NAME: - Meetkumar Vasava

MACHINE LEARNING INTERNSHIP @ BHARAT INTERN

PROJECT NAME - **WINE QUALITY PREDICTION**

Github Link

https://github.com/MeetVasava/Wine_Quality_Predictio

In [27]:

```
import pandas as pd
import matplotlib.pyplot as pt
from sklearn.linear_model import LinearRegression
```

In [28]:

```
import seaborn as sb
```

In [29]:

```
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
```

Gathering, Processing and Cleaning the data

In [30]:

```
wine = pd.read_csv('WineQT.csv')
```

In [31]:

```
wine
```

Out[31]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4
...	...	...	...	...	...	...	...	...	...	...	...
1138	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0
1139	6.8	0.620	0.08	1.9	0.068	28.0	38.0	0.99651	3.42	0.82	9.5
1140	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5
1141	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2
1142	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2

1143 rows × 13 columns

In [32]:

```
wine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1143 entries, 0 to 1142
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1143 non-null   float64
 1   volatile acidity      1143 non-null   float64
 2   citric acid           1143 non-null   float64
 3   residual sugar        1143 non-null   float64
 4   chlorides             1143 non-null   float64
 5   free sulfur dioxide   1143 non-null   float64
 6   total sulfur dioxide  1143 non-null   float64
 7   density               1143 non-null   float64
 8   pH                    1143 non-null   float64
 9   sulphates             1143 non-null   float64
 10  alcohol               1143 non-null   float64
 11  quality               1143 non-null   int64  
 12  Id                    1143 non-null   int64  
dtypes: float64(11), int64(2)
memory usage: 116.2 KB
```

In [33]:

```
wine.pop('Id')
```

Out[33]:

```
0          0
1          1
2          2
3          3
4          4
        ...
1138    1592
1139    1593
1140    1594
1141    1595
1142    1597
Name: Id, Length: 1143, dtype: int64
```

In [34]:

```
wine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1143 entries, 0 to 1142
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1143 non-null   float64
 1   volatile acidity      1143 non-null   float64
 2   citric acid           1143 non-null   float64
 3   residual sugar        1143 non-null   float64
 4   chlorides             1143 non-null   float64
 5   free sulfur dioxide   1143 non-null   float64
 6   total sulfur dioxide  1143 non-null   float64
 7   density               1143 non-null   float64
 8   pH                    1143 non-null   float64
 9   sulphates             1143 non-null   float64
 10  alcohol               1143 non-null   float64
 11  quality               1143 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 107.3 KB
```

In [35]:

```
wine.columns
```

Out[35]:

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

In [36]:

```
y = wine['quality']
x = wine[['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide','density',
       'pH', 'sulphates', 'alcohol']]
```

Plotting

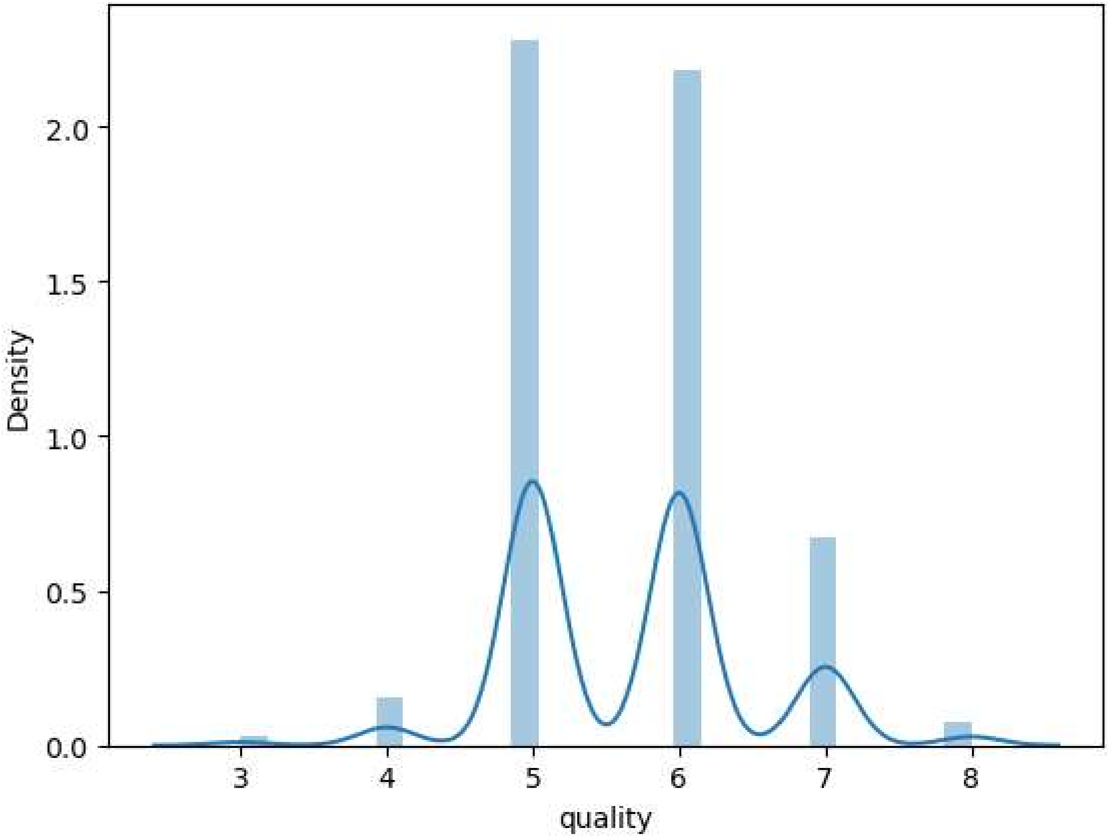
In [37]:

```
sb.distplot(wine['quality'])
```

```
C:\Users\Meet\AppData\Local\Temp\ipykernel_16440\3304093463.py:1: 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

  sb.distplot(wine['quality'])
```

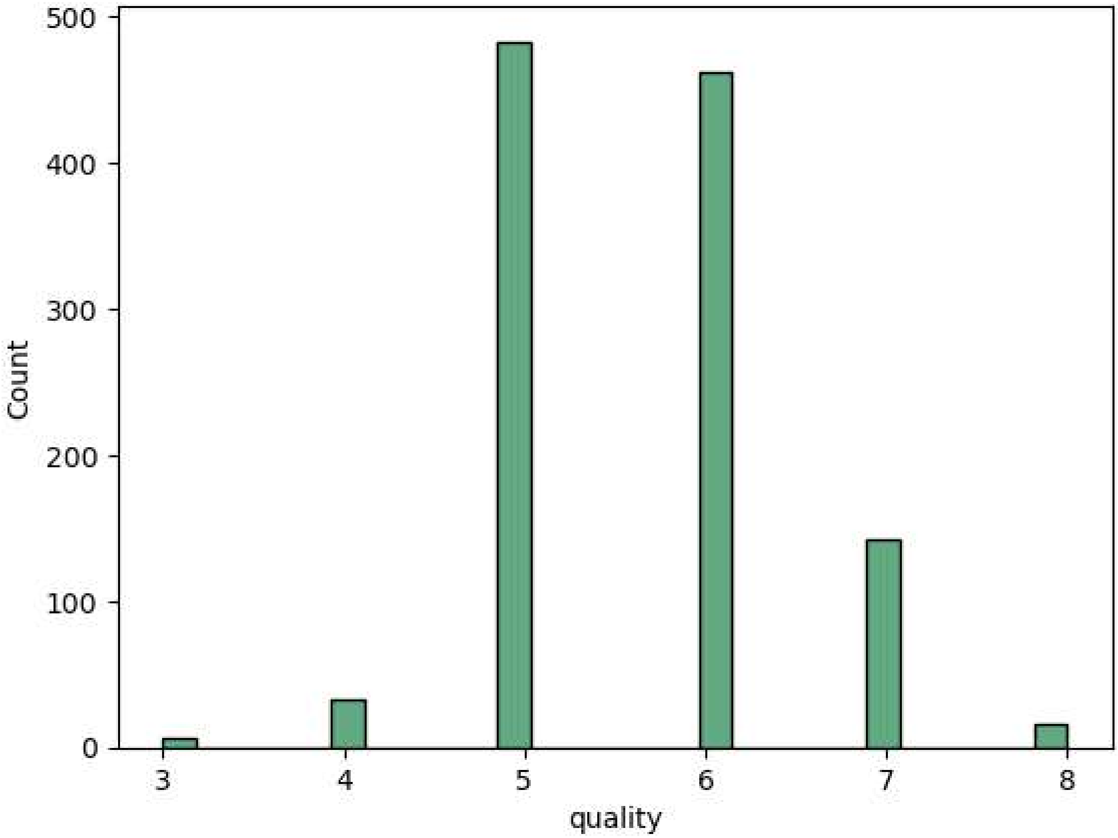
Out[37]:

```
<Axes: xlabel='quality', ylabel='Density'>
```



