Salary project

```
library(readxl)
library(tidyverse)library(ggplot2)
data <-

read_xlsx("C:\\Users\\rexar\\Downloads\\data1.xlsx")
```

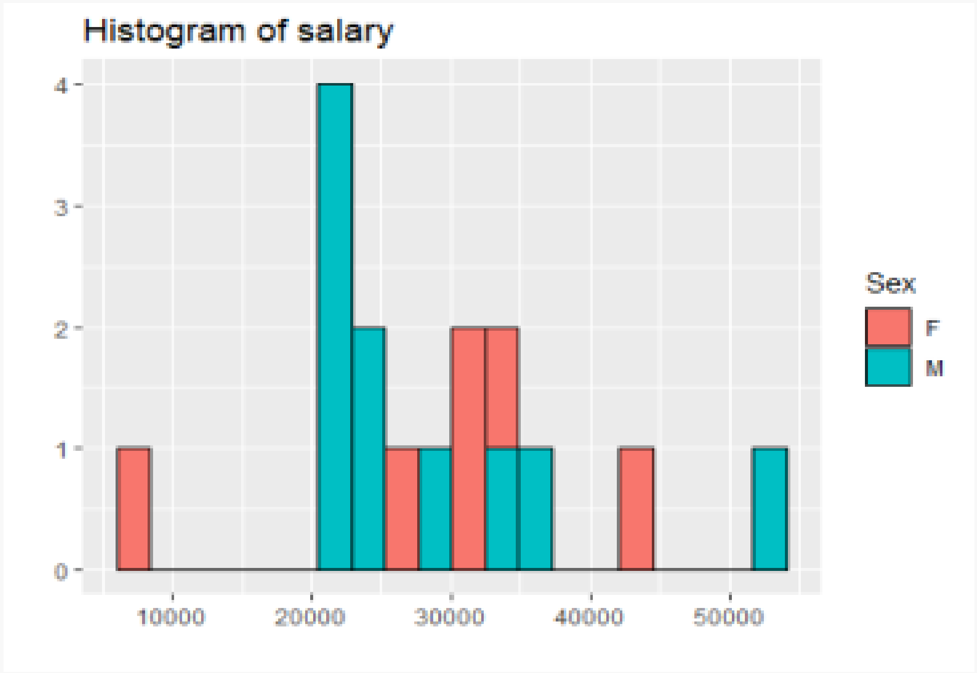
Comm: reading libraries and data file

```
Salary <- data$Salary
Sex <- data$sex
Age <- data$Age
Family <- data$Family
```

Comm: Creating variables to simplify your work.

Exercise 1

```
dan <- data.frame(Salary, Sex)
ggplot(dan, aes(x = Salary, fill = Sex))+
geom_histogram(color = "black", position = "identity", bins =
20)+  labs(title = "Histogram of salary", x = "", y = "")
```



Comm: The histogram shows the salaries of men and women (the colors of men and women can be seen in the graph)

Exercise 2

```
summary_stats <- summary(Salary)
mean_salary <- mean(Salary)
sd_salary <- sd(Salary)
```

```r
min_salary <- min(Salary)
q1_salary <- quantile(Salary, 0.25)
median_salary <- median(Salary)
q3_salary <- quantile(Salary, 0.75)
max_salary <- max(Salary)
print("Statistic:")
```

```
## [1] "Statistic:"
```

```r
print(summary_stats)
```

```
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 7211 22938 27900 29165 33646 52907
```

```r
cat("\n")
```

```r
print(paste("Sample mean: ", mean_salary))
```

```
## [1] "Sample mean: 29164.7223853993"
```

```r
print(paste("Standard deviation: ", sd_salary))
```

```
## [1] "Standard deviation: 9529.26937982479"
```

```r
print(paste("Minimum value: ", min_salary))
```

```
## [1] "Minimum value: 7211.46096370928"
```

```r
print(paste("Q1: ", q1_salary))
```

```
## [1] "Q1: 22937.7510776331"
```

```r
print(paste("Median: ", median_salary))
```

```
## [1] "Median: 27900.3075518158"
```

```r
print(paste("Q3: ", q3_salary))
```

```
## [1] "Q3: 33645.9028023671"
```

```r
print(paste("Maximum value: ", max_salary))
```

```
## [1] "Maximum value: 52906.786968965"
```

