

## Importing the needed packages

```
In [1]: import numpy as np  
import pandas as pd
```

```
In [2]: import matplotlib.pyplot as plt  
import seaborn as sns  
from sklearn.preprocessing import StandardScaler,LabelEncoder  
from sklearn.model_selection import train_test_split,GridSearchCV  
from sklearn.linear_model import LinearRegression  
from sklearn.metrics import r2_score,mean_squared_error,roc_auc_score  
from sklearn.linear_model import LogisticRegression  
from sklearn import metrics  
from sklearn.metrics import accuracy_score,recall_score,precision_score,auc  
from sklearn.metrics import classification_report,confusion_matrix,precision_recall_c  
from sklearn.tree import DecisionTreeClassifier  
from sklearn.ensemble import RandomForestClassifier  
from sklearn.neighbors import KNeighborsClassifier  
from sklearn.model_selection import cross_val_score  
from sklearn.naive_bayes import GaussianNB  
from sklearn.svm import SVC  
import warnings  
warnings.filterwarnings('ignore')
```

## Loading the dataset

```
In [3]: ds=pd.read_csv('QualityPrediction.csv')  
ds
```

Out[3]:

	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
...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0

1599 rows × 12 columns

In [4]: `ds.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

```

## Understanding the distribution of target variable

In [5]: `ds['quality'].value_counts()`

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

```
In [6]: ds['quality'].value_counts(normalize=True)
```

```
Out[6]: 5    0.425891
        6    0.398999
        7    0.124453
        4    0.033146
        8    0.011257
        3    0.006254
Name: quality, dtype: float64
```

- There are two wine qualities majorly- 5 and 6. Others are 3,4,7 and 8.
- Understanding the distribution of all values

```
In [7]: ds.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	
<b>count</b>	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	15
<b>mean</b>	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	
<b>std</b>	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	
<b>min</b>	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	
<b>25%</b>	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	
<b>50%</b>	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	
<b>75%</b>	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	
<b>max</b>	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	

- For a normal distribution, mean, median, mode are same

```
In [8]: sk_data=pd.DataFrame()
sk_data['Features']=ds.columns
sk_data['Mean']=[ds[i].mean() for i in ds]
sk_data['Median']=[ds[i].median() for i in ds]
sk_data['mode']=[ds[i].mode()[0] for i in ds]
sk_data['Mean_median_diff_%']=[(abs(ds[i].median()-ds[i].mean())*100)/ds[i].mean() for i in ds]
```

Out[8]:

	Features	Mean	Median	mode	Mean_median_diff_%
<b>0</b>	fixed acidity	8.319637	7.90000	7.2000	5.043937
<b>1</b>	volatile acidity	0.527821	0.52000	0.6000	1.481661
<b>2</b>	citric acid	0.270976	0.26000	0.0000	4.050405
<b>3</b>	residual sugar	2.538806	2.20000	2.0000	13.345075
<b>4</b>	chlorides	0.087467	0.07900	0.0800	9.679749
<b>5</b>	free sulfur dioxide	15.874922	14.00000	6.0000	11.810589
<b>6</b>	total sulfur dioxide	46.467792	38.00000	28.0000	18.222928
<b>7</b>	density	0.996747	0.99675	0.9972	0.000333
<b>8</b>	pH	3.311113	3.31000	3.3000	0.033620
<b>9</b>	sulphates	0.658149	0.62000	0.6000	5.796385
<b>10</b>	alcohol	10.422983	10.20000	9.5000	2.139341
<b>11</b>	quality	5.636023	6.00000	5.0000	6.458056

