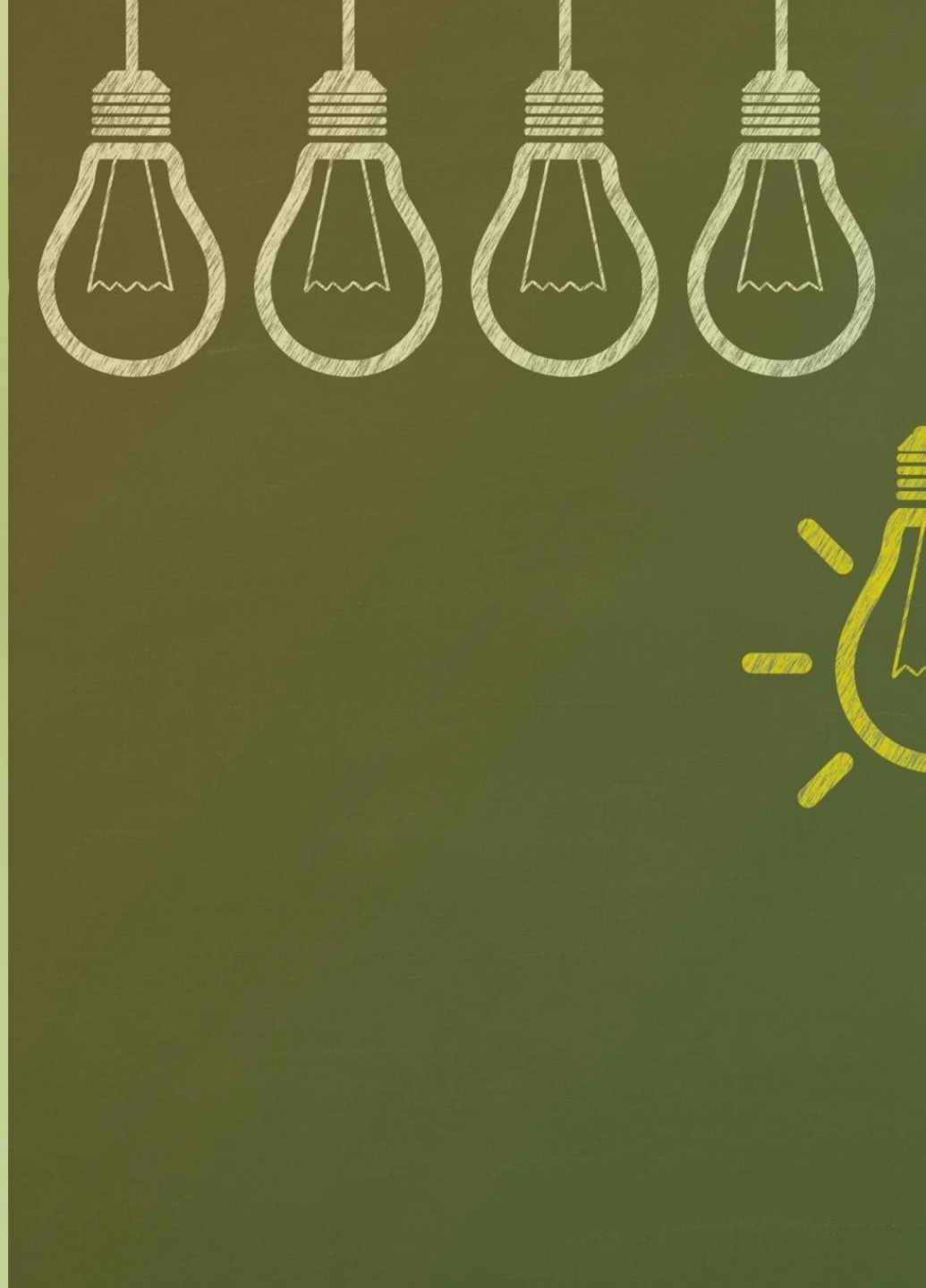
Predicting the Quality of Wine

Maayan Lavi

313374621



Introduction

- The data is concerned with respect to the predicting the quality of wine given its chemical properties.
- There are 2 different kind of wine namely red wine and white wine.
- Totally there are 6497 instances present in the data set and it has 13 features namely Fixed acidity, Volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, PH , sulphates, alcohol, quality and kind .
- All the features are numerical in nature except density and kind which is categorical in nature.
- Density has the categories very high , medium and high. Kind is the kind of wine and can be white wine or red wine.

