1 Batch 5 Task 2

[4]:
```
import pandas as pd
```

[52]:
```
df = pd.read_csv ("C:/Users/prath/Downloads/winequality-blue-black.csv")
```

[53]:
```
df
```

[53]:

	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		
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58		
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56		
…	…	…	…	…	…	…	…	…	…	…		
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58		
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76		
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75		
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71		
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66		

```
2          9.8        5
3          9.8        6
4          9.4        5
…          …        …
1594      10.5        5
1595      11.2        6
1596      11.0        6
1597      10.2        5
1598      11.0        6

[1599 rows x 12 columns]
```

[54]:
```
df.columns
```

[54]:
```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

[55]:
```
import matplotlib.pyplot as plt
```

[56]:
```
print("Original DataFrame:")
print(df)

print("\nRows that are duplicates of a previous row:")
print(df.duplicated())
```

```
Original DataFrame:
      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

      free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                    11.0                  34.0  0.99780  3.51       0.56
1                    25.0                  67.0  0.99680  3.20       0.68
2                    15.0                  54.0  0.99700  3.26       0.65
3                    17.0                  60.0  0.99800  3.16       0.58
4                    11.0                  34.0  0.99780  3.51       0.56
```

```
…                    …                    …       …   …         …
1594                 32.0                    44.0  0.99490  3.45       0.58
1595                 39.0                    51.0  0.99512  3.52       0.76
1596                 29.0                    40.0  0.99574  3.42       0.75
1597                 32.0                    44.0  0.99547  3.57       0.71
1598                 18.0                    42.0  0.99549  3.39       0.66

      alcohol  quality
0         9.4        5
1         9.8        5
2         9.8        5
3         9.8        6
4         9.4        5
…         …        …
1594     10.5        5
1595     11.2        6
1596     11.0        6
1597     10.2        5
1598     11.0        6

[1599 rows x 12 columns]

Rows that are duplicates of a previous row:
0       False
1       False
2       False
3       False
4        True
        …
1594    False
1595    False
1596     True
1597    False
1598    False
Length: 1599, dtype: bool
```

[57]:
```
df.describe()
```

[57]:
```
       fixed acidity  volatile acidity  citric acid  residual sugar  \
count    1599.000000       1599.000000  1599.000000     1599.000000
mean        8.319637          0.527821     0.270976        2.538806
std         1.741096          0.179060     0.194801        1.409928
min         4.600000          0.120000     0.000000        0.900000
25%         7.100000          0.390000     0.090000        1.900000
50%         7.900000          0.520000     0.260000        2.200000
75%         9.200000          0.640000     0.420000        2.600000
max        15.900000          1.580000     1.000000       15.500000
```

```
          chlorides  free sulfur dioxide  total sulfur dioxide      density  \
count  1599.000000          1599.000000           1599.000000  1599.000000
mean      0.087467            15.874922             46.467792     0.996747
std       0.047065            10.460157             32.895324     0.001887
min       0.012000             1.000000              6.000000     0.990070
25%       0.070000             7.000000             22.000000     0.995600
50%       0.079000            14.000000             38.000000     0.996750
75%       0.090000            21.000000             62.000000     0.997835
max       0.611000            72.000000            289.000000     1.003690