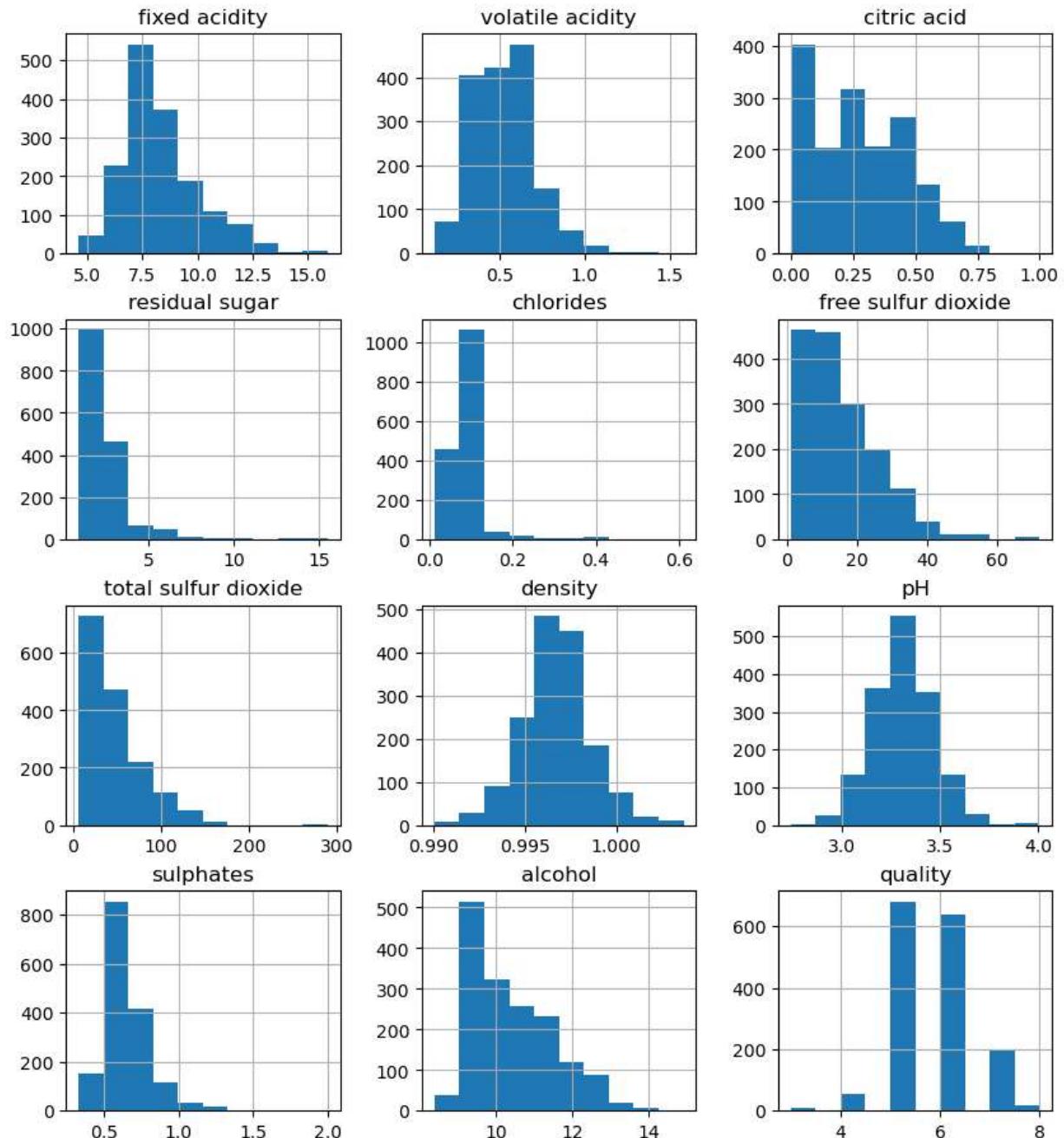
## Visualization

In [ ]:

```
#sns.pairplot(ds, height = 4, aspect = 1, kind='reg',diag_kind = 'kde')
#plt.show()
```

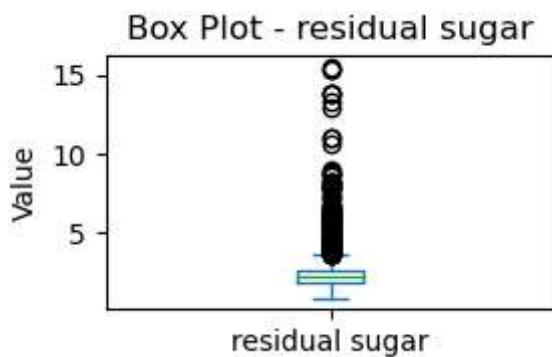
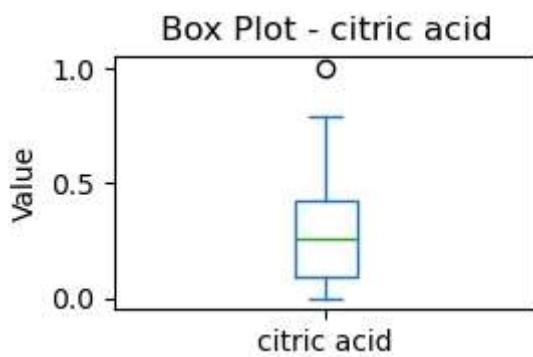
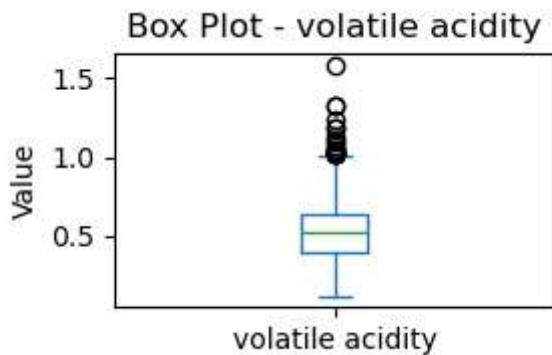
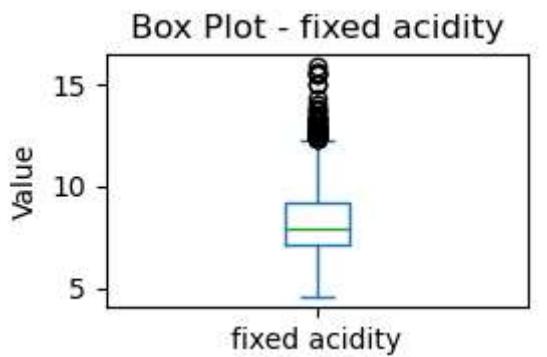
```
ds.hist(bins=10, figsize=(10,11))
plt.suptitle("Data Distribution of all the columns")
plt.show()
```

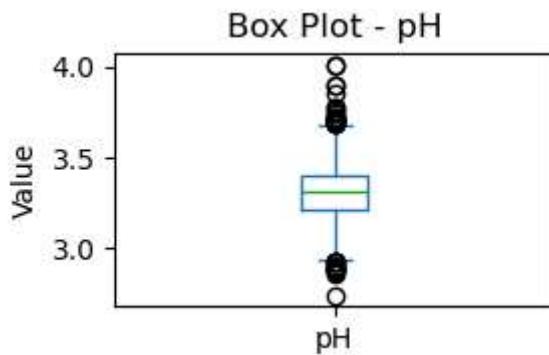
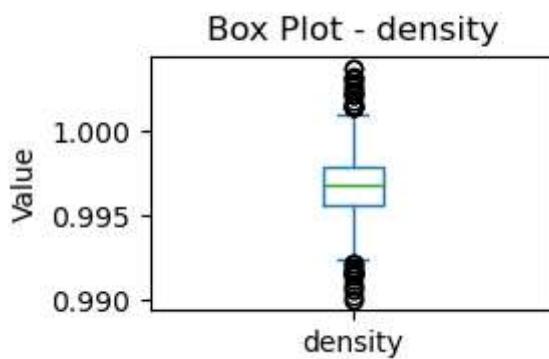
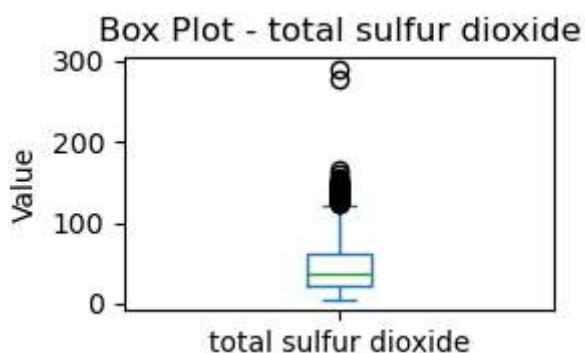
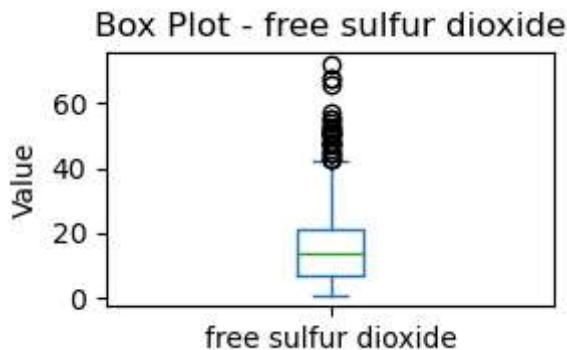
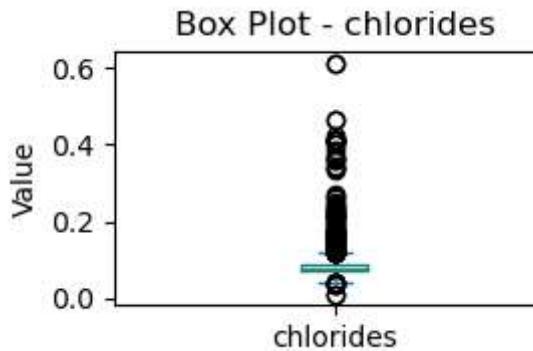
## Data Distribution of all the columns

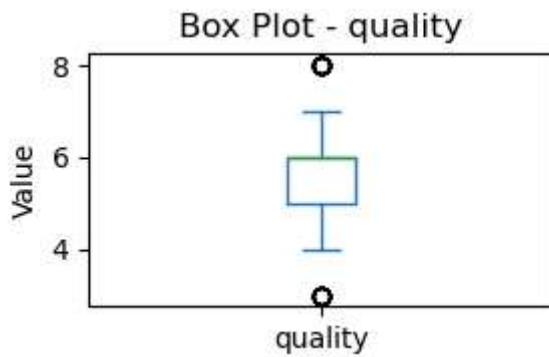
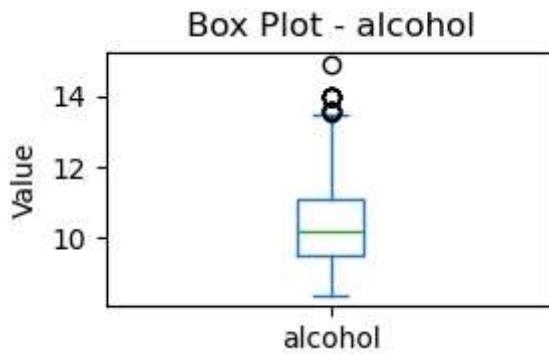
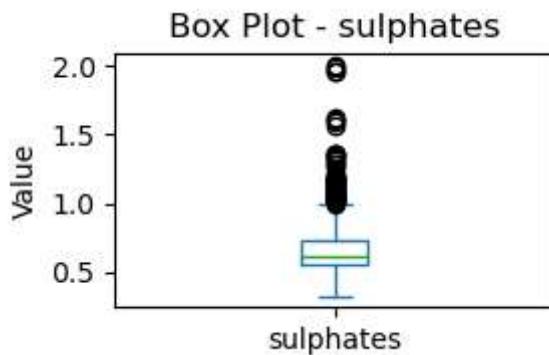


- Most columns are following normal distribution.
- The columns that are skewed are 'residual sugar','chlorides','free sulfur dioxide','total sulfur dioxide' as per the histogram

```
In [11]: # Plot individual box plots for each column
for col in ds.columns:
    plt.figure(figsize=(3, 2))
    ds[col].plot(kind='box')
    plt.title(f"Box Plot - {col}")
    plt.ylabel("Value")
    plt.tight_layout()
    plt.show()
```

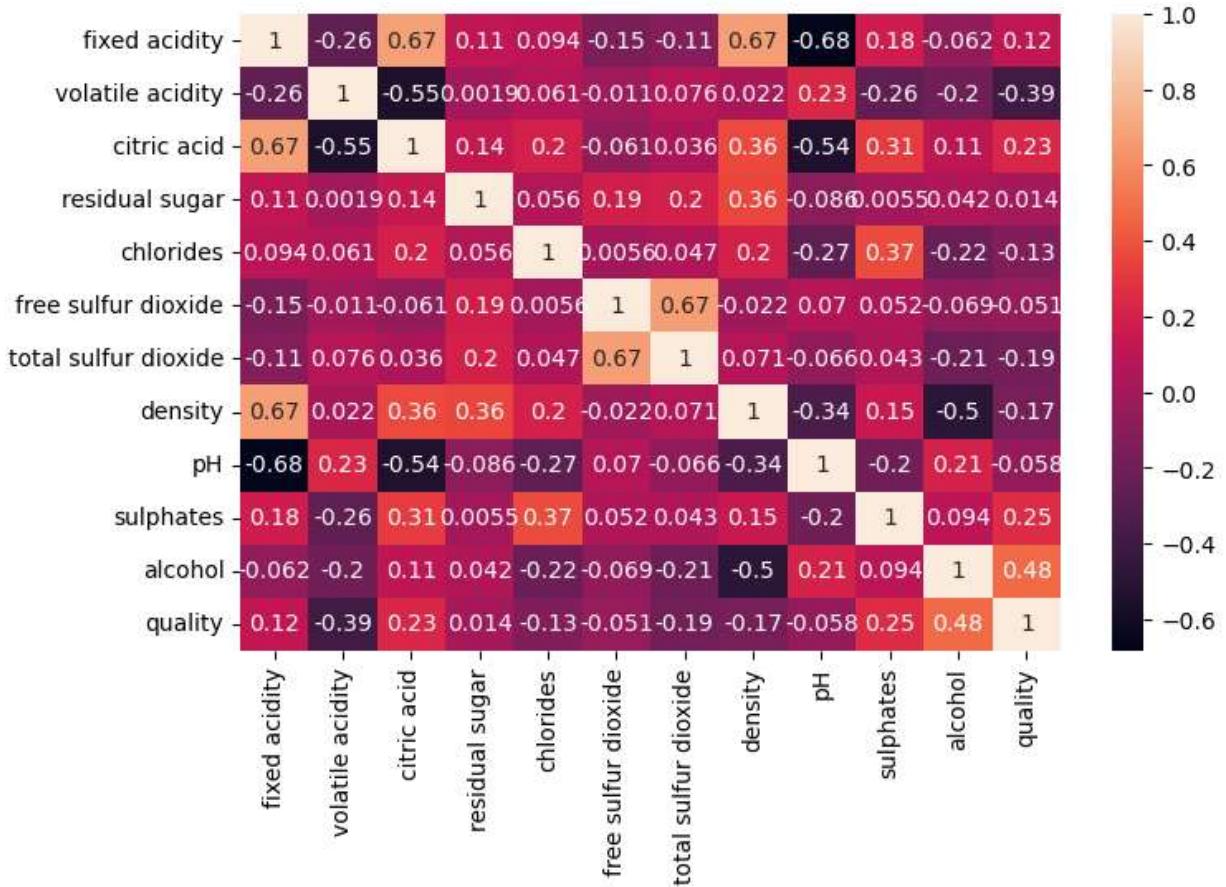






- The box plot shows outliers are present in the data

```
In [12]: plt.figure(figsize = (8, 5))
sns.heatmap(ds.corr(), annot = True)
plt.show()
```



- The heat map shows that Free Sulphur Dioxide and Total Sulphur Dioxide have a positive correlation. Also Fixed Acidity with Citric acid and density. Quality shows the best correlation with Alcohol

## Preprocessing

- The data does not have null values.
- We further check for outliers. Analysis can be done about the cause for these outliers. For our course of study, we maintain the outliers.

```
In [13]: import statistics

def cal_outlier_zscore(col):
    l1=[]
    avg=col.mean()
    sd = statistics.stdev(col)
    for x in col:
        zsc=(x-avg)/sd
        if (zsc>3 or zsc<-3):
            l1.append(x)
    return(l1)

sk_data=pd.DataFrame(columns=['Features','No. of outliers','Outliers'])
sk_data['Features']=ds.columns
idx=0
```

```

for i in ds:

    s=cal_outlier_zscore(ds[i])
    sk_data. loc[idx, 'No. of outliers']= len(s)
    sk_data. loc[idx, 'Outliers']= str(s)
    idx=idx+1
sk_data

```

Out[13]:

	Features	No. of outliers	Outliers
<b>0</b>	fixed acidity	12	[15.0, 15.0, 13.8, 14.0, 13.7, 13.7, 15.6, 14....]
<b>1</b>	volatile acidity	10	[1.13, 1.07, 1.33, 1.33, 1.09, 1.24, 1.185, 1....]
<b>2</b>	citric acid	1	[1.0]
<b>3</b>	residual sugar	30	[10.7, 7.3, 7.2, 7.0, 11.0, 11.0, 7.9, 7.9, 15....]
<b>4</b>	chlorides	31	[0.368, 0.341, 0.332, 0.464, 0.401, 0.467, 0.2....]
<b>5</b>	free sulfur dioxide	22	[52.0, 51.0, 50.0, 68.0, 68.0, 54.0, 53.0, 52....]
<b>6</b>	total sulfur dioxide	15	[148.0, 153.0, 165.0, 151.0, 149.0, 147.0, 148....]
<b>7</b>	density	18	[1.0032, 1.0026, 1.00315, 1.00315, 1.00315, 1....]
<b>8</b>	pH	8	[3.9, 3.85, 2.74, 3.9, 3.78, 3.78, 4.01, 4.01]
<b>9</b>	sulphates	27	[1.56, 1.28, 1.2, 1.28, 1.95, 1.22, 1.95, 1.98....]
<b>10</b>	alcohol	8	[14.0, 14.0, 14.0, 14.0, 14.9, 14.0, 14.0, 14.0]
<b>11</b>	quality	10	[3, 3, 3, 3, 3, 3, 3, 3, 3]

- Residual sugar, chlorides , sulphates and free sulphur dioxide have more than 20 outliers
- We drop duplicate records

In [14]:

```

print(ds.duplicated().sum())
ds.drop_duplicates(inplace=True)

```

240

- We further calculate skewness to understand the distribution.

In [15]:

```

#calculate skewness for each variable
sk_data=pd.DataFrame()
sk_data['Features']=ds.columns
from scipy.stats import skew
sk_data['Skewness']=[skew(ds[i]) for i in ds]
sk_data

```

Out[15]:

	Features	Skewness
<b>0</b>	fixed acidity	0.940002
<b>1</b>	volatile acidity	0.728474
<b>2</b>	citric acid	0.312380
<b>3</b>	residual sugar	4.543132
<b>4</b>	chlorides	5.496412
<b>5</b>	free sulfur dioxide	1.225225
<b>6</b>	total sulfur dioxide	1.538667
<b>7</b>	density	0.044728
<b>8</b>	pH	0.231776
<b>9</b>	sulphates	2.403848
<b>10</b>	alcohol	0.858892
<b>11</b>	quality	0.192194

- Residual sugar, chlorides and sulphates are highly skewed.

## Train Test Split

```
In [16]: x_col=ds.drop('quality',axis=1)
x=ds.drop('quality',axis=1).values
y=ds['quality'].values
```

- We perform oversampling here with the sampling strategy as 'all' indicating equal number of samples for each class.

```
In [17]: """from imblearn.over_sampling import SMOTE
from collections import Counter
from sklearn.datasets import make_classification
from imblearn.over_sampling import RandomOverSampler

# summarize class distribution
print(Counter(y))
# define oversampling strategy
oversample = RandomOverSampler(sampling_strategy='all')
# fit and apply the transform
x, y = oversample.fit_resample(x, y)
# summarize class distribution
print(Counter(y))"""
#should be used in cases where the dataset is not representative of true population
```

```
Out[17]: "from imblearn.over_sampling import SMOTE\nfrom collections import Counter\nfrom sklearn.datasets import make_classification\nfrom imblearn.over_sampling import RandomOverSampler\n\n# summarize class distribution\nprint(Counter(y))\n# define oversampling strategy\noversample = RandomOverSampler(sampling_strategy='all')\n# fit and apply the transform\nx, y = oversample.fit_resample(x, y)\n# summarize class distribution\nprint(Counter(y))"
```

```
In [18]: x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=5)
```

```
In [ ]:
```

```
In [19]: ds.columns
```

```
Out[19]: Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',  
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',  
       'pH', 'sulphates', 'alcohol', 'quality'],  
      dtype='object')
```

## Feature Scaling to apply KNN and SVM since the two models are distance based

- Since many columns follow normal distribution, we follow StandardScaler
- Splitting for Standard Scaler

```
In [20]: from sklearn.preprocessing import StandardScaler  
ss=StandardScaler()  
  
x_train=ss.fit_transform(x_train)  
x_test=ss.transform(x_test)
```

- Multicollinearity will help to check if any two variables are highly correlated.

```
#multicollinearity-check  
from statsmodels.stats.outliers_influence import variance_inflation_factor  
vif_data=pd.DataFrame()  
vif_data['Features']=x_col.columns  
vif_data['VIF']=[variance_inflation_factor(x_train,i) for i in range(x_train.shape[1])]  
vif_data['multicollinear']=[True if vif_data['VIF'][i]>5 else False for i in range(x_train.shape[1])]  
vif_data
```

Out[21]:

	Features	VIF	multicollinear
0	fixed acidity	8.190673	True
1	volatile acidity	1.750200	False
2	citric acid	3.136416	False
3	residual sugar	1.676717	False
4	chlorides	1.527067	False
5	free sulfur dioxide	1.979689	False
6	total sulfur dioxide	2.260255	False
7	density	6.490853	True
8	pH	3.376254	False
9	sulphates	1.485474	False
10	alcohol	3.159173	False

- Since multi collinearity exists and we cannot remove the columns Logistic Regression and Naive Bayes cannot be applied

## Model Building

### Decision tree model with max\_depth=3

```
In [22]: mod_4=DecisionTreeClassifier(random_state=1,max_depth=3)
mod_4.fit(x_train,y_train)
y_pred=mod_4.predict(x_test)
acc=accuracy_score(y_test,y_pred)
print("Accuracy:",acc)
```

Accuracy: 0.5661764705882353

### Decision tree model with max\_depth=4

```
In [23]: mod_4=DecisionTreeClassifier(random_state=1,max_depth=4)
mod_4.fit(x_train,y_train)
y_pred=mod_4.predict(x_test)
acc=accuracy_score(y_test,y_pred)
print(acc)
```

0.5625

```
In [24]: mod_4=DecisionTreeClassifier(random_state=1,max_depth=6)
mod_4.fit(x_train,y_train)
y_pred=mod_4.predict(x_test)
accdt=accuracy_score(y_test,y_pred)
print("Accuracy:",acc)
```

Accuracy: 0.5625

```
In [25]: mod_4=DecisionTreeClassifier(random_state=1,max_depth=5)
mod_4.fit(x_train,y_train)
y_pred=mod_4.predict(x_test)
acc=accuracy_score(y_test,y_pred)
print("Accuracy:",acc)
```

Accuracy: 0.5882352941176471

- max\_depth= 5 showed the best possible results

## Random Forest Classifier

```
In [26]: rf=RandomForestClassifier(random_state=1,class_weight='balanced')
rf.fit(x_train,y_train)
y_pred=rf.predict(x_test)
accrf=accuracy_score(y_test,y_pred)
print("Accuracy:",accrf)
score=cross_val_score(rf,x_train,y_train,cv=10)
print("Accuracy after k-fold validation:",score.mean())
print(classification_report(y_test,y_pred))
```

Accuracy: 0.6507352941176471

Accuracy after k-fold validation: 0.6079595650696568

	precision	recall	f1-score	support
3	0.00	0.00	0.00	1
4	0.00	0.00	0.00	7
5	0.66	0.82	0.73	105
6	0.66	0.64	0.65	119
7	0.62	0.42	0.50	36
8	0.00	0.00	0.00	4
accuracy			0.65	272
macro avg	0.32	0.31	0.31	272
weighted avg	0.62	0.65	0.63	272

## KNN

```
In [27]: knn=KNeighborsClassifier(n_neighbors=15)
knn.fit(x_train,y_train)
y_pred=knn.predict(x_test)
accknn=accuracy_score(y_test,y_pred)
print(accknn)
```

0.5661764705882353

- We use SVM further

```
In [28]: svc=SVC(kernel='rbf',C=1,gamma=0.1)
svc.fit(x_train,y_train)
ypred=svc.predict(x_test)
print(metrics.accuracy_score(y_test,ypred))
```

0.6213235294117647

- Highest accuracy is obtained for Random Forest. Hence we will tune and check the model further.

```
In [29]: imp=rf.feature_importances_
fi=pd.DataFrame()
fi['Feature']=x_col.columns
fi['Score']=imp
fi.sort_values('Score',ascending=False)
```

Out[29]:

	Feature	Score
<b>10</b>	alcohol	0.129426
<b>1</b>	volatile acidity	0.128414
<b>9</b>	sulphates	0.117479
<b>7</b>	density	0.093324
<b>6</b>	total sulfur dioxide	0.091488
<b>4</b>	chlorides	0.091088
<b>2</b>	citric acid	0.076745
<b>8</b>	pH	0.075121
<b>5</b>	free sulfur dioxide	0.067579
<b>0</b>	fixed acidity	0.066734
<b>3</b>	residual sugar	0.062604

- All features are contributing to prediction.

## Hyperparameter tuning for Random Forest with 5 features

```
In [30]: param_dist={'max_depth':[4,5,6],
                  'criterion':['gini','entropy'],
                  'oob_score':[True,False],
                  'max_features':['auto','sqrt','log2'],
                  'bootstrap':[True,False]}
cv_rf=GridSearchCV(rf, cv=10, param_grid=param_dist, verbose=2, n_jobs=1)
cv_rf.fit(x_train,y_train)
rf.set_params(**cv_rf.best_params_)
rf.fit(x_train,y_train)
y_pred=rf.predict(x_test)
```

















































```
[CV] END bootstrap=False, criterion=entropy, max_depth=6, max_features=log2, oob_scoring=False; total time= 0.3s
```

```
In [31]: acc=accuracy_score(y_test,y_pred)
print(acc)
```

```
0.5220588235294118
```

- Hyperparameter tuning showed reduced accuracy ##### Hyperparameter tuning for SVM

```
In [32]: paramm={'C':[0.1,1,10],
            'gamma':[1,0.1,0.01],
            'kernel': ['linear','rbf','poly']}
ms=GridSearchCV(svc,paramm,cv=5,scoring='accuracy',verbose=1)
```

```
In [33]: ms.fit(x_train,y_train)
```

```
Fitting 5 folds for each of 27 candidates, totalling 135 fits
```

```
Out[33]:
```

```
▶ GridSearchCV
  ▶ estimator: SVC
    ▶ SVC
```

```
In [34]: print(ms.best_params_)
```

```
{'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
```

```
In [35]: svcm=SVC(kernel='rbf',C=1,gamma=1)
svcm.fit(x_train,y_train)
ypred=svcm.predict(x_test)
accsvm=metrics.accuracy_score(y_test,ypred)
print(accsvm)
```

```
0.5845588235294118
```

**RandomForestClassifier without hyperparameter tuning and SVM after tuning showed the best accuracy**

```
In [36]: print(classification_report(y_test,ypred))
```

	precision	recall	f1-score	support
3	0.00	0.00	0.00	1
4	0.00	0.00	0.00	7
5	0.57	0.79	0.66	105
6	0.60	0.56	0.58	119
7	0.60	0.25	0.35	36
8	0.00	0.00	0.00	4
accuracy			0.58	272
macro avg	0.30	0.27	0.27	272
weighted avg	0.56	0.58	0.56	272

```
In [37]: cm=confusion_matrix(y_test,ypred)
cm
```

```
Out[37]: array([[ 0,  0,  1,  0,  0,  0],  
   [ 0,  0,  7,  0,  0,  0],  
   [ 0,  0,  83, 22,  0,  0],  
   [ 0,  0,  48, 67,  4,  0],  
   [ 0,  0,  7, 20,  9,  0],  
   [ 0,  0,  2,  2,  0]], dtype=int64)
```

## Accuracy comparison

```
In [38]: accpd=pd.DataFrame()  
accpd['Model']=['Decision Tree', 'Random Forest', 'KNN', 'SVM after hyperparameter tuning'  
accpd['Accuracy']=[accdt,accrf,accknn,accsvm]  
accpd
```

Out[38]:

	Model	Accuracy
0	Decision Tree	0.529412
1	Random Forest	0.650735
2	KNN	0.566176
3	SVM after hyperparameter tuning	0.584559

The accuracy achieved was 89% with Random Forest

In [ ]: