6.1-EDA_On_Wine_Dataset

July 16, 2025

0.1 Exploratory Data Analysis on the Red Wine Quality Dataset

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
import seaborn as sns
```

[2]:
```
df = pd.read_csv('..//0-Dataset//winequality-red.csv', sep =';')
```

[3]:
```
df.head()
```

[3]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

[4]:
```
## summary of data
df.info()
```

```
<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
```

[5]:
```
## Descriptive summary of the data set

df.describe()
```

[5]:

	fixed acidity	volatile acidity	citric acid	residual sugar	\
count	1599.000000	1599.000000	1599.000000	1599.000000	
mean	8.319637	0.527821	0.270976	2.538806	
std	1.741096	0.179060	0.194801	1.409928	
min	4.600000	0.120000	0.000000	0.900000	
25%	7.100000	0.390000	0.090000	1.900000	
50%	7.900000	0.520000	0.260000	2.200000	
75%	9.200000	0.640000	0.420000	2.600000	
max	15.900000	1.580000	1.000000	15.500000	

	chlorides	free sulfur dioxide	total sulfur dioxide	density	\
count	1599.000000	1599.000000	1599.000000	1599.000000	
mean	0.087467	15.874922	46.467792	0.996747	
std	0.047065	10.460157	32.895324	0.001887	
min	0.012000	1.000000	6.000000	0.990070	
25%	0.070000	7.000000	22.000000	0.995600	
50%	0.079000	14.000000	38.000000	0.996750	
75%	0.090000	21.000000	62.000000	0.997835	
max	0.611000	72.000000	289.000000	1.003690	

	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000
mean	3.311113	0.658149	10.422983	5.636023
std	0.154386	0.169507	1.065668	0.807569
min	2.740000	0.330000	8.400000	3.000000

```
25%       3.210000      0.550000      9.500000      5.000000
50%       3.310000      0.620000     10.200000      6.000000
75%       3.400000      0.730000     11.100000      6.000000
max       4.010000      2.000000     14.900000      8.000000
```

[6]:
```
sh = df.shape
print(f"Shape of dataset with duplicate records: {sh}")
```

```
Shape of dataset with duplicate records: (1599, 12)
```

[7]:
```
df.columns
```

[7]:
```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

[8]:
```
df['quality'].unique()
```

[8]:
```
array([5, 6, 7, 4, 8, 3], dtype=int64)
```

[9]:
```
## checking missing values in dataset

df.isnull().sum()
```

[9]:
```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
```

[10]:
```
## Duplicate  records

df[df.duplicated()]
```

[10]:
```
      fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
4               7.4             0.700         0.00            1.90      0.076
11              7.5             0.500         0.36            6.10      0.071
27              7.9             0.430         0.21            1.60      0.106
40              7.3             0.450         0.36            5.90      0.074
```

```
65              7.2             0.725         0.05            4.65      0.086
…               …               …             …               …         …
1563            7.2             0.695         0.13            2.00      0.076
1564            7.2             0.695         0.13            2.00      0.076
1567            7.2             0.695         0.13            2.00      0.076
1581            6.2             0.560         0.09            1.70      0.053
1596            6.3             0.510         0.13            2.30      0.076

      free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
4                    11.0                  34.0  0.99780  3.51       0.56
11                   17.0                 102.0  0.99780  3.35       0.80
27                   10.0                  37.0  0.99660  3.17       0.91
40                   12.0                  87.0  0.99780  3.33       0.83
65                    4.0                  11.0  0.99620  3.41       0.39
…                     …                     …       …   …          …
1563                 12.0                  20.0  0.99546  3.29       0.54
1564                 12.0                  20.0  0.99546  3.29       0.54
1567                 12.0                  20.0  0.99546  3.29       0.54
1581                 24.0                  32.0  0.99402  3.54       0.60
1596                 29.0                  40.0  0.99574  3.42       0.75

      alcohol  quality
4         9.4        5
11       10.5        5
27        9.5        5
40       10.5        5
65       10.9        5
…         …          …
1563     10.1        5
1564     10.1        5
1567     10.1        5
1581     11.3        5
1596     11.0        6

[240 rows x 12 columns]
```

[11]:
```
## Removing the duplicates

df.drop_duplicates(inplace=True)
df.shape
```

[11]:
```
(1359, 12)
```

[12]:
```
print(f"Shape of dataset without duplicate records: {df.shape}")
```

```
Shape of dataset without duplicate records: (1359, 12)
```

[13]:
```
## corelation between the features
df.corr()
```

[13]:

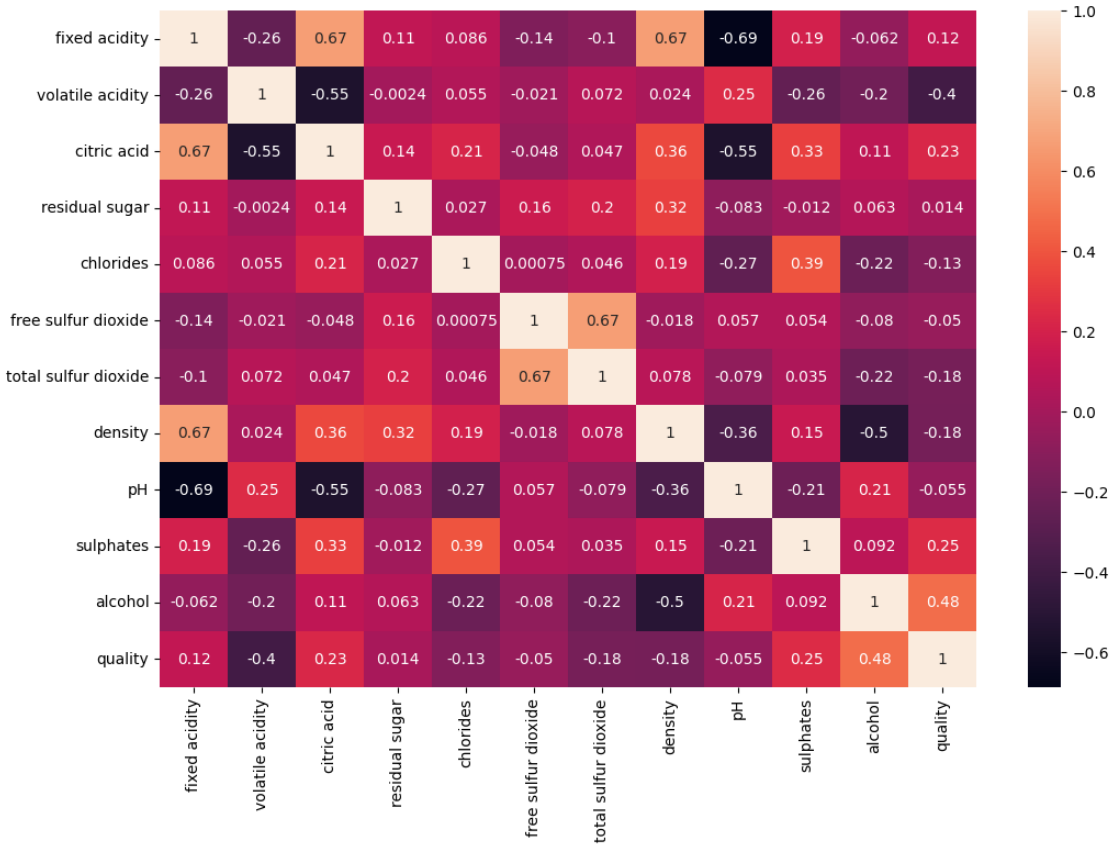
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.000000	-0.255124	0.667437	0.111025	0.085886	-0.140580	-0.103777	0.670195	-0.686685	0.190269	-0.061596	0.119024
volatile acidity	-0.255124	1.000000	-0.551248	-0.002449	0.055154	-0.020945	0.071701	0.023943	0.247111	-0.256948	-0.197812	-0.395214
citric acid	0.667437	-0.551248	1.000000	0.143892	0.210195	-0.048004	0.047358	0.357962	-0.550310	0.326062	0.105108	0.228057
residual sugar	0.111025	-0.002449	0.143892	1.000000	0.026656	0.160527	0.201038	0.324522	-0.083143	-0.011837	0.063281	0.013640
chlorides	0.085886	0.055154	0.210195	0.026656	1.000000	0.000749	0.045773	0.193592	-0.270893	0.394557	-0.223824	-0.130988
free sulfur dioxide	-0.140580	-0.020945	-0.048004	0.160527	0.000749	1.000000	0.667246	-0.018071	0.056631	0.054126	-0.080125	-0.050463
total sulfur dioxide	-0.103777	0.071701	0.047358	0.201038	0.045773	0.667246	1.000000	0.078141	-0.079257	0.035291	-0.217829	-0.177855
density	0.670195	0.023943	0.357962	0.324522	0.193592	-0.018071	0.078141	1.000000	-0.355617	0.146036	-0.504995	-0.184252
pH	-0.686685	0.247111	-0.550310	-0.083143	-0.270893	0.056631	-0.079257	-0.355617	1.000000	-0.214134	0.213418	-0.055245
sulphates	0.190269	-0.256948	0.326062	-0.011837	0.394557	0.054126	0.035291	0.146036	-0.214134	1.000000	0.091621	0.248835
alcohol	-0.061596	-0.197812	0.105108	0.063281	-0.223824	-0.080125	-0.217829	-0.504995	0.213418	0.091621	1.000000	0.228057
quality	0.119024	-0.395214	0.228057	0.013640	-0.130988	-0.050463	-0.177855	-0.184252	-0.055245	0.248835	0.228057	1.000000

```
volatile acidity     -0.197812 -0.395214
citric acid           0.105108  0.228057
residual sugar        0.063281  0.013640
chlorides            -0.223824 -0.130988
free sulfur dioxide  -0.080125 -0.050463
total sulfur dioxide -0.217829 -0.177855
density              -0.504995 -0.184252
pH                    0.213418 -0.055245
sulphates             0.091621  0.248835
alcohol               1.000000  0.480343
quality               0.480343  1.000000
```

0.2 Visualization of Dataset

[14]:
```
plt.figure(figsize=(12,8))
sns.heatmap(df.corr(), annot=True)
```

[14]: <Axes: >

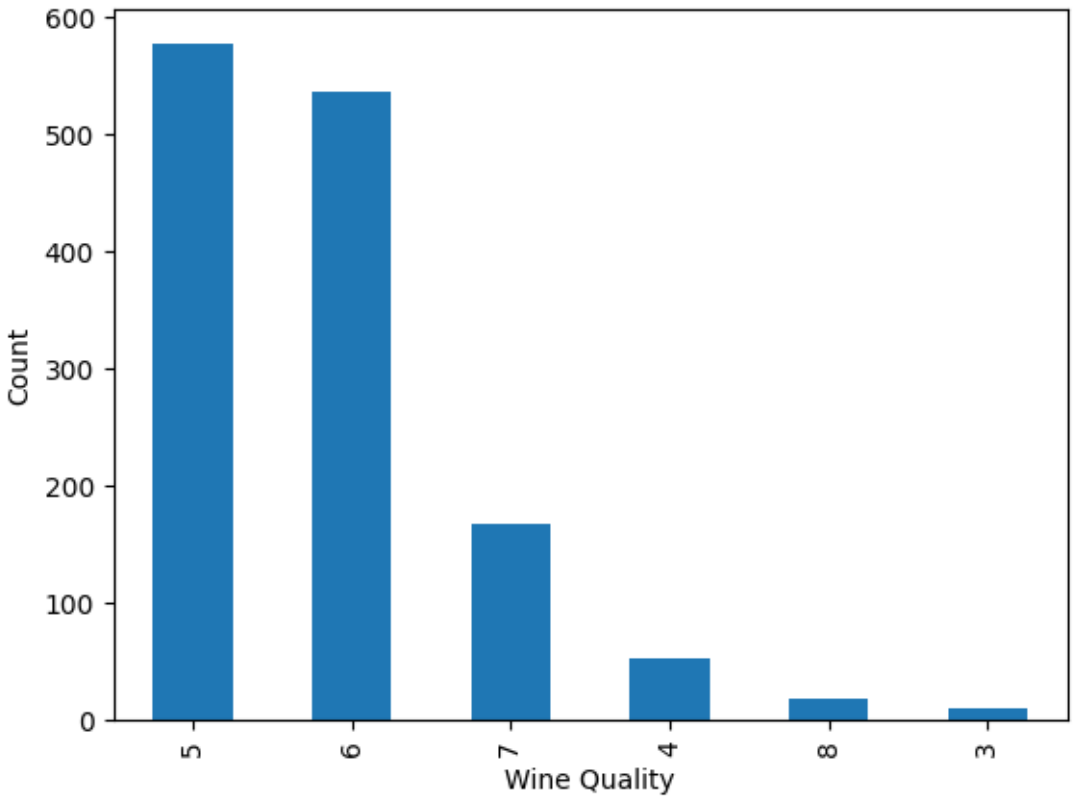


[15]:
```
## it is an imbalance dataset
df.quality.value_counts()
```

[15]:
```
quality
5    577
6    535
7    167
4     53
8     17
3     10
Name: count, dtype: int64
```

[16]:
```
df.quality.value_counts().plot(kind = 'bar')
plt.xlabel('Wine Quality')
plt.ylabel('Count')
```

[16]:
```
Text(0, 0.5, 'Count')
```



[17]:
```
df.head()
```

[17]:

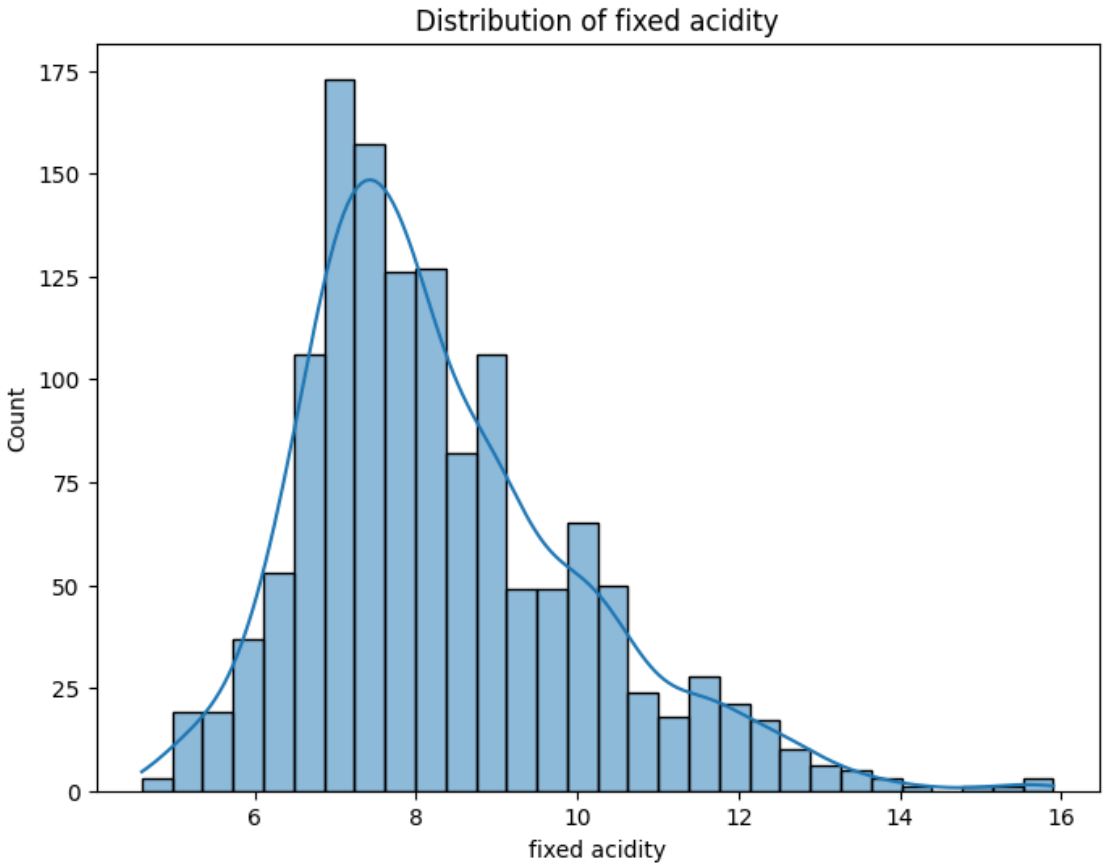
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	7.4	0.70	0.00	1.9	0.076	
1	7.8	0.88	0.00	2.6	0.098	
2	7.8	0.76	0.04	2.3	0.092	
3	11.2	0.28	0.56	1.9	0.075	
5	7.4	0.66	0.00	1.8	0.075	

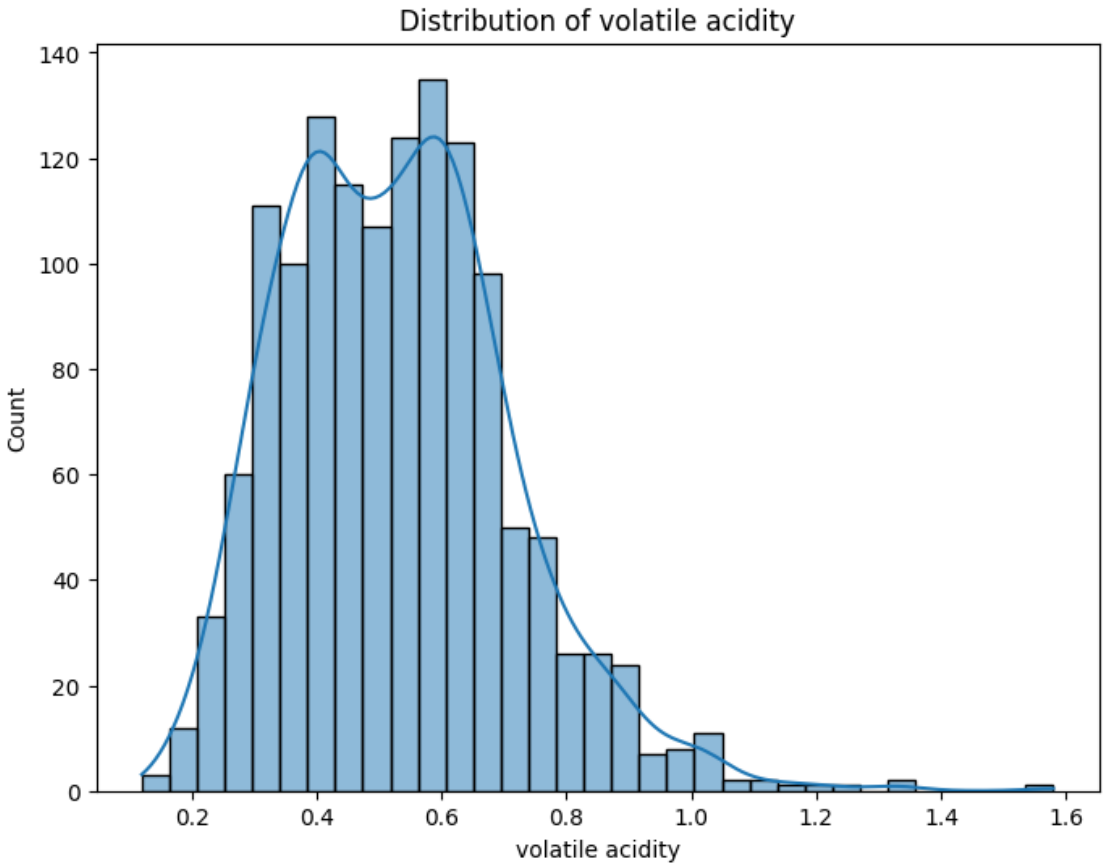
	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	11.0	34.0	0.9978	3.51	0.56	
1	25.0	67.0	0.9968	3.20	0.68	
2	15.0	54.0	0.9970	3.26	0.65	
3	17.0	60.0	0.9980	3.16	0.58	
5	13.0	40.0	0.9978	3.51	0.56	

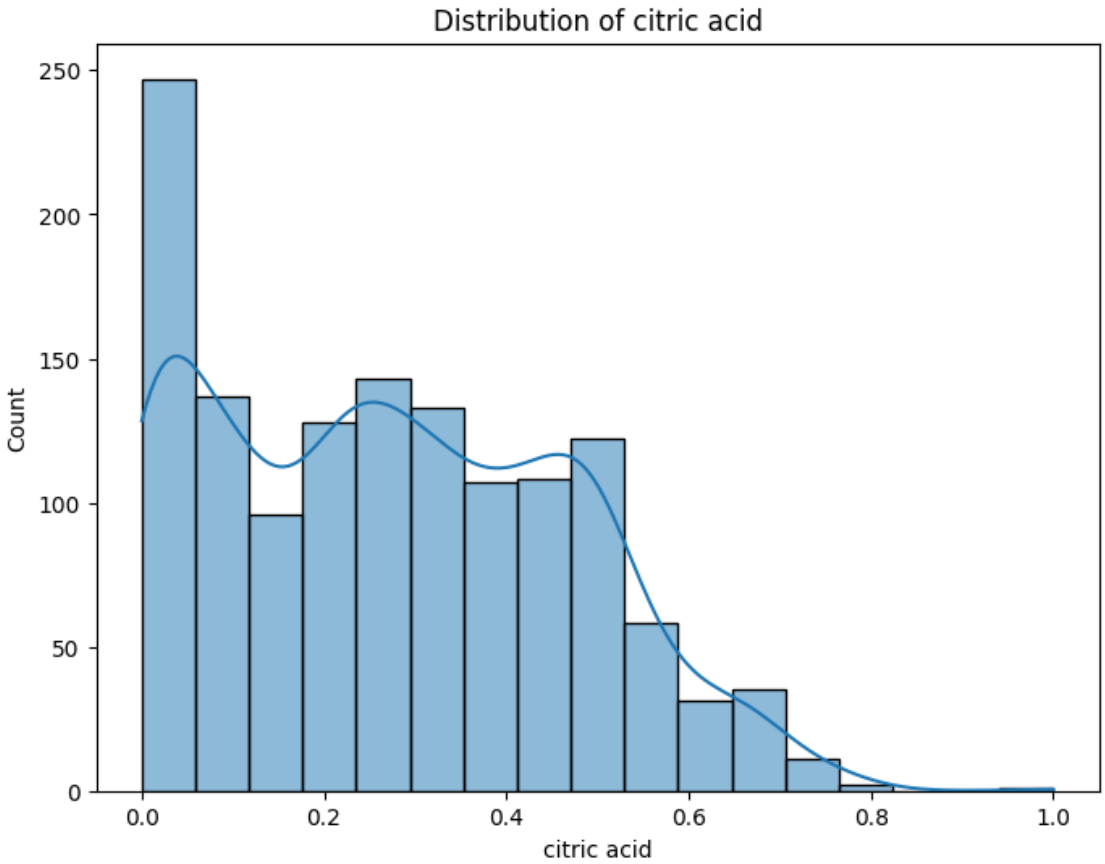
	alcohol	quality
0	9.4	5
1	9.8	5
2	9.8	5
3	9.8	6
5	9.4	5

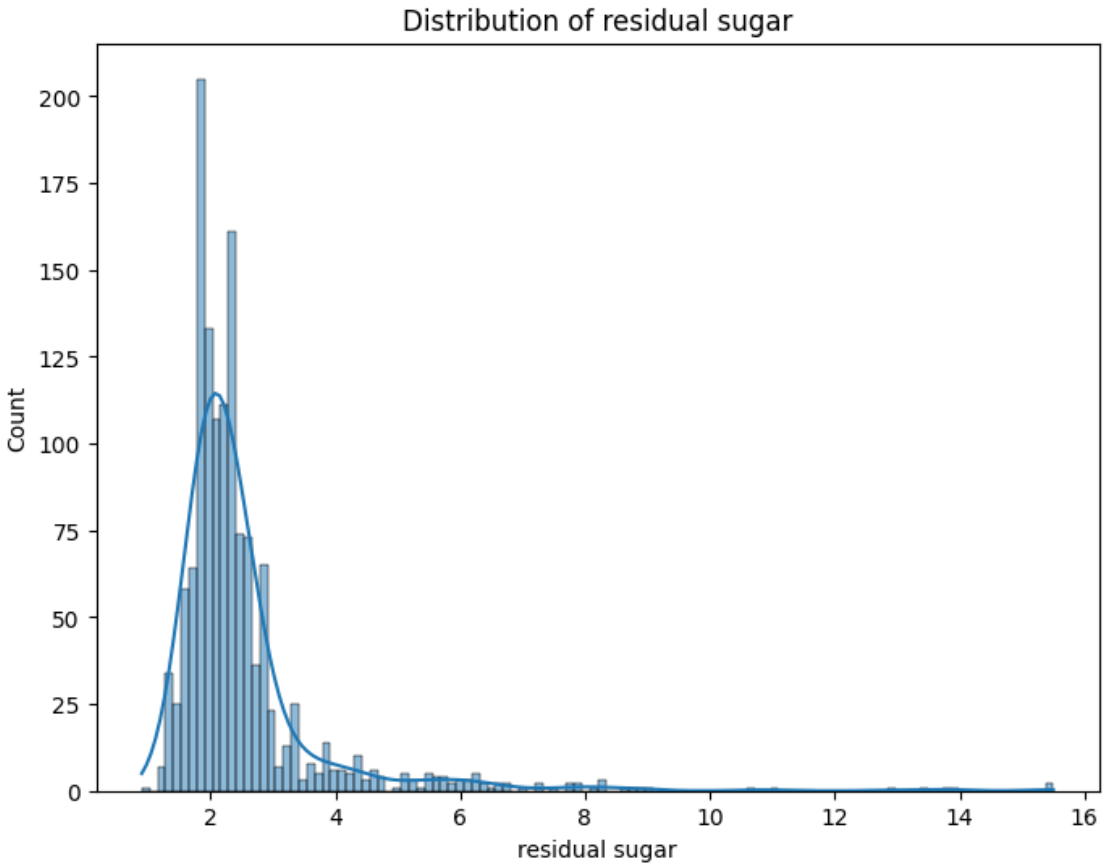
0.3 visualize the columns distribution

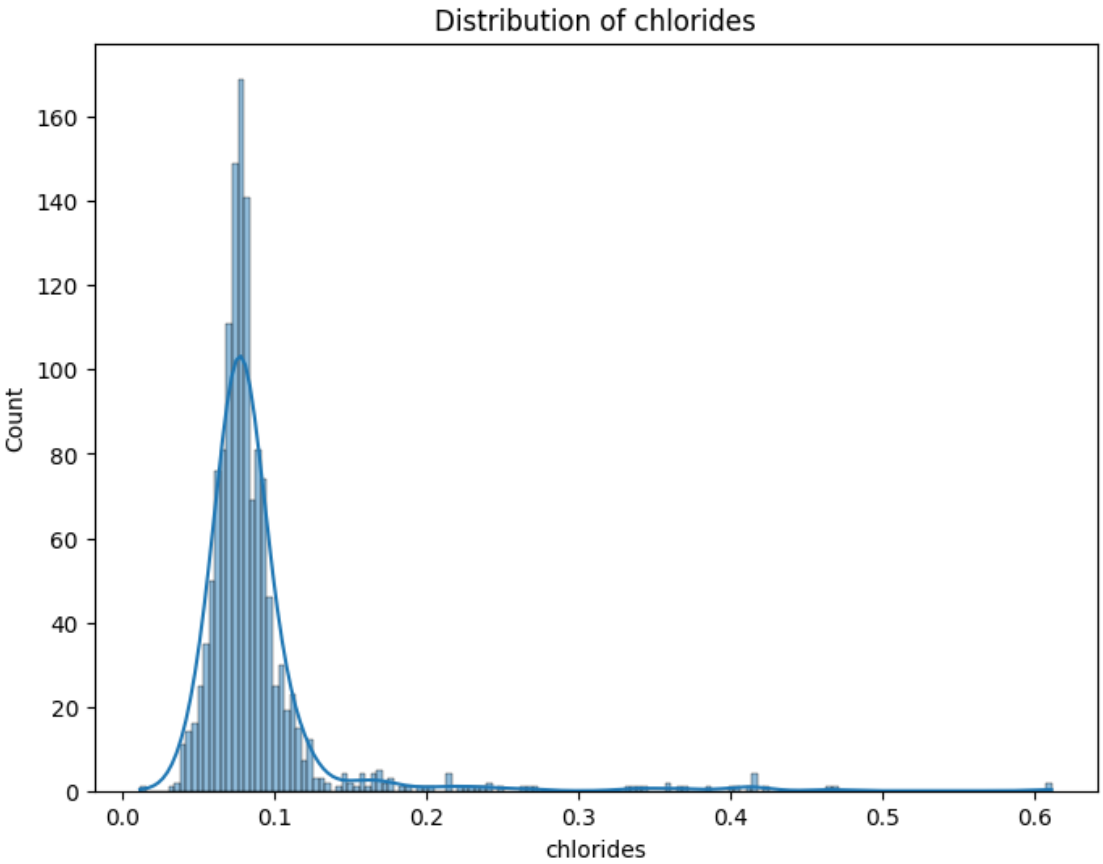
[ ]:
```
for column in df.columns:
    plt.figure(figsize=(8, 6))
    sns.histplot(df[column], kde=True)
    plt.title(f'Distribution of {column}')
    plt.show()
    plt.close()
```

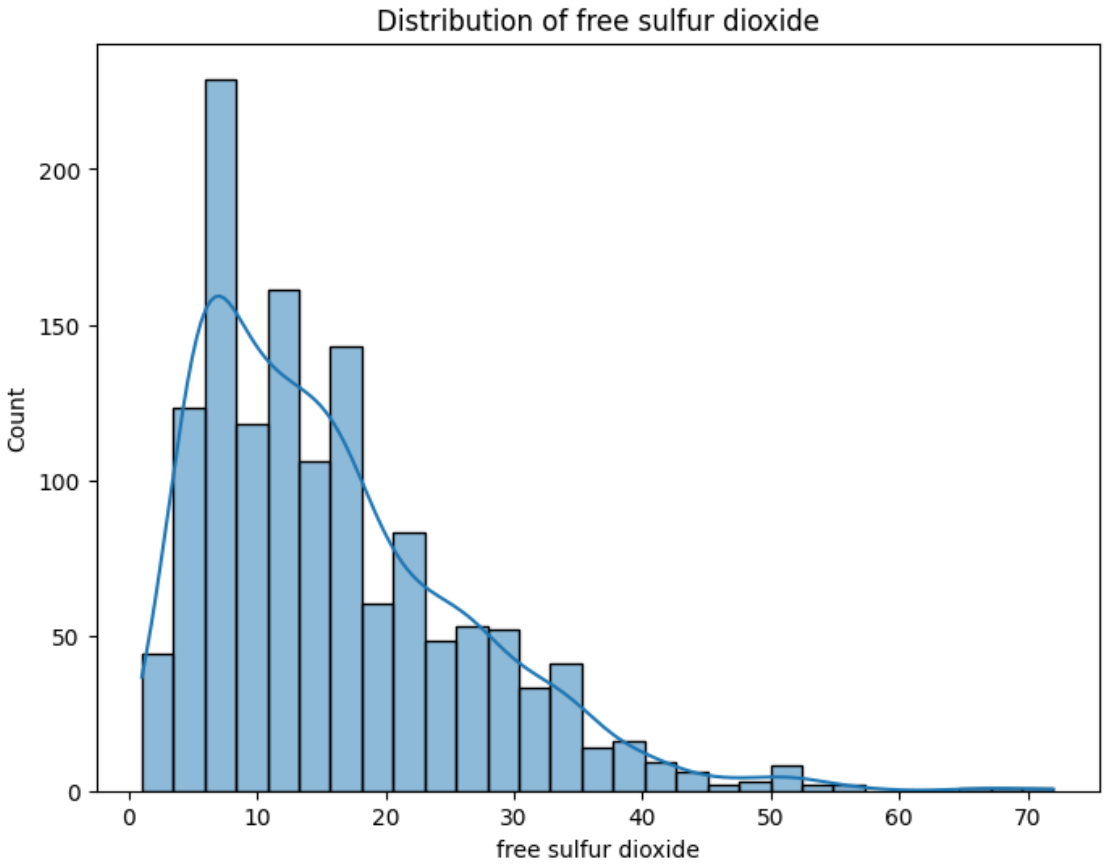

Distribution of fixed acidity
Count
fixed acidity
0
25
50
75
100
125
150
175
6
8
10
12
14
16

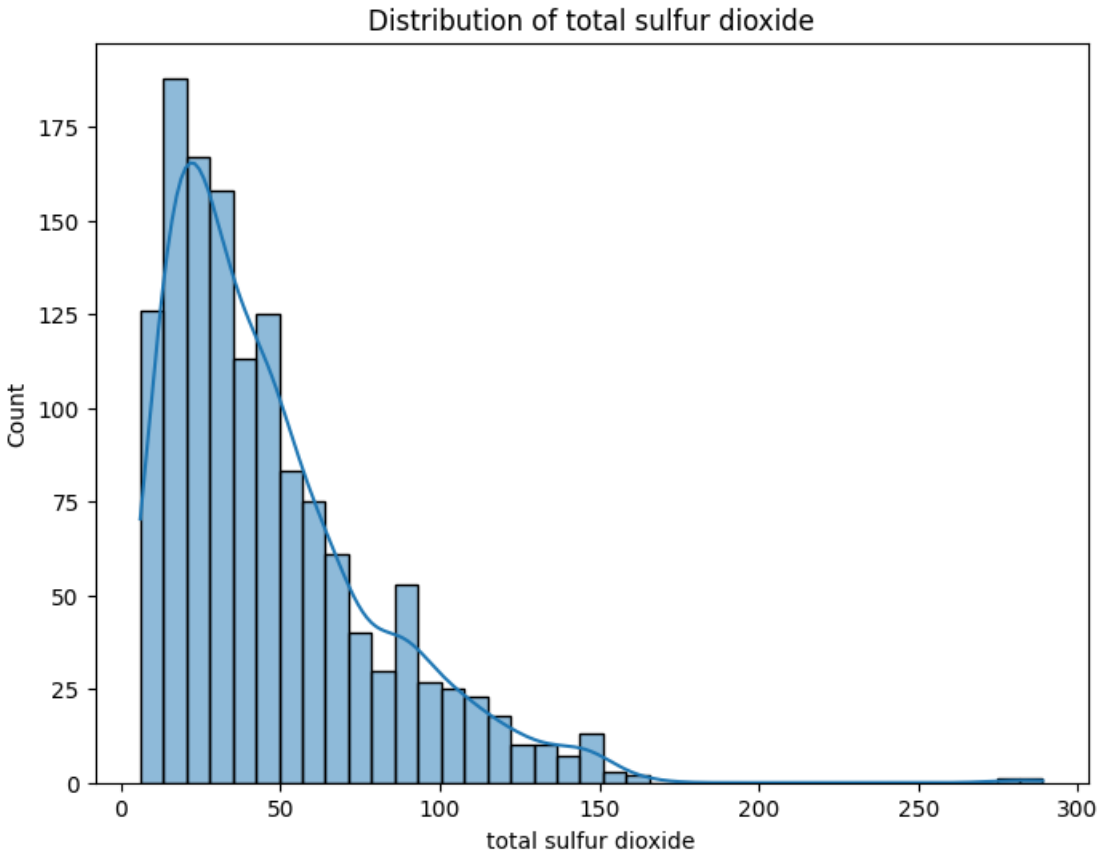

Distribution of volatile acidity
Count
volatile acidity
0
20
40
60
80
100
120
140
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6

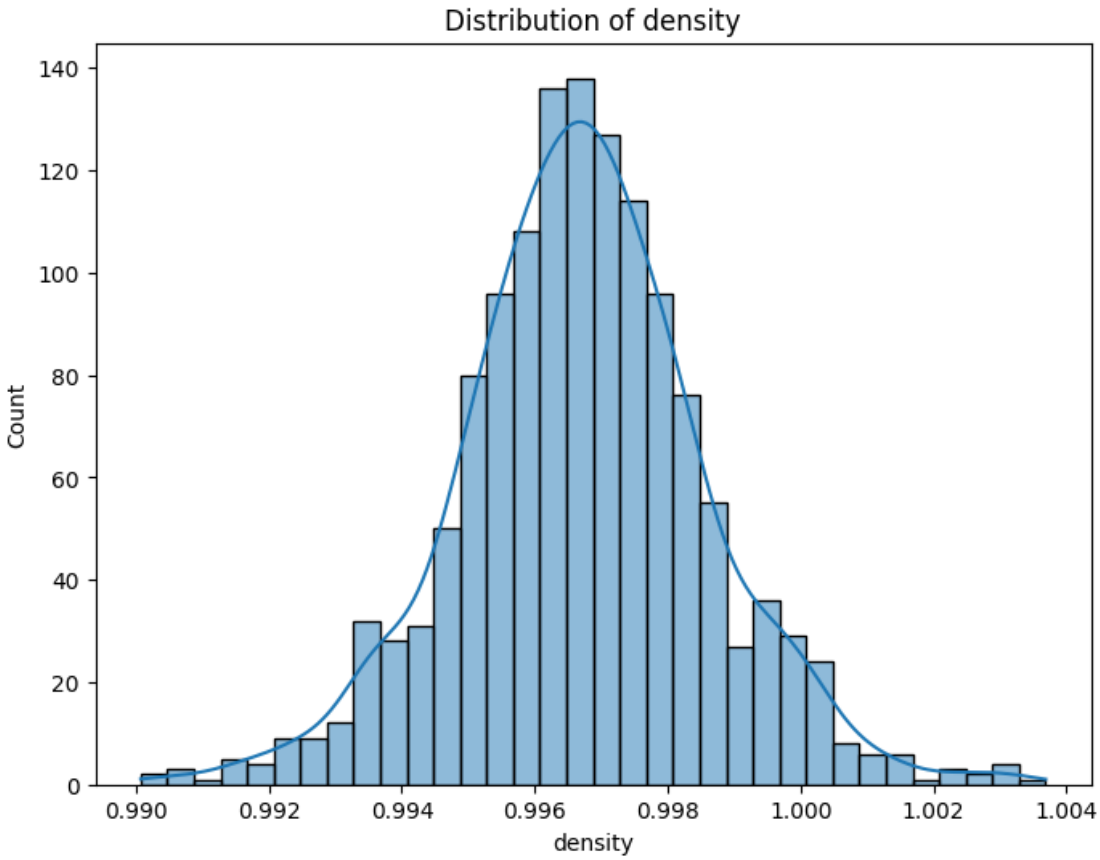

Distribution of citric acid
Count
citric acid
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0

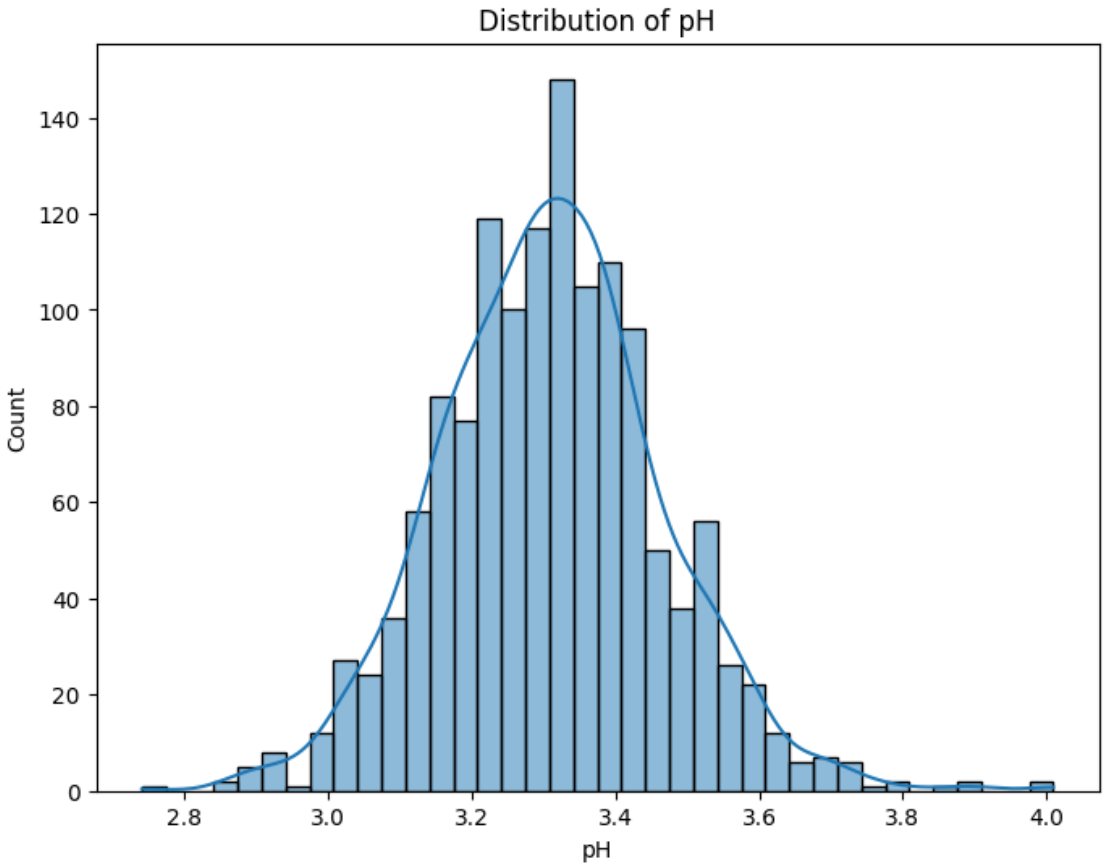

Distribution of residual sugar
Count
residual sugar
0
25
50
75
100
125
150
175
200
2
4
6
8
10
12
14
16

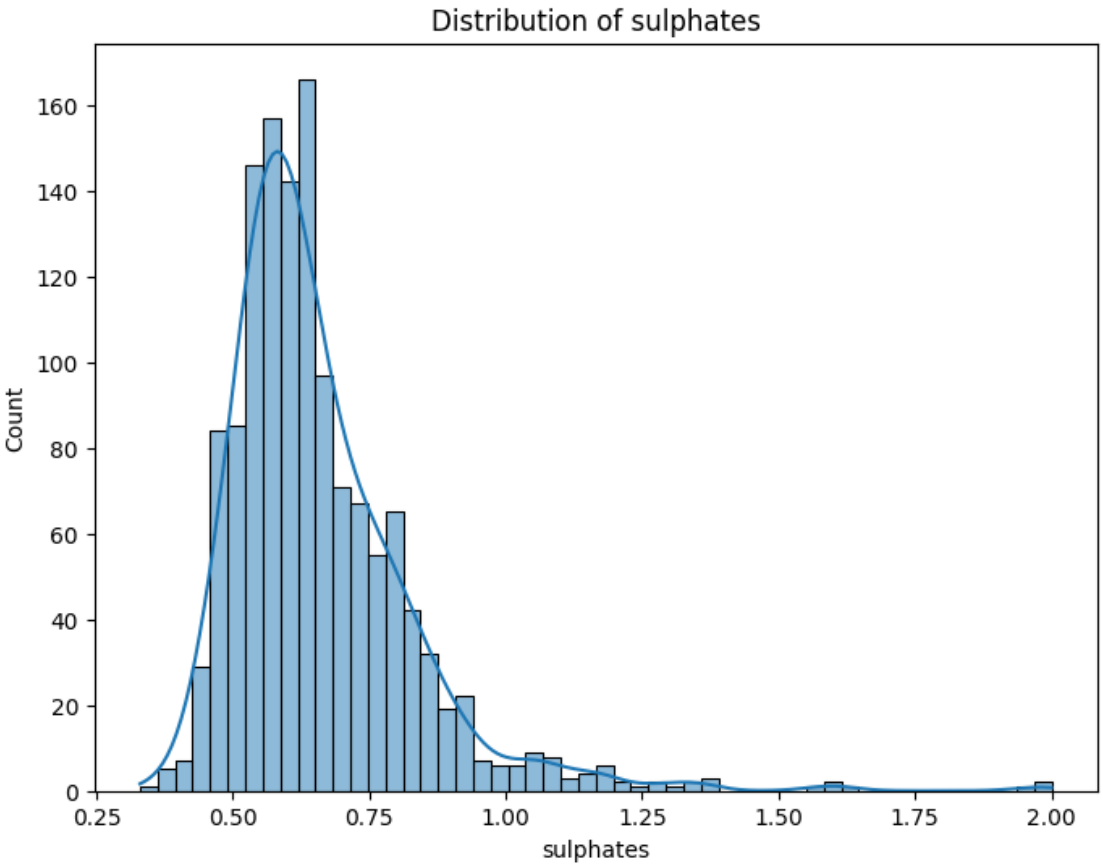

Distribution of chlorides
Count
chlorides
0
20
40
60
80
100
120
140
160
0.0
0.1
0.2
0.3
0.4
0.5
0.6

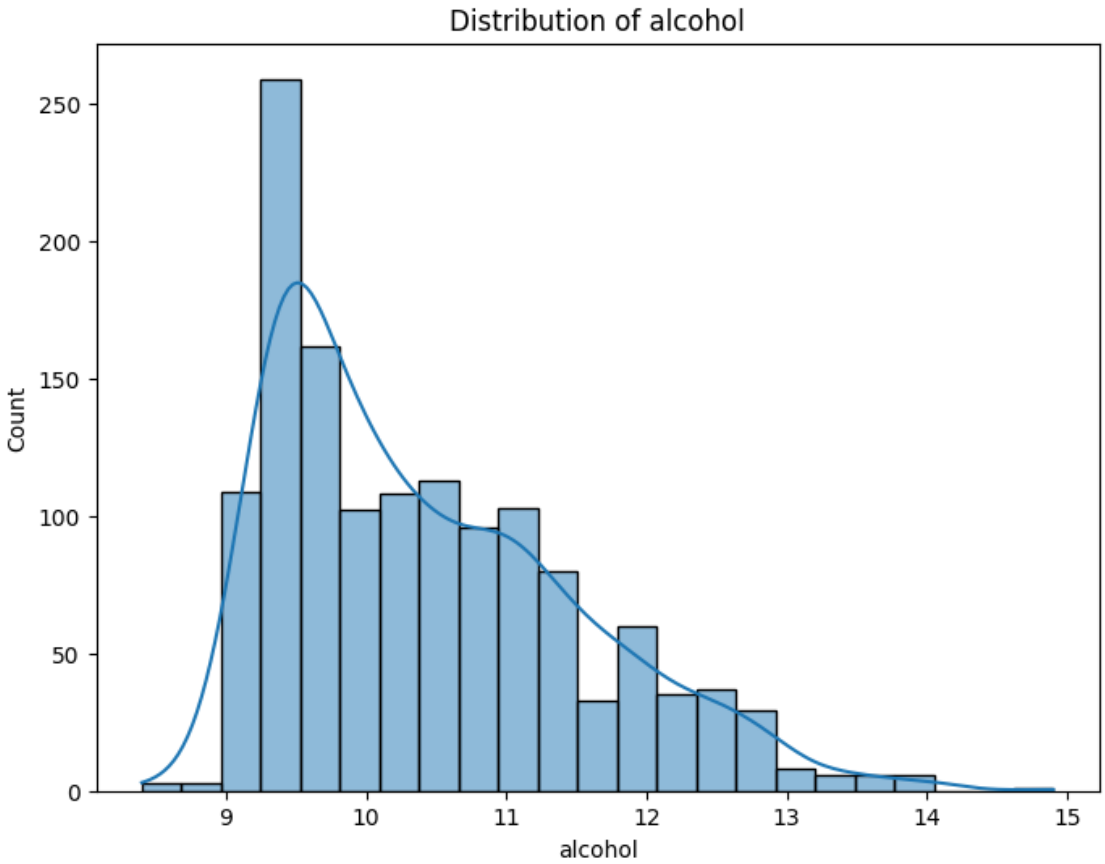

Distribution of free sulfur dioxide
Count
200
150
100
50
0
0
10
20
30
40
50
60
70
free sulfur dioxide

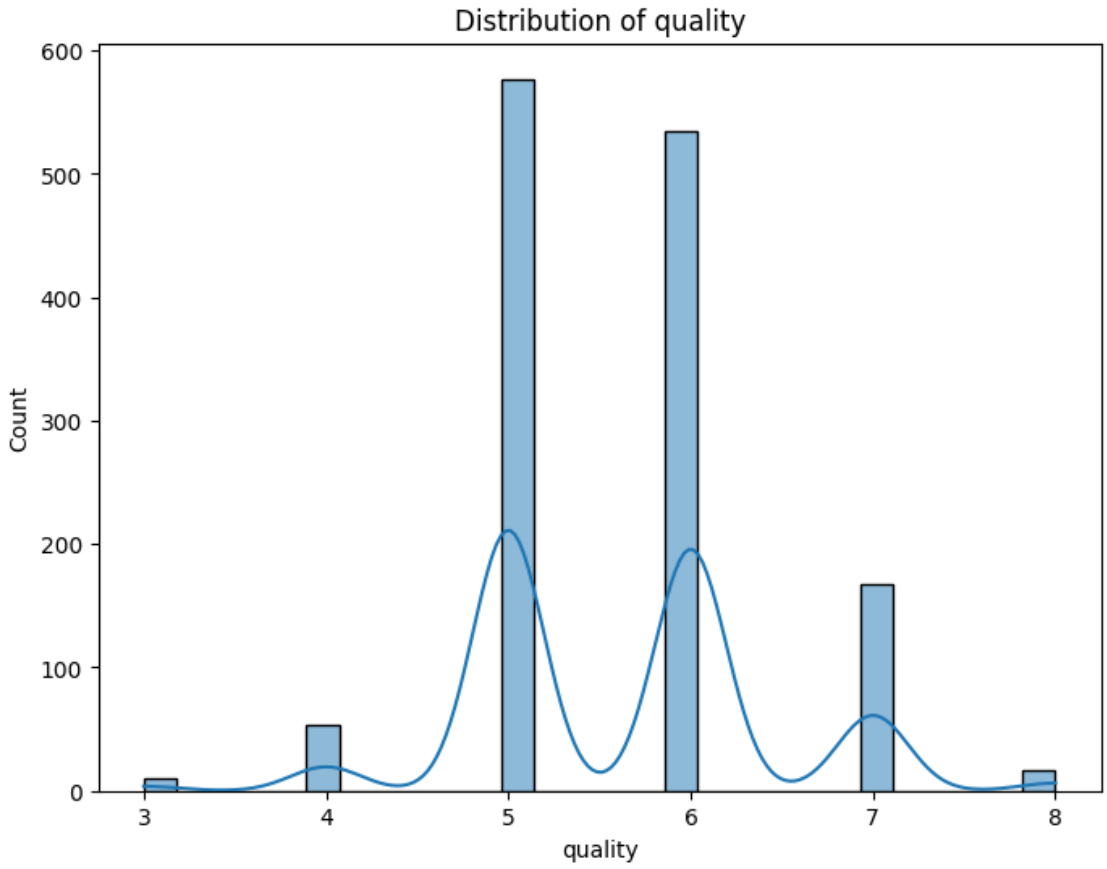

Distribution of total sulfur dioxide
Count
total sulfur dioxide
0
25
50
75
100
125
150
175
0
50
100
150
200
250
300


Distribution of density
Count
density
0
20
40
60
80
100
120
140
0.990
0.992
0.994
0.996
0.998
1.000
1.002
1.004


Distribution of pH
Count
pH
0
20
40
60
80
100
120
140
2.8
3.0
3.2
3.4
3.6
3.8
4.0


Distribution of sulphates
Count
sulphates
0
20
40
60
80
100
120
140
160
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

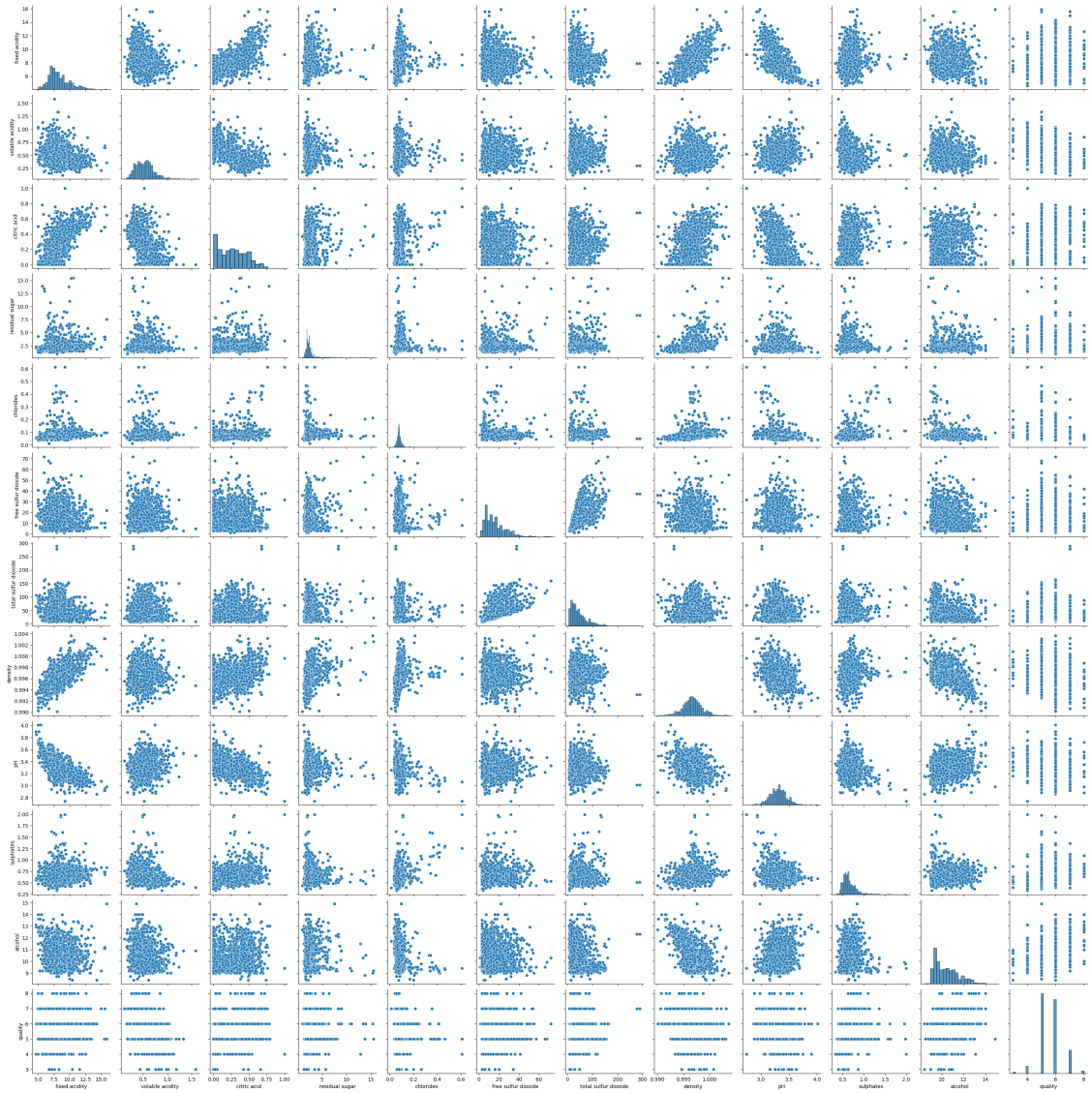

Distribution of alcohol
Count
alcohol
0
50
100
150
200
250
9
10
11
12
13
14
15



0.4 Univariate, bivariate, multivariate analysis

[ ]:
```
sns.pairplot(df)
```

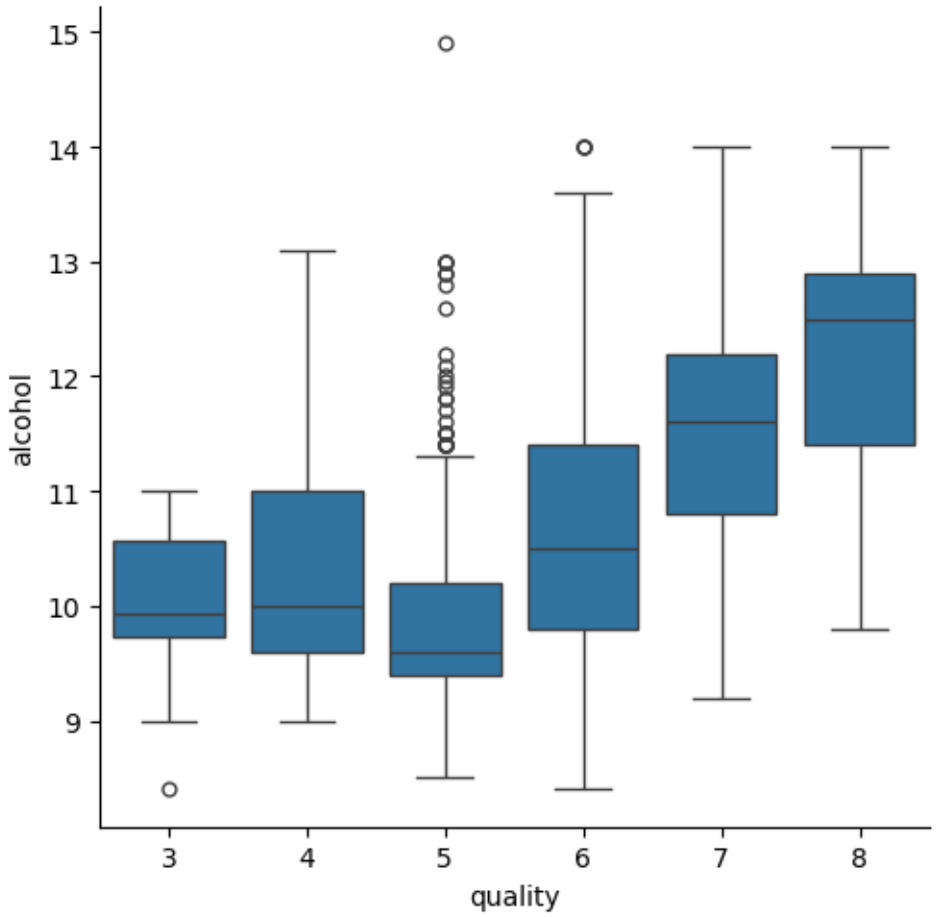
[ ]:
```
<seaborn.axisgrid.PairGrid at 0x1a1497171f0>
```



0.5 Categorical plot

[ ]:
```
sns.catplot(x = 'quality', y = 'alcohol', data = df, kind='box')
```

[ ]:
```
<seaborn.axisgrid.FacetGrid at 0x1a1496e7190>
```

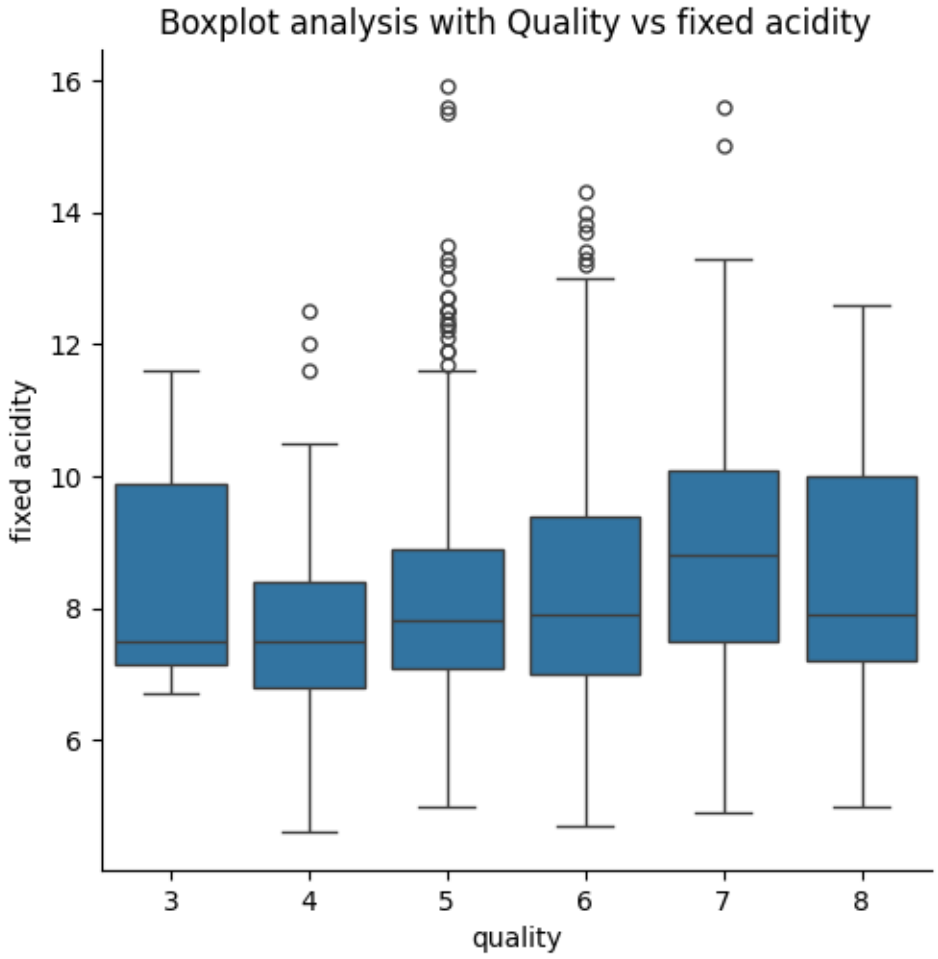


[21]:
```
lst = df.columns
lst
```

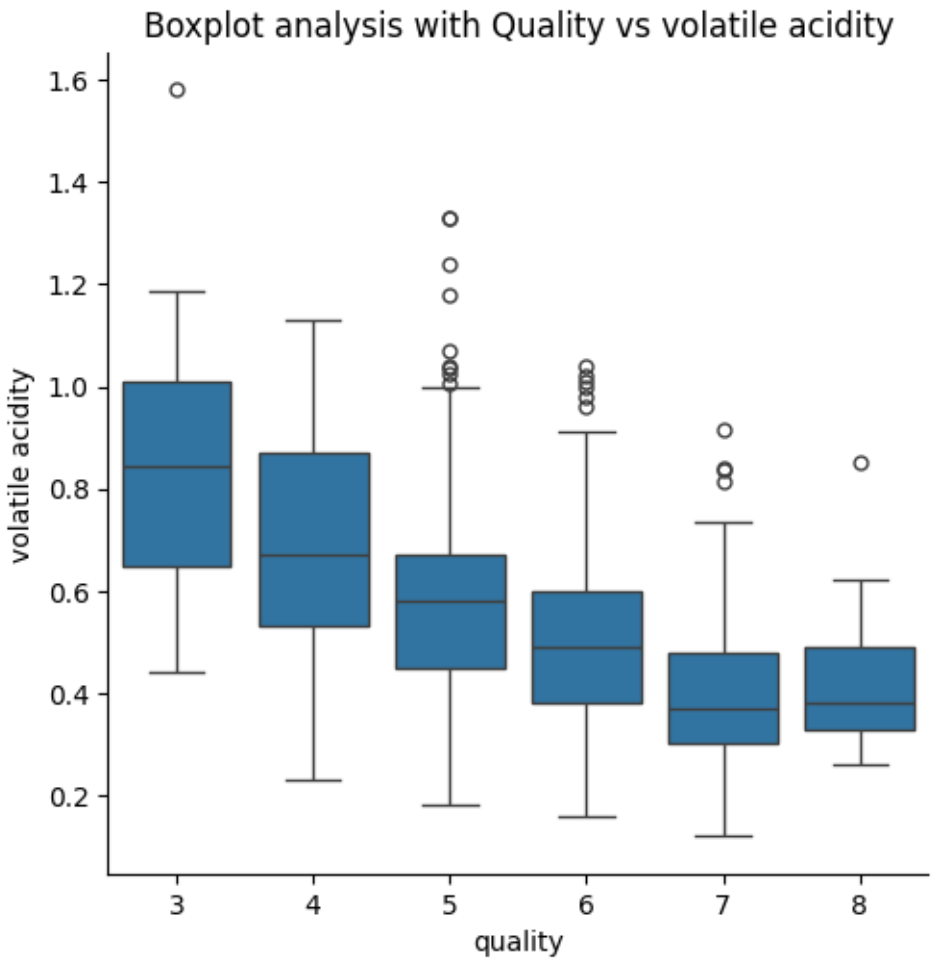
[21]:
```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

[22]:
```
for features in df.columns:
        plt.figure(figsize=(8, 6))
        if features=='quality':
                continue
        else:
                sns.catplot(x = 'quality', y = df[features], data = df,␣
 ↪kind='box')
                plt.title(f'Boxplot analysis with Quality vs {features}')
                plt.show()
                plt.close()
```

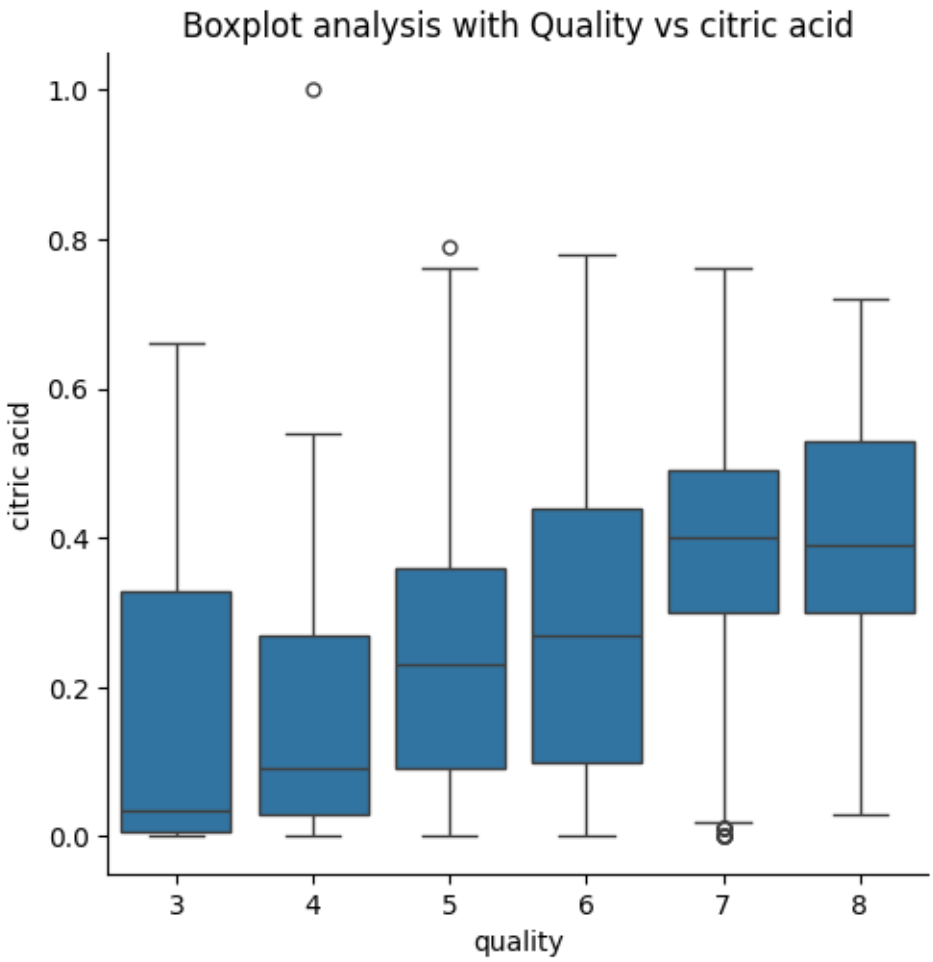
```
<Figure size 800x600 with 0 Axes>
```



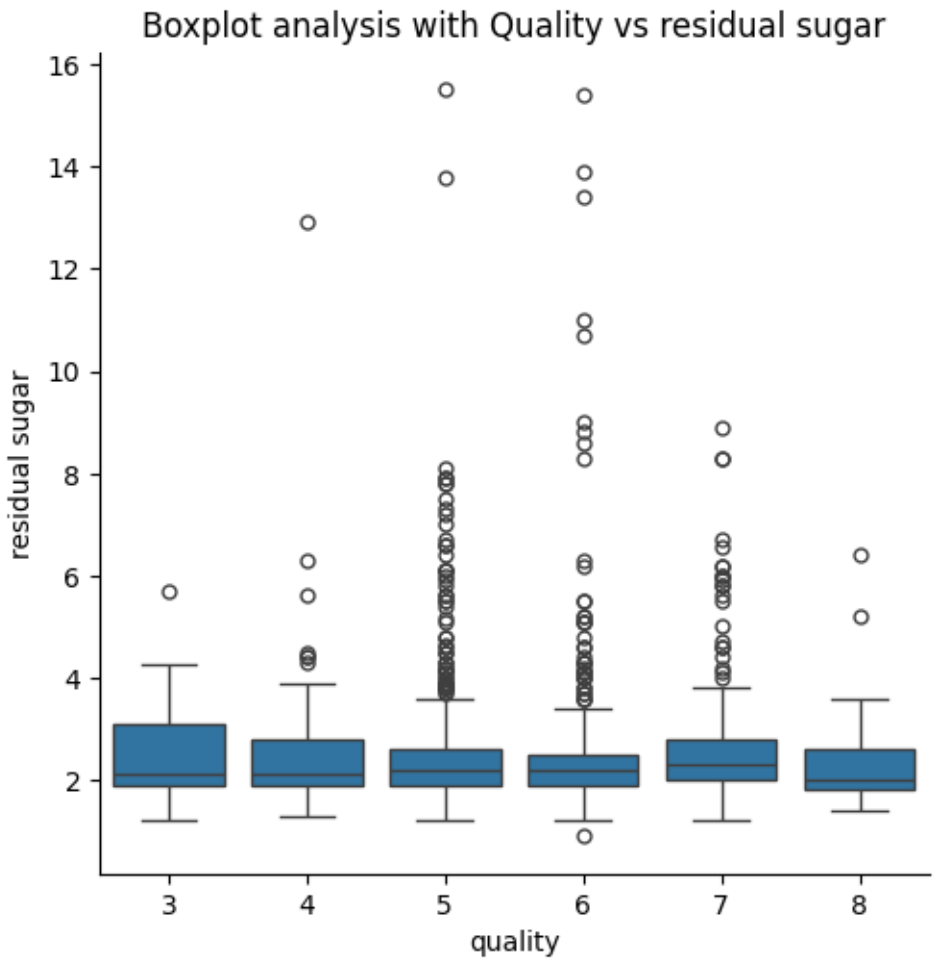

```
<Figure size 800x600 with 0 Axes>
```



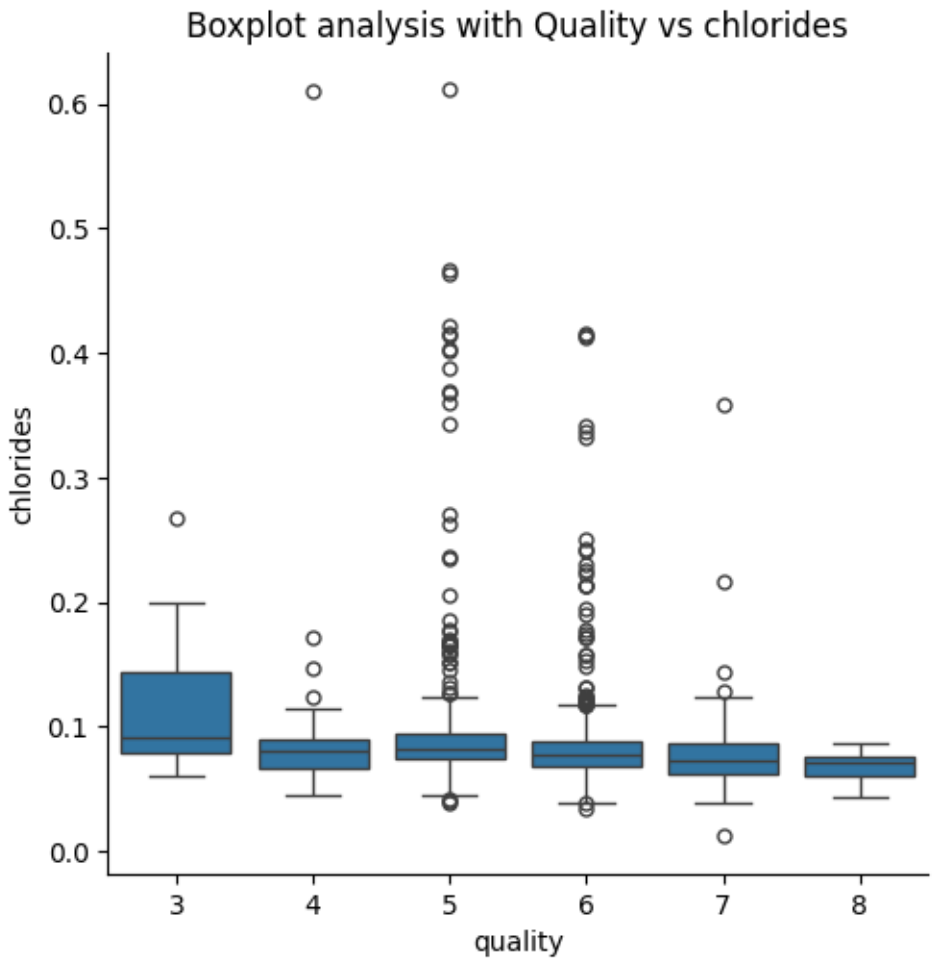

```
<Figure size 800x600 with 0 Axes>
```



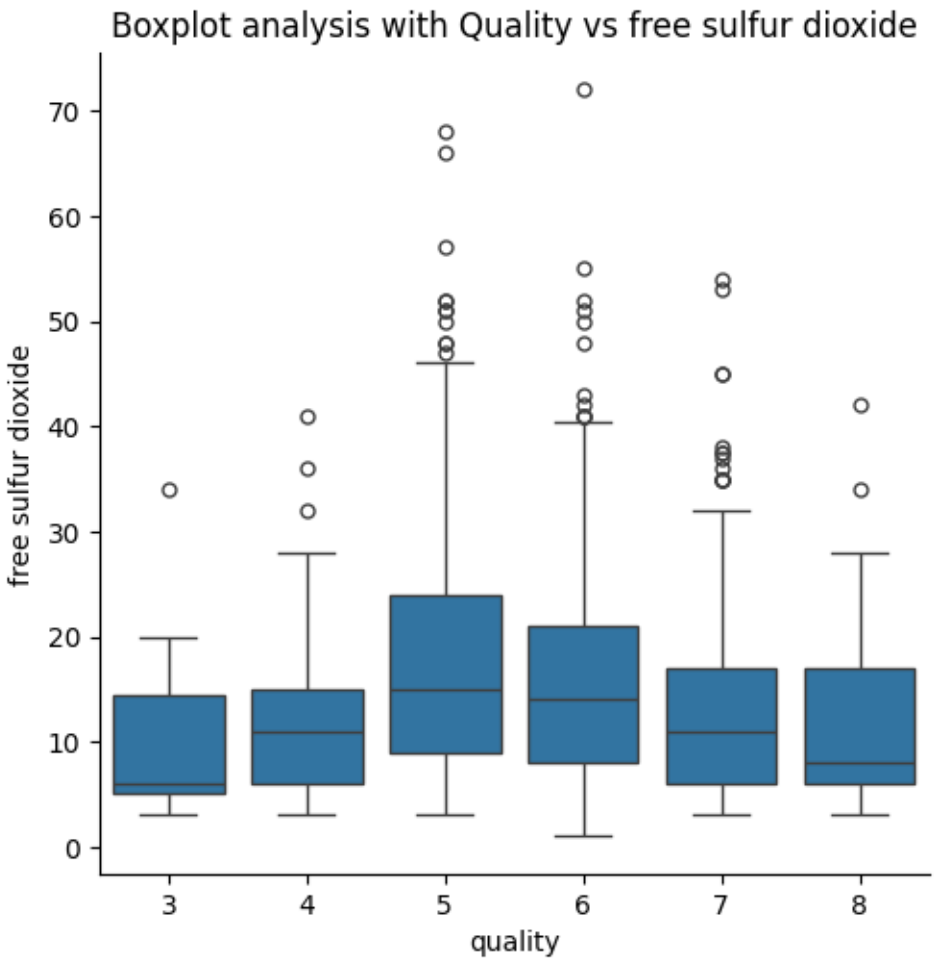

```
<Figure size 800x600 with 0 Axes>
```



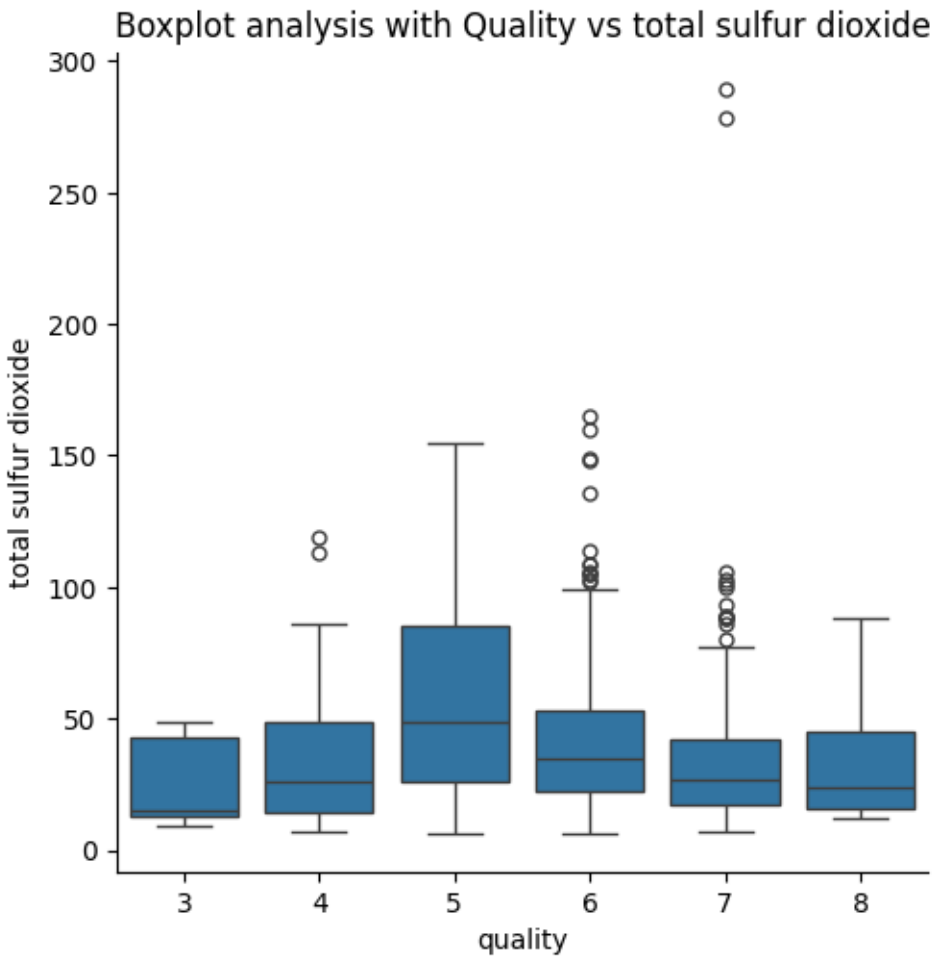

```
<Figure size 800x600 with 0 Axes>
```



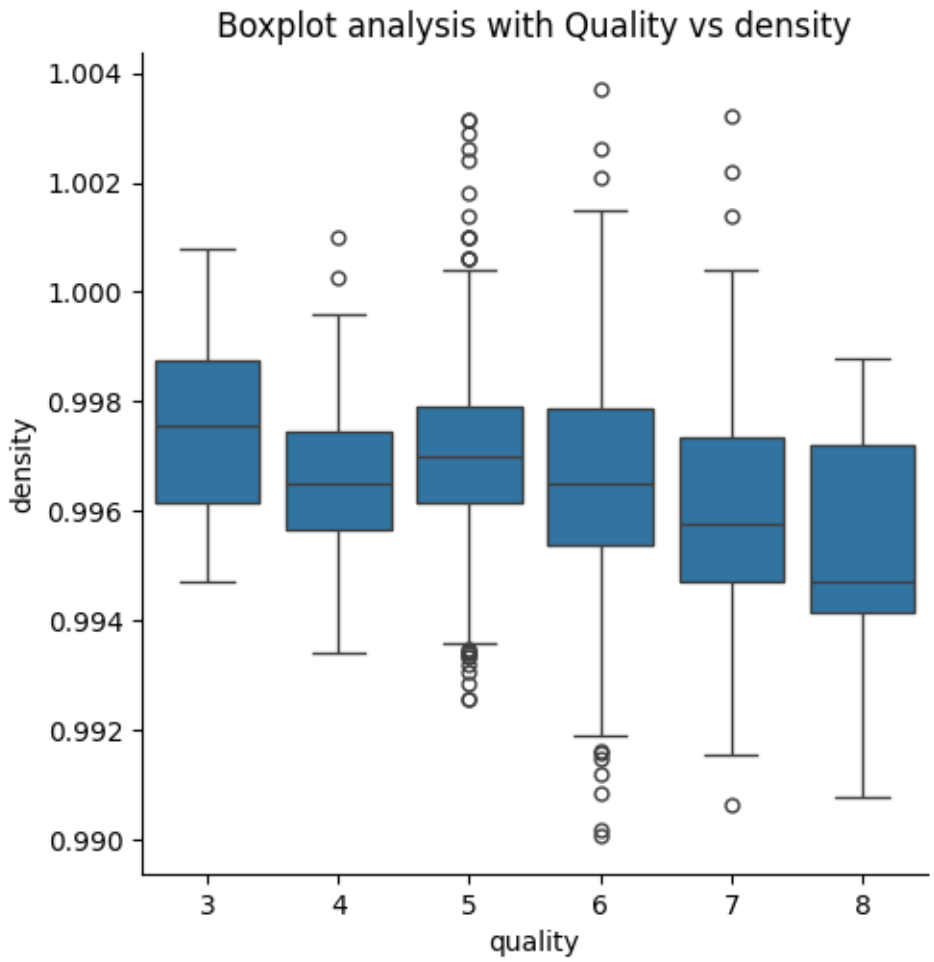

```
<Figure size 800x600 with 0 Axes>
```



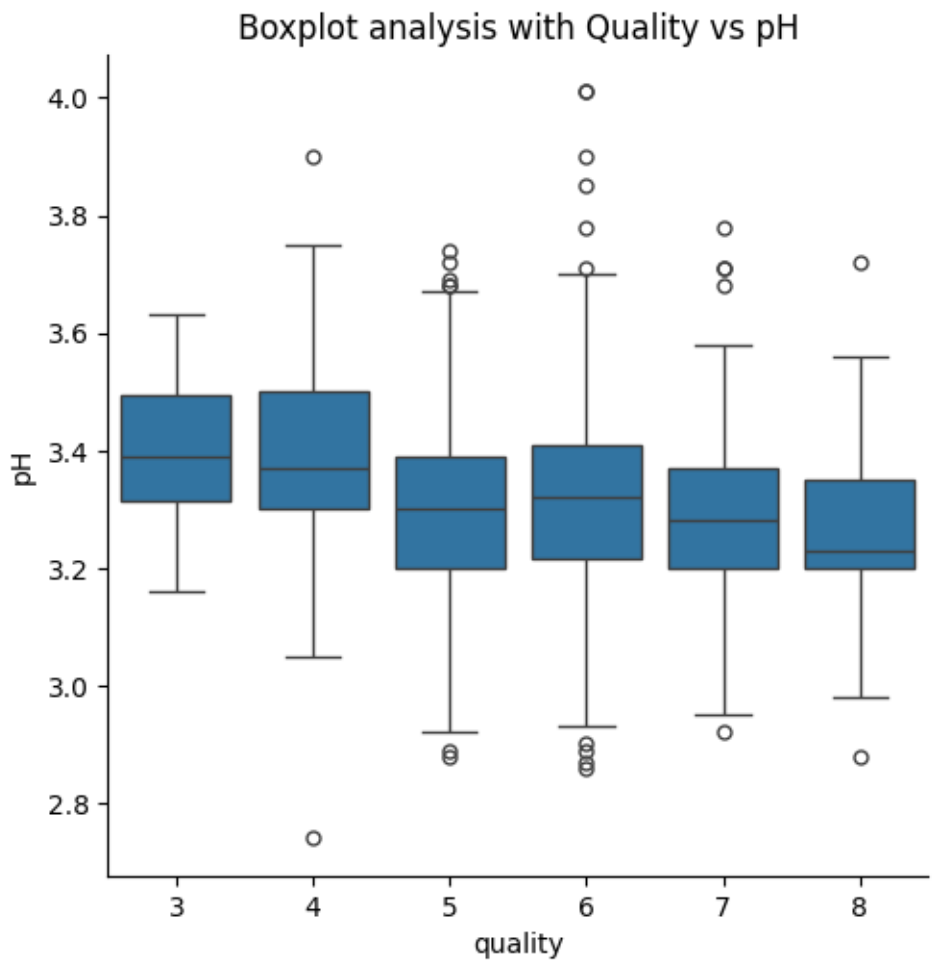

```
<Figure size 800x600 with 0 Axes>
```



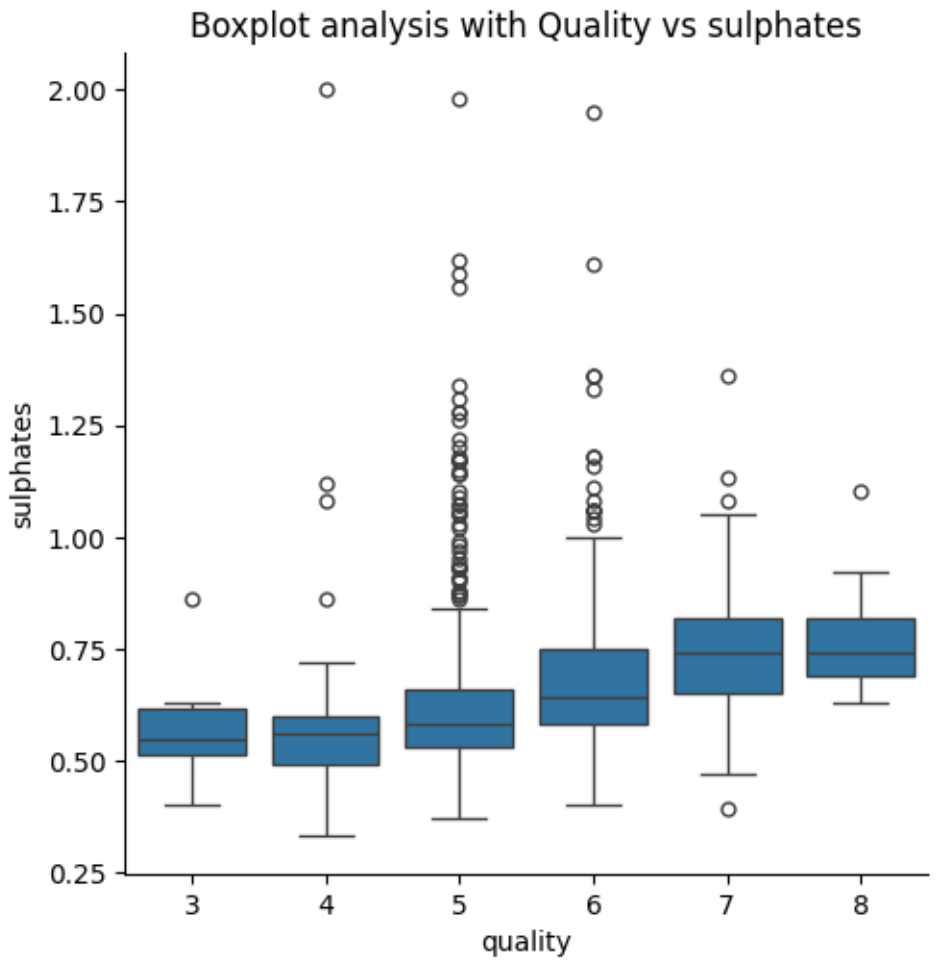

```
<Figure size 800x600 with 0 Axes>
```



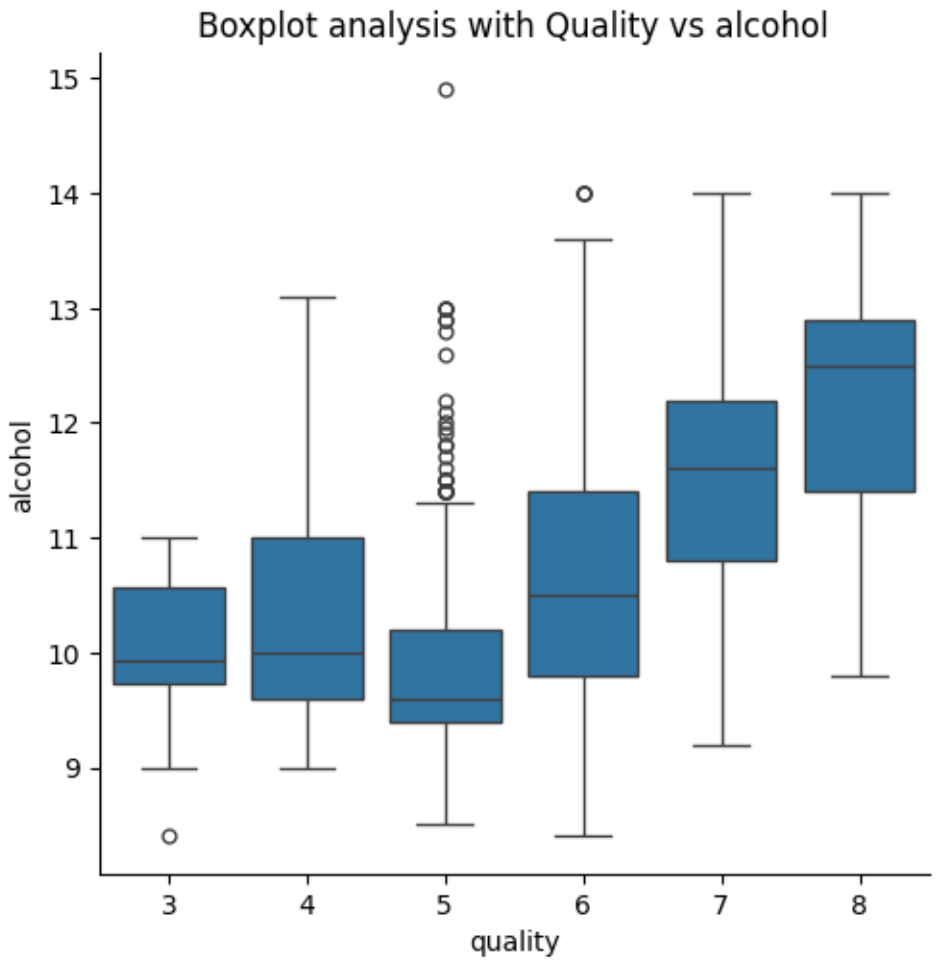
<Figure size 800x600 with 0 Axes>



<Figure size 800x600 with 0 Axes>



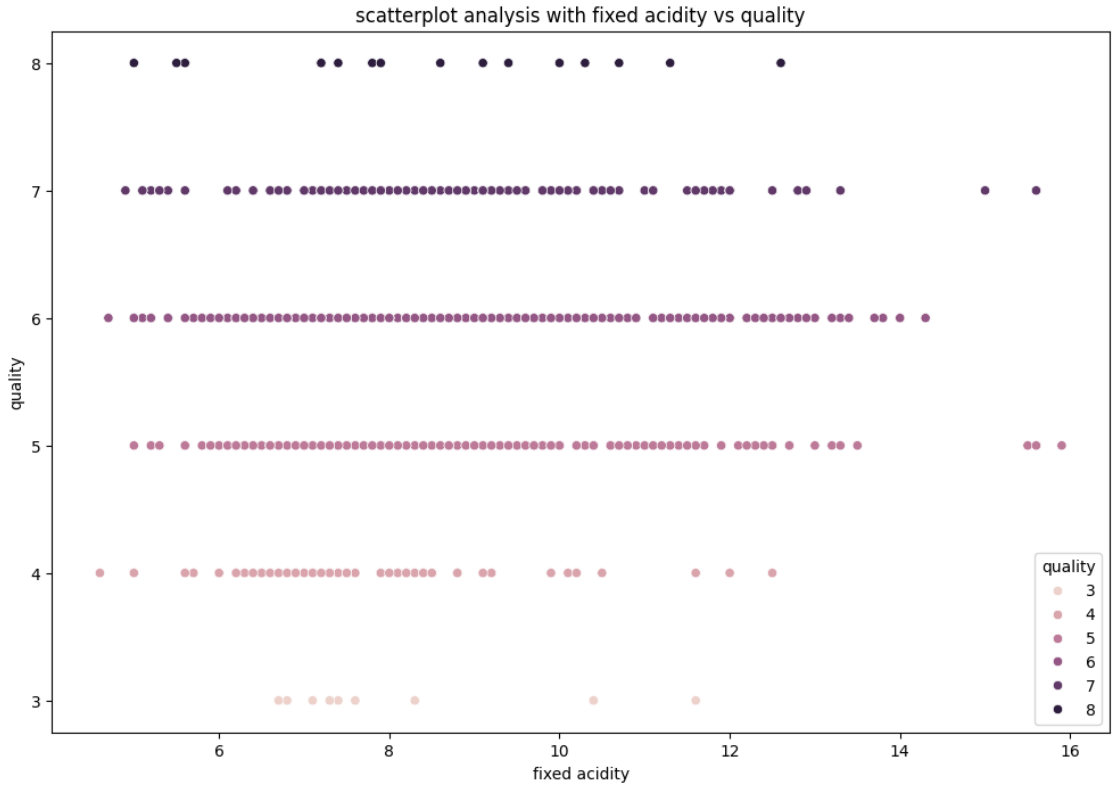

```
<Figure size 800x600 with 0 Axes>
```

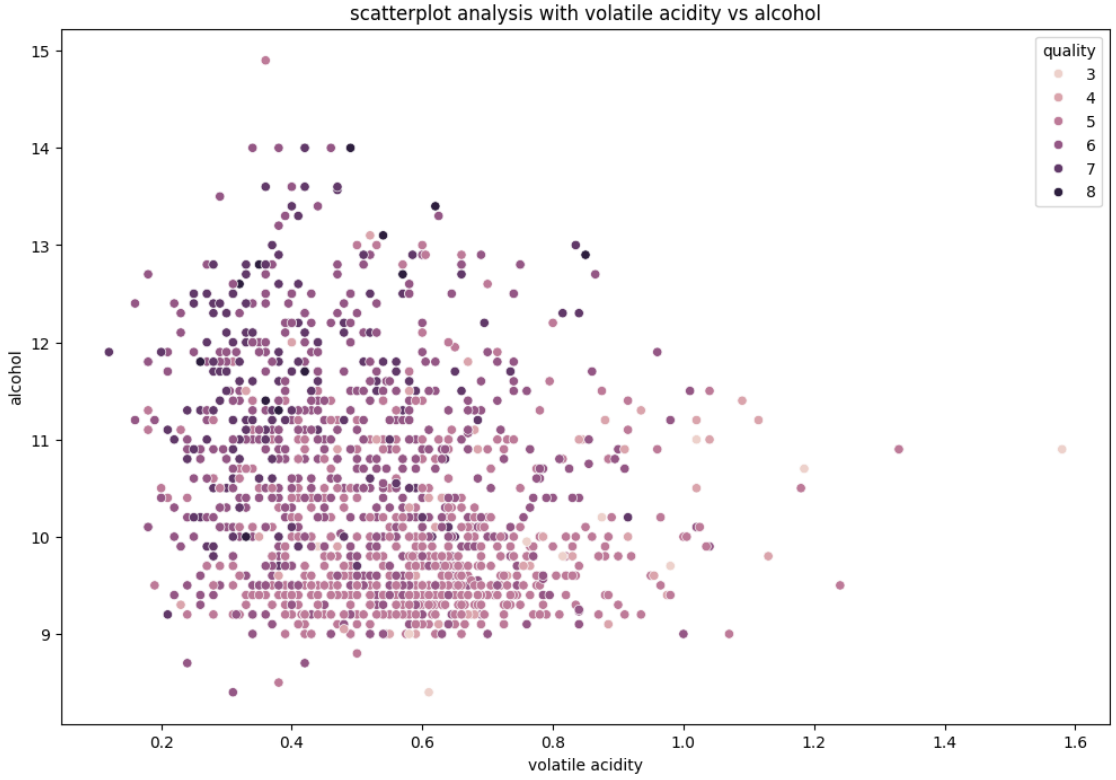


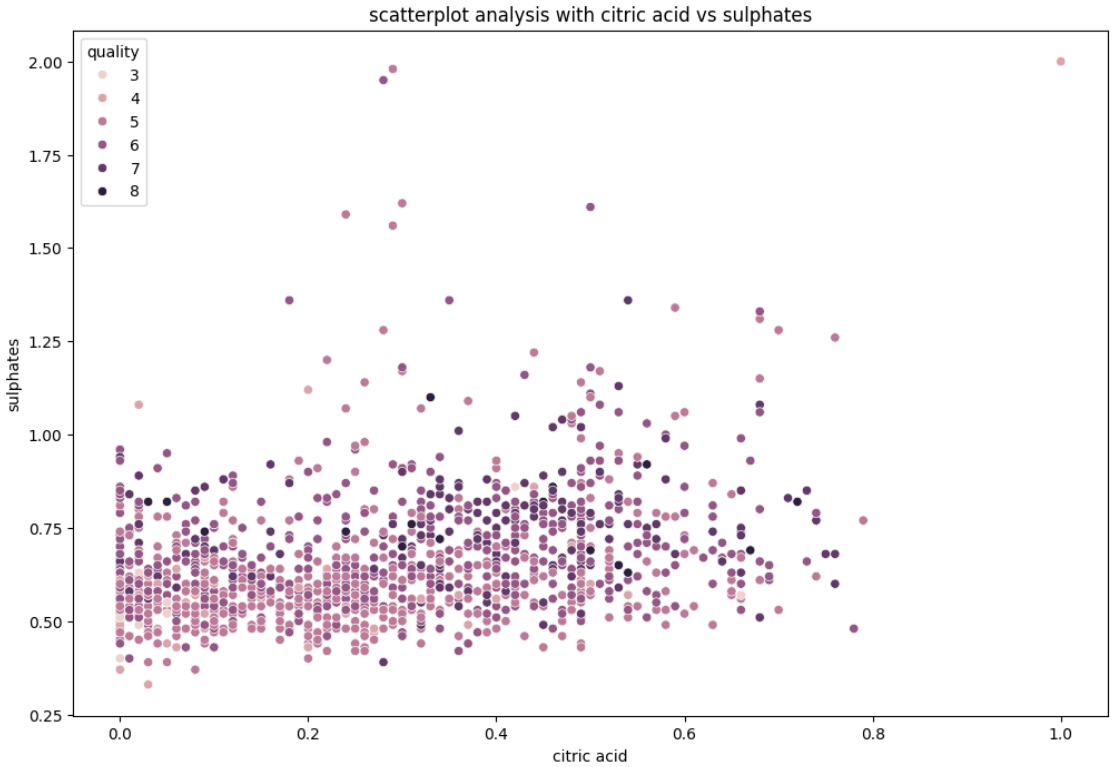

```
<Figure size 800x600 with 0 Axes>
```

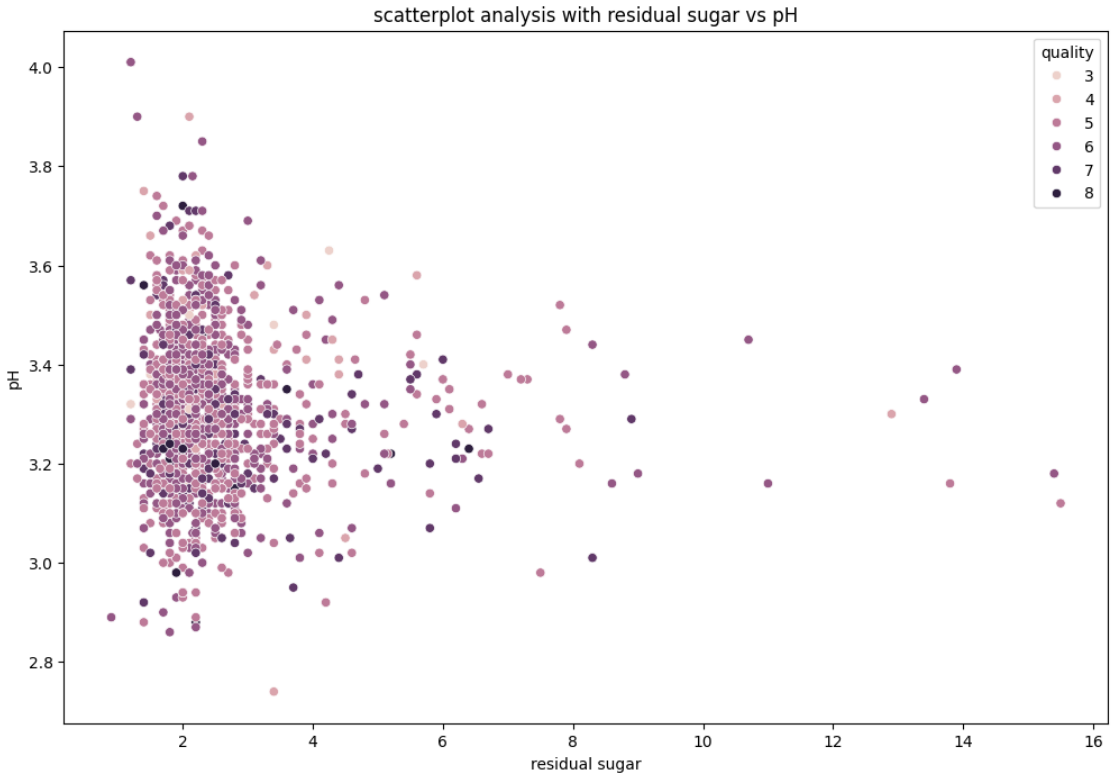
0.6 Scatter plot for the dataset

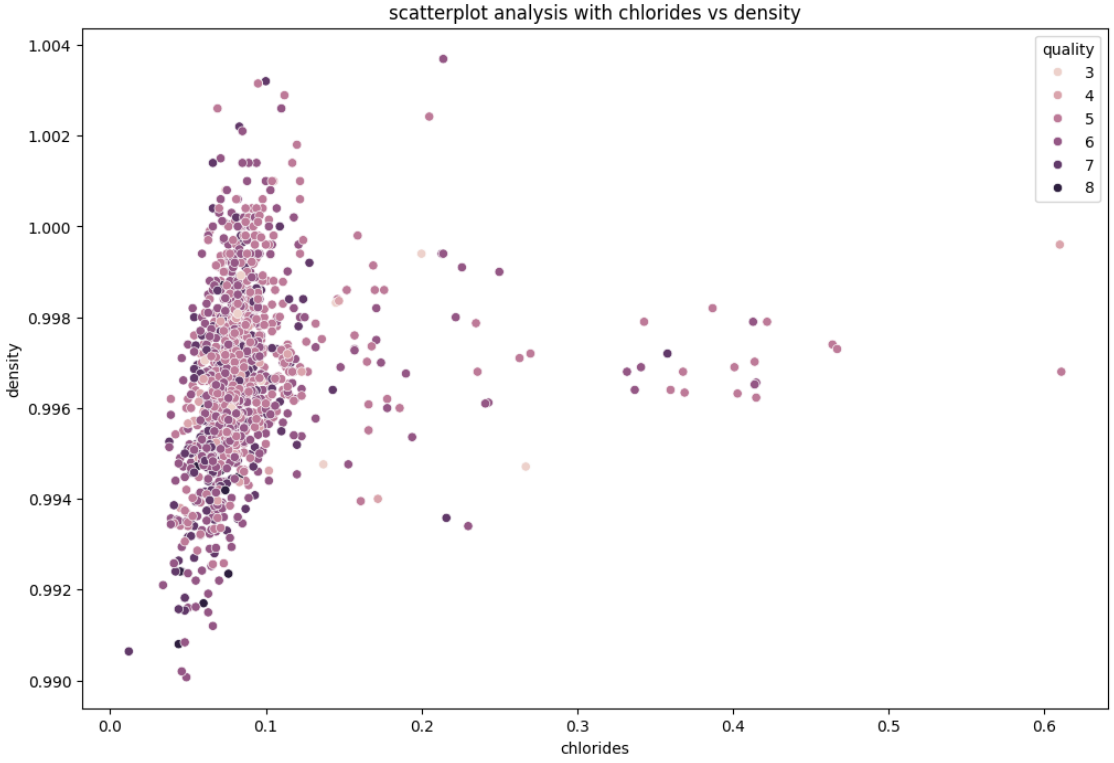
```
[ ]: for feature1, feature2 in zip(df.columns, df.columns[::-1]):
        plt.figure(figsize=(12,8))
        sns.scatterplot(x =df[feature1], y = df[feature2], hue = 'quality',␣
     ↪data = df )
        plt.title(f'scatterplot analysis with {feature1} vs {feature2}')
        plt.show()
        plt.close()
```

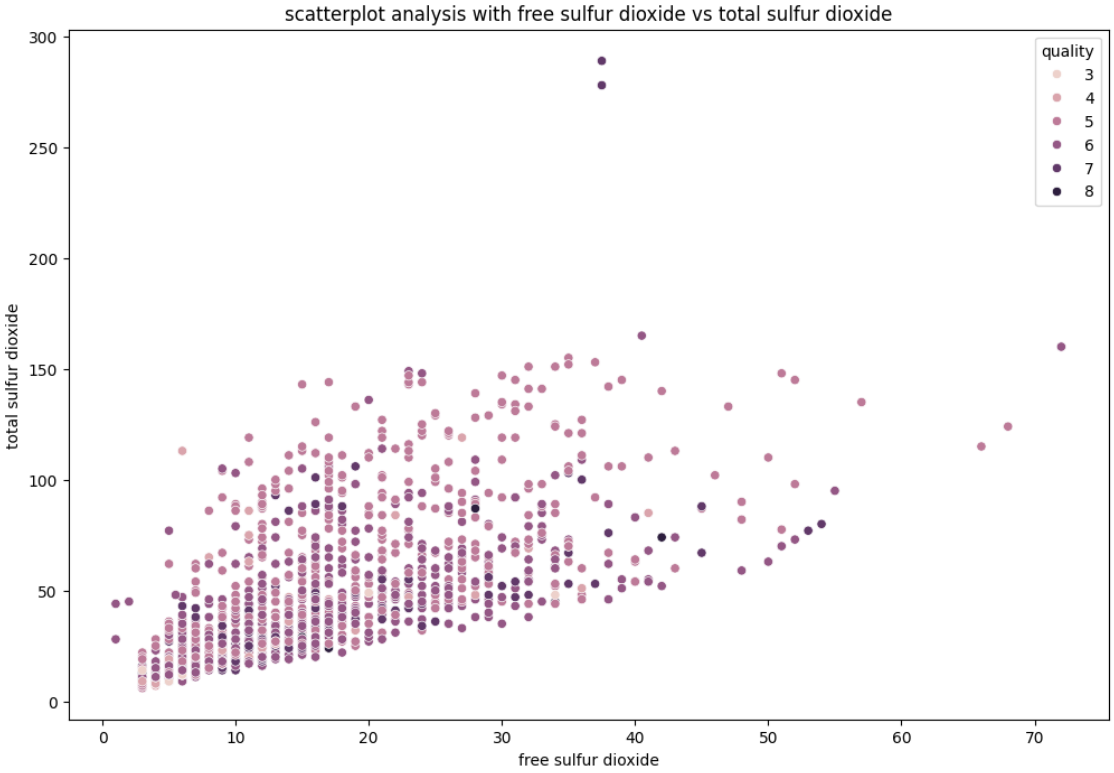

scatterplot analysis with fixed acidity vs quality
quality
fixed acidity
quality
3
4
5
6
7
8

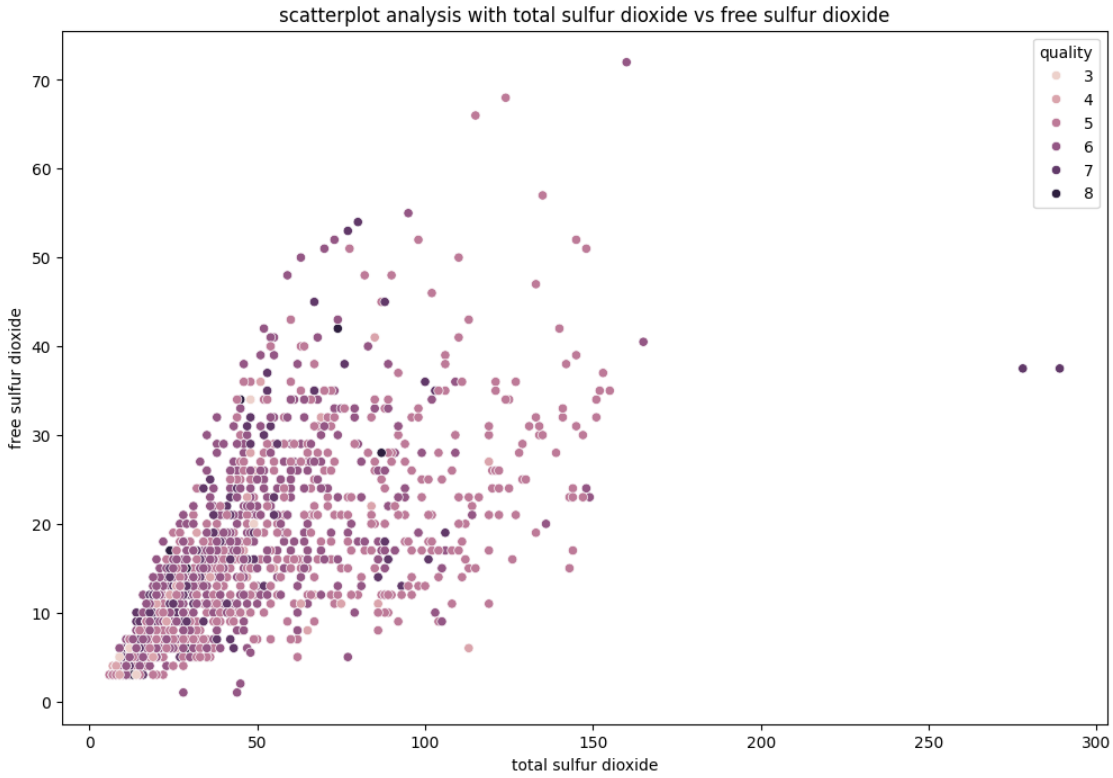

scatterplot analysis with volatile acidity vs alcohol
alcohol
volatile acidity
quality
3
4
5
6
7
8

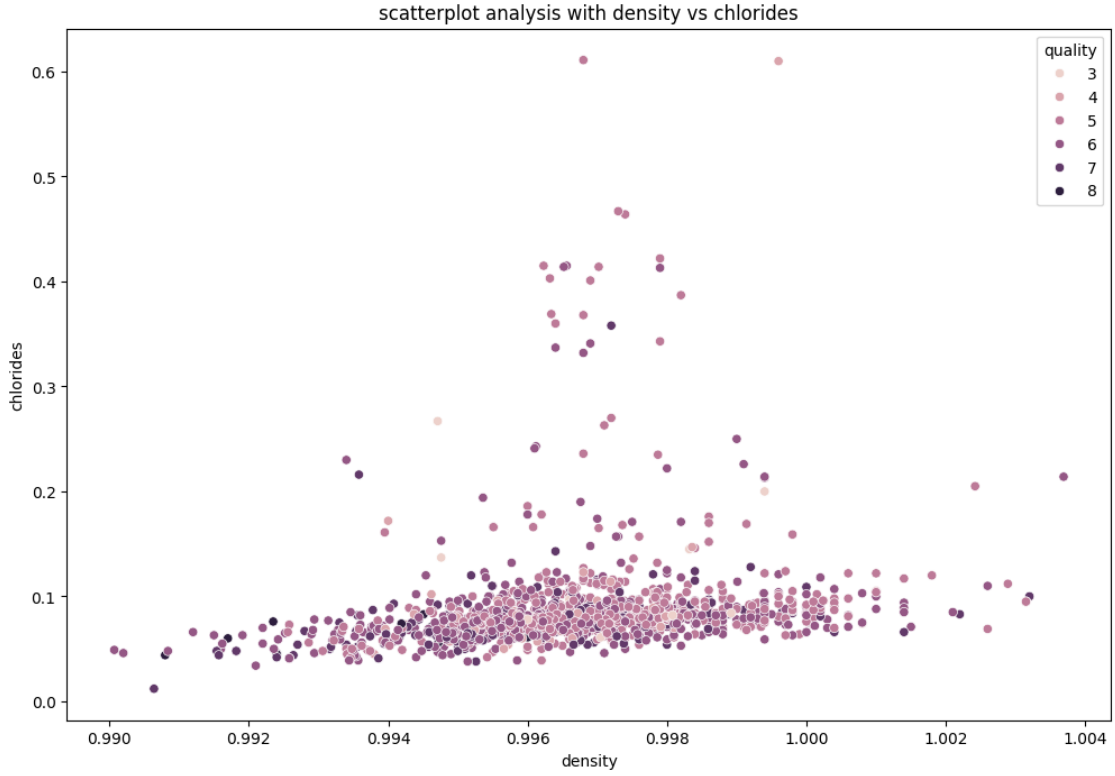

scatterplot analysis with citric acid vs sulphates
quality
3
4
5
6
7
8
sulphates
citric acid
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.0
0.2
0.4
0.6
0.8
1.0

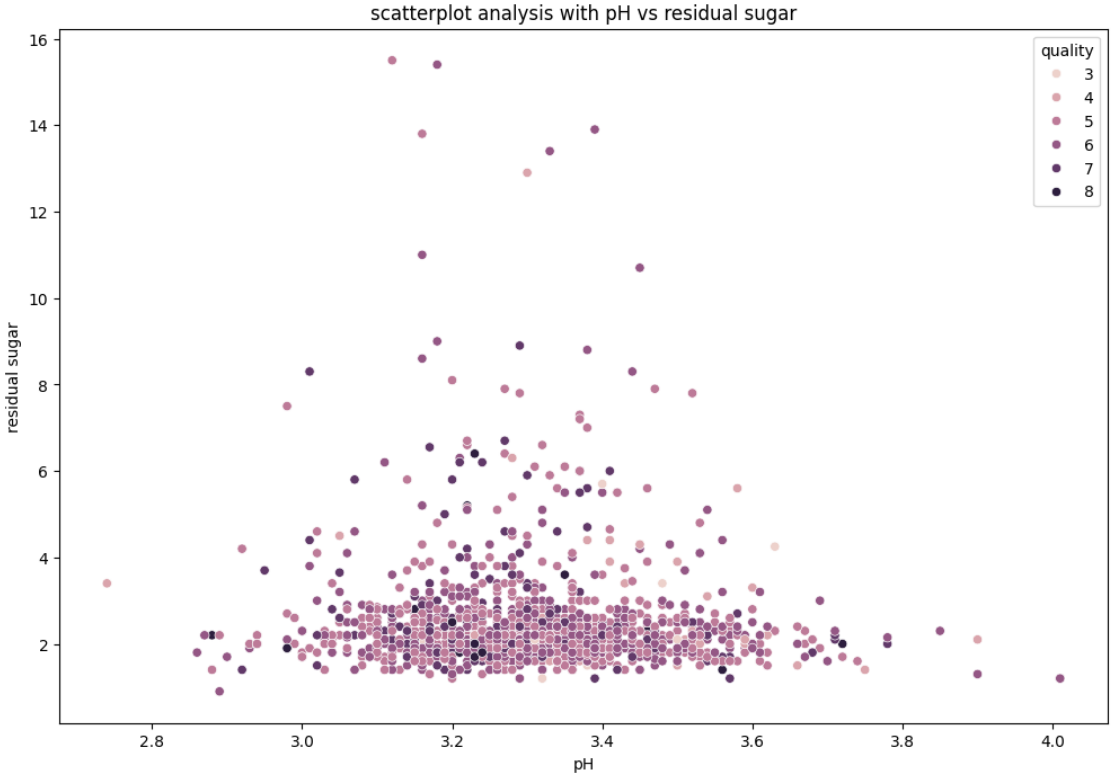

scatterplot analysis with residual sugar vs pH
quality
3
4
5
6
7
8
pH
residual sugar
4.0
3.8
3.6
3.4
3.2
3.0
2.8
2
4
6
8
10
12
14
16

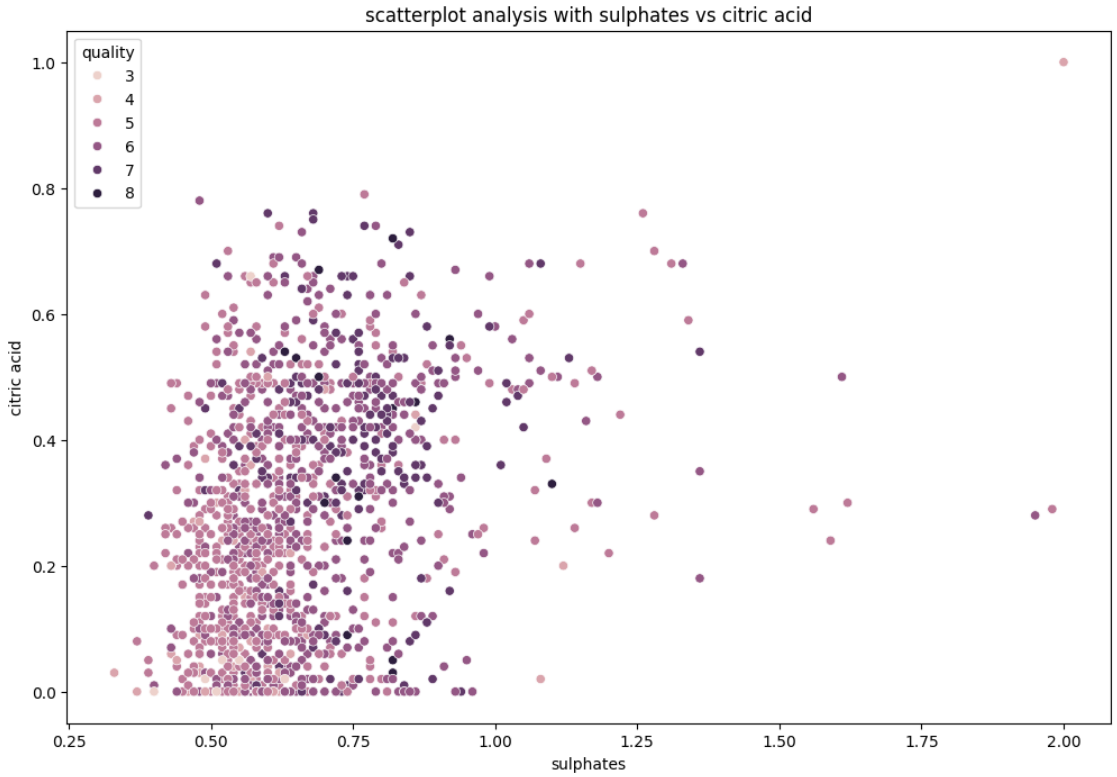

scatterplot analysis with chlorides vs density
quality
3
4
5
6
7
8
density
chlorides
1.004
1.002
1.000
0.998
0.996
0.994
0.992
0.990
0.0
0.1
0.2
0.3
0.4
0.5
0.6

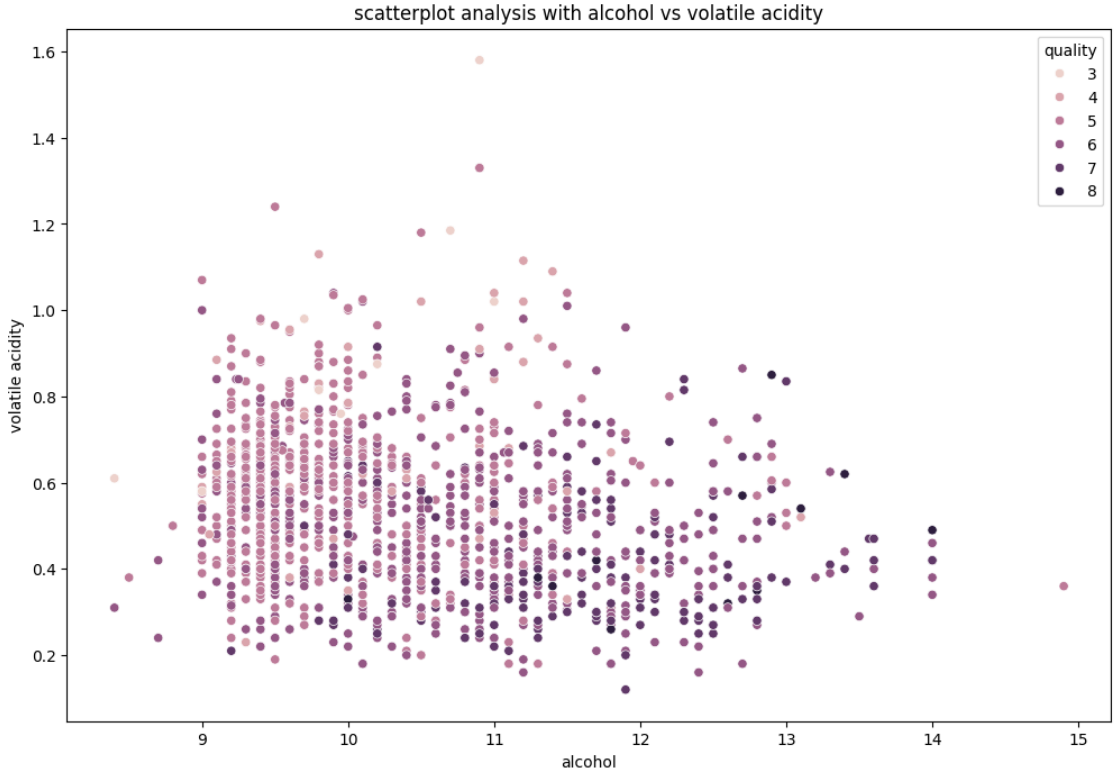

scatterplot analysis with free sulfur dioxide vs total sulfur dioxide
quality
3
4
5
6
7
8
total sulfur dioxide
free sulfur dioxide
0
50
100
150
200
250
300
0
10
20
30
40
50
60
70

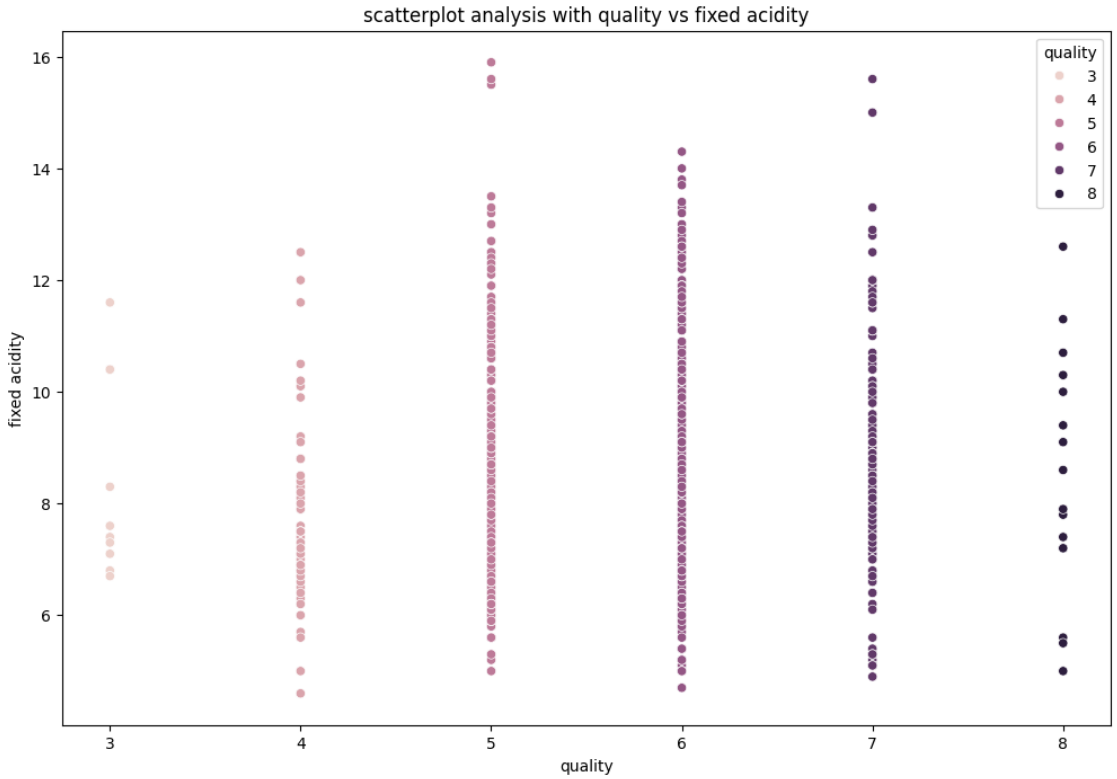

scatterplot analysis with total sulfur dioxide vs free sulfur dioxide
quality
3
4
5
6
7
8
free sulfur dioxide
total sulfur dioxide
0
10
20
30
40
50
60
70
0
50
100
150
200
250
300


scatterplot analysis with density vs chlorides
quality
3
4
5
6
7
8
chlorides
density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.990
0.992
0.994
0.996
0.998
1.000
1.002
1.004


scatterplot analysis with pH vs residual sugar
quality
3
4
5
6
7
8
residual sugar
pH
16
14
12
10
8
6
4
2
2.8
3.0
3.2
3.4
3.6
3.8
4.0


scatterplot analysis with sulphates vs citric acid
quality
3
4
5
6
7
8
citric acid
sulphates
0.0
0.2
0.4
0.6
0.8
1.0
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


scatterplot analysis with alcohol vs volatile acidity
quality
3
4
5
6
7
8
volatile acidity
alcohol
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
9
10
11
12
13
14
15


scatterplot analysis with quality vs fixed acidity
quality
3
4
5
6
7
8
fixed acidity
16
14
12
10
8
6
3
4
5
6
7
8
quality