Wine Type Classification

About Dataset

This data set contains records related to red and white variants of the Portuguese Vinho Verde wine. It contains information from 1599 red wine samples and 4898 white wine samples. Input variables in the data set consist of the type of wine and metrics from objective tests, while the target/output variable is a numerical score based on sensory data from wine experts.

In [3]:
```
# IMPORT LIBRARIES
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

In [5]:
```
# LOAD THE DATASET
data = pd.read_csv('wine-quality-white-and-red.csv')
```

In [7]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   type                  6497 non-null   object 
 1   fixed acidity         6497 non-null   float64
 2   volatile acidity      6497 non-null   float64
 3   citric acid           6497 non-null   float64
 4   residual sugar        6497 non-null   float64
 5   chlorides             6497 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6497 non-null   float64
 10  sulphates             6497 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64  
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

In [9]:
```
data.shape
```

Out[9]: (6497, 13)

In [11]:
```
data.describe()
```

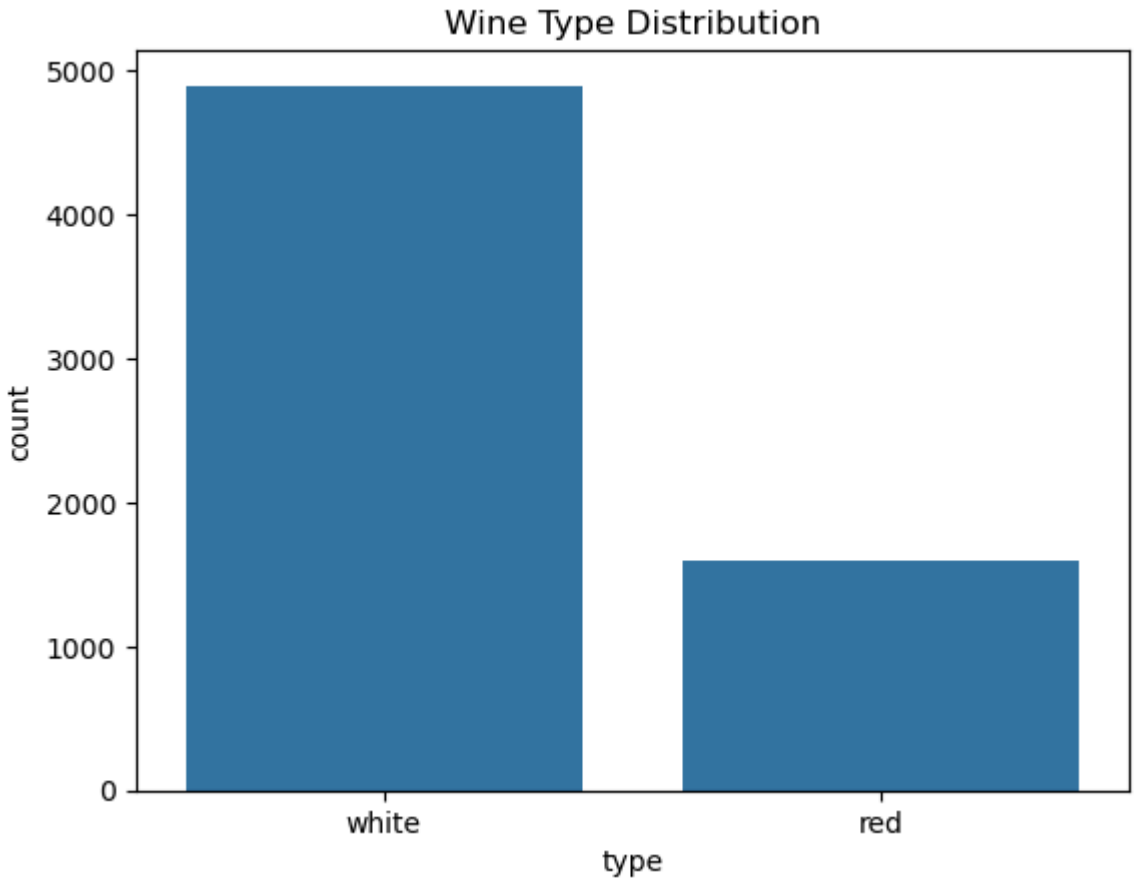
Out[11]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	6
count	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	
mean	7.215307	0.339666	0.318633	5.443235	0.056034	30.525319	
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.749400	
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	

Distribution of wine types

- This shows how many red and white wine samples are present.

In [16]:
```
sns.countplot(data=data, x='type')
plt.title('Wine Type Distribution')
plt.show()
```



- We can see that there are more white wine samples present in this dataset

Average quality by wine type

- Compares the mean quality score of red vs white wines.

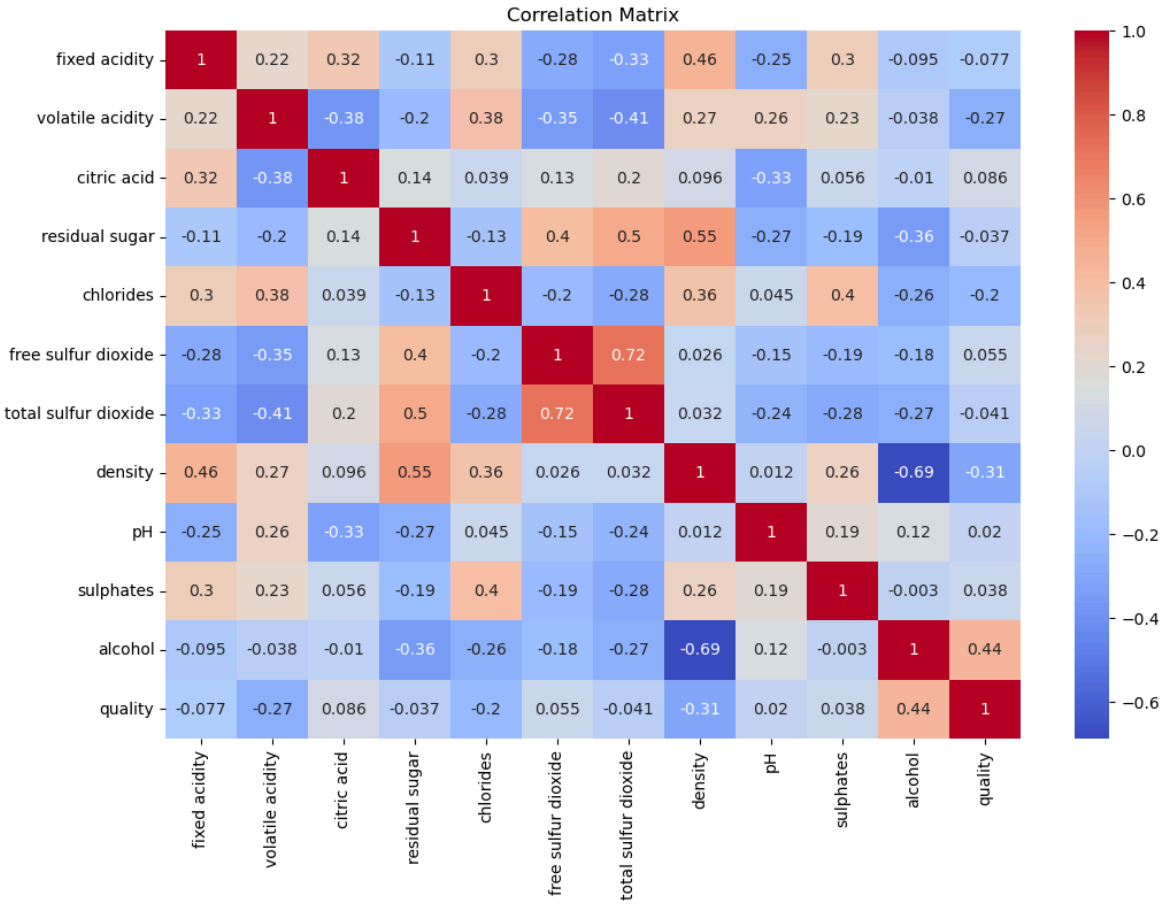
In [24]:
```
print(data.groupby('type')['quality'].mean())
```

