# 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 a measure of 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 after fermentation. Chlorides are salts that are found in wine. Free sulfur dioxide is a type of preservative that is added to wine to prevent the growth of bacteria. Total sulfur dioxide is the sum of free sulfur dioxide 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 substance. Sulphates are salts that are found in wine. Alcohol is the percentage of alcohol by volume in the wine. Quality is a subjective measure of the wine's overall quality.

## Insights on dataset

The dataset contains 1599 instances. The mean value of fixed acidity is 8.319637, the mean value of volatile acidity is 0.527821, and so on. The minimum value of fixed acidity is 4.600000, the minimum value of volatile acidity is 0.120000, and so on.

## Insights on Null Values in the dataset

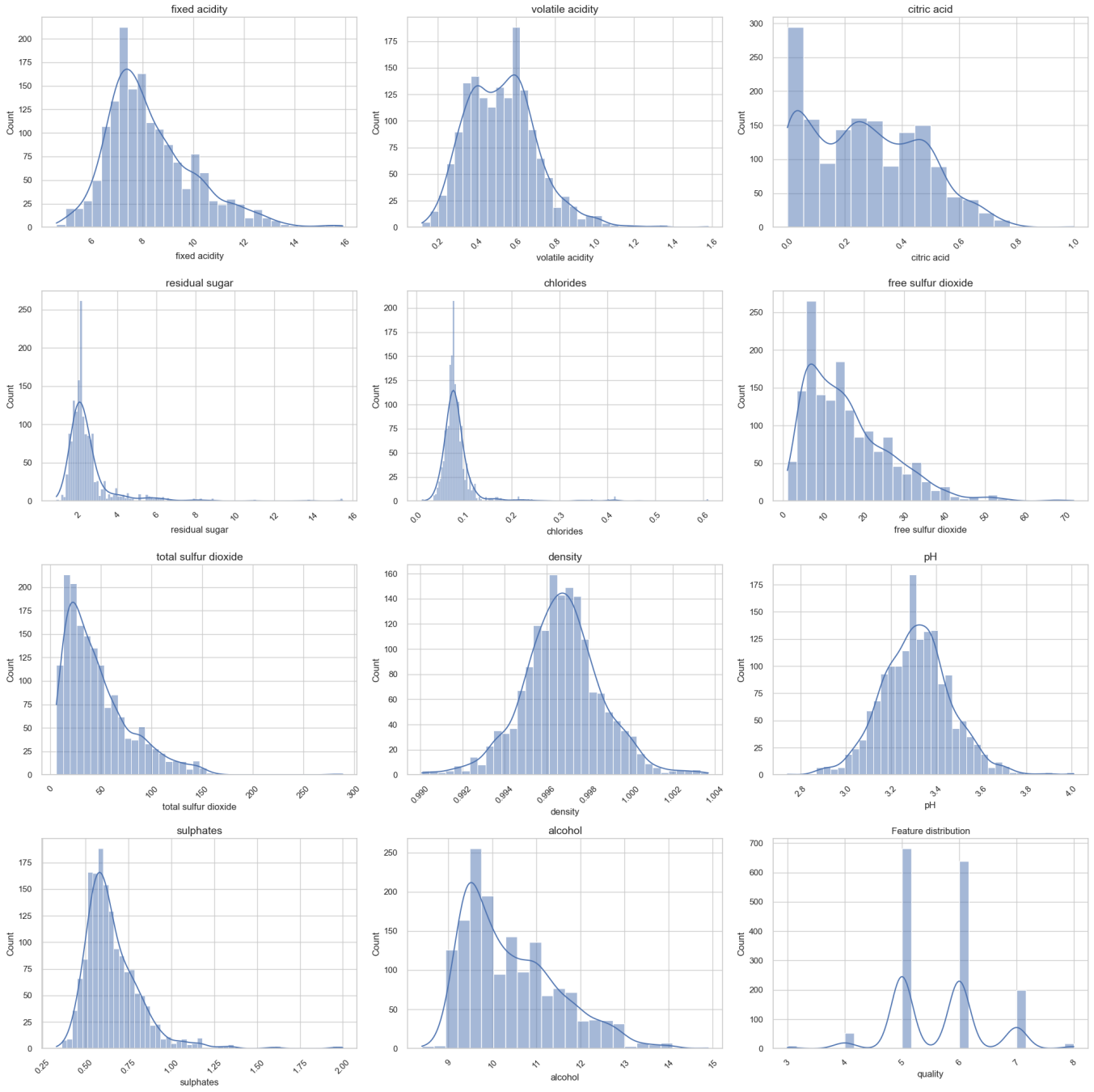
The dataset has 0 null values in each feature. This is an ideal situation as it means that all data is available for analysis. However, it is important to note that this dataset is relatively small and may not be representative of the population as a whole. Additionally, it is possible that some of the values may be inaccurate, as they were not collected in a controlled environment. Therefore, it is important to be cautious when interpreting the results of any analysis conducted on this dataset.

## Feature Distribution

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

Fixed acidity: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Volatile acidity: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Citric acid: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Residual sugar: moderately right skewed. This indicates that the data is more concentrated towards the higher values.
Chlorides: moderately right skewed. This indicates that the data is more concentrated towards the higher values.
Free sulfur dioxide: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Total sulfur dioxide: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Density: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
pH: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Sulphates: moderately right skewed. This indicates that the data is more concentrated towards the higher values.
Alcohol: slightly left skewed. This indicates that the data is more concentrated towards the lower values.
Quality: slightly left skewed. This indicates that the data is more concentrated towards the lower values.

The skewness of the data has several consequences. For example, the mean of a skewed distribution is not a good measure of central tendency, as it is pulled towards the tail of the distribution. Additionally, skewed distributions can make it difficult to interpret statistical tests, as the results may be biased towards the tail of the distribution.




## Outlier Detection

There are a few outliers in the dataset.

For fixed acidity, there are 4 observations below 4.6 and 1 observation above 15.9. These observations may be due to measurement errors or mislabeling.

For volatile acidity, there are 2 observations below 0.12 and 1 observation above 1.58. These observations may be due to measurement errors or mislabeling.

For citric acid, there are 2 observations below 0.0 and 1 observation above 1.0. These observations may be due to measurement errors or mislabeling.

For residual sugar, there are 2 observations below 0.9 and 1 observation above 15.5. These observations may be due to measurement errors or mislabeling.

For chlorides, there are 2 observations below 0.012 and 1 observation above 0.611. These observations may be due to measurement errors or mislabeling.

For free sulfur dioxide, there are 3 observations below 1.0 and 1 observation above 72.0. These observations may be due to measurement errors or mislabeling.

For total sulfur dioxide, there are 3 observations below 6.0 and 1 observation above 289.0. These observations may be due to measurement errors or mislabeling.

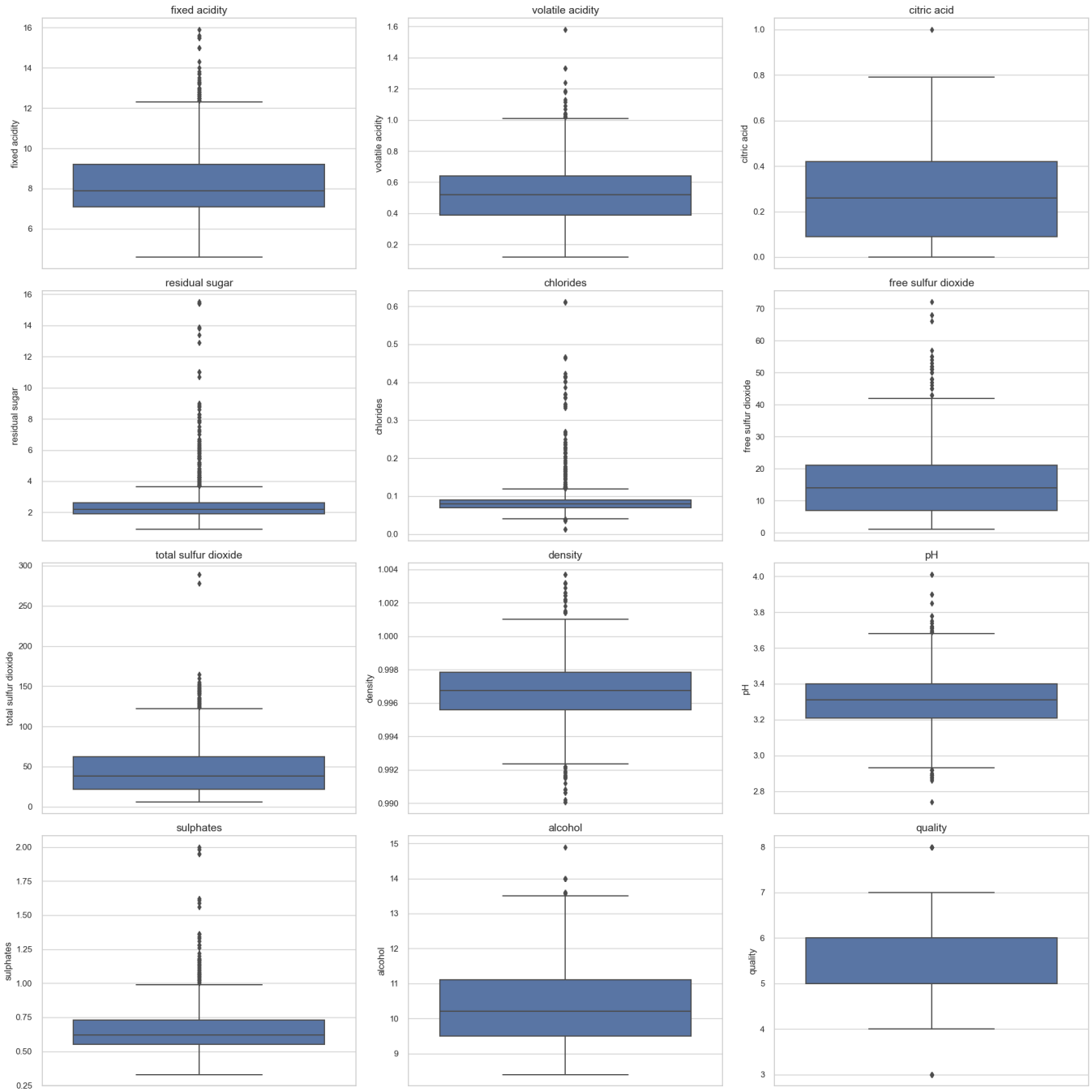
For density, there are 2 observations below 0.99007 and 1 observation above 1.00369. These observations may be due to measurement errors or mislabeling.

For pH, there are 2 observations below 2.74 and 1 observation above 4.01. These observations may be due to measurement errors or mislabeling.

For sulphates, there are 2 observations below 0.33 and 1 observation above 2.0. These observations may be due to measurement errors or mislabeling.

For alcohol, there are 2 observations below 8.4 and 1 observation above 14.9. These observations may be due to measurement errors or mislabeling.

The presence of outliers can have a significant impact on the results of statistical analysis. For example, if an outlier is included in a mean calculation, it can significantly skew the results. Therefore, it is important to carefully consider the presence of outliers when interpreting the results of statistical analysis.



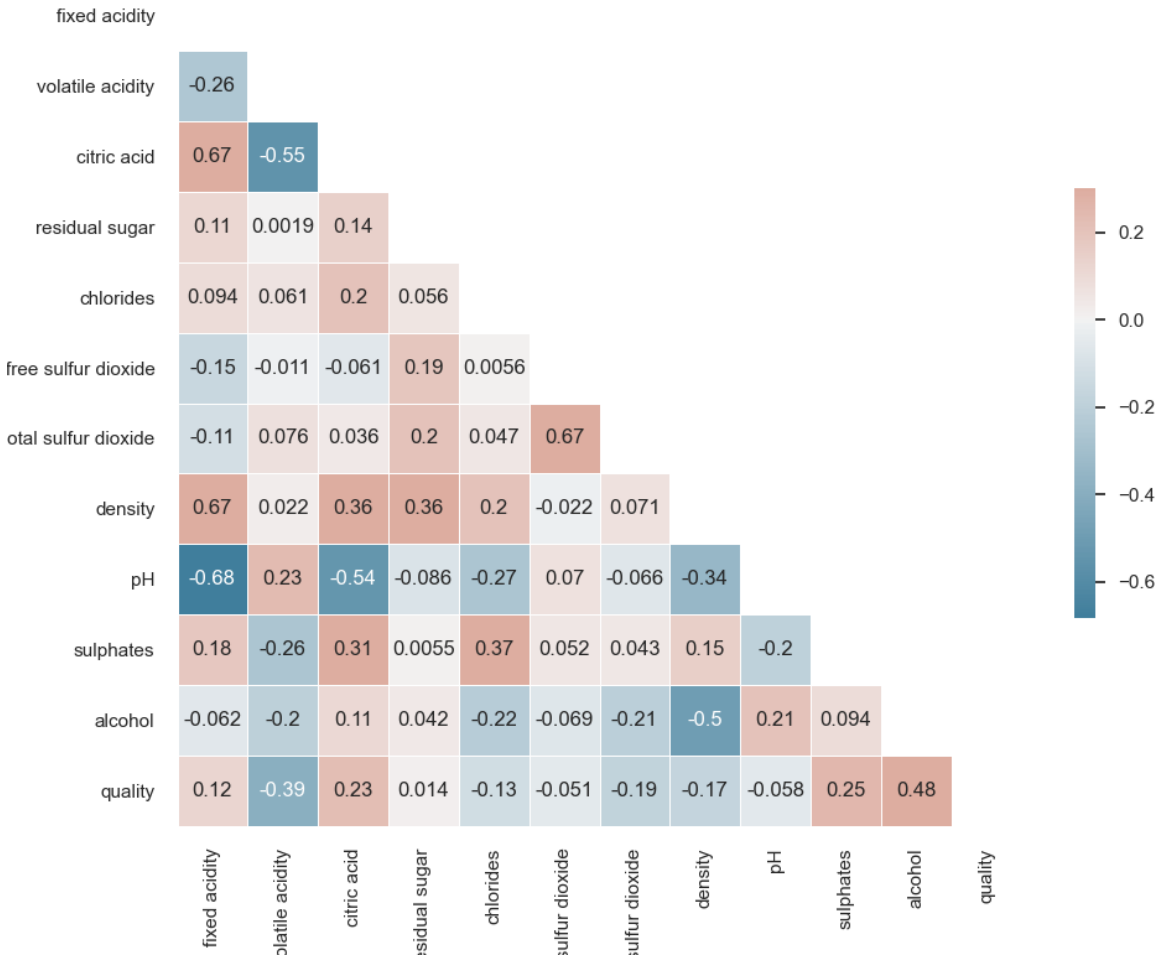

## Correlation between features

The correlation matrix shows the relationships between the 12 features in the dataset.

Fixed acidity is positively correlated with citric acid and negatively correlated with volatile acidity.
Volatile acidity is negatively correlated with citric acid and positively correlated with sulphates.
Citric acid is positively correlated with sulphates and negatively correlated with residual sugar.
Residual sugar is negatively correlated with chlorides and positively correlated with sulphates.
Chlorides is positively correlated with free sulfur dioxide and negatively correlated with total sulfur dioxide.

Free sulfur dioxide is negatively correlated with total sulfur dioxide.
Total sulfur dioxide is negatively correlated with density.
Density is positively correlated with pH and negatively correlated with sulphates.
pH is negatively correlated with sulphates and alcohol.
Sulphates is positively correlated with alcohol and quality.
Alcohol is positively correlated with quality.

Overall, the dataset shows that the features are moderately correlated with each other.




## Machine Learning experiment settings

This experiment is a regression task with the target variable of `quality`. The original data is a 1599x12 dataframe, and after transformation, it becomes a 1599x12 dataframe. The training set has 1119 rows and the test set has 480 rows. The experiment uses 10-fold cross-validation with simple imputation (mean for numeric features and mode for categorical features). The experiment is run on CPU without using GPU.

| Description | Value |
|---|---|
| Target | quality |
| Target type | Regression |
| Original data shape | (1599, 12) |
| Transformed data shape | (1599, 12) |
| Transformed train set shape | (1119, 12) |
| Transformed test set shape | (480, 12) |
| Numeric features | 11 |
| Preprocess | True |

| Description | Value |
|---|---|
| Imputation type | simple |
| Numeric imputation | mean |
| Categorical imputation | mode |
| Fold Generator | KFold |
| Fold Number | 10 |
| CPU Jobs | -1 |
| Use GPU | False |