                pH    sulphates      alcohol      quality
count  1599.000000  1599.000000  1599.000000  1599.000000
mean      3.311113     0.658149    10.422983     5.636023
std       0.154386     0.169507     1.065668     0.807569
min       2.740000     0.330000     8.400000     3.000000
25%       3.210000     0.550000     9.500000     5.000000
50%       3.310000     0.620000    10.200000     6.000000
75%       3.400000     0.730000    11.100000     6.000000
max       4.010000     2.000000    14.900000     8.000000
```

[73]:
```
X=df.iloc[:100,0]
X
```

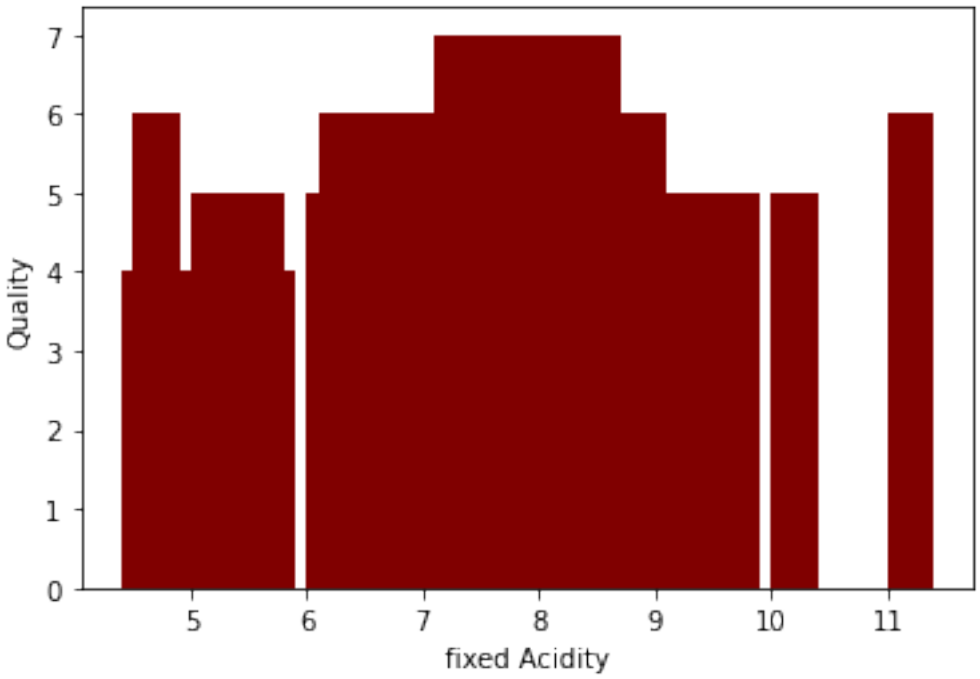
[73]:
```
0      7.4
1      7.8
2      7.8
3     11.2
4      7.4
      ...
95     4.7
96     6.8
97     7.0
98     7.6
99     8.1
Name: fixed acidity, Length: 100, dtype: float64
```

[74]:
```
Y=df.iloc[:100,11]
Y
```

[74]:
```
0     5
1     5
2     5
3     6
4     5
     ..
95    6
```

```
96    5
97    5
98    5
99    6
Name: quality, Length: 100, dtype: int64
```

[76]:
```
import matplotlib.pyplot as plt
plt.bar(X,Y,color="maroon",width=0.4)
plt.xlabel("fixed Acidity")
plt.ylabel("Quality")
plt.show()
```



2 Inference:

The value of Fixed Acidity mainly hovers between 6 and 10. The average value is 8 where the highest Quality is obtained.

[98]:
```
Z=df.iloc[:100,1]
Z
```

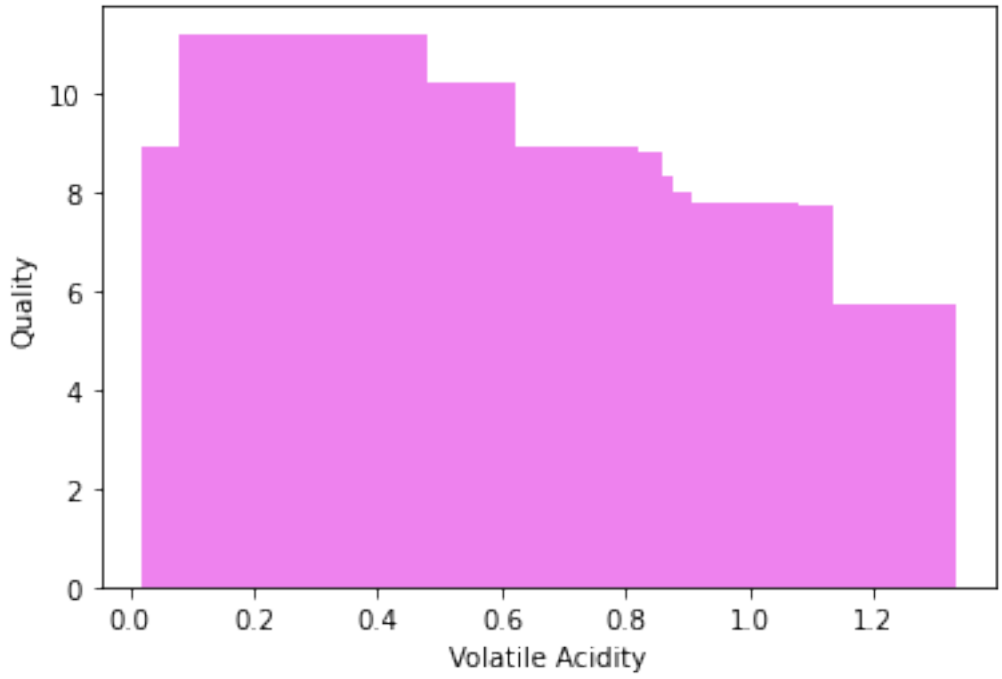
[98]:
```
0     0.700
1     0.880
2     0.760
3     0.280
```

```
4     0.700
      …
95    0.600
96    0.775
97    0.500
98    0.900
99    0.545
Name: volatile acidity, Length: 100, dtype: float64
```

[78]:
```
U=df.iloc[:100,0]
U
```

[78]:
```
0      7.4
1      7.8
2      7.8
3     11.2
4      7.4
      …
95     4.7
96     6.8
97     7.0
98     7.6
99     8.1
Name: fixed acidity, Length: 100, dtype: float64
```

[99]:
```
import matplotlib.pyplot as plt
plt.bar(Z,U,color="violet",width=0.4)
plt.xlabel("Volatile Acidity")
plt.ylabel("Quality")
plt.show()
```



3 Inference:

Volatile Acidity decreases on increasing the quality of Wine.

[80]:
```
A=df.iloc[:100,2]
A
```

[80]:
```
0     0.00
1     0.00
2     0.04
3     0.56
4     0.00
      ...
95    0.17
96    0.00
97    0.25
98    0.06
99    0.18
Name: citric acid, Length: 100, dtype: float64
```

[81]:
```
B=df.iloc[:100,3]
B
```

[81]:
```
0     1.9
1     2.6
2     2.3
3     1.9
4     1.9
     …
95    2.3
96    3.0
97    2.0
98    2.5
99    1.9
Name: residual sugar, Length: 100, dtype: float64
```

[82]:
```
C=df.iloc[:100,4]
C
```

[82]:
```
0     0.076
1     0.098
2     0.092
3     0.075
4     0.076
      …
95    0.058
96    0.102
97    0.070
98    0.079
99    0.080