Comm: These statistics provide descriptive statistics for a data set. Here is how each of them can be interpreted:

Statistic (first block):

Min. (Minimum): 7211 - This is the minimum value in the data set. 1st Qu. (First Quartile): 22938 - 25% of the data is below this value. Median: 27900 - This is the middle of the data set, also known as the second quartile. Mean: 29165 - The mean value (the sum of all values divided by the number of values). 3rd Qu. (Third Quartile): 33646 - 75% of the data is below this value. Max. (Maximum): 52907 - The maximum value in the data set.

Sample mean (second block):

29164.7223853993 - This is the sample mean, similar to what was reported in the previous context.

Standard deviation (third block):

9529.26937982479 - The standard deviation, which measures the spread of values relative to the mean. A larger standard deviation indicates greater variation in the data.

Minimum, Q1, Median, Q3, and Maximum value (last block):

Further clarification regarding the scatter of data in different quartiles. These statistics help you understand the nature of the data, its central tendency (mean, median) and spread (minimum, maximum, standard deviation).

Exercise 3

```
data$Gender <- ifelse(data$sex == "M", 1, 0)
cor_matrix <- cor(data[, c("Age", "Family", "Salary",
"Gender")]) print(cor_matrix)
```

```
## Age Family Salary Gender
## Age 1.00000000 -0.6150601 0.407125054 -0.058570741
## Family -0.61506008 1.0000000 -0.492343944 -0.310684883
## Salary 0.40712505 -0.4923439 1.000000000 -0.009993352
## Gender -0.05857074 -0.3106849 -0.009993352 1.000000000
```

Comm: This is the correlation matrix between your variables. The values in the correlation matrix range from -1 to 1, where:

1: Positive linear correlation (when one variable increases, the other variable also increases linearly).
-1: Negative linear correlation (as one variable increases, the other variable decreases linearly).

0: No linear correlation. The correlation coefficient of -0.61506008 indicates a negative linear correlation between age and family size. This means that as age increases, family size tends to decrease.

Age and Salary:

The correlation coefficient of 0.40712505 indicates a positive linear correlation between age and salary. This means that as age increases, salary tends to increase.

Family and Salary:

The correlation coefficient of -0.492343944 indicates a negative linear correlation between family size and salary. This means that as family size increases, salary tends to decrease.

Gender and other variables:

The correlation coefficients are close to zero, indicating a weak linear relationship between gender and other variables (age, family size, salaries). The correlation results you obtained indicate the degree of association (or lack of association) between age and salary for each group:

Exercise 4

```
less_18_sal <- data$Salary[data$Age < 18]
more_18_sal <- data$Salary[data$Age > 18]
equal_18_sal <- data$Salary[data$Age == 18]

cor_less_18 <- cor(data$Age[data$Age < 18], less_18_sal)
cor_more_18 <- cor(data$Age[data$Age > 18], more_18_sal)
cor_equal_18 <- cor(data$Age[data$Age == 18],

equal_18_sal) cat("For less then 18:", cor_less_18, "\n")
```

```
## For less then 18: 0.2052135
```

```
cat("For more then 18", cor_more_18, "\n")
```

```
## For more then 18 -0.1763593
```

```
cat("For equal 18:", cor_equal_18, "\n")
```

```
## For equal 18: NA
```

Comm:

For age below 18 years (0.2052135): A positive correlation value (closer to 1) means that there is an increase in salaries as age under 18 increases, but the relationship is not very strong.

For age above 18 years (-0.1763593): A negative correlation value (closer to -1) indicates that there is a tendency for salaries to decrease with increasing age over 18, but again the relationship is not very strong.

For equal 18: NA - This means that there is no diversity of values (there is only one age 18) and it is not possible to calculate the difference.

