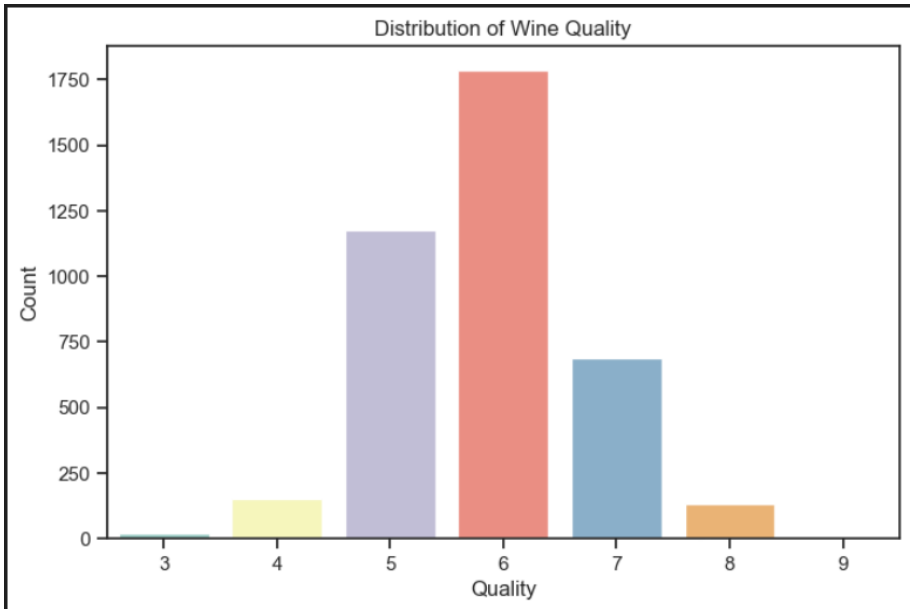
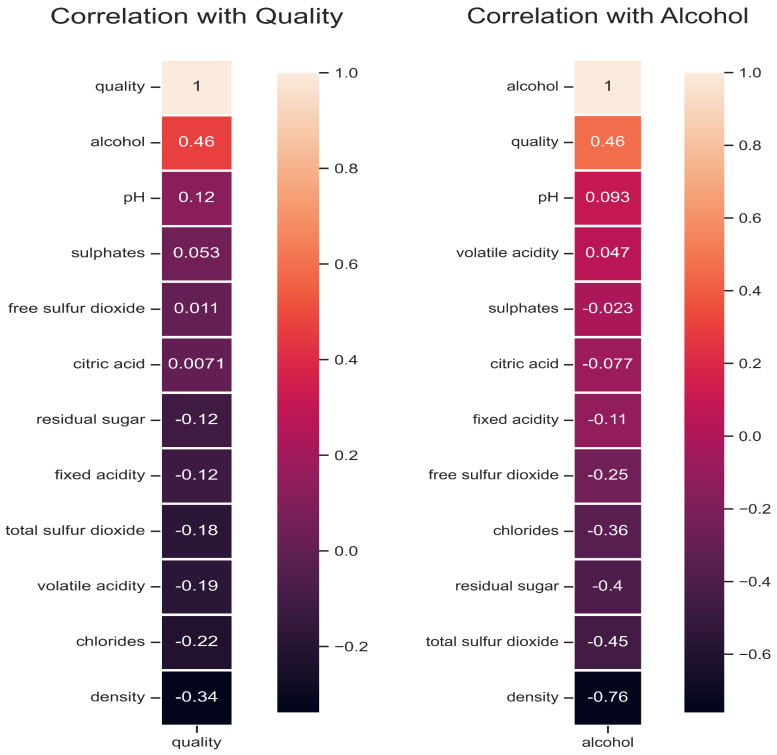
Made by - Rahul Jain

WINE QUALITY PREDICTION

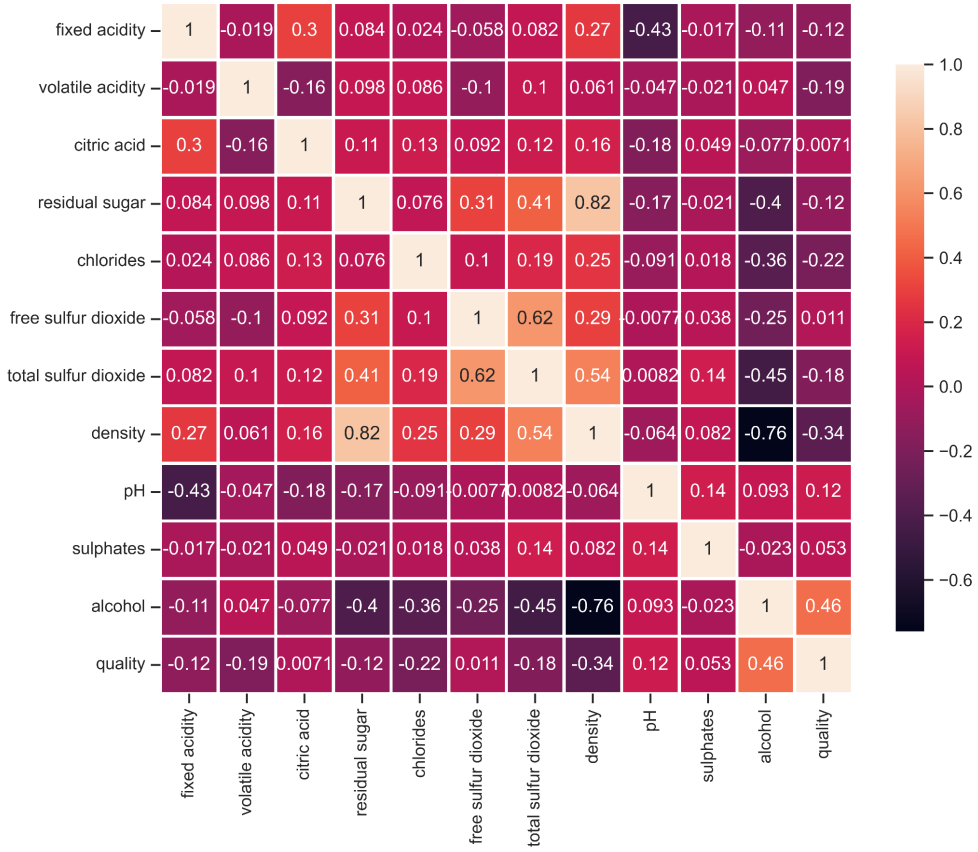
1. What is the distribution of the wine quality scores?



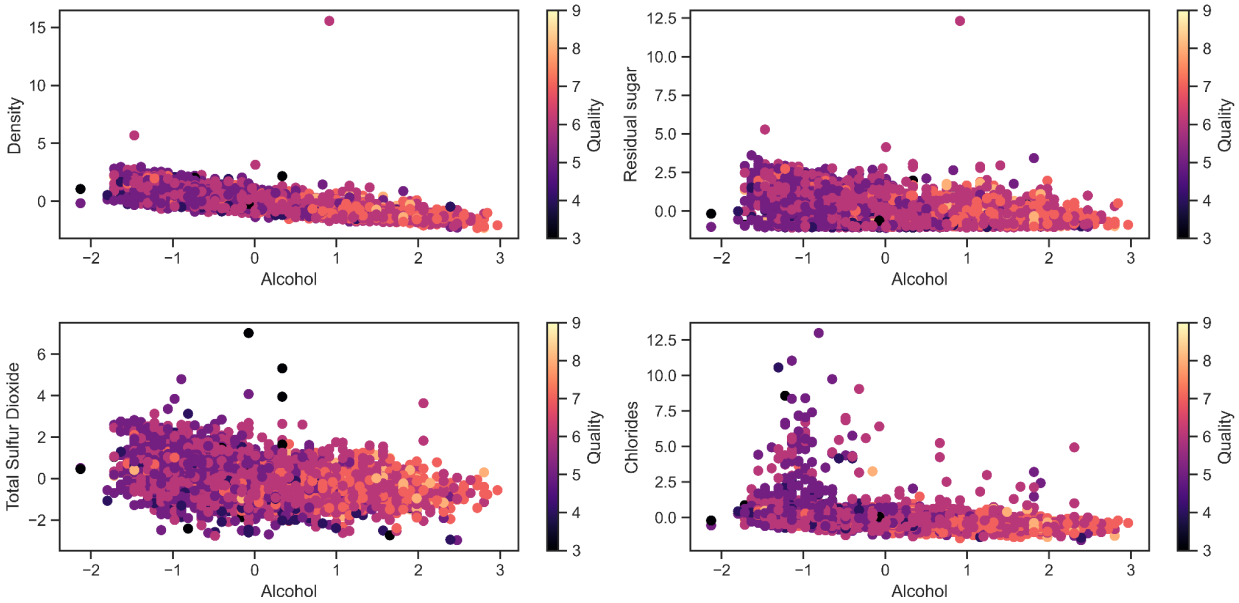
2. What are the relationships between the different features?

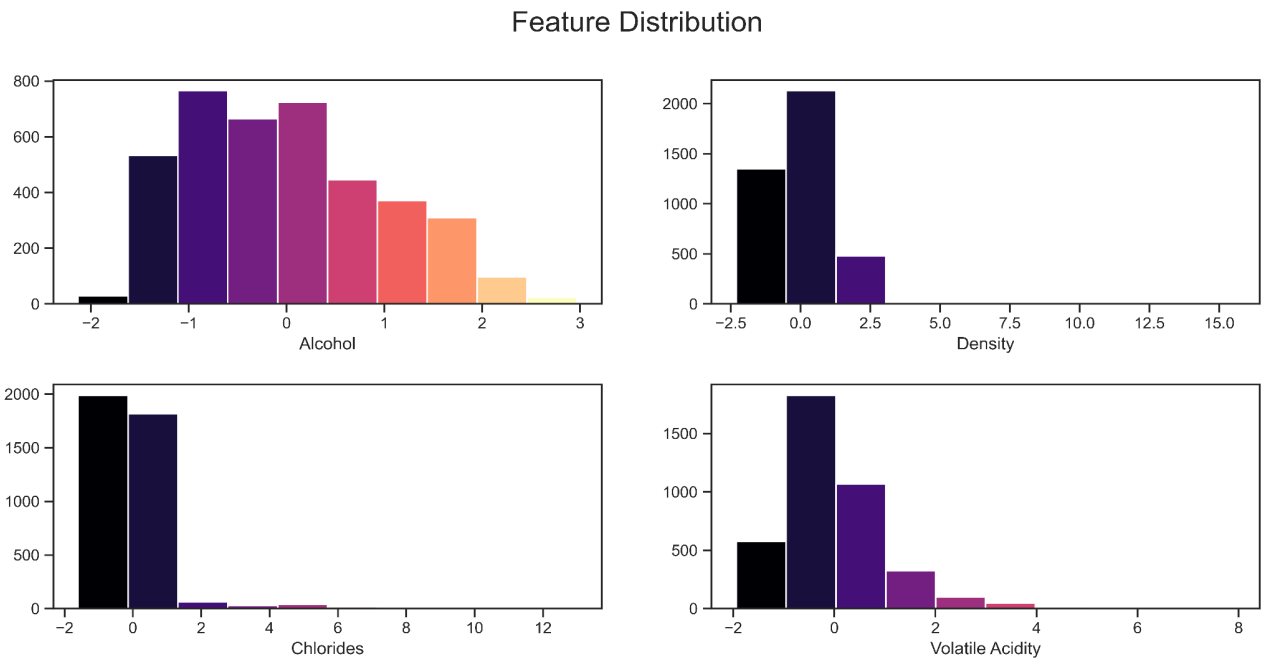


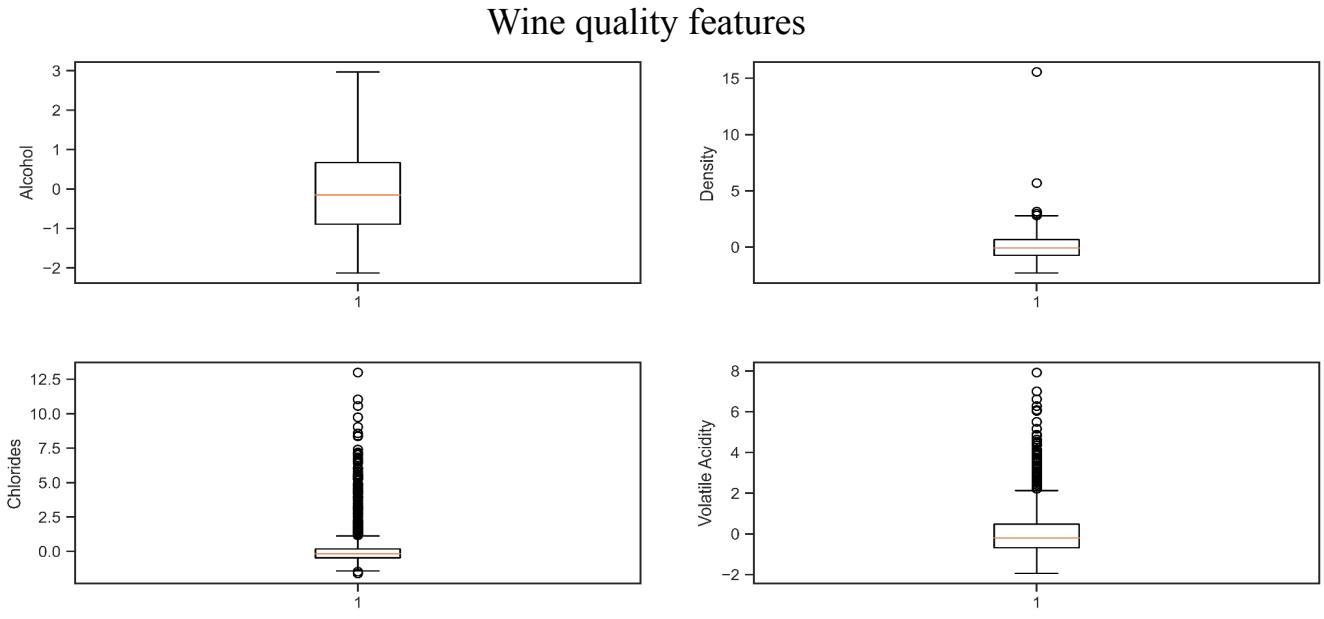
Correlation Matrix



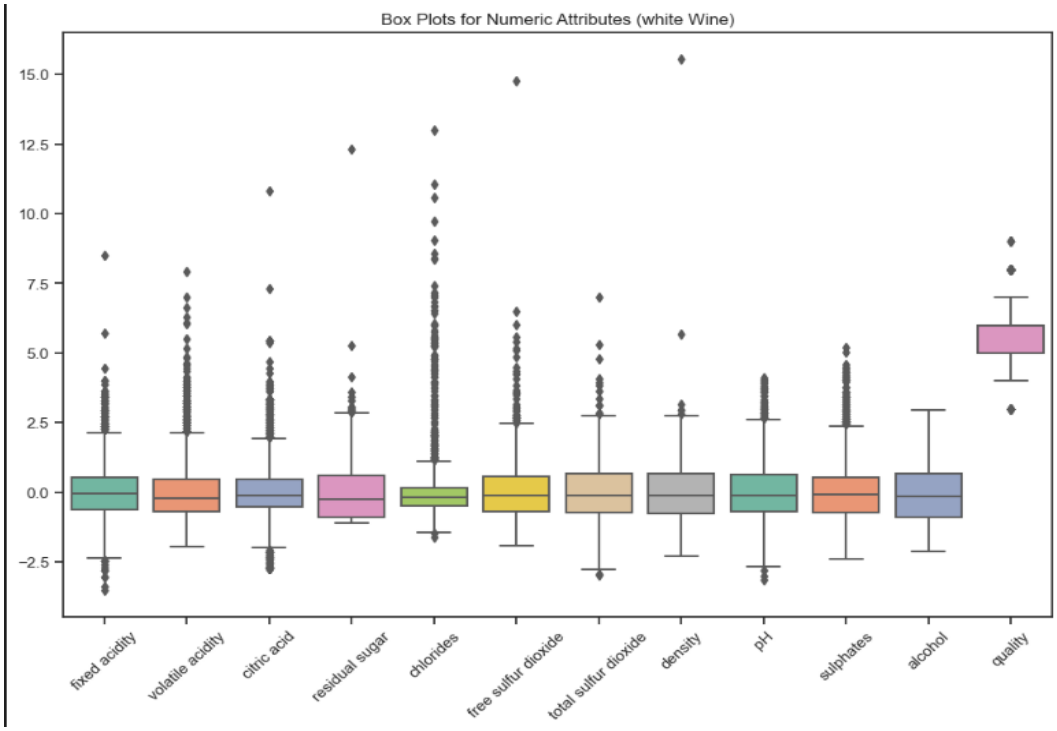
Alcohol Correlation



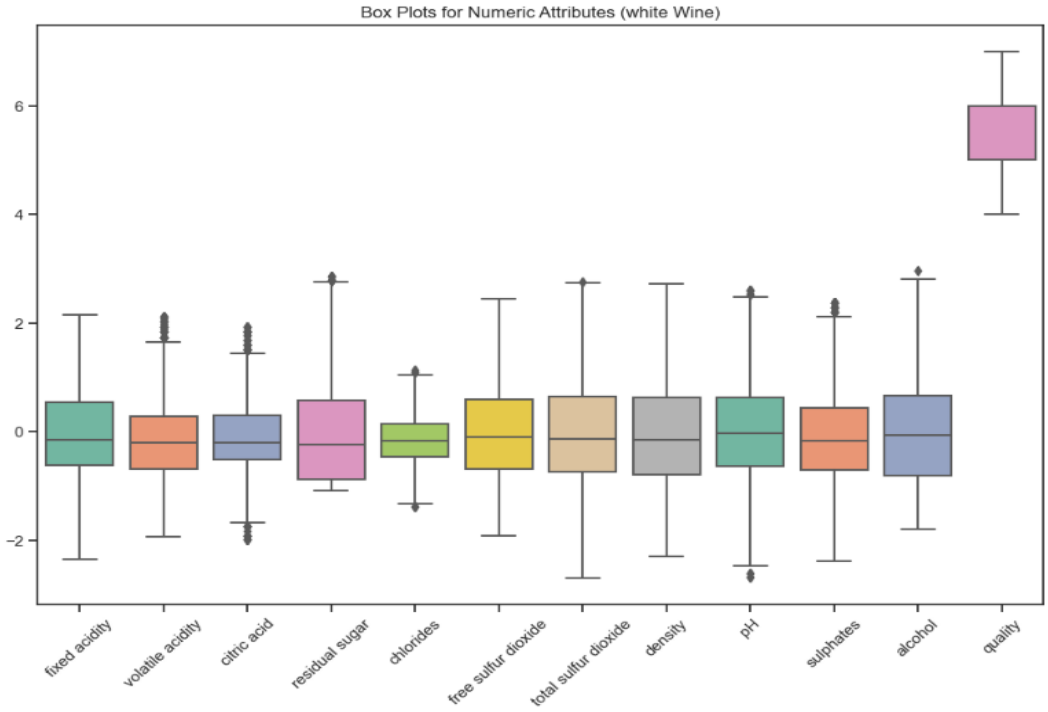

Feature Distribution
Alcohol
Density
Chlorides
Volatile Acidity


Wine quality features
Alcohol
Density
Chlorides
Volatile Acidity

3. Are there any outliers in the data?



Without outliers:



4. **What is the accuracy of the linear regression model?**

```
Training Accuracy : 0.8500159239461068
Testing Accuracy : 0.8873016540112716
```

5. **What are the most important features for the linear regression model?**

```
Feature Ranking:
Feature fixed acidity: 4
Feature volatile acidity: 1
Feature citric acid: 5
Feature residual sugar: 1
Feature chlorides: 7
Feature free sulfur dioxide: 2
Feature total sulfur dioxide: 6
Feature density: 1
Feature pH: 1
Feature sulphates: 3
Feature alcohol: 1
Selected Features: Index(['volatile acidity', 'residual sugar', 'density', 'pH', 'alcohol'], dtype='object')
```

6. **What is the MSE of the linear regression model?**

```
Mean Squared Error (MSE): 0.1146881327254739
```

7. **What is the r-squared of the linear regression model?**

```
R-squared (R2): 0.8873016540112716
```

8. **How can you improve the performance of the linear regression model?**

   The performance of the linear regression model can be improved by:

   1. Feature engineering and selection
   2. Outlier handling
   3. Data preprocessing
   4. Regularization
   5. Feature scaling
   6. Data cleaning

9. **What are the limitations of the linear regression model?**
    a. Linear regression assumes a linear relationship between the features and the target variable. In real-world datasets like wine quality, relationships can be nonlinear, and linear regression may not capture these complexities effectively.
    b. It is sensitive to outliers
    c. Linear regression is designed for numerical features and may not handle categorical variables well without proper encoding.
    d. Linear regression is relatively simple and may not be suitable for modeling complex relationships or capturing interactions between features.
    e. the residuals may deviate from a normal distribution.

10. **What are the implications of your findings for the real-world problem?**

The operations conducted on this wine quality prediction dataset provide valuable insights for real-world problems:

1. Feature Engineering: Effective feature engineering can enhance the predictive power of machine learning models. In real-world scenarios, understanding the domain and the dataset is critical for creating meaningful features.

2. Feature Selection: Identifying the most relevant features is essential for model efficiency and interpretability. In practice, feature selection helps in reducing model complexity and can lead to faster predictions.

3. Outlier Detection: Detecting and handling outliers is crucial for model performance. In real-world applications, outliers may represent exceptional or erroneous data points. Deciding how to handle them depends on the context and potential impact on model predictions.

4. Data Quality: Ensuring data quality and integrity is paramount in real-world applications. Careful preprocessing, data cleaning, and validation are essential to build trustworthy predictive models.