1llodbpid

October 8, 2023

NAME- AKARSHIT MISRA

REG NO. 21BAI1597

0.1 Data Preprocessing

```python
[78]: import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt
      import seaborn as sns
```

```python
[79]: df = pd.read_csv('/content/winequality.csv')
      df.head()
```

```
[79]:     type  fixed acidity  volatile acidity  citric acid  residual sugar  \
      0  white            7.0              0.27         0.36            20.7
      1  white            6.3              0.30         0.34             1.6
      2  white            8.1              0.28         0.40             6.9
      3  white            7.2              0.23         0.32             8.5
      4  white            7.2              0.23         0.32             8.5

         chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  \
      0      0.045                 45.0                 170.0   1.0010  3.00
      1      0.049                 14.0                 132.0   0.9940  3.30
      2      0.050                 30.0                  97.0   0.9951  3.26
      3      0.058                 47.0                 186.0   0.9956  3.19
      4      0.058                 47.0                 186.0   0.9956  3.19

         sulphates  alcohol  quality
      0       0.45      8.8        6
      1       0.49      9.5        6
      2       0.44     10.1        6
      3       0.40      9.9        6
      4       0.40      9.9        6
```

```python
[80]: df.shape
```

```
[80]: (6497, 13)
```

```
[81]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   type                  6497 non-null   object
 1   fixed acidity         6487 non-null   float64
 2   volatile acidity      6489 non-null   float64
 3   citric acid           6494 non-null   float64
 4   residual sugar        6495 non-null   float64
 5   chlorides             6495 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6488 non-null   float64
 10  sulphates             6493 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

```
[82]: df.isnull().any()
```

```
[82]: type                  False
      fixed acidity          True
      volatile acidity       True
      citric acid            True
      residual sugar         True
      chlorides              True
      free sulfur dioxide   False
      total sulfur dioxide  False
      density               False
      pH                     True
      sulphates              True
      alcohol               False
      quality               False
      dtype: bool
```

```
[83]: df.isnull().sum()
```

```
[83]: type                   0
      fixed acidity         10
      volatile acidity       8
      citric acid            3
      residual sugar         2
```

```
chlorides              2
free sulfur dioxide    0
total sulfur dioxide   0
density                0
pH                     9
sulphates              4
alcohol                0
quality                0
dtype: int64
```

[84]: `df.describe()`

[84]:
	fixed acidity	volatile acidity	citric acid	residual sugar
count	6487.000000	6489.000000	6494.000000	6495.000000
mean	7.216579	0.339691	0.318722	5.444326
std	1.296750	0.164649	0.145265	4.758125
min	3.800000	0.080000	0.000000	0.600000
25%	6.400000	0.230000	0.250000	1.800000
50%	7.000000	0.290000	0.310000	3.000000
75%	7.700000	0.400000	0.390000	8.100000
max	15.900000	1.580000	1.660000	65.800000

	chlorides	free sulfur dioxide	total sulfur dioxide	density
count	6495.000000	6497.000000	6497.000000	6497.000000
mean	0.056042	30.525319	115.744574	0.994697
std	0.035036	17.749400	56.521855	0.002999
min	0.009000	1.000000	6.000000	0.987110
25%	0.038000	17.000000	77.000000	0.992340
50%	0.047000	29.000000	118.000000	0.994890
75%	0.065000	41.000000	156.000000	0.996990
max	0.611000	289.000000	440.000000	1.038980

	pH	sulphates	alcohol	quality
count	6488.000000	6493.000000	6497.000000	6497.000000
mean	3.218395	0.531215	10.491801	5.818378
std	0.160748	0.148814	1.192712	0.873255
min	2.720000	0.220000	8.000000	3.000000
25%	3.110000	0.430000	9.500000	5.000000
50%	3.210000	0.510000	10.300000	6.000000
75%	3.320000	0.600000	11.300000	6.000000
max	4.010000	2.000000	14.900000	9.000000

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

[8]: 7

[9]: `df.quality.unique()`

```
[9]: array([6, 5, 7, 8, 4, 3, 9])
```

```
[10]: df.quality.value_counts()
```

```
[10]: 6    2836
      5    2138
      7    1079
      4     216
      8     193
      3      30
      9       5
      Name: quality, dtype: int64
```

```
[11]: df.alcohol.value_counts()
```

```
[11]: 9.500000    367
      9.400000    332
      9.200000    271
      10.000000   229
      10.500000   227
                   …
      10.533333     1
      11.366667     1
      12.333333     1
      14.050000     1
      10.750000     1
      Name: alcohol, Length: 111, dtype: int64
```

```
[12]: df.corr()
```

```
<ipython-input-12-2f6f6606aa2c>:1: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it will
default to False. Select only valid columns or specify the value of numeric_only
to silence this warning.
  df.corr()
```

```
[12]:                       fixed acidity  volatile acidity  citric acid  \
      fixed acidity              1.000000          0.220172     0.323736
      volatile acidity           0.220172          1.000000    -0.378061
      citric acid                0.323736         -0.378061     1.000000
      residual sugar            -0.112319         -0.196702     0.142486
      chlorides                  0.298421          0.377167     0.039315
      free sulfur dioxide       -0.283317         -0.353230     0.133437
      total sulfur dioxide      -0.329747         -0.414928     0.195218
      density                    0.459204          0.271193     0.096320
      pH                        -0.251814          0.260660    -0.328689
      sulphates                  0.300380          0.225476     0.057613
```

```
alcohol                      -0.095603          -0.038248      -0.010433
quality                      -0.077031          -0.265953       0.085706


                    residual sugar  chlorides  free sulfur dioxide  \
fixed acidity             -0.112319   0.298421            -0.283317
volatile acidity          -0.196702   0.377167            -0.353230
citric acid                0.142486   0.039315             0.133437
residual sugar             1.000000  -0.128902             0.403439
chlorides                 -0.128902   1.000000            -0.195042
free sulfur dioxide        0.403439  -0.195042             1.000000
total sulfur dioxide       0.495820  -0.279580             0.720934
density                    0.552498   0.362594             0.025717
pH                        -0.267050   0.044806            -0.145191
sulphates                 -0.185745   0.395332            -0.188489
alcohol                   -0.359706  -0.256861            -0.179838
quality                   -0.036825  -0.200886             0.055463


                    total sulfur dioxide   density        pH  sulphates  \
fixed acidity                  -0.329747  0.459204 -0.251814   0.300380
volatile acidity               -0.414928  0.271193  0.260660   0.225476
citric acid                     0.195218  0.096320 -0.328689   0.057613
residual sugar                  0.495820  0.552498 -0.267050  -0.185745
chlorides                      -0.279580  0.362594  0.044806   0.395332
free sulfur dioxide             0.720934  0.025717 -0.145191  -0.188489
total sulfur dioxide            1.000000  0.032395 -0.237687  -0.275381
density                         0.032395  1.000000  0.011920   0.259454
pH                             -0.237687  0.011920  1.000000   0.191248
sulphates                      -0.275381  0.259454  0.191248   1.000000
alcohol                        -0.265740 -0.686745  0.121002  -0.003261
quality                        -0.041385 -0.305858  0.019366   0.038729


                     alcohol   quality
fixed acidity       -0.095603 -0.077031
volatile acidity    -0.038248 -0.265953
citric acid         -0.010433  0.085706
residual sugar      -0.359706 -0.036825
chlorides           -0.256861 -0.200886
free sulfur dioxide -0.179838  0.055463
total sulfur dioxide -0.265740 -0.041385
density             -0.686745 -0.305858
pH                   0.121002  0.019366
sulphates           -0.003261  0.038729
alcohol              1.000000  0.444319
quality              0.444319  1.000000
```

```python
[13]: df.corr().quality.sort_values(ascending =False)
```

```
<ipython-input-13-faa9332a27f4>:1: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it will
default to False. Select only valid columns or specify the value of numeric_only
to silence this warning.
  df.corr().quality.sort_values(ascending =False)
```

```
[13]: quality                  1.000000
      alcohol                  0.444319
      citric acid              0.085706
      free sulfur dioxide      0.055463
      sulphates                0.038729
      pH                       0.019366
      residual sugar          -0.036825
      total sulfur dioxide    -0.041385
      fixed acidity           -0.077031
      chlorides               -0.200886
      volatile acidity        -0.265953
      density                 -0.305858
      Name: quality, dtype: float64
```

0.2   Visualization

0.2.1   Univariate Analysis

```
[14]: sns.distplot(df.density)
```
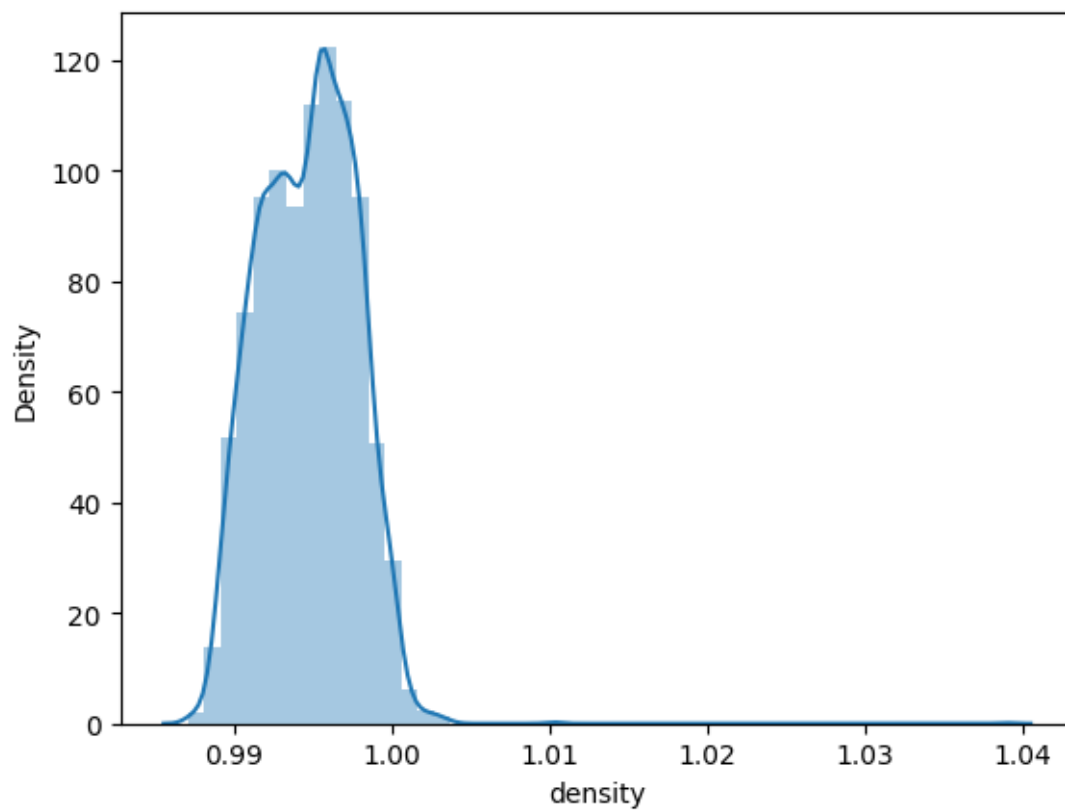
```
<ipython-input-14-d7414a6caaf8>:1: UserWarning:

`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

  sns.distplot(df.density)
```
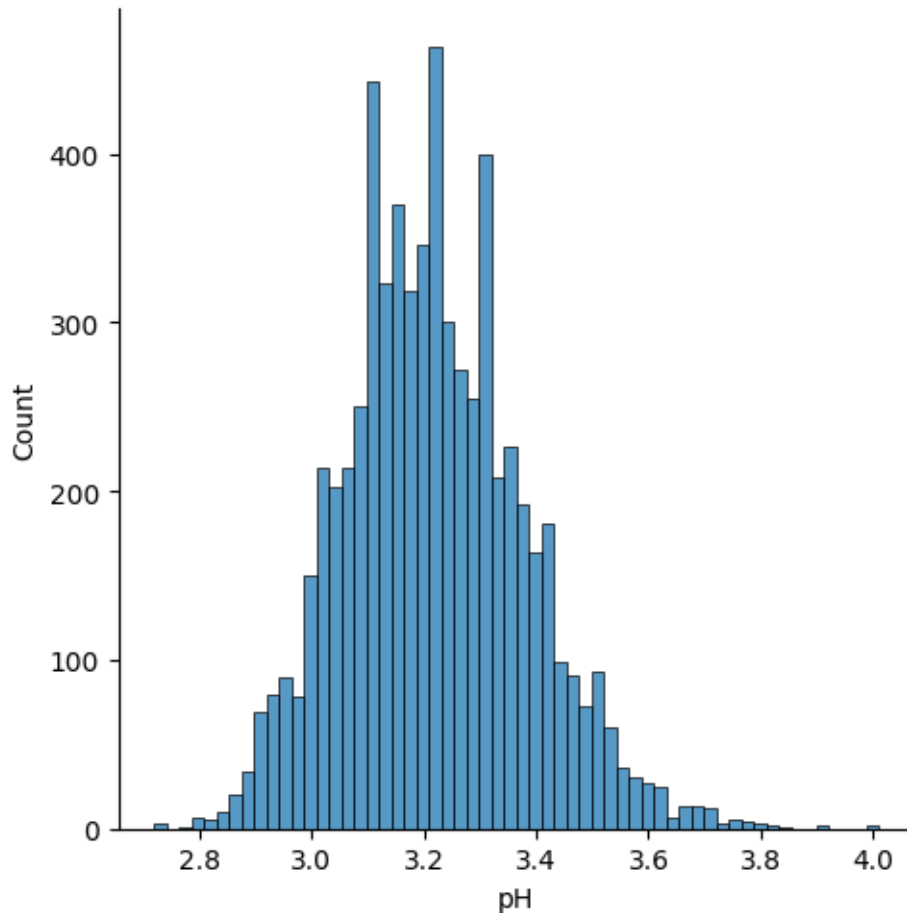
```
[14]: <Axes: xlabel='density', ylabel='Density'>
```

```
[15]: sns.displot(df.pH)
```

[15]: <seaborn.axisgrid.FacetGrid at 0x7ece8e5e7c40>

