Exploratory Data Analysis Using R

“Survival from malignant melanoma” Dataset

Ekemini Michael Useh

Jan. 18, 2024

Table of Contents

1 Introduction

This report presents the exploratory data analysis of the “Survival from malignant melanoma” data set using R (R Core Team. 2020). The data is provided for the Statistics for Data Science & AI coursework and available on Canvas using this link “https://canvas.wlv.ac.uk/courses/41912/assignments/257877”. It is the measurements recorded on patients with malignant melanoma whose tumors were removed at the Department of Plastic Surgery, University Hospital of Odense, Denmark. This procedure was done through surgery and the records were kept from 1962 to 1977.

A total of 7 variables were recorded on 205 patients, according to the data provided.

These variables are:

Time – Survival time (days) after the operation, **Status** – The patients' status when the study ended (1 = died from malignant melanoma, 2 = still alive, 3 = died from other causes not related to malignant melanoma), **Sex** – The patients’ sex (1 = male, 0 = female), **Age** – Age (years) when the operation was done, **Year** – Year of operation, **Thickness** – Tumor thickness (mm), **Ulcer** – To show if ulcer was present or absent (1 = present, 0 = absent).

The data set was imported into R by loading the tidyverse package by (Wickham et al. 2019), ggplot2 package by (Wickham 2016), and ggthemes by (Arnold 2021). A data frame called melanoma_df to import the melanoma.csv file was created. For more information on how to use R programming, readers are directed to ‘R for Data Science (2e).’ (Wickham et al.) and ‘Basic R Guide for NSC statistics.’ (Deanna 2020).

Importing the data into R is implemented using the code below:

```
> melanoma_df <- read_csv("C:/Users/HP/Documents/statistics_assignment/melano
ma.csv")
```

2 Data cleaning

- The data was viewed to see what it contains by simply writing:

```
> melanoma_df
```

According to the data collected, 205 patients were operated on, and the number of variables is 8.

- The unimportant column was removed.

The column on the extreme left of the data is just the serial number and is unimportant in this analysis. Therefore, the column was removed using:

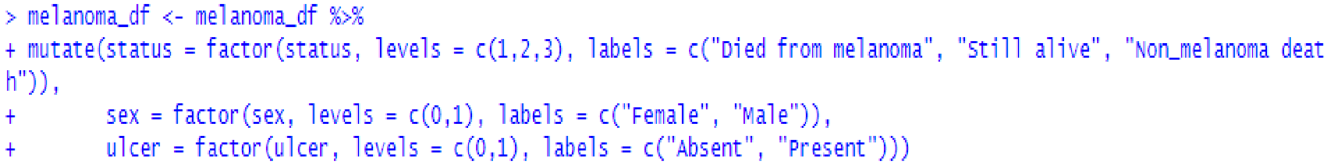
```
> melanoma_df <- melanoma_df %>% select(-1)
```

The data was viewed to check if the command was successfully implemented.

```
> head(melanoma_df)
```

- The data types of variables status, sex, and ulcer were changed to factor.

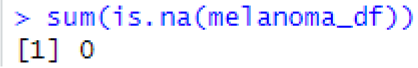
These variables are categorical/ nominal variables. Therefore, the data types were changed from double to factor. Since the data inputs in these columns were represented in numbers, the labels were changed to the actual inputs for easy data exploration.



```
> melanoma_df <- melanoma_df %>%
+ mutate(status = factor(status, levels = c(1,2,3), labels = c("Died from melanoma", "Still alive", "Non_melanoma deat
h")),
+        sex = factor(sex, levels = c(0,1), labels = c("Female", "Male")),
+        ulcer = factor(ulcer, levels = c(0,1), labels = c("Absent", "Present")))
```

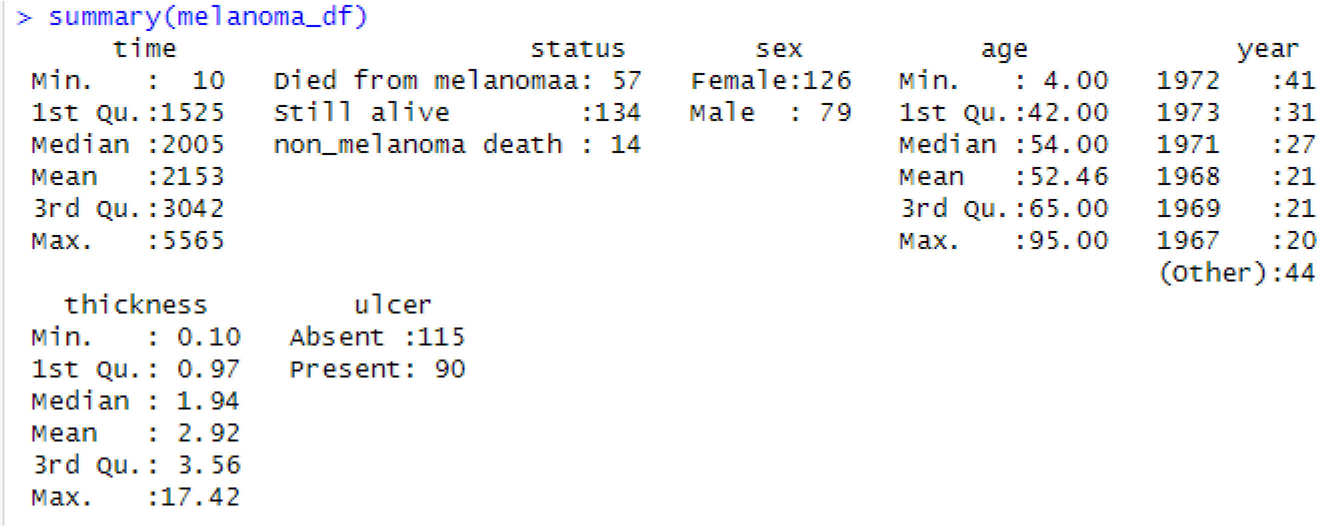
The data was viewed to see if this was implemented successfully.