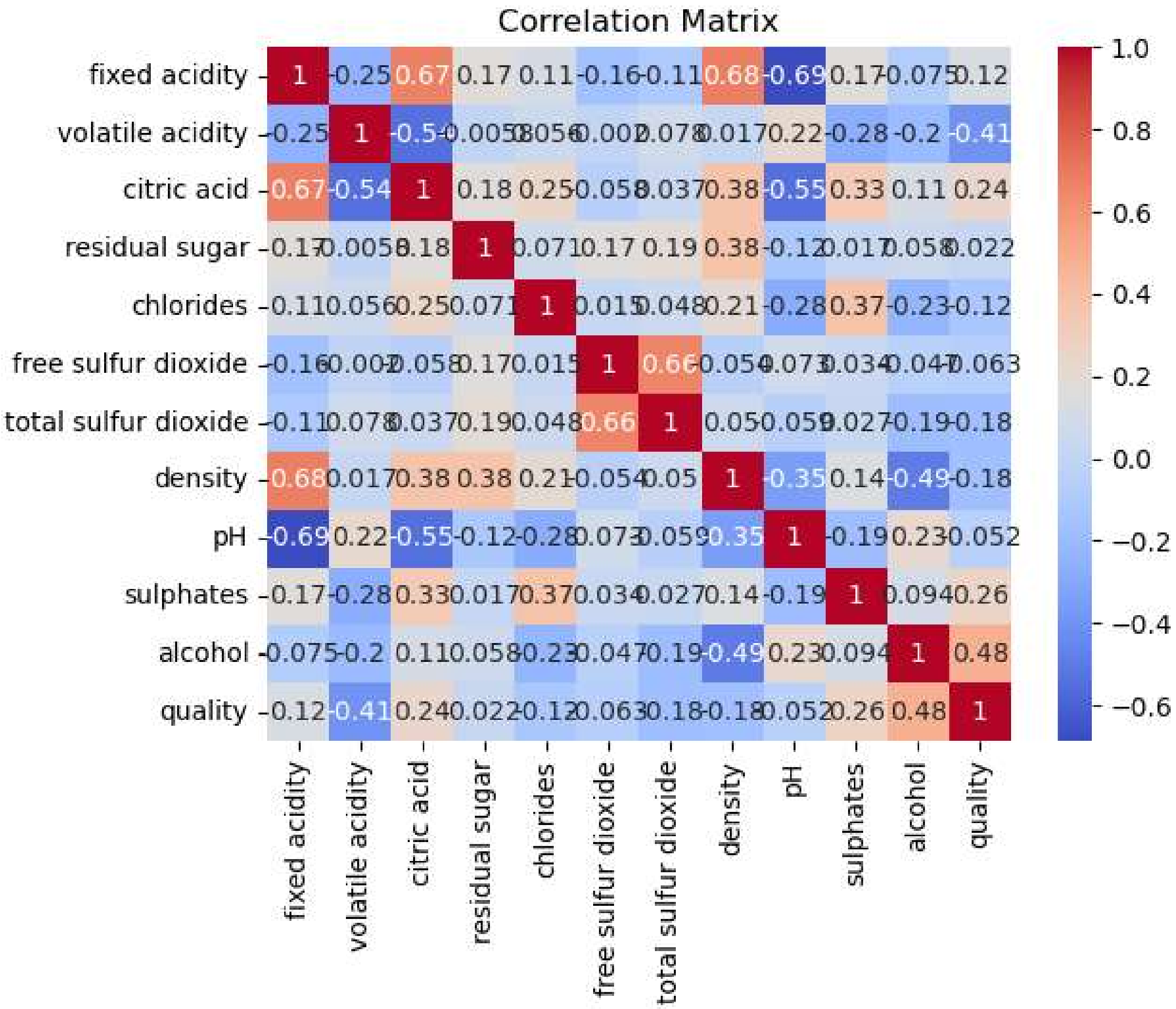
In [38]:

```
sb.histplot(wine['quality'], color = 'seagreen')
pt.show()
```

