EXPLORATORY DATA ANALYSIS OF THE ‘SURVIVAL FROM MALIGNANT MELANOMA’ DATASET FROM KAGGLE USING R .

Samjeh Emmanuel

Introduction

Melanoma is a type of skin cancer that can appear on any part of the body but mostly in areas exposed to the sun. The dataset explored in this analysis is of patients diagnosed with malignant melanoma and had their tumours removed in the department of plastic surgery in University Hospital of Odense in a period of 15 years (between 1962 and 1977). Patients with thick/ulcerated tumours had an increased chance of dying from this cancer.

The dataset consist of 7 variables to include time, status,sex,age,year, thickness and ulcer. The below work will be split into 7 sections, including a discussion at the end, to explore different aspects of the dataset.

A. ***Summary statistics of the dataset***.

A total number of 208 patients were operated on in this hospital in the stated time period. Seven parameters in the columns in this data frame will now be looked at.

A.1 Time

Mean= 2152.8 days. This is the average number of days patients survived after they had their operation done.

Median=2005 days. This indicates the middle point in this set of data. About half the patients lived x number of days above this value and another half of the patients lived x number of days below this value.
Standard Deviation = 1122.061 indicating that the data points do not converge close to the mean, the values of most of the data points are different from the mean value.

A.2 Status
Mean= 1.790244. This value indicates that approximately more patients were still alive after their operation during the period of observation.

A.3 Sex (1=Male, 0=Female)
Summary of 5 numbers=

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.0000 0.0000  0.3854 1.0000   1.0000
```

Looking at the above we see that more females were included in this study than men.
Also, if we can get the mode if we use the table function, we get the following results:

```
  0   1
126  79
```

This confirms the number of women in the study are more than the men.

A.4 Age
Mean= 52.46341. This means the average age of the participants was about 52 years.
Standard Deviation= 16.67171. The dispersion of the ages is generally not far from the mean age of this population.

A.5 Year of Operation
Mode= we can get this by using the table function again(see appendix 1) which gives us the following:

```
1962 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1977
   1    1      11   10      20      21   21   19       27    41      31    1    1
```

We can clearly see from the above that 1972 was the year that had the most operations done.

A.6 Tumour Thickness

Mean= 2.919854. This represents the average thickness of the tumours.
Mode= 1.29mm. This means more patients had a tumour of thickness 1.29mm than any other, in this case 16 patients.

A.7 Ulcer (1=present,0 =absent)

Mode= 0 1
115 90

Ulceration was present in 90 patients compared to it being absent in 115 patients.

(*The codes for all the above operations are present in appendix 1*)

B. Graphical summaries

Using a history to graphically represent the data, we can draw the following insights from looking at the graphs.

B.1 We can see from the graph representing 'Time' that the maximum follow up time for most patients was about 2000 days and the least number of days of observation was about 5000 days.

B.2 The graph on 'Status' shows that most patients were alive by the end of the study (2.0), followed by those who had died before the study ended (1.0) and those who had died from other causes being the least (3.0).

B.3 The graph in the case of 'Sex' showed that the majority of participants in this study were female (0.0).

B.4 The average age of the participants was between 50 and 60 years with the majority of the participants being younger than this average.

B.5 The graph on the 'Years' indicates that the majority of operations were sometime around 1972 with the least number of operations happening at the beginning of the studies.

B.6 The most frequently occurring thickness of tumour in the participants was just above 2.5 mm but below 3 mm with the least occurring being about 14mm.

B.7 Ulceration was absent from the majority participants as can be seen on the graph.

(*graphs and codes can be seen on appendix 2*)

C. Correlation computations and Regression analysis

C.1.1 Correlation between time and thickness= -0.2354087. This is a negative correlation which means as one variable increases in one direction, the other variable goes in the opposite direction. As the time of follow up after operation increased, the thickness of the tumour reduced. (See scatterplot of this in appendix 3.1).

C.1.2 Correlation between time and age= -0.3015179. This also indicates a negative correlation between the 2 variables meaning younger participants had a longer follow up period than older ones. Was this because older subjects died earlier before the end of the study?

C.1.3 Correlation between thickness and age= 0.2124798.
This indicates a weak positive correlation between thickness and age, meaning the thickness of the tumours increased with age.

The code for the above is:

```
     # CORRELATION ANALYSIS
> attach(Melanoma)
> cor(time,thickness,method = 'pearson')
[1] -0.2354087
> plot(time,thickness,main='scatterplot')
> plot(time,thickness,main='scatterplot')
> cor(time,age,method='pearson')
[1] -0.3015179
> cor(thickness,age,method='pearson')
[1] 0.2124798
```

```
> plot(age,thickness,main='scatterplot 3.2')
```

Regression Analysis

C.2.1

```
#regression analysis
> model1=lm(formula = time~thickness)
> model1

Call:
lm(formula = time ~ thickness)

Coefficients:
(Intercept)    thickness
    2413.41       -89.25

> #this will give the regression equation of our model as y=-89x+2413.41. y
here is the thickness and x is time.
> model2=lm(formula = time~age)
> model2

Call:
lm(formula = time ~ age)

Coefficients:
(Intercept)          age
    3217.45       -20.29

> #regression equation here is y=-20x+3217.45. y here is time and x age( the
independent variable)
> model3=lm(formula = thickness~age)
> model3

Call:
lm(formula = thickness ~ age)