Initial Data Analysis

- checking the description of the data set

	FA	VA	CA	RS	chlorides	FSD	TSD	pH	sulphates	alcohol	quality
count	6303.000000	6497.00000	6040.000000	6425.000000	6497.000000	6125.000000	5770.000000	6497.000000	6497.000000	6416.000000	6497.000000
mean	7.183104	0.37178	1.863810	5.443837	0.056034	146.834122	115.802080	3.218501	0.531268	10.487661	5.818378
std	1.378804	0.50821	0.889848	4.763496	0.035034	25.724505	56.473495	0.160787	0.148806	1.191010	0.873255
min	0.110000	0.08000	0.000000	0.600000	0.009000	59.000000	6.000000	2.720000	0.220000	8.000000	3.000000
25%	6.400000	0.23000	1.330000	1.800000	0.038000	129.000000	77.000000	3.110000	0.430000	9.500000	5.000000
50%	7.000000	0.29000	1.860000	3.000000	0.047000	147.000000	118.000000	3.210000	0.510000	10.300000	6.000000
75%	7.700000	0.41000	2.380000	8.100000	0.065000	164.000000	155.000000	3.320000	0.600000	11.300000	6.000000
max	15.900000	13.00000	9.960000	65.800000	0.611000	347.000000	440.000000	4.010000	2.000000	14.900000	9.000000

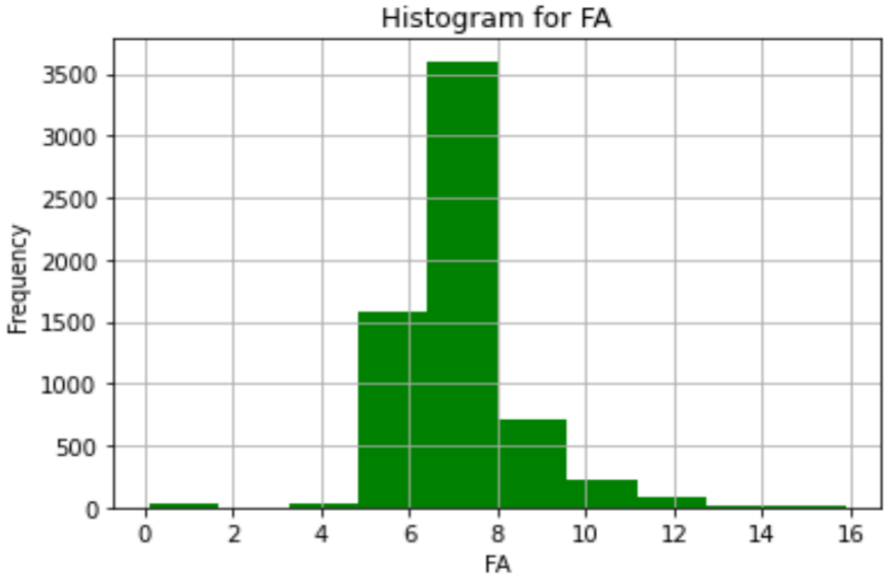
Initial Data Analysis

- Checking for the presence of null values in the data set.
- There were several null values present in the data set.
- The number of missing values for each feature is quite less and we will replace with it respective mean.
- There were a few missing values in the density feature as well. I removed those rows from the data set.

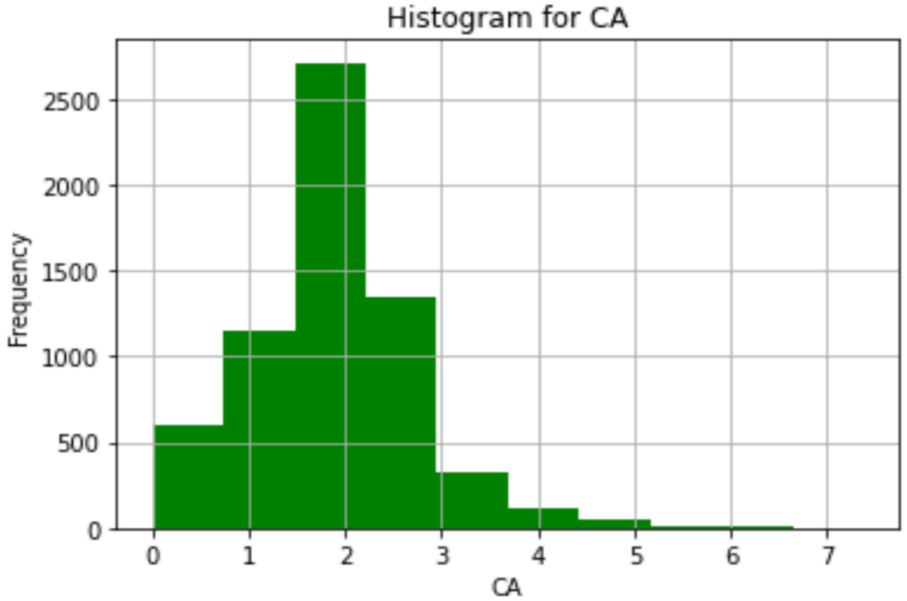
FA	194
VA	0
CA	457
RS	72
chlorides	0
FSD	372
TSD	727
density	192
pH	0
sulphates	0
alcohol	81
quality	0
kind	0

