EXPLORATORY DATA ANALYSIS OF THE MELANOMA DATASET USING R.

2237361

JANUARY, 2024

EXPLORATORY DATA ANALYSIS OF THE MELANOMA DATASET IN R

TESTIMONY DAMILOLA OMOKHAFE (2237361)

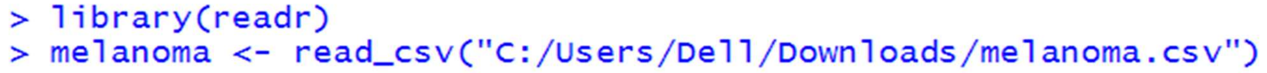
7CS039 STATISTICS FOR AI AND DATA SCIENCE

1. Introduction

In this report we present an exploratory data set conducted using R (RStudio Team 2020)**.** The Survival of Malignant Melanoma dataset which was downloaded from canvas and imported into R studio. The study involved taking measurements from patients who had presented with melanoma and had their tumor removed by the plastic surgery department of University of Odense, Denmark from a period of 1962 to 1977. The data frame contains the following columns: time, status, sex, age, year, thickness and ulcer.

2. **Data Summary**

The melanoma dataset was imported in R.

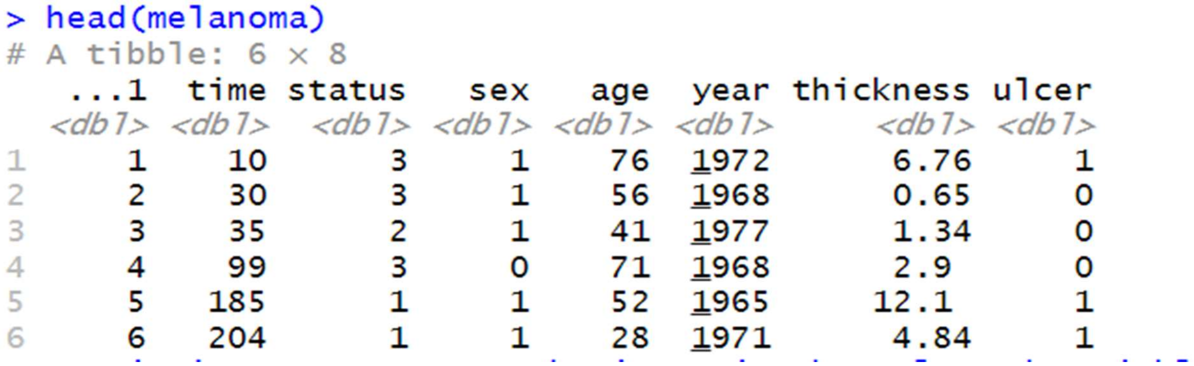


```
> library(readr)
> melanoma <- read_csv("C:/Users/Dell/Downloads/melanoma.csv")
```

Tidy verse (Wickham, Hadley, et al, 2019) was loaded into the R studio to enable certain functions to be carried out in the dataset.

```
> library(tidyverse)
```

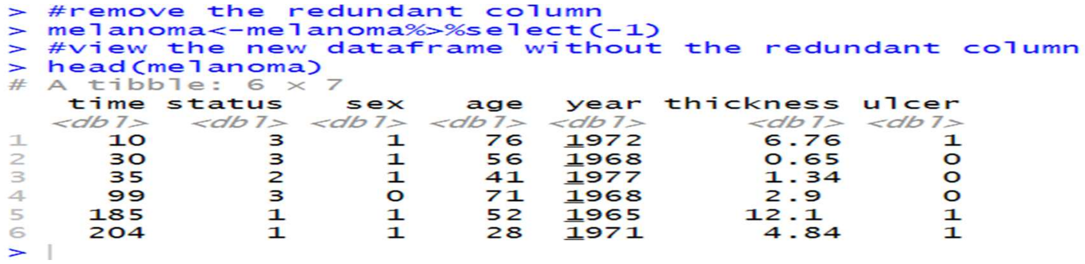
We view the first 5 rows of the data set



```
> head(melanoma)
# A tibble: 6 × 8
   ...1  time status   sex   age  year thickness ulcer
  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>     <dbl> <dbl>
1     1    10      3     1    76  1972      6.76     1
2     2    30      3     1    56  1968      0.65     0
3     3    35      2     1    41  1977      1.34     0
4     4    99      3     0    71  1968      2.9      0
5     5   185      1     1    52  1965     12.1      1
6     6   204      1     1    28  1971      4.84     1
```

The column …1 is the list of the variables numbered 1 to 205, it isn't important in our dataset, we can call it a redundant variable hence we can have it removed by

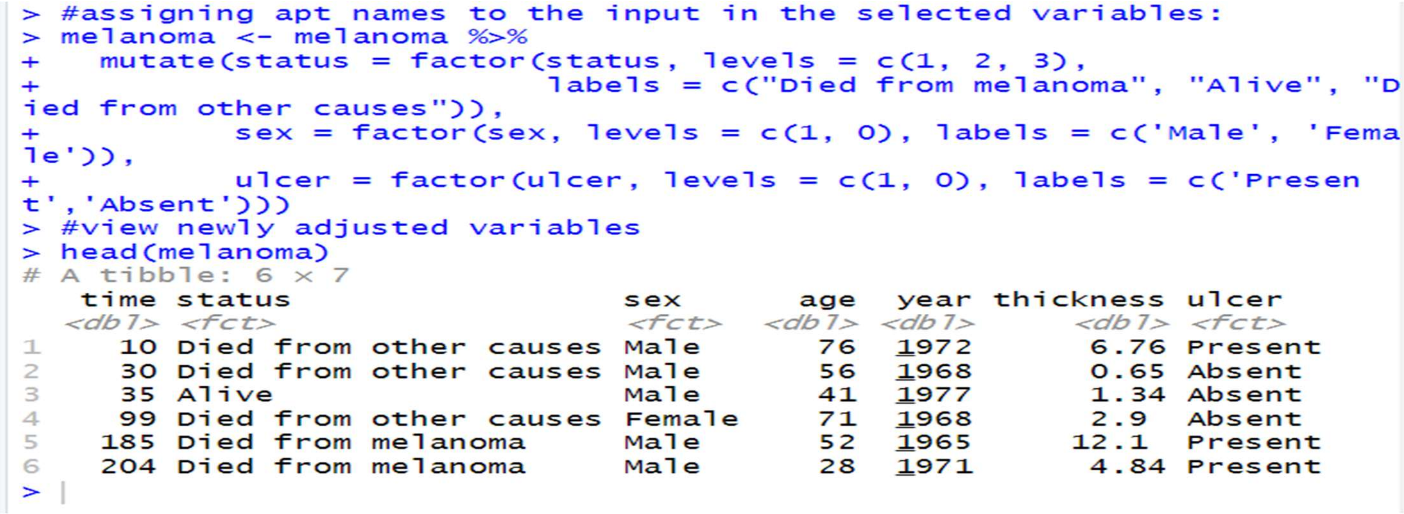
melanoma <-melanoma%>%select(-1)



```
> #remove the redundant column
> melanoma<-melanoma%>%select(-1)
> #view the new dataframe without the redundant column
> head(melanoma)
# A tibble: 6 × 7
   time status   sex   age  year thickness ulcer
  <dbl>  <dbl> <dbl> <dbl> <dbl>     <dbl> <dbl>
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
> 
```

The data frame above shows a removal of the redundant column, however, the inputs of categorical variables: status, sex and ulcer are represented as shown in the introduction title of the report. These inputs seen are categorical for status, sex and ulcer.

The dataset is then seen to transform for what the inputs represents, and the viewed as seen below:

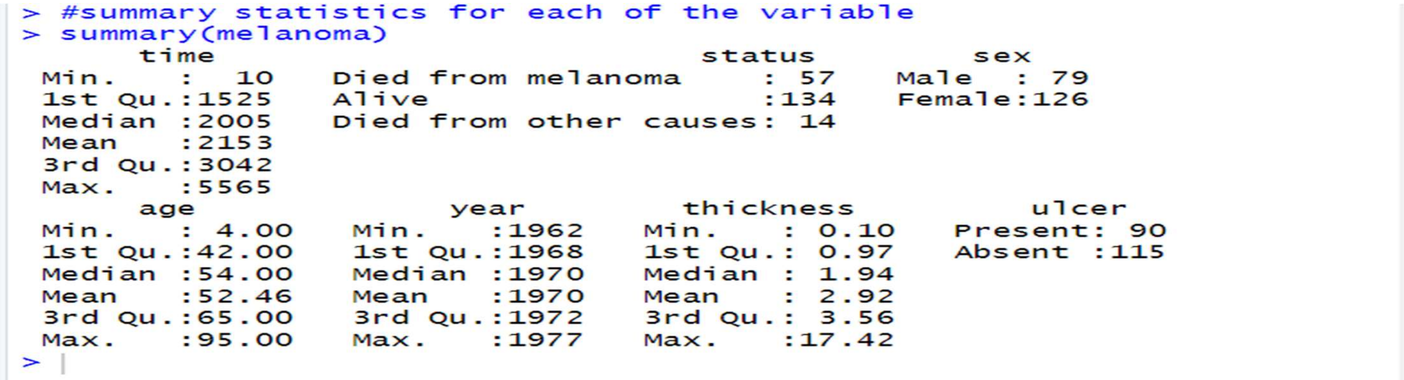


```
> #assigning apt names to the input in the selected variables:
> melanoma <- melanoma %>%
+   mutate(status = factor(status, levels = c(1, 2, 3),
+                          labels = c("Died from melanoma", "Alive", "D
ied from other causes")),
+          sex = factor(sex, levels = c(1, 0), labels = c('Male', 'Fema
le')),
+          ulcer = factor(ulcer, levels = c(1, 0), labels = c('Presen
t','Absent')))
> #view newly adjusted variables
> head(melanoma)
# A tibble: 6 x 7
   time status                 sex      age  year thickness ulcer
  <dbl> <fct>                  <fct>  <dbl> <dbl>     <dbl> <fct>
1    10 Died from other causes Male      76  1972      6.76 Present
2    30 Died from other causes Male      56  1968      0.65 Absent
3    35 Alive                  Male      41  1977      1.34 Absent
4    99 Died from other causes Female    71  1968      2.9  Absent
5   185 Died from melanoma     Male      52  1965     12.1  Present
6   204 Died from melanoma     Male      28  1971      4.84 Present
> |
```

With the transformed dataset now, exploratory analysis can now be carried out on the data set.

2.1. Numerical Summary

The summary of the variables in the dataset can be seen by calling out the summary function



```
> #summary statistics for each of the variable
> summary(melanoma)
      time                           status         sex     
 Min.   :  10   Died from melanoma     : 57   Male  : 79  
 1st Qu.:1525   Alive                  :134   Female:126  
 Median :2005   Died from other causes: 14               
 Mean   :2153                                             
 3rd Qu.:3042                                             
 Max.   :5565                                             
      age             year         thickness          ulcer    
 Min.   : 4.00   Min.   :1962   Min.   : 0.10   Present: 90  
 1st Qu.:42.00   1st Qu.:1968   1st Qu.: 0.97   Absent :115  
 Median :54.00   Median :1970   Median : 1.94                
 Mean   :52.46   Mean   :1970   Mean   : 2.92                
 3rd Qu.:65.00   3rd Qu.:1972   3rd Qu.: 3.56                
 Max.   :95.00   Max.   :1977   Max.   :17.42                
> |
```

Commentary on values:

a. **Time (in days):** we have the minimum survival time after the surgery to be 10days with maximum survival to be 5565days. However, the central tendencies of mean and median are 2153 and 2005 respectively, this shows that the minimum days 10 could be seen in this instance as an outlier in relation to the other observed values.
b. **Status**: seeing the number of persons alive (134) and the deaths unrelated to melanoma (14) compared to the deaths resulting from melanoma (57) it is encouraging to say that the surgery would

overall be a recommendation for patients who have the disease. As deaths from melanoma cover only about 27.80% of the observations.

c. **Sex**: more females (126) than males (79) are seen to be presented with melanoma.
d. **Age** (in years): a maximum age of 95 is seen and minimum of 4 is seen.
e. **Year**: from the summary we can say this dataset covers surgery of melanoma carried out from 1962 to 1977 in said institution.
f. **Thickness** (in mm): the thickness of the tumors range between 0.10(min) and 17.42(max), however, from the dataset, most people would present with tumors ranging between 0.97(Q1) and 3.56(Q3), with an average tumor of about 2.92(mean).
g. **ulcer:** more people had ulcers absent(115) than they did present(90)