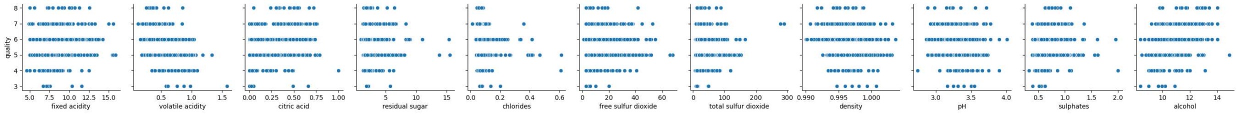


In [39]:

```
co_matrix = wine[['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar'
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide','density',
       'pH', 'sulphates', 'alcohol','quality']].corr()
sb.heatmap(co_matrix, annot = True, cmap = 'coolwarm')
pt.title('Correlation Matrix')
pt.show()
```



In [40]:
```
sb.pairplot(wine, x_vars = ['fixed acidity', 'volatile acidity', 'citric acid', 'resic
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide','density',
       'pH', 'sulphates', 'alcohol'], y_vars = 'quality', kind = 'scatter')
pt.show()
```



Training and Testing

In [41]:
```
from sklearn.model_selection import train_test_split
```

In [42]:
```
xtrain, xtest, ytrain, ytest = train_test_split(x,y,test_size=0.2) #training the model
```

In [43]:
```
winelr = LinearRegression()
```

In [44]:
```
winelr.fit(xtrain, ytrain)
```

```
Out[44]: ▾ LinearRegression
         LinearRegression()
```

```
In [45]: winelr.coef_
```

```
Out[45]: array([ 7.95837635e-03, -1.08098123e+00, -1.17189079e-01, -1.08053493e-02,
                -1.84472543e+00,  2.54184412e-03, -2.23246763e-03, -2.72636291e+00,
                -6.01985470e-01,  7.57757023e-01,  3.07720048e-01])
```

```
In [46]: pd.DataFrame(winelr.coef_,index=x.columns,columns=['mycoef'])
```

Out[46]:

	mycoef
fixed acidity	0.007958
volatile acidity	-1.080981
citric acid	-0.117189
residual sugar	-0.010805
chlorides	-1.844725
free sulfur dioxide	0.002542
total sulfur dioxide	-0.002232
density	-2.726363
pH	-0.601985
sulphates	0.757757
alcohol	0.307720

```
In [47]: pr = winelr.predict(xtest)
```

Metrics

```
In [48]: r2_score(ytest, pr)

Out[48]: 0.39188286605967915
```

```
In [49]: mean_absolute_error(ytest,pr)

Out[49]: 0.5112609861165464
```

```
In [50]: mean_squared_error(ytest, pr)

Out[50]: 0.41275569240616666
```

```
In [51]: x.columns
```

Out[51]:

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

Prediction

In [52]:

```
winelr.predict([[8, 0.5, 0.15, 1.9, 0.07, 23.0, 35.0, 0.92, 3.5, 0.65, 10.2]])
```

```
C:\Users\Meet\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does no
t have valid feature names, but LinearRegression was fitted with feature names
  warnings.warn(
```

Out[52]:

```
array([5.79519792])
```

In [ ]: