IntelliML Report

Sample Datset

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7.8	0.88	0.0	2.6	0.098	25.0	67.0	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.998	3.16	0.58	9.8	6
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

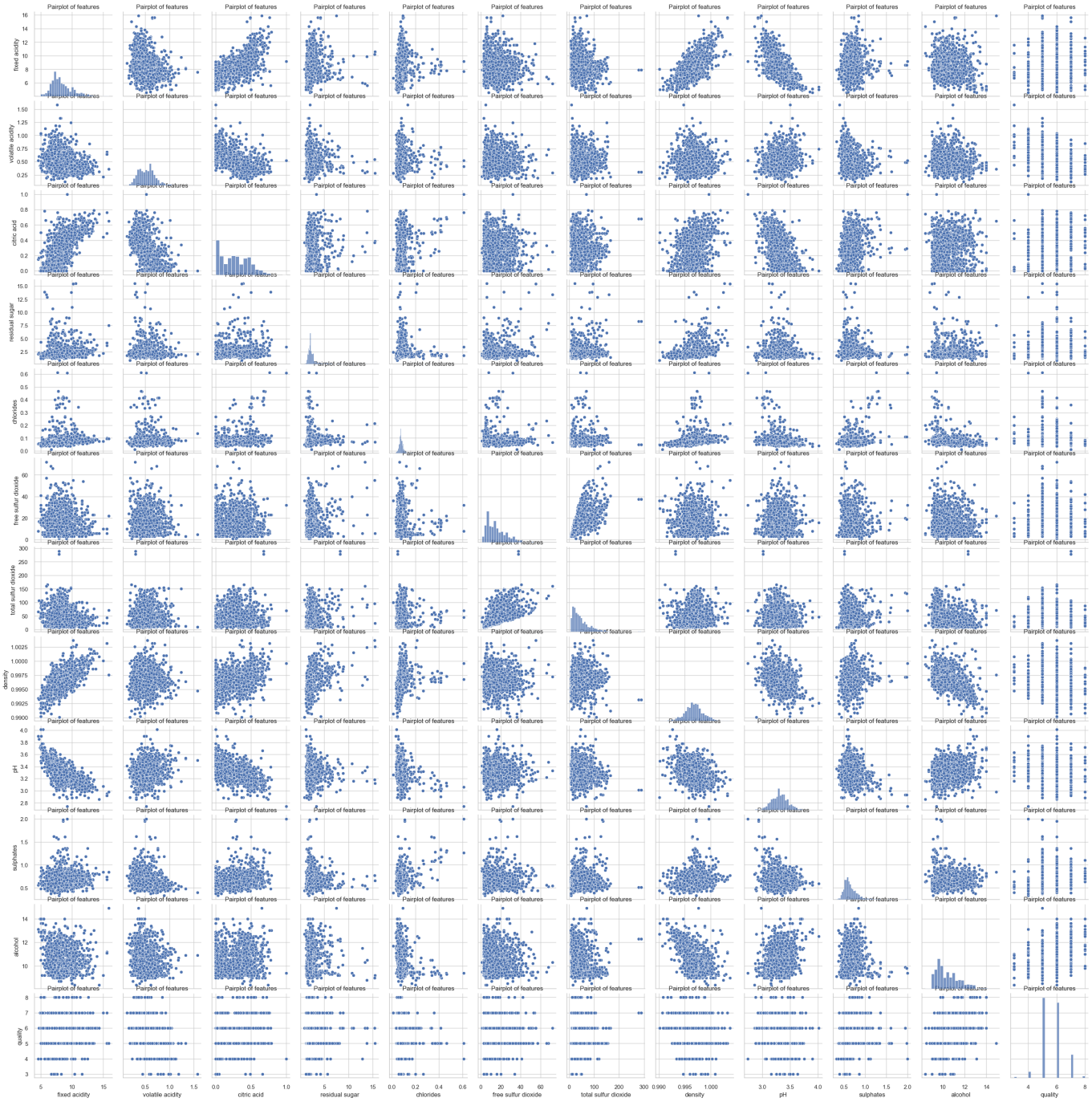
Feature Description

The dataset contains 11 features that describe different aspects of wine.

Fixed acidity is the amount of tartaric acid in the wine. Volatile acidity is a measure of the amount of acetic acid in the wine. Citric acid is a type of acid that is found in citrus fruits. Residual sugar is the amount of sugar that remains in the wine after fermentation. Chlorides are salts that are found in wine. Free sulfur dioxide is a type of preservative that is added to wine. Total sulfur dioxide is the sum of free and bound sulfur dioxide. Density is a measure of the weight of a substance per unit volume. pH is a measure of the acidity or alkalinity of a solution. Sulphates are salts of sulfuric acid. Alcohol is the percentage of alcohol by volume in the wine. Quality is a subjective measure of the wine's taste and overall quality.

Insights on dataset

The dataset contains 1599 instances. The features are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. The mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile, and maximum of each feature are reported.



Insights on Null Values in the dataset

The dataset does not contain any missing values. This is ideal as it means that all of the data is available for analysis. However, it is important to note that this dataset is relatively small, and it is possible that a larger dataset would contain some missing values. If a dataset does contain missing values, it is important to consider how to deal with them. One option is to simply remove the rows or columns that contain missing values, but this can lead to a loss of data. Another option is to impute the missing values, which means to replace them with an estimated value. The best approach to dealing with missing values will depend on the specific dataset and the analysis that is being performed.

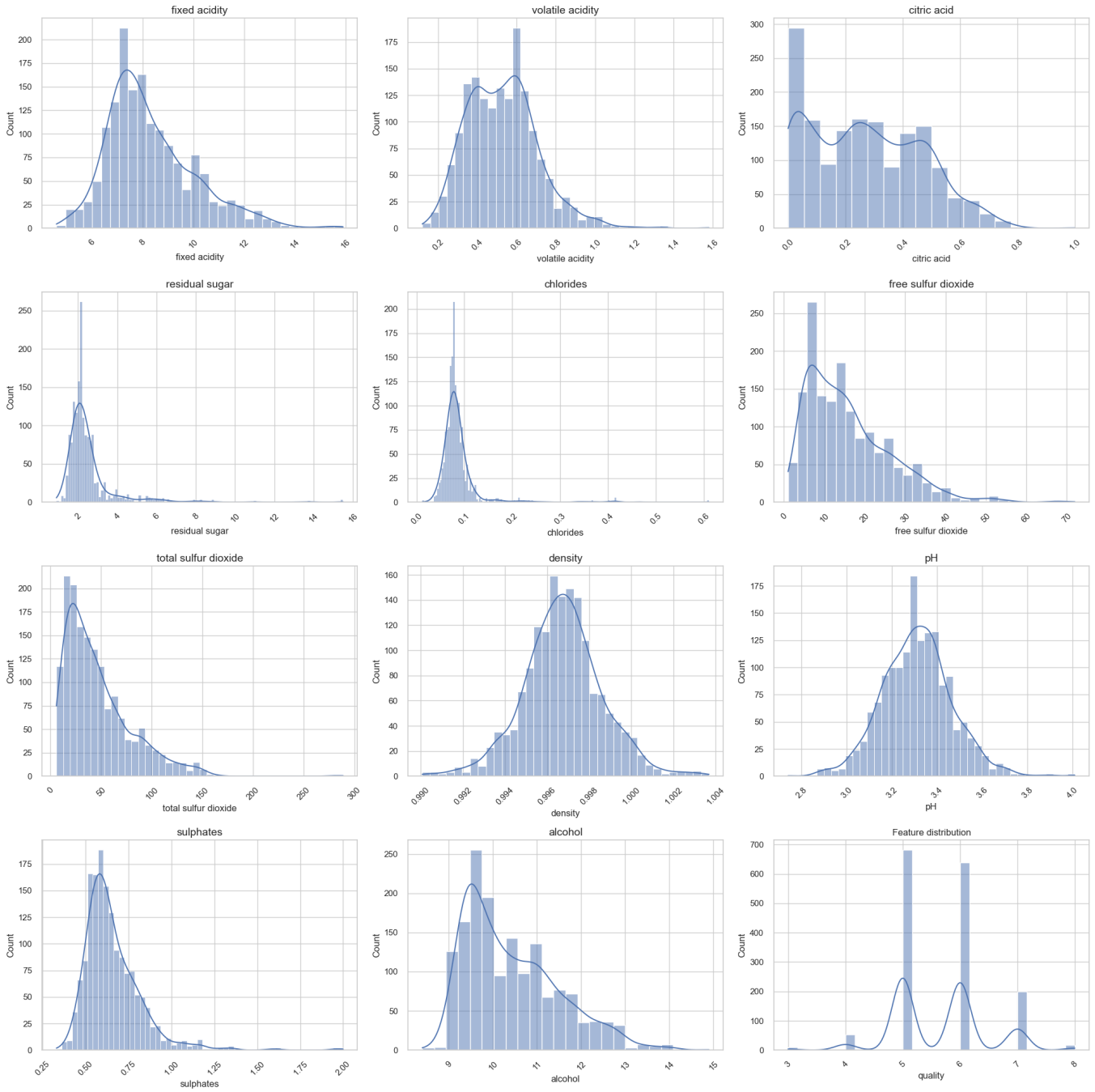
Feature Distribution

The distribution of each feature in the dataset is as follows:

fixed acidity: slightly left skewed
volatile acidity: slightly left skewed
citric acid: slightly left skewed
residual sugar: moderately right skewed
chlorides: moderately right skewed
free sulfur dioxide: slightly left skewed
total sulfur dioxide: slightly left skewed
density: slightly left skewed
pH: slightly left skewed
sulphates: moderately right skewed
alcohol: slightly left skewed

quality: slightly left skewed

The skewness of the data has several consequences. For example, the moderately right skewed distribution of residual sugar means that there are more observations with higher values of residual sugar than lower values. This could make it difficult to identify the outliers in this feature. Similarly, the slightly left skewed distribution of pH means that there are more observations with lower values of pH than higher values. This could make it difficult to identify the outliers in this feature.

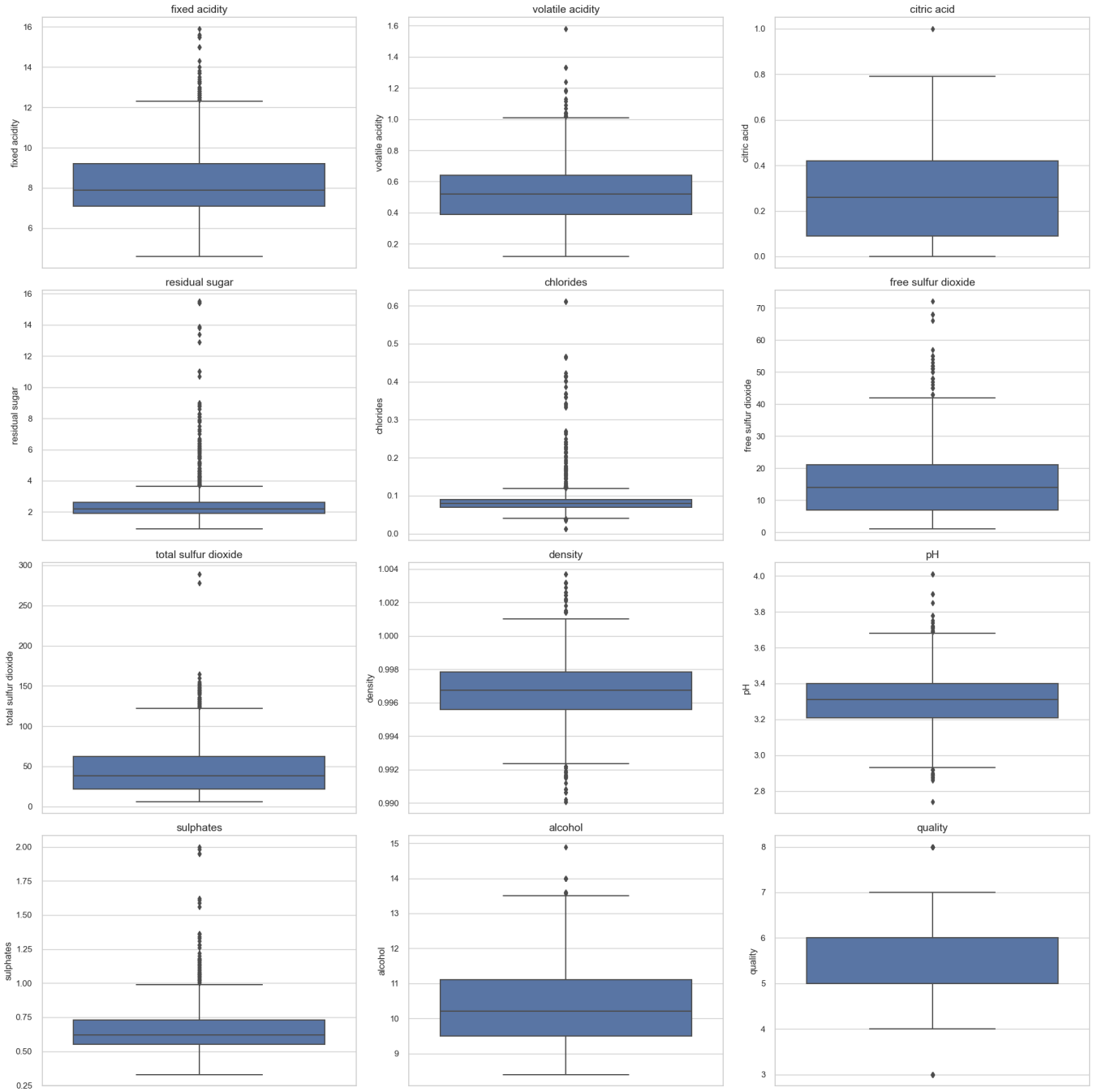


Outlier Detection

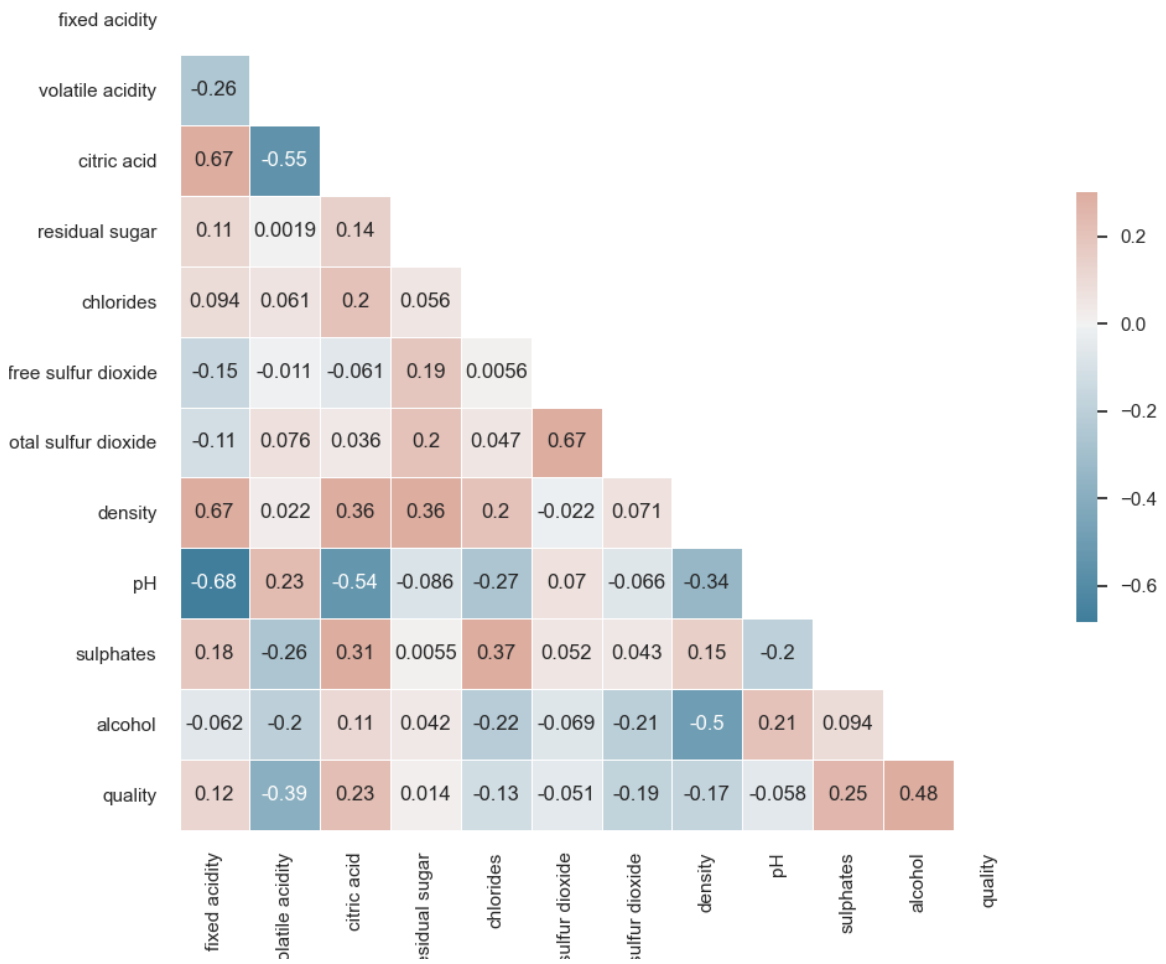
There are a few outliers in the dataset. For fixed acidity, there is a single data point of 15.9, which is much higher than the rest of the data. This could be due to a measurement error or a genuine outlier. For volatile acidity, there are two data points of 1.58 and 0.611, which are both higher than the rest of the data. This could be due to a measurement error or a genuine outlier. For citric acid, there is a single data point of 1.0, which is much lower than the rest of the data. This could be due to a measurement error or a genuine outlier. For residual sugar, there are two data points of 15.5 and 0.9, which are both higher than the rest of the data. This could be due to a measurement error or a genuine outlier. For chlorides, there is a single data point of 0.611, which is much higher than the rest of the data. This could be due to a measurement error or a genuine outlier. For free sulfur dioxide, there are two data points of 72 and 289, which are both much higher than the rest of the data. This could be due to a measurement error or a genuine outlier. For total sulfur dioxide, there is a single data point of 289, which is much higher than the rest of the data. This could be due to a measurement error or a genuine outlier. For density, there is a single data point of 1.00369, which is much higher than the rest of the data. This could be due to a measurement error or a genuine outlier. For pH, there is a single data point of 4.01, which is much lower than the rest of the data. This could be due to a measurement error or a genuine outlier. For sulphates, there is a single data point of 2.0, which is much higher than the rest of the data. This could be due to a measurement error or a genuine outlier. For alcohol, there are two data points of 14.9 and 8.0, which are both higher than the rest of the data. This could be due to a measurement error or a genuine outlier.

The presence of outliers in the dataset could have a number of consequences. For example, outliers could skew the results of statistical analyses, making it difficult to draw accurate conclusions. Outliers could also make it difficult to identify trends in the data. Additionally, outliers could make it difficult to develop models that accurately predict future outcomes.

It is important to take steps to address outliers in the dataset. One way to address outliers is to remove them from the dataset. However, this should only be done if the outliers are clearly erroneous. Another way to address outliers is to transform the data so that the outliers are less extreme. Finally, it is also possible to develop models that are robust to outliers.



Correlation between features



The correlation matrix shows that there are strong positive correlations between fixed acidity and citric acid (0.6717), citric acid and sulphates (0.3127), and alcohol and quality (0.4762). There are also strong negative correlations between volatile acidity and citric acid (-0.5525), volatile acidity and sulphates (-0.2609), and pH and sulphates (-0.5419).

Fixed acidity is the amount of tartaric acid in wine, and citric acid is a type of acid found in grapes. Both of these acids contribute to the tartness of wine. Sulphates are a type of mineral found in wine, and they can contribute to the dryness of wine. Alcohol is a type of alcohol found in wine, and it is responsible for the intoxicating effects of wine. Quality is a subjective measure of how good a wine is, and it is often based on the taste of the wine.

The strong positive correlations between fixed acidity and citric acid, citric acid and sulphates, and alcohol and quality suggest that these features are related to each other. For example, wines with high levels of fixed acidity and citric acid are likely to be dry and tart, while wines with high levels of alcohol are likely to be more intoxicating. The strong negative correlations between volatile acidity and citric acid, volatile acidity and sulphates, and pH and sulphates suggest that these features are inversely related to each other. For example, wines with high levels of volatile acidity and sulphates are likely to be sour and bitter, while wines with low levels of pH are likely to be acidic.

Overall, the correlation matrix provides a useful overview of the relationships between the features in the dataset. This information can be used to help understand the data and to make predictions about the quality of wine.