Q1

August 11, 2025

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

0.0.1 1. Load & Inspect

Load CSV file into a DataFrame.

Show first 10 rows.

Display shape, column names, and data types.

[2]:
```
df= pd.read_csv('/Users/sivakumar/projects/ml_project/DotKonnet/Assignment2/
 ↪wine+quality/winequality-white.csv',sep=';')

#Downloaded Wine Dataset from here https://archive.ics.uci.edu/dataset/186/
 ↪wine+quality
```

[3]:
```
df.head(10)
```

[3]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40
5	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44
6	6.2	0.32	0.16	7.0	0.045					
7	7.0	0.27	0.36	20.7	0.045					
8	6.3	0.30	0.34	1.6	0.049					
9	8.1	0.22	0.43	1.5	0.044					

```
6                30.0                 136.0   0.9949  3.18       0.47
7                45.0                 170.0   1.0010  3.00       0.45
8                14.0                 132.0   0.9940  3.30       0.49
9                28.0                 129.0   0.9938  3.22       0.45

   alcohol  quality
0      8.8        6
1      9.5        6
2     10.1        6
3      9.9        6
4      9.9        6
5     10.1        6
6      9.6        6
7      8.8        6
8      9.5        6
9     11.0        6
```

[4]:
```
#df.shape

print(f'No.Of Rows:{df.shape[0]}')
print(f'No.of Columns:{df.shape[1]}')
```

```
No.Of Rows:4898
No.of Columns:12
```

[5]:
```
df.columns
```

[5]:
```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

[6]:
```
df.dtypes
```

[6]:
```
fixed acidity           float64
volatile acidity        float64
citric acid             float64
residual sugar          float64
chlorides               float64
free sulfur dioxide     float64
total sulfur dioxide    float64
density                 float64
pH                      float64
sulphates               float64
alcohol                 float64
quality                   int64
dtype: object
```

0.0.2 2. Summary Statistics

.describe() for numeric columns.

Count missing values per column.

Fill missing numeric values with the mean.

[7]:
```
df.describe().round(3)
```

[7]:
```
       fixed acidity  volatile acidity  citric acid  residual sugar  \
count       4898.000          4898.000     4898.000        4898.000
mean           6.855             0.278        0.334           6.391
std            0.844             0.101        0.121           5.072
min            3.800             0.080        0.000           0.600
25%            6.300             0.210        0.270           1.700
50%            6.800             0.260        0.320           5.200
75%            7.300             0.320        0.390           9.900
max           14.200             1.100        1.660          65.800

       chlorides  free sulfur dioxide  total sulfur dioxide   density  \
count   4898.000              4898.000              4898.000  4898.000
mean       0.046                35.308               138.361     0.994
std        0.022                17.007                42.498     0.003
min        0.009                 2.000                 9.000     0.987
25%        0.036                23.000               108.000     0.992
50%        0.043                34.000               134.000     0.994
75%        0.050                46.000               167.000     0.996
max        0.346               289.000               440.000     1.039

             pH  sulphates   alcohol   quality
count  4898.000   4898.000  4898.000  4898.000
mean      3.188      0.490    10.514     5.878
std       0.151      0.114     1.231     0.886
min       2.720      0.220     8.000     3.000
25%       3.090      0.410     9.500     5.000
50%       3.180      0.470    10.400     6.000
75%       3.280      0.550    11.400     6.000
max       3.820      1.080    14.200     9.000
```

[8]:
```
for col, missing in df.isnull().sum().items():
    print(f'No of Missing values in {col}: {missing}')
```

```
No of Missing values in fixed acidity: 0
No of Missing values in volatile acidity: 0
No of Missing values in citric acid: 0
No of Missing values in residual sugar: 0
No of Missing values in chlorides: 0
No of Missing values in free sulfur dioxide: 0
No of Missing values in total sulfur dioxide: 0
```

```
No of Missing values in density: 0
No of Missing values in pH: 0
No of Missing values in sulphates: 0
No of Missing values in alcohol: 0
No of Missing values in quality: 0
```

[9]:
```
df.fillna(df.mean(numeric_only=True), inplace= True)
```

0.0.3 3. Filter & Sort

Filter rows by a numeric condition (e.g., Age > 30).

Sort dataset by a column in descending order.

[10]:
```
filter_df= df[df['quality']>6]
filter_df
```

[10]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
13	6.6	0.16	0.40	1.5	0.044	48.0	143.0	0.99120	3.54	0.52	12.40	7
15	6.6	0.17	0.38	1.5	0.032	28.0	112.0	0.99140	3.25	0.55	11.40	7
17	6.2	0.66	0.48	1.2	0.029	29.0	75.0	0.98920	3.33	0.39	12.80	8
20	6.2	0.66	0.48	1.2	0.029	29.0	75.0	0.98920	3.33	0.39	12.80	8
21	6.4	0.31	0.38	2.9	0.038	19.0	102.0	0.99120	3.17	0.35	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…
4870	6.1	0.32	0.28	6.6	0.021	29.0	132.0	0.99188	3.15	0.36	…	…
4876	6.2	0.38	0.42	2.5	0.038	34.0	117.0	0.99132	3.36	0.59	…	…
4886	6.2	0.21	0.28	5.7	0.028	45.0	121.0	0.99168	3.21	1.08	…	…
4887	6.2	0.41	0.22	1.9	0.023	5.0	56.0	0.98928	3.04	0.79	…	…
4896	5.5	0.29	0.30	1.1	0.022	20.0	110.0	0.98869	3.34	0.38	…	…

```
21      11.00        7
…       …       …
4870    11.45        7
4876    11.60        7
4886    12.15        7
4887    13.00        7
4896    12.80        7

[1060 rows x 12 columns]
```

[11]:
```
df.sort_values(by= ['pH', 'quality'], ascending=[True, False])
```

[11]:
```
      fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
1900           10.0             0.230         0.27           14.10      0.033
1214            9.7             0.240         0.45            1.20      0.033
2162            9.9             0.490         0.23            2.40      0.087
1959            8.5             0.170         0.31            1.00      0.024
1960            8.5             0.170         0.31            1.00      0.024
…               …                 …            …               …         …
2321            4.6             0.445         0.00            1.40      0.053
2036            5.7             0.270         0.32            1.20      0.046
2771            6.3             0.200         0.24            1.70      0.052
1255            6.4             0.220         0.34            1.80      0.057
1250            5.3             0.260         0.23            5.15      0.034

      free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
1900                 45.0                 166.0  0.99880  2.72       0.43
1214                 11.0                  59.0  0.99260  2.74       0.47
2162                 19.0                 115.0  0.99480  2.77       0.44
1959                 13.0                  91.0  0.99300  2.79       0.37
1960                 13.0                  91.0  0.99300  2.79       0.37
…                     …                     …       …   …         …
2321                 11.0                 178.0  0.99426  3.79       0.55
2036                 20.0                 155.0  0.99340  3.80       0.41
2771                 36.0                 135.0  0.99374  3.80       0.66
1255                 29.0                 104.0  0.99590  3.81       0.57
1250                 48.0                 160.0  0.99520  3.82       0.51

      alcohol  quality
1900      9.7        6
1214     10.8        6
2162      9.4        6
1959     10.1        5
1960     10.1        5
…         …       …
2321     10.2        5
2036     10.2        6
```

```
2771     10.8        6
1255     10.3        6
1250     10.5        7

[4898 rows x 12 columns]
```

0.0.4 4. Group & Aggregate

Group by a categorical column, calculate mean of a numeric column.

[12]:
```
alcohol_avg= df.groupby('quality')['alcohol'].mean()
alcohol_avg
```

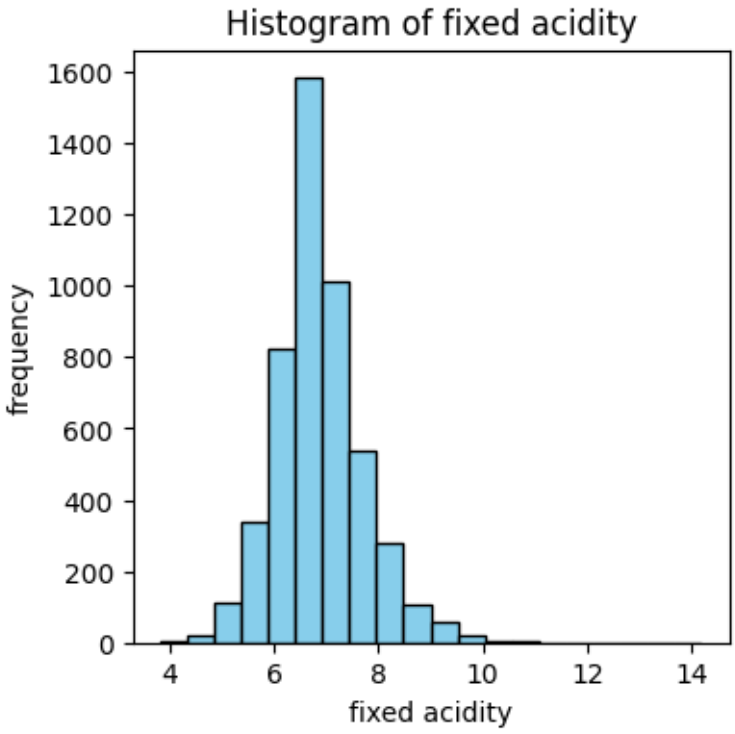
[12]:
```
quality
3    10.345000
4    10.152454
5     9.808840
6    10.575372
7    11.367936
8    11.636000
9    12.180000
Name: alcohol, dtype: float64
```

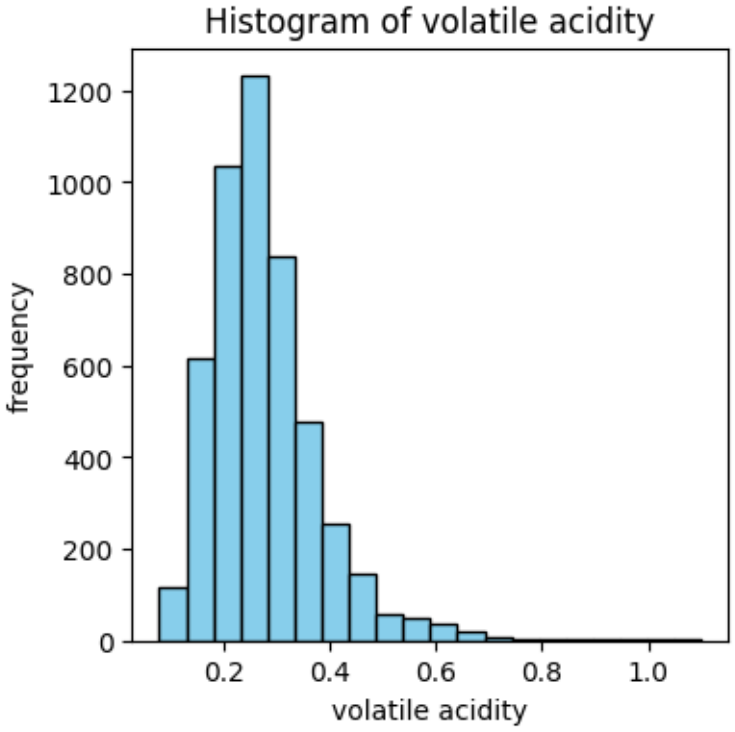
0.0.5 5. Visualize

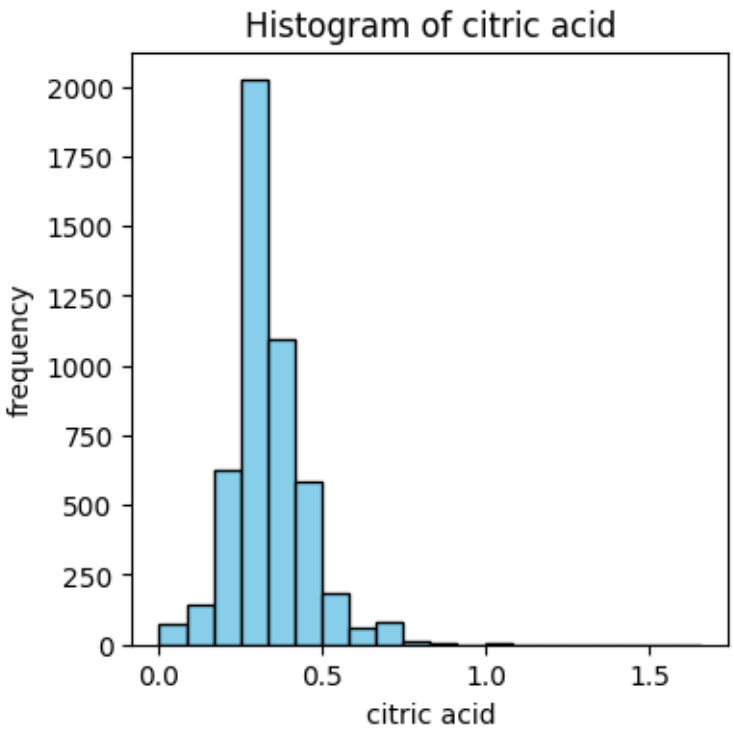
Create a histogram for a numeric column.

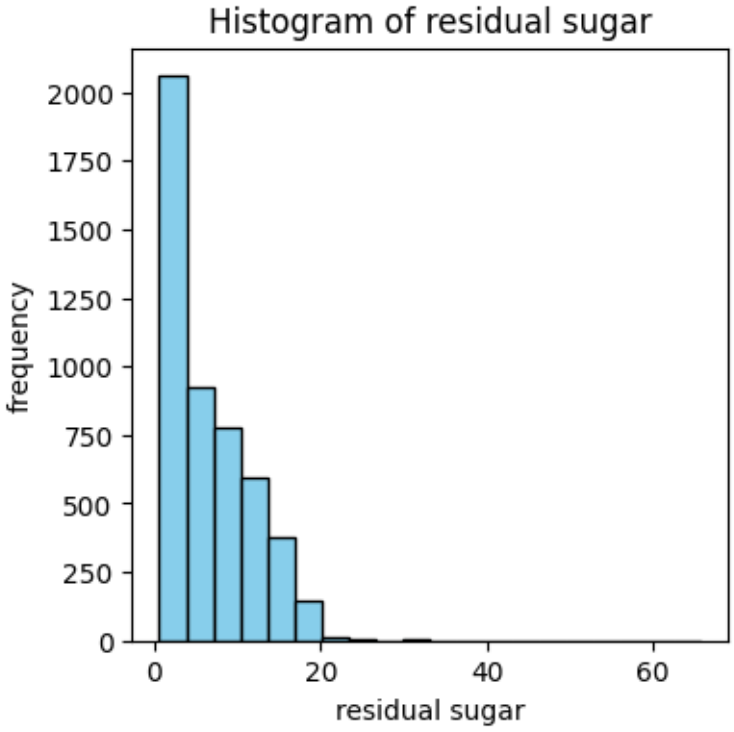
Create a bar chart of group averages.

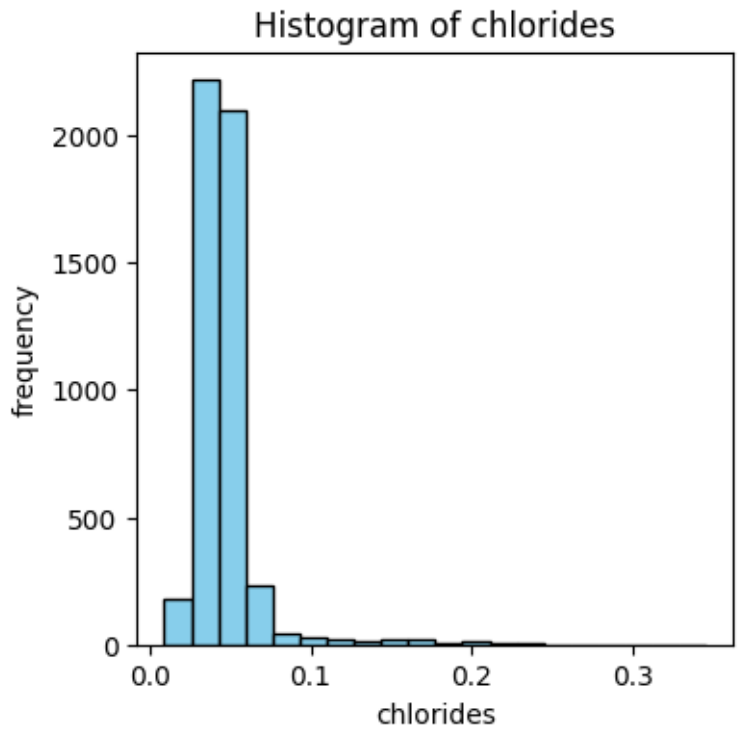
[13]:
```
for col in df.columns:
    plt.figure(figsize=(4,4))
    plt.hist(df[col], bins=20, color='skyblue', edgecolor='black')
    plt.title(f'Histogram of {col}')
    plt.xlabel(col)
    plt.ylabel('frequency')
    plt.show()
```

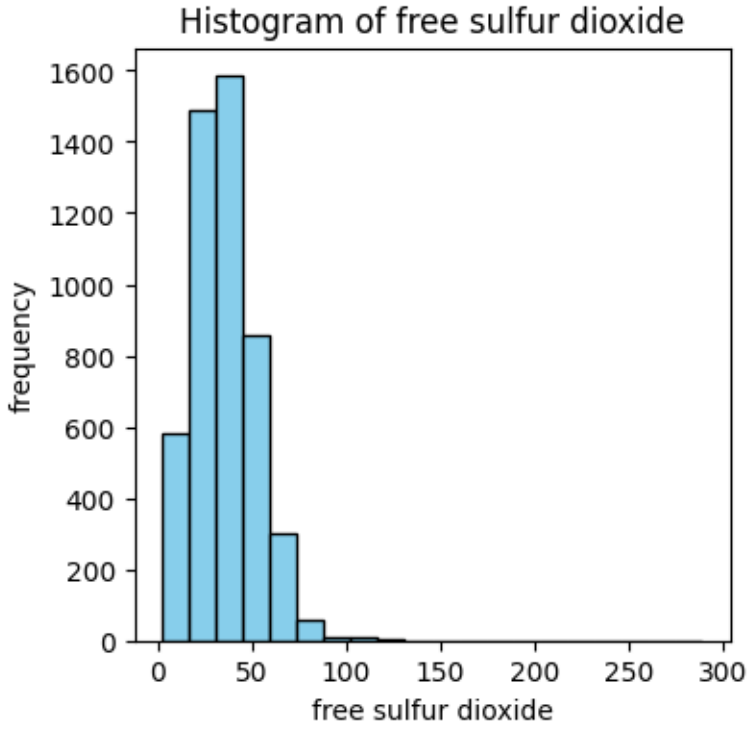

Histogram of fixed acidity
frequency
fixed acidity
0
200
400
600
800
1000
1200
1400
1600
4
6
8
10
12
14

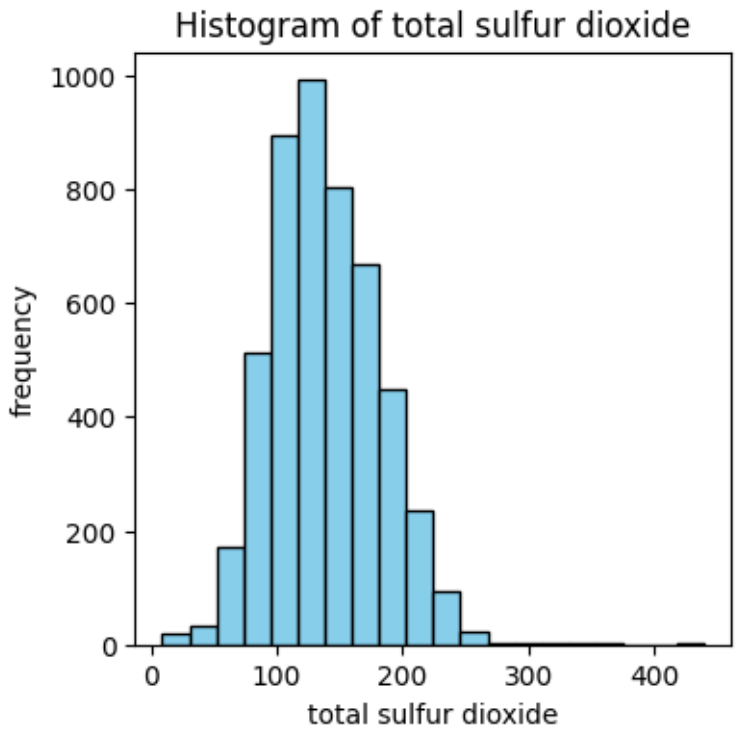

Histogram of volatile acidity
frequency
volatile acidity
0
200
400
600
800
1000
1200
0.2
0.4
0.6
0.8
1.0

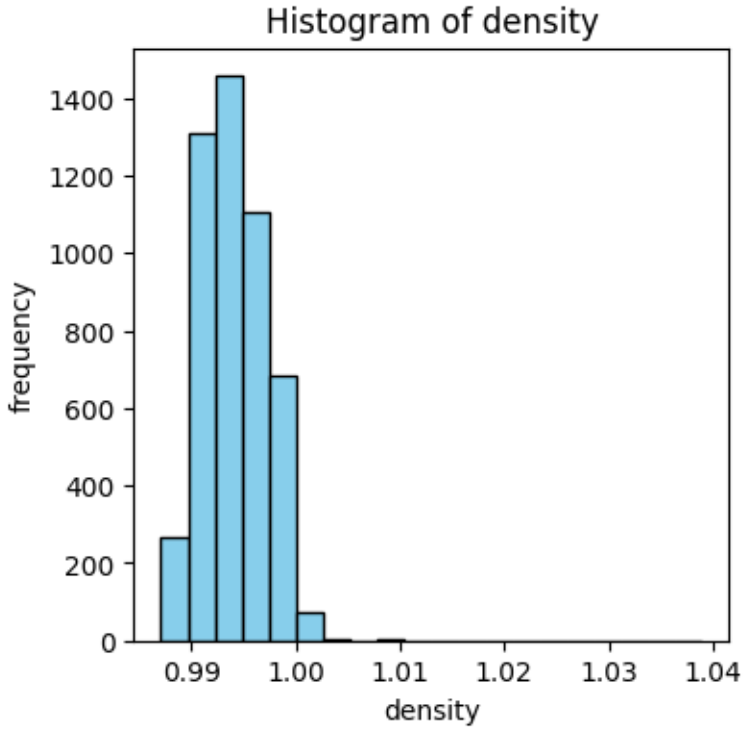

Histogram of citric acid
frequency
2000
1750
1500
1250
1000
750
500
250
0
0.0
0.5
1.0
1.5
citric acid

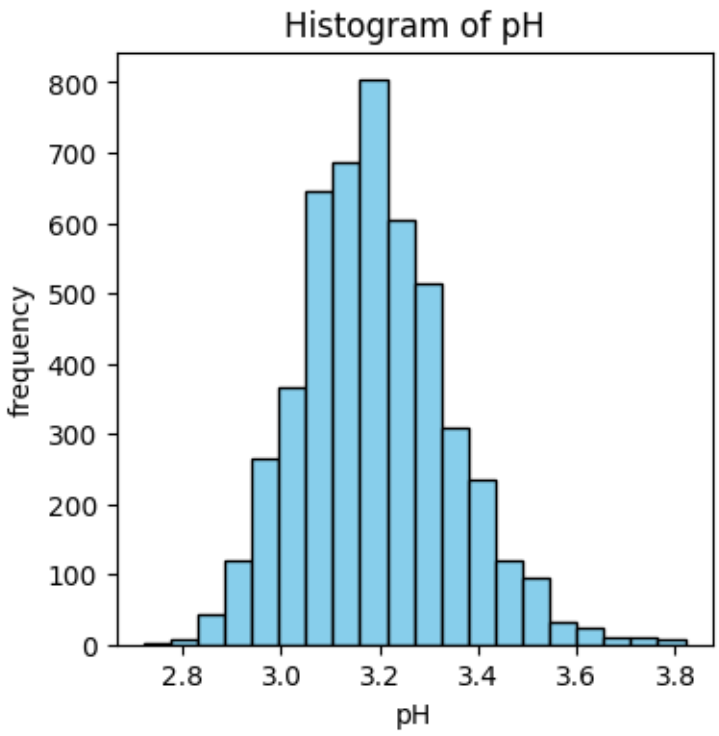

Histogram of residual sugar
frequency
2000
1750
1500
1250
1000
750
500
250
0
0
20
40
60
residual sugar

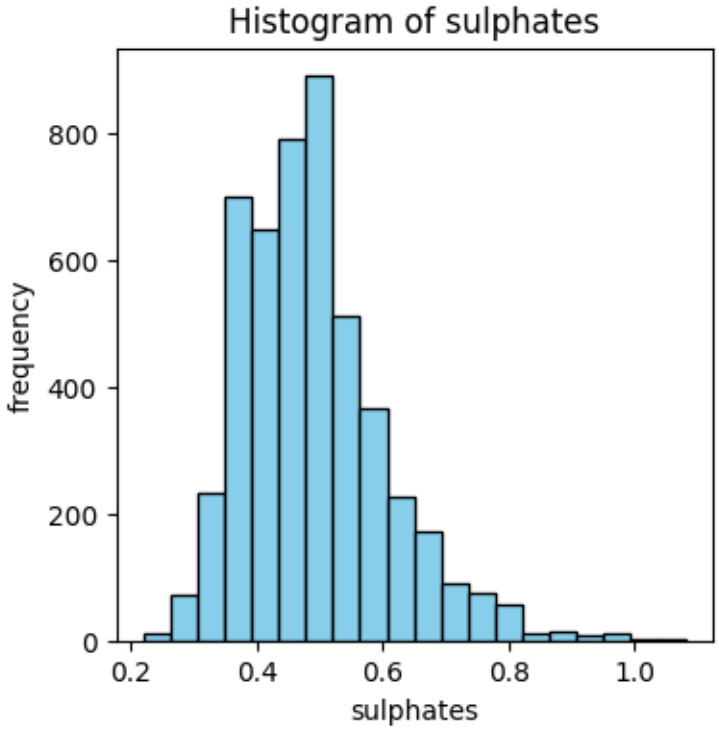

Histogram of chlorides
frequency
chlorides
0
500
1000
1500
2000
0.0
0.1
0.2
0.3

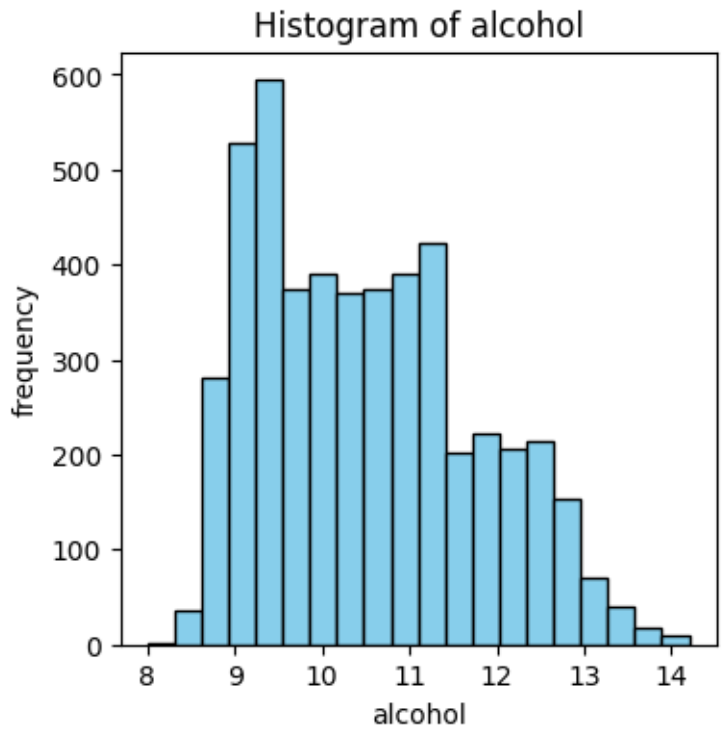

Histogram of free sulfur dioxide
frequency
