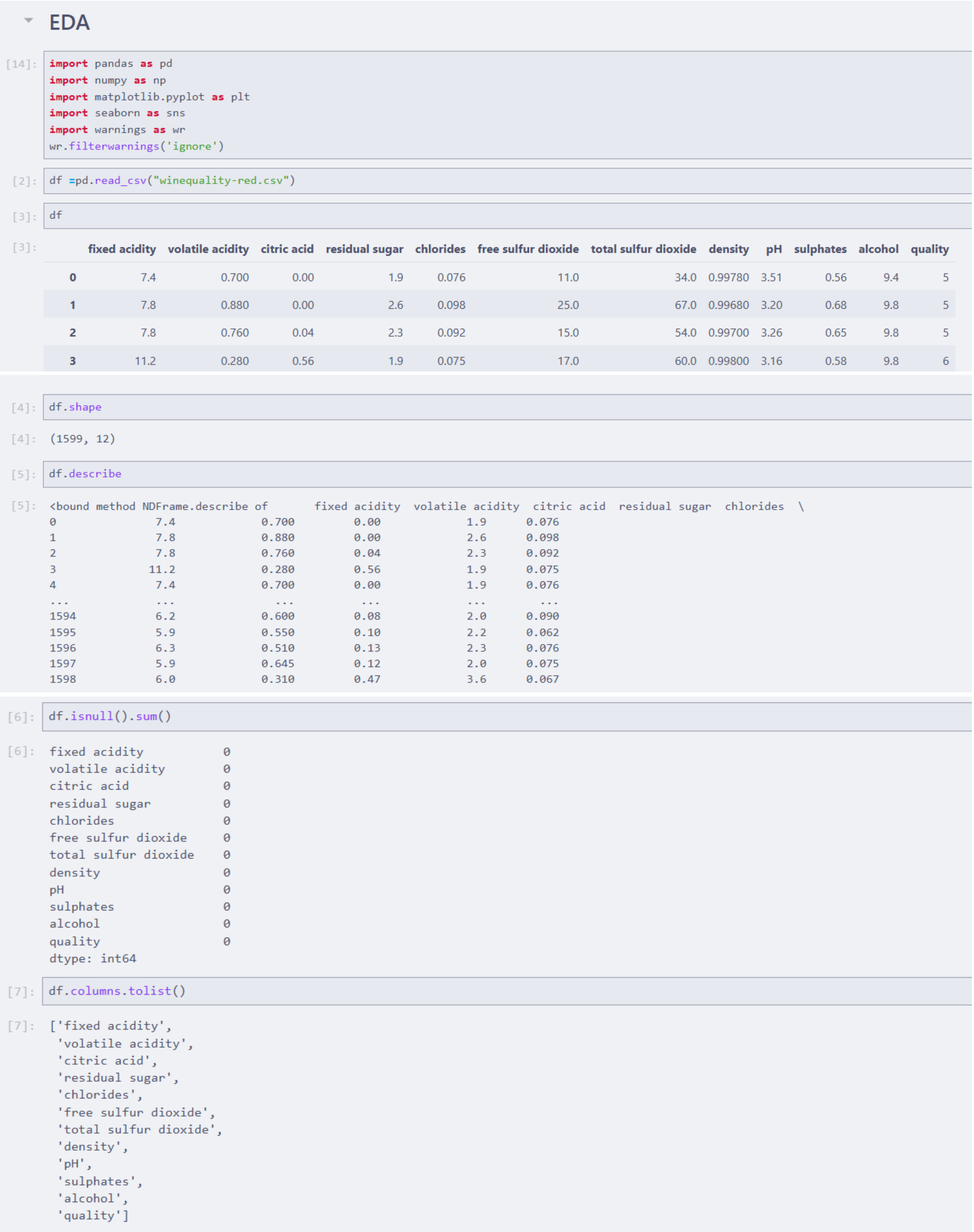
Practical 4: EDA



EDA

```
[14]: import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt
      import seaborn as sns
      import warnings as wr
      wr.filterwarnings('ignore')
```

```
[2]: df =pd.read_csv("winequality-red.csv")
```

```
[3]: df
```

[3]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6

```
[4]: df.shape
```

```
[4]: (1599, 12)
```

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

```
[5]: <bound method NDFrame.describe of       fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
     0               7.4             0.700         0.00             1.9      0.076
     1               7.8             0.880         0.00             2.6      0.098
     2               7.8             0.760         0.04             2.3      0.092
     3              11.2             0.280         0.56             1.9      0.075
     4               7.4             0.700         0.00             1.9      0.076
     ...             ...               ...          ...             ...        ...
     1594            6.2             0.600         0.08             2.0      0.090
     1595            5.9             0.550         0.10             2.2      0.062
     1596            6.3             0.510         0.13             2.3      0.076
     1597            5.9             0.645         0.12             2.0      0.075
     1598            6.0             0.310         0.47             3.6      0.067
```

```
[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
```

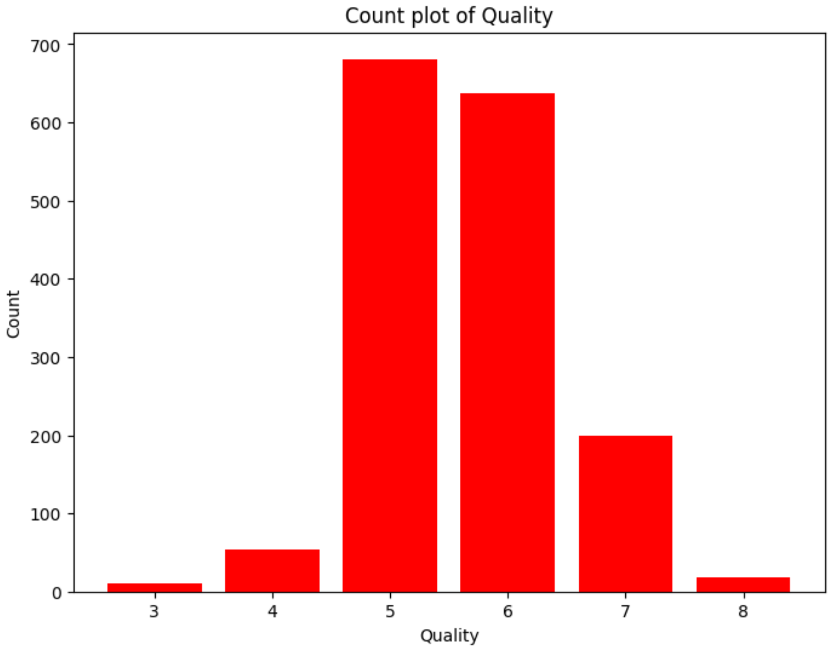
```
[7]: df.columns.tolist()
```

```
[7]: ['fixed acidity',
      'volatile acidity',
      'citric acid',
      'residual sugar',
      'chlorides',
      'free sulfur dioxide',
      'total sulfur dioxide',
      'density',
      'pH',
      'sulphates',
      'alcohol',
      'quality']
```

```
[8]: df.nunique()
```

