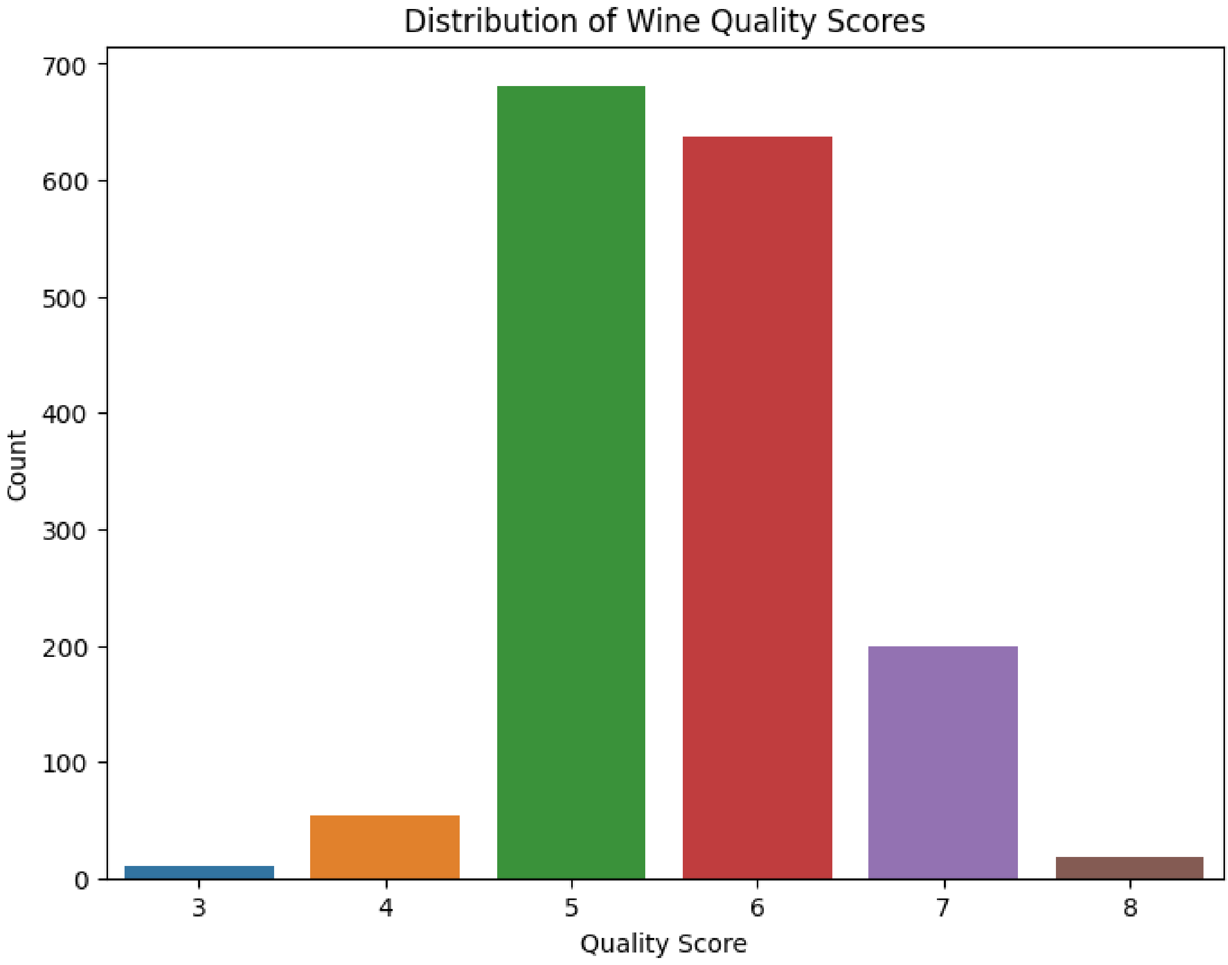
# ASSIGNMENT 2: PREDICTING WINE QUALITY WITH LINEAR REGRESSION

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

   **Code:**

   ```
   plt.figure(figsize=(8, 6))
   sns.countplot(x='quality', data=df)
   plt.title('Distribution of Wine Quality Scores')
   plt.xlabel('Quality Score')
   plt.ylabel('Count')
   plt.show()
   ```

   **output:**



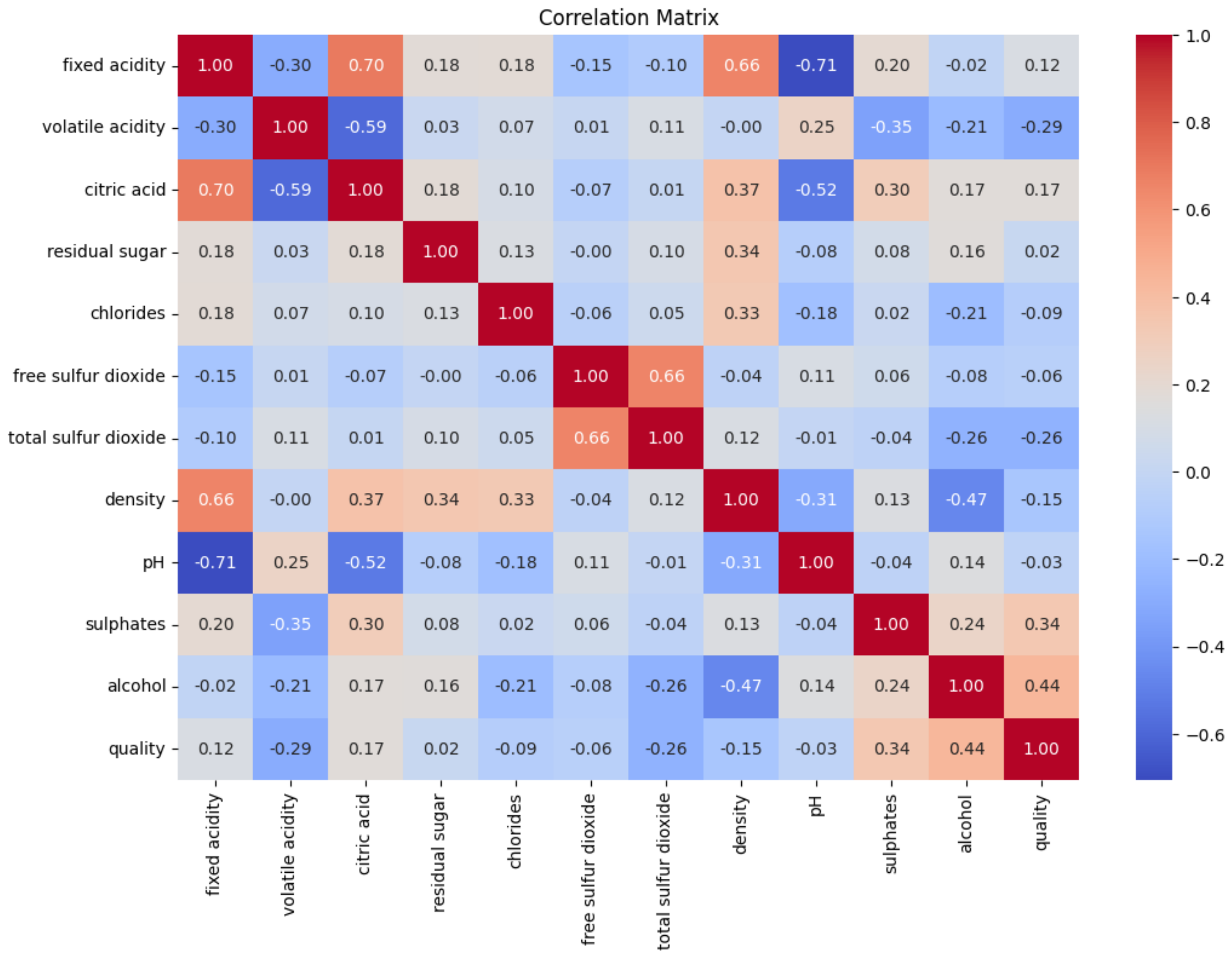
The distribution is across 3,4,5,6,7,8 . the highest distribution is for quality 5.

## 2. What are the relationships between the different features?

**Code:**

```
import seaborn as sns
corr_matrix = df.corr()
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```

**output:**



| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fixed acidity | 1.00 | -0.30 | 0.70 | 0.18 | 0.18 | -0.15 | -0.10 | 0.66 | -0.71 | 0.20 | -0.02 | 0.12 |
| volatile acidity | -0.30 | 1.00 | -0.59 | 0.03 | 0.07 | 0.01 | 0.11 | -0.00 | 0.25 | -0.35 | -0.21 | -0.29 |
| citric acid | 0.70 | -0.59 | 1.00 | 0.18 | 0.10 | -0.07 | 0.01 | 0.37 | -0.52 | 0.30 | 0.17 | 0.17 |
| residual sugar | 0.18 | 0.03 | 0.18 | 1.00 | 0.13 | -0.00 | 0.10 | 0.34 | -0.08 | 0.08 | 0.16 | 0.02 |
| chlorides | 0.18 | 0.07 | 0.10 | 0.13 | 1.00 | -0.06 | 0.05 | 0.33 | -0.18 | 0.02 | -0.21 | -0.09 |
| free sulfur dioxide | -0.15 | 0.01 | -0.07 | -0.00 | -0.06 | 1.00 | 0.66 | -0.04 | 0.11 | 0.06 | -0.08 | -0.06 |
| total sulfur dioxide | -0.10 | 0.11 | 0.01 | 0.10 | 0.05 | 0.66 | 1.00 | 0.12 | -0.01 | -0.04 | -0.26 | -0.26 |
| density | 0.66 | -0.00 | 0.37 | 0.34 | 0.33 | -0.04 | 0.12 | 1.00 | -0.31 | 0.13 | -0.47 | -0.15 |
| pH | -0.71 | 0.25 | -0.52 | -0.08 | -0.18 | 0.11 | -0.01 | -0.31 | 1.00 | -0.04 | 0.14 | -0.03 |
| sulphates | 0.20 | -0.35 | 0.30 | 0.08 | 0.02 | 0.06 | -0.04 | 0.13 | -0.04 | 1.00 | 0.24 | 0.34 |
| alcohol | -0.02 | -0.21 | 0.17 | 0.16 | -0.21 | -0.08 | -0.26 | -0.47 | 0.14 | 0.24 | 1.00 | 0.44 |
| quality | 0.12 | -0.29 | 0.17 | 0.02 | -0.09 | -0.06 | -0.26 | -0.15 | -0.03 | 0.34 | 0.44 | 1.00 |

**Observation:**

**Positive Correlations:**

-"Alcohol" has a strong positive correlation with "Quality." This suggests that wines with higher alcohol content tend to have higher quality scores.

- "Citric Acid" has a moderate positive correlation with "Fixed Acidity." This indicates that wines with higher fixed acidity tend to have slightly higher citric acid content.

- "Free Sulfur Dioxide" and "Total Sulfur Dioxide" have a strong positive correlation, which is expected since the total sulfur dioxide includes free sulfur dioxide.

**Negative Correlations**:

- "Volatile Acidity" has a moderate negative correlation with "Quality." Lower volatile acidity is associated with higher-quality wines.
- "pH" has a negative correlation with "Fixed Acidity." Wines with higher fixed acidity tend to have lower pH values.
- "Citric Acid" and "Volatile Acidity" have a negative correlation, suggesting that higher citric acid content is associated with lower volatile acidity.

**Low Correlations**:

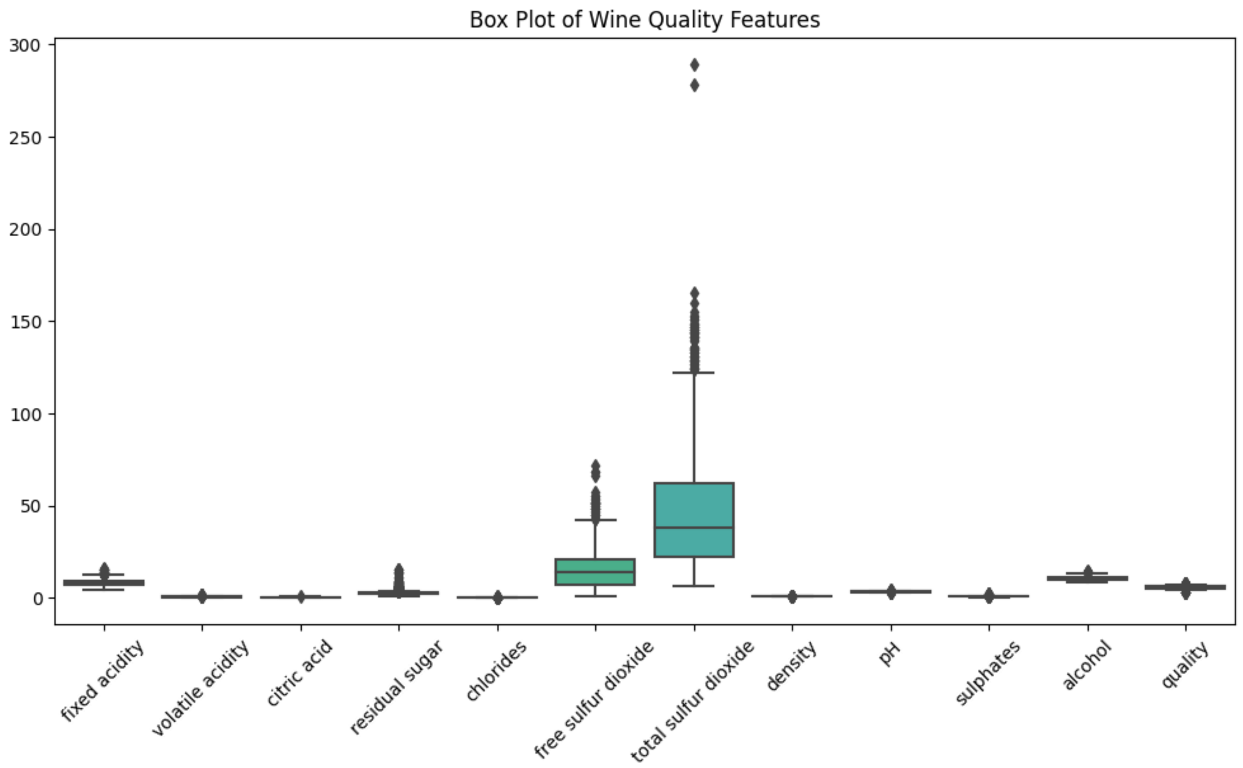
- Features like "Chlorides," "Density," and "Residual Sugar" show relatively low correlations with other features, indicating that they may not strongly influence wine quality.

3. **Are there any outliers in the data?**

**Code:**

```
plt.figure(figsize=(12, 6))
sns.boxplot(data=df)
plt.title("Box Plot of Wine Quality Features")
plt.xticks(rotation=45)
plt.show()
```

**output**:



**Observation:**

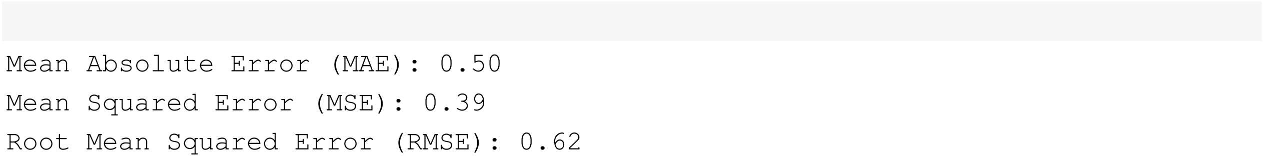
There are outliers in the data.

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

**Code:**

```
X = df.drop('quality', axis=1)
y = df['quality']
# training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
# Display evaluation metrics
print(f'Mean Absolute Error (MAE): {mae:.2f}')
print(f'Mean Squared Error (MSE): {mse:.2f}')
print(f'Root Mean Squared Error (RMSE): {rmse:.2f}')
```

**output:**



```
Mean Absolute Error (MAE): 0.50
Mean Squared Error (MSE): 0.39
Root Mean Squared Error (RMSE): 0.62
```

**Observation:**

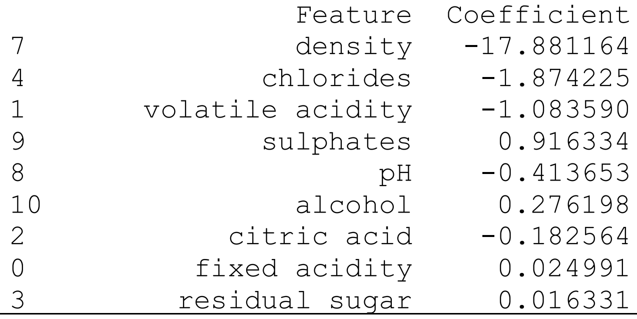
The linear regression model exhibits moderate predictive accuracy, with Mean Absolute Error (MAE) of 0.50 indicating an average prediction error of 0.50 quality score points. However, the Root Mean Squared Error (RMSE) of 0.62 suggests a moderate level of variability in predictions compared to actual wine quality scores.

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

**Code:**

```
model = LinearRegression()
model.fit(X, y)
coefficients = model.coef_
feature_names = X.columns
feature_importance = pd.DataFrame({'Feature': feature_names, 'Coefficient': coefficients})
feature_importance = feature_importance.reindex(feature_importance['Coefficient'].abs().sort_values(ascending=False).index)
print(feature_importance)
```

**output:**



```
                Feature  Coefficient
7               density   -17.881164
4             chlorides    -1.874225
1      volatile acidity    -1.083590
9             sulphates     0.916334
8                    pH    -0.413653
10              alcohol     0.276198
2           citric acid    -0.182564
0         fixed acidity     0.024991
3        residual sugar     0.016331
```

```
5    free sulfur dioxide     0.004361
6   total sulfur dioxide    -0.003265
```

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

   **Code:**

```
X = df.drop('quality', axis=1)
y = df['quality']
# training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
print(f'Mean Squared Error (MSE): {mse:.2f}')
```

   **output:**

```
Mean Squared Error (MSE): 0.31
```

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

   **Code:**

```
X = df.drop('quality', axis=1)
y = df['quality']
# training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
print(f'Root Mean Squared Error (RMSE): {rmse:.2f}')
```

   **Output:**

```
Root Mean Squared Error (RMSE): 0.56
```

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

   1. Feature Engineering
   2. Regularization
   3. Data Normalization
   4. Explore Nonlinear Relationships

**9. What are the limitations of the linear regression model?**

The limitations of the linear regression model include its assumption of a linear relationship between features and the target variable, sensitivity to outliers, independence of predictor variables, and the assumption of constant error variance. It may not handle nonlinearity, complex interactions, and categorical data effectively. Additionally, it can suffer from overfitting or underfitting without regularization, and its predictive performance may be limited in cases where these assumptions do not hold.

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

The findings from the analysis of wine quality data have several implications for the real-world problem of wine quality assessment. Understanding the factors that influence wine quality, as identified through feature importance analysis and regression coefficients, can guide winemakers and viticulturists in making informed decisions to improve wine production. For instance, the identification of "Density," "Chlorides," and "Volatile Acidity" as influential factors suggests that maintaining optimal levels of these attributes can positively impact wine quality. Moreover, recognizing the limitations of linear regression in modeling wine quality highlights the need for more complex models that can capture nonlinear relationships and interactions. Overall, these findings contribute to enhancing the quality control and production processes in the wine industry, ultimately leading to better wine quality for consumers.