```
[68]: plt.pie(df.quality.value_counts(),[0,0,0,0.1,0.1,0.1],labels =␣
        ↪[5,6,7,4,8,3],autopct ='%1.1f%%',type = quality )
      plt.title('QUALITY')
      plt.show()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-68-c1163dbbc575> in <cell line: 1>()
----> 1 plt.pie(df.quality.value_counts(),[0,0,0,0.1,0.1,0.1],labels =␣
      ↪[5,6,7,4,8,3],autopct ='%1.1f%%',type = quality )
      2 plt.title('QUALITY')
      3 plt.show()

NameError: name 'quality' is not defined
```
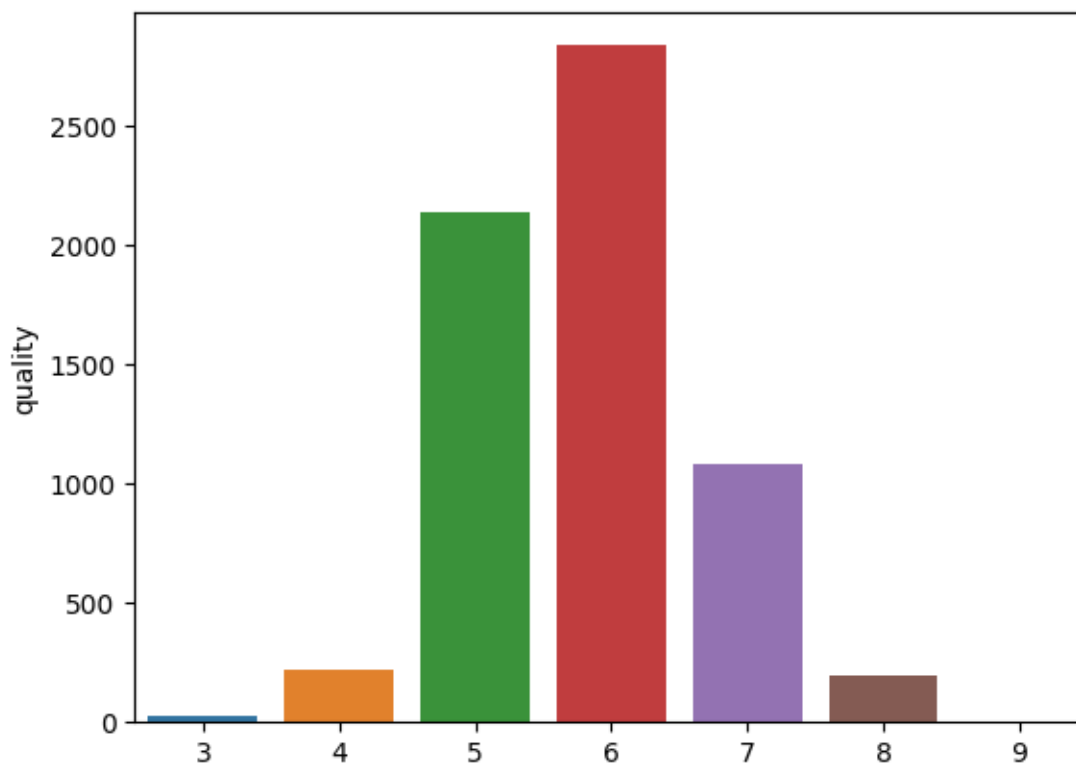
```
[17]: sns.barplot(x =df.quality.value_counts().index,y =df.quality.value_counts() )
```
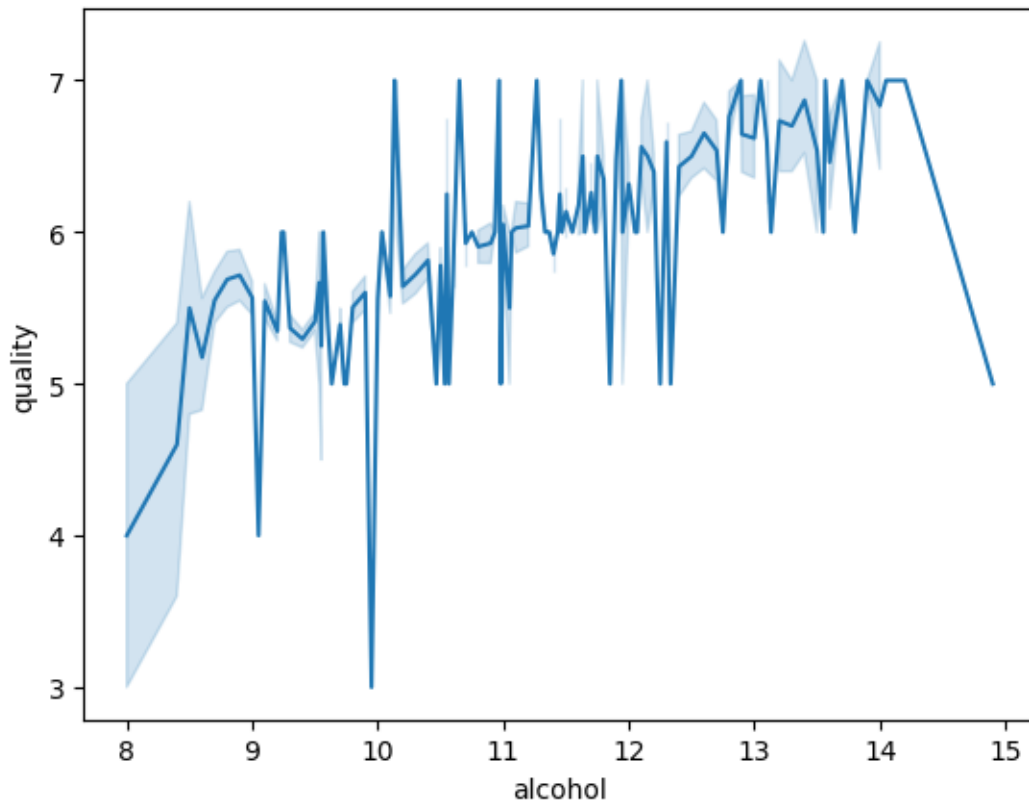
[17]: <Axes: ylabel='quality'>

0.2.2 Bivariate Analysis

```
[18]: sns.lineplot(x = df.alcohol,y=df.quality)
```
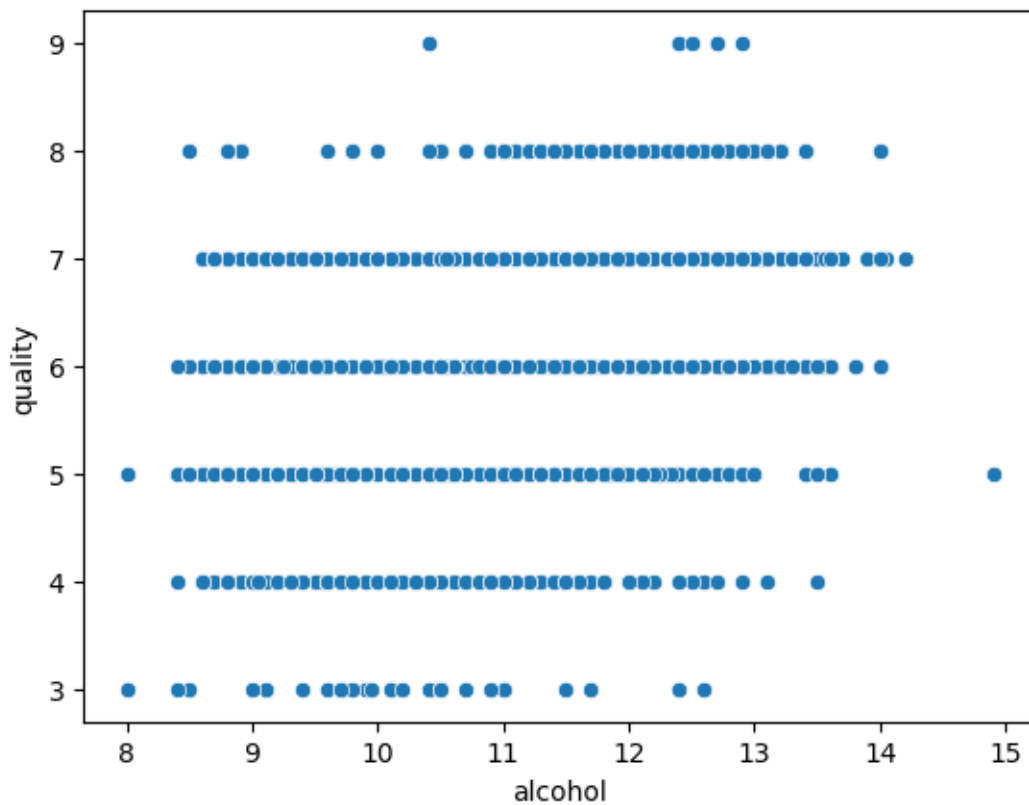
[18]: <Axes: xlabel='alcohol', ylabel='quality'>

```
[19]: sns.scatterplot(x = df.alcohol,y=df.quality)
```

```
[19]: <Axes: xlabel='alcohol', ylabel='quality'>
```

0.2.3 Multivariate Analysis

```
[20]: sns.pairplot(df)
```

```
[20]: <seaborn.axisgrid.PairGrid at 0x7ece8e68d0c0>
```
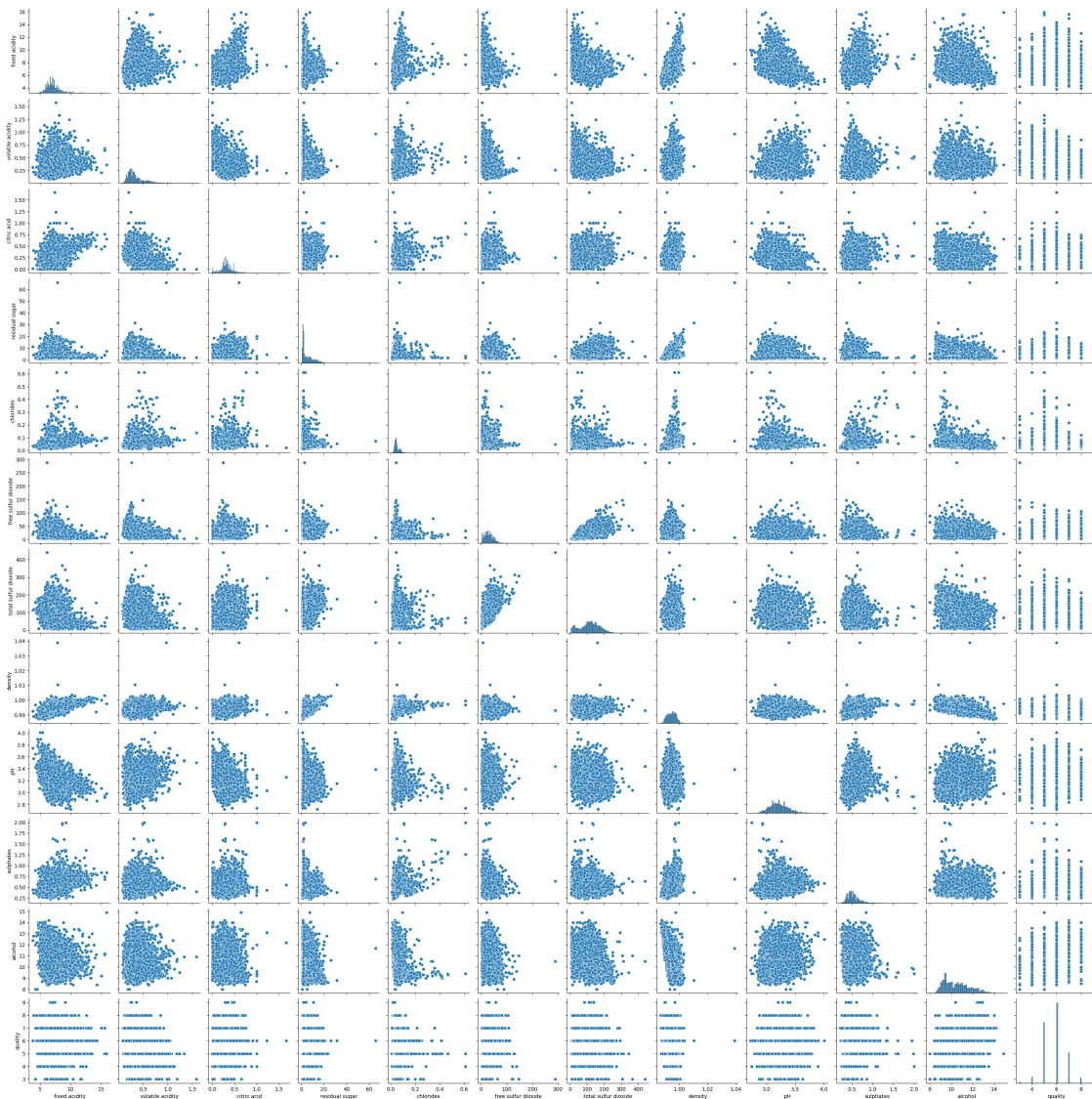
```
[21]: sns.heatmap(df.corr(),annot=True)
```

```
<ipython-input-21-8df7bcac526d>:1: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it will
default to False. Select only valid columns or specify the value of numeric_only
to silence this warning.
  sns.heatmap(df.corr(),annot=True)
```

```
[21]: <Axes: >
```
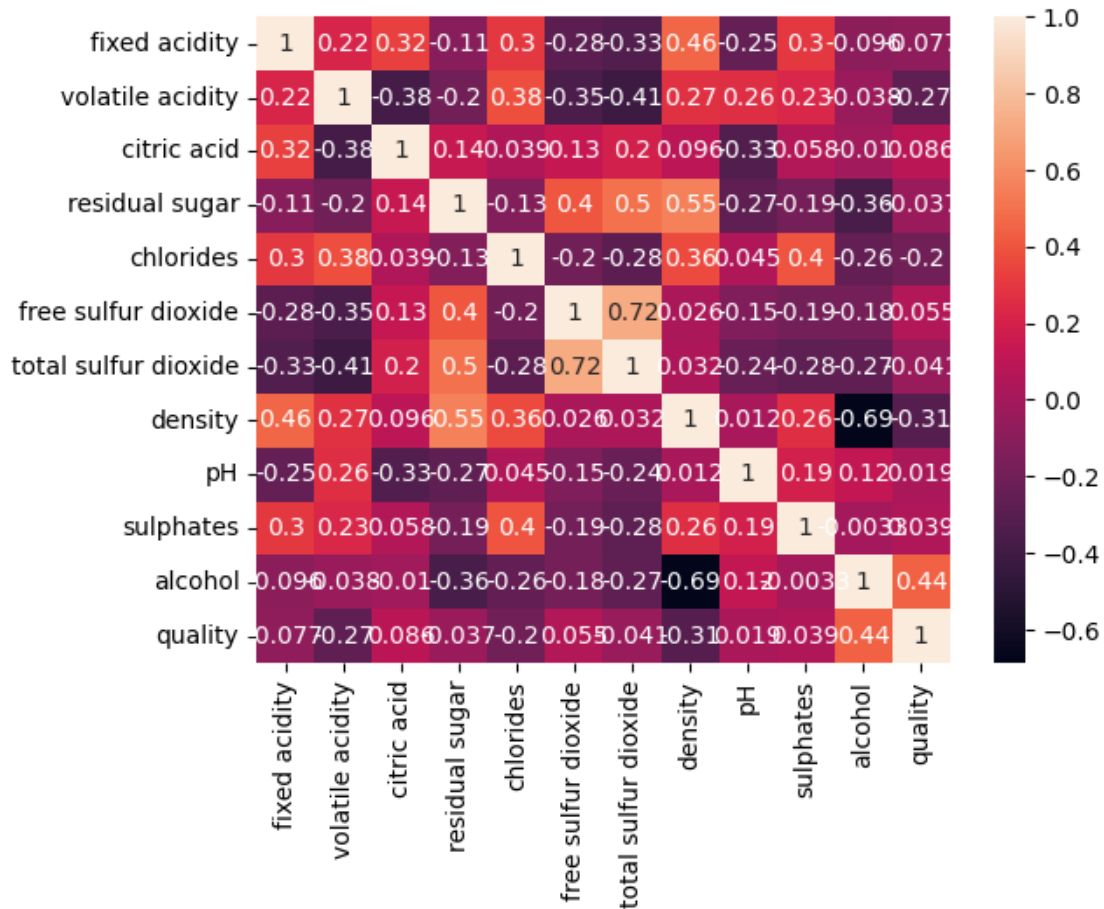
0.2.4 Outlier detection

```
[22]: df.head()
```

```
[22]:      type  fixed acidity  volatile acidity  citric acid  residual sugar  \
      0  white            7.0              0.27         0.36            20.7
      1  white            6.3              0.30         0.34             1.6
      2  white            8.1              0.28         0.40             6.9
      3  white            7.2              0.23         0.32             8.5
      4  white            7.2              0.23         0.32             8.5

         chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  \
      0      0.045                 45.0                 170.0   1.0010  3.00
      1      0.049                 14.0                 132.0   0.9940  3.30
      2      0.050                 30.0                  97.0   0.9951  3.26
      3      0.058                 47.0                 186.0   0.9956  3.19
      4      0.058                 47.0                 186.0   0.9956  3.19
```

```
     sulphates  alcohol  quality
0         0.45      8.8        6
1         0.49      9.5        6
2         0.44     10.1        6
3         0.40      9.9        6
4         0.40      9.9        6
```

[23]: `sns.boxplot(df.quality)`

[23]: `<Axes: >`
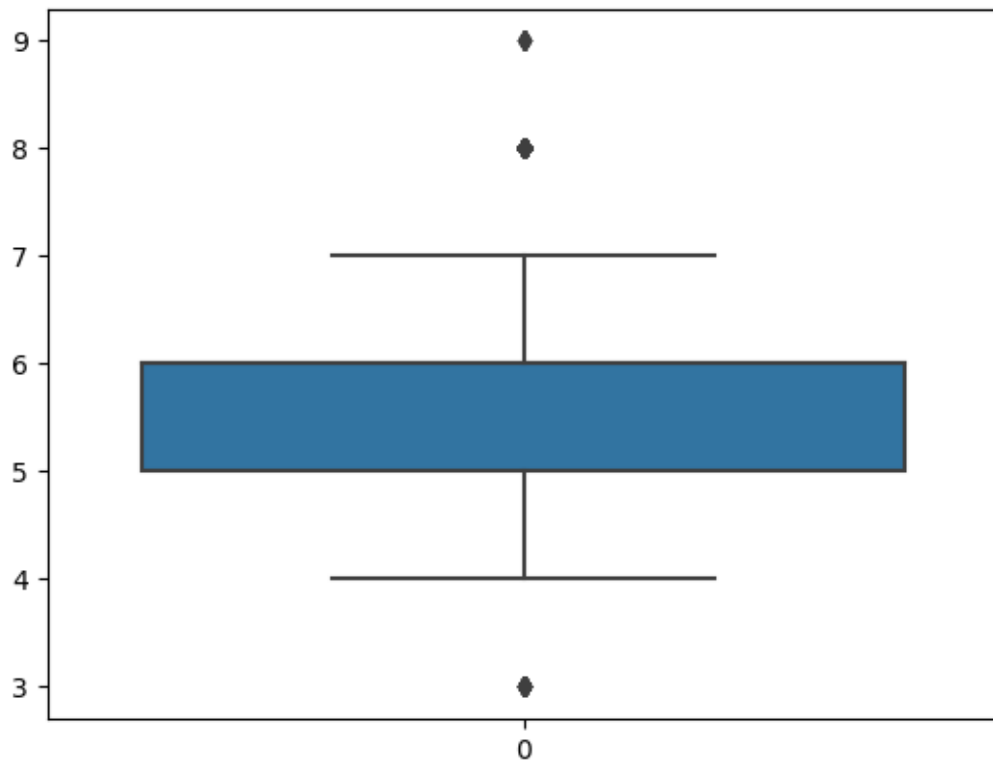
0.2.5 Outlier removal using IQR

[24]:
```
q1 = df.quality.quantile(0.25)
q3 = df.quality.quantile(0.75)
```

[25]: `IQR=q3-q1`

[26]:
```
upper_limit = q3+1.5*IQR
upper_limit
```

[26]: `7.5`

```
[27]: df.median()
```

<ipython-input-27-6d467abf240d>:1: FutureWarning: The default value of
numeric_only in DataFrame.median is deprecated. In a future version, it will
default to False. In addition, specifying 'numeric_only=None' is deprecated.
Select only valid columns or specify the value of numeric_only to silence this
warning.
  df.median()

```
[27]: fixed acidity           7.00000
      volatile acidity        0.29000
      citric acid             0.31000
      residual sugar          3.00000
      chlorides               0.04700
      free sulfur dioxide    29.00000
      total sulfur dioxide  118.00000
      density                 0.99489
      pH                      3.21000
      sulphates               0.51000
      alcohol                10.30000
      quality                 6.00000
      dtype: float64
```

