Data Analyst Exam Project

Contents

[Introduction 1](#_Toc134250500)

[Initial Assumptions and Hypothesis 1](#_Toc134250501)

[EDA 2](#_Toc134250502)

[Data Cleaning 2](#_Toc134250503)

[Descriptive Statistics 2](#_Toc134250504)

[Trends, patterns, and Anomalies 18](#_Toc134250505)

[Scatter plots – Insights 18](#_Toc134250506)

# Introduction

As an assistant to a restaurant chain manager, we have been given a task to perform a report about the quality of variants of Portuguese Vinho Verde. To give a scientific rating of the chemical components of the wines, we have analyzed a dataset based on a study performed by Cortez et al. (2009). After a thorough EDA-analysis, we have discovered various patterns and trends in the dataset, including hypothesis tests and validating our statements about the dataset. The results of this analysis will help the chemists choose the best wine type from the wine producer’s offers. In this report we will present our insight and conclusions in a logical and clear matter, to help the restaurant chain manager in making an informed decision and ensure they make the best possible investment.

# Initial Assumptions and Hypothesis

After reading through and gaining an understanding of our given dataset we have formed a few assumptions and hypotheses about the chemical components of the variants of Portuguese Vinho Verde. Our first assumption is that wines with higher levels of Citric Acid can have a less pleasant taste and therefore lower quality. Another assumption is that wines with higher levels of sulfur dioxide can have higher quality because of their antimicrobial and antioxidant abilities. Our hypothesis is that there is a correlation between alcohol percentage and higher quality, where wines with a moderate alcohol percentage have a higher quality than wines with a low or high alcohol percentage.

# EDA

## Data Cleaning

To start this EDA, we need to perform Data cleaning of the dataset. We need to find out if there are any missing values or errors in our dataset before starting the analysis. To check for missing values, we use the COUNTIFBLANK formula in Excel and to check for errors we use the COUNT(IF(ISERROR formula. After using these formulas, we can observe no missing values and errors in the dataset.

### Descriptive Statistics

Here we have created a table in Excel that contains the descriptive statistics of the dataset.

A picture containing text, font, line, number

Description automatically generated

Median – Measure of central tendency & less affected by outliers.

STD – Standard deviation shows how spread out the dataset is and how much the values differ from the average. The higher the STD, the higher spread is in the dataset.

Mode – Most frequently used number for each variable.

Mean – The mean represents the average, or the sum of all values divided by the number of values.

Skewness – This shows the symmetry of the dataset. If the distribution is skewed to the left or has a skewness value of less than -0.5, it is considered to be negatively skewed. If the distribution is skewed to the right or has a skewness value above 0.5, it is considered to be positively skewed.

Kurtosis – Measures how much the values are clustered around the mean. The higher the kurtosis, the more extreme values (heavier tails) in the variable. The lower the kurtosis, the more spread out (flatter) the distribution is.

We can see that the mean for quality is 5.64 and the mode is 5. This means that the majority of the wines are of average quality. The skewness is 0.22, which means that the distribution is positively skewed to the right. The kurtosis is moderately positive with a value of 0.3, which indicates that the distribution of the variable quality is flat and contains fewer extreme values or outliers. The standard deviation is 0.8, which means that on average each of the wines represented in the dataset is 0.8 away from the mean, 5.64.

For now, we will not analyze any of the remaining variables. Rather, if our analysis needs any support or more information, we can refer to our descriptive statistics.

#### Scatter plots – correlation between Quality and chemical components

Below we have found the correlation between the quality of wine and each chemical in the wine. We have created a scatter plot to visualize the correlation between the chemicals in wine and its quality.

First, we have Alcohol. From this scatter plot, we can see there is a correlation between having a higher percentage of alcohol in the wine and quality. And the lower percentage of alcohol one has, the higher chances are for lower-quality wine. To ensure the highest quality of wine, the wine must contain at least 9.8 % alcohol.

Here we can see the relationship between Sulphates and the quality of wine. We can see that the highest quality wine has a Sulphate between 0.63 and 1.1. If the wine contains more than 1.1 Sulphate, the quality can drop to 7, 6, and 5. And if the level of Sulphate is less than 1.1 and less than 0.6, it tends to drop to a quality of 3. To conclude, from this scatter plot, we can see that high levels of sulphate tend to make the quality of wine higher, but sulphate levels higher than 1.1 is not needed to make the highest quality wine. Additionally, lower levels of wine tend to make the quality of wine lower.

Here we can see the relationship between the quality of wine and pH levels. From this scatter plot, it is hard to see any insights into the relationship between quality in pH. If we compare quality 8 with 7, 6, and 5, we can see that the higher the wine’s pH levels, the lower the quality is. But they are practically the same if we compare 8 and 4. Therefore, we cannot say that the higher the pH levels are, the lower the quality of wine is. For now, we need further analysis on this topic to be able to draw any insights or conclusions.

Here we can see the relationship between the level of Density and the quality of the wine. From this scatter plot, we can see that the wine with the highest quality has a Density ranging between 0.9908 and 0.9988. If the Density goes above this, it can drop all the way down to a quality of 3, but the quality tends to only drop to 5, 4, or 3 only if the Density is above 0.9908. Overall, we can see that the higher the Density of wine, the lower the quality of wine is.

Here we can see the relationship between Total Sulfur Dioxide and the quality of wine. Overall, we can see that the higher the total sulfur dioxide there is in wine, the lower quality of the wine. The only exception from this is when the wine has a quality of 3. We can argue that the reason for this having a quality of 3 is that most are very low and with only a few outliers. The lowest sulfur dioxide for a wine with a quality of 8 is 12.8. We can argue that if the sulfur dioxide goes below this, it will drop and it has the potential to drop to 3. To conclude, there is a correlation between lower-quality wine and higher amounts of sulfur dioxide, but if the wine contains less than 12.8 sulfur dioxide it has a high potential to drop in and can even drop down to a quality of 3.

Here we can see the relationship between Volatile acidity and the quality of wine. Here there is a clear correlation between having a higher level of volatile acidity and a lower quality of wine. The best quality wine has a volatile acidity that tends to be between 0.26 and 0.62, with one outlier that has 0.85.

Here we can see the relationship between Free sulfur dioxide and the quality of wine. There are not many insights or conclusions that can be drawn from this scatter plot. If we compare a wine with a quality of 8 with 7, 6, and 5, we can see a correlation between lower-quality of wine and higher levels of free sulfur dioxide. But they are practically the same if we compare 8 with 4 and 3. Therefore, we cannot say that there is a correlation between higher levels of free sulfur dioxide and lower quality of the wine. For now, further analysis is needed to draw any insights or conclusions about the relationship between free sulfur dioxide and the quality of wine.

Here we can see the relationship between Chlorides and the quality of wine. We can see there is a clear correlation between lower-quality wine and higher levels of Chlorides. The wine with a quality of 8 has a level of chloride between 0.0556 and 0.083 and an outlier of 0.044.

Here we can see the relationship between Citric acid and the quality of wine. From this scatter plot, we can see that the wines with a quality of 5 to 8, overall have higher levels of citric acid compared to the wines with a quality of 3 and 4. Looking at the wines with a quality of 7 and 8, we see that most are scattered around the higher end with levels of citric acid, while especially 3 and 4 are scattered around the lower ends of citric acid. This indicates that there is a correlation between higher levels of citric acid and higher-quality wine.

Here we can see the relationship between the residual sugar and the quality of the wine. Comparing the wine with the quality of 8 with 7, 6, and 5 we can see that there is a correlation between higher levels of residual sugar and lower quality of wine. But by comparing wines with a quality of 8 with 3 and 4, we see they are very similar to each other. Therefore, we cannot draw any conclusions about residual sugar and the quality of the wine.

Here we can see the relationship between fixed acidity and the quality of the wine. We can see that wines with a quality of 5 to 8 tend to have higher levels of fixed acidity compared to 3 and 4. Therefore, we can say there is a correlation between higher levels of fixed acidity and higher quality of the wine.

##### Correlation Matrix

Here we have created a correlation matrix with colors where green indicates that two variables have a positive correlation, yellow has little correlation and red has a negative correlation.

The variable that has the highest positive correlation with the quality of wine is alcohol. This indicates that the higher levels of alcohol the higher quality of the wine will be. Volatile acidity is the chemical component that has the highest negative correlation with the quality of wine, meaning the more volatile acidity there is, the lower quality of wine there is as well.

The five variables of alcohol, sulphates, citric acid, fixed acidity, and residual sugar all have a positive correlation with the quality of the wine. If one wants to make a good quality wine, these are all good chemical components. The six variables of volatile acidity, total sulfur dioxide, density, chlorides, pH, and free sulfur dioxide all have a negative correlation with the quality of the wine.

A screenshot of a graph

Description automatically generated with low confidence

Now we need to compare this correlation matrix with our scatter plots showing the relationship between the quality of wine and each of its chemical components.

Alcohol:

* The correlation matrix supports our scatter plot in showing that there is a correlation between having higher levels of alcohol in wine and higher quality of the wine. Our scatter plot also shows that one does not need an alcohol percentage higher than 14 to make the highest quality wine.

Sulphates:

* The correlation matrix supports our scatter plot in showing that there is a correlation between having higher levels of alcohol in wine and higher quality of wine. Our scatter plot shows that to make the highest quality of wine, sulphate levels above 1.1 is not needed.

pH

* In our scatter plot for pH, we could not draw any insights or conclusions from it. Our correlation matrix is showing there is a -0.057 negative correlation between pH and the quality of wine. This indicates that the lower the pH levels in wine, the lower quality. But we must take into consideration that a negative correlation of -0.0057 is very low. The correlation between pH and the quality of wine is almost null and our scatter plot shows very similar levels of pH levels independent of the quality of wine. Therefore, we can argue that the pH level in wine does not affect the quality of wine. But still, we cannot draw any conclusions yet, further analysis must be done.

Density

* When analyzing the correlation between the density and quality of wine, we concluded by saying that the higher the density of wine, the lower quality. Looking at the correlation matrix, we can see that it has a negative correlation of -0.174, meaning the higher density of wine, the lower quality. And therefore, we can see that our correlation matrix supports our conclusion of the scatter plot.

Total Sulfur Dioxide

* In our scatter plot we concluded by saying there is a negative correlation between total sulfur dioxide and quality of wine and the total sulfur dioxide drop below a certain amount, it has the potential to drop a quality of 3. Looking at our correlation matrix, we see that it also has a negative correlation, -0.185.

Volatile acidity

* In our scatter plot we concluded by saying there is a clear negative correlation between volatile acidity and the quality of wine. Our correlation matrix also supports this statement by giving it a negative correlation of -0.395, the highest negative correlation of all chemical components in wine.

Free Sulfur Dioxide

* In our scatter plot we concluded that we did not see any clear insights or conclusions about the correlation between free sulfur dioxide and wine and that we needed further analysis. Looking at the correlation matrix there is a negative correlation of -0.050. This indicates that the higher amounts of free sulfur dioxide, the lower quality of wine. But we need to consider that a negative correlation of -0.050 is very low and we must argue that there is a possibility that the amount of free sulfur dioxide does not affect the quality of wine so much. Before making any conclusions, we must do further analysis of this topic.

Chlorides

* In our scatter plot we concluded by saying that there is a correlation between higher levels of chloride and lower-quality wine. Our correlation matrix supports this statement, as it has given a negative correlation of -0.128 between chlorides and the quality of wine.

Citric Acid

* In our scatter plot we concluded by saying there is a positive correlation between higher levels of citric acid and the quality of wine. Our correlation matrix supports this statement, as it has given a positive correlation of 0.226 between the relationship between citric acid and the quality of wine.

Residual Sugar

* In our scatter plot we concluded by saying we cannot draw any conclusion between the relationship of residual sugar and the quality of the wine. Our correlation matrix gives a positive correlation of 0.0137 between the variables. This indicates that higher levels of residual sugar in wine give a higher quality of wine.

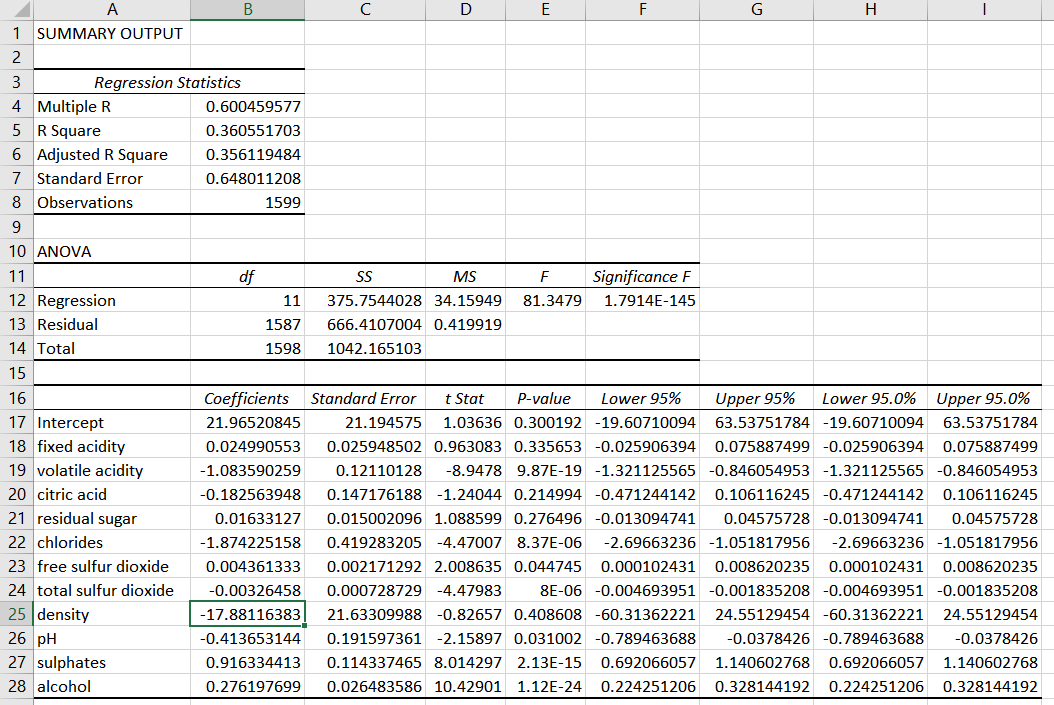
Fixed Acidity

* In our scatter plot we concluded by saying there is a positive correlation between fixed acidity and the quality of the wine. Our correlation matrix supports this statement and has given a positive correlation of 0.124.

###### Statistical hypotheses test – Regression model

Here we have created a regression model where quality is the dependent variable (Y), and the remaining 11 variables represent the chemical components are the independent variable (X). This Regression model is intended to answer the following question.

How do the 11 chemical components in wine affect the quality of the wine?



Here is a summary of the Regression model:

* Fixed Acidity – Positive coefficient, 2.50%.
* Volatile Acidity – Negative coefficient, -108.36%.
* Citric Acid – Negative coefficient, -18.26%.
* Residual Sugar – Positive coefficient, 1.63%.
* Chlorides – Negative coefficient, -187.42%.
* Free Sulfur Dioxide – Positive coefficient, 0.44%.
* Total Sulfur Dioxide – Negative coefficient, -0.33%
* Density – Negative coefficient, 1788.12%.
* pH – Negative coefficient, -41.37%.
* Sulphates – Positive coefficient, 91.63%.
* Alcohol – Positive coefficient, 27.62%.

A positive coefficient means that the variable has a positive impact on the quality of the wine. A negative coefficient means the variable has a negative impact on the quality of the wine.

We can see that the variables with a positive impact on the quality of the wine are Fixed Acidity, Residual Sugar, Free Sulfur Dioxide, Sulphates, and Alcohol. The variables with a negative impact on the quality of the wine are Volatile Acidity, Citric Acid, Chlorides, Total Sulfur Dioxide, and pH.

The variables with a P-value less than 5% are Volatile Acidity, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, pH, Sulphates, and Alcohol. Therefore, we have a 95% confidence level to say that the coefficient values for these 7/11 variables are true and we can reject the null hypothesis and say that the model is statistically significant. Additionally, we cannot trust the coefficient values for the remaining 4 variables, Fixed Acidity, Citric Acid, Residual Sugar, and Density. Therefore, we cannot reject the null hypothesis and can conclude by saying these four variables are not statistically significant.

The R square is 36%. This means that 36% of the quality of wine can be explained by the 11 independent variables included in the regression model. 64% of the quality of wine cannot be explained by these 11 independent variables.

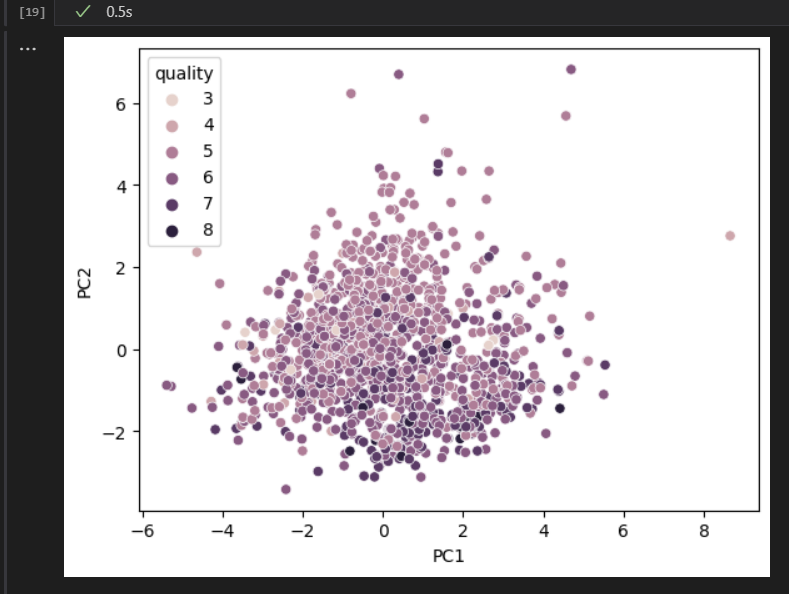
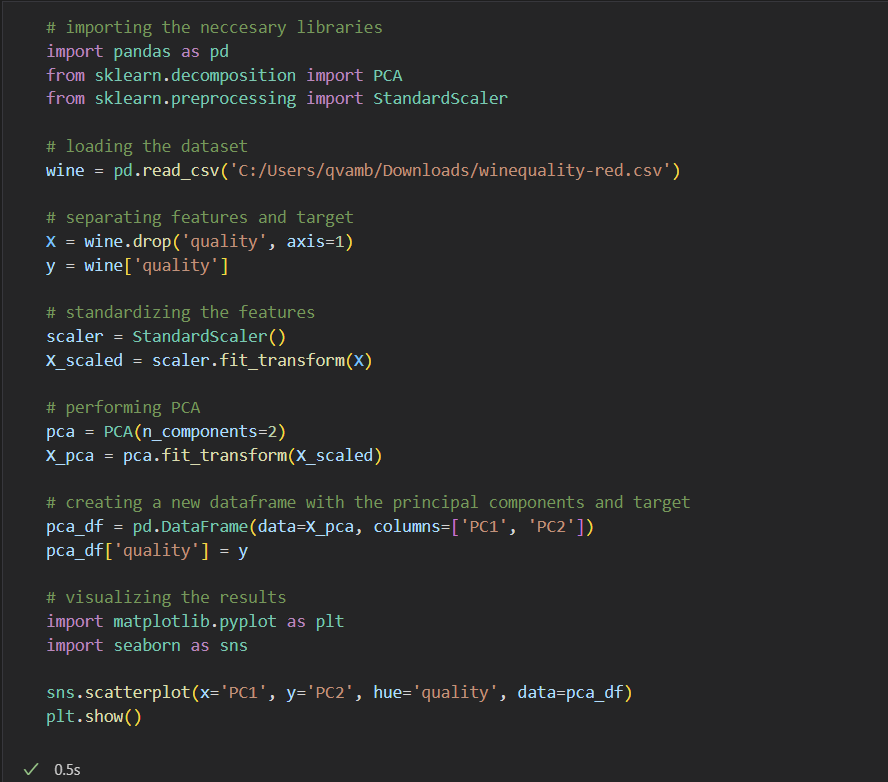
The regression model has a high F score of 81.34 and a low Significance F score of 1.7914E-145. This indicates that there is a significant correlation between the independent variables and the dependent variable and that this model is statistically significance. Therefore, we can conclude by saying that this model is a good fit for our data.

To conclude, our regression model is statistically significant and is a good representation of our data. According to our model, 36% of the quality of the wine can be explained by its 11 chemical components. 7 out of the 11 variables are statistically significant and 4 of the variables are not. Out of the seven significant variables, 4 of them is showing a negative coefficient which indicates that these variables have a poor impact on the quality of the wine. The remaining 3 variables have a positive coefficient and therefore a good impact on the quality of the wine. From this regression model, we can better understand which chemical components are beneficial for the wine and which ones are bad for the wine.

The following code is performing a component analysis (PCA) on the dataset to reduce the dimensionality from the 12 variables that represent the chemical components of wine. They are reduced to PC1 and PC2. The scatter plot shows the result between PC1 and PC2.

PCA - Principal Component Analysis

PCA is a technique that reduces the dimensionality and the size of the dataset and transforms the variables into a lower set, called PC1 and PC2.

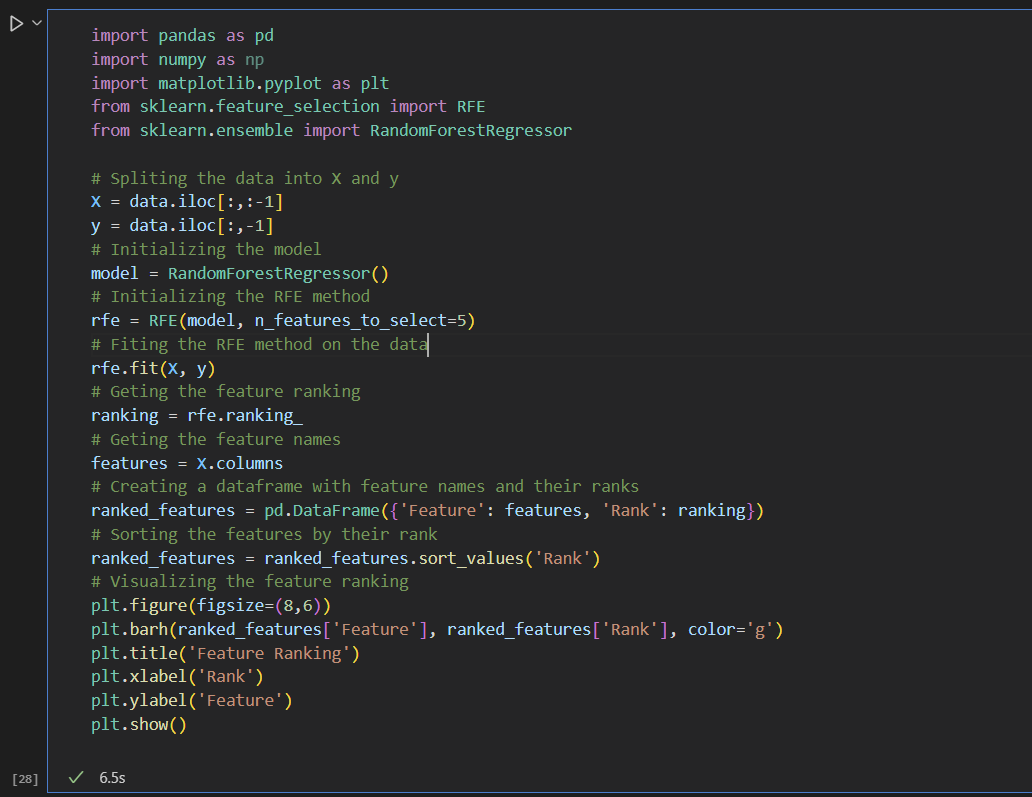


We can see that this scatter plot shows the quality levels from 3 to 8 are separated from each other, especially along the PC1 axis. This indicates that certain chemical components of the wine affect the wine to have a lower or higher quality.

In the PC2 axis, we can see that there is a separation between the quality levels and a high degree of clustering or overlapping. This indicates that there are certain chemical components in the wine that do not have such a big difference in the quality of wine.

Recursive feature elimination - RFE

Here we have created a recursive feature elimination. RFE is a feature selection method is used to fit a model and remove the weakest features in the model until specified numbers of features is reached. One of the benefits of using RFE is that it is good at removing the least important elements and selecting the most important ones so that it can make a better prediction of the target variable.



This chart shows the feature ranking for all 11 chemical components of the wine. A rank of 5 is considered to have the least importance of the quality of wine, while a rank of 1 has the most importance.

As we can see the chemical components of volatile acidity, chlorides, total sulfur dioxide, sulphates, and alcohol are the most important components that affect the quality of the wine. While free sulfur dioxide, citric acid, and fixed acidy has the least importance.

A picture containing text, display, screenshot, software

Description automatically generated

Feature importance ranking - FIR

Here we have created a feature importance ranking using RandomForestClassifier from sklearn.esemble and we used matplotlib.pyplot as pls to plot a bar chart of the results.

Feature importance ranking – FIR, measures the contributions of input features (variables) to the performance of a supervised learning model. This supervised learning model helps us to better understand which variables are important and less important in correlation with the target variable (quality).

A picture containing text, screenshot, display, software

Description automatically generated

Looking at our bar chart, we can see which variables or features of the target variable are important and less important in relation to the target variables (quality). The three most important ones are alcohol, sulphates, and volatile acidity. The three least important ones are free sulfur dioxide, residual sugar, and citric acid.

A screenshot of a graph

Description automatically generated with medium confidence

# Trends, patterns, and Anomalies

## Scatter plots – Insights

Our scatter plots show how relationship between each variable (chemical component) and its target variable (quality).

To summarize our findings from our scatter plots we will list which of the variables that have a positive, negative, and somewhat natural relationship with our target variable quality.

These are the chemical components of wine that have a positive relationship with the variable wine: Alcohol, Sulphates, Density, Citric Acid, and Fixed Acidity. If the wine contains higher amounts of these chemical components, the quality tends to be higher.

These are the chemical components that have a negative relationship with the quality of wine: Total Sulfur Dioxide, Volatile Acidity, and Chlorides.

These are the chemical components where we did not observe any difference in the quality of wine: pH, Free Sulfur Dioxide, and Residual Sugar.

Overall, from our scatter plots 5/11 of chemical components have a positive correlation with the quality of wine, 3/11 has a negative correlation, and 3/11 did not have any correlation between positive or negative.

### Correlation Matrix – Insights

Introduction -

Our correlation matrix showed 5 positive correlations: Alcohol, Sulphates, Citric Acid, Fixed Acidity, and Residual Sugar.

Our correlation matrix showed 6 negative correlations: Density, Total Sulfur Dioxide, Volatile Acidity, pH, Chloride, and Free Sulfur Dioxide.

That being said three variables had a correlation value less than 0.0, Residual Sugar, pH, and Free Sulfur Dioxide. When the correlation value is less than 0.0, we can argue that these variables whether positive or negative, are pretty much neutral and that they do not affect the quality of wine too much. But it can still be valuable to know which variables have a positive or negative value, even if it is very little.

Comparatively, findings from our scatter plot shared similar outcomes as our correlation matrix. It is important to see that our scatter plot and matrix correlation are similar, as it indicates that we can trust the result of these approaches and that our conclusions and findings are not by chance.

#### Statistical Hypothesis Test – Regression Model

To sum up, the findings from our Regression Model, the chemical components with a positive coefficient on the quality of wine are Alcohol, Sulphates, Citric Acid, Fixed Acidity, and Residual Sugar. The chemical components with a negative coefficient on the quality of wine are Density, Total Sulfur Dioxide, Volatile Acidity, pH, Chloride, and Free Sulfur Dioxide.

Comparatively to our correlation matrix, the results are similar. Both show the same variables with a positive correlation, and both show the same variable which has a negative correlation. Another similarity is the variables of Residual Sugar, pH, and Free Sulfur dioxide, they were very low just like their correlation value in the correlation matrix were also very low. This again indicates that these three chemical components have little effect on the quality of the wine.

Our Regression models help to support the statements we made in our matrix correlation and scatter plot, as it has similar results and findings.

##### PCA – Principal Component Analysis

In our PCA analysis, we observed that certain chemical components in wine make the quality better or worse. We also observed that certain components have little effect on the quality of wine.

These observations correlate with the findings from our scatter plots, correlation matrix, and regression model, where we also observed variables with positive correlation, variables with negative correlation, and variables with very low effects on the quality.

Our PCA analysis shows that there are positive, negative, and natural chemical components to wine. In our scatter plots, correlation matrix, and regression model, we can observe which chemical components are positive, negative, or neutral.

###### Recursive Feature Elimination – RFE

Our recursive feature elimination model is the first model that challenges some of the points we have made so far. According to our RFE model Alcohol, Sulphates, Volatile Acidity, Total Sulfur Dioxide, and Chlorides are the variables with the most impact on the quality of the wine. This correlates with our findings so far. But when it comes to the least important ones Free Sulfur Dioxide, Citric Acid, and Fixed acidity are the chemical components with the least effect on the quality of wine.

On the other hand, our scatter plots, correlation matrix, and regression model all indicated that pH, Free Sulfur Dioxide, and Residual Sugar have the least impact on the quality of wine.

Feature Importance Ranking – FIR

If we look at the most important chemical components to look out for when creating wine are alcohol, Sulphates, Volatile Acidity, and Total Sulfur Dioxide. The least important ones are Free Sulfur Dioxide, Residual Sugar, and Citric Acid.

Conclusion

Now we will sum up the findings from all our statistical analysis of the data.

We can say with great confidence that the chemical components that affect the quality of wine the most are Alcohol, Sulphates, and Volatile Acidity. Our Scatter plots, Correlation Matrix, Regression Model,