# CS4300 - Project Proposal

Andrea Smith

September 27, 2019

# Contents

# List of Tables



# List of Figures

# 1 Dataset

The dataset for this project applies a rating to wine quality based on physico-chemical inputs, such as acidity and alcohol content. By applying a regression analysis to the dataset, we can determine if the quality of the wine has a relationship with its physical and chemical properties. The data all relates to a specific type of wine, but without the specifics of the type of grape used, the brand, the price, etc.

The dataset can be found here:
https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/version/1

# 2 Output Data Visualization

The output data is a numeric rating of the wine quality. Within this datset, the output ranges from 3 to 8 and is represented as integer values. The lower values represent poor quality wine, while higher values represent excellent quality. Normal or average wines will measure in the 5 or 6 rating. The ratings in this dataset are not uniformly distributed, as average and normal wines are over-represented within the dataset. However, the distribution is within the standard bell-curve that would be realistic.

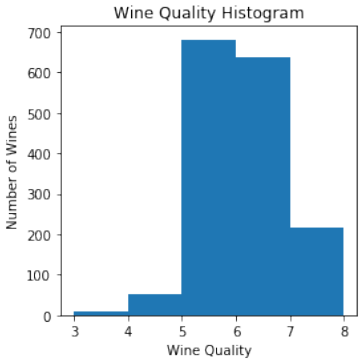


Figure 1: Wine Quality Distribution

# 3 Input Data Visualization

From a visual representation, there is no distinct relationship between a single input variable and the output variable.

Table 1: Input Feature Statistics

| | Min | Max | Median | Mean |
|---|---|---|---|---|
| Fixed Acidity | 4.60 | 15.90 | 7.90 | 8.32 |
| Volatile Acidity | 0.12 | 1.58 | 0.52 | 0.53 |
| Citric Acid | 0.00 | 1.00 | 0.26 | 0.27 |
| Residual Sugar | 0.90 | 15.50 | 2.20 | 2.54 |
| Chlorides | 0.01 | 0.61 | 0.08 | 0.09 |
| Free Sulfur Dioxide | 1.00 | 72.00 | 14.00 | 15.87 |
| Total Sulfur Dioxide | 6.00 | 289.00 | 38.00 | 46.47 |
| Density | 0.99 | 1.00 | 1.00 | 1.00 |
| pH | 2.74 | 4.01 | 3.31 | 3.31 |
| Sulphates | 0.33 | 2.00 | 0.62 | 0.66 |
| Alcohol | 8.40 | 14.90 | 10.20 | 10.42 |

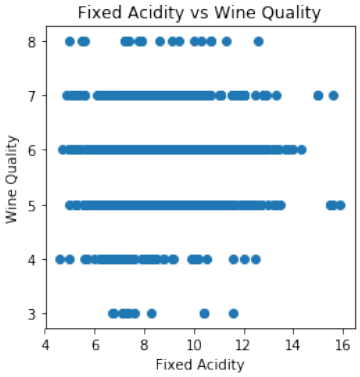


Figure 2: Fixed Acidity vs Wine Quality

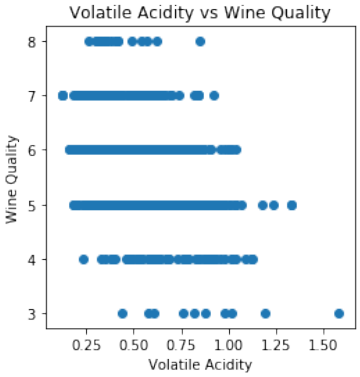


Figure 3: Volatile Acidity vs Wine Quality

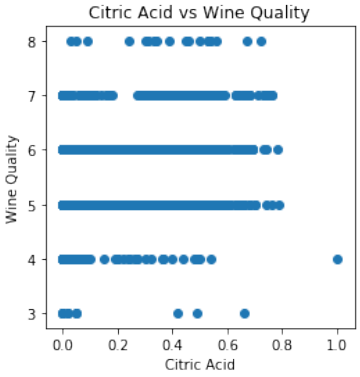


Figure 4: Citric Acid vs Wine Quality

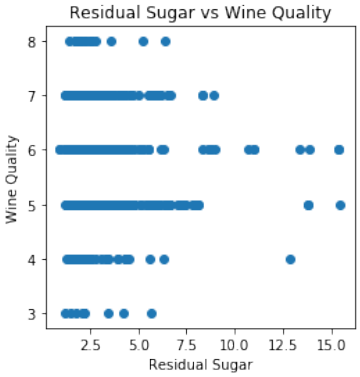


Figure 5: Residual Sugar vs Wine Quality

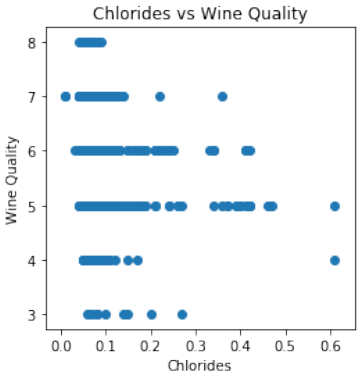


Figure 6: Chlorides vs Wine Quality

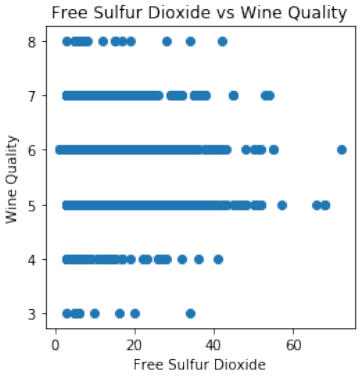


Figure 7: Free Sulfur Dioxide vs Wine Quality

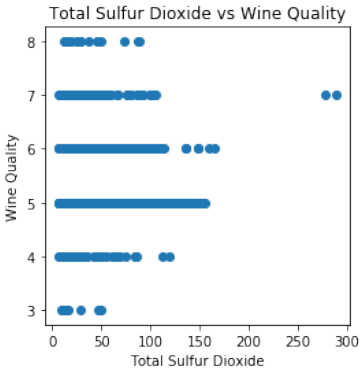


Figure 8: Total Sulfur Dioxide vs Wine Quality

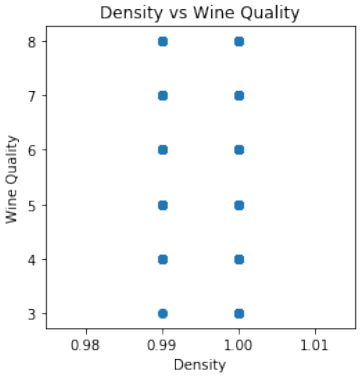


Figure 9: Density vs Wine Quality

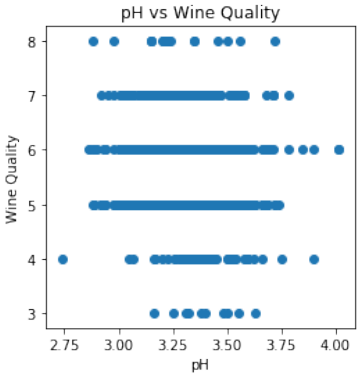


Figure 10: Ph. vs Wine Quality

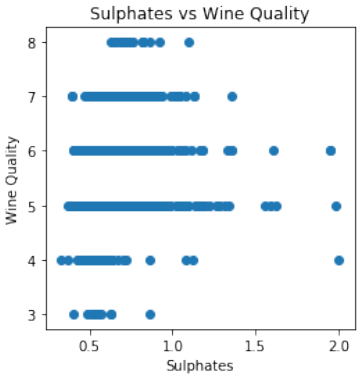


Figure 11: Sulphates vs Wine Quality

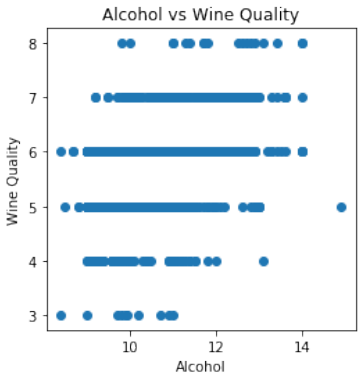


Figure 12: Alcohol vs Wine Quality

# 4 Data Analysis

## 4.1 Normalization

In order to normalize the data, each of the input variables will be normalized based on their range, known as Min-Max Normalization. For this method, the following equation can be employed on each input variable:

$$x^{'} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

## 4.2 Data Splitting

The data consists of 1600 entries to be split across the training data, validation data, and test data. Half of the data will be used for training, an eighth of the data will be used for validation, and the remainder will be used for testing. This pushes the bulk of the data into the training phase, to help develop the most accurate model.

## 4.3 Neural Networks

We will start with a feed-forward network based on multivariate linear regression. Our loss function will be mean squared error, which will also be used to evaluate the test set results.

## 4.4 Model Evaluation and Improvement

The test set will use mean squared error to evaluate how well the model represents the data and it's ability to accurately predict the outputs. A low mean squared error between predicted and actual outputs of the dataset implies that the model is an accurate representation of the dataset.

To improve the model, we can change the data normalization process to a Z-score (standardization) method. This method is more complex than the aforementioned Min-Max method, but is also more widely used in machine learning. Additionally, we can move more of the data into the training and validation sets, and a smaller proportion in the test sets.