Exploratory Data Analysis

- **Plotting of histograms –**



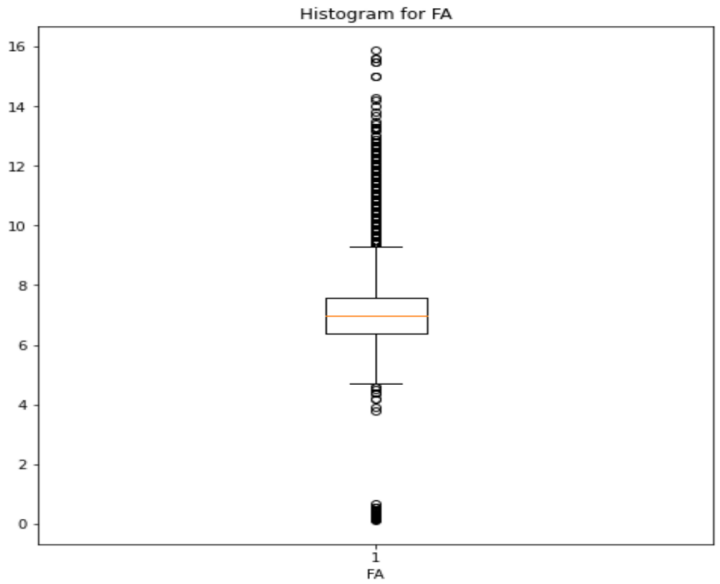
Histogram of Fixed Acidity



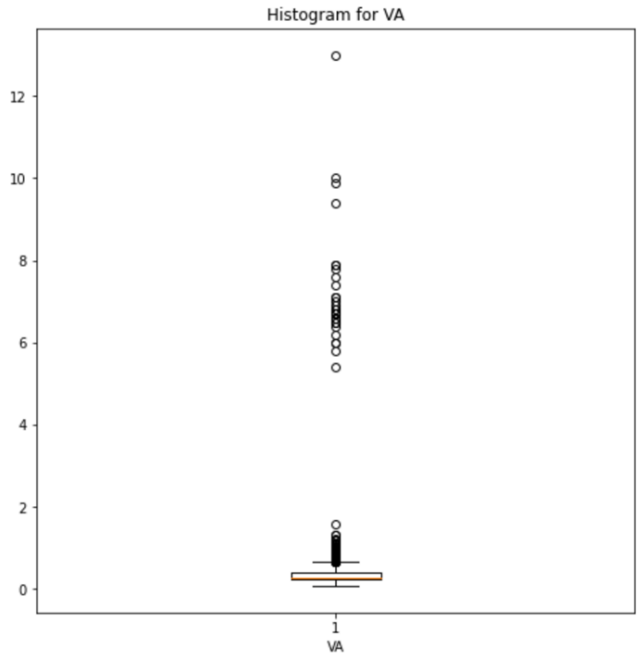
Histogram for Citric Acid

Exploratory Data Analysis

- **Plotting for boxplots -**



Histogram of Fixed Acidity

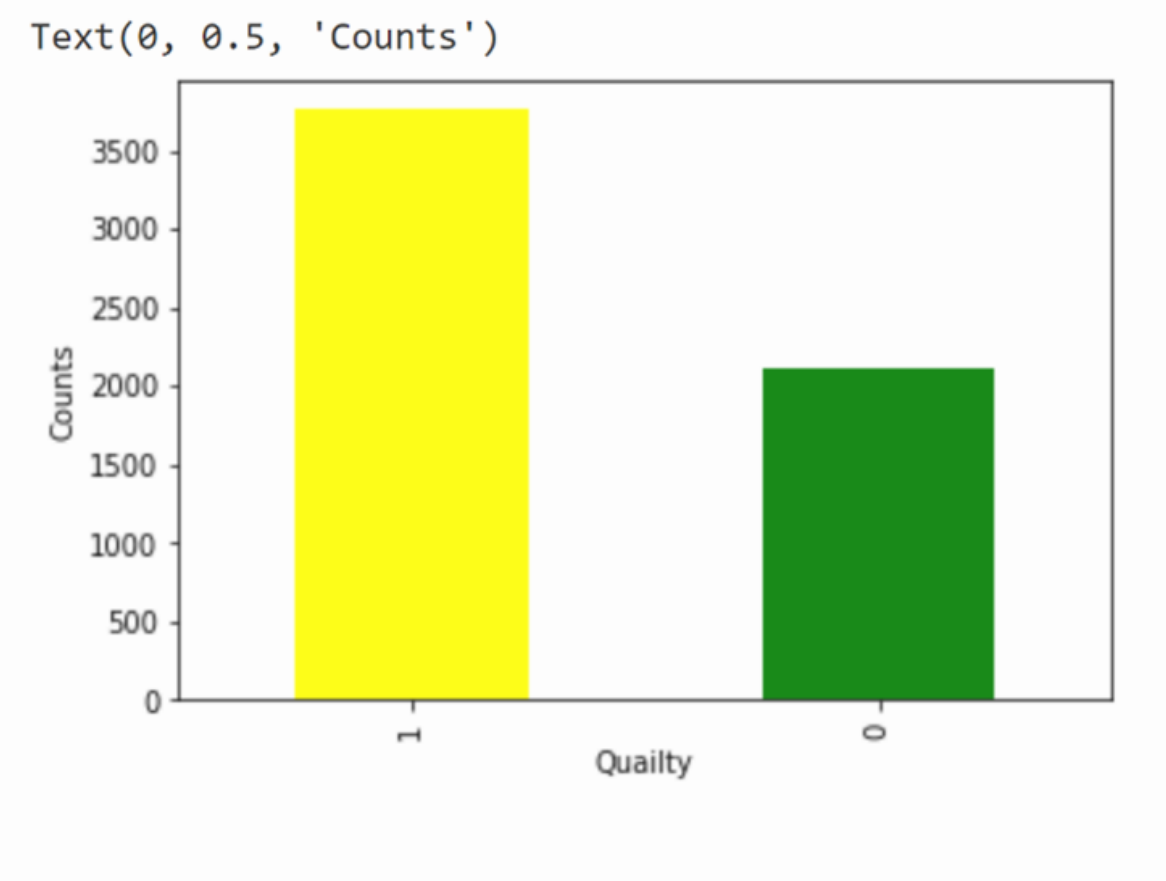


Histogram for Volatile Acidity

There are several outliers present in the data set which might affect the performance of prediction. So, we go ahead and remove the outliers from the features. The black dots in the boxplots indicates that they are outliers. After the removal of outliers we are left with 5865 instances in the data set

Converting to a classification problem

- We are predicting the quality of wine and it can be any value between 1 and 10 with 1 being the lowest quality and 10 being the highest quality wine. We can convert the problem into a classification problem as follows: If quality is less than or equal to 5 then classify it low quality wine else classify it as high-quality wine.
- We can see that there are around 3700 high quality wine instances compared to around 2000 low quality wine instances.



