Assignment 2

Title: Exploring Data with Pandas

Name: Pranjal Rane

NUID: 002756852

In [1]:

```
# To get the Data uncomment and run below given 3 cells. If you already have the data, no
need to run the below 3 cells
```

In [2]:

```
# !wget https://archive.ics.uci.edu/static/public/186/wine+quality.zip
```

In [3]:

```
# !unzip wine+quality.zip
```

In [4]:

```
# !wget http://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires
.csv
```

In [5]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Objective 1

Wine Quality Dataset

In [6]:

```
data_red_wine = pd.read_csv('wine+quality/winequality-red.csv', sep=';')
data_white_wine = pd.read_csv('wine+quality/winequality-white.csv', sep=';')
```

In [7]:

```
data_red_wine['wineType'] = 'red'
data_white_wine['wineType'] = 'white'
```

In [8]:

```
data = pd.concat([data_red_wine, data_white_wine], ignore_index=True)
data.sample(5)
```

Out[8]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	wineType
5860	6.0	0.31	0.27	2.30	0.042	19.0	120.0	0.98952	3.32	0.41	12.7	7	white
5036	6.7	0.40	0.22	8.80	0.052	24.0	113.0	0.99576	3.22	0.45	9.4	5	white
4387	6.9	0.40	0.42	6.20	0.066	41.0	176.0	0.99552	3.12	0.54	9.4	5	white
4943	6.9	0.38	0.29	13.65	0.048	52.0	189.0	0.99784	3.00	0.60	9.5	6	white

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur	total sulfur	density	pH	sulphates	alcohol	quality	wineType
3737	6.6	0.24	0.32	6.70	0.060	19.0	102.0	0.99400	3.27	0.31	10.4	5	white

Objective 1.1

Summary Statistics Compute and display summary statistics for each feature available in the dataset. These must include: 1) minimum value 2) maximum value 3) mean 4) range 5) standard deviation 6) variance 7) count 8) 25:50:75 percentiles.

In [9]:

```
summary_statistics = data.describe(include='number')
numeric_data = data.select_dtypes(include='number')
summary_statistics.loc['range'] = numeric_data.max() - numeric_data.min()
summary_statistics.loc['variance'] = numeric_data.var()

summary_statistics
```

Out[9]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	I
count	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.0000
mean	7.215307	0.339666	0.318633	5.443235	0.056034	30.525319	115.744574	0.994697	3.2185
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.749400	56.521855	0.002999	0.1607
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.7200
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.1100
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.2100
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.3200
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980	4.0100
range	12.100000	1.500000	1.660000	65.200000	0.602000	288.000000	434.000000	0.051870	1.2900
variance	1.680740	0.027105	0.021117	22.636696	0.001227	315.041192	3194.720039	0.000009	0.0258