Name: chlorides, Length: 100, dtype: float64
```

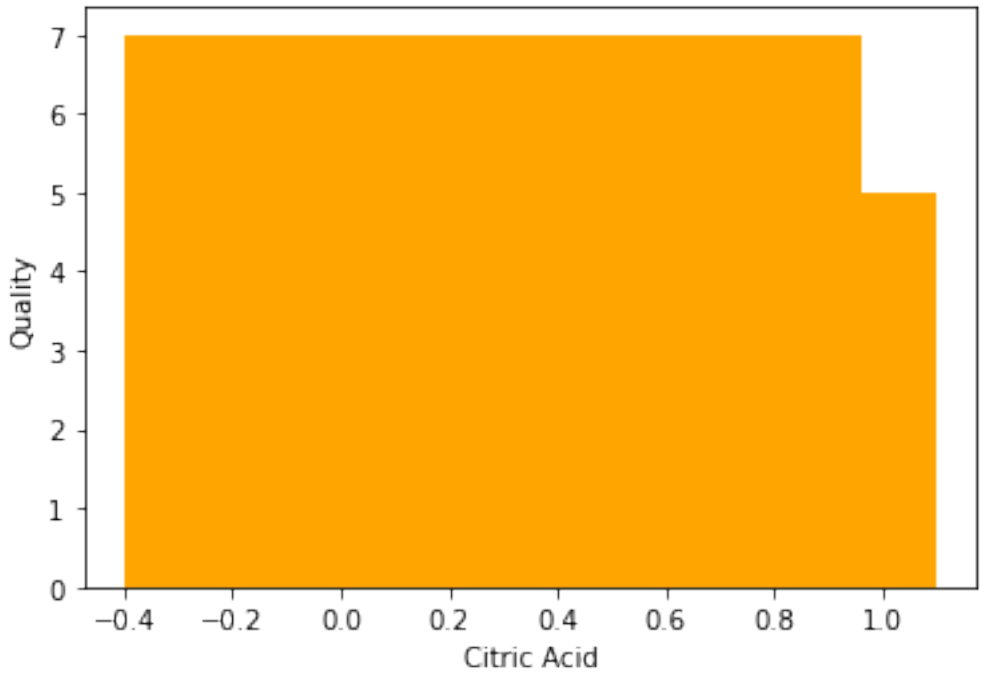
[94]:
```
D=df.iloc[:100,7]
D
```

[94]:
```
0     0.9978
1     0.9968
2     0.9970
3     0.9980
4     0.9978
       …
95    0.9932
96    0.9965
97    0.9963
98    0.9967
99    0.9972
Name: density, Length: 100, dtype: float64
```

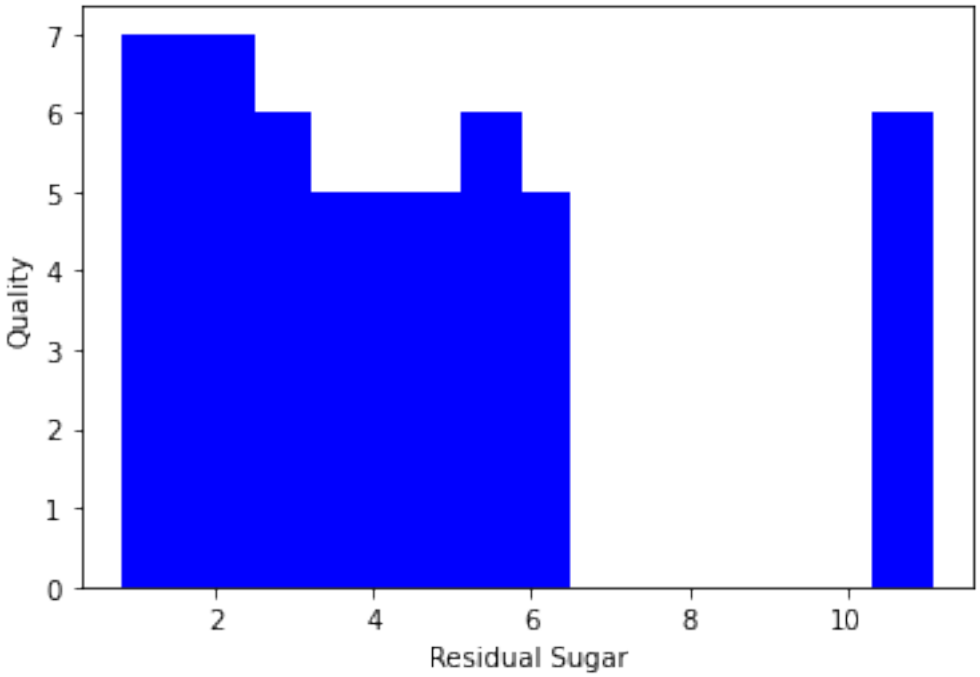
[95]:
```
E=df.iloc[:100,8]
E
```

[95]:
```
0     3.51
1     3.20
2     3.26
3     3.16
4     3.51
      …
95    3.85
96    3.45
97    3.25
98    3.39
99    3.30
Name: pH, Length: 100, dtype: float64
```

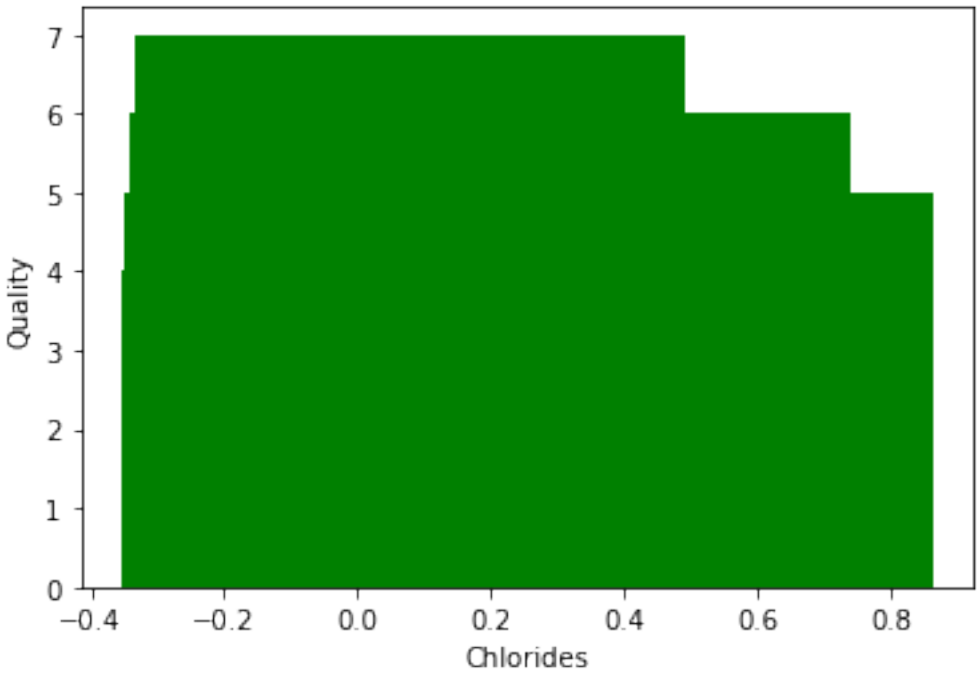
[88]:
```
import matplotlib.pyplot as plt
plt.bar(A,Y,color="orange")
plt.xlabel("Citric Acid")
plt.ylabel("Quality")
plt.show()
```



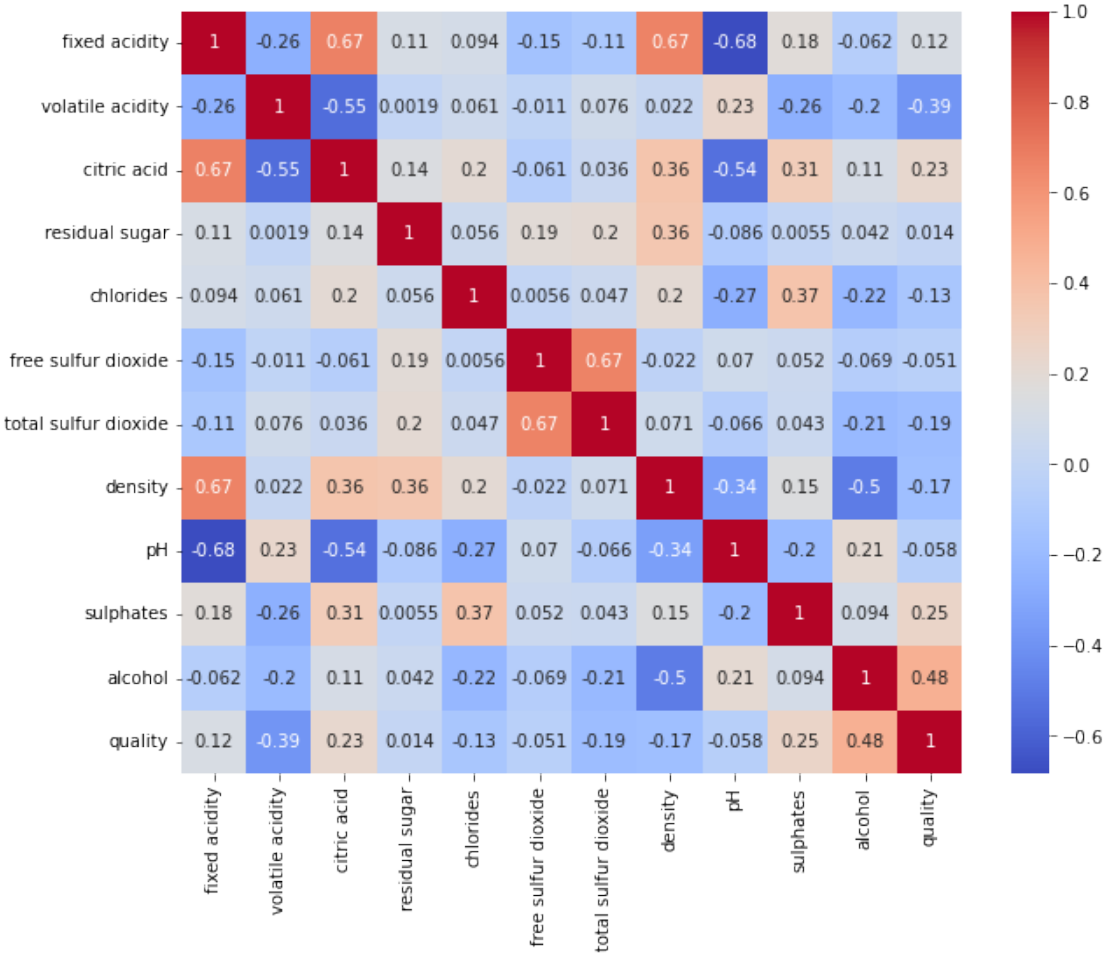
[89]:
```
import matplotlib.pyplot as plt
plt.bar(B,Y,color="blue")
plt.xlabel("Residual Sugar")
plt.ylabel("Quality")
plt.show()
```



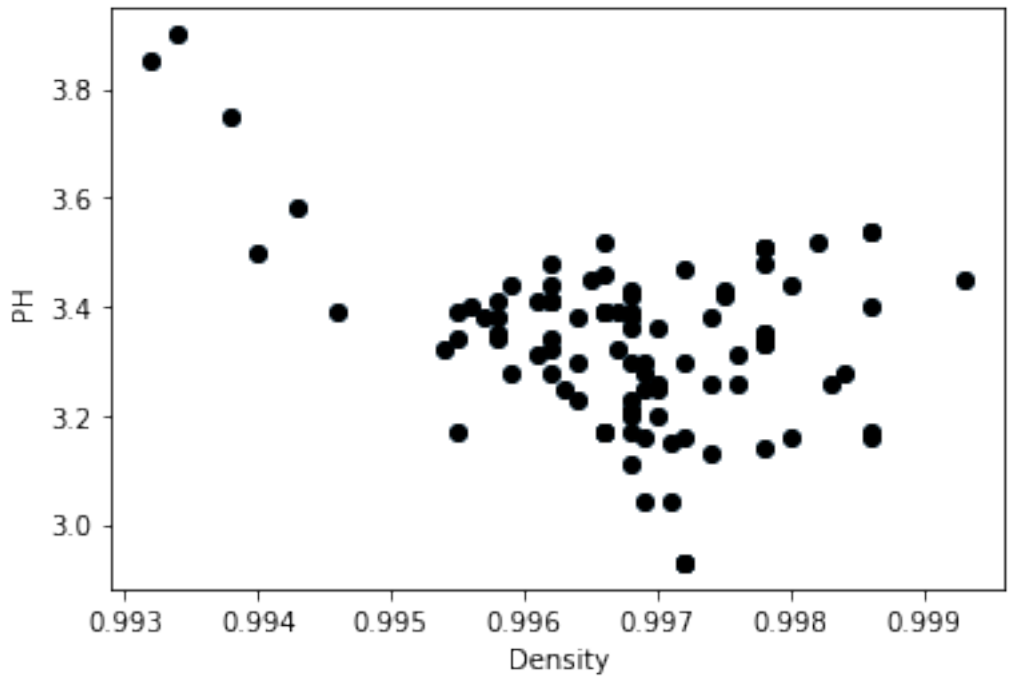
[91]:
```
import matplotlib.pyplot as plt
plt.bar(C,Y,color="Green")
plt.xlabel("Chlorides")
plt.ylabel("Quality")
plt.show()
```



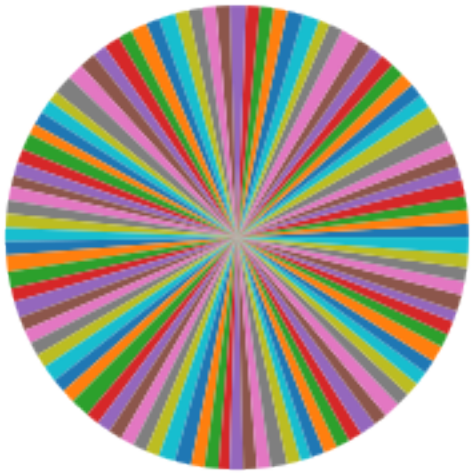
[93]:
```
import seaborn as sns
# Calculate the correlation matrix
corr = df.corr()

# Display the correlation matrix as a heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.show()
```



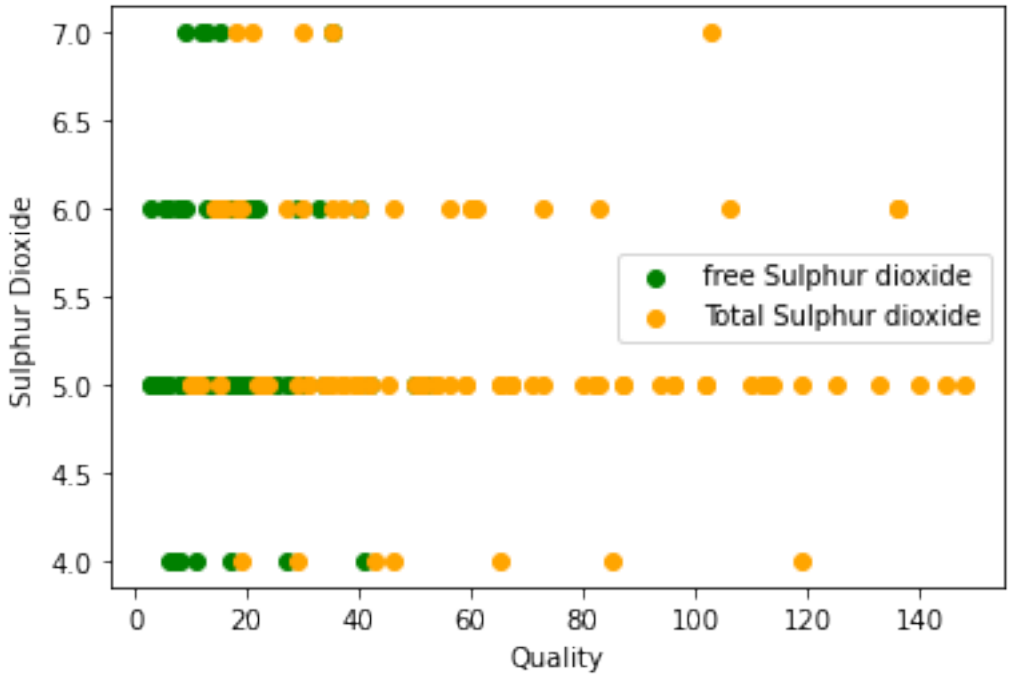
[97]:
```
plt.scatter(D,E)
plt.scatter(D,E,color="black")
plt.xlabel("Density")
plt.ylabel("PH")
plt.show()
```



[100]:
```
plt.pie(Y)
plt.show()
```



[106]:
```
M=df.iloc[:100,5]
N=df.iloc[:100,6]
plt.scatter(M,Y,label="free Sulphur dioxide",color="Green")
plt.scatter(N,Y,label="Total Sulphur dioxide",color="orange")
plt.legend()
plt.xlabel("Quality")
plt.ylabel("Sulphur Dioxide")
plt.show()
```



4 Conclusion: We studied the data on wine quality. Various parameters were studied such as acidity, sugar level, chlorides, sugar, citric acid etc.

5 The plotted graphs show a few very visible trends.

6 The value of Fixed Acidity mainly hovers between 6 and 10. The average value is 8 where the highest Quality is obtained.

7 Volatile Acidity decreases on increasing the quality of Wine.

8 The quality of wine decreases as chlorides and citric acid increases.

9 Fixed acidity and Ph are the most inversely correlated.