2.2 Graphical summary

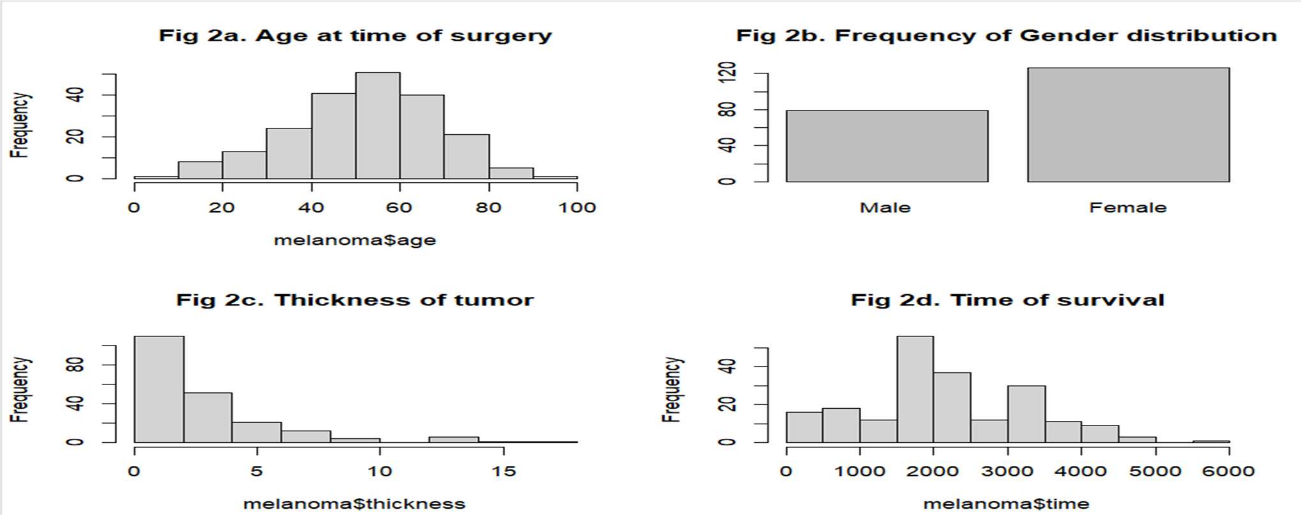


Fig 2a. Age at time of surgery

Fig 2b. Frequency of Gender distribution

Fig 2c. Thickness of tumor

Fig 2d. Time of survival

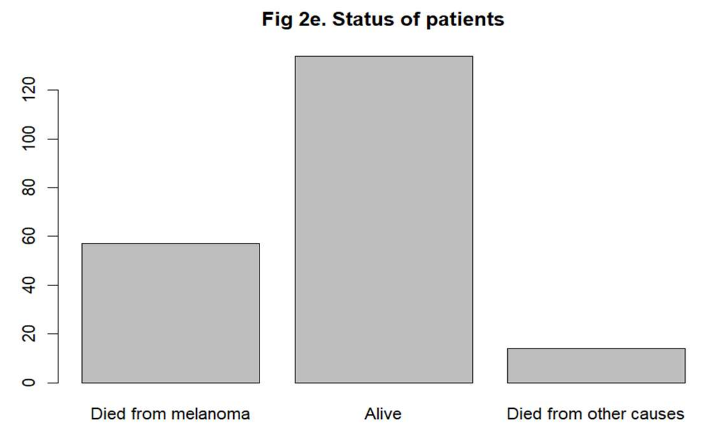


Fig 2e. Status of patients

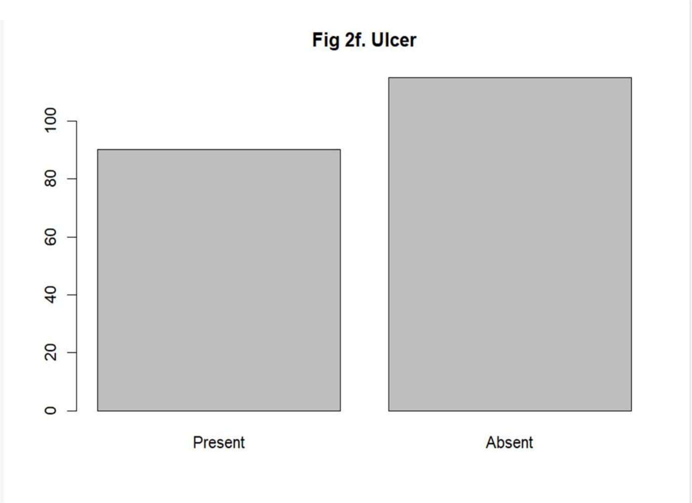


Fig 2f. Ulcer

From our data collected we can see a trend in the following

- (Fig 2a.)The age distribution shows a normal age distribution amongst patients from the histogram.
- More females than males present with the disease. (Fig 2b)
- The thickness distribution shows a left skewed, showing that the distribution is not a normal distribution (Fig 2c)

- (Fig 2d) shows the graphical summary of the time of survival; highest number of survival is seen as the highest bar and the least number of survival on the lowest bar. It also appears to be fairly normally distributed.
- (Fig 2e) shows the graphical summary for the status of patients and we have more patients alive, fewer died from melanoma and lowest ranking of the bar for those that died from other causes.
- (Fig 2f) shows most patients don't have ulcer (absent) and fewer have ulcers (present).

3.1 Regression and Correlation

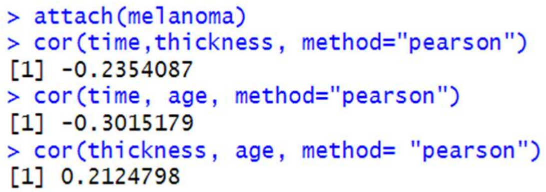
I use the attach function in R to call our data (melanoma), as this allows me to call each variable alone.

```
> attach(melanoma)
```

In the data melanoma, I will be performing correlation between these sets of variables:

- time ~thickness
- time ~ age
- thickness ~ age

We input the codes for each after the attachment as seen



```
> attach(melanoma)
> cor(time,thickness, method="pearson")
[1] -0.2354087
> cor(time, age, method="pearson")
[1] -0.3015179
> cor(thickness, age, method= "pearson")
[1] 0.2124798
```

Correlation values range from -1 to +1(negative to positive). This forms the baseline for assessing relationships between variables. The closer the correlation between two relationships are to any of the extremes, the stronger the relationship and vice versa.

From this we can see that,

- time ~ thickness: (-0.2354087) share a weak negative relationship
- time ~ age: (-0.3015179) share a weak negative relationship
- thickness ~ age: (0.212478) share a weak positive relationship

REGRESSION ANALYSIS

Haven gotten the correlations between each set of variables above, computations of the relationships between each set of variables can be computed using the regression model: **y=mx +c.**

Fig 3a. Relationship between Time and Thickness

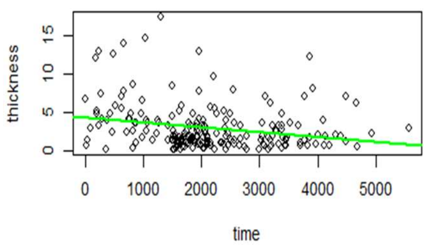


Fig 3b. Relationship between Time and Age

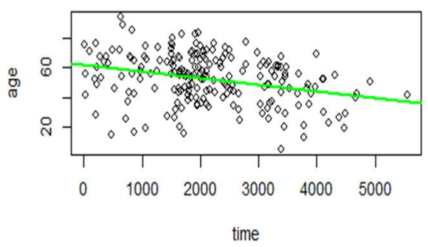
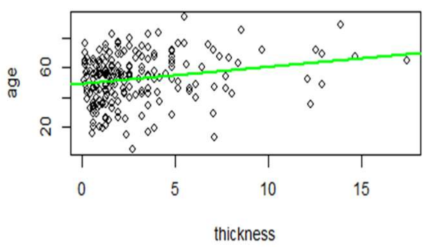
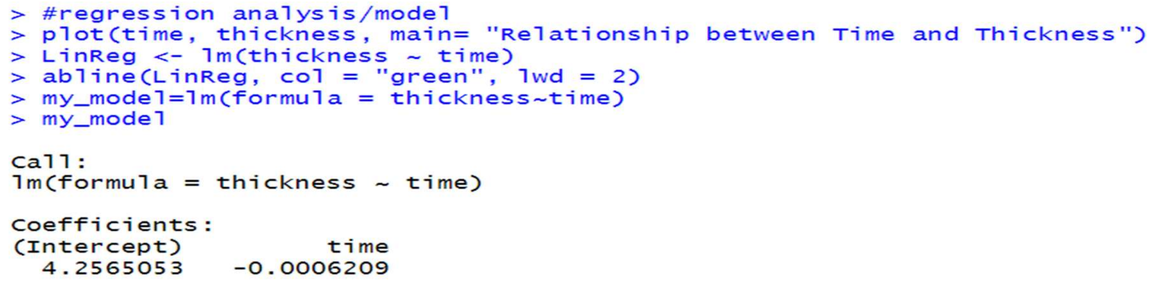


Fig 3c. Relationship between Thickness and Age



For time and thickness: (fig 3a)



```
> #regression analysis/model
> plot(time, thickness, main= "Relationship between Time and Thickness")
> LinReg <- lm(thickness ~ time)
> abline(LinReg, col = "green", lwd = 2)
> my_model=lm(formula = thickness~time)
> my_model

Call:
lm(formula = thickness ~ time)

Coefficients:
(Intercept)         time  
  4.2565053   -0.0006209  
```

So the regression analysis of our model: y=mx+c where y= thickness and x= time.

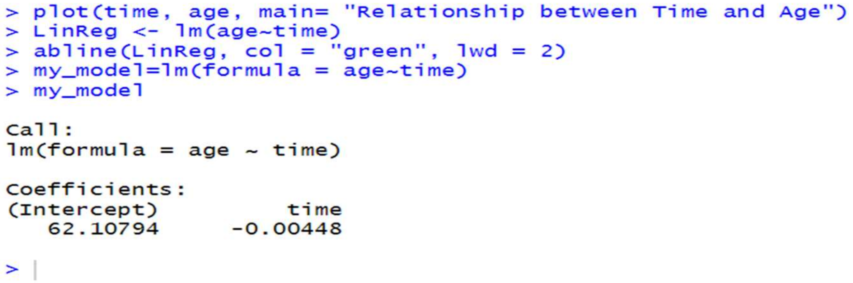
Therefore, from our model,

y= (-0.0006209) x + 4.2565053; or for better organization;

thickness = 4.2565053+ (-0.0006209) time.

That is for every increase in thickness of tumor there is a 0.0006209 decrease in time of survival.

- **For time and age(fig 3b)**



```
> plot(time, age, main= "Relationship between Time and Age")
> LinReg <- lm(age~time)
> abline(LinReg, col = "green", lwd = 2)
> my_model=lm(formula = age~time)
> my_model

Call:
lm(formula = age ~ time)

Coefficients:
(Intercept)         time  
   62.10794     -0.00448  

> |
```

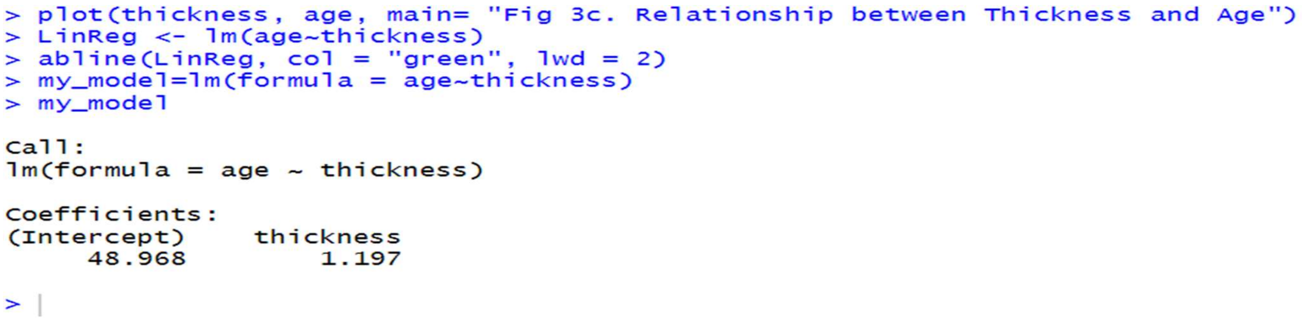
So the regression analysis of our model: y=mx+c where y= age and x= time.

y= (-0.00448) x + 62.10794; or for better organization;

age= 62.10794+ (-0.00448)time

Therefore, from our model, an increase in age will cause a 0.00448 decrease in time of survival

- **For thickness and age(fig 3c)**



```
> plot(thickness, age, main= "Fig 3c. Relationship between Thickness and Age")
> LinReg <- lm(age~thickness)
> abline(LinReg, col = "green", lwd = 2)
> my_model=lm(formula = age~thickness)
> my_model

Call:
lm(formula = age ~ thickness)

Coefficients:
(Intercept)    thickness  
     48.968        1.197  

> |
```

So the regression analysis of our model: y=mx+c where y= age and x= thickness

Therefore, from our model,

y= (1.197) x + 48.968; or for better organization; **age= 48.968+ (1.197)thickness**

This shows that for an increase in age there is likely to be a 1.197 increase in thickness of tumor.

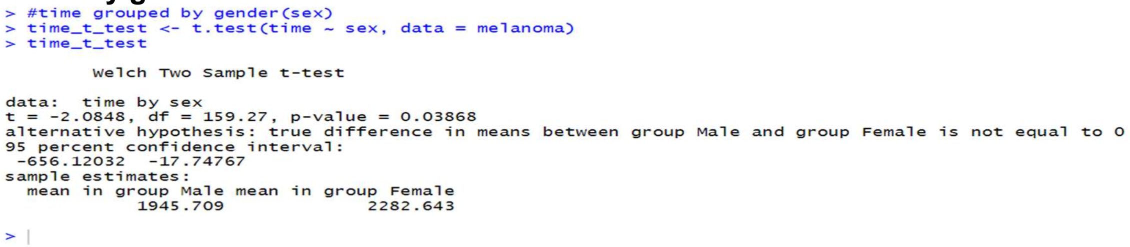
3.2 **Commentary on Observed Relationships between the variables above**

1. Fig 3a (Time and Thickness) shows a downward curve, which is in line with the correlation value (-0.2354087) between the pair of variables time and thickness. This shows a weak inverse(negative) linear relationship between both variables, i.e. time of survival of patients decreases with the increase in thickness of the tumor of the patients and vice versa.
2. Fig 3b (Time and age) shows a similar trend with Fig 3a. A correlation value of -0.3015179. A weak inverse (negative) linear relationship exists between both variables i.e. time of survival of patients decreases as the age of patient's increases and vice versa.
3. Fig 3c (Thickness and age) shows an upward curve which is indicative of a positive relationship between both variables. A positive correlation value is also seen (0.2124798). From the slope direction and the correlation value we can say a weak positive linear relationship exists between both variables, i.e. the more aged the patient is the thicker the tumor size they present with or likely to present with (in prospective).

4. Appropriate Two sample significant tests for Variables grouped by gender

A. T-TEST

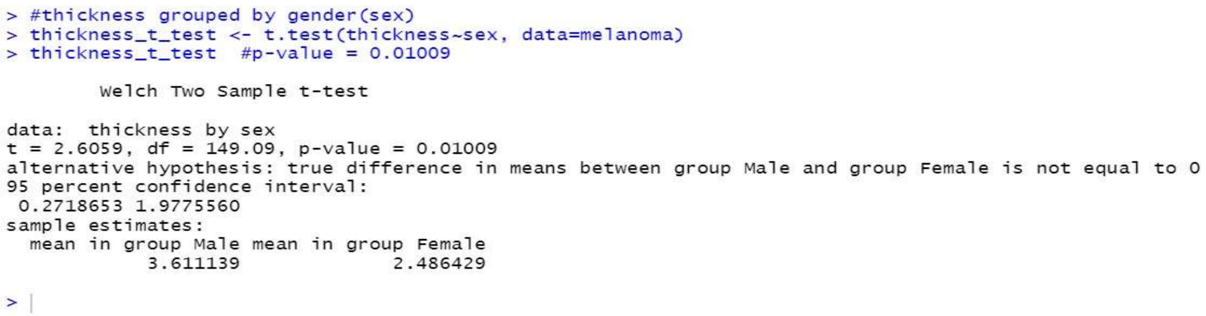
- Time by gender



```
> #time grouped by gender(sex)
> 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 Male and group Female is not equal to 0
95 percent confidence interval:
 -656.12032  -17.74767
sample estimates:
  mean in group Male mean in group Female 
            1945.709             2282.643 

> |
```

From the p-value above (0.03868): the p value: is less than 0.05. Therefore we can reject the null hypothesis and conclude that there is evidence that the true mean time of survival is different for both gender types.

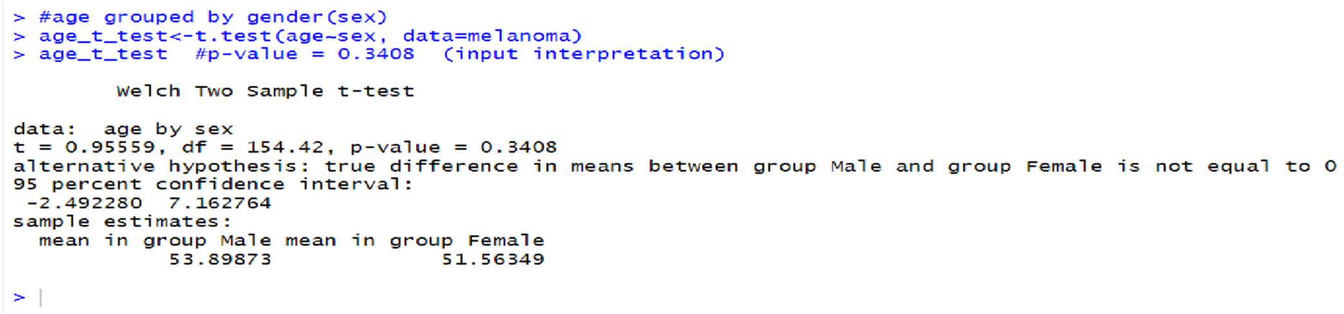
- **Thickness by gender**



```
> #thickness grouped by gender(sex)
> thickness_t_test <- t.test(thickness~sex, data=melanoma)
> thickness_t_test  #p-value = 0.01009

	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 Male and group Female is not equal to 0
95 percent confidence interval:
 0.2718653 1.9775560
sample estimates:
  mean in group Male mean in group Female
            3.611139             2.486429

> |
```

From the p-value above (0.01009): the p value: is less than 0.05. Therefore we can reject the null hypothesis and conclude that there is evidence that the true mean thickness in mm is different for both gender types.

