Practical Machine Learning: Exploratory Data Analysis

Russell Todd[1]

[1]University of Wyoming

Abstract

1 Introduction

In this exercise, I will be looking at raw datasets before any machine learning is applied to them. I will familiarize myself with the data while looking for potential problems and obstacles for applying machine learning. To see how preprocessing might be helpful, I will select a few preprocessing methods and explore their application to the datasets.

The two datasets I will be looking at are the White Wine Quality dataset and the Primary Tumor dataset which are described below.

1.1 White Wine Quality Dataset

The famous White Wine Quality dataset is comprised of 12 features. The first 11 are measurements of various physical qualities of each wine (fixed acidity, citric acid, residual sugar, pH, alcohol, etc.) and the 12th is the output variable which is a score (0-10) denoting the quality of each wine. There are 4,897 entries (wines) in the dataset and no entries are missing data for any feature. All features are positive numeric values.

1.2 Primary Tumor Dataset

The tumor dataset, which was recommended to me by Dr. Lars Kotthoff, consists of 339 entries with 18 features. The first 17 features describe the status of a human participant (age, sex, potential tumor sites, etc.) and the 18th is called binaryClass and it denotes whether that entry tested positive or negative. All entries are complete with the exception of 67 entries that lack histologicType data and 155 entries that lack degreeOfDifferentiation data. 5 features are categorical variables that are either boolean or ordinal in nature (age, sex, histologicType, degreeOfDifferentiation, and binaryClass) with the rest being boolean variables. The exact meaning of any of these features that correspond to sites on the body being positive or negative is unclear to me. My assumption is that it denotes whether there is a tumor present at that site. I further assume that binaryClass denotes whether that tumor is cancerous or not. I couldn't find any information on the dataset to clear up these questions, so some of the exploration of this dataset is aimed at answering those questions.

1.3 Selected Methods

For both datasets I tried to follow a similar process. The first step is alsways getting the data properly loaded into a pandas dataframe and ensuring that the datatypes of each feature is appropriate. Afterwards, several summary statistics panes are produced using built-in functions of the pandas dataframe class like info(), dtypes, and describe(). I then use these to double check that the data loading process was done correctly. If not, I then corrected them manually.

I was interested in the differences between the different feature transformation methods (scalers) and so decided to do a comparison of 5 of the most popular ones (StandardScaler, MinMax Scaler, MaxAbs Scaler, Quantile Transformer, and Robust Scaler). I then produced 3 plots (histogram, correlation heatmap, and pairplot) for the raw dataset and then each of the scaled datasets. While

this worked fairly well for the wine dataset, it didn't seem to be very meaningful for the tumor dataset. I then visually compared the graphs, looking for differences notable enough to be seen by the human eye.

While I was interested in the feature transformation functions, I now believe that exploring undersampling and oversampling would have been a better approach. Undersampling and oversampling would have meaningful applications for both datasets despite their different datatypes.

In the case of the tumor dataset, I also produced some plots intended to find the meaning of the positive and negative values for the site features and binaryClass. This was mainly done by splitting the dataset into a binaryClass positive dataset and a binaryClass negative dataset. I then produced frequency plots to see if there was something I could notice.

The results of these methods, namely the plots and associated observations are included below.

2 Results of Analysis

2.1 White Wine Quality Dataset

2.1.1 Before Preprocessing. The White Wine Quality dataset is quite imbalanced with relatively few entries for low quality and high quality wines. It has two variables, density and residual sugar, that are highly correlated so it might be worth dropping density before training models as it is less correlated to the score than residual sugar.

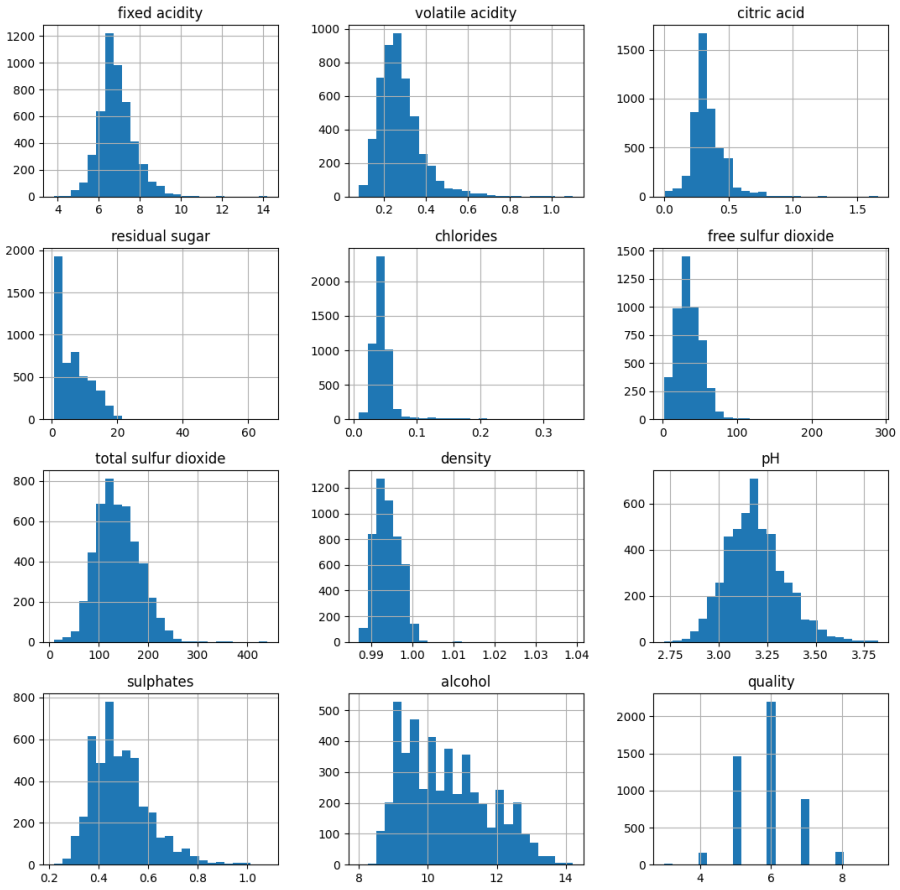

Figure 1: Histogram of Wine dataset features

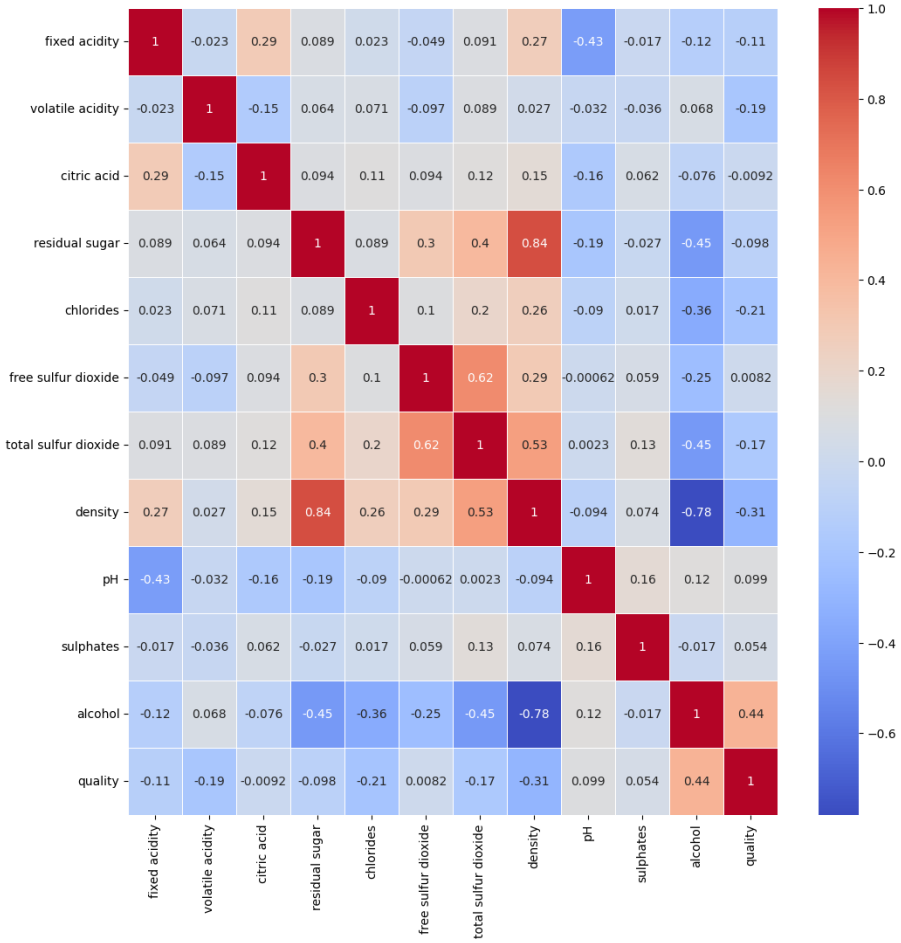

Figure 2: Correlation Heatmap of Wine dataset features

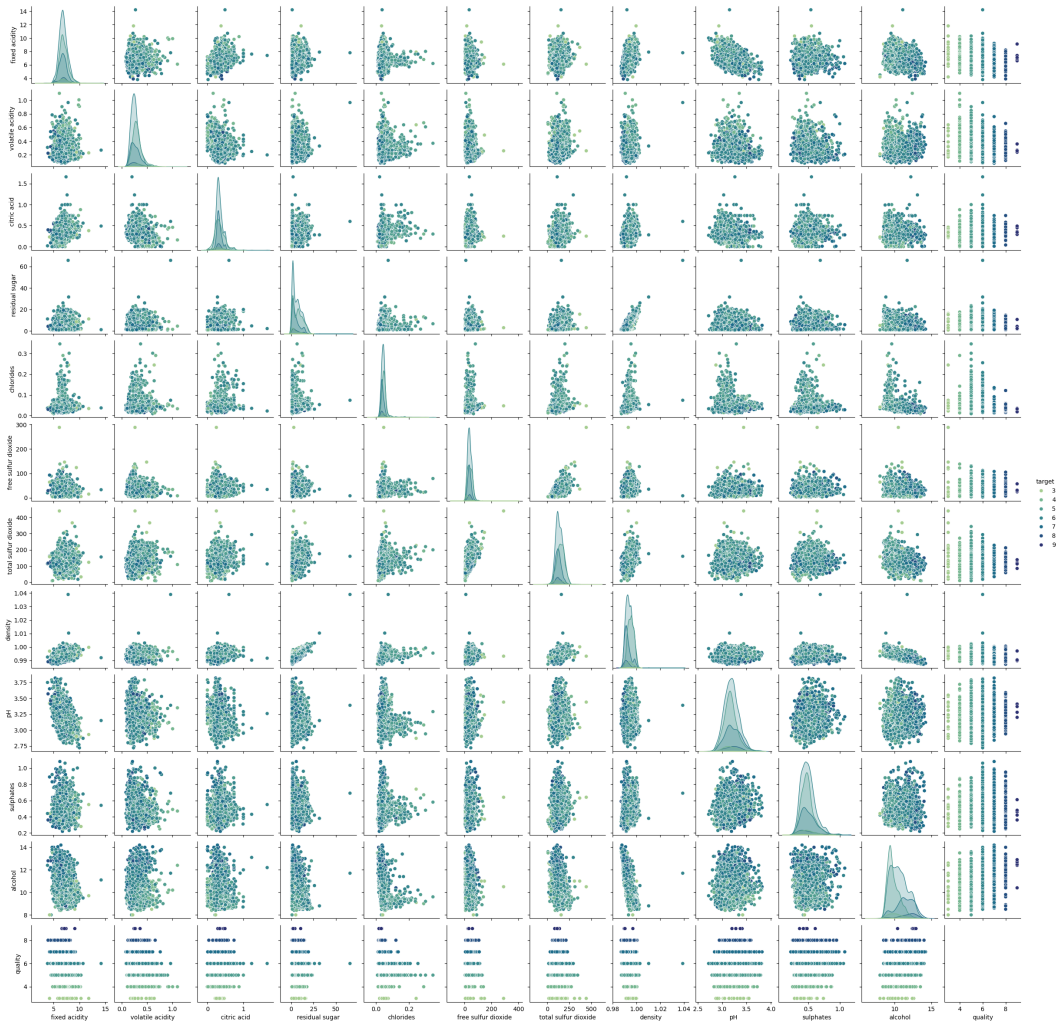

Figure 3: Pairplot of Wine dataset features

2.1.2 **After Preprocessing**. For all three plots (histogram, correlation heatmap, pairplot), the only scaler that produces a notably visually different output was the Quantile Scaler which is shown below. All of the other scalers produced changes that were more subtle or at least indistinguishable in these plot formats.

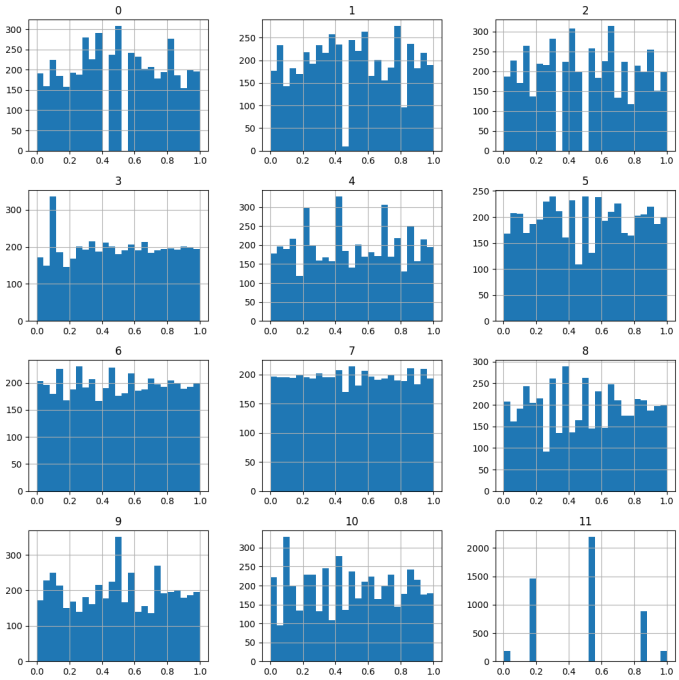

Figure 4: Quantile Scaler Histogram

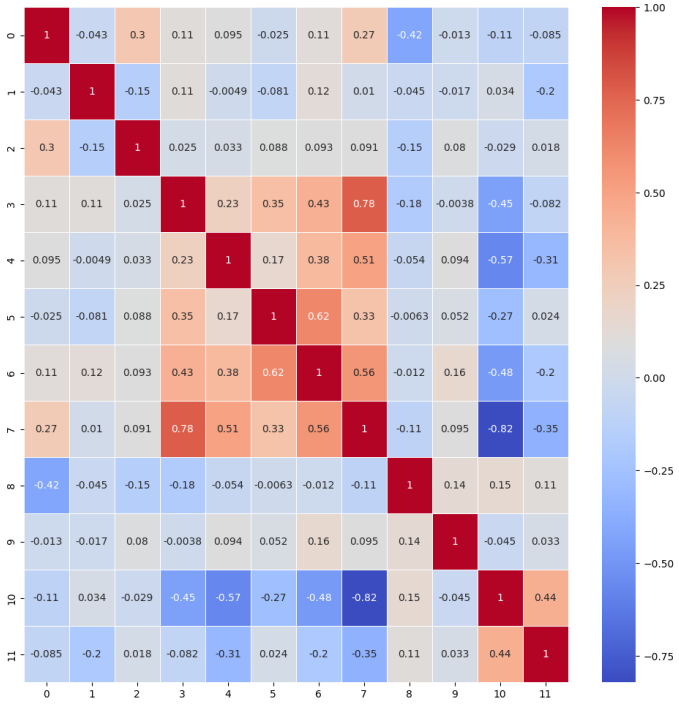

Figure 5: Quantile Scaler Correlation Heatmap

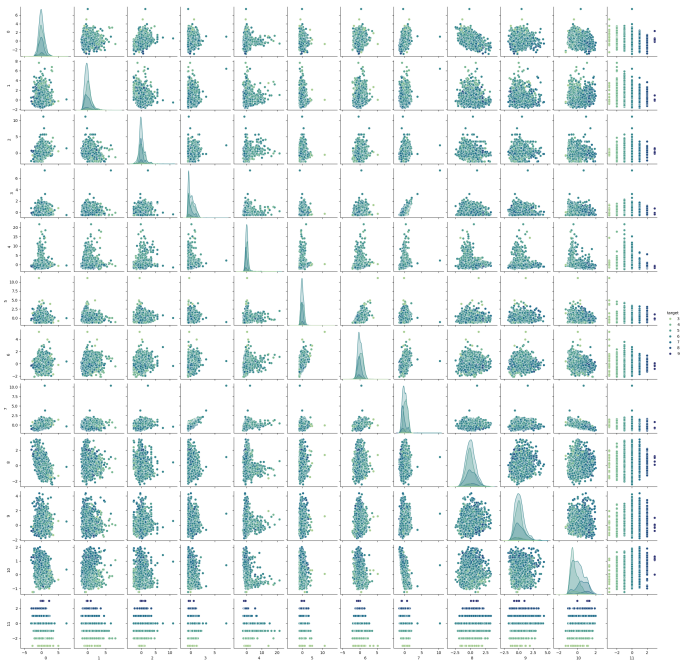

Figure 6: Robust Scaler Pairplot

2.2 Primary Tumor Dataset

2.2.1 Before Preprocessing. The tumor dataset is fairly imbalanced with far more binaryClass negative entries than positive ones. There are only a few features that are loosely correlated, namely abdominal and liver which makes sense considering how close those two sites are to each other. Mediastinum was the feature most closely correlated to the binaryClass.

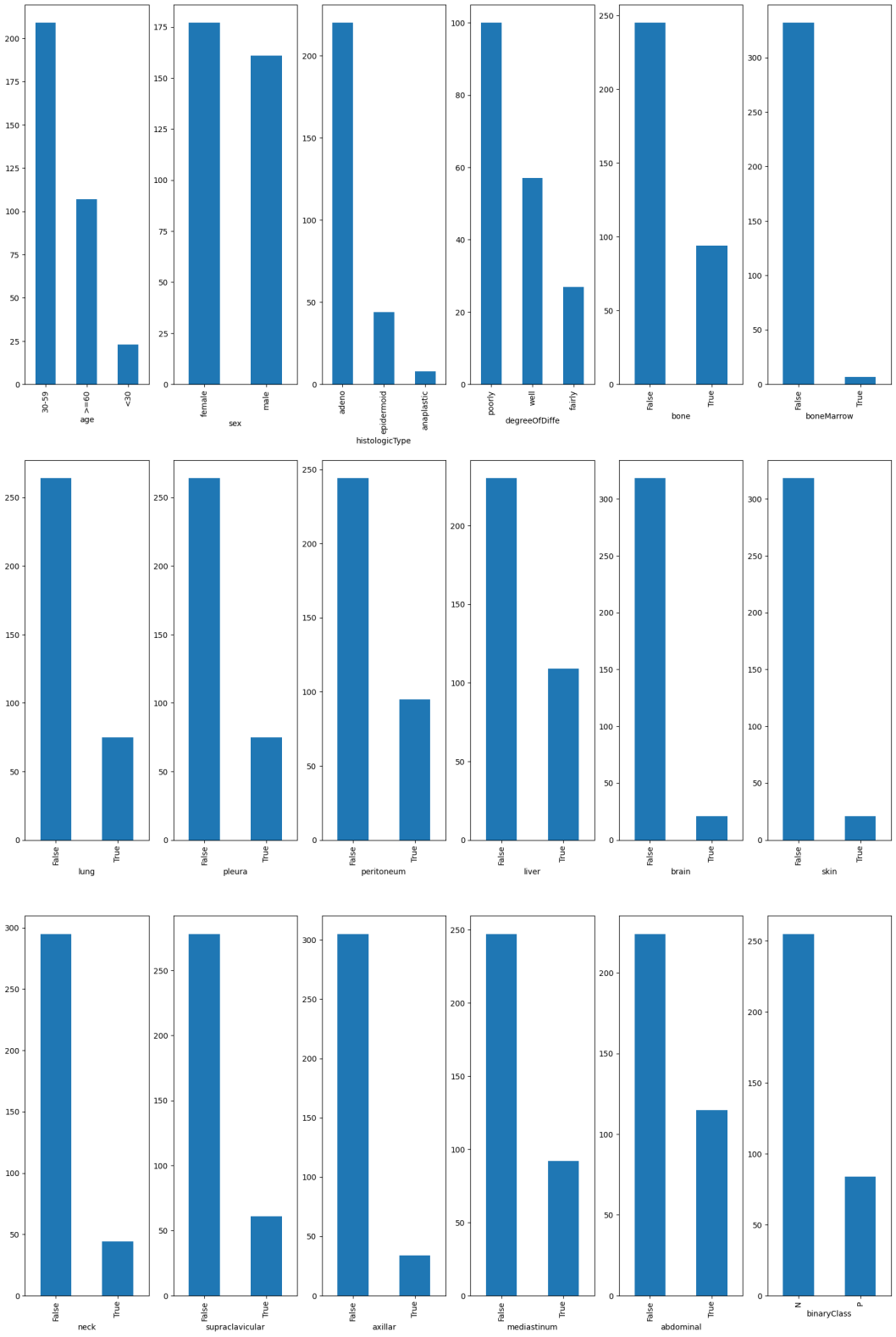

Figure 7: Frequency chart of Tumor dataset features

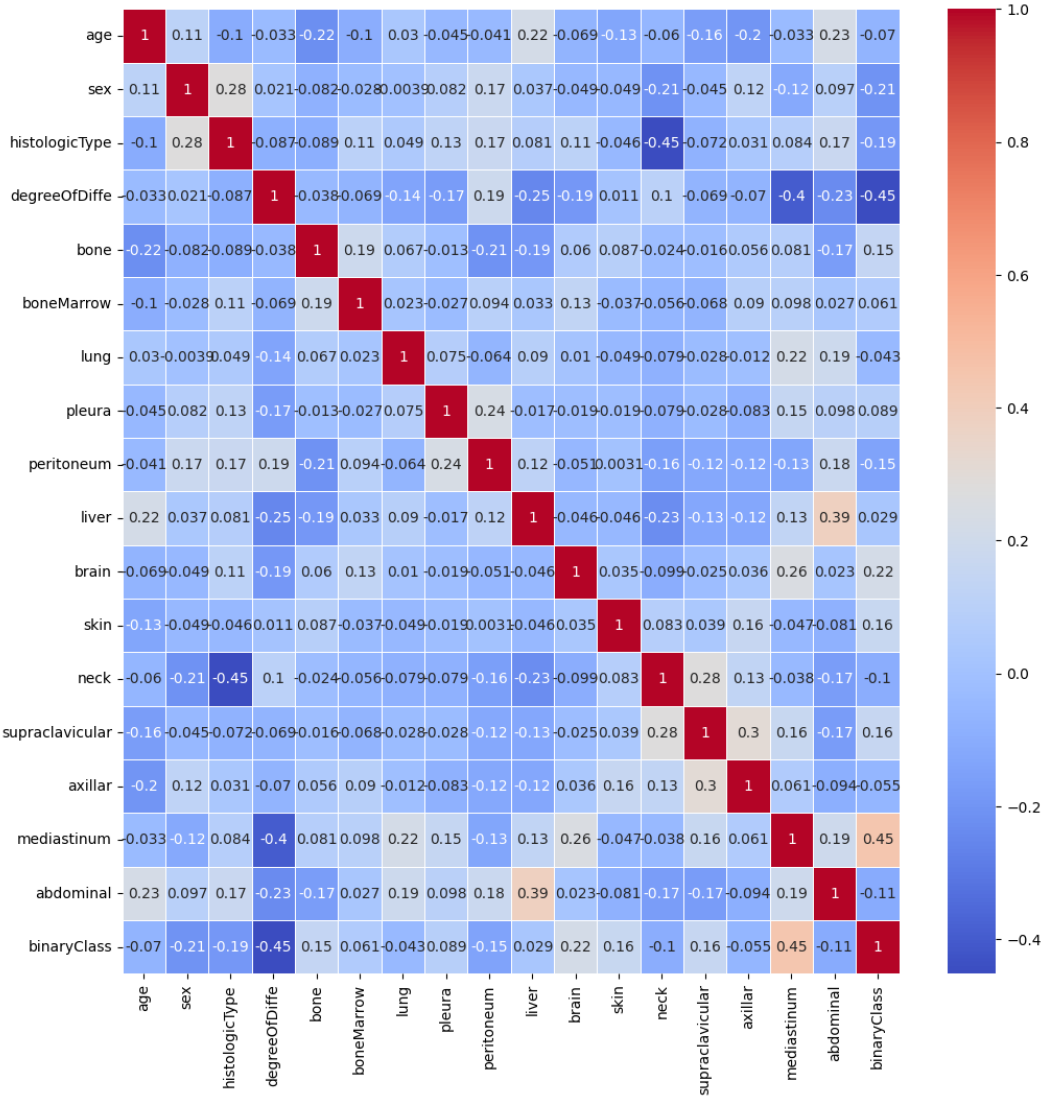

Figure 8: Correlation Heatmap of Wine dataset features

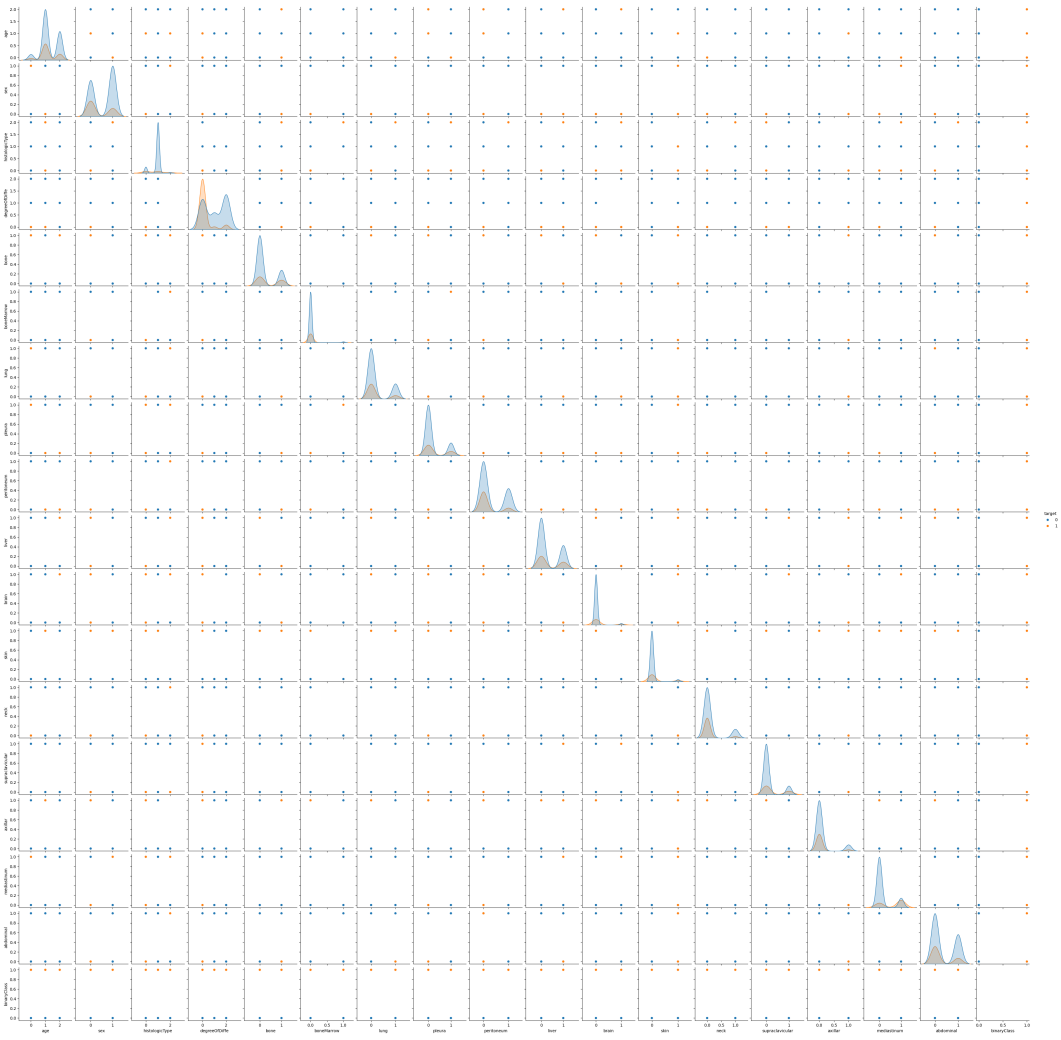

Figure 9: Pairplot of dataset features

2.2.2 **After Preprocessing**. For the tumor dataset, once again, the quantile scaler was the only feature transformation tool that produced a visually different graph. However, this time it was only the histogram that was visually different. The heatmaps were all nearly identical and were so omitted. The pairplots were not meaningful and were also omitted. However, frequency plots for the positive and negative entries are shown below. They are what led me to believe that the positive/negative state of the binaryClass variable denoted whether the tumor was cancerous as opposed to whether or not the entry was from a patient that had a tumor or not.

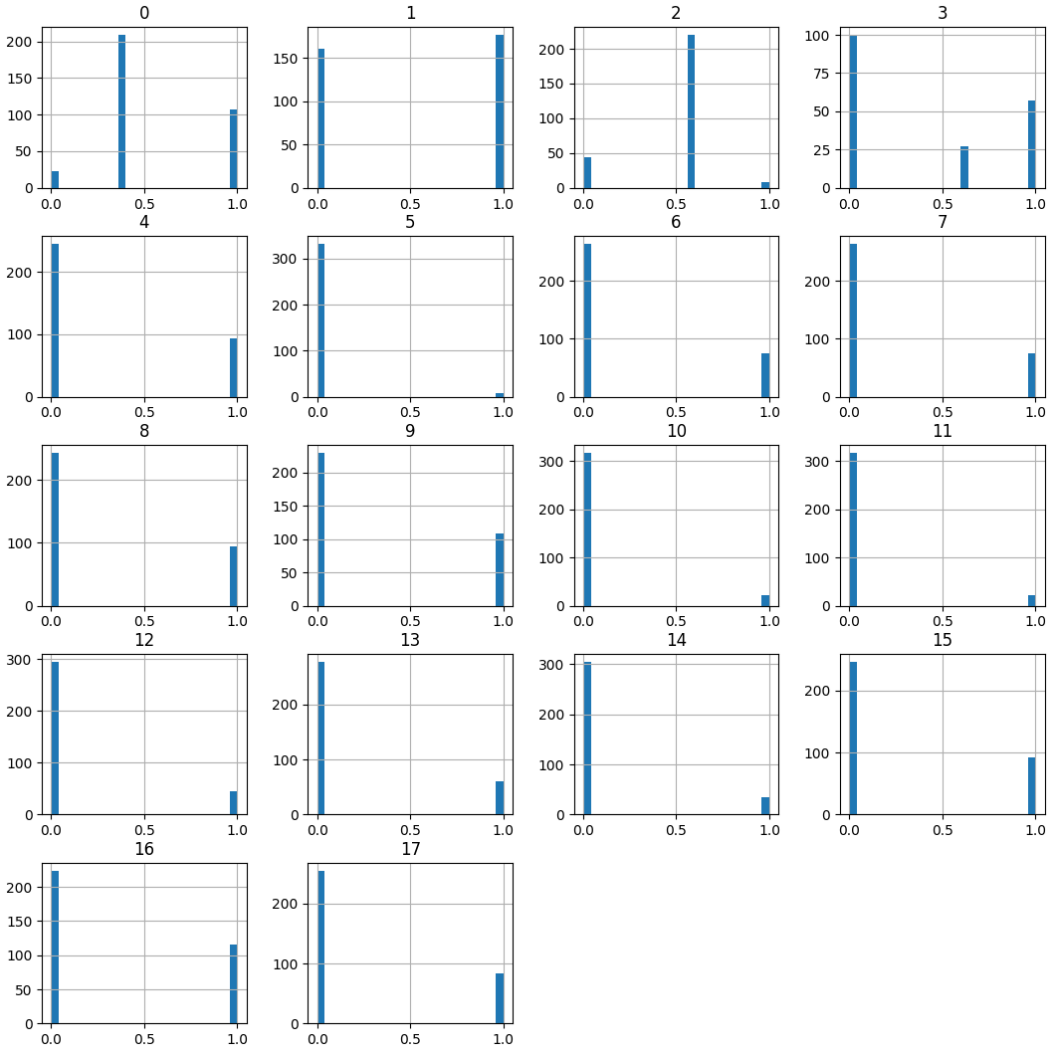

Figure 10: Quantile Scaler Histogram

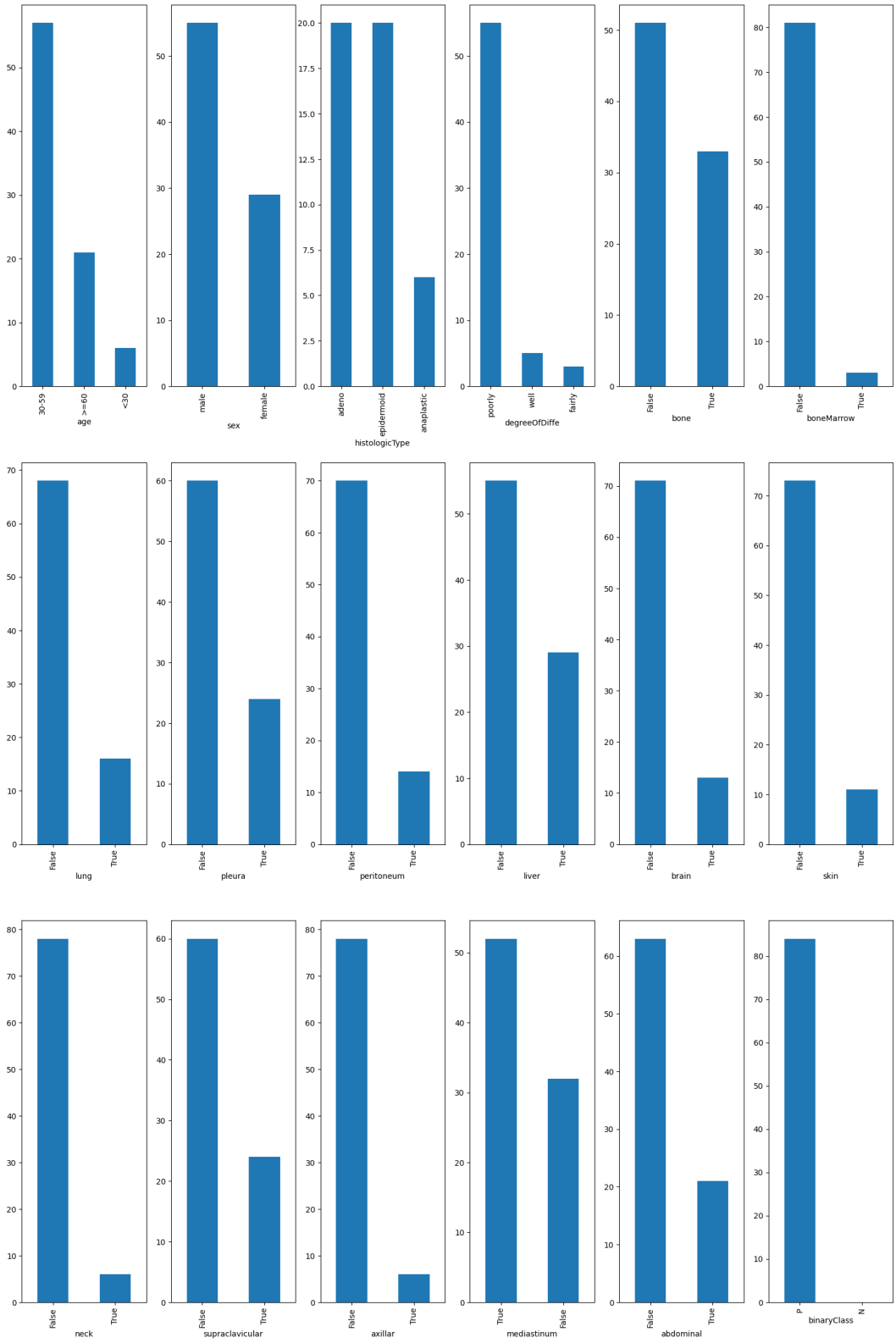

Figure 11: Positive Frequency plot

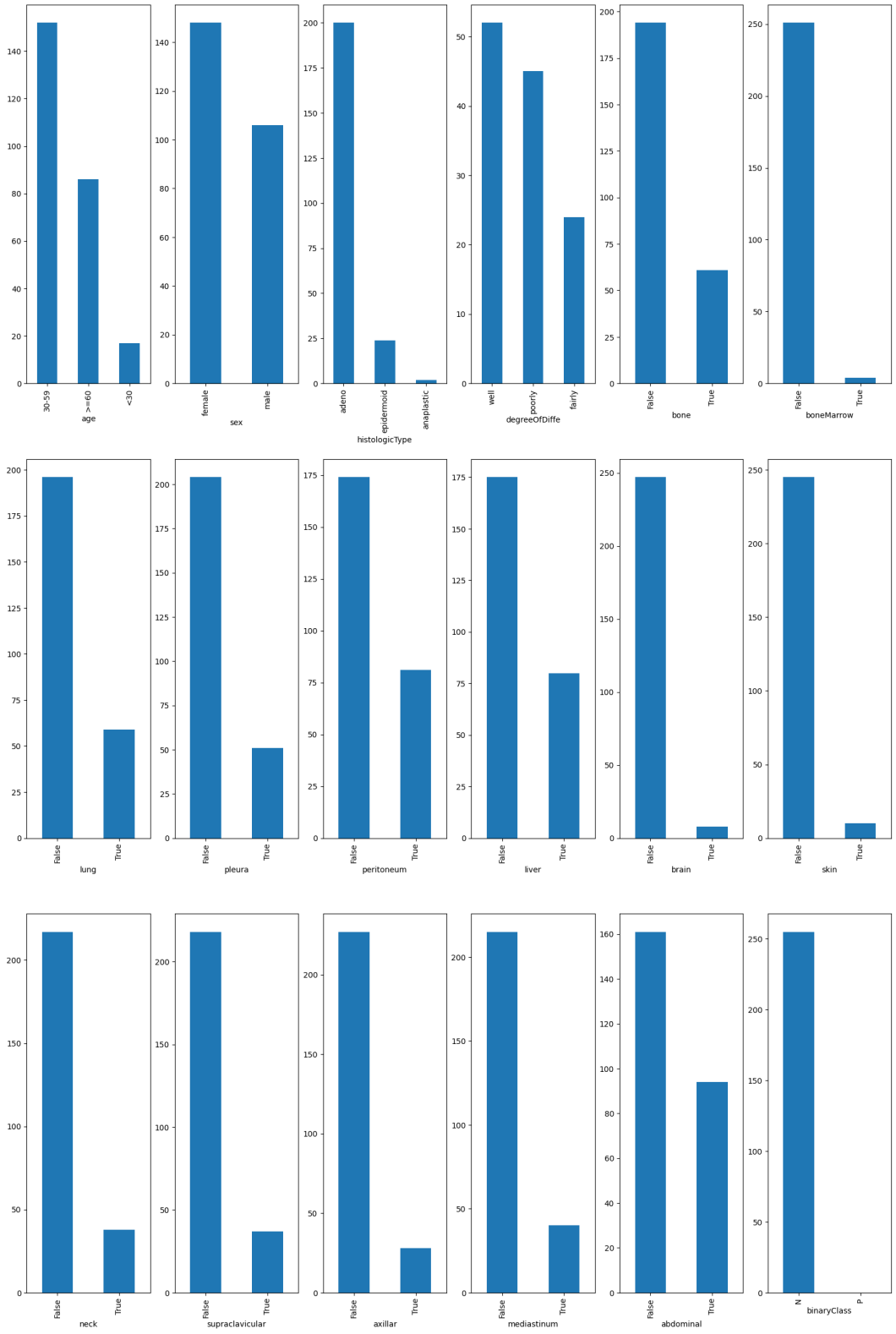

Figure 12: Negative Frequency plot