Pattern Recognition – Project 3: K-Nearest Neighbors

-Prathamesh Patil (025910428)

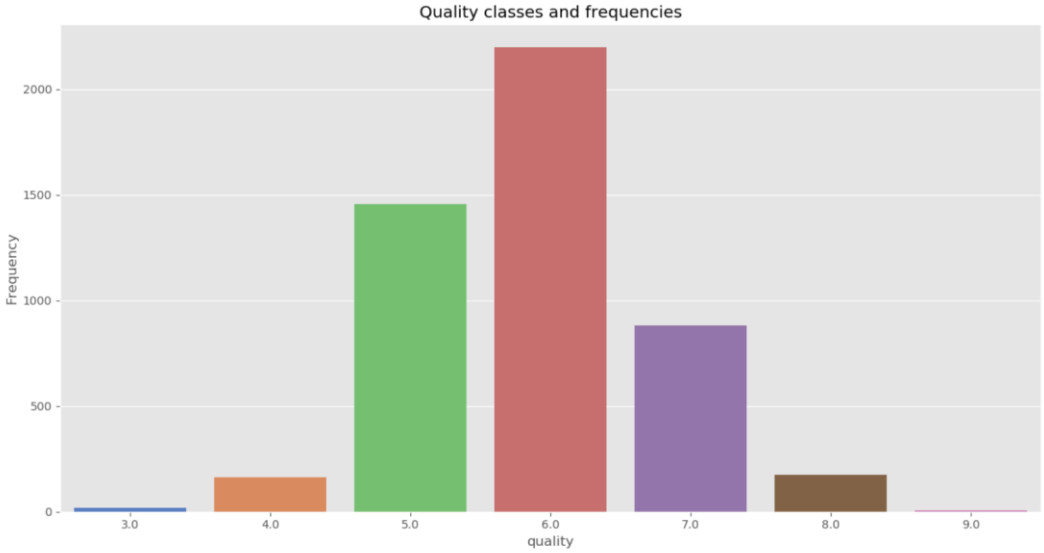
1) Loading and analyzing data:

Data statistics =>

```
Data statistics=>       fixed acidity  volatile acidity  citric acid  residual sugar  \
count    4901.000000       4901.000000  4901.000000     4901.000000
mean        6.853101          0.278199     0.354468        6.389196
std         0.849422          0.100845     1.428792        5.072222
min         0.000000          0.000000     0.000000        0.000000
25%         6.300000          0.210000     0.270000        1.700000
50%         6.800000          0.260000     0.320000        5.200000
75%         7.300000          0.320000     0.390000        9.900000
max        14.200000          1.100000   100.000000       65.800000

         chlorides  free sulfur dioxide  total sulfur dioxide      density  \
count  4901.000000          4901.000000           4901.000000  4901.000000
mean      0.045751            35.298715            138.306162     0.993621
std       0.021861            17.020541             42.577100     0.020300
min       0.000000             0.000000              0.000000     0.000000
25%       0.036000            23.000000            108.000000     0.991720
50%       0.043000            34.000000            134.000000     0.993730
75%       0.050000            46.000000            167.000000     0.996100
max       0.346000           289.000000            440.000000     1.038980

               pH    sulphates      alcohol      quality
count  4901.000000  4901.000000  4901.000000  4900.000000
mean      3.186982     0.489618    10.510178     5.877755
std       0.164122     0.114536     1.248521     0.892437
min       0.000000     0.000000     0.000000     0.000000
25%       3.090000     0.410000     9.500000     5.000000
50%       3.180000     0.470000    10.400000     6.000000
75%       3.280000     0.550000    11.400000     6.000000
max       3.820000     1.080000    14.200000    11.000000
```

```
Missing Values =>
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                 1
dtype: int64


Classes and their frequencies=>
6.0      2198
5.0      1457
7.0       880
8.0       175
4.0       163
3.0        20
9.0         5
11.0        1
0.0         1
NaN         1
Name: quality, dtype: int64
```

