Chanel Thorpe
DS219 Spark-Seprep
30 October 2024

Data Analysis Report

Importing Necessary Libraries and Data

I imported pandas primarily for the data frame creation and analysis. I imported various different packages from the sklearn package such as the train_test_split, and cross_val_score in order to complete the predictions for hypothesis two. I imported metrics to calculate the model score on the validation and test sets for hypothesis two. I also imported mean_squared_error and mean_absolute_error to calculate those based off of my model to gain a sense of accuracy level of my numbers and the overall model score. Finally, I imported matplotlib.pyplot and seaborn in order to create some of the visualizations I included such as the histograms, the box plots, and the correlation matrix. I started off my analysis by taking a look at the dataframe for both red and white wine to get a sense of what variables I'd want to take a look at by using the .head() function. Although I could've gotten rid of the columns I knew I wasn't going to need, I decided to keep them since I knew I wanted to look at a correlation matrix which would tell me a lot more about the data if all the columns were included.

Exploratory Data Analysis

For this part of my analysis, I decided to split the two dataframes and take a look at them separately. I first displayed the info using .info(). I will say it didn't really tell me anything except for the data type of each column, the total amount of items, and the amount of null values which were none for each column. Because of this, I knew I needed to look at the data in a different way. I decided to use .describe() to find some basic statistical information about each column. By doing so, I found the count, mean, std, min, 25%, 50% (median), 75%, and the max. This told me a lot more about what I was looking at and it also gave me a better idea of what I

would be interested in for the hypothesis section. For the visualizations, I decided to look at histograms and a boxplot. The histograms were comparing different variables I chose on the x axis and the count (for each value on the x axis) was on the y axis for each graph. It was interesting to see what the most common amounts of things were in wine. For red wine, the most common amount of residual sugar used at a count of about 260 was slightly under 2.5 (figure 1). The density varied quite a bit for the red wines and actually seemed to have a bit of a gaussian distribution which was the similar case for the alcohol however it looks like it is skewed to the left a bit. The box plot for red wine is very symmetrical, looking surprisingly. Looking at the quality distribution histogram, it makes sense. There are only a few numerical values for quality and a vast majority of them are 5 and 6. The median being 6 makes perfect sense, seeing how it is also the 75%, there is no line in the middle of the box plot like we would expect. What's interesting about the residual sugar for the white wine is that it is on a different scale than the red wine so it looks like the residual sugar is lower than the red wine. This actually isn't necessarily true. While it looks like red wine might have more, white wine actually has more. The x scale for red wine goes just above 15 while the scale for white wine goes above 60 (figure 2). The concentration of the most values is between 0 and 20 which may imply that it is the same as red wine. By looking at it, you can see that there is much more variation but also a lot more residual sugar in white wine based on the data. The density score is actually pretty similar, it just looks different because of the scale. The quality box blot for white wine is actually pretty much the same; it just has more outliers.