- **Age by gender**

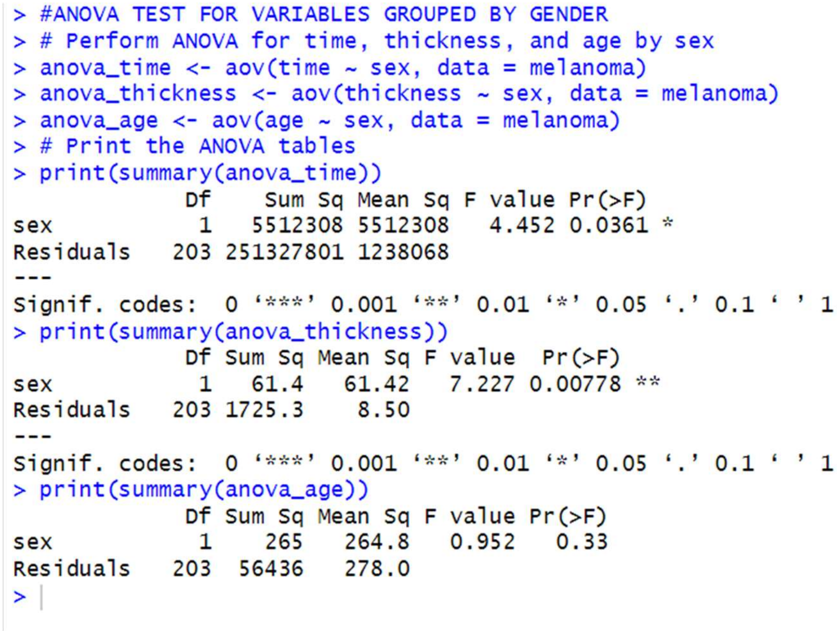


```
> #age grouped by gender(sex)
> age_t_test<-t.test(age~sex, data=melanoma)
> age_t_test  #p-value = 0.3408  (input interpretation)

	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 Male and group Female is not equal to 0
95 percent confidence interval:
 -2.492280  7.162764
sample estimates:
  mean in group Male mean in group Female
            53.89873             51.56349

> |
```

From the p-value above (0.3408): the p value: is more than 0.05. Therefore we may not reject the null hypothesis and conclude that there is evidence that the true mean age is the same for both gender types.

B. ANOVA TEST

For the variables time, thickness and age grouped by gender,



```
> #ANOVA TEST FOR VARIABLES GROUPED BY GENDER
> # Perform ANOVA for time, thickness, and age by sex
> anova_time <- aov(time ~ sex, data = melanoma)
> anova_thickness <- aov(thickness ~ sex, data = melanoma)
> anova_age <- aov(age ~ sex, data = melanoma)
> # Print the ANOVA tables
> print(summary(anova_time))
             Df    Sum Sq Mean Sq F value Pr(>F)  
sex           1   5512308 5512308   4.452 0.0361 *
Residuals   203 251327801 1238068                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> print(summary(anova_thickness))
             Df Sum Sq Mean Sq F value  Pr(>F)   
sex           1   61.4   61.42   7.227 0.00778 **
Residuals   203 1725.3    8.50                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> print(summary(anova_age))
             Df Sum Sq Mean Sq F value Pr(>F)
sex           1    265   264.8   0.952   0.33
Residuals   203  56436   278.0               
> |
```

From the above table we have the p-values of the ANOVA tests for the variables by gender

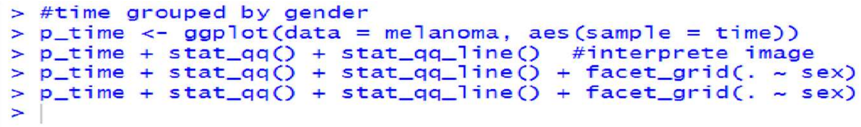
- **For time**: p-value is 0.0361: The ANOVA summary output indicates that there is a significant difference in the means of the 'time' variable between the two levels of the categorical variable 'sex'. The p-value (Pr(>F)) is less than the commonly used significance level of 0.05 (indicated by the asterisk '*'), we can reject the null hypothesis of equal means. Therefore, we conclude that there is a significant difference in the mean 'time' between both genders.

-**For thickness**: p-value is 0.00778: The ANOVA summary output for the 'thickness' variable here indicates a significant difference in means between the two levels of the categorical variable 'sex'. The p-value (Pr(>F)) is less than 0.05, therefore we can reject the null hypothesis of equal means. Therefore, we conclude that there is a significant difference in the mean 'thickness' between both genders.

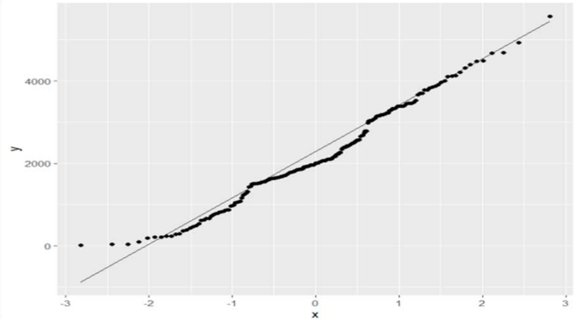
- **For age:** p-value is 0.33: The ANOVA summary output for the 'age' variable indicates that there is no significant difference in means between the two levels of the categorical variable 'sex'. The p-value (Pr(>F)) is 0.33, which is greater than 0.05. We fail to reject the null hypothesis of equal means. Therefore, we would not consider the difference in mean 'age' between the different levels of 'sex' to be statistically significant.

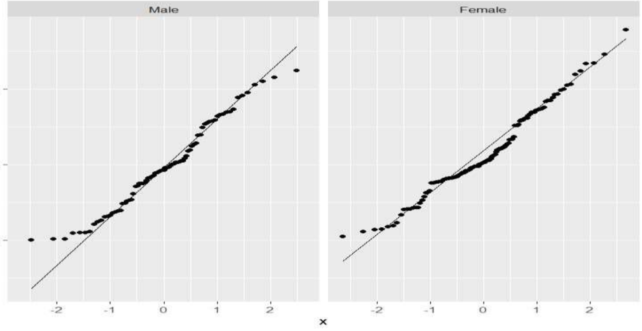
5. QQ-Plots and commentary about underlying distribution of variables

- **Time and gender**



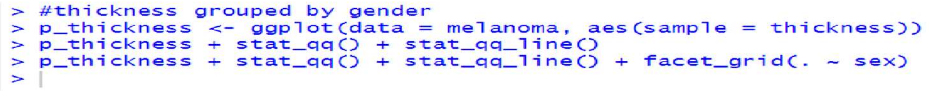
```
> #time grouped by gender
> p_time <- ggplot(data = melanoma, aes(sample = time))
> p_time + stat_qq() + stat_qq_line()   #interprete image
> p_time + stat_qq() + stat_qq_line() + facet_grid(. ~ sex)
> p_time + stat_qq() + stat_qq_line() + facet_grid(. ~ sex)
> 
```