Coefficients:
(Intercept)          age
    0.94105      0.03772

> #regression formula y=0.03772x+0.94105. y=thickness and x=age
> summary(model1)

Call:
lm(formula = time ~ 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 ***
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

> summary(model2)

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

> summary(model3)

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 ‘ ’ 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
```

D. Looking at the above variables, we can see that thickness of the tumour has a negative correlation to time. This may indicate that following the operation, the tumours reduced with time such that the patients that stayed on the study longest saw a reduction in the tumour thickness recorded, while those who dropped out earlier (died from the disease or from some other cause) saw a thicker tumour size recorded. This could indicate the treatment worked as healing post-operation happened with time.

   Also, time and age have a negative correlation meaning study subjects who were younger stayed longer on the study than those who were older. More of the older subjects possibly died before the end of the study.

   Finally, thickness of tumour and age have a positive correlation, meaning the older the study subject was, the thicker their tumour equally was.

E. TWO SAMPLE SIGNIFICANCE TESTS
   R codes for these calculations are on Appendix 4.

F. QQ-PLOTS VARIABLES TIME, THICKNESS, AND AGE GROUPED BY GENDER.

   (SEE APPENDIX 5 FOR PLOTS AND CODE)

The qqplots confirm the earlier analysis of the data more broadly but as pertaining to the 3 variables in question, we see that most male patients had about the same follow up time post-surgery (from analysis with the sample population). As time went by, the tumour thickness in women reduced, which is the same observation that can be made for male tumour thickness.

G. **Discussion**

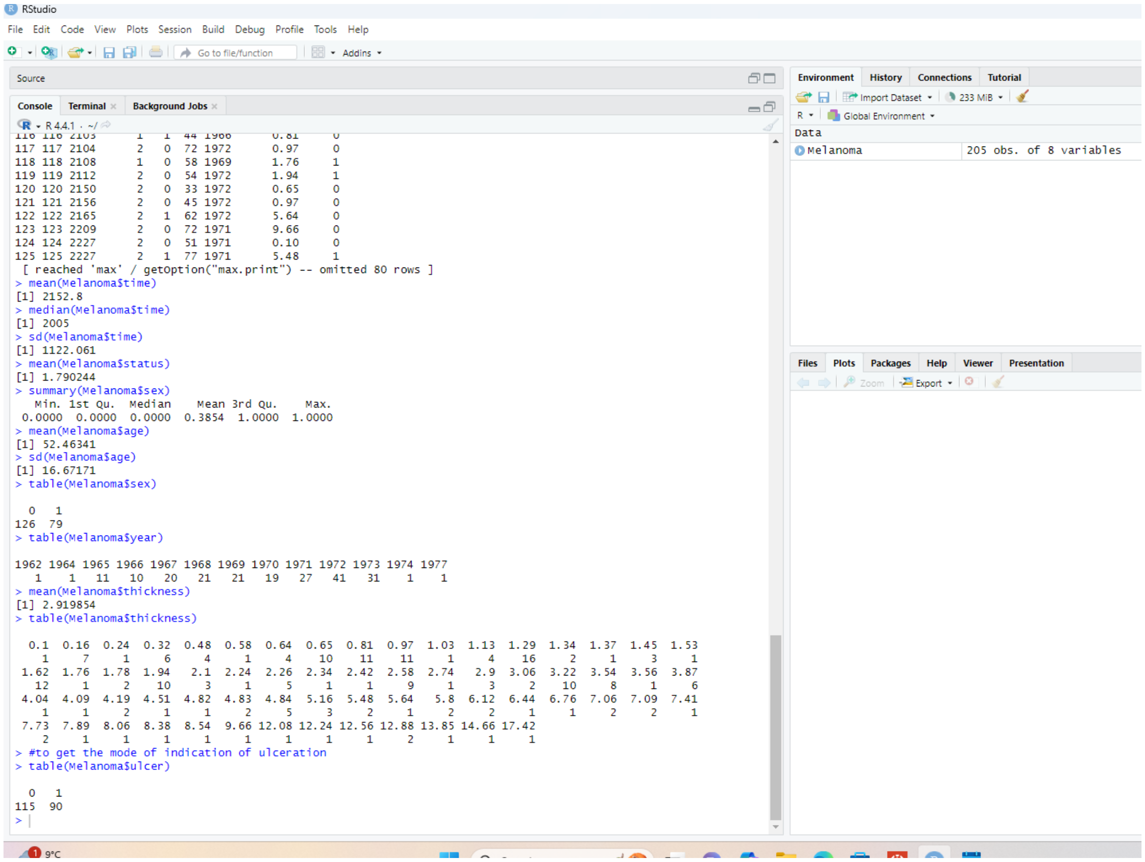
The dataset analysed shows that more women participated in this study than men. This did not influence the overall analysis which mainly showed that post surgery, the thickness of the tumour reduced meaning the attempted treatment worked for most of patients followed up to the end of the study. Does Melanoma affect more women than men? Is there a predisposing factor in women? Or are men just less likely to come forward for such studies? Some more research can be done to answer this.

