```python
import pandas as pd

# Assuming the file paths are 'path_to_red_wine_data.csv' and
'path_to_white_wine_data.csv'
# Please replace these with the actual file paths or names
red_wine_data = pd.read_csv('C:/Users/dande/Desktop/DMT/Assignment
2/wine+quality/winequality-red.csv', sep=';')
white_wine_data = pd.read_csv('C:/Users/dande/Desktop/DMT/Assignment
2/wine+quality/winequality-white.csv', sep=';')

# Add a 'wineType' column to each DataFrame
red_wine_data['wineType'] = 'red'
white_wine_data['wineType'] = 'white'

# Combine the datasets
combined_wine_data = pd.concat([red_wine_data, white_wine_data],
ignore_index=True)

# Check the data types again
print(combined_wine_data)
```

```
      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
4               7.4              0.70         0.00             1.9
0.076
...             ...               ...          ...             ...
...
6492            6.2              0.21         0.29             1.6
0.039
6493            6.6              0.32         0.36             8.0
0.047
6494            6.5              0.24         0.19             1.2
0.041
6495            5.5              0.29         0.30             1.1
0.022
6496            6.0              0.21         0.38             0.8
0.020

      free sulfur dioxide  total sulfur dioxide  density    pH
sulphates  \
0                    11.0                  34.0  0.99780  3.51
0.56
```

```
1                     25.0                     67.0  0.99680  3.20
0.68
2                     15.0                     54.0  0.99700  3.26
0.65
3                     17.0                     60.0  0.99800  3.16
0.58
4                     11.0                     34.0  0.99780  3.51
0.56
...                    ...                      ...      ...   ...
...
6492                  24.0                     92.0  0.99114  3.27
0.50
6493                  57.0                    168.0  0.99490  3.15
0.46
6494                  30.0                    111.0  0.99254  2.99
0.46
6495                  20.0                    110.0  0.98869  3.34
0.38
6496                  22.0                     98.0  0.98941  3.26
0.32

      alcohol  quality wineType
0         9.4        5      red
1         9.8        5      red
2         9.8        5      red
3         9.8        6      red
4         9.4        5      red
...       ...      ...      ...
6492     11.2        6    white
6493      9.6        5    white
6494      9.4        6    white
6495     12.8        7    white
6496     11.8        6    white

[6497 rows x 13 columns]

# Select only numeric columns for the operations
numeric_data = combined_wine_data.select_dtypes(include=[np.number])

# Computing summary statistics for numeric features in the dataset
summary_stats = numeric_data.describe()

# Directly calculating the range for each numeric feature
range_values = numeric_data.max() - numeric_data.min()

# Calculating the variance for each numeric feature
variance_values = numeric_data.var()

# Adding the range and variance to the summary statistics DataFrame
summary_stats.loc['range'] = range_values
```

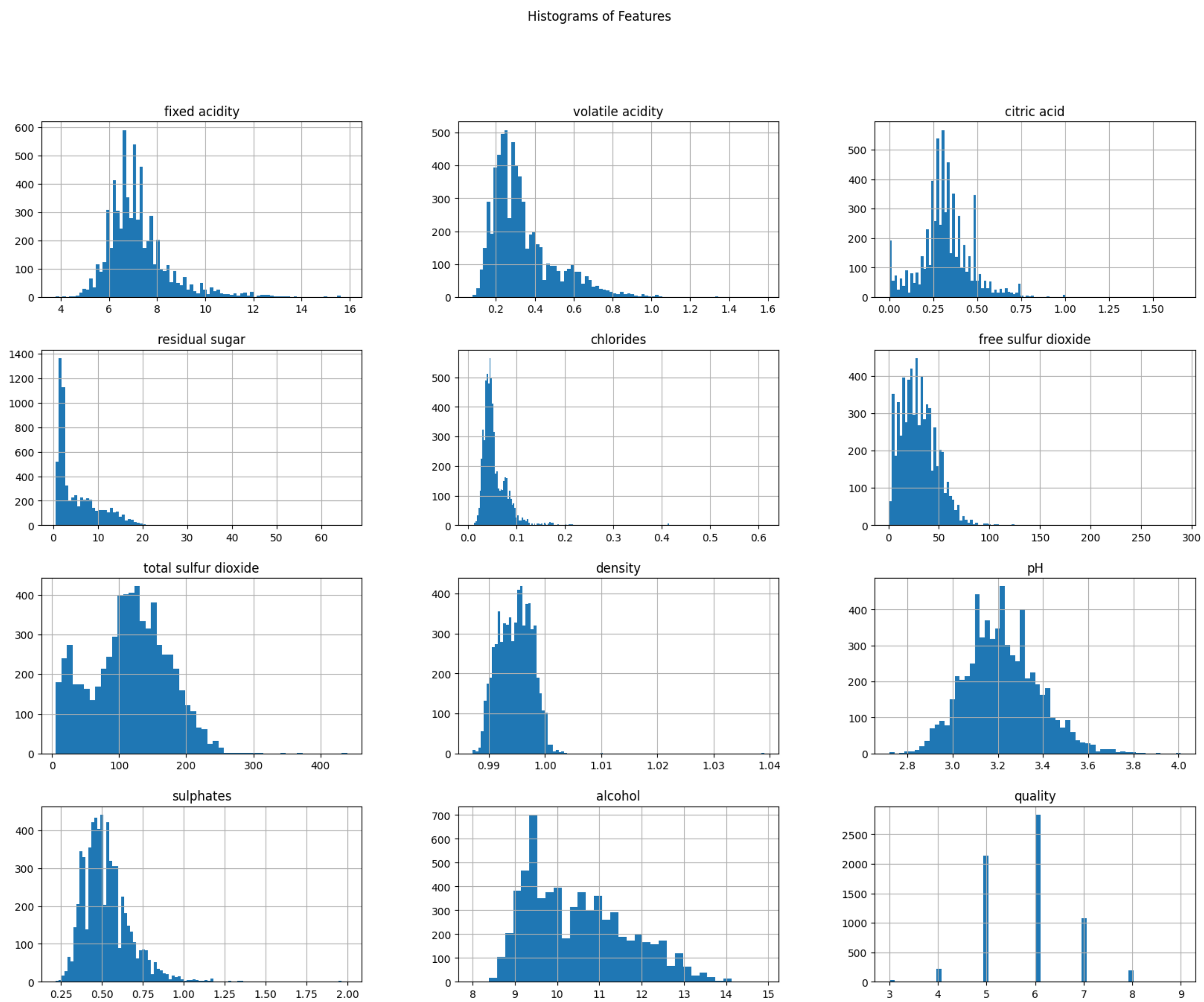
```python
summary_stats.loc['variance'] = variance_values

# Displaying the summary statistics
print(summary_stats)
```

```
          fixed acidity  volatile acidity  citric acid  residual sugar  \
count       6497.000000       6497.000000  6497.000000     6497.000000

mean           7.215307          0.339666     0.318633        5.443235

std            1.296434          0.164636     0.145318        4.757804

min            3.800000          0.080000     0.000000        0.600000

25%            6.400000          0.230000     0.250000        1.800000

50%            7.000000          0.290000     0.310000        3.000000

75%            7.700000          0.400000     0.390000        8.100000

max           15.900000          1.580000     1.660000       65.800000

range         12.100000          1.500000     1.660000       65.200000

variance       1.680740          0.027105     0.021117       22.636696


            chlorides  free sulfur dioxide  total sulfur dioxide
density  \
count     6497.000000          6497.000000           6497.000000
6497.000000
mean         0.056034            30.525319            115.744574
0.994697
std          0.035034            17.749400             56.521855
0.002999
min          0.009000             1.000000              6.000000
0.987110
25%          0.038000            17.000000             77.000000
0.992340
50%          0.047000            29.000000            118.000000
0.994890
75%          0.065000            41.000000            156.000000
0.996990
max          0.611000           289.000000            440.000000
1.038980
range        0.602000           288.000000            434.000000
0.051870
variance     0.001227           315.041192           3194.720039
0.000009
```

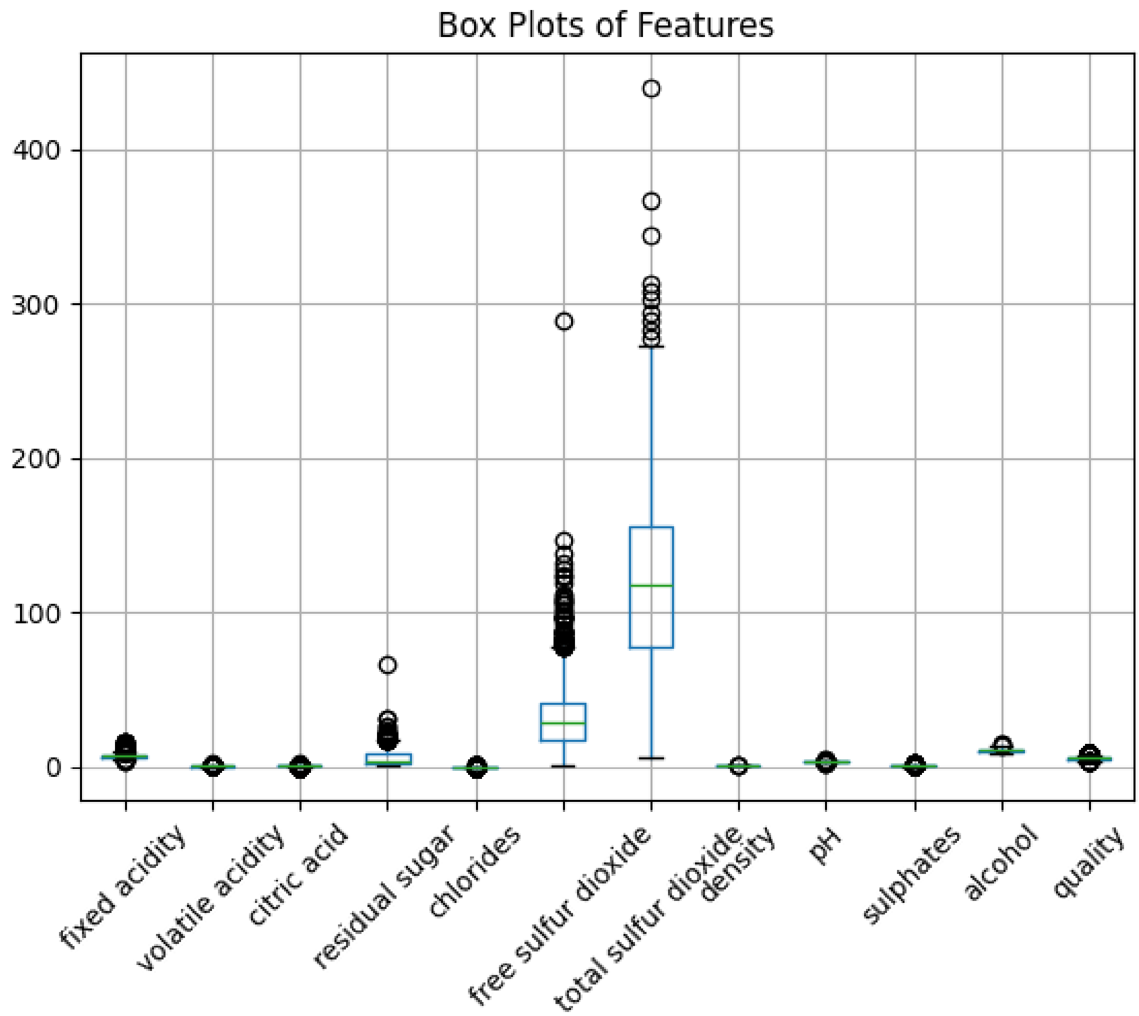
	pH	sulphates	alcohol	quality
count	6497.000000	6497.000000	6497.000000	6497.000000
mean	3.218501	0.531268	10.491801	5.818378
std	0.160787	0.148806	1.192712	0.873255
min	2.720000	0.220000	8.000000	3.000000
25%	3.110000	0.430000	9.500000	5.000000
50%	3.210000	0.510000	10.300000	6.000000
75%	3.320000	0.600000	11.300000	6.000000
max	4.010000	2.000000	14.900000	9.000000
range	1.290000	1.780000	6.900000	6.000000
variance	0.025853	0.022143	1.422561	0.762575

```python
import matplotlib.pyplot as plt
import seaborn as sns
from itertools import combinations

# 1. Histograms for each feature
def plot_histograms(data):
    data.hist(bins='auto', figsize=(20, 15))
    plt.suptitle('Histograms of Features')
    plt.show()

# 2. Box Plots for each feature
def plot_boxplots(data):
    plt.figure(figsize=(7, 5))
    data.boxplot()
    plt.title('Box Plots of Features')
    plt.xticks(rotation=45)
    plt.show()

# 3. Pairwise Scatter Plot
def plot_pairwise_scatter(data):
    features = data.select_dtypes(include=[np.number]).columns
    pair_combinations = combinations(features, 2)
    for pair in pair_combinations:
        plt.figure(figsize=(4, 3))
        sns.scatterplot(data=data, x=pair[0], y=pair[1])
        plt.title(f'Scatter Plot of {pair[0]} vs {pair[1]}')
        plt.show()

# 4. Class-wise Visualization (Pairwise plots for each class)
def plot_classwise_pairwise(data, class_column):
    classes = data[class_column].unique()
    features = data.select_dtypes(include=[np.number]).columns
    pair_combinations = combinations(features, 2)

    for cls in classes:
        class_data = data[data[class_column] == cls]
        for pair in pair_combinations:
            plt.figure(figsize=(4, 3))
```

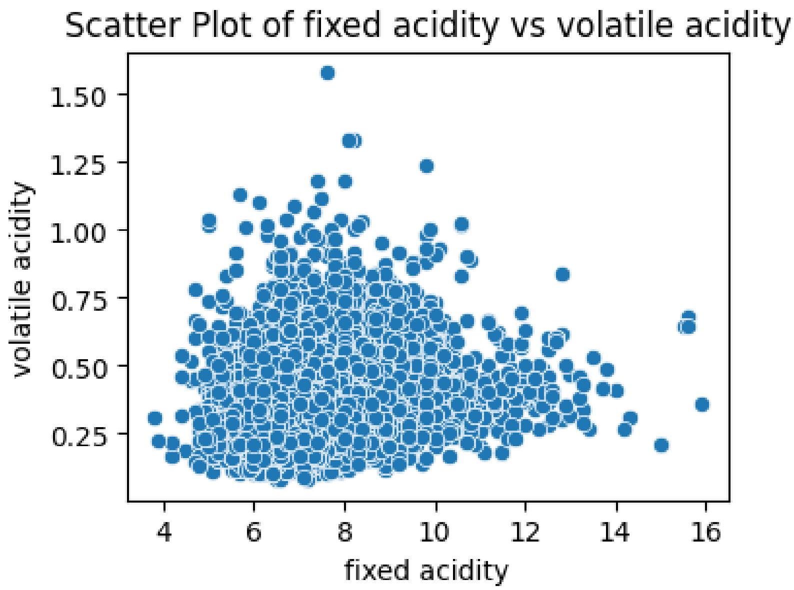
```
            sns.scatterplot(data=class_data, x=pair[0], y=pair[1])
            plt.title(f'Class {cls}: Scatter Plot of {pair[0]} vs {pair[1]}')
            plt.show()

# Call the functions to plot the graphs
plot_histograms(combined_wine_data)
```

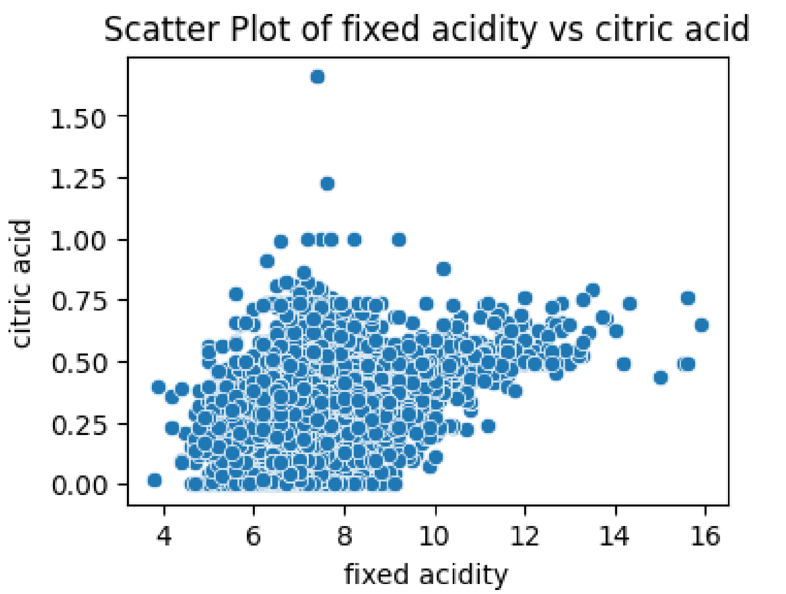


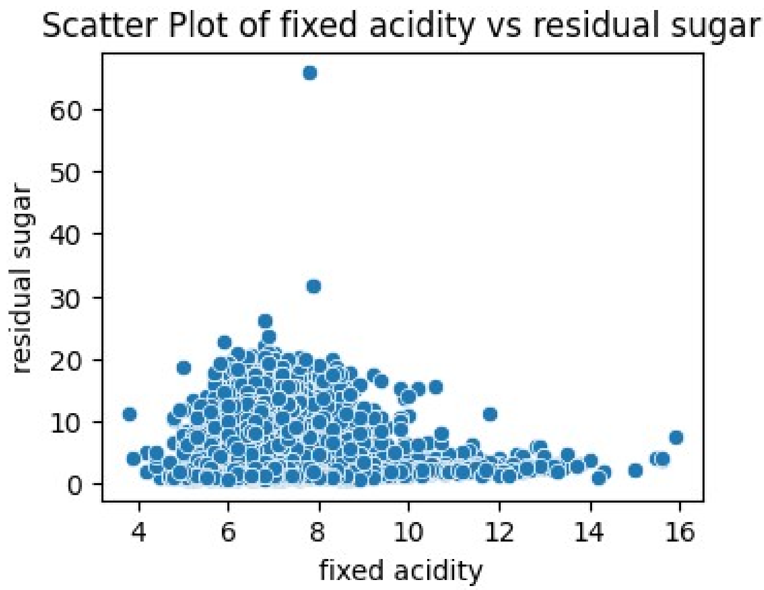

```
plot_boxplots(combined_wine_data)
```

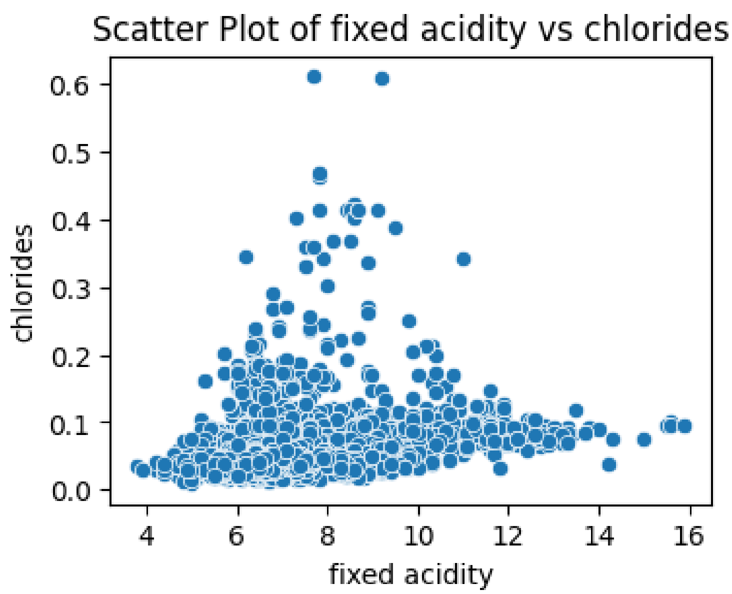


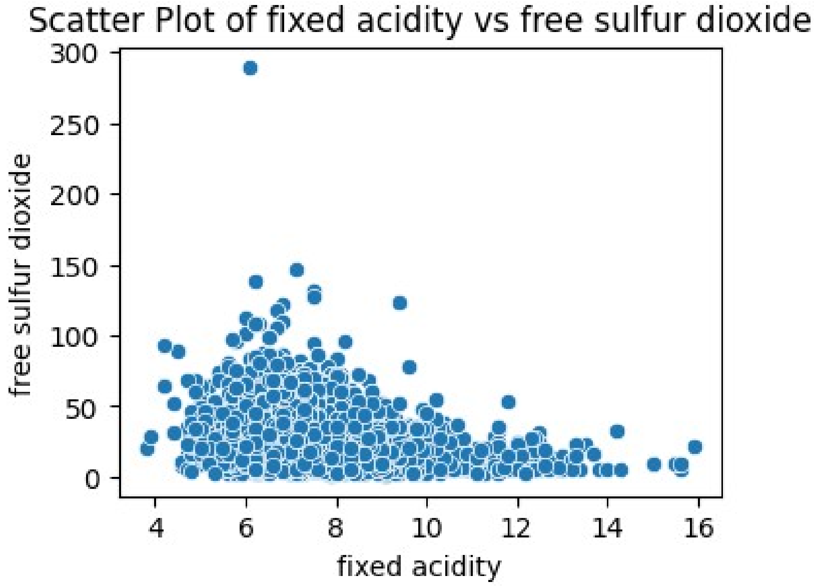

```
plot_pairwise_scatter(combined_wine_data)
```

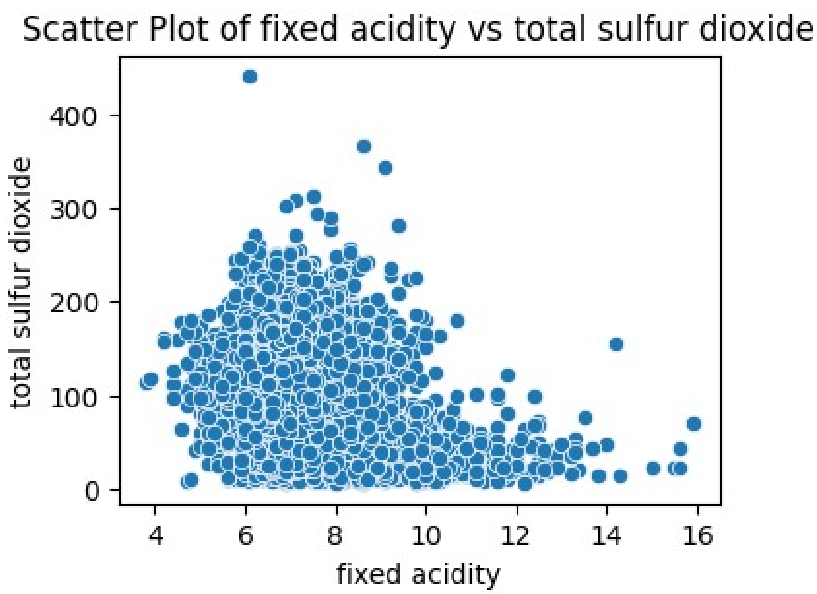

Scatter Plot of fixed acidity vs volatile acidity
volatile acidity
fixed acidity
1.50
1.25
1.00
0.75
0.50
0.25
4
6
8
10
12
14
16

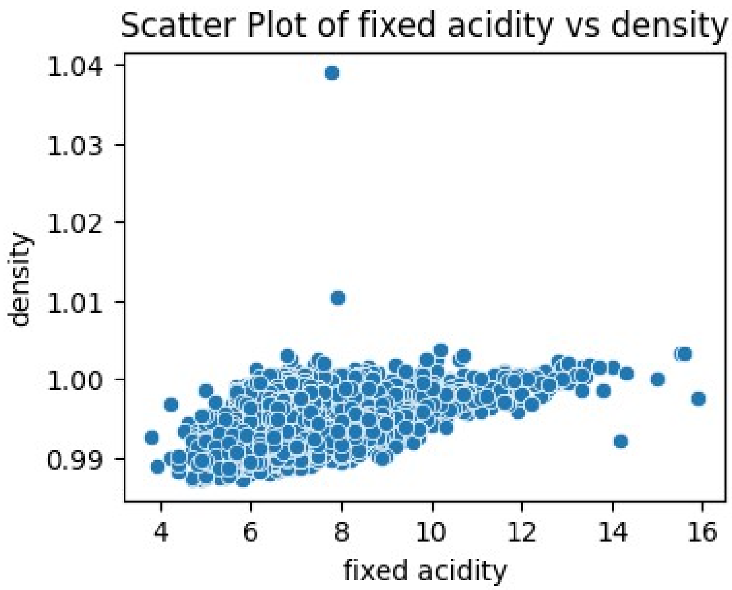

Scatter Plot of fixed acidity vs citric acid
citric acid
fixed acidity
1.50
1.25
1.00
0.75
0.50
0.25
0.00
4
6
8
10
12
14
16

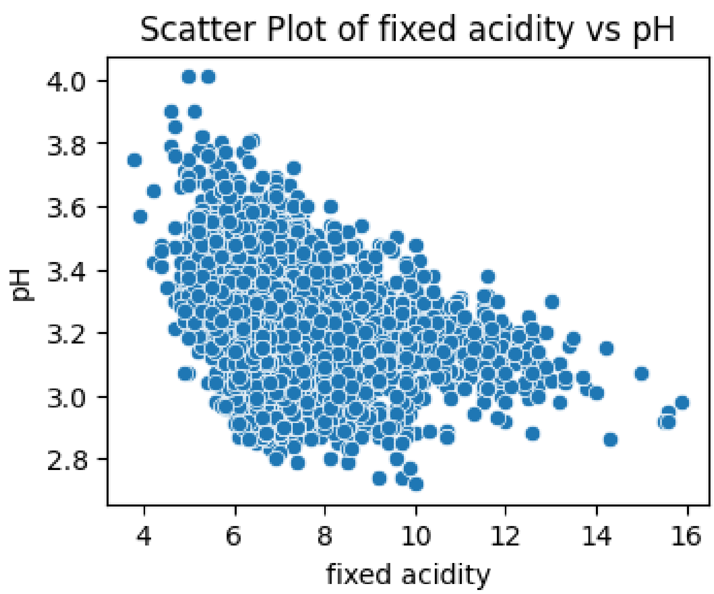

Scatter Plot of fixed acidity vs residual sugar
residual sugar
fixed acidity

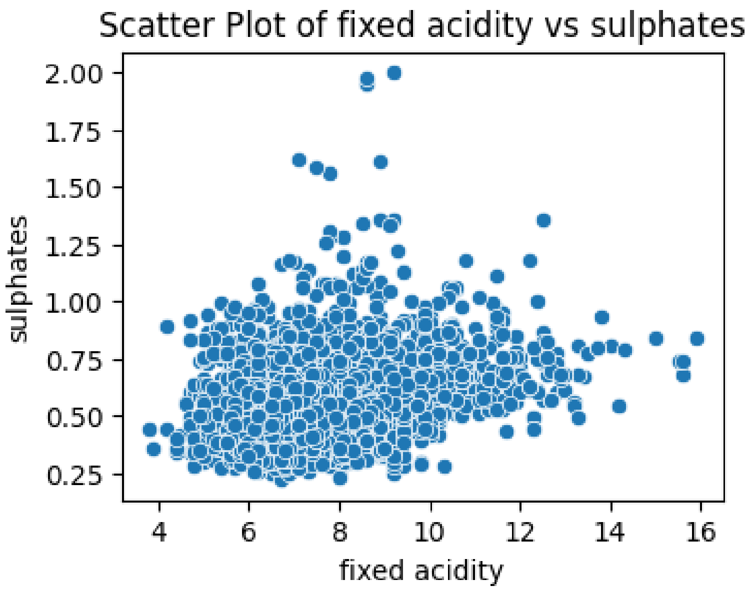

Scatter Plot of fixed acidity vs chlorides
chlorides
fixed acidity

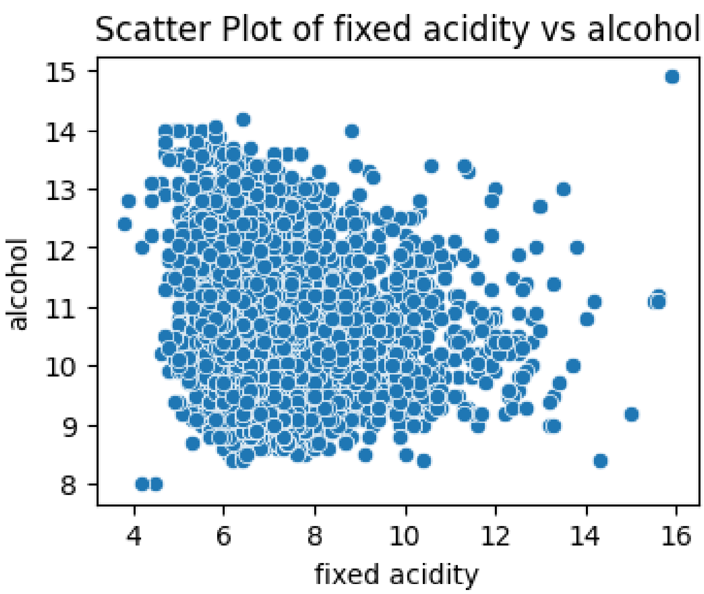

Scatter Plot of fixed acidity vs free sulfur dioxide
free sulfur dioxide
fixed acidity
0
50
100
150
200
250
300
4
6
8
10
12
14
16

