EDA with Red Wine Data Quality - Fazal Rehman

In [ ]:
```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:
```
df=pd.read_csv("C:/Users/fazal/OneDrive/Desktop/winequality-red.csv",sep=";")
```

In [5]:
```
df.head()
```

Out[5]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	a
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	

In [6]:
```
df.describe() #for descriptive statistics
```

Out[6]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	tota
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

In [9]:
```
df.shape
```

Out[9]:
```
(1599, 12)
```

In [11]:
```
df.columns
```

Out[11]:
```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

In [13]:
```
df["quality"].unique()
```

Out[13]:
```
array([5, 6, 7, 4, 8, 3], dtype=int64)
```

In [15]:
```
df["quality"].value_counts()
```

Out[15]:
```
quality
5    681
6    638
7    199
4     53
8     18
3     10
Name: count, dtype: int64
```

In [17]:
```
df.isnull().sum() #missing values
```

Out[17]:
```
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
```

In [18]:
```
df.duplicated() #to check duplicacy
```

Out[18]:
```
0       False
1       False
2       False
3       False
4        True
        ...
1594    False
1595    False
1596     True
1597    False
1598    False
Length: 1599, dtype: bool
```

In [19]:
```
df[df.duplicated()]
```

Out[19]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphate
4	7.4	0.700	0.00	1.90	0.076	11.0	34.0	0.99780	3.51	0.5
11	7.5	0.500	0.36	6.10	0.071	17.0	102.0	0.99780	3.35	0.8
27	7.9	0.430	0.21	1.60	0.106	10.0	37.0	0.99660	3.17	0.9
40	7.3	0.450	0.36	5.90	0.074	12.0	87.0	0.99780	3.33	0.8
65	7.2	0.725	0.05	4.65	0.086	4.0	11.0	0.99620	3.41	0.3
...	...	...	...	...	...	...	...	...	...	.
1563	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.29	0.5
1564	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.29	0.5
1567	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.29	0.5
1581	6.2	0.560	0.09	1.70	0.053	24.0	32.0	0.99402	3.54	0.6
1596	6.3	0.510	0.13	2.30	0.076	29.0	40.0	0.99574	3.42	0.7

240 rows × 12 columns

In [20]:
```
df.drop_duplicates(inplace=True) #remove duplicate
```

In [21]:
```
df.shape
```

Out[21]:
```
(1359, 12)
```

In [22]:
```
df.corr()
```

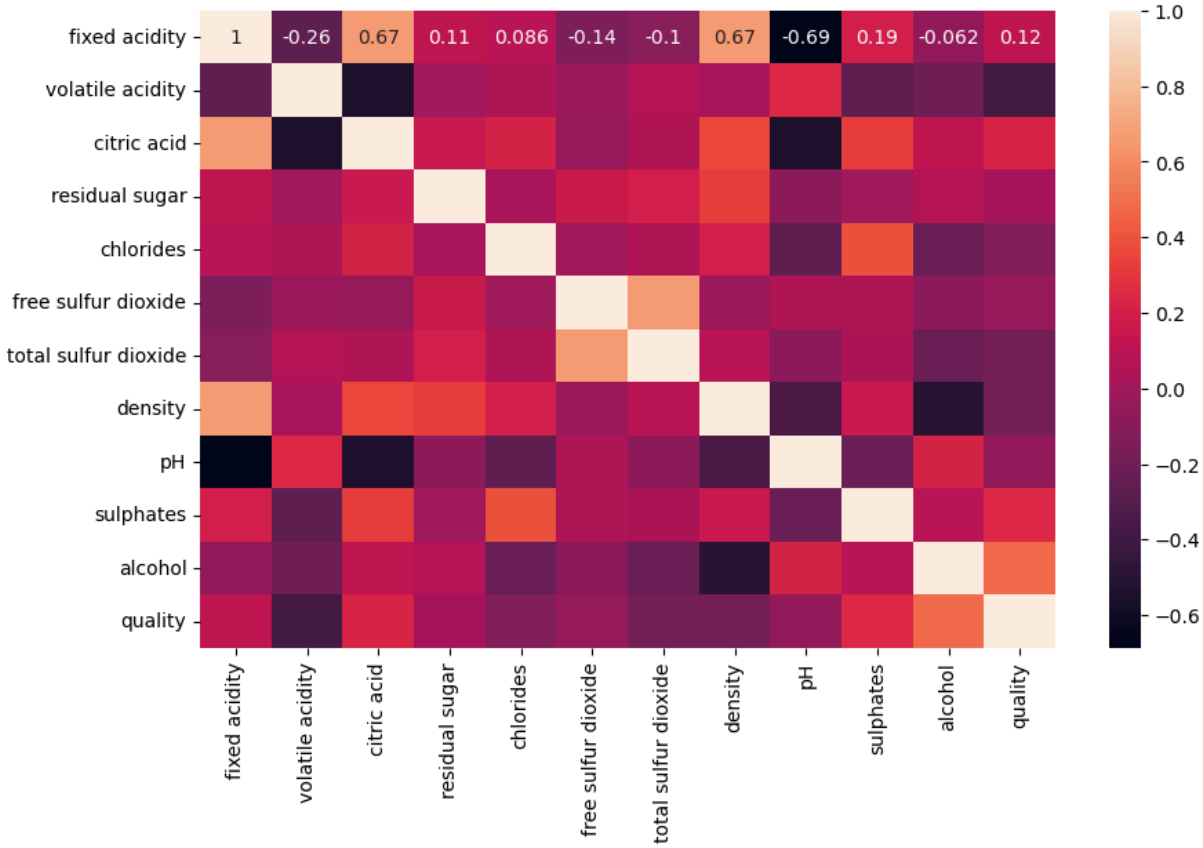
Out[22]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	den
fixed acidity	1.000000	-0.255124	0.667437	0.111025	0.085886	-0.140580	-0.103777	0.67
volatile acidity	-0.255124	1.000000	-0.551248	-0.002449	0.055154	-0.020945	0.071701	0.02
citric acid	0.667437	-0.551248	1.000000	0.143892	0.210195	-0.048004	0.047358	0.35
residual sugar	0.111025	-0.002449	0.143892	1.000000	0.026656	0.160527	0.201038	0.32
chlorides	0.085886	0.055154	0.210195	0.026656	1.000000	0.000749	0.045773	0.19
free sulfur dioxide	-0.140580	-0.020945	-0.048004	0.160527	0.000749	1.000000	0.667246	-0.01
total sulfur dioxide	-0.103777	0.071701	0.047358	0.201038	0.045773	0.667246	1.000000	0.07
density	0.670195	0.023943	0.357962	0.324522	0.193592	-0.018071	0.078141	1.00
pH	-0.686685	0.247111	-0.550310	-0.083143	-0.270893	0.056631	-0.079257	-0.35
sulphates	0.190269	-0.256948	0.326062	-0.011837	0.394557	0.054126	0.035291	0.14
alcohol	-0.061596	-0.197812	0.105108	0.063281	-0.223824	-0.080125	-0.217829	-0.50
quality	0.119024	-0.395214	0.228057	0.013640	-0.130988	-0.050463	-0.177855	-0.18

In [24]:

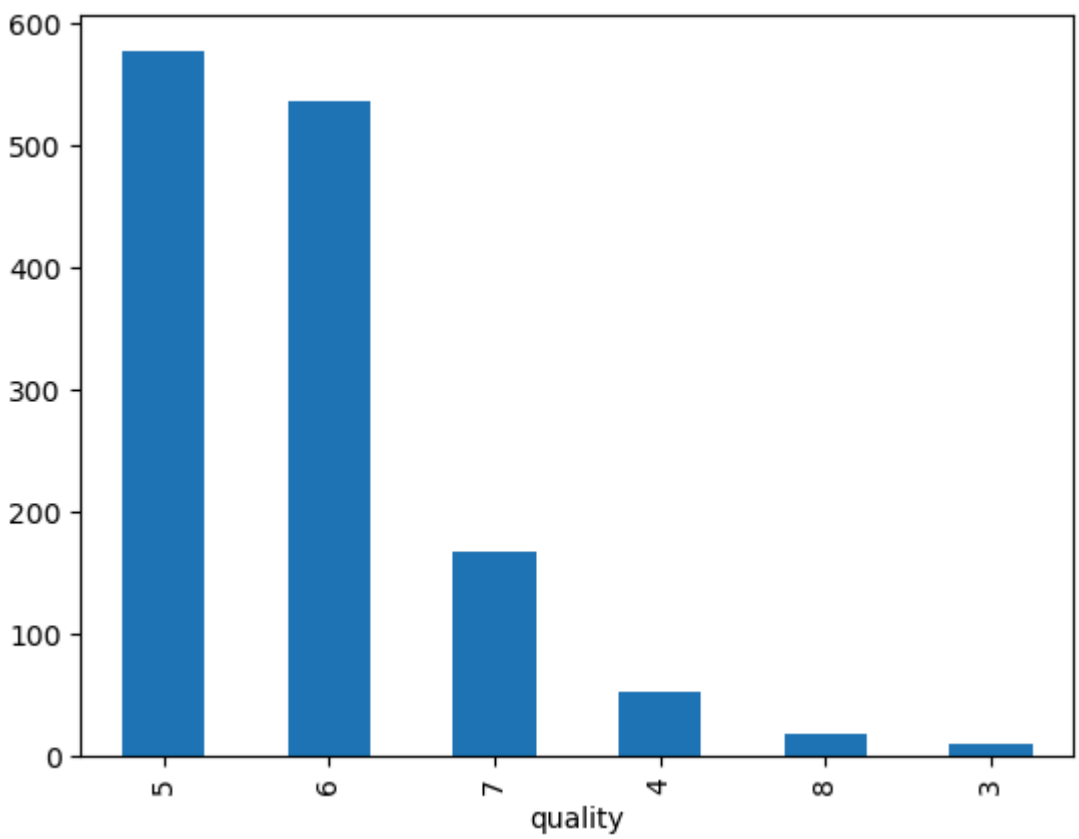
```
plt.figure(figsize=(10,6))
sns.heatmap(df.corr(),annot=True)
```

Out[24]: <Axes: >



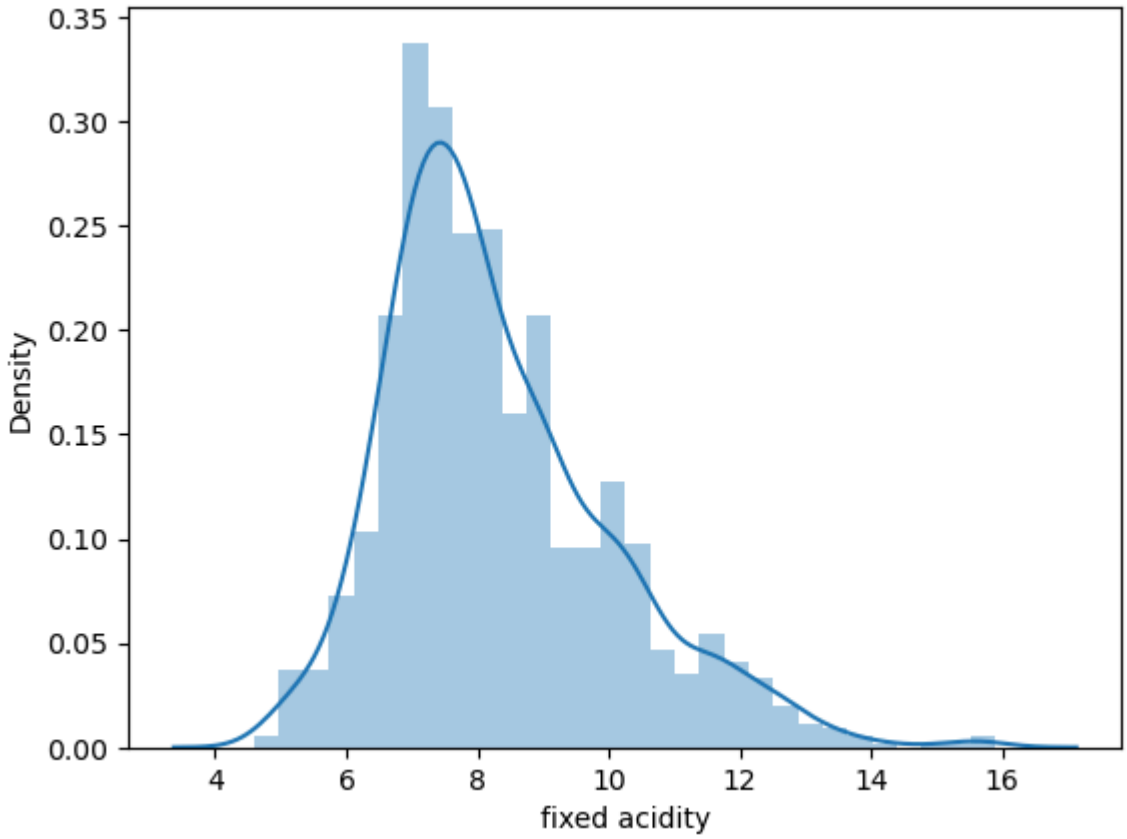
In [25]:
```
df.quality.value_counts().plot(kind="bar")
```

Out[25]: <Axes: xlabel='quality'>



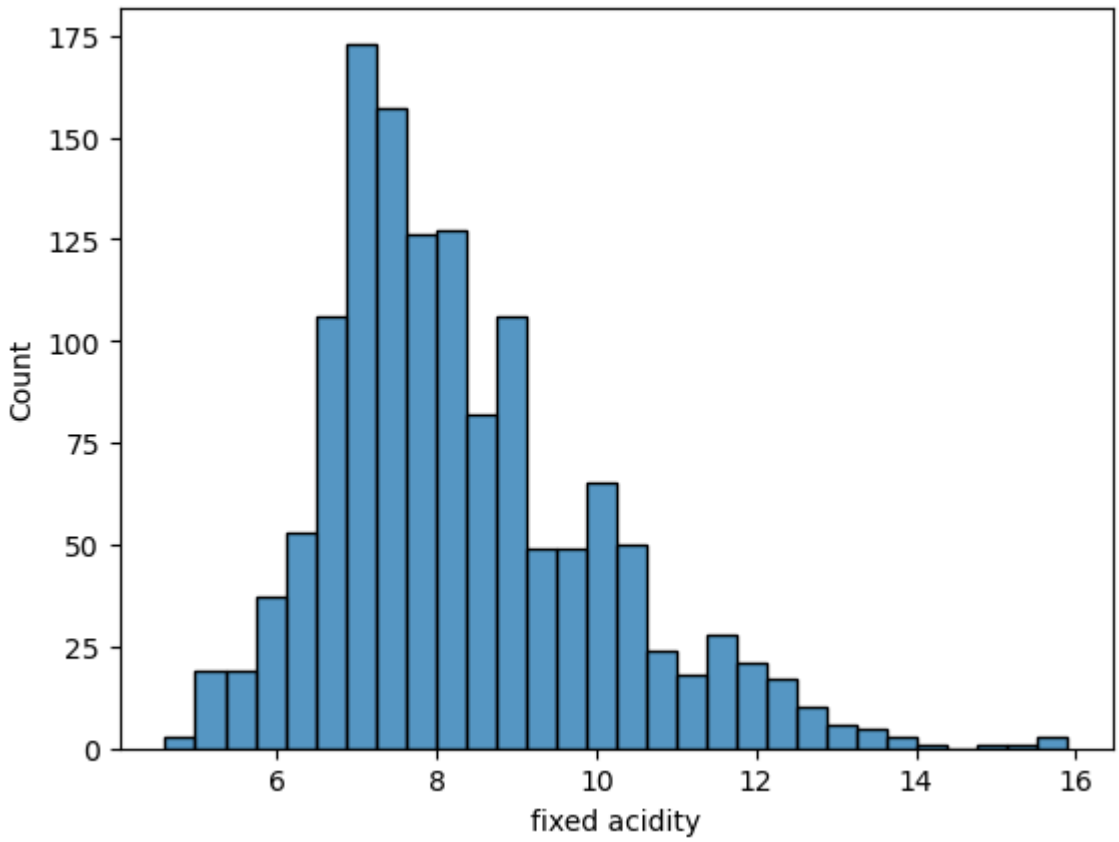
In [29]:
```
sns.distplot(df["fixed acidity"])
```

Out[29]:
```
<Axes: xlabel='fixed acidity', ylabel='Density'>
```

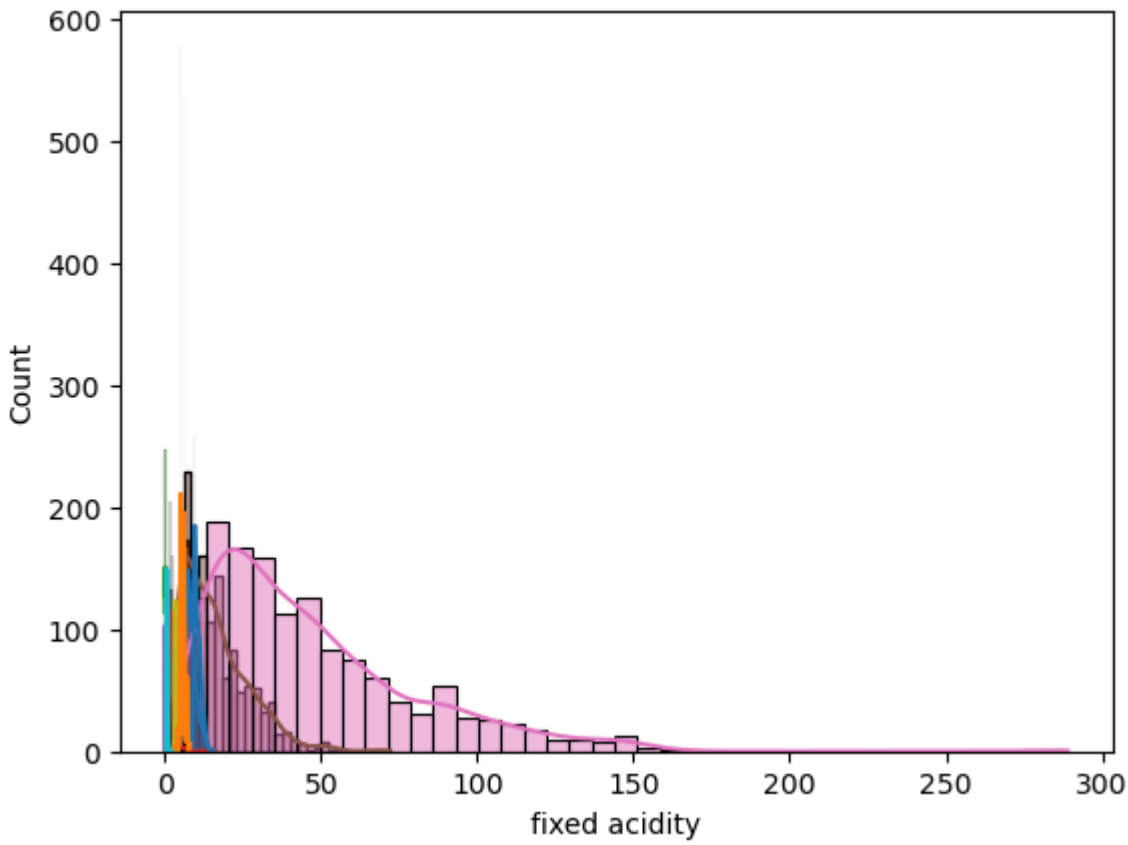


In [30]:
```
sns.histplot(df["fixed acidity"])
```

Out[30]:
```
<Axes: xlabel='fixed acidity', ylabel='Count'>
```

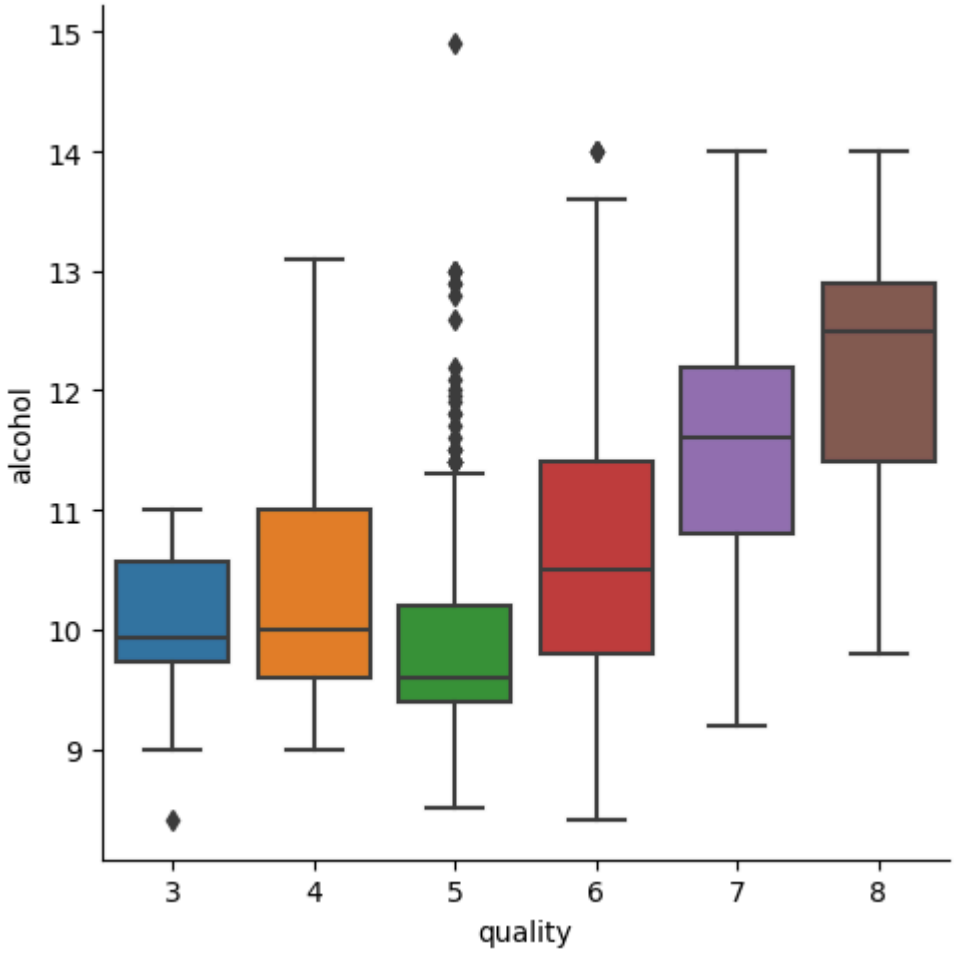


In [31]:
```
for i in df.columns:
    sns.histplot(df[i],kde=True)
```



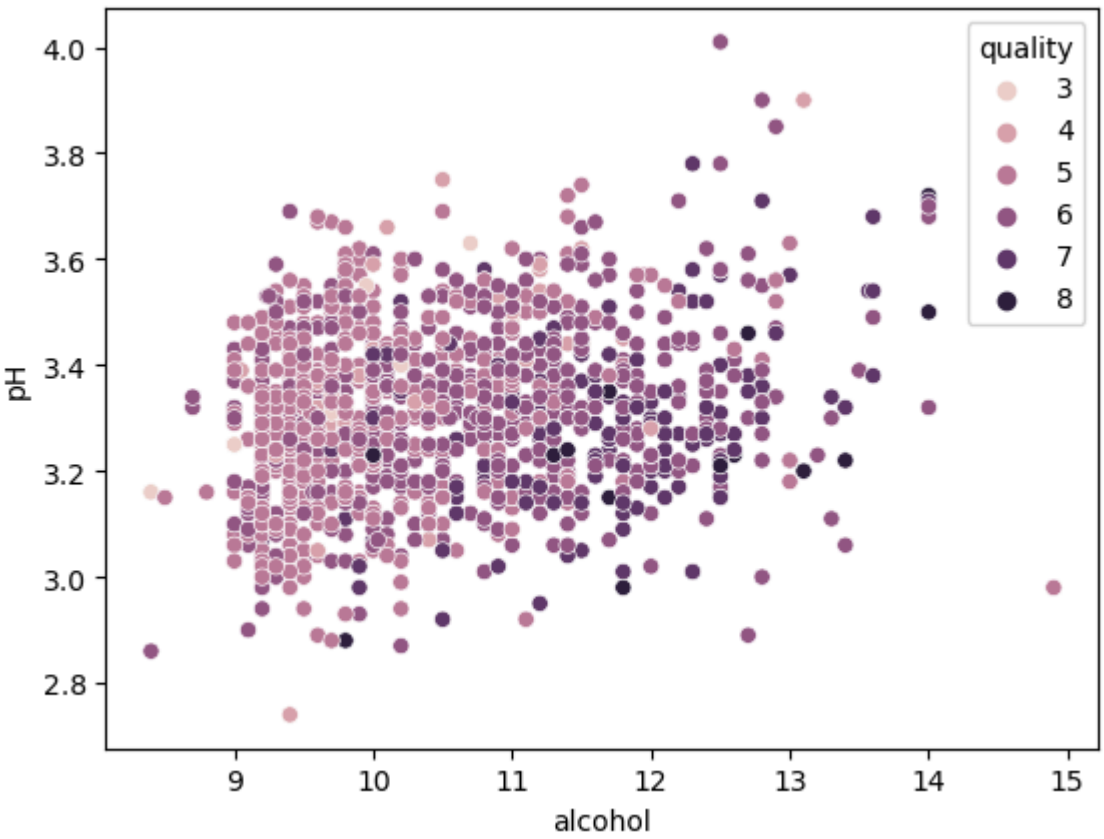
In [34]:
```
sns.catplot(x="quality",y="alcohol",data=df,kind="box") #categorical plot
```

Out[34]:
```
<seaborn.axisgrid.FacetGrid at 0x1d7254b7550>
```



In [36]:
```
sns.scatterplot(x="alcohol",y="pH",hue="quality",data=df)
```

Out[36]:
```
<Axes: xlabel='alcohol', ylabel='pH'>
```


quality
3
4
5
6
7
8
pH
4.0
3.8
3.6
3.4
3.2
3.0
2.8
9
10
11
12
13
14
15
alcohol