Wine Quality Prediction

By: Twisha Jain (tj2481), Rakshith Kamath (rk3165), Rahulraj Singh (rs4211), Yue Zhang (yz4155), Kechengjie Zhu (kz2407)

Introduction

Through this project, we will be aiming to predict the quality of wine based on the proportion of the ingredients that are used to ferment and make the wine. We will be studying 11 feature variables, all of which directly affect the quality of wine. The output, that is, wine quality is rated on a scale of 1 to 10 with 1 representing a low quality and 10 representing highest quality.

- In the initial stages of data exploration and cleaning we have studied relationships between the variables and their respective affect on the target variable.
- We have identified the missing entries in the dataset and performed steps to impute the missing values.
- Later into the project, we will be exploring various regression techniques to make predictions of the wine quality.

Feature Variables in the Dataset

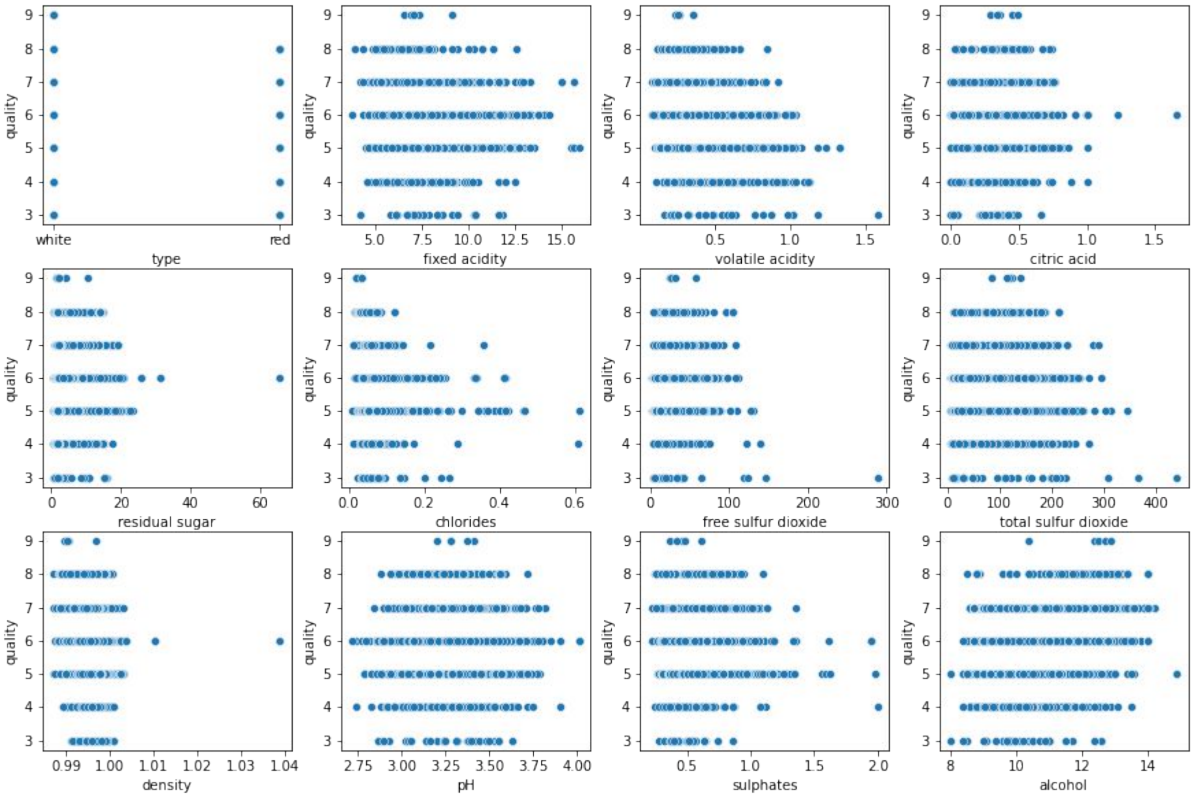
The dataset that we are exploring in this project consists of 6497 rows, inclusive of data from both Red and White wine variants. This information is sourced from “Vinho Verde”, a Portuguese wine.

The feature variables in the dataset are mentioned below:

- **Fixed Acidity** - Nonvolatile acids that do not evaporate readily
- **Volatile Acidity** - The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- **Citric Acid -** Found in small quantities, citric acid can add 'freshness' and flavor to wines.
- **Residual Sugar -** The amount of sugar remaining after fermentation stops. Generally, wines contain at least 1 gram/liter.
- **Chlorides -** Amount of salts contained in the wine.
- **Free Sulfur Dioxide -** Free-form sulphur dioxide.
- **Total Sulfur Dioxide -** Total amount of free and bound sulphur dioxide in the wine.
- **Density** - The density of water is close to that of water depending on the percent alcohol and sugar content.
- **pH** - Scales from 0 (very acidic) to 14 (very basic); most wines are between 3-4.
- **Sulphates** - A wine additive which can contribute to sulfur dioxide levels, and acts as an antimicrobial
- **Alcohol** - Concentration of alcohol present in the wine.

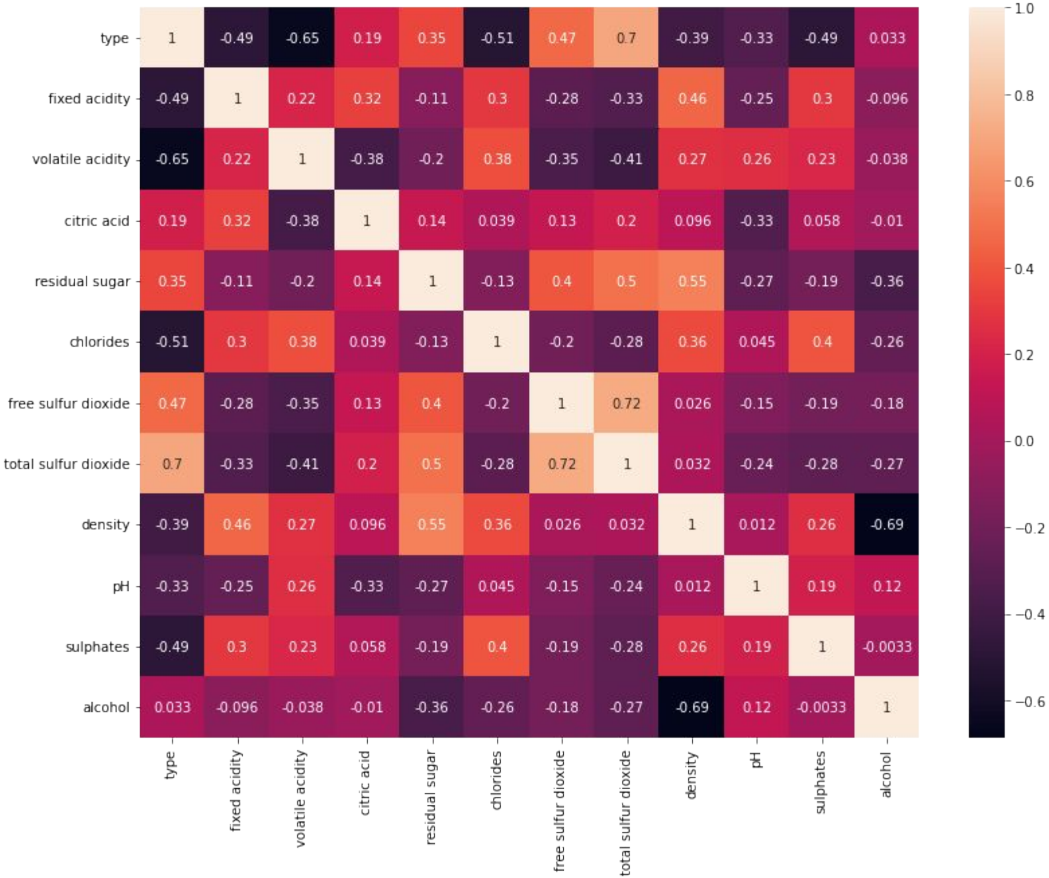
Initial Data Exploration

- An initial exploration of the data involved plotting the relationship between the quality of the wine and all other features.
- We do not see a clear linear relationship between any single feature and wine quality.
- We see that red and white wine exists of all qualities, but we do not have red wine of quality 9.



Correlation between Feature Variables

- We encode the type variable (red = 0, white = 1) and plot the correlation matrix of all features
- No high correlation between feature variables observed; all correlation coefficients are below 0.75
- Highest correlation spotted between ‘free sulfur dioxide’ and ‘total sulfur dioxide’. Considering the difference in definition, we decide not to drop any of them for now