```
type
red      5.636023
white    5.877909
Name: quality, dtype: float64
```

- We can see that both wines are almost similarly rated as there is no much difference.

Correlation matrix

In [32]:
```
corr = data.corr(numeric_only=True)
plt.figure(figsize=(12, 8))
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



- Alcohol has the strongest positive correlation with wine quality (+0.44).
- Volatile acidity has a moderate negative correlation with quality (-0.27) — higher acidity often means lower quality.
- Density is negatively correlated with quality (-0.31) — higher density may indicate poorer wines.
- Sulphates show a slight positive correlation with quality (+0.038) — could mildly enhance wine quality.

- Citric acid and quality have a slight positive correlation (+0.086) — more citric acid may indicate fresher wines.
- Fixed acidity, chlorides, free sulfur dioxide, and total sulfur dioxide have very weak or negligible correlations with quality.
- Residual sugar has almost no impact on quality (+0.037) — sweet wines aren't necessarily better.
- pH has almost zero correlation (+0.02) — acidity level alone doesn't determine quality.
- Total and free sulfur dioxide are strongly correlated with each other (+0.72) — likely due to chemical dependency.
- Density and alcohol are strongly negatively correlated (-0.69) — more alcohol means less dense wine.

Top features correlated with quality

- Sorts and shows which features most affect wine quality which can be crucial in building ML models.

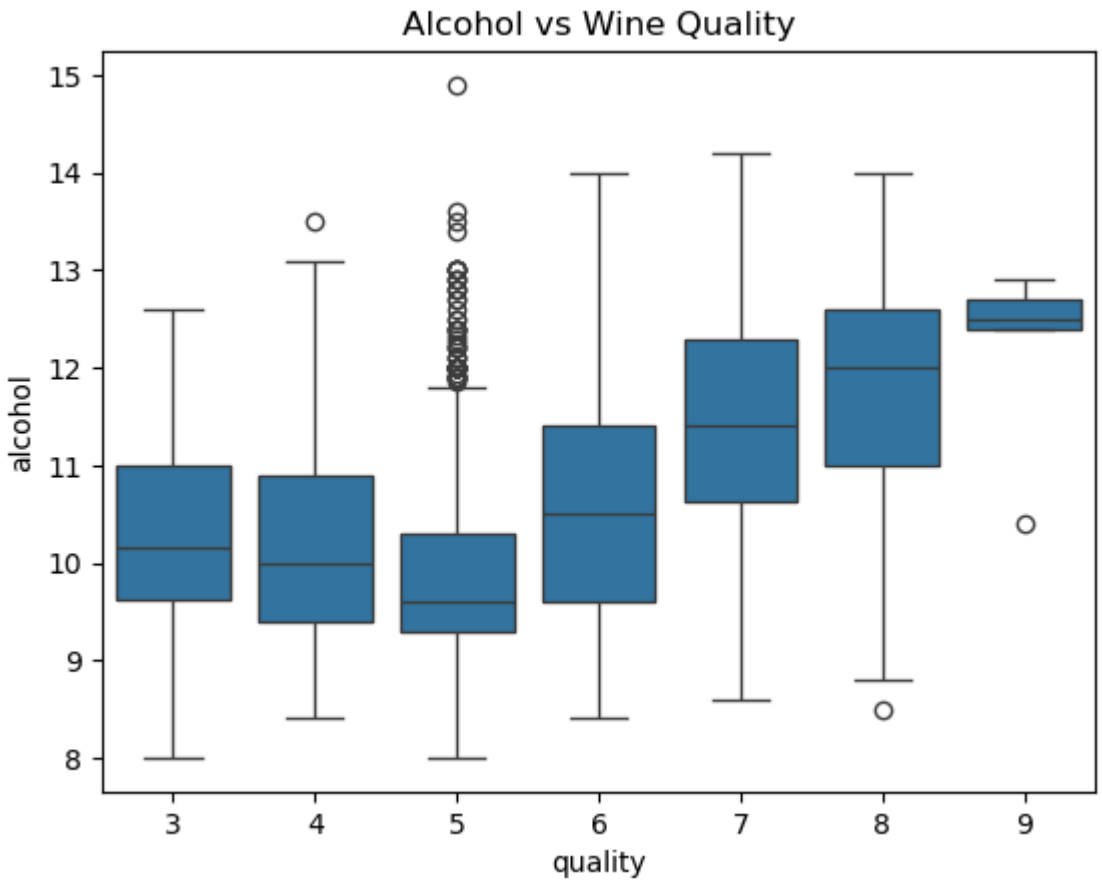
In [36]:
```
top_corr = corr['quality'].drop('quality').sort_values(ascending=False)
print("Top correlated features with quality:\n", top_corr)
```

```
Top correlated features with quality:
 alcohol                0.444319
citric acid             0.085532
free sulfur dioxide     0.055463
sulphates               0.038485
pH                      0.019506
residual sugar         -0.036980
total sulfur dioxide   -0.041385
fixed acidity          -0.076743
chlorides              -0.200666
volatile acidity       -0.265699
density                -0.305858
Name: quality, dtype: float64
```

Alcohol vs quality

- Checks how alcohol content varies across wine quality levels.

