dataMining02-data_exploration-wines

October 29, 2019

1 From UCI Machine Learning Repository

1.1 Wine Quality Dataset

1.1.1 Read data from archive.

In this case, it is a csv with header In this case, it is a csv with header, separator is ';' The download url is http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv

Use the read_csv() method of pandas dataframe https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

Use `df` as the dataframe name

In this dataset the column names are already included in the .csv file

1.1.2 Show column names

Use the `columns` attribute of pandas on `df`

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

1.1.3 Show portion of data

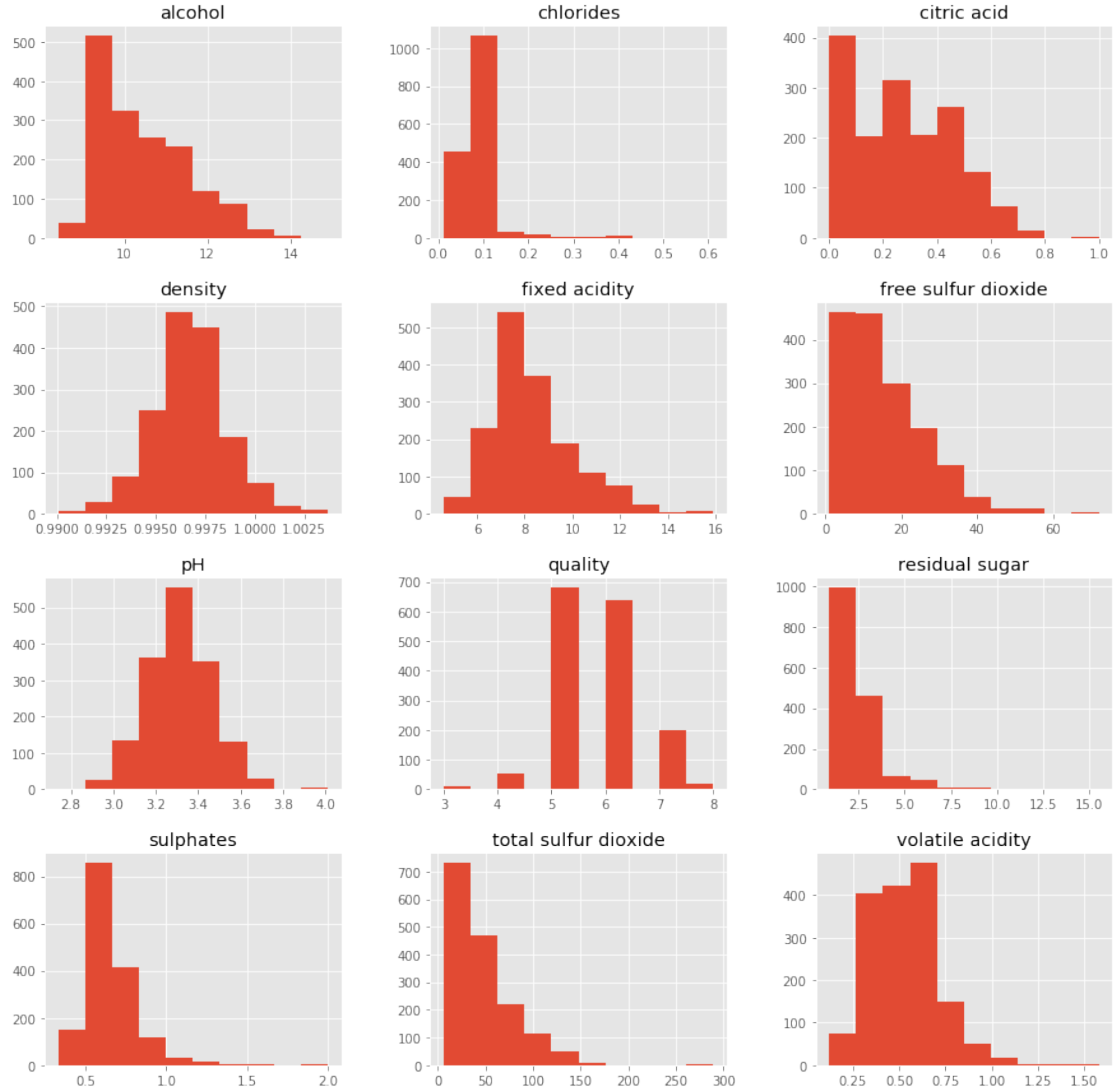
Use the `head` method of pandas dataframe

```
[4]:    fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
     0            7.4              0.70         0.00             1.9      0.076
     1            7.8              0.88         0.00             2.6      0.098
     2            7.8              0.76         0.04             2.3      0.092
     3           11.2              0.28         0.56             1.9      0.075
     4            7.4              0.70         0.00             1.9      0.076

        free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
     0                 11.0                  34.0   0.9978  3.51       0.56
     1                 25.0                  67.0   0.9968  3.20       0.68
```

```
2                 15.0                    54.0   0.9970  3.26        0.65
3                 17.0                    60.0   0.9980  3.16        0.58
4                 11.0                    34.0   0.9978  3.51        0.56

   alcohol  quality
0      9.4        5
1      9.8        5
2      9.8        5
3      9.8        6
4      9.4        5
```

1.1.4 Show histograms for all numeric values

Use the `DataFrame.hist` method of Pandas. You can set the `figsize` parameter to adjust size



1.1.5 Show synthetic description

Use the describe method of Pandas

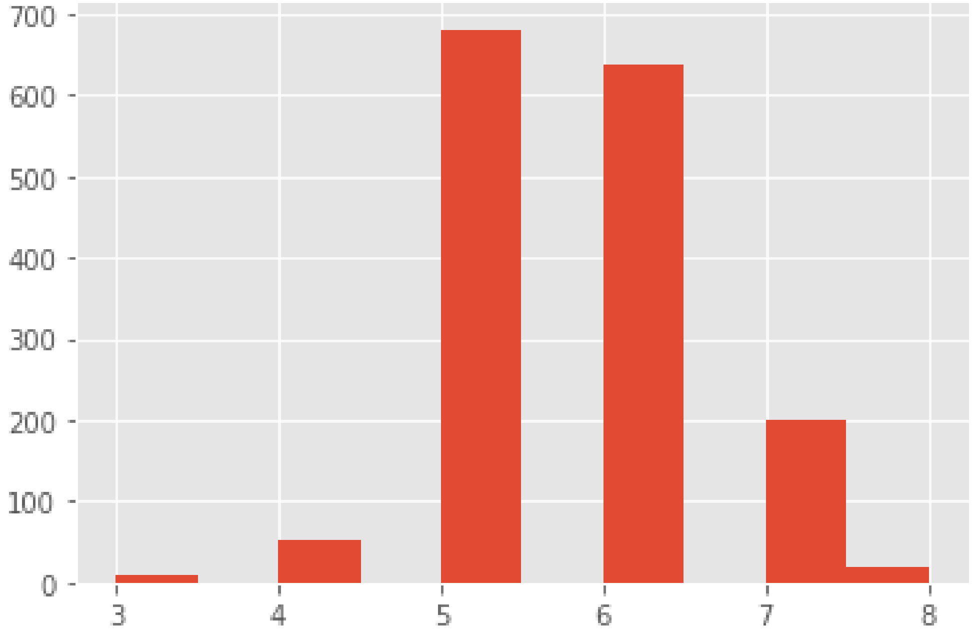
```
[6]:        fixed acidity  volatile acidity  citric acid  residual sugar  \
     count    1599.000000       1599.000000  1599.000000     1599.000000
     mean        8.319637          0.527821     0.270976        2.538806
     std         1.741096          0.179060     0.194801        1.409928
     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
```

Quality is the target class in this dataset. The **describe** method of pandas dataframes gives a short summary

```
[7]: count    1599.000000
     mean        5.636023
     std         0.807569
     min         3.000000
     25%         5.000000
     50%         6.000000
     75%         6.000000
     max         8.000000
     Name: quality, dtype: float64
```

1.1.6 Plot an histogram for "quality"

Use the `hist` method of `matplotlib.pyplot` applied to the `quality` column of `df`


700
600
500
400
300
200
100
0
3
4
5
6
7
8