Introduction

Melanoma is **a type of skin cancer that develops when melanocytes (the cells that give the skin its tan or brown colour) start to grow out of control**. Cancer starts when cells in the body begin to grow out of control. Cells in nearly any part of the body can become cancer, and can then spread to other areas of the body. After surgery for the removal of the tumour, some patients survive and some do not. It is important to understand the factors that contribute to the death after surgery, so that adequate measures can be taken to protect patients' lives and increase their chances of survival.

The data set used for this analysis consists of measurements made on patients with malignant melanoma. Each patient had their tumour removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark during the period 1962 to 1977. The surgery consisted of complete removal of the tumour together with about 2.5cm of the surrounding skin. Among the measurements taken were the thickness of the tumour and whether it was ulcerated or not. These are thought to be important prognostic variables in that patients with a thick and/or ulcerated tumour have an increased chance of death from melanoma. Patients were followed until the end of 1977.

Melanoma dataset

The melanoma data frame has 205 rows and 7 columns, This data frame contains the following columns:

Column	Description
sex	The patients sex; 1=male, 0=female.
Age	Age in years at the time of the operation.
Thickness	Tumour thickness in mm.
Year	Year of operation.
Survival time	in days since the operation, possibly censored.
Ulcer	Indication of ulceration 1=present, 0=absent.
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 3 indicates that they had died from causes unrelated to their melanoma.