In [43]:
```
sns.boxplot(x='quality', y='alcohol', data=data)
plt.title('Alcohol vs Wine Quality')
plt.show()
```

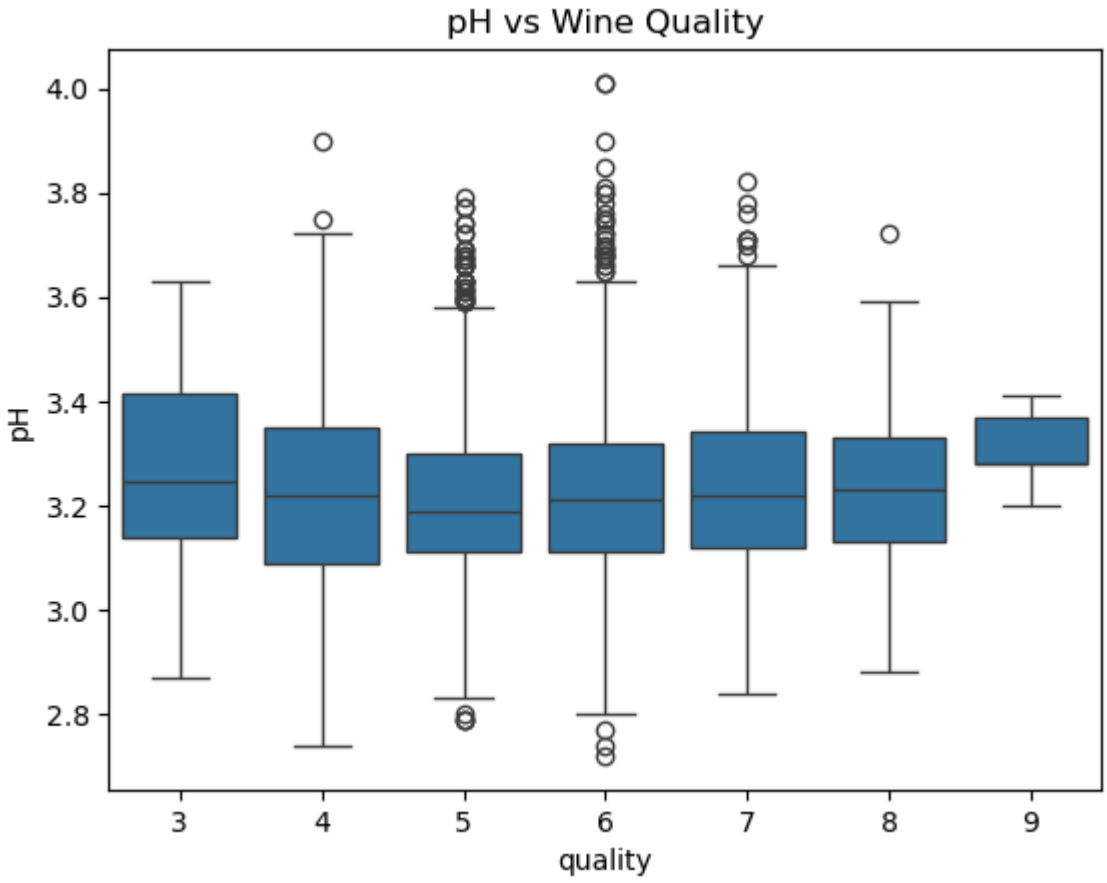


- Alcohol content is positively correlated with wine quality—higher alcohol often indicates a better-rated wine.
  - Higher-quality wines (especially scores 7–9) tend to have higher alcohol content.
  - Lower-quality wines (scores 3–5) generally have lower alcohol levels.
  - The median alcohol level increases steadily with quality score.
  - Quality 9 wines show very tight alcohol range, indicating consistency.
  - Outliers are more common in mid-quality (5–6) wines, showing greater variability.

pH vs quality

- Acidity (pH) can impact the taste and preservation of wine.

In [50]:
```
sns.boxplot(x='quality', y='pH', data=data)
plt.title('pH vs Wine Quality')
plt.show()
```

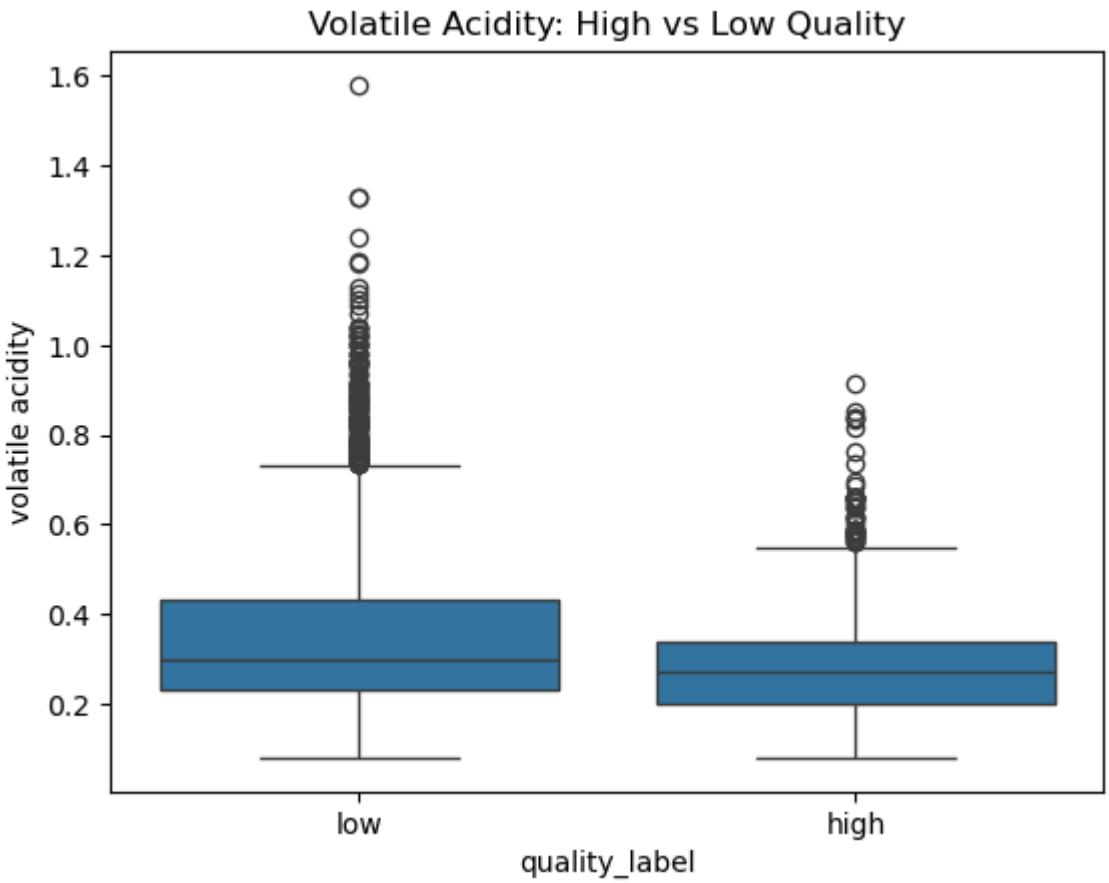


- pH (acidity) values are fairly consistent across all wine qualities.
- High-quality wines (8–9) tend to have slightly lower pH, meaning higher acidity, which can enhance taste and preservation.
- Outliers are mostly present in low to mid-quality wines (3–6), showing less acidity control.

