wine-quality-test

February 22, 2025

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
df=pd.read_csv(r'C:\Users\kmuba\Downloads\Wine Quality Dataset.csv ')
```

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

[3]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

[4]:
```
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
```

[5]:
```
df.describe()
```

[5]:

	fixed acidity	volatile acidity	citric acid	residual sugar
count	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415
std	0.843868	0.100795	0.121020	5.072058
min	3.800000	0.080000	0.000000	0.600000
25%	6.300000	0.210000	0.270000	1.700000
50%	6.800000	0.260000	0.320000	5.200000
75%	7.300000	0.320000	0.390000	9.900000
max	14.200000	1.100000	1.660000	65.800000

	chlorides	free sulfur dioxide	total sulfur dioxide	density
count	4898.000000	4898.000000	4898.000000	4898.000000
mean	0.045772	35.308085	138.360657	0.994027
std	0.021848	17.007137	42.498065	0.002991
min	0.009000	2.000000	9.000000	0.987110
25%	0.036000	23.000000	108.000000	0.991723
50%	0.043000	34.000000	134.000000	0.993740
75%	0.050000	46.000000	167.000000	0.996100
max	0.346000	289.000000	440.000000	1.038980

	pH	sulphates	alcohol	quality
count	4898.000000	4898.000000	4898.000000	4898.000000
mean	3.188267	0.489847	10.514267	5.877909
std	0.151001	0.114126	1.230621	0.885639
min	2.720000	0.220000	8.000000	3.000000
25%	3.090000	0.410000	9.500000	5.000000
50%	3.180000	0.470000	10.400000	6.000000
75%	3.280000	0.550000	11.400000	6.000000
max	3.820000	1.080000	14.200000	9.000000

[6]:
```
df.isnull().sum()
```

[6]:
```
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
```

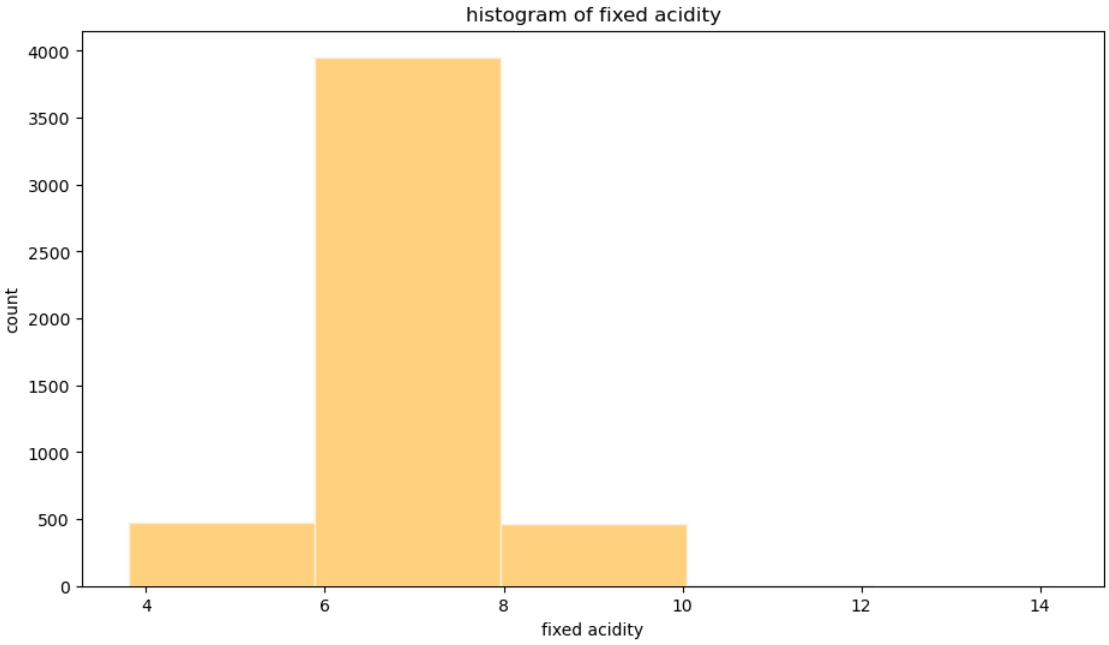
[7]:
```
df.shape
```

[7]:
```
(4898, 12)
```

[8]:
```
df.columns
```

[8]:
```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