```
[28]: df['quality']=np.where(df['quality']>upper_limit,6,df['quality'])
```

```
[29]: sns.boxplot(df.quality)
```

```
[29]: <Axes: >
```
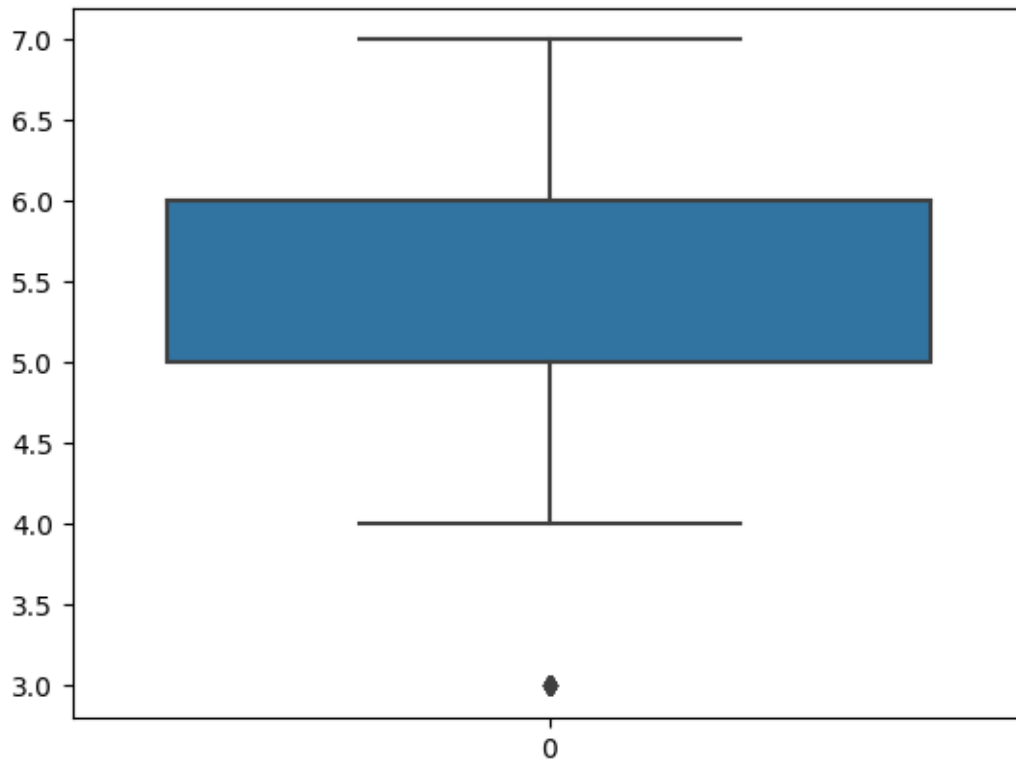
```
[30]: df.shape
```

```
[30]: (6497, 13)
```

0.2.6 Z-Score method

```
[31]: from scipy import stats
```

```
[32]: quality_zscore=stats.zscore(df.quality)
      quality_zscore
```

```
[32]: 0        0.311447
      1        0.311447
      2        0.311447
      3        0.311447
      4        0.311447
                  …
      6492    -0.968421
      6493     0.311447
      6494     0.311447
      6495    -0.968421
      6496     0.311447
```

16

```
Name: quality, Length: 6497, dtype: float64
```

[33]:
```
df_z=df[np.abs(quality_zscore)<=3]
```

[34]:
```
sns.boxplot(df_z.quality)
```

[34]: <Axes: >
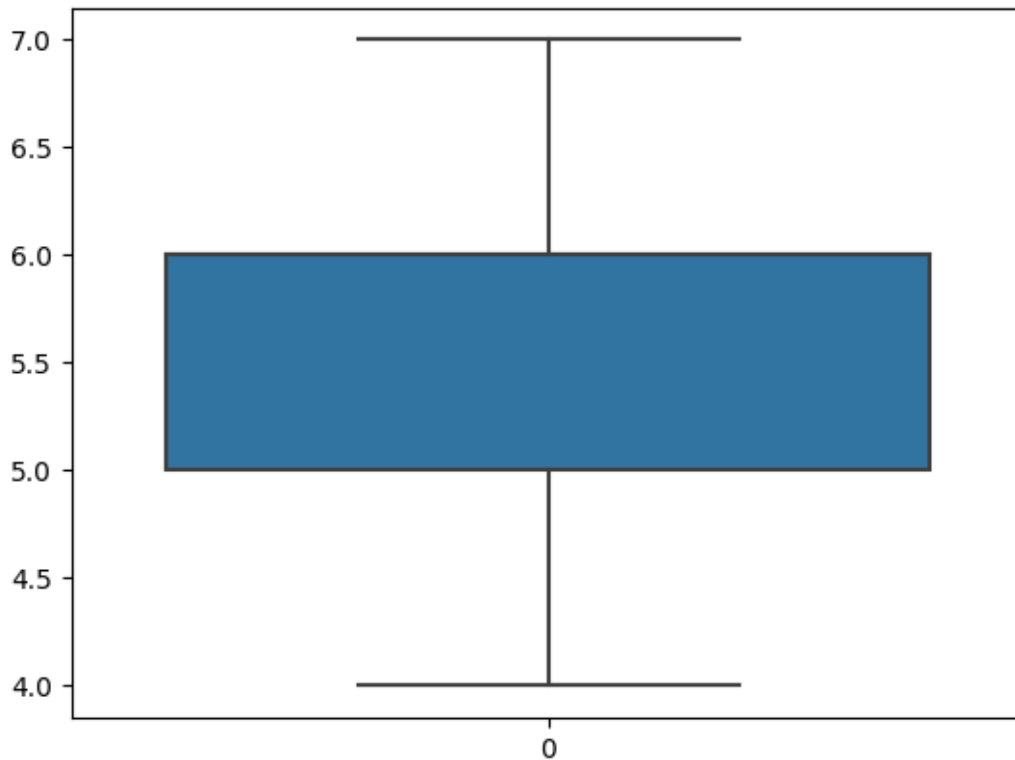
0.2.7 Label Encoding

[35]:
```python
from sklearn.preprocessing import LabelEncoder
```

[36]:
```python
le = LabelEncoder()
```

[37]:
```python
df.alcohol = le.fit_transform(df.alcohol)
df.quality = le.fit_transform(df.quality)
```

[38]:
```python
df.head()
```

[38]:

	type	fixed acidity	volatile acidity	citric acid	residual sugar	\
0	white	7.0	0.27	0.36	20.7	
1	white	6.3	0.30	0.34	1.6	

```
2   white               8.1                 0.28                0.40                6.9
3   white               7.2                 0.23                0.32                8.5
4   white               7.2                 0.23                0.32                8.5

   chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  \
0     0.045                 45.0                  170.0   1.0010  3.00
1     0.049                 14.0                  132.0   0.9940  3.30
2     0.050                 30.0                   97.0   0.9951  3.26
3     0.058                 47.0                  186.0   0.9956  3.19
4     0.058                 47.0                  186.0   0.9956  3.19

   sulphates  alcohol  quality
0       0.45        5        3
1       0.49       15        3
2       0.44       29        3
3       0.40       25        3
4       0.40       25        3
```

0.2.8 One Hot encoding

```python
[39]: df_main = pd.get_dummies(df,columns =['quality'])
      df_main.head()
```