- The data was checked for null values and there were none.



```
> sum(is.na(melanoma_df))
[1] 0
```

3 Summary statistics



```
> summary(melanoma_df)
      time                        status          sex            age              year    
 Min.   :  10   Died from melanomaa: 57   Female:126   Min.   : 4.00   1972   :41  
 1st Qu.:1525   Still alive        :134   Male  : 79   1st Qu.:42.00   1973   :31  
 Median :2005   non_melanoma death : 14                Median :54.00   1971   :27  
 Mean   :2153                                          Mean   :52.46   1968   :21  
 3rd Qu.:3042                                          3rd Qu.:65.00   1969   :21  
 Max.   :5565                                          Max.   :95.00   1967   :20  
                                                                       (Other):44  
   thickness        ulcer    
 Min.   : 0.10   Absent :115  
 1st Qu.: 0.97   Present: 90  
 Median : 1.94                
 Mean   : 2.92                
 3rd Qu.: 3.56                
 Max.   :17.42                
```

- Time

From the summary statistics for time,

Minimum days survived since operation = 10 days. Maximum days survived since operation = 5565 days. 1st Quartile = 1525 days. 3rd Quartile = 3042 days. Median = 2005 days. Mean (Average days survived) = 2152.8 = 2153 days

- Status

The status variable is categorical and has a factor type. Due to this, the summary is the total observations in each category.

The total number of deaths from melanoma = 57. The total still alive = 134. Total number of deaths from causes not related to their melanoma = 14.

- Sex

This is a categorical variable with ‘factor’ as the data type. Therefore, the summary statistics is the total for each category. Female = 126, Male = 79

- Age

From the summary statistics for age,
Minimum age = 4 years. Maximum age = 95 years. 1st Quartile = 42 years. 3rd Quartile = 65 years. The median of the patient’s age (Middle age) = 54 years. Average (Mean) = 52 years.

- Year

The year variable was changed to a categorical variable using “factor” as the data type to arrive at this.

- Thickness

The smallest tumor thickness (Minimum thickness) = 0.10 mm. 1st Quartile = 0.97 mm. 3rd Quartile = 3.56 mm. Median (Middle thickness) = 1.94 mm. Average tumor thickness (Mean) = 2.92 mm. The presence of outliers makes the maximum thickness 17.42 mm. Using a boxplot, the maximum is seen to be approximately 7.50 mm. The variance of tumor thickness = 8.758242. The standard deviation of thickness = 2.959433

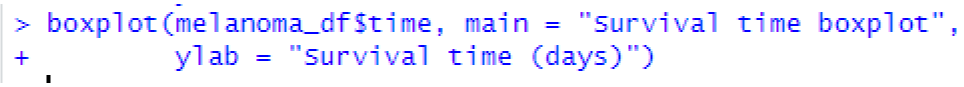
- Ulcer

Ulcer is a categorical variable with ‘factor’ as the data type. Therefore, each category's total number of observations is computed as summary statistics.
Tumor ulcerations absent = 115, Tumor ulcerations present = 90

4 Graphical summaries for each of the variables

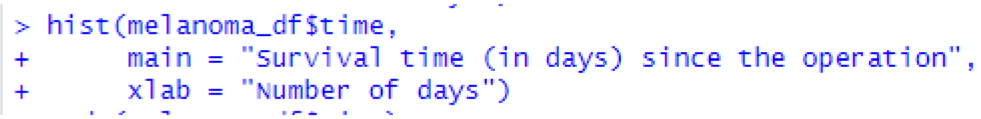
- Time

A boxplot of the time variable (Fig 5.).

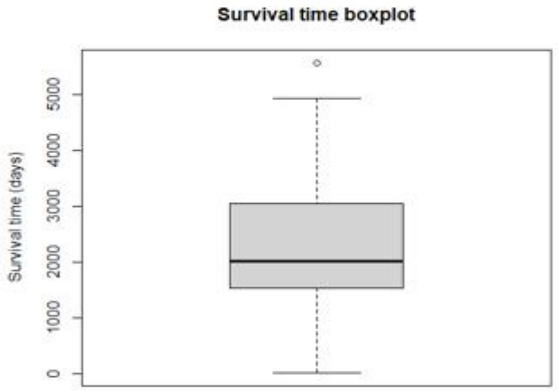


```
> boxplot(melanoma_df$time, main = "Survival time boxplot",
+         ylab = "Survival time (days)")
```

The histogram of the time variable (Fig 4.).



```
> hist(melanoma_df$time,
+      main = "Survival time (in days) since the operation",
+      xlab = "Number of days")
```



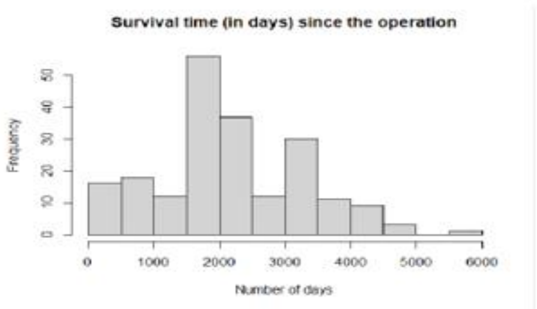


Fig 5. Survival time boxplot

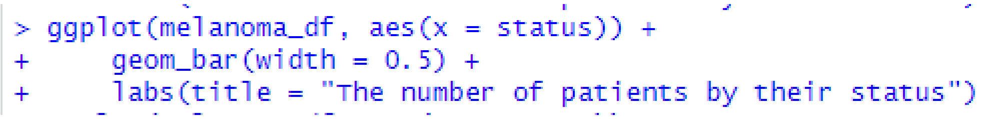
Fig 4. Histogram of time variable

From the box plot (Fig 5), there is the presence of an outlier in the time variable. This is equally observed in the histogram (Fig 4). This outlier is likely the maximum value of time previously observed to be 5565 days in the summary statistics of time.

The histogram (Fig 4) is a multimodal distribution with modes at approximately 750 days, 1750 days, and 3250 days.

- <u>Status</u>

Fig 3 is a bar graph showing the number of patients according to their status.



```
> ggplot(melanoma_df, aes(x = status)) +
+     geom_bar(width = 0.5) +
+     labs(title = "The number of patients by their status")
```

The bar graph (Fig 3) shows the number of patients by status. From the graph, those who were still alive were the highest in number (134), deaths from their melanoma (57), while deaths from other things unrelated to their melanoma were the least in number (14).

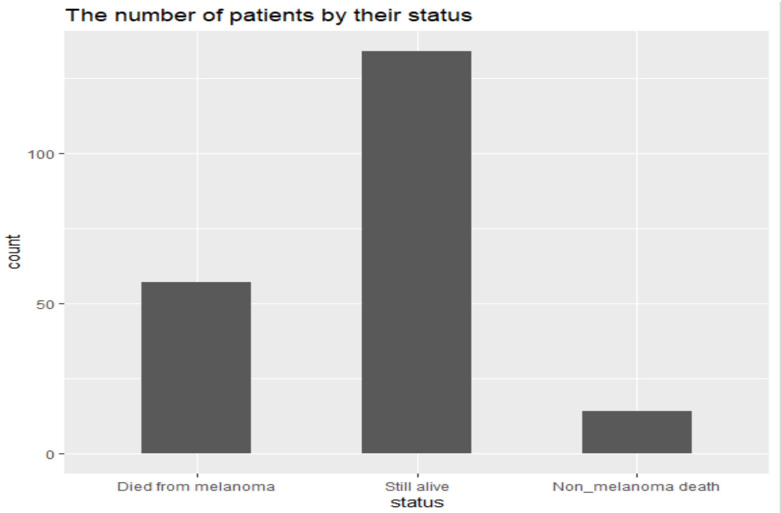
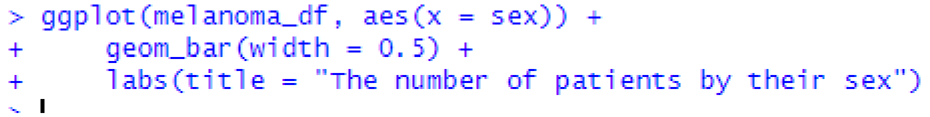


Fig 3. Number of patients by status

- Sex

A bar graph (Fig 6) showing the number of patients by sex.



```
> ggplot(melanoma_df, aes(x = sex)) +
+     geom_bar(width = 0.5) +
+     labs(title = "The number of patients by their sex")
```

From the bar graph (Fig 6), more females were operated up on (126) while the males were 79 in number.

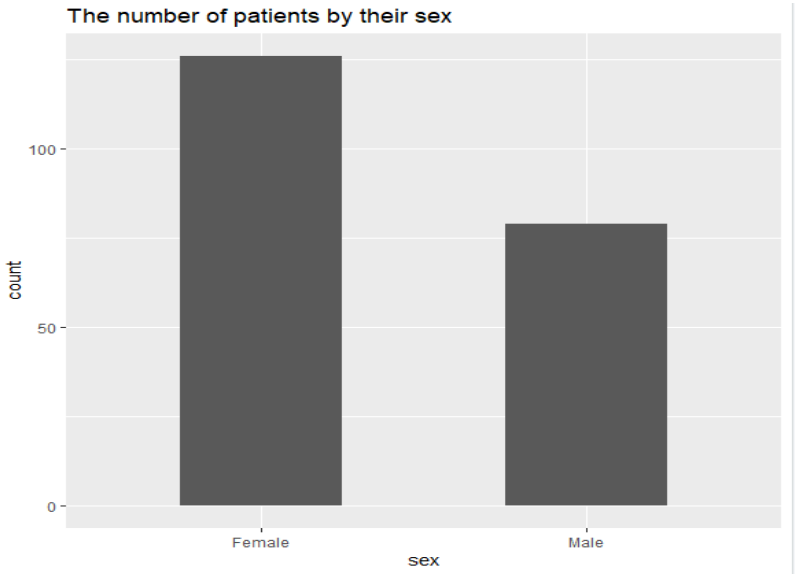
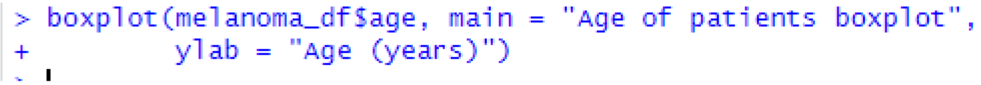


Fig 6. Number of patients by sex

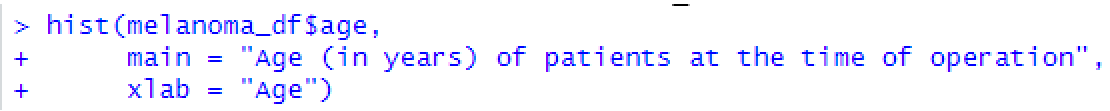
- Age

A boxplot of the age variable (Fig 7.).



```
> boxplot(melanoma_df$age, main = "Age of patients boxplot",
+         ylab = "Age (years)")
```

A histogram of age variable (Fig 8.).



```
> hist(melanoma_df$age,
+      main = "Age (in years) of patients at the time of operation",
+      xlab = "Age")
```

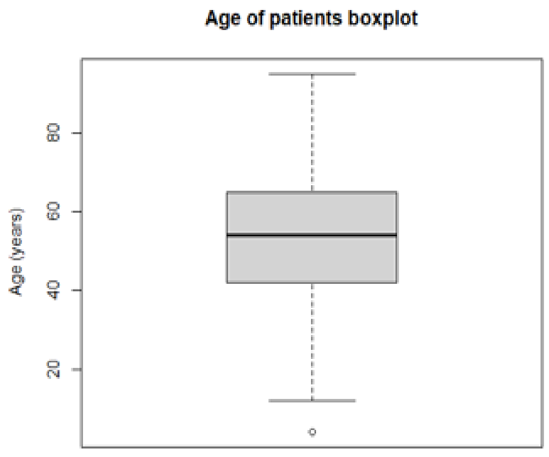


Fig 7. Boxplot of age variable

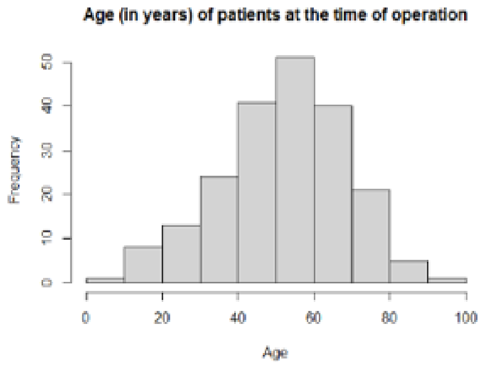


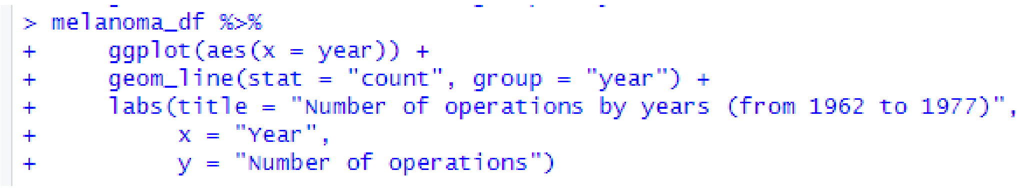
Fig 8. Histogram of age variable

From the boxplot (Fig 7), There is a presence of an outlier which is observed in the summary statistics of age as the minimum value of 4 years.

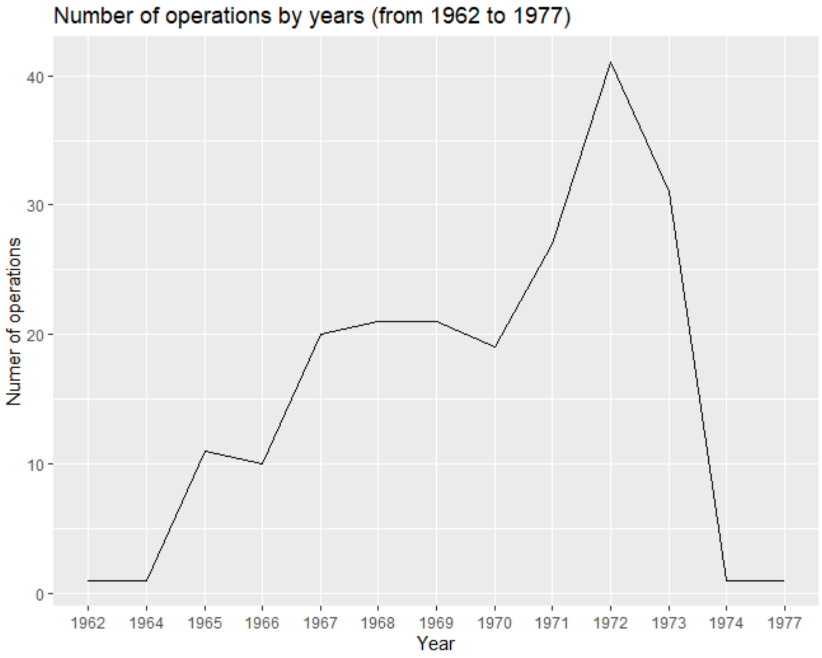
The age histogram (Fig 8.) is a uni-modal distribution. It has one peak. The most occurring age is approximately 56 years, The histogram is slightly skewed to the left.

- <u>Year</u>

A line plot of the year variable.



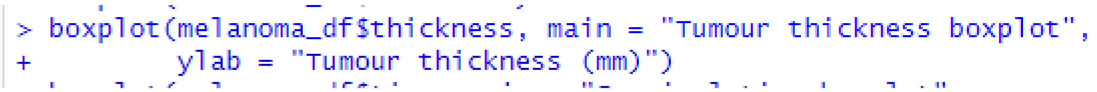
```
> melanoma_df %>%
+     ggplot(aes(x = year)) +
+     geom_line(stat = "count", group = "year") +
+     labs(title = "Number of operations by years (from 1962 to 1977)",
+          x = "Year",
+          y = "Number of operations")
```



This record indicates that the highest number of melanoma surgeries were done in the year 1972 with a total of 41 surgeries. While 1963, 1975, and 1976 had no melanoma surgery records.

- Thickness

A boxplot of the thickness variable (Fig 9.).



```
> boxplot(melanoma_df$thickness, main = "Tumour thickness boxplot",
+         ylab = "Tumour thickness (mm)")
```

Fig 10. is the thickness histogram.

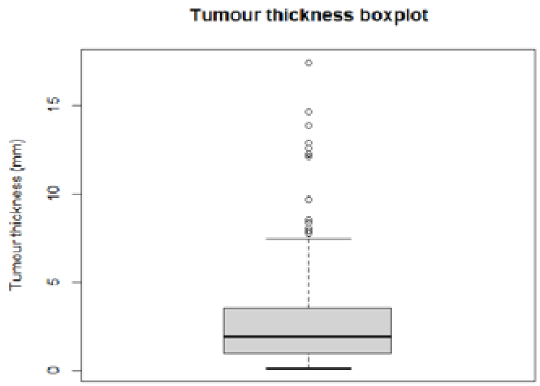


Fig 9. Boxplot of thickness variable

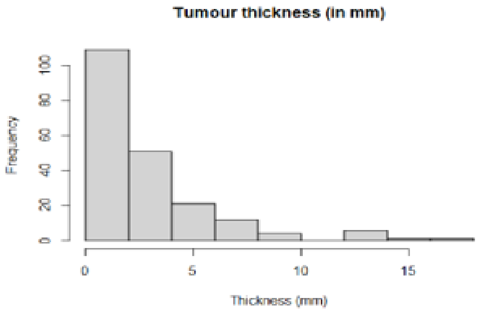
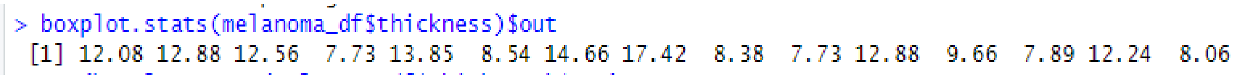


Fig 10. Histogram of tumor thickness

The boxplot (Fig 9) indicates that outliers are present in the data on tumor thickness. The outliers are shown below:

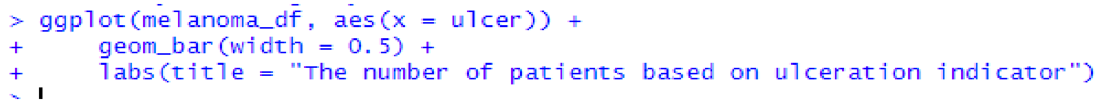


```
> boxplot.stats(melanoma_df$thickness)$out
 [1] 12.08 12.88 12.56  7.73 13.85  8.54 14.66 17.42  8.38  7.73 12.88  9.66  7.89 12.24  8.06
```

The histogram (Fig 10) is skewed to the right and is not normally distributed.

- Ulcer

A bar graph showing the number of patients in groups whose tumors had ulcerations, or not.



```
> ggplot(melanoma_df, aes(x = ulcer)) +
+     geom_bar(width = 0.5) +
+     labs(title = "The number of patients based on ulceration indicator")
```

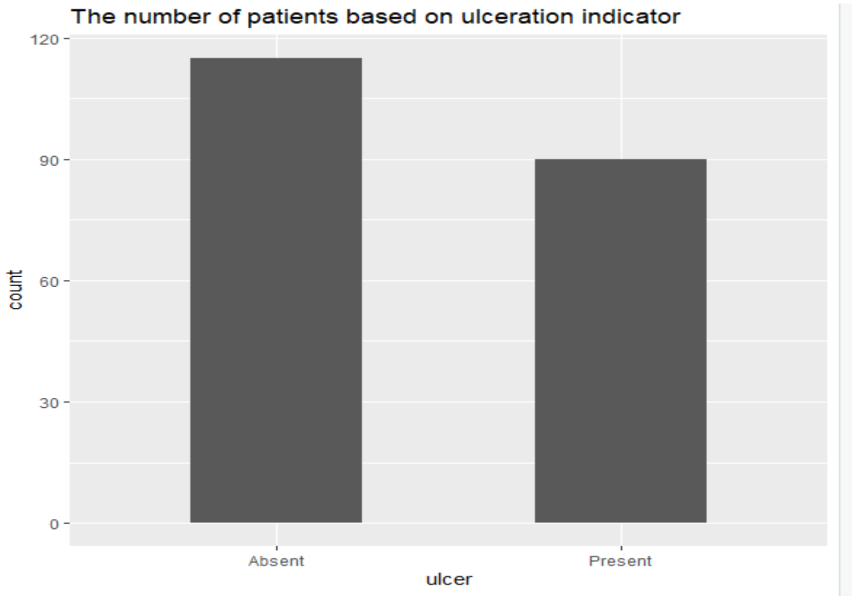


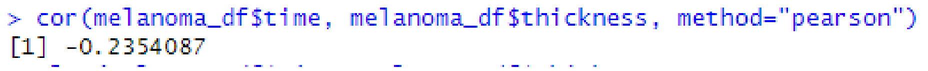
Fig 11. Bar graph of total patients by ulcer indicator

From the bar graph (Fig 11), out of the total patients operated on (205), those with ulcerations absent were more in number than those with ulcerations.

5 Regression analysis and correlation computations

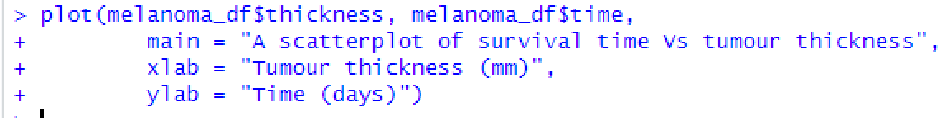
- <u>The relationship between **time ~ thickness**</u>**.**

Using Pearson's correlation, the relationship has a weak negative correlation of -0.2354087. This means as the tumor thickness gets bigger, the smaller the survival time a patient lives after the procedure.

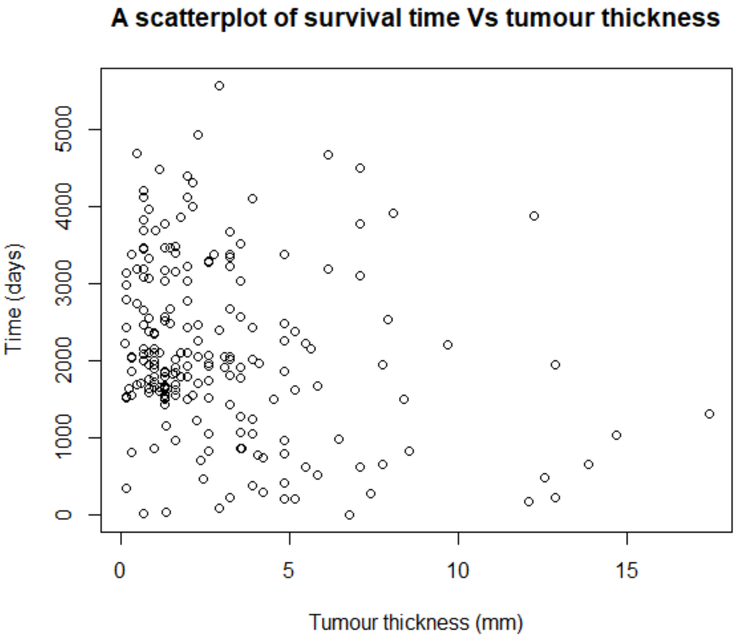


```
> cor(melanoma_df$time, melanoma_df$thickness, method="pearson")
[1] -0.2354087
```

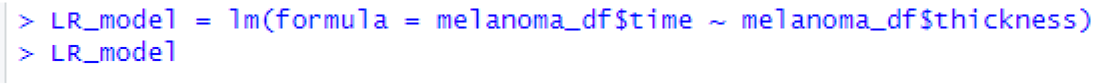
The scatter plot of this relationship is represented in Fig 12.



```
> plot(melanoma_df$thickness, melanoma_df$time,
+         main = "A scatterplot of survival time vs tumour thickness",
+         xlab = "Tumour thickness (mm)",
+         ylab = "Time (days)")
```



Regression analysis of **time ~ thickness.**



```
> LR_model = lm(formula = melanoma_df$time ~ melanoma_df$thickness)
> LR_model
```

The line of best fit is:

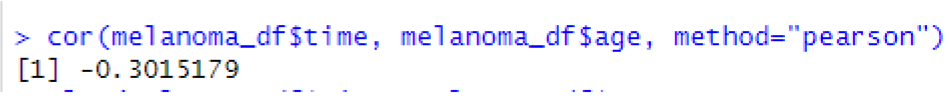
time = 2413.41 – 89.25 thickness

here, the y-intercept is 2413.41 when the thickness is equal to zero. The gradient is -89.25. For every unit increase in the value of thickness, time is predicted to decrease by 89.25, on average.

The residual standard error is 1093 and the R-squared is 0.05542. Due to the unreliability of this model, using it for any predictive analysis of the relationship should be done with caution.

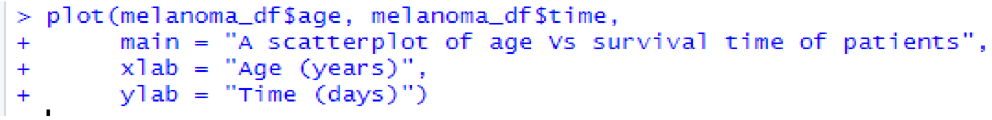
- The relationship between **time ~ age**

The relationship has a weak negative correlation of -0.3015179. This means that as age increases, a patient lives fewer days after the operation.

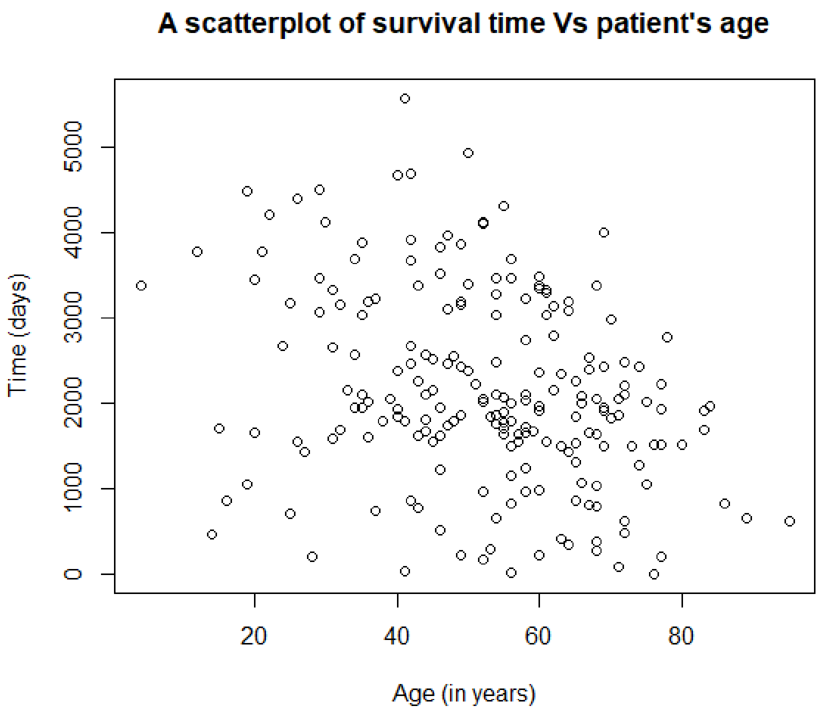


```
> cor(melanoma_df$time, melanoma_df$age, method="pearson")
[1] -0.3015179
```

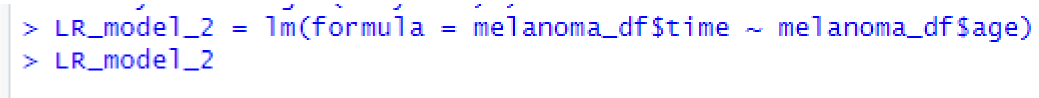
The scatter plot of this relationship is represented in Fig 13.



```
> plot(melanoma_df$age, melanoma_df$time,
+      main = "A scatterplot of age Vs survival time of patients",
+      xlab = "Age (years)",
+      ylab = "Time (days)")
```



Regression analysis of **time ~ age**



```
> LR_model_2 = lm(formula = melanoma_df$time ~ melanoma_df$age)
> LR_model_2
```

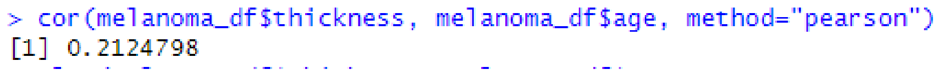
The line of best fit is:

time = 3217.45 – 20.29 age

here, the y-intercept is 3217.45 when the age is equal to zero. The gradient is -20.29. For every unit increase in the value of age, time is predicted to decrease by 20.29, on average. The residual standard error is 1072 and the R-squared is 0.09091. Due to the unreliability of this model, using it for any predictive analysis of the relationship should be done with caution.

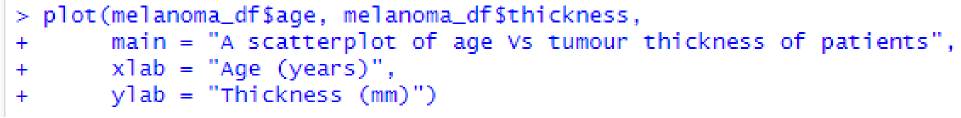
- <u>The relationship between **thickness ~ age**</u>

The relationship is a weak positive correlation of 0.2124798. This means as the age increases, the bigger the tumor thickness.

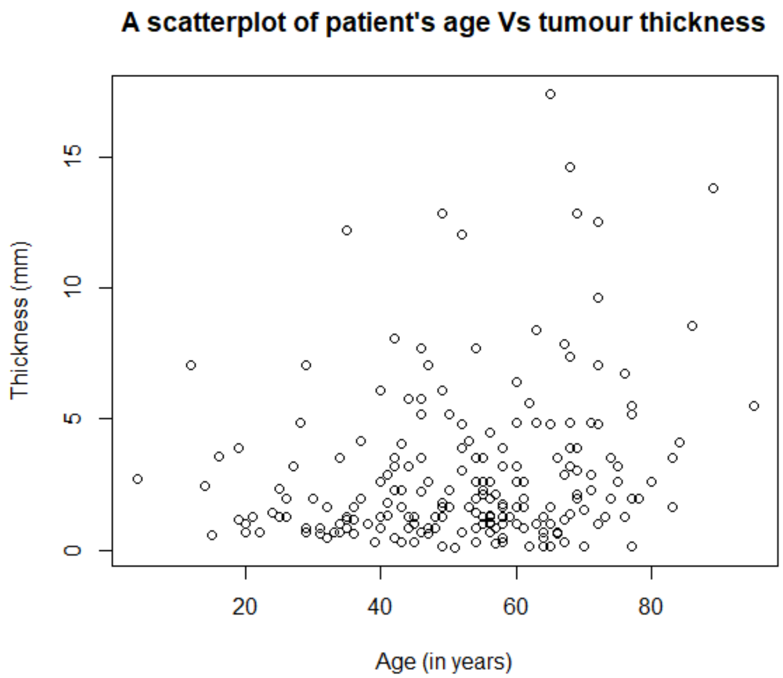


```
> cor(melanoma_df$thickness, melanoma_df$age, method="pearson")
[1] 0.2124798
```

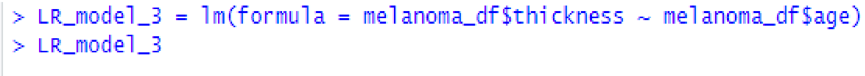
The scatter plot of this relationship is represented in Fig 14.



```
> plot(melanoma_df$age, melanoma_df$thickness,
+      main = "A scatterplot of age Vs tumour thickness of patients",
+      xlab = "Age (years)",
+      ylab = "Thickness (mm)")
```



Regression analysis of **thickness ~ age**



```
> LR_model_3 = lm(formula = melanoma_df$thickness ~ melanoma_df$age)
> LR_model_3
```

The line of best fit is:

thickness = 0.94105 + 0.03772 age

here, the y-intercept is 0.94105 when the age is equal to zero. The gradient is 0.03772. For every unit increase in the value of age value, thickness is predicted to increase by 0.03772, on average. The residual standard error is 2.899 and the R-squared is 0.04515. Due to the unreliability of this model, using it for any predictive analysis of the relationship should be done with caution.

6 Two sample significance tests (grouped by sex)

Using t-test

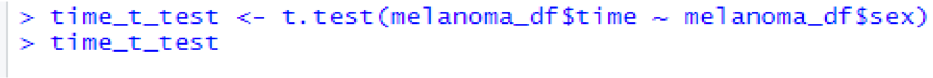
Where *H*o = Null hypothesis, *H*1 = alternative hypothesis

The default level of significance α = 0.05

- Time grouped by sex.

*H*o: The mean survival time after the operation is the same for both sexes

*H*1: The mean survival time after the operation is different in both sexes.

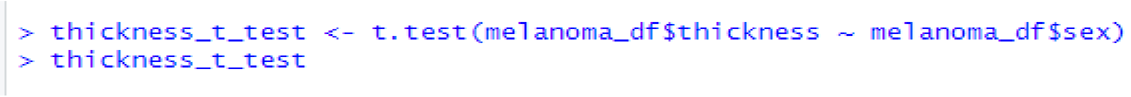


```
> time_t_test <- t.test(melanoma_df$time ~ melanoma_df$sex)
> time_t_test
```

The *p*-value = 0.03868 which is smaller than α = 0.05. Therefore, we can reject *H*o and conclude that there is evidence that the true mean survival time after the operation is different depending on whether the patient is male or female.

- Thickness grouped by sex.

*H*o: The mean thickness of the tumor is the same for both sexes
*H*1: The mean thickness of the tumor is different in both sexes.

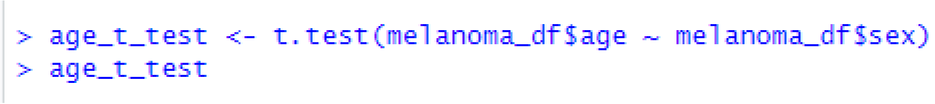


```
> thickness_t_test <- t.test(melanoma_df$thickness ~ melanoma_df$sex)
> thickness_t_test
```

The *p*-value = 0.01009 which is smaller than α = 0.05. Therefore, we can reject *H*o and conclude that there is evidence that the true mean thickness of the tumor is different depending on whether the patient is male or female.

- Age grouped by sex.

*H*o: The mean age of the patients is the same in both sexes
*H*1: The mean age of the patients is different depending in both sexes.



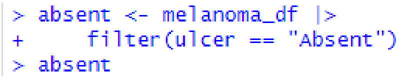
```
> age_t_test <- t.test(melanoma_df$age ~ melanoma_df$sex)
> age_t_test
```

The *p*-value = 0.3408 which is greater than α = 0.05. Therefore, we **cannot** reject *H*o. We **can** conclude that there is evidence that the true mean age of patients is the same in male and female.

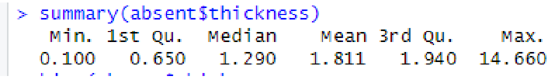
7 Observations

The findings gathered from the data show that malignant melanoma is likely to have ulcerations present depending on its tumor thickness. This implies that a larger tumor thickness is likely ulcerated. The following summary statistics show this:

Extracting the data of those with ulceration absent

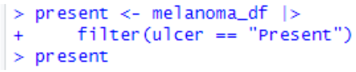


```
> absent <- melanoma_df |>
+     filter(ulcer == "Absent")
> absent
```

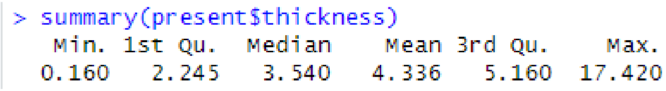


```
> summary(absent$thickness)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.100   0.650   1.290   1.811   1.940  14.660 
```

Extracting the data of those with ulceration present



```
> present <- melanoma_df |>
+     filter(ulcer == "Present")
> present
```



```
> summary(present$thickness)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.160   2.245   3.540   4.336   5.160  17.420 
```

A table to compare the summary statistics of tumors with ulcerations present and absent.

Summary statistics	Ulceration present (mm)	Ulceration absent (mm)
Minimum	0.160	0.100
1st Quartile	2.245	0.650
Median	3.540	1.290
Mean	4.336	1.811
3rd Quartile	5.160	1.940
Maximum	17.420	14.660

More people survive after getting their melanoma removed which means that fewer people are likely going to die due to their melanoma.

The risk of dying from melanoma increases when the tumor becomes ulcerated due to the increase in thickness of the tumor.

8 Recommendations

People need to seek proper medical care and advice at the sight of any skin infection as it could be melanoma. This will hinder the growth and spread of such infection. The possibility of ulceration, if detected early, will decrease and cured without the risk of fatality.

References

- R Core Team. 2020. R: A language and environment for statistics computing. R Foundation for Statistical Computing, Vienna, Austria. "https://www.R-project.org.
- Arnold, Jeffrey B. 2021. "Ggthemes: Extra Themes, Scales and Geoms for 'Ggplot2'." https://CRAN.R-project.org/package=ggthemes.
- Deanna L. 2020. "Basic R Guide for NSC Statistics." https://bookdown.org/dli/rguide/r-and-rstudio.html.
- Wickham, Hadley. 2016. "Ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.
- Wickham Hadley, Çetinkaya-Rundel Mine, Grolemund Garrett. "R for Data Science (2e)." https://r4ds.hadley.nz/.
- Wickham Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse" 4: 1686. https://doi.org/10.21105/joss.01686.