Wine Quality Classification using Random Forest

Data 230 - Data Visualization

Group 6: Abdul Sohail Ahmed, Nandini Sreekumaran Nair, Nghi Nguyen, Poojan Gagrani, Priya Varahan

Abstract

The wine sector is very competitive, and comprehending the aspects that play a role in determining the quality of wine is crucial for triumph. The main purpose of this project is to investigate a dataset on wine quality and identify the main factors that influence it. The dataset contains chemical measurements of red and white wines, and the corresponding ratings for the wine's quality. The project applies Random Forest algorithm to classify wine quality into three categories: high, average, and low. The accuracy score of the model is about 61%, which can be explained by the unbalanced target distribution and not strong correlations between independent and dependent variables. The data is also explored using descriptive statistics and data visualization techniques such as violin plots, swam plots, density plots, scatterplots, heat maps, and bar charts. The purpose of the dashboards is to evaluate the Random Forest's result and provide valuable insights about the factors that contribute to wine quality, such as wine categories, or the levels of density, alcohol, residual sugar, and so forth. Moreover, winemakers and marketers can make use of the visualizations to decide production, pricing, and marketing.

1. Introduction

The wine industry has shown exponential growth in recent years as social drinking has increased. Today, industry players use product quality certification to price, validate, and promote their products. This is a time-consuming and expensive process, requiring human expert evaluation. Similarly, the pricing of wine is also based on an abstract notion of wine evaluation by tasters, and tasters' opinions can vary greatly, thereby generating the need for a standard evaluation process. Physicochemical testing is another important component of wine certification and quality assessment. Factors such as acidity, pH, sugar content, and other chemical characteristics are considered during physicochemical testing. The wine market would be interesting if we could relate the quality of human taste to the chemical properties of wine. In this project, the aim is to determine which characteristics are the best indicators of wine quality and to gain an understanding of each of these factors in a model which will assist manufacturers in determining wine quality based on a standard evaluation process and in classifying wine prices accordingly.

1.1 Project Background

Humans have made and drank wine for thousands of years, and it has been a significant part of many cultures and societies. Wine is a complex sensory experience that features a variety of textures, aromas, and tastes in addition to being an alcoholic beverage. There are “wine enthusiasts” who respect wine and enjoy learning about its creation, history, and tasting notes. So, it becomes really imperative to maintain the wine's high quality. Firstly, to ensure distinctive flavors and qualities are maintained, and secondly, to meet customers' palates.

1.2 Problem Definition

Wine is considered a complex beverage with many chemical constituents that add to its aroma and flavor. Understanding wine's chemical constitution and how various constituents interact is essential for assessing its quality. But, the problem here is, it is difficult for even the wine taster with years of experience to determine the exact chemical composition of wine and since wine quality testing is most often based on sensory assessment, which can be subjective and greatly varies between tasters it is possible that different tasters provide different evaluations which could result in discrepancies.

1.3 Objective

The purpose of our project is to apply data visualization and machine learning methodologies to gain insights of the wine quality from its physicochemical characteristics. To be more specific, the machine learning model has the ability to classify wine into three categories: low, average, and high. We predict based on available attributes such as acidity, sugar, chloride, sulfur dioxide, density, pH, sulphates, and alcohol level. Apart from the model, data visualizations show clear relationships between wine quality and other physicochemical features. The project can help winemakers determine the quality of wine batches and decide on their price. The winemakers, therefore, don't need to taste all the wine barrels to determine their grades. Additionally, a reasonable pricing strategy increases not only the customers' satisfaction but also the sellers' revenue.

1.4 Literature Survey

The paper "A Review of Machine Learning Methods for Wine Quality Assessment" by N. Singh and M. Kaur (2018) offers a thorough analysis of the many machine learning approaches applied to wine quality assessment. In addition to outlining machine learning approaches including decision trees, support vector machines, artificial neural networks, and ensemble methods, the authors also explore the significance of wine quality rating.The study highlights the benefits and drawbacks of various strategies after reviewing multiple studies that have used these techniques to assess wine quality. The authors examine the usefulness of ensemble techniques in merging several models for enhanced accuracy as well as the significance of feature selection in enhancing the accuracy of classification and regression models.The article also offers an overview of the various datasets utilized for assessing the quality of wines as well as the performance measures employed to assess the model's accuracy.

The study "Data Mining and Wine Quality" by M. Cortez and A. Cerdeira (2009) explores the application of data mining methods for determining wine quality. The authors' main area of interest is the prediction of wine quality based on physicochemical characteristics.The importance of wine quality assessment and the difficulties posed by conventional sensory evaluation techniques are covered by the authors in the opening paragraphs. After that, they give an outline of the physicochemical characteristics of wine and how they could affect wine quality. The study analyzes a number of studies that evaluated wine quality using data mining methods and focuses on machine learning methods such as support vector machines, decision trees, and random forests. In order to assess the models correctness, the authors use performance metrics including mean squared error and mean absolute error on a sample of red and white Portuguese wines. The authors also go through the significance of feature selection in enhancing the models accuracy and illustrate the connections between various physicochemical characteristics and wine quality using scatter plots and parallel coordinate plots, among other data visualization approaches.

Wine quality evaluation using data mining techniques is thoroughly reviewed in the paper "Wine Quality Assessment Using Data Mining Techniques: A Review" by P. O. Afolabi and O. O. Oladele(2021). The authors talk about the value of evaluating wine quality and give an overview of the several data mining methods that are employed in this area, such as feature selection, clustering, classification, and regression.The authors stress the value of feature selection in raising the precision of regression and classification models.They go over clustering's function in spotting trends and connections in the data as well.The report also offers an overview of the many data mining software programs used for determining wine quality, including Weka, RapidMiner, and Orange. The authors come to the conclusion that data mining approaches are helpful for determining the major elements that influence wine quality.To investigate the possibilities of these strategies in enhancing wine production and quality control, they contend that more study is necessary.

The application of data visualization techniques for wine quality analysis is the main topic of the study "Data Visualization for Wine Quality Analysis" by L. Pagano, F. L. Ricci, and P. C. Masiero(2019). The authors talk about the value of wine quality analysis in the wine business as well as the drawbacks of using conventional data analysis techniques.The study gives an overview of different data visualization methods, including scatter plots, heatmaps, parallel coordinate plots, and boxplots, as well as their possible uses in analyzing wine quality. These methods are used by the writers to illustrate the connections between various physicochemical characteristics of wines and how they affect wine quality.In reviewing many studies that have applied data visualization methods to the examination of wine quality, the paper highlights the value of interactive visualization tools that enable users to explore the data and spot trends and linkages.The authors also talk about how data visualization approaches might help with quality control and wine production by letting winemakers see where they can improve and make wise decisions regarding blending and production.

The use of a Random Forest classifier to predict wine quality is suggested in the 2018 article "Application of Random Forest Classifier for Wine Quality Prediction" by Yingting Liu and Wenjing Fang. The chemical characteristics of red and white wines and the related quality ratings are included in the dataset used by the writers. Before training the classifier, the dataset is preprocessed and features are chosen. The research demonstrates that in terms of accuracy, precision, and recall, the Random Forest classifier performs better than other classification methods. Also, the authors examine the effectiveness of several feature selection techniques and demonstrate that the Recursive Feature Elimination (RFE) technique outperforms others.

2. CRISP DM Methodology

We have followed a hybrid model of the waterfall and CRISP-DM methodology to develop the proposed solution. This methodology enables us to plan everything as we can have a complete picture of the time duration of each task and how to split the work among each team member equally, as shown in Figure 1. The CRISP-DM methodology consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase is composed of several tasks that must be completed before moving on to the next phase. The methodology is iterative.

The hybrid model of the waterfall and CRISP-DM methodology combines the strengths of both approaches to create a structured and efficient approach to software development and data mining projects. The hybrid model starts with the linear, sequential approach of the waterfall model, which provides a clear and well-defined plan for the project.

Business Understanding

This is the initial phase of the project development, where the team will bring in their ideas and literature surveys so that the best problem statement can be derived from it. We brainstormed by reading various literature on the topics we have short-listed. Finally, to plan the solution for the problem statement, teammates researched and found various papers on Online learning vs. Offline learning and visualizations. Various types of visualizations can be performed to demonstrate the various data types. We also planned the steps for Data mining.

Data Understanding

In this phase, we collected the datasets. The team has decided to make the comparison with historical data. Therefore, we decided to collect the data collected from Kaggle and UCI Machine Learning Repository. After collecting the data, we performed each dataset's EDA and data quality checks separately for each dataset. We have used some visualization techniques, such as boxplots, pie charts, etc., to know the quality of the datasets.

Data Preparation

In Data Preparation, we have performed various steps of data pre-processing. We had to perform the data cleaning for the dataset as it contained duplicate values, missing values, outliers and various other anomalies and treat those.

Data Modeling

In Data modeling we used random forest to classify the dataset to different categories. Random forest is a machine learning algorithm and an ensemble of groups of decision trees .We did not specify any parameters for the random forest function as we want a basic model. The difference between random forest and decision trees is that random forests only select a subset of the independent variables while decision trees consider all the possible variables.

We've also carried out interactive visualizations that reveal useful details about the datasets we gathered. During this phase, other patterns are evaluated as well.

Data Evaluation

During this phase, accuracy score (performance metric) was evaluated and analyzed using Tableau and Python. We plotted the plots created using calculated fields in Tableau . The results were evaluated using this technique.

Deployment

Deployment is the last phase, in which we used Tableau and Python for the visualizations. All Python codes and Tableau visualizations were published on Github and Tableau Public for users' visibility. Github and Tableau Public are dynamic environments where anyone can access and visualize the works that are uploaded.

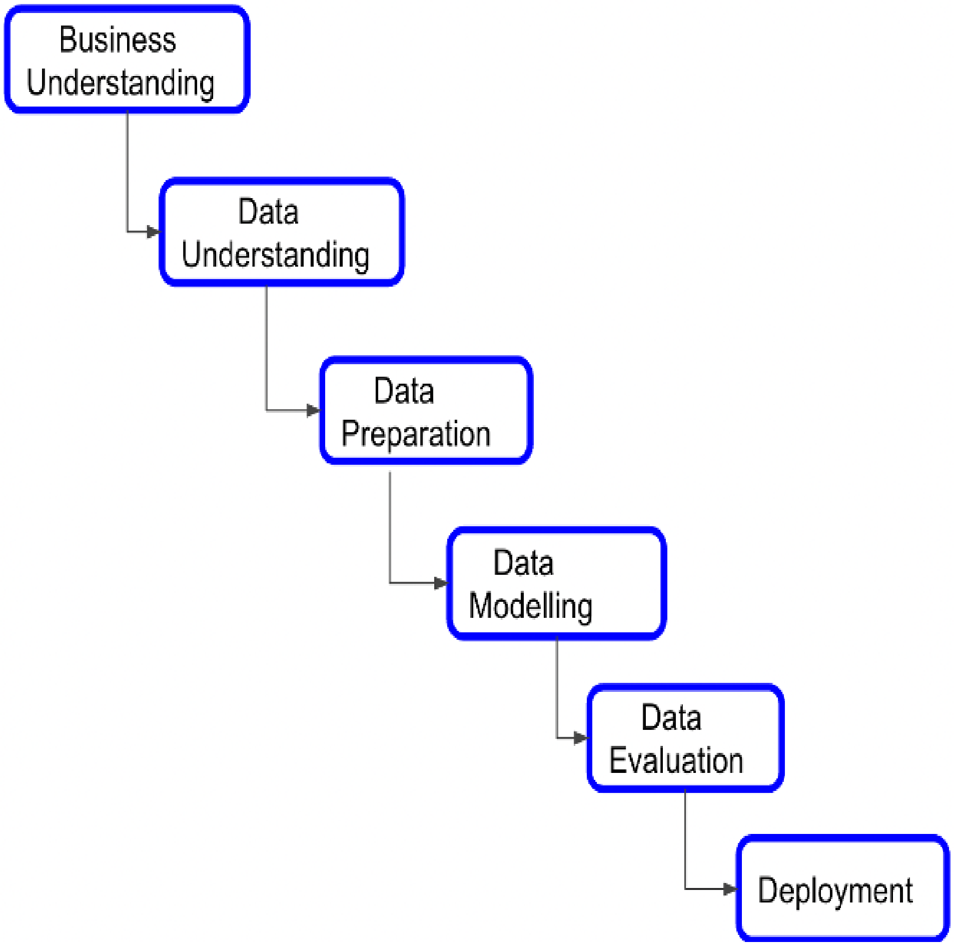


Figure 1:Hybrid Waterfall & CRISP-DM Methodology

3. Data Flow Diagram

The Diagram below shows the complete process of our project. First of all we collected the data and learned all the features' characteristics. Then, we performed exploratory data analysis using the raw dataset to see if there was any pattern or abnormal point. Next, we conducted data cleaning on duplicates, missing values, outliers, and other potential issues. We used the cleaned data for modeling without feature selection because Random Forest function includes feature selection itself. After modeling, we created many visualizations from cleaned data to show the relationships between the independent variables and the target, and the relationships between the independent variables themself. The visualizations were made by either Python or Tableau. For deployment, we published our Python code on Github and Tableau dashboards on Tableau Public.

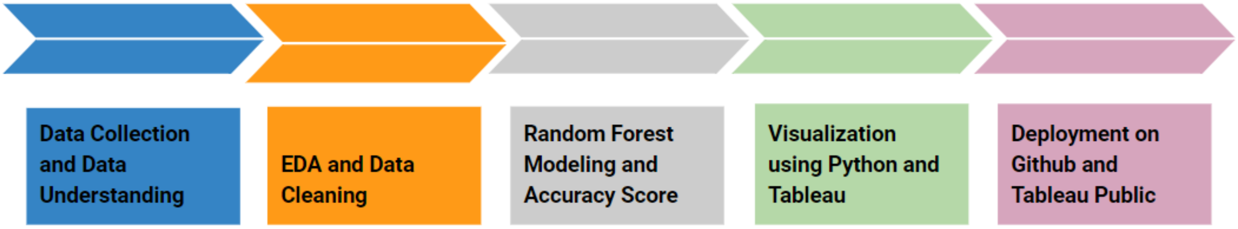

Data Collection and Data Understanding
EDA and Data Cleaning
Random Forest Modeling and Accuracy Score
Visualization using Python and Tableau
Deployment on Github and Tableau Public

Figure 2: Data Flow Diagram

4. Data Collection

One of the most important stages in the project is to gather meaningful data in order to generate relevant insights. For the purpose of this project we have selected the Wine Dataset from the UCI Machine Learning Repository. Our dataset consists of the two datasets related to red and white variants of the Portuguese "Vinho Verde" wine. These Dataset will be used to generate analysis and for Modeling. We merge the two datasets and have a column “category” to indicate if a row illustrates red or white wine. There are a total of 13 attributes in the combined dataset, which are category, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality (score between 0 and 10)

Dataset Name: Wine Quality

Dataset Source: UCI Machine Learning Repository

Source URL: https://archive.ics.uci.edu/ml/datasets/wine+quality

5. Exploratory Data Analysis (EDA)

While performing data cleaning on the dataset we simultaneously performed exploratory data analysis to get valuable insights about the raw data. Firstly, we determined the correlation between existing columns of the dataset by plotting a correlogram. Figure 3 illustrates the correlation matrix of the dataset. It is quite evident from the matrix that there is a high correlation between total and free sulfer dioxide that is 0.72. Followed by high correlation between free sulfur dioxide and residual sugar & total sulfur dioxide and residual sugar.

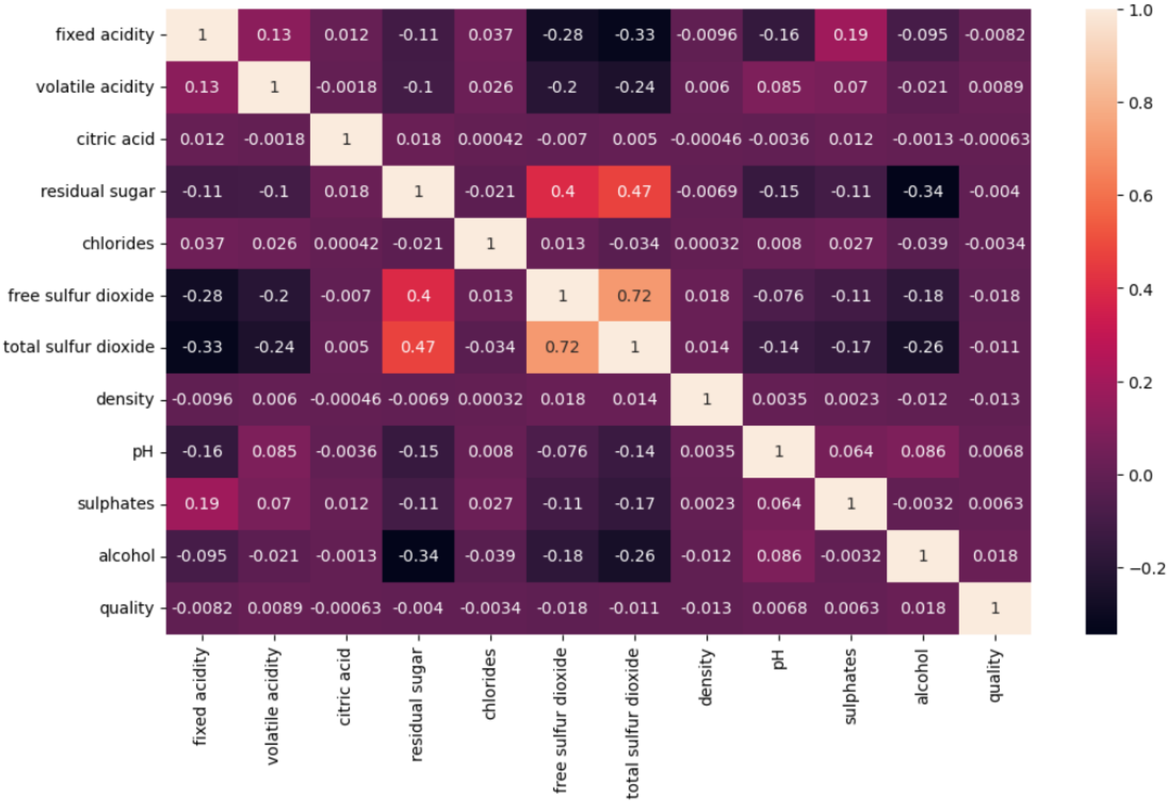


Figure 3: Correlogram or Correlation matrix of Raw Data

This information is useful for winemakers because in order to obtain desired results in terms of flavor, fragrance, and other wine qualities, they can modify their manufacturing techniques accordingly. We also created a box and whisker plot on all the numeric columns in order to determine the outliers present in them. Figure 4 illustrates all the box plots obtained for each column.

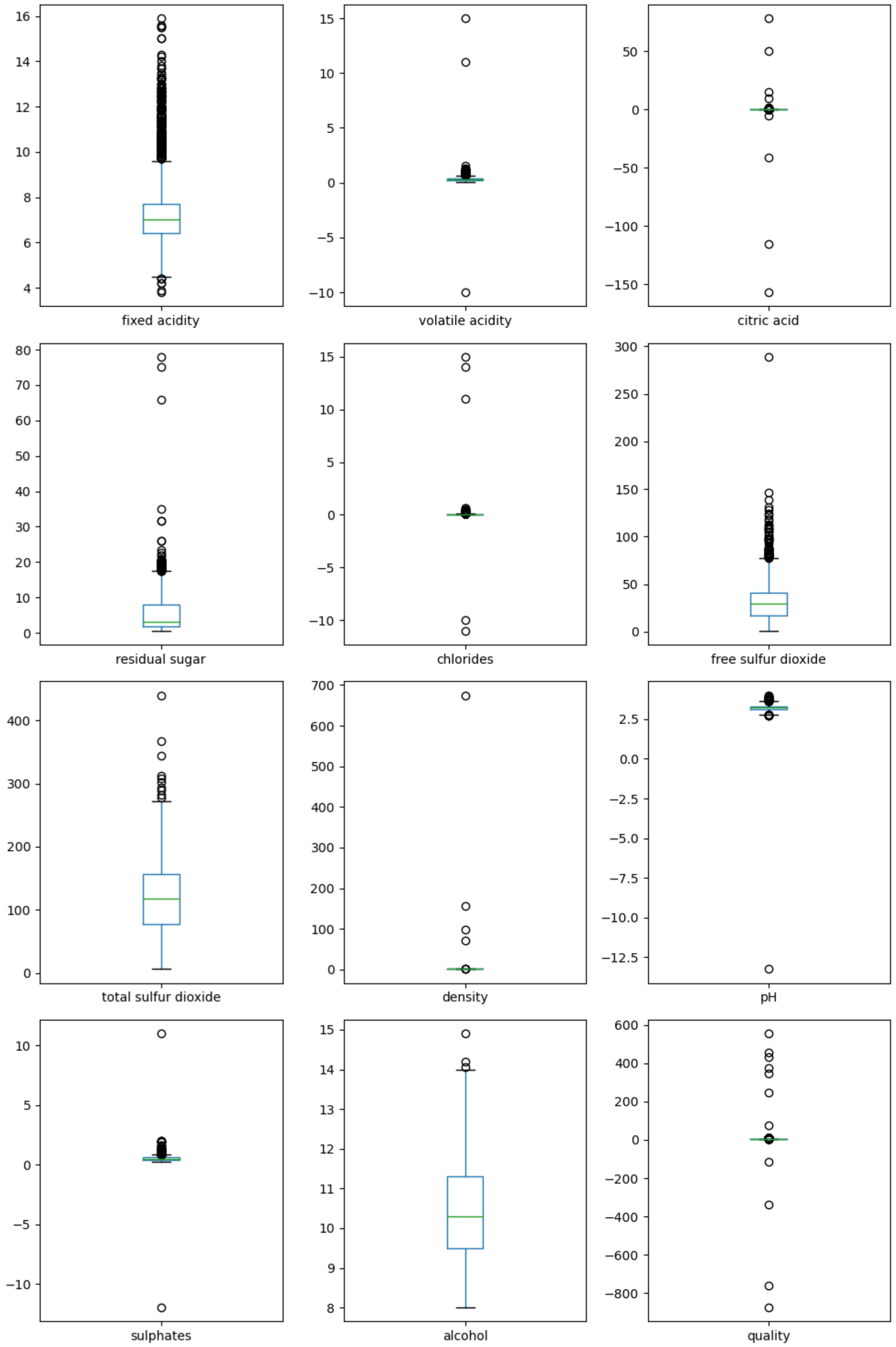


Figure 4: Box Plot of Wine Characteristics

It is quite evident from the boxplots that there are a lot of data points that are acting as outliers for various columns. Here, a data point is considered an outlier if it falls outside the 1.5 times the Interquartile Range (IQR) above the third quartile or 1.5 times the IQR below the first quartile. Only the column ‘Alcohol’ has an insignificant amount of outliers, the rest have a lot of outliers.

It is an important piece of information because this might imply that the wine's alcohol content is a more consistent variable than the other elements that influence the wine's quality. However, there are many outliers in the other columns, showing that these variables are more variable and may be influenced by outside variables like weather, grape quality, or production procedures

Wine producers should concentrate on enhancing the quality and consistency of the variables that have a high number of outliers by doing this. This could result in higher-quality wines and more consumer satisfaction. Additionally, it can aid in determining the elements that have the greatest influence on wine quality, enabling producers to take the necessary steps to preserve or raise the wine's quality.

In order to get a better understanding about the skewness in data we have created a distplot or density plot. This plot will help us to better understand the distribution of data for each variable and spot patterns in the data distribution. When examining the link between the independent numerical variables and the dependent variable, this was helpful as a starting point. Figure 5 illustrates the density plots created using the data present in each variable.

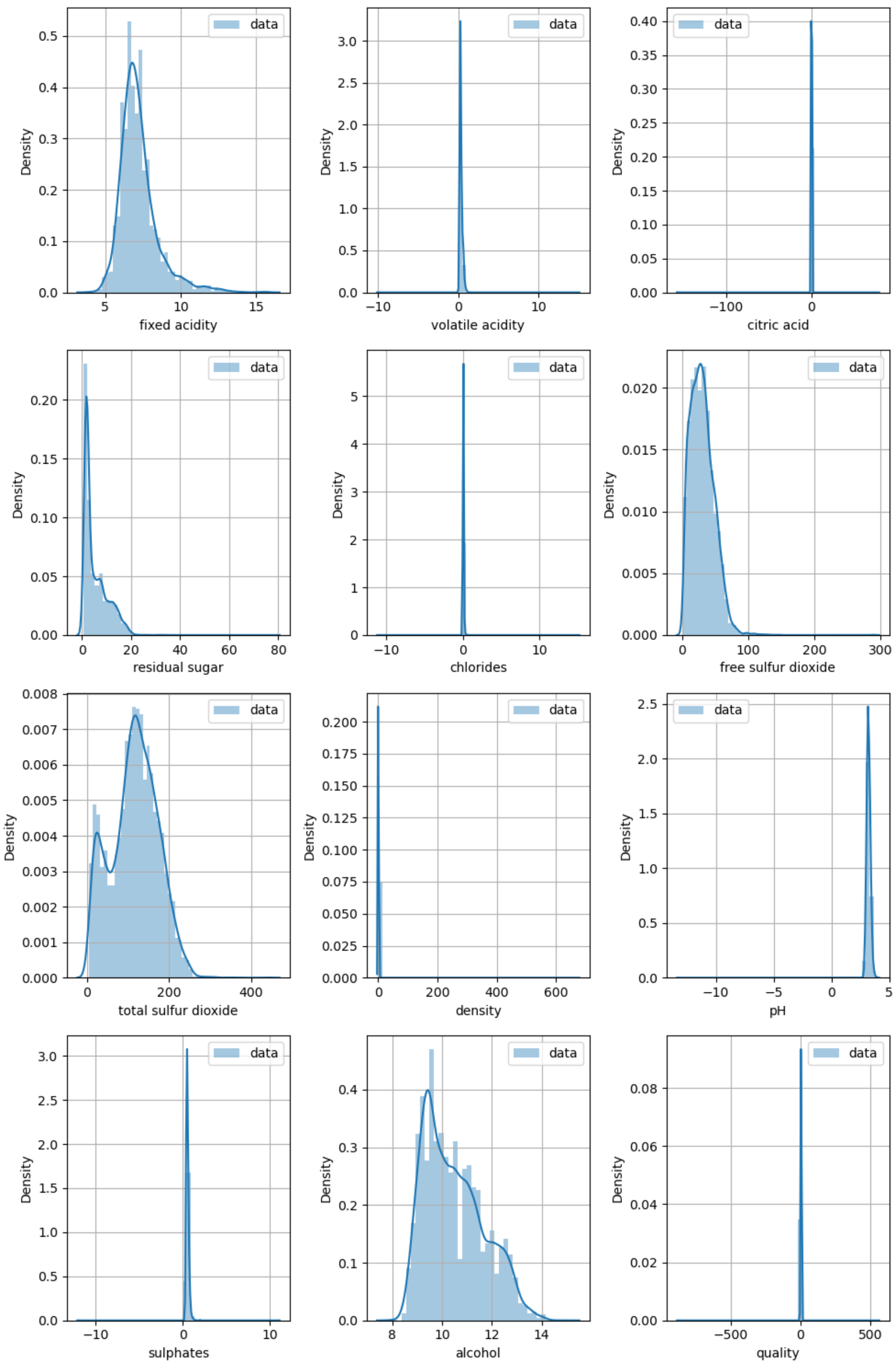


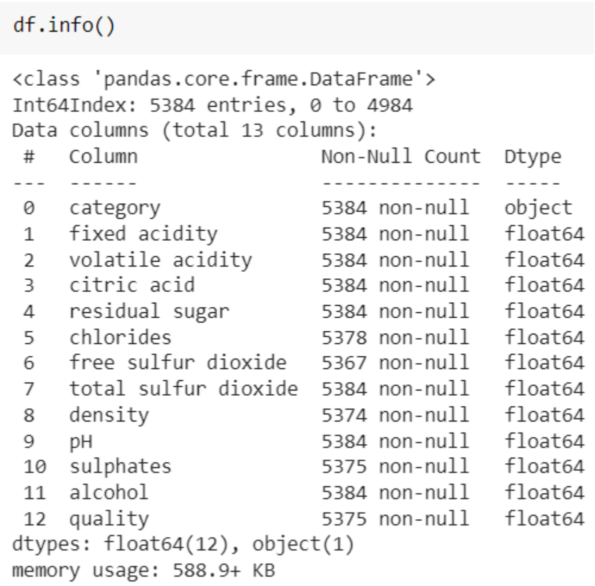
Figure 5: Density Plot of Wine Characteristics

We can observe that most of the alcohol content typically lies between 8 to 14% and pH levels lie between 2.5 to 4. Also the residual sugar content is between 0 to 20 and chloride i.e., salt content is prevalent at 0.1. It is also quite evident that variables “fixed acidity” and “pH” are approximately normally distributed. The other variables like “residual sugar”, “volatile acidity”, “free sulfur dioxide” and “sulphates” seems to be left or negatively skewed. The "total sulfur dioxide" variable shows a bimodal distribution, indicating the presence of two subpopulations with distinct sulfur dioxide levels.

This plot provides the useful information that since most values lie within a particular range, alcohol content plays a significant role in determining the quality of wine. Similarly, pH values are crucial since they impact how acidic a wine will be. The fact that there are two subpopulations in the "total sulfur dioxide" variable suggests that the amounts of sulfur dioxide in the wine may be influenced by various procedures or circumstances. Some variables' left- or negatively-skewed distributions suggest that they might have many low values and few high ones, which could affect how much of a contribution they make to the wine's overall quality.

6. Data cleaning

The first step in data cleaning is checking for duplicates, and it turns out that the dataset has 1264 duplicate values. After dropping the duplicates, there are 5384 rows left. Figure shows the dataset information at this point.

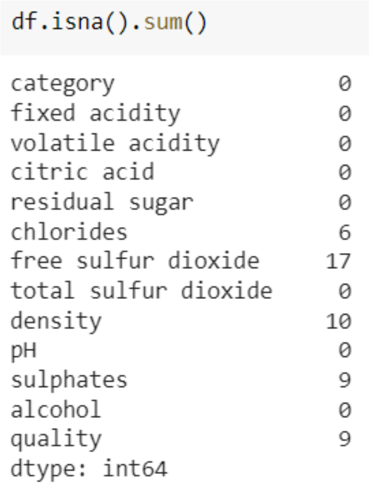


```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5384 entries, 0 to 4984
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   category              5384 non-null   object 
 1   fixed acidity         5384 non-null   float64
 2   volatile acidity      5384 non-null   float64
 3   citric acid           5384 non-null   float64
 4   residual sugar        5384 non-null   float64
 5   chlorides             5378 non-null   float64
 6   free sulfur dioxide   5367 non-null   float64
 7   total sulfur dioxide  5384 non-null   float64
 8   density               5374 non-null   float64
 9   pH                    5384 non-null   float64
 10  sulphates             5375 non-null   float64
 11  alcohol               5384 non-null   float64
 12  quality               5375 non-null   float64
dtypes: float64(12), object(1)
memory usage: 588.9+ KB
```

Figure 6: Dataset information after dropping duplicates

By looking at figure 6, we can see that there are some missing values in a couple of columns. Figure 7 below provides more details about the missing values.



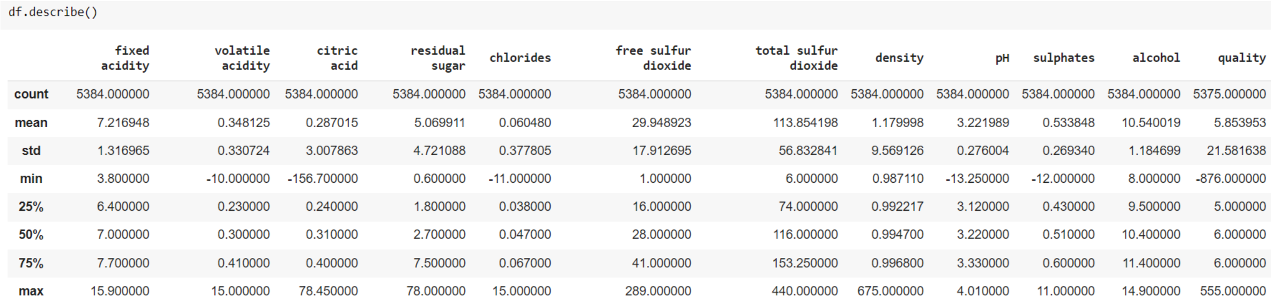
```
df.isna().sum()
```

```
category                 0
fixed acidity            0
volatile acidity         0
citric acid              0
residual sugar           0
chlorides                6
free sulfur dioxide     17
total sulfur dioxide     0
density                 10
pH                       0
sulphates                9
alcohol                  0
quality                  9
dtype: int64
```

Figure 7: Missing values

We handle independent variables first and leave quality to handle in the outlier section. For chlorides, free sulfur dioxide, and density, we fill in with their median value as all these features are continuous. And for sulphates, we fill missing spots with its mode value because this feature is a discrete value.

Next, we work on outliers. Figure 8 provides general statistics of all variables of the dataset. “Category” is a categorical variable, so it is not listed in the figure. Observe that the features having abnormal statistics are volatile acidity (negative minimum and significantly large maximum), citric acid (negative minimum and significantly large maximum), residual sugar (significantly large maximum), chlorides (negative minimum and significantly large maximum), free sulfur dioxide (significantly large maximum), density (significantly large maximum), pH (negative minimum), sulphates (negative minimum), and quality (negative minimum and significantly large maximum).



```
df.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5375.000000
mean	7.216948	0.348125	0.287015	5.069911	0.060480	29.948923	113.854198	1.179998	3.221989	0.533848	10.540019	5.853953
std	1.316965	0.330724	3.007863	4.721088	0.377805	17.912695	56.832841	9.569126	0.276004	0.269340	1.184699	21.581638
min	3.800000	-10.000000	-156.700000	0.600000	-11.000000	1.000000	6.000000	0.987110	-13.250000	-12.000000	8.000000	-876.000000
25%	6.400000	0.230000	0.240000	1.800000	0.038000	16.000000	74.000000	0.992217	3.120000	0.430000	9.500000	5.000000
50%	7.000000	0.300000	0.310000	2.700000	0.047000	28.000000	116.000000	0.994700	3.220000	0.510000	10.400000	6.000000
75%	7.700000	0.410000	0.400000	7.500000	0.067000	41.000000	153.250000	0.996800	3.330000	0.600000	11.400000	6.000000
max	15.900000	15.000000	78.450000	78.000000	15.000000	289.000000	440.000000	675.000000	4.010000	11.000000	14.900000	555.000000

Figure 8: Dataset statistics after handling duplicates and missing values of chlorides, free sulfur dioxide, density, and sulphates.

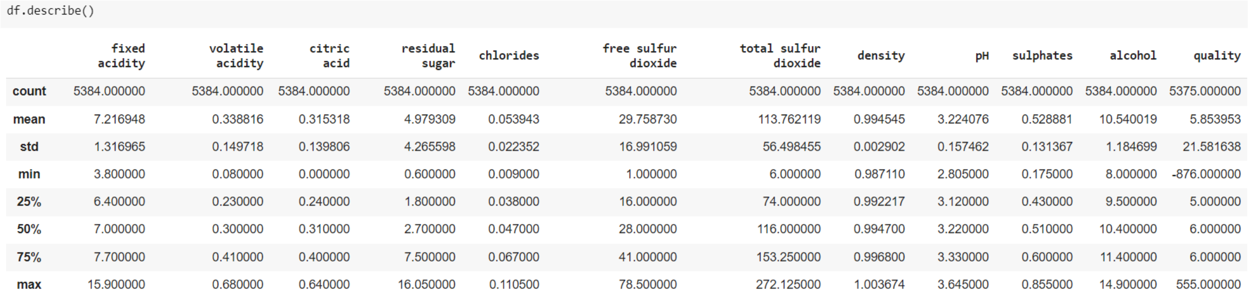
We process the outliers of independent variables as follows:

Let Q1 is the 25th percentile value, Q3 is the 75th percentile, and IQR = Q3-Q1.

Upper = Q3+1.5*IQR

Lower = Q1-1.5*IQR

For any variable, if any value is less than Lower, then replace it by Lower, and if any value is greater than Upper, replace it by Upper. Figure 9 illustrates the data statistics after handling outliers of the independent features.



```
df.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5384.000000	5375.000000
mean	7.216948	0.338816	0.315318	4.979309	0.053943	29.758730	113.762119	0.994545	3.224076	0.528881	10.540019	5.853953
std	1.316965	0.149718	0.139806	4.265598	0.022352	16.991059	56.498455	0.002902	0.157462	0.131367	1.184699	21.581638
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.805000	0.175000	8.000000	-876.000000
25%	6.400000	0.230000	0.240000	1.800000	0.038000	16.000000	74.000000	0.992217	3.120000	0.430000	9.500000	5.000000
50%	7.000000	0.300000	0.310000	2.700000	0.047000	28.000000	116.000000	0.994700	3.220000	0.510000	10.400000	6.000000
75%	7.700000	0.410000	0.400000	7.500000	0.067000	41.000000	153.250000	0.996800	3.330000	0.600000	11.400000	6.000000
max	15.900000	0.680000	0.640000	16.050000	0.110500	78.500000	272.125000	1.003674	3.645000	0.855000	14.900000	555.000000

Figure 9: Dataset statistics after handling outliers of independent variables.

The next step is to deal with missing values and outliers of the target feature "quality." The figure below shows the distribution of the target variable. It can be seen that the data is not very balanced.

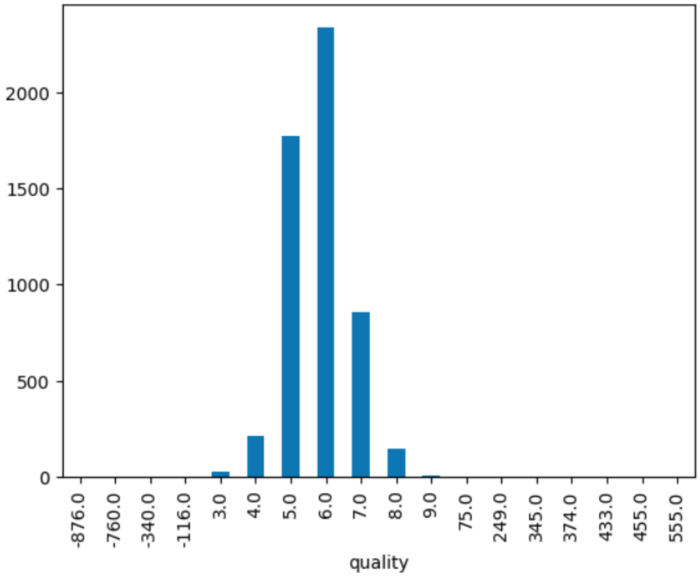


Figure 10: Target Feature Distribution.

We convert any value that is less than 0 to 0 and all values greater than 10 to 10 because it is the official range of the attribute. After that, we re-group the classes as follows: from 0 to 5 is poor quality (class 0), 6 is average quality (class 1), and from 7 to 10 is high quality (class 2). Figure 0 represents the distribution of the new target groups.

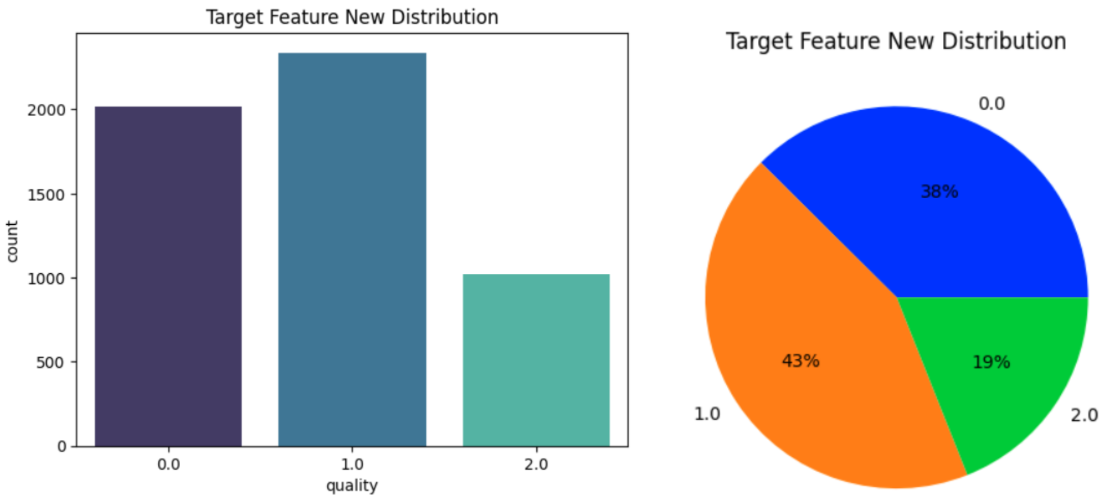


Figure 11: Target Feature New Distribution.

The “Category” column includes “Red” and “White”, which are respectively converted to 0 and 1 for the modeling purpose.

7. Data Modeling

After data cleaning, the dataset is splitted into training and testing datasets with the portion of 80-20. We implement random forest to classify the dataset to three different categories: 0 means low quality, 1 means average quality, and 2 means high quality. We don’t specify any parameters for the random forest function as we want a basic model. Random forest is a machine learning algorithm and an ensemble of groups of decision trees. The difference between random forest and decision trees is that random forests only select a subset of the independent variables while decision trees consider all the possible variables.

We don’t apply feature selection before model implementation because it is not necessary for random forest whose algorithm already includes feature selection. Figure 12 shows the target prediction distribution of the model. The portion looks quite reasonable as the amount of category 0 is a bit lower than category 1 and significantly higher than category 2.

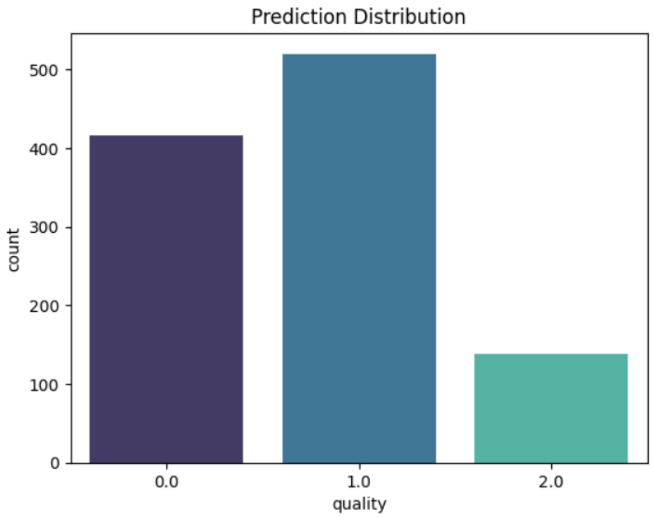


Figure 12: Target Prediction Distribution.

Figure 13 shows the confusion matrix of the basic random forest model when checking with the testing dataset. If we look at the diagonal from left to right, we can see the accuracy score for each group. For example, 66% of class 0 (low quality) is predicted correctly while that score is 64% for class 1 (average quality) and 46% for class 2 (high quality). On average, the accuracy score is about 61%, which is not high. The reason can be the unbalanced distribution of the target feature and it's possible that we need more important features in the dataset.

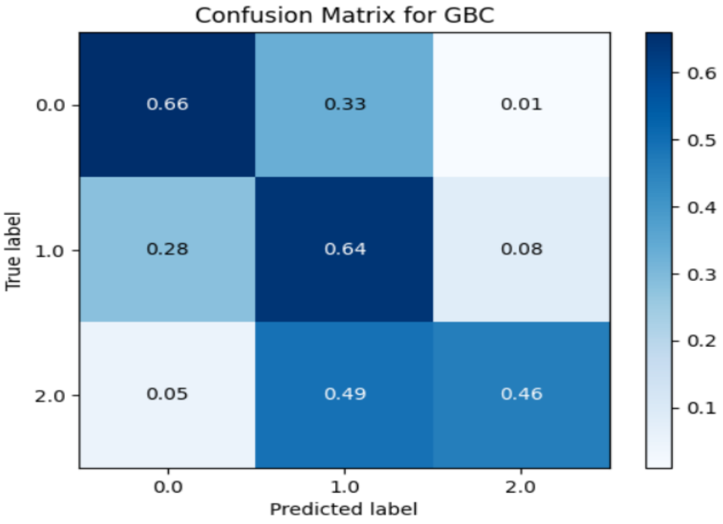


Figure 13: Confusion Matrix.

8. Visualizations

Figure 14 is the correlogram of the cleaned data. Observe that Alcohol and Density is highly correlated with each other with the index of (-0.7). However, the connection is not too strong to eliminate one of them. Alcohol has the highest correlation with our target feature, about 0.47, and density has a negative correlation of (-0.34) with quality.

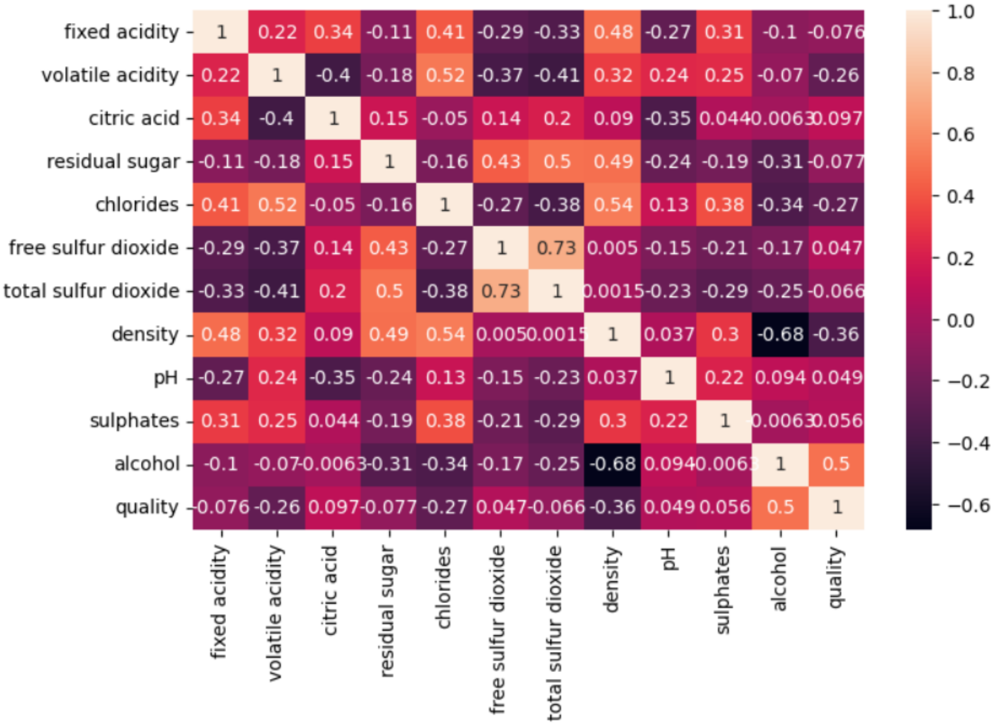


Figure 14: Correlogram or Correlation matrix of Cleaned Data

The countplot in figure 15 helps us understand the difference between count with respect to quality between red and white wine. It’s obvious that the count of the quality of white wine is more than twice that of red wine.

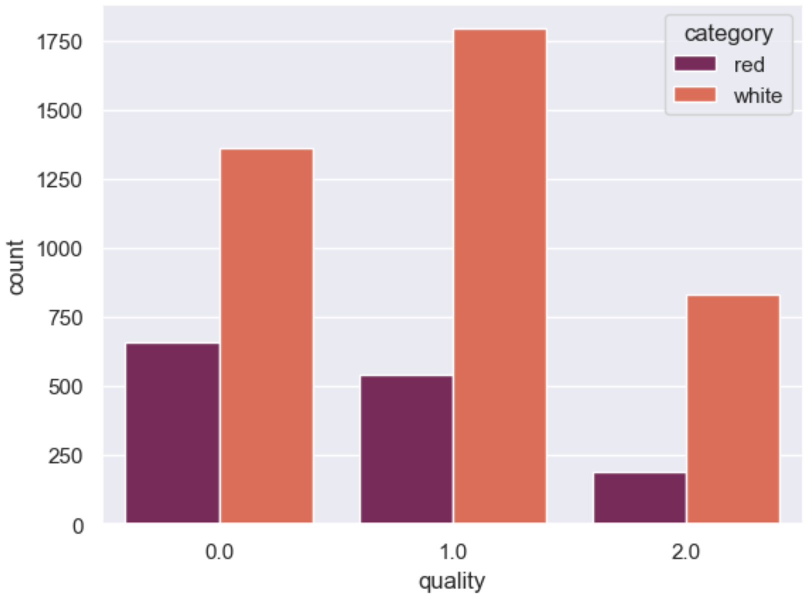


Figure 15: Countplot based on wine quality and type

The grouped bar chart in figure 16 helps us to visualize mean values of different chemical properties of wine grouped by their category (red/white). Here, it is quite evident that white wine has higher composition of total sulphur dioxide, free sulphur dioxide and residual sugar than red wine whereas red wine has higher composition of fixed acidity than white wine.

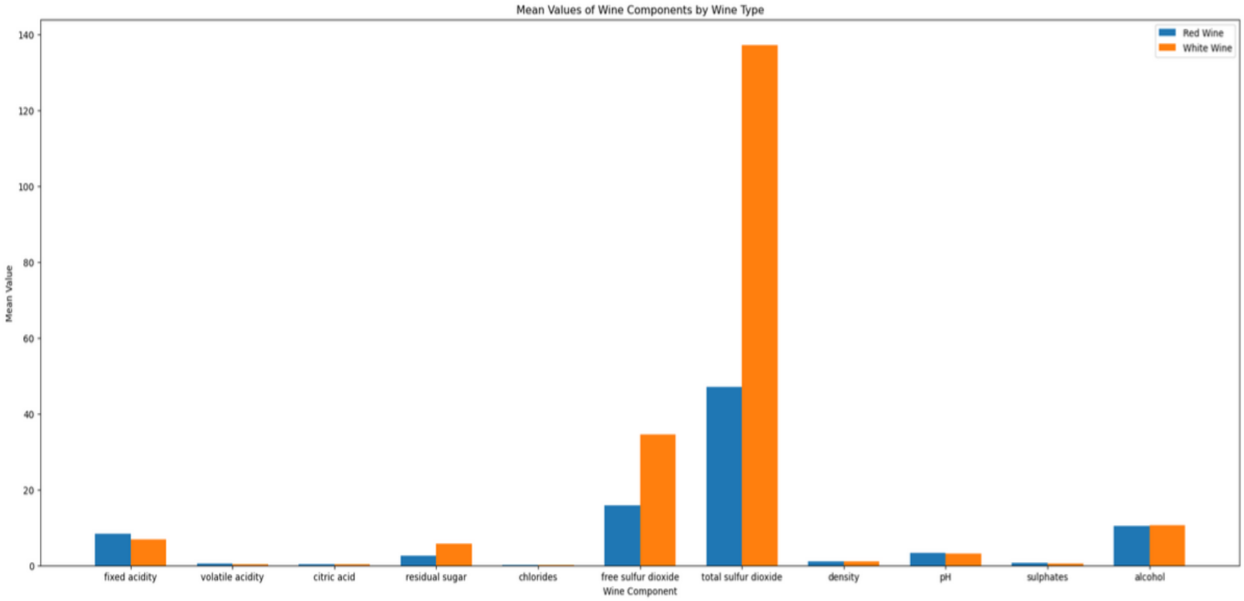


Figure 16: Mean Value of Wine Components by wine type

Figure 17 shows a violin plot that represents the alcohol content in different qualities of wines (red and white). It is easy to infer from this graph that alcohol content in the wine increases with increase in quality of wine. That is, the wine which is graded as ‘high’ quality is having higher alcohol content than the one which is graded as ‘average’ quality wine.

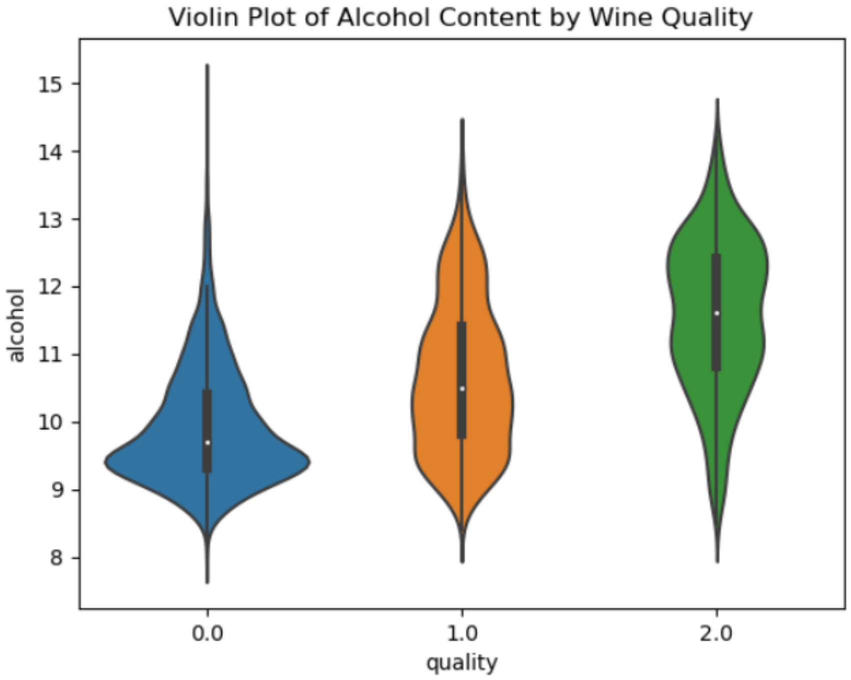


Figure 17: Violin Plot of Alcohol Content by Wine Quality

The facetgrid in figure 18 helps us visualize the distribution of quality of wine based on their category i.e, Red and White Wine. It’s evident that the quality of red wine is poor when compared to white wine as more of the quality for red wine is prevalent at 0. Whereas, white wine has prevalent average quality. Overall white wine is superior to red wine in terms of quality.

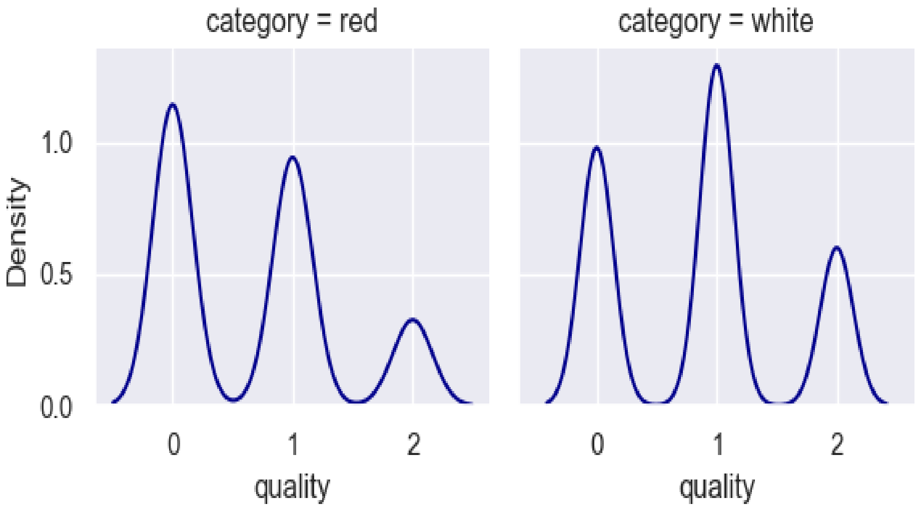


Figure 18: Wine quality based on type of wine

Figure 19 shows a scatter plot that represents how the wine density varies with change in alcohol content. We can clearly see that with the increase in alcohol content the wine density decreases. This shows the high negative correlation between alcohol and density.

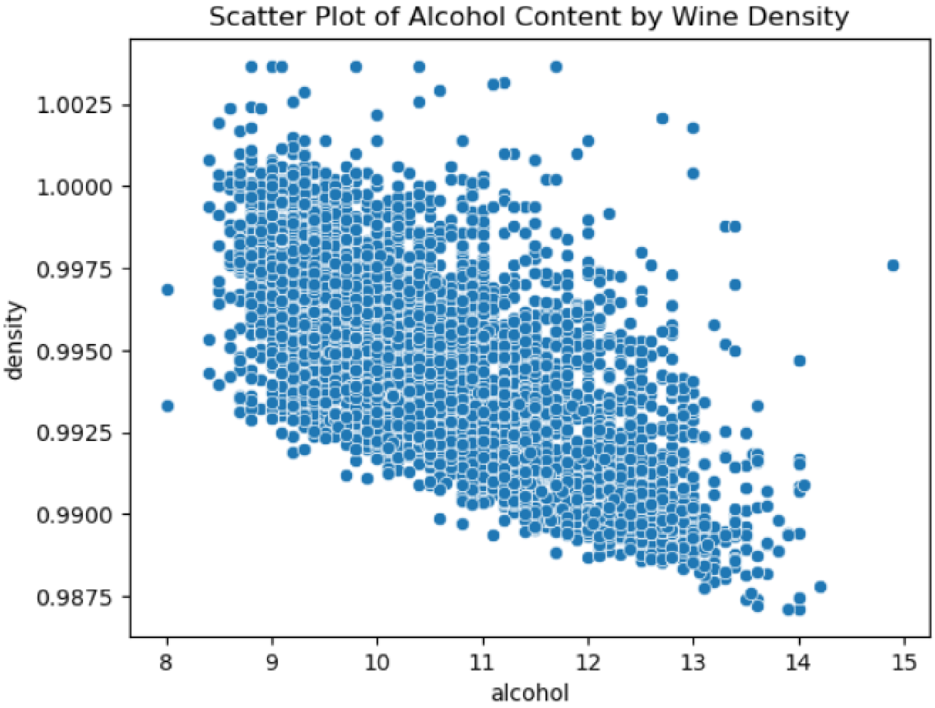


Figure 19: Scatter plot of Alcohol Content by Wine Density

The Swarm plot in figure 20 represents how volatile acidity affects the quality of wine. It is quite interesting to see that with the increase in the volatile acidity of wine the quality decreases. That is higher quality wine has less volatile acidity. This shows the negative correlation between volatile acidity and wine quality.

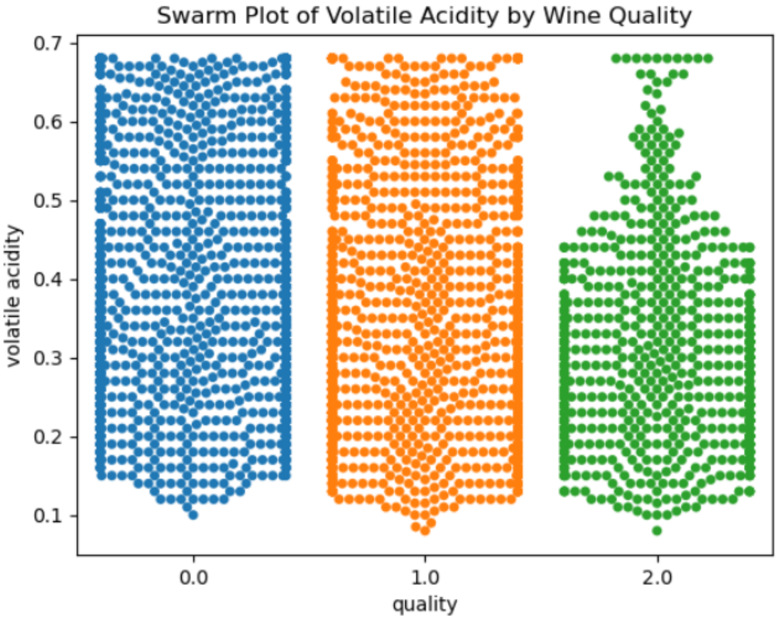


Figure 20: Swarm plot of Volatile Acidity by Wine Quality

The Violin plot in the figure 21 represents how the Sulphates affects the quality of wine. It’s quite evident that presence of Sulphates is directly proportional to the quality, meaning the more Sulphates are present the better the quality.

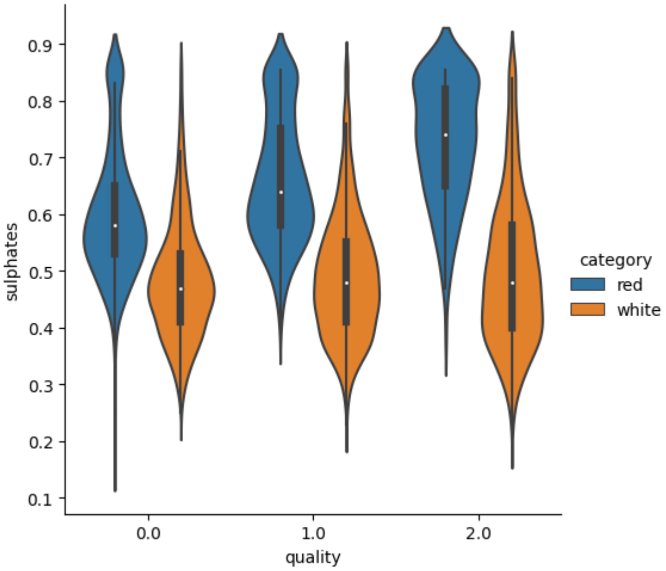


Figure 21: Violin plot for Sulphates vs. Quality

The Violin plot in the figure 22 represents how the volatile acidity affects the quality of wine. It’s important to note that the Volatile acidity is inversely proportional to the quality which is clearly visible in the plot, meaning that with the decrease in volatile acidity the quality of wine increases and vice versa.

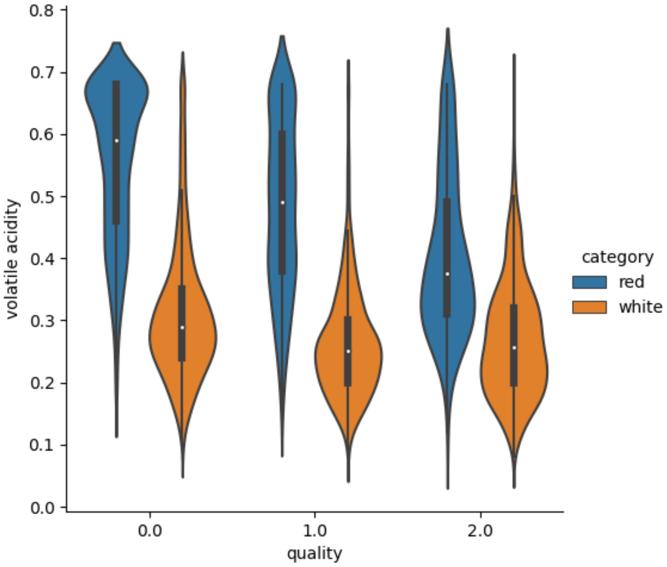


Figure 22: Violin plot for Volatile acidity vs. Quality

The box plot in the figure 23 represents how the residual sugar affects the quality of wine. It's obvious that the sugar is directly proportional to the quality, As also we know from our prior knowledge that the more the sugar content the better the quality.

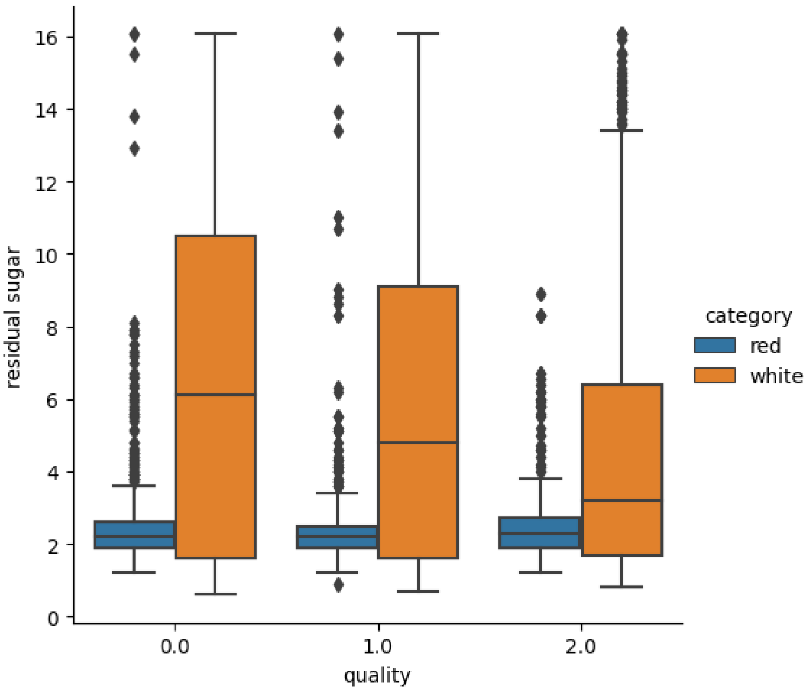


Figure 23: Box plot for Residual sugar vs. Quality

The Violin plot in the figure 24 represents how the chloride affects the quality of wine. It shares the similarities with the volatile acidity meaning is inversely proportional to the quality which is clearly visible in the plot, meaning that with the decrease in chlorides the quality of wine increases and vice versa.

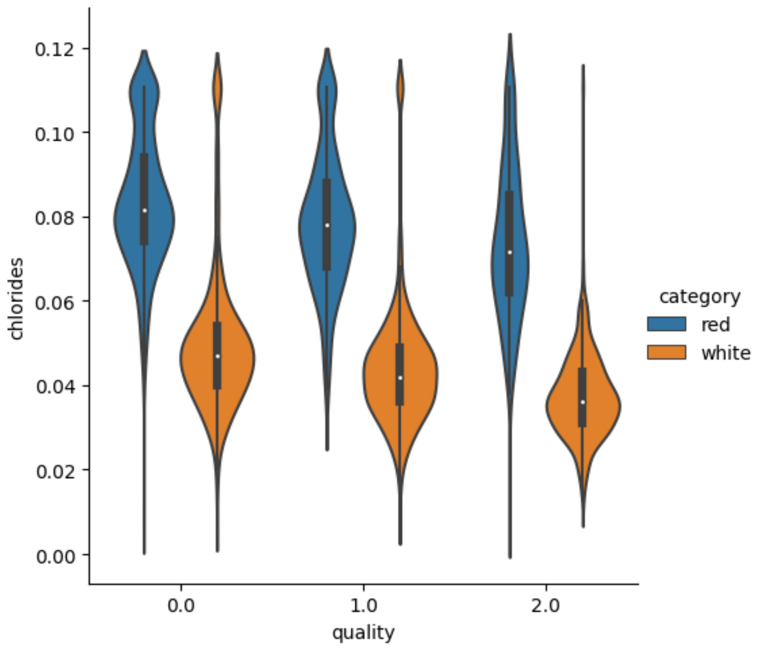


Figure 24: Violin plot for Chlorides vs. Quality

The Violin plot in the figure 25 represents how the free sulphur dioxide affects the quality of wine. It has positive correlation meaning that with the increase in free sulfur dioxide the quality increases which we can see here especially in the case of white wine.

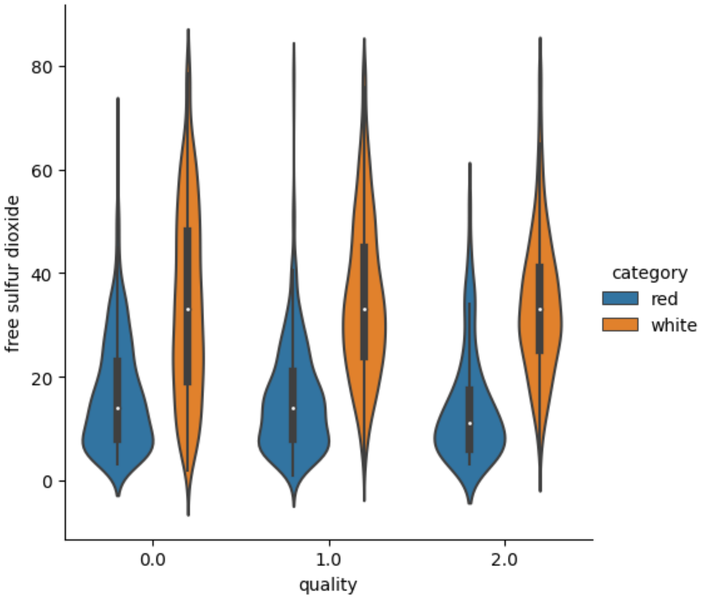


Figure 25: Violin Plot for Free Sulfur dioxide vs. Quality

The Violin plot in figure 26 represents how the total sulfur dioxide affects the quality of wine. Unlike free sulfur dioxide the total sulfur dioxide does the opposite, meaning it is inversely proportional to the quality which is clearly visible in the plot, meaning that with the decrease in total sulfur dioxide the quality of wine increases and vice versa. However, It's quite interesting to note that white wine has more sulfur dioxide than red wine.

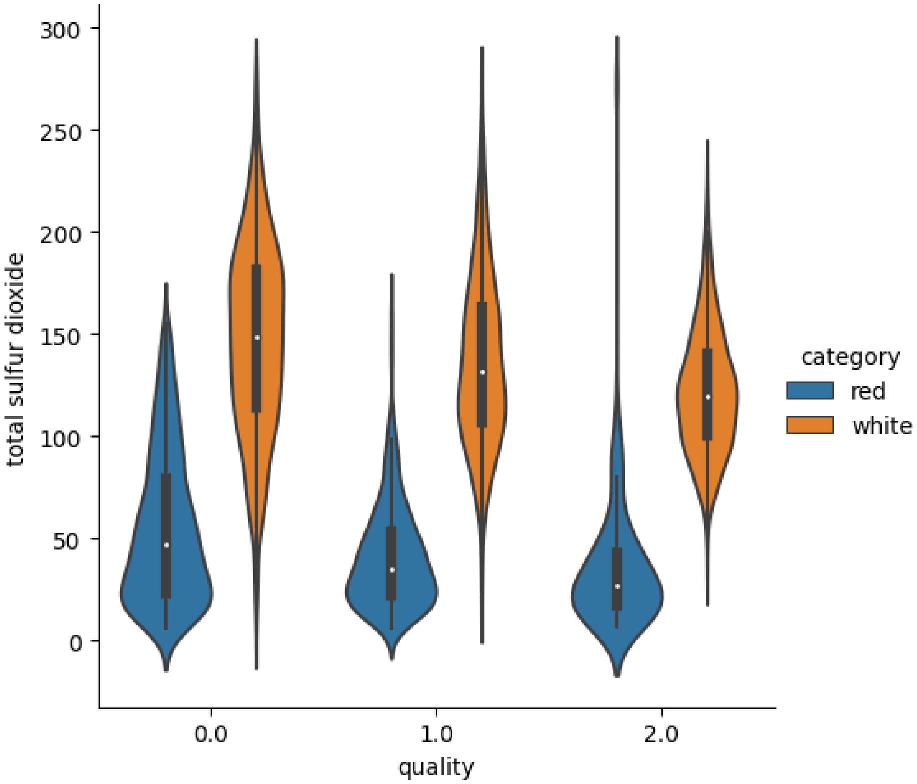


Figure 26: Violin Plot for Total Sulfur dioxide vs. Quality

9. Discussion

Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol are common chemical parameters used to evaluate wine quality. We examined the data and discovered that several of these variables have a considerable impact on wine quality.

Alcohol concentration is a major component that influences wine quality. A higher alcohol percentage usually suggests a higher grade of wine. We discovered that high-quality wines contain more alcohol than low-quality wines in the wine quality dataset. This suggests that alcohol concentration is a key component in determining wine quality.

The volatile acidity of volatile acids such as acetic acid is measured. A high quantity of volatile acidity in wine denotes a vinegar taste. In the dataset it was observed that the quality of wine decreases with the increase in volatile acidity. But this is not always true, it depends on the type of wine as well, some wine varieties, such as Pinot Noir and Chardonnay, have naturally high amounts of volatile acidity, which gives them a particular flavor and aroma. As a result, it is not always true that low volatile acidity is good for wine quality.

Sulphates are a type of preservative commonly used in winemaking to prevent oxidation and bacterial spoilage. In general the presence of sulphates in wine can affect the taste, aroma, and quality of the wine i.e. the more the sulphates, the better the quality. But there was a contradiction observed where the excessive use of sulphates can also give wine a bitter or chemical taste.

However, the optimal level of sulphates in wine depends on several factors such as the grape variety, the winemaking process, and the intended style of wine. Winemakers need to carefully balance the use of sulphates with other factors to achieve the desired quality and style of wine.

10. Conclusion

Through the collection of data from various sources we were able to consolidate various types of trends and this helped us gain useful insights. It was interesting to observe that the presence of Sulphates can enhance the aroma ,flavor which improves the quality of wine.

The level of volatile acidity in wine is an important factor in determining its quality, as excessive levels can result in unpleasant flavors and aromas. There are other components such as chlorides,Sulfur dioxide ,sugar etc which play a significant role in achieving the desired quality and style of wine .

Also the Residual sugar is a crucial component the more the sugar content the stronger and better the quality of wine.

By understanding the factors that contribute to wine quality, winemakers can make informed decisions about how to optimize their processes and improve the quality of their wines.

11. Future Scope

As a future scope we would like to conduct the survey to collect data on consumer preferences, wine tasting experiences, and other factors that contribute to wine quality. The advancement of sensor technology allows winemakers to monitor and modify their winemaking operations in real-time as critical chemical and sensory characteristics of wine are measured. One of the important future scope is investigating how climate change affects wine quality, and can aid in the creation of environmentally friendly winemaking techniques. In addition to evaluating the wine itself, a wine quality analysis is also dependent on factors such as the vineyard location, grape variety, climate, soil type, and winemaking processes used to produce the wine, hence All this data should additionally be considered.

12. Deployment

Youtube video for project demo: https://www.youtube.com/watch?v=h-nAQ38tBW0

Our Python code is deployed on Github and Tableau Dashboards to Tableau Public. The links are mentioned below:

1. Github link: https://github.com/poojan243/winequalityanalysis
2. Tableau public link:

➔ Basic Information Dashboard:

https://public.tableau.com/app/profile/abdul.sohail.ahmed/viz/BasicInformation_16840419846110/BasicInformation

➔ Variation in Quality of Wine (High Correlated):

https://public.tableau.com/app/profile/abdul.sohail.ahmed/viz/VariationinQualityofWineHighCorrelated/VariationinQualityofWineHighCorrelated?publish=yes

13. References

[1] Bhardwaj, P., Tiwari, P., Olejar, K., Parr, W., & Kulasiri, D. (2022). A machine learning application in wine quality prediction. *Machine Learning with Applications*, *8*, 100261. https://doi.org/10.1016/j.mlwa.2022.100261

[2] Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Using data mining for Wine Quality Assessment. *Discovery Science*, 66–79. https://doi.org/10.1007/978-3-642-04747-3_8

[3] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

[4] Hwang, J., & Yoon, Y. (2021a). *Data Analytics and Visualization in Quality Analysis Using Tableau*. https://doi.org/10.1201/9781003157694

[5] Wine Quality prediction using random forest classifier. (2021). *Strad Research*, *8*(6). https://doi.org/10.37896/sr8.6/013