[9]:
```
plt.figure(figsize=(11,6))
sns.histplot(data=df,x="fixed acidity",color="orange",edgecolor="linen",alpha=0.
↪5,bins=5)
plt.title("histogram of fixed acidity")
plt.xlabel('fixed acidity')
plt.ylabel("count")
plt.show()
```



[10]:
```
df["fixed acidity"].mean()
```

[10]: 6.854787668436097

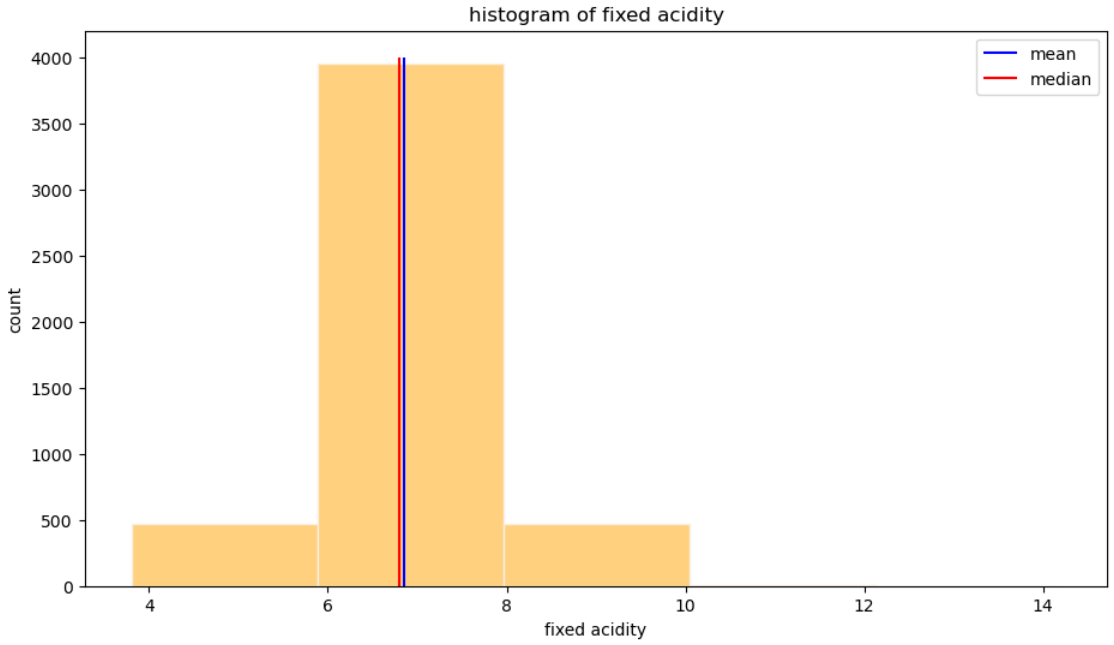
[11]:
```
round(df["fixed acidity"].mean(),2)
```

[11]: 6.85

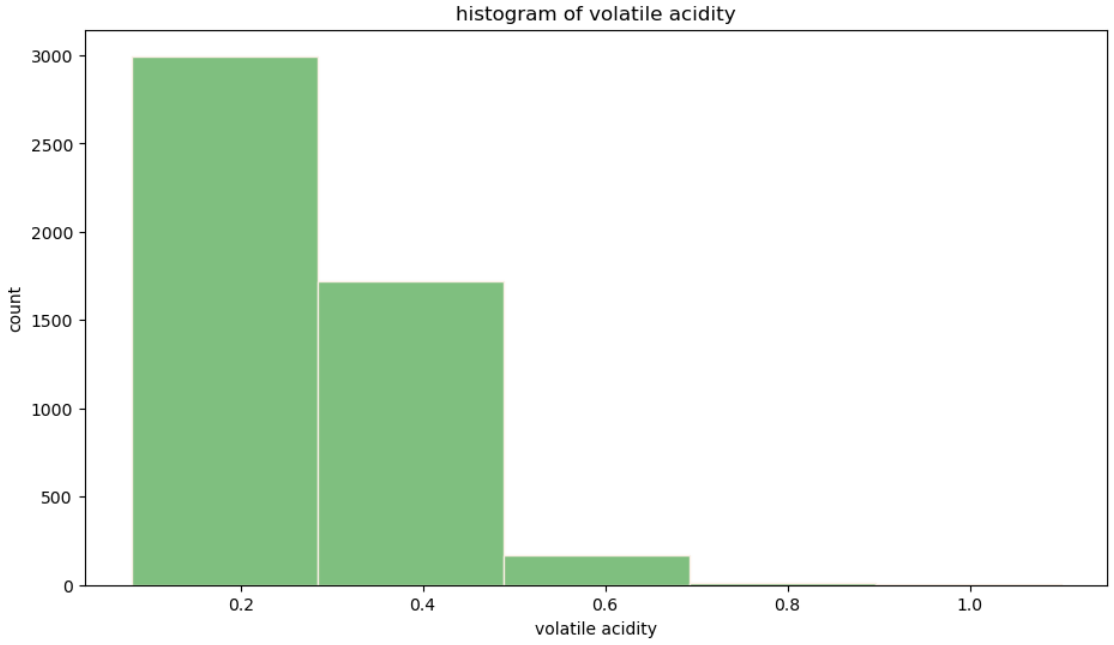
[12]:
```
df["fixed acidity"].median()
```

[12]: 6.8

[13]:
```
plt.figure(figsize=(11,6))
sns.histplot(data=df,x="fixed acidity",color="orange",edgecolor="linen",alpha=0.
  ↪5,bins=5)
plt.title("histogram of fixed acidity")
plt.xlabel('fixed acidity')
plt.ylabel("count")
plt.vlines(df["fixed acidity"].
  ↪mean(),ymin=0,ymax=4000,colors="blue",label="mean")
plt.vlines(df["fixed acidity"].
  ↪median(),ymin=0,ymax=4000,colors="red",label="median")
plt.legend()
plt.show()
```

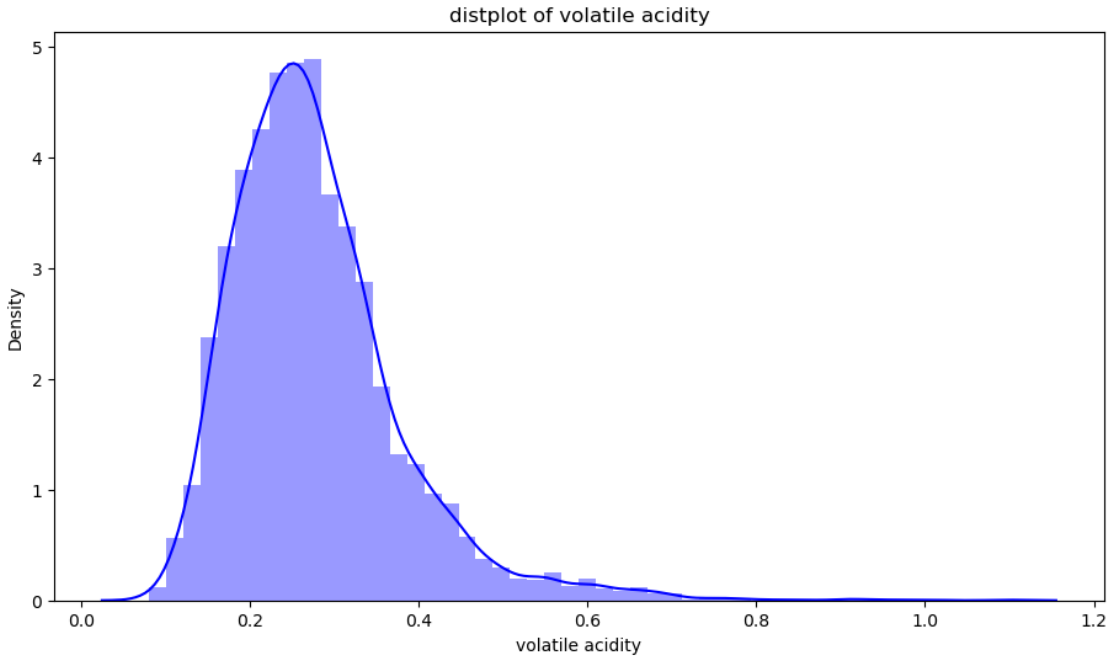


[14]:
```
plt.figure(figsize=(11,6))
sns.histplot(data=df,x="volatile␣
 ↪acidity",color="green",edgecolor="linen",alpha=0.5,bins=5)
plt.title("histogram of volatile acidity")
plt.xlabel('volatile acidity')
plt.ylabel("count")
plt.show()
```



[15]:
```
plt.figure(figsize=(11,6))
sns.distplot(df["volatile acidity"],color="blue")
plt.title("distplot of volatile acidity")
plt.xlabel('volatile acidity')
plt.ylabel("Density")
plt.show()
```

```
C:\Users\kmuba\AppData\Local\Temp\ipykernel_40544\1865294773.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df["volatile acidity"],color="blue")
```



[16]:
```
df['volatile acidity'].mean()
```

[16]:
```
0.27824111882400976
```

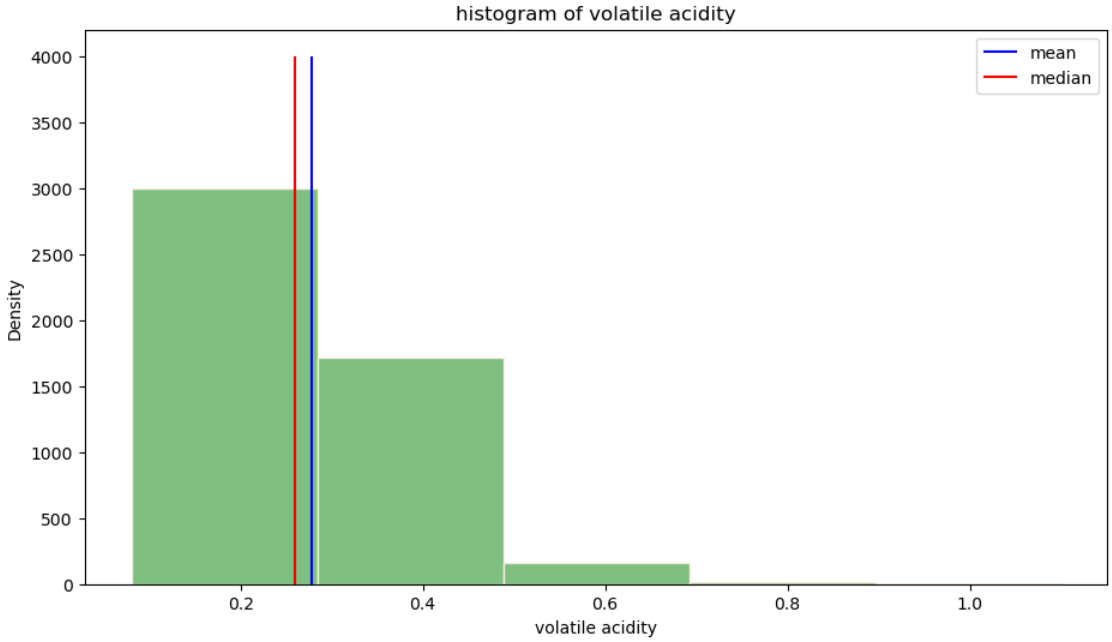
[21]:
```
df['volatile acidity'].skew()
```