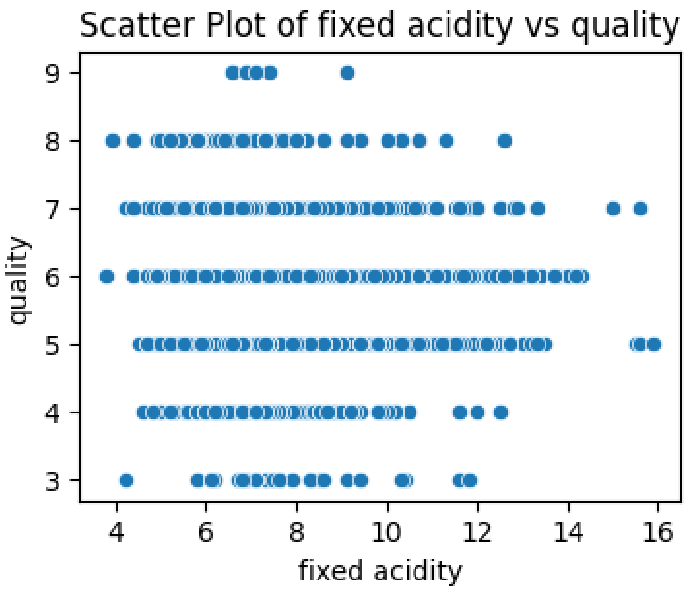

Scatter Plot of fixed acidity vs total sulfur dioxide
total sulfur dioxide
fixed acidity
0
100
200
300
400
4
6
8
10
12
14
16

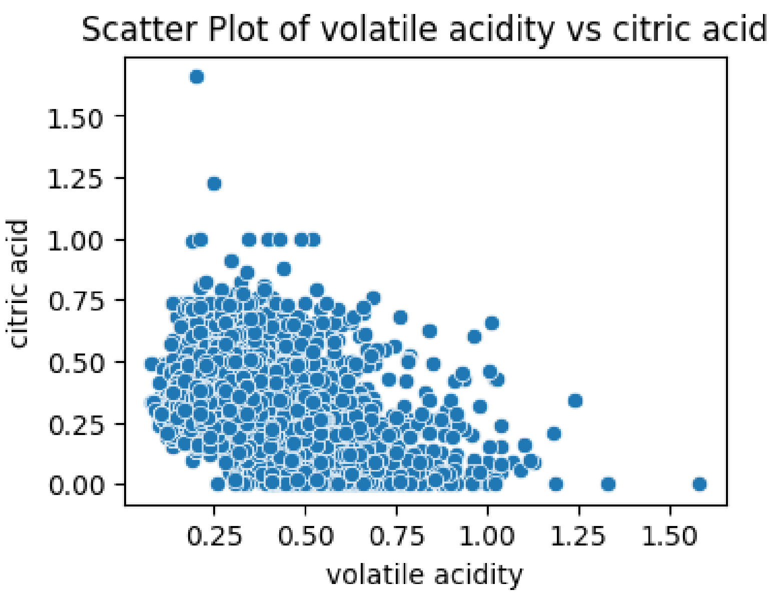

Scatter Plot of fixed acidity vs density
1.04
1.03
1.02
1.01
1.00
0.99
density
4
6
8
10
12
14
16
fixed acidity

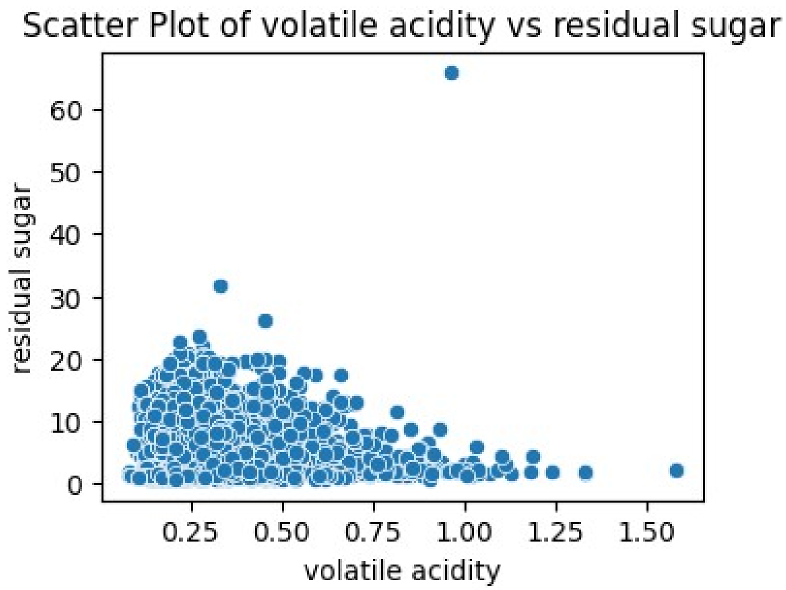

Scatter Plot of fixed acidity vs pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
pH
4
6
8
10
12
14
16
fixed acidity

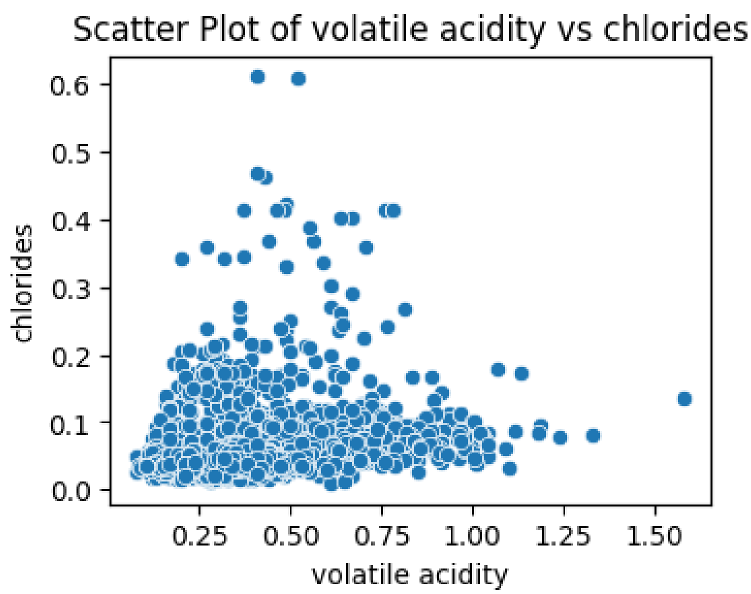

Scatter Plot of fixed acidity vs sulphates
sulphates
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
4
6
8
10
12
14
16
fixed acidity

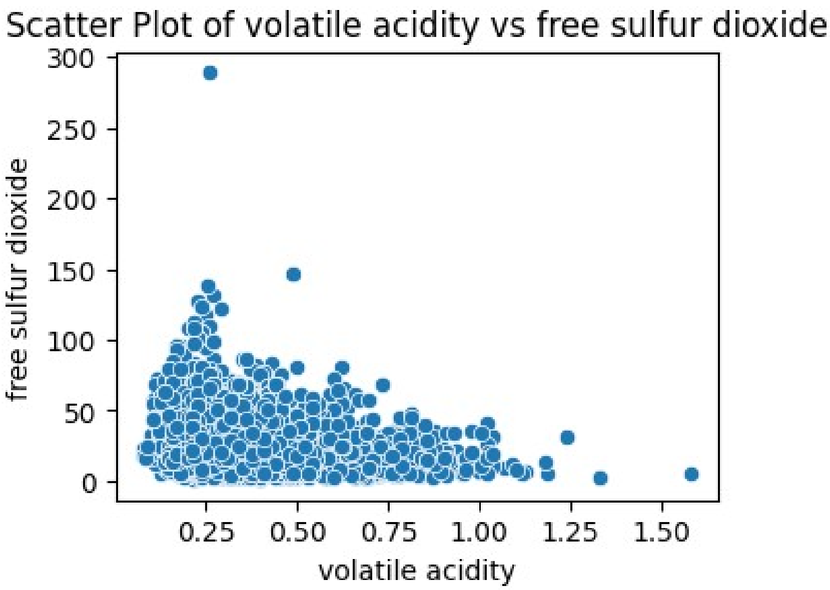

Scatter Plot of fixed acidity vs alcohol
alcohol
15
14
13
12
11
10
9
8
4
6
8
10
12
14
16
fixed acidity

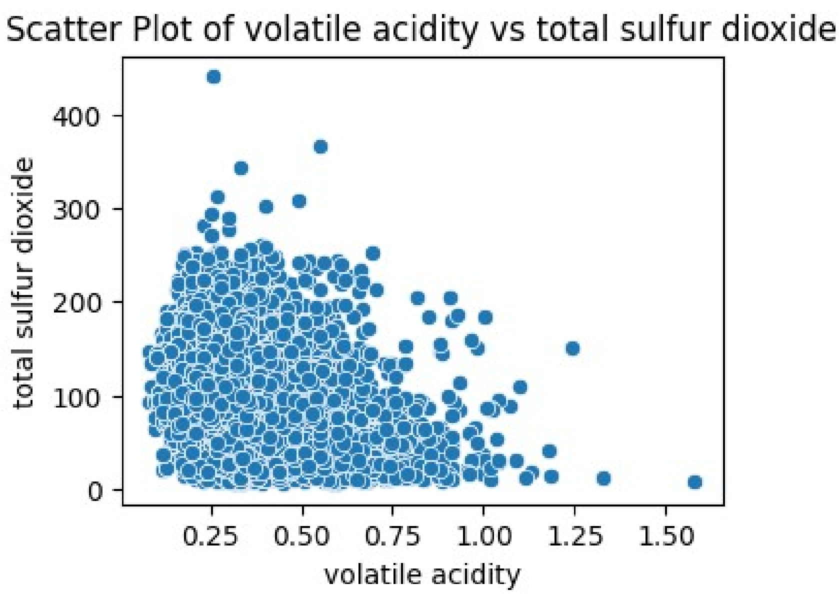

Scatter Plot of fixed acidity vs quality
quality
fixed acidity
9
8
7
6
5
4
3
4
6
8
10
12
14
16

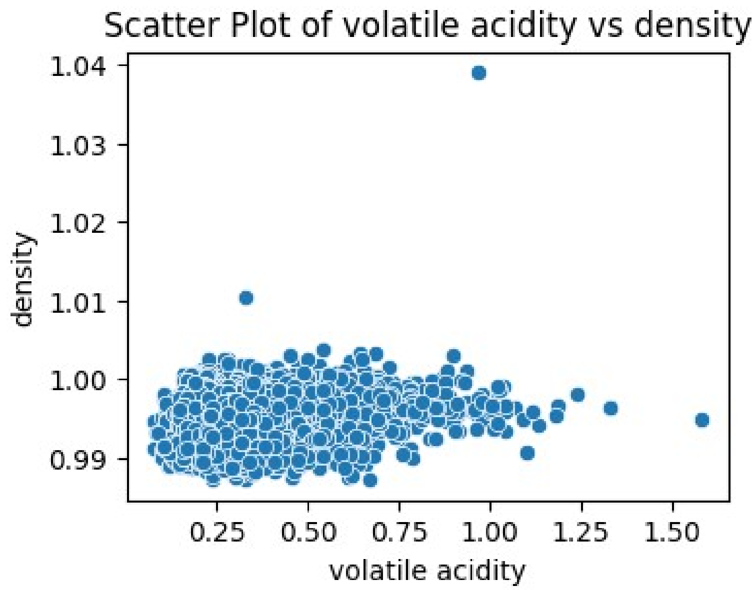

Scatter Plot of volatile acidity vs citric acid
citric acid
volatile acidity
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50

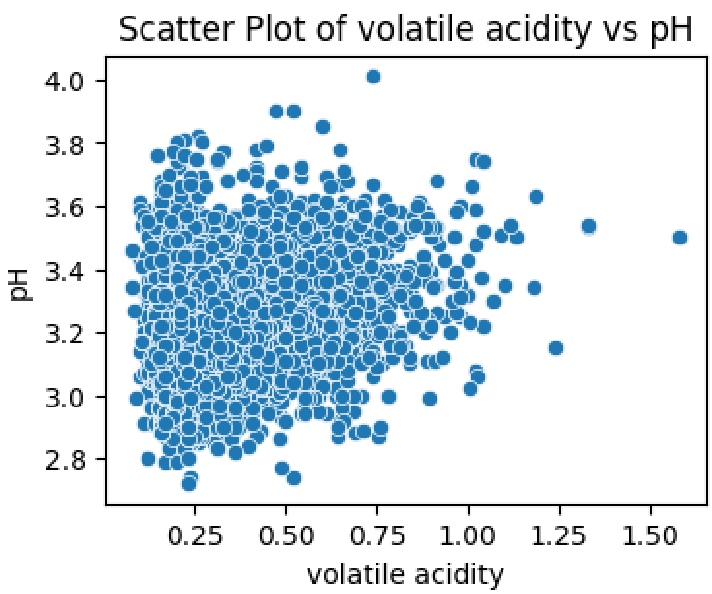

Scatter Plot of volatile acidity vs residual sugar
residual sugar
volatile acidity
0
10
20
30
40
50
60
0.25
0.50
0.75
1.00
1.25
1.50

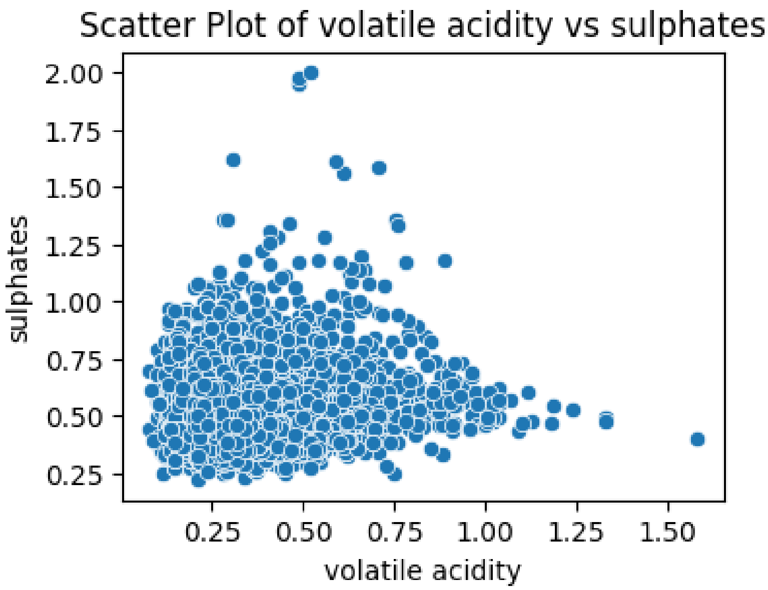

Scatter Plot of volatile acidity vs chlorides
chlorides
volatile acidity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.25
0.50
0.75
1.00
1.25
1.50

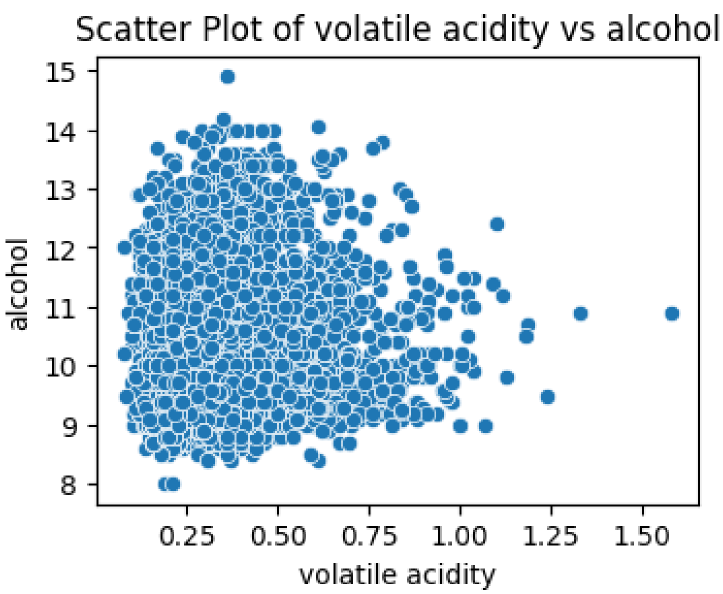

Scatter Plot of volatile acidity vs free sulfur dioxide
free sulfur dioxide
volatile acidity
0
50
100
150
200
250
300
0.25
0.50
0.75
1.00
1.25
1.50

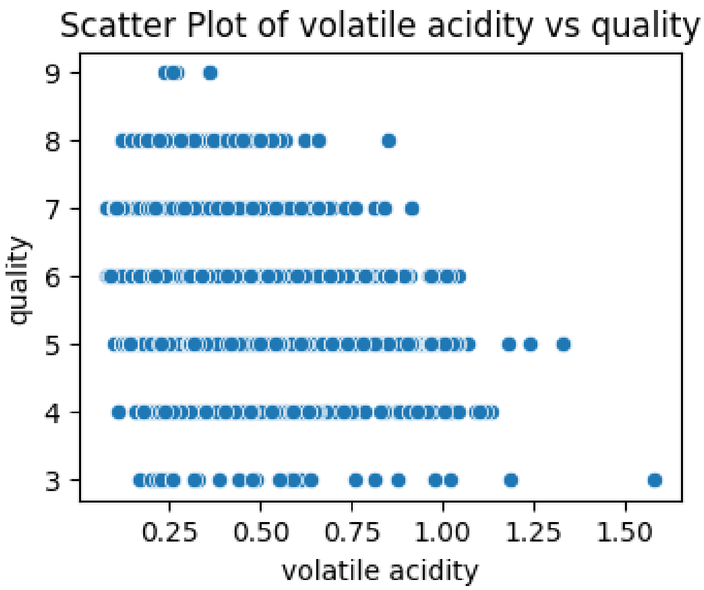

Scatter Plot of volatile acidity vs total sulfur dioxide
total sulfur dioxide
volatile acidity
0
100
200
300
400
0.25
0.50
0.75
1.00
1.25
1.50

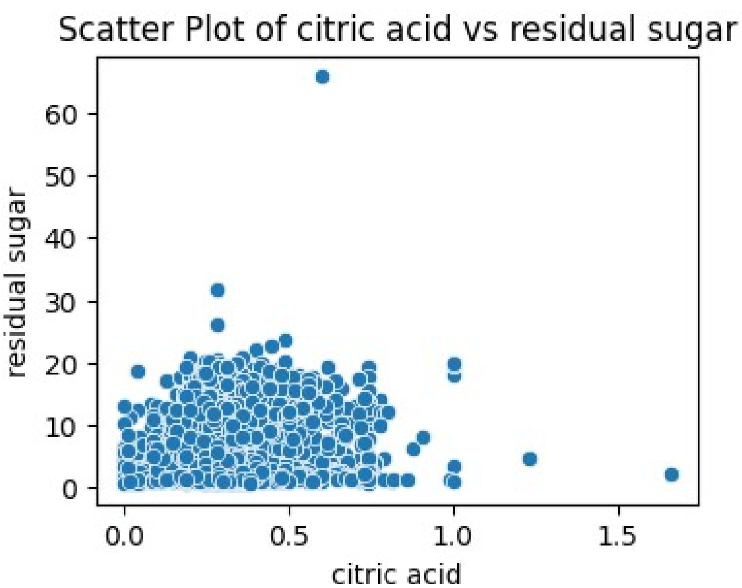

Scatter Plot of volatile acidity vs density
density
1.04
1.03
1.02
1.01
1.00
0.99
0.25
0.50
0.75
1.00
1.25
1.50
volatile acidity

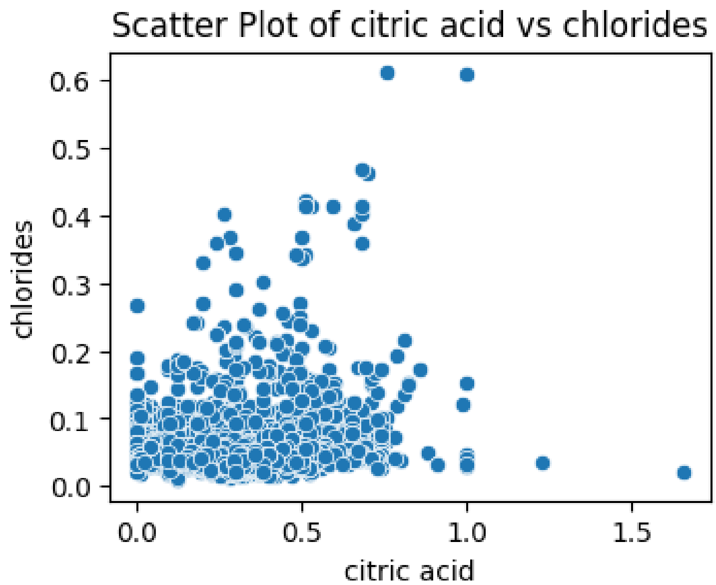

Scatter Plot of volatile acidity vs pH
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
0.25
0.50
0.75
1.00
1.25
1.50
volatile acidity

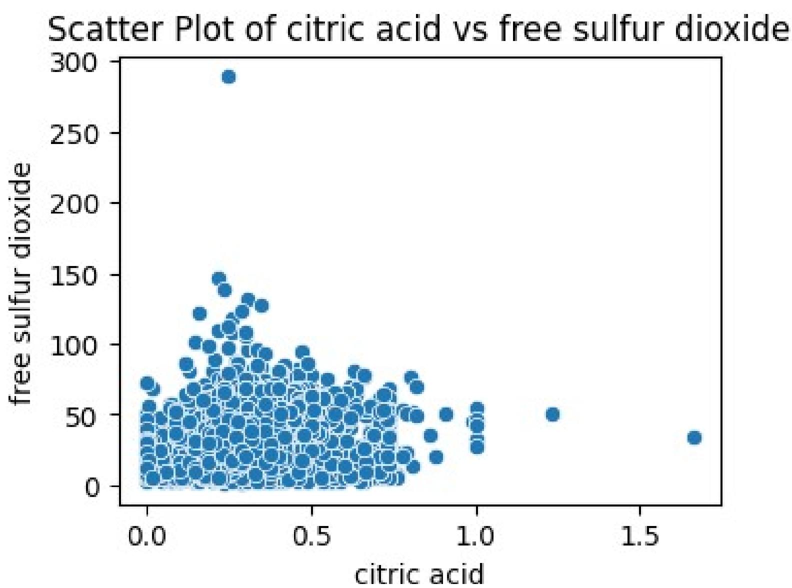

Scatter Plot of volatile acidity vs sulphates
sulphates
volatile acidity
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.25
0.50
0.75
1.00
1.25
1.50

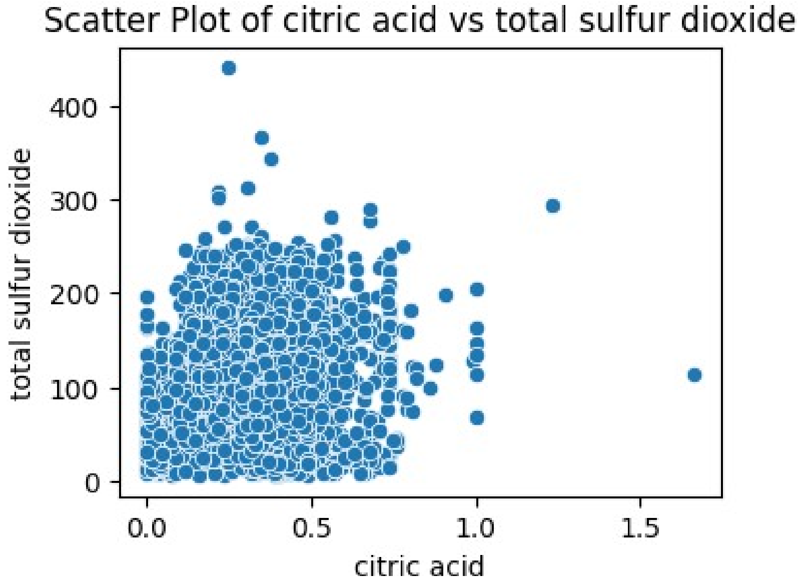

Scatter Plot of volatile acidity vs alcohol
alcohol
volatile acidity
15
14
13
12
11
10
9
8
0.25
0.50
0.75
1.00
1.25
1.50

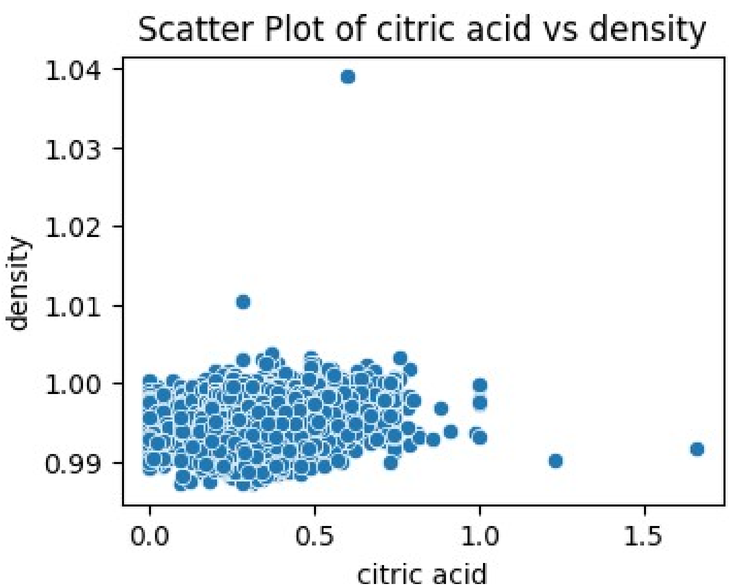

Scatter Plot of volatile acidity vs quality
quality
9
8
7
6
5
4
3
0.25
0.50
0.75
1.00
1.25
1.50
volatile acidity

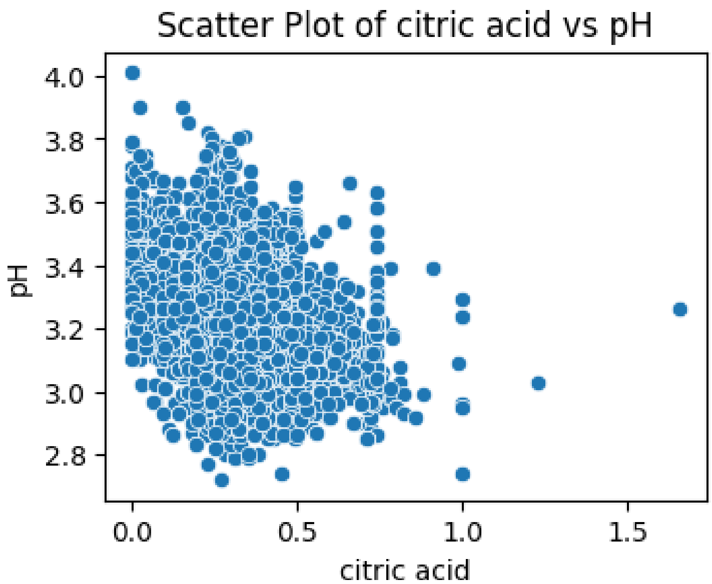

Scatter Plot of citric acid vs residual sugar
residual sugar
60
50
40
30
20
10
0
0.0
0.5
1.0
1.5
citric acid


Scatter Plot of citric acid vs chlorides
chlorides
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.5
1.0
1.5
citric acid

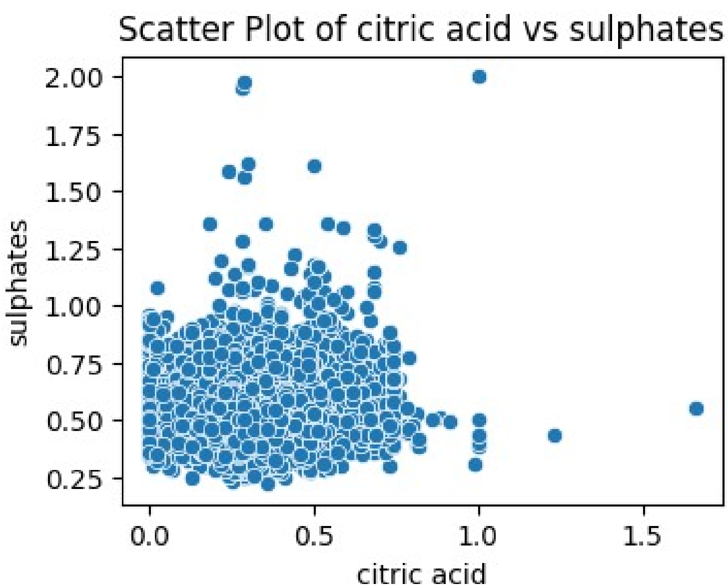

Scatter Plot of citric acid vs free sulfur dioxide
free sulfur dioxide
0
50
100
150
200
250
300
0.0
0.5
1.0
1.5
citric acid


