**Introduction**

Malignant Melanoma is a type of skin cancer that arises from melanocytes. Melanocytes are cells that produce melanin which are responsible for the color of the skin. Malignant melanoma, unlike other forms of melanoma, can be a very deadly form of skin cancer because it spreads very fast to other organs (Skin Cancer Foundation, 2020). Melanoma can be treated successfully if it is discovered early.

This report conducts an exploratory data analysis (EDA) of the ‘Survival from Malignant Melanoma’ dataset using ‘R’ software (R Core Team, 2023) to glean insights and inform further research in the field. It is believed that the thickness of the tumor, and whether it is ulcerated influences the survival chances of patients that have been diagnosed with melanoma (the Skin Cancer Foundation, 2020). The data was collected from a study carried out on patients that had their tumor removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark during the period 1962 to 1977.

**The dataset**

**A number of people with black text

Description automatically generated with medium confidence**

**Fig 1: Updated Melanoma Data**

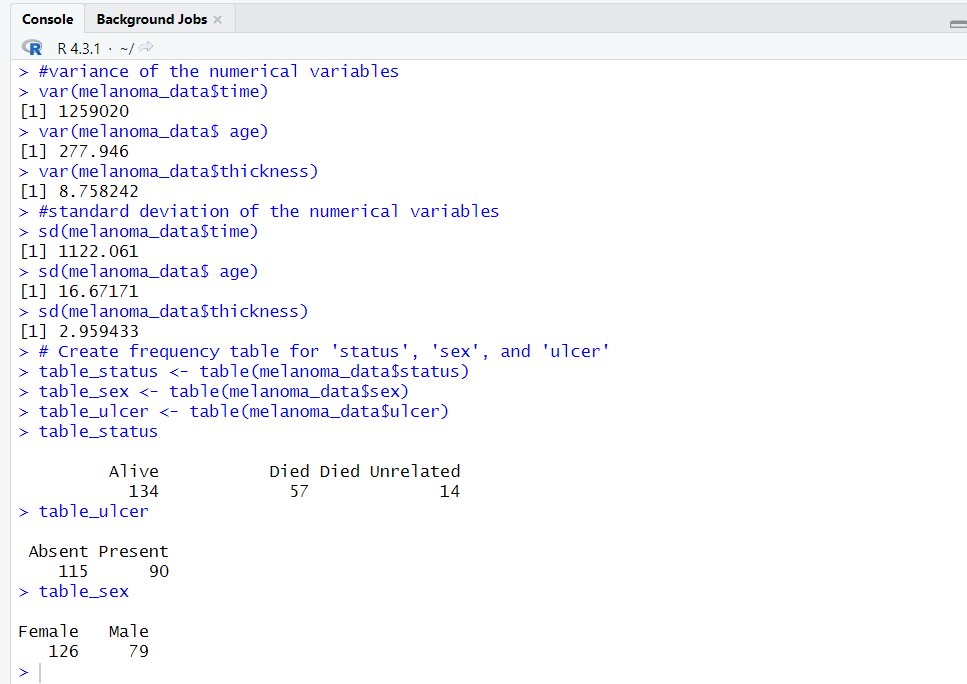
The data consists of measurements made on patients with malignant melanoma. The dataset has 7 columns and 205 rows. Fig 1 shows an updated version of the data after some transformation was done. The columns ‘sex’, ‘ulcer’ and ‘status’ were recoded to reflect what they represent. In the initial dataset, they were represented with numbers. The columns include :

* time - Survival time in days since the operation.
* status - The patients’ status at the end of the study.
* sex - The patients’ sex.
* age - Age in years at the time of the operation.
* year - Year of operation.
* thickness - Tumour thickness in mm.
* ulcer - Indicator of ulceration.

**Numerical Summary**

**A screenshot of a computer

Description automatically generated**

****

**Fig 2: Descriptive Statistics of the data**

This section presents the numerical summary of the dataset.

* Time: The minimum time a patient has survived for is 10 days, while the maximum time is 5565 days. This shows that some of the patients who had their operation at the time the study spanned from, have survived and are still alive. The mean time is 2153 days while the median time is 2005 days. The standard deviation and variance are 1122.06 and 1259020 days respectively. Standard deviation is the average distance between the data point and the mean (Khan Academy, 2018). It measures how each data point deviates from the mean. A standard deviation of 1122.06 days suggests that, on average, each time value in the dataset is about 1122.06 days away from the mean. Variance measures how much each data point differs from the mean (Khan Academy, 2018).
* Age: The age for the patients in the dataset ranges from 4 to 95 years. The mean age is 52.46 while the median age is 54. This distribution is relatively symmetric. Standard deviation of the age is 16.67171while the variance is 277.946 years.
* Year: The study covers 15 years, from 1962 to 1977 which is reflected in the min and max years. The mean year is 1970, and the median year is also 1970.
* Thickness: The thickness of the melanoma is between 0.10mm and 17.42mm. The mean thickness is 2.92mm and a median of 1.94mm. Standard deviation of the thickness is 2.959433, while variance is 8.758242mm.

**Graphical Summary**

A group of graph of melanoma

Description automatically generated

**Fig 3: Histogram of year, time, thickness and age variables in the melanoma data**

**A screenshot of a graph

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA screenshot of a computer screen

Description automatically generated**

**Fig 4: Bar chart showing distribution of Ulcer, sex and Status variables.**

**Fig. 3** shows thatthe distribution of the histogram showing melanoma\_data$thickness is right skewed. This shows that most people in the dataset have lower thickness values with a few concentrated on higher thickness values. The histogram showing the variables, age, data and time in the melanoma data are relatively symmetrical. The frequencies are evenly distributed with the data not being concentrated on any value.

**Fig 4** shows the distribution of the ulcer, sex and status variables. It shows that a larger number of people have ulcer absent than present, there are more females than males, and more patients are alive than dead, and a few patients have died due to unrelated causes.

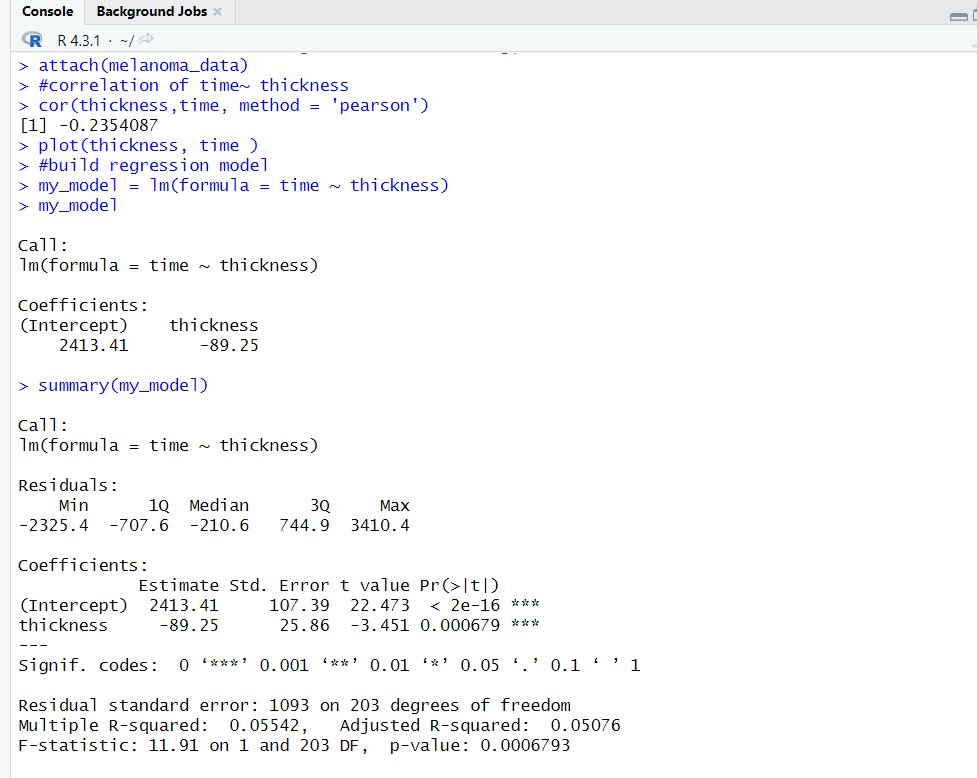
**Correlation and Regression**

**Time ~ Thickness**

**Figure 5** below shows that time has a negative correlation of -0.2354087 with thickness. This indicates that they have a very weak negative linear relationship since the correlation value is close to zero. As thickness increases, time decreases, and vice versa.

Using y = mx+c, the result shows that time = thickness ( -89.25) + 2413.41. 2413.41 is the intercept. This means that for every time the thickness is zero, there is an increase of 2413.41 in the time of survival, and for every unit increase in thickness, there is an estimated decrease of 89.25 in time.