More patients had an operation in 1972 than in any other year during this study. Was this just a coincidence or was something causing more skin cancers of Melanoma type in that particular year? Or did the news of the study being carried out attract more patients that particular year to join, if so, why did the numbers then drop the very next year? Another research to clarify this may be helpful.

It is also observed from the data that the ages of participants tend to reduce as we approach the end of the study. This could mean older patients did not survive long after the operation. That said, it can equally be observed that a good number of patients, mainly those above 50 died from melanoma compared to those that died from other causes before the end of the study.

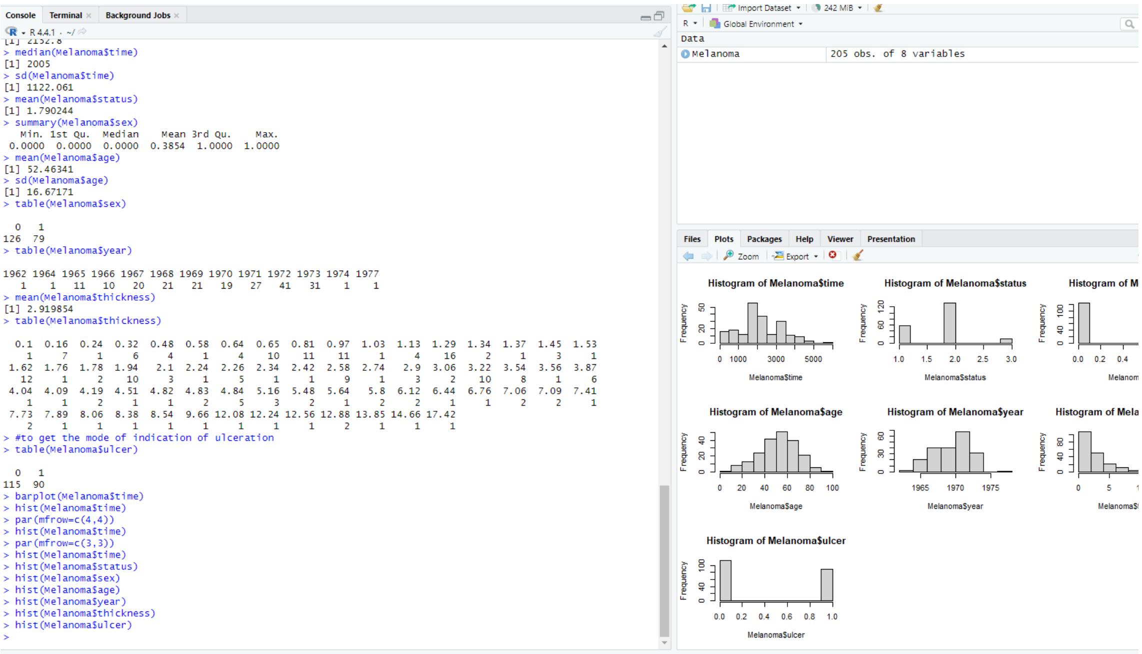
Given the study ends after a number of years after surgery, is there any way of knowing if this treatment made a long-term difference for the study subjects? Was any other treatment required to ensure the patients were completely cancer free or was the operation on its own effective.

According to the NHS, surgery still remains the main form of treatment for melanoma with chemotherapy, radiotherapy and medications used sometimes.



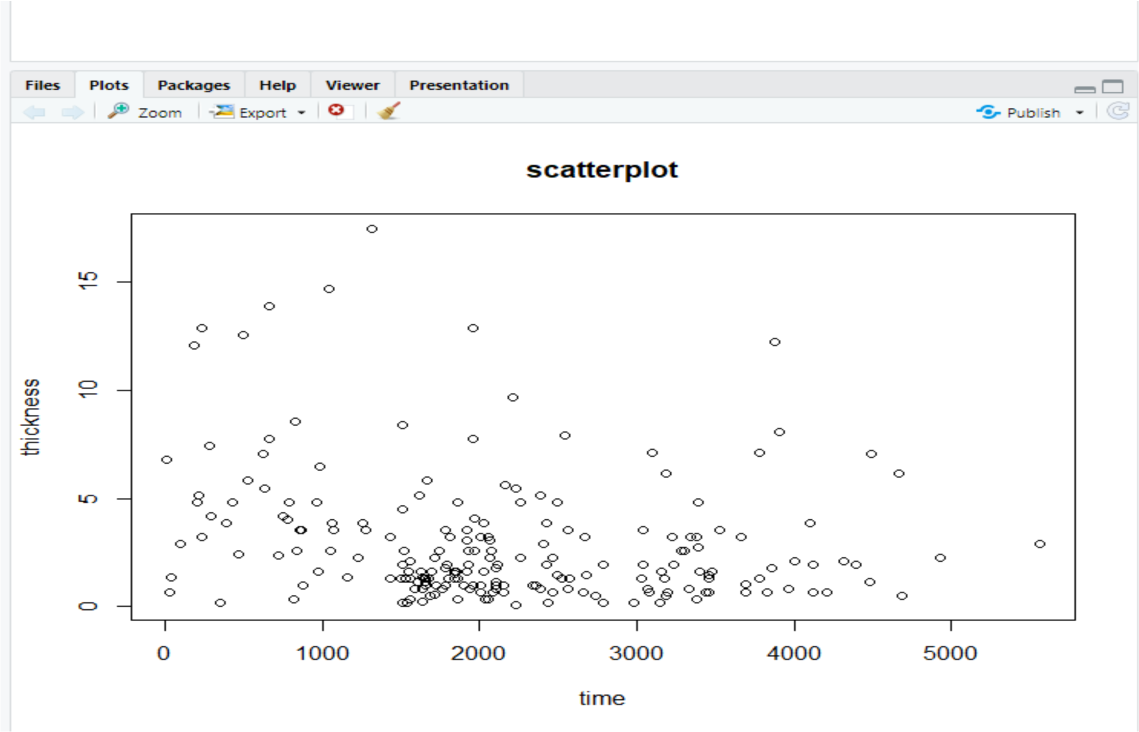
```
116 116 2103       1   1  44 1966       0.81      0
117 117 2104       2   0  72 1972       0.97      0
118 118 2108       1   0  58 1969       1.76      1
119 119 2112       2   0  54 1972       1.94      1
120 120 2150       2   0  33 1972       0.65      0
121 121 2156       2   0  45 1972       0.97      0
122 122 2165       2   1  62 1972       5.64      0
123 123 2209       2   0  72 1971       9.66      0
124 124 2227       2   0  51 1971       0.10      0
125 125 2227       2   1  77 1971       5.48      1
 [ reached 'max' / getOption("max.print") -- omitted 80 rows ]
> mean(Melanoma$time)
[1] 2152.8
> median(Melanoma$time)
[1] 2005
> sd(Melanoma$time)
[1] 1122.061
> mean(Melanoma$status)
[1] 1.790244
> summary(Melanoma$sex)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.3854  1.0000  1.0000 
> mean(Melanoma$age)
[1] 52.46341
> sd(Melanoma$age)
[1] 16.67171
> table(Melanoma$sex)