Volatile Acidity: High vs Low Quality

- Volatile acidity gives a vinegary taste if high.

In [57]:
```
data['quality_label'] = data['quality'].apply(lambda x: 'high' if x >= 7 else 'l
sns.boxplot(x='quality_label', y='volatile acidity', data=data)
plt.title('Volatile Acidity: High vs Low Quality')
plt.show()
```

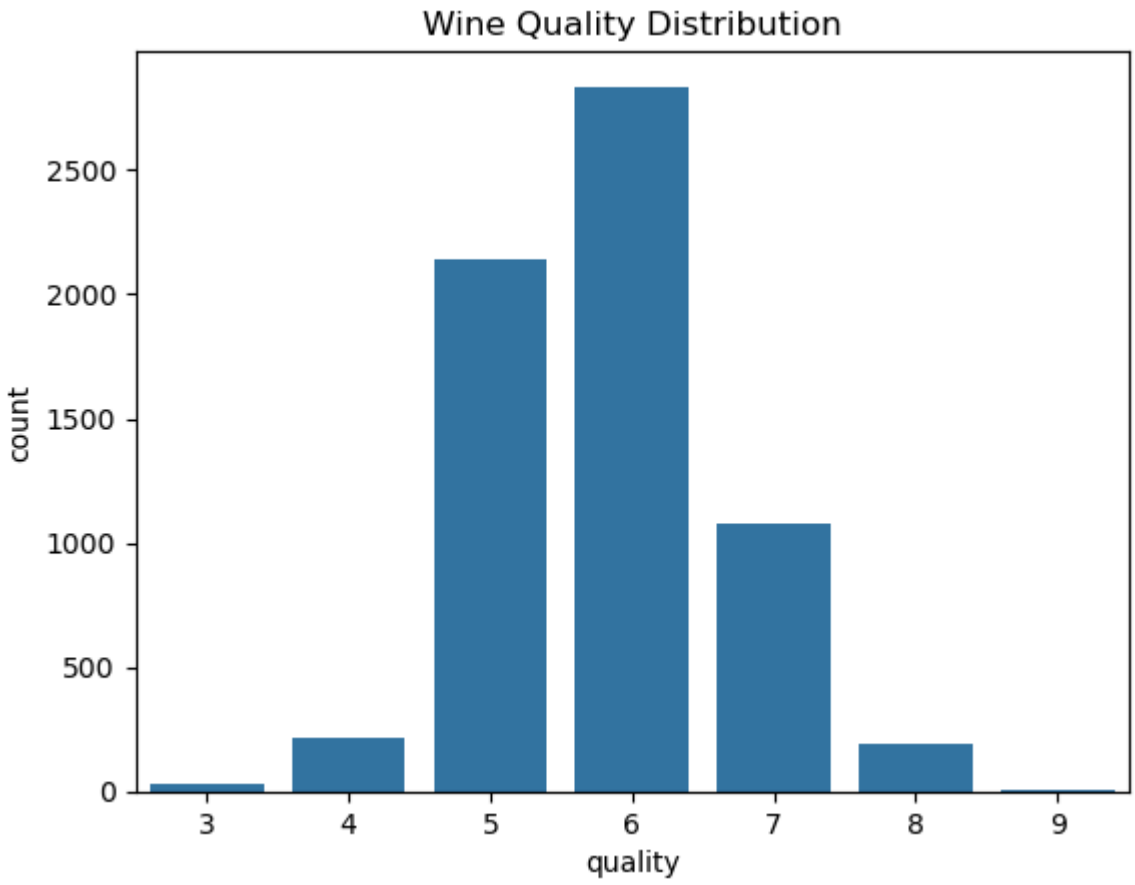


- Low-quality wines tend to have higher volatile acidity, which can create a vinegary taste.
- High-quality wines generally have lower and tighter ranges of volatile acidity.
- Volatile acidity is a key negative indicator of wine quality.

Quality distribution

- Shows how wine quality ratings are distributed.

In [62]:
```
sns.countplot(x='quality', data=data)
plt.title('Wine Quality Distribution')
plt.show()
```

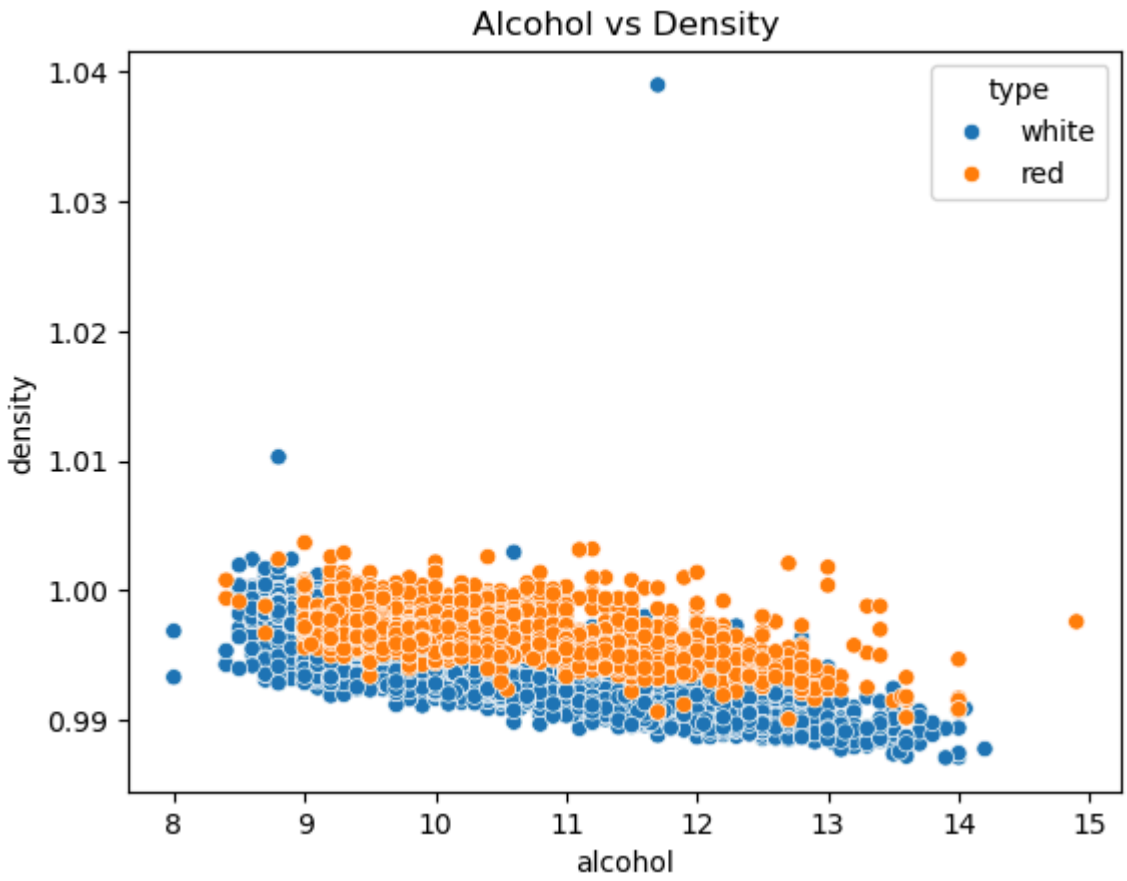


- Most wines are rated between 5 and 6, indicating average quality dominates.
- Very few wines are rated extremely high (8–9) or very low (3–4).

Alcohol vs Density

- Density is related to sugar and alcohol levels.

In [68]:
```
sns.scatterplot(data=data, x='alcohol', y='density', hue='type')
plt.title('Alcohol vs Density')
plt.show()
```

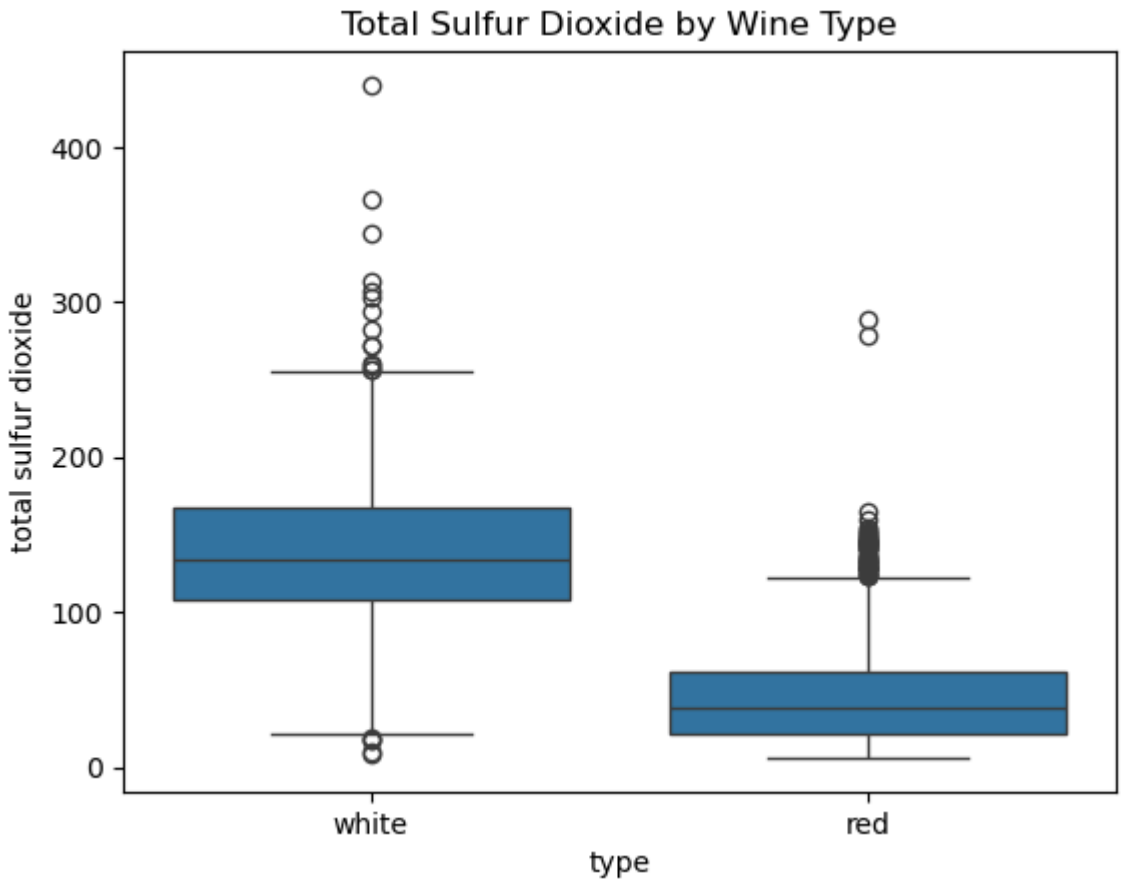


- As alcohol increases, density decreases—consistent with the chemistry of wine.
- Red wines (orange) are more spread across lower densities, while white wines (blue) cluster slightly higher.
- Helps identify that alcohol and sugar content are inversely related to density.

Total sulfur dioxide comparison

- Sulfur dioxide preserves wine but can affect taste.

In [74]:
```
sns.boxplot(x='type', y='total sulfur dioxide', data=data)
plt.title('Total Sulfur Dioxide by Wine Type')
plt.show()
```

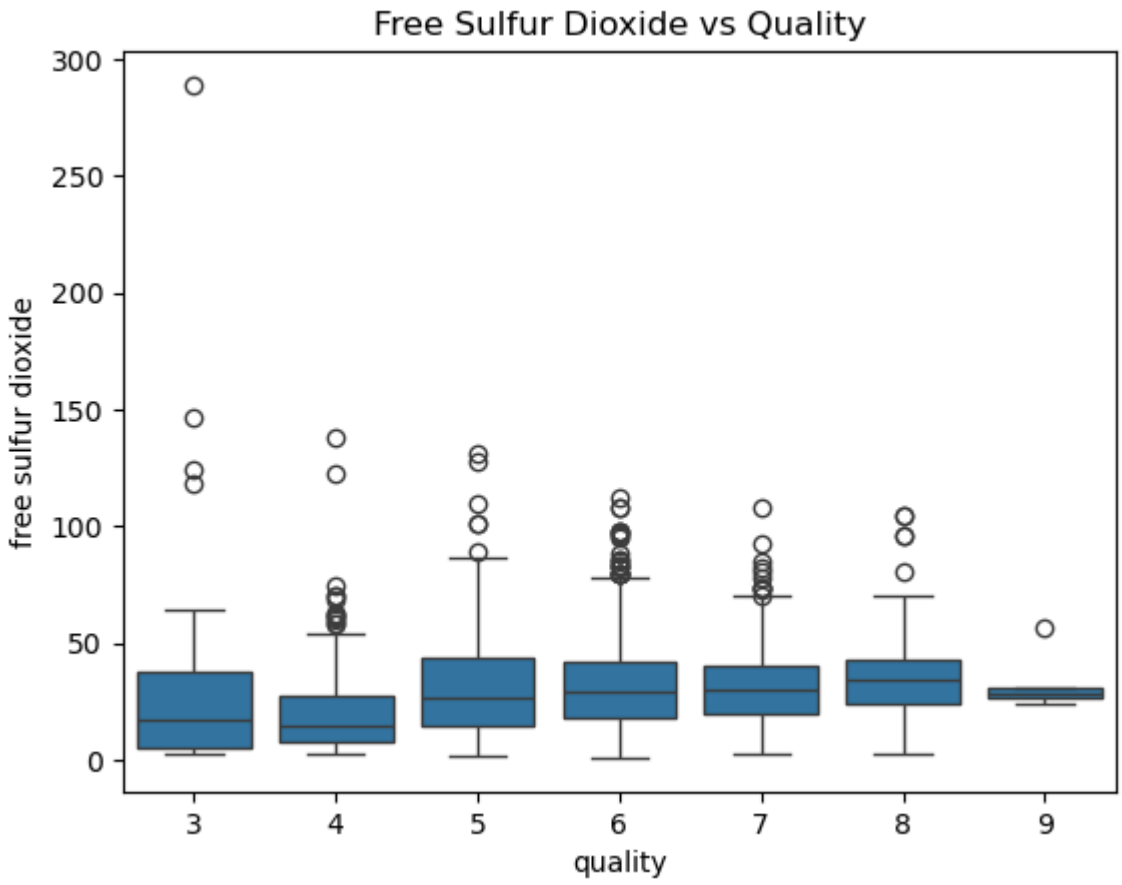


- White wines contain significantly more total sulfur dioxide than red wines.
- This is expected since white wines need more preservatives due to lower tannin content.
- Red wines show a tighter sulfur dioxide range, indicating better stability with lower preservatives.

Free sulfur dioxide vs quality

- Looks at how preservative level varies with wine quality.

In [77]:
```
sns.boxplot(x='quality', y='free sulfur dioxide', data=data)
plt.title('Free Sulfur Dioxide vs Quality')
plt.show()
```

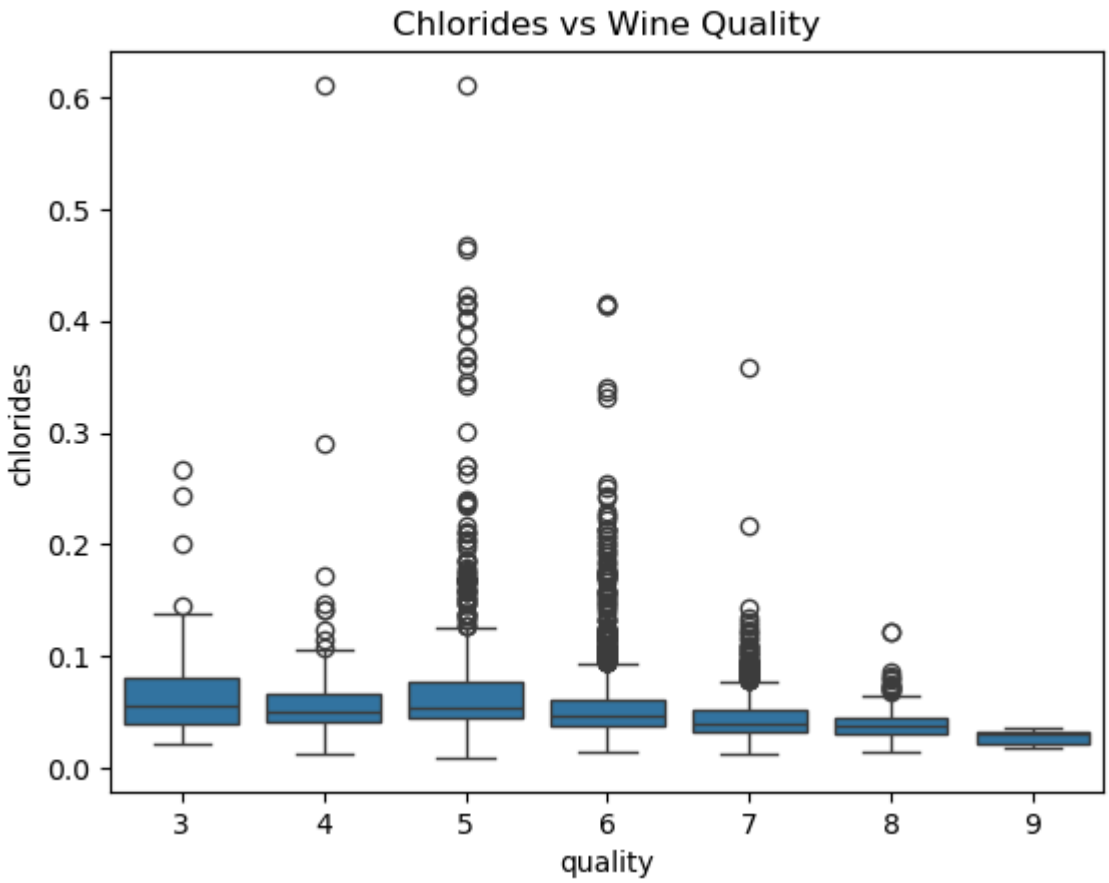


- Higher wine quality generally corresponds to slightly lower levels of free sulfur dioxide.
- Quality 9 wines have tightly packed low sulfur levels.
- Suggests that less preservative might be used in high-quality wines.

Chlorides vs quality

- Chloride content affects salinity/taste.

In [80]:
```
sns.boxplot(x='quality', y='chlorides', data=data)
plt.title('Chlorides vs Wine Quality')
plt.show()
```

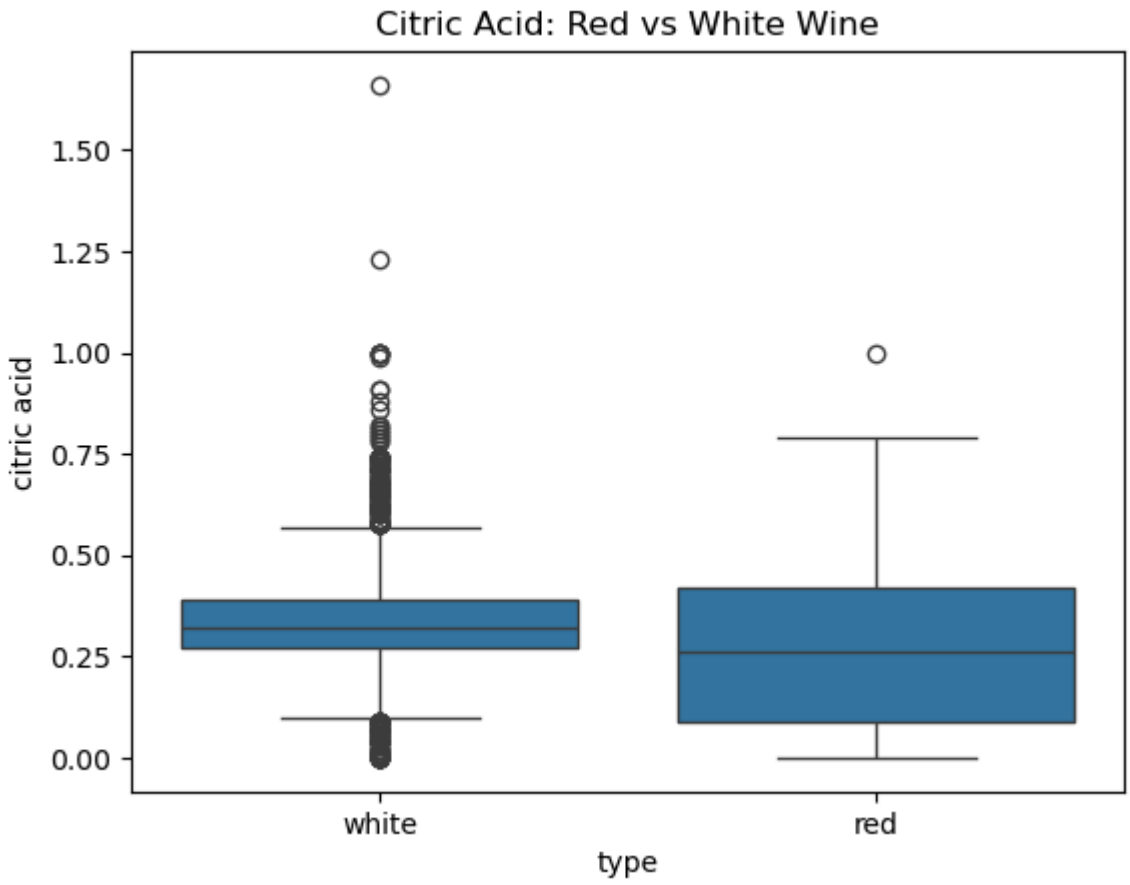


- Chloride content, which affects salinity, tends to be lower in higher-quality wines.
- Lower chlorides may contribute to better taste and higher quality perception.
- Indicates an inverse relationship between chlorides and quality.

Citric acid comparison

- Citric acid adds freshness/tang to wine.

In [85]:
```
sns.boxplot(x='type', y='citric acid', data=data)
plt.title('Citric Acid: Red vs White Wine')
plt.show()
```

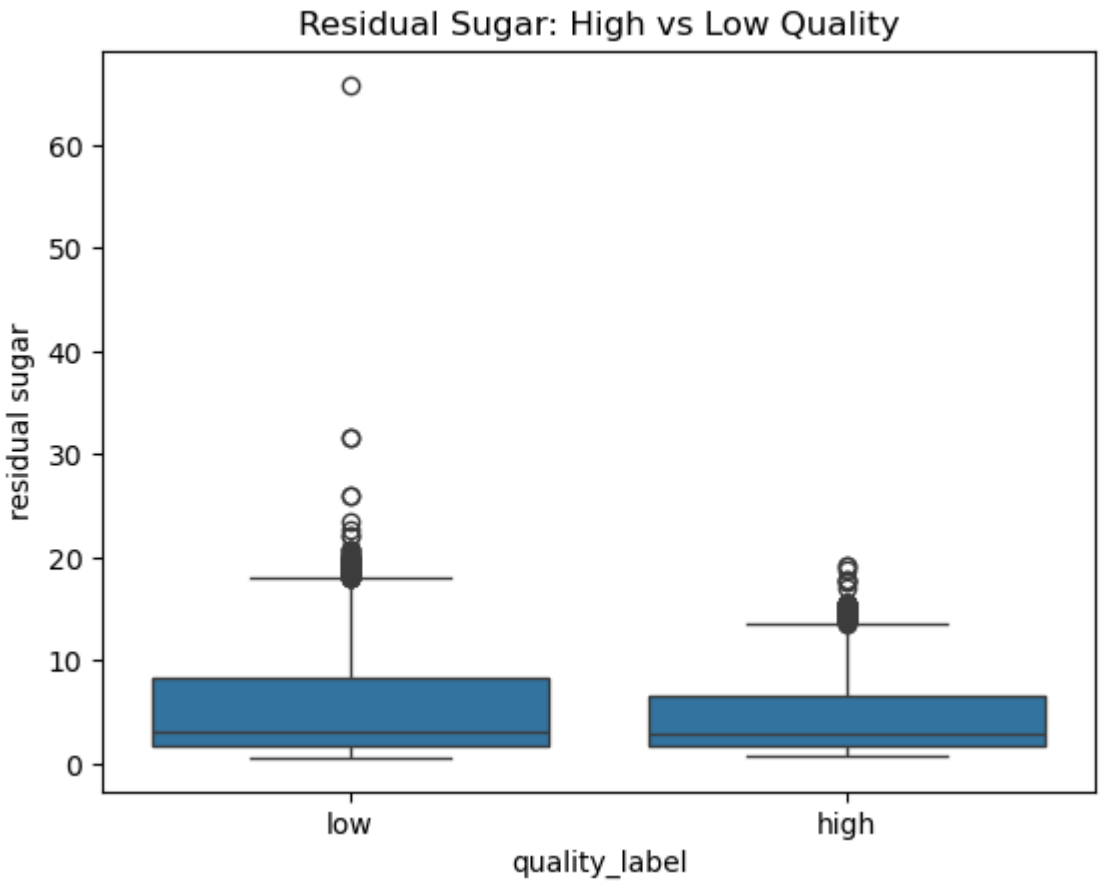


- Red wines show a broader distribution of citric acid values.
- White wines have a higher median citric acid level.
- Citric acid adds freshness and tang; hence, white wines are often perceived as tangier.

Residual sugar in high vs low quality

- Checks whether sweet wines are rated better.

In [91]:
```
sns.boxplot(x='quality_label', y='residual sugar', data=data)
plt.title('Residual Sugar: High vs Low Quality')
plt.show()
```

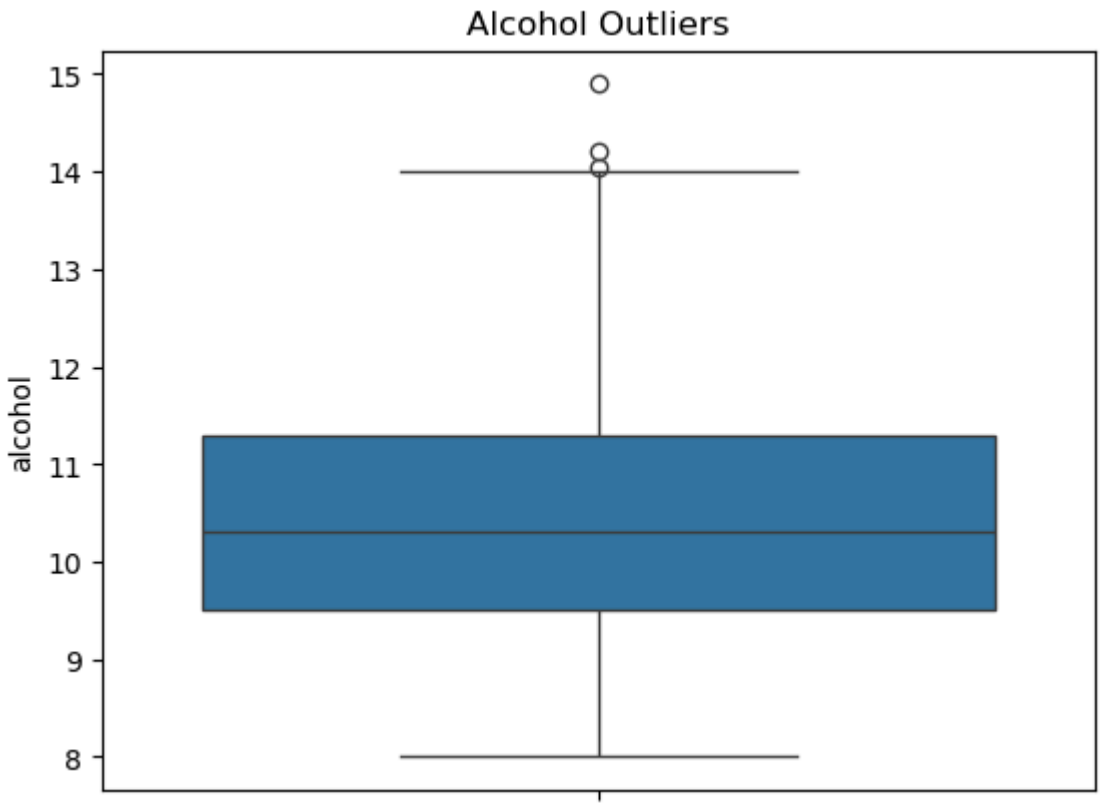


- Low-quality wines tend to have higher residual sugar.
- High-quality wines generally have less sugar, suggesting sweetness does not guarantee better ratings.
- Indicates that drier wines are perceived to be of higher quality.

Outlier detection in alcohol

- Identifies outlier wines with extreme alcohol content.

In [95]:
```
sns.boxplot(data=data, y='alcohol')
plt.title('Alcohol Outliers')
plt.show()
```



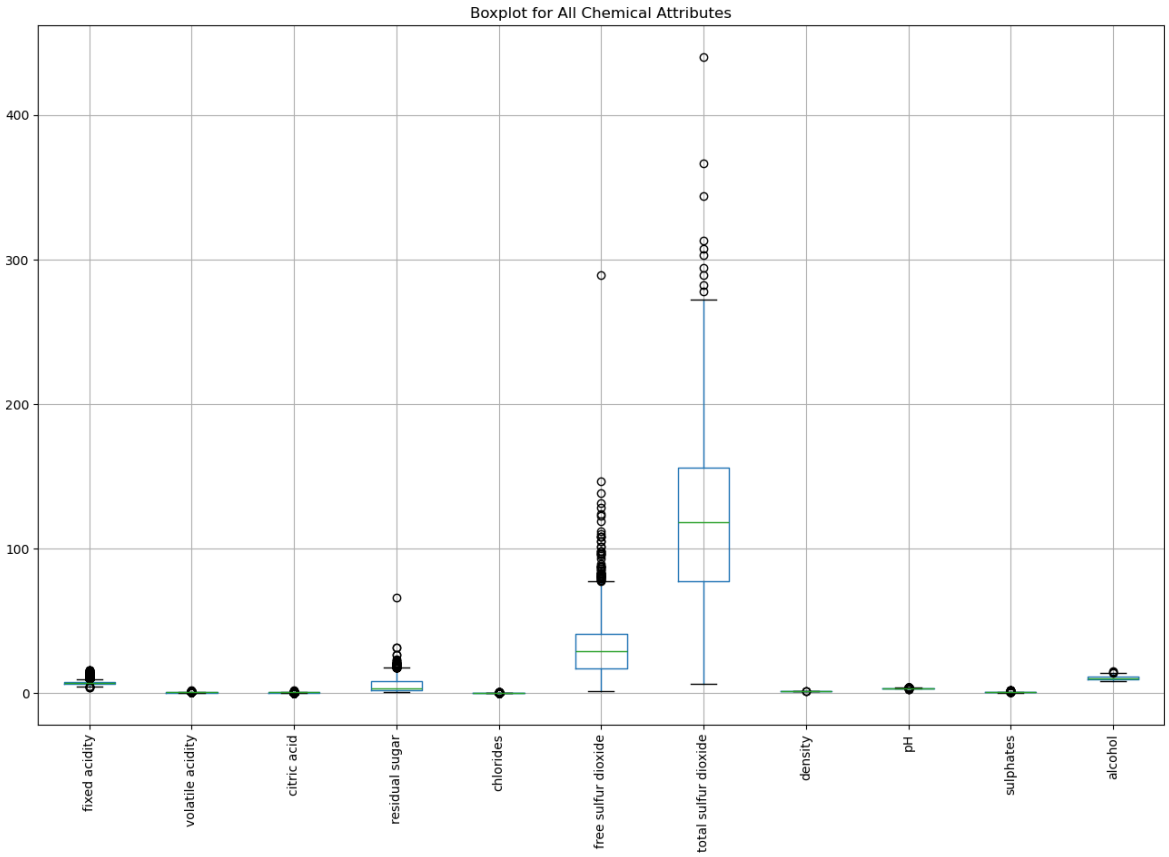
- Most wines have alcohol between ~9% and ~12.5%.
- Some outliers exist above 14%, which could indicate specialty or fortified wines.
- Useful for identifying extreme samples that may need special attention.

Boxplot for all features

- Visualizes distribution and range for all chemical features.

In [121…

```
plt.figure(figsize=(16, 10))
data.drop(['quality', 'type', 'quality_label'], axis=1).boxplot(rot=90)
plt.title("Boxplot for All Chemical Attributes")
plt.show()
```



- High variance in sulfur dioxide (free and total), and residual sugar.
- Most other features are within a tighter range.
- Strong case for normalization or outlier handling before modeling.

Machine Learning

In [101…
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
```

In [105…
```
X = data.drop(['quality', 'quality_label', 'type'], axis=1)
y = data['quality_label']
```

In [107…
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_
```

In [109…
```
clf = RandomForestClassifier()
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
```

In [111…
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

        high       0.79      0.53      0.64       253
         low       0.89      0.97      0.93      1047

    accuracy                           0.88      1300
   macro avg       0.84      0.75      0.78      1300
weighted avg       0.87      0.88      0.87      1300
```

Model Evaluation

- Overall Accuracy: 88%
- High-quality wine recall is low (0.53), meaning many high-quality wines are misclassified.
- Low-quality wine precision and recall are high → The model is biased toward predicting low-quality wines.