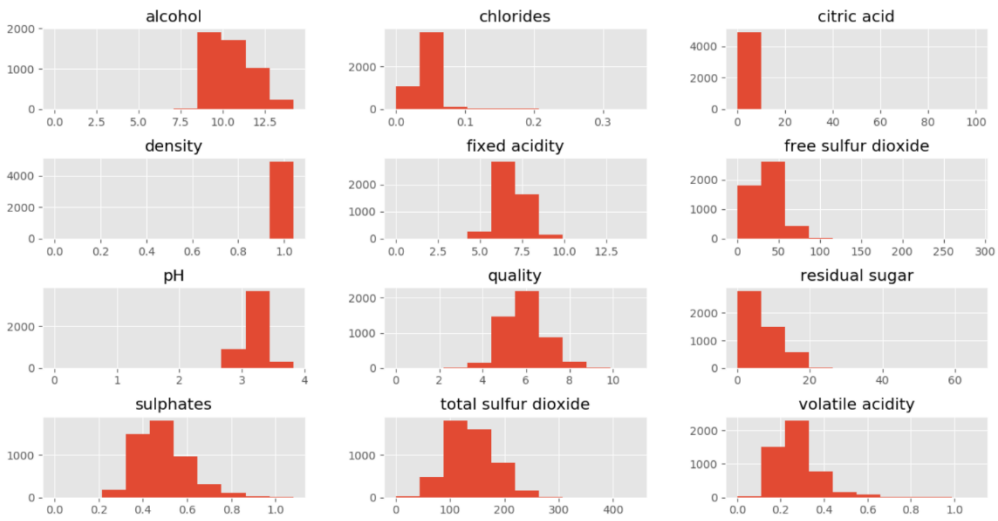
Class value frequencies before removing anomalies=>



2) Data Preprocessing:

```
(Quality)Classes and their frequencies after removing anomalies=>
6.0    2198
5.0    1457
7.0     880
8.0     175
4.0     163
3.0      20
9.0       6
Name: quality, dtype: int64
```

Distribution of all features with their frequencies =>

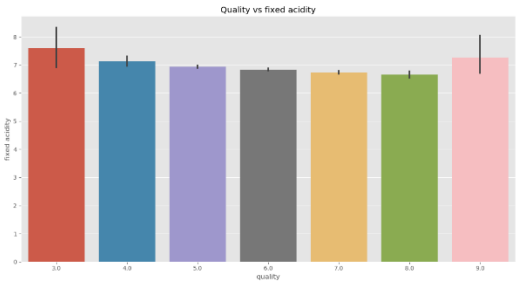


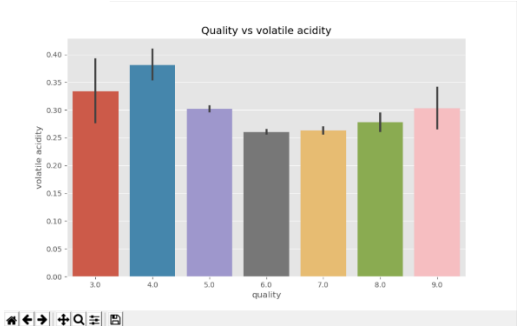
- We can notice that range of predictor variables: 'volatile acidity' ranges from 0.0 to 1.0 and 'free sulfur dioxide' ranges from 0 to ~120. KNN algorithm which is based on distances between data points may thus focus unfairly on variables with larger range. Thus we scale our data so that features that have different units/scale are rescaled for KNN to perform optimally.

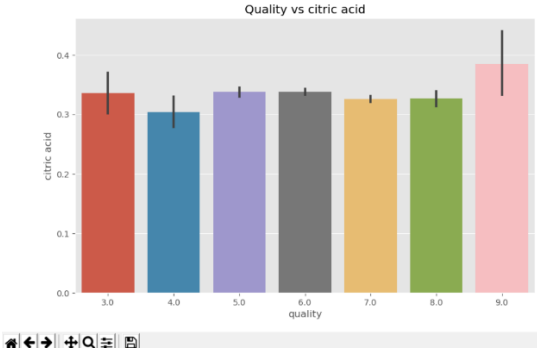
Feature Correlations =>

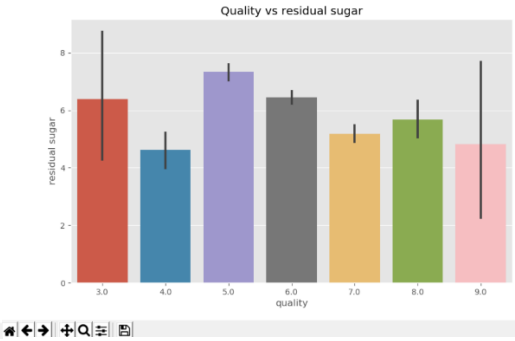
- The correlation coefficient ranges from –1 to 1. When it is close to 1, it means that there is a strong positive correlation.
- And from the heat map we can infer that as compared to other features, quality has maximum correlation with alcohol. Which implies when value of ‘alcohol’ increases, ‘quality’ increases with it.
- Similarly when correlation value is negative, it would imply that when ‘density’ increases, ‘quality’ decreases with it.
- Finally, value of correlation closer to 0 implies there is not much effect of that feature on the quality of the wine. For example – correlation between ‘sulphates’ and ‘quality’ = -0.02

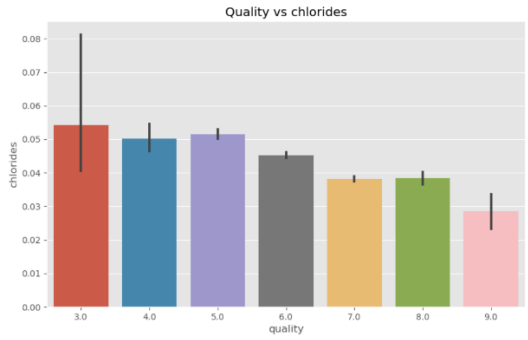
Individual barplots to depict relationships =>

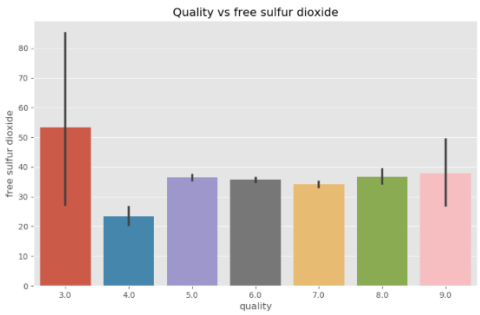


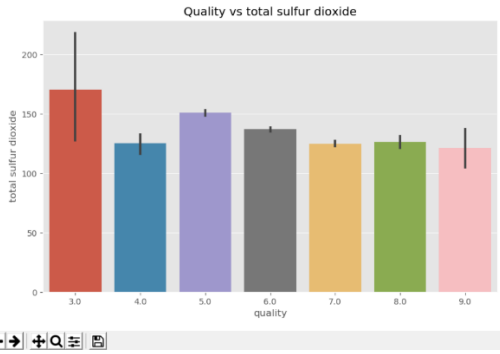


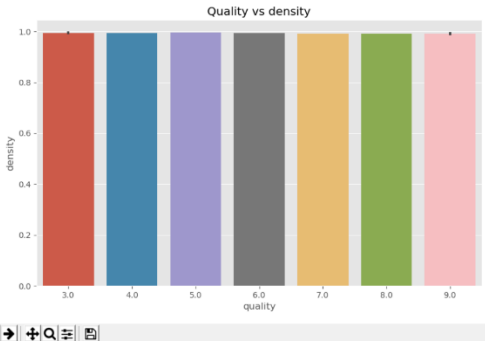


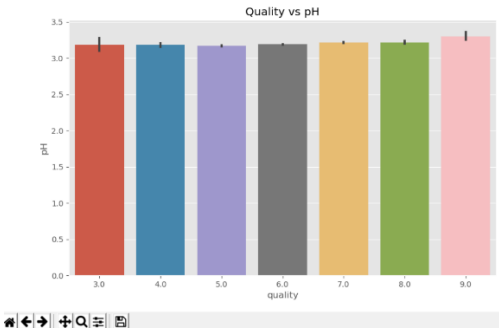


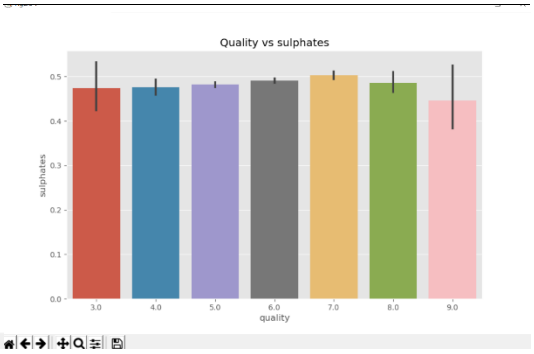


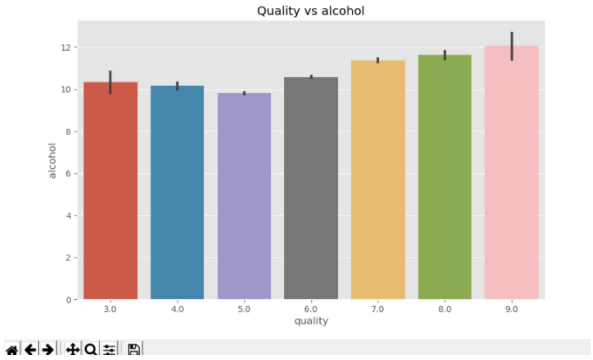












3) Applying KNN, cross validation and hyper-tuning parameters

- From this we can observe that KNN algorithm has produced maximum accuracy= 56% when k= 1

```
KNN model trained and tested with Accuracy =  63.87755102040816 at k =  1


Cross validation with number of groups =  5

Cross validation scores => [0.51020408 0.5255102  0.46632653 0.53571429 0.53217569]

cv_scores mean:0.5139861583039752


Hyper tuning parameters with number of groups for cross validation =  5  and Range for optimizing k => [ 1 , 25 ]

Best Param => {'n_neighbors': 1}

Best Score => 0.56766296304017

Process finished with exit code 0
```