  0   1 
126  79 
> table(Melanoma$year)

1962 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1977 
   1    1   11   10   20   21   21   19   27   41   31    1    1 
> mean(Melanoma$thickness)
[1] 2.919854
> table(Melanoma$thickness)

  0.1  0.16  0.24  0.32  0.48  0.58  0.64  0.65  0.81  0.97  1.03  1.13  1.29  1.34  1.37  1.45  1.53 
    1     7     1     6     4     1     4    10    11    11     1     4    16     2     1     3     1 
 1.62  1.76  1.78  1.94   2.1  2.24  2.26  2.34  2.42  2.58  2.74   2.9  3.06  3.22  3.54  3.56  3.87 
   12     1     2    10     3     1     5     1     1     9     1     3     2    10     8     1     6 
 4.04  4.09  4.19  4.51  4.82  4.83  4.84  5.16  5.48  5.64   5.8  6.12  6.44  6.76  7.06  7.09  7.41 
    1     1     2     1     1     2     5     3     2     1     2     2     1     1     2     2     1 
 7.73  7.89  8.06  8.38  8.54  9.66 12.08 12.24 12.56 12.88 13.85 14.66 17.42 
    2     1     1     1     1     1     1     1     1     2     1     1     1 
> #to get the mode of indication of ulceration
> table(Melanoma$ulcer)

  0   1 
115  90 
> 
```

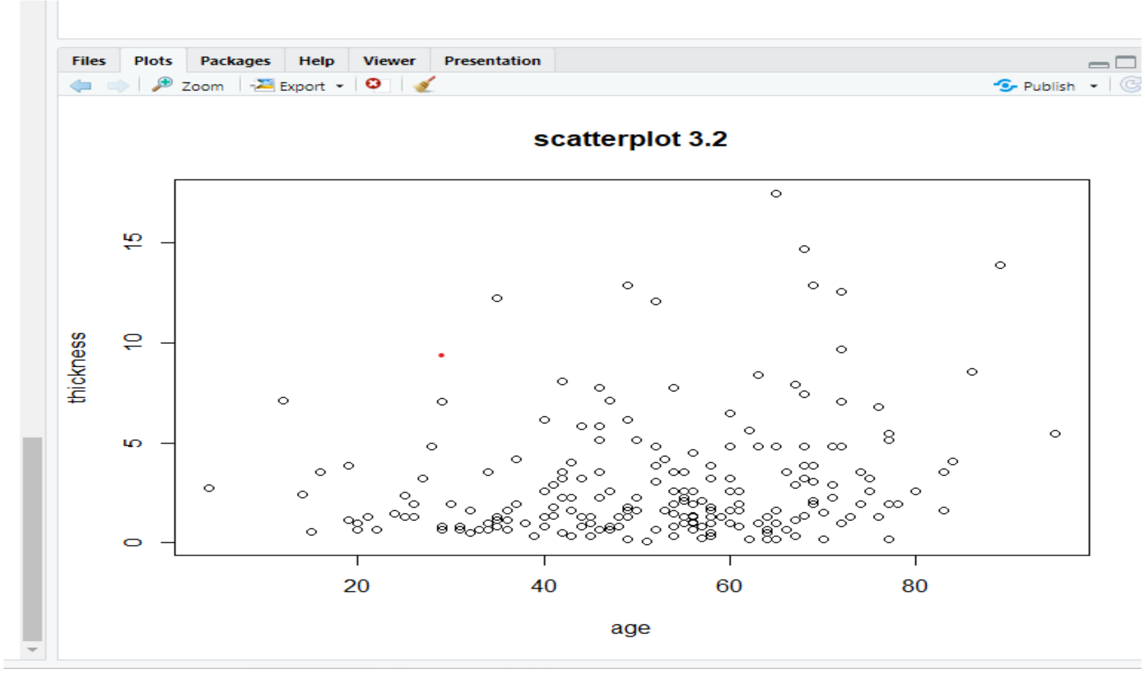
```
Data
Melanoma          205 obs. of 8 variables
```