Scatter Plot of citric acid vs total sulfur dioxide
total sulfur dioxide
400
300
200
100
0
0.0
0.5
1.0
1.5
citric acid

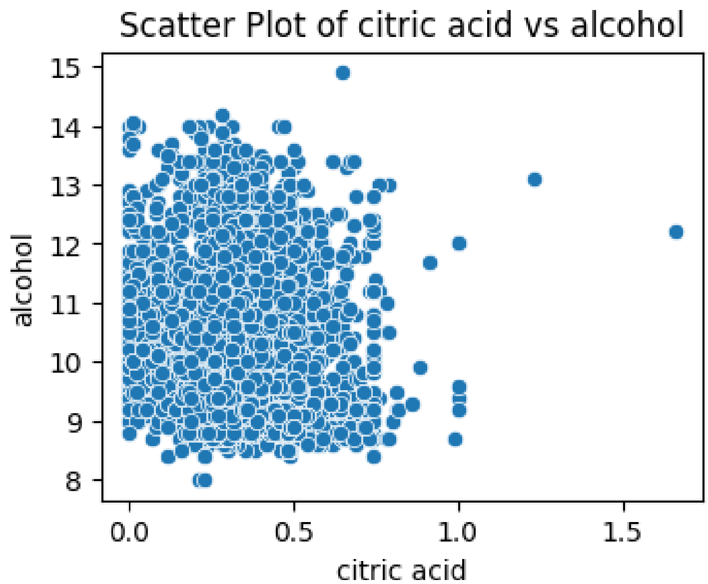

Scatter Plot of citric acid vs density
density
1.04
1.03
1.02
1.01
1.00
0.99
0.0
0.5
1.0
1.5
citric acid

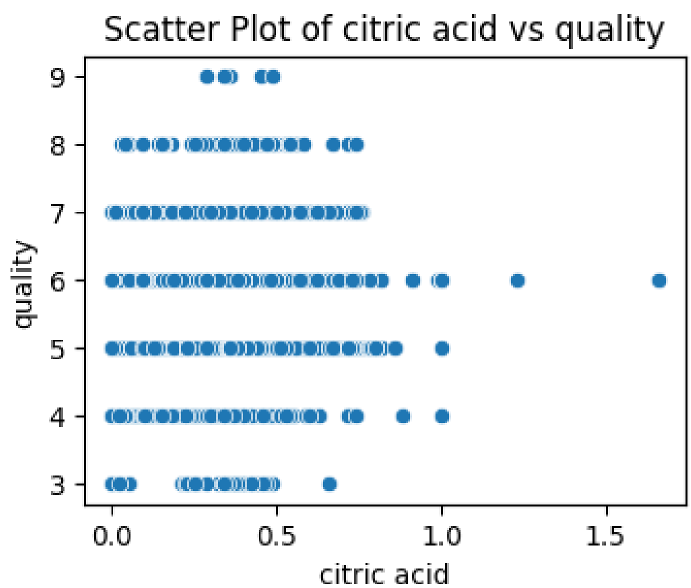


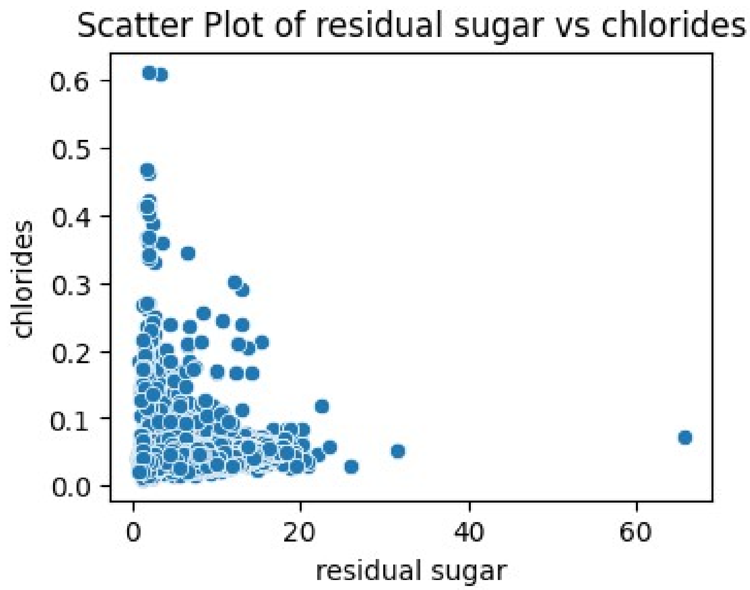
Scatter Plot of citric acid vs pH
pH
citric acid

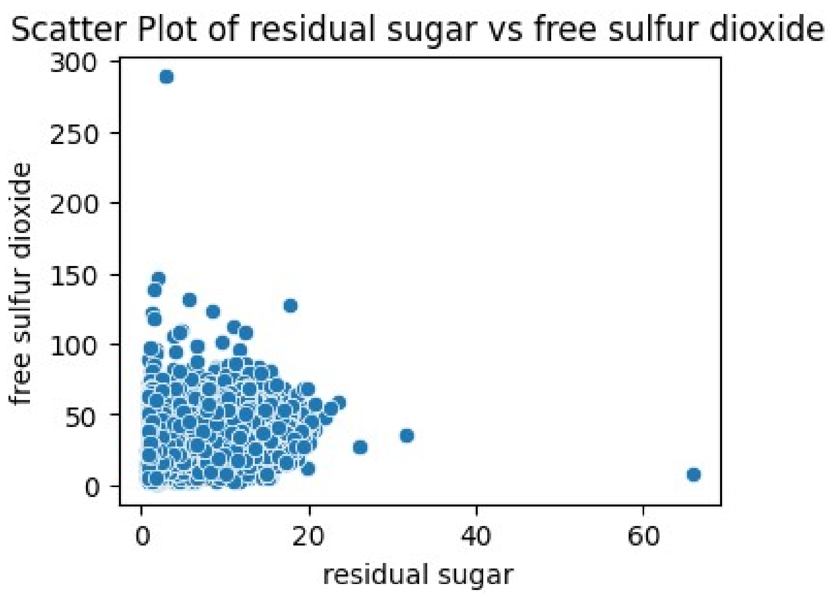


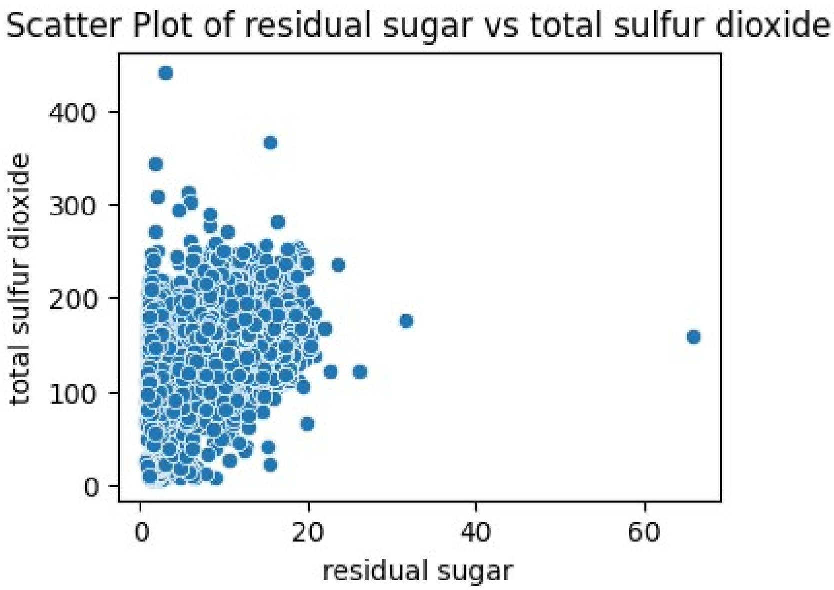
Scatter Plot of citric acid vs sulphates
sulphates
citric acid

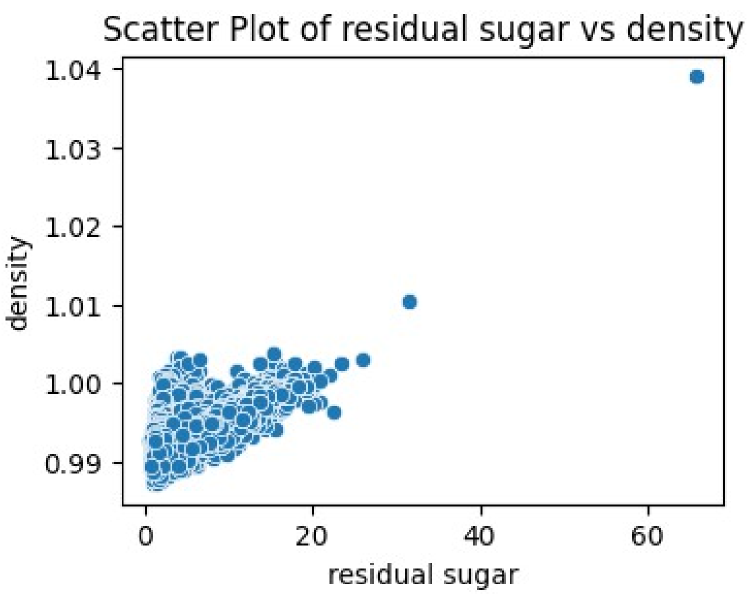

Scatter Plot of citric acid vs alcohol
alcohol
citric acid
15
14
13
12
11
10
9
8
0.0
0.5
1.0
1.5

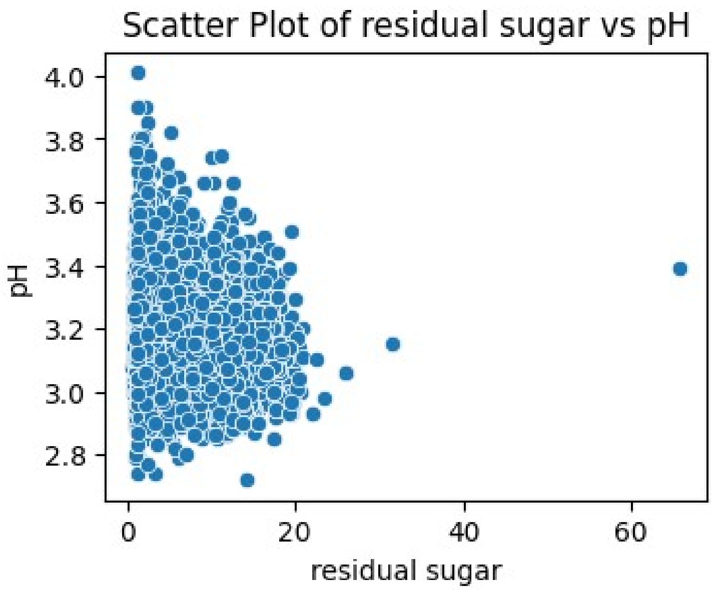

Scatter Plot of citric acid vs quality
quality
citric acid
9
8
7
6
5
4
3
0.0
0.5
1.0
1.5

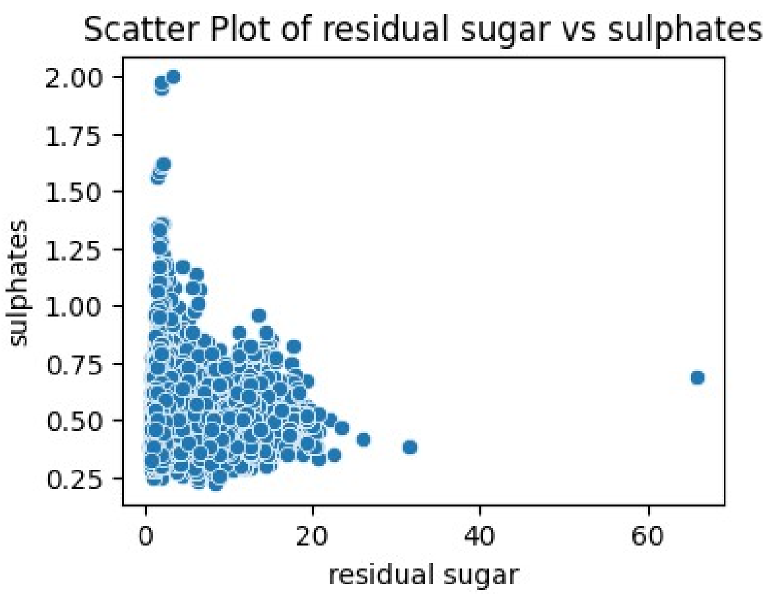

Scatter Plot of residual sugar vs chlorides
0.6
0.5
0.4
0.3
0.2
0.1
0.0
chlorides
0
20
40
60
residual sugar

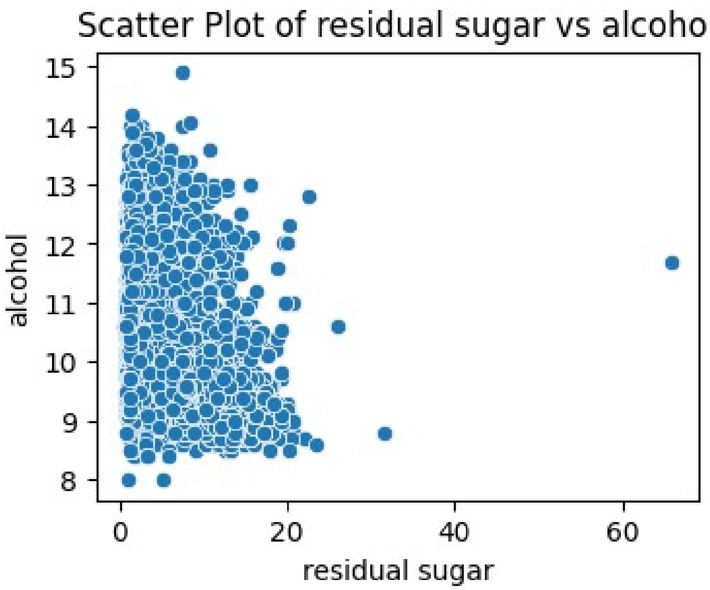

Scatter Plot of residual sugar vs free sulfur dioxide
300
250
200
150
100
50
0
free sulfur dioxide
0
20
40
60
residual sugar

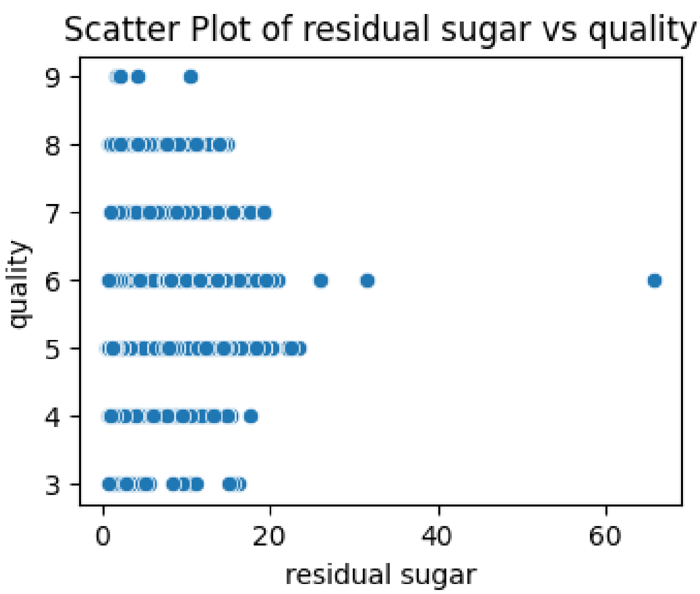

Scatter Plot of residual sugar vs total sulfur dioxide
total sulfur dioxide
residual sugar
0
100
200
300
400
0
20
40
60

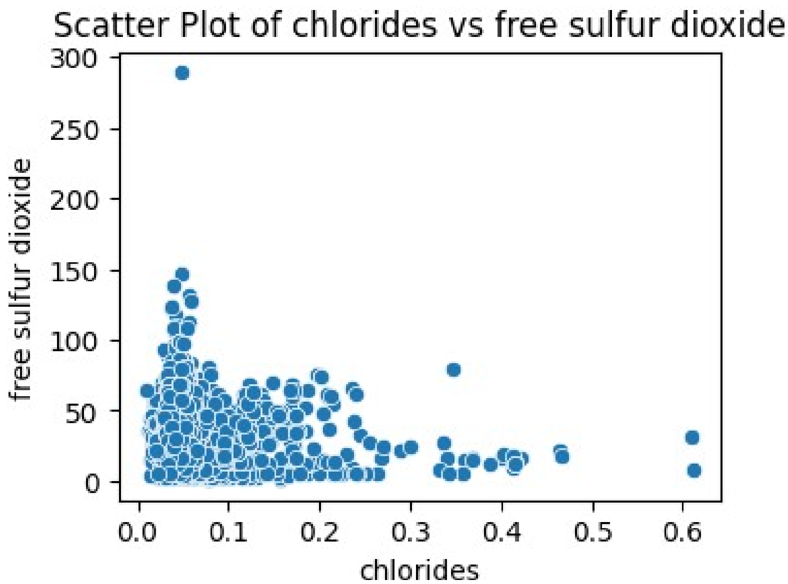

Scatter Plot of residual sugar vs density
density
residual sugar
0.99
1.00
1.01
1.02
1.03
1.04
0
20
40
60

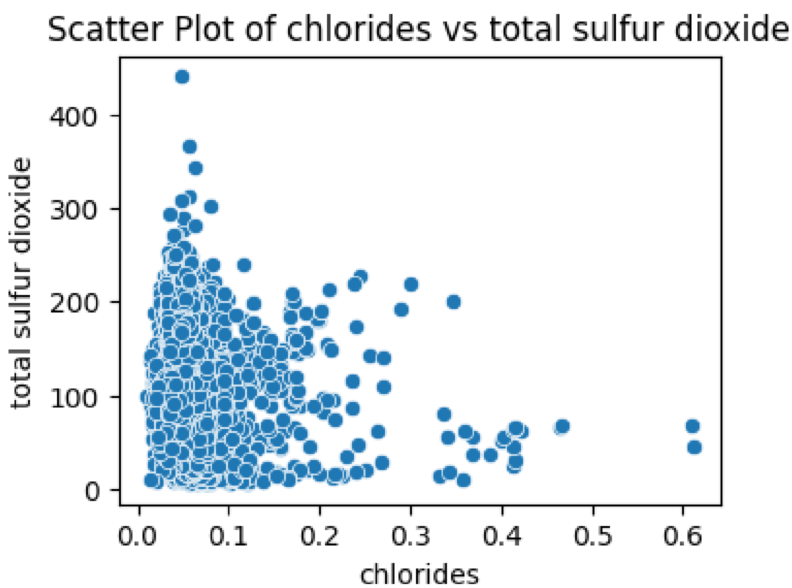

Scatter Plot of residual sugar vs pH
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
0
20
40
60
residual sugar

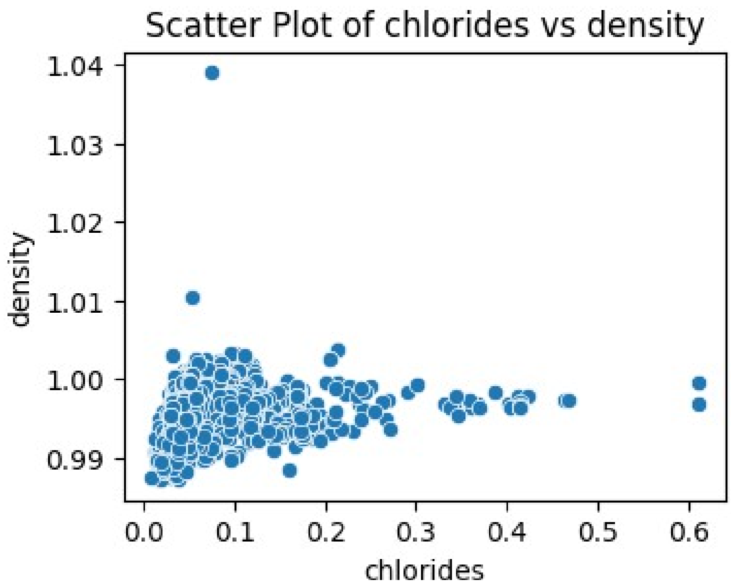

Scatter Plot of residual sugar vs sulphates
sulphates
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
20
40
60
residual sugar

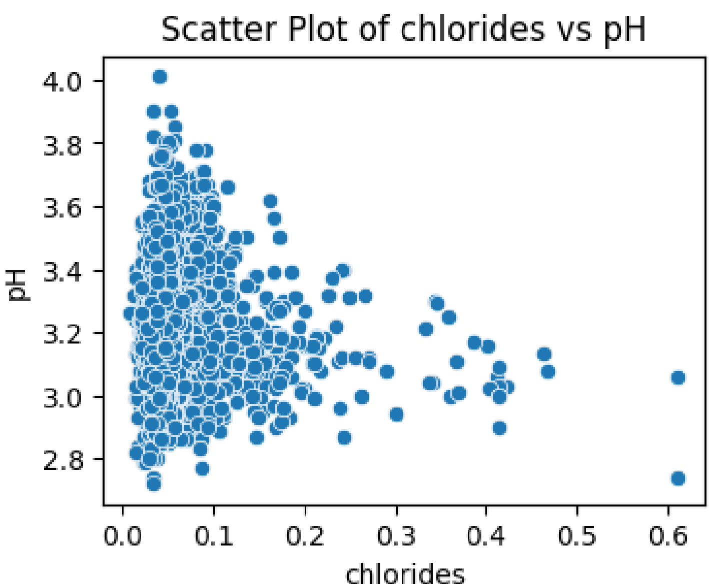

Scatter Plot of residual sugar vs alcohol
alcohol
residual sugar
15
14
13
12
11
10
9
8
0
20
40
60

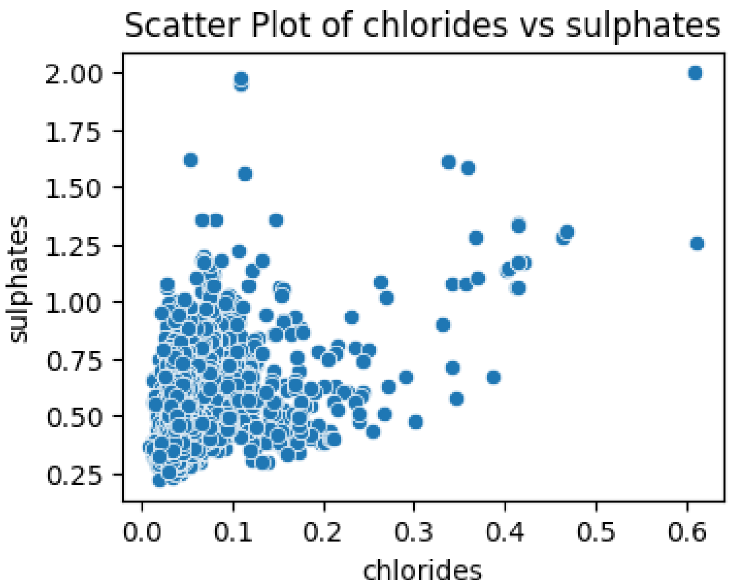

Scatter Plot of residual sugar vs quality
quality
residual sugar
9
8
7
6
5
4
3
0
20
40
60

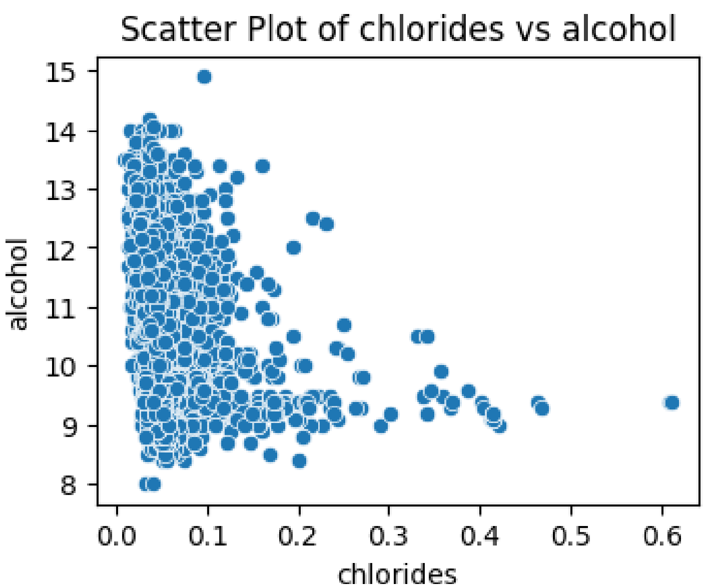

Scatter Plot of chlorides vs free sulfur dioxide
free sulfur dioxide
0
50
100
150
200
250
300
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides

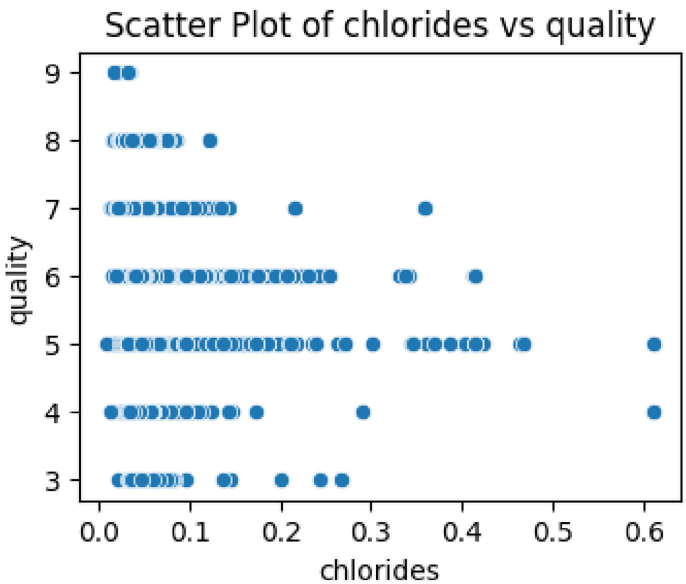

Scatter Plot of chlorides vs total sulfur dioxide
total sulfur dioxide
0
100
200
300
400
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides


Scatter Plot of chlorides vs density
1.04
1.03
1.02
1.01
1.00
0.99
density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides

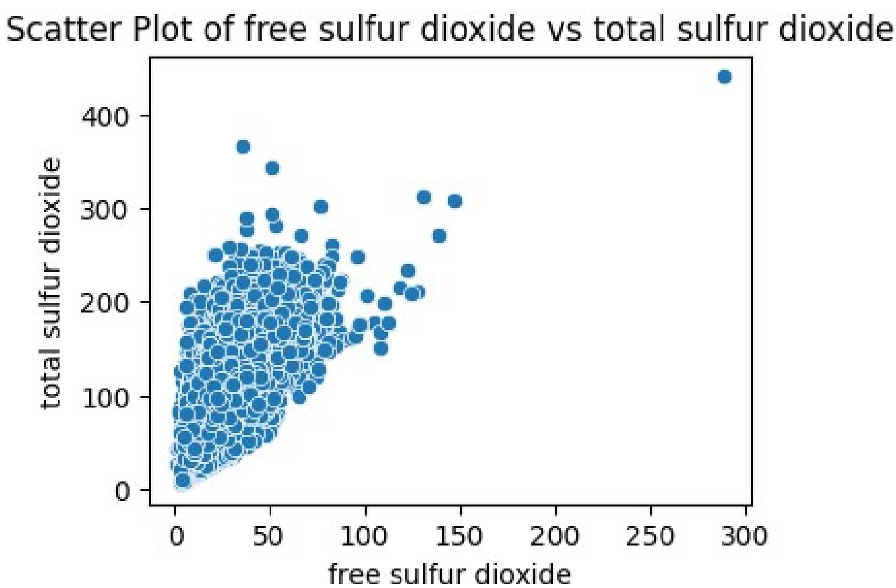

Scatter Plot of chlorides vs pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
pH
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides


Scatter Plot of chlorides vs sulphates
sulphates
chlorides

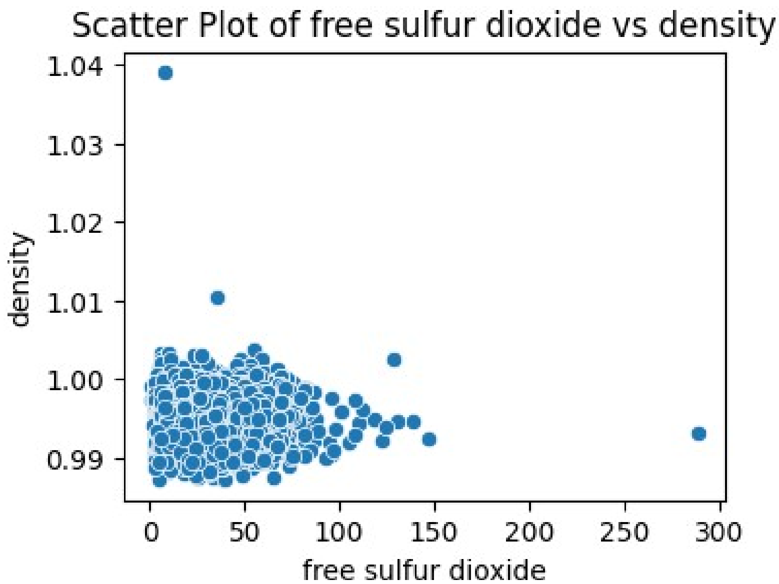

Scatter Plot of chlorides vs alcohol
alcohol
chlorides

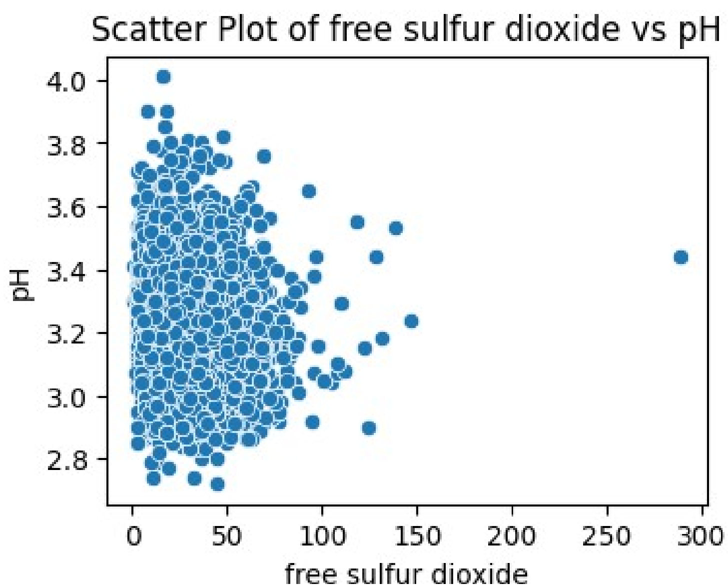


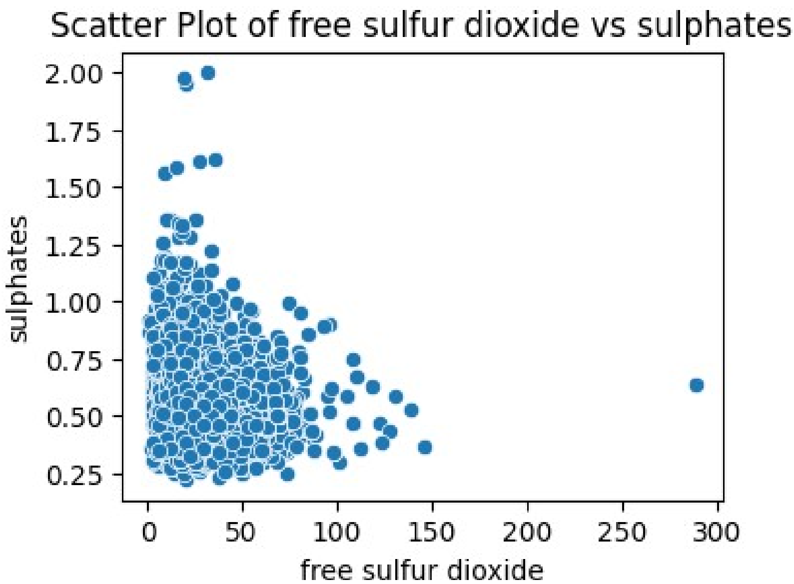
Scatter Plot of chlorides vs quality
quality
chlorides

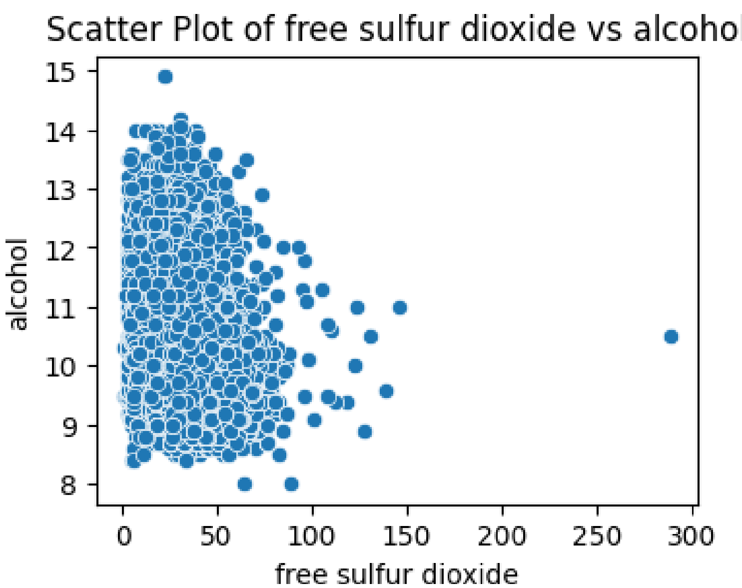


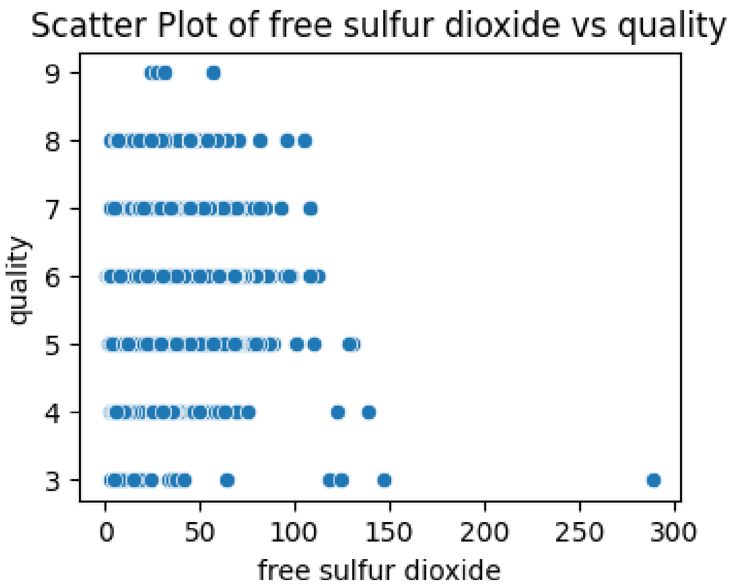
Scatter Plot of free sulfur dioxide vs total sulfur dioxide
total sulfur dioxide
free sulfur dioxide


Scatter Plot of free sulfur dioxide vs density
density
1.04
1.03
1.02
1.01
1.00
0.99
0
50
100
150
200
250
300
free sulfur dioxide

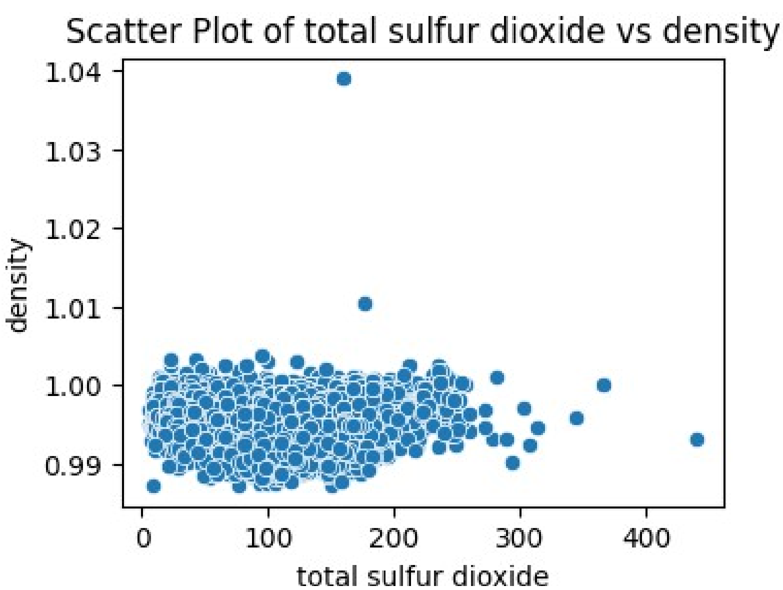

Scatter Plot of free sulfur dioxide vs pH
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
0
50
100
150
200
250
300
free sulfur dioxide


Scatter Plot of free sulfur dioxide vs sulphates
sulphates
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
50
100
150
200
250
300
free sulfur dioxide

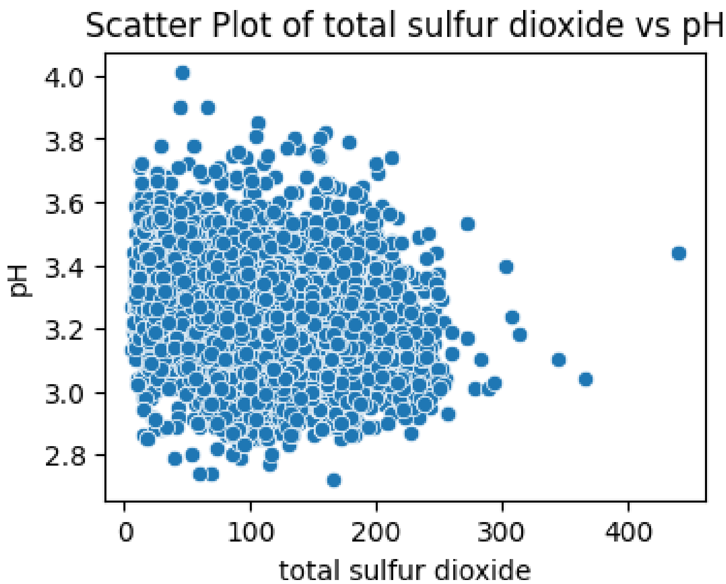

Scatter Plot of free sulfur dioxide vs alcohol
alcohol
15
14
13
12
11
10
9
8
0
50
100
150
200
250
300
free sulfur dioxide

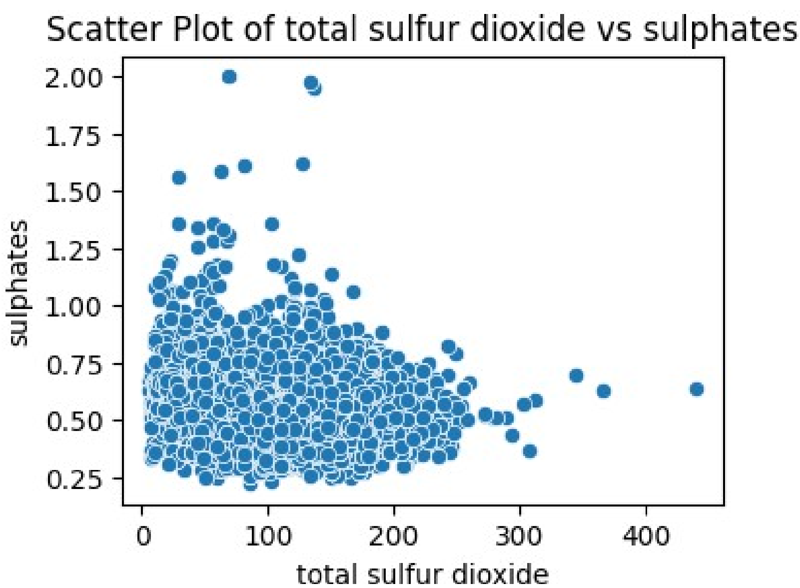


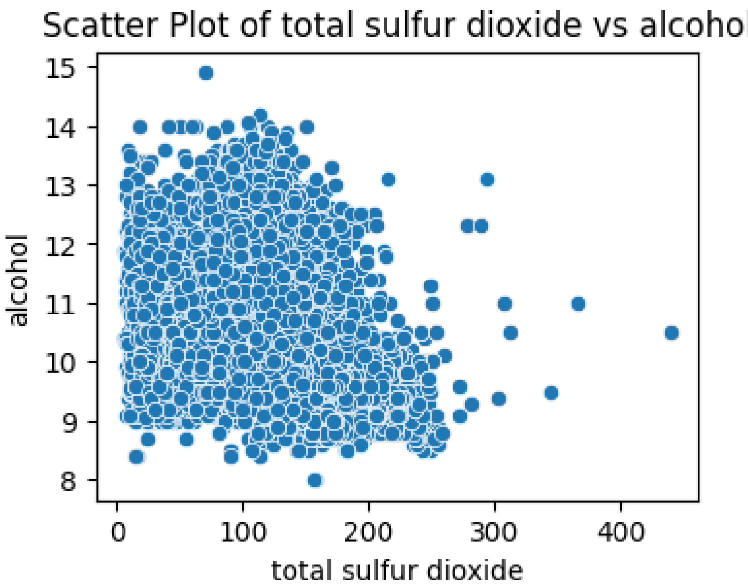
Scatter Plot of free sulfur dioxide vs quality
quality
9
8
7
6
5
4
3
0
50
100
150
200
250
300
free sulfur dioxide

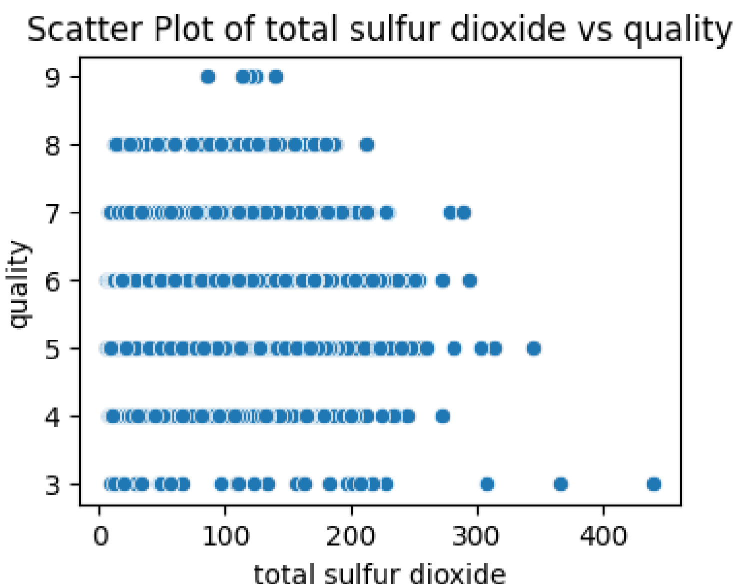


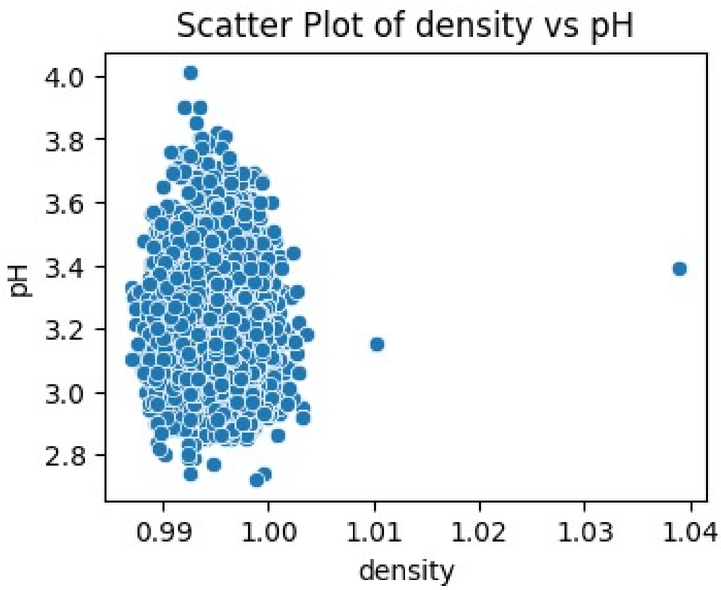
Scatter Plot of total sulfur dioxide vs density
density
1.04
1.03
1.02
1.01
1.00
0.99
0
100
200
300
400
total sulfur dioxide

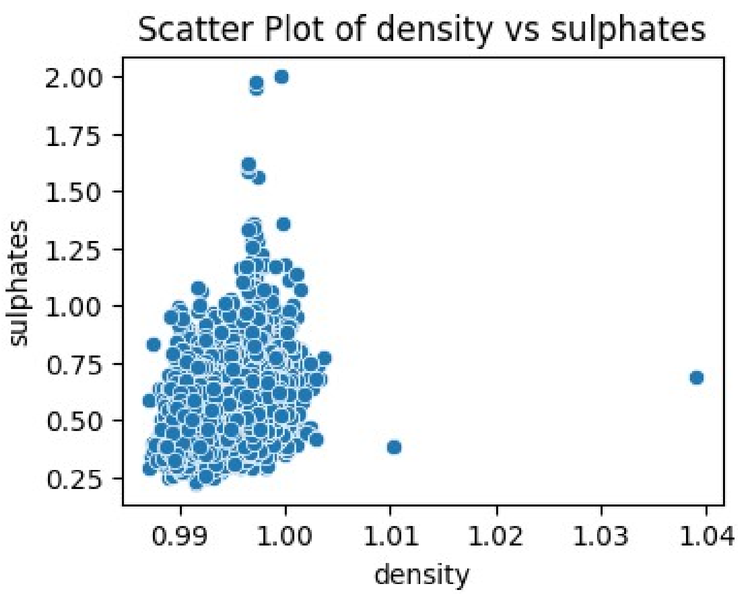

Scatter Plot of total sulfur dioxide vs pH
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
0
100
200
300
400
total sulfur dioxide

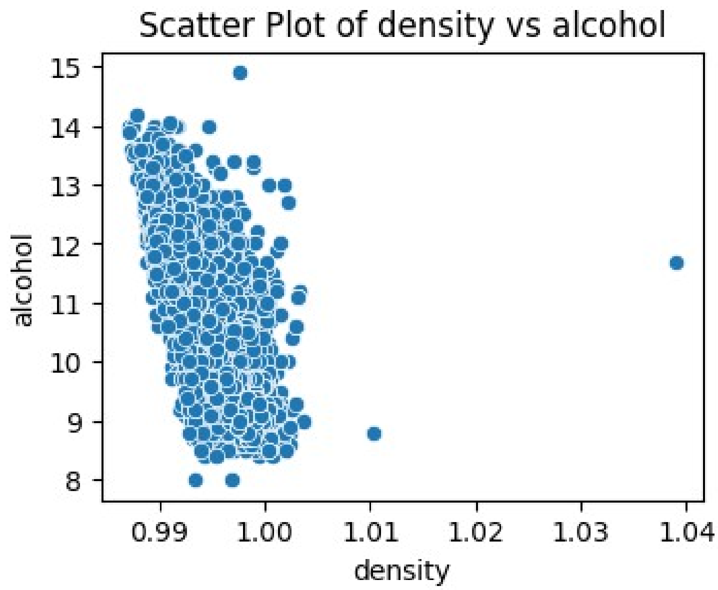

Scatter Plot of total sulfur dioxide vs sulphates
sulphates
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
100
200
300
400
total sulfur dioxide

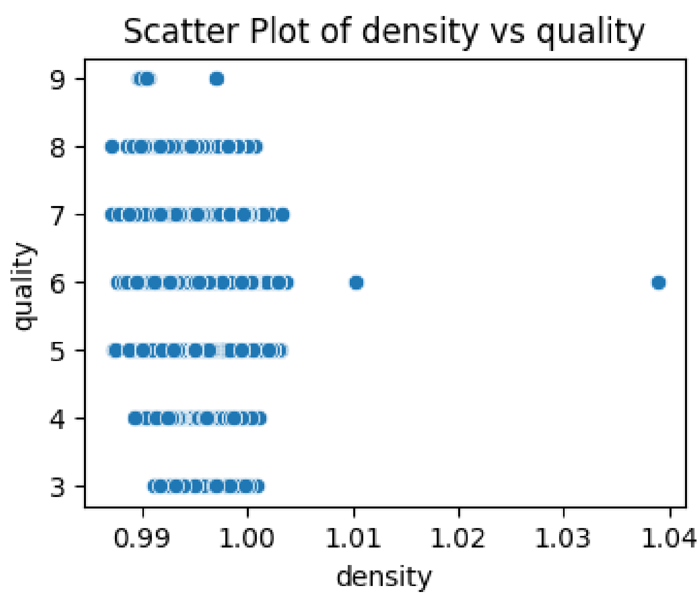

Scatter Plot of total sulfur dioxide vs alcohol
alcohol
total sulfur dioxide
15
14
13
12
11
10
9
8
0
100
200
300
400

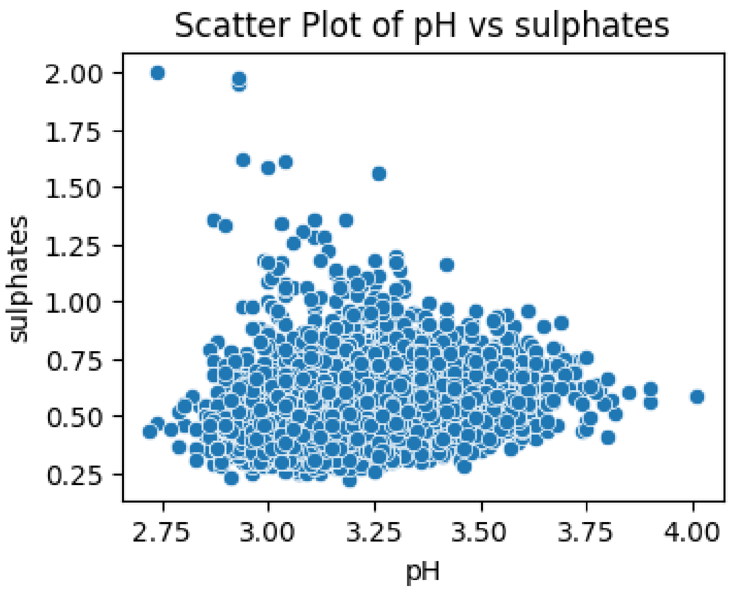

Scatter Plot of total sulfur dioxide vs quality
quality
total sulfur dioxide
9
8
7
6
5
4
3
0
100
200
300
400

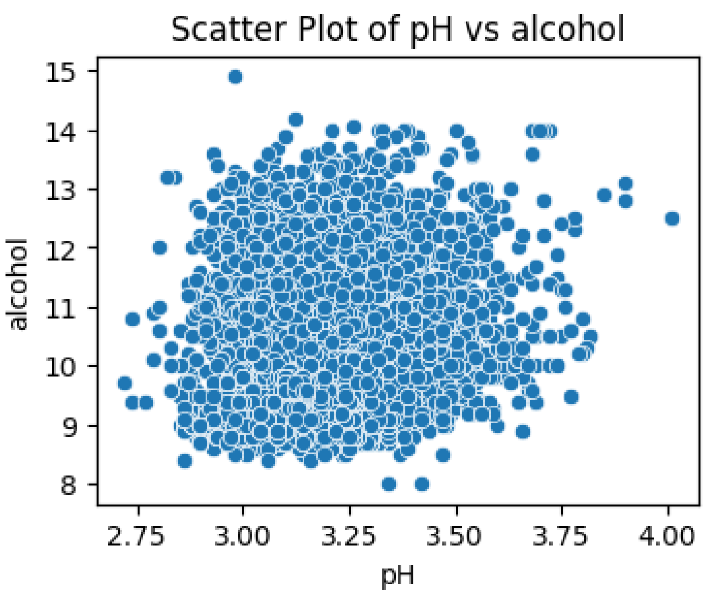

Scatter Plot of density vs pH
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
0.99
1.00
1.01
1.02
1.03
1.04
density

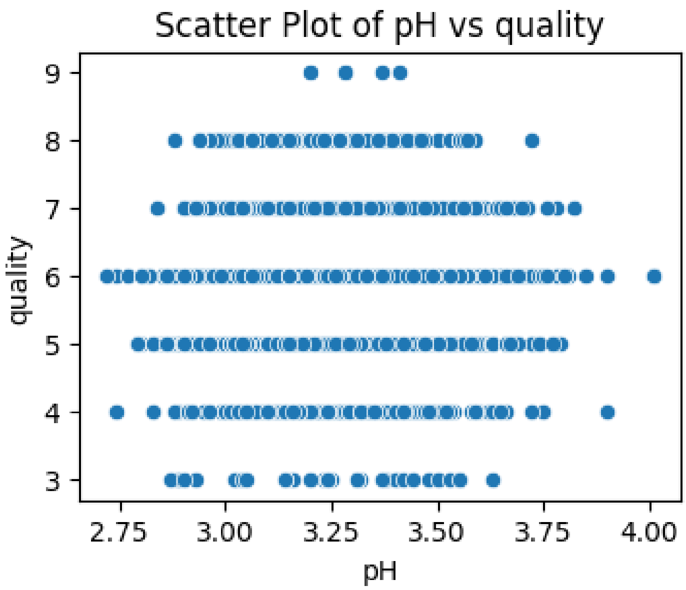

Scatter Plot of density vs sulphates
sulphates
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.99
1.00
1.01
1.02
1.03
1.04
density


Scatter Plot of density vs alcohol
alcohol
density
15
14
13
12
11
10
9
8
0.99
1.00
1.01
1.02
1.03
1.04

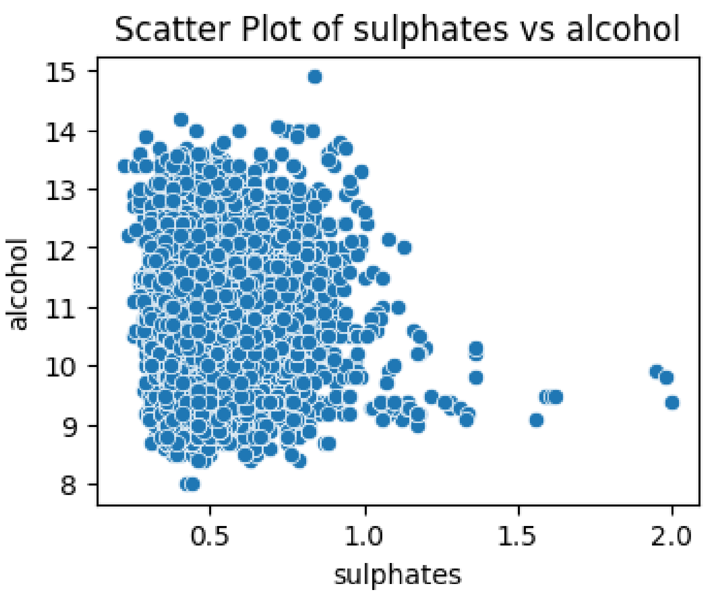

Scatter Plot of density vs quality
quality
density
9
8
7
6
5
4
3
0.99
1.00
1.01
1.02
1.03
1.04


Scatter Plot of pH vs sulphates
sulphates
pH
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
2.75
3.00
3.25
3.50
3.75
4.00

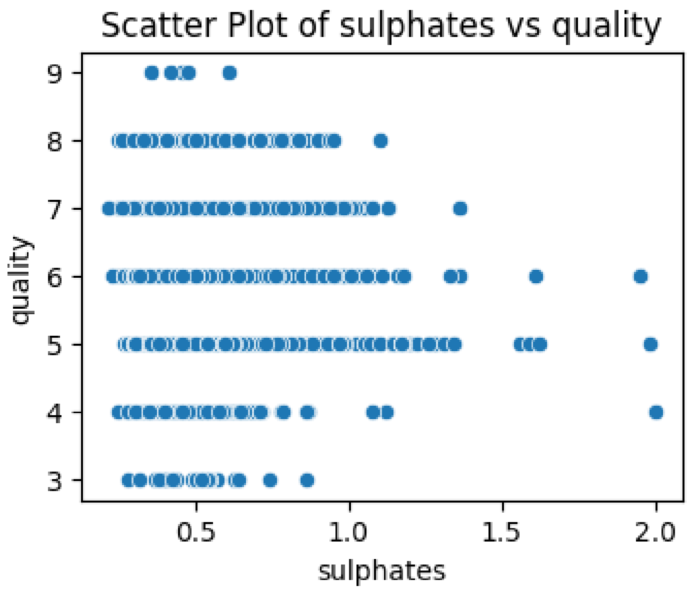

Scatter Plot of pH vs alcohol
alcohol
pH
15
14
13
12
11
10
9
8
2.75
3.00
3.25
3.50
3.75
4.00



Scatter Plot of pH vs quality
quality
9
8
7
6
5
4
3
2.75
3.00
3.25
3.50
3.75
4.00
pH

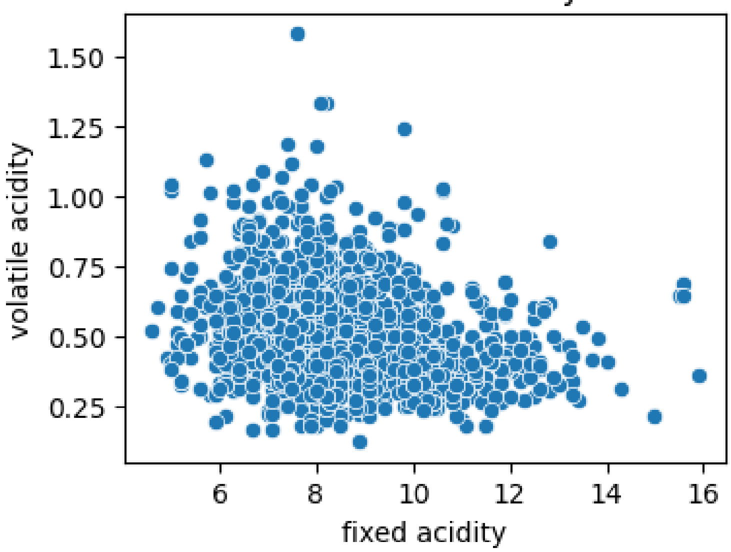


Scatter Plot of sulphates vs alcohol
alcohol
15
14
13
12
11
10
9
8
0.5
1.0
1.5
2.0
sulphates

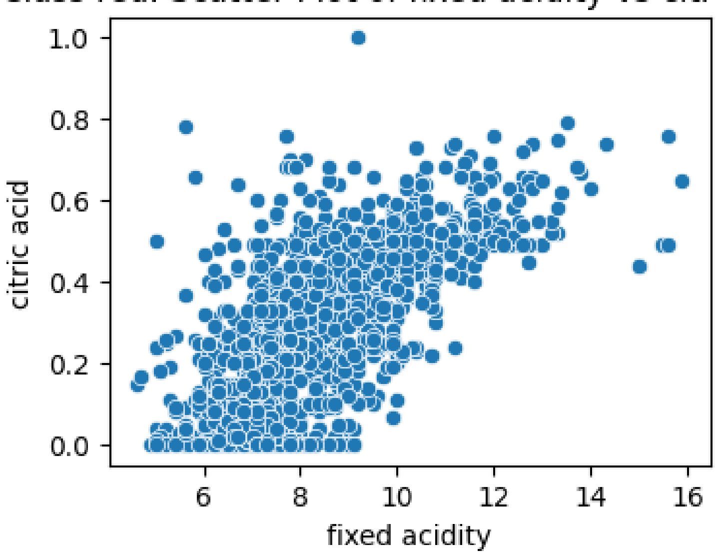


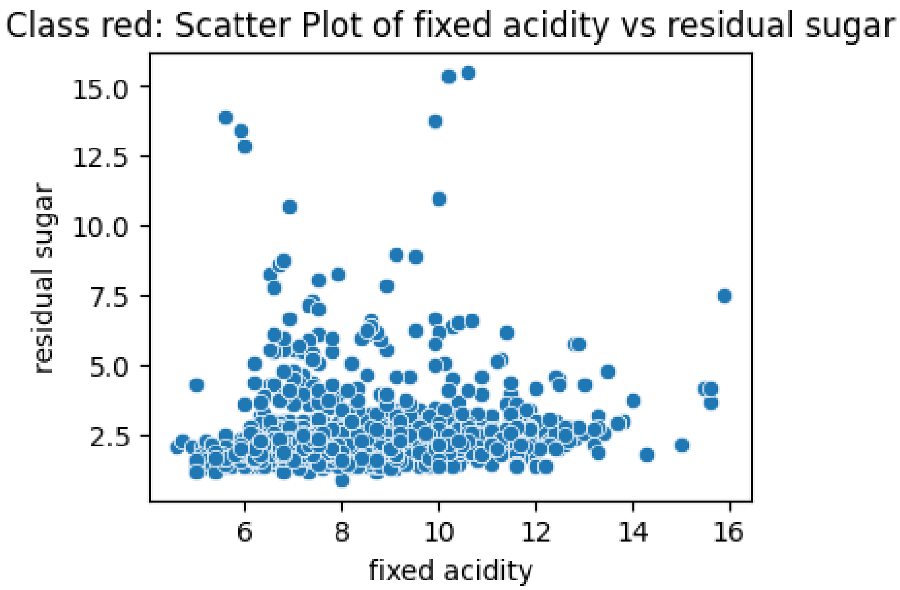

```
plot_classwise_pairwise(combined_wine_data, 'wineType')
```

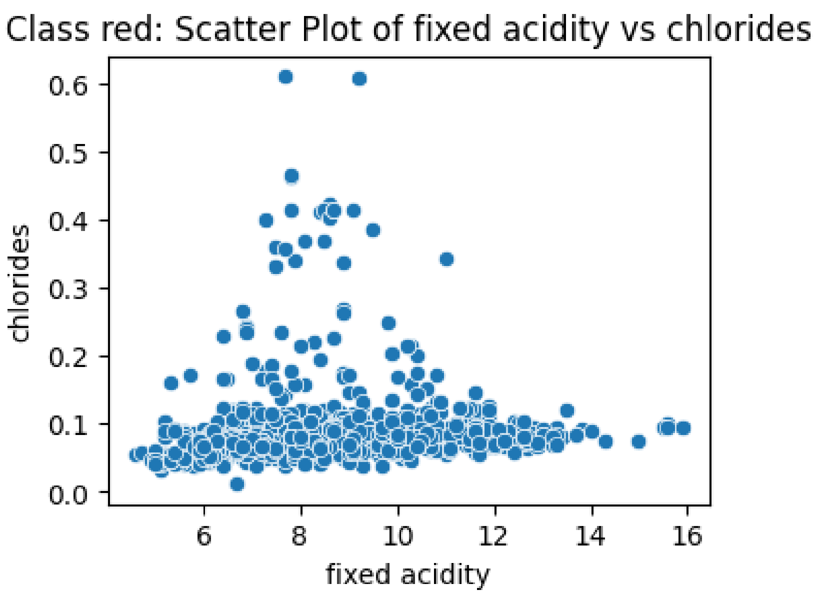
Class red: Scatter Plot of fixed acidity vs volatile acidity



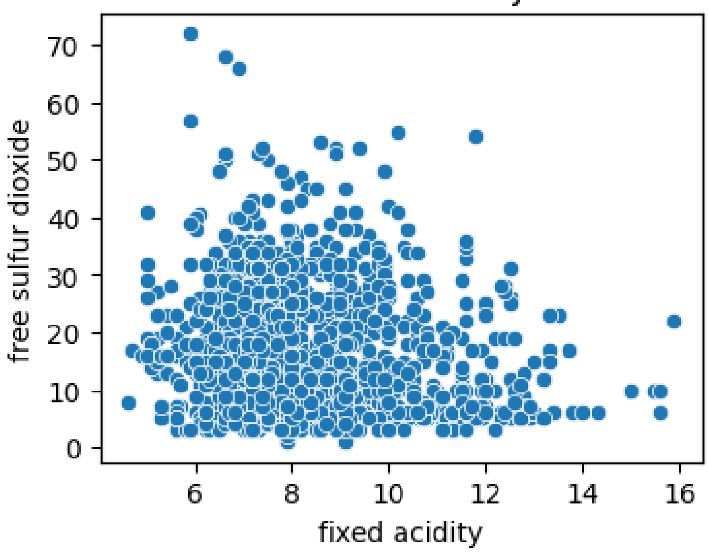
Class red: Scatter Plot of fixed acidity vs citric acid



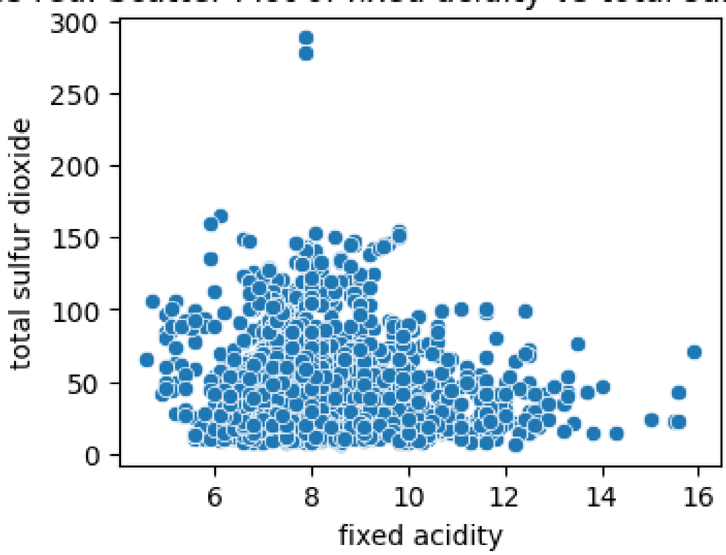

Class red: Scatter Plot of fixed acidity vs residual sugar
residual sugar
15.0
12.5
10.0
7.5
5.0
2.5
6
8
10
12
14
16
fixed acidity

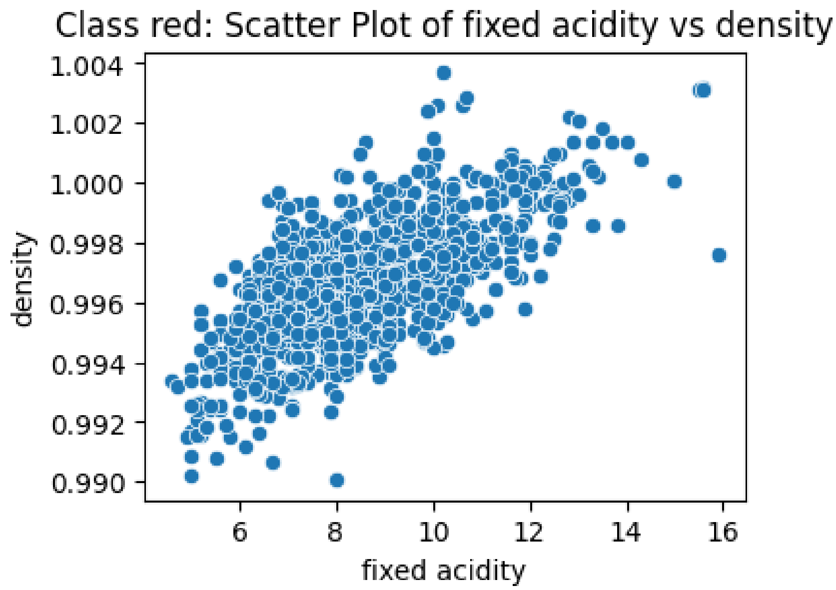

Class red: Scatter Plot of fixed acidity vs chlorides
chlorides
0.6
0.5
0.4
0.3
0.2
0.1
0.0
6
8
10
12
14
16
fixed acidity

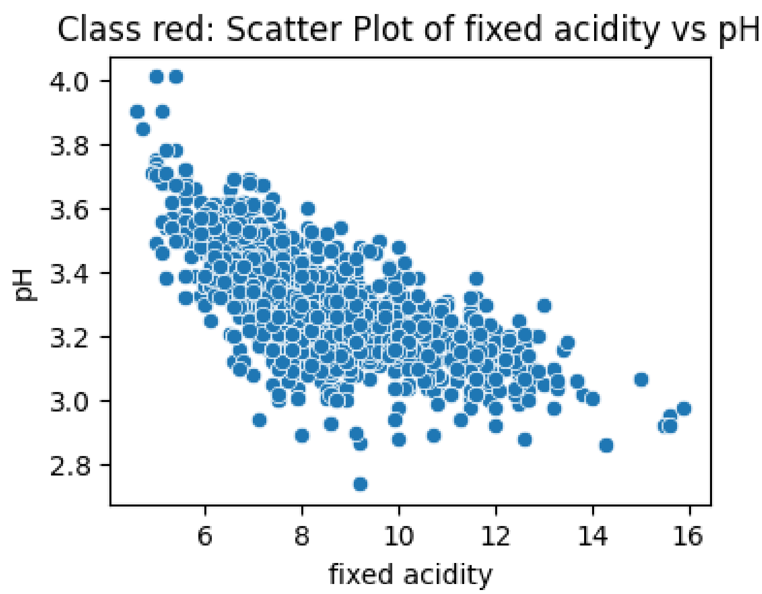
Class red: Scatter Plot of fixed acidity vs free sulfur dioxide

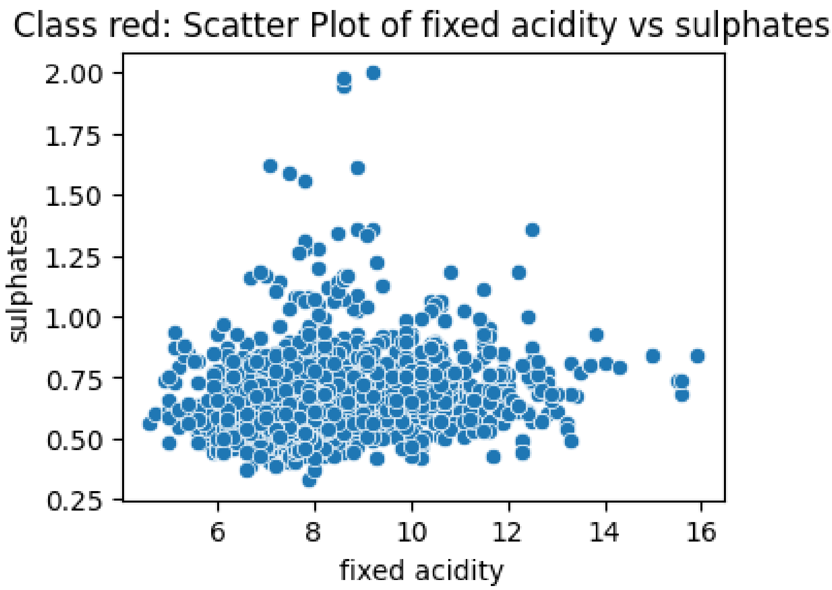


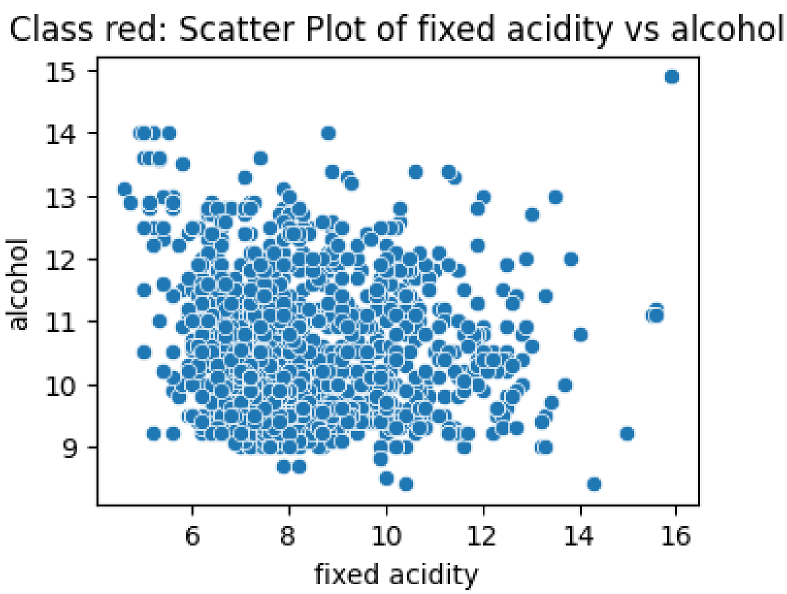
Class red: Scatter Plot of fixed acidity vs total sulfur dioxide



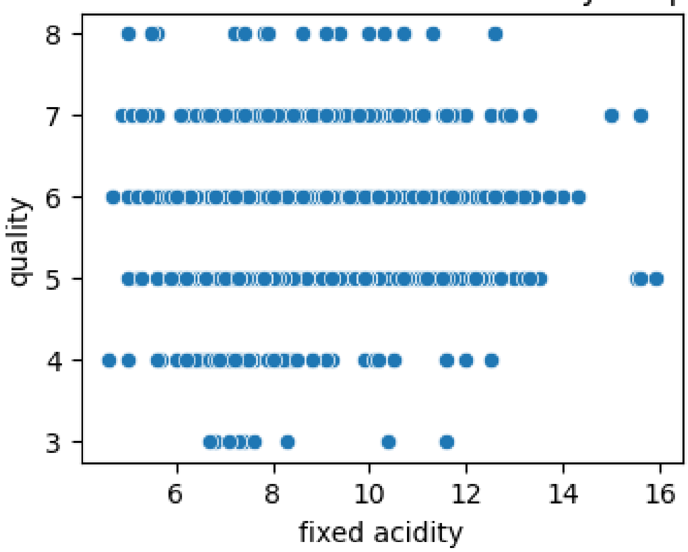

Class red: Scatter Plot of fixed acidity vs density
density
fixed acidity
1.004
1.002
1.000
0.998
0.996
0.994
0.992
0.990
6
8
10
12
14
16


Class red: Scatter Plot of fixed acidity vs pH
pH
fixed acidity
4.0
3.8
3.6
3.4
3.2
3.0
2.8
6
8
10
12
14
16

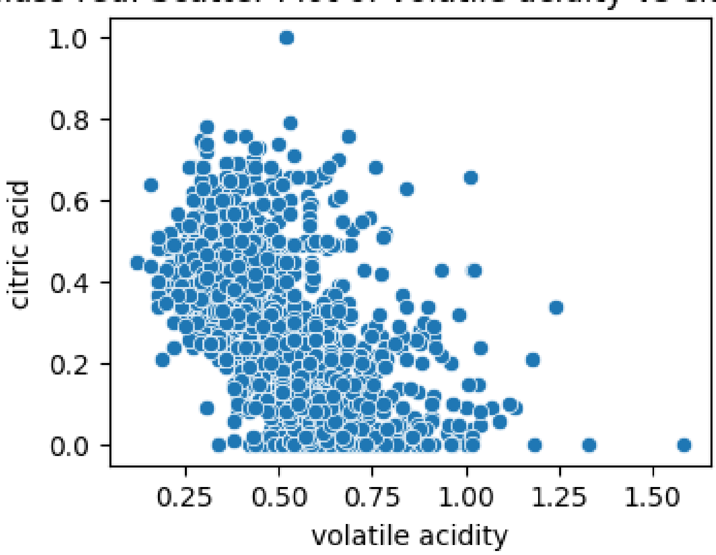

Class red: Scatter Plot of fixed acidity vs sulphates
sulphates
fixed acidity
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
6
8
10
12
14
16


Class red: Scatter Plot of fixed acidity vs alcohol
alcohol
fixed acidity
15
14
13
12
11
10
9
6
8
10
12
14
16

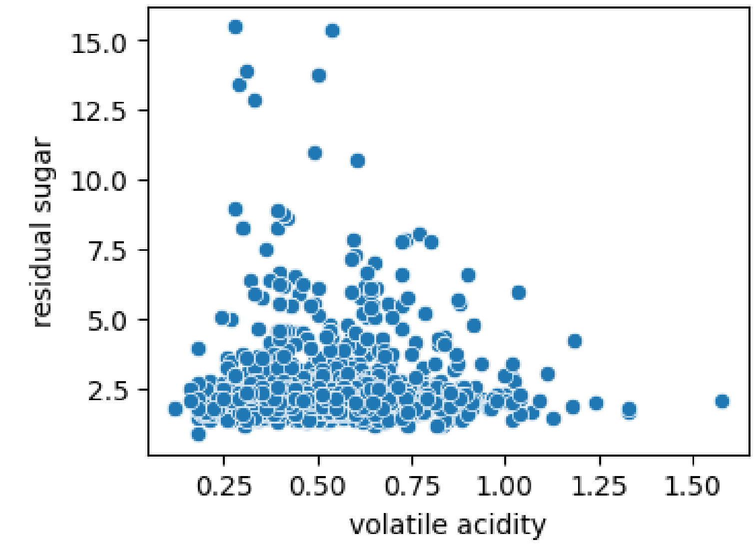
Class red: Scatter Plot of fixed acidity vs quality



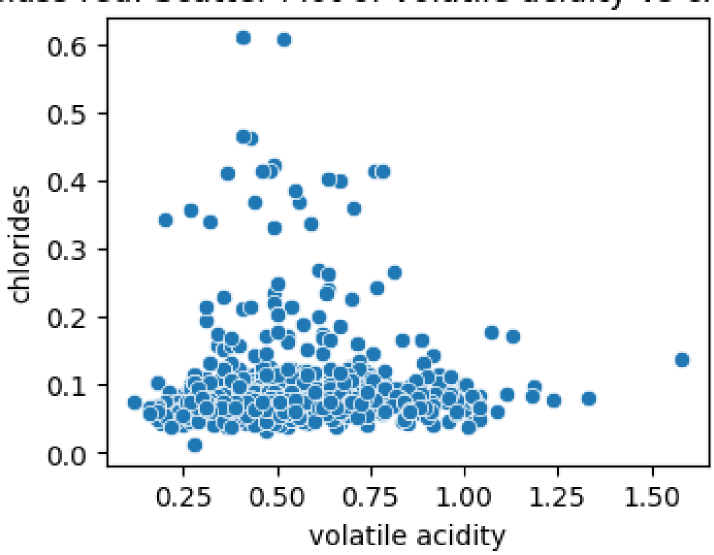
Class red: Scatter Plot of volatile acidity vs citric acid



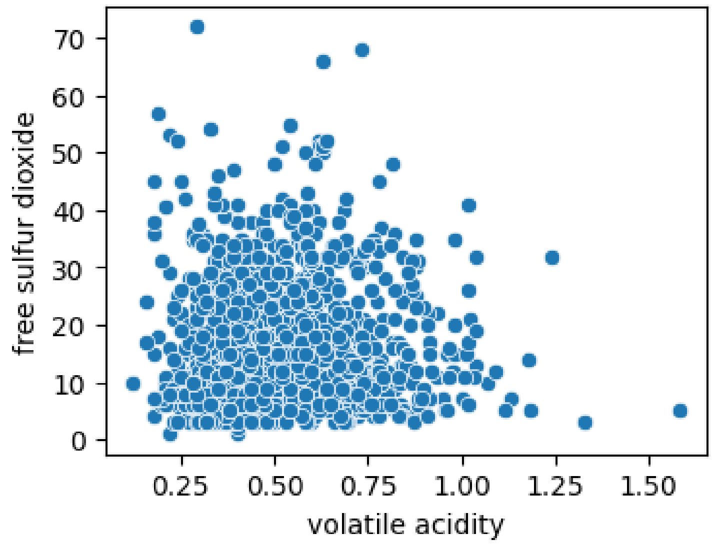
Class red: Scatter Plot of volatile acidity vs residual sugar



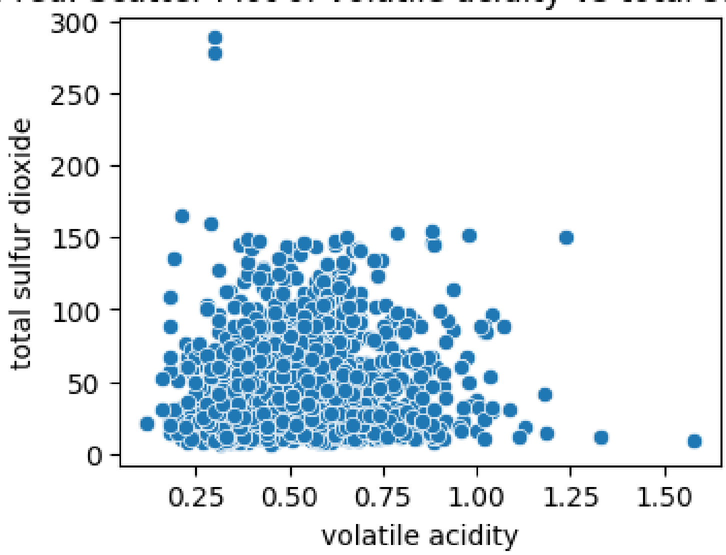
Class red: Scatter Plot of volatile acidity vs chlorides

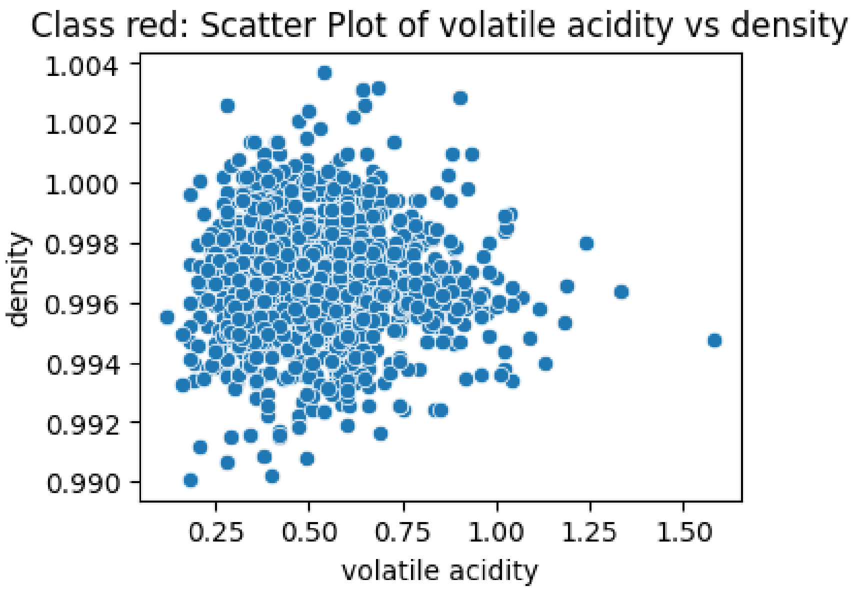


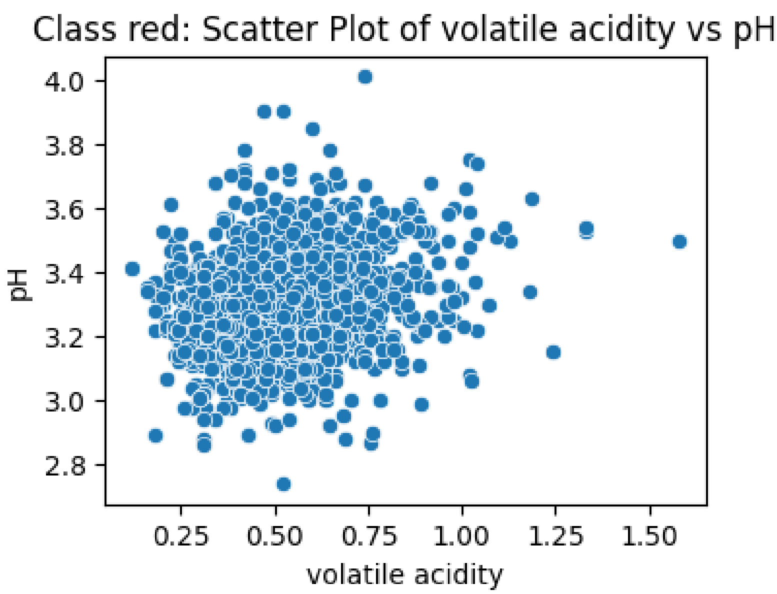
Class red: Scatter Plot of volatile acidity vs free sulfur dioxide



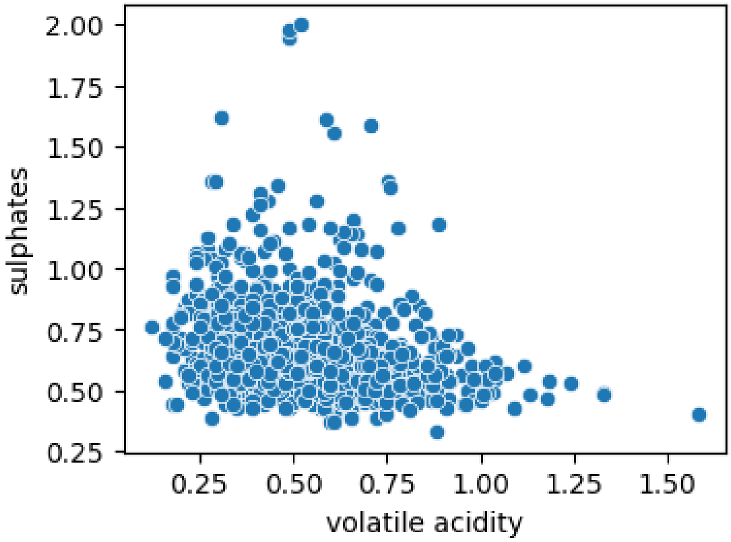
Class red: Scatter Plot of volatile acidity vs total sulfur dioxide



