SURVIVAL FROM MALIGNANT MELANOMA

OLALEKAN OLAOLUWAKITAN BABALOLA (2374470)

7CS039 STATISTICS FOR AI & DATA SCIENCE

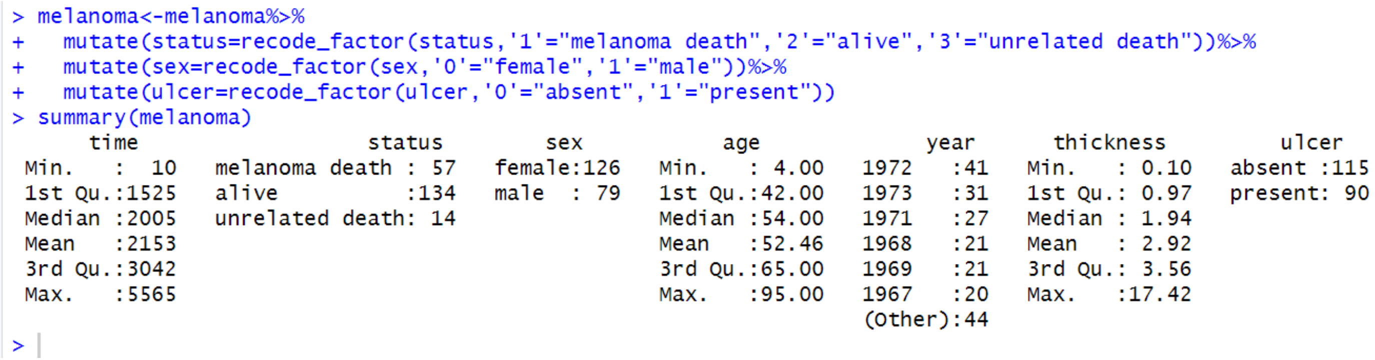
INTRODUCTION

In this report, data analysis exploration using R [1] of the Malignant Melanoma dataset adapted from [2]. The study focuses on 205 patients treated for melanoma between 1962 and 1977 at the department of Plastic Surgery, University Hospital of Odense, Denmark. The study measures a set of variables listed below:

- Time: Survival time (in days) since the operation.
- Status: The patient's status at the end of the study. 1 indicates that they had died from melanoma, 2 indicates that they were still alive and 3 indicates that they had died from causes unrelated to their melanoma.
- Sex: The patients where 1 is male, 0 is female.
- Age: Age (in years) at the time of the operation.
- Year: Year of operation.
- Thickness: Tumour thickness (in mm).
- Ulcer: Indicator of ulceration (1 is present, 0 is absent).

i. SUMMARY STATISTICS

Due to the presence of nominal categories in the data set, data set has been labeled and simplified to show precise values as shown below.



```
> melanoma<-melanoma%>%
+   mutate(status=recode_factor(status,'1'="melanoma death",'2'="alive",'3'="unrelated death"))%>%
+   mutate(sex=recode_factor(sex,'0'="female",'1'="male"))%>%
+   mutate(ulcer=recode_factor(ulcer,'0'="absent",'1'="present"))
> summary(melanoma)
      time                 status        sex           age             year      thickness         ulcer    
 Min.   :  10   melanoma death : 57   female:126   Min.   : 4.00   1972   :41   Min.   : 0.10   absent :115  
 1st Qu.:1525   alive          :134   male  : 79   1st Qu.:42.00   1973   :31   1st Qu.: 0.97   present: 90  
 Median :2005   unrelated death: 14                Median :54.00   1971   :27   Median : 1.94                
 Mean   :2153                                      Mean   :52.46   1968   :21   Mean   : 2.92                
 3rd Qu.:3042                                      3rd Qu.:65.00   1969   :21   3rd Qu.: 3.56                
 Max.   :5565                                      Max.   :95.00   1967   :20   Max.   :17.42                
                                                                   (Other):44                                

> 
```

COMMENTS

Time: The minimum patient's survival time is only 10 days which could mean that the patients were already old and therefore didn't survive long after surgery. Average survival time is 2,153 days.

Status: From a total of 205 patients, 134 were alive at the end of the study in 1977, a combined total of 35% died of melanoma (57) and unrelated causes (14).

Sex: 126 (61%) patients were female, while 79 (39%) were male. This suggests that females are more susceptible to melanoma than males.

Age: The average patient age was approximately 53 years, most common age (median) is 54 years with youngest being 4 and the oldest being 95.

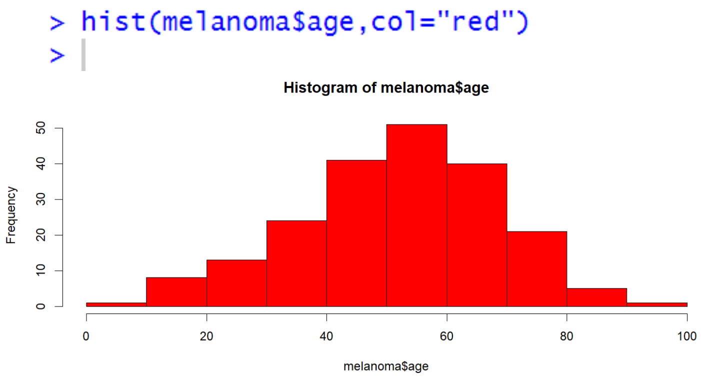
Year: Only one operation was conducted in 1962, the first year of the study. The highest number of operations (41) were conducted in 1972, exactly 10 years after the study began.

Thickness: The tumor thickness ranged from 0.1mm to 17.42mm, interquartile range is over 2.5mm, with the most common size being 1.94m.