```
[1] 2152.8
> median(Melanoma$time)
[1] 2005
> sd(Melanoma$time)
[1] 1122.061
> mean(Melanoma$status)
[1] 1.790244
> summary(Melanoma$sex)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.3854  1.0000  1.0000 
> mean(Melanoma$age)
[1] 52.46341
> sd(Melanoma$age)
[1] 16.67171
> table(Melanoma$sex)

  0   1 
126  79 
> table(Melanoma$year)

1962 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1977 
   1    1   11   10   20   21   21   19   27   41   31    1    1 
> mean(Melanoma$thickness)
[1] 2.919854
> table(Melanoma$thickness)

  0.1  0.16  0.24  0.32  0.48  0.58  0.64  0.65  0.81  0.97  1.03  1.13  1.29  1.34  1.37  1.45  1.53 
    1     7     1     6     4     1     4    10    11    11     1     4    16     2     1     3     1 
 1.62  1.76  1.78  1.94   2.1  2.24  2.26  2.34  2.42  2.58  2.74   2.9  3.06  3.22  3.54  3.56  3.87 
   12     1     2    10     3     1     5     1     1     9     1     3     2    10     8     1     6 
 4.04  4.09  4.19  4.51  4.82  4.83  4.84  5.16  5.48  5.64   5.8  6.12  6.44  6.76  7.06  7.09  7.41 
    1     1     2     1     1     2     5     3     2     1     2     2     1     1     2     2     1 
 7.73  7.89  8.06  8.38  8.54  9.66 12.08 12.24 12.56 12.88 13.85 14.66 17.42 
    2     1     1     1     1     1     1     1     1     2     1     1     1 
> #to get the mode of indication of ulceration
> table(Melanoma$ulcer)

  0   1 
115  90 
> barplot(Melanoma$time)
> hist(Melanoma$time)
> par(mfrow=c(4,4))
> hist(Melanoma$time)
> par(mfrow=c(3,3))
> hist(Melanoma$time)
> hist(Melanoma$status)
> hist(Melanoma$sex)
> hist(Melanoma$age)
> hist(Melanoma$year)
> hist(Melanoma$thickness)
> hist(Melanoma$ulcer)
> 
```

```
Data
Melanoma        205 obs. of 8 variables
```





TWO SAMPLE SIGNIFICANCE TESTS

```
library(tidyverse)
── Attaching core tidyverse packages
────────────────────────────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.5.1     ✓ tibble    3.2.1
✓ lubridate 1.9.3      ✓ tidyr     1.3.1
✓ purrr     1.0.2
── Conflicts
─────────────────────────────────────────────────────────────────────
tidyverse_conflicts() ──
✗ dplyr::filter() masks stats::filter()
✗ dplyr::lag()    masks stats::lag()
i Use the conflicted package to force all conflicts to become errors
> Melanoma<-as_tibble(Melanoma::Melanoma)
Error in loadNamespace(x) : there is no package called ‘Melanoma’
> Melanoma2<-as_tibble(Melanoma2)
> Melanoma2
# A tibble: 205 × 8
      X  time status   sex   age  year thickness ulcer
  <int> <int>  <int> <int> <int> <int>     <dbl> <int>
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
 7     7   210      1     1    77  1972      5.16     1
 8     8   232      3     0    60  1974      3.22     1
 9     9   232      1     1    49  1968     12.9      1
10    10   279      1     0    68  1971      7.41     1
# ℹ 195 more rows
# ℹ Use `print(n = ...)` to see more rows
> Melanoma2<-Melanoma2%>%mutate(sex=recode_factor(sex,'1='male','o'=female'))
Error: unexpected symbol in "Melanoma2<-
Melanoma2%>%mutate(sex=recode_factor(sex,'1='male"

> Melanoma2<-Melanoma2%>%mutate(sex=recode_factor(sex,'1'="male",'o'="female"))
> Melanoma2<-Melanoma2%>%mutate(sex=recode_factor(sex,'1'="male",'0'="female"))
> Melanoma2
# A tibble: 205 × 8
       X  time status sex     age  year thickness ulcer
   <int> <int>  <int> <fct> <int> <int>     <dbl> <int>
 1     1    10      3 male     76  1972      6.76     1
 2     2    30      3 male     56  1968      0.65     0
 3     3    35      2 male     41  1977      1.34     0
 4     4    99      3 NA       71  1968      2.9      0
 5     5   185      1 male     52  1965     12.1      1
 6     6   204      1 male     28  1971      4.84     1
 7     7   210      1 male     77  1972      5.16     1
 8     8   232      3 NA       60  1974      3.22     1
 9     9   232      1 male     49  1968     12.9      1
10    10   279      1 NA       68  1971      7.41     1
```

```
library(ggplot2)
> head(Melanoma2)
# A tibble: 6 × 8
      X  time status sex     age  year thickness ulce  <int> <int>  <int> <fct>
<int> <int>  <dbl> <int>
1     1    10      3 male     76  1972      6.76     1
2     2    30      3 male     56  1968      0.65     0
3     3    35      2 male     41  1977      1.34     0
4     4    99      3 NA       71  1968      2.9      0
5     5   185      1 male     52  1965     12.1      1
6     6   204      1 male     28  1971      4.84     1


> #using only sample from male participants


> head(Melanoma2,n=10)


# A tibble: 10 × 8
   X  time status   sex     age     year     thickness      ulcer

<int> <int>  <int> <fct> <int> <int>     <dbl> <int>
 1     1    10      3 male     76  1972      6.76     1
 2     2    30       3 male     56  1968      0.65     0
 3     3    35       2 male     41  1977      1.34     0
 4     4    99       3 NA       71  1968      2.9      0
 5     5   185       1 male     52  1965     12.1      1
 6     6   204       1 male     28  1971      4.84     1
 7     7   210       1 male     77  1972      5.16     1
 8     8   232       3 NA       60  1974      3.22     1
 9     9   232       1 male     49  1968     12.9      1
10    10   279       1 NA       68  1971      7.41     1

> #we have a sample size of 10. Null hypothesis is mean of time=mean of
thickness. Alternative hypothesis= mean

> of time is not = mean of thickness
Error: unexpected symbol in "of time"
> #of time is not = mean of thickness
> time<-c(10,30,35,185,204,210,232)
> thickness<-(6.76,0.65,1.34,12.1,4.84,5.16,12.9)
Error: unexpected ',' in "thickness<-(6.76,"
> thickness<-c(6.76,0.65,1.34,12.1,4.84,5.16,12.9)
> t.test(x=time,y=thicness)
Error: object 'thicness' not found
> t.test(x=time,y=thickness)

      Welch Two Sample t-test

data:  time and thickness
t = 3.2903, df = 6.0281, p-value = 0.0165
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  31.67616 214.68099
sample estimates:
mean of x mean of y
 129.4286    6.2500

> #with the p-value = 0.0165, which is below 0.05, we can reject the null
hypothesis with regards to the make participants.

> Null hypothesis is mean of time=mean of age. Alternative hypothesis= mean
Error: unexpected symbol in "Null hypothesis"

> #Null hypothesis is mean of time=mean of age. Alternative hypothesis is mean of
time is not equal to mean of #age.

> time<-c(10,30,35,185,204,210,232)
> age<-c(76,56,41,52,28,77,48)
> t.test(x=time,y=age)

      Welch Two Sample t-test

data:  time and age
t = 1.9853, df = 6.3882, p-value = 0.09143
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.18751 167.04465
sample estimates:
mean of x mean of y 
 129.4286   54.0000 


> #because the p-value is the least (0.05) value to reject the null hypothesis, 
in this case we cannot reject
> #HO for the male population.
> #
> #Null hypothesis is mean of age=mean of thickness. Alternative hypothesis is 
mean of age is not equal to mean of thickness in the male subjects of this study.


> age<-c(76,56,41,52,28,77,48)
> thickness<-c(6.76,0.65,1.34,12.1,4.84,5.16,12.9)
> t.test(x=age,y=thickness)

      Welch Two Sample t-test

data:  age and thickness
t = 6.8525, df = 6.8621, p-value = 0.000264
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 31.20542 64.29458
sample estimates:
mean of x mean of y 
    54.00      6.25 

> #with the p value being 0.000264, we can reject the null hypothesis of this 
sample.
```

```
 > #to do the 2 sample significance test for the female sample, we will use 
 a sample size of 20.
 > head(Melanoma2,n=20)
 # A tibble: 20 × 8
        X  time status sex     age  year thickness ulce   <int> <int>  <int>
 <fct> <int> <int>     <dbl> <int>                                      
  1      1    10      3 male     76  1972      6.76     1
  2      2    30      3 male     56  1968      0.65     0
  3      3    35      2 male     41  1977      1.34     0
  4      4    99      3 NA       71  1968      2.9      0
  5      5   185      1 male     52  1965     12.1      1
  6      6   204      1 male     28  1971      4.84     1
  7      7   210      1 male     77  1972      5.16     1
  8      8   232      3 NA       60  1974      3.22     1
  9      9   232      1 male     49  1968     12.9      1
 10     10   279      1 NA       68  1971      7.41     1
 11     11   295      1 NA       53  1969      4.19     1
 12     12   355      3 NA       64  1972      0.16     1
 13     13   386      1 NA       68  1965      3.87     1
```

```
14    14   426       1 male     63  1970       4.84      1
15    15   469       1 NA       14  1969       2.42      1
16    16   493       3 male     72  1971      12.6       1
17    17   529       1 male     46  1971       5.8       1
18    18   621       1 male     72  1972       7.06      1
19    19   629       1 male     95  1968       5.48      1
20    20   659       1 male     54  1972       7.73      1

> #for the female subjects of this study, with NA=FEMALE on our Melanoma2
dataset
> #null hypothesis is mean of time=mean of thickness, alternative
hypothesis is mean of time is not equal thicknes>

> tail(Melanoma2,N=10)
# A tibble: 6 × 8
      X  time status sex     age  year thickness ulcer
  <int> <int>  <int> <fct> <int> <int>     <dbl> <int>
1   200  4479      2 NA       19  1965      1.13     1
2   201  4492      2 male     29  1965      7.06     1
3   202  4668      2 NA       40  1965      6.12     0
4   203  4688      2 NA       42  1965      0.48     0
5   204  4926      2 NA       50  1964      2.26     0
6   205  5565      2 NA       41  1962      2.9      0

> time<-c(4479,4668,4688,4929,5565)
> thickness<-(1.13,6.12,0.48,2.26,2.9)
Error: unexpected ',' in "thickness<-(1.13,"
> thickness<-c(1.13,6.12,0.48,2.26,2.9)
> t.test(x=time,y=thickness)

      Welch Two Sample t-test

data:  time and thickness
t = 25.753, df = 4.0002, p-value = 1.35e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4338.916 5387.528
sample estimates:
mean of x mean of y
 4865.800     2.578

> #with the above p-value, we cannot reject the null hypothesis
> #H0 is mean of time is not equal to mean of age and H1 is mean of time =
mean age
> time<-c(4479,4668,4688,4929,5565)
> age<-c(19,40,42,50,41)
> t.test(x=time,y=age)

      Welch Two Sample t-test

data:  time and age
t = 25.554, df = 4.006, p-value = 1.375e-05
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4303.203 5351.597
sample estimates:
mean of x mean of y
   4865.8      38.4
```

```
> #p-value is greater than 0.05, therefore we cannot reject the null
hypothesis.
> #H0 is mean of thickness = mean of age. H1 is mean of thickness is not
equal to mean of age.
> age<-c(19,40,42,50,41)
> thickness<-c(1.13,6.12,0.48,2.26,2.9)
> t.test(x=age,y=thickness)

      Welch Two Sample t-test

data:  age and thickness
t = 6.8158, df = 4.2884, p-value = 0.001874
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 21.60885 50.03515
sample estimates:
mean of x mean of y
   38.400     2.578
```

```
> #we can reject the null hypothesis in this case as the p-value=0.001874,
so H1 holds in this situation.
```

```
> library(ggplot2)
>
> par(mfrow=c(2,2))
> library(tidyverse)
```

```
> melanoma<-as_tibble(melanoma)
> melanoma<-melanoma%>%mutate(sex=recode_factor(sex,'1'=male,'0'=female))

melanoma<-melanoma%>%mutate(sex=recode_factor(sex,'1'='male','0'='female'))
> head(melanoma,n=10)
# A tibble: 10 × 8
       X  time status sex      age  year thickness ulcer
   <int> <int>  <int> <fct>  <int> <int>     <dbl> <int>
 1     1    10      3 male      76  1972      6.76     1
 2     2    30      3 male      56  1968      0.65     0
 3     3    35      2 male      41  1977      1.34     0
 4     4    99       3 female    71  1968      2.9       0
 5     5   185       1 male      52  1965     12.1       1
 6     6   204       1 male      28  1971      4.84      1
 7     7   210       1 male      77  1972      5.16      1
 8     8   232       3 female    60  1974      3.22      1
 9     9   232       1 male      49  1968     12.9       1
10    10   279       1 female    68  1971      7.41      1
> maleTime<-c(10,30,35,185,204,210,232)
> maleThickness<-c(6.76,0.65,1.34,12.1,4.84,5.16,12.9)
> maleThickness
[1]  6.76  0.65  1.34 12.10  4.84  5.16 12.90
> maleTime
[1]  10  30  35 185 204 210 232
```

```
tail(melanoma)
# A tibble: 6 × 8
      X  time status sex      age  year thickness ulcer
  <int> <int>  <int> <fct>  <int> <int>     <dbl> <int>
1   200  4479      2 female    19  1965      1.13     1
2   201  4492      2 male      29  1965      7.06     1
3   202  4668      2 female    40  1965      6.12     0
4   203  4688      2 female    42  1965      0.48     0
5   204  4926      2 female    50  1964      2.26     0
6   205  5565      2 female    41  1962      2.9      0
> FemaleAge<-c(19,40,42,50,41)
> FemaleThickness<-c(1.13,6.12,0.48,2.26,2.9)
```

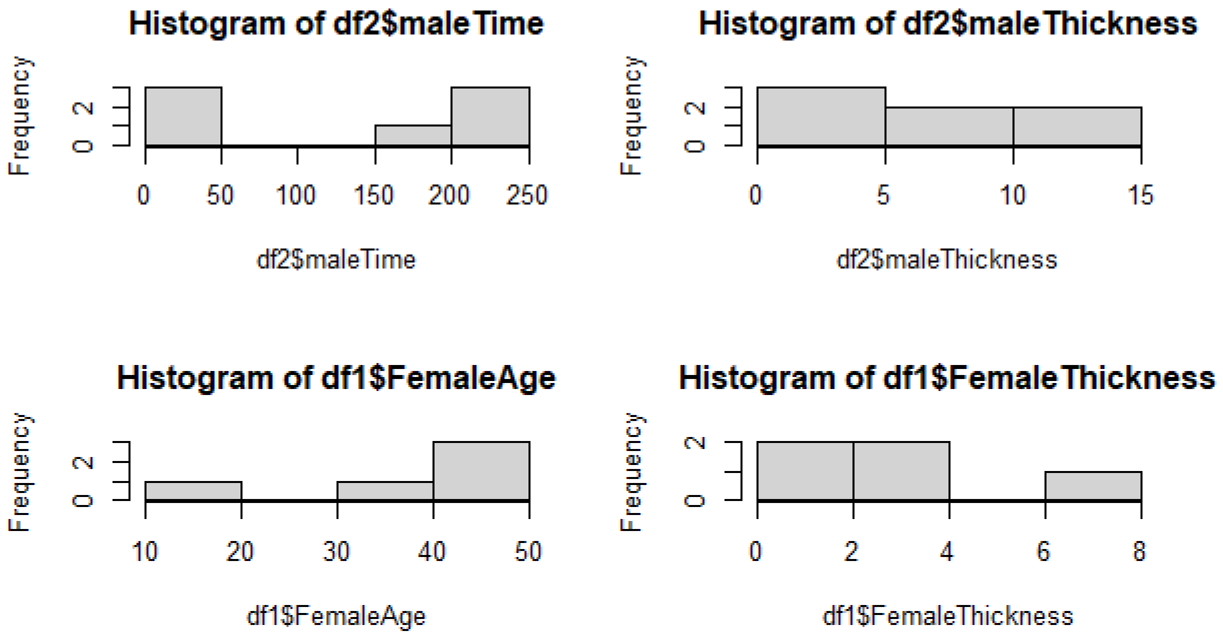
```
df1<-data.frame(FemaleAge,FemaleThickness)
```

```
> df1
```

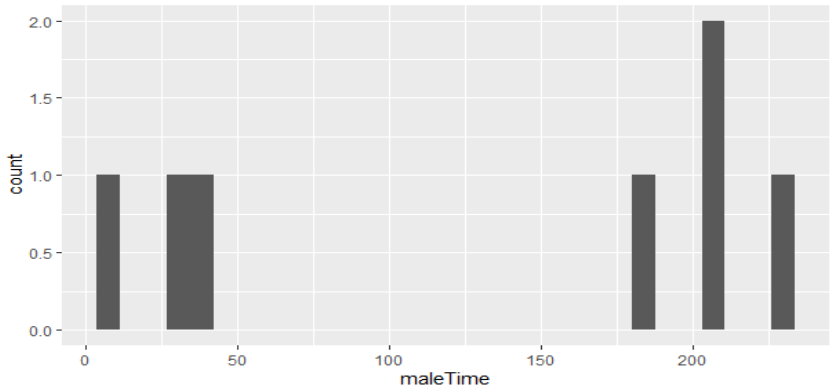
```
  FemaleAge FemaleThickness
1        19            1.13
2        40            6.12
3        42            0.48
4        50            2.26
5        41            2.90
```

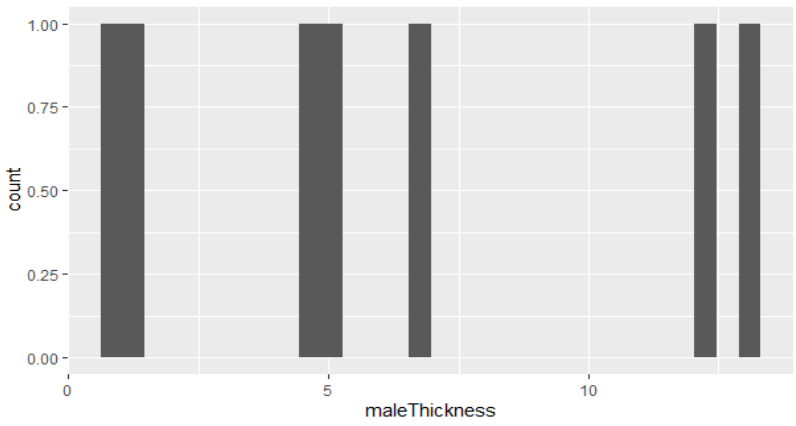
```
df2<-data.frame(maleTime,maleThickness)
> df2
  maleTime maleThickness
1       10          6.76
2       30          0.65
3       35          1.34
4      185         12.10
5      204          4.84
6      210          5.16
7      232         12.90
```

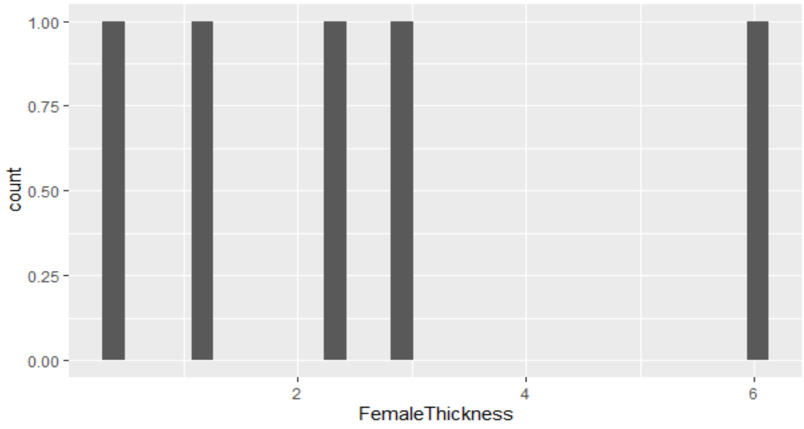
```
hist(df2$maleThickness)
> hist(df1$FemaleAge)
> hist(df1$FemaleThickness)
> ggplot(df1$FemaleAge)
```


Histogram of df2$maleTime
Frequency
0 2
0 50 100 150 200 250
df2$maleTime
Histogram of df2$maleThickness
Frequency
0 2
0 5 10 15
df2$maleThickness
Histogram of df1$FemaleAge
Frequency
0 2
10 20 30 40 50
df1$FemaleAge
Histogram of df1$FemaleThickness
Frequency
0 2
0 2 4 6 8
df1$FemaleThickness

GGPLOTS


2.0
1.5
1.0
0.5
0.0
count
0 50 100 150 200
maleTime





Codes for the above histograms

```
  ggplot(data=df1)+geom_histogram(mapping = aes(x=FemaleThickness))
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> ggplot(data=df2)+geom_histogram(mapping = aes(x=maleThickness))
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> ggplot(data=df2)+geom_histogram(mapping = aes(x=maleTime))
```