

Lead concentration

Project Biomedical Statistics

Group G

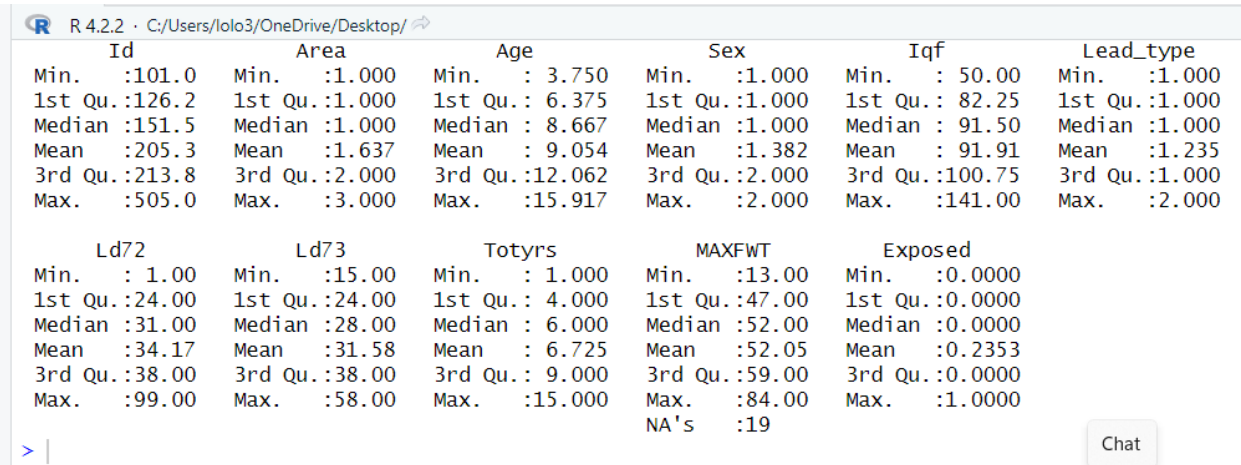
Islam Nabil	202000325
Doaa Maged	202000840
Lujain Mohamed	202002670
Shady El-sherif	202002137

Dr. Mohamed Maysara

1. Descriptive statistics

First, we read the data → RDATA using `load("lead.RData")` and `get()` functions, multiple functions such as `str()`, `summary()` for :

- Data identification and exploration
- Gaining insights into the data columns' data types.
- Data summarization: computing the average, middle value, lowest and highest values, as well as the first and third quartiles.



R 4.2.2 · C:/Users/lolo3/OneDrive/Desktop/

Id	Area	Age	Sex	Iqf	Lead_type
Min. :101.0	Min. :1.000	Min. : 3.750	Min. :1.000	Min. : 50.00	Min. :1.000
1st Qu.:126.2	1st Qu.:1.000	1st Qu.: 6.375	1st Qu.:1.000	1st Qu.: 82.25	1st Qu.:1.000
Median :151.5	Median :1.000	Median : 8.667	Median :1.000	Median : 91.50	Median :1.000
Mean :205.3	Mean :1.637	Mean : 9.054	Mean :1.382	Mean : 91.91	Mean :1.235
3rd Qu.:213.8	3rd Qu.:2.000	3rd Qu.:12.062	3rd Qu.:2.000	3rd Qu.:100.75	3rd Qu.:1.000
Max. :505.0	Max. :3.000	Max. :15.917	Max. :2.000	Max. :141.00	Max. :2.000

Ld72	Ld73	Totyrs	MAXFWT	Exposed
Min. : 1.00	Min. :15.00	Min. : 1.000	Min. :13.00	Min. :0.0000
1st Qu.:24.00	1st Qu.:24.00	1st Qu.: 4.000	1st Qu.:47.00	1st Qu.:0.0000
Median :31.00	Median :28.00	Median : 6.000	Median :52.00	Median :0.0000
Mean :34.17	Mean :31.58	Mean : 6.725	Mean :52.05	Mean :0.2353
3rd Qu.:38.00	3rd Qu.:38.00	3rd Qu.: 9.000	3rd Qu.:59.00	3rd Qu.:0.0000
Max. :99.00	Max. :58.00	Max. :15.000	Max. :84.00	Max. :1.0000

NA's :19

Chat

- **Creating frequency tables** for categorical data as → Area, Sex, Lead type, and Exposed columns.

```
> table(myData$Sex)
  Male Female 
   63     39 

> table(myData$Exposed)
 0  1 
78 24 

> table(myData$Lead_type)
 1  2 
78 24 

> table(myData$Area)
 1  2  3 
52 35 15 

> |
```

- **Calculating correlation coefficient** for (MAXWT and Ld72) and (MAXWT and Ld73) :
had been done considering the following information:

- The correlation coefficient (MAXWT and Ld72) :

```
cor(myData$MAXFWT, myData$Ld72)
```

- b. The correlation coefficient (MAXWT and Ld73)

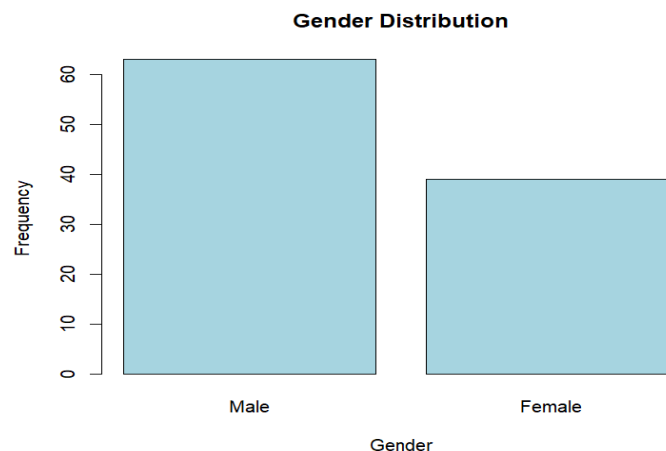
```
cor(myData$MAXFWT, myData$Ld73)
```

```
[1] NA
> cor(myData$MAXFWT, myData$Ld72, use = "complete.obs")
[1] -0.2655747
> cor(myData$MAXFWT, myData$Ld73, use = "complete.obs")
[1] -0.341128
```

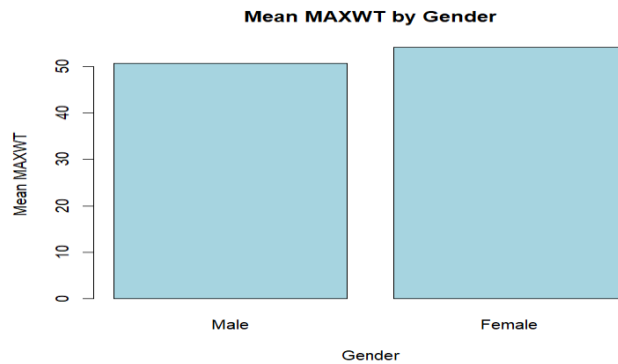
→ So, there is a (Weak negative correlation)