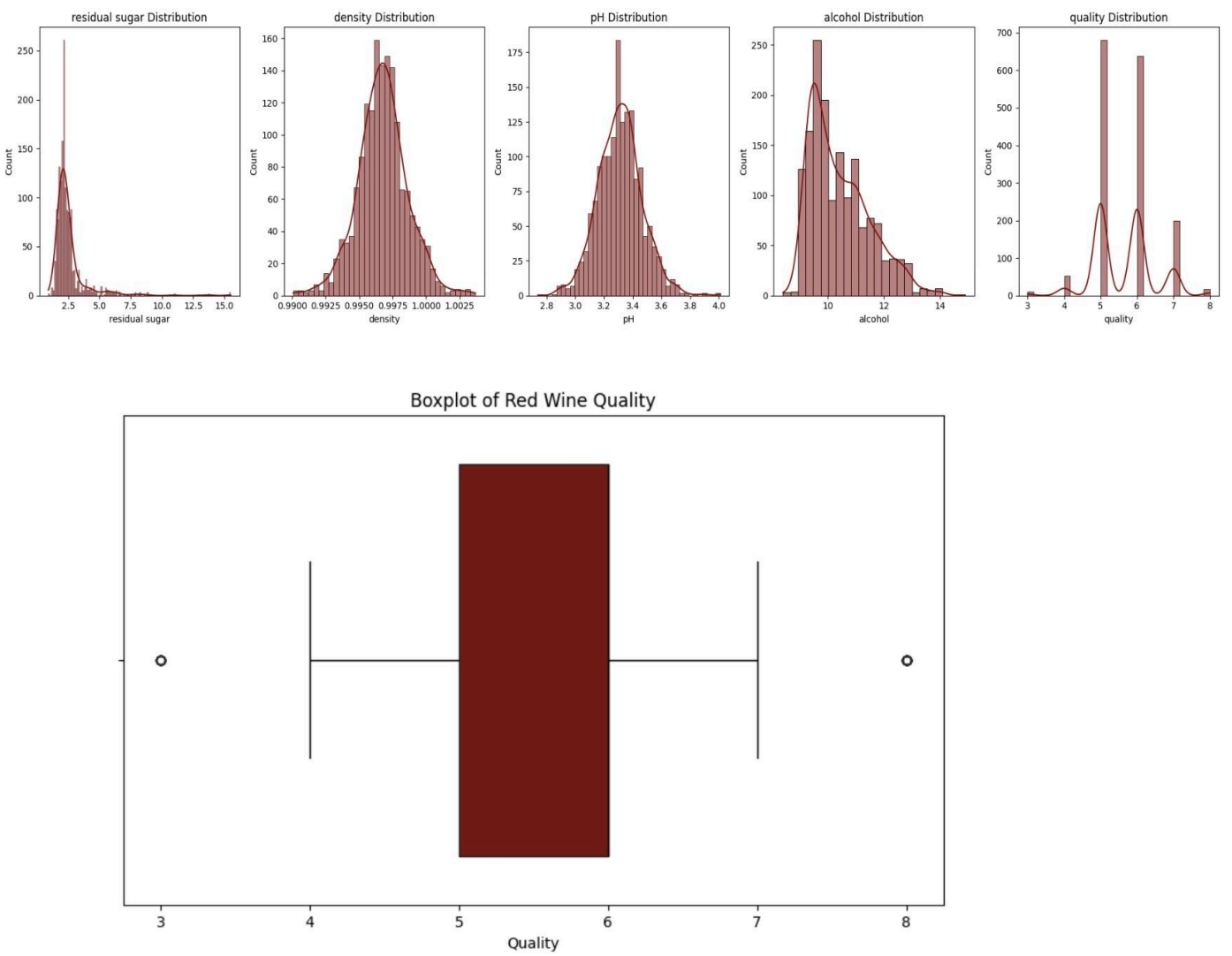

Figure 1: Histograms and Boxplots for Red Wine Dataframe

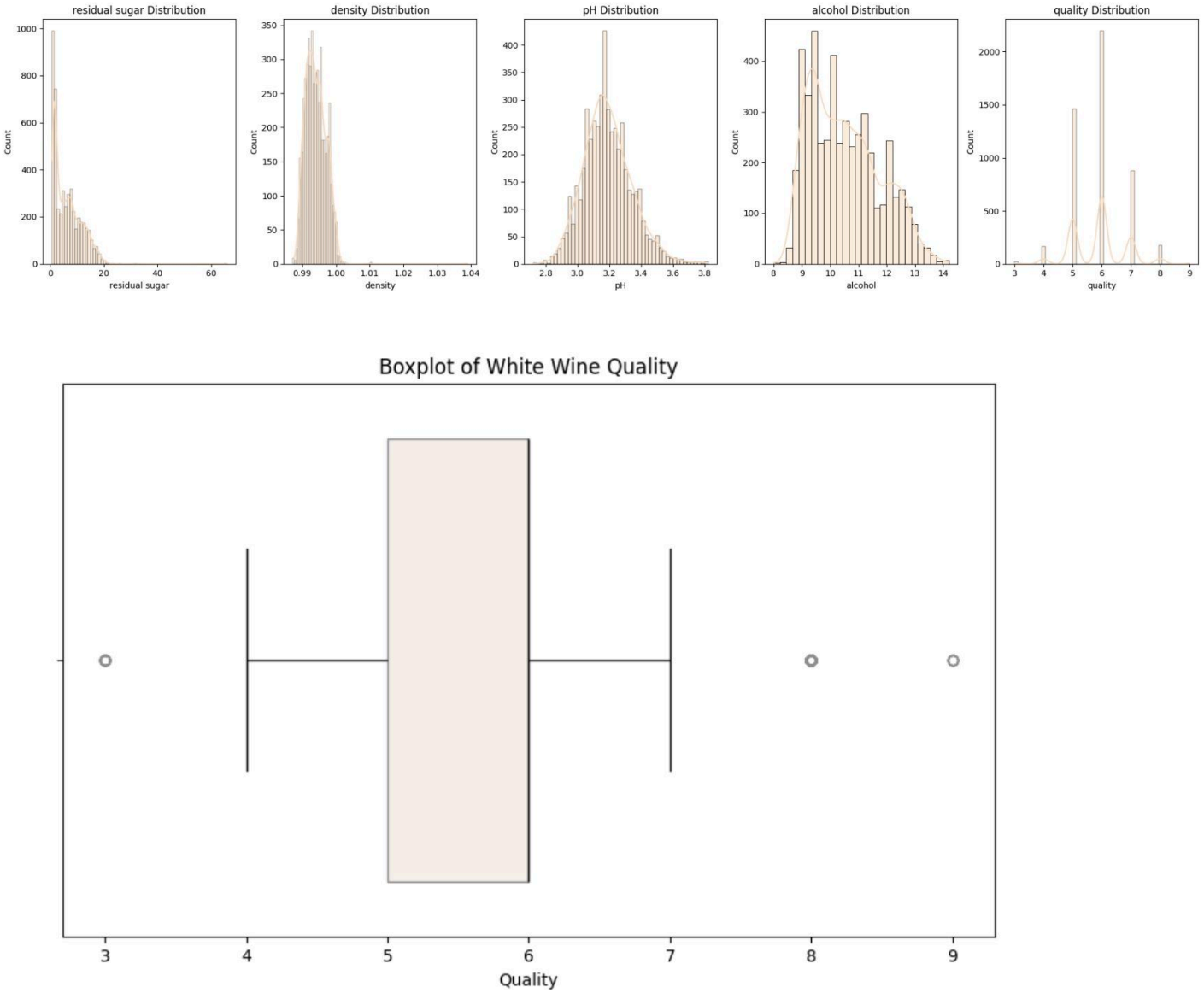

Figure 2: Histograms and Boxplots for White Wine Dataframe

Hypothesis Formulation

There is a correlation between residual sugar and density in white wine. I came up with this hypothesis based heavily on the correlation matrix for the white wine. I decided to look into this because the correlation score on the matrix is very close to one. **The wine quality is greater the more alcohol content is in the wine for both red and white wine.** This hypothesis is based on both the box plot and the correlation matrix. The correlation score for both red wine and white wine for alcohol content and quality is about 48% and 44% respectively. The box plot shows the average of the alcohol for both red wine and white wine to be about 10.5 and the quality in the description analysis to be about 5.6 for red wine and 5.8 for white wine.

Hypothesis Testing

In order to test my hypothesis of "there is a correlation between residual sugar and density in white wine", I decided to build a correlation matrix. It;s the simplest and easiest way to show whether or not my hypothesis is true. By looking at figure 3, you'll find that the correlation score between density and residual sugar is 0.84 indicating that there is a strong correlation between the two. While looking at the correlation matrix, I noticed that other variables that may be interesting to look at were the alcohol and quality of the wine. Although I had constructed my hypothesis before, I wasn't sure how I was going to test this hypothesis. I figured the most useful way would be to construct a prediction model.

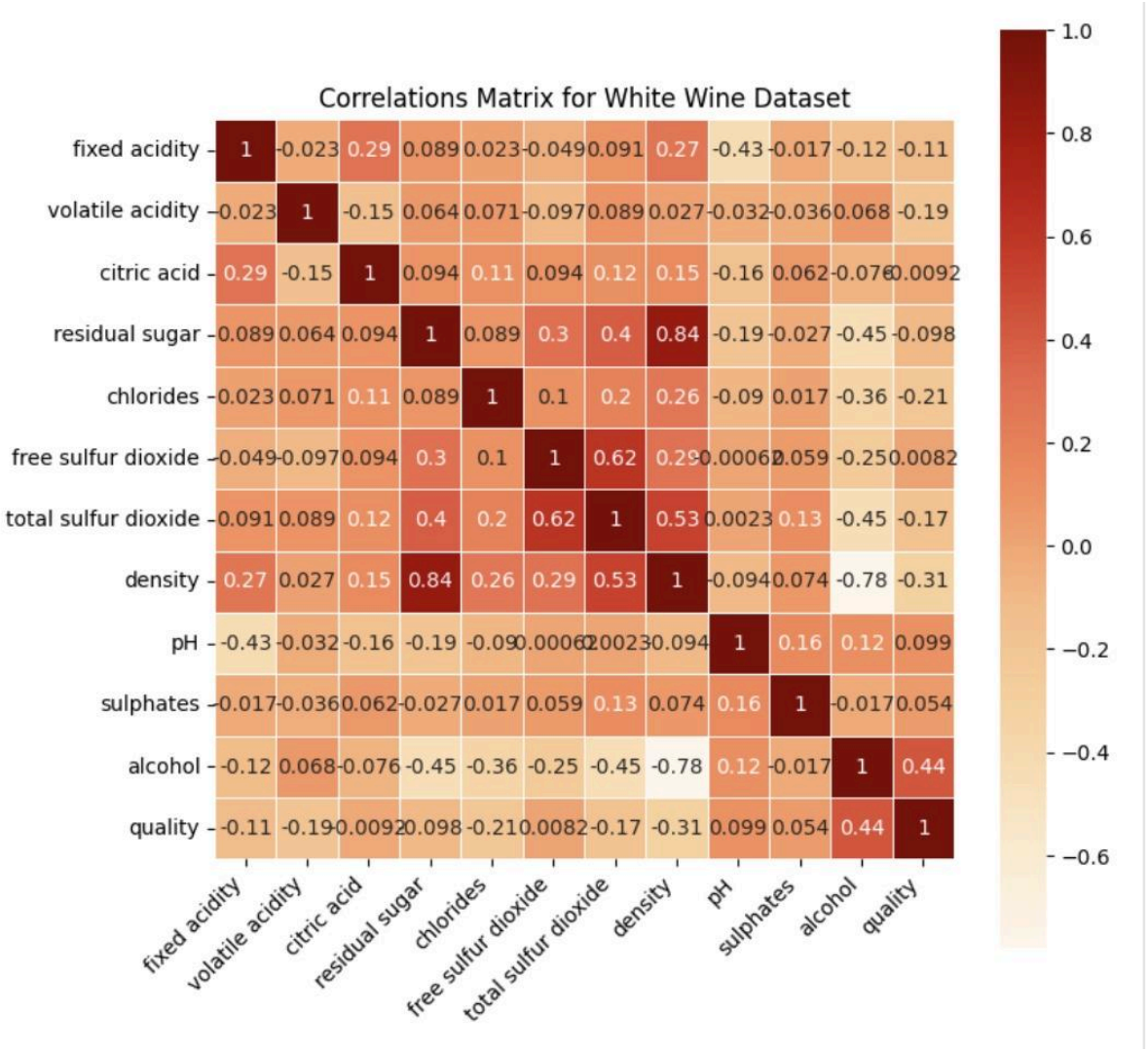

Figure 3: A correlation matrix of the white wine dataframe

For the second hypothesis, "the wine quality is greater the more alcohol content is in the wine for both red and white wine," I decided to make a prediction model that would tell me pretty much whether or not the quality would improve if the alcohol content was higher. For both of the wines, I took the averages of each of the columns and put it as the values for the new row values except for the alcohol. For both the red and white wine, I put a random high number for alcohol value. For red wine, the predicted quality value would be 6.2, for white wine, it predicted 6.81. Both of these values are quite high but they do not surpass the max number that we saw

earlier in the description analysis of the tables. The max number of quality for red wine was 8 and white wine was 9. This could be for several reasons but to paint a better picture, I printed out the scores for the validation set and the test set as well as the MSE and the MAE for both red and white wine. For red wine, the validation set got a score of .478 which is pretty low. The test score got an even lower score of .216. This indicates that the model itself isn't really learning well nor is it performing well. The MSE and the MAE are quite low however with scores of 28.59% and 39.83% respectively, indicating that the error of the model is not terrible. I used both MSE and MAE because I know MSE can be a little tricky and hard to interpret sometimes. The white wine prediction performed slightly worse than the red wine with MSE and MAE scores of 45.89% and 48.49%. But the model score for both the validation set and test set was a bit higher.

Conclusion

This analysis demonstrates a comprehensive approach to understanding and predicting wine quality based on the attributes, residual sugar, density, and alcohol content. Through hypothesis formulation and testing, we found significant correlations—such as between residual sugar and density in white wine. However I don't think the same could be said for the second hypothesis. The prediction model did not really support the hypothesis and it leaves me with more questions. If I were to do this experiment again, I would look at various different attributes and how they may play into the quality of wine. I do not think it is just alcohol that impacts the quality of the wine which is probably why the quality score didn't end up surpassing the max scores for the two wine colors.