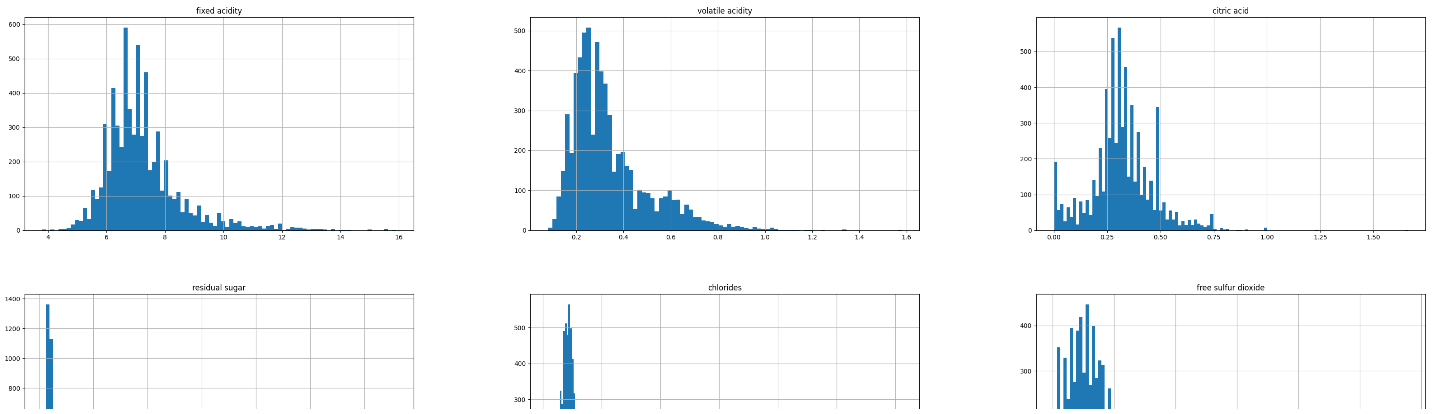
Objective 1.2

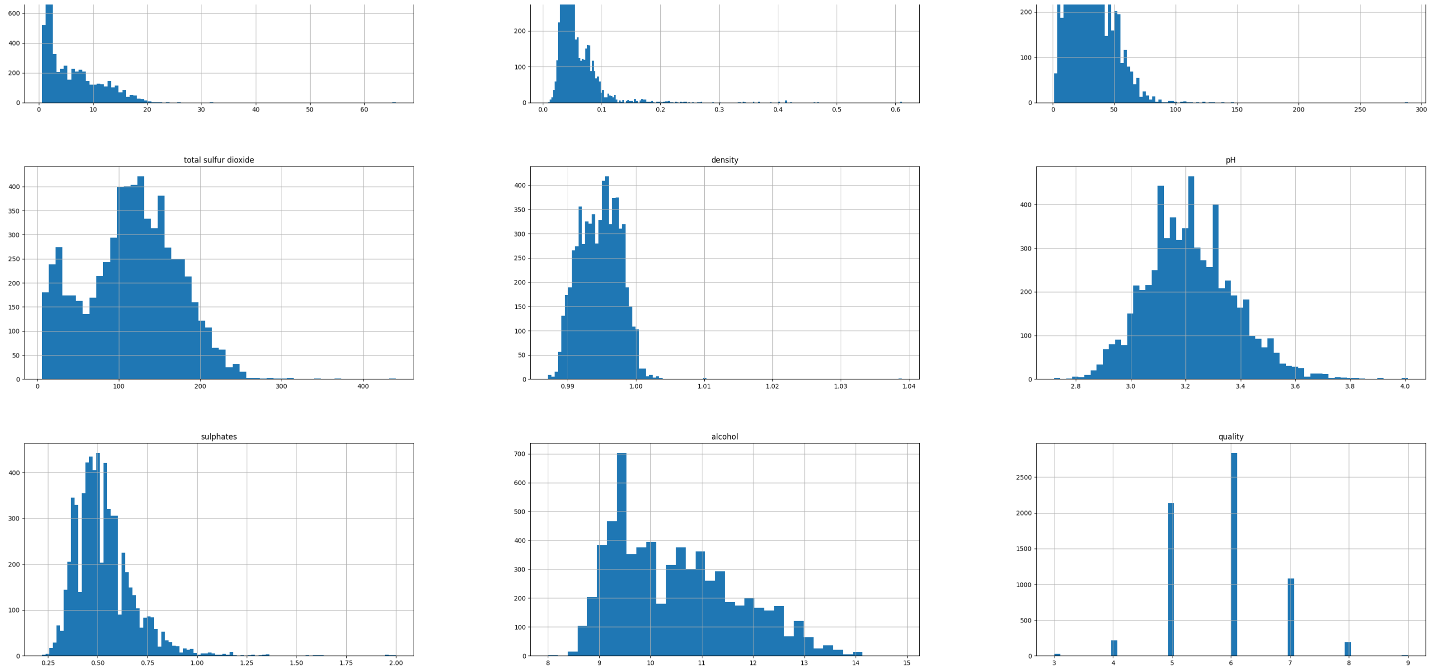
Data Visualization

Histograms: To illustrate the feature distributions, create a histogram for each feature in the dataset. You may plot each histogram individually or combine them all into a single plot. When generating histograms for this assignment, use the default number of bins. Recall that a histogram provides a graphical representation of the distribution of the data.

