I. Introduction

As of 2022, the United States of America was named to have consumed the largest volume of wine worldwide, with 34 million hectoliters (Conway, 2023). It was also said that Merlot, which is a kind of red wine, to be the most popular variety in the US by 2018 (Conway, 2023). So what makes red wine a popular drink? There are many things to consider, it has some health benefits, it's versatile, it has aroma, and most importantly, the taste. When looking for red wine, people are sure to look for the ones with the best quality. Of course, when it comes to quality of wines, they would think about the taste, smell, color, depth, balance, and any other qualities that wine connoisseurs would think of. But what if we look at the chemical properties of wine? What properties affect the quality of red wine anyway?

II. Motivation

With the question in mind, we want to explore the relationship between the physicochemical properties of wines and their sensory quality ratings. By analyzing these properties, we aim to gain insights into the factors that influence wine quality and understand how different chemical properties contribute to the overall sensory perception of wines. This study has practical implications for winemakers, sommeliers, and wine enthusiasts in understanding the key factors that impact wine quality hence, improving wine production processes and decisions.

III. Data Source

The data for this study will be obtained from the Wine Quality Dataset, which is available from https://archive.ics.uci.edu/ml/datasets/wine+quality. The original data contains information on the physicochemical properties and sensory quality of red and white variants of Portuguese Vinho Verde wine but we chose red wine as our dataset to classify the quality of wine.

IV. The Dataset

Variables	Descriptions
Fixed acidity	the non-volatile acids in wine
Volatile acidity	the amount of acetic acid in wine
Citric acid	the amount of citric acid in wine
Residual sugar	the amount of sugar remaining after fermentation.
Chlorides	the amount of salt in wine
Free sulfur dioxide	the amount of sulfur dioxide as a preservative in wine
Total sulfur dioxide	the total amount of sulfur dioxide in wine
Density	the density of wine
pH	the acidity or basicity of wine
Sulphates	the amount of sulfur dioxide bound to potassium in wine
Alcohol	the alcohol content of wine
Quality	the quality rating of wine (scored from 0 to 10)

The dataset contains 1599 rows with 12 columns and it is a multivariate dataset. The target outcome/variable is the Quality variable which has a score rating of 0 to 10.

V. Pre-Processing The Data:

The dataset used for this project was relatively clean, as we found no instances of missing, null, or duplicate values that could potentially skew the results of our analyses. We facilitated ease of analysis by coding all predictor variables as numeric, allowing for straightforward computations and more accurate interpretations of the findings. Our examination of the dataset revealed that there were more wines classified as normal (quality in the range of 5-6) compared to those classified as excellent (quality>6) or poor (quality<5).

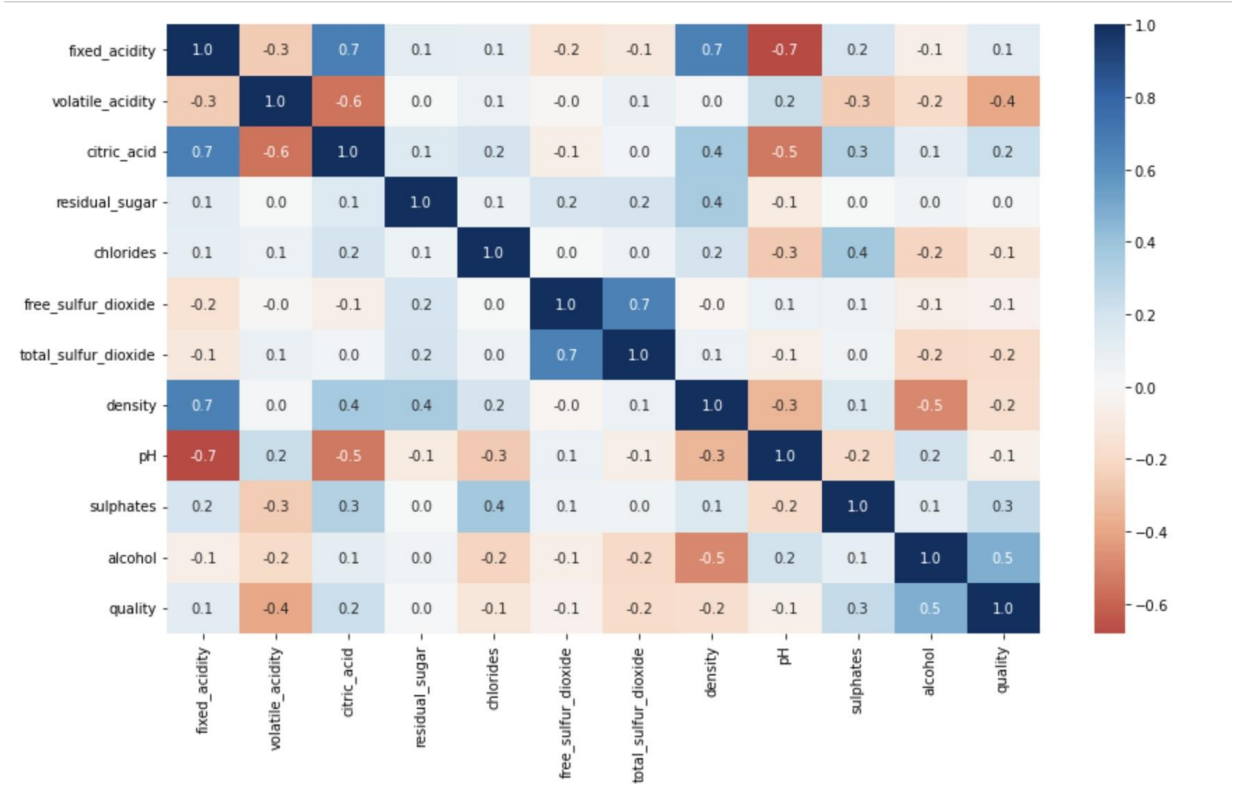
Although we did observe some outliers in the dataset, we decided to retain them in our analyses due to the imbalanced nature of the data. We believe that these outliers could still provide valuable insights into the factors that contribute to the overall quality of wines.

To standardize the column names and make them more convenient for future analysis, we used the strip() and replace() methods of the pandas dataframe to modify the column labels of our dataset. By doing so, we ensured that all column names adhere to a standardized format, making it easier to work with the data in the future. This modification will help avoid potential syntax errors or naming issues that could arise from having spaces in column names.

VI. Exploratory Data Analysis

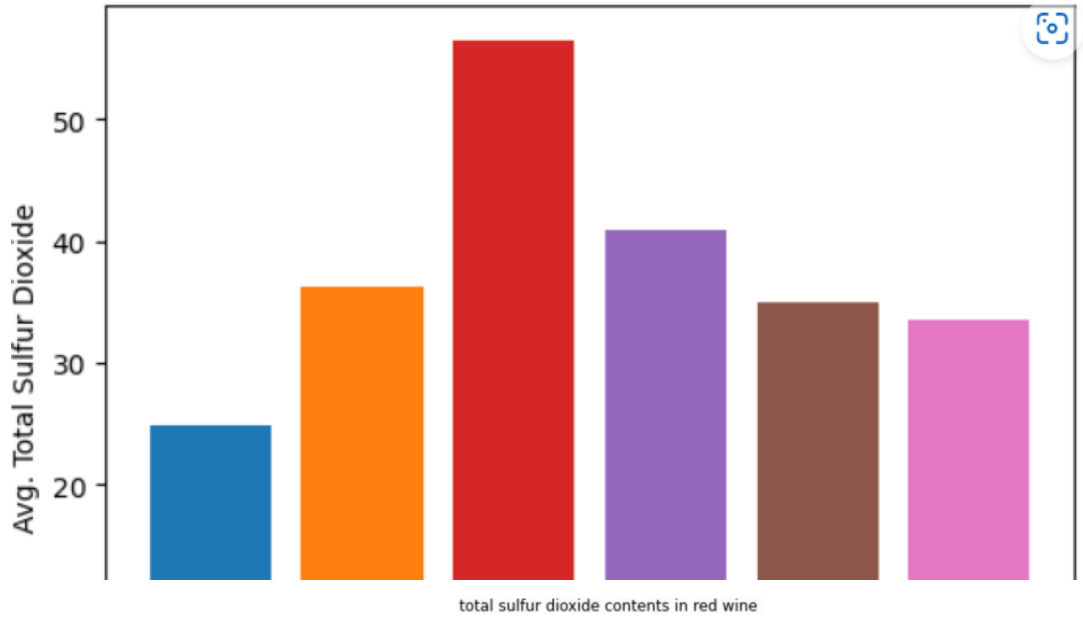
Exploratory Data Analysis (EDA) is a critical step in any data mining project as it provides valuable insights into the data, helps identify patterns and trends, and guides decision-making. In our EDA of the wine dataset, we found several key findings that can help businesses in the wine industry optimize their production and quality control processes.

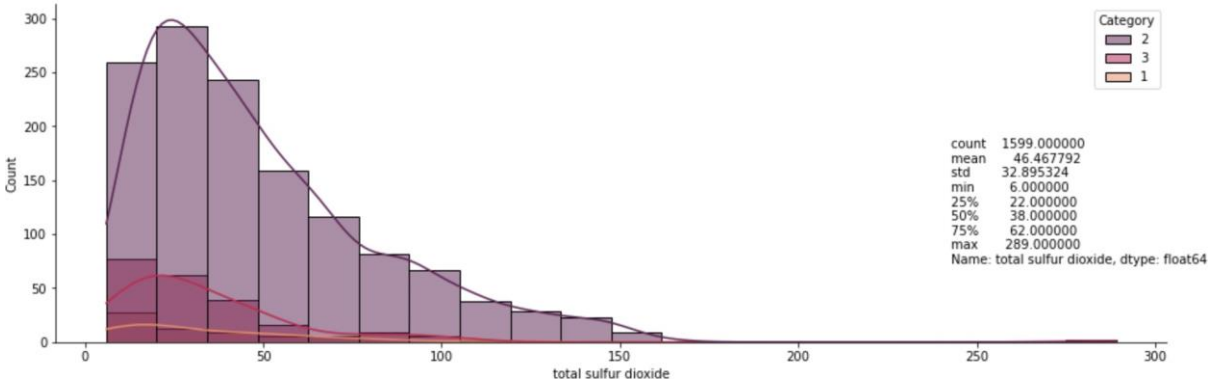
Firstly, we discovered that the quality of wine is significantly correlated with three variables: alcohol content, acetic acid quantity, and sulphates.



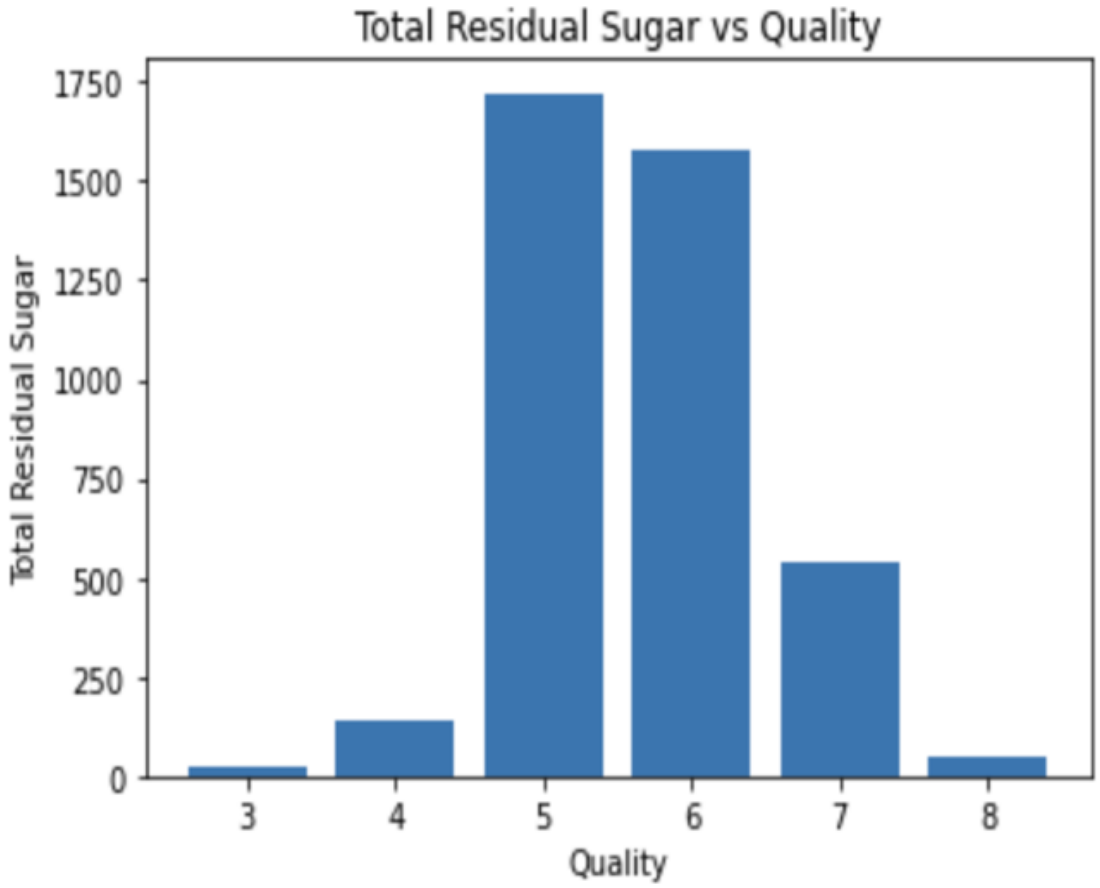
This finding is crucial for businesses as it suggests that controlling these variables could lead to better quality wines. For instance, optimizing the alcohol content to be within the range of 11-13% and minimizing acetic acid concentration to 0.3-0.5 can lead to higher quality wines.

We also found that the total sulfur dioxide content in wine is critical for maintaining its optimal taste and quality. Our analysis showed that sulfur dioxide levels above 50 ppm can spoil the taste and smell of the wine, while levels below 25 ppm can degrade the quality of wine. Therefore, it is essential to keep the sulfur dioxide content in the range of about 30-35 ppm for optimal quality.

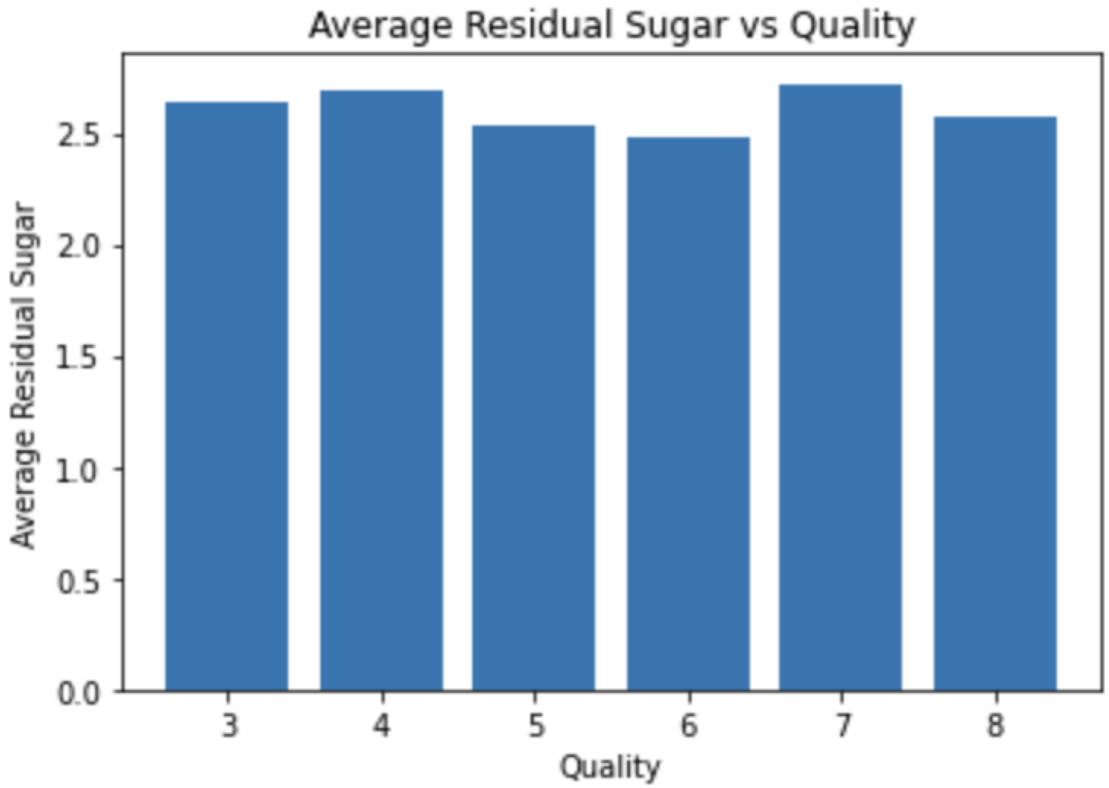




Another critical finding from our EDA is that normal quality wines have more than three times the total residual sugar than excellent or poor wines.

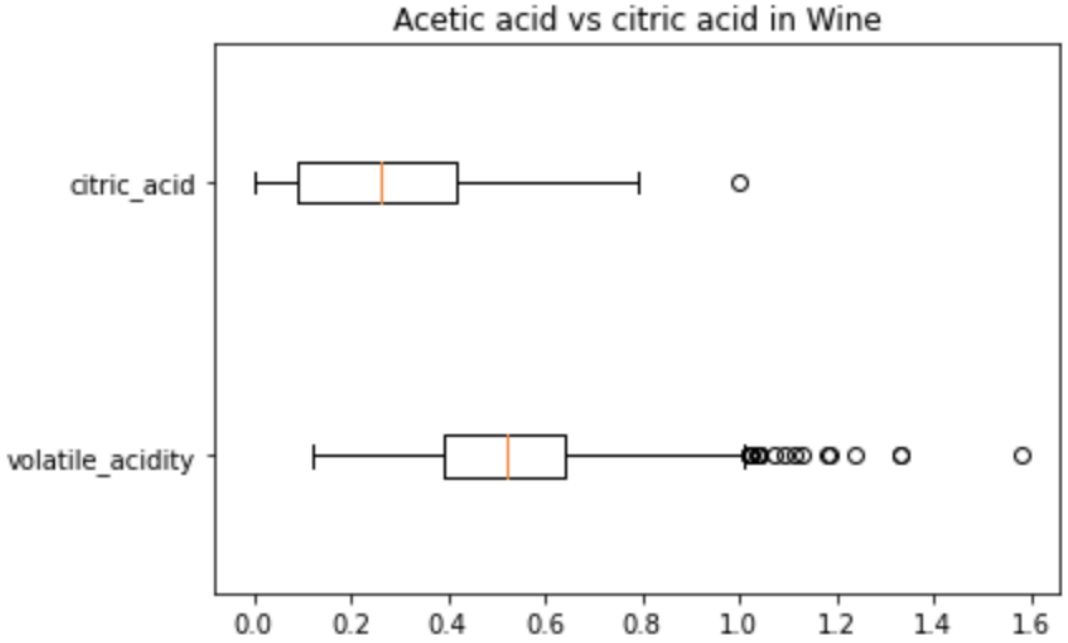


However, when we plotted the graph for average total residual sugar, we found that it is almost equal for all qualities of wine.



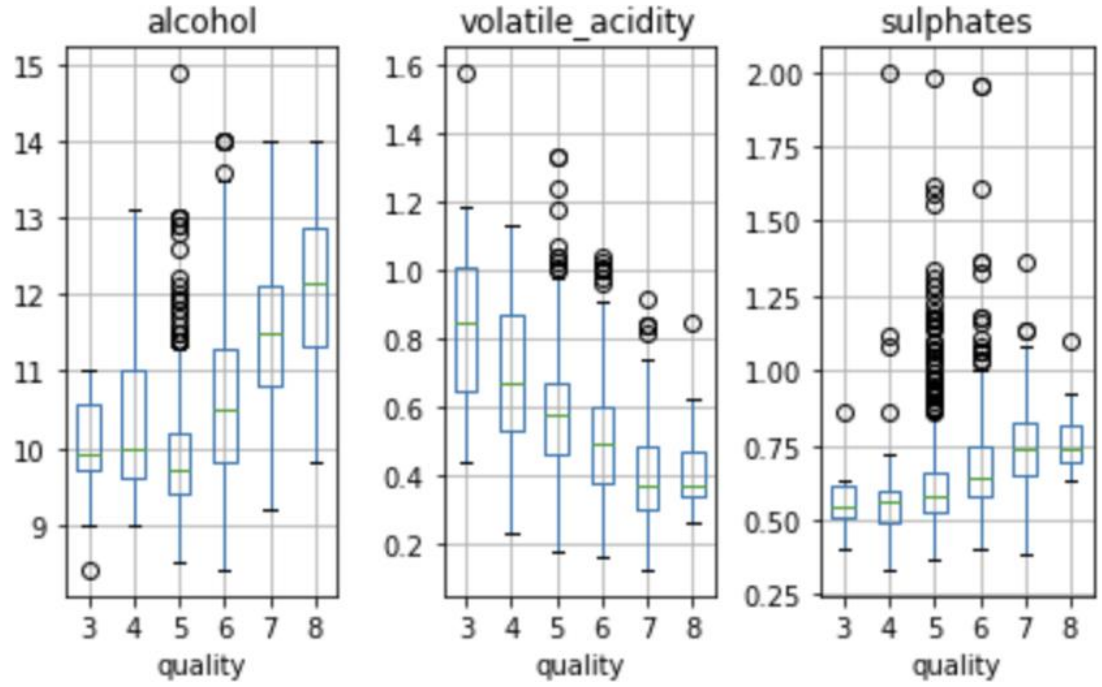
This suggests that the data is concentrated around wines with normal quality. Businesses can leverage this insight by adjusting their production processes to improve the quality of wines with normal ratings and potentially expand their customer base.

Additionally, we found that the amount of citric acid in wines, which adds freshness and flavor, is only half compared to acetic acid.

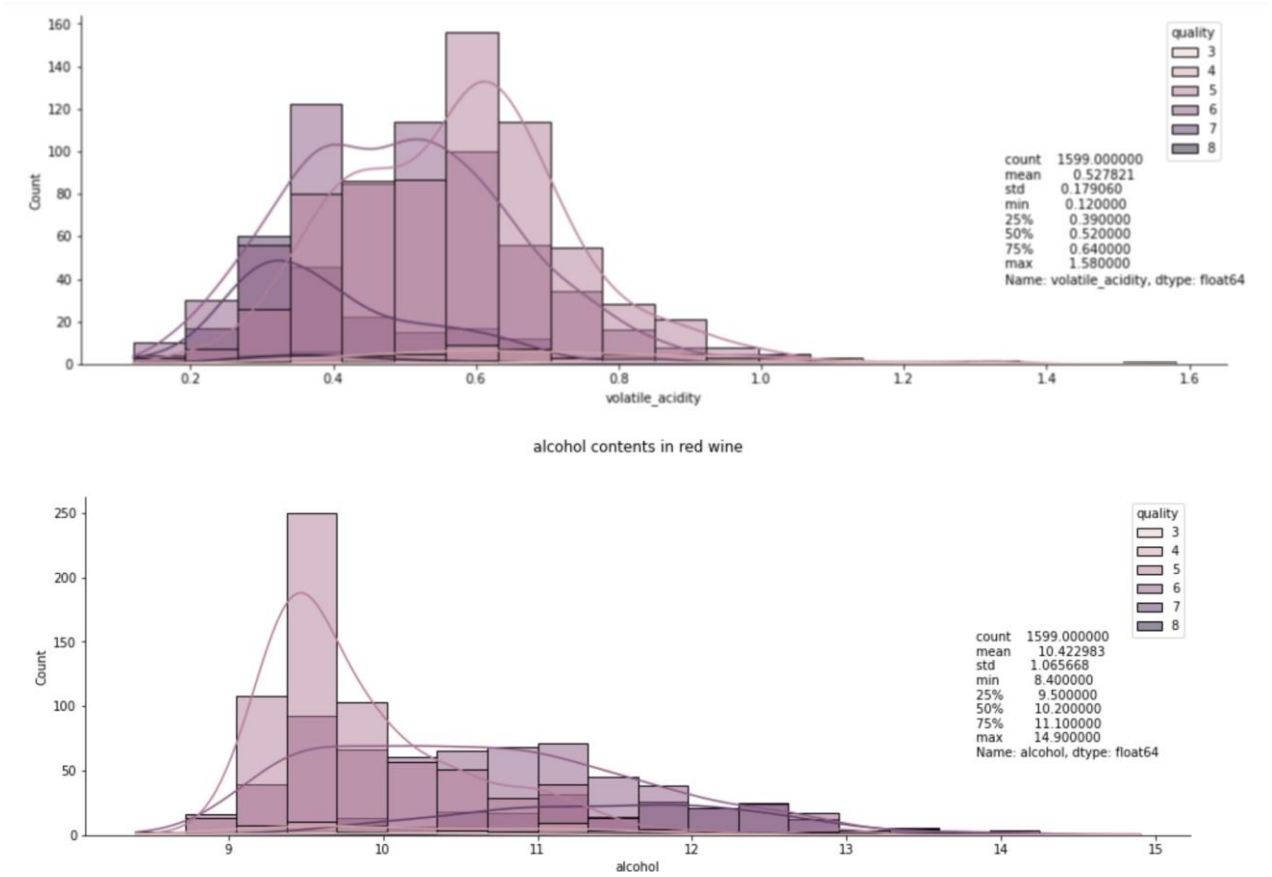


This finding can guide businesses in controlling the amount of these acids during wine production, potentially leading to more favorable flavors and customer satisfaction.

Finally, we found that better quality red wines generally have higher alcohol content and lower volatile acidity, with the best quality wines having almost 25% more alcohol content and 50% less acetic acid than the poorest quality wines.



However, it is important to note that this does not necessarily mean that higher alcohol and lower acidity will always increase the quality of wine. There is a range of values within which this is true.



As we can see, the best quality wines have a volatile acidity amount of around 0.3ppm and alcohol content in the range of 11.5% to 13%. Winemakers can use this insight to make better quality wines.

In conclusion, the insights gained from our EDA can guide businesses in the wine industry to optimize their production processes, improve the quality of their wines, and ultimately increase customer satisfaction and sales.

VII. Multiple Linear Regression

Linear regression is a popular statistical modeling technique used to establish a relationship between a dependent variable and one or more independent variables. It is widely employed in various fields to predict, analyze, and understand the impact of different factors on a target variable. In this report, we present a comparative analysis of four different linear regression models: Exhaustive Search, Backward, Forward, and Stepwise. The performance of each model is evaluated using several regression statistics, including Mean Error (ME), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE), and Mean Absolute Percentage Error (MAPE). The aim of this analysis is to identify the most effective model in terms of predictive accuracy and error metrics.

Regression Statistics

Exhaustive Search:

```
Regression statistics

                    Mean Error (ME) : -0.0073
         Root Mean Squared Error (RMSE) : 0.6488
              Mean Absolute Error (MAE) : 0.5088
            Mean Percentage Error (MPE) : -1.4889
Mean Absolute Percentage Error (MAPE) : 9.2676
```

Backward:

```
Regression statistics

                    Mean Error (ME) : -0.0073
         Root Mean Squared Error (RMSE) : 0.6488
              Mean Absolute Error (MAE) : 0.5088
            Mean Percentage Error (MPE) : -1.4889
Mean Absolute Percentage Error (MAPE) : 9.2676
```

Forward:

```
Regression statistics

                    Mean Error (ME) : -0.0112
         Root Mean Squared Error (RMSE) : 0.6474
              Mean Absolute Error (MAE) : 0.5076
            Mean Percentage Error (MPE) : -1.5552
Mean Absolute Percentage Error (MAPE) : 9.2451
```

Stepwise:

```
Regression statistics

                        Mean Error (ME) : -0.0112
     Root Mean Squared Error (RMSE) : 0.6474
          Mean Absolute Error (MAE) : 0.5076
        Mean Percentage Error (MPE) : -1.5552
Mean Absolute Percentage Error (MAPE) : 9.2451
```

Discussion and Analysis:

Exhaustive Search variables: fixed acidity, volatile acidity, chlorides, density, sulphates, alcohol

The exhaustive search method involves systematically testing all possible combinations of predictor variables. In this case, the model yielded an ME of -0.00073, indicating a slight negative bias in the predictions. The RMSE value of 0.6488 suggests a moderate level of overall prediction error. The MAE of 0.5088 indicates the average absolute difference between the predicted and actual values. The MPE of -1.4889 indicates a slight underestimation bias. The MAPE of 9.2676 represents the average percentage difference between the predicted and actual values. Overall, the exhaustive search model performs reasonably well but may benefit from further refinement.

Backward variables: fixed acidity, volatile acidity, chlorides, density, sulphates, alcohol

The backward method starts with a model that includes all predictor variables and iteratively removes variables with the least significance. The backward model exhibits an ME of -0.0073, indicating a slight negative bias. The RMSE of 0.6488 suggests a similar level of prediction error compared to the exhaustive search model. The MAE of 0.5088 which is similar to the exhaustive search model. The MPE of -1.4889 represents a moderate negative bias. The MAPE of 9.2676 suggests a similar level of percentage difference as observed in the exhaustive search model. Overall, the backward model performs similarly to the exhaustive search model.

Forward variables: alcohol, volatile acidity, sulphates, chlorides, pH

The forward method starts with an empty model and iteratively adds predictor variables based on their significance. The forward model exhibits an ME of -0.0112, suggesting a slightly larger negative bias compared to the previous two models. The RMSE of 0.6474 is comparable to the other models, indicating a moderate level of prediction error. The MAE of 0.5076 is similar to the exhaustive search and backward models, indicating consistent accuracy. The MPE of -1.5552 represents a moderate negative bias in the predictions. The MAPE of 9.2451 is consistent with the other models, suggesting similar levels of

percentage difference between predicted and actual values. Overall, the forward model performs well and provides a good balance between accuracy and complexity.

Stepwise variables: alcohol, volatile acidity, sulphates, chlorides, pH

The stepwise method combines the forward and backward approaches by iteratively adding and removing predictor variables based on their significance. The stepwise model exhibits an ME of -0.0112, indicating a similar negative bias as observed in the forward model. The RMSE of 0.6474 is identical to the forward model, suggesting comparable overall prediction error. The MAE of 0.5076 is also consistent with the forward and other models, indicating similar accuracy. The MPE of -1.5552 represents a moderate negative bias, consistent with the forward model. The MAPE of 9.2451 is also similar to the forward and other models, indicating a consistent level of percentage difference. Overall, the stepwise model performs on par with the forward model, suggesting the effectiveness of the combined feature selection strategy.

VIII. - K-Nearest Neighbors (KNN) For Prediction

The KNN algorithm is a non-parametric classification algorithm that classifies an unlabeled sample based on the classes of its nearest neighbors. In this report, we present the results of applying the KNN algorithm on a continuous variable.

Results:

The table below shows the best k (6) with their respective RMSE values.

```
RMSE value for k=  1 is: 0.7974568953868291
RMSE value for k=  2 is: 0.7120612684313057
RMSE value for k=  3 is: 0.6813487441179524
RMSE value for k=  4 is: 0.6734356873525489
RMSE value for k=  5 is: 0.6683935966180405
RMSE value for k=  6 is: 0.6584124425042744
RMSE value for k=  7 is: 0.662401321106836
RMSE value for k=  8 is: 0.6642554866634524
RMSE value for k=  9 is: 0.6686602137718989
```

Observations:

The results suggest that increasing the value of k beyond a certain threshold may not necessarily improve the model's RMSE. Therefore, it is essential to select the optimal value of k based on the specific dataset and problem at hand. In this case, k = 6.

Data Information:

The dataset used for training and testing the KNN model consists of several features and a target variable (quality). The features are preprocessed and normalized using z-scores. Here are the details of the features:

- Fixed Acidity (zfixed acidity)
- Volatile Acidity (zvolatile acidity)
- Residual Sugar (zresidual sugar)
- Chlorides (zchlorides)
- Free Sulfur Dioxide (zfree sulfur dioxide)
- Density (zdensity)
- pH (zpH)
- Sulphates (zsulphates)
- Alcohol (zalcohol)

Additionally, the dataset contains the target variable "quality," which represents the quality rating of the wine samples.

Nearest Neighbors:

For a specific scenario, where k=6, the distances and indices of the nearest neighbors for a test sample are provided:

```
[5]
Distances [[0.         0.62350434 0.67014874 0.72576342 0.75080402 0.76951908]]
Indices [[ 37 420  22 344 486  11]]
```

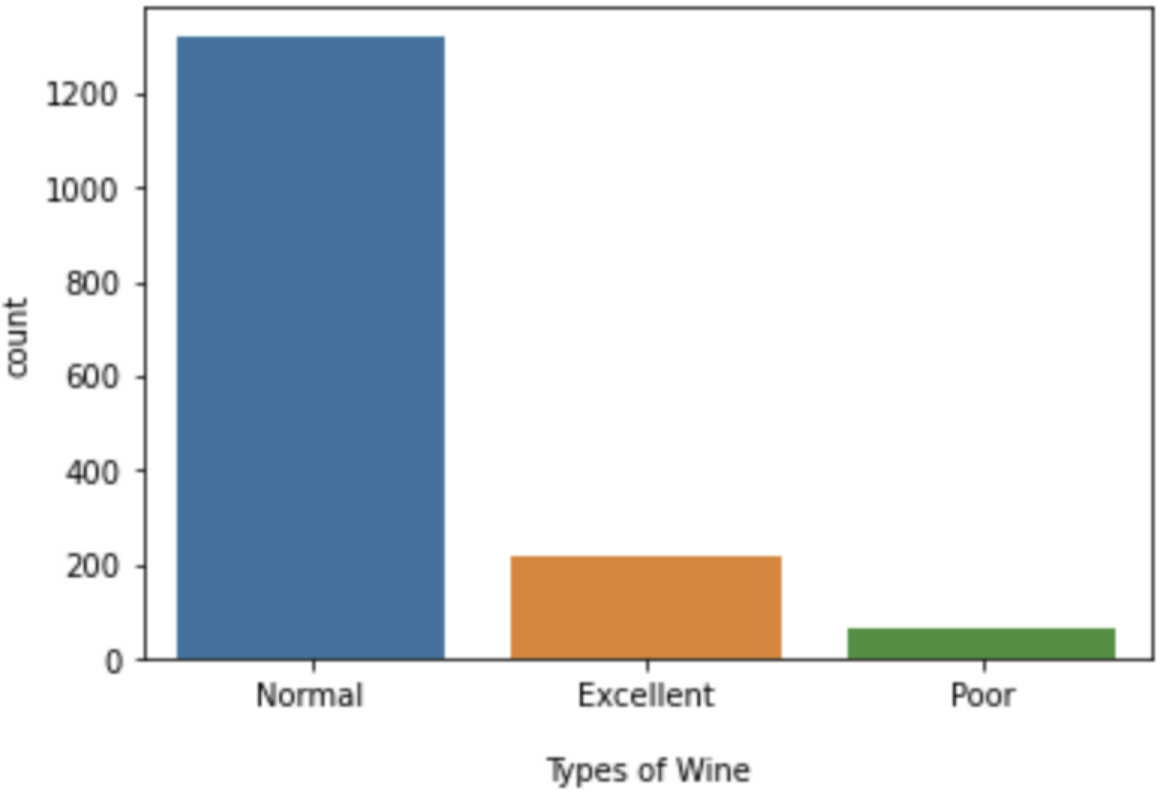
It is worth noting that the presented report is based on a specific dataset and specific values of k. For further analysis or generalization of these findings, additional experiments and evaluations may be required.

IX. Classification Models

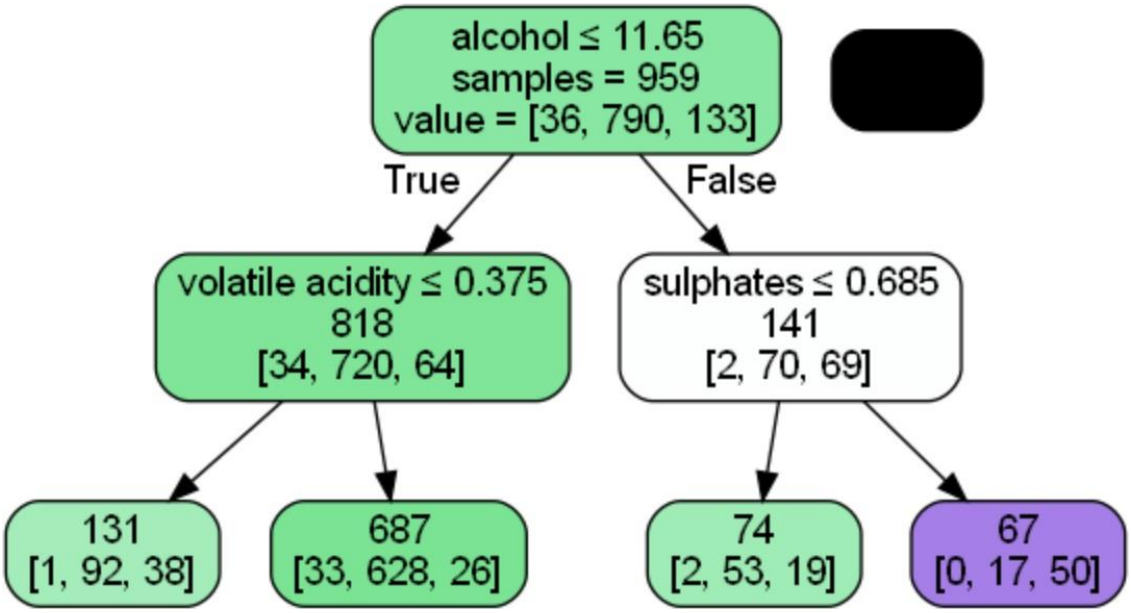
Multi Class Classification

Grouping the wine dataset into three categories

- Poor ranges from 0 to 4 and consists of 63 records
- Normal ranges from 5 to 6 and consists of 1319 records
- Excellent ranges from 7 to 10 and consists of 217 records



Decision Tree Classifier



The inference from the decision tree says alcohol and volatile acidity plays a predominant role in classifying wine quality.

Classification Report

	precision	recall	f1-score	support
1	0.00	0.00	0.00	27
2	0.85	0.97	0.91	529
3	0.55	0.26	0.35	84
accuracy			0.83	640
macro avg	0.47	0.41	0.42	640
weighted avg	0.78	0.83	0.79	640

Confusion matrix

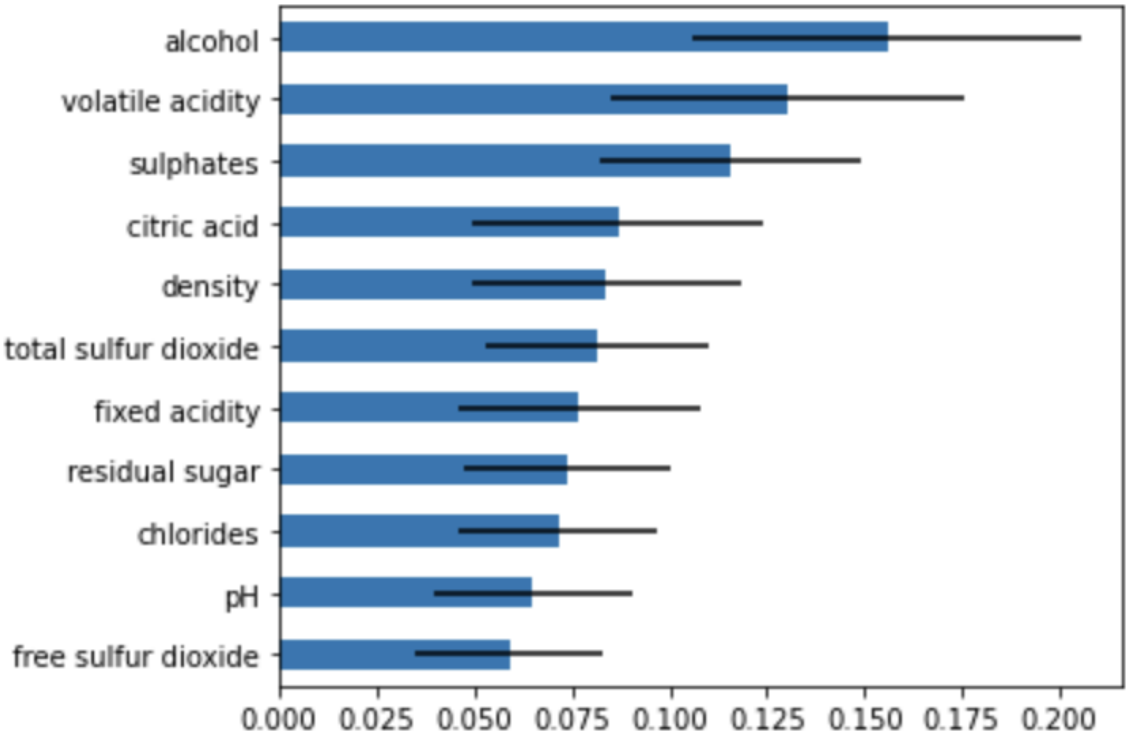
Confusion Matrix (Accuracy 0.8328)

	Prediction		
Actual	0	1	2
0	0	27	0
1	0	511	18
2	0	62	22

The above results are obtained after grid search and the decision tree model gave us an accuracy of 83% with max depth of 2. All values in the poor category has been predicted as normal wine and 85% precision and 97% recall score for normal wine and 55% precision and 26% recall score for excellent wine. Overall our model predicts well for the normal wines.

Random Forest:

Important Features:



From the horizontal bar plot above, the top 5 important features that help us in predicting quality are,

- Alcohol
- Volatile Acidity
- Sulphates
- Citric Acid
- Density

Random forest model results in the validation data

Classification Report:

	precision	recall	f1-score	support
1	0.50	0.04	0.07	27
2	0.88	0.96	0.92	529
3	0.67	0.50	0.57	84
accuracy			0.86	640
macro avg	0.68	0.50	0.52	640
weighted avg	0.84	0.86	0.84	640

Confusion matrix:

Confusion Matrix (Accuracy 0.8594)

Actual \ Prediction	0	1	2
0	1	26	0
1	1	507	21
2	0	42	42

The random model gives a 50% precision and 4% recall score for the poor wines, 88% precision and 96% recall for the normal wines and 67% precision and 50% recall for the excellent wines. Overall our model does a good job of classifying the normal wines better than the other wines with a very good recall score of 96%.

Boosted Trees:

Classification Report:

	precision	recall	f1-score	support
1	0.40	0.15	0.22	27
2	0.88	0.95	0.91	529
3	0.62	0.45	0.52	84
accuracy			0.85	640
macro avg	0.63	0.52	0.55	640
weighted avg	0.83	0.85	0.83	640

Confusion Matrix:

Confusion Matrix (Accuracy 0.8484)

Actual \ Prediction	0	1	2
0	4	22	1
1	6	501	22
2	0	46	38

Overall Boosted trees does a good job in classifying poor and excellent wines and classified normal wines with 88% precision and 95% recall scores.

KNN Classification:

We applied KNN to various accuracy as below,

	k	accuracy
0	1	0.821875
1	2	0.812500
2	3	0.826562
3	4	0.832812
4	5	0.828125
5	6	0.817187
6	7	0.820312
7	8	0.826562
8	9	0.826562
9	10	0.826562
10	11	0.831250
11	12	0.832812
12	13	0.831250
13	14	0.823438
14	15	0.825000
15	16	0.828125
16	17	0.828125
17	18	0.829688
18	19	0.831250

and found k=11 gives us the highest accuracy of 83.28% and below are the results of the KNN model,

Confusion Matrix:

Confusion Matrix (Accuracy 0.8313)

Actual \ Prediction	0	1	2
0	0	26	1
1	0	501	28
2	0	53	31

Classification Report:

	precision	recall	f1-score	support
1	0.00	0.00	0.00	27
2	0.86	0.95	0.90	529
3	0.52	0.37	0.43	84
accuracy			0.83	640
macro avg	0.46	0.44	0.44	640
weighted avg	0.78	0.83	0.80	640

KNN model does a good job in classifying normal wines and for excellent wine classifies with 52 % precision and 37% recall scores but performs badly on poor wines.

Logistic Regression:

Confusion Matrix:

Confusion Matrix (Accuracy 0.8313)

Actual \ Prediction	0	1	2
0	0	26	1
1	0	509	20
2	0	61	23

Classification Report:

	precision	recall	f1-score	support
1	0.00	0.00	0.00	27
2	0.85	0.96	0.90	529
3	0.52	0.27	0.36	84
accuracy			0.83	640
macro avg	0.46	0.41	0.42	640
weighted avg	0.77	0.83	0.80	640

The Logistic Regression model does a very good job in classifying normal wines and for excellent wine classifies with 52 % precision and 27% recall scores but performs badly on poor wines.

Evaluating models with 5 fold cross validation

	Model	CV Score
1	Logistic Regression	84.98 %
2	Decision Tree Classifier	78.73 %
3	Random Forest Classifier	85.82 %
4	Boosting Classifier	84.57 %
5	KNN Classifier	83.11 %

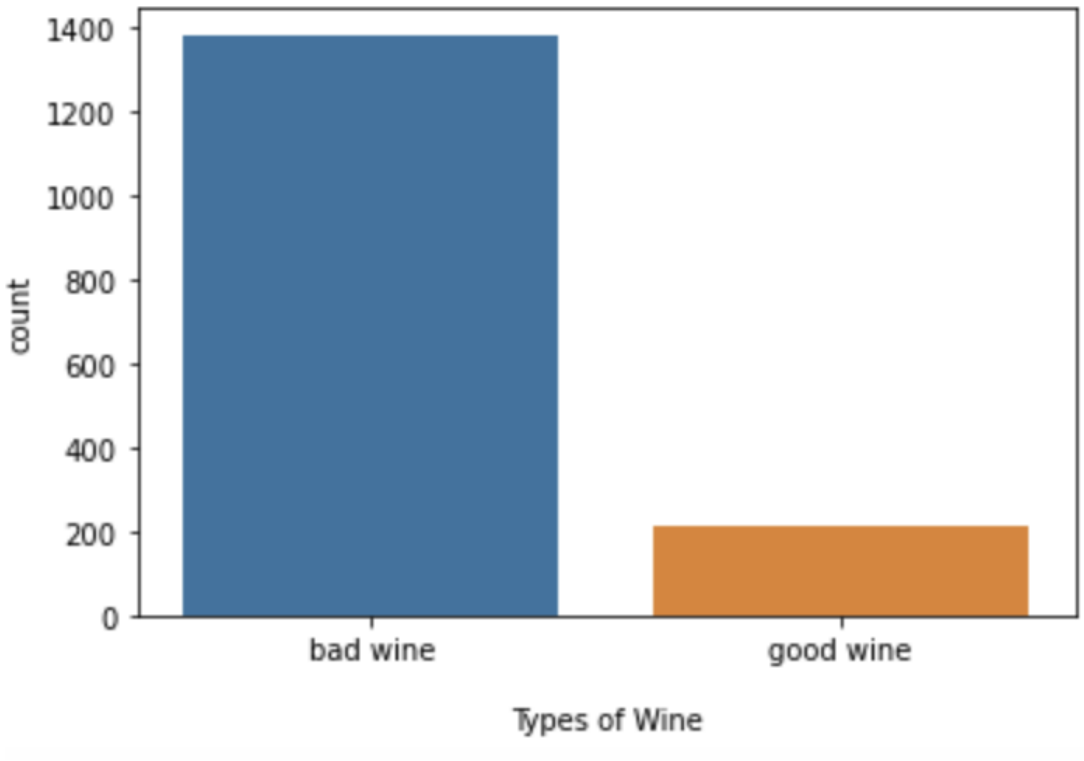
Of all the 5 models Random Forest Classifier gives the highest accuracy of 85.82%

2 class classification models

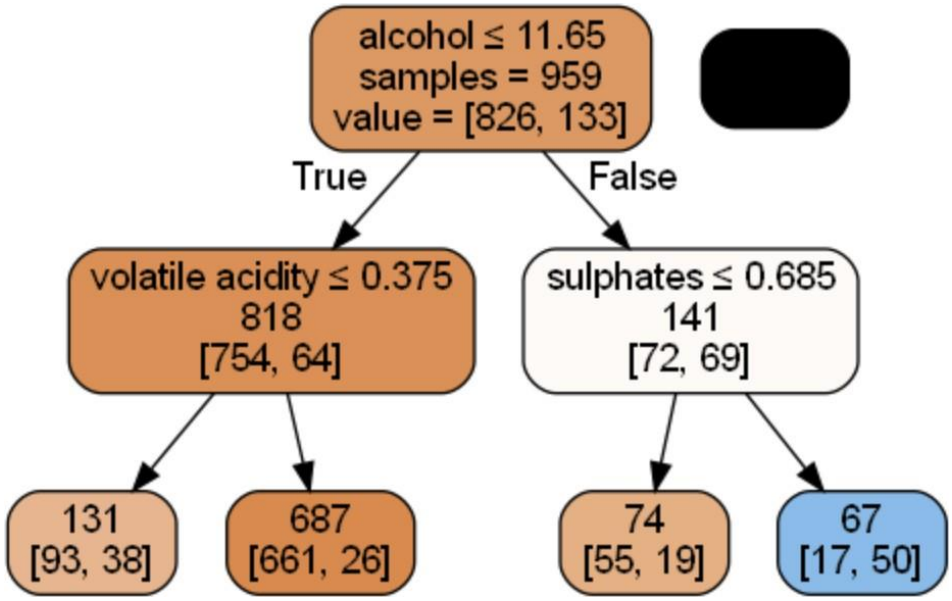
Also tried two class classification models.

Grouping the wine dataset into two categories

- Bad wine ranges from 0 to 6 and consists of 1382 records (86%)
- Excellent ranges from 7 to 10 and consists of 217 records (14%)



Decision Tree



The inference from the decision tree says alcohol and volatile acidity plays a predominant role in classifying wine quality.

Confusion matrix:

```
Confusion Matrix (Accuracy 0.8750)

       Prediction
Actual   0   1
     0 538  18
     1  62  22
```

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.97	0.93	556
1	0.55	0.26	0.35	84
accuracy			0.88	640
macro avg	0.72	0.61	0.64	640
weighted avg	0.85	0.88	0.86	640

The above results are obtained after grid search and the decision tree model gave us an accuracy of 88% with max depth of 10. The decision tree model gave us good results for the **bad wine** with 90% precision and 97% recall scores and the **good wine** did a moderate job in classification with 55% precision and 26% recall scores.

Random Forest:

Confusion Matrix:

```
Confusion Matrix (Accuracy 0.9062)

       Prediction
Actual   0   1
     0 539  17
     1  43  41
```

Classification Report

	precision	recall	f1-score	support
0	0.93	0.97	0.95	556
1	0.71	0.49	0.58	84
accuracy			0.91	640
macro avg	0.82	0.73	0.76	640
weighted avg	0.90	0.91	0.90	640

The above results are obtained after grid search and the decision tree model gave us an accuracy of 91%. The decision tree model gave us good results for the **bad wine** with 93% precision and 97% recall scores and **good wine** did a reasonably good job in classification with 71% precision and 49% recall scores.

Boosted Tree

Confusion Matrix

Confusion Matrix (Accuracy 0.8844)

Actual \ Prediction	0	1
0	530	26
1	48	36

Classification Report

	precision	recall	f1-score	support
0	0.92	0.95	0.93	556
1	0.58	0.43	0.49	84
accuracy			0.88	640
macro avg	0.75	0.69	0.71	640
weighted avg	0.87	0.88	0.88	640

The above results are obtained after grid search and the decision tree model gave us an accuracy of 88%. The decision tree model gave us good results for the **bad wine** with 92% precision and 95% recall scores and **good wine** did a reasonably good job in classification with 58% precision and 43% recall scores.

Logistic Regression

Confusion Matrix:

Confusion Matrix (Accuracy 0.8719)

Actual	Prediction 0	1
0	535	21
1	61	23

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.96	0.93	556
1	0.52	0.27	0.36	84
accuracy			0.87	640
macro avg	0.71	0.62	0.64	640
weighted avg	0.85	0.87	0.85	640

The above results are obtained after grid search and the decision tree model gave us an accuracy of 87%. The decision tree model gave us good results for the **bad wine** with 90% precision and 96% recall scores and **good wine** did a reasonably good job in classification with 52% precision and 27% recall scores.

KNN Classification:

Confusion Matrix:

Confusion Matrix (Accuracy 0.8734)

Actual	Prediction 0	1
0	532	24
1	57	27

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.96	0.93	556
1	0.53	0.32	0.40	84
accuracy			0.87	640
macro avg	0.72	0.64	0.66	640
weighted avg	0.85	0.87	0.86	640

Evaluated 2 class models with 5 fold cross validation

	Model	CV Score
1	Logistic Regression	88.43 %
2	Decision Tree Classifier	86.23 %
3	Random Forest Classifier	90.09 %
4	Boosting Classifier	90.2 %
5	KNN Classifier	87.17 %

Of all the 5 models Boosting Classifier gives the highest accuracy of 90.2%

Comparing 2 class classification models with the kaggle dataset scores below,

	Model	CV Score
1	Logistic Regression	81.72 %
2	Decision Tree Classifier	87.12 %
3	Random Forest Classifier	92.47 %
4	Xgboost Classifier	93.29 %

From the kaggle dataset results we observe that the Logistic Regression and Decision Tree Classifier performed better than the kaggle results. However the random forest model's accuracy is 1% higher in kaggle, but we believe by performing class imbalance techniques like oversampling the accuracy can be improved.

X. Recommendations

- At the conclusion of our study, we have identified key recommendations based on our findings. In the context of multiple linear regression, we suggest utilizing the following variables to effectively predict wine quality: alcohol, volatile acidity, sulfates, and pH. These factors have demonstrated significant predictive power in determining the quality of wine.

- For classification purposes, we have identified several important features that play a crucial role in classifying wines as either good or bad. These features include alcohol, volatile acidity, sulfates, citric acid, and density. By considering these variables, winemakers can accurately categorize their wines based on quality.

- To ensure the maintenance of high-quality wines, we advise winemakers to pay close attention to the levels of total sulfur dioxide. Keeping the sulfur dioxide concentration within the recommended range of 30 to 35 ppm is crucial for preserving the desired quality attributes of the wines.

- With a comprehensive understanding of which physicochemical properties are indicative of wine quality, winemakers can now shift their focus to pricing strategies. Armed with knowledge about the factors that contribute to wine quality, winemakers can confidently price their products based on their inherent value.

- It is worth noting that the dataset used in our study exhibits an imbalance, with more data available for normal quality wines compared to poor and excellent quality wines. To address this class imbalance, we recommend employing oversampling techniques. By oversampling the minority classes, a more representative and balanced dataset can be obtained, leading to more accurate modeling and analysis results.

Sources:

Conway, J. (2023, May 2). *Global wine consumption by country 2022*. Statista. https://www.statista.com/statistics/858743/global-wine-consumption-by-country/#:~:text=The%20United%20States%20consumes%20the,leading%20consumer%20of%20wine%20worldwide.

Wine Quality Dataset. UCI Machine Learning Repository: Wine quality data set. (n.d.). https://archive.ics.uci.edu/ml/datasets/wine+quality