```
[39]:    type  fixed acidity  volatile acidity  citric acid  residual sugar  \
0   white            7.0              0.27         0.36            20.7
1   white            6.3              0.30         0.34             1.6
2   white            8.1              0.28         0.40             6.9
3   white            7.2              0.23         0.32             8.5
4   white            7.2              0.23         0.32             8.5

   chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  \
0     0.045                 45.0                  170.0   1.0010  3.00
1     0.049                 14.0                  132.0   0.9940  3.30
2     0.050                 30.0                   97.0   0.9951  3.26
3     0.058                 47.0                  186.0   0.9956  3.19
4     0.058                 47.0                  186.0   0.9956  3.19

   sulphates  alcohol  quality_0  quality_1  quality_2  quality_3  quality_4
0       0.45        5          0          0          0          1          0
1       0.49       15          0          0          0          1          0
2       0.44       29          0          0          0          1          0
3       0.40       25          0          0          0          1          0
4       0.40       25          0          0          0          1          0
```

```python
[40]: df_main.corr()
```

```
<ipython-input-40-b764c75a6398>:1: FutureWarning: The default value of
```

```
numeric_only in DataFrame.corr is deprecated. In a future version, it will
default to False. Select only valid columns or specify the value of numeric_only
to silence this warning.
  df_main.corr()
```

[40]:

	fixed acidity	volatile acidity	citric acid
fixed acidity	1.000000	0.220172	0.323736
volatile acidity	0.220172	1.000000	-0.378061
citric acid	0.323736	-0.378061	1.000000
residual sugar	-0.112319	-0.196702	0.142486
chlorides	0.298421	0.377167	0.039315
free sulfur dioxide	-0.283317	-0.353230	0.133437
total sulfur dioxide	-0.329747	-0.414928	0.195218
density	0.459204	0.271193	0.096320
pH	-0.251814	0.260660	-0.328689
sulphates	0.300380	0.225476	0.057613
alcohol	-0.098798	-0.043591	-0.000616
quality_0	0.033473	0.073398	-0.017692
quality_1	0.010350	0.133305	-0.059262
quality_2	0.060872	0.213069	-0.053040
quality_3	-0.043064	-0.155956	0.036879
quality_4	-0.030183	-0.137636	0.049299

	residual sugar	chlorides	free sulfur dioxide
fixed acidity	-0.112319	0.298421	-0.283317
volatile acidity	-0.196702	0.377167	-0.353230
citric acid	0.142486	0.039315	0.133437
residual sugar	1.000000	-0.128902	0.403439
chlorides	-0.128902	1.000000	-0.195042
free sulfur dioxide	0.403439	-0.195042	1.000000
total sulfur dioxide	0.495820	-0.279580	0.720934
density	0.552498	0.362594	0.025717
pH	-0.267050	0.044806	-0.145191
sulphates	-0.185745	0.395332	-0.188489
alcohol	-0.344258	-0.260449	-0.172378
quality_0	-0.004357	0.040817	0.033354
quality_1	-0.050313	0.021742	-0.103325
quality_2	0.052973	0.172449	-0.011363
quality_3	0.018451	-0.073414	0.045233
quality_4	-0.066606	-0.137205	-0.002594

	total sulfur dioxide	density	pH	sulphates
fixed acidity	-0.329747	0.459204	-0.251814	0.300380
volatile acidity	-0.414928	0.271193	0.260660	0.225476
citric acid	0.195218	0.096320	-0.328689	0.057613
residual sugar	0.495820	0.552498	-0.267050	-0.185745
chlorides	-0.279580	0.362594	0.044806	0.395332

free sulfur dioxide	0.720934	0.025717	-0.145191	-0.188489
total sulfur dioxide	1.000000	0.032395	-0.237687	-0.275381
density	0.032395	1.000000	0.011920	0.259454
pH	-0.237687	0.011920	1.000000	0.191248
sulphates	-0.275381	0.259454	0.191248	1.000000
alcohol	-0.257704	-0.683388	0.110382	-0.012222
quality_0	0.007579	0.023791	0.016652	-0.011392
quality_1	-0.040397	0.008411	0.015269	-0.032747
quality_2	0.063129	0.269154	-0.027682	-0.022592
quality_3	-0.003292	-0.085370	-0.001131	-0.000818
quality_4	-0.057214	-0.233773	0.026076	0.047432

	alcohol	quality_0	quality_1	quality_2	quality_3 \
fixed acidity	-0.098798	0.033473	0.010350	0.060872	-0.043064
volatile acidity	-0.043591	0.073398	0.133305	0.213069	-0.155956
citric acid	-0.000616	-0.017692	-0.059262	-0.053040	0.036879
residual sugar	-0.344258	-0.004357	-0.050313	0.052973	0.018451
chlorides	-0.260449	0.040817	0.021742	0.172449	-0.073414
free sulfur dioxide	-0.172378	0.033354	-0.103325	-0.011363	0.045233
total sulfur dioxide	-0.257704	0.007579	-0.040397	0.063129	-0.003292
density	-0.683388	0.023791	0.008411	0.269154	-0.085370
pH	0.110382	0.016652	0.015269	-0.027682	-0.001131
sulphates	-0.012222	-0.011392	-0.032747	-0.022592	-0.000818
alcohol	1.000000	-0.013644	-0.049315	-0.386320	0.131039
quality_0	-0.013644	1.000000	-0.012631	-0.047700	-0.063752
quality_1	-0.049315	-0.012631	1.000000	-0.129874	-0.173578
quality_2	-0.386320	-0.047700	-0.129874	1.000000	-0.655529
quality_3	0.131039	-0.063752	-0.173578	-0.655529	1.000000
quality_4	0.338339	-0.030395	-0.082757	-0.312537	-0.417708

	quality_4
fixed acidity	-0.030183
volatile acidity	-0.137636
citric acid	0.049299
residual sugar	-0.066606
chlorides	-0.137205
free sulfur dioxide	-0.002594
total sulfur dioxide	-0.057214
density	-0.233773
pH	0.026076
sulphates	0.047432
alcohol	0.338339
quality_0	-0.030395
quality_1	-0.082757
quality_2	-0.312537
quality_3	-0.417708
quality_4	1.000000