In [10]:

```
data.hist(bins='auto', figsize=(40,30))
plt.show()
```

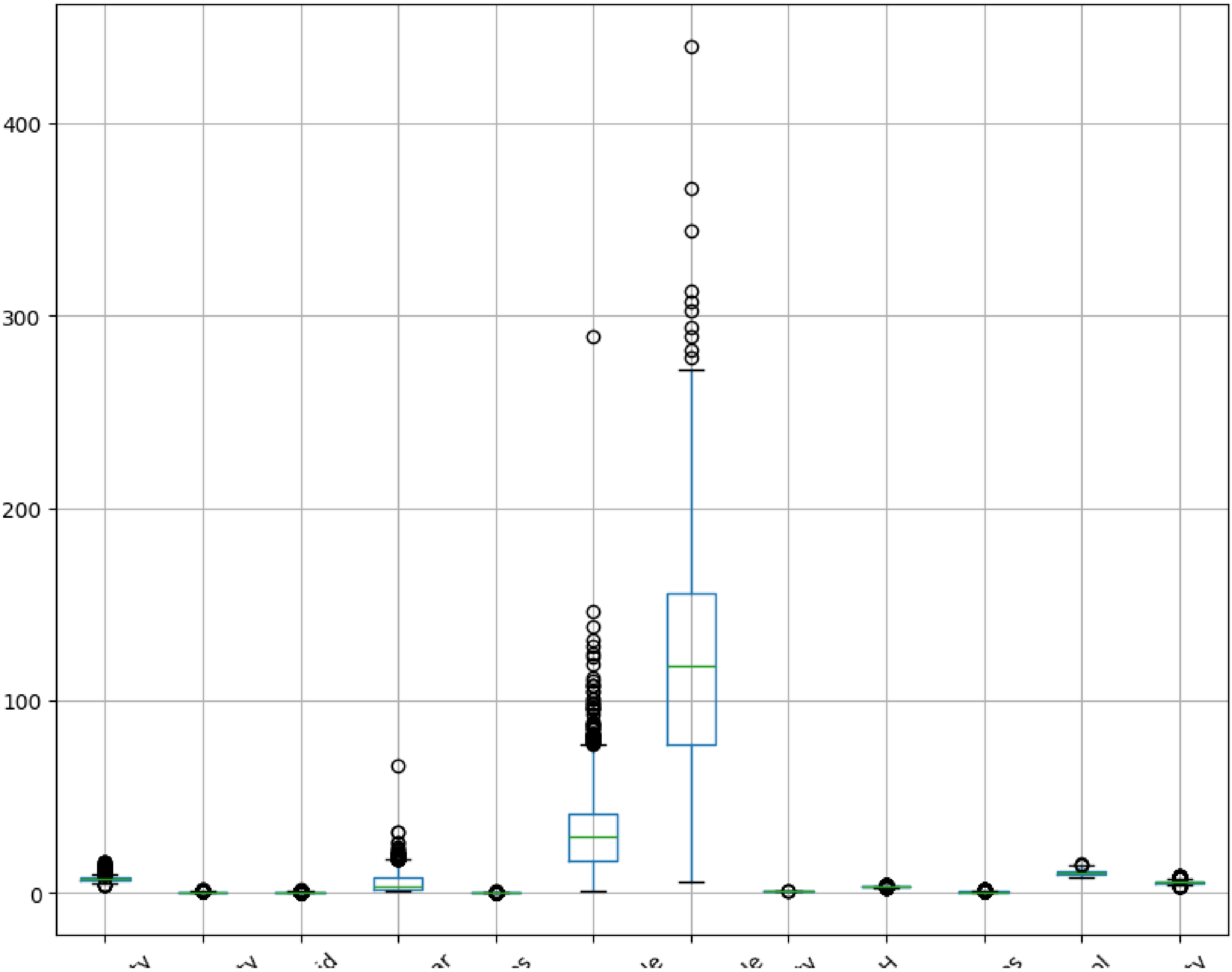


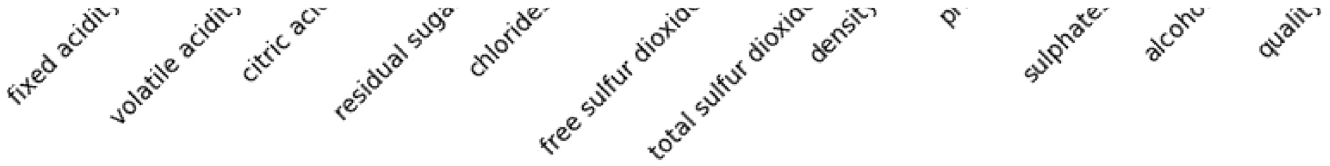


Box Plots: To further assess the data, create a boxplot for each feature in the dataset. All of the boxplots will be combined into a single plot. Recall that a boxplot provides a graphical representation of the location and variation of the data through their quartiles; they are especially useful for comparing distributions and identifying outliers.

In [11]:

```
plt.figure(figsize=(10, 8))
data.boxplot()
plt.xticks(rotation=45)
plt.show()
```

