PROJECT BIOMEDICAL STATISTICS

LEAD CONCENTRATION

Dr. Mohamed Maysara

ISLAM NABIL 202000325
DOAA MAGED 202000840
LUJAIN MOHAMED 202002670
SHADY EL-SHERIF 202002137

CONTENT

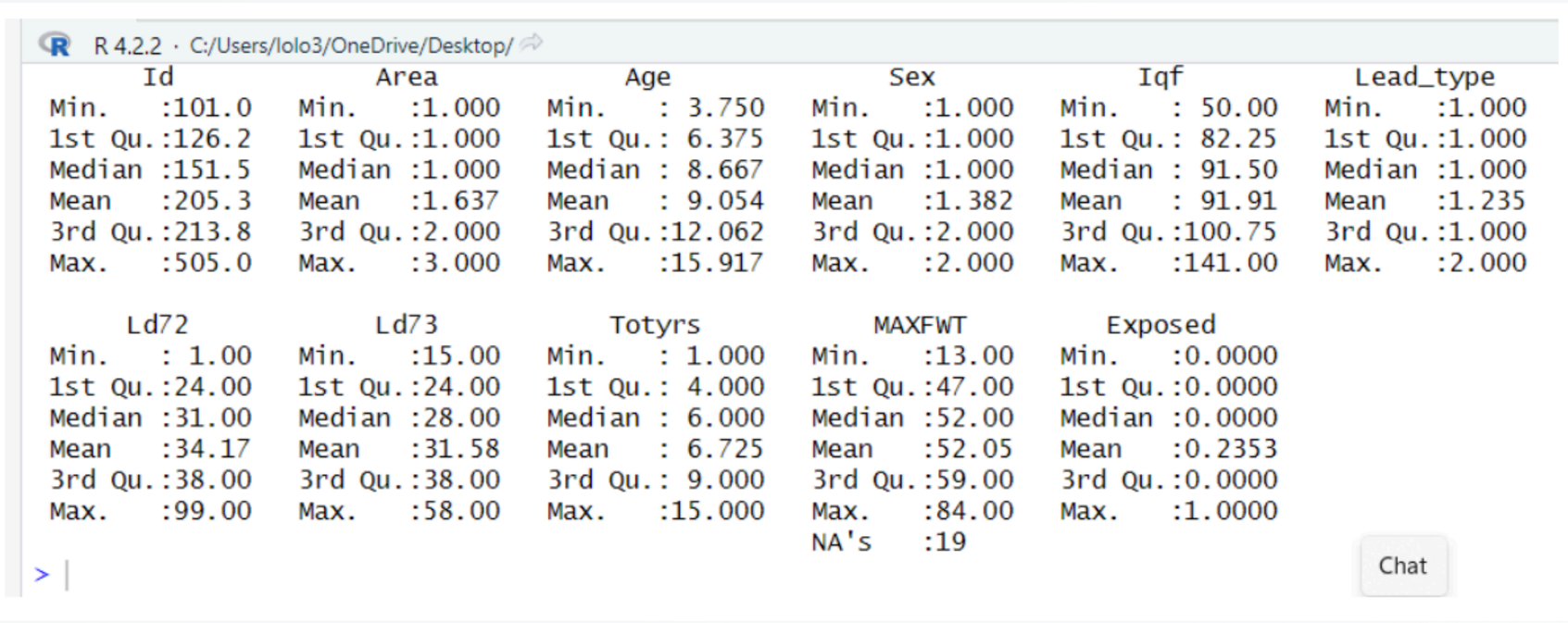
DESCRIPTIVE STATISTICS

FIRST, WE READ THE DATA → RDATA USING LOAD("LEAD.RDATA") AND GET() FUNCTIONS, MULTIPLE FUNCTIONS SUCH AS STR(), SUMMARY() FOR

A. DATA IDENTIFICATION AND EXPLORATION

B. GAINING INSIGHTS INTO THE DATA COLUMNS' DATA TYPES.

C. DATA SUMMARIZATION: COMPUTING THE AVERAGE, MIDDLE VALUE, LOWEST AND HIGHEST VALUES, AS WELL AS THE FIRST AND THIRD QUARTILES



DESCRIPTIVE STATISTICS

CREATING FREQUENCY TABLES

```
> 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
>
```

DESCRIPTIVE STATISTICS

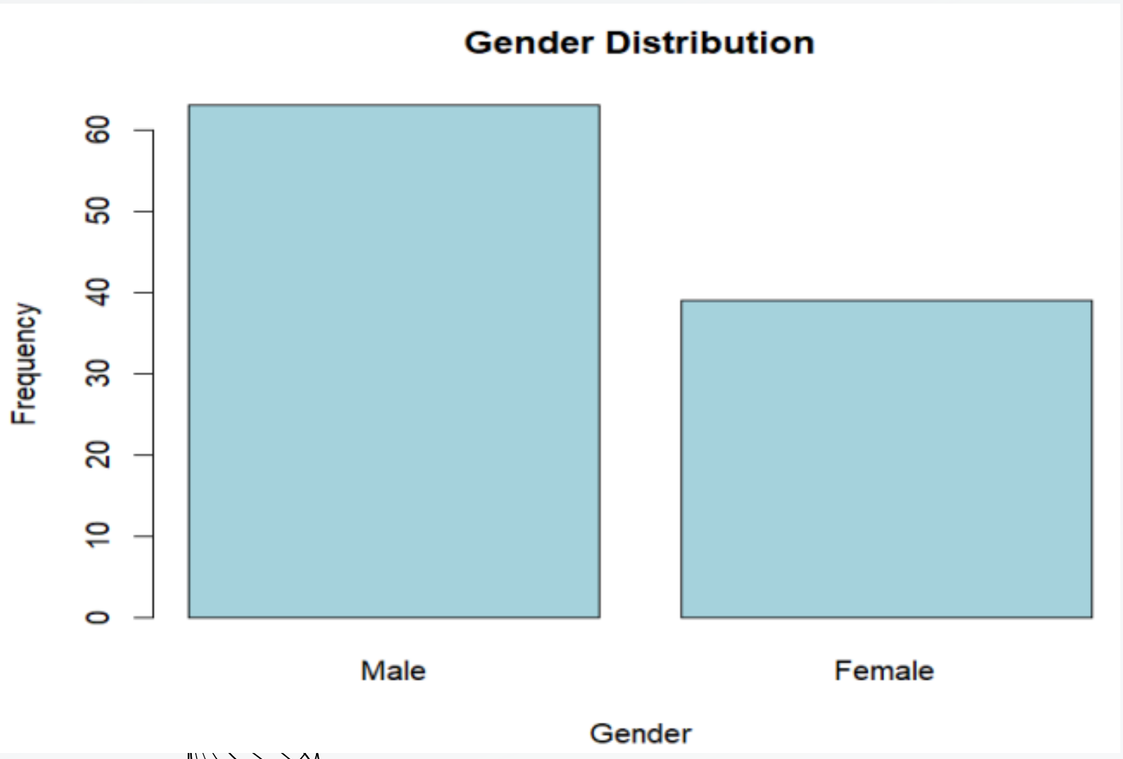
CALCULATING CORRELATION COEFFICIENT

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

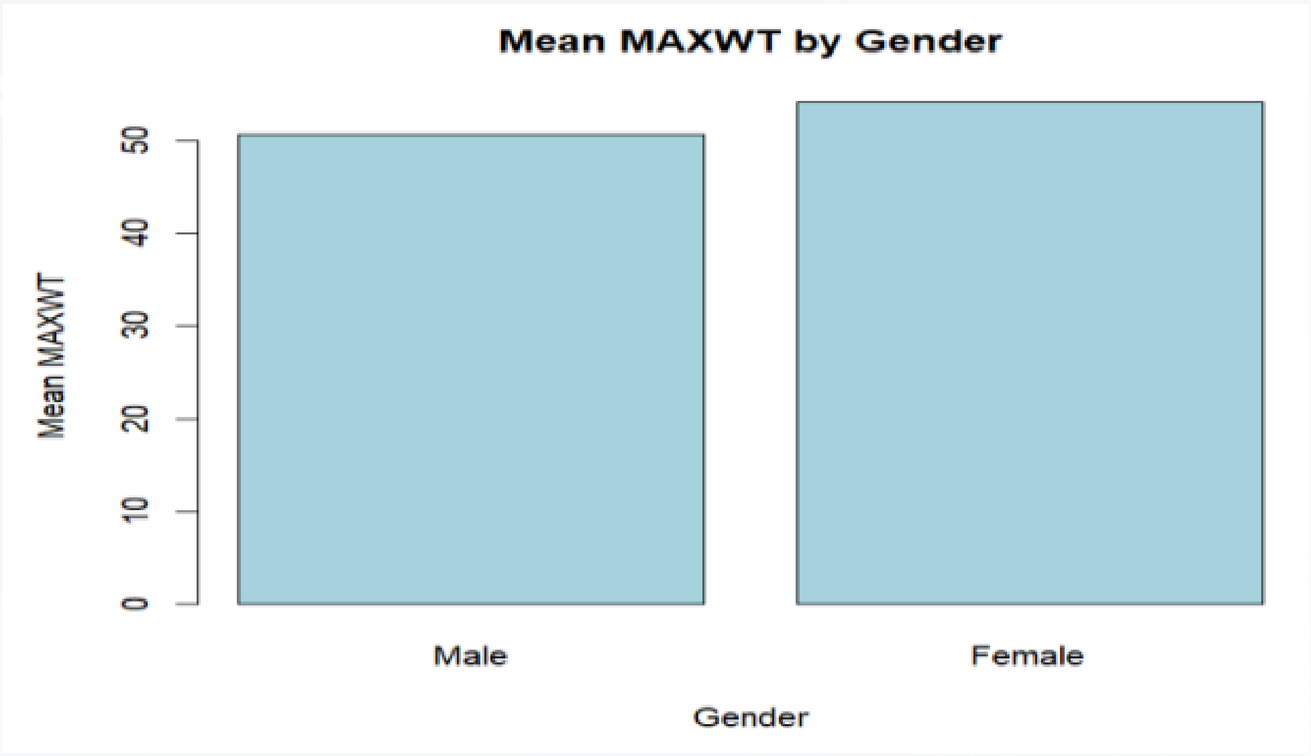
SO WE CAN CONCLUDE THAT THERE IS A WEAK NEGATIVE CORRELATION

GRAPHICS

Generate a bar chart of a categorical variable for the gender (Sex parameter).

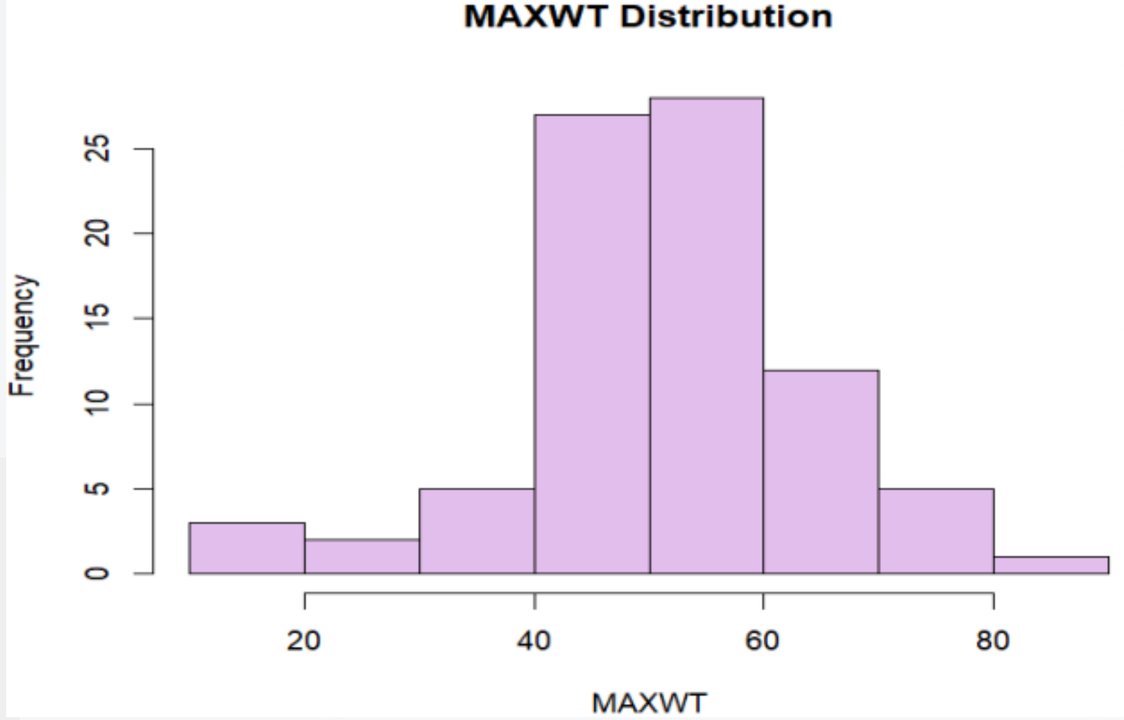


Generate a bar chart graph with mean MAXWT in males and females.



GRAPHICS

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.

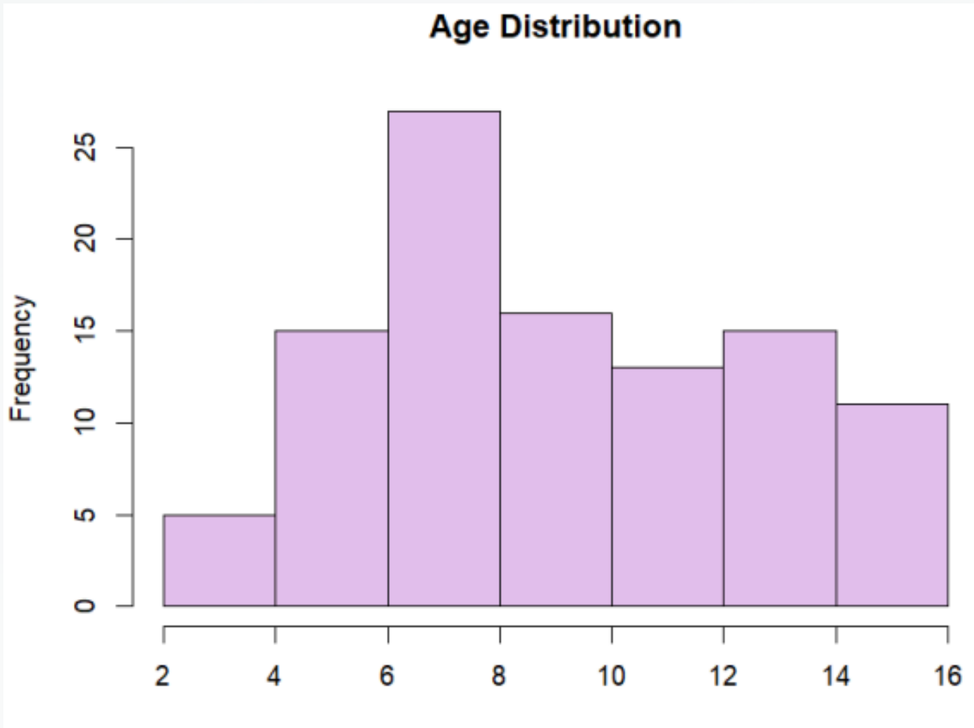
b. Age → Right Skewed

c. MAXFWT → left 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

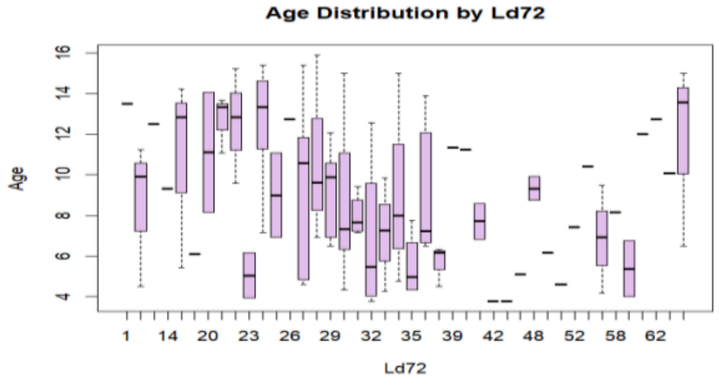
2. Median = 52

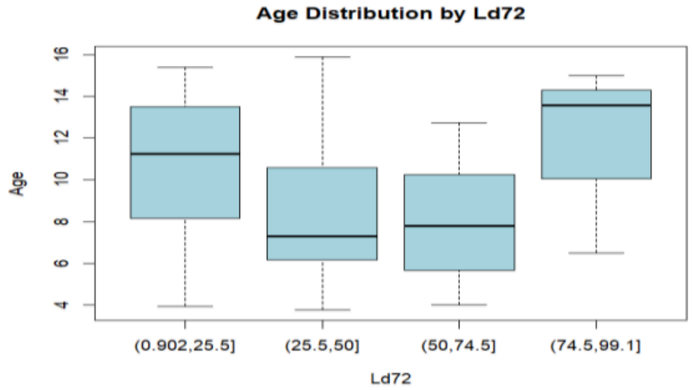


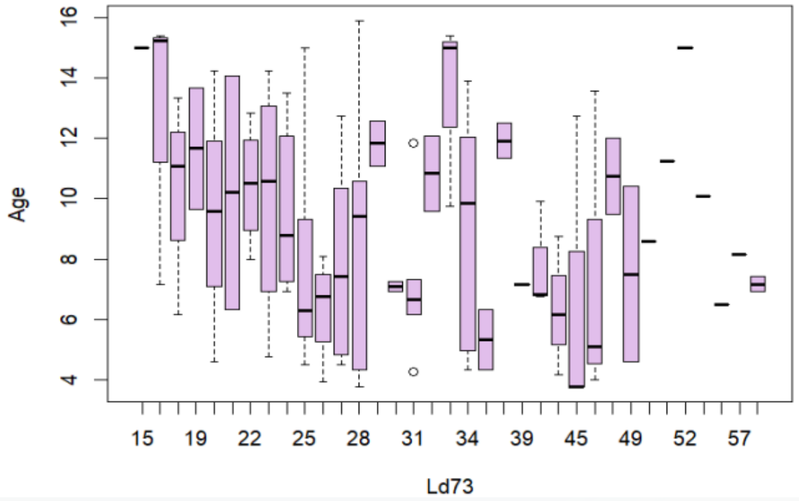
GRAPHICS

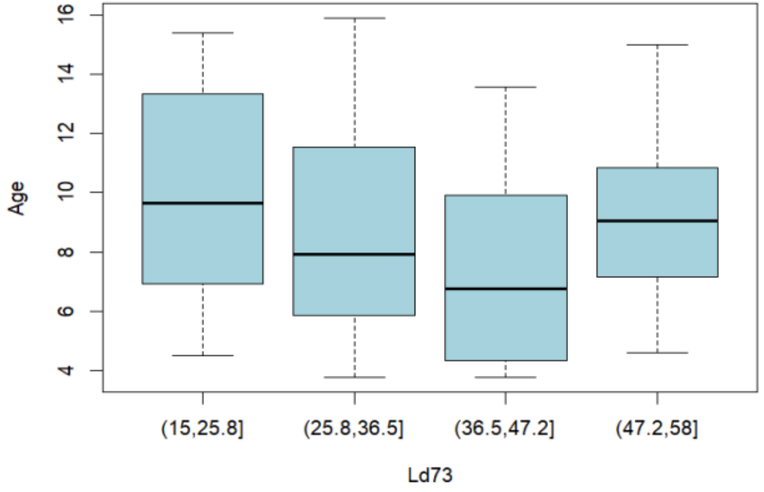
Make a boxplot of age and separate boxplots per Ld72 and per Ld73 (as.factors).

TO ASSESS THE VARIATION OF THE DATA AND SEPARATE IT INTO 4 INTERVALS TO SEE THE VARIATION. (SEEMS THAT → NOT EQUAL VARIANCE)



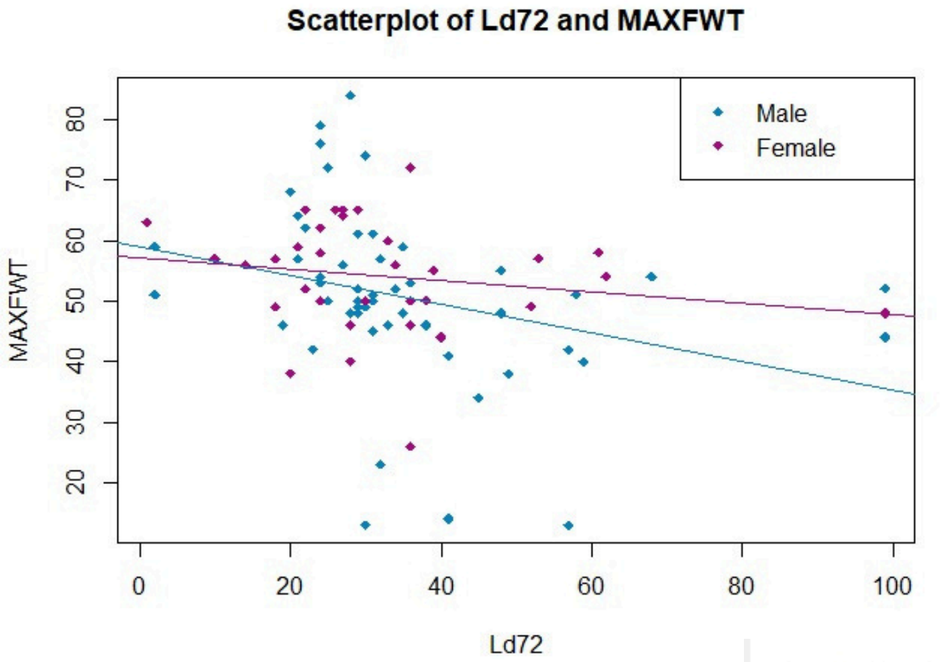






GRAPHICS

Make a scatterplot of 2 continuous variables Ld72 and MAXWT and add the regression lines for each gender.

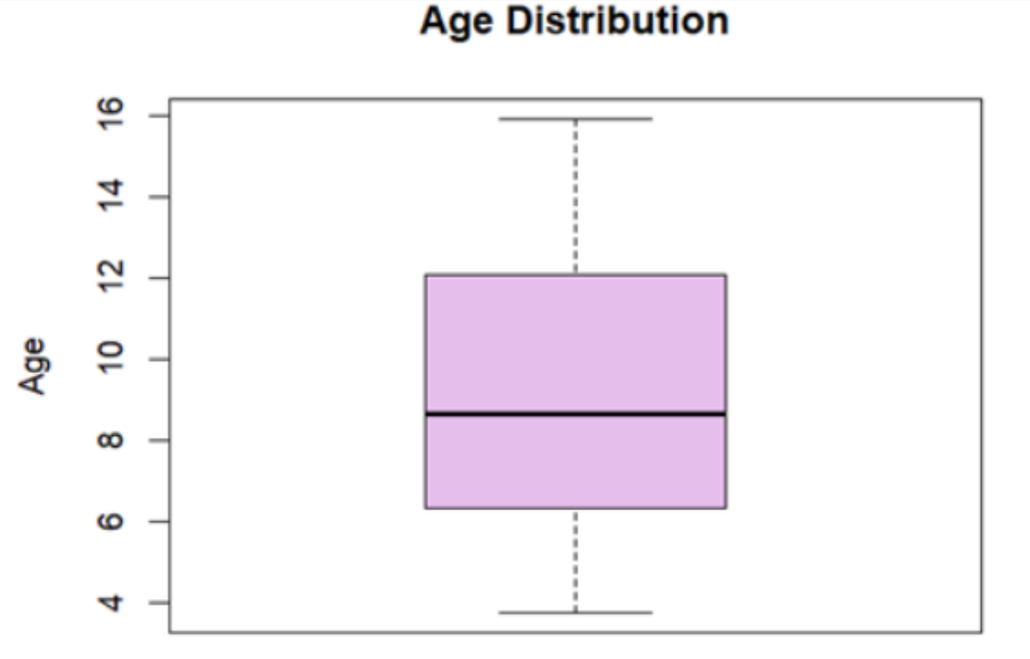



```r
# Make a scatterplot of 2 continuous variables Ld72 and MAXWT, and add the regression lines for each gender
plot(myData$Ld72, myData$MAXFWT, xlab = "Ld72", ylab = "MAXFWT", main = "Scatterplot of Ld72 and MAXFWT",pch=18)
points(subset(myData, Sex == "Male")$Ld72, subset(myData, Sex == "Male")$MAXFWT, col = "#0983b8",pch=18)
points(subset(myData, Sex == "Female")$Ld72, subset(myData, Sex == "Female")$MAXFWT, col = "#a30b82",pch=18)

# Add regression lines for each gender
abline(lm(MAXFWT ~ Ld72, data = subset(myData, Sex == "Male")), col = "#0983b8")
abline(lm(MAXFWT ~ Ld72, data = subset(myData, Sex == "Female")), col = "#a30b82")
legend("topright", legend = c("Male", "Female"), col = c("#0983b8", "#a30b82"), pch = 18)
```

GRAPHICS

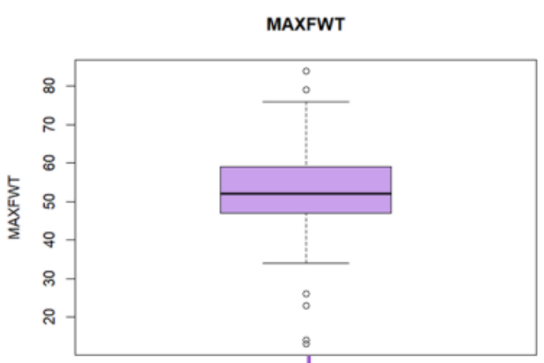
- TO SUMMARIZE ALL GRAPHS, IN HISTOGRAM (X-AXES): V.DATA & (Y-AXES): THE FREQUENCY FROM THE DISTRIBUTIONS, WE CAN SEE THAT THE DATA SEEMS TO BE NOT NORMAL MAY BE DUE TO SOME OUTLIERS DUE TO ITS SENSITIVITY SO, WE WILL GO FURTHER AND USE Q-Q PLOTS AND TESTS TO MAKE SURE.
- ALSO, OUTLIERS MAY BE AFFECTED ON TESTS. IN THE SCATTERPLOT, WE SAID THAT THERE IS A NEGATIVE RELATION. IN THE BOXPLOT, WE SAW THE VARIATION.



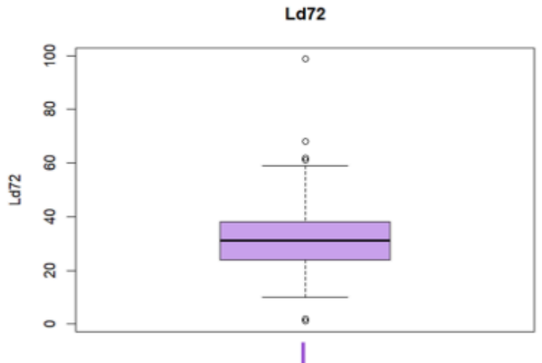
AS WE SEE THERE ARE NO OUTLIERS

3-OUTLIER DETECTION

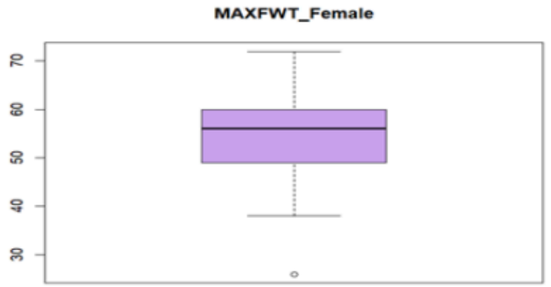
There are some outliers in most variables except some variables such as Ld73 & Age. Outliers could affect some tests we use.



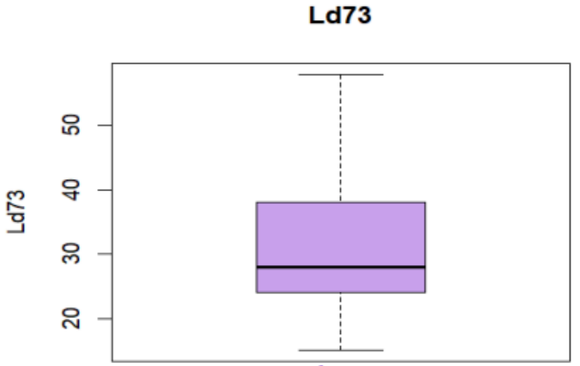

```
> outliers
[1] 84 23 26 13 79 14 13
```



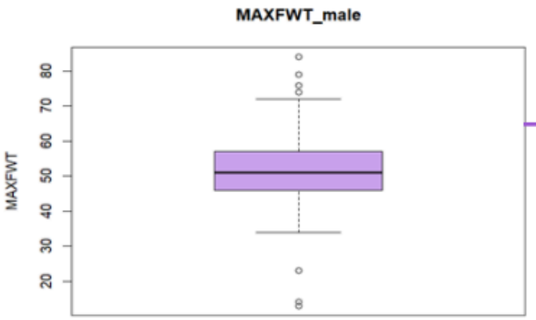

```
> outlier2
[1]  2  1  2 68 62 61 99 99 99  2
```




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



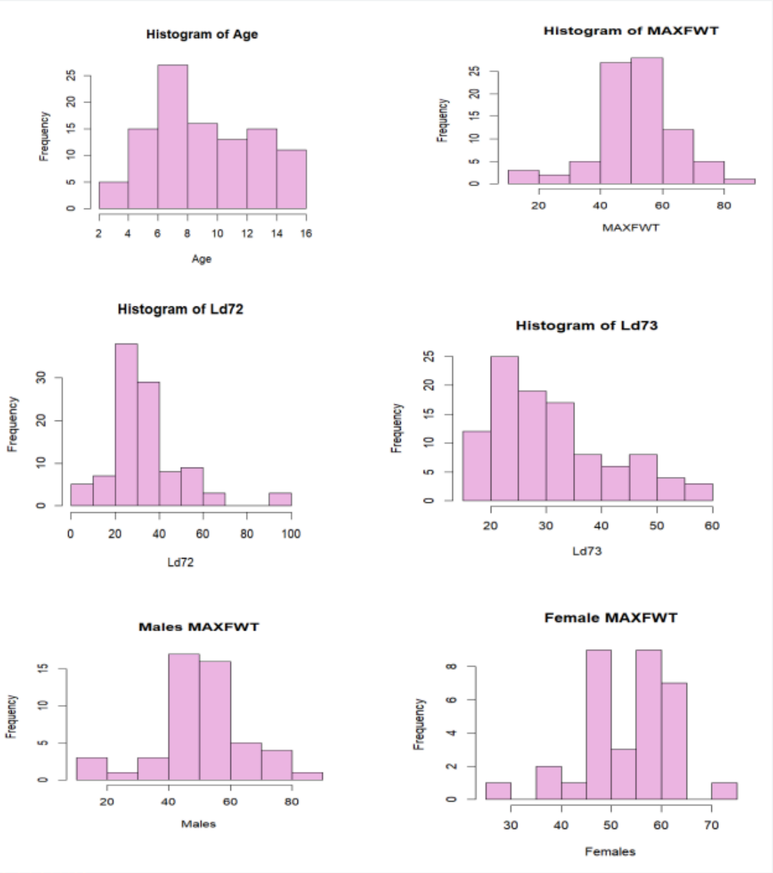

```
> outlier3 <- boxplot.stats(myData$Ld73)$out
> outlier3
integer(0)
```



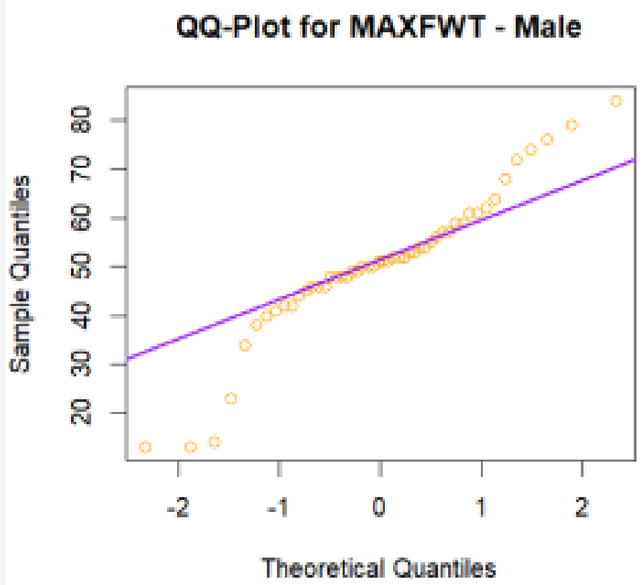

```
> boxplot(myData[myData$Sex == "Male",]$MAXFWT, main = "MAXFWT_male", ylab = "MAXFWT",col="#cba4ed
> outlier4 <- boxplot.stats(myData[myData$Sex == "Male",]$MAXFWT)$out
> outlier4
[1] 74 84 23 76 13 79 14 13
>
```

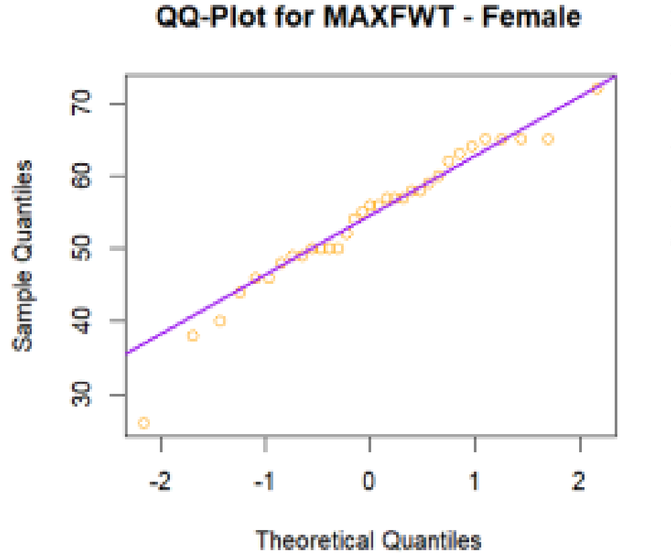
4- TESTING FOR NORMALITY/ HOMOSCEDASTICITY

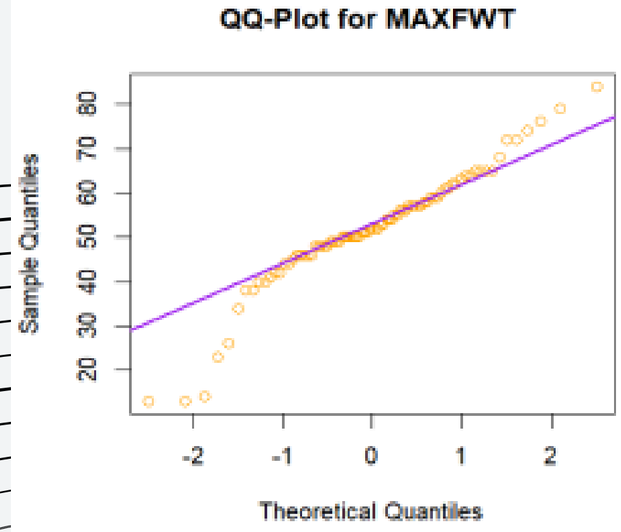
We used some methods to test normality as (Histogram & Shapiro test & Q-Q plot) : So, we visualize using histograms to see the normality of the data to choose the test : Note → In Histogram (x-axes): v.data & (y-axes): the frequency from the distributions, we can see that the data seems to be not normal 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.

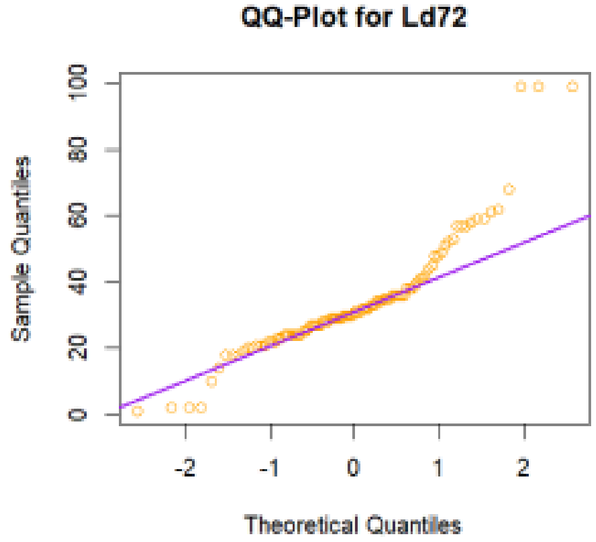


Also , Q-Q plot


QQ-Plot for MAXFWT - Male
Sample Quantiles
Theoretical Quantiles
20
30
40
50
60
70
80
-2
-1
0
1
2


QQ-Plot for MAXFWT - Female
Sample Quantiles
Theoretical Quantiles
30
40
50
60
70
-2
-1
0
1
2


QQ-Plot for MAXFWT
Sample Quantiles
Theoretical Quantiles
20
30
40
50
60
70
80
-2
-1
0
1
2


QQ-Plot for Ld72
Sample Quantiles
Theoretical Quantiles
0
20
40
60
80
100
-2
-1
0
1
2

Shapiro : Normality: We Assumed that :

Null : sample distribution is normal

Alternative : sample distribution is not normal

So we have enough evidence to reject the null hypothesis in all except **"Female" → MAXFWT**

The p-value for Age is 0.0004677

The p-value for MAXFWT is 0.0005636

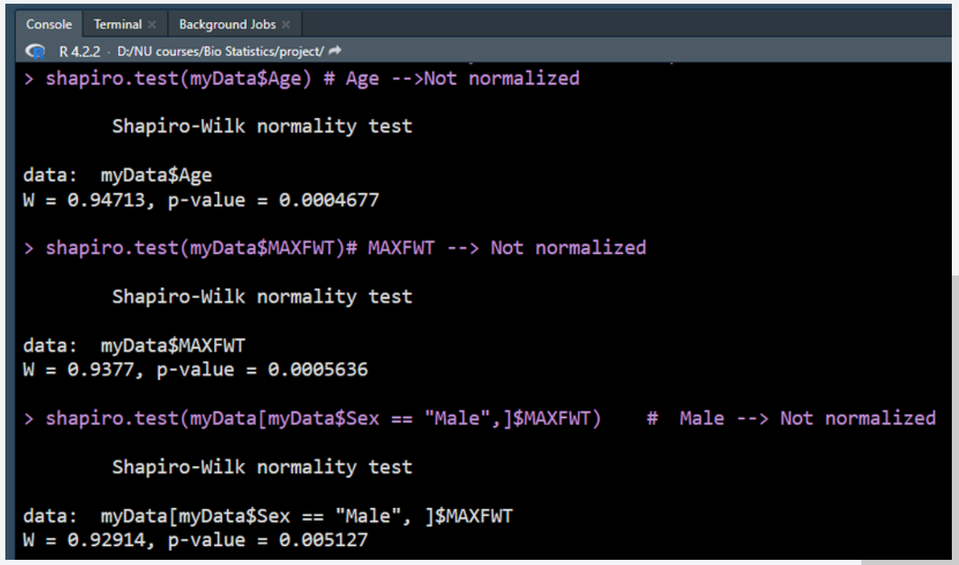
For the gender "Male" → MAXFWT, the p-value is 0.005127

For the gender "Female" → MAXFWT, the p-value is 0.2299

The p-value for Ld72 is 3.188e-08

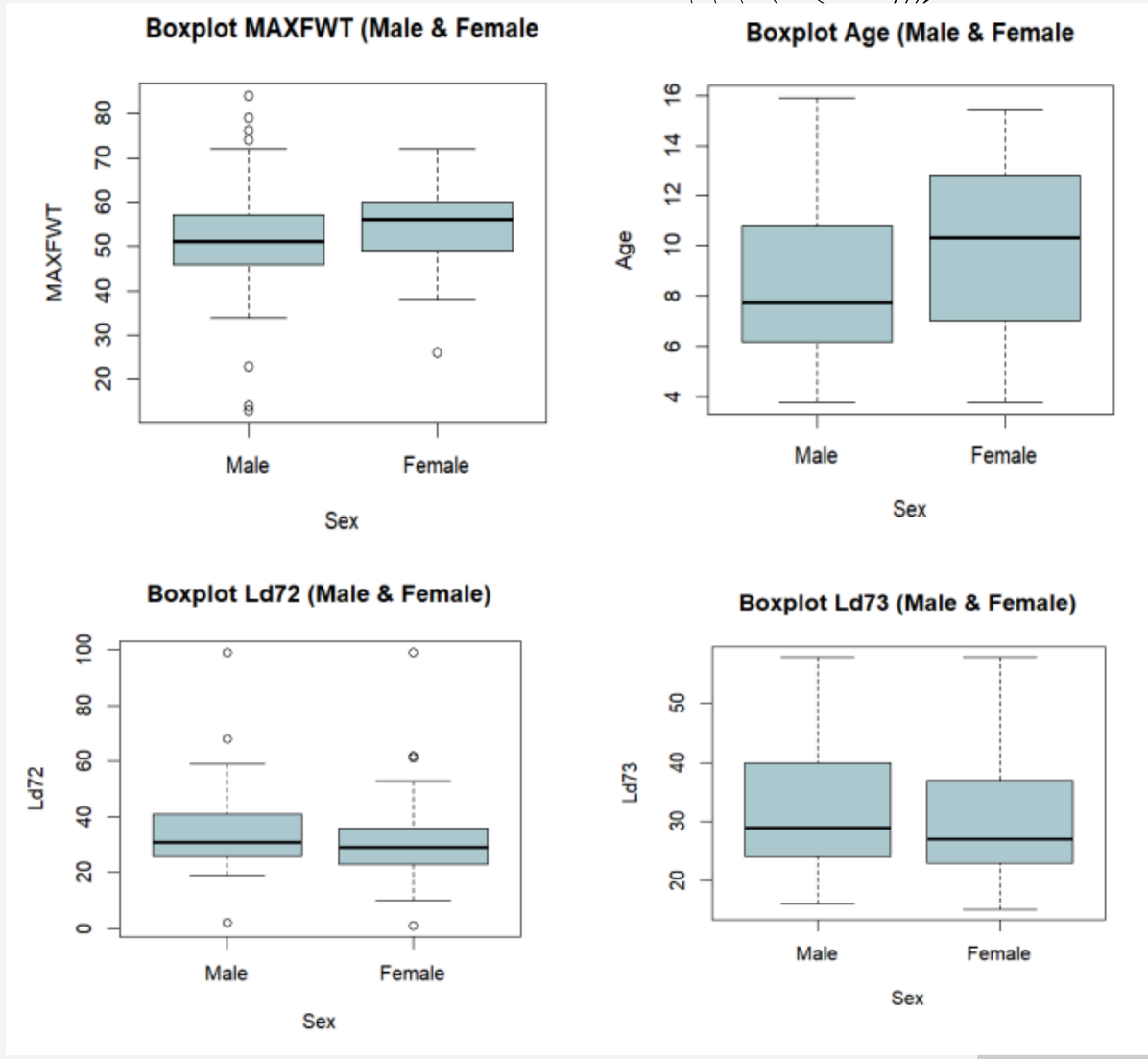
The p-value for Ld73 is 3.515e-05

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



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



Levene Test : Assumes:

a. Null : All groups' variances are equal
b. Alternative : The variance is not the same for the all the groups

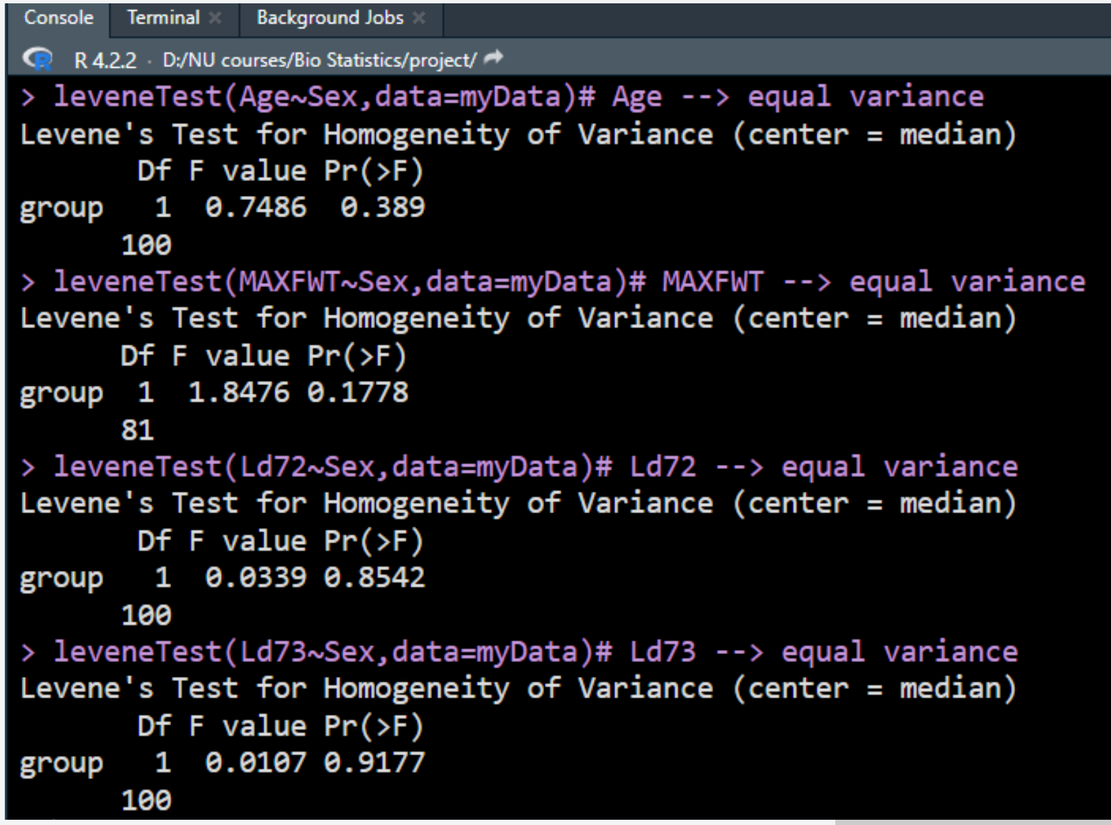
For MAXFWT, the p-value (pr(f-value)) is 0.1778

For Age (both genders) , the p-value (pr(f-value)) is 0.389

The p-value (pr(f-value)) for Ld72 for both genders is 0.8542

The p-value (pr(f-value)) for Ld73 for both genders is 0.9177

The p-value (pr(f-value)) for Ld73 for both genders is 0.9177

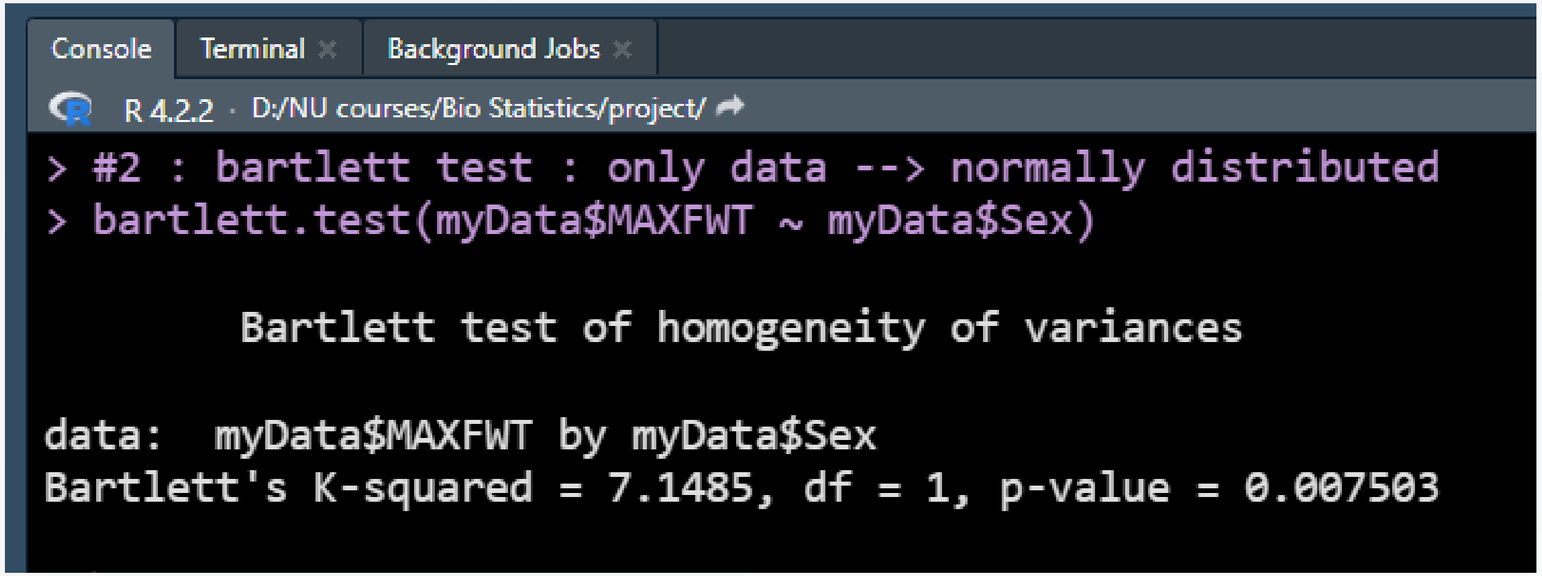


Results-> Regarding homoscedasticity, the results from Levene's test indicate that the variances of these variables are approximately equal across groups. So, this will make us think more about the hypotheses and which tests to use.

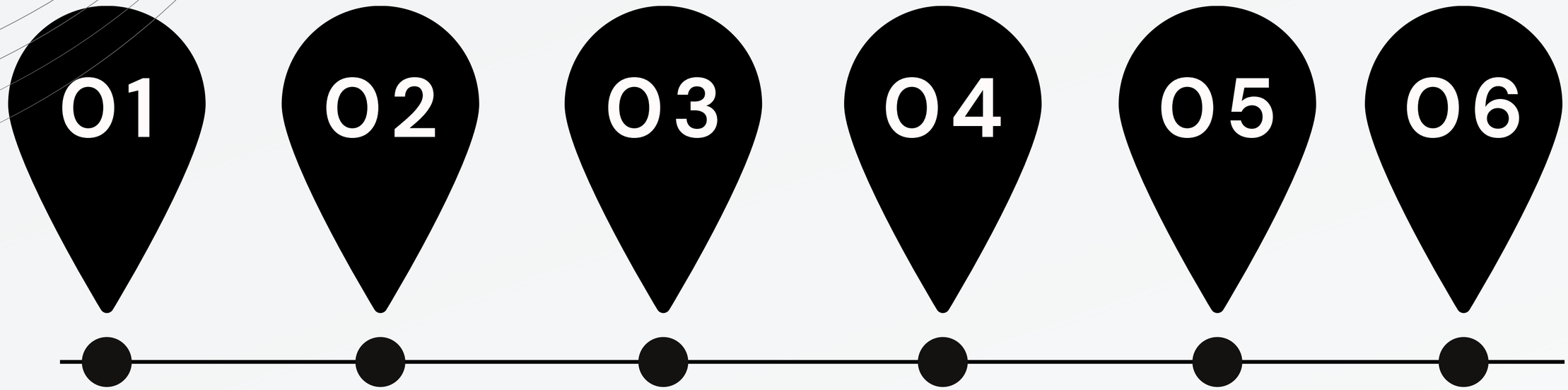
BARTLETT TEST

It tests the null hypothesis that the variances of the given variables are equal across different groups or categories. (Data that is normally distributed)

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



Statistical Inference



90 percent confidence interval → Male : 47.19258 : 54.16742

90 percent confidence interval → Female : 51.36321 : 56.87921

95 percent confidence interval → Male : 46.49985 : 54.86015

95 percent confidence interval → Female : 50.80466 : 57.43776

99 percent confidence interval → Male : 45.10538 : 56.25462

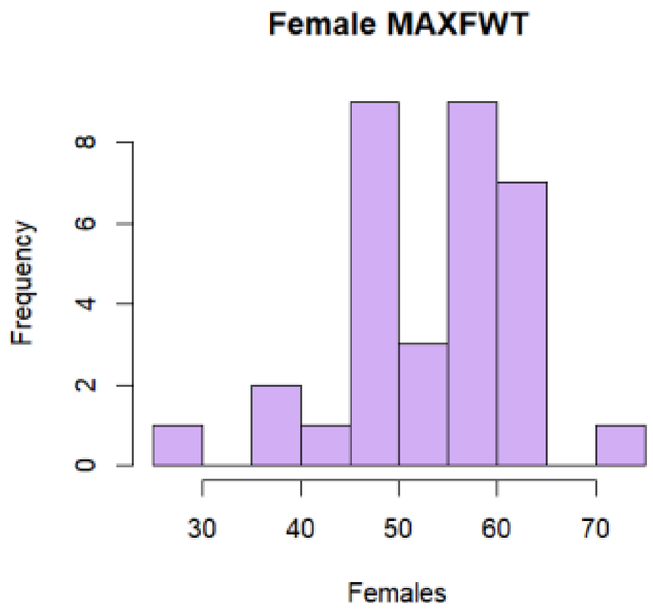
99 percent confidence interval → Female : 49.66240 : 58.58003

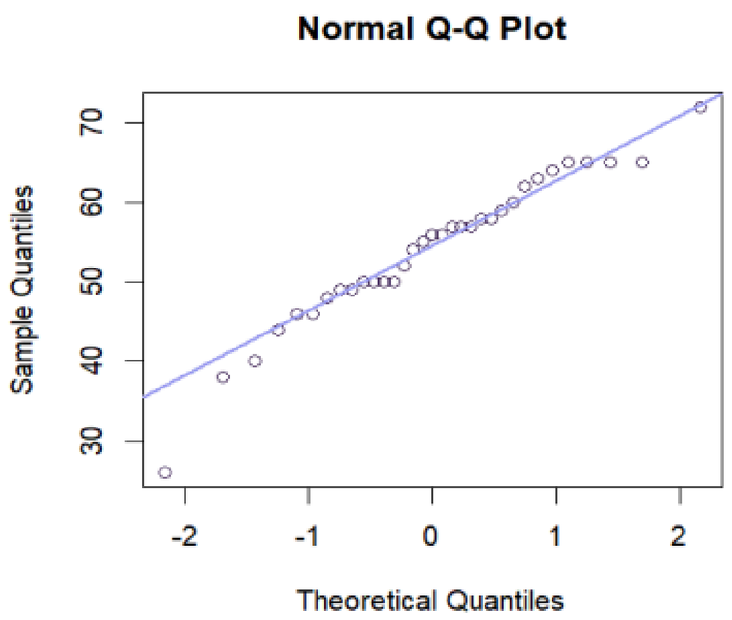
Hypothesis Testing

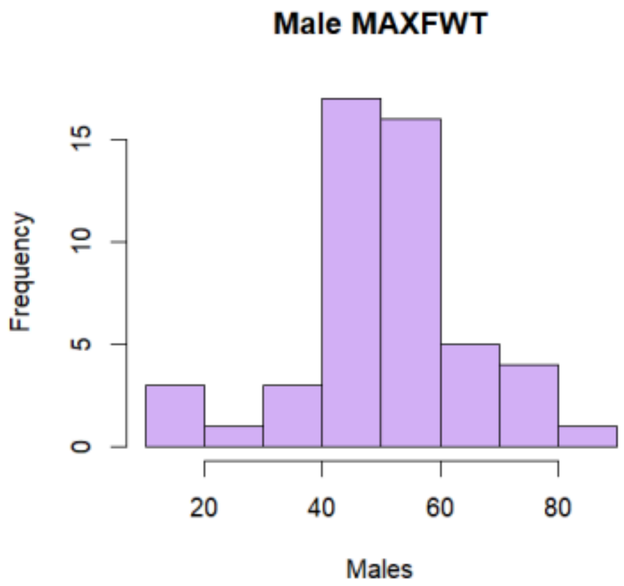
Our hypothesis is that there is a difference in MAXWT between males and females.

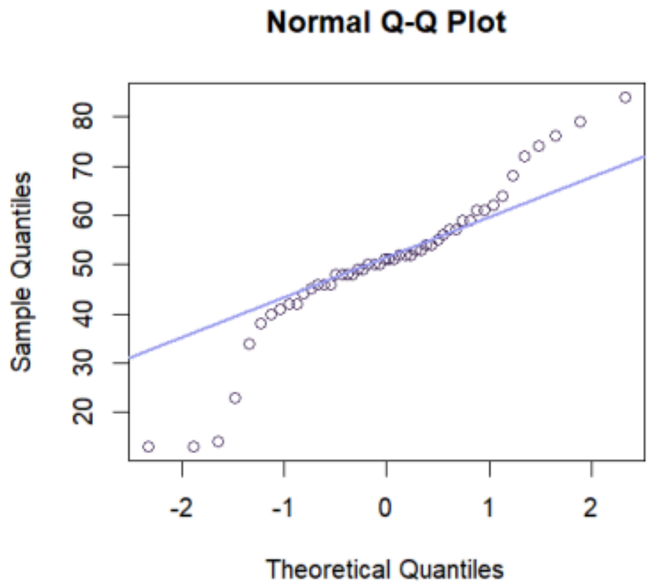
1. We start by stating the research question: Is there a difference in MAXWT between males and females?
2. Next, we convert the research question to a statistical one: Does the mean MAXWT for males differ from the mean MAXWT for females?
3. Stating the null and alternative hypotheses for our test:
   - Null hypothesis: The mean MAXWT for males is equal to the mean MAXWT for females.
   - Alternative hypothesis: The mean MAXWT for males is not equal to the mean MAXWT for females.
     - **Assuming normality and homoscedasticity, we use a 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.

Hypothesis Testing



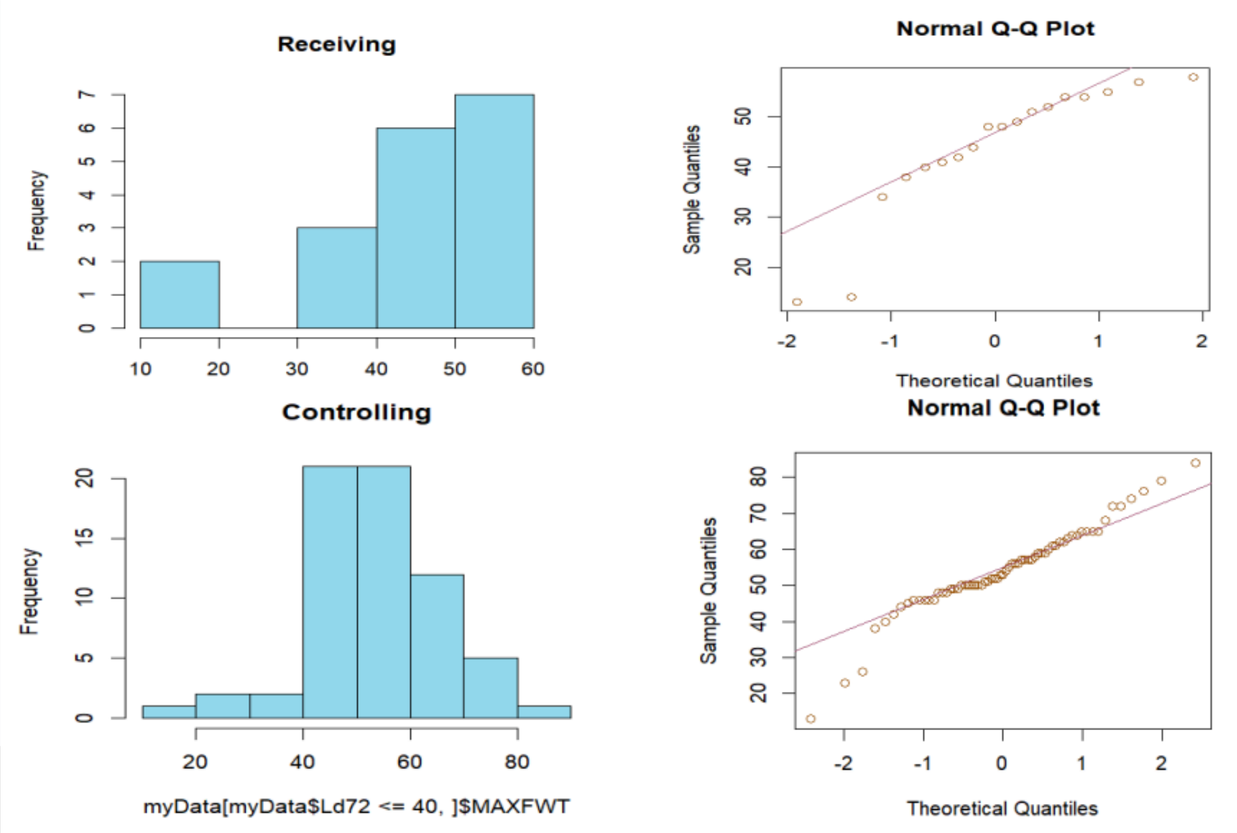






In summary, we assessed the assumptions for the two-sample t-test and found that normality assumption was not met for males, while it was met for females. We also confirmed homoscedasticity assumption for both groups. Due to the non-normality of the data and not meeting all assumptions, we used the Wilcox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

Hypothesis Testing



In summary, the assumptions for the test were not fully met, as the data was not normal but was homoscedastic. However, since the **Wilcox 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.

Hypothesis Testing

Our hypothesis is that there is a difference in MAXWT between the different Lead types with different genders.

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:

- **Null hypothesis: There is no difference in the mean MAXWT between the groups of Lead types and gender.**
- **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 :

- **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.**
- **However, the 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.**
- **The 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.**

Hypothesis Testing

5 Next, we performed a Tukey test to examine the pairwise differences between the means of different groups.

- 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.
- The p-adjusted values for Male:2-Male:1→ 0.0036208 and Male:1-Female:20.0044057 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 2with female lead type, 1 was different in terms of MAXFWT.
- Furthermore, the confidence intervals for these two comparisons did not contain the value of, 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.

ANOVA TEST

Console | Terminal | Background Jobs

R 4.2.2 · D:/NU courses/Bio Statistics/project/

```
> summary(v1)
             Df Sum Sq Mean Sq F value  Pr(>F)
Lead_type     1   1596  1596.1  10.943 0.00142 **
Sex           1    138   138.1   0.947 0.33340
Lead_type:Sex 1    380   379.7   2.604 0.11060
Residuals    79  11522   145.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
19 observations deleted due to missingness
```

R 4.2.2 · D:/NU courses/Bio Statistics/project/

```
> report(v1)
The ANOVA (formula: MAXFWT ~ Lead_type * Sex) suggests that:

  - The main effect of Lead_type is statistically significant and medium (F(1, 79) = 10.94, p = 0.001; Eta2
(partial) = 0.12, 95% CI [0.03, 1.00])
  - The main effect of Sex is statistically not significant and small (F(1, 79) = 0.95, p = 0.333; Eta2 (partial) =
0.01, 95% CI [0.00, 1.00])
  - The interaction between Lead_type and Sex is statistically not significant and small (F(1, 79) = 2.60, p =
0.111; Eta2 (partial) = 0.03, 95% CI [0.00, 1.00])
```

Tukey test

Console | Terminal | Background Jobs

R 4.2.2 · D:/NU courses/Bio Statistics/project/

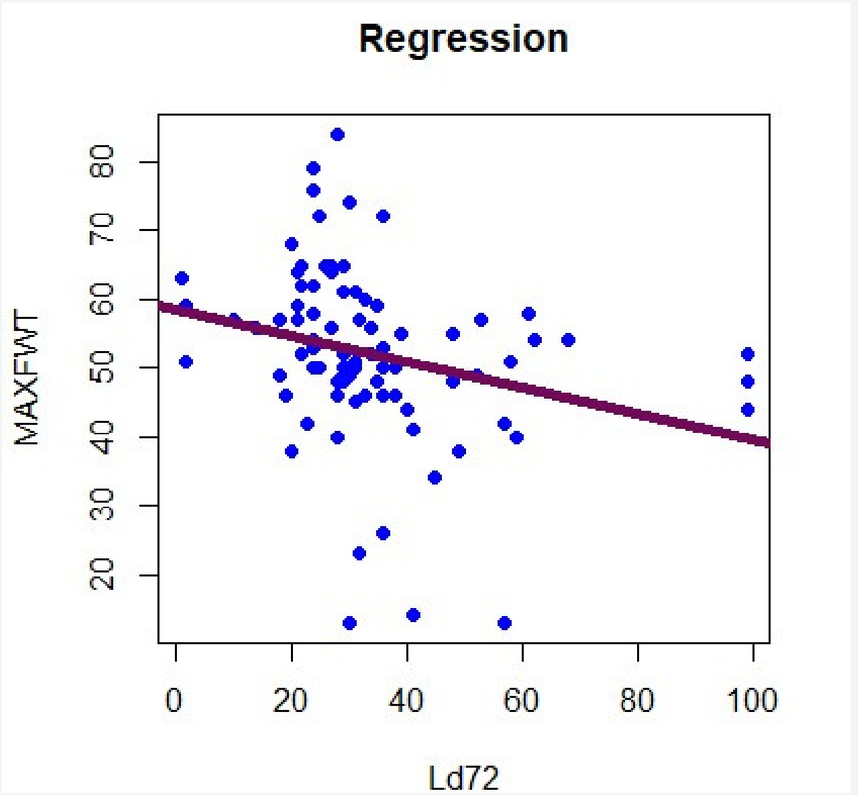
```
$`Lead_type:Sex`
                                             diff        lwr       upr     p adj
Lead_type2:Male-Lead_type1:Male      -13.8087318 -24.027962 -3.589502 0.0036208
Lead_type1:Female-Lead_type1:Male      0.3963964  -7.626163  8.418956 0.9992144
Lead_type2:Female-Lead_type1:Male     -2.6036036 -16.553258 11.346051 0.9611403
Lead_type1:Female-Lead_type2:Male     14.2051282   3.505169 24.905088 0.0044057
Lead_type2:Female-Lead_type2:Male     11.2051282  -4.438418 26.848674 0.2449552
Lead_type2:Female-Lead_type1:Female   -3.0000000 -17.305570 11.305570 0.9461804
```

Linear Model

- Simple linear regression describes the linear relationship between a predictor (explanatory) variable, plotted on the x-axis (Ld72), and a response variable, plotted on the y- axis (MAXFWT).

To fit the model, it is better to firstly draw the data to visualize the relationship between the regressors. Then using the (lm) function to build the model, where the (MAXFWT) is the response and the (Ld72) is the explanatory.

```r
#Graph for the data
plot(myData$Ld72,myData$MAXFWT,col="blue",main="Regression",ylab = "MAXFWT",xlab = "Ld72")
#Do the regression
myData.regression <- lm(MAXFWT~Ld72 , data=myData)
#Look at the R^2, F-value and p-value
abline(myData.regression,col="#73075a")
```



Interpreting the model

```
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
```

p-value: 0.01524<0.05

- The p-value is significant, which means that I have enough evidence to reject the null hypothesis "The null hypothesis asserts that the slope = 0", which means that the explanatory can’t predict the response.
- The slope (= -0.1887 which is not equal to zero) & the p-value is significant, SO, the model can use the explanatory (Ld72) to predict the response(MAXFWT)

Min, Max, Median

- The Median = 0.160 which is NOT good as it is NOT closer to the zero.
- Min & Max values are NOT mirrored values as they are NOT closer to each other as the (Min=-39.821 & Max=30.802).

Multiple R-squared

Multiple R-squared: 0.07053

- It shows that the variations in Y (MAXWHT) can be explained only by 7.053%, which indicates that this model is NOT good enough to predict the response.

Estimates

- It show that the intercept = 58.4812 measured by the unit of response (MAXWHT), and this means that the predicted value of Y = 58.4812 when X = zero.

The slope = -0.1887, which means that the rate of change in Y (MAXWHT) = -0.1887 as x (Ld72) changes by one unit.

The unit that measures the slope is (MAXFWT pre one unit of Ld72).

Bonus

The Confidence Interval

95% C.I. = β1 ± 2 x SE

95% C.I. = -0.1887 ± 2 x 0.0761

```
model <- lm(MAXFWT~Ld72 , data=myData)
summary(model)
confint(model, 'Ld72', level=0.95)
```

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

Bonus

The Prediction

```
> Estimate <- predict (model, newdata = data.frame(Ld73=100))
> Estimate
       1
24.98319
```

It means that when the Ld73=100 mg/100ml, the MAXFWT=24.98319


Thank
you!