```
[41]: plt.figure(figsize=(12,10))
      sns.heatmap(df_main.corr(),annot =True)
```

<ipython-input-41-65b4d4cae710>:2: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it will
default to False. Select only valid columns or specify the value of numeric_only
to silence this warning.
    sns.heatmap(df_main.corr(),annot =True)

[41]: <Axes: >
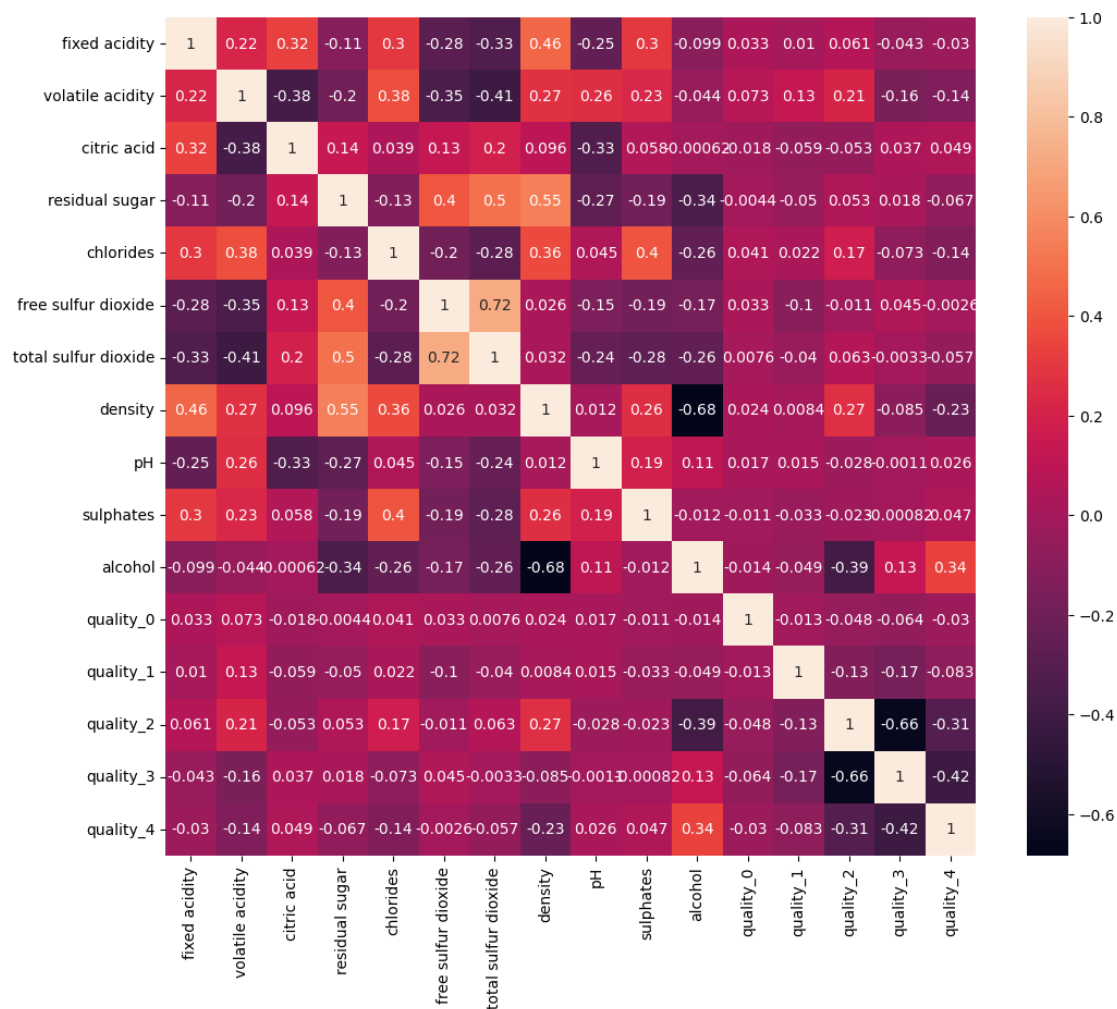

```
[42]: df_main.corr().alcohol.sort_values(ascending=False)
```

<ipython-input-42-3bc757a85a29>:1: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it will
default to False. Select only valid columns or specify the value of numeric_only

```
    to silence this warning.
      df_main.corr().alcohol.sort_values(ascending=False)
```

[42]:
```
alcohol                  1.000000
quality_4                0.338339
quality_3                0.131039
pH                       0.110382
citric acid             -0.000616
sulphates               -0.012222
quality_0               -0.013644
volatile acidity        -0.043591
quality_1               -0.049315
fixed acidity           -0.098798
free sulfur dioxide     -0.172378
total sulfur dioxide    -0.257704
chlorides               -0.260449
residual sugar          -0.344258
quality_2               -0.386320
density                 -0.683388
Name: alcohol, dtype: float64
```

[43]: `df_main.head()`

[43]:

	type	fixed acidity	volatile acidity	citric acid	residual sugar \
0	white	7.0	0.27	0.36	20.7
1	white	6.3	0.30	0.34	1.6
2	white	8.1	0.28	0.40	6.9
3	white	7.2	0.23	0.32	8.5
4	white	7.2	0.23	0.32	8.5

	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH \
0	0.045	45.0	170.0	1.0010	3.00
1	0.049	14.0	132.0	0.9940	3.30
2	0.050	30.0	97.0	0.9951	3.26
3	0.058	47.0	186.0	0.9956	3.19
4	0.058	47.0	186.0	0.9956	3.19

	sulphates	alcohol	quality_0	quality_1	quality_2	quality_3	quality_4
0	0.45	5	0	0	0	1	0
1	0.49	15	0	0	0	1	0
2	0.44	29	0	0	0	1	0
3	0.40	25	0	0	0	1	0
4	0.40	25	0	0	0	1	0

0.2.9 x and y split

```
[44]: x =df.drop(columns =['quality'],axis =1)
```

```
[45]: x.head()
```

```
[45]:     type  fixed acidity  volatile acidity  citric acid  residual sugar  \
      0  white            7.0              0.27         0.36            20.7
      1  white            6.3              0.30         0.34             1.6
      2  white            8.1              0.28         0.40             6.9
      3  white            7.2              0.23         0.32             8.5
      4  white            7.2              0.23         0.32             8.5

         chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  \
      0      0.045                 45.0                 170.0   1.0010  3.00
      1      0.049                 14.0                 132.0   0.9940  3.30
      2      0.050                 30.0                  97.0   0.9951  3.26
      3      0.058                 47.0                 186.0   0.9956  3.19
      4      0.058                 47.0                 186.0   0.9956  3.19

         sulphates  alcohol
      0       0.45        5
      1       0.49       15
      2       0.44       29
      3       0.40       25
      4       0.40       25
```

```
[46]: y =df.quality
      y.head()
```

```
[46]: 0    3
      1    3
      2    3
      3    3
      4    3
      Name: quality, dtype: int64
```

```
[47]: from sklearn.preprocessing import MinMaxScaler
      scale =MinMaxScaler()
```

0.2.10 Model Building

```
[71]: from sklearn.linear_model import LogisticRegression
      model = LogisticRegression()
```

0.2.11 Evaluation for classification model

```
[72]: from sklearn.metrics import accuracy_score, confusion_matrix,␣
       ↪classification_report, roc_auc_score, roc_curve
```

0.2.12 Random Forest Classifier

```
[73]: from sklearn.ensemble import RandomForestClassifier
      model =RandomForestClassifier(criterion='entropy')
```

```
[74]: df.head()
```

```
[74]:      type  fixed acidity  volatile acidity  citric acid  residual sugar  \
      0  white            7.0              0.27         0.36            20.7
      1  white            6.3              0.30         0.34             1.6
      2  white            8.1              0.28         0.40             6.9
      3  white            7.2              0.23         0.32             8.5
      4  white            7.2              0.23         0.32             8.5

         chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  \
      0      0.045                 45.0                 170.0   1.0010  3.00
      1      0.049                 14.0                 132.0   0.9940  3.30
      2      0.050                 30.0                  97.0   0.9951  3.26
      3      0.058                 47.0                 186.0   0.9956  3.19
      4      0.058                 47.0                 186.0   0.9956  3.19

         sulphates  alcohol  quality
      0       0.45        5        3
      1       0.49       15        3
      2       0.44       29        3
      3       0.40       25        3
      4       0.40       25        3
```

```
[77]: from sklearn.linear_model import LinearRegression
      lr = LinearRegression()
```