[21]:
```
1.5769795029952025
```

[26]:
```
df['volatile acidity'].median()
```

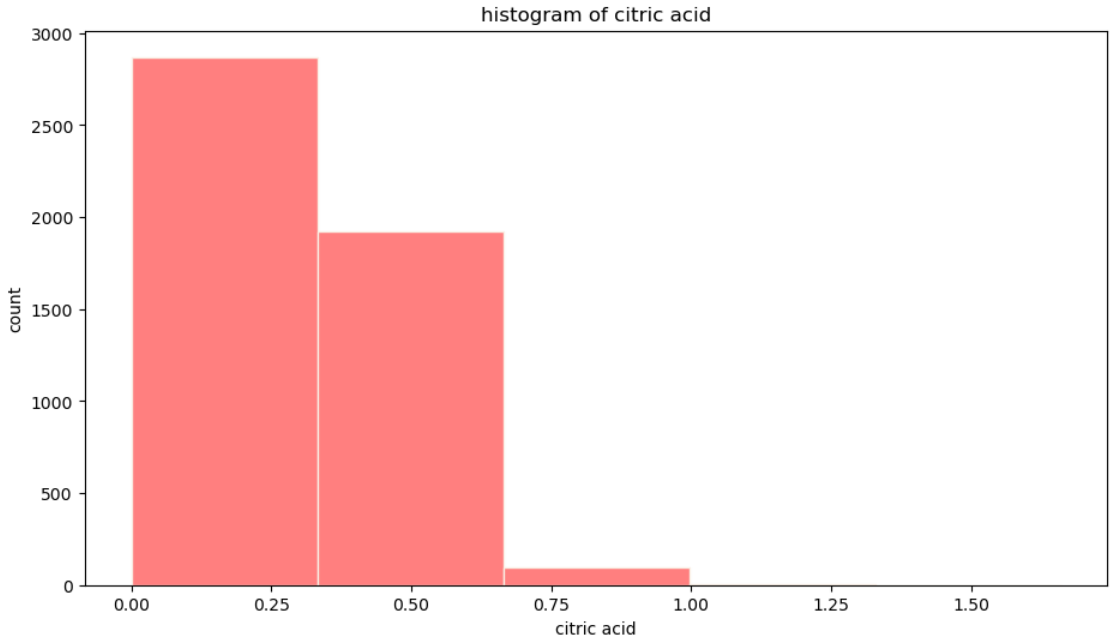
[26]:
```
0.26
```

[27]:
```
plt.figure(figsize=(11,6))
sns.histplot(data=df,x="volatile␣
 ↪acidity",color="green",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="blue",label="mean")
plt.vlines(df["volatile acidity"].
 ↪median(),ymin=0,ymax=4000,colors="red",label="median")
plt.legend()
plt.show()
```



[32]:
```
plt.figure(figsize=(11,6))
sns.histplot(data=df,x="citric acid",color="red",edgecolor="linen",alpha=0.
 ↪5,bins=5)
plt.title("histogram of citric acid")
plt.xlabel('citric acid')
```

```
plt.ylabel("count")
plt.show()
```



