In [1]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

red_wine_data = pd.read_csv('winequality-red.csv', delimiter=';')
white_wine_data = pd.read_csv('winequality-white.csv', delimiter=';')
```

In [2]:
```python
# Basic exploration of the red wine dataset
red_wine_data.head()
red_wine_data.info()
red_wine_data.describe()

# Basic exploration of the white wine dataset
white_wine_data.head()
white_wine_data.info()
white_wine_data.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         4898 non-null   float64
 1   volatile acidity      4898 non-null   float64
 2   citric acid           4898 non-null   float64
 3   residual sugar        4898 non-null   float64
 4   chlorides             4898 non-null   float64
 5   free sulfur dioxide   4898 non-null   float64
 6   total sulfur dioxide  4898 non-null   float64
 7   density               4898 non-null   float64
 8   pH                    4898 non-null   float64
 9   sulphates             4898 non-null   float64
 10  alcohol               4898 non-null   float64
 11  quality               4898 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

Out[2]:

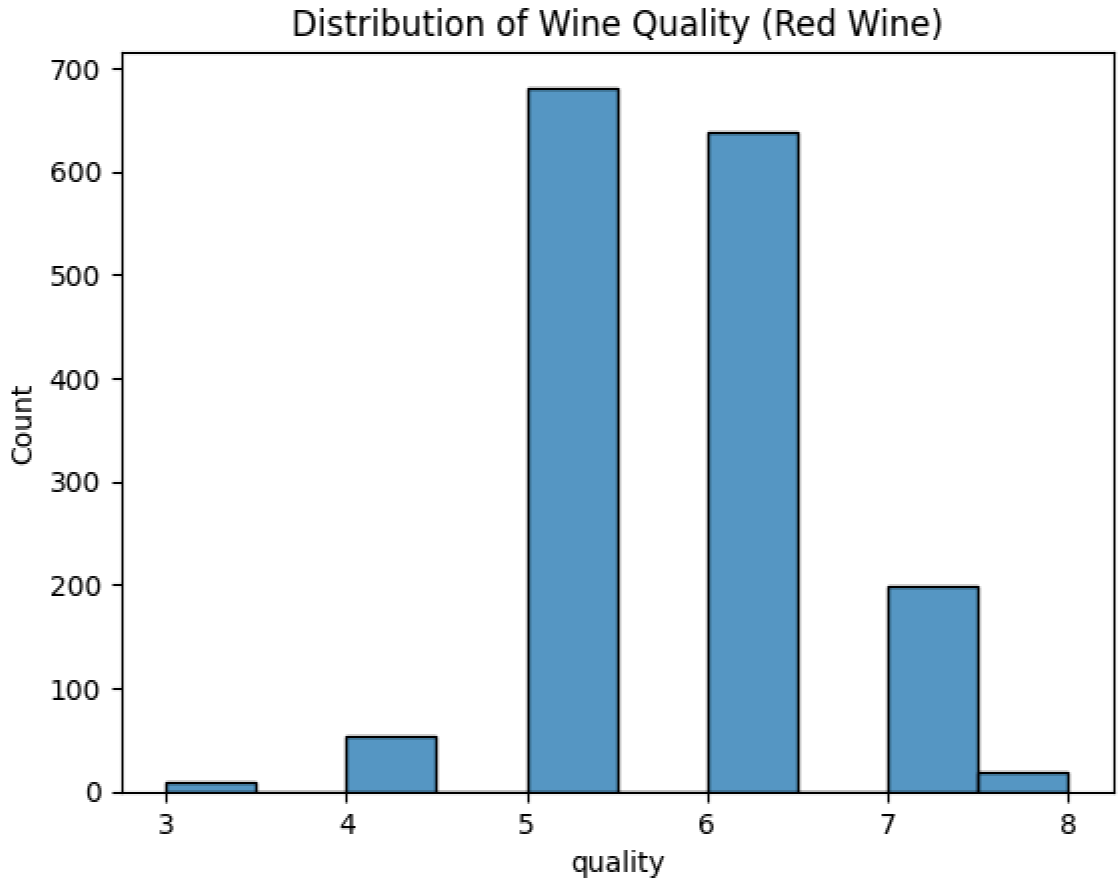
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	to
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	489
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	13
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	4
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	10
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	13
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	16
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	44

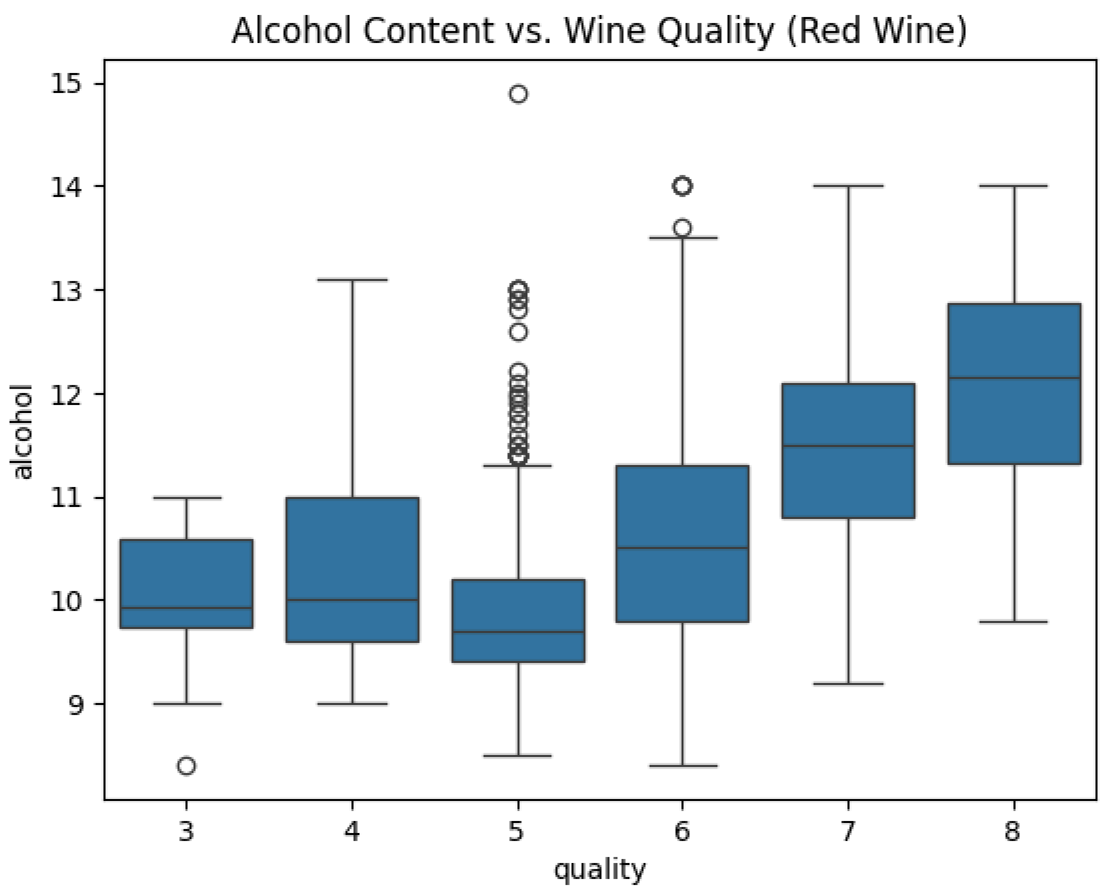
In [13]:

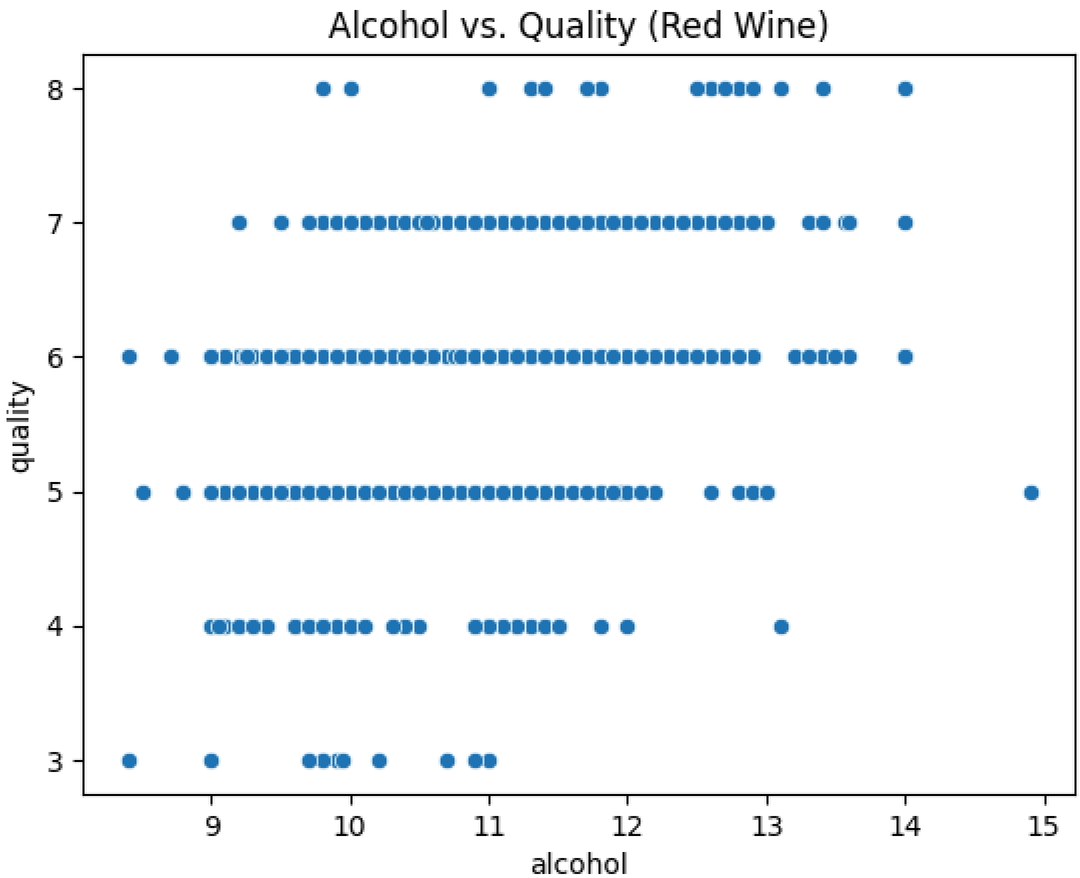
```
# Visualizing distributions of important features for red wine
sns.histplot(red_wine_data['quality'], kde=False, bins=10)
plt.title('Distribution of Wine Quality (Red Wine)')
plt.show()

# Boxplot of alcohol content
sns.boxplot(data=red_wine_data, x='quality', y='alcohol')
plt.title('Alcohol Content vs. Wine Quality (Red Wine)')
plt.show()

# Scatter plot to show correlation between alcohol and quality
sns.scatterplot(data=red_wine_data, x='alcohol', y='quality')
plt.title('Alcohol vs. Quality (Red Wine)')
plt.show()
```


Distribution of Wine Quality (Red Wine)
Count
0
100
200
300
400
500
600
700
3
4
5
6
7
8
quality


Alcohol Content vs. Wine Quality (Red Wine)
alcohol
9
10
11
12
13
14
15
3
4
5
6
7
8
quality

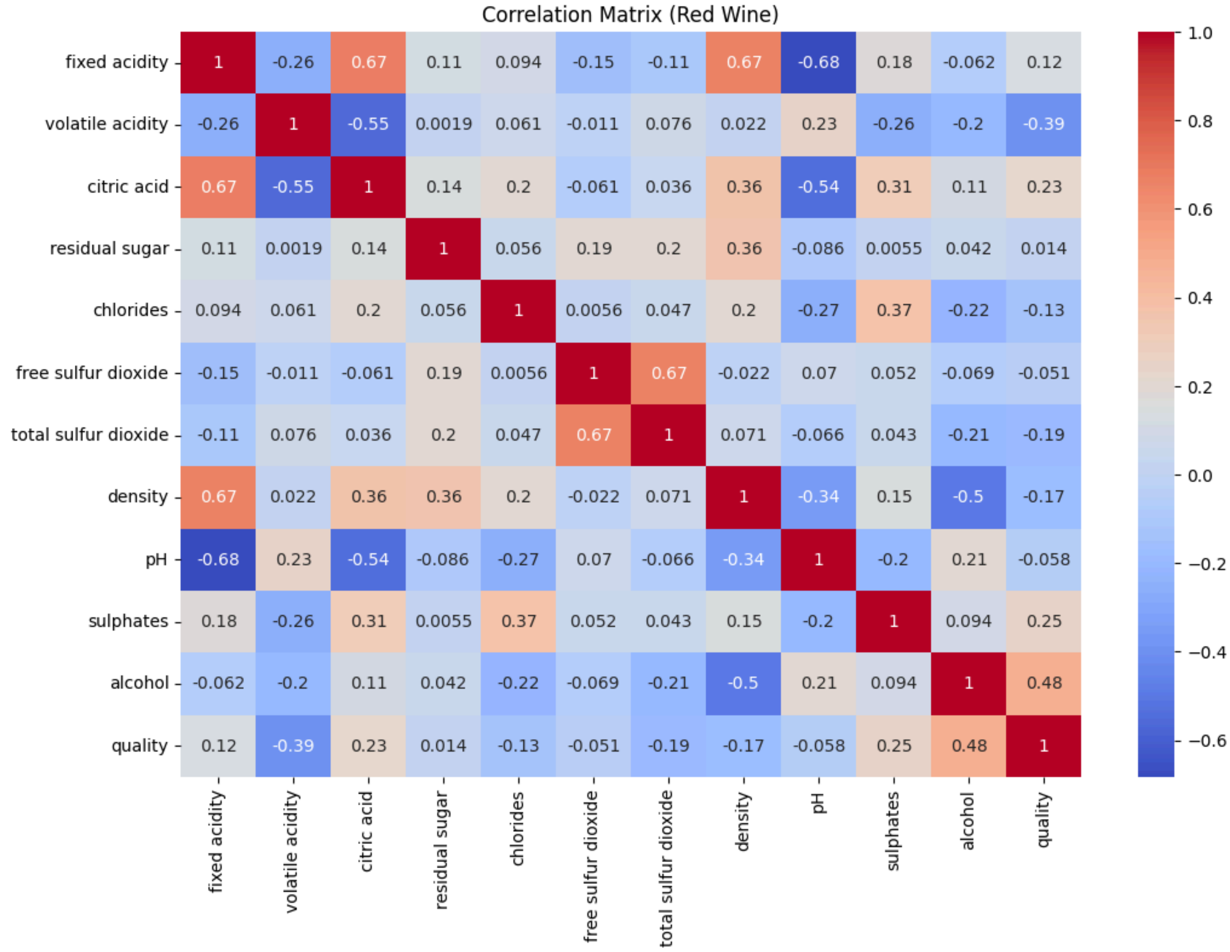


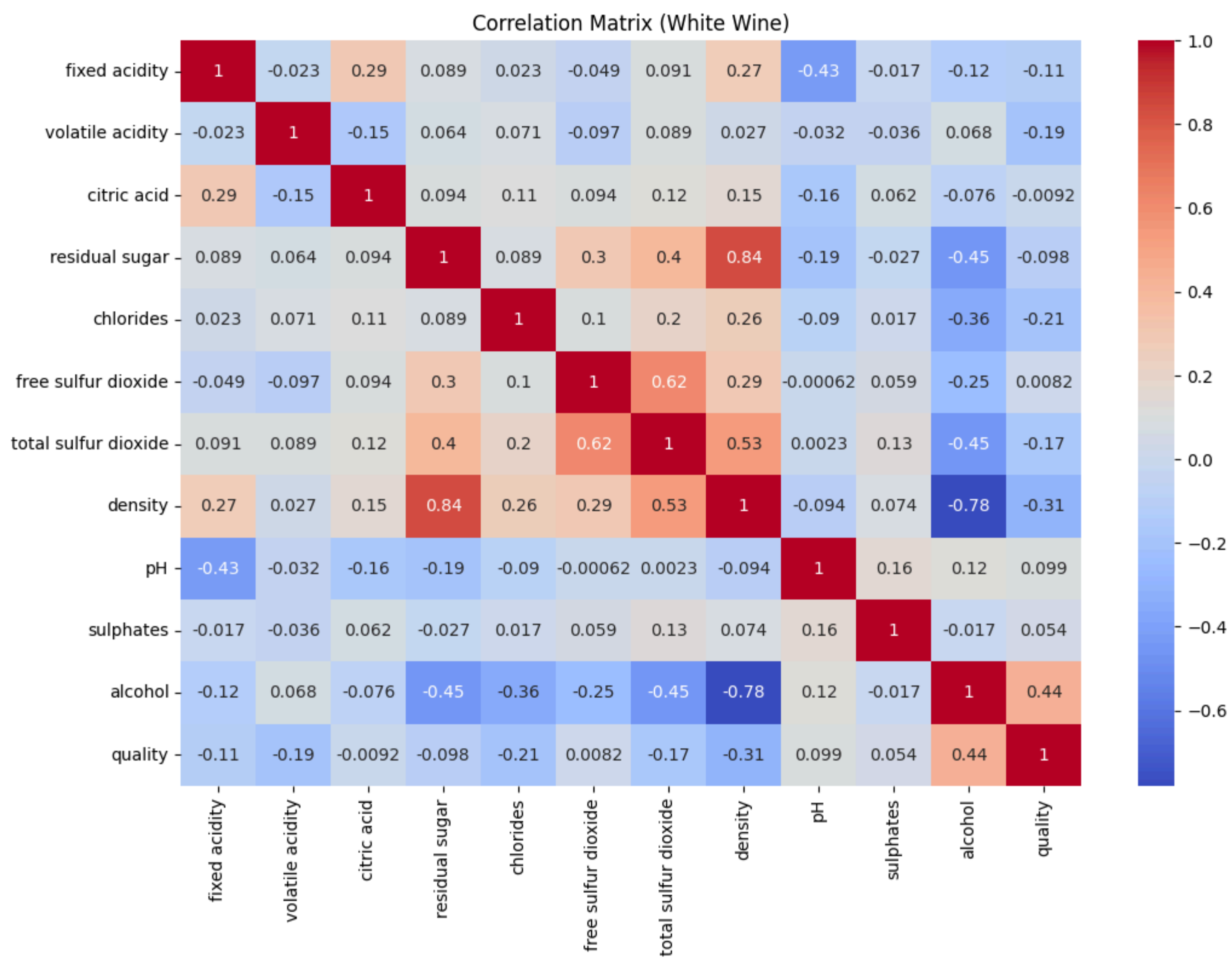
Hypothesis Formulation

- Hypothesis 1: There is a positive correlation between alcohol content and wine quality.
- Hypothesis 2: Higher levels of volatile acidity are associated with lower wine quality.

In [9]:
```
# Correlation matrix for red wine
plt.figure(figsize=(12, 8))
corr_matrix_red = red_wine_data.corr()
sns.heatmap(corr_matrix_red, annot=True, cmap="coolwarm")
plt.title('Correlation Matrix (Red Wine)')
plt.show()

# Correlation matrix for white wine
plt.figure(figsize=(12, 8))
corr_matrix_white = white_wine_data.corr()
sns.heatmap(corr_matrix_white, annot=True, cmap="coolwarm")
plt.title('Correlation Matrix (White Wine)')
plt.show()
```


Correlation Matrix (Red Wine)
fixed acidity 1 -0.26 0.67 0.11 0.094 -0.15 -0.11 0.67 -0.68 0.18 -0.062 0.12
volatile acidity -0.26 1 -0.55 0.0019 0.061 -0.011 0.076 0.022 0.23 -0.26 -0.2 -0.39
citric acid 0.67 -0.55 1 0.14 0.2 -0.061 0.036 0.36 -0.54 0.31 0.11 0.23
residual sugar 0.11 0.0019 0.14 1 0.056 0.19 0.2 0.36 -0.086 0.0055 0.042 0.014
chlorides 0.094 0.061 0.2 0.056 1 0.0056 0.047 0.2 -0.27 0.37 -0.22 -0.13
free sulfur dioxide -0.15 -0.011 -0.061 0.19 0.0056 1 0.67 -0.022 0.07 0.052 -0.069 -0.051
total sulfur dioxide -0.11 0.076 0.036 0.2 0.047 0.67 1 0.071 -0.066 0.043 -0.21 -0.19
density 0.67 0.022 0.36 0.36 0.2 -0.022 0.071 1 -0.34 0.15 -0.5 -0.17
pH -0.68 0.23 -0.54 -0.086 -0.27 0.07 -0.066 -0.34 1 -0.2 0.21 -0.058
sulphates 0.18 -0.26 0.31 0.0055 0.37 0.052 0.043 0.15 -0.2 1 0.094 0.25
alcohol -0.062 -0.2 0.11 0.042 -0.22 -0.069 -0.21 -0.5 0.21 0.094 1 0.48
quality 0.12 -0.39 0.23 0.014 -0.13 -0.051 -0.19 -0.17 -0.058 0.25 0.48 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Correlation Matrix (White Wine)
fixed acidity 1 -0.023 0.29 0.089 0.023 -0.049 0.091 0.27 -0.43 -0.017 -0.12 -0.11
volatile acidity -0.023 1 -0.15 0.064 0.071 -0.097 0.089 0.027 -0.032 -0.036 0.068 -0.19
citric acid 0.29 -0.15 1 0.094 0.11 0.094 0.12 0.15 -0.16 0.062 -0.076 -0.0092
residual sugar 0.089 0.064 0.094 1 0.089 0.3 0.4 0.84 -0.19 -0.027 -0.45 -0.098
chlorides 0.023 0.071 0.11 0.089 1 0.1 0.2 0.26 -0.09 0.017 -0.36 -0.21
free sulfur dioxide -0.049 -0.097 0.094 0.3 0.1 1 0.62 0.29 -0.00062 0.059 -0.25 0.0082
total sulfur dioxide 0.091 0.089 0.12 0.4 0.2 0.62 1 0.53 0.0023 0.13 -0.45 -0.17
density 0.27 0.027 0.15 0.84 0.26 0.29 0.53 1 -0.094 0.074 -0.78 -0.31
pH -0.43 -0.032 -0.16 -0.19 -0.09 -0.00062 0.0023 -0.094 1 0.16 0.12 0.099
sulphates -0.017 -0.036 0.062 -0.027 0.017 0.059 0.13 0.074 0.16 1 -0.017 0.054
alcohol -0.12 0.068 -0.076 -0.45 -0.36 -0.25 -0.45 -0.78 0.12 -0.017 1 0.44
quality -0.11 -0.19 -0.0092 -0.098 -0.21 0.0082 -0.17 -0.31 0.099 0.054 0.44 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

In [14]:
```
from scipy.stats import pearsonr, spearmanr

# Pearson correlation for red wine
pearson_corr, _ = pearsonr(red_wine_data['alcohol'], red_wine_data['quality'])
print(f'Pearson correlation between alcohol and quality (Red Wine): {pearson_c

# Pearson correlation for white wine
pearson_corr_white, _ = pearsonr(white_wine_data['alcohol'], white_wine_data['
print(f'Pearson correlation between alcohol and quality (White Wine): {pearson_
```

```
Pearson correlation between alcohol and quality (Red Wine): 0.4761663240011360
7
Pearson correlation between alcohol and quality (White Wine): 0.43557471546137
33
```

In [15]:
```
from scipy.stats import ttest_ind

# Example: Split red wine data into high alcohol (>10%) and low alcohol (<10%)
high_alcohol = red_wine_data[red_wine_data['alcohol'] > 10]
low_alcohol = red_wine_data[red_wine_data['alcohol'] <= 10]

# t-test
t_stat, p_value = ttest_ind(high_alcohol['quality'], low_alcohol['quality'])
print(f'T-test between high and low alcohol content: T-stat={t_stat}, P-value=
```

```
T-test between high and low alcohol content: T-stat=17.15894925897105, P-value
=1.0584977770451115e-60
```

In [16]:
```
# Pearson correlation between volatile acidity and quality for red and white w
# Red wine
vol_acid_quality_corr_red = red_wine_data['volatile acidity'].corr(red_wine_da
# White wine
vol_acid_quality_corr_white = white_wine_data['volatile acidity'].corr(white_w

(vol_acid_quality_corr_red, vol_acid_quality_corr_white)
```

Out[16]:
```
(-0.390557780264007, -0.19472296892113414)
```

Hypothesis 1: There is a positive correlation between alcohol content and wine quality.

- Red Wine: A moderate positive correlation (0.476) was found, meaning higher alcohol content is associated with better wine quality. The t-test confirms a significant difference in quality between high- and low-alcohol wines.
- White Wine: A moderate positive correlation (0.436) also supports the hypothesis that higher alcohol content improves quality.

Hypothesis 2: Higher levels of volatile acidity are associated with lower wine quality.

- Red Wine: A moderate negative correlation (-0.391) shows that higher volatile acidity leads to lower wine quality.
- White Wine: A weak negative correlation (-0.195) suggests volatile acidity has a smaller negative impact on white wine quality.

Both hypotheses are supported, with stronger effects in red wines for both alcohol and volatile acidity.

In [18]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

X = red_wine_data.drop('quality', axis=1)
y = red_wine_data['quality']
# Train/test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, randor

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
print(f'Mean Squared Error: {mse}')
```

```
Mean Squared Error: 0.3900251439639545
```

The linear regression model achieved a Mean Squared Error (MSE) of 0.39, indicating reasonably accurate predictions of wine quality based on chemical properties. While the model performs well, there is room for improvement through advanced models or feature engineering to further enhance accuracy. Overall, the model provides a good foundation for understanding the relationship between wine quality and its chemical features.