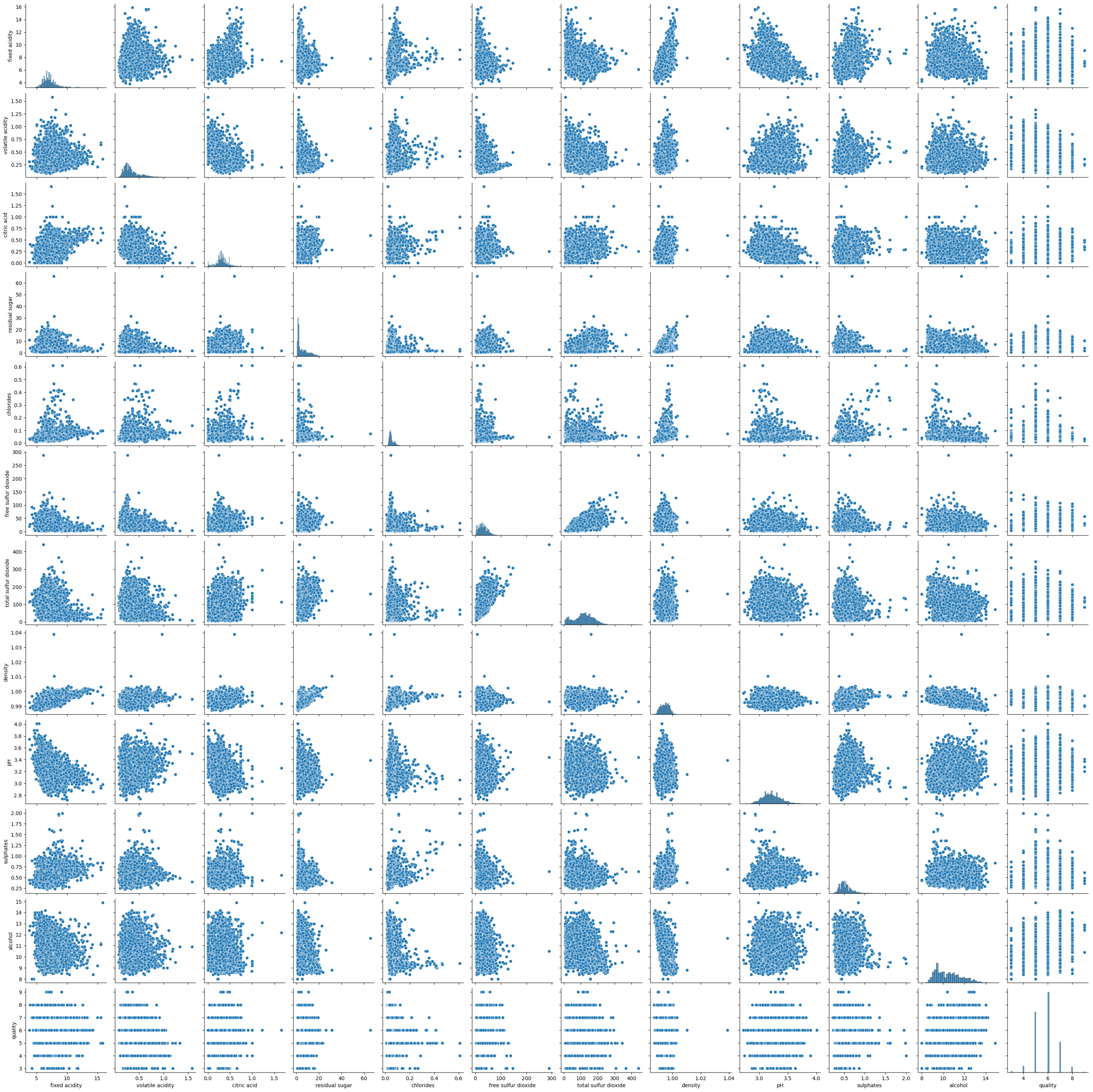




Pairwise Plot: To understand the relationship between the features, create scatter plot for each pair of the features. If there are n features in the dataset, there should be nC2 plots.

In [12]:

```
sns.pairplot(numeric_data)
plt.show()
```



Class-wise Visualization: Create pairwise plots for each pair of features in a similar way for each of the different classes present in the data.

In [ ]:

```
sns.pairplot(data, hue="wineType")
plt.show()
```

Objective 1.3

Conceptual Questions

1. **How many features are there? What are the types of the features (e.g., numeric, nominal, discrete, continuous)?**

   There are a total of 13 features in the given wine dataset. They are:

   - **fixed acidity - Continuous**
   - **volatile acidity - Continuous**
   - **citric acid - Continuous**
   - **residual sugar - Continuous**
   - **chlorides - Continuous**
   - **free sulfur dioxide - Continuous**
   - **total sulfur dioxide - Continuous**
   - **density - Continuous**
   - **pH - Continuous**
   - **sulphates - Continuous**
   - **alcohol - Continuous**
   - **quality - Discrete**
   - **wineType - Nominal**

1. **What can you conclude from the histograms about the distribution of the features in the dataset? Are there any features that are approximately normally distributed? Are there any features that are highly skewed?**

   Histograms provide crucial insights into the distribution of features in a dataset. They allow us to understand the nature of the distribution, be it normal or skewed, by examining the arrangement of the data. The width or range of the histogram sheds light on the data's variability or spread. Moreover, histograms are instrumental in identifying outliers, which are the unusual values that stand apart from the bulk of the data.

   The distribution of features based on the histograms -

   - **chlorides - Highly Skewed (Right skewed)**
   - **pH - Normally Distributed**
   - **residual sugar - Highly Skewed (Right skewed)**
   - **free sulfur dioxide - Highly Skewed (Right skewed)**

1. **Based on the box plots, are there any features that appear to have many outliers? Are there any features that appear to have a similar spread of values across different quality ratings? Are there any features that appear to have different spreads of values across different quality ratings?**

   Features that are exhibiting a notable abundance of outliers are:

   - **sulfates**
   - **chlorides**
   - **citric acid**
   - **free sulfur dioxide**
   - **fixed acidity**
   - **residual sugar**
   - **volatile acidity.**

   Features that are displaying comparable ranges of values are:

   - **citric acid**
   - **sulfates**
   - **fixed acidity**
   - **pH.**

   Features that have dissimilar ranges of values are:

   - **free sulfur dioxide**

- residual sugar
- quality
- chlorides alcohol

1. **Based on the pairwise plots, which features appear to be highly correlated? Are there any features that do not appear to be correlated with any other features?**

   Features that are exhibiting strong correlations are:

   - **density and residual sugar**
   - **free sulfur dioxide and total sulfur dioxide**

   Features that are not exhibiting any significant correlations with other features are:

   - **residual sugar**
   - **alcohol**
   - **total sulfur dioxide**

1. **Based on the class-wise visualizations, are there any pairs of features that appear to be more correlated for certain wine types than for others?**

   For red wine, these pairs exhibit higher correlations compared to white wine:

   - **Citric acid and fixed acidity**
   - **Density and fixed acidity**
   - **Chlorides and sulfates**

   On the other hand, for red wine, the following pairs are more correlated than for white wine:

   - **Density and residual sugar**
   - **Total sulfur dioxide and density**
   - **Total sulfur dioxide and residual sugar**

Objective 2

Forest Fires Dataset

In [ ]:

```
data_2 = pd.read_csv('forestfires.csv')
data.head()
```

Objective 2.1

Summary Statistics

In [ ]:

```
summary_statistics_2 = data_2.describe(include='number')
numeric_data_2 = data_2.select_dtypes(include='number')
summary_statistics_2.loc['range'] = numeric_data_2.max() - numeric_data_2.min()
summary_statistics_2.loc['variance'] = numeric_data_2.var()

summary_statistics_2
```

Objective 2.2

Data Visualization

As done in Section 1, create histograms and boxplots for the dataset. Now, create another boxplot without the outliers. You can use showfliers=False to remove outliers from the boxplots. You are expected to present two Boxplots in total.

Boxplots in total.

In [ ]:

```
data_2.hist(bins='auto', figsize=(20,16))
plt.show()
```

In [ ]:

```
data_2.boxplot(figsize=(10, 8), showfliers=True)
plt.title('Boxplot of Features')
plt.xticks(rotation=45)
plt.show()
```

In [ ]:

```
data_2.boxplot(figsize=(10, 8), showfliers=False)
plt.title('Boxplot of Features')
plt.xticks(rotation=45)
plt.show()
```

Objective 2.3

Conceptual Questions

1. **From the boxplot without outliers, which features has a significantly different distribution from others? Why?**

   Features that have a significantly different distribution from others are :

   - **X**
   - **Y**
   - **DC**
   - **rain**
   - **area**

Outliers significantly impact the distribution of a feature and its respective boxplot representation. They can induce skewness, extend the boxplot's range, and shift the median, resulting in marked differences in distribution compared to features with fewer outliers. Features with a higher occurrence of outliers may exhibit wider ranges, extended whiskers, and increased variances, indicating potential substantial departures from the trends seen in other features.

1. **What does the outlier in the features mean? If you remove the outliers from the dataset, what problems might arise?**

   Outliers, which stand out significantly from the rest of the data, can profoundly influence analyses and model building. They have the potential to warp statistical measures, alter distribution shapes, influence correlations and regression lines, and affect the performance of predictive models. However, indiscriminately discarding these outliers can introduce bias, alter the perceived variability of the data, and lead to the loss of critical information. Not every outlier represents a mistake; some are indicative of rare occurrences or substantial variations within the data. Therefore, it's crucial to carefully evaluate the nature of outliers before deciding to remove them. This approach helps retain essential information and ensures a thorough comprehension of the dataset. Outliers can challenge the assumptions underlying many models, potentially undermining the validity of the analyses that follow. Yet, eliminating these outliers can also limit the model's ability to generalize effectively when confronted with new data that may include similar outliers.

1. **Create a histogram for only FFMC after removing all the values outside of range [88, 96].**

In [ ]:

```
required_data = data_2[(data_2['FFMC'] >= 88) & (data_2['FFMC'] <= 96)]['FFMC']
plt.hist(required_data)
plt.show()
```

1. **What distribution does the new histogram follow?**

   The new histogram plotted for FFMC (range between 88 and 96) clearly follows normal distribution.