INTRODUCTION

This Jupyter notebook is part of my learning experience in the study of central tendency. I will be working with a simple dataset that contains details of wine quality. In this exercise, I will perform the following tasks:

Load and study the data: I will import the dataset and analyze its structure and properties. View the distributions of the various features in the dataset and calculate their central tendencies: I will visualize the distributions of different features present in the dataset and compute their central tendencies, including measures such as mean, median, and mode. Create a new Pandas Series that contains the details of the representative factor for quality: I will generate a new Pandas Series specifically dedicated to representing the factor that influences wine quality. *By completing these tasks, I aim to gain a deeper understanding of the central tendencies exhibited by the wine quality dataset, enabling me to better comprehend its overall composition and characteristics.

Task 1 - Load and study the data:

To begin, we will load the dataset and examine its features, including:

*Fixed acidity

*Volatile acidity

*Citric acid

Now, let's proceed with the implementation.

In [1]:
```
# Load "numpy" and "pandas" for manipulating numbers and data frames
# Load "matplotlib.pyplot" and "seaborn" for data visualisation

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

*Now that we have imported the necessary libraries, we can continue with our analysis.

In [2]:
```
# Read the "winequality-white.csv" file and create a Pandas DataFrame

df = pd.read_csv(r'E:\PORTFOLIO-PROJECTS\STATISTICS - Data Analysis (Wine Quality)\Pyt
```

In [4]:
```
# Displaying a preview of the dataset

df.head()
```

Out[4]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	qua
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	

*By examining the first few rows of the DataFrame, we can get a glimpse of the data and its structure.

In [8]:
```
# Retrieving the dimensions of the dataframe

df.shape
```

Out[8]:
```
(4898, 12)
```

*The dataframe contains {4898} rows and {12} columns. This information will help us understand the size and structure of the dataset.

In [10]:
```
# Retrieving the row names of the dataframe

df.index
```

Out[10]:
```
RangeIndex(start=0, stop=4898, step=1)
```

In [12]:
```
# Retrieving the column names of the dataframe

df.columns
```

Out[12]:
```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

*The dataframe includes the following column names: {'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol', 'quality'}. These names represent the different features or attributes present in the dataset.

In [13]:
```
# Obtaining basic information about the dataframe

df.info()
```

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

*The dataframe provides essential details such as the column names, data types, and the number of non-null values. This information is helpful in understanding the structure and integrity of the dataset.

In [14]:
```
# Checking for null values in the dataframe

df.isna().sum()
```

Out[14]:
```
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
```

*The dataframe contains the following number of null values for each column: {0}. This information allows us to identify any missing values that may need to be addressed in our analysis.

In [15]:
```
df.describe()
```

Out[15]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	48
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	

Observations from Task 1-->

Upon examining the data, the following observations can be made:

The dataset comprises 4898 rows and 12 columns, indicating that we have information for 4898 instances of white wine samples. Each row represents the details of various types of acids present in white wine, along with the corresponding quality assessment. Specifically, the dataset includes features such as different acids (e.g., fixed acidity, volatile acidity, citric acid) and their associated quality ratings.

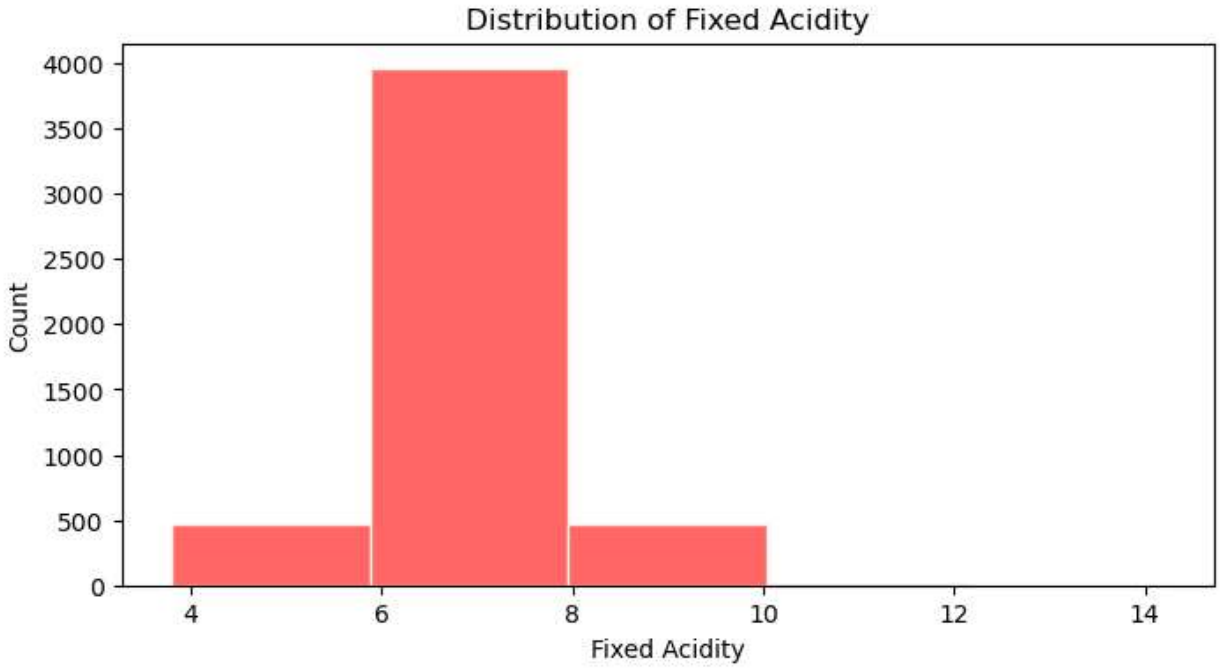
These initial observations lay the foundation for our further analysis and exploration of the dataset.

Task 2 - View the distributions of the various features in the dataset and calculate their central tendencies:

In this task, we will explore the distributions of the different features present in the dataset. Additionally, we will calculate appropriate measures of central tendency to gain insights into the data. Let's proceed with the analysis.

In [19]:
```
# Creating a histogram of the "Fixed acidity" feature

plt.figure(figsize = (8,4))
sns.histplot(data = df, x = df['fixed acidity'], color='Red', alpha = 0.6, bins = 5, e
plt.title('Distribution of Fixed Acidity')
plt.xlabel('Fixed Acidity')
plt.ylabel('Count')
plt.show()
```



*The histogram provides a visual representation of the distribution of the "Fixed acidity" feature in the dataset. It allows us to observe the frequency of different acidity levels present in the white wine samples.

Observations from Task 2-->

Upon analyzing the histogram of the "Fixed acidity" feature, the following observations can be made:

The distribution appears to follow a normal distribution pattern, with a peak in the range of 6 to 8 on the x-axis. The majority of white wine samples in the dataset have fixed acidity levels within this range. These observations provide us with an initial understanding of the distribution and concentration of fixed acidity values in the dataset.

Examining Measures of Central Tendency:

Now, let's calculate and explore the measures of central tendency for the dataset's features, including the mean, median, and mode. By doing so, we can gain insights into the typical or representative values within each feature.

In [21]:
```
# Calculating the mean of the "Fixed acidity" feature

mean_fixed_acidity = df['fixed acidity'].mean()
mean_fixed_acidity
```

Out[21]:
```
6.854787668436075
```

*The mean of the "Fixed acidity" feature is {6.854787668436075}. This value represents the average or typical value for fixed acidity in the white wine samples.

In [22]:
```
# Calculating the rounded mean of the "Fixed acidity" feature

round_mean_fixed_acidity = round(df['fixed acidity'].mean(),2)
round_mean_fixed_acidity
```

Out[22]:
```
6.85
```

*The rounded mean of the "Fixed acidity" feature is {6.85}. This value represents the average or typical value for fixed acidity in the white wine samples, rounded to two decimal places.

In [23]:
```
# Calculating the median of the "Fixed acidity" feature

median_fixed_acidity = df['fixed acidity'].mean()
median_fixed_acidity
```

Out[23]:
```
6.854787668436075
```

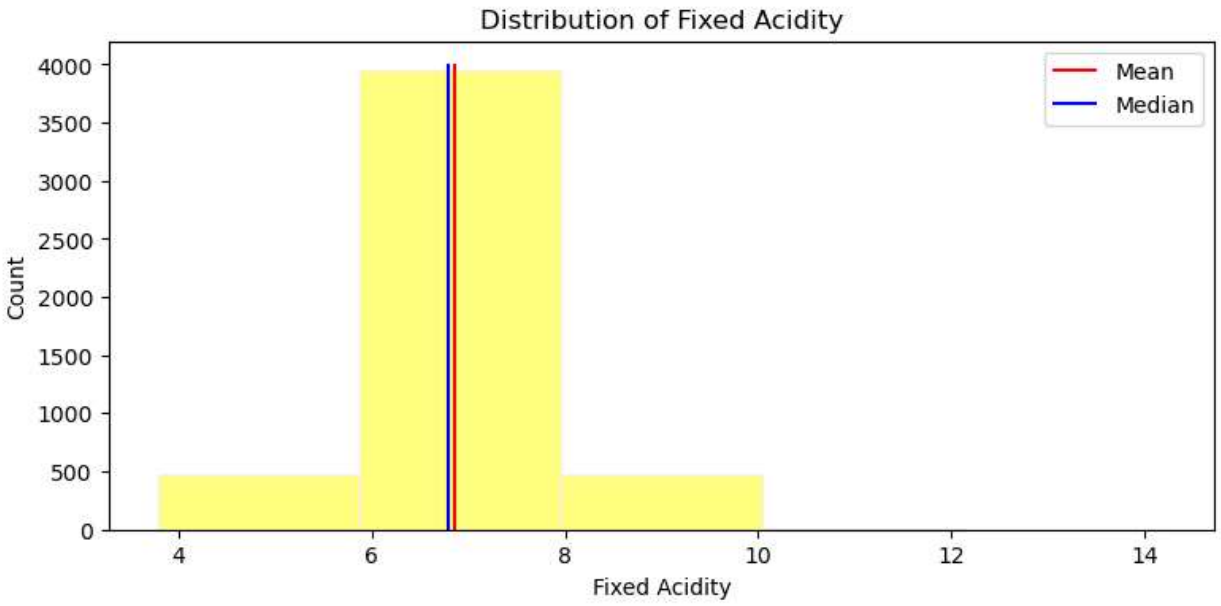
*The median of the "Fixed acidity" feature is {6.8}. This value represents the middle value when the fixed acidity values are arranged in ascending order.

In [28]:
```
# Creating a histogram of the "Fixed acidity" feature with mean and median

plt.figure(figsize = (9,4))
sns.histplot(data = df, x = 'fixed acidity', color = 'Yellow', edgecolor = 'linen', al
plt.title("Distribution of Fixed Acidity")
plt.xlabel('Fixed Acidity')
plt.ylabel('Count')

# Adding vertical lines for mean and median

plt.vlines(df['fixed acidity'].mean(), ymin = 0, ymax = 4000, colors = 'Red', label =
plt.vlines(df['fixed acidity'].median(), ymin = 0, ymax = 4000, colors = 'Blue', label
plt.legend()
plt.show()
```



*The histogram displays the distribution of the "Fixed acidity" feature, with vertical lines indicating the position of the mean and median. This visualization allows us to compare the central tendency measures with the overall distribution of fixed acidity values.

Observations-->

After analyzing the histogram and considering the measures of central tendency, the following observations can be made:

The mean and median values of the "Fixed acidity" feature are close to each other, indicating that they are both representative measures of central tendency. This suggests that the distribution of fixed acidity values is relatively symmetric around the central value. Therefore, we can choose either the mean or median as a suitable measure of central tendency for this feature. These observations allow us to better understand the representative value of fixed acidity in the white wine samples and its distribution characteristics.

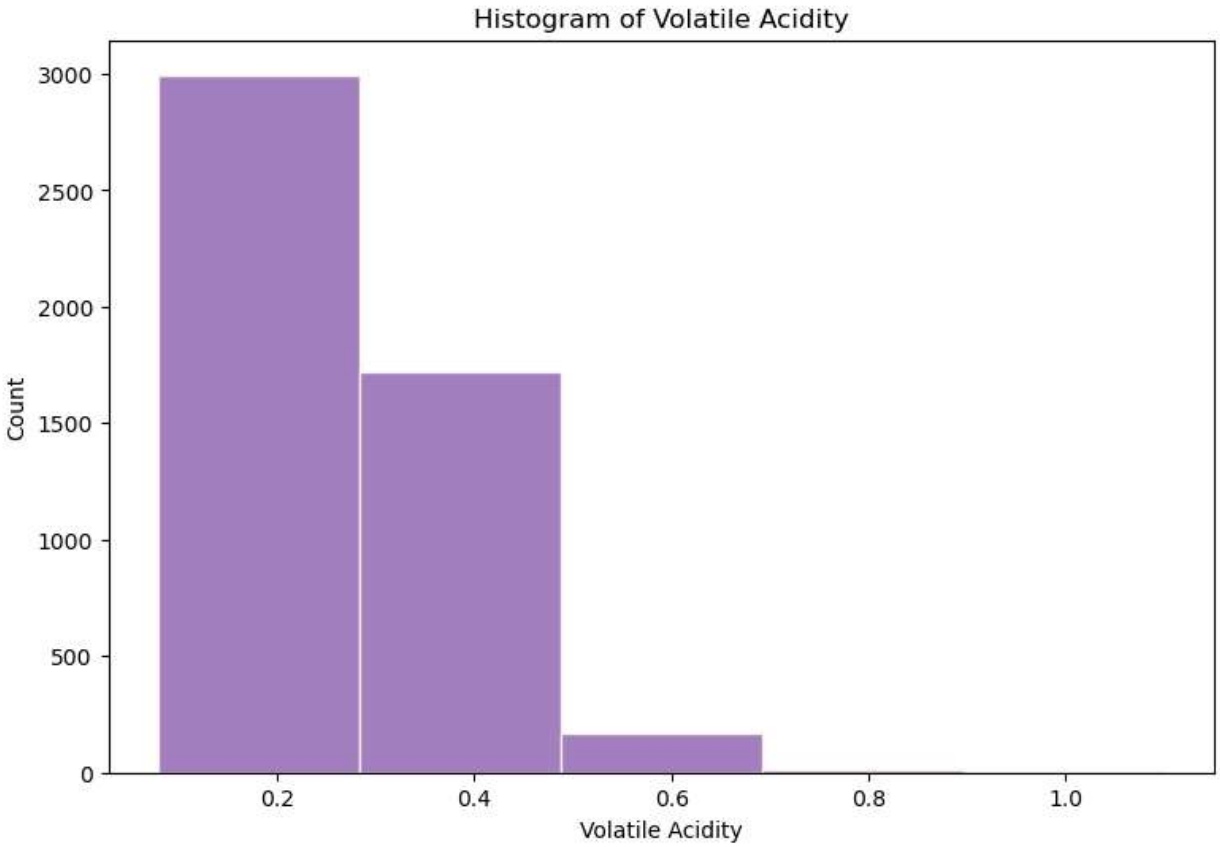
Histogram of the "Volatile Acidity" Feature:

Now, let's create a histogram to visualize the distribution of the "volatile acidity" feature in the dataset. This histogram will provide insights into the frequency and concentration of volatile acidity values in the white wine samples.

Let's proceed with the implementation.

In [36]:
```
# Create a histogram of the "volatile acidity" feature

plt.figure(figsize = (9,6))
sns.histplot(data = df, x =df['volatile acidity'], color = 'indigo', edgecolor = 'line
plt.title('Histogram of Volatile Acidity')
plt.xlabel('Volatile Acidity')
plt.ylabel('Count')
plt.show()
```



Observations -->

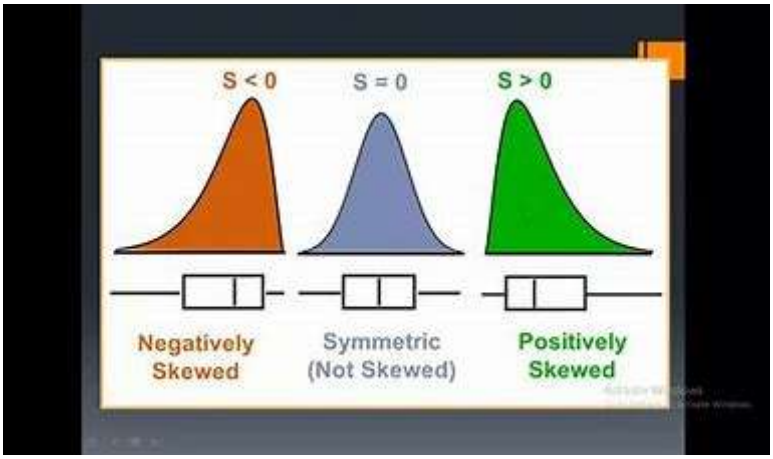
Upon analyzing the histogram of the "Volatile Acidity" feature, the following observations can be made:

The distribution appears to be slightly skewed towards the right, indicating that there is a concentration of lower volatile acidity values. This skewed distribution suggests that the majority of white wine samples have relatively lower volatile acidity levels. To further examine the distribution and visualize any potential skewness, we can utilize the distplot function.

Certainly! Skewness is a statistical measure that helps us understand the symmetry of a distribution.

Here's how to interpret skewness:

*If the skewness (S) is equal to 0, the distribution is considered to be normally distributed, meaning it has a symmetric shape.*If the skewness (S) is greater than 0, the distribution is positively skewed, indicating that the tail of the distribution extends towards higher values. *If the skewness (S) is less than 0, the distribution is negatively skewed, indicating that the tail of the distribution extends towards lower values. By considering the skewness value, we can determine the direction and degree of skewness in a distribution, which provides valuable insights into the shape and characteristics of the data.



Distribution Plot (Distplot) of the "Volatile Acidity" Feature:

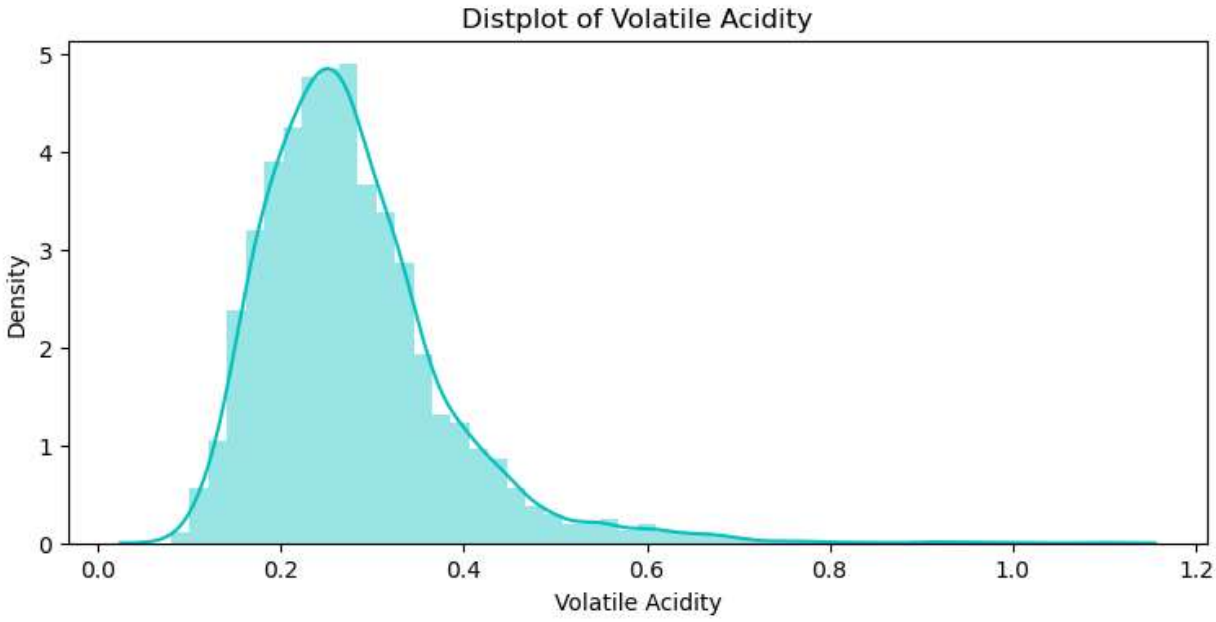
To gain a better understanding of the distribution and potential skewness of the "volatile acidity" feature, we will create a distribution plot (distplot). This plot will provide a visual representation of the density of volatile acidity values in the white wine samples.

Let's proceed with plotting the distplot.

In [39]:
```
# Distribution Plot (Distplot) of the "Volatile Acidity" Feature:

plt.figure(figsize = (9,4))
sns.distplot(df['volatile acidity'], color = 'c')
plt.title('Distplot of Volatile Acidity')
plt.xlabel('Volatile Acidity')
plt.ylabel('Density')
plt.show()
```

```
C:\Users\HOSHANGI\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWa
rning: `distplot` is a deprecated function and will be removed in a future version. P
lease adapt your code to use either `displot` (a figure-level function with similar f
lexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



Observation-->

Upon analyzing the distribution plot (distplot) of the "Volatile Acidity" feature, the following observations can be made:

*The plot indicates that the distribution of volatile acidity values appears to follow a normal distribution.*A normal distribution is characterized by several key properties, including: *The mean and median of the distribution are equal, indicating a symmetric shape.*The distribution has only one mode, representing the most frequent value. *It is symmetric, meaning that the values decrease equally on both sides of the center.*The shape of the plot resembles a "bell curve," which is a common representation of a normal distribution. *These observations provide insights into the distribution characteristics of the volatile acidity feature and suggest that it follows a normal distribution pattern.

In [40]:
```
# Calculate skewness of 'Volatile Acidity'
round(df['volatile acidity'].skew(),2)
```

Out[40]:
```
1.58
```

Observation-->

Upon calculating the skewness of the "Volatile Acidity" feature, we find that the skewness value is greater than 1. This indicates that the distribution of volatile acidity values is positively skewed.

A positive skewness suggests that the tail of the distribution extends towards higher values, while the majority of the data is concentrated towards lower values.

This observation aligns with our earlier analysis, where we noted a slight skewness towards the right in the histogram and the distribution plot.

In [41]:
```
# Calculate the mean "Volatile Acidity" feature

df['volatile acidity'].mean()
```

Out[41]:
```
0.27824111882401087
```

In [42]:
```
# Calculate the median "Volatile Acidity" feature

df['volatile acidity'].median()
```

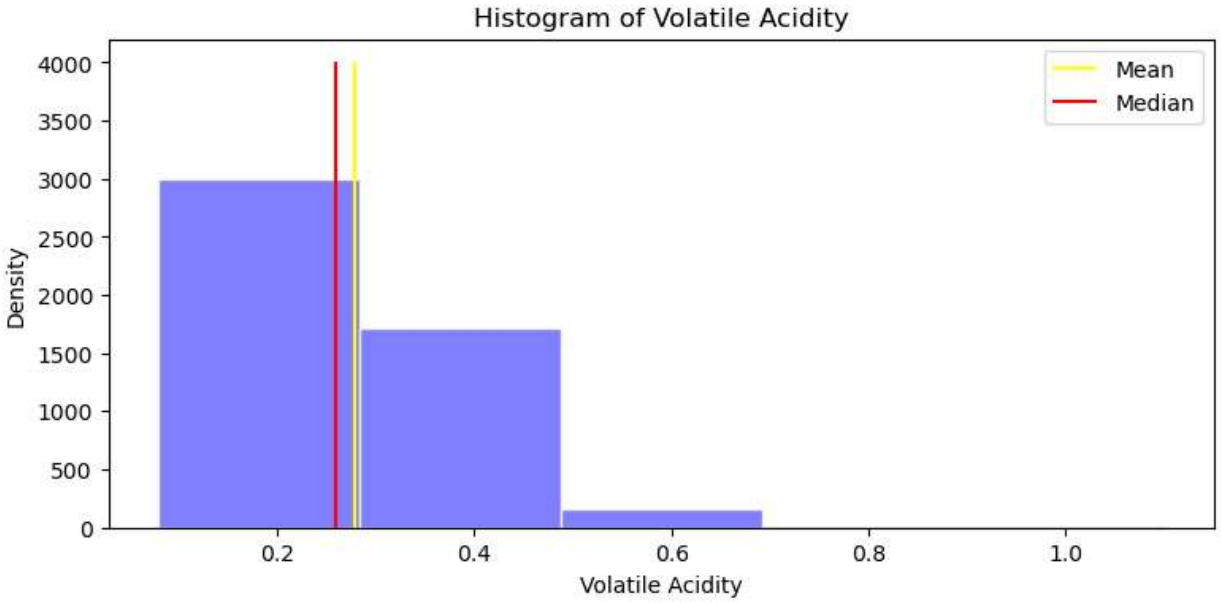
Out[42]:
```
0.26
```

In [47]:
```
# Create a histogram of the "Volatile Acidity" feature and also show the mean and the

plt.figure(figsize = (9,4))

sns.histplot(data = df ,x = 'volatile acidity', color = 'blue',
             edgecolor = 'linen', alpha = 0.5, bins = 5)
```

```
plt.title("Histogram of Volatile Acidity")
plt.xlabel('Volatile Acidity')
plt.ylabel('Density')
plt.vlines(df['volatile acidity'].mean(), ymin = 0, ymax = 4000, colors='yellow', labe
plt.vlines(df['volatile acidity'].median(), ymin = 0, ymax = 4000, colors='red', label
plt.legend()
plt.show()
```

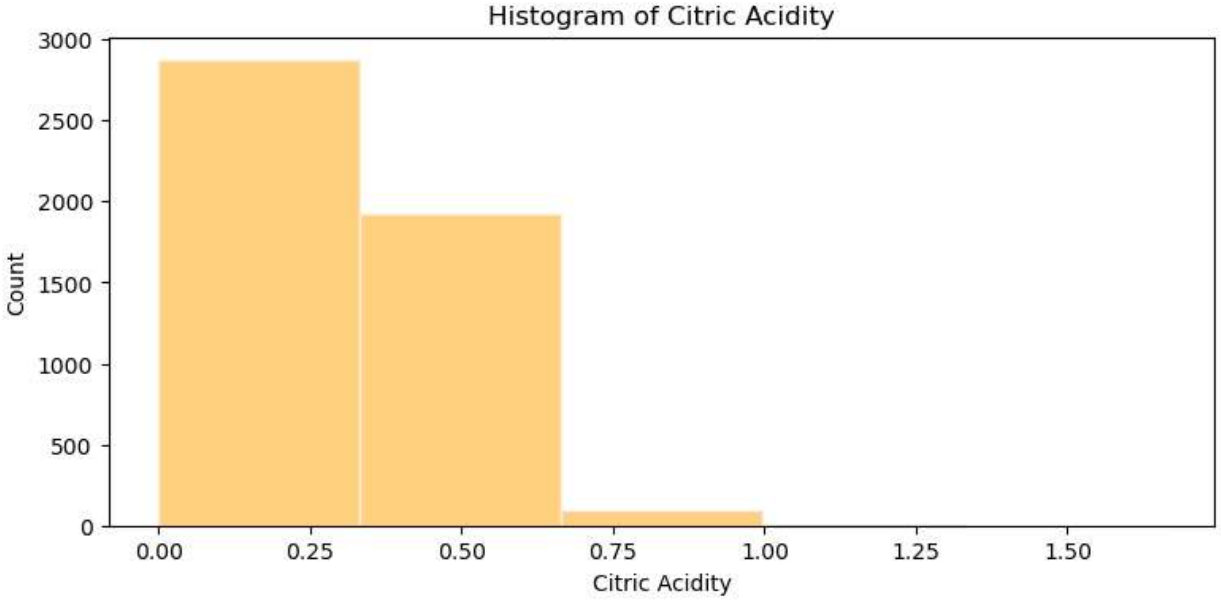


In [49]:
```
# Create a histogram of the "citric acid" feature

plt.figure(figsize = (9,4))

sns.histplot(data = df ,x = 'citric acid', color = 'orange',
             edgecolor = 'linen', alpha = 0.5, bins = 5)

plt.title("Histogram of Citric Acidity")
plt.xlabel('Citric Acidity')
plt.ylabel('Count')
plt.show()
```



Observation-->

Upon analyzing the histogram of the "Volatile Acidity" feature, we observe that the distribution is slightly skewed towards the right. This indicates that there is a concentration of lower volatile acidity values in the white wine samples.

The skewed distribution suggests that the majority of the white wine samples have relatively lower volatile acidity levels. It is important to consider this skewness when analyzing and interpreting the data, as it may have implications for the overall flavor and quality of the wines.

Understanding the distribution characteristics of volatile acidity is valuable for assessing its impact on the sensory profile and perceived quality of the white wines.

In [50]:
```
# Calculate the mean "Citric Acid" feature

df['citric acid'].mean()
```

Out[50]:
```
0.33419150673743736
```

In [51]:
```
# Calculate the median "Citric Acid" feature

df['citric acid'].median()
```

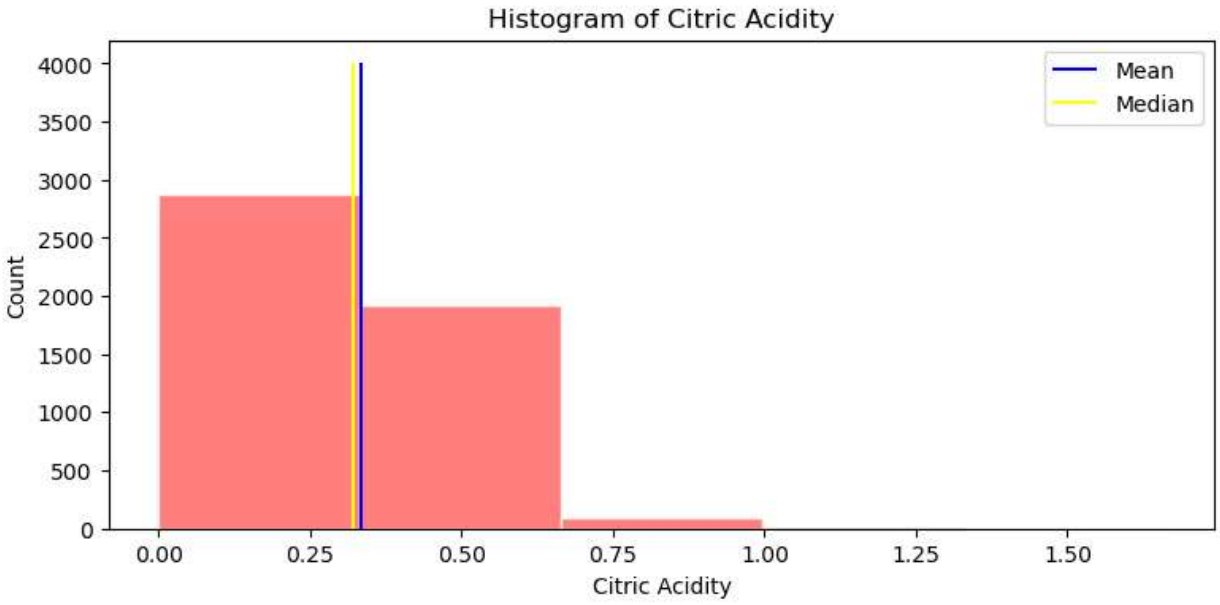
Out[51]:
```
0.32
```

In [54]:
```
# Create a histogram of the "Citric Acid" feature and also show the mean and the media

plt.figure(figsize = (9,4))

sns.histplot(data = df ,x = 'citric acid', color = 'red',
             edgecolor = 'linen', alpha = 0.5, bins = 5)

plt.title("Histogram of Citric Acidity")
plt.xlabel('Citric Acidity')
plt.ylabel('Count')
plt.vlines(df['citric acid'].mean(), ymin = 0, ymax = 4000, colors='blue', label='Mean
plt.vlines(df['citric acid'].median(), ymin = 0, ymax = 4000, colors='yellow', label='
plt.legend()
plt.show()
```



Observation-->

Upon examining the mean and median of the "Volatile Acidity" feature, we find that they are close to each other and exhibit a small difference. This suggests a relatively symmetrical distribution of the data.

Given the proximity of the mean and median, we can confidently select the mean as an appropriate measure of central tendency for this feature. The mean provides an estimate of the typical or average value of volatile acidity in the white wine samples.

By considering the mean as a representative measure, we gain valuable insights into the central tendency of the volatile acidity and its overall influence on the sensory attributes and quality of the white wines.

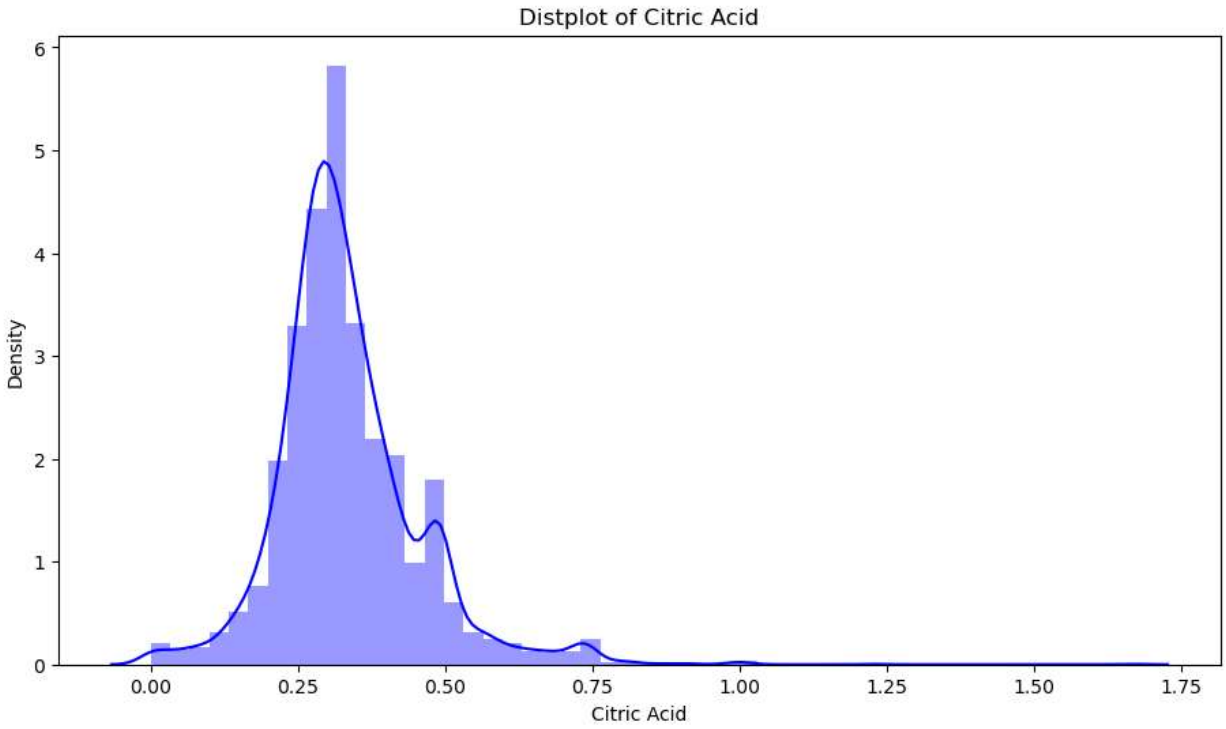
In [55]:
```
# Calculate distplot using 'Citric Acidity' feature

plt.figure(figsize = (11,6))

sns.distplot(df['citric acid'], color = 'blue')

plt.title("Distplot of Citric Acid")
plt.xlabel('Citric Acid')
plt.ylabel('Density')
plt.show()
```

```
C:\Users\HOSHANGI\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWa
rning: `distplot` is a deprecated function and will be removed in a future version. P
lease adapt your code to use either `displot` (a figure-level function with similar f
lexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



In [56]:
```
quality = pd.DataFrame(df['quality'].value_counts())
quality
```

Out[56]:

	quality
6	2198
5	1457
7	880
8	175
4	163
3	20
9	5

In [57]:
```
q_index=quality.index
q_index
```

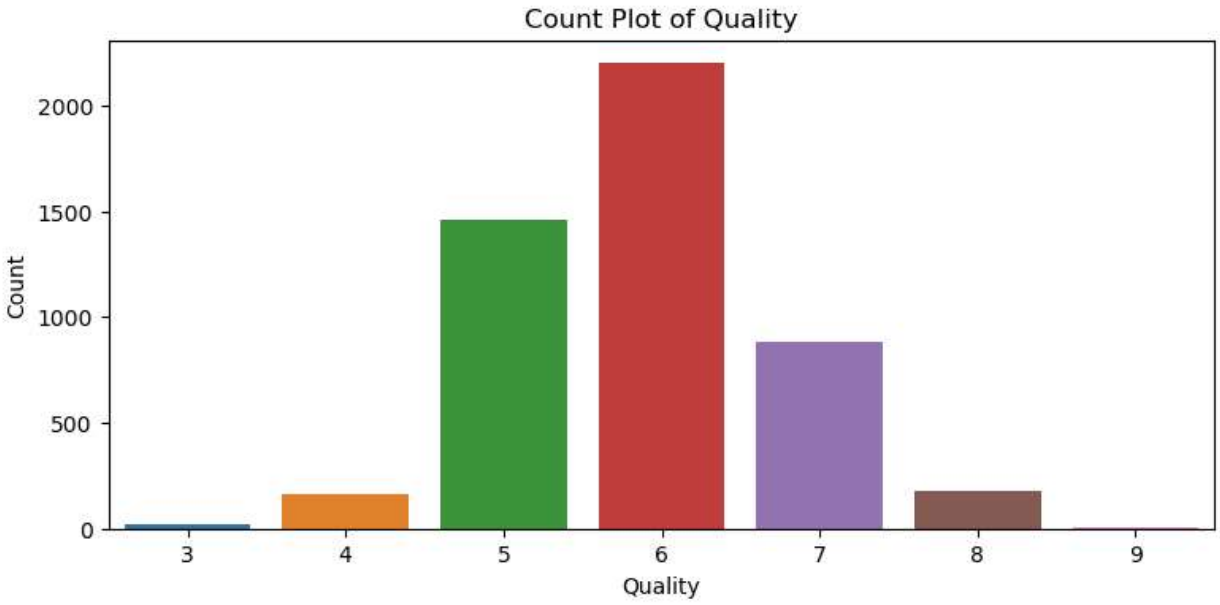
Out[57]:
```
Int64Index([6, 5, 7, 8, 4, 3, 9], dtype='int64')
```

In [58]:
```
# Create a count plot of the "Quality" feature

plt.figure(figsize = (9,4))

sns.barplot(data= df, x= quality.index, y= quality.quality)

plt.title("Count Plot of Quality")
plt.xlabel('Quality')
plt.ylabel('Count')
plt.show()
```



Observation-->

Based on the count plot of the "Quality" feature, it is evident that the highest count of quality ratings is 6, indicating that wines with a quality rating of 6 are the most common in the dataset. On the other hand, the count of wines with a quality rating of 9 is negligible, suggesting that wines with this rating are relatively rare in the dataset.

This observation highlights the distribution and frequency of quality ratings among the white wines. Understanding the distribution of quality ratings is crucial for assessing the overall quality profile of the wines and identifying any potential patterns or trends.

In [59]:
```
df['quality'].mode()
```

Out[59]:
```
0    6
Name: quality, dtype: int64
```

Observations from Task 2-->

During the analysis of the wine dataset, the following observations were made:

Distributions: Plots such as histograms and distribution plots were used to visualize the distributions of various features. It was observed that some features exhibited normal distributions, while others displayed slight skewness.

Central Tendency Measures: Measures of central tendency, including mean, median, and mode, were calculated for the different features. It was found that for most features, the mean and median values were similar, indicating a relatively symmetrical distribution. In such cases, the mean was considered a suitable representative measure.

Mode for Quality: In the case of the "Quality" feature, the mode was chosen as a representative value. The count plot revealed that wines with a quality rating of 6 had the highest frequency,

while a rating of 9 was significantly less common.

These observations provide insights into the distribution characteristics and central tendencies of the wine dataset. They help in understanding the typical values and variability of the features, which are crucial for further analysis and interpretation.

Task 3 - Create a new Pandas Series that contains the details of the acid types for a quality:

In this task, we will create a Pandas Series that contains the representative values for each of the acid types based on the quality of the wine. This will allow us to analyze the variations in acid types across different quality ratings.

Let's proceed with creating the Pandas Series.

In [61]:
```
# Create a new Pandas Series called "rep_acid" that contains the details of the repres

rep_acid = pd.DataFrame(index= ['fixed acidity','volatile acidity','citric acid','qual
                        data = [df['fixed acidity'].mean(),df['volatile acidity'].mean(),
                                df['citric acid'].mean(),df['quality'].value_counts().inc
```

In [62]:
```
# Print the "rep_acid" series

rep_acid =rep_acid.rename(columns= {0:'Mean'})
rep_acid
```

Out[62]:

	Mean
fixed acidity	6.854788
volatile acidity	0.278241
citric acid	0.334192
quality	6.000000

Observations from Task 3-->

After creating the Pandas Series to represent acid types based on quality, the following observations were made:

Fixed Acidity: The mean value of fixed acidity for the given quality is approximately 6.854.

Volatile Acidity: The mean value of volatile acidity for the given quality is approximately 0.2782.

Citric Acid: The mean value of citric acid for the given quality is approximately 0.3341.

Quality: The representative value for quality is 6.

These observations provide insights into the average values of different acid types based on the given quality rating. Such analysis allows us to understand the relationship between acid characteristics and wine quality, helping in further evaluation and decision-making.

Final Conclusions:

In this analysis of the wine dataset, several important findings and conclusions have been drawn:

Data Distribution: Visualizations such as histograms and distribution plots were utilized to gain insights into the distribution of the data. These visualizations provided a clear understanding of how the data points were spread across different features.

Measures of Central Tendency: Measures such as mean, median, and mode were employed to represent a group of observations or summarize the central values. The selection of a specific measure depended on the type and distribution of the data.

Choosing the Appropriate Measure: The choice of the appropriate measure of central tendency was based on the characteristics of the data. In cases where the data exhibited a symmetrical distribution, the mean and median were comparable and could be used interchangeably. The mode was particularly useful for representing the most frequently occurring value, as observed in the "Quality" feature.

These findings highlight the significance of data visualization and central tendency measures in understanding and summarizing the dataset. They provide a foundation for further analysis and decision-making processes in the domain of wine quality assessment.