Table 1: **Melanoma dataset**

Brief visualisation of dataset

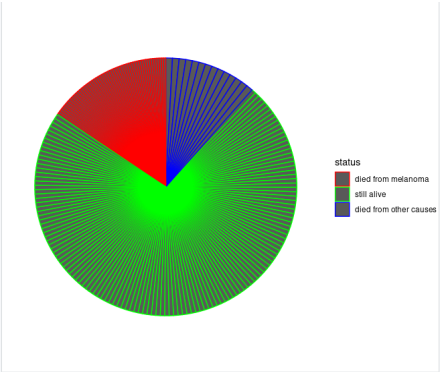

Chart 1: Status

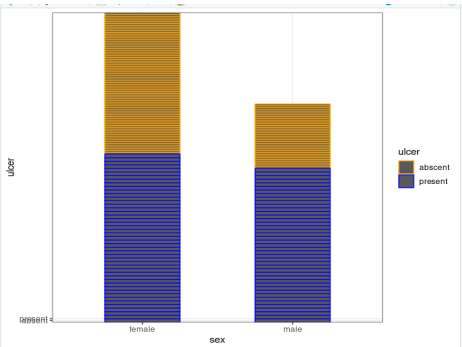

Chart 2: **ulcer between F/M**

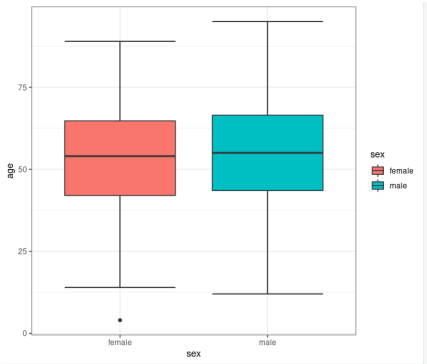

Chart 3: **Age between F/M**

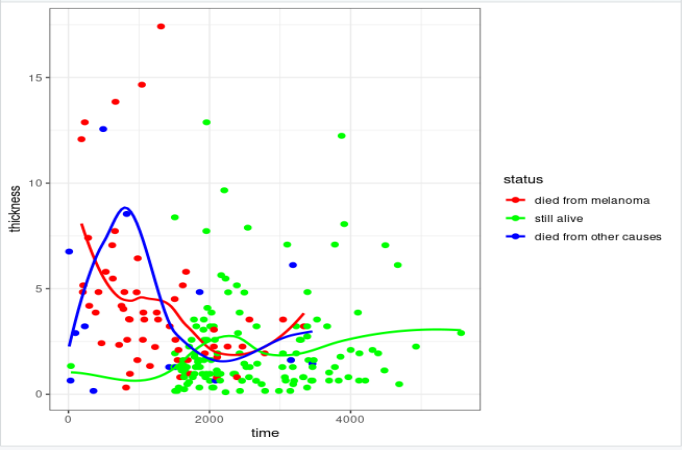

Chart 4: **thickness over time for different status**

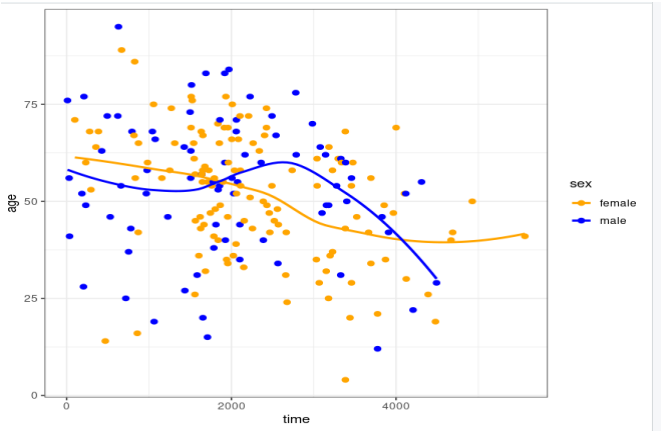

Chart 5: **Age distribution over time for different sex**

Summary of dataset

```
> summary(data)
      time                      status       sex          age            year       thickness       ulcer
 Min.   :  10   died from melonema    : 57   female:126   Min.   : 4.00   Min.   :1962   Min.   : 0.10   absent :115
 1st Qu.:1525   still alive          :134   male  : 79   1st Qu.:42.00   1st Qu.:1968   1st Qu.: 0.97   present: 90
 Median :2005   died from other causes: 14                Median :54.00   Median :1970   Median : 1.94
 Mean   :2153                                            Mean   :52.46   Mean   :1970   Mean   : 2.92
 3rd Qu.:3042                                            3rd Qu.:65.00   3rd Qu.:1972   3rd Qu.: 3.56
 Max.   :5565                                            Max.   :95.00   Max.   :1977   Max.   :17.42
>
```

Summary of dataset (Excluding recorded status which was created for analysis)

Analysis using Kaplan Meier method

A new column was created (recorded status) in order to group the status into two observations event and censored, the event are those that died from melanoma, while those that are alive and those that died from other causes are considered as censored observations, we are assuming that those died from other causes are censored observations, because we are not able to ascertain how long the would have survived before dying from melanoma, other features such as sex and ulcer are converted to factors.

The data is fitted into the surv function, and fitted into the Kaplan meier curve using the Survfit function, the results are as follows

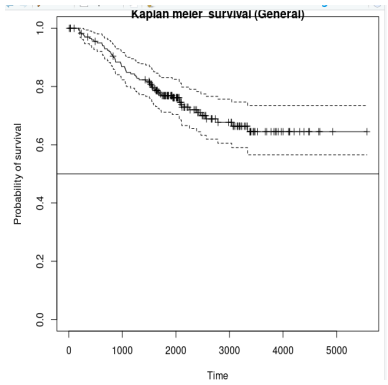

Chart 6: **Kaplan meier(general)**

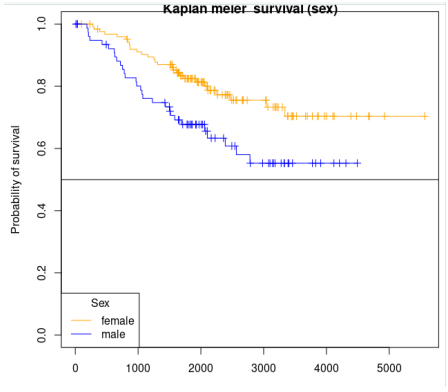

Chart 7: **Kaplan meier(sex)**

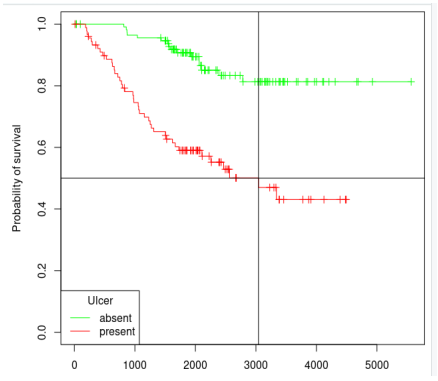

Chart 8: **Kaplan meier(Ulcer)**

Feature	n	Event	Median	0.95LCL	0.95UCL
General	205	57	NA	NA	NA
Sex (Male)	79	29	NA	2388	NA
sex (female)	126	28	NA	NA	NA
Ulcer(absent)	115	16	NA	NA	NA
Ulcer(present)	90	41	3042	1726	NA

Table 2: **showing median survival time for ulcer and sex**

From the survival charts above, we can assume that females have a higher survival rate than males and that people with ulcer absent have a better survival rate than those with ulcer present. Also, this could be also due to the fact that a higher proportion of males have ulcer present compared to females, as seen in **Chart 2.** However; this assumptions need to be verified using the p value, to be sure if we are going to accept the null hypothesis or reject it so as to avoid a type 1 error.
In addition,It can also be observed that most of the of data point for both sex and ulcer fall above the median line except for those with ulcer present. The median line is the point on the probability of survival axis where the value is 0.5, this gives us an insight of how many people with the feature indicated have a survival probability above or below 50%. 18% of those with ulcer present have less than 50% chance of survival.
Using the survdiff function of the survival library we calculate to obtain the P value of the features above and the results are as follow

Feature	P Value
sex	1%
Ulcer	0.000005%

Table 3: **P value for sex and Ulcer (log rank test)**

Conclusion of the KM model

For both features the p < 5%, which means that we can reject the null hypothesis, and conclude that Females have higher survival rate than the males and that those with ulcer present have a lower survival rate than those with ulcer absent.

Cox model

The cox model was fitted for six different scenarios in order to ascertain the best model to be used for the analysis

model	Features	P - values respectively
1	Sex + age+ year+ thickness + ulcer	9.3% - 5% - 9.2% - 0.8% - 0.01%
2	Sex + age+ thickness + ulcer	10% - 14% - 0.03% - 0.01%
3	Sex + thickness + ulcer	8.4% -0.2% - 0.02%
4	thickness + ulcer	0.16% - 0.008%
5	ulcer1	0.0000629%
6	thickness	0.0000296%

Table 4: **Models used for cox proportional hazard**

The Anova test was carried out on all the models and the best model was chosen to be **model 4,** which included only the thickness and the ulcer, the test showed significant difference between the 2 models. The final model chosen is shown below

	coef	exp(coef)	exp(-coef)	Lower 0.95	Upper 0.95	P
Thickness	0.1140	1.121	0.8922	1.044	1.203	0.00159
Ulcer	1.2180	3.380	0.2958	1.845	6.195	**0.0000812**

Table 5: **summary of data from the cox model chosen (Bestfit)**

The assumptions for the cox model are as follows

For 1 unit additional thickness of tumour, the risk of death is increased by 12.1% (1.12) with a confidence interval between 4.4% (1.044) and 20.3% (1.203), also having ulcer can increase the risk of death by 238% (3.380), with a confidence interval of between 84.5% (1.845) and 519.5% (6.195).

However, we need to check if these assumptions are valid. First we check for the residual using the deviance and martingale methods

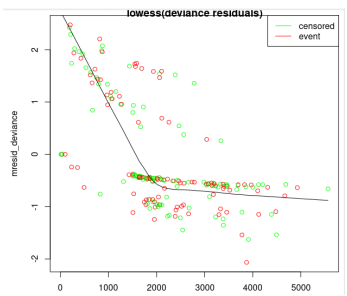

Chart 9: **deviance(time)**

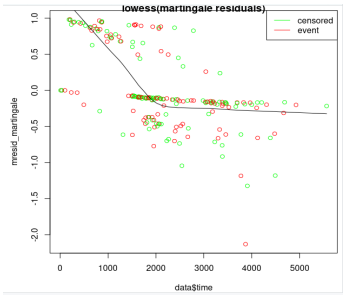

Chart 10: **martingale(time)**

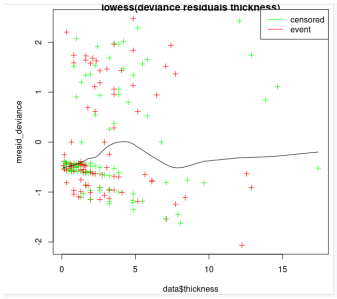

Chart 11: **deviance(thickness)**

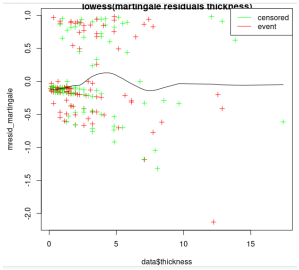

Chart 12: **martingale(thickness)**

No clear trend in residuals, so we check for proportional hazard assumption using cox.zph function

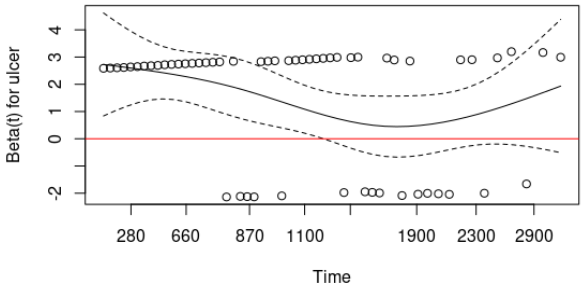

Chart 13: **proportional hazards (Ulcer)**

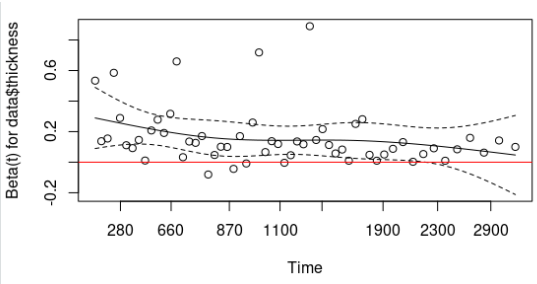

Chart 14: **proportional hazards (thickness)**

The proportional hazard ratio does not seem to be fully met, indicating that the proportional hazard changes sometimes over time.

Conclusion for the cox model

For 1 unit additional thickness of tumour, the risk of death is increased by 12.1% (1.121), and Having ulcer can increase the risk of death by 238% or (3.380).