Statistical Analysis of Skin Cancer Risk Factors

Mark Zusman

1. Introduction

Cancer is the second most common disease in America (*CDC)*. Out of all cancers, skin cancer is the most popular type. The most popular types of skin cancer are squamous cell carcinoma, basal cell carcinoma, and melanoma. Although melanoma is not the most common it has the highest probability to invade nearby tissue and spread to other parts of the body. This leads to melanoma being responsible for the most amount of deaths out of all other skin cancers (*Skin Cancer*, n.d.).

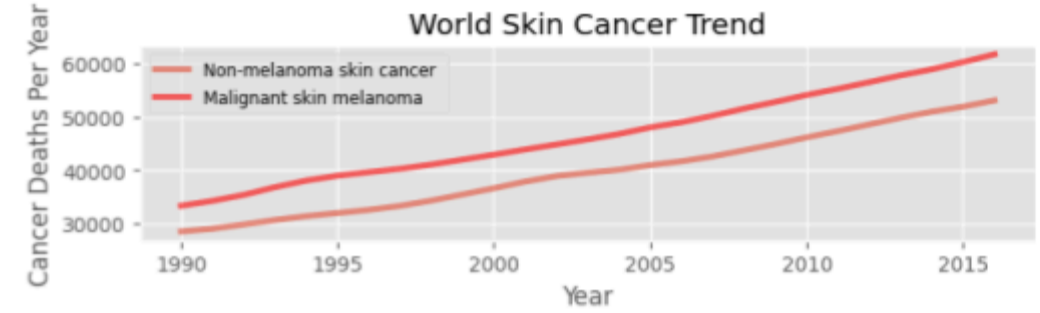
The ultimate goal of this report is to help understand the skin cancer numbers, what could cause cancer, and how the smoothness of tumors plays a role in whether the tumors are malignant or benign. Understanding certain health patterns that could be linked to skin cancer is the most important way to prevent cancer.

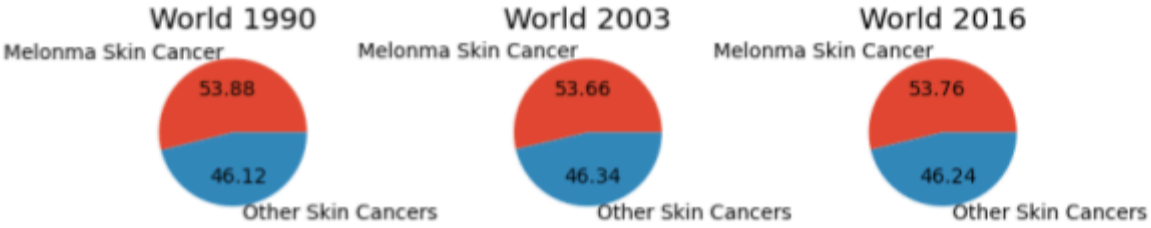
2. Analysis

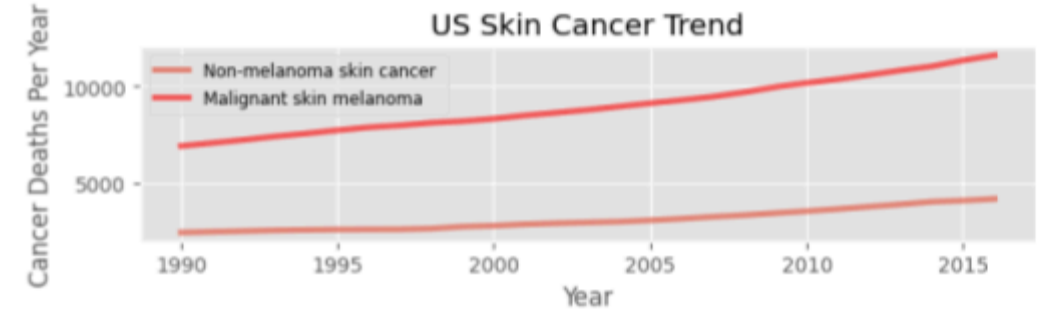
This report contains an analysis of three different datasets. The different datasets have been broken down into 3 parts.

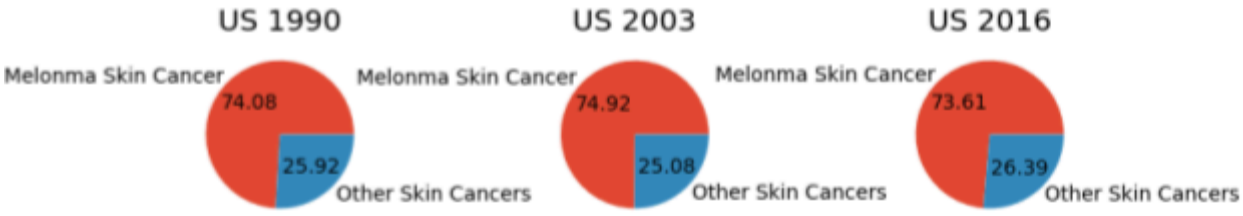
Part 1

In part one the dataset was a breakdown count of different types of cancers in almost every country in the world. The dataset can be found on Kaggle[1.] (see Appendix B). It contained information about 18 different types of cancers, however only the skin cancer columns and years were used for this report. This dataset provides a good description of the cancer breakdown in the world vs. the US in recent years.









'igure 1a (row #1): Line chart showing the world skin cancer trend vs the year (1990-2016)

'igure 1b,c,d (row #2): Pie charts showing the world skin cancer percentage in the years (1990, 2003, 2016)

'igure 1e (row #3): Line chart showing the US skin cancer trend vs the year (1990-2016)

'igure 1f,g,h (row #4): Pie charts showing the US skin cancer percentage in the years (1990, 2003, 2016)

Part 2

In part two, the ultimate goal was to attempt to recognize patterns to detect skin cancer. The first dataset used is about the health patterns of Americans. The dataset was created by the Centers for Disease Control and Prevention (CDC). The dataset was based on telephone surveys and it aims to get the most popular health-related risk behaviors. In part two, the data set was from Kaggle[2] (see Appendix B)

The second dataset is from a health survey with many categories of categorical data. One of the columns indicated if a person was diagnosed with skin cancer or not. To understand the dataset there will be a preliminary analysis understanding the distributions of the data.

In the dataset, the three categorical columns were SleepTime, Physical Health, and Mental Health. The physical health and mental health columns both indicated a numeric rating scale for the person who felt they had poor physical or mental health that day. Both columns had more than 68,700 samples that ranged from 0 to 30. The average response for physical health was 3.35 with a standard deviation of 7.9. The mental health column was similar as it had an average response of 3.93 with a standard deviation of 8.0. The sleep column rendered different data. The sleep column had a numeric value indicating the average number of hours the respondent slept in 24 hours. This column had values that ranged from 1 to 24 with an average response of 7.10 hours. The standard deviation for this column was 1.4.

In *Figure 1*, we can see the distribution of the frequency of skin cancer. Overall, there are 68,795 data points in this dataset. As we can see most of the people surveyed in the dataset did not have skin cancer.

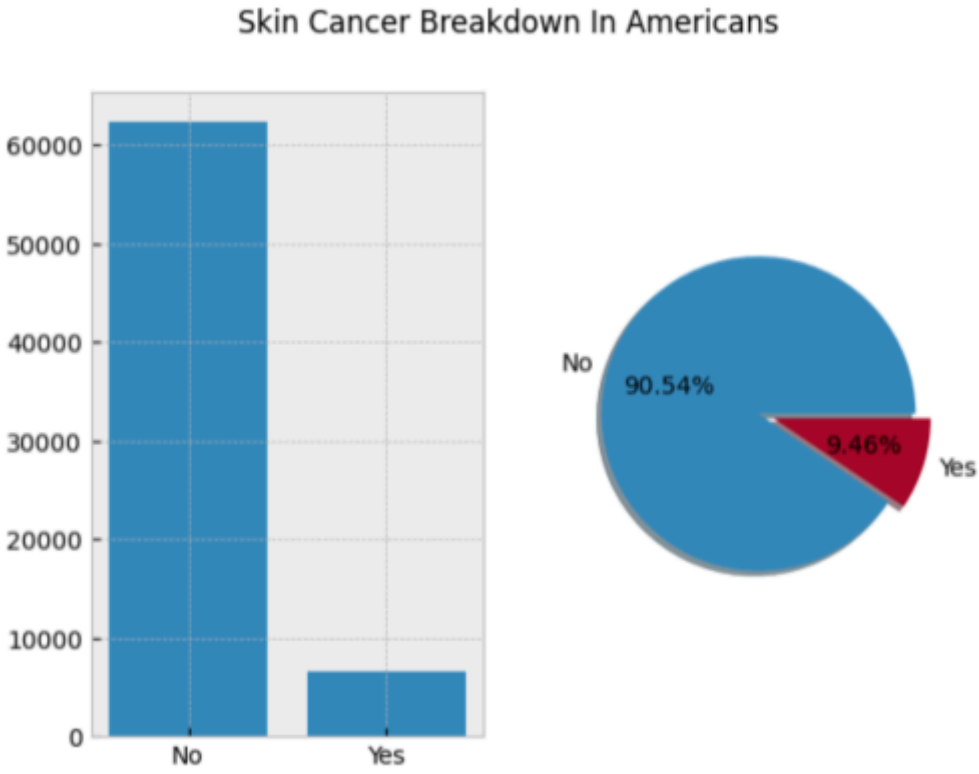


Figure 2a (left): Bar chart showing the count distribution of surveyed people with skin cancer

Figure 2b (right) : Pie chart showing the percentage distribution of surveyed people with skin cancer

Next, understanding the distribution of some variables will help with understanding the dataset. The age category distribution can be seen here in this bar chart.

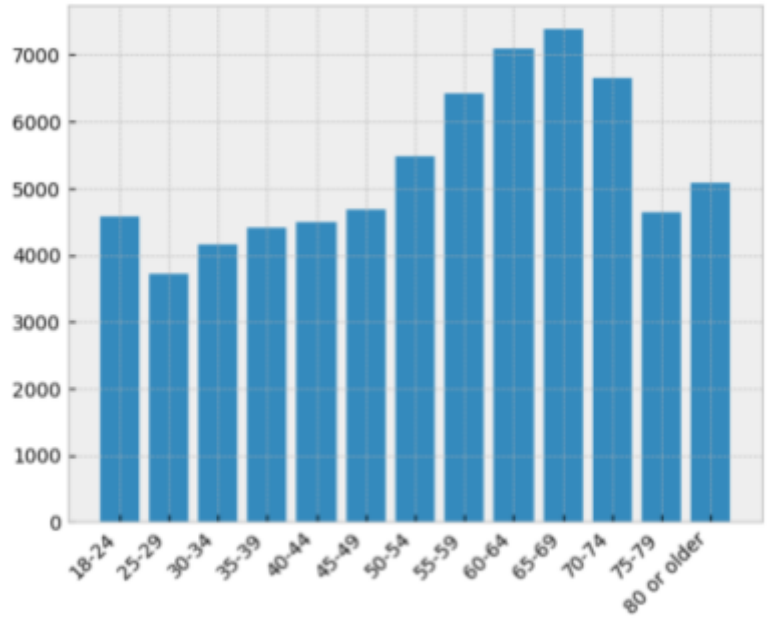
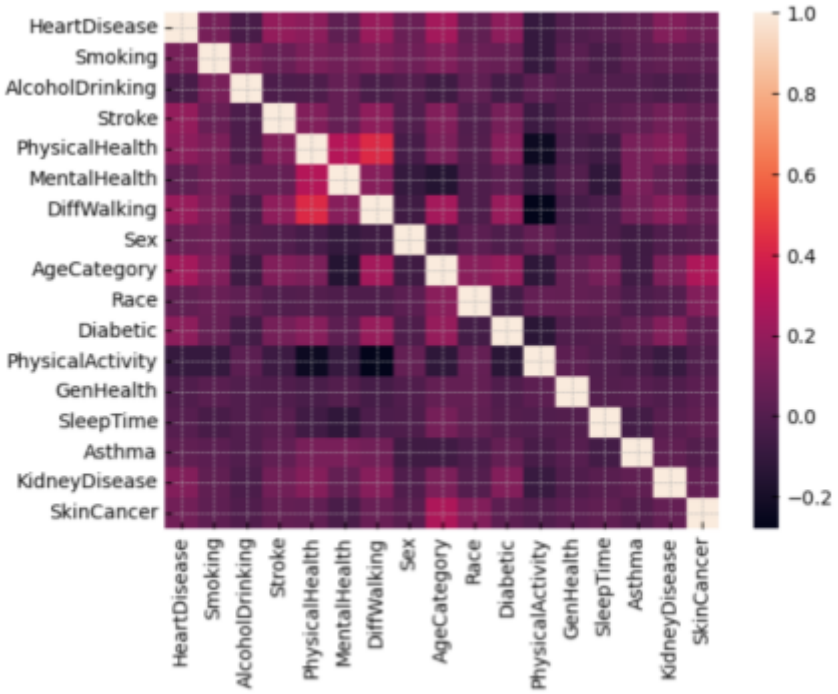


Figure 3: Bar chart showing the age range distribution breakdown



To start our analysis let's look at the correlation of all of the variables amongst each other. Overall, there is not a The age category that has the highest correlation with the skin cancer category.

Figure 4: A Seaborn heatmap that shows the Pearson r correlation between all categories.

Since there is some correlation between whether a person has skin cancer or not and their age, more analysis to see how these are correlated is necessary. A Mann-Whitney U test was performed. This test tests whether the distributions of the population are identical or not. The null hypothesis states that they are identical and the alternative hypothesis states they are not identical. After running the test the P-value for the Mann-Whitney U test was 0.0000. This p-value indicates that we should reject the null hypothesis and accept that the alternative hypothesis (that these groups are not identical) is true. To see where the difference in the groups was I created a box and whisker plot to compare the distributions. If a person had skin cancer this plot indicated that typically they are older.

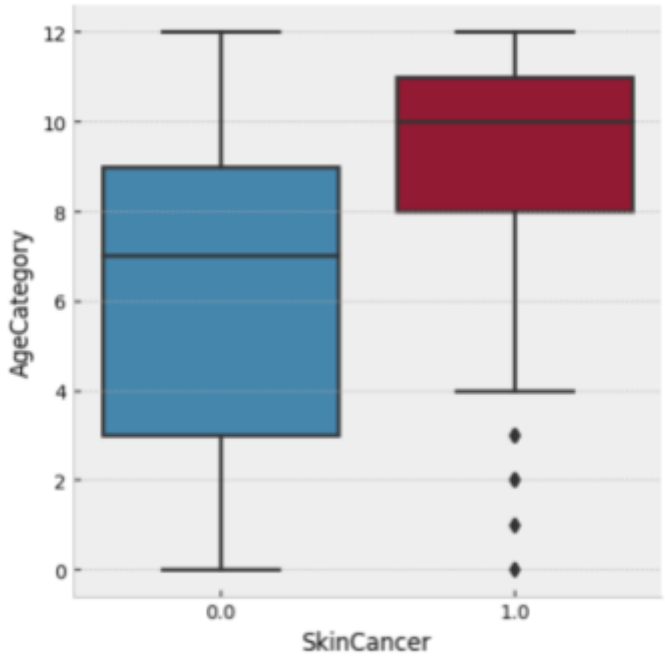


Figure 5: A Seaborn box and whisker plot that shows the categorical distribution if a person's age and if they had skin cancer or not.

The breakdown of skin cancer against each of the other columns was analyzed with a Chi-Square Test of Independence. A Chi-Square test tests the correlation between columns of categorical data. In the Chi-squared test, the null hypothesis assumes the two categories are independent of each other. Conversely, the alternative hypothesis for the test is that the two variables are related. The hypothesis accepted is determined by the p-value run from the Chi-Square test. If the p-value is less than .05 then we would reject the null hypothesis and accept the alternative hypothesis. The graphs show the contingency (raw count) subtracted from the expected values. When the result is positive it indicates there were more values in the contingency than expected. Alternatively, when the result is negative, it indicates that there were more expected values than the contingency (raw count).

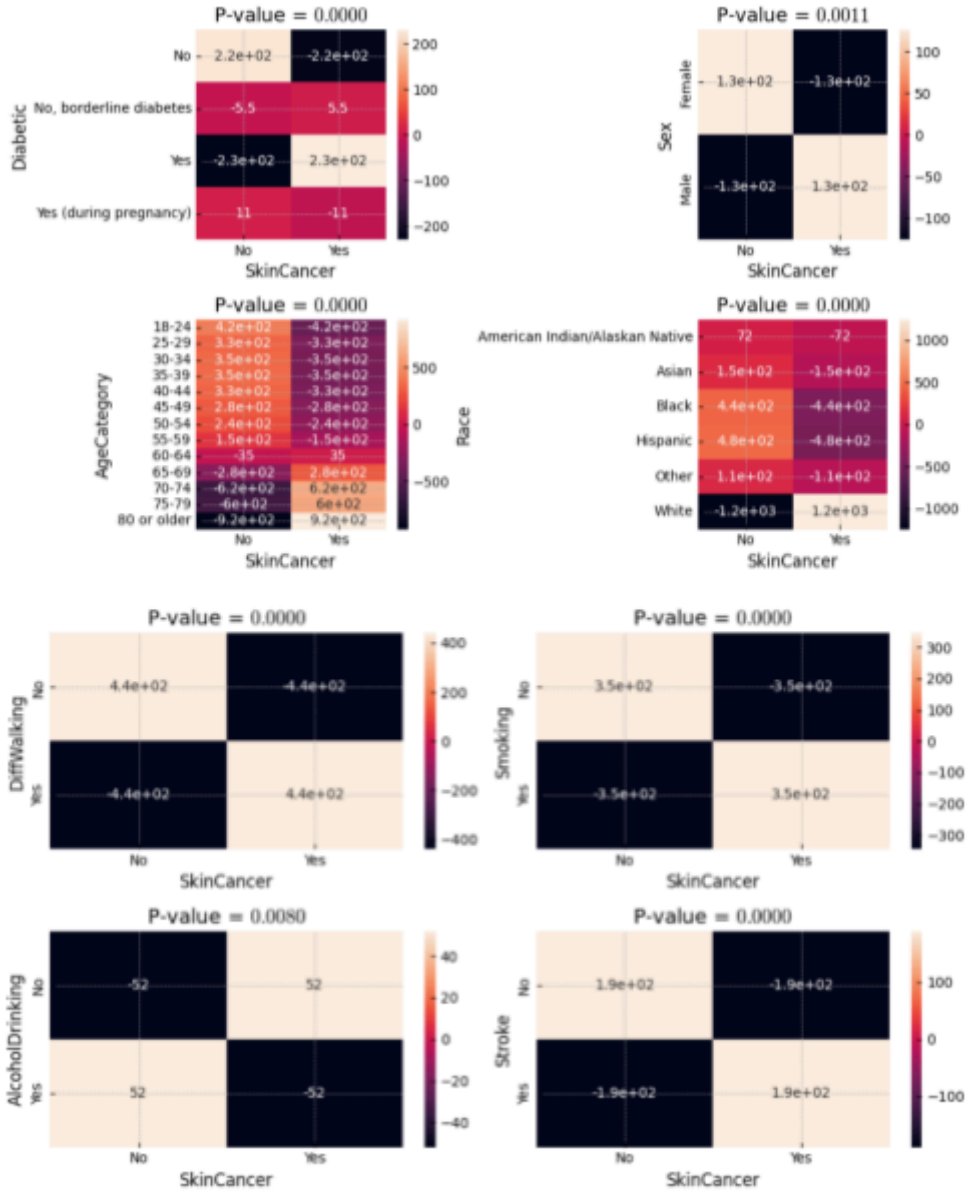


Figure 6a: Chi-Square test Contingency (count) - expected tables.

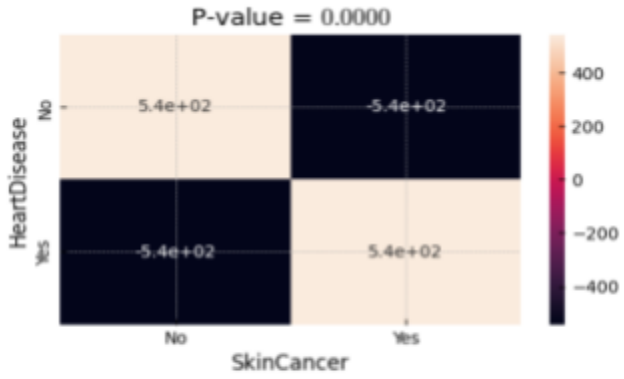


Figure 6b: Chi-Square test Contingency (count) - expected tables.

Part 3

In part 3, the data set was also found on Kaggle.com[3] (see Appendix B). This dataset had tumor measurements for all types of cancers. The goal of the analysis in this section was to find out if the smoothness of a tumor was related to the type of tumor it was.

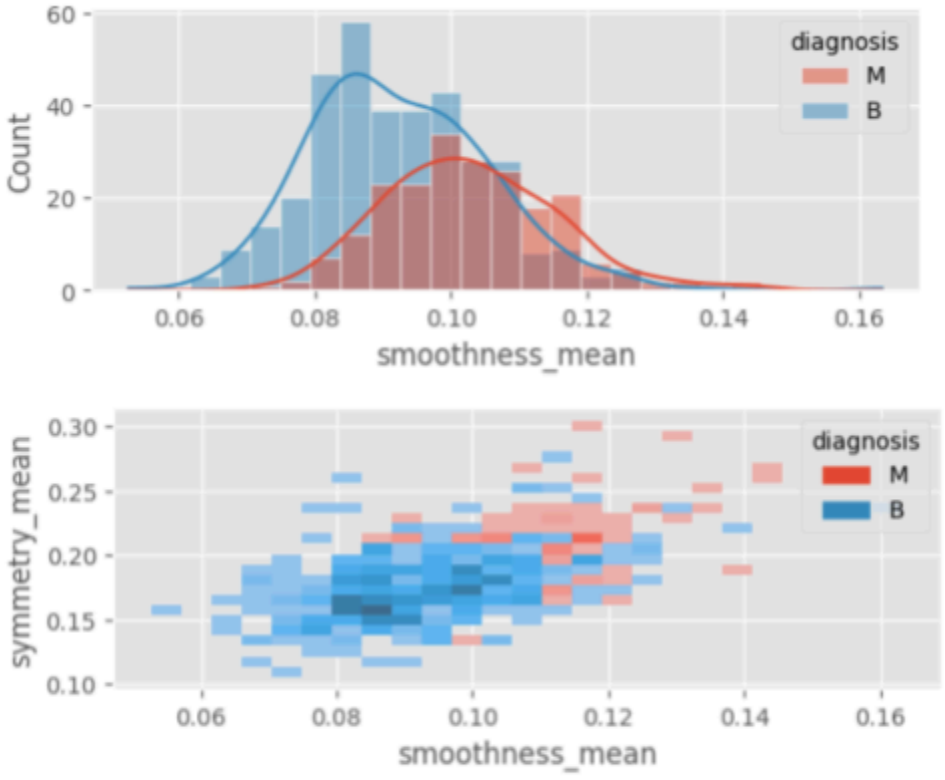


Figure 7: Histogram and 2D histogram of smoothness vs count and tumor symmetry (mean). Broken down into benign and malignant tumors.

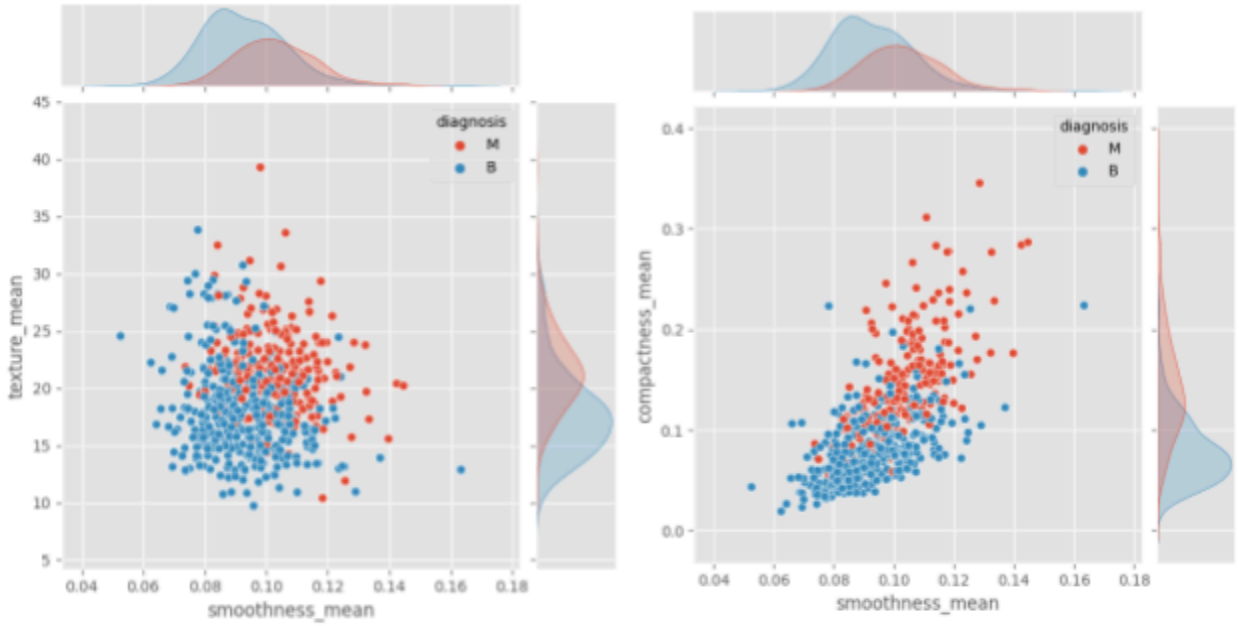


Figure 9a and 9b: Scatter plot with a kernel density estimate plot on each axis

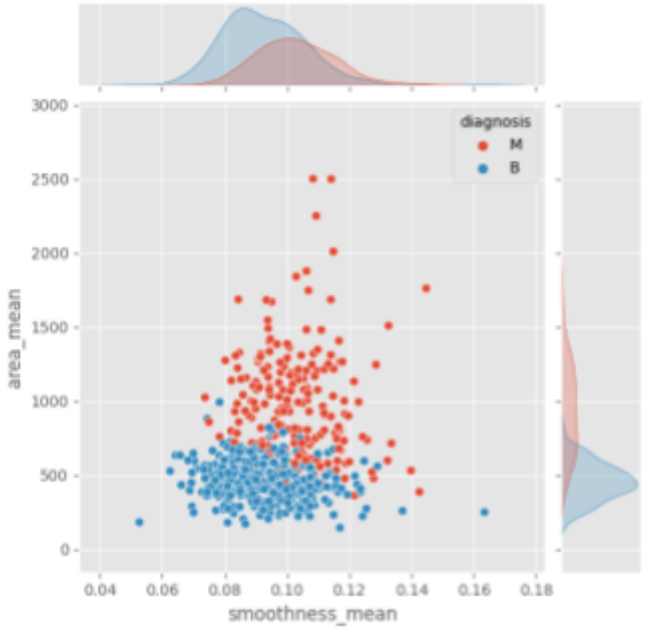


Figure 9c: Scatter plot with a kernel density estimate plot on each axis

Next, to determine if smoothness affected tumor diagnosis a one-way ANOVA test was performed. This test determined the difference in smoothness between benign and malignant tumors. The p-value of the test was significant with a p-value of 0.0000.

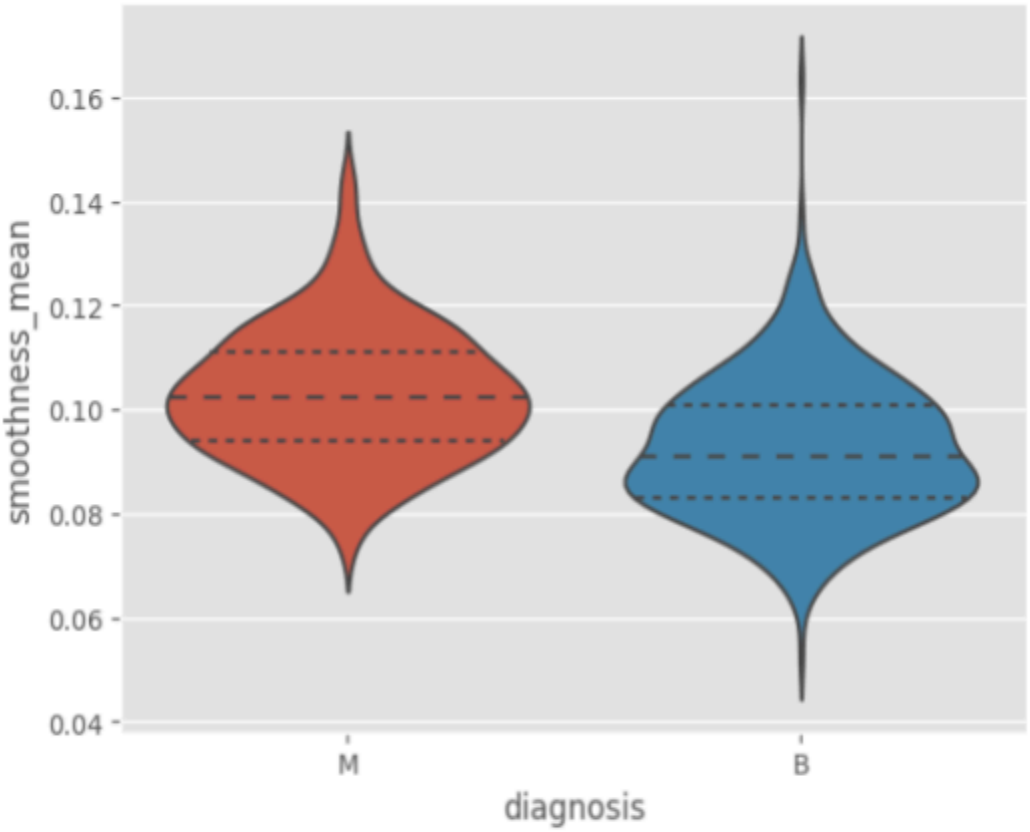


Figure 10: A violin plot of the smoothness vs. the diagnosis. The dashed indicate the mean and IQR

This report was able to delve into what are the relationships between overall health and skin cancer. Preventing cancer by building good habits is one of the important things one should do to stay healthy. Age is one of the largest indicators of whether a person has skin cancer or not. If a person is then diagnosed with cancer this report analyzes tumor measurements and diagnosis as well. Tumor diagnosis displayed a correlation with smoothness. Typically, the more smooth a tumor was the more dangerous it would be. Cancer is a common disease and analyzing the disease can help understand it and prevent it.

5. References

1. Skin cancer . (n.d.). [cgvCancerTypeHome]. Retrieved April 23, 2024, from https://www.cancer.gov/types/skin
2. Faststats. (2024, January 17). https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

6. Appendices

Appendix A

1. project4_dsc440.ipynb A file that contains the raw analysis from this report. It includes the datasets 1, 2, and 3 from Appendix B.

Appendix B

1. https://www.kaggle.com/datasets/antimoni/cancer-deaths-by-country-and-type-1990-2016/data
2. https://www.kaggle.com/datasets/hassaneskikri/brfss-samplecsv/data
3. https://www.kaggle.com/datasets/erdemtaha/cancer-data