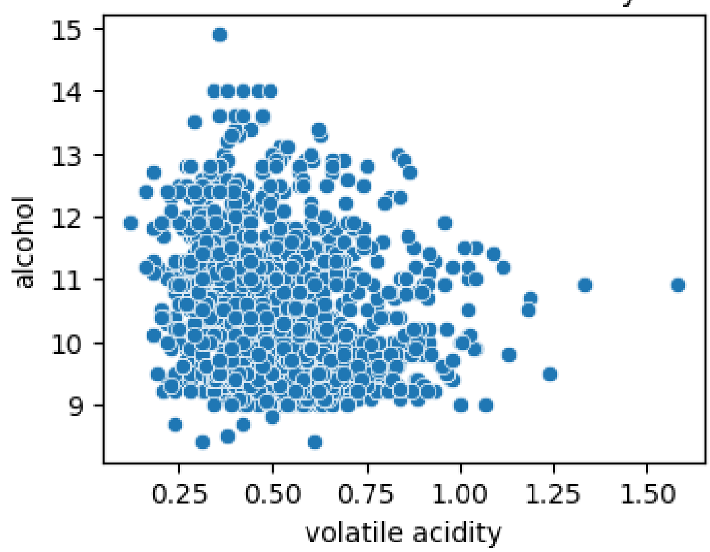

Class red: Scatter Plot of volatile acidity vs density
density
1.004
1.002
1.000
0.998
0.996
0.994
0.992
0.990
0.25
0.50
0.75
1.00
1.25
1.50
volatile acidity


Class red: Scatter Plot of volatile acidity vs pH
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
0.25
0.50
0.75
1.00
1.25
1.50
volatile acidity

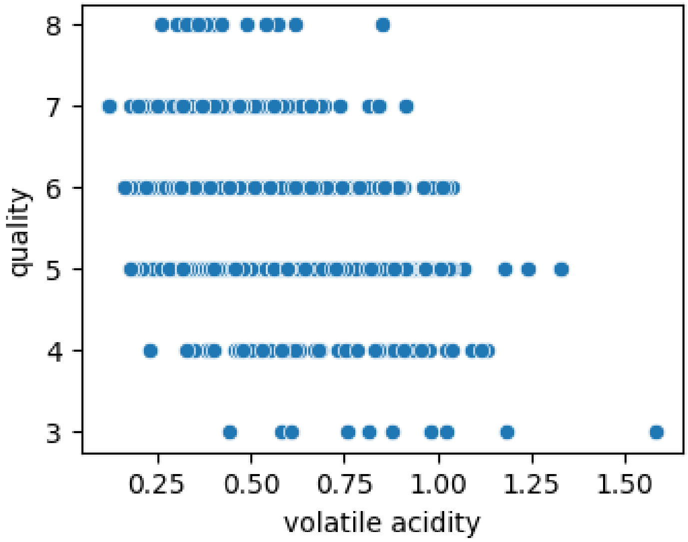
Class red: Scatter Plot of volatile acidity vs sulphates



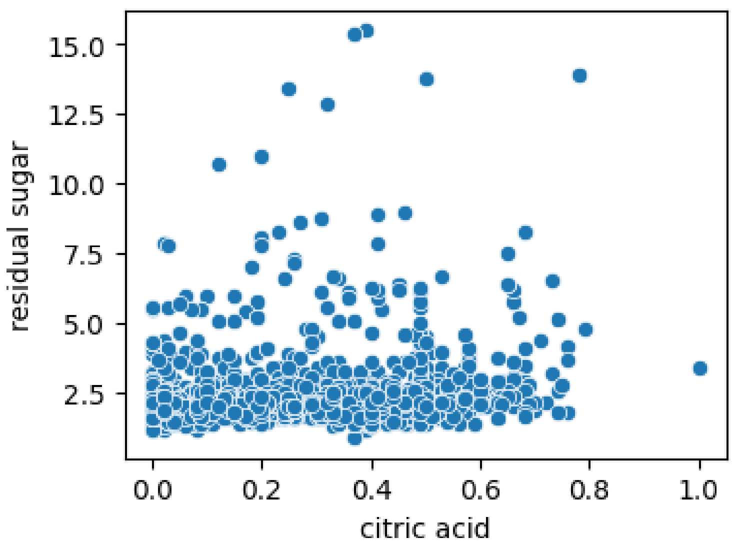
Class red: Scatter Plot of volatile acidity vs alcohol



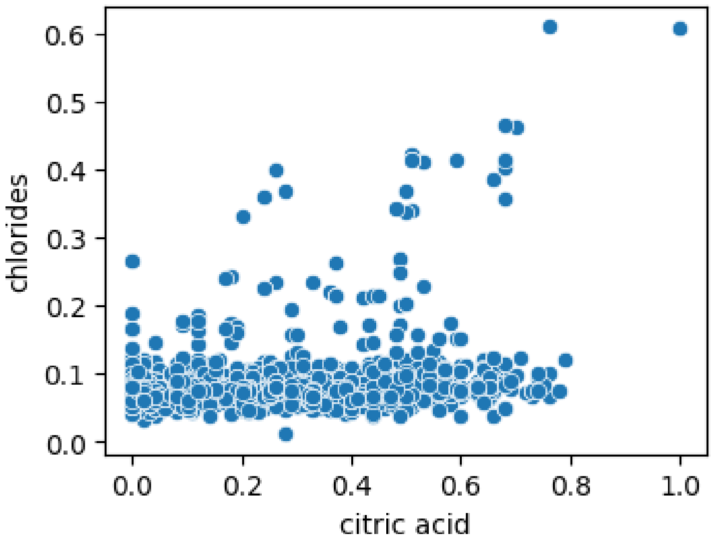
Class red: Scatter Plot of volatile acidity vs quality



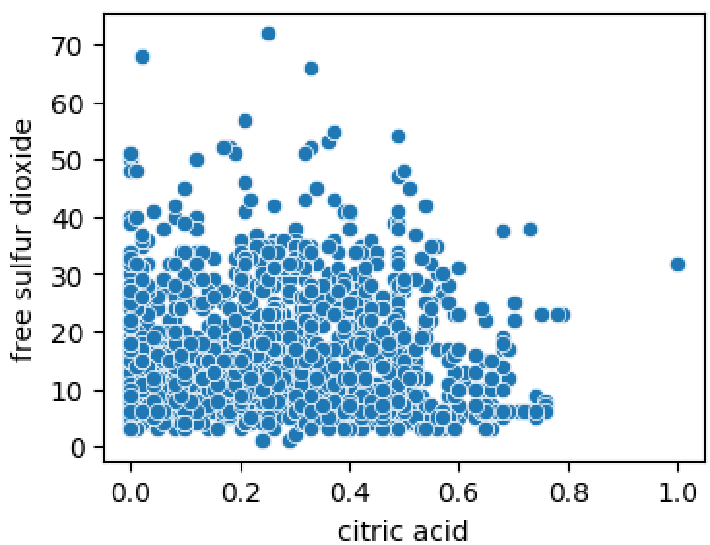
Class red: Scatter Plot of citric acid vs residual sugar



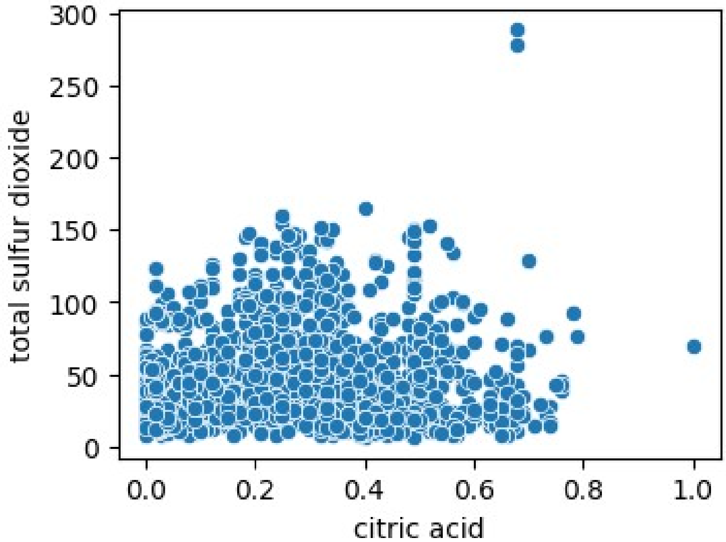
Class red: Scatter Plot of citric acid vs chlorides



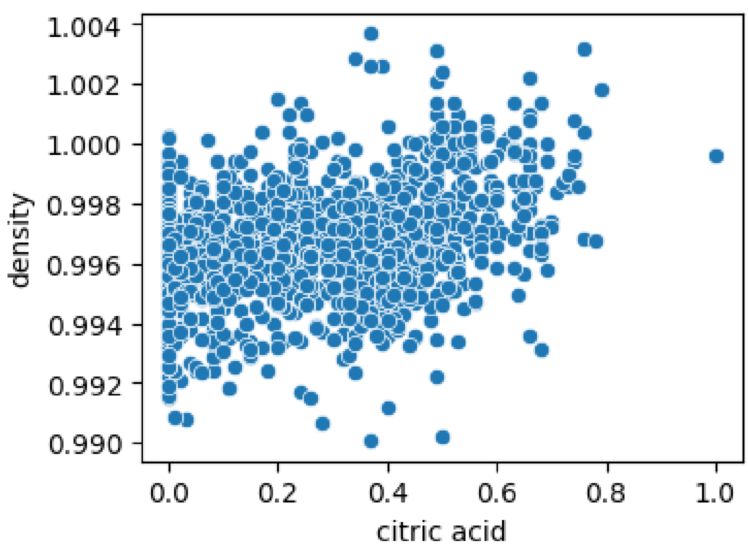
Class red: Scatter Plot of citric acid vs free sulfur dioxide

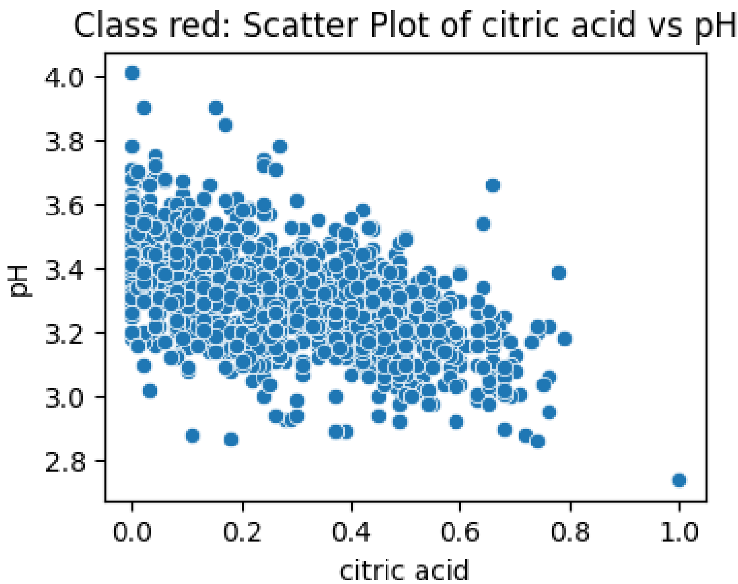


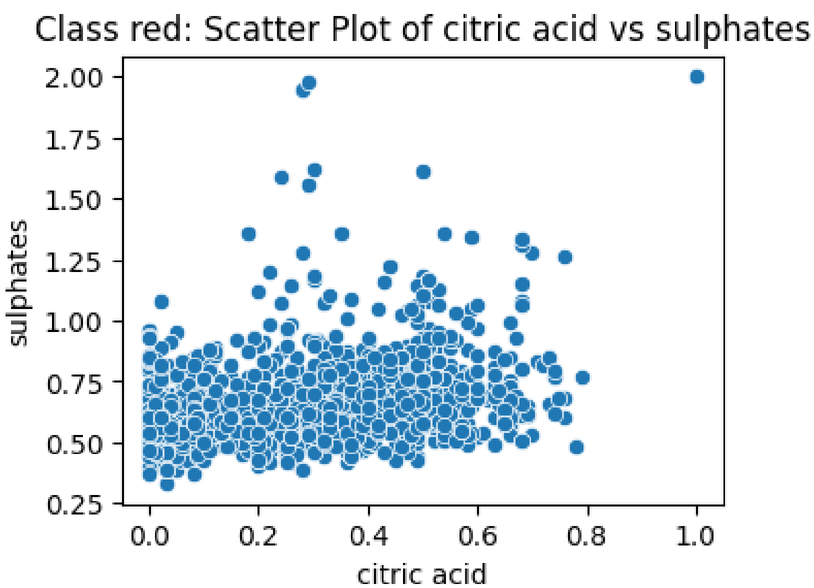
Class red: Scatter Plot of citric acid vs total sulfur dioxide



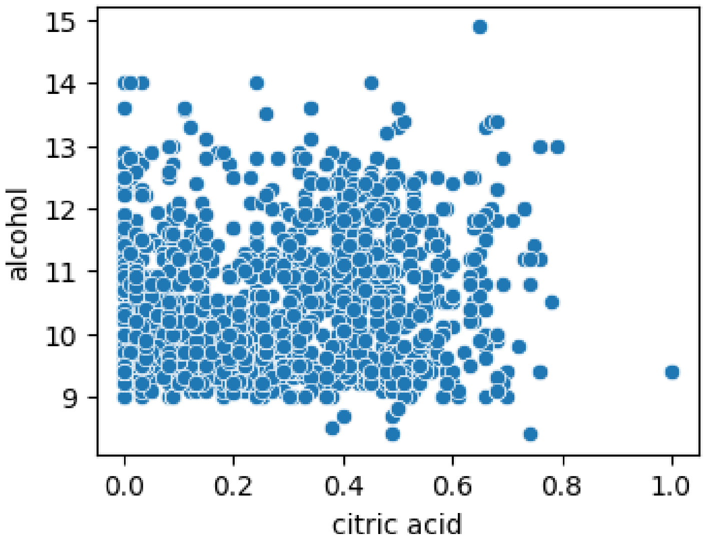
Class red: Scatter Plot of citric acid vs density



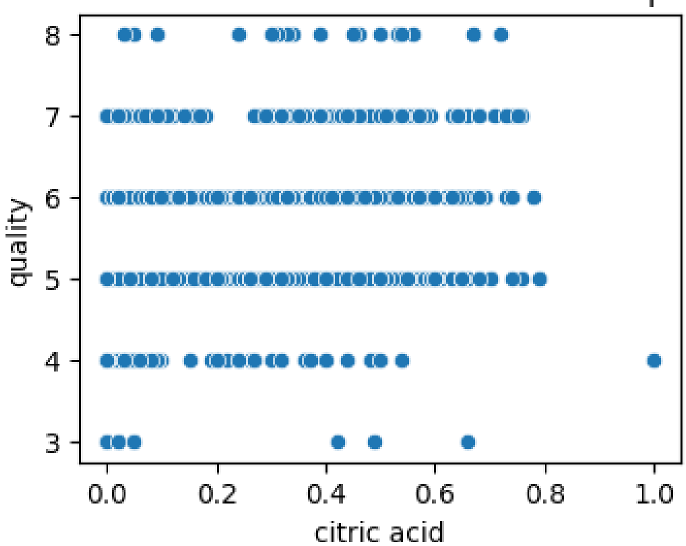

Class red: Scatter Plot of citric acid vs pH
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
0.0
0.2
0.4
0.6
0.8
1.0
citric acid


Class red: Scatter Plot of citric acid vs sulphates
sulphates
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
citric acid

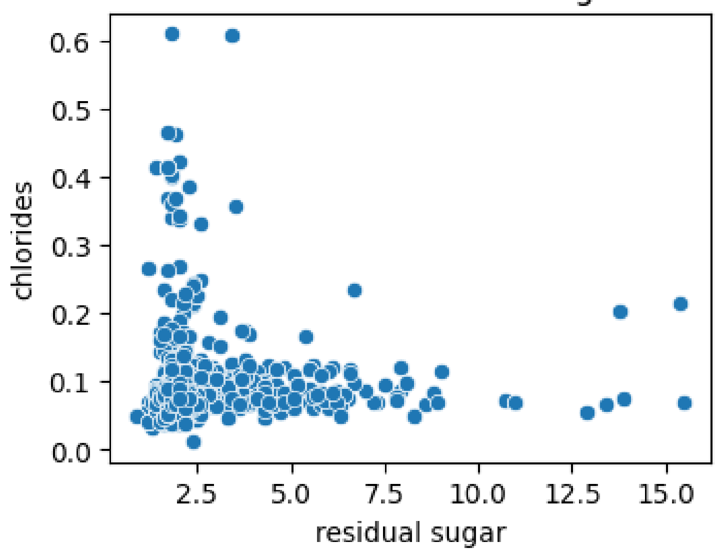
Class red: Scatter Plot of citric acid vs alcohol



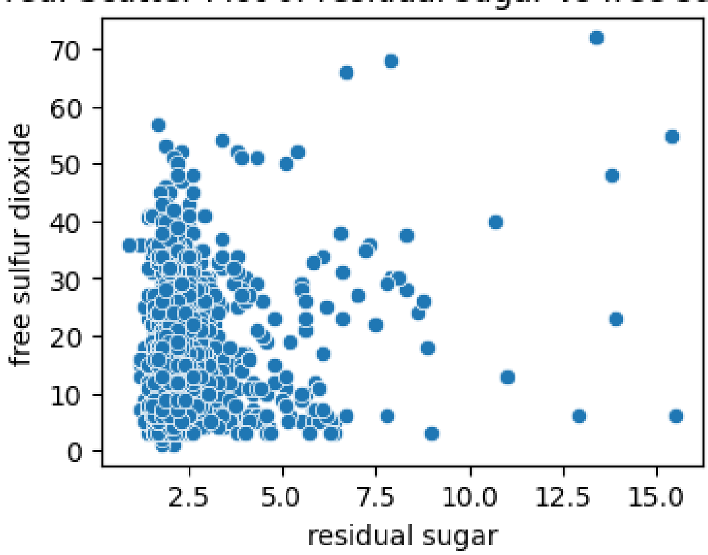
Class red: Scatter Plot of citric acid vs quality

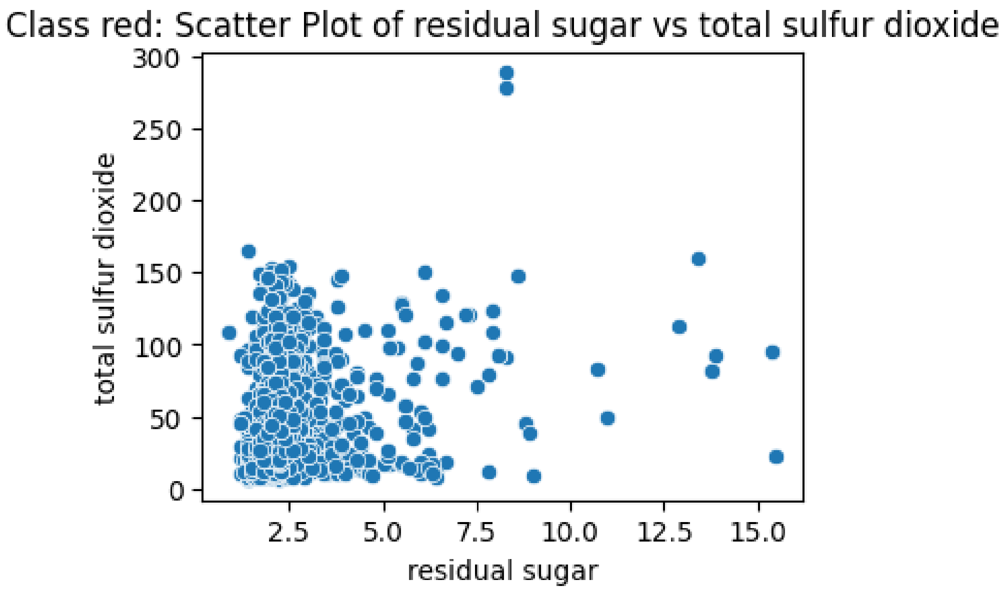


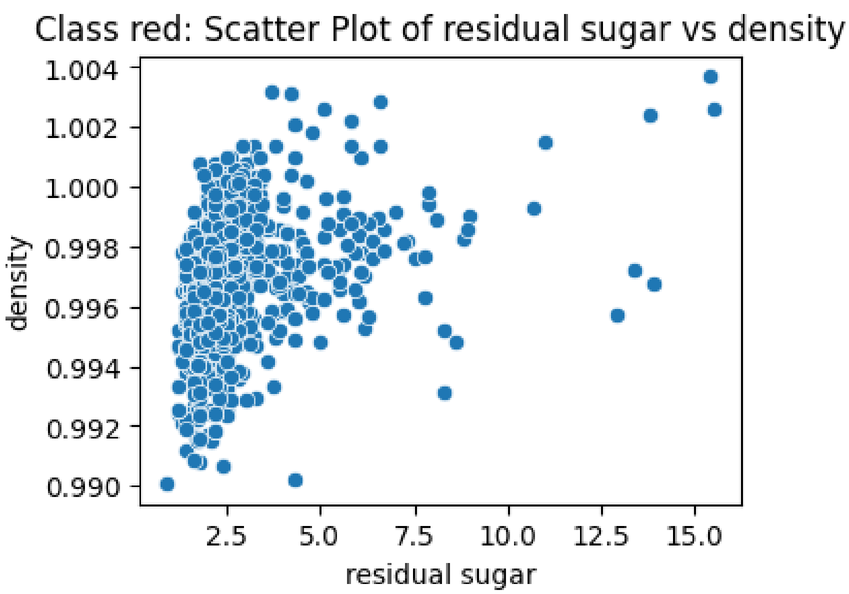
Class red: Scatter Plot of residual sugar vs chlorides



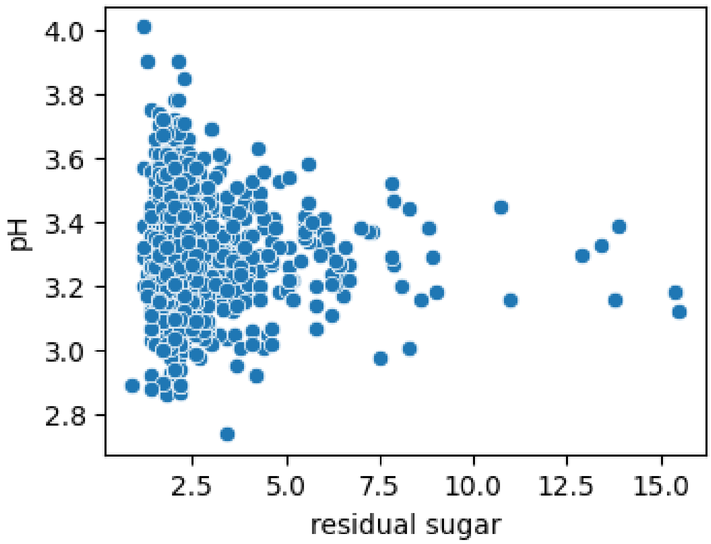
Class red: Scatter Plot of residual sugar vs free sulfur dioxide



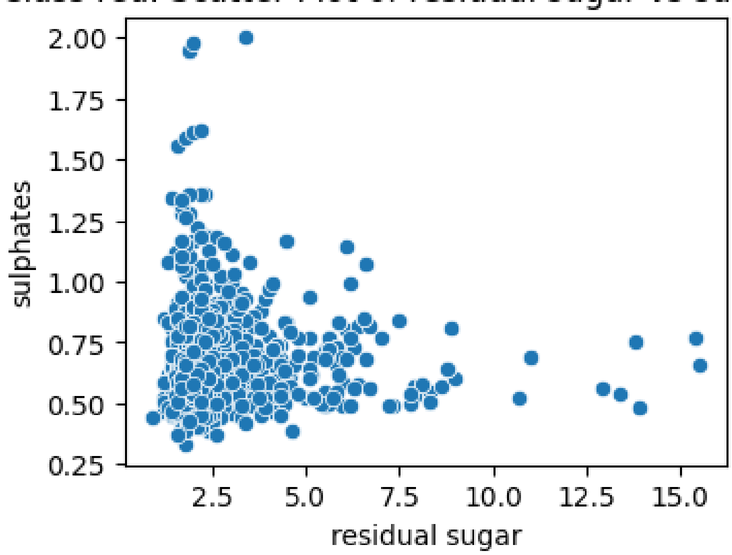

Class red: Scatter Plot of residual sugar vs total sulfur dioxide
total sulfur dioxide
residual sugar
0
50
100
150
200
250
300
2.5
5.0
7.5
10.0
12.5
15.0


Class red: Scatter Plot of residual sugar vs density
density
residual sugar
0.990
0.992
0.994
0.996
0.998
1.000
1.002
1.004
2.5
5.0
7.5
10.0
12.5
15.0

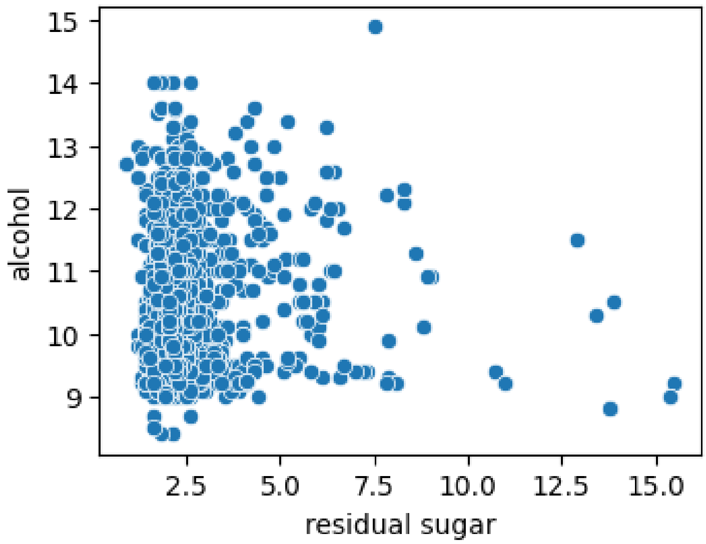
Class red: Scatter Plot of residual sugar vs pH



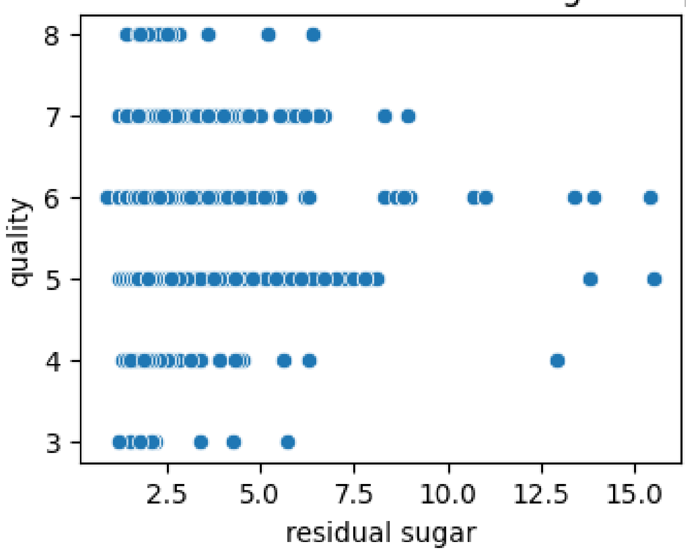
Class red: Scatter Plot of residual sugar vs sulphates



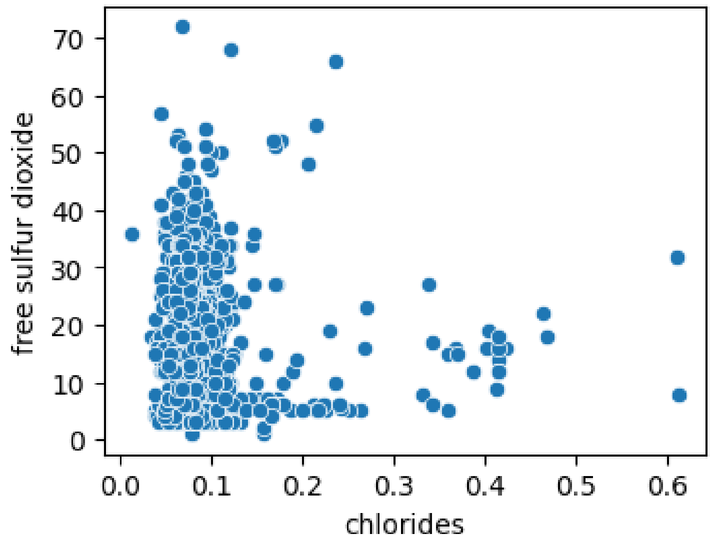
Class red: Scatter Plot of residual sugar vs alcohol