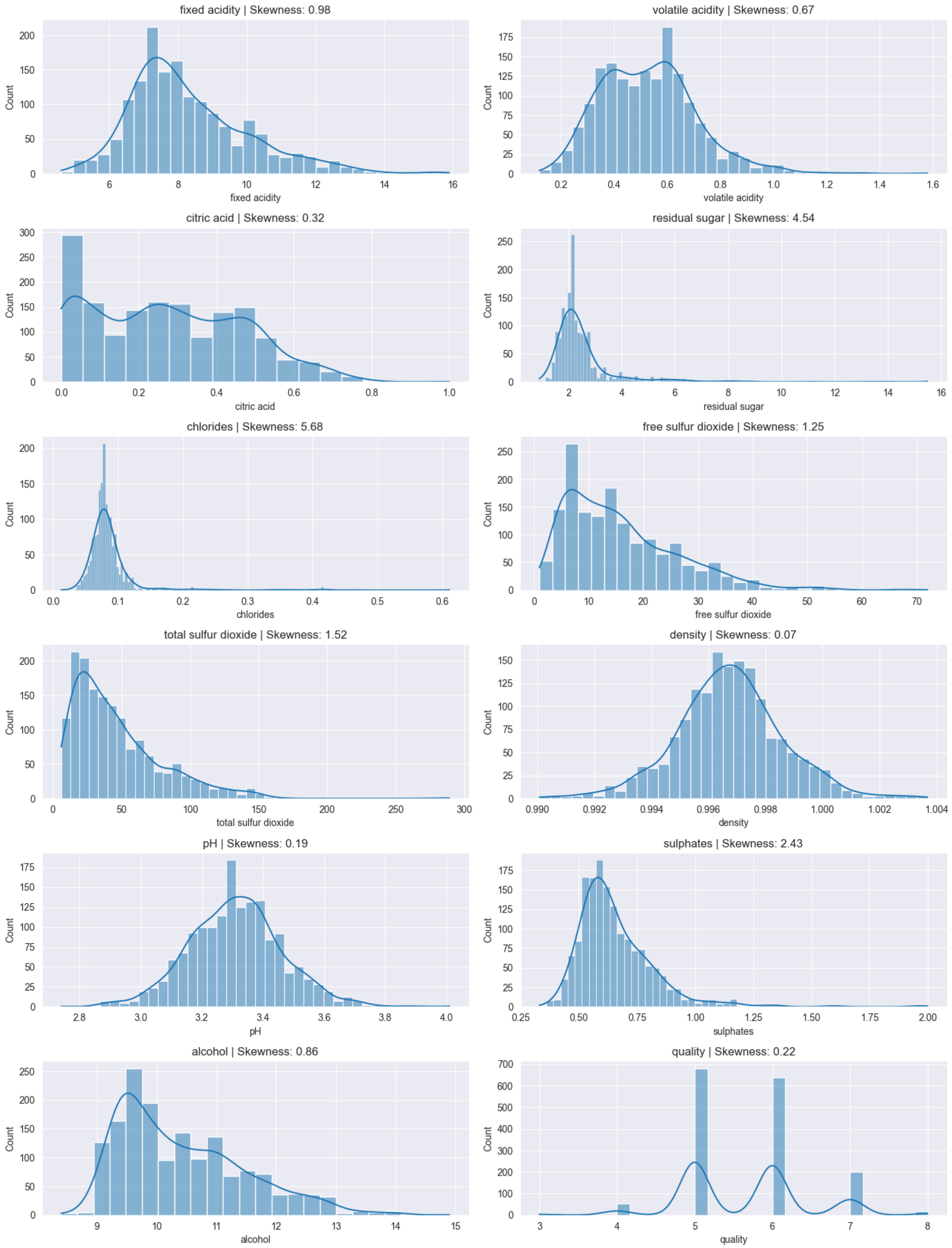
```
[8]: fixed acidity            96
     volatile acidity        143
     citric acid              80
     residual sugar           91
     chlorides               153
     free sulfur dioxide      60
     total sulfur dioxide    144
     density                 436
     pH                       89
     sulphates                96
     alcohol                  65
     quality                   6
     dtype: int64
```

```
[9]: qc =df['quality'].value_counts()

     plt.figure(figsize=(8,6))
     plt.bar(qc.index,qc,color="red")
     plt.title("Count plot of Quality")
     plt.xlabel("Quality")
     plt.ylabel("Count")
     plt.show()
```



Kernel Density Plots

```
[11]: sns.set_style("darkgrid")
      numerical_columns = df.select_dtypes(include=["int64", "float64"]).columns

      plt.figure(figsize=(14, len(numerical_columns) * 3))
      for idx, feature in enumerate(numerical_columns, 1):
       plt.subplot(len(numerical_columns), 2, idx)
       sns.histplot(df[feature], kde=True)
       plt.title(f"{feature} | Skewness: {round(df[feature].skew(), 2)}")

      plt.tight_layout()
      plt.show()
```