2. Graphics

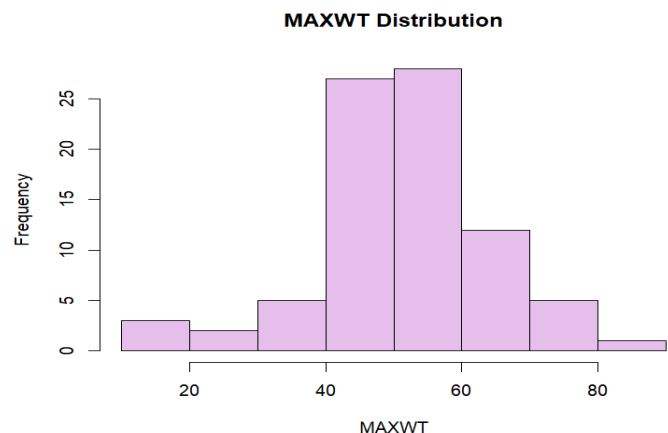
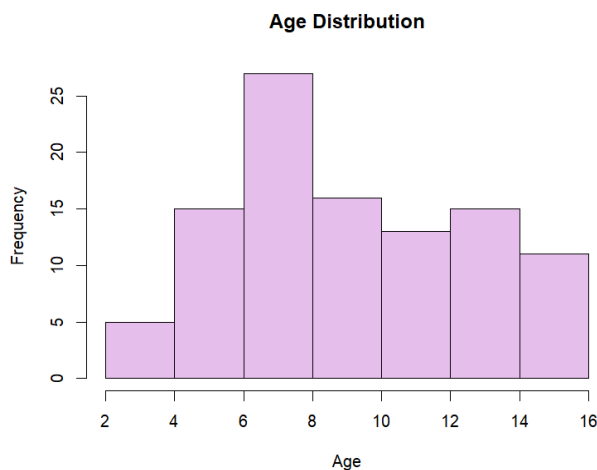
- **Generate a bar chart of a categorical variable for the gender (Sex parameter).**
 - a. To determine the gender with the highest frequency, compare the number of males (gender 1) to the number of females (gender 2).
- Males > Females



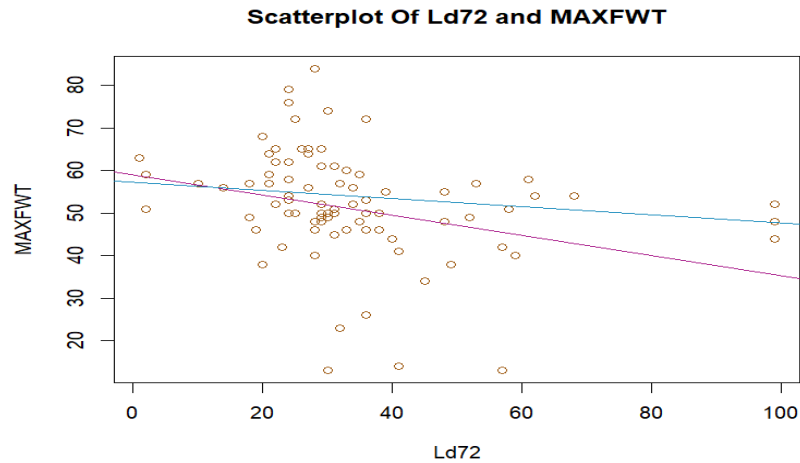
- **Generate a bar chart graph with mean MAXWT in males and females.**
 - a. To visualize and determine which gender has a higher mean of finger tapping, compare the mean MAXFWT (maximum finger tapping) of females to the mean MAXFWT of males.
- females have a higher average finger-tapping performance.



- **Make a histogram of a continuous variable: “age” as well as “MAXWT”.**
 - a. To assess the spread and normality of these variables, it can be observed that both histograms indicate that the data does not appear to follow a normal distribution.
 - b. Age → Right Skewed
 - c. MAXFWT → Right Skewed
 - d. To make sure, we calculated the **mean and median** for Age:
 1. mean = 9.053922
 2. median= 8.666667
 - e. The **mean and median** for MAXFWT :
 1. Mean = 52.04819
 2. Median = 52

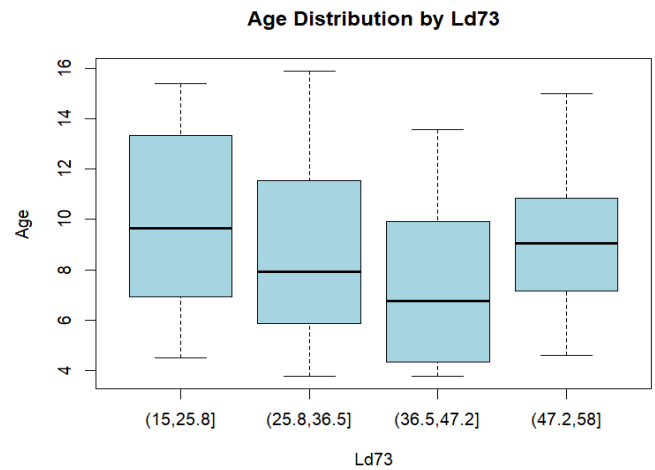
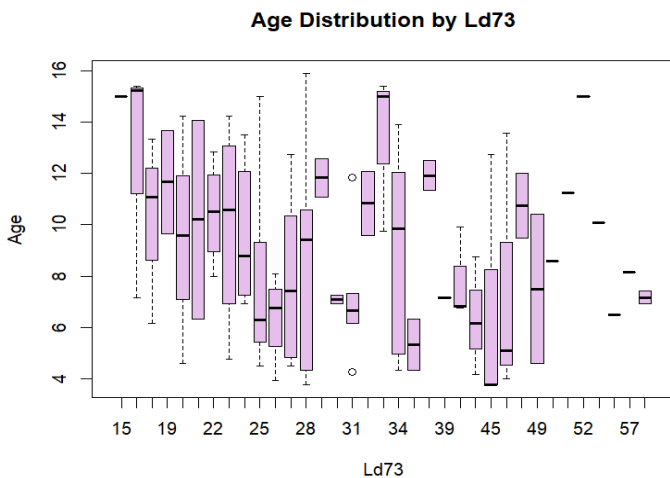
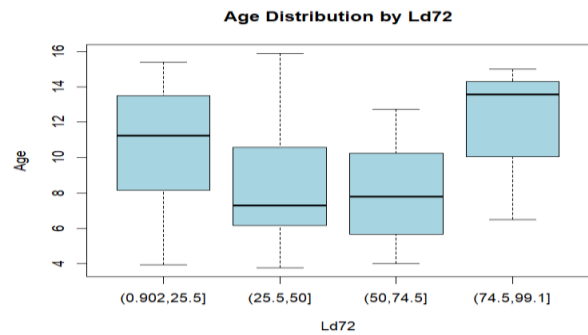
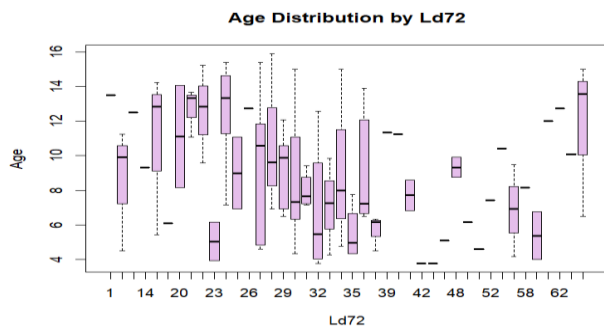


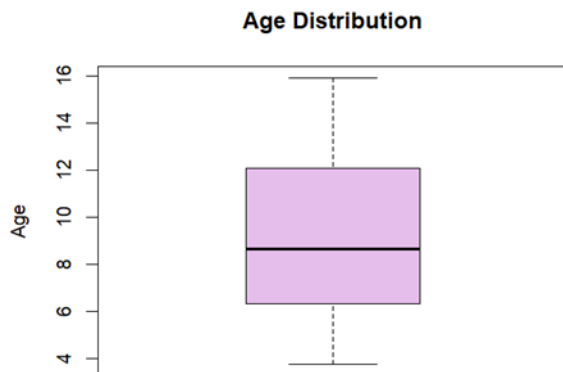
- **Make a scatterplot of 2 continuous variables Ld72 and MAXWT and add the regression lines for each gender.**
 - a. Based on the observed data, there appears to be a negative relationship (negative slope) between the dependent variable MAXWT and the independent variable Ld72. Specifically, as Ld72 increases, MAXWT tends to decrease.



- Make a boxplot of age and separate boxplots per Ld72 and per Ld73 (as.factors).

a. To assess the variation of the data and separate it into 4 intervals to see the variation. (Seems that → not equal variance)





Here, there are no outliers.

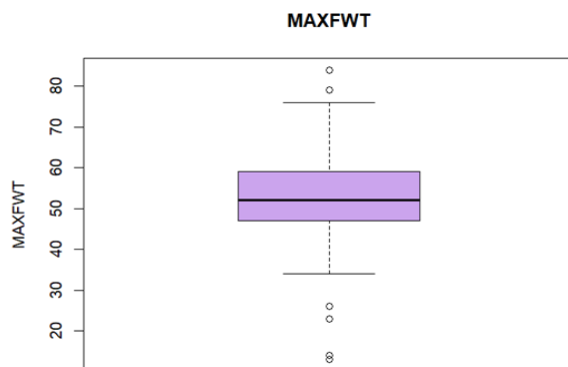
To summarize all graphs, In Histogram (x-axes): v.data & (y-axes): the frequency from the distributions, we can see that the data seems to be not normal may be due to some outliers due to its sensitivity so, we will go further and use Q-Q plots and tests to make sure.

- **Also, outliers may be affect on tests.**

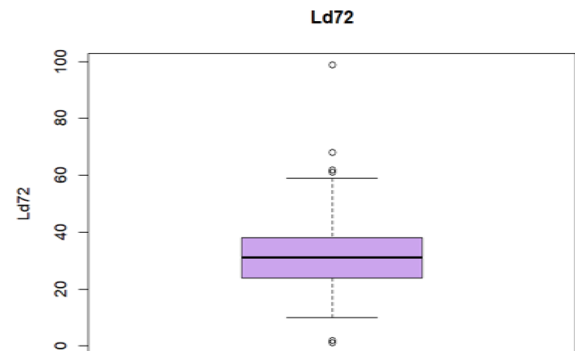
In the scatterplot, we said that there is a negative relation.

In the boxplot, we saw the variation.

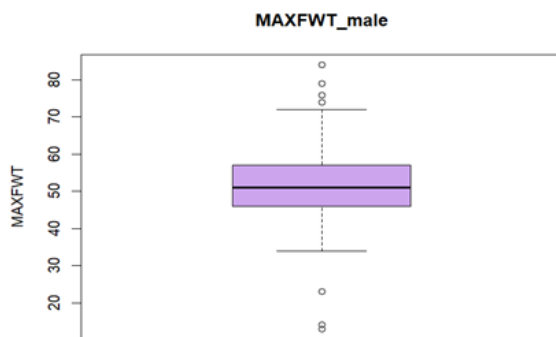
3. Outlier detection



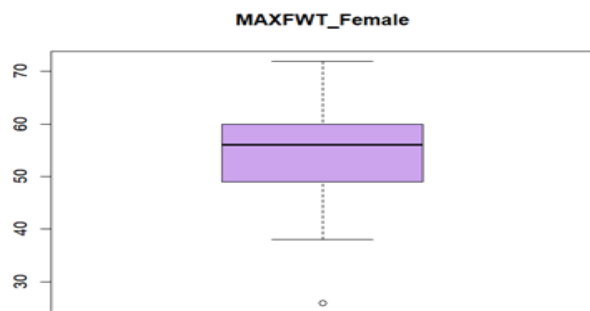
```
> outliers
[1] 84 23 26 13 79 14 13
```



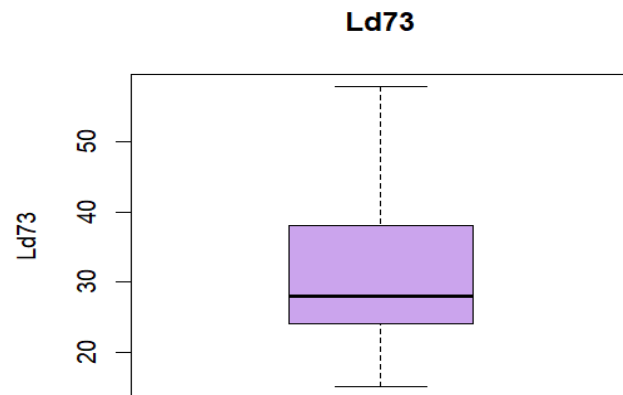
```
> outlier2
[1] 2 1 2 68 62 61 99 99 99 2
```



```
> boxplot(myData[myData$Sex == "Male",]$MAXFWT, main = "MAXFWT_male", ylab = "MAXFWT", col="#c4a4ed")
> outlier4 <- boxplot.stats(myData[myData$Sex == "Male",]$MAXFWT)$out
> outlier4
[1] 74 84 23 76 13 79 14 13
```



```
> outlier5 <- boxplot.stats(myData[myData$Sex == "Female",]$MAXFWT)$out
> outlier5
[1] 26
> |
```



```
> outlier3 <- boxplot.stats(myData$Ld73)$out
> outlier3
integer(0)
```

→ There are some outliers in most variables except some variables such as Ld73 & Age.

Outliers could affect on some tests we use.

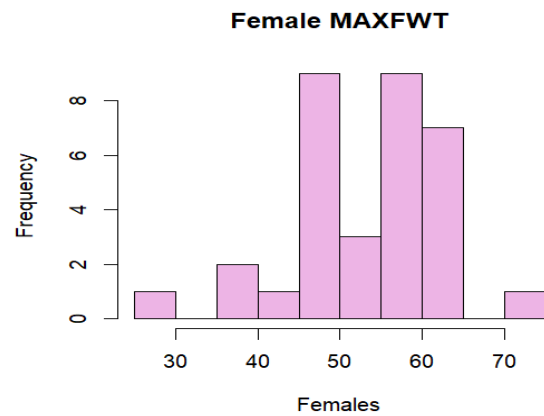
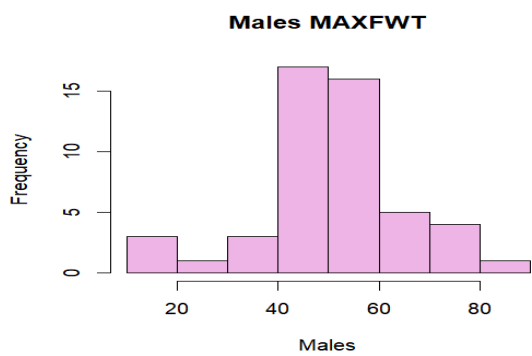
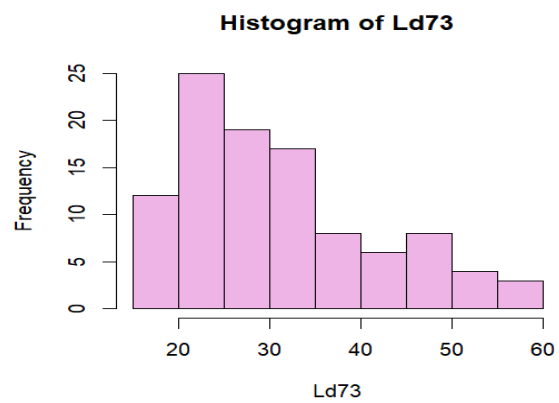
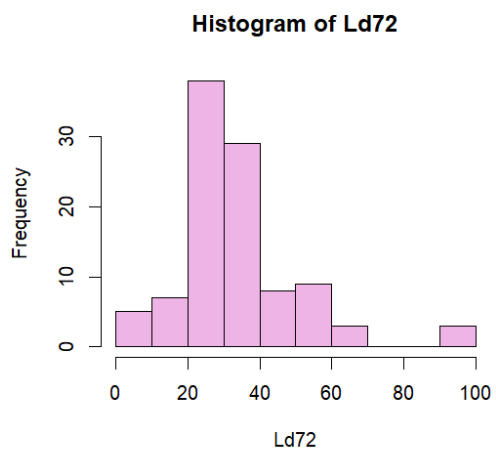
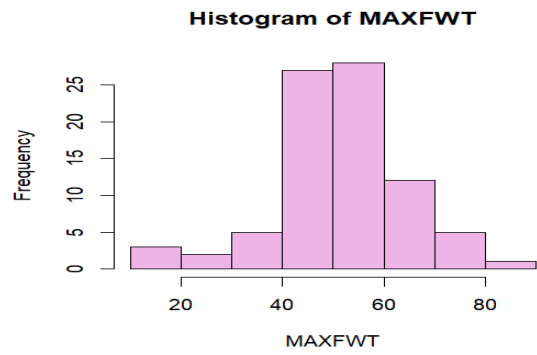
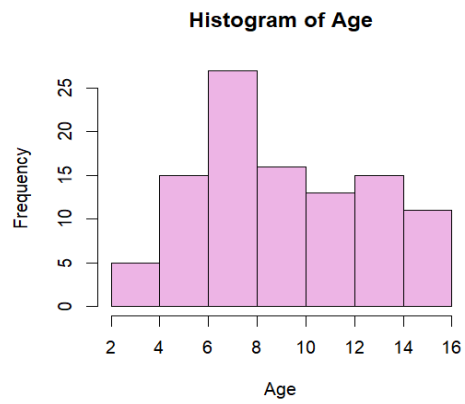
4. Testing for normality/ homoscedasticity

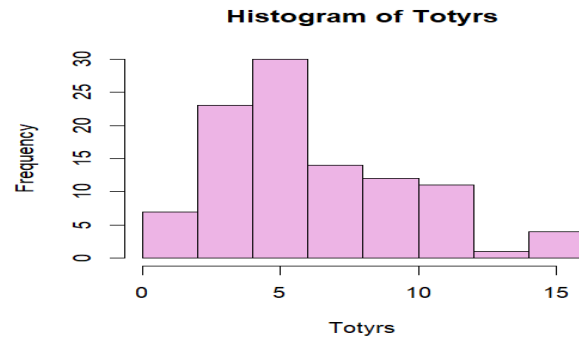
a. We used some methods to test normality as (Histogram & Shapiro test & Q-Q plot) :

So, we visualize using histograms to see the normality of the data to choose the test :

Note → In Histogram (x-axes): v.data & (y-axes): the frequency from the distributions, we can see that the data seems to be not normal may be due to some outliers due to its sensitivity so, we will go further and use Q-Q plots and tests to make sure.

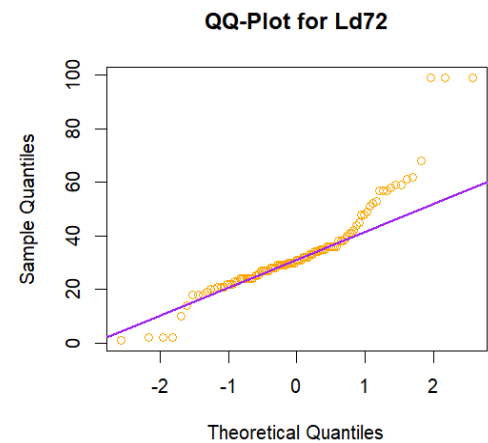
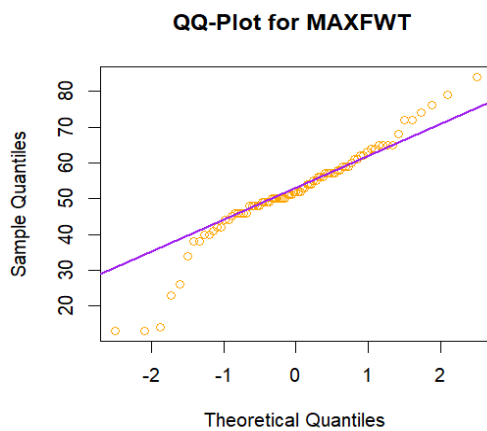
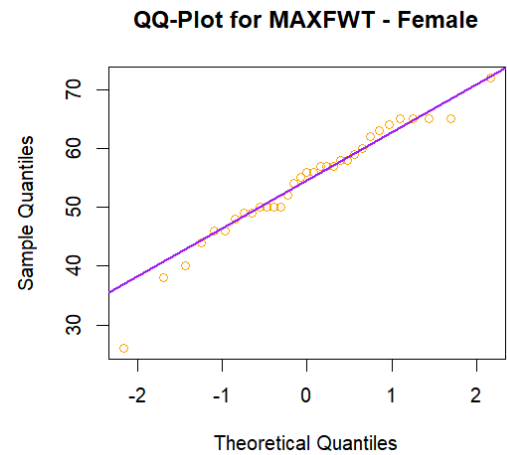
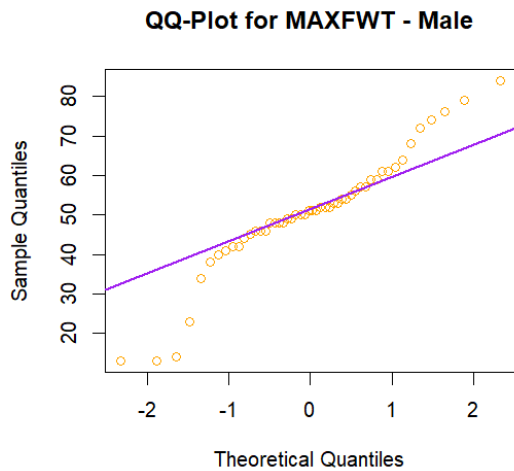
- Also, outliers may be affect on tests.

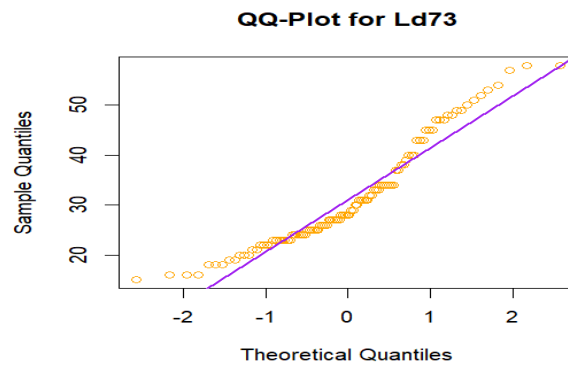




➔ Right Skewed : Age , MAZFWT , Ld72 , Ld73

Also , Q-Q plot :





- All data: Not normal except MAXFWT (Females)

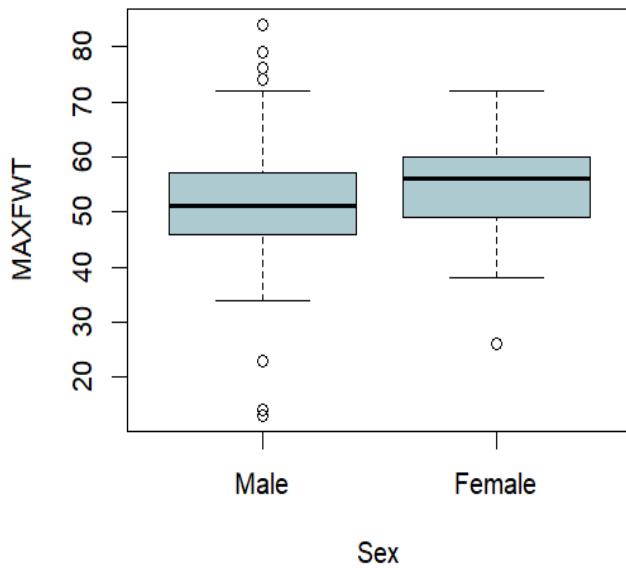
→ **Shapiro : Normality: We Assumed that :**

- a. **Null** : sample distribution is normal
 - b. **Alternative** : sample distribution is not normal
1. The p-value for Age is **0.0004677** → we have enough evidence to reject the null hypothesis. Therefore, the distribution is **not normal**.
 2. The p-value for MAXFWT is **0.0005636** → we have enough evidence to reject the null hypothesis. Therefore, the distribution is **not normal**.
 3. For the gender "Male" → MAXFWT, the p-value is **0.005127**, we have enough evidence to reject the null hypothesis. Therefore, the distribution is **not normal**.
 4. For the gender "Female" → MAXFWT, the p-value is **0.2299**, we **do not** have enough evidence to reject the null hypothesis. Therefore, the distribution is **normal**.
 5. The p-value for Ld72 is **3.188e-08**, we have enough evidence to reject the null hypothesis. Therefore, the distribution is **not normal**.
 6. The p-value for Ld73 is **3.515e-05**, we have enough evidence to reject the null hypothesis. Therefore, the distribution is **not normal**.

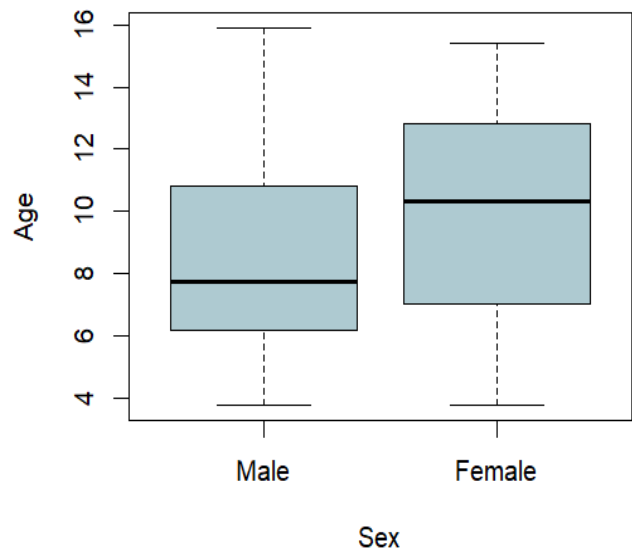
→ **Homoscedasticity: We used (Boxplot & Levene's Test & bartlett):**

- a. We visualize using Boxplot to see if there are differences to choose the test :
- b. Seems equal

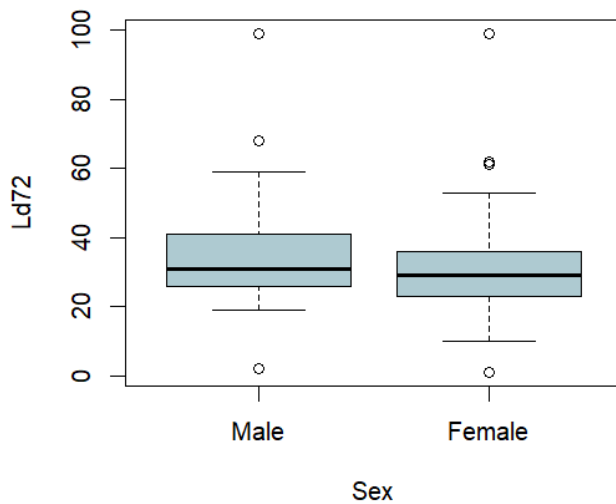
Boxplot MAXFWT (Male & Female)



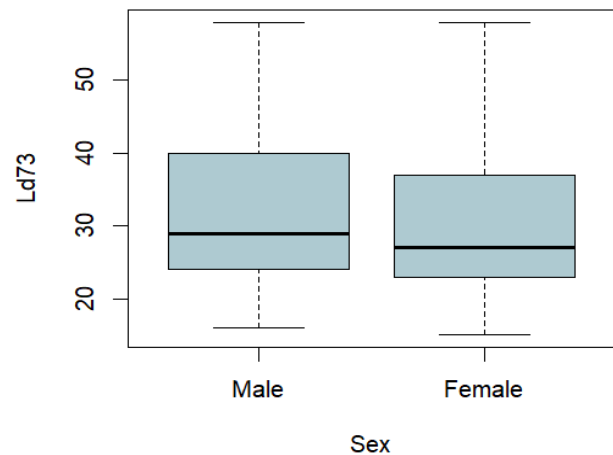
Boxplot Age (Male & Female)



Boxplot Ld72 (Male & Female)



Boxplot Ld73 (Male & Female)



→ **Bartlett Test** : It tests the null hypothesis that the variances of the given variables are equal across different groups or categories. (Data that normally distributed)

1. For MAXFWT, the **p-value** = **0.007503** : less than the conventional significance level of 0.05.

2. As a check we see the **p-value = 0.6986** for Ld72 for both genders are greater than the conventional significance level of 0.05
3. And also, the **p-value = 0.9501** for Ld73 for both genders are greater than the conventional significance level of 0.05

→ **Levene Test : We Assumed:**

- a. **Null** : All groups' variances are equal
 - b. **Alternative** : The variance is not the same for the all the groups
1. For MAXFWT, the p-value (pr(f-value)) is **0.1778**, we **do not** have enough evidence to reject the null hypothesis. Therefore, the variances of all groups are **equal**.
 2. For Age (both genders) , the p-value (pr(f-value)) is **0.389**, we **do not** have enough evidence to reject the null hypothesis. Therefore, the variances of all groups are **equal**.
 3. The p-value (pr(f-value)) for Ld72 for both genders is **0.8542**, we **do not** have enough evidence to reject the null hypothesis. Therefore, the variances of all groups are **equal**.
 4. The p-value (pr(f-value)) for Ld73 for both genders is **0.9177**, we **do not** have enough evidence to reject the null hypothesis. Therefore, the variances of all groups are **equal**.
 5. The p-value (pr(f-value)) for Ld73 for both genders is **0.9177**, we **do not** have enough evidence to reject the null hypothesis. Therefore, the variances of all groups are **equal**.
 6. **Finally**, → Based on the results, it appears that several variables, including Age, MAXFWT, MAXFWT for males, Ld72, and Ld73, **do not follow a normal distribution**. This suggests that the data for these variables may not meet the assumption of normality in statistical analyses that rely on this assumption. **Regarding homoscedasticity**, the results from Levene's test indicate that the variances of these variables are approximately equal across groups. So, this will make us think more about the hypotheses and which tests to use. **Overall**, the departures from normality and homoscedasticity observed in the dataset highlight the importance of careful data analysis and appropriate statistical modeling techniques that can handle these characteristics.
-

5. Statistical Inference

1. 90 percent confidence interval → Male : **47.19258 : 54.16742**

2. 90 percent confidence interval → Female : 51.36321 : 56.87921
3. 95 percent confidence interval → Male : 46.49985 : 54.86015
4. 95 percent confidence interval → Female : 50.80466 : 57.43776
5. 99 percent confidence interval → Male : 45.10538 : 56.25462
6. 99 percent confidence interval → Female : 49.66240 : 58.58003

Applying the 95 percent, the analysis suggests that we can be 95 percent confident that the range of MAXFWT for males is between 46.49985 & 54.86015.

On the other hand, for females, the range is estimated to be between 50.80466 & 57.43776.

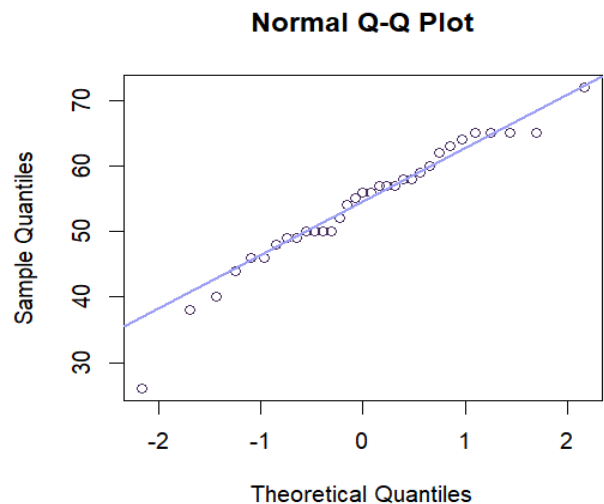
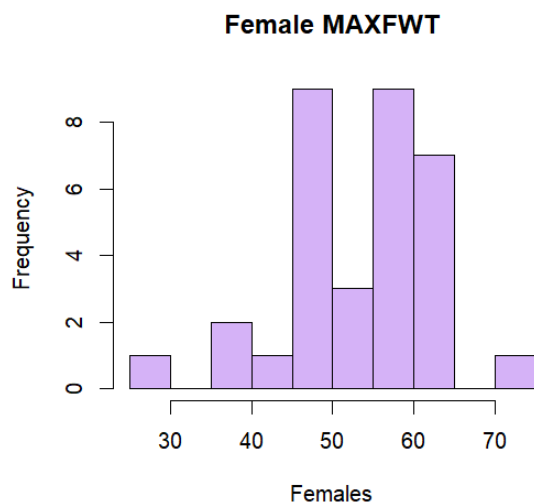
These ranges indicate that females tend to have better MAXFWT values compared to males, as higher scores are considered better.

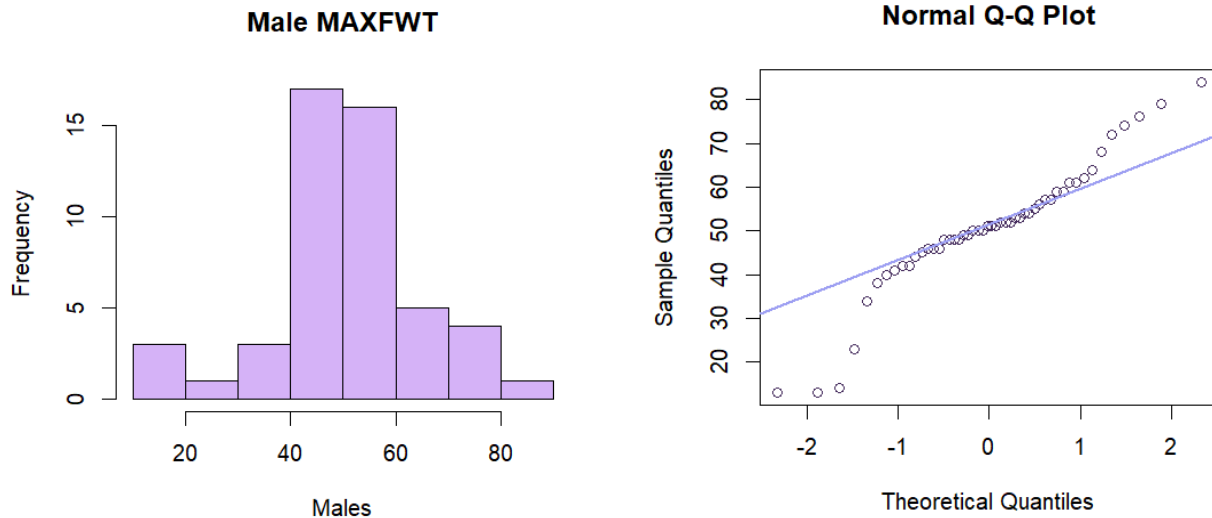
- It's important to note that as the confidence level increases, the range becomes larger. This means that with a higher confidence level, the estimated range becomes broader, allowing for more variability in the data.

6. Hypothesis Testing

- Our hypothesis is that there is a difference in MAXWT between males and females.
- To test this hypothesis using a **statistical hypothesis framework**, we follow the steps below:
 1. We start by stating the **research question**: Is there a difference in MAXWT between males and females?
 2. Next, we convert the research question to a **statistical one**: Does the mean MAXWT for males differ from the mean MAXWT for females?
 3. Stating the null and alternative hypotheses for our test:
 - a. **Null hypothesis**: The mean MAXWT for males is equal to the mean MAXWT for females.
 - b. **Alternative hypothesis**: The mean MAXWT for males is not equal to the mean MAXWT for females.
- **Assuming normality and homoscedasticity**, we use two sample t-test to calculate the p-value.

4. Based on our calculations, we obtain a p-value of **0.2364**, which is **greater than** the significance level (alpha) of **0.05**. Therefore, our result is not statistically significant, and we **do not** have enough evidence to reject the null hypothesis. In other words, we do not have evidence to say that there is a difference in MAXWT between males and females.
5. To determine whether the assumptions for the two-sample t-test were met, we conducted several tests.
 - a. **Firstly**, we assessed **normality** for both males and females using QQ plots, histograms, and the Shapiro-Wilk test. We observed that the **male group** did not meet the normality assumption, as evidenced by the deviation of the QQ plot's line from data points, a non-normal histogram, and a Shapiro-Wilk test p-value of **0.005127**, which was lower than the alpha. **Conversely, the female group** met the normality assumption, with no deviation from data points in the QQ plot, an almost normal histogram, and a Shapiro-Wilk test result of **0.2299**, which was greater than the alpha. Therefore, we assumed that our sex data was not normally distributed.





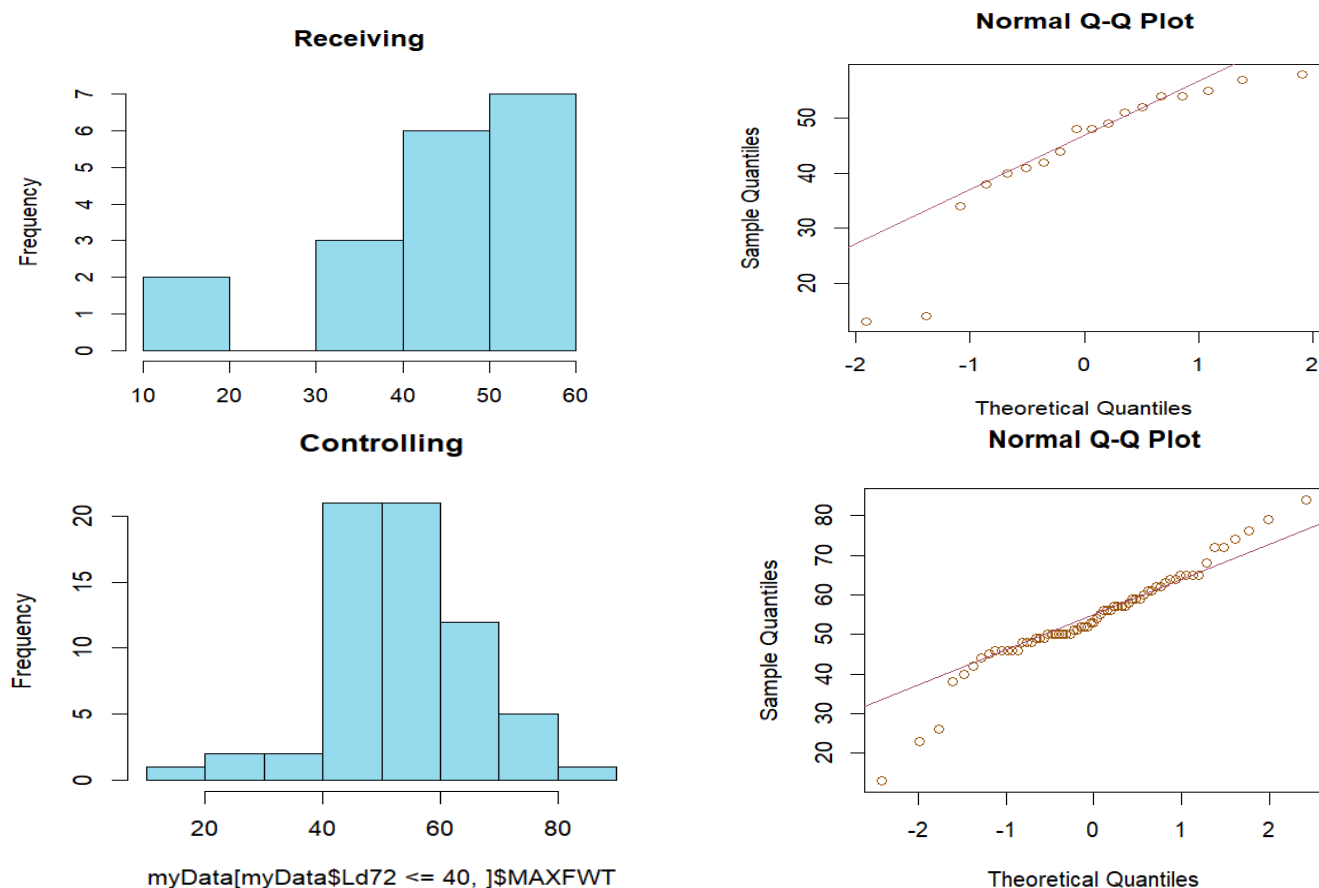
- b. Next, we checked for **homoscedasticity** using the Levene test. The p-value we obtained was **0.1778**, which was greater than the alpha. Thus, we did not have enough evidence to reject the null hypothesis, indicating that the data was homoscedastic.
- c. Since the data was not normally distributed, and not all assumptions were met, we used the Wilcoxon test instead of the two-sample t-test. After performing this test, we obtained a p-value of **0.1399**, which was **greater than** the significance level (alpha) of **0.05**. Therefore, our result was not statistically significant, and we did not have enough evidence to reject the null hypothesis. This indicates that there is no significant difference in MAXWT between males and females.

In summary, we assessed the assumptions for the two-sample t-test and found that normality assumption was not met for males, while it was met for females. We also confirmed homoscedasticity assumption for both groups. Due to the non-normality of the data and not meeting all assumptions, we used the Wilcoxon test and found no significant difference in MAXWT between males and females.

➤ Our hypothesis is that the MAXWT is "lower" in the group receiving Ld72 > 40 compared to the control Ld72 ≤ 40. To test this hypothesis, assuming heteroscedasticity, we follow the steps below:

1. We start by stating the **research question**: Does MAXWT is "lower" in the group receiving Ld72 > 40 compared to the control Ld72 ≤ 40?

2. Next, we convert the research question to a **statistical one**: Does the mean MAXWT for the group receiving Ld72 > 40 differ from the mean MAXWT for the control group Ld72 ≤ 40?
3. We then state the null and alternative hypotheses for our test:
 - a. **Null hypothesis**: The mean MAXWT for the group receiving Ld72 > 40 is not "lower" than the mean MAXWT for the control group Ld72 ≤ 40.
 - b. **Alternative hypothesis**: The mean MAXWT for the group receiving Ld72 > 40 is "lower" than the mean MAXWT for the control group Ld72 ≤ 40.
4. We test normality using QQ plots, histograms, and the Shapiro-Wilk test for the two groups: the group receiving Ld72 > 40 and the control group Ld72 ≤ 40. We observed that both groups did not meet the normality assumption, as evidenced by the deviation of the QQ plot's line from datapoints, non-normal histograms, and Shapiro-Wilk test p-values of **0.005391 (Receiving)** and **0.009785(Controlling)**, respectively.
5. Receiving → Left Skewed, Controlling → Left Skewed



6. Assuming heteroscedasticity and non-normality, we used the Wilcoxon test to calculate the p-value. Upon conducting this test, we obtained a p-value of **0.003036**, which was **smaller than** the significance level (alpha) of **0.05**. Thus, we **had enough** evidence to reject the null hypothesis, indicating that the mean MAXWT for the group receiving Ld72 > 40 was "lower" than the mean MAXWT for the control group Ld72 ≤ 40.
7. To **assess the assumptions for the test**, we conducted a Levene test on the variance of Ld72 that does not assume normality & Pr(>F) value was **0.7382**, which was greater than the alpha. Therefore, we **did not** have enough evidence to reject the null hypothesis, indicating that the data was homoscedastic.

In summary, the assumptions for the test were not fully met, as the data was not normal but was homoscedastic. However, since the Wilcoxon test is appropriate for non-normal and homoscedastic data, we were able to proceed with this test to determine whether there was a difference in MAXWT between the group receiving Ld72 > 40 and the control group Ld72 ≤ 40.

We also used the boxplot → page 11

-
- Our hypothesis is that there is a difference in MAXWT between the different Lead types with different genders. To test this hypothesis, we follow the steps below:
 1. We start by stating the **research question**: Is "MAXWT" different between the different Lead types with the different genders?
 2. We convert the research question to a **statistical one**: Does the mean of MAXWT differ between the groups of Lead types and gender?
 3. We then state the null and alternative hypotheses for our test:
 - a. **Null hypothesis**: There is no difference in the mean MAXWT between the groups of Lead types and gender.
 - b. **Alternative hypothesis**: There is a difference in the mean MAXWT between the groups of Lead types and gender.
 4. Then due to assuming normality and homoscedasticity :
 - a. We performed an **ANOVA** analysis on MAXFWT with the assumption of normality. The Pr(>F) value for Sex was **0.33340**, which was **greater than** the alpha level of **0.05**. Therefore, the result was not significant, and we **did not have**

enough evidence to reject the null hypothesis. Hence, we concluded that there was no significant difference in MAXFWT between different genders.

- b.** However, the $\text{Pr}(> F)$ value for Lead Type was **0.00142**, which was **lower than** the alpha level of **0.05**, indicating that the result was significant. Therefore, we **had enough** evidence to reject the null hypothesis and concluded that there was a significant difference in MAXFWT between different lead types.
 - c.** The $\text{Pr}(> F)$ value for the interaction between **Sex and Lead Type** was **0.11060**, which was **greater than** the alpha level of **0.05**. Hence, the result was not significant, and we **did not have** enough evidence to reject the null hypothesis. Therefore, we concluded that there was no significant difference in MAXFWT between different genders and different lead types.
- 5.** Next, we performed a Tukey test to examine the pairwise differences between the means of different groups.
 - a.** The test included the difference between the means of both cases, the lower and upper bounds of the confidence interval, and the corrected p-value.
 - b.** The p-adjusted values for Male:2-Male:1 \rightarrow **0.0036208** and Male:1-Female:2 **0.0044057** were **less** than the alpha level of 0.05, indicating that the results were significant. Therefore, we **had enough evidence** to reject the null hypothesis, and we concluded that male lead type 2 with male lead type 1 and male lead type 2 with female lead type 1 were different in terms of MAXFWT.
 - c.** Furthermore, **the confidence intervals** for these two comparisons did not contain the value of 0, this means that there is a significant difference between the means of the corresponding groups. The remaining comparisons had p-adjusted values greater than the alpha level of 0.05, indicating no significant differences in terms of MAXFWT. Additionally, **the confidence intervals** for these comparisons contained the value of 0, further supporting the high p-adjusted values.



7. Linear Model

To analyze the relationship between Ld72 and MAXFWT, we fit a linear regression model to the data. Here is a summary of our analysis:

- a. We started by plotting the data and fitting a linear regression line to the plot for the Ld72 column with MAXFWT.
- b. In our linear regression model, the residual median was **0.160**, which is relatively small, indicating that the model is significant.
- c. The estimates for the intercept and slope were **58.4812** and **-0.1887**, respectively, which means that our equation is $y = 58.4812 - 0.1887x$.
- d. The **t-test** value for the slope was **0.0152**, which was **smaller than** the significance level (alpha) of **0.05**. This result was significant, indicating that there is a relationship between our two variables, and **we have enough evidence** to reject the null hypothesis that the slope of the variable is equal to 0.
- e. The multiple R-squared (R^2) value was **0.07053**, which is very low. This value indicates that Ld72 explains only **7.053%** of the variability in MAXFWT.
- f. The residual standard error was **12.51**.
- g. A 95% confidence interval of the regression slope \rightarrow Ld72 will increase between -0.3400905 and -0.03725285

```
> summary(myData.regression)

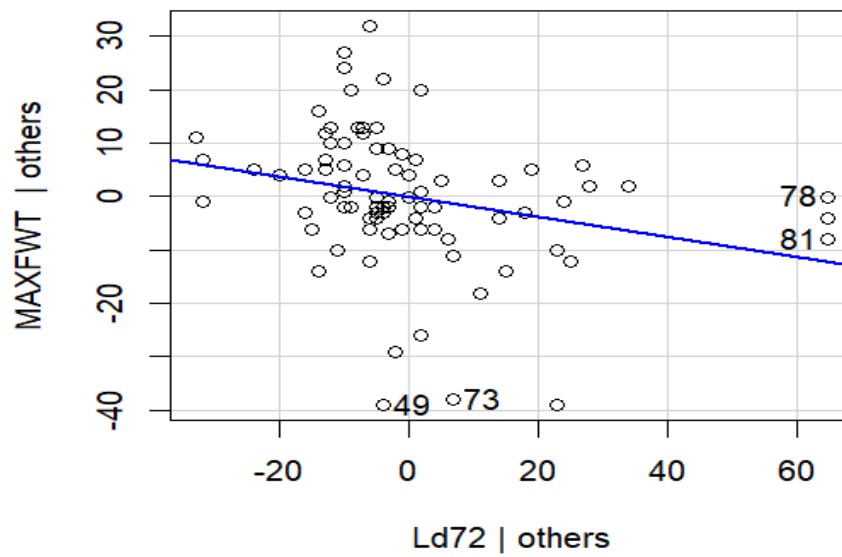
Call:
lm(formula = MAXFWT ~ Ld72, data = myData)

Residuals:
    Min       1Q   Median       3Q      Max
-39.821  -5.500   0.160   7.868  30.802

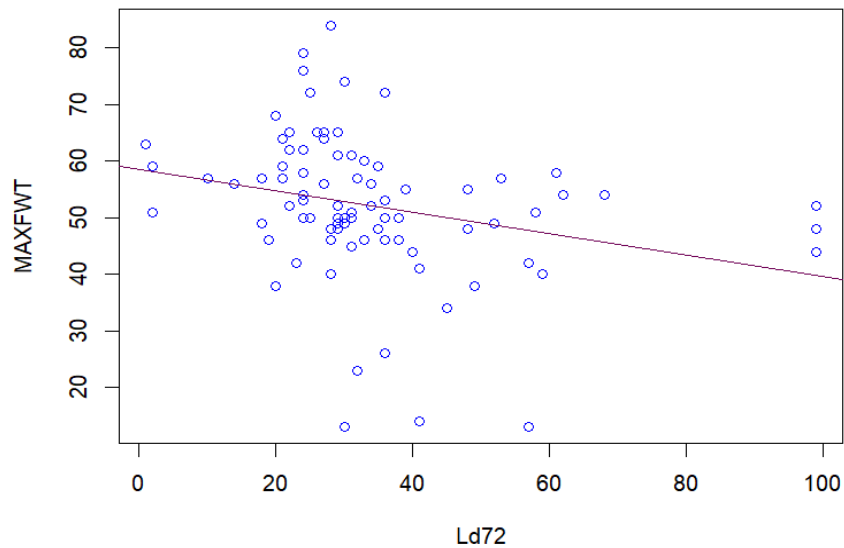
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.4812     2.9357  19.921  <2e-16 ***
Ld72        -0.1887     0.0761  -2.479   0.0152 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.51 on 81 degrees of freedom
(19 observations deleted due to missingness)
Multiple R-squared:  0.07053,    Adjusted R-squared:  0.05905
F-statistic: 6.146 on 1 and 81 DF,  p-value: 0.01524
```

```
> confint(myData.regression, 'Ld72', level = 0.95)
                2.5 %      97.5 %
Ld72 -0.3400905 -0.03725285
>
```



Regression



→ Workload :

1. Hypothesis testing : **All the team**
2. Outlier detection : **Islam & Shady**
3. Normality/ homoscedasticity : **Islam & Shady**
4. Graphics and descriptive data : **All the team**
5. Statistical Inference : **Doaa & Lujain**
6. Linear Model : **Doaa & Lujain**