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.

```
df.sum <- df %>%
  select(time, age, thickness) %>% # Select specific quantitative variables to summarize
  summarise_each(funs(Minimum = min,
                      Q1 = quantile(., 0.25),
                      Median = median,
                      Q3 = quantile(., 0.75),
                      Maximum = max,
                      Mean = mean,
                      SD = sd))

df.stats.tidy <- df.sum %>% gather(stat, val) %>% # Reshaping using tidyr
  separate(stat, into = c("var", "stat"), sep = "_") %>%
  spread(stat, val) %>%
  select(var, Minimum, Q1, Median, Q3, Maximum, Mean, SD) # reorder columns

print(df.stats.tidy)
# Print statistical properties of variables
```

From the above image, we can establish some pattern

- 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.

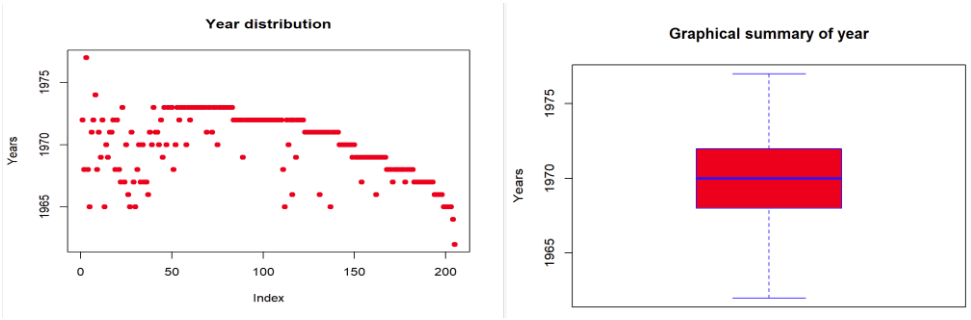
Graphical summary

```
boxplot(rio_csv$year,
        col="red", # Color of box
        border= "blue", # Color of borders around the box
        main= "Graphical summary of year",
        xlab= " ",
        ylab= "Years") # Y-axis label

plot(rio_csv$year,
     col="red",
     pch= 19, # Solid circles for points
     main= "Year distribution",
     xlab= "Index", # X-axis label
     ylab= "Years") # Y-axis label
```

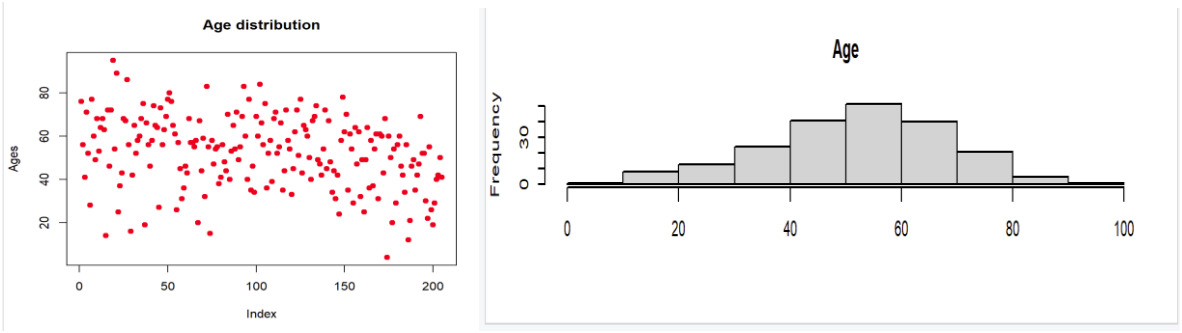
The above code prints out a boxplot and a scatter plot for the **year** in our data frame. A statistical summary is also presented for context:

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



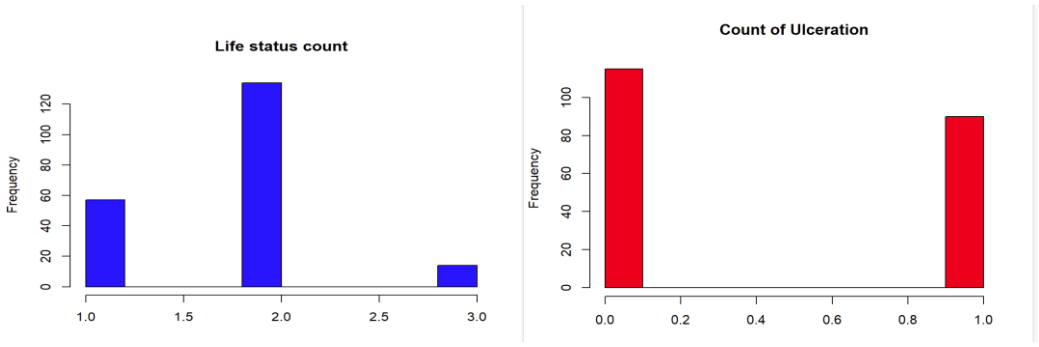
The above image has the following deductions, from the *year distribution and the graphical summary of year;*

- Most of the operations performed were between 1967 to 1972
- Only 1 operation was performed in 1962 which was also the earliest year while 1 operation in 1977 which was the latest year an operation was performed.



From the above image, we can conclude that;

- Ages 40 to 70 had the highest number of operations

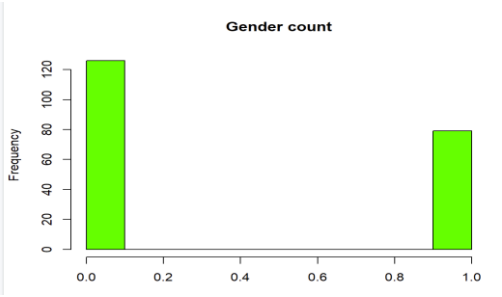


From the *Status* data in our dataset, we know that 1 indicates that the patient died from melanoma, 2 indicates that they were still alive and 3 indicates that they had died from causes unrelated to their melanoma. Hence, we can deduct the following;

- > 55 people and < 60 have died from Melanoma since operation.
- >120 people are alive since operation.

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

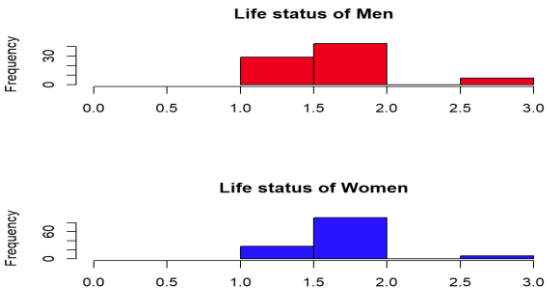
- More than 85 people but less than 90 have skin ulcer while over 100 people do not




```
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
```

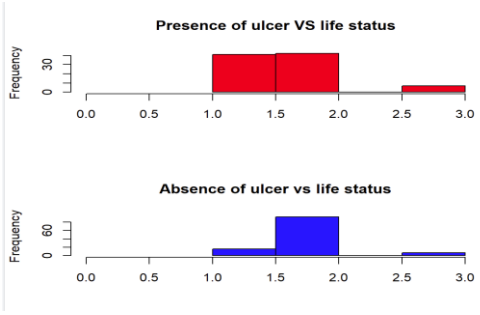
From our dataset, we established the patients sex 1=male, 0=female.

- There are significantly more female patients than there are male patients




```
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")
94
```

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

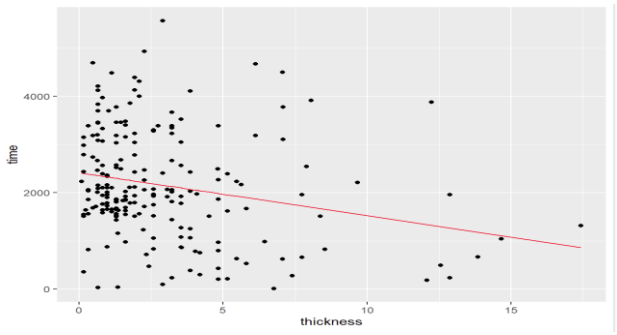



```
97  par(mfrow = c(2,1)) # combining two histograms in one plot
98  hist(rio_csv$status [rio_csv$ulcer == 1],
99       xlim= c(0,3),
100      breaks= 3,
101      main = "Presence of ulcer VS life status",
102      xlab = "",
103      col = "red")
104
105
106 hist(rio_csv$status [rio_csv$ulcer == 0],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

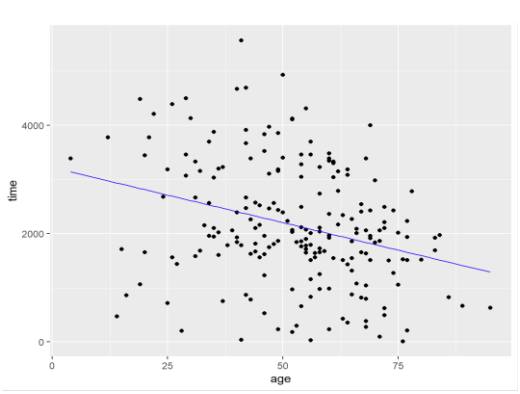
The image above image shows there are higher death rates of patients with ulcer, over two times greater of patients without ulcer.

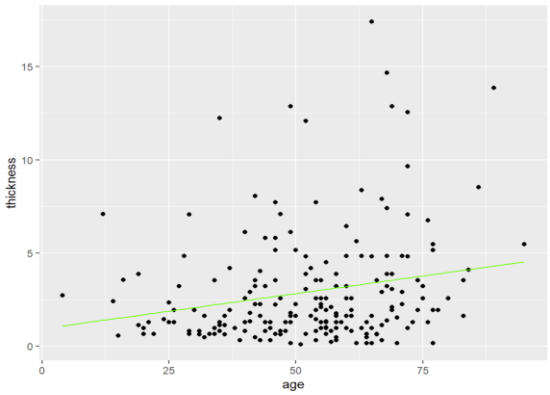
Regression