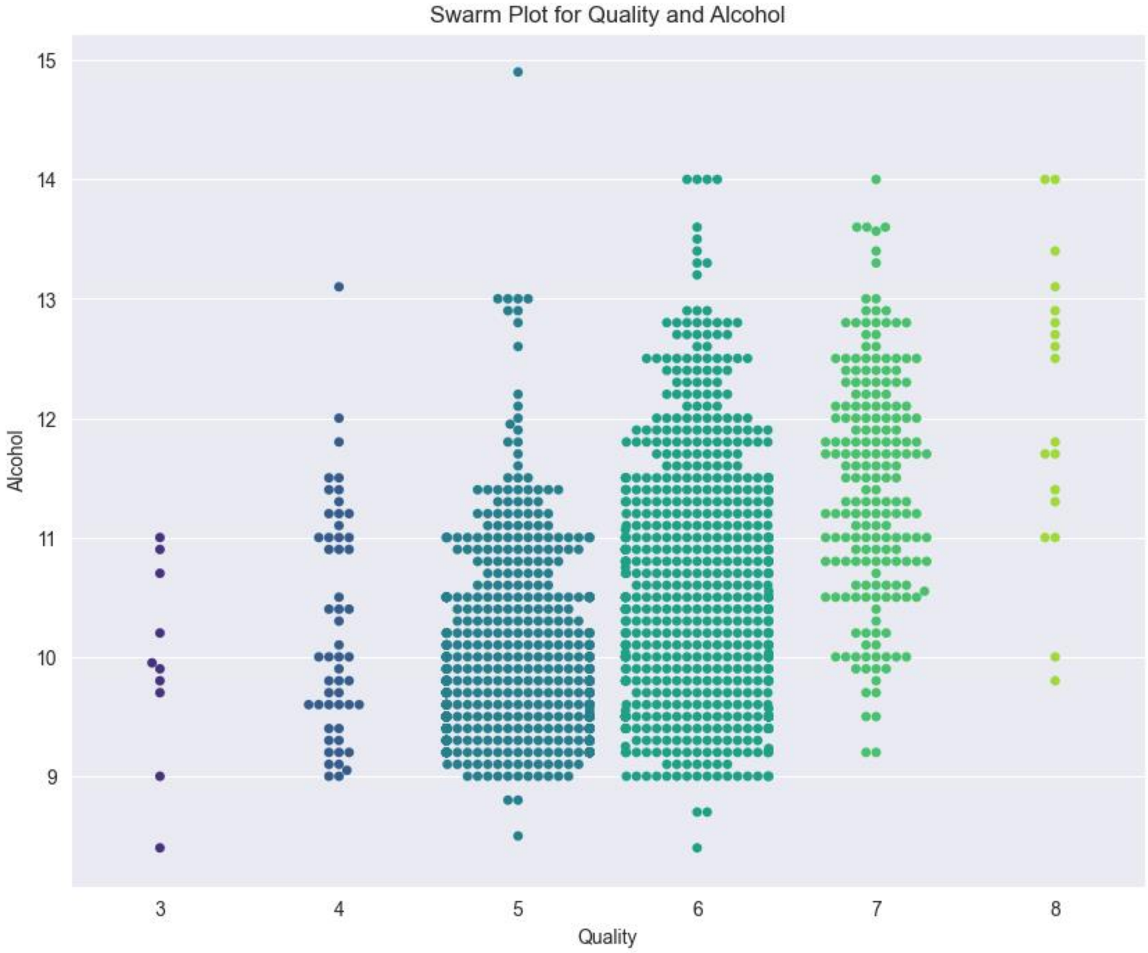


fixed acidity | Skewness: 0.98
Count
fixed acidity
0
50
100
150
200
6
8
10
12
14
16
volatile acidity | Skewness: 0.67
Count
volatile acidity
0
25
50
75
100
125
150
175
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
citric acid | Skewness: 0.32
Count
citric acid
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
residual sugar | Skewness: 4.54
Count
residual sugar
0
50
100
150
200
250
2
4
6
8
10
12
14
16
chlorides | Skewness: 5.68
Count
chlorides
0
50
100
150
200
0.0
0.1
0.2
0.3
0.4
0.5
0.6
free sulfur dioxide | Skewness: 1.25
Count
free sulfur dioxide
0
50
100
150
200
250
0
10
20
30
40
50
60
70
total sulfur dioxide | Skewness: 1.52
Count
total sulfur dioxide
0
50
100
150
200
0
50
100
150
200
250
300
density | Skewness: 0.07
Count
density
0
25
50
75
100
125
150
0.990
0.992
0.994
0.996
0.998
1.000
1.002
1.004
pH | Skewness: 0.19
Count
pH
0
25
50
75
100
125
150
175
2.8
3.0
3.2
3.4
3.6
3.8
4.0
sulphates | Skewness: 2.43
Count
sulphates
0
25
50
75
100
125
150
175
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
alcohol | Skewness: 0.86
Count
alcohol
0
50
100
150
200
250
9
10
11
12
13
14
15
quality | Skewness: 0.22
Count
quality
0
100
200
300
400
500
600
700
3
4
5
6
7
8

Swarm plot

[15]:

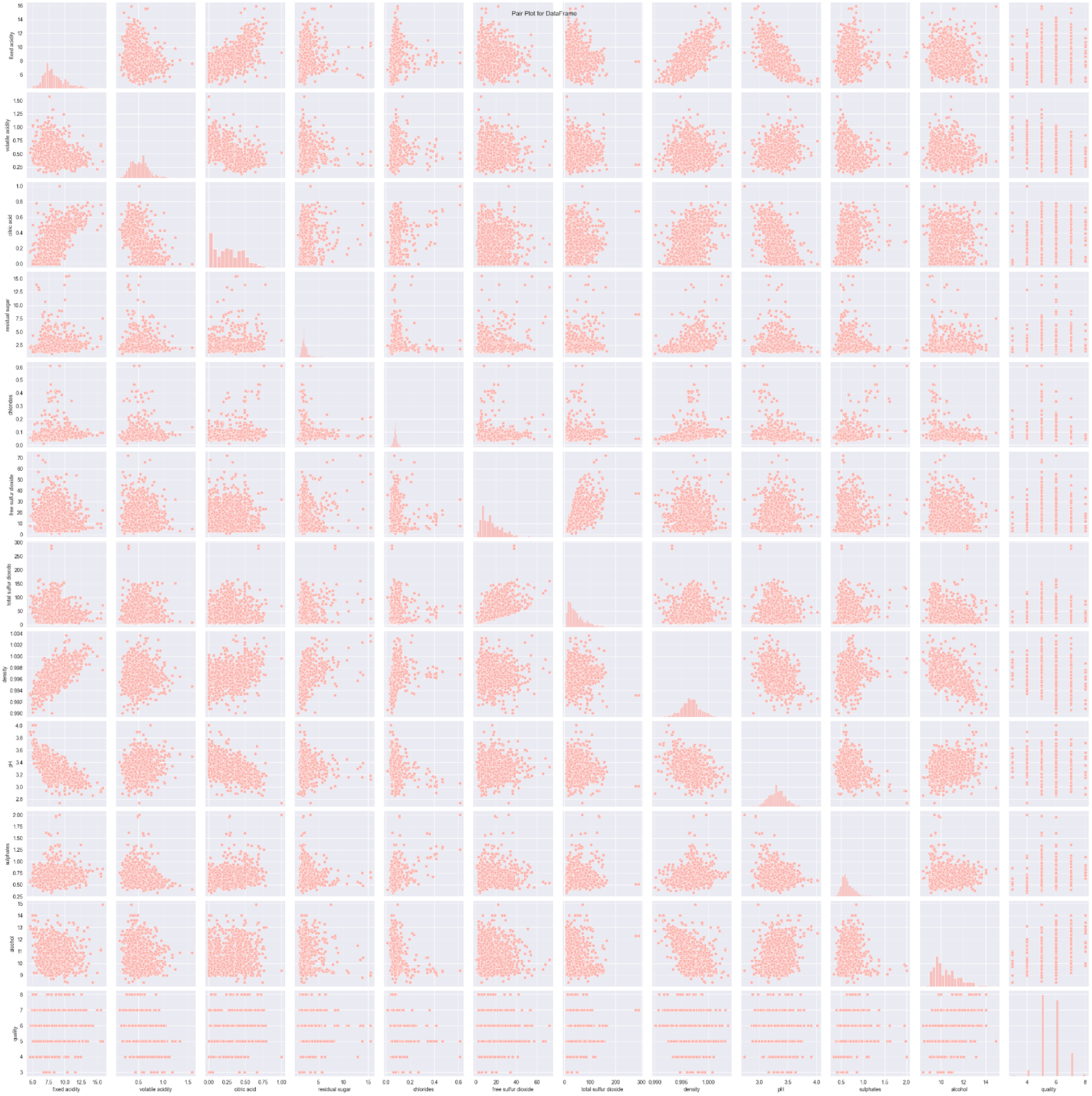
```
plt.figure(figsize=(10, 8))

sns.swarmplot(x="quality", y="alcohol", data=df, palette='viridis')
plt.title('Swarm Plot for Quality and Alcohol')
plt.xlabel('Quality')
plt.ylabel('Alcohol')
plt.show()
```

