Melanoma Dataset

The dataset:

- 3 quantitative variables (**Time**, **age**, **thickness**)
- 3 categorical variables (**status**, **sex**, **ulcer**), The **year** is a continuous interval variable and for the purpose of this dataset will be assigned as **neutral** (quantitative and categorical)

Statistical summary

Summaries are used to summarize a data frame, a way of deriving statistical measures of our data.

- The statistical variables were chosen from the dataset **(Time**, **age** and **thickness**)
- The chosen statistical variables are all quantitative.

	var	Minimum	Q1	Median	Q3	Maximum	Mean	SD
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	age	4	42	54	65	95	52.5	16.7
2	thickness	0.1	0.97	1.94	3.56	17.4	2.92	2.96
3	time	10	1525	2005	3042	5565	2153.	1122.

The above image denotes the following:

1. Youngest person at time of operation is 4**(age)** and oldest is 95**(age)** while the average age of a person at time of operation is 52.
2. Largest tumour has a size of 17.4mm and the smallest tumour a size of 0.1mm
3. On average, a person lived for 2153 days since operation day. The least number of days lived by a person since operation is 10 and the highest is 5565.

The variable *year* from our dataset will be illustrated with both *graphical and statistical summaries* for a better illustration.

Graphical summary

The above code prints out a boxplot and a scatter plot for the **year** in our data frame. A statistical summary is also presented for context:

```
> summary(rio_csv$year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1962    1968    1970    1970    1972    1977
```

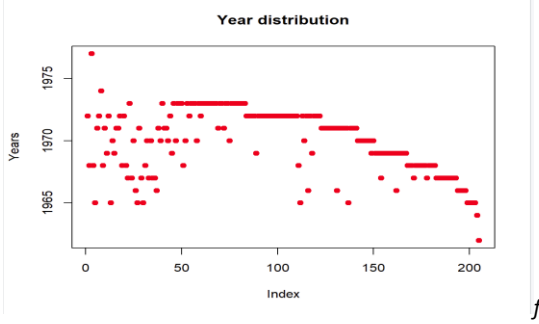


fig 1

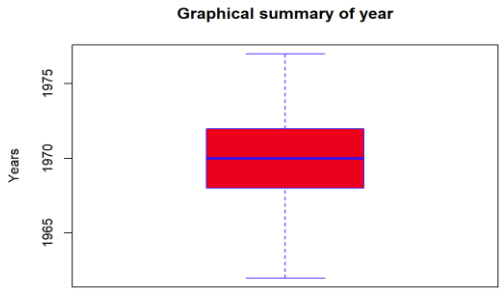


fig 1.1

The above images has the following deductions, from the *year distribution and the graphical summary of year;*

- Most of the operations performed were between 1967 to 1973 (see *fig 1 and fig 2*)
- From *fig 1 and 2*, one operation was performed in 1962 which was also the earliest year while 1 operation in 1977 which was the latest year of an operation both being outliers.

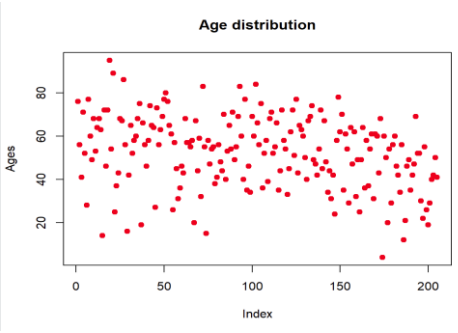


fig 2

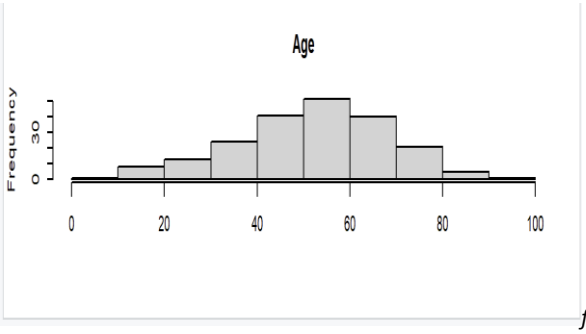


fig 2.1

From the above image, we can conclude that;

- Ages 40 to 70 had the highest number of operations as shown in *fig 2.1*

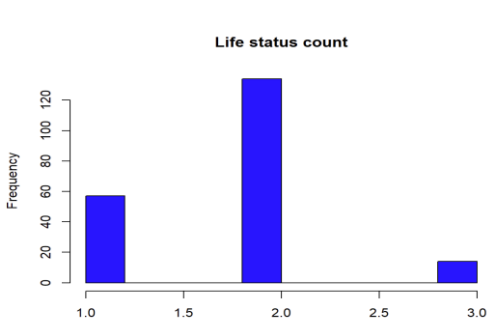


fig 3

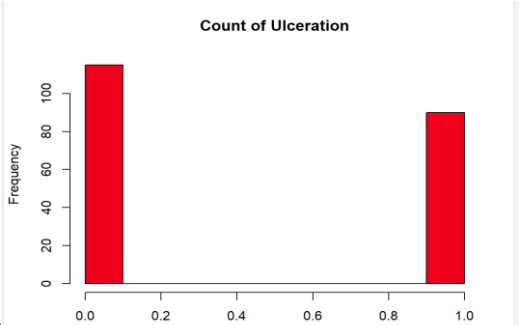


fig 3.1

From the *Status* data in our dataset, we know that 1 indicates that the patient died from melanoma, 2 indicates that they were still alive and 3 indicates that they had died from causes unrelated to their melanoma. Hence, we can deduct the following;

- *Fig 3* shows more than 55 people and less 60 have died from Melanoma since operation.
- More than 120 people are alive since operation.

In terms of *ulceration*, our dataset Indicates 1=present, 0=absent. Hence, the *fig 3.1* shows the following:

- More than 80 people but less than 90 have skin ulcer while over 100 people do not

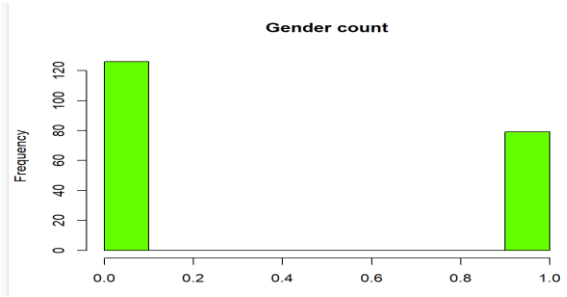
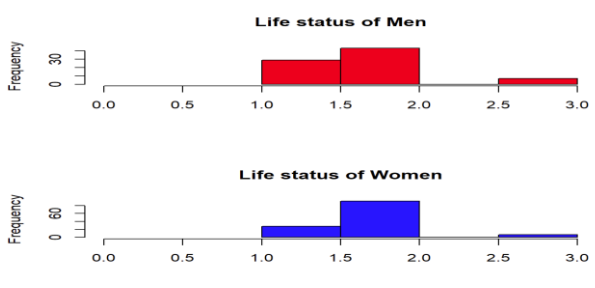


fig 4

```
63
64
65  hist(rio_csv$status,
66       main = "Life status count",
67       col = "blue",
68       xlab="")
69  hist(rio_csv$ulcer,
70       main = "Count of Ulceration",
71       col = "red",
72       xlab="")
73  hist(rio_csv$sex,
74       main = "Gender count",
75       col = "green",
76       xlab="")
77  # Histogram for categorical variables
78
```

From our dataset, we established the patients sex 1=male, 0=female.

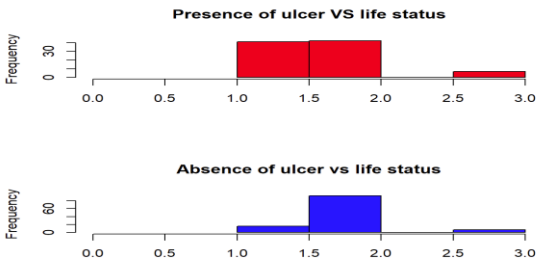
- There are significantly more female patients than there are male patients (see *fig 4*)




```
80  par(mfrow = c(2,1)) # combining two histograms in one plot
81  hist(rio_csv$status [rio_csv$sex == 1],
82       xlim= c(0,3),
83       breaks= 3,
84       main = "Life status of Men",
85       xlab = "",
86       col = "red")
87
88  hist(rio_csv$status [rio_csv$sex == 0],
89       xlim= c(0,3),
90       breaks= 3,
91       main = "Life status of Women",
92       xlab = "",
93       col = "blue")
```

fig 5

I. From the above image we can conclude that more men died from Melanoma than women
II. More female patients since operation are alive compared to their male counterparts

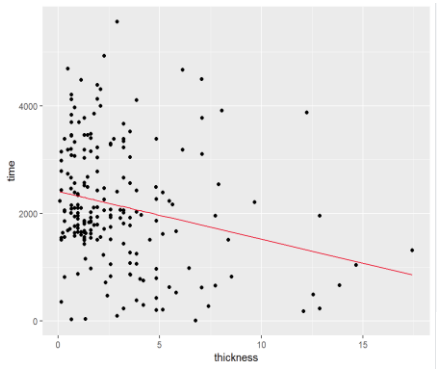



```
97  par(mfrow = c(2,1)) # combining two histograms in one plot
98  hist(rio_csv$status [rio_csv$ulcer == 1],
99       xlim= c(0,3),
100      breaks= 3,
101      main = "Presence of ulcer VS life status",
102      xlab = "",
103      col = "red")
104
105
106 hist(rio_csv$status [rio_csv$ulcer == 0],
107      xlim= c(0,3),
108      breaks= 3,
109      main = "Absence of ulcer vs life status",
110      xlab = "",
111      col = "blue")
```

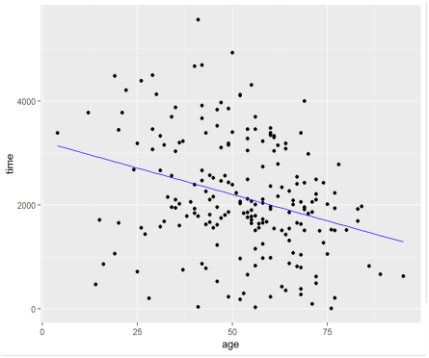
fig 6

The image above image shows there are higher death rates of patients with ulcer, over two times greater of patients without ulcer.

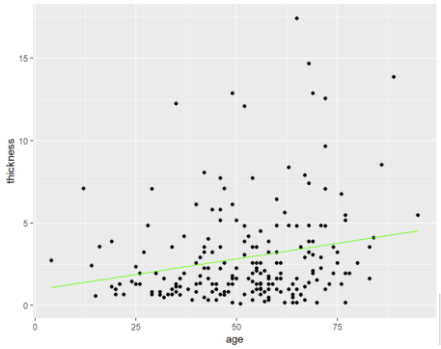
Regression




```
> cor(x,y)
[1] -0.2354087
```




```
> cor(x,y)
[1] -0.3015179
```




```
> cor(x,y)
[1] 0.2124798
```

1. **Time ~ thickness** has a negative regression line *–0.2354087* as thickness increases, time reduces.
2. **Time ~ age** also negative *-0.3015179.* As age increases, time reduces
3. **Thickness ~ age** being the only positive regression line and highest correlation of all 3-variable pair *0.2124798* The older a patient, the thicker the tumour.

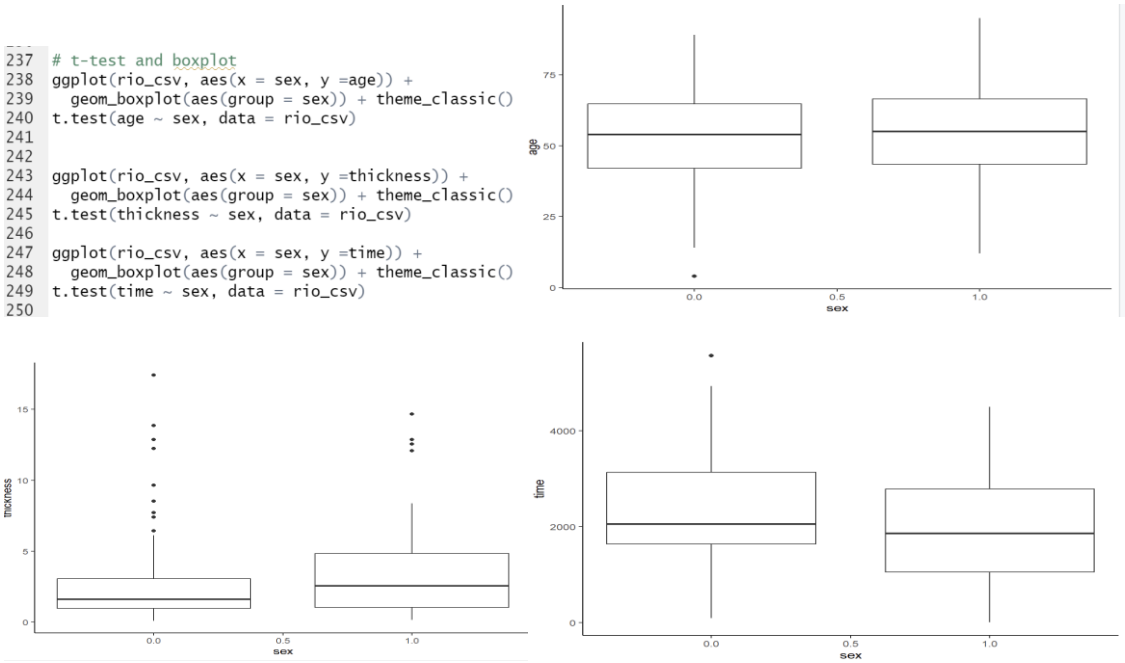
Two sample significance test – Grouped by gender(sex)

```r
# t-test and boxplot
ggplot(rio_csv, aes(x = sex, y =age)) +
  geom_boxplot(aes(group = sex)) + theme_classic()
t.test(age ~ sex, data = rio_csv)


ggplot(rio_csv, aes(x = sex, y =thickness)) +
  geom_boxplot(aes(group = sex)) + theme_classic()
t.test(thickness ~ sex, data = rio_csv)

ggplot(rio_csv, aes(x = sex, y =time)) +
  geom_boxplot(aes(group = sex)) + theme_classic()
t.test(time ~ sex, data = rio_csv)

```



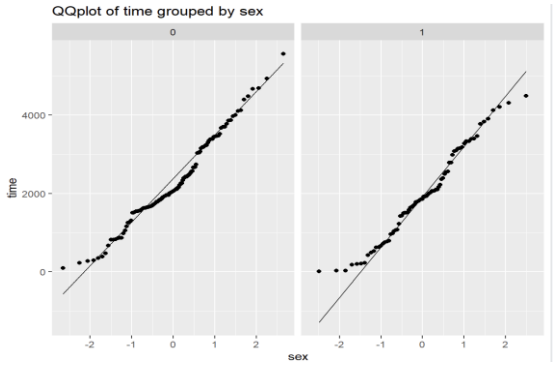
The above boxplots show us the outliers and the skewness of the variables grouped by gender

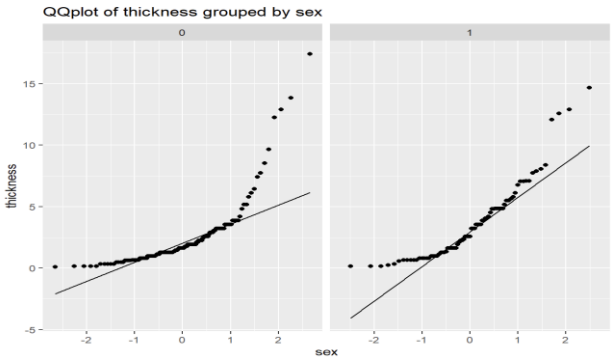
QQplots – Grouped by gender(sex)

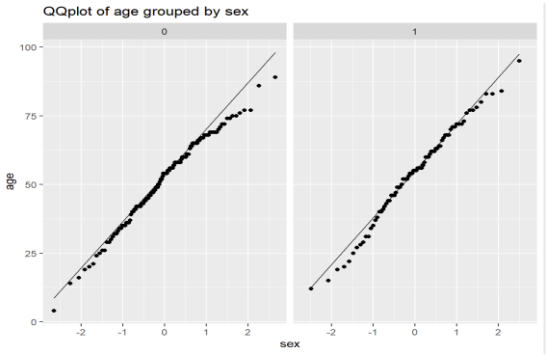
```r
#QQplots
ggplot(data = rio_csv, aes(sample = time)) +
  geom_qq() +
  stat_qq_line() +
  facet_wrap(~ sex) +
  labs(title = "QQplot of time grouped by sex",
       x = "sex",
       y = "time")

ggplot(data = rio_csv, aes(sample = thickness)) +
  geom_qq() +
  stat_qq_line() +
  facet_wrap(~ sex) +
  labs(title = "QQplot of thickness grouped by sex",
       x = "sex",
       y = "thickness")

ggplot(data = rio_csv, aes(sample = age)) +
  geom_qq() +
  stat_qq_line() +
  facet_wrap(~ sex) +
  labs(title = "QQplot of age grouped by sex",
       x = "sex",
       y = "age")
```







1. **Time** grouped by gender from the QQ plot above has a *thin tailed data distribution*
2. **Thickness** grouped by gender from the QQ plot above has a data distribution *skewed to the right*
3. **Age** grouped by gender from QQ plot above has a data distribution *skewed to the left*

Discussion

Fig 7

The above histogram shows the tumour thickness greater than the mean value and less than the mean value in patients. Patients with tumour thickness greater than 2.92 have significantly higher death rate than patients who have their tumour thickness below the mean value.

Conclusion

"These are thought to be important prognostic variables in that patient with a thick and/or ulcerated tumour have an increased chance of death from melanoma".

- Patients with an ulcerated tumour have an increased chance of death from melanoma (see *fig 6*)
- Patients with higher tumour thickness have an increased chance of death from melanoma (see *fig 7*)

Recommendation:

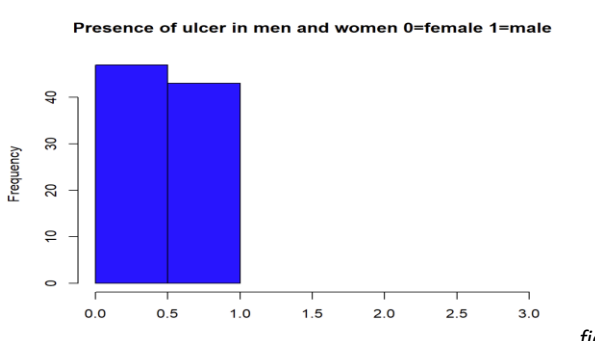
Patients who have both ulcerated tumour and thickness higher than mean value should be analysed.

Further discussion

We have established that;

1. There are almost twice as many women than men in the data set, from *fig 4*.
2. There's a higher number of deaths from Melanoma in men compared to women, *fig 5*
3. *Fig 6* shows presence of ulcer is a main symptom of death in Melanoma patients

The chart below shows higher presence of ulcer in female patients than in men.




```
278 hist(rio_csv$sex [rio_csv$ulcer == 1],
279     xlim= c(0,3),
280     breaks= 3,
281     main = paste("Presence of ulcer in men and women",
282                  "0=female 1=male"),
283     xlab = "",
284     col = "blue")
285
```

fig 8

If more female patients have ulcer and ulcer being a major attribute of melanoma leading to death, then perhaps female patients should have the higher death rate and not the male patients.