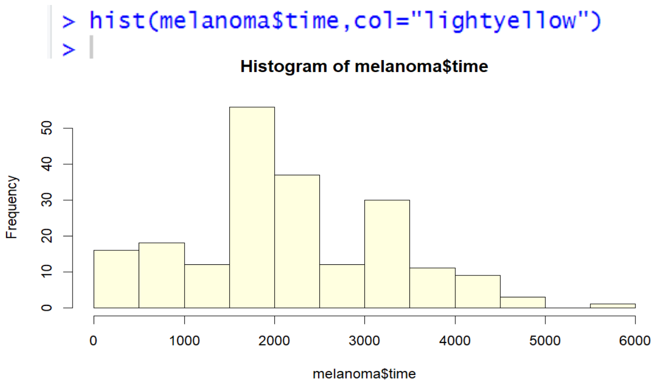
Ulcer: Ulceration was present in 56% of patients, marginally higher than patients where ulcers were absent (44%)

ii. **GRAPHICAL SUMMARY**

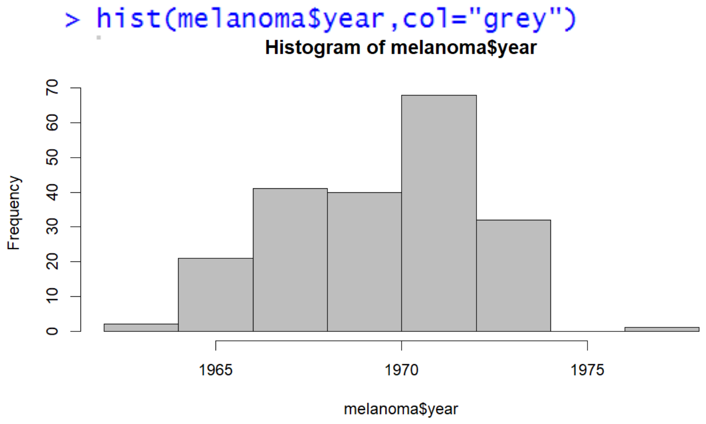
Graphical summaries of the variables are outlined below:



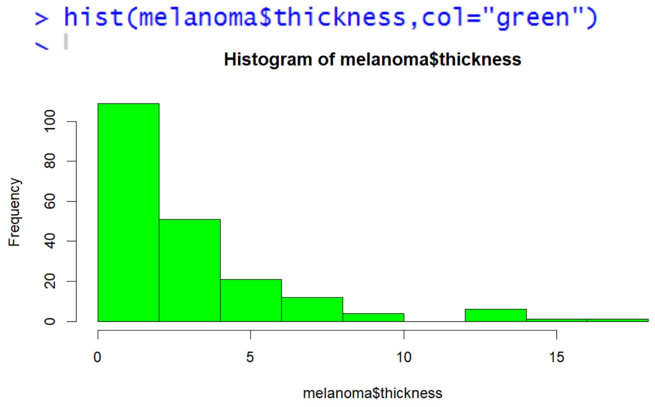
```
> hist(melanoma$age,col="red")
> 
```



```
> hist(melanoma$time,col="lightyellow")
> 
```

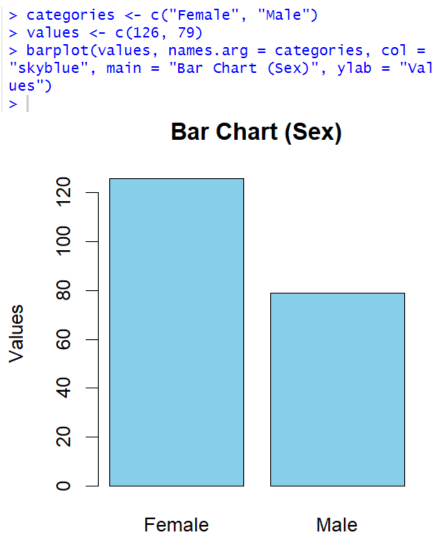


```
> hist(melanoma$year,col="grey")
```

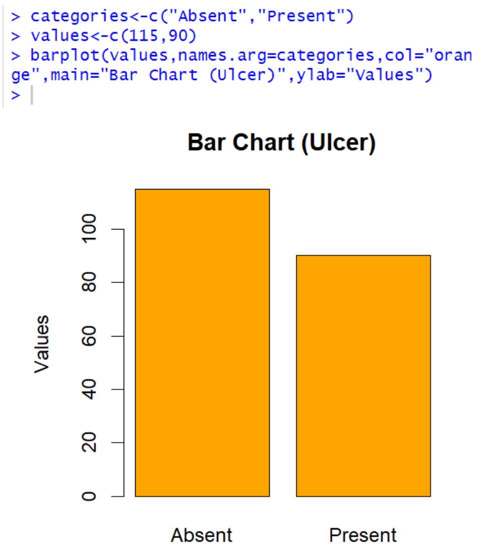


```
> hist(melanoma$thickness,col="green")
> 
```

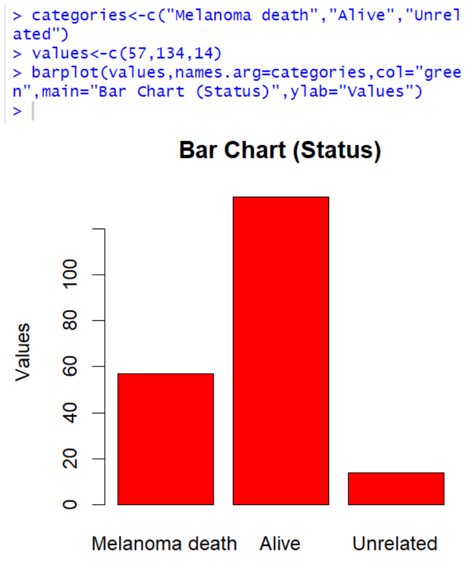
Bar charts have been used for the graphical illustrations of the nominal categories:



```
> categories <- c("Female", "Male")
> values <- c(126, 79)
> barplot(values, names.arg = categories, col = 
"skyblue", main = "Bar Chart (Sex)", ylab = "Val
ues")
> 
```



```
> categories<-c("Absent","Present")
> values<-c(115,90)
> barplot(values,names.arg=categories,col="oran
ge",main="Bar Chart (Ulcer)",ylab="Values")
> 
```



```
> categories<-c("Melanoma death","Alive","Unrel
ated")
> values<-c(57,134,14)
> barplot(values,names.arg=categories,col="gree
n",main="Bar Chart (Status)",ylab="Values")
> 
```