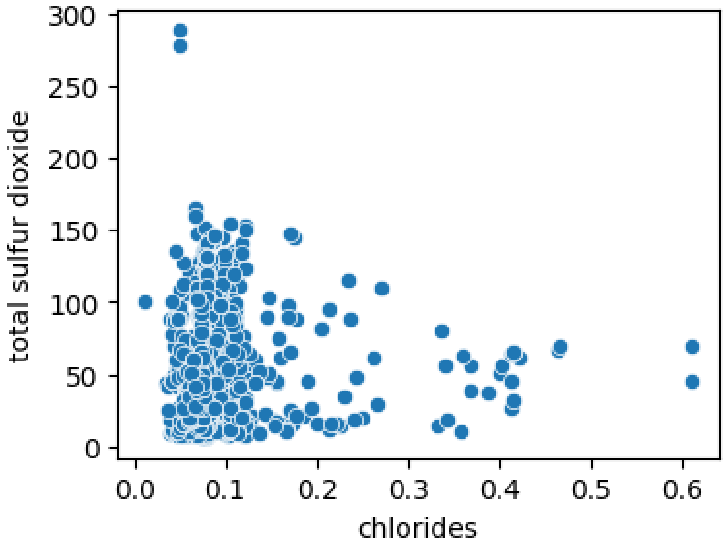
Class red: Scatter Plot of residual sugar vs quality

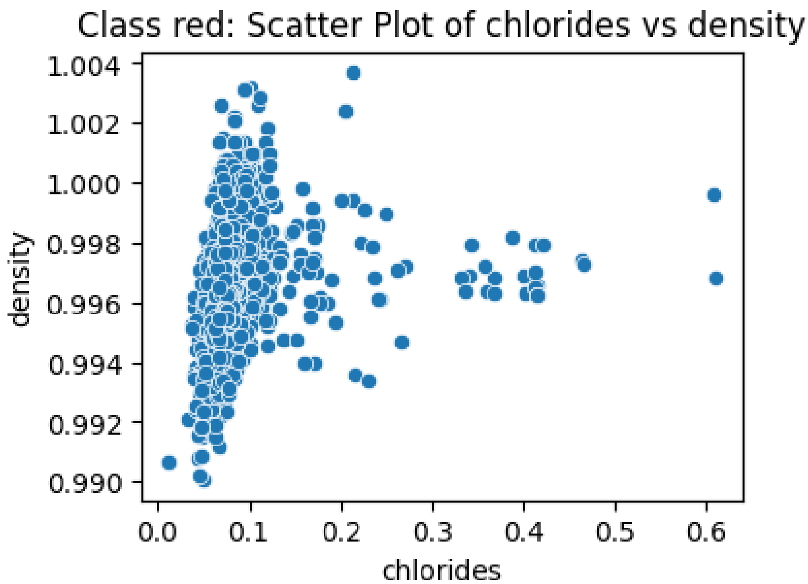


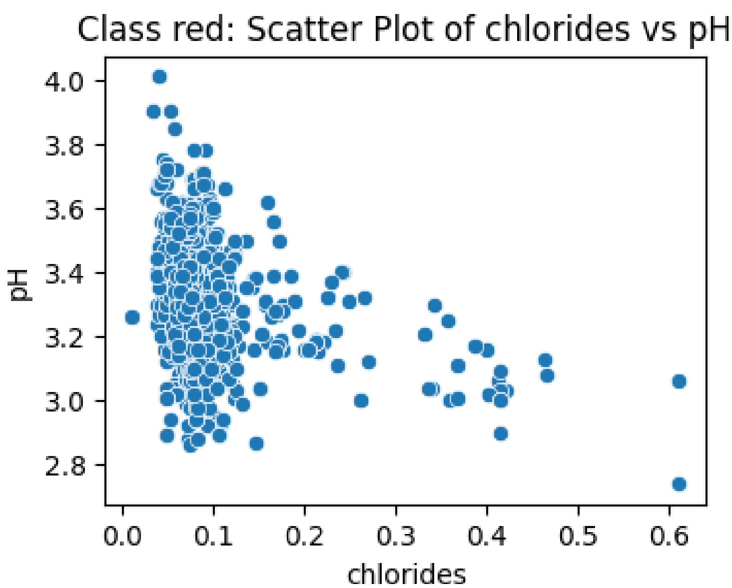
Class red: Scatter Plot of chlorides vs free sulfur dioxide



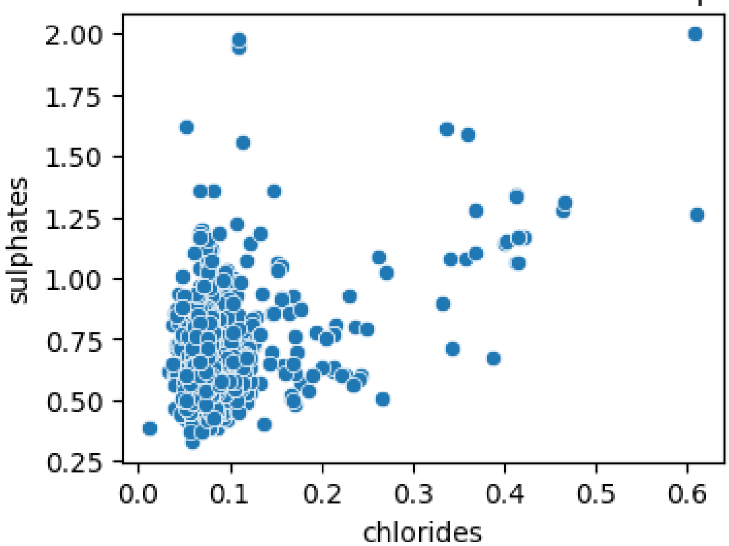
Class red: Scatter Plot of chlorides vs total sulfur dioxide



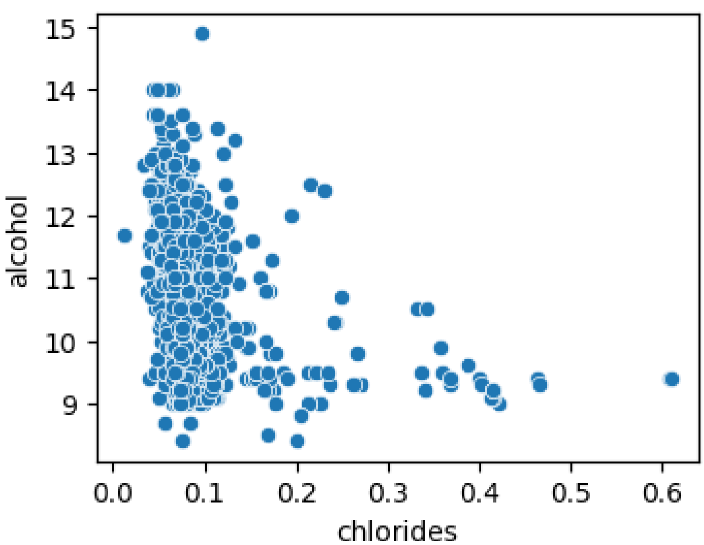

Class red: Scatter Plot of chlorides vs density
density
chlorides
1.004
1.002
1.000
0.998
0.996
0.994
0.992
0.990
0.0
0.1
0.2
0.3
0.4
0.5
0.6


Class red: Scatter Plot of chlorides vs pH
pH
chlorides
4.0
3.8
3.6
3.4
3.2
3.0
2.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6

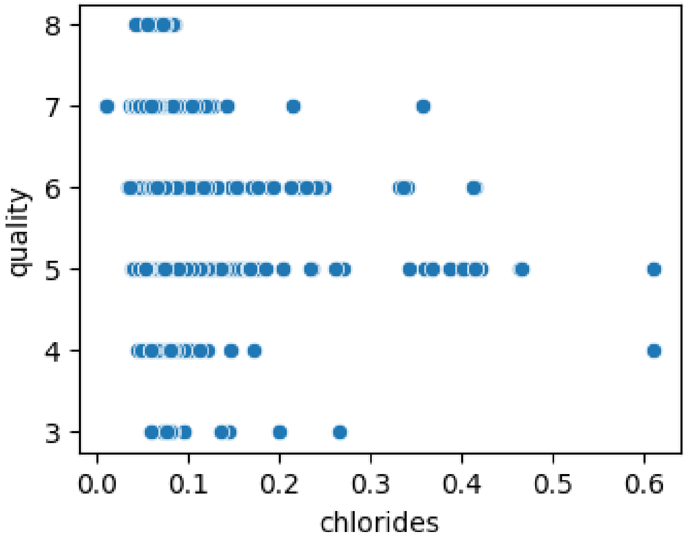
Class red: Scatter Plot of chlorides vs sulphates



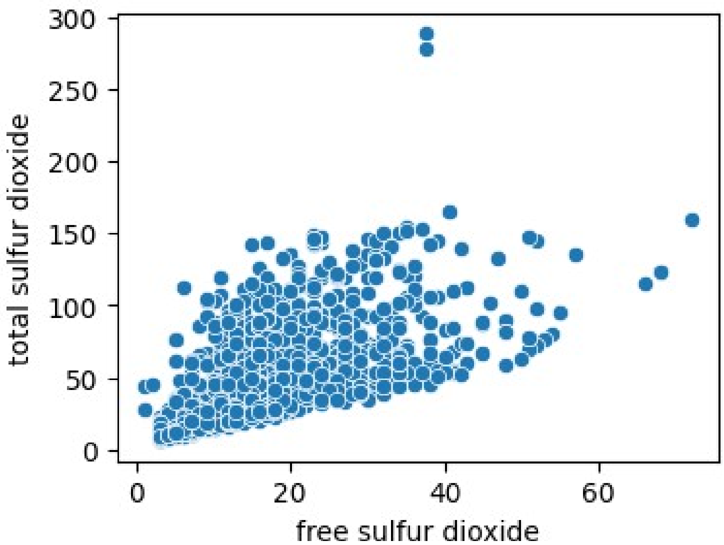
Class red: Scatter Plot of chlorides vs alcohol



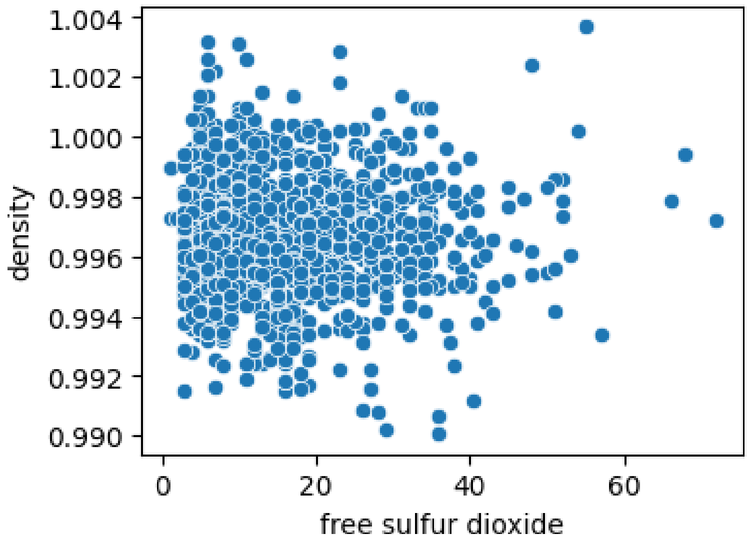
Class red: Scatter Plot of chlorides vs quality



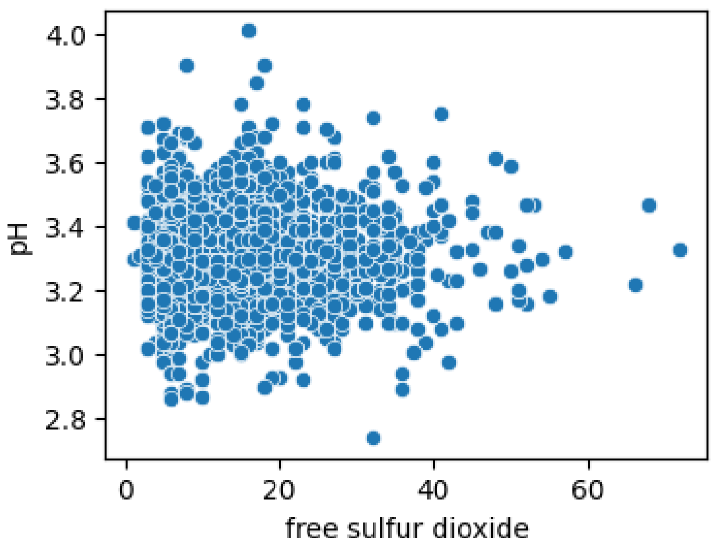
Class red: Scatter Plot of free sulfur dioxide vs total sulfur dioxide



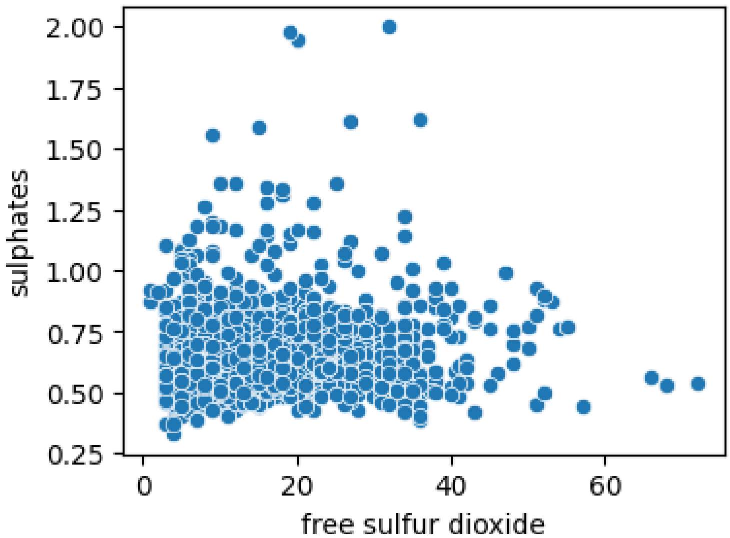
Class red: Scatter Plot of free sulfur dioxide vs density



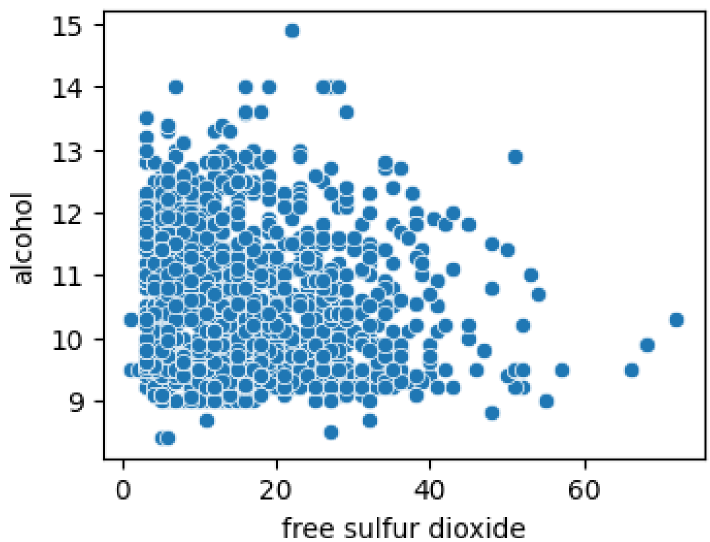
Class red: Scatter Plot of free sulfur dioxide vs pH



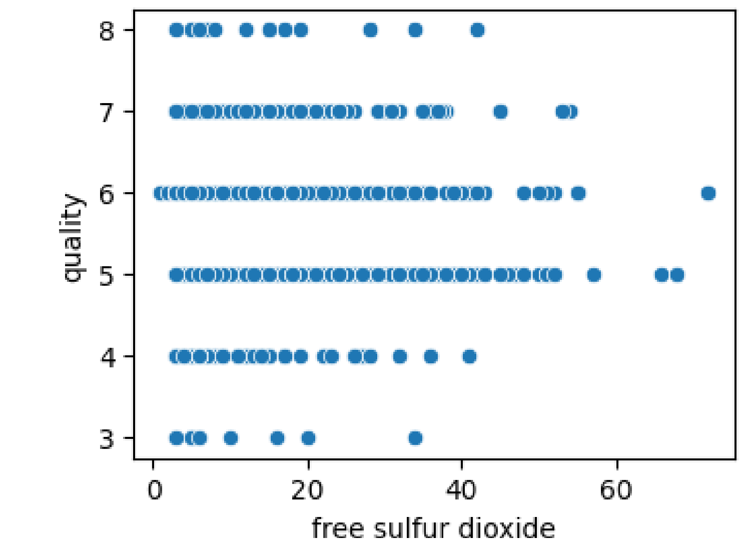
Class red: Scatter Plot of free sulfur dioxide vs sulphates



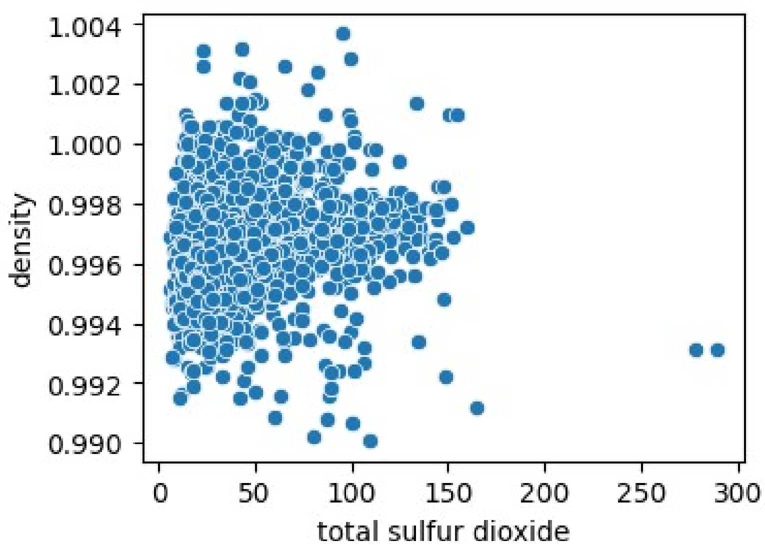
Class red: Scatter Plot of free sulfur dioxide vs alcohol



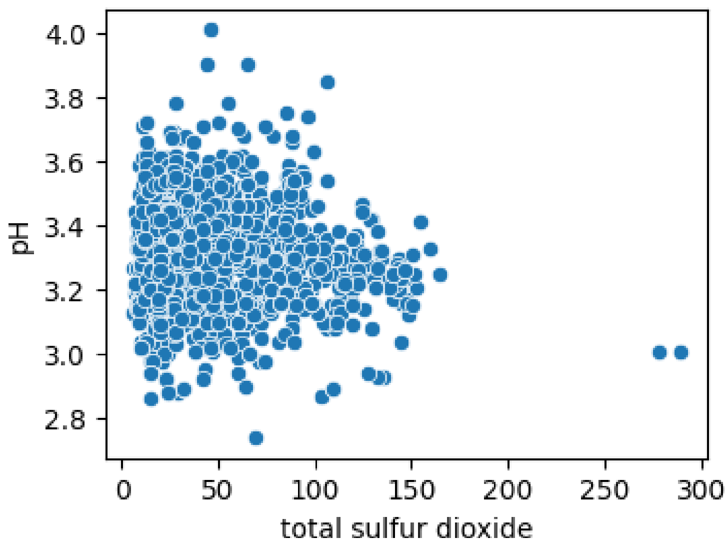
Class red: Scatter Plot of free sulfur dioxide vs quality



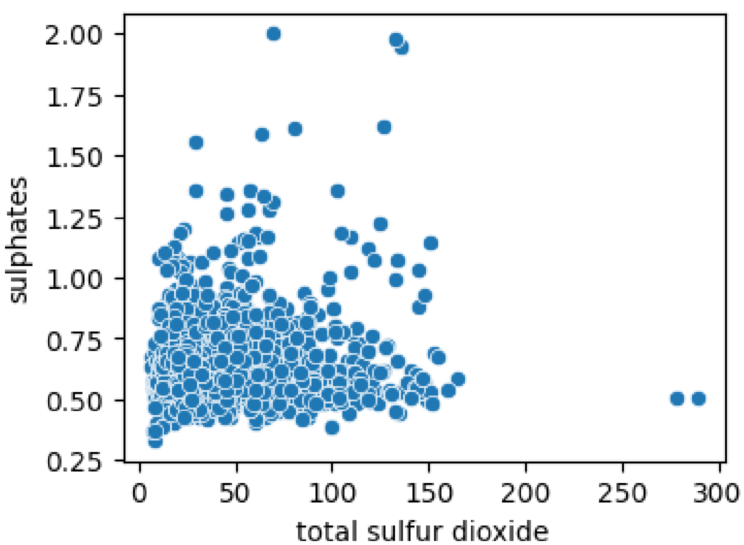
Class red: Scatter Plot of total sulfur dioxide vs density



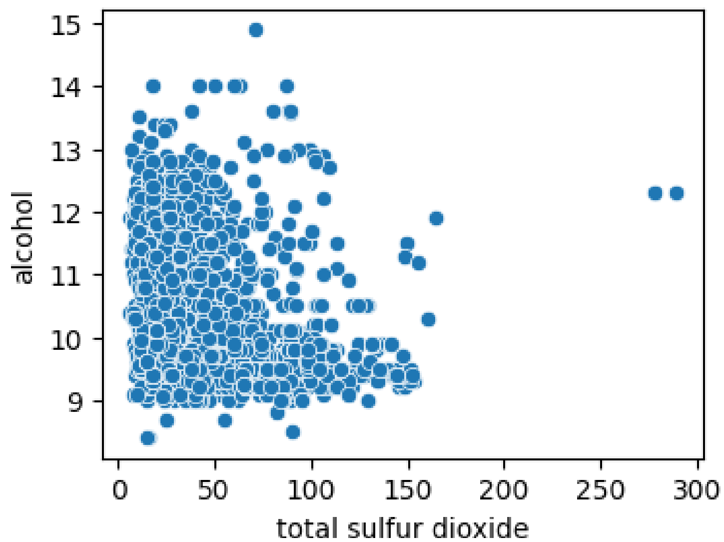
Class red: Scatter Plot of total sulfur dioxide vs pH



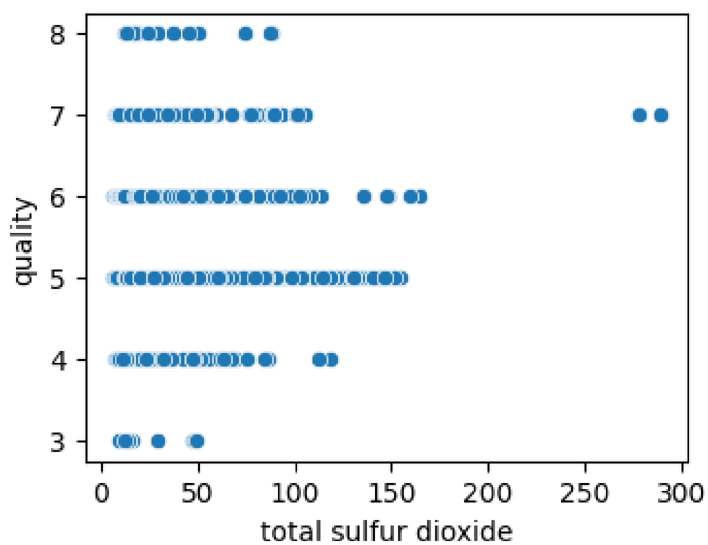
Class red: Scatter Plot of total sulfur dioxide vs sulphates

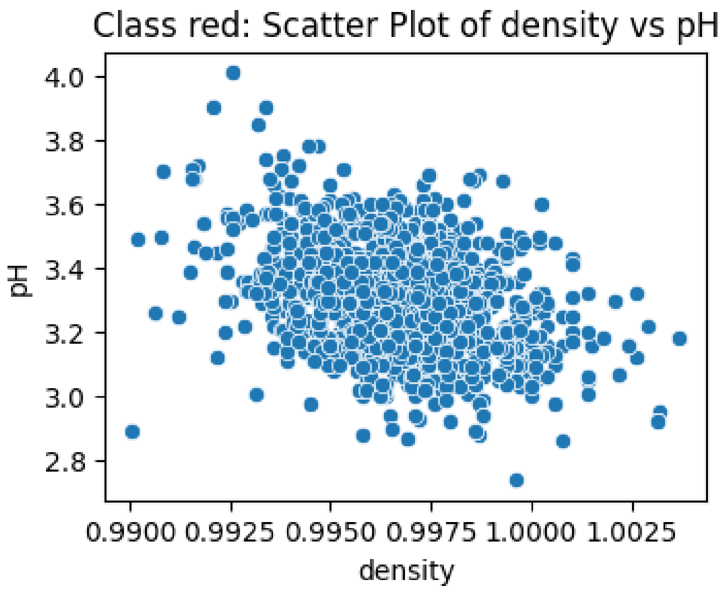


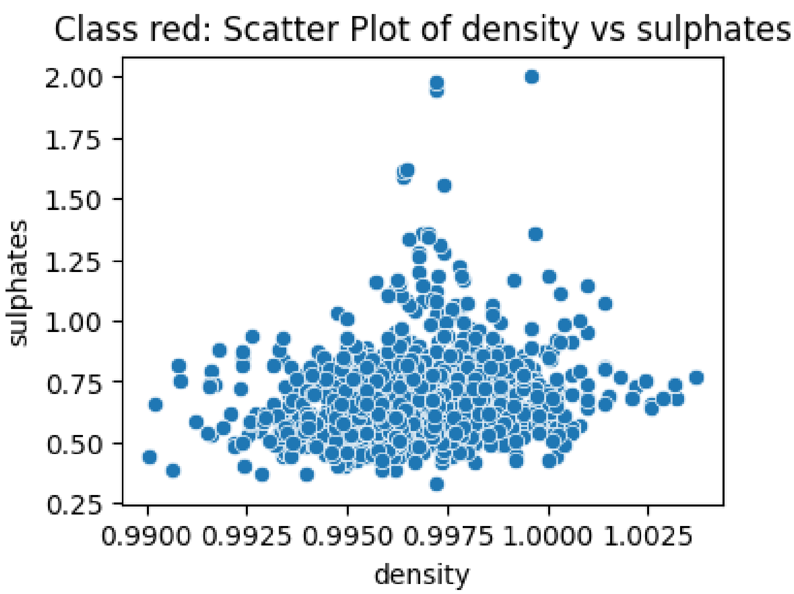
Class red: Scatter Plot of total sulfur dioxide vs alcohol

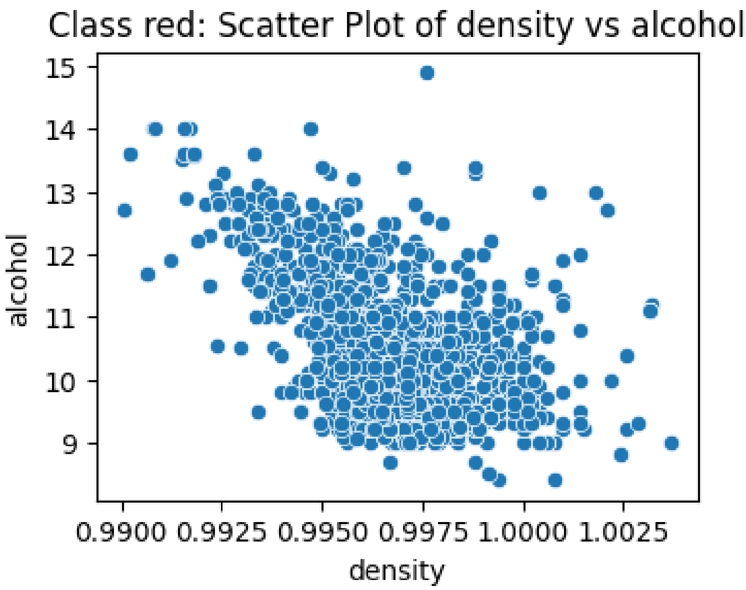


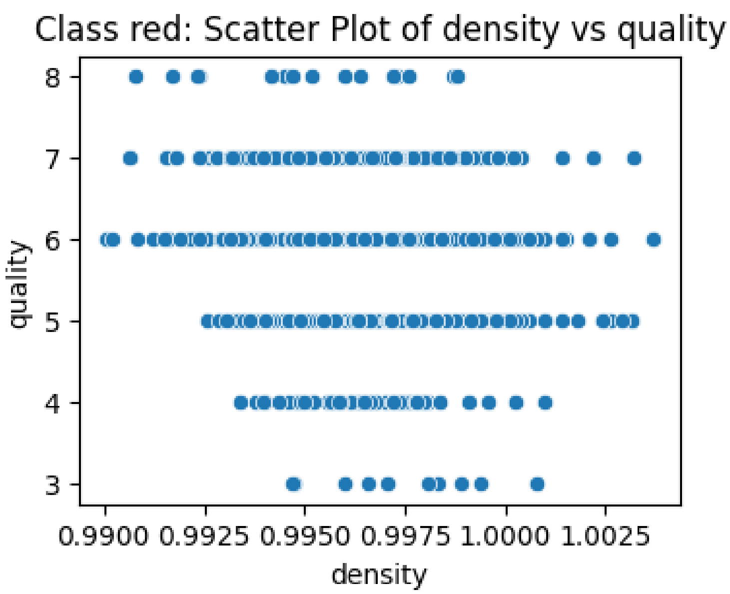
Class red: Scatter Plot of total sulfur dioxide vs quality

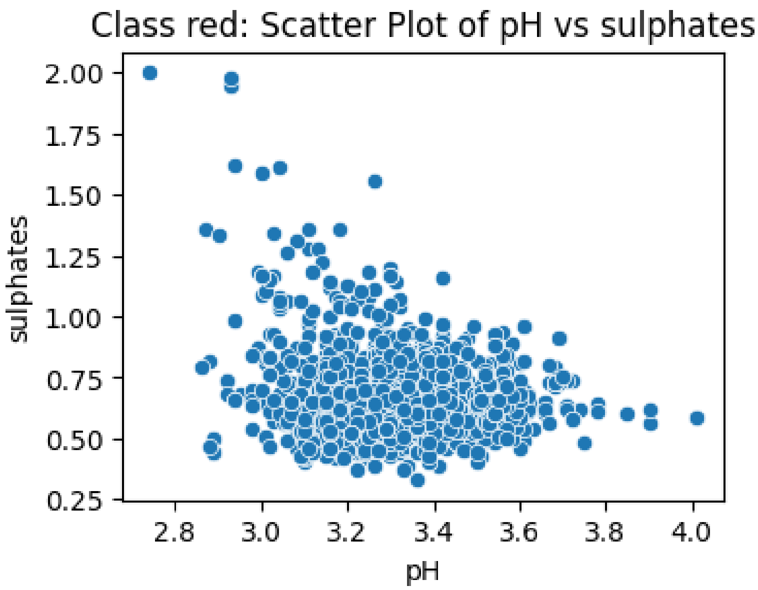


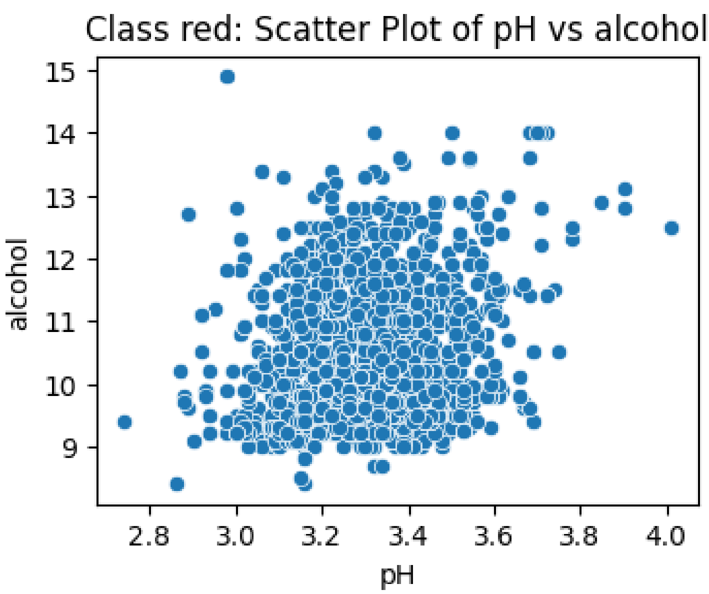

Class red: Scatter Plot of density vs pH
pH
density
4.0
3.8
3.6
3.4
3.2
3.0
2.8
0.9900
0.9925
0.9950
0.9975
1.0000
1.0025

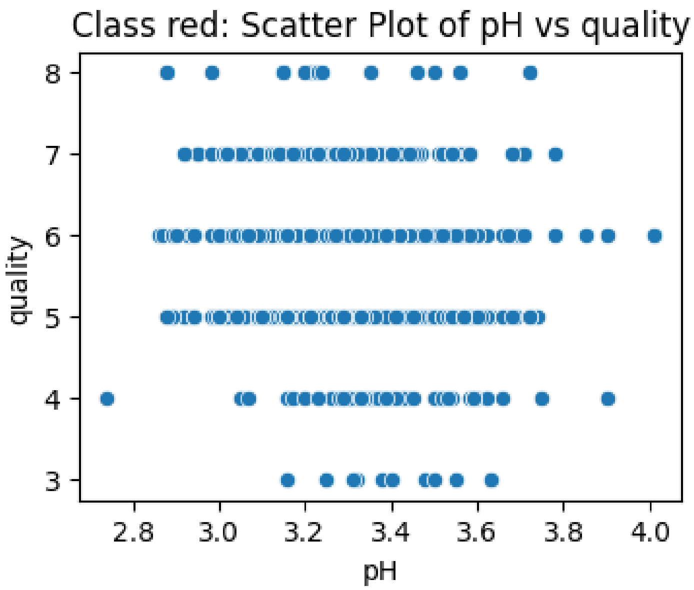

Class red: Scatter Plot of density vs sulphates
sulphates
density
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.9900
0.9925
0.9950
0.9975
1.0000
1.0025

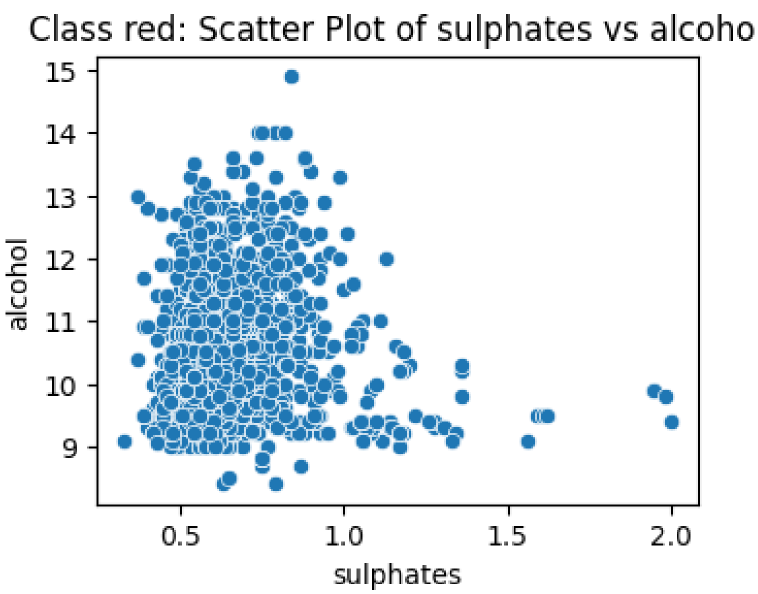

Class red: Scatter Plot of density vs alcohol
alcohol
density
15
14
13
12
11
10
9
0.9900
0.9925
0.9950
0.9975
1.0000
1.0025

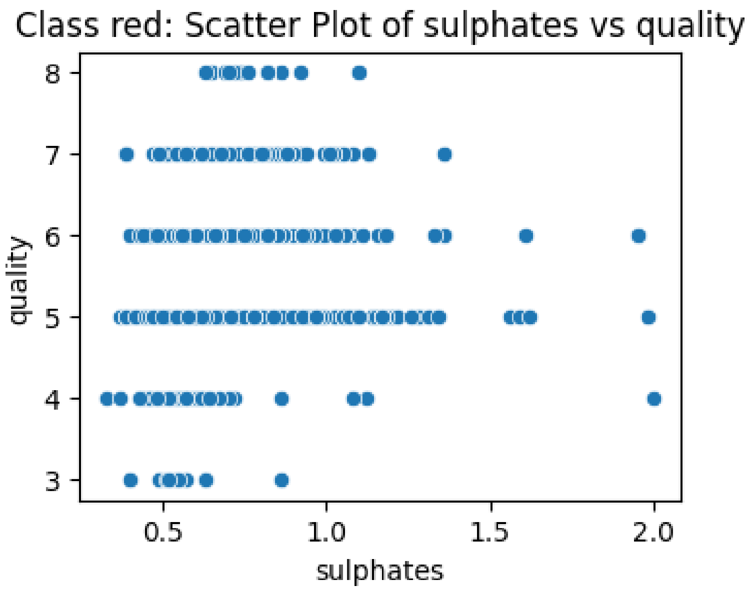

Class red: Scatter Plot of density vs quality
quality
density
8
7
6
5
4
3
0.9900
0.9925
0.9950
0.9975
1.0000
1.0025

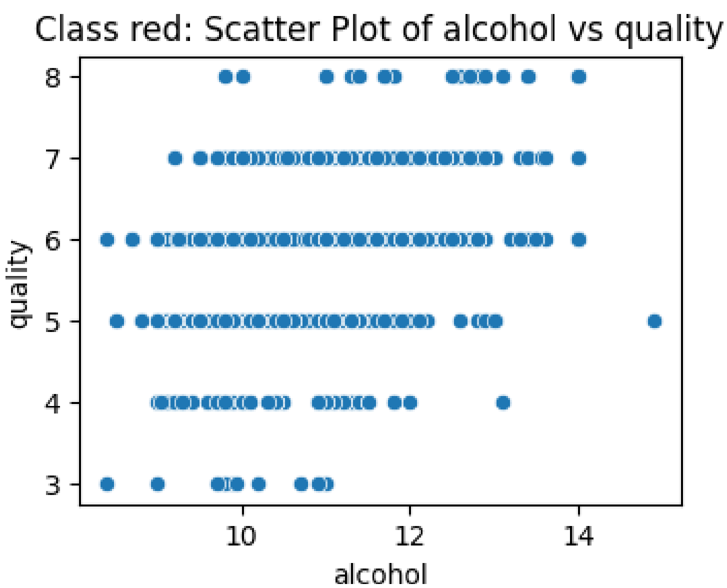

Class red: Scatter Plot of pH vs sulphates
sulphates
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
2.8
3.0
3.2
3.4
3.6
3.8
4.0
pH


Class red: Scatter Plot of pH vs alcohol
alcohol
15
14
13
12
11
10
9
2.8
3.0
3.2
3.4
3.6
3.8
4.0
pH


Class red: Scatter Plot of pH vs quality
quality
8
7
6
5
4
3
2.8
3.0
3.2
3.4
3.6
3.8
4.0
pH


Class red: Scatter Plot of sulphates vs alcohol
alcohol
15
14
13
12
11
10
9
0.5
1.0
1.5
2.0
sulphates





1. Number and Types of Features • Number of Features: The dataset contains 12 features. • Types of Features: • Numeric Features: Most of the features are numeric and continuous, including 'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', and 'alcohol'. • Discrete Feature: 'quality' is a numeric but discrete feature as it represents quality ratings. • Nominal Feature: 'wineType' is a nominal feature as it categorizes the wine into red or white.

2. Conclusions from Histograms • Normal Distribution: None of the features show a perfect normal distribution. However, 'alcohol' and 'density' exhibit distributions that are somewhat closer to normal but not exactly. • Skewness: Several features are skewed. For instance, 'residual sugar', 'chlorides', 'free sulfur dioxide', and 'total

sulfur dioxide' show a right (positive) skew. 'fixed acidity' and 'volatile acidity' also display some skewness.

3. Insights from Box Plots • Outliers: Features like 'residual sugar', 'chlorides', 'free sulfur dioxide', and 'total sulfur dioxide' have many outliers, as indicated by the points outside the whiskers of the box plots. • Spread of Values Across Quality Ratings: It's difficult to conclusively say which features have a similar spread across different quality ratings without overlaying quality on the box plots. However, the spread of 'alcohol', 'sulphates', and 'citric acid' seems to vary with different quality ratings, suggesting a possible correlation between these features and wine quality. • Different Spreads Across Quality Ratings: Features like 'alcohol', 'sulphates', and 'density' might show different spreads across different quality ratings, indicating their potential influence on the quality.

4. Observations from Pairwise Plots • High Correlation: Certain pairs of features, like 'free sulfur dioxide' and 'total sulfur dioxide', exhibit a high degree of correlation. 'fixed acidity' and 'citric acid' also appear to be somewhat correlated. • Low or No Correlation: Features like 'pH' and 'residual sugar' do not show a strong correlation with other features.

5. Class-wise Visualization Analysis • Correlation Differences by Wine Type: When examining the class-wise scatter plots (separating red and white wines), some pairs of features may exhibit different correlation patterns depending on the wine type. For instance, the relationship between 'total sulfur dioxide' and 'free sulfur dioxide' might be more pronounced in white wines than in red wines. Similarly, the relationship between 'fixed acidity' and 'citric acid' might vary between red and white wines.

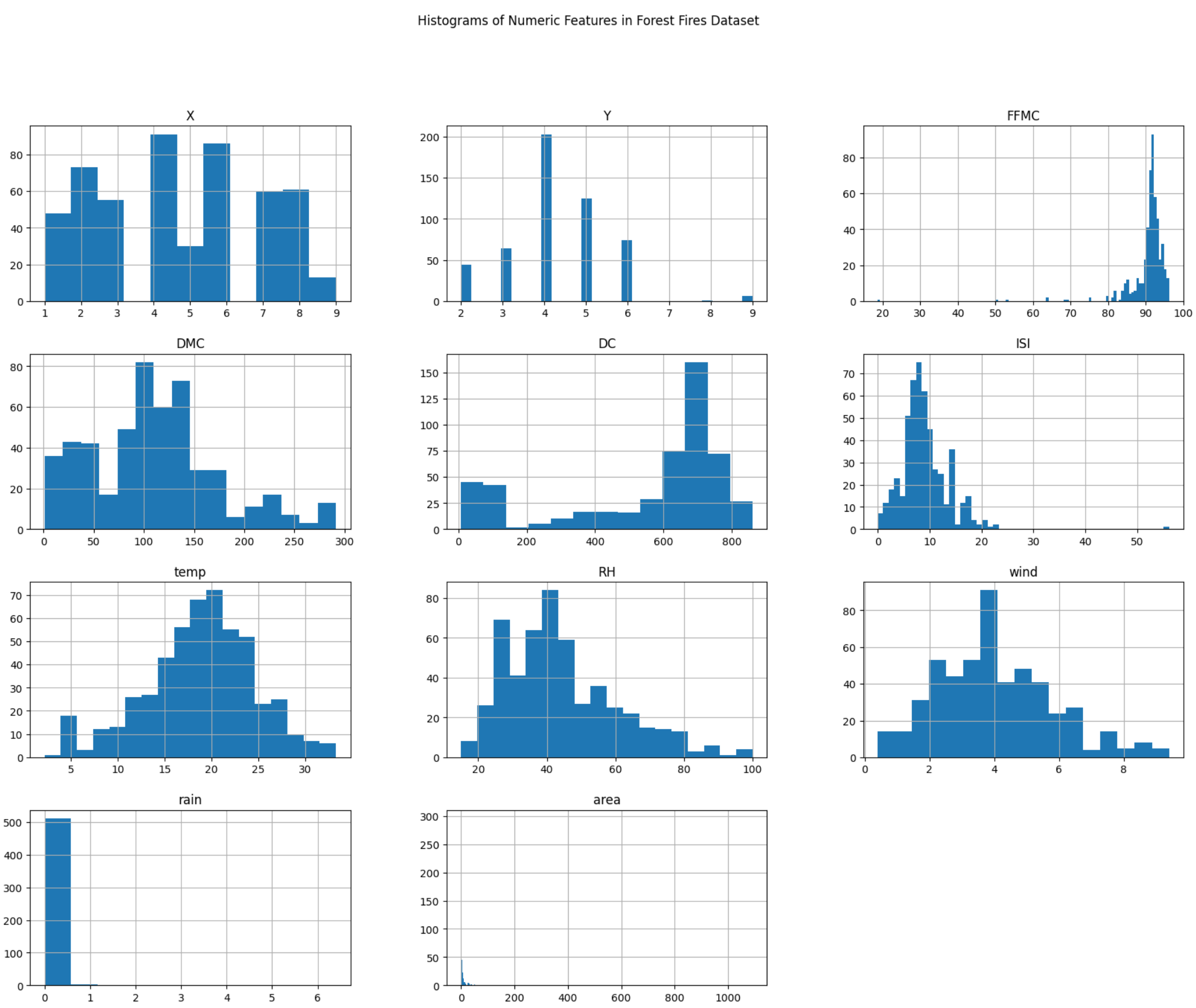
```python
# Adjust the file path as necessary
file_path = 'C:/Users/dande/Desktop/DMT/Assignment 
2/forest+fires/forestfires.csv'

# Try reading the file with a comma as the separator
forest_fires_data = pd.read_csv(file_path, sep=',')

# Check the data types with the new separator
print("Data Types with Comma Separator:")
print(forest_fires_data.dtypes)
```

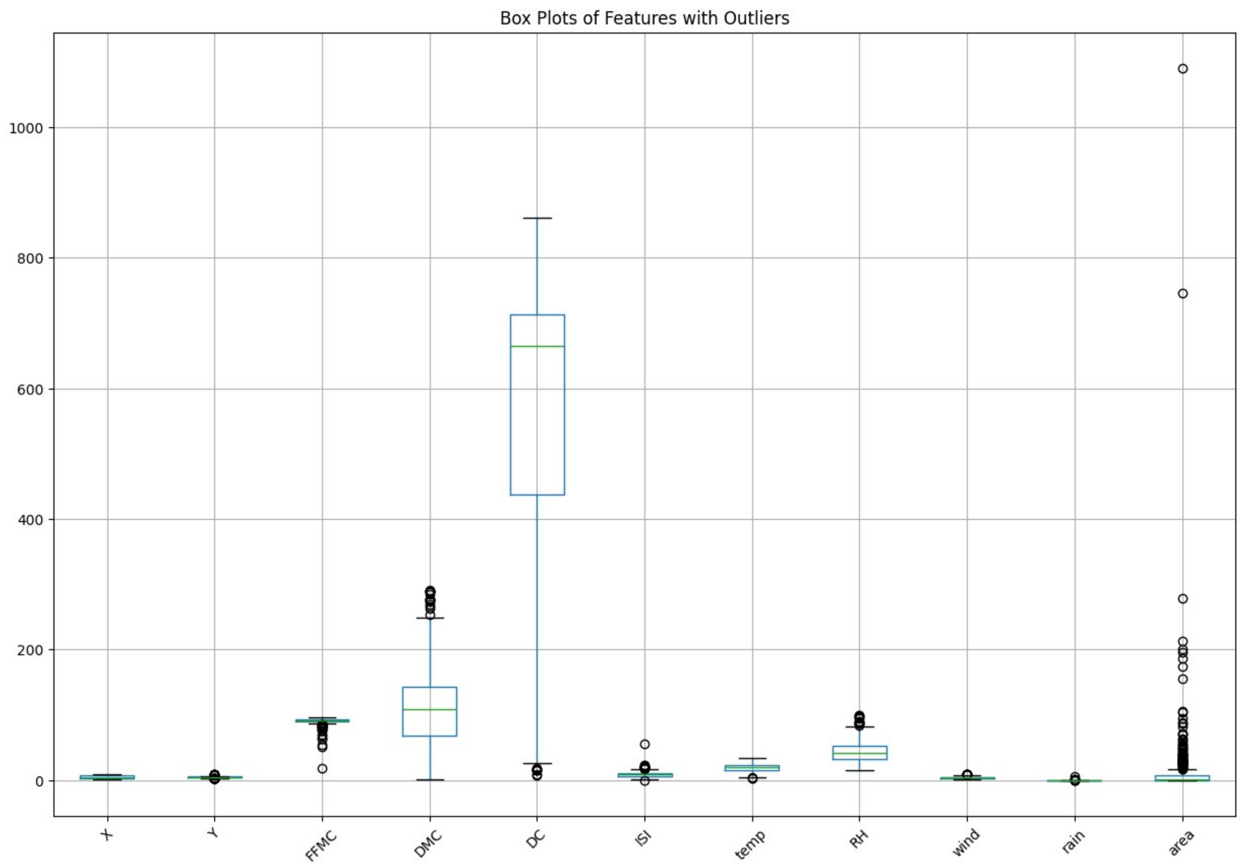
```
Data Types with Comma Separator:
X           int64
Y           int64
month      object
day        object
FFMC      float64
DMC       float64
DC        float64
ISI       float64
```

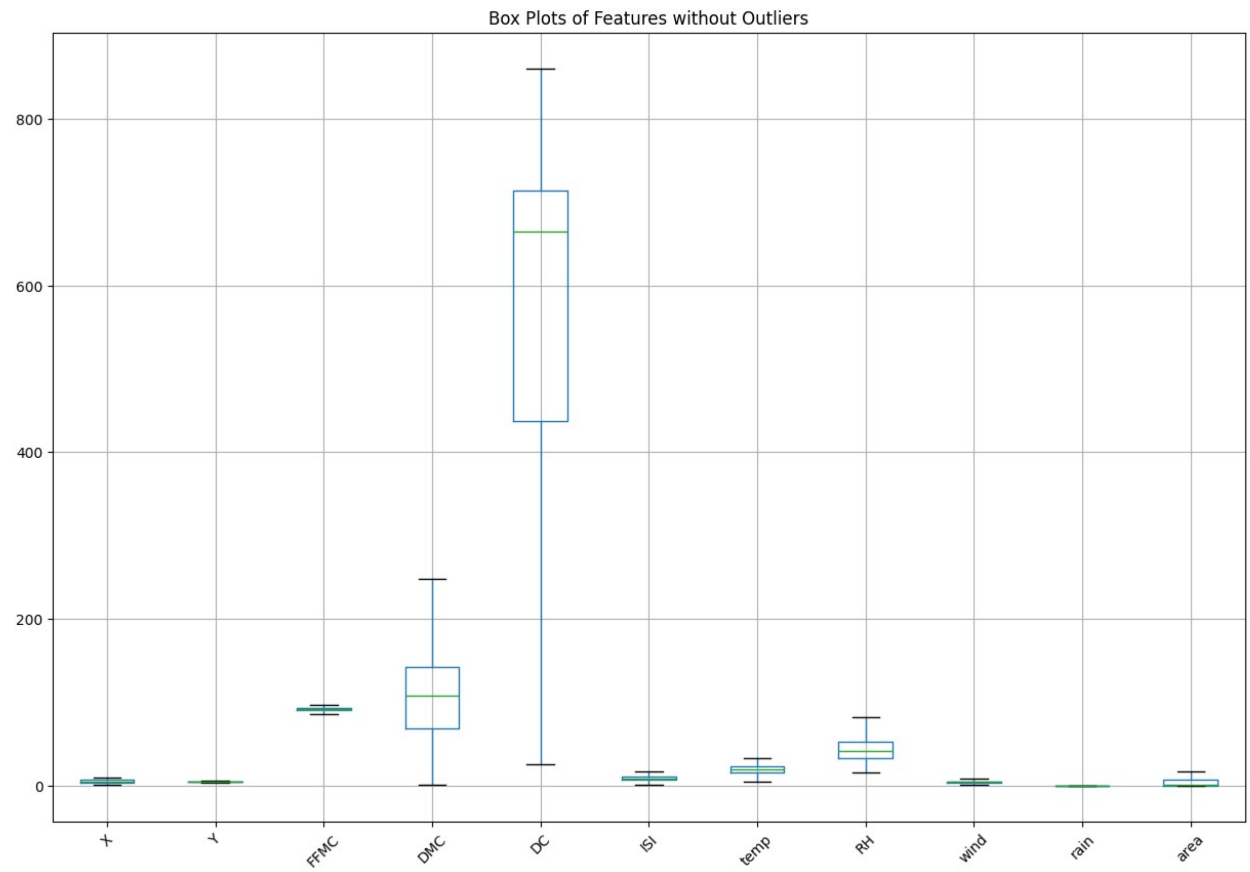
```
temp      float64
RH          int64
wind      float64
rain      float64
area      float64
dtype: object

# Select only numeric columns for the operations in the forest fires
dataset
numeric_data_forest_fires =
forest_fires_data.select_dtypes(include=[np.number])

# Computing summary statistics for numeric features in the dataset
summary_stats_forest_fires = numeric_data_forest_fires.describe()

# Directly calculating the range for each numeric feature
range_values_forest_fires = numeric_data_forest_fires.max() -
numeric_data_forest_fires.min()

# Calculating the variance for each numeric feature
variance_values_forest_fires = numeric_data_forest_fires.var()

# Adding the range and variance to the summary statistics DataFrame
summary_stats_forest_fires.loc['range'] = range_values_forest_fires
summary_stats_forest_fires.loc['variance'] =
variance_values_forest_fires

summary_stats_forest_fires

                   X           Y         FFMC          DMC
DC  \
count     517.000000  517.000000  517.000000   517.000000
517.000000
mean        4.669246    4.299807   90.644681   110.872340
547.940039
std         2.313778    1.229900    5.520111    64.046482
248.066192
min         1.000000    2.000000   18.700000     1.100000
7.900000
25%         3.000000    4.000000   90.200000    68.600000
437.700000
50%         4.000000    4.000000   91.600000   108.300000
664.200000
75%         7.000000    5.000000   92.900000   142.400000
713.900000
max         9.000000    9.000000   96.200000   291.300000
860.600000
range       8.000000    7.000000   77.500000   290.200000
852.700000
variance    5.353568    1.512655   30.471624  4101.951889
```

```
61536.835467

                 ISI        temp          RH        wind       
rain  \
count     517.000000  517.000000  517.000000  517.000000  517.000000

mean        9.021663   18.889168   44.288201    4.017602    0.021663

std         4.559477    5.806625   16.317469    1.791653    0.295959

min         0.000000    2.200000   15.000000    0.400000    0.000000

25%         6.500000   15.500000   33.000000    2.700000    0.000000

50%         8.400000   19.300000   42.000000    4.000000    0.000000

75%        10.800000   22.800000   53.000000    4.900000    0.000000

max        56.100000   33.300000  100.000000    9.400000    6.400000

range      56.100000   31.100000   85.000000    9.000000    6.400000

variance   20.788832   33.716898  266.259802    3.210019    0.087592


                 area
count      517.000000
mean        12.847292
std         63.655818
min          0.000000
25%          0.000000
50%          0.520000
75%          6.570000
max       1090.840000
range     1090.840000
variance  4052.063225

import matplotlib.pyplot as plt
import seaborn as sns

# Creating histograms for each numeric feature
def plot_histograms(data):
    data.hist(bins='auto', figsize=(20, 15))
    plt.suptitle('Histograms of Numeric Features in Forest Fires 
Dataset')
    plt.show()

# Creating box plots with outliers
def plot_boxplots_with_outliers(data):
    plt.figure(figsize=(15, 10))
    data.boxplot()
```

```python
    plt.title('Box Plots of Features with Outliers')
    plt.xticks(rotation=45)
    plt.show()

# Creating box plots without outliers
def plot_boxplots_without_outliers(data):
    plt.figure(figsize=(15, 10))
    data.boxplot(showfliers=False)
    plt.title('Box Plots of Features without Outliers')
    plt.xticks(rotation=45)
    plt.show()

# Call the functions to plot the graphs
plot_histograms(numeric_data_forest_fires)
plot_boxplots_with_outliers(numeric_data_forest_fires)
plot_boxplots_without_outliers(numeric_data_forest_fires)
```




Box Plots of Features with Outliers
1000
800
600
400
200
0
X
Y
FFMC
DMC
DC
ISI
temp
RH
wind
rain
area



1. Boxplot Analysis Without Outliers In the boxplot without outliers, certain features may exhibit significantly different distributions compared to others. These differences can be due to the range, median, and interquartile range (IQR) of these features. For instance:

- Features like `FFMC` (Fine Fuel Moisture Code) might have a narrower spread and higher median, indicating a concentration of values in a specific range.
- `DC` (Drought Code) might show a wider spread, suggesting more variability in the data.
- Features like `rain` might have a lower median and a smaller IQR, potentially indicating that rain is a less common event.

A significantly different distribution suggests that the feature behaves differently in terms of its central tendency and variability, which could be due to the nature of the variable itself or the conditions under which the data was collected.

1. Implications of Removing Outliers Outliers in a dataset can represent:

- Extreme but valid data points that are crucial for certain analyses, especially in understanding risk or rare events.
- Errors or anomalies in the data collection process.

Removing outliers might lead to:

- Loss of valuable information, especially if the outliers are genuine extreme values that are significant for the analysis.

- A biased dataset that does not accurately represent the real-world scenario the data is supposed to model.
- Poor performance of predictive models if they are trained on data that do not include these extreme but possible scenarios.

Hence, it's important to carefully consider the context and purpose of the analysis before deciding to remove outliers.

```
import matplotlib.pyplot as plt

# Filter FFMC values in the range [88, 96]
ffmc_filtered = forest_fires_data[(forest_fires_data['FFMC'] >= 88) &
(forest_fires_data['FFMC'] <= 96)]

# Plot histogram for FFMC in the specified range
plt.figure(figsize=(10, 6))
plt.hist(ffmc_filtered['FFMC'], bins='auto', color='blue',
edgecolor='black')
plt.title('Histogram of FFMC (Range 88 to 96)')
plt.xlabel('FFMC')
plt.ylabel('Frequency')
plt.show()
```