Statistical Significance: the p-value is 0.0006793 which is very low, suggesting that thickness is a statistical predictor of time.

****

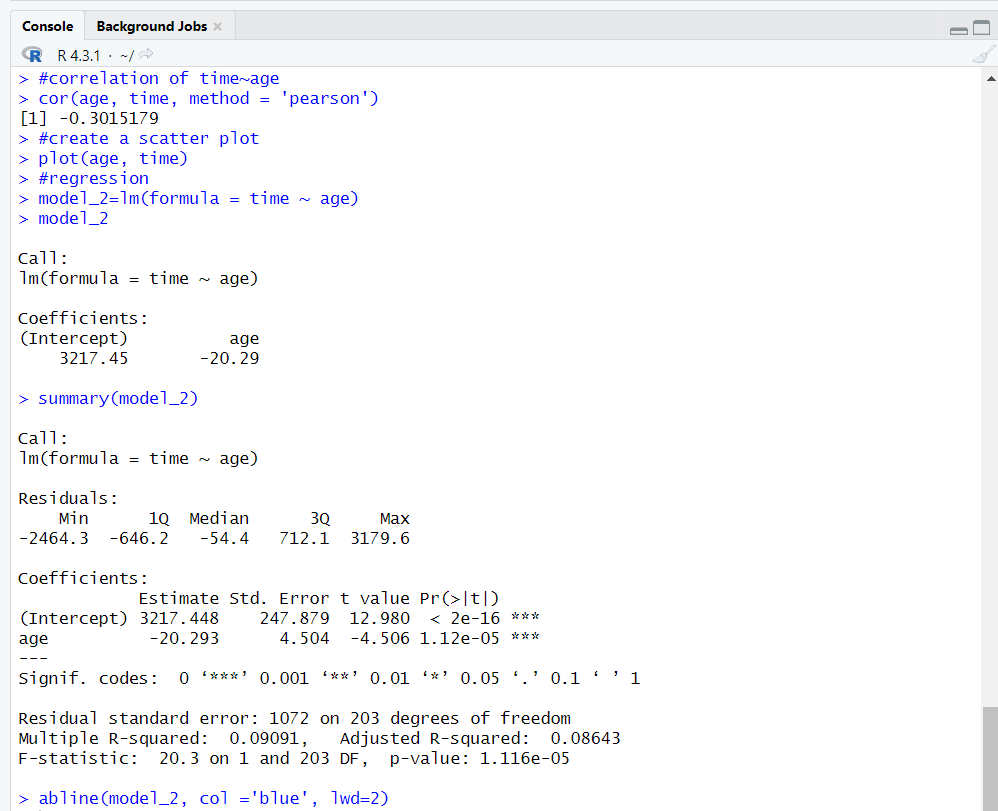
**Fig 5: Correlation results showing relationship between time and thickness.**

**Time ~ Age**

**Figure 6** below shows that time has a negative correlation of -0.3015179 with age. This indicates that they have a very weak negative linear relationship since the correlation value is close to zero. As time increases, age decreases, and vice versa.

Using y = mx+c, the result shows that time = age ( -20.29) + 3217.45. 3217.45 is the intercept. This means that for every time the age is zero, there is an increase of 3217.45 in the time of survival, and for every unit increase in time, there is an estimated decrease of 20.29 in age.

Statistical Significance: the p-value is 1.116e-05 which is very low, suggesting that age is a statistical predictor of time.



**Fig. 6: Correlation results showing relationship between time and age.**

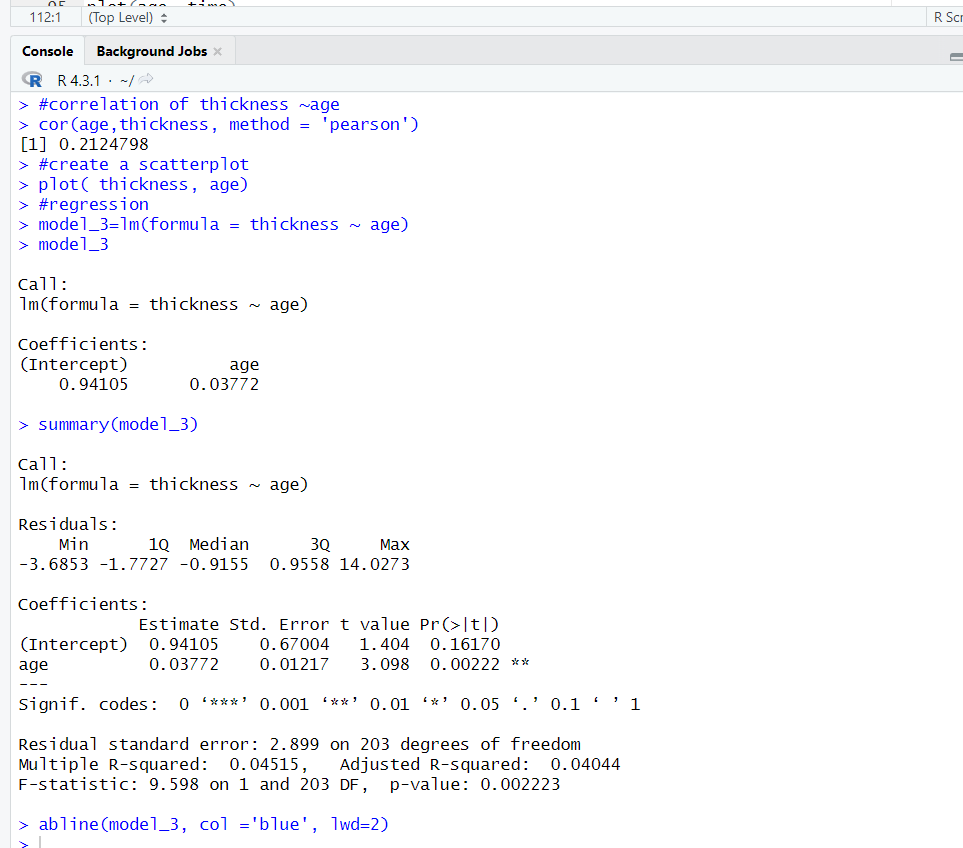
**Thickness ~ Age**

**Figure 7** below shows that thickness has a positive correlation of 0.2124798 with age. This indicates that they have a very weak positive linear relationship since the correlation value is close to zero. As thickness increases, age increases, and vice versa.

The result shows that the equation for the regression model is y = 0.03772*x* + c, where y = thickness and x = age. Thus;

thickness = age (0.03772) + 0.94105. 0.94105 is the intercept. This means that for every time the age is zero, there is an increase of 0.94105 in thickness, and for every unit increase in thickness, there is an estimated increase of in age.

Statistical Significance: the p-value is 0.002223 which is very low, suggesting that thickness is a statistical predictor of age.



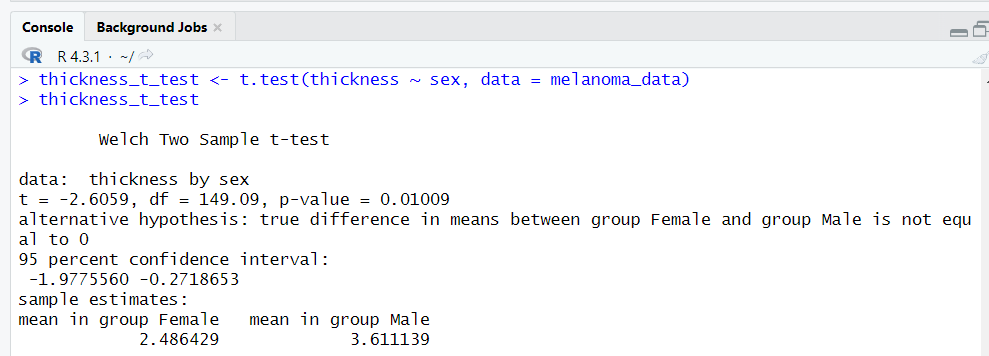
**Fig. 7: Correlation results showing relationship between thickness and age.**

**Two Sample Significance Tests**

**Thickness**

Ho: The true difference in mean ‘thickness’ between the Female and Male groups is equal to 0.

Alternative Hypothesis (H1): The true difference in mean ‘thickness’ between the female and male groups is not equal to 0.



**Fig 8: T-test result for thickness grouped by sex.**

The default level of significance is α = 0.05 and we have a p-value which is much smaller  
than this. Therefore, we can reject the null hypothesis Ho and conclude that there is evidence that the true mean thickness is different depending on if the person is male or female. The t-value of -2.6059 indicates how many standard errors the observed difference in means is from zero. In this case, the negative t-value suggests that the mean "thickness" for the Female group is lower than that for the Male group.

**Time**

Ho: The true difference in mean ‘time’ between the female and male groups is equal to 0.

Alternative Hypothesis (H1): The true difference in mean ‘time’ between the female and male groups is not equal to 0.

A screenshot of a computer

Description automatically generated

**Fig 9 : T-test result for time grouped by sex.**

The default level of significance is α = 0.05 and we have a p-value which is smaller  
than this. Therefore, we can reject H0 and conclude that there is evidence that the true  
mean ‘time’ values are significantly different between the male and female groups. The t-value of 2.0848 indicates how many standard errors the observed difference in means is from zero. In this case, the negative t-value suggests that the mean "time" for the female group is higher than that for the male group.

**Age**

Ho: The true difference in mean ‘age’ between the female and male groups is equal to 0.

Alternative Hypothesis (H1): The true difference in mean ‘age’ between the female and male groups is not equal to 0.

A screenshot of a computer

Description automatically generated

**Fig 10: T-test result for age grouped by sex.**

The default level of significance is α = 0.05 and we have a p-value which is relatively higher than 0.05. Therefore, this suggests that we do not have enough evidence to reject the null hypothesis that the mean ages between the female and male groups are equal, thus we may not reject it. The t-value of -0.95559 indicates how many standard errors the observed difference in means is from zero. In this case, the negative t-value suggests that the mean "age " for the female group is lower than that for the male group.

**Q-Q Plots**

**Q-Q Plots of Thickness by Sex**

**Figure 11** below shows the two Q-Q plots for the male and female groups appearing as a concave curve. From the plots, the smallest and largest thickness values are larger than would be expected from a normal distribution plot. The points are above the line on the Q-Q plot. The lower end of the data distribution has been reduced relative to normal distribution, while the upper end has been extended relative to normal distribution.

A screenshot of a graph

Description automatically generated

**Fig 11: Q-Q Plots of Thickness by Sex**

**Q-Q Plots of Age by Sex**

F**igure 12** below shows that the two Q-Q plots for the male and female are normally distributed. The distribution is relatively symmetrical as the largest and smallest observations are not far from the distribution line. The distribution has no skew.

A screenshot of a graph

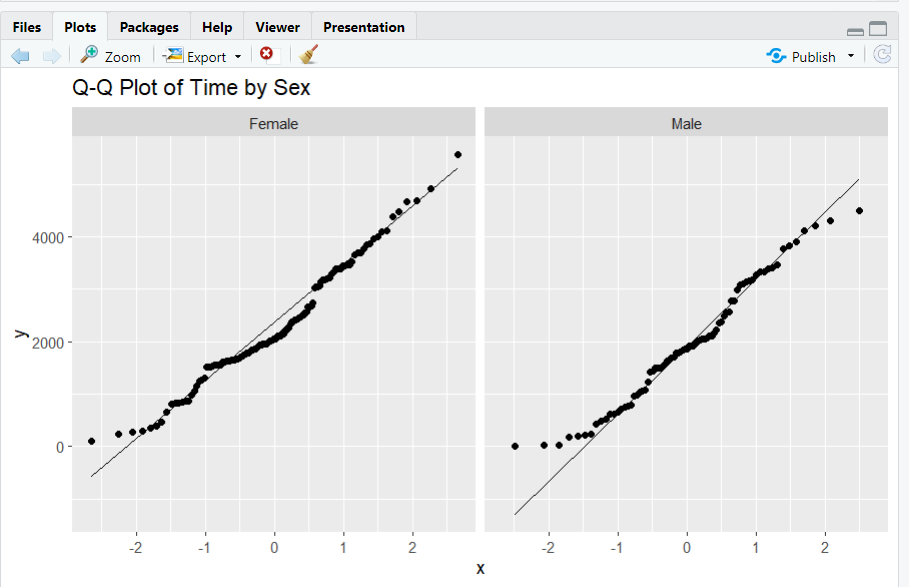
Description automatically generated

**Fig 12:** **Q-Q Plots of Age by Sex**

**Q-Q Plots of Time by Sex**

**Figure 13** below shows that the Q-Q plots for the female is normally distributed. The distribution is relatively symmetrical as the smallest observations is not far from the distribution line while the largest observation is on the distribution line.

The Q-Q plot for the male can be said to be under dispersed. The smallest observations are larger than you would expect from a normal distribution (i.e. the points are above the line on the Q-Q plot), while the largest observations are less than what would be expected from a normal distribution (distribution (i.e. the points are below the line on the Q-Q plot), This means the lower and upper tails of the data’s distribution has been reduced, relative to a normal distribution.



**Fig 13: Q-Q Plots of Time by Sex**

**Discussion and Recommendations**

An exploratory data analysis (EDA) was done using the melanoma dataset. The results of the analysis give insights to the data which is summarised below.

1. Even though the time variable gave valuable insights to the survival of patients after melanoma surgery, the wide range of 10 to 5565 days suggests considerable variability in outcomes. This could be related to the year in which the surgery took place. Some surgeries took place just immediately before the data was taken, which could be responsible for the large variability. The Minimum value for time of survival is 10 days which could be that the person just did the surgery before the data was collected. The mean and median survival time of 2153 and 2005 days respectively also indicates variability which suggests varied outcomes for different patients. Thus, a personalized treatment plan is essential.

2. The analysis suggests a relationship between the thickness of the tumor and survival time, with a negative linear correlation of -0.2354. This implies that the thicker the tumor, the decreased chance of survival after the surgery. The regression model further supports this, indicating a statistically significant predictor role for thickness in survival.

3. The analysis also reveals that age has a relationship with thickness of the tumor on survival outcomes. Age has a positive correlation 0.2124798 with thickness and a negative linear correlation of -0.3015179 with time of survival. This indicates that younger patients have less thick tumors and usually survive longer than older patients. As age decreases, tumor thickness is low, thus time of survival increases and vice versa. This highlights the importance of considering the age of the patient in treatment decisions.

4. The analysis reveals that the average "thickness” and ‘age’ for the females is lower than males, while the average time of survival is higher in females than males. This suggests that females tend to have lower thickness and higher survival times compared to males.

5. The right-skewed distribution of tumor thickness suggests that most patients have lower thickness values, aligning with a more favorable prognosis. Contrastingly, the relatively symmetrical distributions of age, year, and time indicate a more even spread of values.

**Recommendations**

1. A more detailed investigation can be done to explore the impact of ulceration on survival outcomes. The time of survival should also be compared between patients that have ulceration or not. An investigation could also be done to see if there is a relationship between the thickness of the tumor and whether an ulceration is present. This can be categorised by sex to explore if the results are peculiar to men or women. This would help in specialised treatment plans and clinical management. For instance, according to Bønnelykke-Behrndtz and Steiniche, 2017) patients with ulceration have their survival chances increased if treated with adjuvant immunotherapy.

2. The age can be grouped into different age groups to understand if specific age groups exhibit unique survival patterns. This would help in tailoring interventions for the different age groups.

3. An analysis could also be done based on year to investigate if there exist unique patterns based on survival outcomes for each year. This is to see if survival rates improved in any particular year and investigate factors that may be responsible for it.

4. Explore the time of survival of the patients based on the year the surgery was carried out to identify any distinct characteristics the long-term survivors have in common.

5. The dataset can be expanded to include additional variables such as sathe of the melanoma at diagnosis or genetic factors. This approach could enhance the predictions for survival.

**Conclusion**

This exploratory data analysis sheds light on factors influencing the survival of patients with malignant melanoma. The insights derived from thickness, age, time of survival and sex serves as a basis to aid in customizing treatment plans for each patient. Further investigations into ulceration, trends, and comprehensive age analyses are recommended to reveal more insights. By exploring these aspects, valuable contributions can be made in the study of melanoma, especially to increase chances of survival for patients.

**References**

1. Bønnelykke-Behrndtz, M.L. and Steiniche, T. (2017). Ulcerated Melanoma: Aspects and Prognostic Impact. [online] PubMed. Available at: https://www.ncbi.nlm.nih.gov/books/NBK481861/.Khan Academy (2018). Population and sample standard deviation review. [online] Khan Academy. Available at: https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/variance-standard-deviation-sample/a/population-and-sample-standard-deviation-review.
2. Per Kragh Andersen, Borgan, Ø., Gill, R. and Niels Keiding (1993). Introduction. Springer series in statistics, pp.1–44. doi:https://doi.org/10.1007/978-1-4612-4348-9\_1.
3. R Core Team (2023). R: A Language and Environment for Statistical Computing. [online] R Foundation for Statistical Computing. Available at: https://www.r-project.org/.
4. The Skin Cancer Foundation. (2020). Melanoma Stages. [online] Available at: https://www.skincancer.org/skin-cancer-information/melanoma/the-stages-of-melanoma/#:~:text=The%20thicker%20the%20melanoma%20measures [Accessed 17 Jan. 2024].