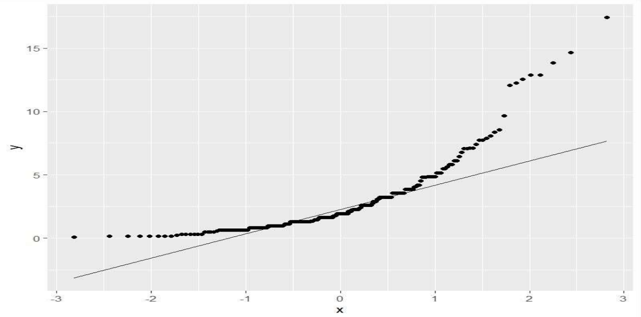


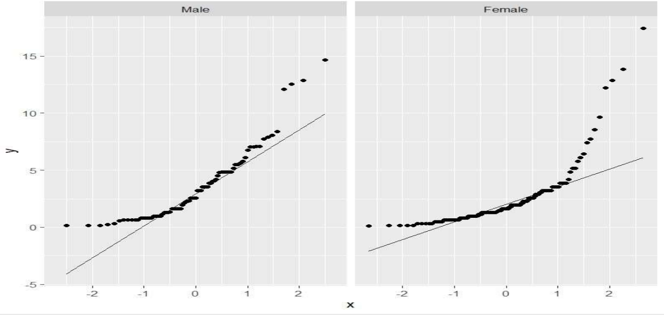
We want to see our data lying pretty close to the straight line. This is a good indication that our data is normally distributed. And for both genders (male and female) grouped by Time, we can see that it is normally distributed.

- **Thickness by Gender**



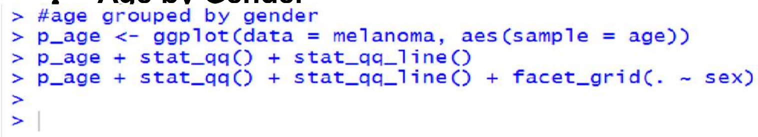
```
> #thickness grouped by gender
> p_thickness <- ggplot(data = melanoma, aes(sample = thickness))
> p_thickness + stat_qq() + stat_qq_line()
> p_thickness + stat_qq() + stat_qq_line() + facet_grid(. ~ sex)
> 
```



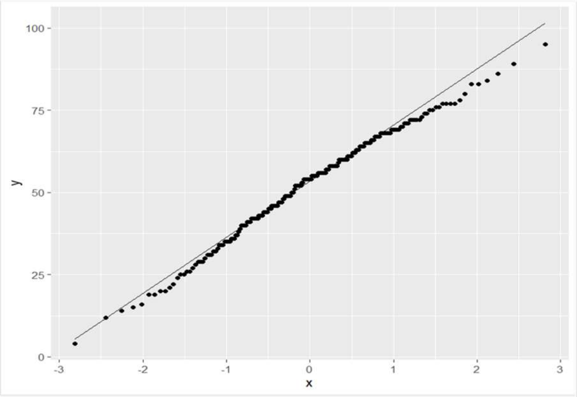


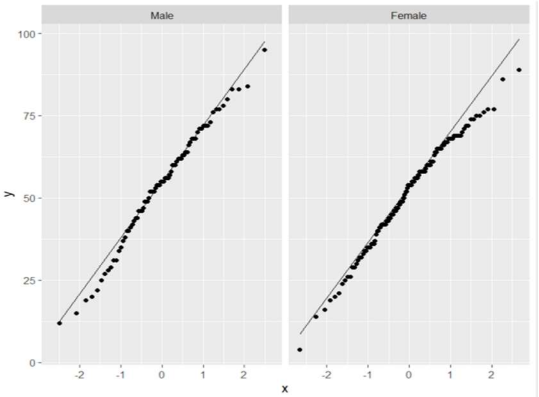
We want to see our data lying pretty close to the straight line. This is a good indication that our data is normally distributed. However, in this case, the data lies away from the straight line for both genders (male and female) grouped by thickness, so we can say that it is not normally distributed.

- **Age by Gender**



```
> #age grouped by gender
> p_age <- ggplot(data = melanoma, aes(sample = age))
> p_age + stat_qq() + stat_qq_line()
> p_age + stat_qq() + stat_qq_line() + facet_grid(. ~ sex)
>
> |
```





We want to see our data lying pretty close to the straight line. This is a good indication that our data is normally distributed. And for both genders (male and female) grouped by age, we can see that it is normally distributed.

5. DISCUSSION ON INSIGHTS GENERATED FROM THE DATA

Melanoma is a type of cancer that arises from melanocytes, which are cells that produce pigment. (Schwartzman RM, 1962) Malignant melanoma is another term for the same malignancy. Although they seldom happen in the mouth, intestines, or eyes. Melanomas usually develop in the skin. For women, they usually appear on the legs, and for men, they usually appear on the back. UV radiation is the main factor contributing to melanoma in people whose melanin pigment levels are low. The UV rays could originate from the sun or other objects, such tanning beds. People who have a lot of moles, a family history of the condition, and weakened immune systems are more vulnerable. The process of diagnosing involves taking a biopsy and analysing any skin lesion that exhibits indications of being malignant. The most deadly kind of skin cancer is melanoma. In 2012, it happened for the first time in 232,000 persons worldwide. 3.1 million individuals had an active illness in 2015, which led to 59,800 fatalities.(Vos T. et al, 2016). May of every year, is set aside for melanoma awareness creation.

Our data consists of variables such as time of survival, patient status, and age of patient, tumor thickness and presence and absence of ulcer. These were all analyzed in this report to give several valuable insights. Here are some of the insights one could derive.

Survival analysis: from our data, the mean survival time is 2153 days (approximately 6 years), for patients who have had the surgery for melanoma. This is likely to be influenced by factors. What could make both patients have the surgery and different prognostic values?

From our data:

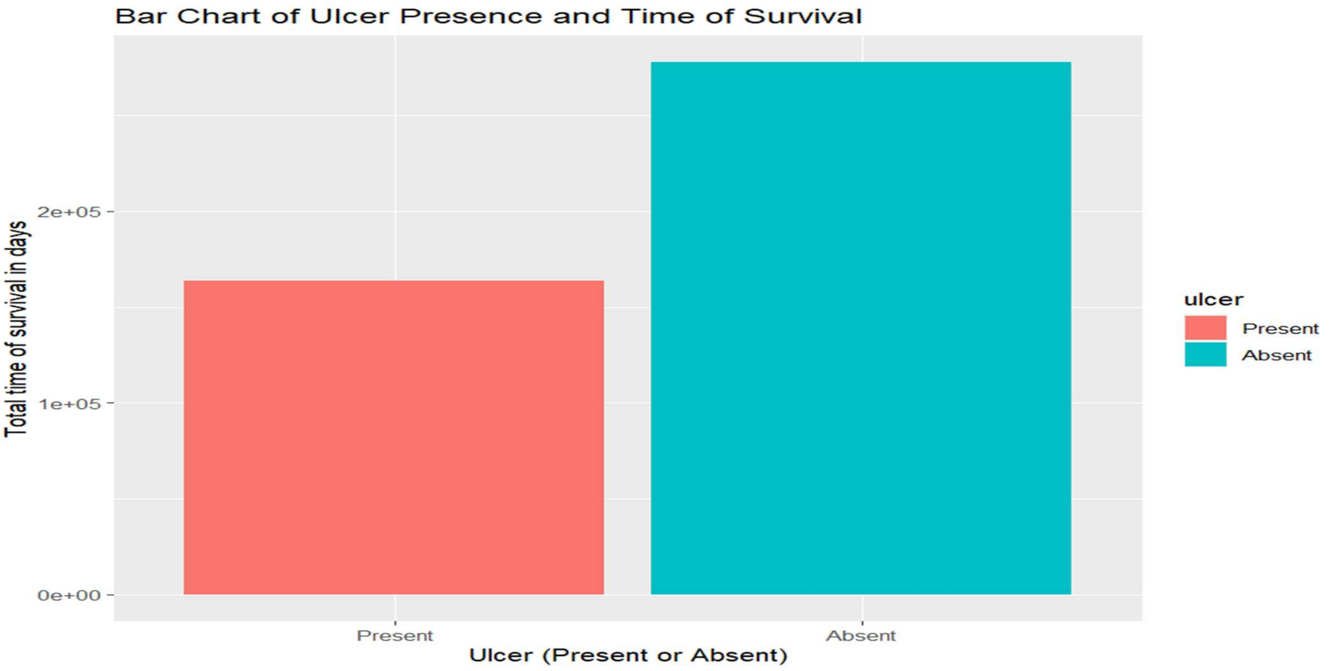
- **Impact of age**: the correlation between survival times of patients and the age of patients at time of surgery shared an inverse relationship (as seen from our correlation: weak negative relationship, value: -0.31), this is to say generally older patients are less likely to have lengthened survival time post-op than younger patients. In general, older individuals have a worse prognosis (Canadian Cancer Society, 2024). The older a patient the less likely the he is to have long survival time and vice versa. Age is key prognostic factor from our data.

- **Tumour thickness**: There's an inverse relationship seen between the thickness of the tumor and survival time from the correlation analysis (weak negative, value: -0.24). The thickness of the primary tumour is an important prognostic factor. It helps predict the risk that the cancer will spread. The prognosis is worse with a thicker tumour (Canadian Cancer Society, 2024).

.

- **Status of patient**: The death status of patients from the data are into two- death from melanoma and death from other causes. Survival rates can give you an idea of what percentage of people with the same type and stage of cancer are still alive a certain amount of time (usually 5 years) after they were diagnosed. Although they cannot predict your exact life span, they can help you better appreciate the likelihood that your therapy will be effective (American Cancer Society, 2023). The death toll from patients who died from melanoma are fewer (57) than those alive (134), however, resulting death from other causes when compared to deaths from melanoma appeared fewer (14). What does this tell us? That more people would have good prognosis from surgery (e.g. alive), death may still result from the melanoma for some (this could be influenced by other factors such as how progressive the disease was, etc.).

- **Ulcer:** Ulceration is a breakdown of the skin over the melanoma. Melanomas that are ulcerated tend to have a worse outlook (American Cancer Society, 2023). For the sake of this discuss I'll be inserting a bar chart to give insight on how presence of ulcers give an interpretation to its input on time of survival.



The bar chart shows the overall number of persons having ulcer present and absent and how they measure against time of survival in days. The total time of survival represents the sum time of survival for each patient identified with presence or absence of ulcers. It is observed that those who have ulcers present are seen to have fewer days of survival while those who have no ulcers present display a longer survival times. This is quite insightful as it shows that an ulcerated melanoma confers good prognosis (based on survival times) on patients and vice, versa.

In conclusion, one can use a number of phrases to characterize a prognosis, such as excellent (showing a very high possibility of recovery), guarded (expressing uncertainty), unfavorable (representing a negative outlook), or favorable (indicating a decent chance of recovery). It is crucial to realize that a prognosis is an educated guess based on the information at hand and medical experience, not a promise. However, collaborations of these findings with clinicians and oncologists would further enhance the interpretation and applications of these insights in a real world clinical setting. The goal of providing a prognosis is to help patients and their families understand what to expect and make informed decisions about treatment options and planning for the future.

RECOMMENDATIONS

It would be okay to encourage patients to perform the surgery for good prognostic outcomes based on available data, but a comparative data exploration for patients who presented with melanoma but never had the surgery as I believe comparison could give full closure on how clinicians can best guide their patients on actions to take. A short sample size is one potential drawback of the Melanoma dataset. More than 205 patients could ideally have been surveyed, given the years covered by the study are 15 years, from 1962 to 1977. With such a small sample size, it could be difficult to ascertain whether any other significant associations exist. Another issue with this dataset is that there are unequal numbers of observations for each year, making it unsuitable for us to create a time series that looks at the trends in the future. If we have the most recent data, it is also nice to examine the state of affairs.

REFERENCES

American Cancer Society. (2023). Treating Melanoma Skin Cancer. Cancer.org. Available at: https://www.cancer.org/cancer/types/melanoma-skin-cancer/treating.html (Accessed: 14 January 2024).

Canadian Cancer Society. (2024) 'Prognosis and Survival in Skin Melanoma', Canadian Cancer Society. https://cancer.ca/en/cancer-information/cancer-types/skin-melanoma/prognosis-and-survival **(Accessed: 14 January 2024).**

RStudio Team. “RStudio | Open Source & Professional Software for Data Science Teams.” Rstudio.com, 2020, www.rstudio.com/ (Accessed Jan. 15th 2024)

Schwartzman RM, Orkin M (1962). A Comparative Study of Diseases of Dog and Man. Springfield, IL: Thomas. p. 85. The term 'melanoma' in human medicine indicates a malignant growth; the prefix 'malignant' is redundant.

Wickham, Hadley, et al. “Welcome to the Tidyverse.” Journal of Open Source Software, vol. 4, no. 43, 21 Nov. 2019, p. 1686, joss.theoj.org/papers/10.21105/joss.01686, https://doi.org/10.21105/joss.01686. Accessed 14 Jan. 2024.

Vos T, Allen C, Arora M, Barber RM, Bhutta ZA, Brown A, et al. (GBD 2015 Mortality Causes of Death Collaborators) (October 2016). "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015". Lancet. 388 (10053): 1459–1544. doi:10.1016/s0140-6736(16)31012-1. PMC 5388903. PMID 27733281.