	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
type	1	-0.49	-0.65	0.19	0.35	-0.51	0.47	0.7	-0.39	-0.33	-0.49	0.033
fixed acidity	-0.49	1	0.22	0.32	-0.11	0.3	-0.28	-0.33	0.46	-0.25	0.3	-0.096
volatile acidity	-0.65	0.22	1	-0.38	-0.2	0.38	-0.35	-0.41	0.27	0.26	0.23	-0.038
citric acid	0.19	0.32	-0.38	1	0.14	0.039	0.13	0.2	0.096	-0.33	0.058	-0.01
residual sugar	0.35	-0.11	-0.2	0.14	1	-0.13	0.4	0.5	0.55	-0.27	-0.19	-0.36
chlorides	-0.51	0.3	0.38	0.039	-0.13	1	-0.2	-0.28	0.36	0.045	0.4	-0.26
free sulfur dioxide	0.47	-0.28	-0.35	0.13	0.4	-0.2	1	0.72	0.026	-0.15	-0.19	-0.18
total sulfur dioxide	0.7	-0.33	-0.41	0.2	0.5	-0.28	0.72	1	0.032	-0.24	-0.28	-0.27
density	-0.39	0.46	0.27	0.096	0.55	0.36	0.026	0.032	1	0.012	0.26	-0.69
pH	-0.33	-0.25	0.26	-0.33	-0.27	0.045	-0.15	-0.24	0.012	1	0.19	0.12
sulphates	-0.49	0.3	0.23	0.058	-0.19	0.4	-0.19	-0.28	0.26	0.19	1	-0.0033
alcohol	0.033	-0.096	-0.038	-0.01	-0.36	-0.26	-0.18	-0.27	-0.69	0.12	-0.0033	1

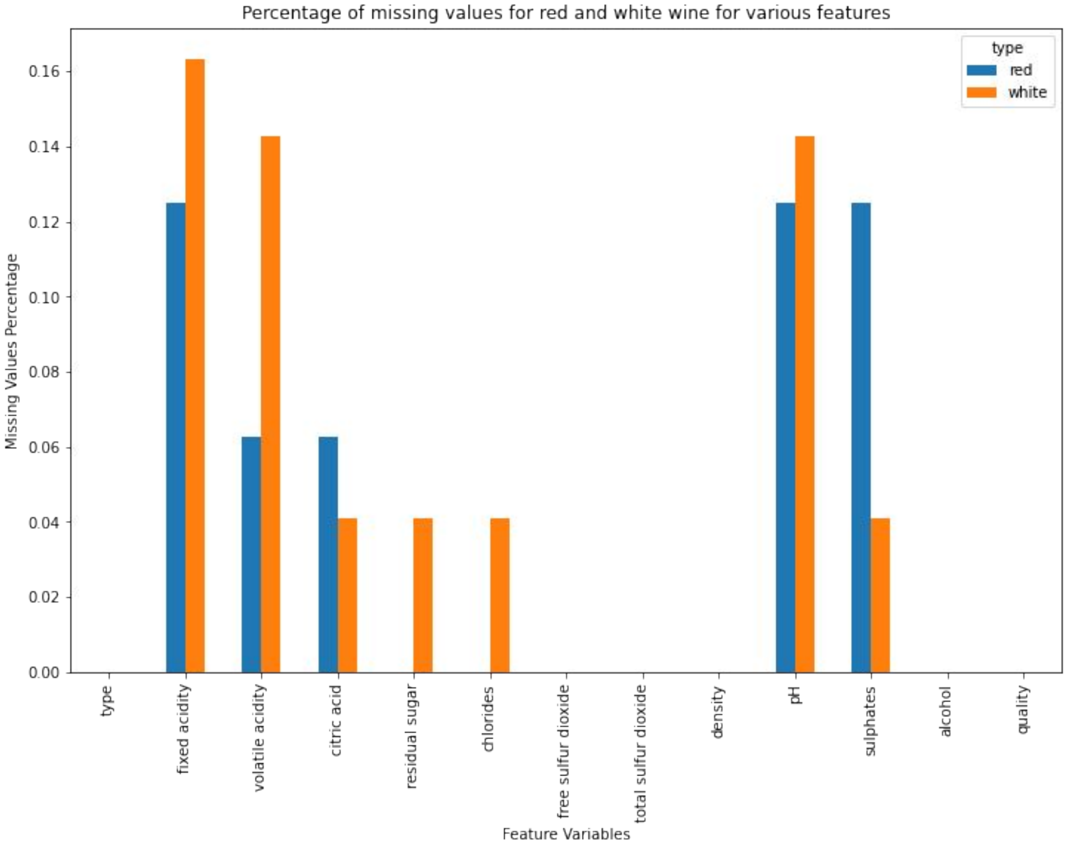
Study of the Relationship between Feature and Target Variable

From the pair plots and the calculation of correlation, we see the following relations between the feature and target variables.

1. The feature variables affecting the outcome of wine quality the most are volatile acidity, chlorides, density and alcohol.
2. Alcohol shows a strong positive correlation while the other variables represent strong negative correlations.
3. Fixed Acidity, Residual Sugar, and Total Sulfur Dioxide have weak negative correlation with wine quality.
4. Citric Acid, Free Sulfur Dioxide, pH and Sulphates have significantly weak positive correlation with the quality of wine.

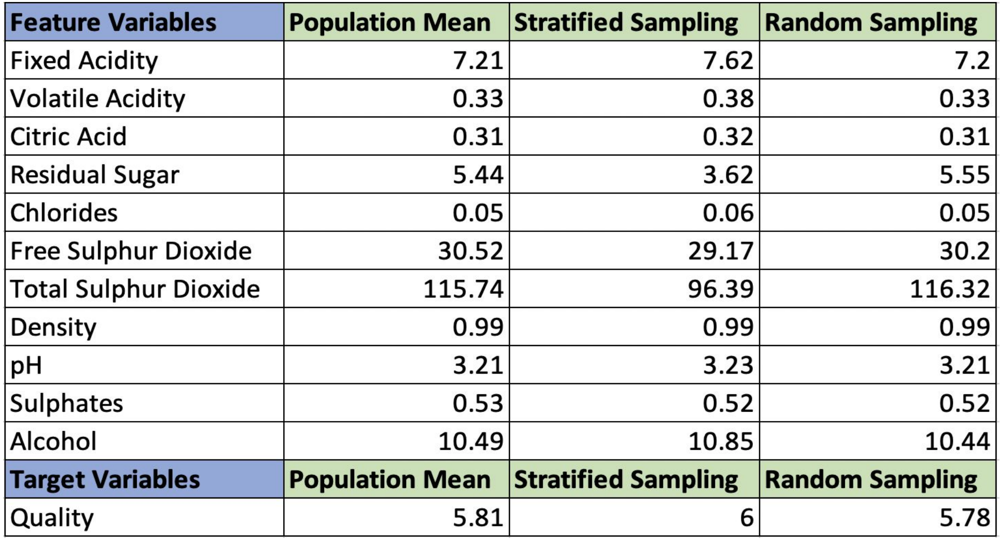
Cleaning and Missing Values Imputation

- Average of roughly 0.1 % of the values are missing for the given set of features.
- In certain cases both red and white wine have missing values such as pH and fixed acidity, but we also notice only white wine having missing values such as chlorides and residual sugar.
- We have replaced the missing values in this case with the mean from its own class i.e. either red or white wine. This makes sure more appropriate values are appended rather than the generic mean.



Sampling and Mean Calculation

To perform sampling on the dataset, we experimented with two sampling methods - Stratified Sampling and Random Sampling. We extracted 45% of the data into the new samples and calculated the mean of the new samples to compare the selected data to its population mean parameter.



Feature Variables	Population Mean	Stratified Sampling	Random Sampling
Fixed Acidity	7.21	7.62	7.2
Volatile Acidity	0.33	0.38	0.33
Citric Acid	0.31	0.32	0.31
Residual Sugar	5.44	3.62	5.55
Chlorides	0.05	0.06	0.05
Free Sulphur Dioxide	30.52	29.17	30.2
Total Sulphur Dioxide	115.74	96.39	116.32
Density	0.99	0.99	0.99
pH	3.21	3.23	3.21
Sulphates	0.53	0.52	0.52
Alcohol	10.49	10.85	10.44
Target Variables	**Population Mean**	**Stratified Sampling**	**Random Sampling**
Quality	5.81	6	5.78

We see that Random Sampling generated sampling statistics closer to the population parameters in comparison to stratified sampling.

Machine Learning Techniques Proposed for Implementation

We will be solving the wine quality problem in two distinct approaches. Predicting the quality of wine by giving it a numerical rating (from 1 to 10) is a regression problem. On the other hand, giving wine quality a standard predictor of ‘Good’ or ‘Bad’ is a classification problem.

- To solve the regression problem, we will be using the following algorithms:
    - Linear Regression
    - Ridge Regression
    - Lasso Regression
    - Elastic-Net Regression

- To solve the classification problem, we will be using the following algorithms:
    - Logistic Regression
    - Support Vector Machines
    - Random Forest

- The idea will be to compare the performance of the various algorithms listed above and create an ensemble the gives the best prediction performance.

Proposed Ensembling Techniques for Prediction

To address the issue of data imbalance in the dataset we also propose to apply techniques such as:

1. Bagging
2. Boosting
    a. Adaboost
    b. XGBoost
    c. LightGBM

Finally, we will run simulations using a combination of these methods to arrive at a specific model that performs best on the test data for prediction.