Exercise 5

```
Salary_m <- data$Salary[1:10]
Salary_f <- data$Salary[11:20]
t_test_result_salary <- t.test(Salary_m, Salary_f)
print(t_test_result_salary)
```

```
##
## Welch Two Sample t-test
##
## data: Salary_m and Salary_f
## t = -0.0424, df = 18, p-value = 0.9666
## alternative hypothesis: true difference in means is not equal to
0 ## 95 percent confidence interval:
## -9383.857 9012.584
## sample estimates:
```

```
## mean of x mean of y
## 29071.90 29257.54
```

Comm: This finding presents the results of Welch's two-sample t-test in the context of analyzing the differences between the two samples represented by the variables Salary_m and Salary_f.

t = -0.0424: This is the value of the t-statistic, which measures the standardized difference between the mean values of the two groups. In this case, the t-statistic is close to zero, which may indicate that there is no statistically significant difference between the mean values.

df = 18: This is the number of degrees of freedom used in the t distribution. The higher the value of df, the more accurate the test is considered to be. In this case, the df is 18.

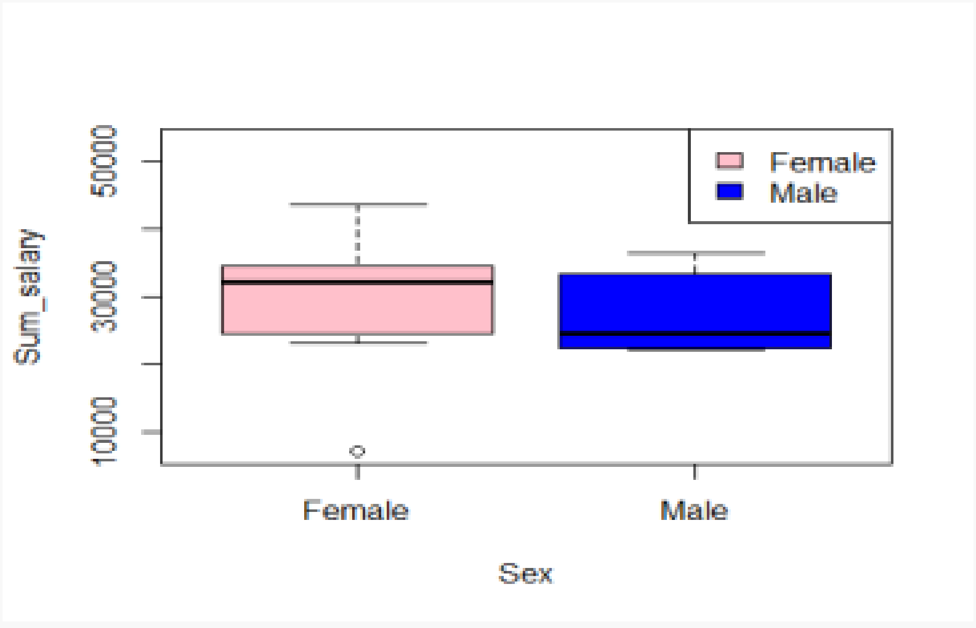
p-value = 0.9666: This is a p-value that indicates the probability of getting the observed data (or more extreme) if the null hypothesis of equality of mean values is true. In this case, a high p-value (close to 1) indicates that there is no statistically significant difference between the mean values of Salary_m and Salary_f.

Alternative hypothesis: This is a formulation of the alternative hypothesis, which says that the true difference between the mean values is not zero.

95 percent confidence interval: The confidence interval for the difference between the mean values is presented here. In this case, the interval from -9383.857 to 9012.584 includes zero, which also emphasizes that there are no statistically significant differences.

Sample estimates: The mean values for each of the samples (mean of x for Salary_m and mean of y for Salary_f) are presented here. In this case, the mean values are close to each other, which also supports the conclusion that the differences are not significant.

```
dan3 <- data.frame(Sum_salary = c(Salary_m, Salary_f),
 Sex = factor(rep(c("Male", "Female"), each = 10)))
boxplot(Sum_salary ~ Sex, dan3, col = c("pink", "blue"))
 legend("topright", legend = levels(dan3$Sex), fill = c("pink", "blue"))
```



Comm: This is the figure for the fifth task, which visualizes the difference between the salaries of a man and a woman.