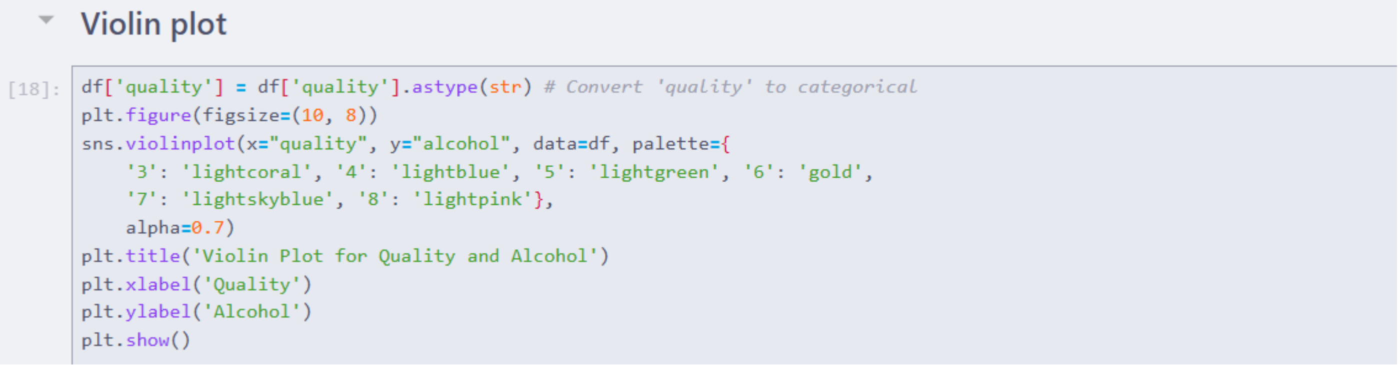


Pair Plots

[17]:

```
sns.set_palette("Pastel1")
plt.figure(figsize=(10, 6))
sns.pairplot(df)
plt.suptitle('Pair Plot for DataFrame')
plt.show()
```

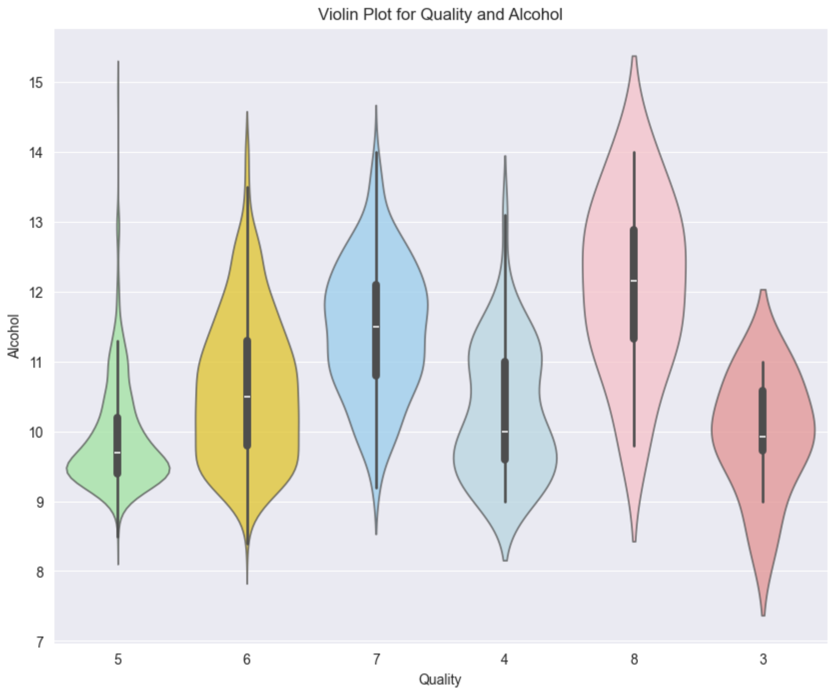




Violin plot

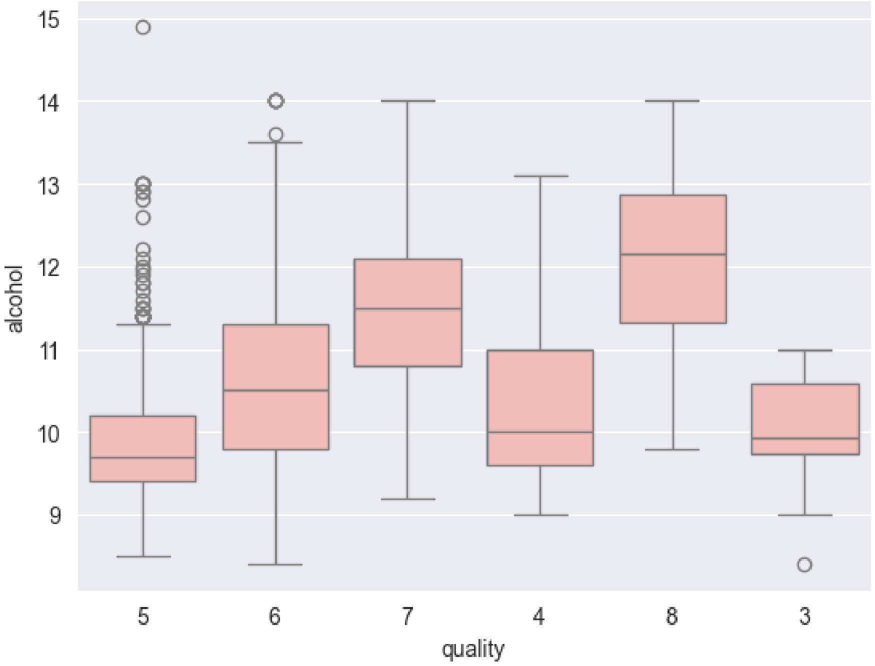
[18]:

```
df['quality'] = df['quality'].astype(str) # Convert 'quality' to categorical
plt.figure(figsize=(10, 8))
sns.violinplot(x="quality", y="alcohol", data=df, palette={
    '3': 'lightcoral', '4': 'lightblue', '5': 'lightgreen', '6': 'gold',
    '7': 'lightskyblue', '8': 'lightpink'},
    alpha=0.7)
plt.title('Violin Plot for Quality and Alcohol')
plt.xlabel('Quality')
plt.ylabel('Alcohol')
plt.show()
```



Box Plot

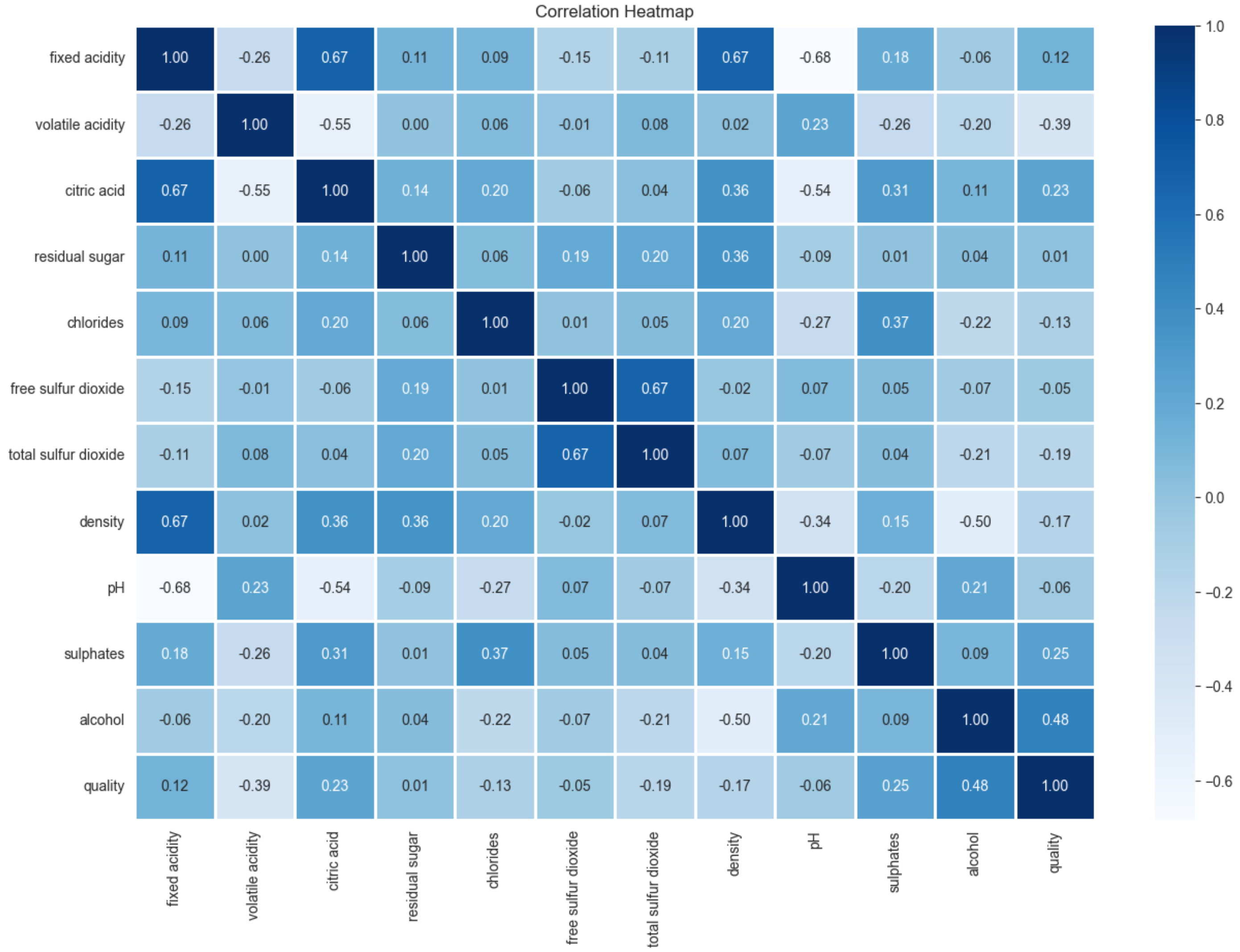
```
[20]: sns.boxplot(x='quality', y='alcohol', data=df)
```



Correlation Matrix

[25]:

```
plt.figure(figsize=(15, 10))
sns.heatmap(df.corr(), annot=True, fmt='.2f', cmap='Blues', linewidths=2)
plt.title('Correlation Heatmap')
plt.show()
```



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12
volatile acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.20	-0.39
citric acid	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	-0.09	0.01	0.04	0.01
chlorides	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	-0.02	0.07	0.05	-0.07	-0.05
total sulfur dioxide	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	-0.07	0.04	-0.21	-0.19
density	0.67	0.02	0.36	0.36	0.20	-0.02	0.07	1.00	-0.34	0.15	-0.50	-0.17
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1.00	-0.20	0.21	-0.06
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.20	1.00	0.09	0.25
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00	0.48
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05	-0.19	-0.17	-0.06	0.25	0.48	1.00