```
161 # Regression
162 data <- rio_csv
163
164 # Define variable groups
165 x <-data[,7]
166 y <-data[,2]
167
168 reg1 <- lm(time ~ thickness,
169            data = rio_csv)
170
171 reg1
172 summary(reg1)
173 confint(reg1)
174 coef(reg1)
175 library(stats)
176 cor(x,y)
177 hist(residuals(reg1))
178
179 predictions <- predict(reg1, data)
180
181
182 # visualize the model
183 ggplot(data, aes(x = thickness, y = time)) +
184   geom_point() +
185   geom_line(aes(y = predict(reg1)), color = "red")
186
```



The above code is a step-by-step regression analysis. The same process was repeated for other variables.






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

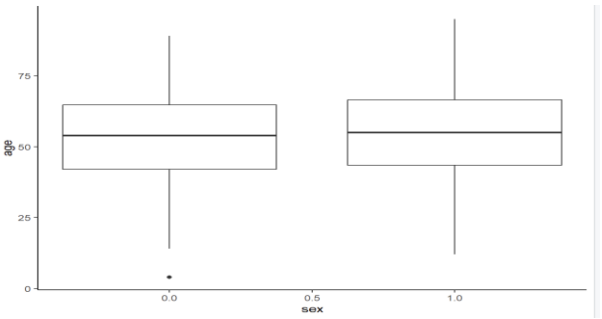
```
> cor(x,y)
[1] -0.3015179
```

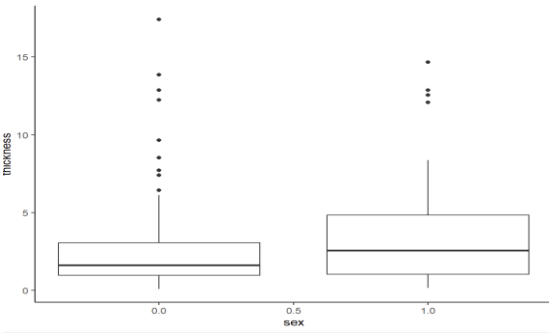
```
> cor(x,y)
[1] 0.2124798
```

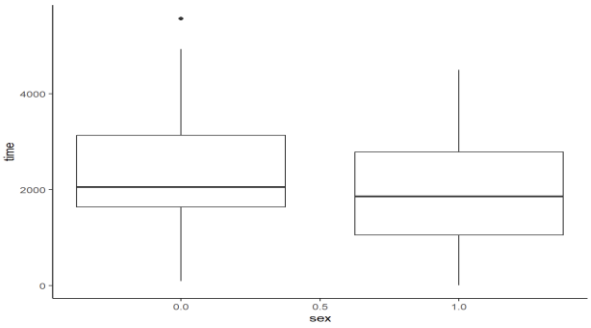
1. **Time ~ thickness** has a negative regression line *–0.2354087* as thickness increases, time reduces.
2. **Time ~ age** also negative *-0.3015179.* As age increases, time reduces
3. **Thickness ~ age** being the only positive regression line and highest correlation of all 3-variable pair *0.2124798*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

Two sample significance test – Grouped by gender(sex)

```
237 # t-test and boxplot
238 ggplot(rio_csv, aes(x = sex, y =age)) +
239   geom_boxplot(aes(group = sex)) + theme_classic()
240 t.test(age ~ sex, data = rio_csv)
241
242
243 ggplot(rio_csv, aes(x = sex, y =thickness)) +
244   geom_boxplot(aes(group = sex)) + theme_classic()
245 t.test(thickness ~ sex, data = rio_csv)
246
247 ggplot(rio_csv, aes(x = sex, y =time)) +
248   geom_boxplot(aes(group = sex)) + theme_classic()
249 t.test(time ~ sex, data = rio_csv)
250
```





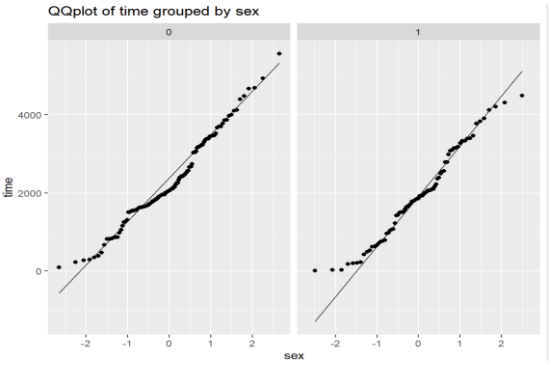


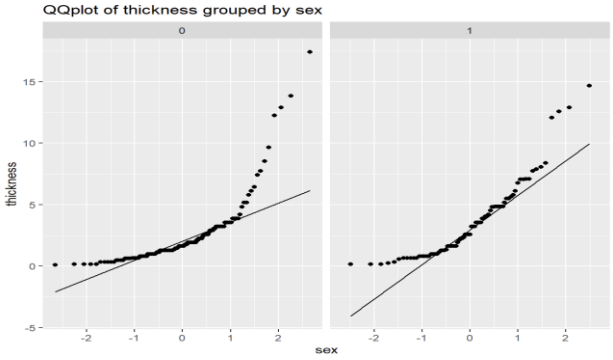
QQplots – Grouped by gender(sex)

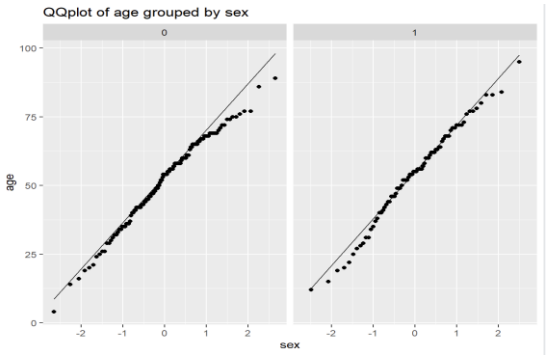
```r
#QQplots
ggplot(data = rio_csv, aes(sample = time)) +
  geom_qq() +
  stat_qq_line() +
  facet_wrap(~ sex) +
  labs(title = "QQplot of time grouped by sex",
       x = "sex",
       y = "time")

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")
```







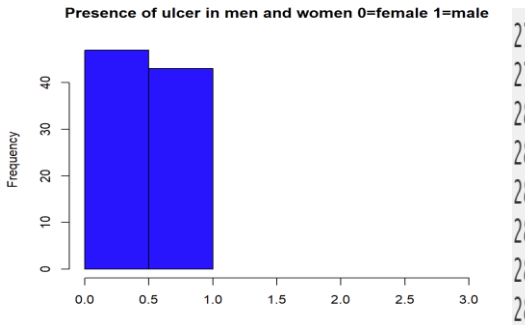
1. **Time** grouped by gender from the QQ plot above has a *thin tailed data distribution*
2. **Thickness** grouped by gender from the QQ plot above has a data distribution *skewed to the right*
3. **Age** grouped by gender from QQ plot above has a data distribution *skewed to the left*

Discussion

We have established that

1. There are almost twice as many women than men in the data set
2. There's a higher number of deaths from Melanoma in men compared to women
3. Presence of ulcer is a main symptom of death in Melanoma patients

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




```r
hist(rio_csv$sex [rio_csv$ulcer == 1],
     xlim= c(0,3),
     breaks= 3,
     main = paste("Presence of ulcer in men and women",
                  "0=female 1=male"),
     xlab = "",
     col = "blue")
```

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.