free sulfur dioxide
0
200
400
600
800
1000
1200
1400
1600
0
50
100
150
200
250
300

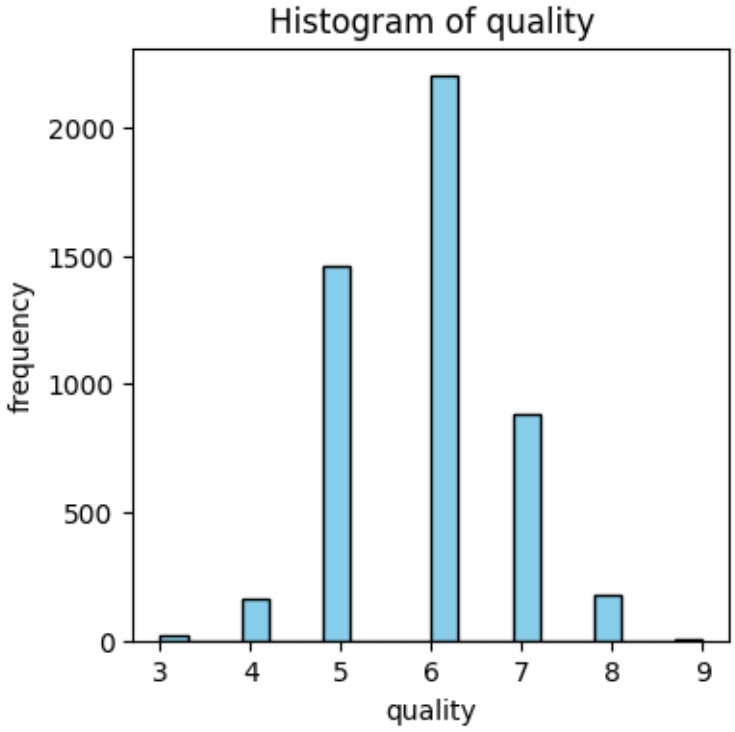

Histogram of total sulfur dioxide
frequency
total sulfur dioxide
0
200
400
600
800
1000
0
100
200
300
400


Histogram of density
frequency
density
0
200
400
600
800
1000
1200
1400
0.99
1.00
1.01
1.02
1.03
1.04


Histogram of pH
frequency
pH
0
100
200
300
400
500
600
700
800
2.8
3.0
3.2
3.4
3.6
3.8


Histogram of sulphates
frequency
sulphates
0
200
400
600
800
0.2
0.4
0.6
0.8
1.0


Histogram of alcohol
frequency
alcohol
0
100
200
300
400
500
600
8
9
10
11
12
13
14


Histogram of quality
frequency
quality
0
500
1000
1500
2000
3
4
5
6
7
8
9

0.0.6 Extra Challenge: Save the cleaned dataset as processed_data.csv.

[14]:
```
df.to_csv('/Users/sivakumar/projects/ml_project/DotKonnet/Assignment2/
↪wine+quality/processed_data.csv')
```

[ ]:

[ ]:

[ ]: