UNIVERSITY OF WOLVERHAMPTON

SCHOOL OF MATHEMATICS AND COMPUTER SCIENCE

Statistics For AI and Data Science

ABIDEMI ALEEM
2128712

TASK 1

Statistical Summary for Patients “Survival from Malignant Melanoma”

1.1 Introduction:

This data set [2] contained malignant melanoma patients’ measurements. It records the complete removal of the tumor (205 Patients) in Denmark at the University Hospital of Odense, from 1962 to 1977 with seven variables. In this report, we present an exploratory data analysis conducted using R [1]

1.2 Data Summary:

The data set contained seven variables, and each of the variables is explained and summarised below.

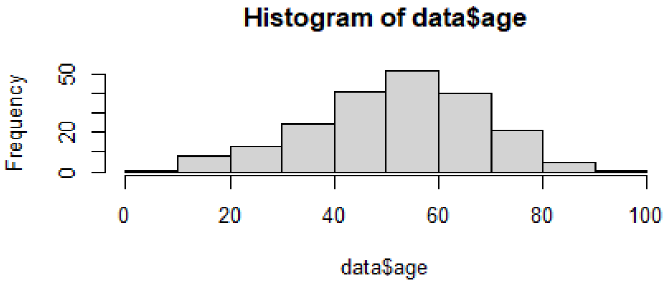
- **Time**: Records the Survival time in days since the operation, having 5565 days as the highest value and 10 days minimum survival time after the operation.
- **Status**: This shows the patient condition at the end of the study. 57 patients died from Malignant Melanoma, 134 patients survived while 14 death recorded from an unrelated cause.
- **Sex**: The data set consists of 126 females and 79 males.
- **Age**: The patient’s age at the time of the operation. The minimum and maximum ages are 4 years and 95 years respectively, these are two extreme numbers that require further investigation.
- **Year**: The period covered by the operation was 1962–1977. There were no operations conducted in the years 1963, 1964, 1975, or 1976.
- **Thickness**: The minimum and maximum values are 0.10 and 17.42 respectively, the mean is 2.92.
- **Ulcer**: 115 patients do not have ulcer while 90 patients suffer from ulcer

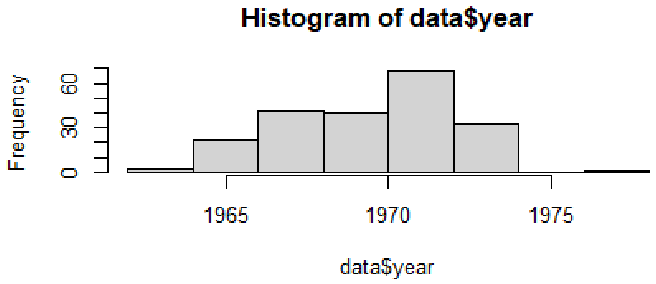
1.3 Numerical Summary:

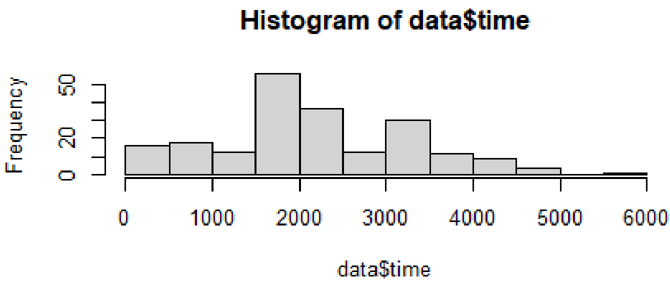
```
> summary(data)
      time                   status          sex            age             year        thickness          ulcer
 Min.   :  10   died           : 57   female:126   Min.   : 4.00   Min.   :1962   Min.   : 0.10   absent :115
 1st Qu.:1525   survived       :134   male  : 79   1st Qu.:42.00   1st Qu.:1968   1st Qu.: 0.97   present: 90
 Median :2005   unrelated death: 14                Median :54.00   Median :1970   Median : 1.94
 Mean   :2153                                      Mean   :52.46   Mean   :1970   Mean   : 2.92
 3rd Qu.:3042                                      3rd Qu.:65.00   3rd Qu.:1972   3rd Qu.: 3.56
 Max.   :5565                                      Max.   :95.00   Max.   :1977   Max.   :17.42
```

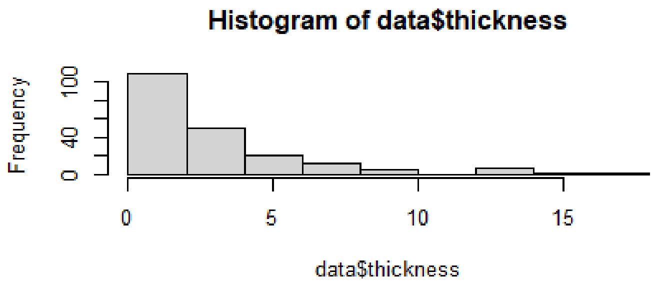
TASK 2

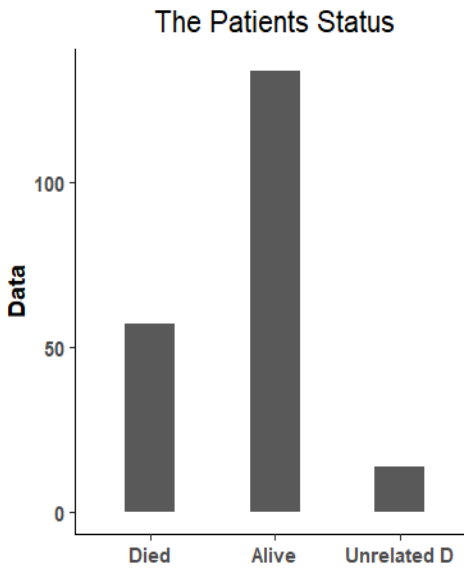
2.1 Visualization of our Data Set "Survival from Malignant Melanoma" using different graphic illustrations

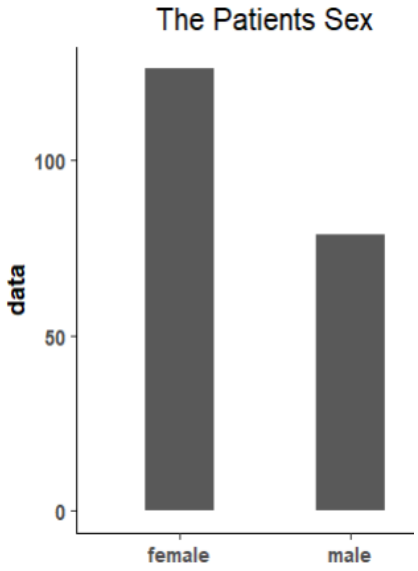


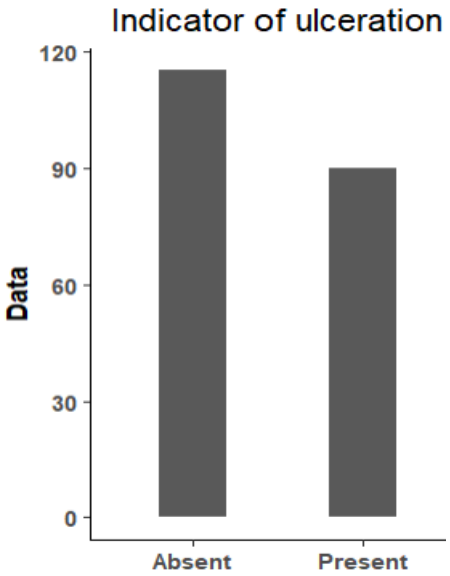


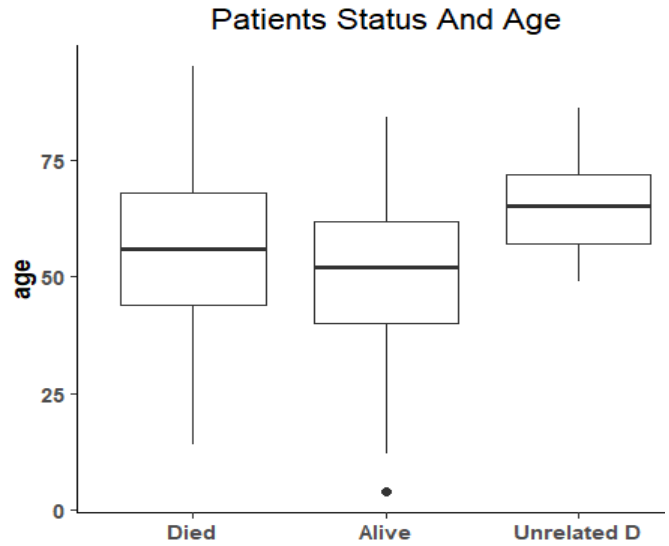


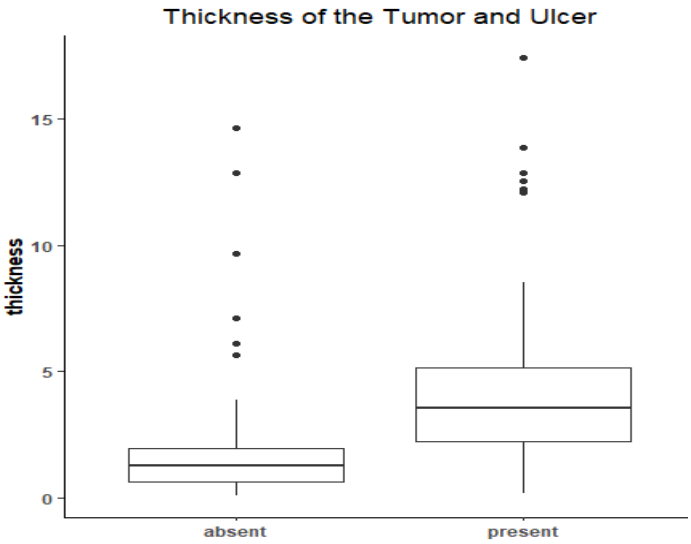


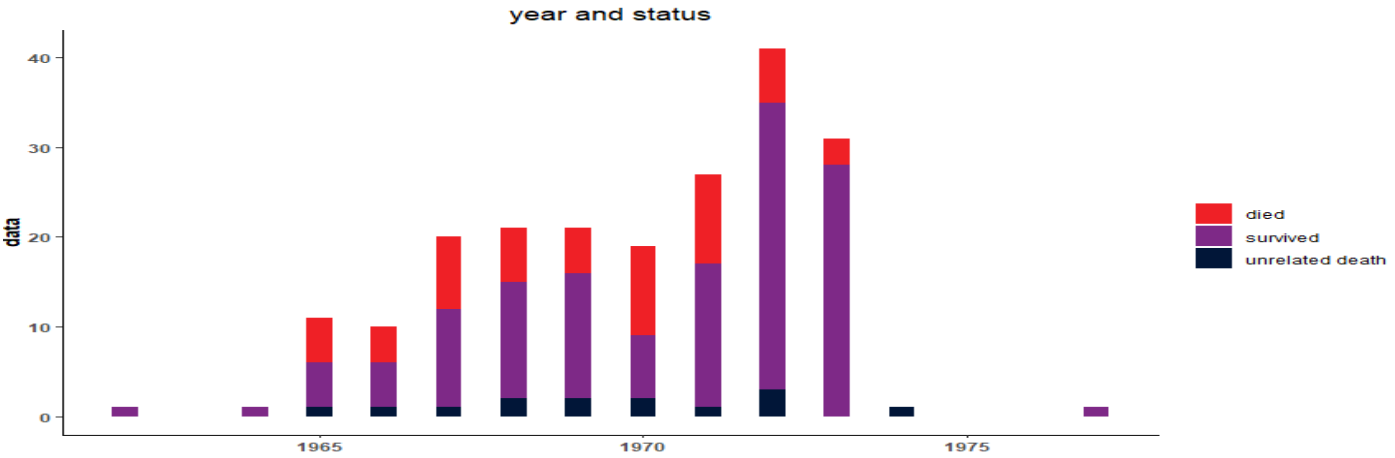












2.2 Commentary:

- The histogram of age shows a symmetric distribution, histogram of thickness is skewed to the right while year is skewed to the left.
- In the histogram for year we observe that most of the surgery performed was in the year 1972 (41patients) followed by 1973 (31 patients) we are mindful of the drastic reduction in the following year (1974 and 1977 had 1 patient each) and the years no operation was performed as this may have undue influence on our analysis.
- The boxplot reveals an interesting relationship between patient age and their status. The mean value of Patient cause-death and unrelated-death (55 and 65 respectively) is higher than the mean

value of survived patients (50). This should be considered when determining factors responsible for patients' status.

- The thickness of the tumour whether ulcerated as seen in the boxplot above shows a statistically significant value when compared with the histogram (indicator of ulceration), patients with ulcer record 4.34 mean value and 1.81 mean value for patients without ulcer.

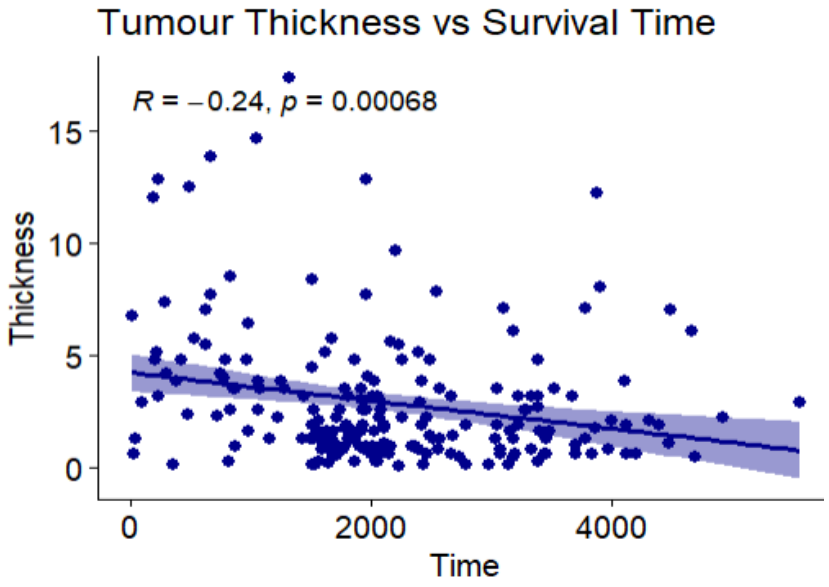
TASK 3

Regression Analysis of The Data Set "Survival from Malignant Melanoma"

H0: r = 0 (No correlation)

H1: r≠ 0 (There is correlation)

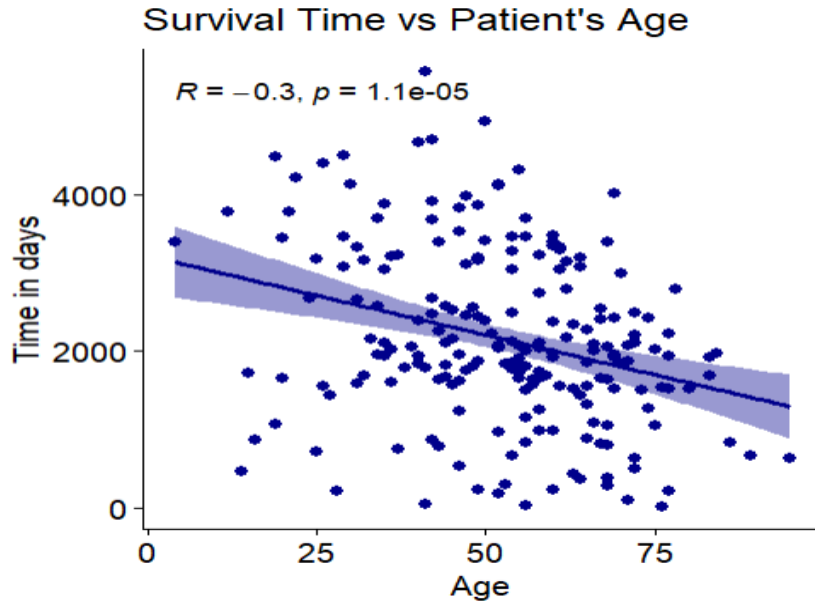
time~thickness




```
> summary(my_model)

Call:
lm(formula = melanoma.2$time ~ melanoma.2$thickness)

Residuals:
    Min      1Q  Median      3Q     Max
-2325.4  -707.6  -210.6   744.9  3410.4

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          2413.41     107.39  22.473  < 2e-16 ***
melanoma.2$thickness   -89.25      25.86  -3.451 0.000679 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1093 on 203 degrees of freedom
Multiple R-squared:  0.05542,   Adjusted R-squared:  0.05076
F-statistic: 11.91 on 1 and 203 DF,  p-value: 0.0006793
```

The p-value is 0.00068 which is < 0.05 level of significance, we reject H0: No correlation, accept H1 and conclude that there is a (weak negative) correlation between variables time and thickness.

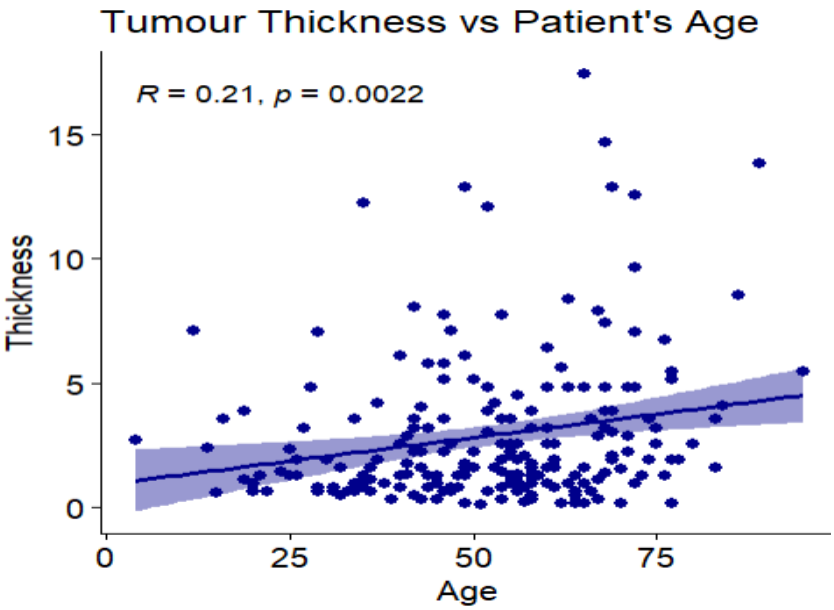
time~age




```
> summary(my_model)

Call:
lm(formula = melanoma.2$time ~ melanoma.2$age)

Residuals:
    Min      1Q  Median      3Q     Max
-2464.3  -646.2   -54.4   712.1  3179.6

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3217.448    247.879  12.980  < 2e-16 ***
melanoma.2$age  -20.293      4.504  -4.506 1.12e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1072 on 203 degrees of freedom
Multiple R-squared:  0.09091,   Adjusted R-squared:  0.08643
F-statistic:  20.3 on 1 and 203 DF,  p-value: 1.116e-05
```

At α=0.05 level of significance, our p-value is 1.116e-05 smaller we reject H0: No correlation, accept H1 and conclude that there is a (weak negative) correlation between variables time and age.

thickness~age




```
Call:
lm(formula = melanoma.2$thickness ~ melanoma.2$age)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6853 -1.7727 -0.9155  0.9558 14.0273

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.94105    0.67004   1.404  0.16170
melanoma.2$age 0.03772    0.01217   3.098  0.00222 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.899 on 203 degrees of freedom
Multiple R-squared:  0.04515,   Adjusted R-squared:  0.04044
F-statistic: 9.598 on 1 and 203 DF,  p-value: 0.002223
```

The p-value is 0.002223 which is < 0.05 level of significance we reject H0: No correlation, accept H1 and conclude that there is a (weak positive) correlation between variables thickness and age.

TASK 4

Commentary On The Observed Relationships between Variables Stated Above

- **Time versus Thickness**

Our correlation coefficient r = -0.235, shows a weak negative correlation as it slopes downward. This implies that an increase in the value of time will lead to a decrease in the value of thickness and vice-versa. The intercept = 2413.41 and the slope which is the coefficient for thickness is -89.25. The multiple R-Square which is the square of Pearson R =0.05542 (known as the coefficient of determination), shows that 5.542% of the variation in time is explained by the variation in thickness which is relatively low. The Residual standard error = 1093.

- **Time versus Age**

The value of our correlation coefficient r = -0.3015, shows a weak negative correlation as it slopes downward. This implies that an increase in the value of time will lead to a decrease in the value of age and vice-versa. The intercept = 3217.45 and the slope which is the coefficient of age is -20.29. The multiple R-Square which is the square of Pearson R =0.09091 (known as the coefficient of determination), shows that 9.1% of the variation in time is explained by the variation in thickness which is relatively low. The Residual standard error =1072.

- **Thickness versus Age**

The value of our correlation coefficient is r = 0.212, which shows a weak positive correlation as it slopes upward. This implies that an increase in the value of time will lead to an increase in the value of thickness. The intercept = 0.941 and the coefficient for age is 0.0377, which is the slope. The multiple R-Square which is the square of Pearson R =0.04515 (known as the coefficient of determination), shows that 4.51% of the variation in time is explained by the variation in thickness which is relatively low. The Residual standard error =2.899 represent the variability of the prediction error.

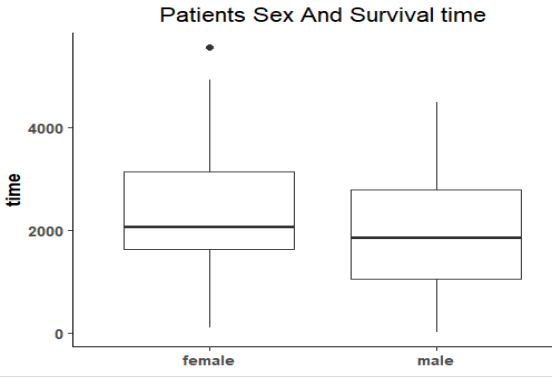
TASK 5

1. Two Sample Test For The Variable 'Time' Grouped by Female and Male.

H0: The mean_time value of female and male are the same.

H1: The mean_time value of female is different from male.

```
        Welch Two Sample t-test

data:  time by sex
t = 2.0848, df = 159.27, p-value = 0.03868
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
  17.74767 656.12032
sample estimates:
mean in group Female   mean in group Male
            2282.643            1945.709
```



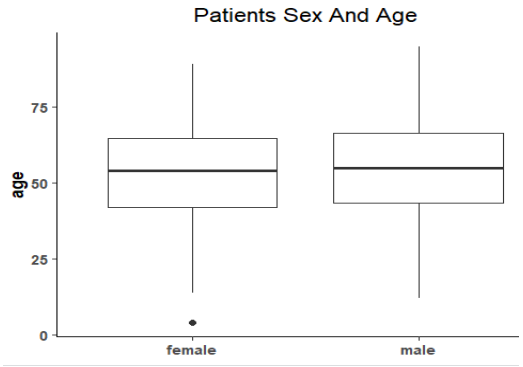
The level of significance is α = 0.05 and we have a smaller p-value of 0.038. Therefore, we reject H0, accept H1 and conclude that the true mean time differs depending on whether grouped by Female or Male as evident in the boxplot above.

2. Two Sample Test For The Variable 'Age' Grouped by Female and Male.

H0: The mean age of female and male are the same.

H1: The mean Age of female is different from the mean age of male.

```
        Welch Two Sample t-test

data:  age by sex
t = -0.95559, df = 154.42, p-value = 0.3408
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
 -7.162764  2.492280
sample estimates:
mean in group Female   mean in group Male
            51.56349            53.89873
```



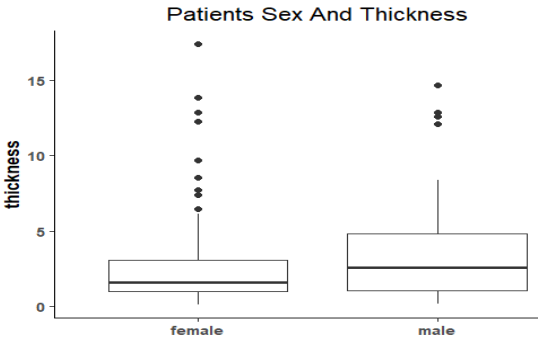
The p-value is 0.34 which is > 0.05(level of significance), therefore we accept H0 that the mean age value of female and male are the same this is also represented in the boxplot above, we reject H1.

3. Two Sample Test For The Variable 'Thickness' Grouped by Female and Male.

H0: The mean Thickness of female and male are the same.

H1: The mean Thickness of female is different from male.

```
        Welch Two Sample t-test

data:  thickness by sex
t = -2.6059, df = 149.09, p-value = 0.01009
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
 -1.9775560 -0.2718653
sample estimates:
mean in group Female   mean in group Male
            2.486429            3.611139
```

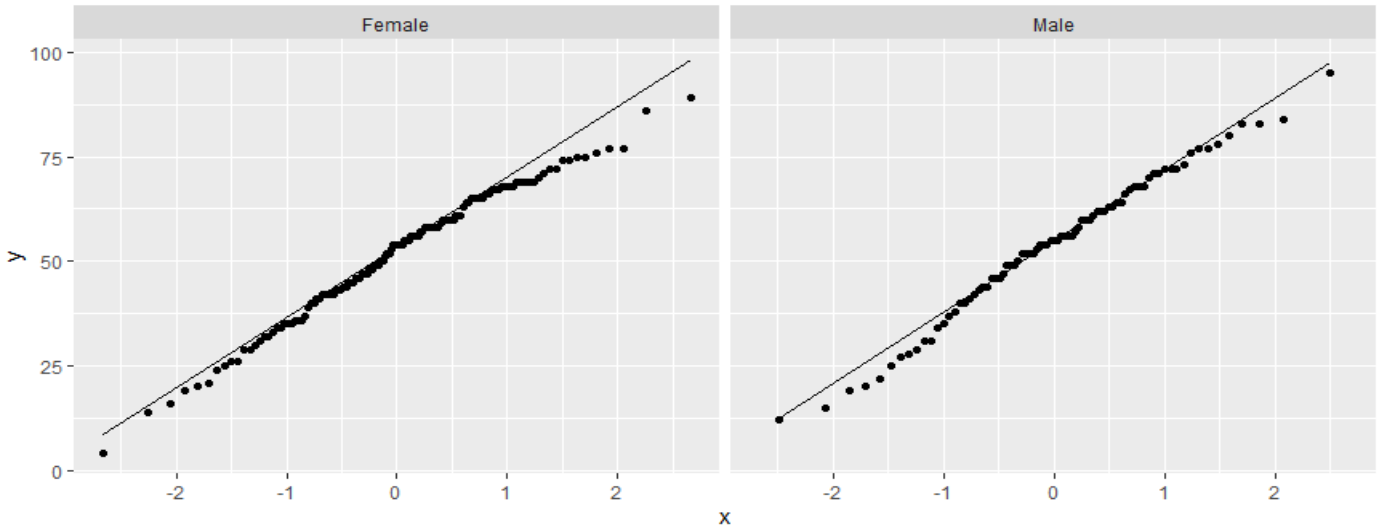


The level of significance is α = 0.05 and we have a p-value 0.010 which is smaller. Therefore, we reject H0, accept H1 and conclude that the true mean Thickness of female different from mean thickness of Male as evident above.

TASK 6

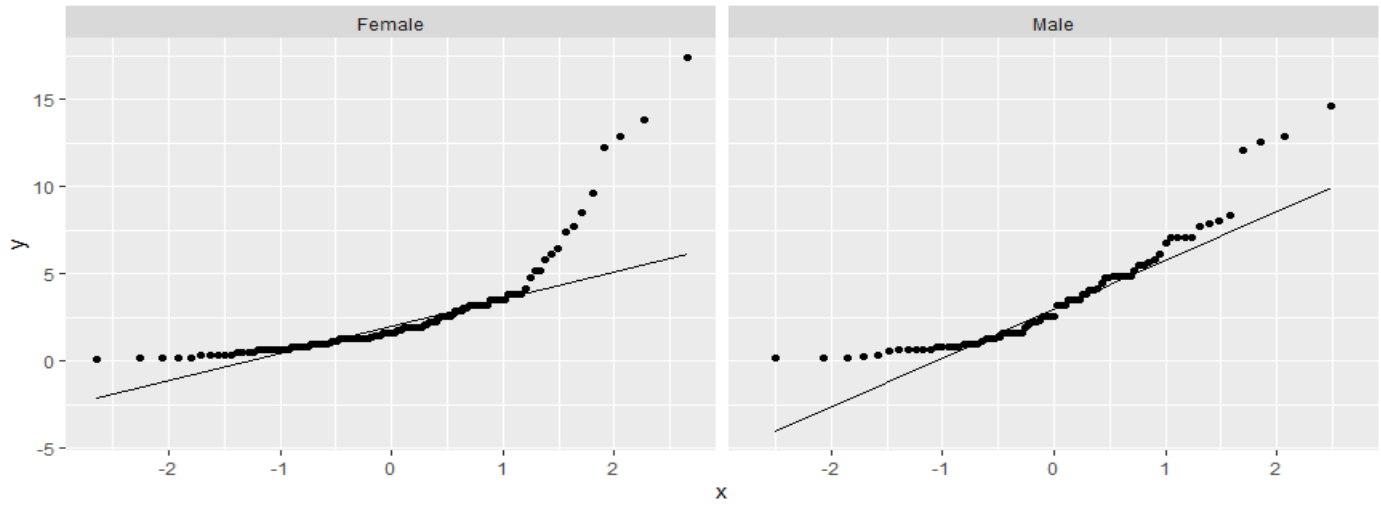
Tests for Normality Using the QQ-plot (quantile-quantile plot).

Sex and age



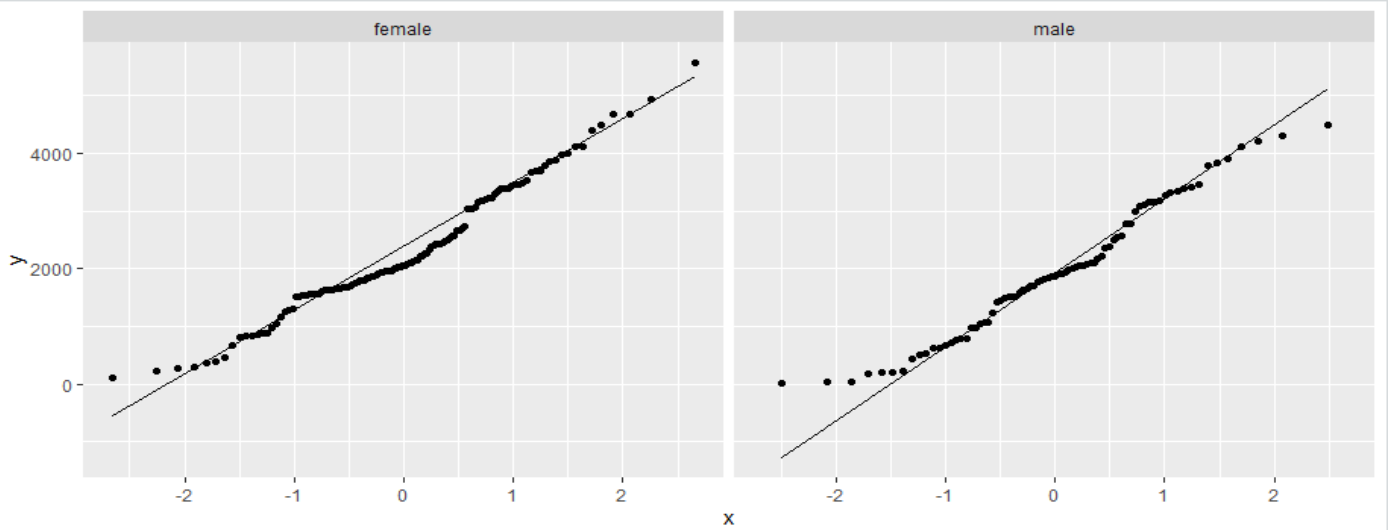
Grouped by gender, the variable Age lies pretty close to the straight line. This is a good indication that our data is normally distributed.

Sex and thickness



This variable is not normally distributed as the data lies far from the line towards the end.

Sex and Time



Grouped by gender, the variable Time lies pretty close to the straight line. This shows that our data is normally distributed.

TASK 7

Findings and Recommendations

- It was observed that only 1 patient (female, 41 years) had 5565 survival time while 1 patient (male 76 years) survived 10 days after the operation due to an unrelated cause this is relevant for further study.
- The data sets show more females had the surgery than males (with 61.5% of females against 38.5% of males) there is need for further investigation to prove whether Malignant Melanoma tumor is common in females than males.
- It is important to investigate factors that might influence patients' status and check what correlation exists (if any) between the variable age and status to determine its impact on patients' status.
- It was observed that children (4,12,14 and 15 years,) are included in datasets that may have an undue influence on our analysis
- It is important to investigate the mean value of patients that died from melanoma and those that died from causes unrelated to melanoma using our time variable. To carry out two sample significance tests for the variable "Time" when grouped by died and unrelated death and determine whether they are the same or different.
- Furthermore, the QQ-plot above (Sex ~ Thickness) indicates a slight caution that the variable thickness might not be normally distributed, therefore there is a need for further investigation.

References

[1] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

URL https://www.R-project.org/

[2] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. Springer-Verlag "Survival from Malignant Melanoma".

URL R: Survival from Malignant Melanoma (ethz.ch)