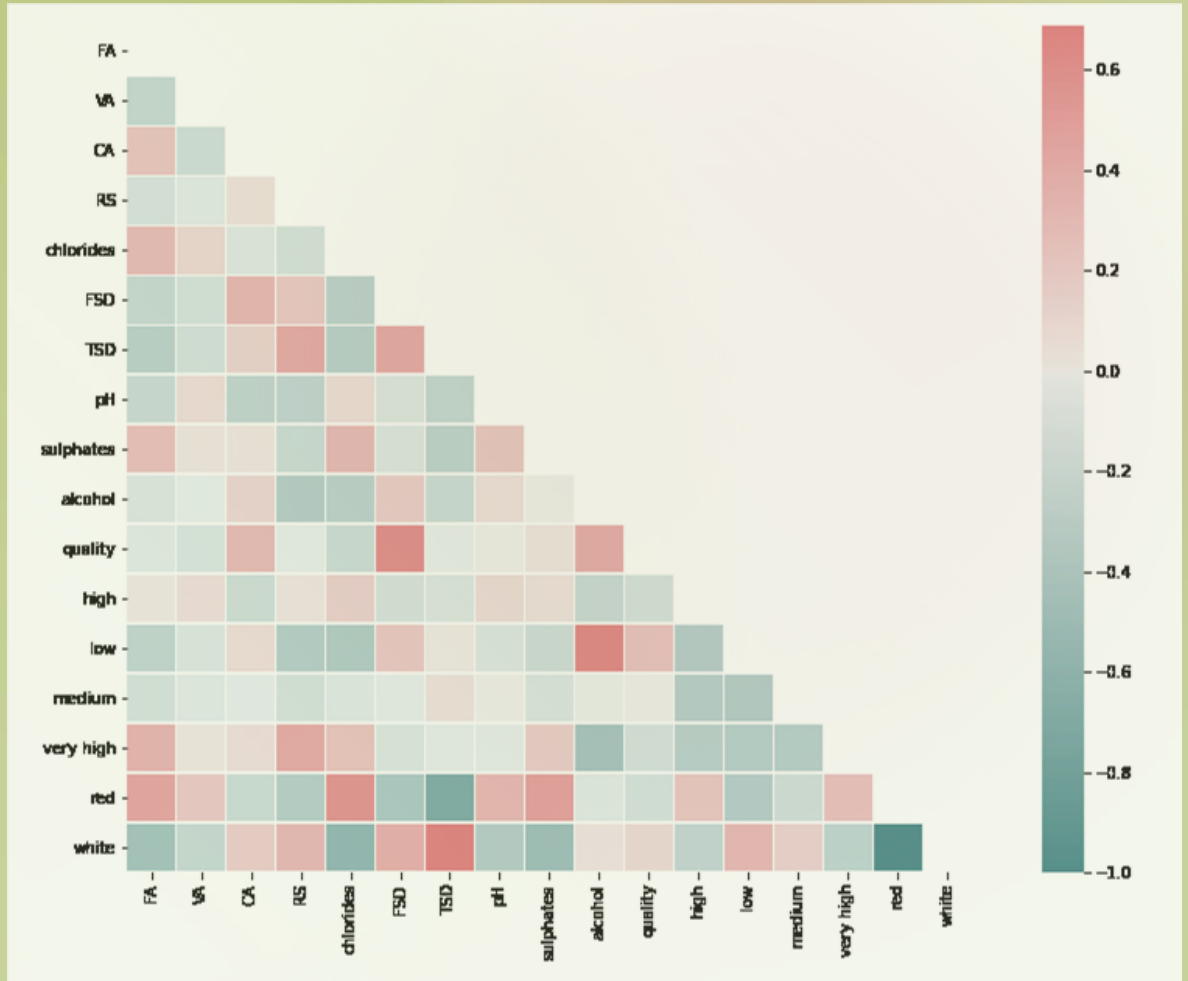
Finding the correlation among features

- Finding the correlation among features to find the important features.
- We will use a correlation matrix for the same purpose.

	FA	VA	CA	RS	chlorides	FSD	TSD	pH	sulphates	alcohol	quality	high	low	medium	very high	red	white
FA	1.000000	-0.235644	0.259100	-0.121842	0.318792	-0.229392	-0.320969	-0.213252	0.290203	-0.094625	-0.062965	0.041764	-0.271115	-0.133159	0.377182	0.465210	-0.465210
VA	-0.235644	1.000000	-0.167317	-0.057765	0.141951	-0.139895	-0.149230	0.109960	0.059575	-0.023071	-0.105505	0.098261	-0.081709	-0.060212	0.048182	0.233748	-0.233748
CA	0.259100	-0.167317	1.000000	0.087391	-0.071685	0.350487	0.165348	-0.274567	0.060450	0.149967	0.325197	-0.183229	0.112121	-0.025834	0.094565	-0.201989	0.201989
RS	-0.121842	-0.057765	0.087391	1.000000	-0.137453	0.229317	0.425759	-0.265373	-0.197020	-0.334723	-0.020095	0.058108	-0.334502	-0.140405	0.434084	-0.336936	0.336936
chlorides	0.318792	0.141951	-0.071685	-0.137453	1.000000	-0.290081	-0.309868	0.104672	0.330730	-0.288974	-0.195744	0.180102	-0.366699	-0.070612	0.272708	0.573684	-0.573684
FSD	-0.229392	-0.139895	0.350487	0.229317	-0.290081	1.000000	0.429562	-0.089426	-0.084691	0.207360	0.593122	-0.125752	0.236153	-0.026680	-0.091299	-0.392098	0.392098
TSD	-0.320969	-0.149230	0.165348	0.425759	-0.309868	0.429562	1.000000	-0.243146	-0.286312	-0.191846	-0.012835	-0.090194	0.037942	0.076692	-0.027834	-0.687193	0.687193
pH	-0.213252	0.109960	-0.274567	-0.265373	0.104672	-0.089426	-0.243146	1.000000	0.242536	0.090107	0.002468	0.108971	-0.083889	0.007670	-0.030415	0.347311	-0.347311
sulphates	0.290203	0.059575	0.060450	-0.197020	0.330730	-0.084691	-0.286312	0.242536	1.000000	0.002503	0.057091	0.081933	-0.180421	-0.102380	0.210730	0.493155	-0.493155
alcohol	-0.094625	-0.023071	0.149967	-0.334723	-0.288974	0.207360	-0.191846	0.090107	0.002503	1.000000	0.397222	-0.213711	0.624465	-0.005089	-0.428590	-0.064638	0.064638
quality	-0.062965	-0.105505	0.325197	-0.020095	-0.195744	0.593122	-0.012835	0.002468	0.057091	0.397222	1.000000	-0.137526	0.266732	0.003163	-0.142037	-0.130579	0.130579
high	0.041764	0.098261	-0.183229	0.058108	0.180102	-0.125752	-0.090194	0.108971	0.081933	-0.213711	-0.137526	1.000000	-0.337318	-0.332738	-0.312362	0.256892	-0.256892
low	-0.271115	-0.081709	0.112121	-0.334502	-0.366699	0.236153	0.037942	-0.083889	-0.180421	0.624465	0.266732	-0.337318	1.000000	-0.354957	-0.333220	-0.341092	0.341092
medium	-0.133159	-0.060212	-0.025834	-0.140405	-0.070612	-0.026680	0.076692	0.007670	-0.102380	-0.005089	0.003163	-0.332738	-0.354957	1.000000	-0.328696	-0.180386	0.180386
very high	0.377182	0.048182	0.094565	0.434084	0.272708	-0.091299	-0.027834	-0.030415	0.210730	-0.428590	-0.142037	-0.312362	-0.333220	-0.328696	1.000000	0.282149	-0.282149
red	0.465210	0.233748	-0.201989	-0.336936	0.573684	-0.392098	-0.687193	0.347311	0.493155	-0.064638	-0.130579	0.256892	-0.341092	-0.180386	0.282149	1.000000	-1.000000
white	-0.465210	-0.233748	0.201989	0.336936	-0.573684	0.392098	0.687193	-0.347311	-0.493155	0.064638	0.130579	-0.256892	0.341092	0.180386	-0.282149	-1.000000	1.000000