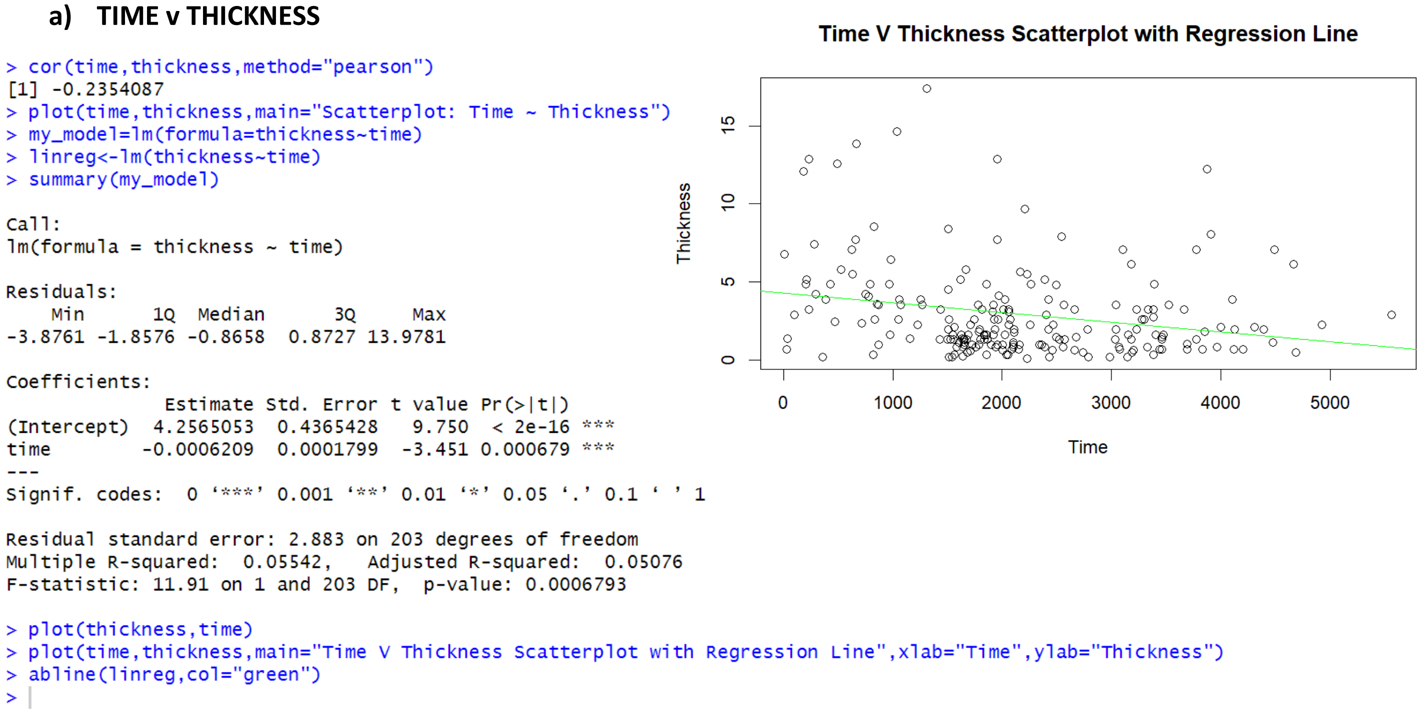
COMMENTS

In the histogram for thickness, we can observe a very high value. This is probably the outlier (max - 17.42mm) observed in the summary statistics, indicative of a very large tumour removed from a patient.

In the bar chart for status, 28% of patients died from melanoma deaths. This is inconsistent with the marginal presence/ absence of ulceration. However, the 7% who died of unrelated causes could be attributed to old age. The gender distribution of patients also suggests that females are more susceptible to melanoma than males.

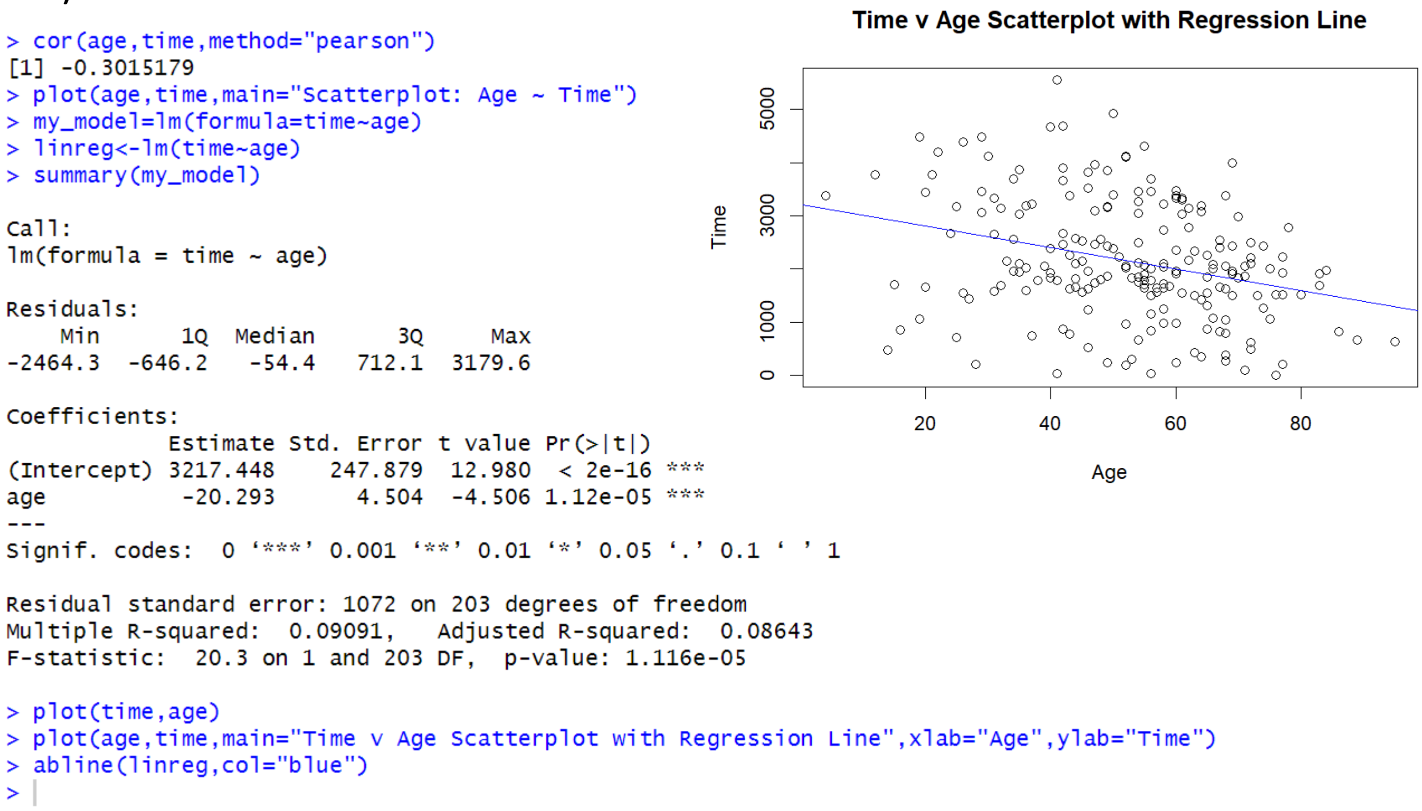
iii. REGRESSION ANALYSIS AND CORRELATION COMPUTATIONS

a) TIME v THICKNESS



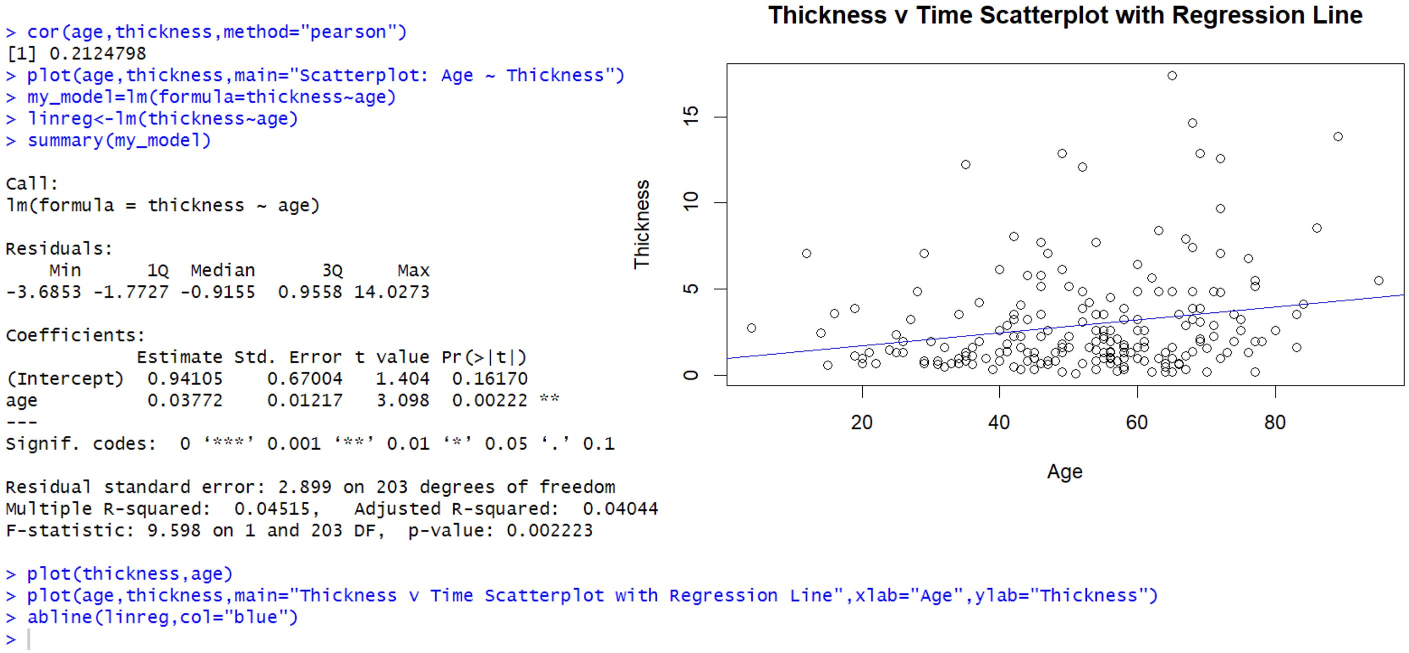
```
> cor(time,thickness,method="pearson")
[1] -0.2354087
> plot(time,thickness,main="Scatterplot: Time ~ Thickness")
> my_model=lm(formula=thickness~time)
> linreg<-lm(thickness~time)
> summary(my_model)

Call:
lm(formula = thickness ~ time)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8761 -1.8576 -0.8658  0.8727 13.9781 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.2565053  0.4365428   9.750  < 2e-16 ***
time        -0.0006209  0.0001799  -3.451 0.000679 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.883 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

> plot(thickness,time)
> plot(time,thickness,main="Time V Thickness Scatterplot with Regression Line",xlab="Time",ylab="Thickness")
> abline(linreg,col="green")
> 
```

b) TIME V AGE



```
> cor(age,time,method="pearson")
[1] -0.3015179
> plot(age,time,main="Scatterplot: Age ~ Time")
> my_model=lm(formula=time~age)
> linreg<-lm(time~age)
> summary(my_model)

Call:
lm(formula = time ~ 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 ***
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

> plot(time,age)
> plot(age,time,main="Time v Age Scatterplot with Regression Line",xlab="Age",ylab="Time")
> abline(linreg,col="blue")
> 
```

c) THICKNESS V AGE



```
> cor(age,thickness,method="pearson")
[1] 0.2124798
> plot(age,thickness,main="Scatterplot: Age ~ Thickness")
> my_model=lm(formula=thickness~age)
> linreg<-lm(thickness~age)
> summary(my_model)

Call:
lm(formula = thickness ~ 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   
age          0.03772    0.01217   3.098  0.00222 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

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

> plot(thickness,age)
> plot(age,thickness,main="Thickness v Time Scatterplot with Regression Line",xlab="Age",ylab="Thickness")
> abline(linreg,col="blue")
> 
```

iv. COMMENTARY ON iii

According to Pearson’s correlation coefficient, when ***r* < *0*** indicates a negative correlation, ***r* = *0*** indicates zero correlation, and ***r* > *0*** indicates a positive correlation. The results of the analysis are as follows:

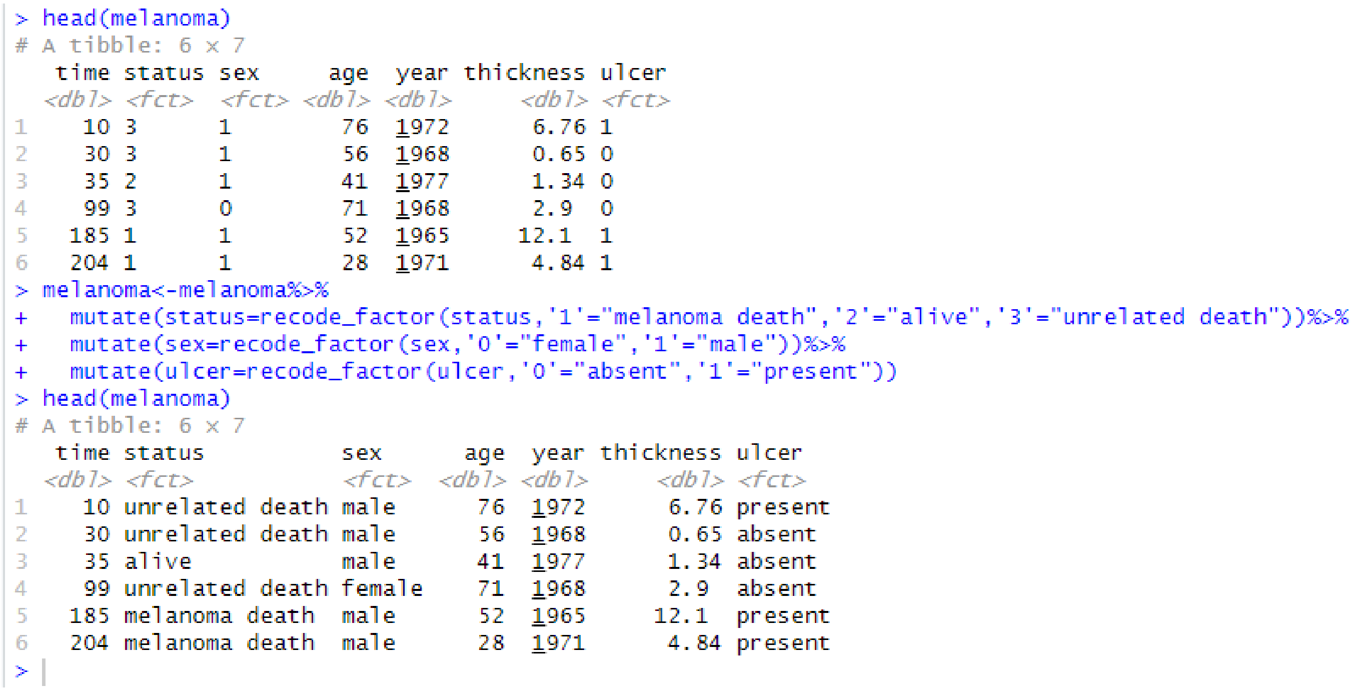
- **Time v Thickness**: Tumour thickness reduces with survival time, indicating a negative correlation (r = -0.235). **P-value is 0.0006793**, therefore there is weak negative linear relationship between survival time and thickness.

- **Time v Age:** Survival time reduces with patient age; this is a negative correlation (r = -0.301). **P-value is 0.00001116**, indicating that there is weak linear relationship between survival time and patient age. i.e. the younger the patient, the less the survival time and vice versa. However, outliers exist in the scatterplot to suggest that there are some rare cases where the situation is reversed.

- **Thickness v Age:** Tumor thickness increases with age, indicating a positive correlation (r = 0.212). **P-value is 0.002223**, therefore there is a strong linear relationship between thickness and age. This is consistent with the notion that in most melanoma cases, the age of the patient can determine or influence the thickness of the tumour [3]. Anomalies are also present in the scatterplot that indicate larger tumour thickness among patients of all ages, this could be as a result of unknown variables that are not present in the dataset.

v. SAMPLE SIGNIFICANCE TESTS BY GENDER

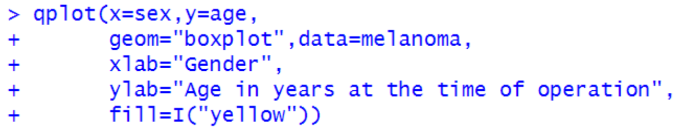
Using a tibble, the data set has been re-labeled to reflect the actual values of the nominal variables (status, sex and ulcer).



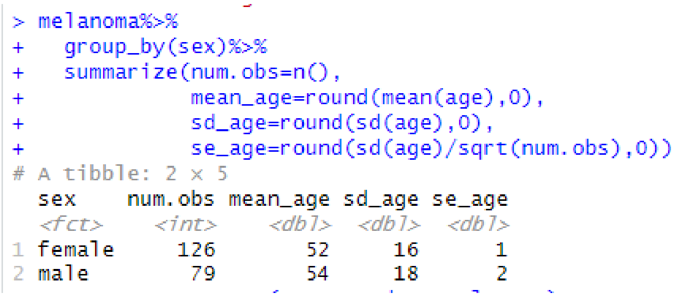
```
> head(melanoma)
# A tibble: 6 x 7
   time status sex    age  year thickness ulcer
  <dbl> <fct>  <fct> <dbl> <dbl>     <dbl> <fct>
1    10 3      1        76  1972      6.76 1
2    30 3      1        56  1968      0.65 0
3    35 2      1        41  1977      1.34 0
4    99 3      0        71  1968      2.9  0
5   185 1      1        52  1965     12.1  1
6   204 1      1        28  1971      4.84 1
> melanoma<-melanoma%>%
+   mutate(status=recode_factor(status,'1'="melanoma death",'2'="alive",'3'="unrelated death"))%>%
+   mutate(sex=recode_factor(sex,'0'="female",'1'="male"))%>%
+   mutate(ulcer=recode_factor(ulcer,'0'="absent",'1'="present"))
> head(melanoma)
# A tibble: 6 x 7
   time status          sex     age  year thickness ulcer
  <dbl> <fct>           <fct>  <dbl> <dbl>     <dbl> <fct>
1    10 unrelated death male      76  1972      6.76 present
2    30 unrelated death male      56  1968      0.65 absent
3    35 alive           male      41  1977      1.34 absent
4    99 unrelated death female    71  1968      2.9  absent
5   185 melanoma death  male      52  1965     12.1  present
6   204 melanoma death  male      28  1971      4.84 present
> 
```

The Welch Two Sample t- test was used for significance testing of the variables by gender groups:

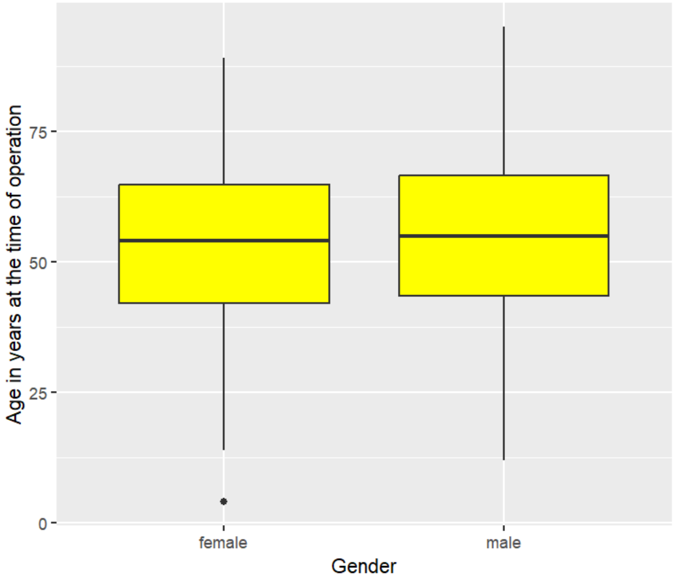
a) AGE GROUPED BY GENDER



```
> qplot(x=sex,y=age,
+       geom="boxplot",data=melanoma,
+       xlab="Gender",
+       ylab="Age in years at the time of operation",
+       fill=I("yellow"))
```



```
> melanoma%>%
+   group_by(sex)%>%
+   summarize(num.obs=n(),
+             mean_age=round(mean(age),0),
+             sd_age=round(sd(age),0),
+             se_age=round(sd(age)/sqrt(num.obs),0))
# A tibble: 2 x 5
  sex    num.obs mean_age sd_age se_age
  <fct>    <int>    <dbl>  <dbl>  <dbl>
1 female     126       52     16      1
2 male        79       54     18      2
```

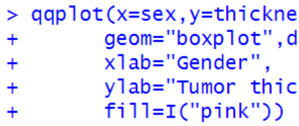


```
> age_t_test<-t.test(age~sex,data=melanoma)
> age_t_test

        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 
```

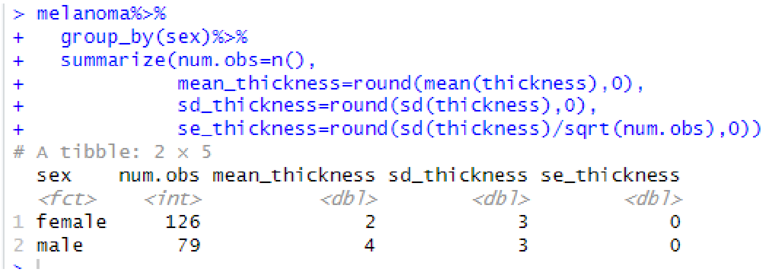
COMMENT

The average patient age in males is slightly higher than in females, as **p-value = 0.3408, therefore p > 0.05**, we accept the null hypothesis that females are more prone than males of the same age. Evident by the lower mean female age (52) and larger female patient size (126) and the single outlier on the female boxplot (the 4-year-old girl).

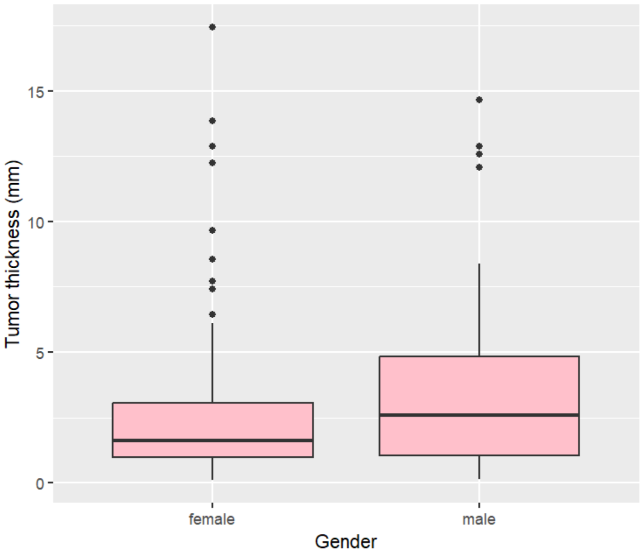
b) THICKNESS GROUPED BY GENDER



```
> qqplot(x=sex,y=thickness,
+        geom="boxplot",data=melanoma,
+        xlab="Gender",
+        ylab="Tumor thickness (mm)",
+        fill=I("pink"))
```



```
> melanoma%>%
+   group_by(sex)%>%
+   summarize(num.obs=n(),
+             mean_thickness=round(mean(thickness),0),
+             sd_thickness=round(sd(thickness),0),
+             se_thickness=round(sd(thickness)/sqrt(num.obs),0))
# A tibble: 2 x 5
  sex    num.obs mean_thickness sd_thickness se_thickness
  <fct>    <int>          <dbl>        <dbl>        <dbl>
1 female     126              2            3            0
2 male        79              4            3            0
```

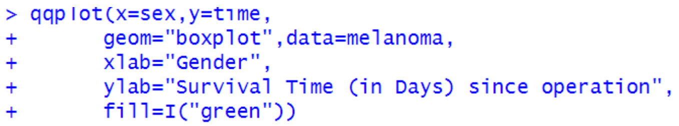


```
> thickness_t_test<-t.test(thickness~sex,data=melanoma)
> thickness_t_test

        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 
```

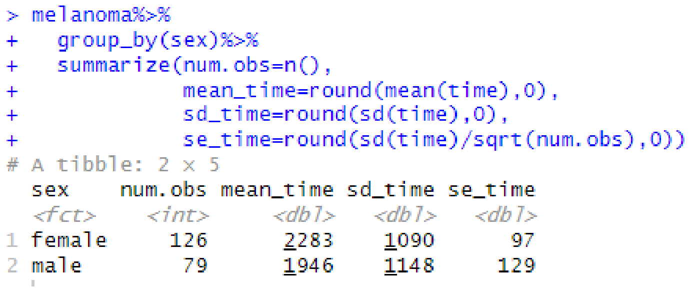
COMMENT

P-value = 0.01009, therefore p < 0.05, we reject the null hypothesis that tumor thickness is higher in males than in females, as the boxplots shows a larger number of outliers on the female boxplot than the male boxplot.

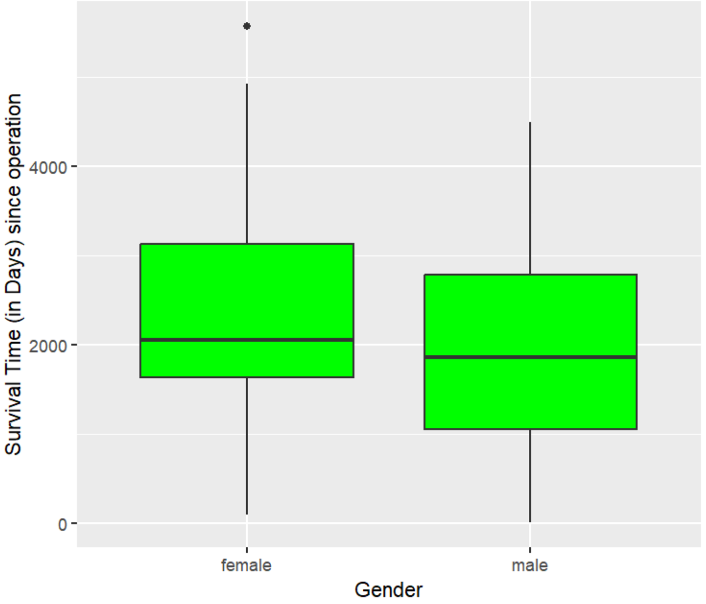
c) TIME GROUPED BY GENDER



```
> qqplot(x=sex,y=time,
+        geom="boxplot",data=melanoma,
+        xlab="Gender",
+        ylab="Survival Time (in Days) since operation",
+        fill=I("green"))
```



```
> melanoma%>%
+   group_by(sex)%>%
+   summarize(num.obs=n(),
+             mean_time=round(mean(time),0),
+             sd_time=round(sd(time),0),
+             se_time=round(sd(time)/sqrt(num.obs),0))
# A tibble: 2 x 5
  sex    num.obs mean_time sd_time se_time
  <fct>    <int>     <dbl>   <dbl>   <dbl>
1 female     126      2283    1090      97
2 male        79      1946    1148     129
```



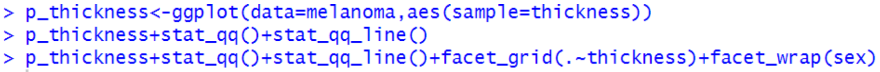
```
> time_t_test<-t.test(time~sex,data=melanoma)
> time_t_test

        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 
```

COMMENT

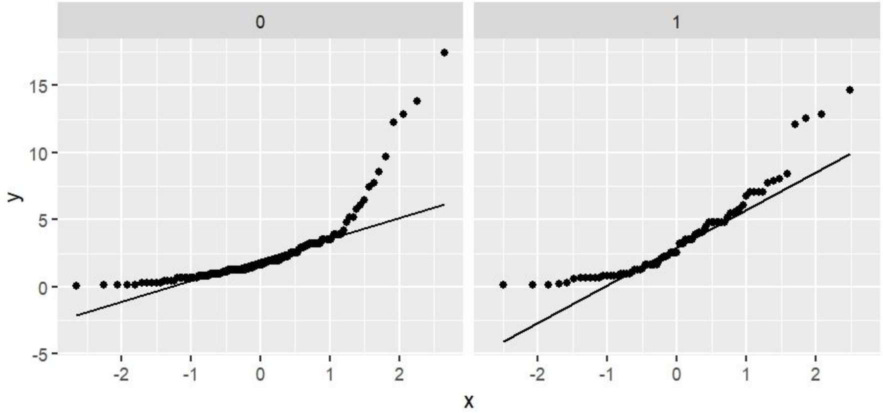
P-value = 0.03868, therefore p < 0.05, we reject the null hypothesis that survival time is higher in males than in females, the female boxplot outlier is the first female to undergo the surgery in 1962 and survived the longest (a total of 5,565 days)

vi. QQPLOTS BY GENDER

a) THICKNESS VS SEX



```
> p_thickness<-ggplot(data=melanoma,aes(sample=thickness))
> p_thickness+stat_qq()+stat_qq_line()
> p_thickness+stat_qq()+stat_qq_line()+facet_grid(.~thickness)+facet_wrap(sex)
```

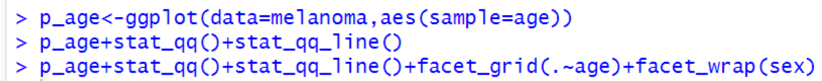


COMMENT
The data for the relationship between thickness and sex indicates an abnormal distribution with more outliers present on the female plot than make plot.

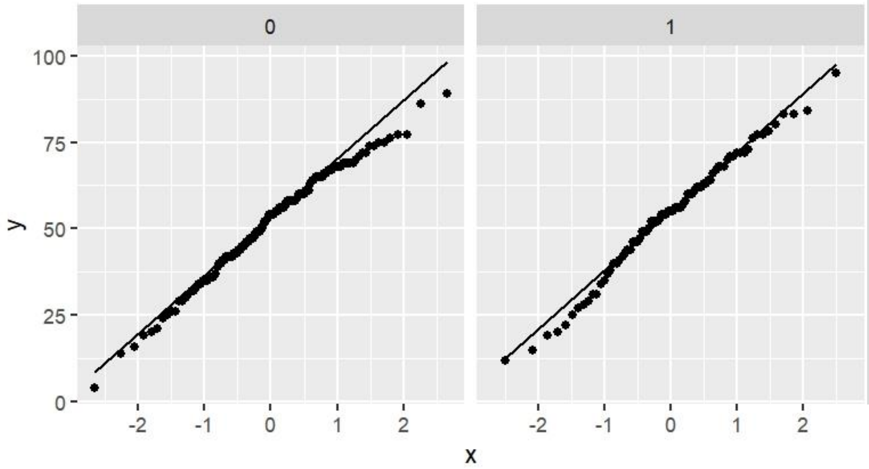
LEGEND

Y - axis	Thickness
X - axis	Sex (0/1)

b) AGE VS SEX



```
> p_age<-ggplot(data=melanoma,aes(sample=age))
> p_age+stat_qq()+stat_qq_line()
> p_age+stat_qq()+stat_qq_line()+facet_grid(.~age)+facet_wrap(sex)
```

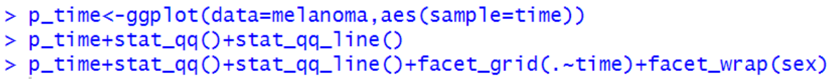


COMMENT
This data lies close to the straight line which is a good indication that the data is normally distributed. However, both possess outliers that the age falls below the line.

