PREDICT THE QUALITY OF WINE

Report

APRIL 26, 2024

VIJAY KUMAR TAMADA
ARAVIND REDDY KASIREDDY

Introduction, Motivation

a. General background information about the topic: Wine, particularly red wine, has been cherished for centuries, not only for its taste but also for its cultural significance. It's a complex beverage influenced by numerous factors including grape variety, terroir, and production methods. Understanding what contributes to the quality of red wine is not only of interest to connoisseurs and sommeliers but also to researchers and enthusiasts delving into the realms of data science and machine learning.

b. What motivated you to work on this topic? The allure of unraveling the mysteries behind red wine quality through data analysis and machine learning techniques is what motivated our exploration. We were captivated by the idea of leveraging mathematical models to decode the intricate relationship between the chemical composition of wine and its perceived quality. Furthermore, the challenge of applying regression and classification algorithms to predict wine quality intrigued us, prompting us to delve into this project with enthusiasm. After seeing a dataset trying to guess how good red wine is using numbers, I got curious. Even though I don't drink much alcohol, I wanted to know what I could do with this data. I had questions like: Can I use the math tricks I learned on this wine data? Will the models I learned work well even though this dataset only has numbers, not categories? I heard Random Forest is great, but will it be great for this wine dataset I found? Also, I wanted to know which numbers are most important for predicting wine quality in different models.

c. Who cares about this project? This project holds significance for various stakeholders: - Wine Enthusiasts: Individuals passionate about wine, whether as consumers or producers, seek insights into what makes a wine exceptional. Understanding the factors influencing wine quality can enhance their appreciation and production processes. - Researchers: Scholars in enology, viticulture, and data science alike are interested in uncovering patterns within wine datasets. Their findings contribute to both scientific understanding and practical applications within the wine industry. - Industry Professionals: Winemakers, marketers, and distributors rely on insights into consumer preferences and quality determinants to optimize their products and strategies.

By addressing these aspects, we aim to shed light on the significance of our project and its relevance to diverse audiences.

Problem Description

a. What are you trying to do? Business Questions: Our primary goal is to leverage data analysis and machine learning techniques to predict the quality of red wine based on its chemical attributes. This endeavor stems from a curiosity to understand how mathematical methods can be applied to decipher patterns within the complex world of wine. Specifically, we aim to address questions such as: - Can mathematical techniques be used to assess the quality of red wine effectively? - Do the models we've learned in data science courses translate well to this real-world dataset composed solely of numerical features? - Is Random Forest, a widely acclaimed algorithm, suitable for predicting wine quality in this

context? - Which chemical attributes play the most significant role in determining the quality of red wine across different predictive models?

b. Explain your objectives using absolutely no jargon: We want to use data analysis to figure out how to predict whether a red wine is of good quality or not. We're curious if we can apply the math skills we've learned to analyze this dataset. We also want to know if the models we've learned about in our data science studies will work well with this dataset, which only has numbers. We've heard about Random Forest being a powerful tool, and we want to see if it's useful for predicting wine quality in this dataset. Lastly, we aim to identify which chemical factors are most important for determining wine quality using different models.

c. Type of problem. Why? This problem falls under the domain of predictive modeling, specifically regression and classification tasks. Regression is appropriate because we're trying to predict a continuous variable (wine quality score), while classification comes into play when we categorize wines into qualitative labels (e.g., good quality vs. poor quality). This problem type is suitable because we have labeled data (quality scores) and want to train models to predict these labels based on input features (chemical attributes).

Variables

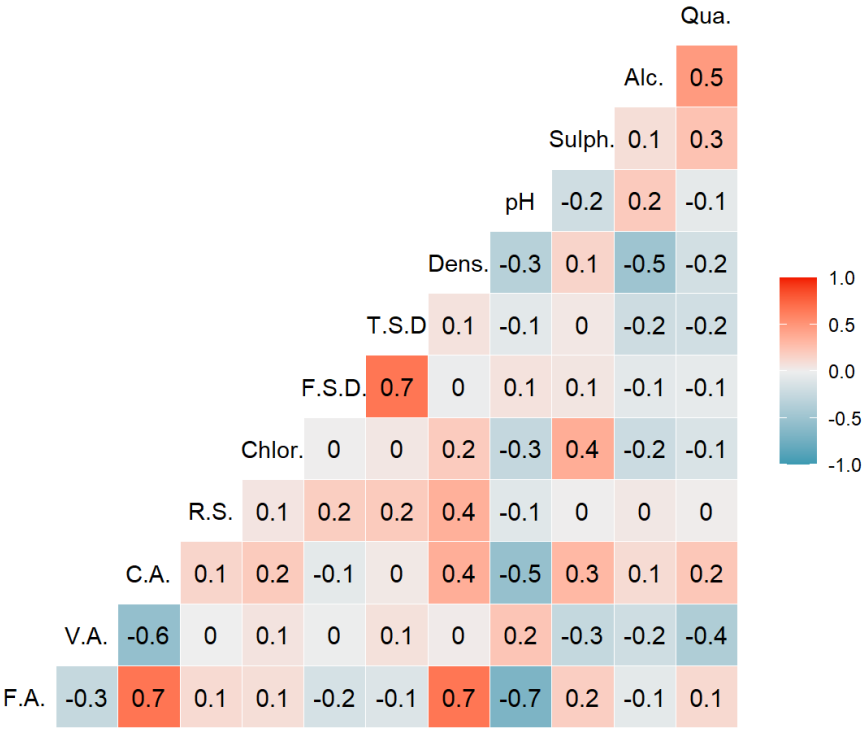
a. What is your response variable? Is it quantitative or qualitative? - The response variable in our analysis is the quality of red wine. It is a quantitative variable represented by a numerical score ranging from 0 to 10, where higher values indicate better quality.

b. What are your predictors? Which of them are quantitative? - The predictors, also known as independent variables, are the various chemical attributes of red wine. These include: 1. Fixed Acidity 2. Volatile Acidity 3. Citric Acid 4. Residual Sugar 5. Chlorides 6. Free Sulfur Dioxide 7. Total Sulfur Dioxide 8. Density 9. pH 10. Sulphates 11. Alcohol Among these, all variables except for the last one, "Alcohol," are quantitative. "Alcohol" is also quantitative but expressed as a percentage by volume.

c. What predictors do you expect to be key predictors? Can you prove that they are key predictors, e.g., using significance? - We expect certain predictors to be key determinants of wine quality based on domain knowledge and prior research. For instance, alcohol content, acidity levels, and sulfur dioxide concentrations are commonly cited factors influencing wine quality. To validate their significance, statistical tests such as linear regression coefficients or feature importance scores from machine learning models can be utilized. These tests help ascertain the relative importance of predictors in explaining variations in wine quality.

d. Is there any collinearity (or multicollinearity) between predictors? - Collinearity refers to the correlation between predictor variables in a regression analysis. Multicollinearity occurs when two or more predictors are highly correlated with each other. In our dataset, it's essential to examine collinearity to ensure the reliability of regression results. Techniques such as calculating correlation coefficients or variance

inflation factors (VIFs) can identify collinearity among predictors.



Factors related to Quality:

- Three key positive relationships between Quality and Citric.Acid, Alcohol and Sulphates.
- Three key negative relationships between Quality and pH, Density and Volatile.Acidity.
- Other variables don't have any significant relationships with Quality.

Other interesting factors:

- Alcohol has a weak positive correlation with pH value.
- Alcohol has a strong negative correlation with Density.
- Density and Citric.Acid have a strong positive correlation with Fixed.Acidity.
- pH value has a negative correlation with Sulphates, Citric.Acid, Fixed.Acidity and Density

Overview of Data

a. Data Source: The dataset used in this analysis was obtained from Kaggle and is titled "Red Wine Quality" by Cortez et al., 2009. The dataset comprises various physicochemical attributes of red wine samples along with their quality ratings.

***Meaning of the variables (based on physicochemical tests)*:**

- **1. Fixed.Acidity:** tartaric acid, measured in g/dm^3 – most of the acids involved with wine are fixed acids,[1] non-volatile.
- **2. Volatile.Acidity:** acetic acid, measured in $g/dm^3//$ – high levels can lead to unpleasant vinegar taste called *vinegar taint*, contributes to many wine spoilage yeasts and bacteria.
- **3. Citric.Acid:** measured in g/dm^3 – is usually found in small quantities, one of the *three primary* acids, adds 'freshness' and flavor to wines.[2]
- **4. Residual.Sugar:** measured in g/dm^3 – the amount of sugar remaining after the fermantation process, it is rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet.[3]
- **5. Chlorides:** sodium chloride, measured in g/dm^3 – the amount of salt in wine.[4]
- **6. Free.Sulfur.Dioxide:** measured in mg/dm^3 – free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.[@Jenny:2019]
- **7. Total.Sulfur.Dioxide:** measured in mg/dm^3 – amount of free and bound forms of SO_2. In low concentrations, SO_2 is mostly undetectable in wine, but at free SO_2 concentrations over 50 ppm, SO_2 becomes evident in the nose and taste of wine.[@Maureen:2018]
- **8. Density:** measured in g/cm^3 – depends on percentage of the alcohol and sugar content.[5]
- **9. pH** the level of acidity – range between 0 (very acidic) and 14 (very basic), most wines are in a range 3-4 on the pH scale.[6]
- **10. Sulphates:** potassium sulphate, measured in g/dm^3 – wine additive/food preservative, contributes to Sulfur Dioxide Gas (SO_2).[@Rachael:2019]
- **11. Alcohol:** % by volume

---

[1] Acids in wine

[2] Citric acid

[3] What is Residual Sugar in Wine

[4] Chloride concentration in red wines: influence of *terroir* and grape type

[5] Measurement of Density of wine

[6] Acidity and pH

- **12. Quality:** range between 0 and 10

b. Data Cleaning: 1. **Verifying No Near-Zero Variance Predictors:** We checked if any of the predictors exhibited near-zero variance, which could affect certain classification models. However, no near-zero variance predictors were found, indicating that the dataset has sufficient variability for analysis.

2. **Checking Missing Values:** No missing values were detected in any of the predictor variables.

3. **Adjusting Target Variable for Classification Models:** The quality ratings in the dataset range from 3 to 8. To facilitate classification modeling, we categorized the quality ratings into two classes: "high" (ratings 6 and above) and "low" (ratings below 6). This categorization aimed to balance the classes and avoid class imbalance issues.

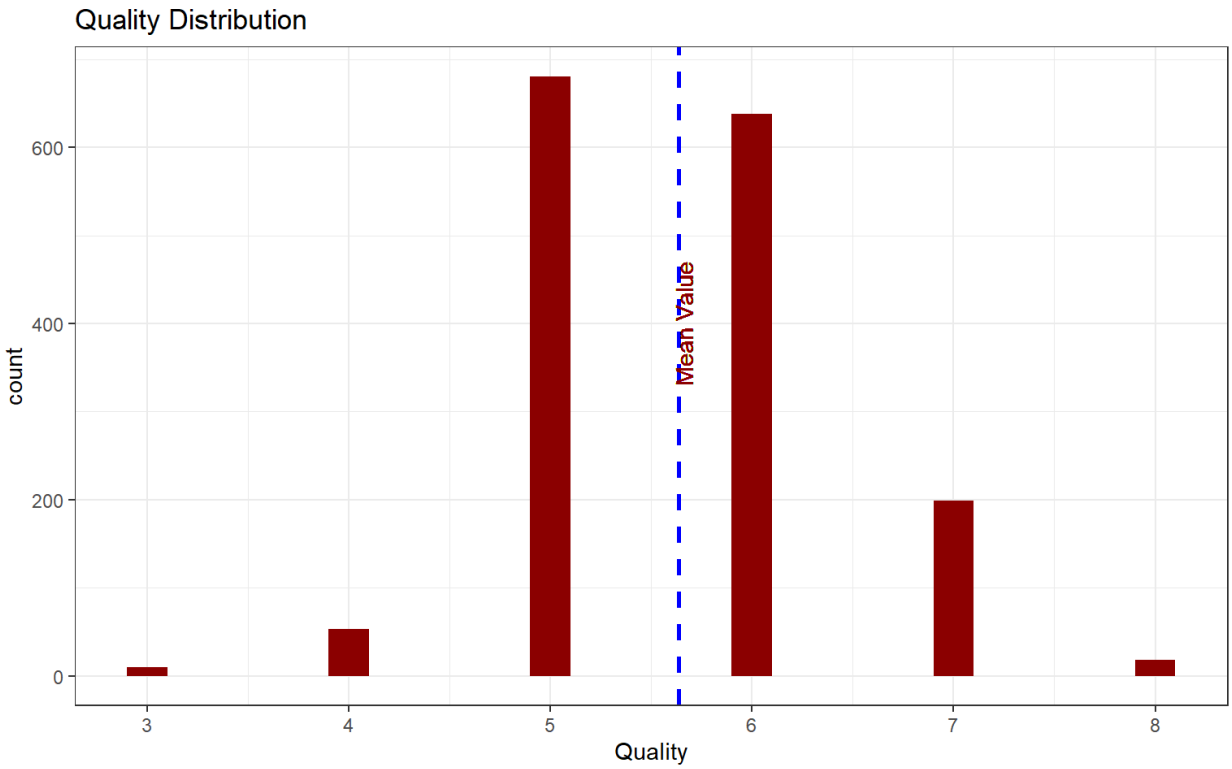
4. **Cross-validation / Train-Test Splitting:** The dataset was split into training and testing sets using an 80:20 ratio, ensuring that the proportions of the "high" and "low" quality classes were maintained in both sets.

EDA and Data Visualization: -

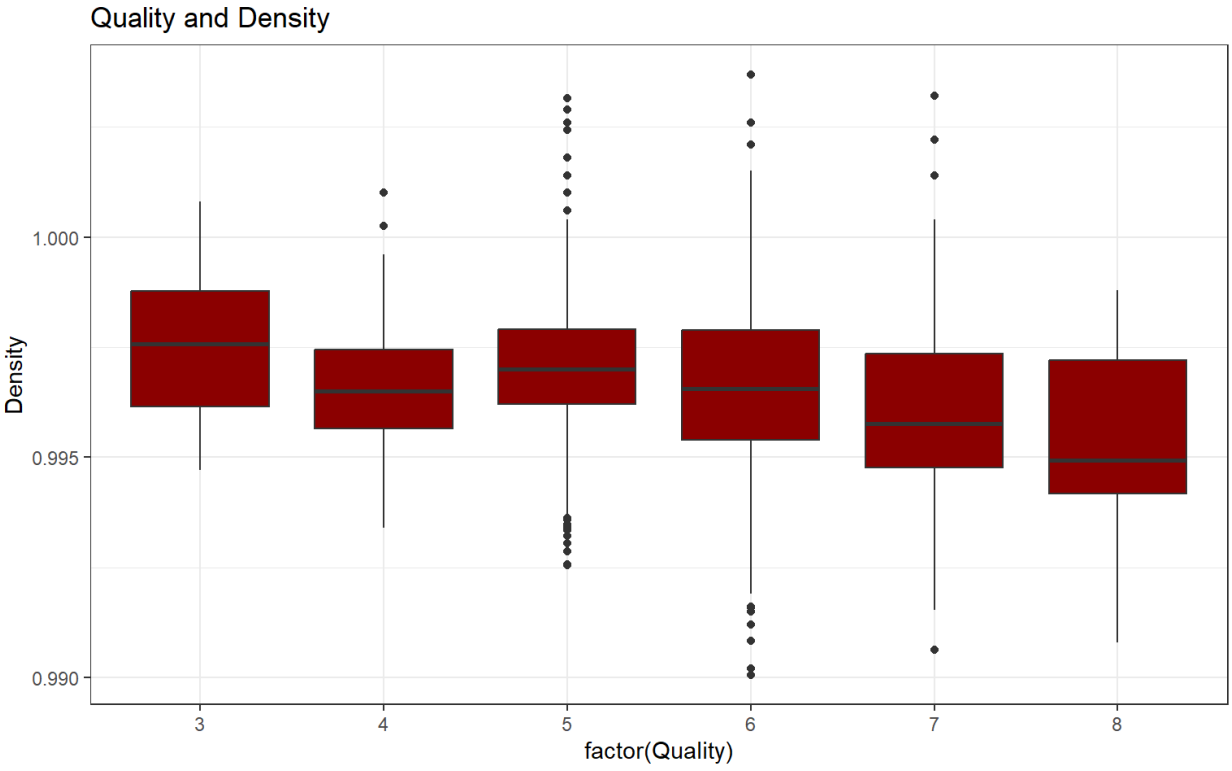
Overall Summary Statistics: Summary statistics were generated for all predictor variables, revealing key insights into their distributions and ranges. Outliers were noted in certain predictors such as fixed acidity, total sulfur dioxide, free sulfur dioxide, and sulphates. –

Univariate Plots: Exploratory data analysis (EDA) involved visualizing the characteristics of the data through histograms, boxplots, and density plots. These plots provided further insights into the distributions and potential outliers in the predictor variables.

c. Some figures that show the characteristics of data: Exploratory analysis revealed various characteristics of the data: - Some predictors exhibit outliers, as indicated by notably larger maximum values compared to the mean, median, or third quartile. - Variables like density and pH demonstrate relatively normal distributions, with maximum values aligning closely with the median or third quartile of the data. - While outliers may need to be addressed in subsequent analyses, for the current exploration, the data will be analyzed as is to observe its impact on predictive models.

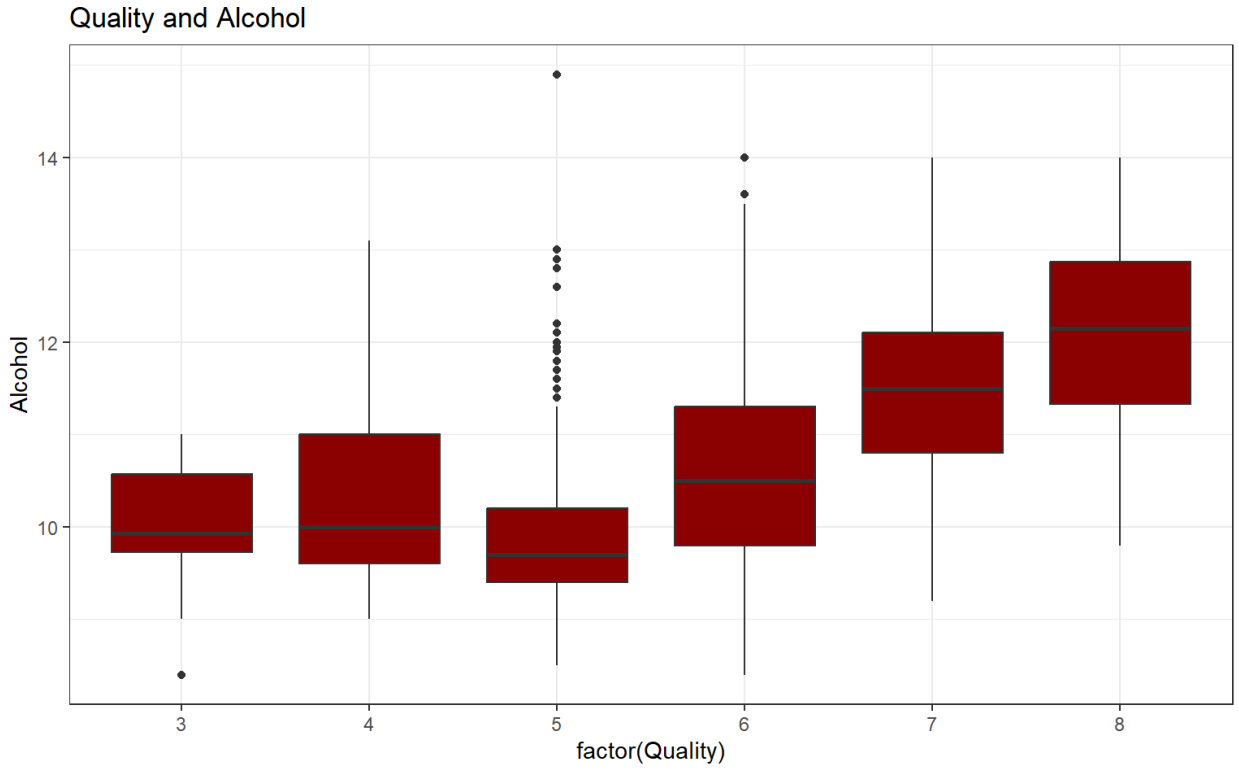


As we can see, vast majority of the quality of wine is around 5 and 6, which is 82.4890557%. This also means that our data set is unbalanced. This makes it harder for us to pin point factor that could affect Quality in any possible way.

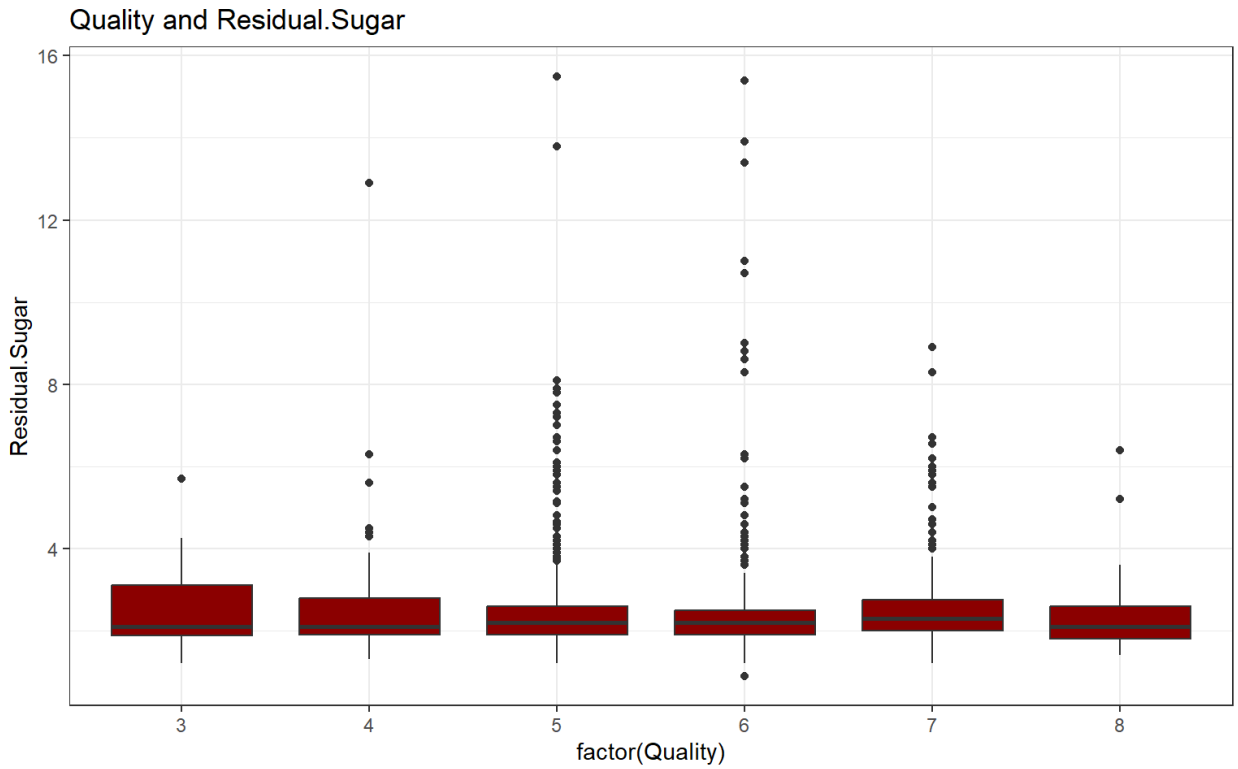


name	type	na	mean	disp	median	mad	min	max
Density	numeric	0	0.9967467	0.0018873	0.99675	0.0016753	0.99007	1.00369

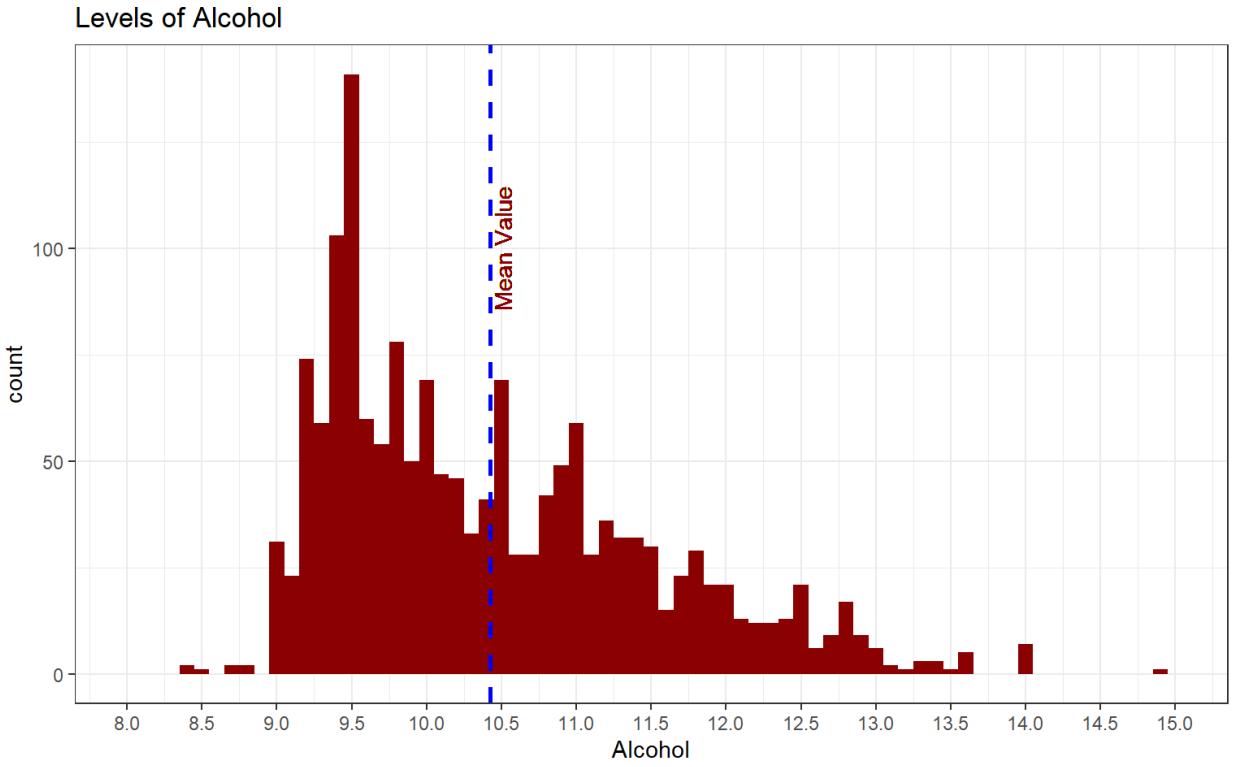
Clearly Density, doesn't have a big effect on the Quality, but there are clear outliers.



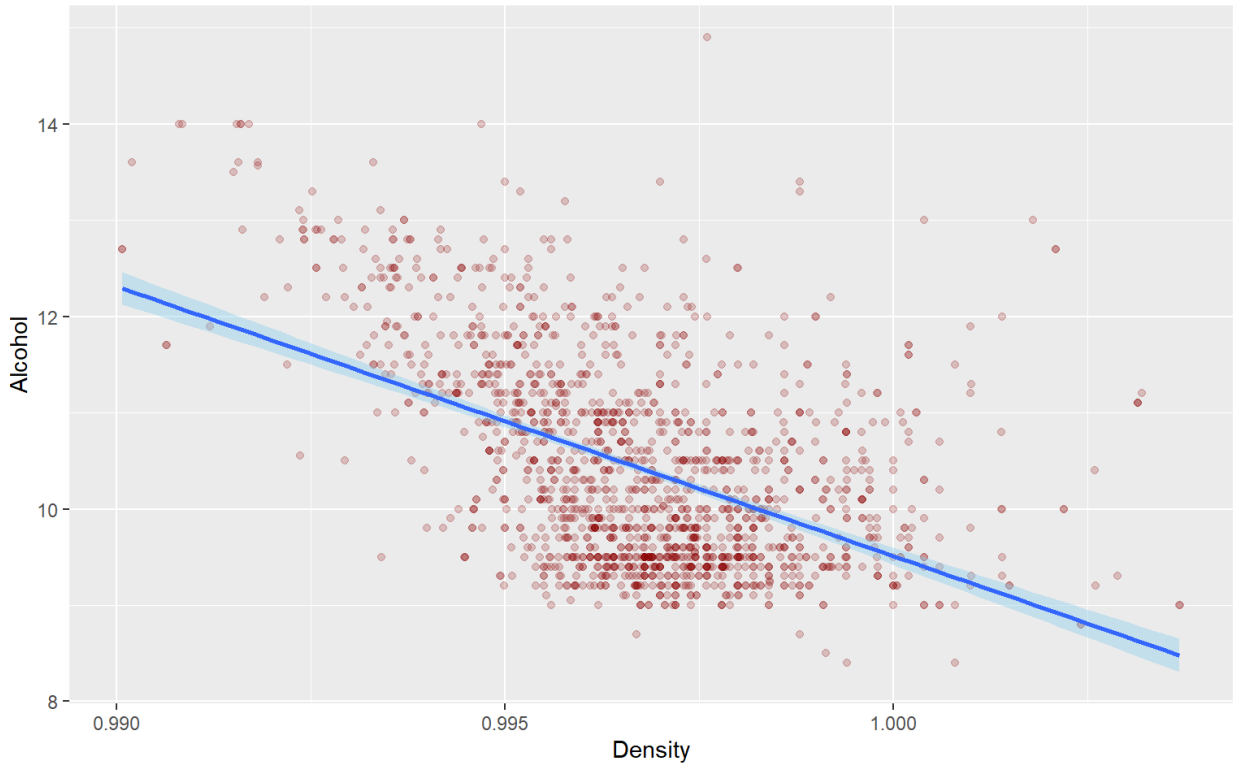
We can see that the percentage of alcohol is positively correlated with the quality. Higher quality wines have a bigger percentage of alcohol in them.



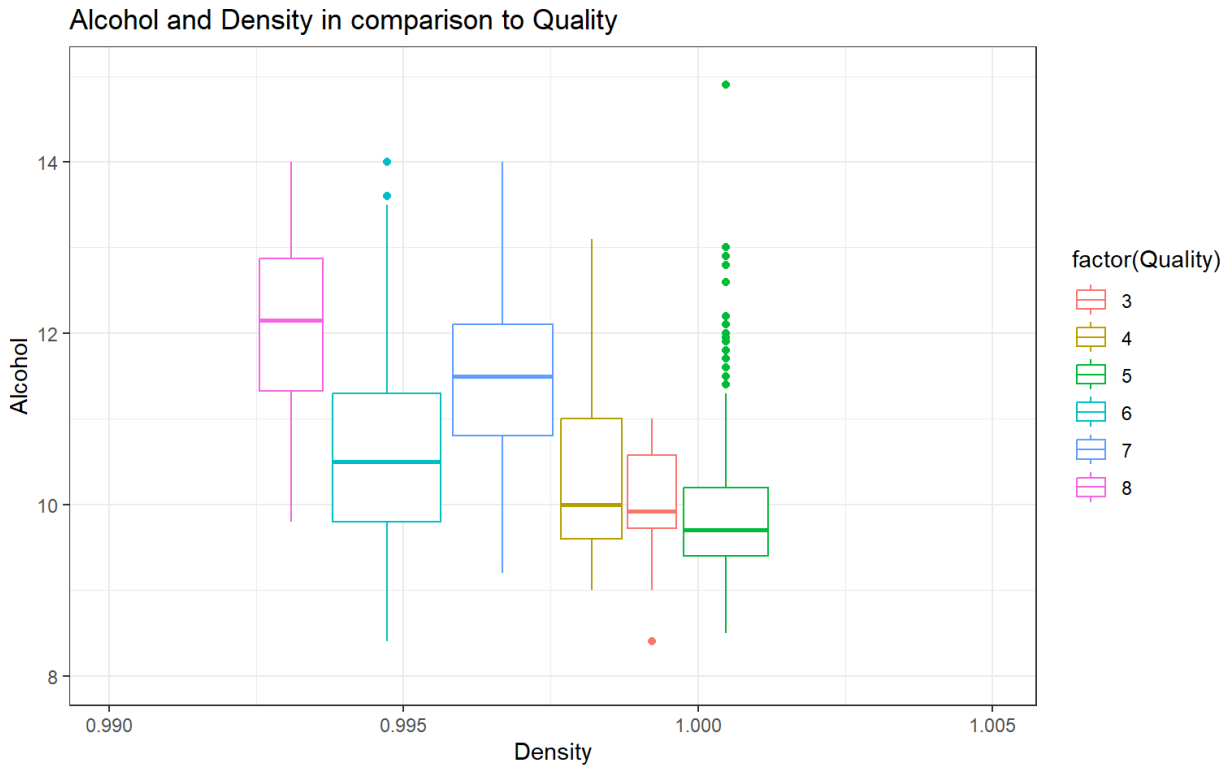
We can see that there is no significant relationship between Quality and Residual.Sugar, also the outlier values are not impacting the Quality, which means we can safely discard Residual.Sugar in the rest of our analysis.



It looks like levels of alcohol are skewed, this could be because the data set is relatively small. Here we can see that most frequently wines have 9.5% of alcohol in them, mean is 10.42%. We already saw that Density doesn't have an important effect on Quality, but let's see is there any meaningful information we can find when we look at Density and Alcohol:



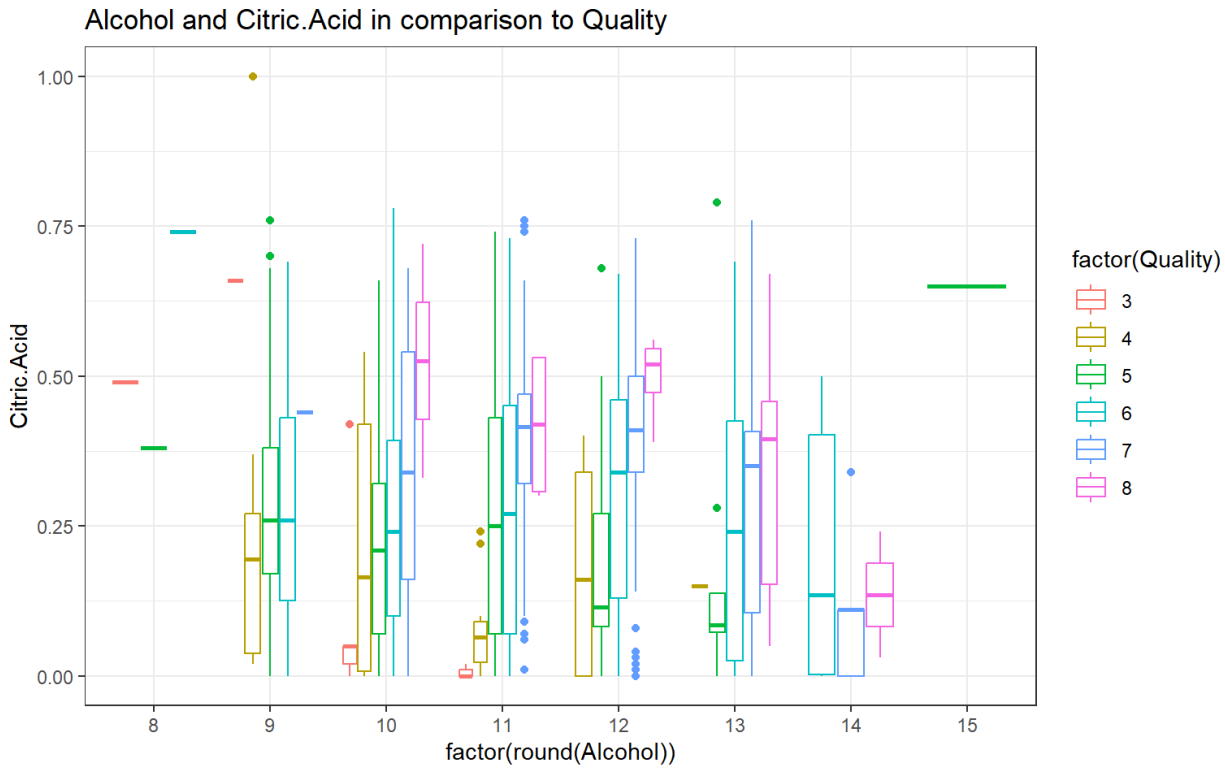
We can't really conclude anything without making a correlation with Quality. So now we will combine Alcohol and Density as we know they are in a strong relationship and see how they together correlate with Quality.



name	type	na	mean	disp	median	mad	min	max	nlevs
Density	numeric	0	0.9967467	0.0018873	0.99675	0.0016753	0.99007	1.00369	0
Alcohol	numeric	0	10.4229831	1.0656676	10.20000	1.0378200	8.40000	14.90000	0

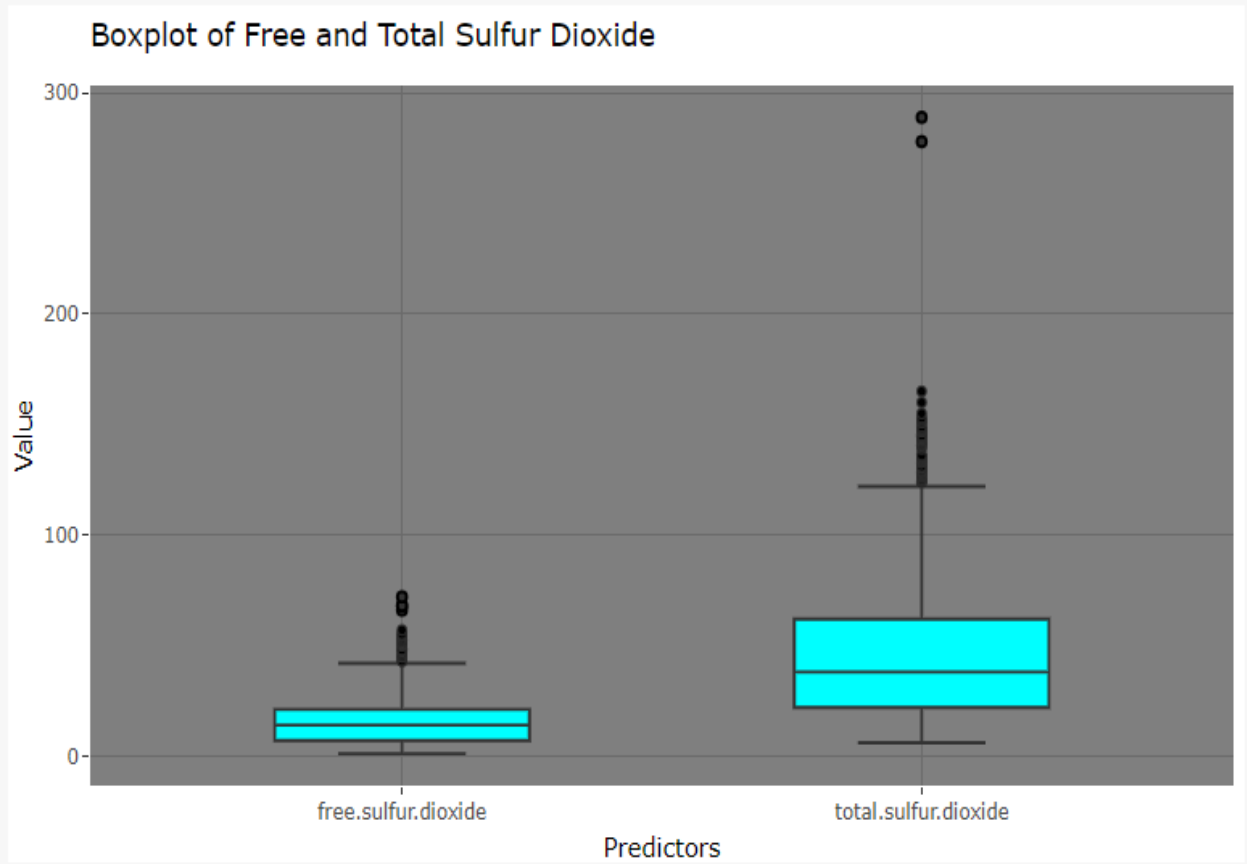
As Density increases, Quality decreases. It's not clear from this examination, how much Density actually affects the quality of wine because Alcohol has the reverse effect on Quality to a similar degree. Because we know Alcohol causes change in Density it would be wise to say that Alcohol affects both Density and Quality.

Alcohol vs Citric.Acid/Chlorides/Volatile.Acidity/Sulphates in correlation to Quality.



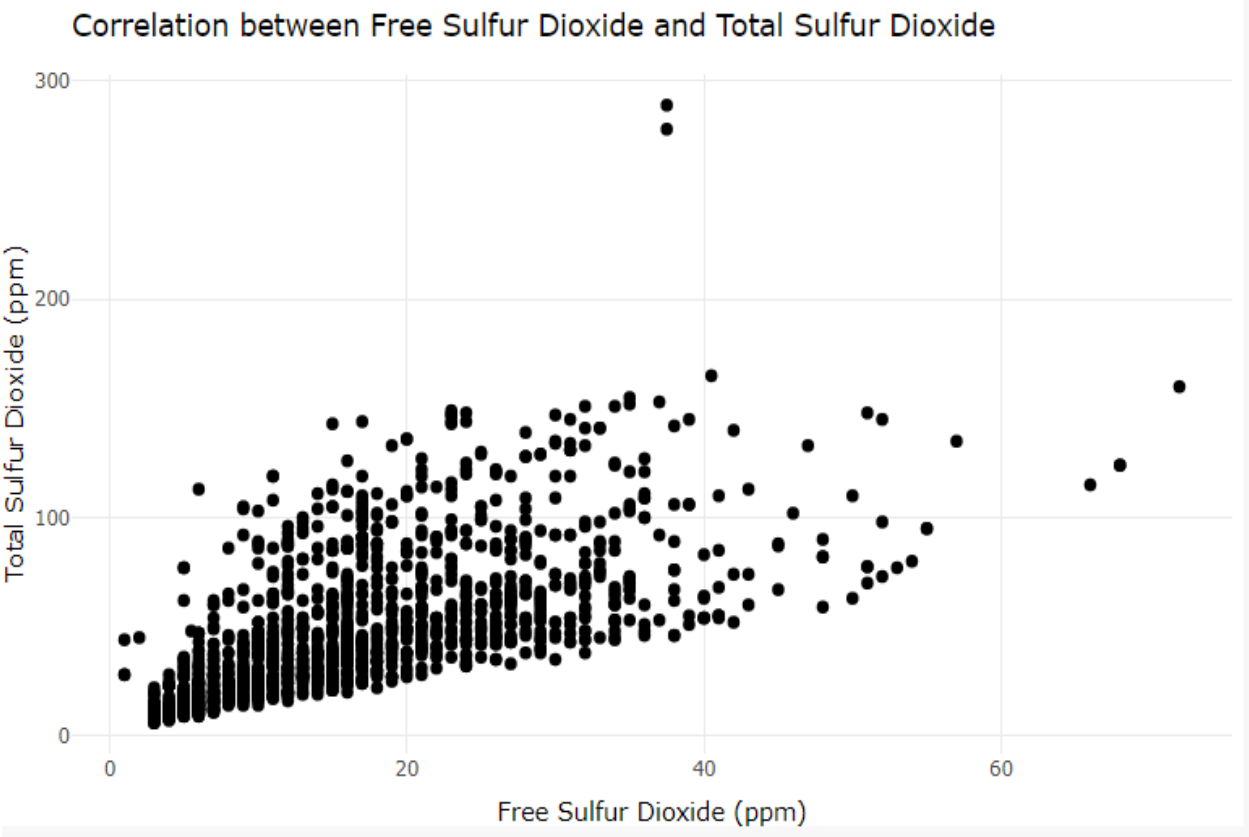
name	type	na	mean	disp	median	mad	min	max	nlevs
Citric.Acid	numeric	0	0.2709756	0.1948011	0.26	0.252042	0.0	1.0	0
Alcohol	numeric	0	10.4229831	1.0656676	10.20	1.037820	8.4	14.9	0

Firstly, we can see that all wines with percentage level below 14 have a positive correlation between Citric.Acid and Alcohol, which means as the level of Citric.Acid increases so does the Quality of the wine. Conversely, lower-quality of wines have low values of Citric.Acid.



The presence of numerous outliers among most predictors suggests caution against their removal, as doing so may lead to significant information loss. Notably, predictors such as alcohol and citric acid exhibit relatively normal and consistent distributions, with few outliers and median values positioned centrally within their respective boxplots. This suggests that these predictors may offer reliable and valuable insights into the dataset without the need for extensive data manipulation.

Relationship b/w Free Sulfur Dioxide and Total Sulfur Dioxide:



While a discernible trend line suggests a positive correlation between the two variables, the scattered spread of data points indicates considerable randomness. As such, despite potential linear relationships, it appears unnecessary to prioritize special feature engineering for these variables.

Models

a. What model(s) are you using for this project. Why? For this project, we employed several regression and classification models to predict the quality of red wine based on its chemical attributes. The models used include:

1. **Linear Regression:** This model is chosen for its simplicity and interpretability. It helps us understand the linear relationship between the predictor variables and the response variable, wine quality.
2. **Decision Tree:** Despite its tendency to overfit, decision trees are intuitive and easy to interpret. They partition the data based on feature splits to predict the target variable.
3. **Random Forest:** Random Forest is an ensemble learning technique that combines multiple decision trees to improve predictive accuracy and reduce overfitting. It's known for handling high-dimensional data well and is robust against outliers and noise.

b. What part of the dataset are you using to fit the model? We used the training set to fit the models. The training set comprises a randomly selected 80% of the entire dataset, ensuring that observations from both high and low-quality wines are represented in the training data. This approach helps the models learn patterns and relationships between the predictor variables and the target variable.

c. What part of the dataset are you using to validate the model? The test set is utilized to validate the performance of the models. It consists of the remaining 20% of the dataset that was not included in the training set. The test set serves as an independent dataset to assess how well the models generalize to unseen data. By evaluating the models on the test set, we can estimate their performance on new, unseen wine samples.

d. Can you provide some model parameters? For the Linear Regression model, the coefficients of predictors are provided. Here's an example of the model parameters:

```
lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol + fixed.acidity,
    data = wine_test)
```

Coefficients: - Intercept: 1.773247 - Volatile Acidity: -1.215154 - Chlorides: -0.256637 - Free Sulfur Dioxide: 0.009265 - Total Sulfur Dioxide: -0.006150 - pH: 0.112568 - Sulphates: 1.047150 - Alcohol: 0.284235 - Fixed Acidity: 0.075909

Evaluation & Results

a. How do you measure the success of your model?

- **Regression Models:** The success of regression models is typically measured using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or R-squared. These metrics quantify the difference between the predicted values and the actual values of the response variable (wine quality).

- **Classification Models:** For classification models, success is often measured using metrics such as accuracy, precision, recall, and F1-score. These metrics evaluate the model's ability to correctly classify instances into their respective classes.

b. Comparison of Different Models:

- **Regression Models:** - Linear Regression: Adjusted R-squared of 44.59%. - Decision Tree: Not suitable due to errors in leaf nodes. - Random Forest: Outperformed other regression models with a score of 47.49% in explaining quality.

- **Classification Models:** - Decision Tree: Achieved 75.00% accuracy and 79.35% precision. - Random Forest: Outperformed other classification methods with 83.12% accuracy and 81.28% precision.

c. Interpretation of Model Results:

- **Regression Models:** - Linear Regression: Indicates that about 44.59% of the variation in wine quality can be explained by the predictor variables included in the model. However, caution is advised due to heteroscedasticity. - Decision Tree: Found to be unfit for estimating wine quality due to errors in leaf nodes. - Random Forest: Provides the best performance among regression models, explaining approximately 47.49% of the variability in wine quality.

Linear Model using All Predictors (as baseline):

```
## lm(formula = quality ~ ., data = wine_train %>% select(-quality_high))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60102 -0.37142 -0.06258  0.45407  1.97898
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.739e+01  2.330e+01   0.746 0.455659
## fixed.acidity        -5.110e-03  2.883e-02  -0.177 0.859325
## volatile.acidity     -1.013e+00  1.385e-01  -7.316 4.51e-13 ***
## citric.acid          -1.343e-01  1.680e-01  -0.799 0.424307
## residual.sugar        1.565e-02  1.600e-02   0.978 0.328300
## chlorides            -2.136e+00  4.522e-01  -4.723 2.59e-06 ***
## free.sulfur.dioxide   4.217e-03  2.418e-03   1.744 0.081371 .
## total.sulfur.dioxide -2.961e-03  8.079e-04  -3.665 0.000257 ***
## density              -1.239e+01  2.378e+01  -0.521 0.602442
## pH                   -6.175e-01  2.154e-01  -2.868 0.004205 **
## sulphates             8.816e-01  1.304e-01   6.760 2.10e-11 ***
## alcohol               2.778e-01  2.963e-02   9.377  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6476 on 1267 degrees of freedom
## Multiple R-squared:  0.3422, Adjusted R-squared:  0.3365
## F-statistic: 59.93 on 11 and 1267 DF,  p-value: < 2.2e-16
```

Interpretations from this simple model indicate:

- Variables such as volatile acidity, chlorides, total sulfur dioxide, sulphates, and alcohol exhibit high significance based on their p-values.

- Free sulfur dioxide and pH, while not as highly significant, still hold significance in the model.
- The adjusted R-squared value, at 35.61%, suggests a relatively low level of explanatory power, indicating room for improvement by eliminating unnecessary predictors.
- Currently, no variables demonstrate perfect separation or singular predictive power for the target variable, as evidenced by p-values that are not extremely insignificant or close to 1. Thus, solely relying on machine learning for prediction based on these variables would not be advisable.

Model using All Predictors (as baseline):

```
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol + fixed.acidity,
##     data = wine_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25444 -0.39954  0.01312  0.39674  1.81988
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.773247   1.259682   1.408 0.160220
## volatile.acidity     -1.215154   0.223025  -5.449 1.03e-07 ***
## chlorides            -0.256637   1.188869  -0.216 0.829233
## free.sulfur.dioxide   0.009265   0.004887   1.896 0.058934 .
## total.sulfur.dioxide -0.006150   0.001636  -3.759 0.000204 ***
## pH                    0.112568   0.325102   0.346 0.729386
## sulphates             1.047150   0.231117   4.531 8.39e-06 ***
## alcohol               0.284235   0.035207   8.073 1.52e-14 ***
## fixed.acidity         0.075909   0.028401   2.673 0.007921 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6378 on 311 degrees of freedom
## Multiple R-squared:  0.4598, Adjusted R-squared:  0.4459
## F-statistic: 33.09 on 8 and 311 DF,  p-value: < 2.2e-16
```

Prediction using Test Data: Given that our model meets most assumptions, rather than predicting the test data separately, we can leverage the predictors from the best model identified during our comparisons and assumption verification process. Surprisingly, the

model's performance on the test data was marginally better than on the training data, with an explained variance of approximately 44.59% for the quality variable.

Conclusions: Our top-performing simple linear regression model, derived through both manual and stepwise approaches, includes predictors such as volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol, and fixed acidity. While the adjusted R-squared for the training data stands at 33.73%, it notably improves to 44.59% on the test data. Although the model falls short of meeting all assumptions, failing only one, further exploration is warranted to determine if objectively superior models exist.

Model Evaluation RF Regression Model:

```
## Random Forest
##
## 1279 samples
##   11 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold, repeated 3 times)
## Summary of sample sizes: 1023, 1023, 1023, 1024, 1023, 1023, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##    2    0.5886977  0.4635493  0.4416328
##    6    0.5867287  0.4587700  0.4332473
##   11    0.5877175  0.4558971  0.4321778
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 6.
```

```
wine_forest_reg$finalModel
```

```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##           Mean of squared residuals: 0.3317218
##                     % Var explained: 47.49
```

In the context of Random Forest modeling, "mtry" denotes the number of predictors utilized in fitting a new Decision Tree model. After multiple iterations, our analysis reveals that the algorithm selected an "mtry" value of 6, as it corresponds to the lowest Root Mean Square Error (RMSE), representing an estimate of model error.

Random Forest incorporates its own cross-validation technique known as Out-Of-Bag (OOB) Error. This approach involves randomly segregating data and evaluating the model's performance on unseen data, akin to manual validation procedures. Consequently, the use of traditional cross-validation or train-test splitting techniques becomes unnecessary when employing Random Forest.

Despite Random Forest's model accuracy being assessed at 47.49%, which may seem relatively low, it represents an improvement compared to our initial linear regression model. This improvement, albeit modest, underscores the efficacy of Random Forest in capturing complex relationships within the data.

- **Classification Models:** - Decision Tree: Achieves reasonable accuracy and precision, making it interpretable but potentially less accurate than other methods. - Random Forest: Offers the highest accuracy and precision among classification methods, making it a robust choice for predicting wine quality.

Model Evaluation Using Customized Parameters: DT

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 115  48
##          1  34 123
##
##                Accuracy : 0.7438
##                  95% CI : (0.6922, 0.7907)
##     No Information Rate : 0.5344
##     P-Value [Acc > NIR] : 1.044e-14
##
##                   Kappa : 0.4882
##
##  Mcnemar's Test P-Value : 0.1511
##
##             Sensitivity : 0.7193
##             Specificity : 0.7718
##          Pos Pred Value : 0.7834
##          Neg Pred Value : 0.7055
##              Prevalence : 0.5344
##          Detection Rate : 0.3844
##    Detection Prevalence : 0.4906
##       Balanced Accuracy : 0.7456
##
##        'Positive' Class : 1
```

Accuracy: 74.38%

Precision / Positive Predictive Value: 78.34%

Although slightly lower than the default parameters, this configuration offers improved readability.

Conclusions:

Comparing our Decision Tree model with default parameters to a pruned alternative, the former demonstrates superior metrics. The predictors - alcohol, volatile acidity, chlorides, sulphates, and total sulfur dioxide - contribute to its effectiveness. Despite this, we've developed a pruned model with customized parameters for enhanced readability, featuring fewer leaf nodes. The accuracy of our best Decision Tree model stands at 75.00%, with a precision of 79.35%.

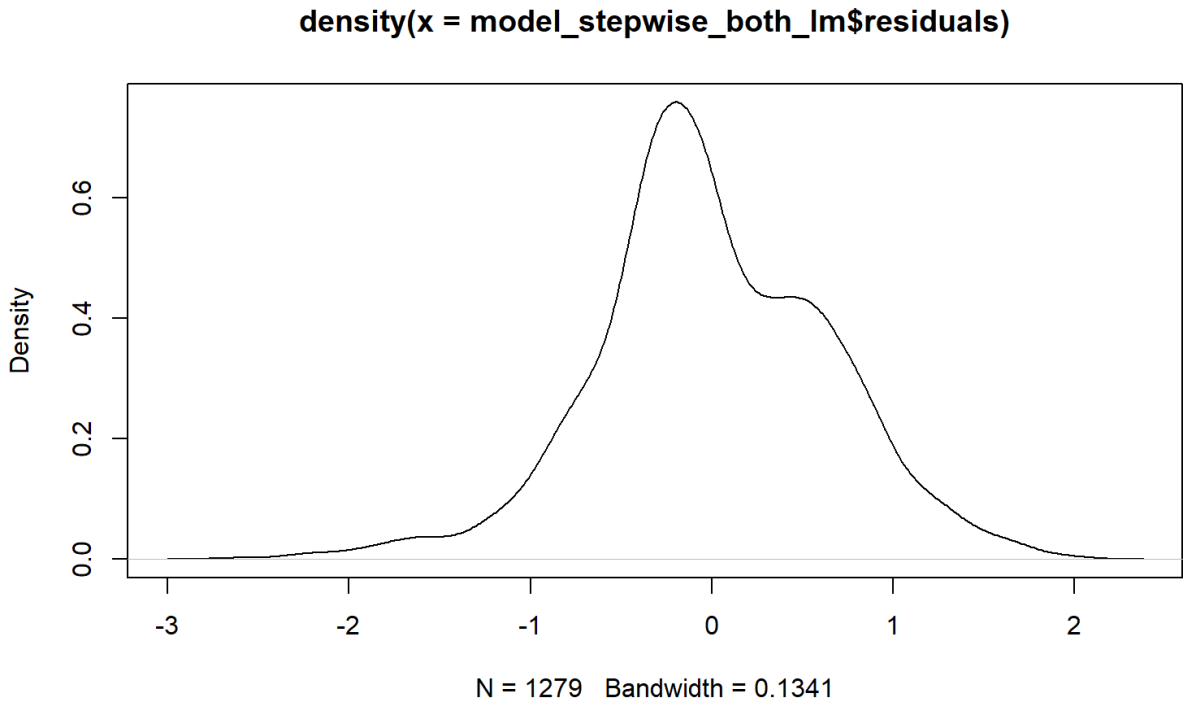
Model Evaluation RF Model

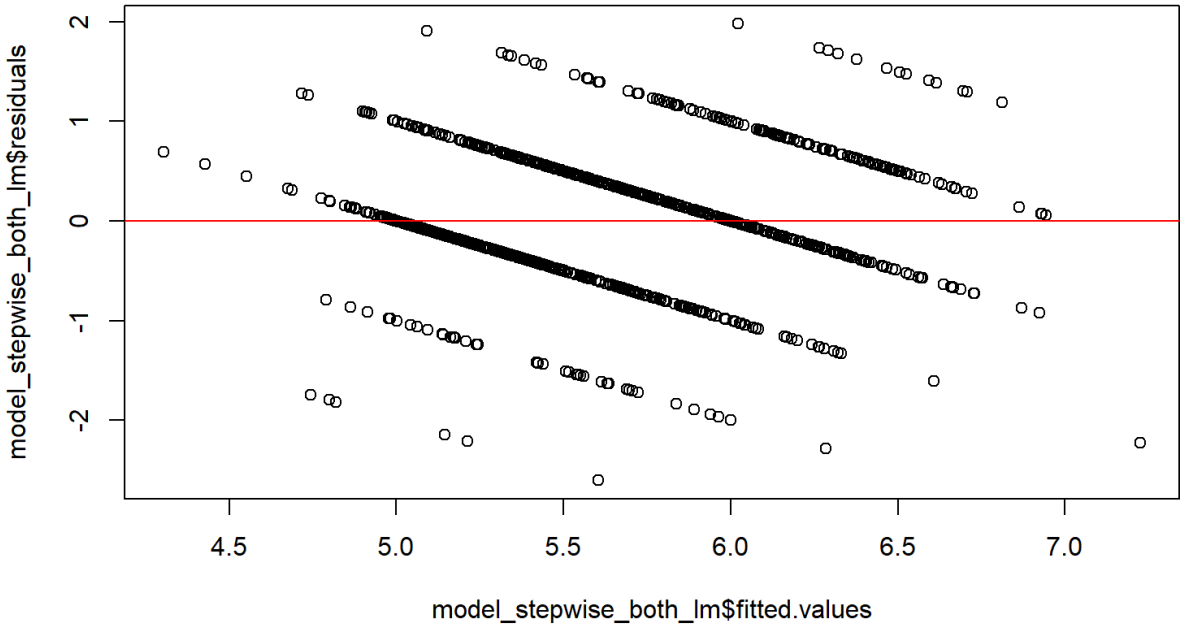
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 114  19
##          1  35 152
##
##                Accuracy : 0.8312
##                  95% CI : (0.7856, 0.8706)
##     No Information Rate : 0.5344
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.6585
##
##  Mcnemar's Test P-Value : 0.04123
##
##             Sensitivity : 0.8889
##             Specificity : 0.7651
##          Pos Pred Value : 0.8128
##          Neg Pred Value : 0.8571
##              Prevalence : 0.5344
##          Detection Rate : 0.4750
##    Detection Prevalence : 0.5844
##       Balanced Accuracy : 0.8270
##
##        'Positive' Class : 1
```

With an impressive accuracy of 83.12% and a precision of 81.28%, the Random Forest model surpasses previous performance benchmarks, marking the first instance where any prediction models, be it regression or classification, achieve an accuracy rate exceeding 80%. This unequivocally positions Random Forest as one of the most reliable models for predicting target variables. In regression, the model demonstrates a capability to explain 47.49% of the target variable, leaving 52.51% attributed to predictors outside those utilized. However, in classification, its accuracy and precision metrics objectively outshine all other models, solidifying Random Forest as the top performer across both domains.

d. Figures Showing Model Results: - Figures such as confusion matrices, ROC curves, or scatter plots comparing predicted vs. actual values can visually represent the performance of the models. These visualizations help in understanding how well the models are performing in terms of classification accuracy or regression fit.

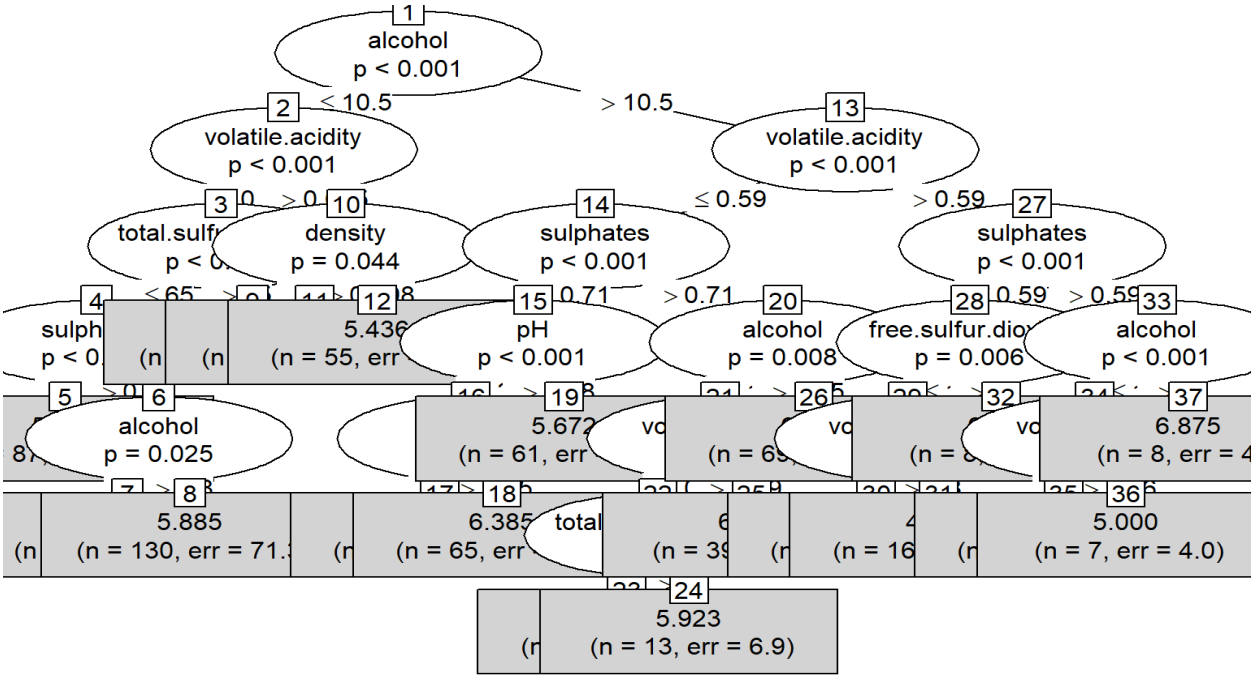
Linear model: it's not your usual normal distribution bell curve, but I would think that this is close enough, that we can pass this assumption.





The dataset exhibits patterns that suggest it may not be well-suited for linear regression analysis, corroborating the results of the BP Test.

Decesion tree plot:



Discussion/Future Work

a. Conclusion of Your Project: The analysis of the Red Wine Quality dataset revealed several key insights into the factors influencing wine quality. The examination of various regression and classification models highlighted the importance of predictors such as Alcohol, Volatile Acidity, Sulphates, and Total Sulfur Dioxides in determining wine quality. Among the models explored, Random Forest emerged as the most robust method, offering a balance between accuracy, precision, and resource utilization. However, it's essential to note that the quality levels of wines in the dataset were unbalanced, which likely impacted the overall results. Additionally, factors like the year of production, brew time, location, and wine brand could have significant effects on wine quality but were not included in the analysis due to data limitations.

b. Potential Improvements: To enhance this work, several avenues for future exploration and improvement can be considered:

1. **Addressing Data Imbalance:** Strategies to mitigate the impact of data imbalance on model performance should be explored. Techniques such as oversampling, undersampling, or using advanced algorithms designed for imbalanced data could be implemented.

2. **Feature Engineering:** Further feature engineering techniques could be employed to derive new predictors or transform existing ones to better capture the underlying patterns in the data.

3. **Incorporating Additional Variables:** Gathering additional data on factors like the year of production, brew time, location, and wine brand could provide valuable insights into their impact on wine quality.

4. **Model Tuning:** Fine-tuning the hyperparameters of the models, particularly the Random Forest model, could potentially improve its performance without significantly increasing computational demands.

By addressing these areas, future iterations of the analysis could provide more accurate and comprehensive insights into the factors influencing red wine quality, ultimately contributing to better-informed decision-making in the wine industry.

References

- Hadley Wickham, Garrett Grolemund. 2016. "R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'reilly Media." https://books.google.si/books?id=I6y3DQAAQBAJ. Link, Rachael. 2019. "What Are Sulfites in Wine? Everything You Need to Know." https://www.healthline.com/nutrition/sulfites-in-wine.
- Malik, Samartha. 2019. "Data Analysis and Visualisations Using r." https://towardsdatascience.com/data-analysis-and-visualisations-using-r-955a7e90f7dd.

- Moroney, Maureen. 2018. “Total Sulfur Dioxide - Why It Matters, Too!” https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too.
- Rashid, Miadad. 2015. “Wine Quality Exploration.” http://rstudio-pubs-static.s3.amazonaws.com/80458_5000e31f84df449099a872ccf40747b7.html.
- Savits, Jenny. 2019. “Sulfur Dioxide Measurement and Overestimation in Red Wine.” https://www.extension.iastate.edu/wine/sulfur-dioxide-measurement-and-overestimation-red-wine.
- Woods, David. 2017. “R Correlation Tutorial.” https://www.datacamp.com/community/blog/r-correlation-tutorial.