Finding the correlation among features

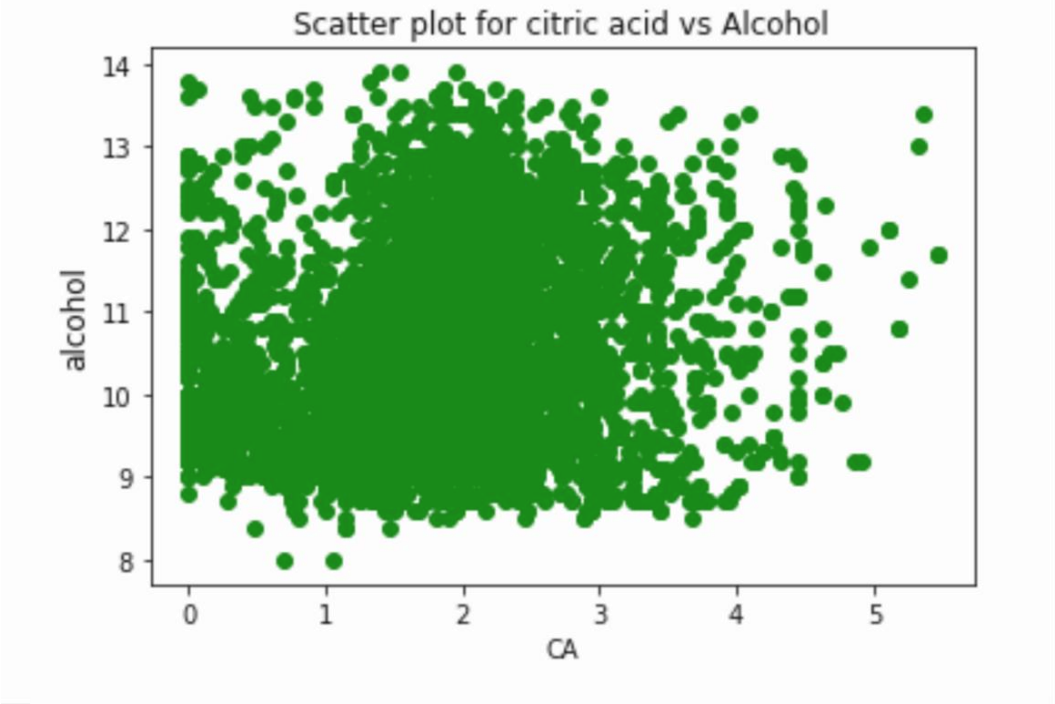
- We can see that no features are particularly highly correlated with each other. We will use all the features except kind. Intuitively it seems kind categorical feature does not give much information while predicting the target feature.



Classification Model

a) Gaussian Naïve Bayes

- We will use permutation importance package form scikit learn to find the best 2 features for the GNB classifier. On using the permutation package, we get Citric Acid and Alcohol as the 2 best features for GNB classifier. The scatter plot between them is :



```
The classification report of Baseline Decision Tree Algorithm is
              precision    recall  f1-score   support

           0       0.83      0.82      0.82       570
           1       0.90      0.90      0.90      1007

    accuracy                           0.87      1577
   macro avg       0.86      0.86      0.86      1577
weighted avg       0.87      0.87      0.87      1577
```

Classification Model

b) Decision Tree Baseline Model

- On using the Decision Tree Baseline Model, we get the classification Report. An accuracy of around 87 percent was achieved.

```
The classification report of Decision tree Algorithm using modified data set  is
              precision    recall  f1-score   support

           0       0.80      0.81      0.81       524
           1       0.89      0.89      0.89       943

    accuracy                           0.86      1467
   macro avg       0.85      0.85      0.85      1467
weighted avg       0.86      0.86      0.86      1467
```

Classification Report on using Decision Tree on Modified Data Set

Classification Model

c) Decision Tree using modified Data Set

- On modelling using Decision on our modified Data Set we get around the same performance. Performance couldn't be improved significantly as compared to the baseline decision tree model.

Summary

Finally, we were able to predict the quality of wine in an effective manner.

Decent results were obtained.

Since we were having only around 6500 instances, with more data, the model will be able to learn more effectively and its performance can be increased significantly.

Moreover different feature engineering and hyperparameter tuning technique can be tried out to increase the performance of the model. Currently the Decision Tree algorithm on the baseline model and the Decision Tree Algorithm on the Modified Data Set performs equivalently. Gaussian Naïve bayes algorithm give accuracy of around 82 percent.

Some of the insights from the analysis is there are more high-quality wine than low quality wines.

Moreover, the features present are somewhat independent of each other.

The average amount of alcohol content present in the wine is around 10 percent.