it means that the more alcohol u consume the longer your life expect
```

No, it's important to be cautious with the interpretation of correlation in terms of causation. If points on a scatter plot tend to move upward from left to right, it indicates a positive correlation between the two variables (in this case, alcohol consumption and life expectancy). A positive correlation means that as one variable increases, the other variable also tends to increase, and vice versa.

However, correlation does not imply causation. Just because there is a positive correlation between alcohol consumption and life expectancy does not mean that consuming more alcohol directly causes an increase in life expectancy. The correlation might be influenced by a third variable that affects both alcohol consumption and life expectancy. For example, socioeconomic factors, healthcare access, or lifestyle choices could be confounding variables. There could be other factors at play, and correlation alone does not establish a cause-and-effect relationship.

For this point we can clarify that alcohol consumption is not independent on life expectancy.

TASK 4

```
life_expectancy <- file[, c("Country", "Year", "Life.expectancy", "Alcohol", "Status", "infant.deaths", "percentage.expendit
ure", "Total.expenditure", "Population", "GDP")]
```

Problem statement

In this part we will consider two hypotheses:

H_0: *There is no positive correlation between life expectancy and general government expenditure.*

H_1: *There is a positive correlation between life expectancy and general government expenditure.*

So, now we will hold a research in order to determine whether to reject th null hypothesis or not.

To begin with, we need to clear our data in order to get rid of unnecessary information. We will live only *Life.expectancy*, *Total.expenditure* and *percentage.expenditure* (just in case). Also, we will avoid data which is NA, as they have no impact on our results.

```
# Step 1: Setting Up Data
data <- na.omit(file)
subset_data <- data[, c(
  "Life.expectancy",
  "percentage.expenditure",
  "Total.expenditure")]
```

Now we can finally perform the Pearson correlation test and find out the answer for the question we made at the very beginning: is there a positive correlation between life expectancy and government expenditure?

The value we are interested in the most is the **p-value** as it will decide whether we reject the null hypothesis or not.

```
correlation_test_percentage_expenditure <- cor.test(
  subset_data$Life.expectancy,
  subset_data$percentage.expenditure,
  method = "pearson"
)

print(correlation_test_percentage_expenditure)
```

```
##
##  Pearson's product-moment correlation
##
## data:  subset_data$Life.expectancy and subset_data$percentage.expenditure
## t = 18.223, df = 1647, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3686483 0.4490240
## sample estimates:
##       cor
## 0.4096308
```

As we can see, the **p-value** is heading straight towards 0 and is almost equal to it. If we take the significance level $\alpha = 0.05$ for both cases, we will make the following conclusion: **reject the H_0 as p-value is <= α** .

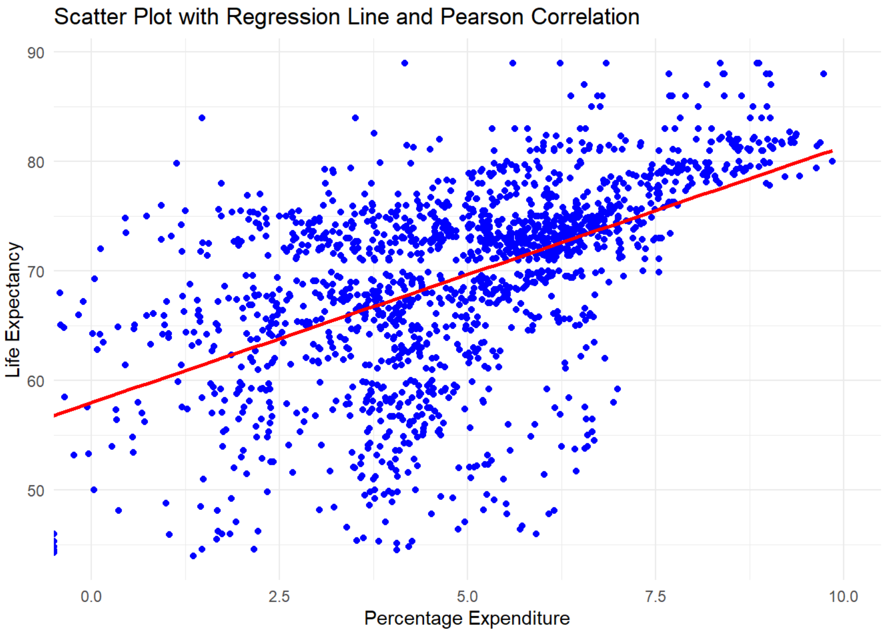
In addition to the **p-value**, we would also give some attention to **correlation** results (because it was the initial goal of the research). We got two values: 0.175 for *Total.expenditure* and 0.41 for *percentage.expenditure*.

So, we have *negligible* correlation for *Total.expenditure*, meaning that the *Life.expectancy* won't change significantly if we moderate it. However, for *percentage.expenditure* we have the opposite result - if we change it, the *Life.expectancy* will change significantly for their **high correlation**.

```
correlation <- cor(log(subset_data$percentage.expenditure), subset_data$Life.expectancy, method = 'pearson')

x_limits <- c(0, 10)

gg <- ggplot(subset_data, aes(x = log(percentage.expenditure), y = Life.expectancy))
gg <- gg + geom_point(color = 'blue')
gg <- gg + geom_smooth(method = "lm", se = FALSE, color = "red")
gg <- gg + labs(title = 'Scatter Plot with Regression Line and Pearson Correlation',
             x = 'Percentage Expenditure', y = 'Life Expectancy')
gg <- gg + theme_minimal()
gg <- gg + coord_cartesian(xlim = x_limits)

print(gg)
```

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

```
## Warning: Removed 5 rows containing non-finite values (`stat_smooth()`).
```



This plot is a visualization of correlation. As we can see, the farther we go, the less is the difference between the red line and the blue dots. One might try to make some assumptions about the relation between the expenditure and life expectancy; but it's crucial to remember that other factors which we do not include directly also influence the final results.

With that said, the only valid assumption we can safely make is that the government expenditure and life expectancy are not independent of each other.