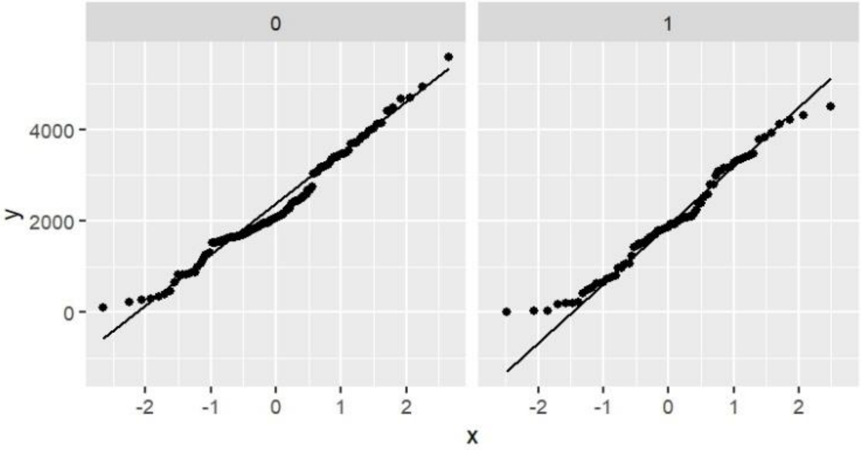
LEGEND

Y - axis	Age
X - axis	Sex (0/1)

TIME VS SEX



```
> p_time<-ggplot(data=melanoma,aes(sample=time))
> p_time+stat_qq()+stat_qq_line()
> p_time+stat_qq()+stat_qq_line()+facet_grid(.~time)+facet_wrap(sex)
```



COMMENT
The time/sex data lies close to the straight line which is a good indication that the data is normally distributed. Again, outliers exist for both genders, suggesting divergent survival times from the normal distribution.

LEGEND

Y - axis	Time
X - axis	Sex (0/1)

vii. DISCUSSION.

INSIGHTS:

Results of the study over the years show that malignant melanoma is treatable and improved over time. It also establishes the fact that females are more susceptible to melanoma than males. The positive correlation between tumour thickness and age is proof that thickness is directly related to the patient's age, therefore the earlier the treatment, the higher the chances of survival.
The only consistent relationship between variables is age and thickness [3], as observed across the various scatterplots, boxplots and charts. There is a definitely strong relationship between the patients' age and the size of the tumor removed from their bodies. However, it is important to recognize the anomalies in the study such as:

- The 4-year-old girl with melanoma, which could be a result of genetic predisposition to the disease, or environmental factors – both of which are not included in the study.
- The longest surviving 41-year-old woman (who was also the very first and only patient of the study in 1962).
- The shortest survival time of 10 day goes to a 76-year-old man. This may seem consistent with his age, a deeper evaluation would reveal new information that was previously unknown.

These 3 examples above are identified as outliers as they may not always conform to the status quo.

The study was moderately successful given the time it was conducted as evident by the averages of survival time and patients 'ages. Due to the dated nature of the dataset, it may be inferred that some scientific approaches, lifestyle habits, behaviors, genetics etc., may have been overlooked or ignored to preserve the integrity of the patients. However, in the 21st century, there are areas for improvement going forward. Hence, a list of recommendations below.

RECOMMENDATIONS:

The study does not factor some elements such as patient weight, alcohol consumption, tumour location, and smoking in the study. It is believed that these variables could add a deeper level of analysis to the adapted dataset [2] and provide more concise conclusions:

- **Tumour location** would've been a good way to further analyze the data set, it is known that the location of a wound would largely determine the survival of the patient, post-surgery. The likelihood of a tumour removed from the extremities (arms or legs) resulting in death is very low, compared to removing a tumour from the head or the torso. [6]
- The frequency and volume of **alcohol consumption** would be a good variable to include in this dataset for the analysis of pre-existing behavioral traits of the patients to determine whether this factor contributed to their condition or not. A 2017 Harvard Health blog [5] revealed a 20% melanoma increase in drinkers compared to occasional or non-drinkers.
- It is recommended that **smoking** is included as a dataset variable to add considerably more detail in the analysis given the relationship between smoking and cancer. As suggested by Venosa, A (2019) "melanoma patients with a history of smoking cigarettes are 40% less likely to survive the disease than those who have never smoked" [4].
- The patient's **weight** is another variable that could've contributed to a more concise analysis. Sergentanis, T, etal (2012) concluded that "overweight and obesity are associated with increased risk of malignant melanoma among males. Meticulous assessment of sunlight exposure is needed especially in women, since self-limited public sun exposure may be prevalent among overweight or obese females".

CONCLUSION

With the never-ending advancement of technology in the field of medicine, new strategies become more available to manage life threatening diseases. However, without proper data analysis to provide understand and solutions to these challenges, a lot of progress would not have been made.
The devil as they say, is in the details.

REFERENCES

[1] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL: https://www.R-project.org/

[2] Drezewiecki, K. (1982). ‘Survival With Malignant Melanoma: A Regression Analysis Of Prognostic Factors’. American Cancer Society Journals.
URL: https://doi.org/10.1002/1097-0142(19820601)49:11<2414::AID-CNCR2820491132>3.0.CO;2-V

[3] Radu, S. (2020). ‘Relationship of Patient Age To Tumor Factors And Outcomes Among Patients Undergoing Sentinel Node Biopsy For Melanoma**’.** The American Journal for Surgery. Vol. 219 (5), p.836-840.

[4] Venosa, A. (2019). ‘History Of Smoking Affects How The Body Fights Melanoma’. Skin Cancer Foundation.
URL: https://www.skincancer.org/blog/history-of-smoking-affects-how-the-body-fights-melanoma/#:~:text=The%20study%2C%20funded%20by%20Cancer,those%20who%20have%20never%20smoked.

[5] Harvard Health Blog 2017 ‘Is There A Link Between Alcohol And Skin Cancer?’
URL: https://www.health.harvard.edu/blog/loose-link-alcohol-skin-cancer-2017120812861

[6] Hemo, Y, Gutman, M, Klausner, J, (1999). ‘Anatomic Site of Primary Melanoma is Associated with Depth of Invasion'. JAMA Surgery Network.
URL: https://jamanetwork.com/journals/jamasurgery/fullarticle/390208

[7] Naughton, L. (2023) ‘Week 1-5’. 7CSO39: Statistics for Data Science. Available at: https://canvas.wlv.ac.uk/courses/36899/modules (Accessed: 18 December 2023)