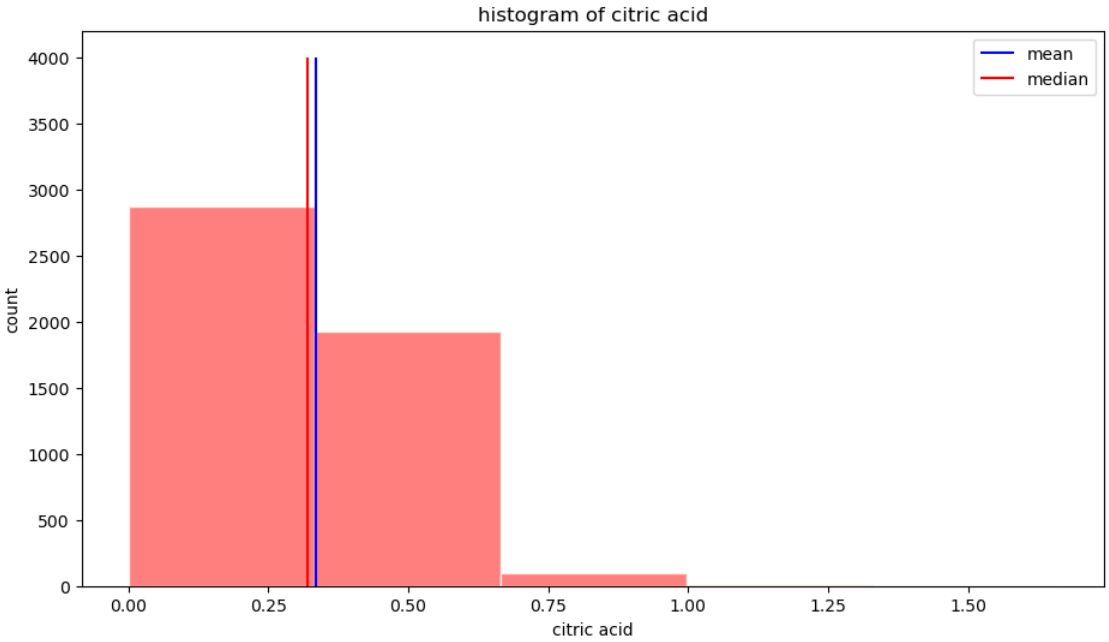
[44]:
```
df['citric acid'].mean()
```

[44]: 0.33419150673744386

[46]:
```
df['citric acid'].median()
```

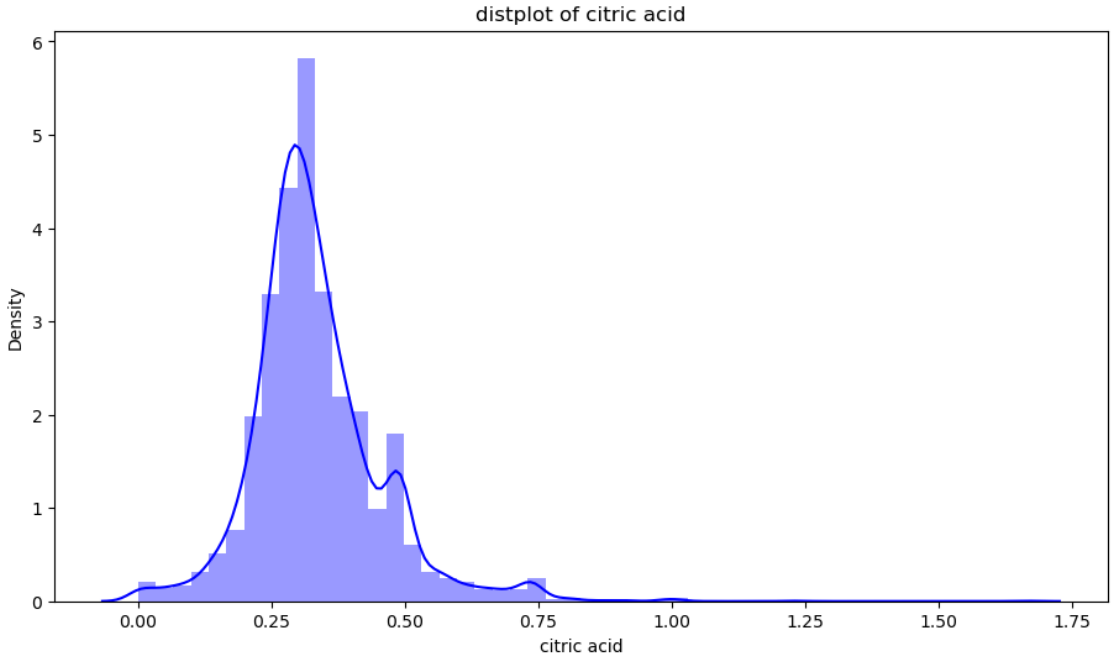
[46]: 0.32

[48]:
```
plt.figure(figsize=(11,6))
sns.histplot(data=df,x="citric acid",color="red",edgecolor="linen",alpha=0.
  ↪5,bins=5)
plt.title("histogram of citric acid")
plt.xlabel('citric acid')
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="red",label="median")
plt.legend()
plt.show()
```

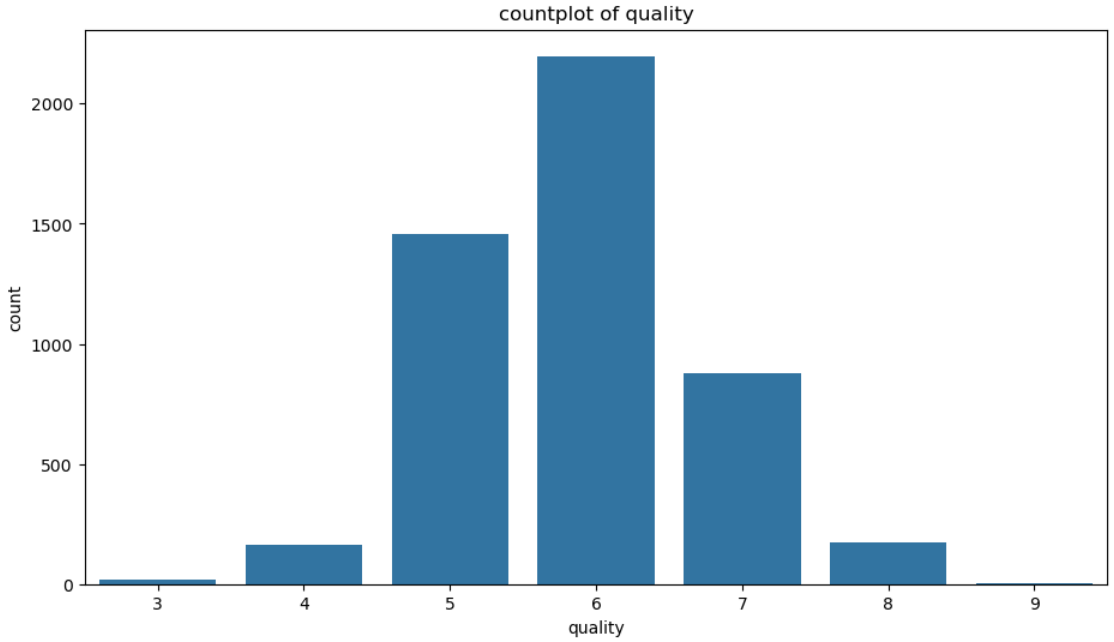


[50]:
```
plt.figure(figsize=(11,6))
sns.distplot(df["citric acid"],color="blue")
plt.title("distplot of citric acid")
plt.xlabel('citric acid')
plt.ylabel("Density")
plt.show()
```

```
C:\Users\kmuba\AppData\Local\Temp\ipykernel_40544\2817031774.py:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

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



[9]:
```
plt.figure(figsize=(11,6))
sns.countplot(x=df["quality"])
plt.title("countplot of quality")
plt.xlabel('quality')
plt.ylabel("count")
plt.show()
```



[56]:
```
df['quality'].value_counts()
```

[56]:
```
quality
6    2198
5    1457
7     880
8     175
4     163
3      20
9       5
Name: count, dtype: int64
```

[60]:
```
df['quality'].value_counts().index[0]
```

[60]:
```
6
```

[74]:
```
df.columns
```

[74]:
```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

[91]:
```
rep_acid=pd.Series(index=["fixed acidicity",'volatile acidity','citric␣
↪acide','qality'],
                   data=[df['fixed acidity'].mean(),df['volatile acidity'].
↪mean(),
                    df['citric acid'].mean(),df['quality'].value_counts().
↪index[0]])
```

[95]:
```
rep_acid
```

[95]:
```
fixed acidicity     6.854788
volatile acidity    0.278241
citric acide        0.334192
qality              6.000000
dtype: float64
```

[ ]:
```

```