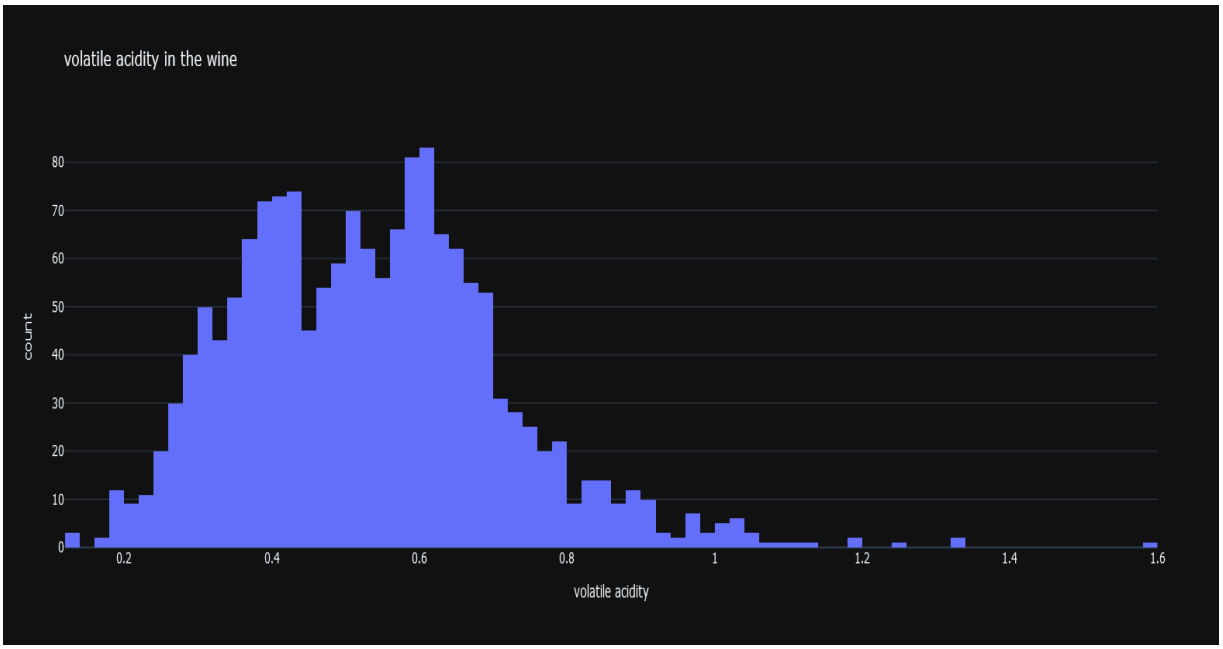
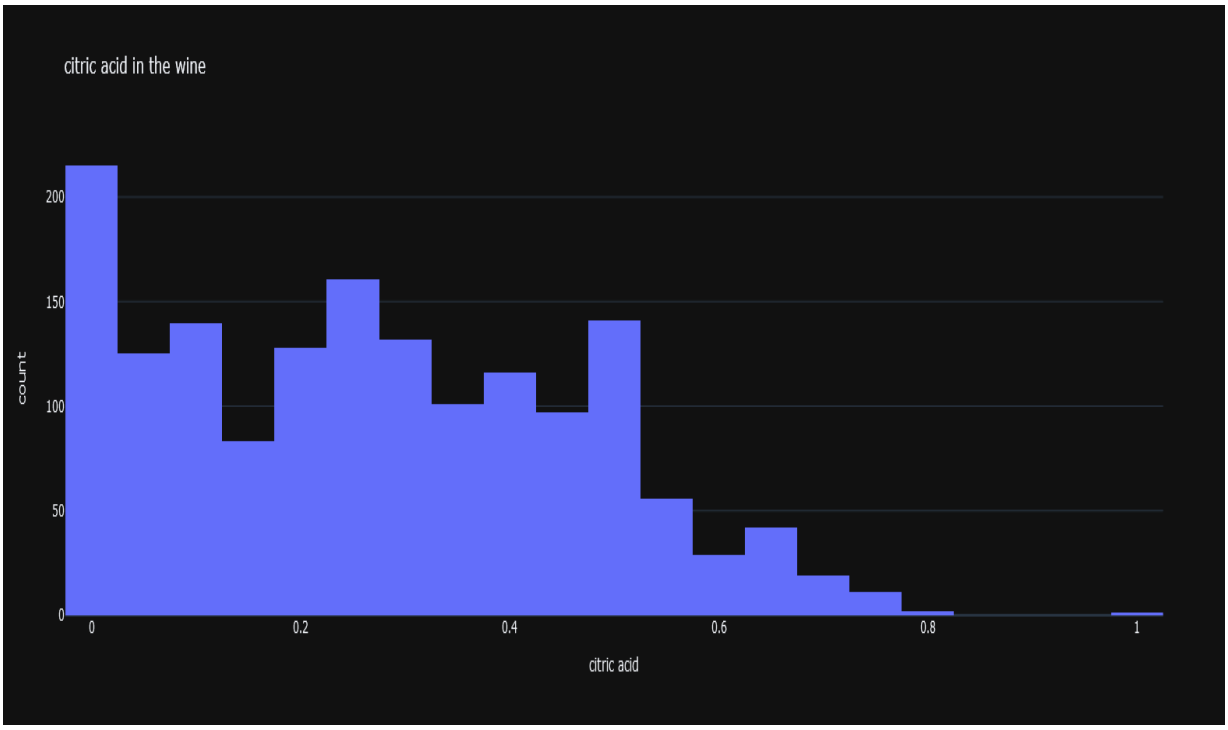
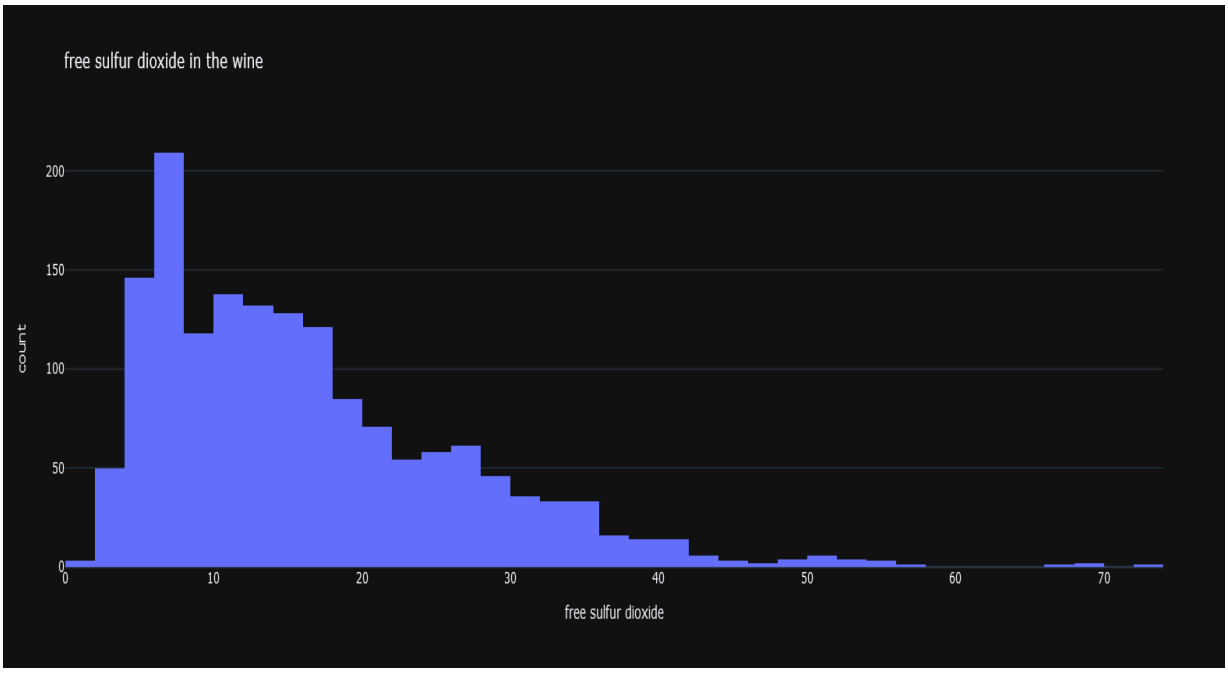
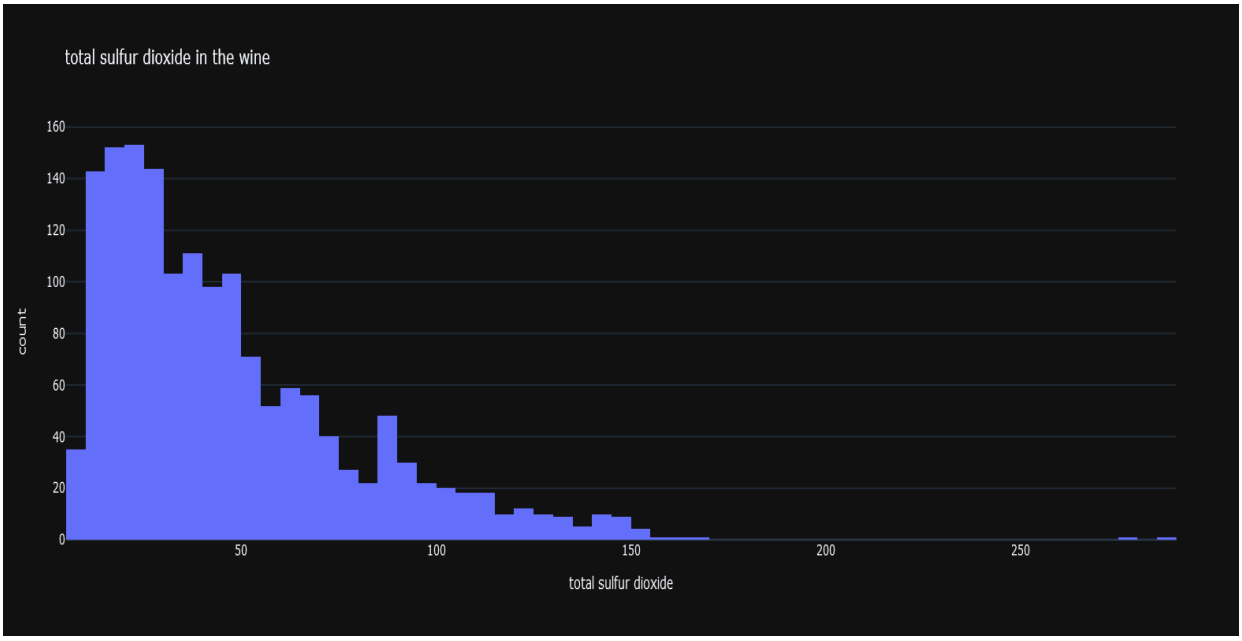
- Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation.
    - The main objective of this model is to predict the quality of wine
- Brief description of the data set you chose and a summary of its attributes.
    - The dataset chosen for this report was red wine quality from UCI repository
    - There are 12 attributes: -
        - fixed acidity
        - volatile acidity
        - citric acid
        - residual sugar
        - chlorides
        - free sulfur dioxide
        - total sulfur dioxide
        - density
        - pH
        - sulphates
        - alcohol
        - Output variable (based on sensory data): quality (score between 0 and 10)
- Brief summary of data exploration and actions taken for data cleaning and feature engineering.
    - There were no null values and the data set was perfectly clean

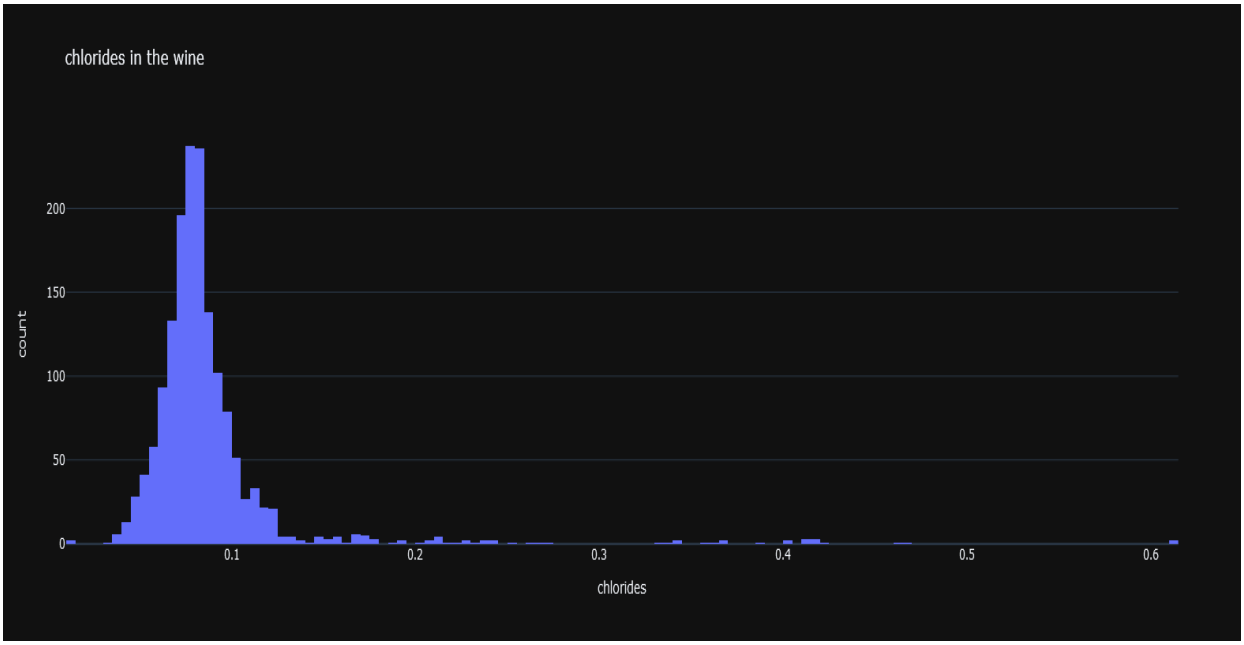
- Univariate analysis

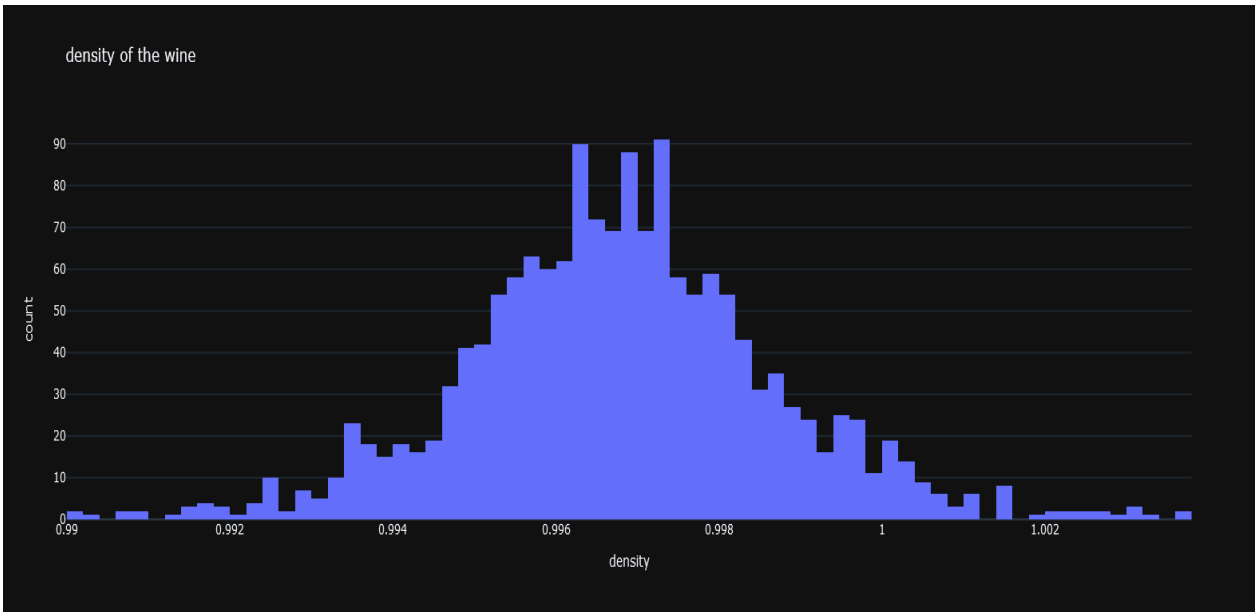


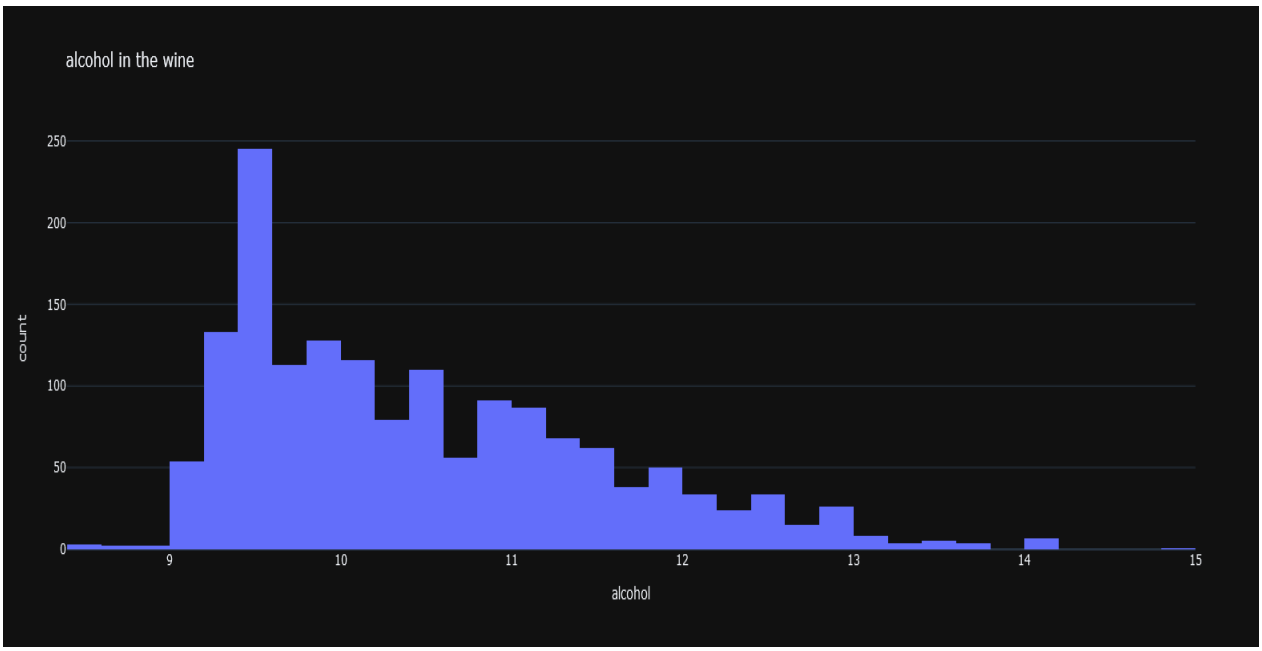

citric acid in the wine
count
200
150
100
50
0
0
0.2
0.4
0.6
0.8
1
citric acid

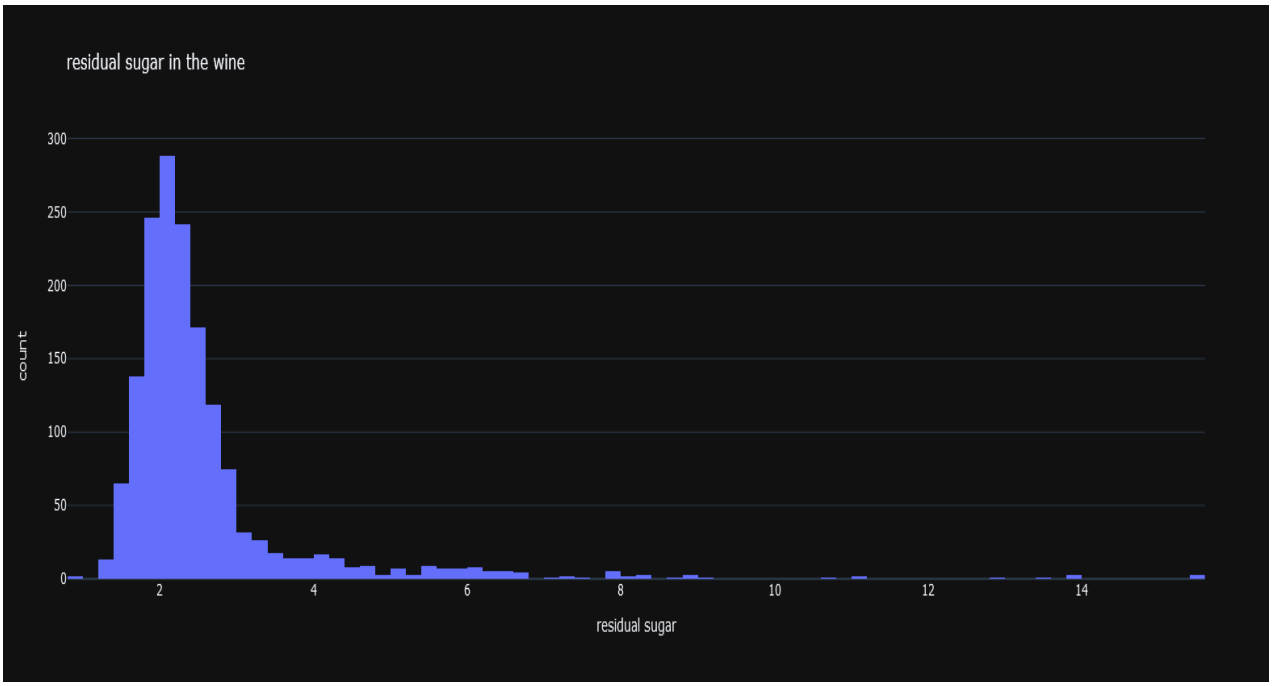

free sulfur dioxide in the wine
count
200
150
100
50
0
0
10
20
30
40
50
60
70
free sulfur dioxide

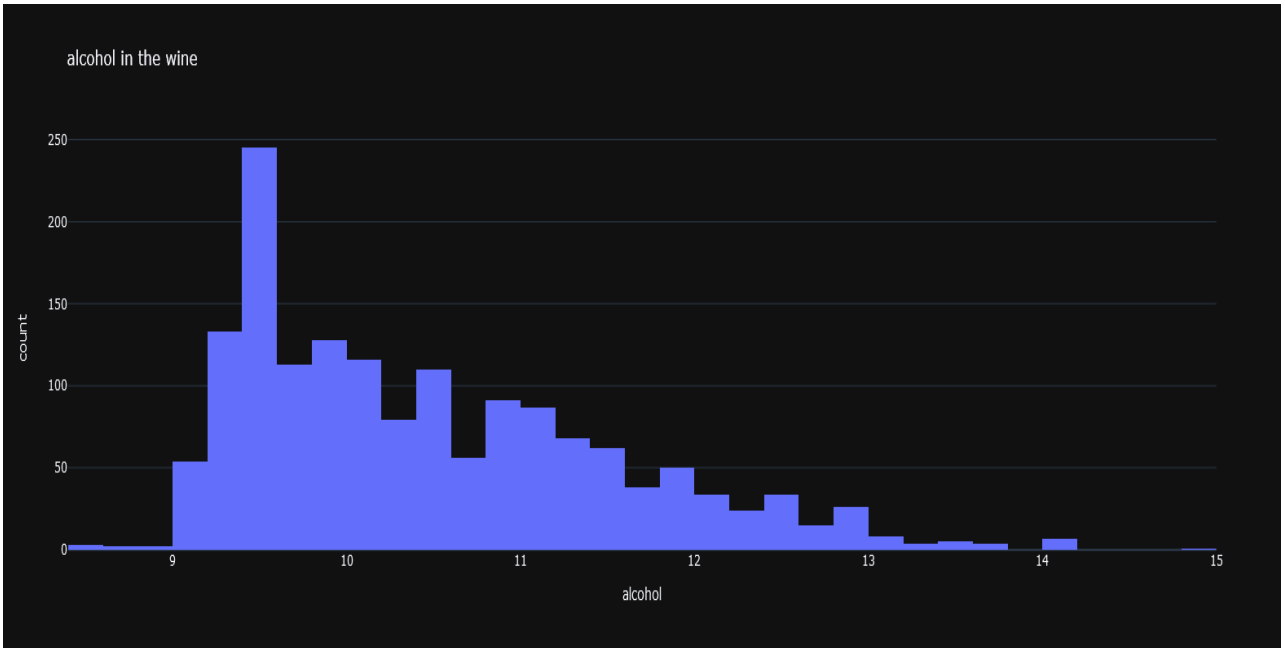

total sulfur dioxide in the wine
160
140
120
100
80
60
40
20
0
count
50
100
150
200
250
total sulfur dioxide

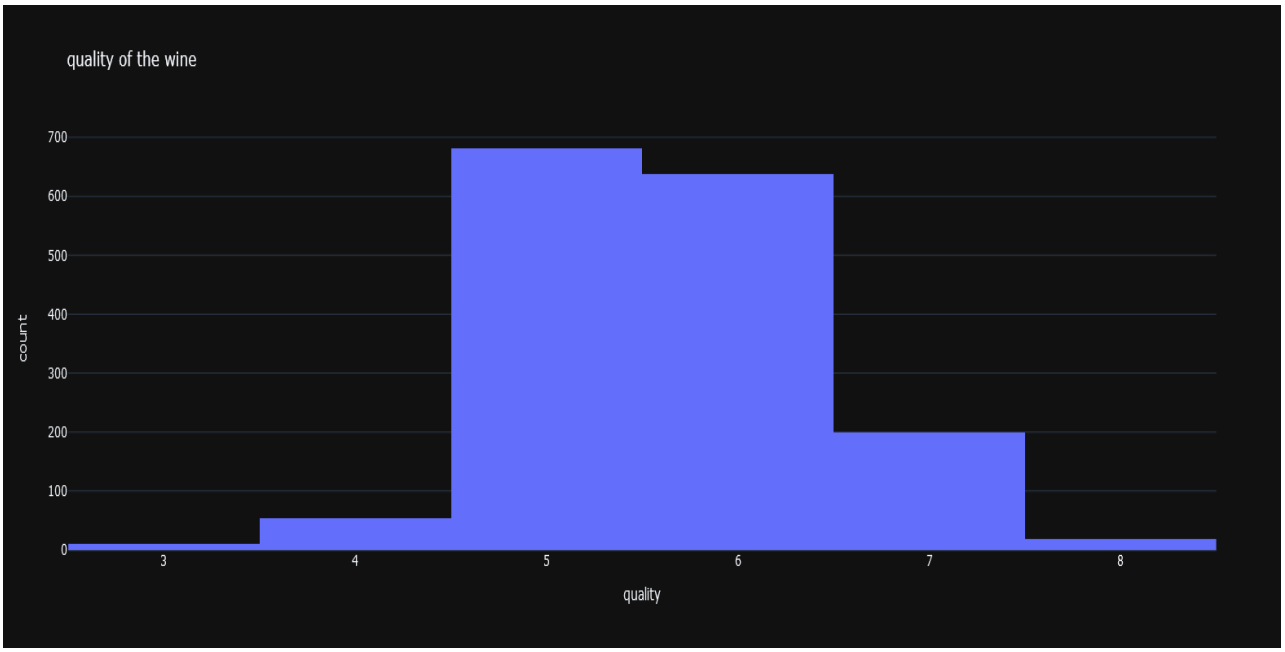

chlorides in the wine
200
150
100
50
0
count
0.1
0.2
0.3
0.4
0.5
0.6
chlorides


density of the wine
90
80
70
60
50
40
30
20
10
0
count
0.99
0.992
0.994
0.996
0.998
1
1.002
density


alcohol in the wine
250
200
150
100
50
0
count
9
10
11
12
13
14
15
alcohol

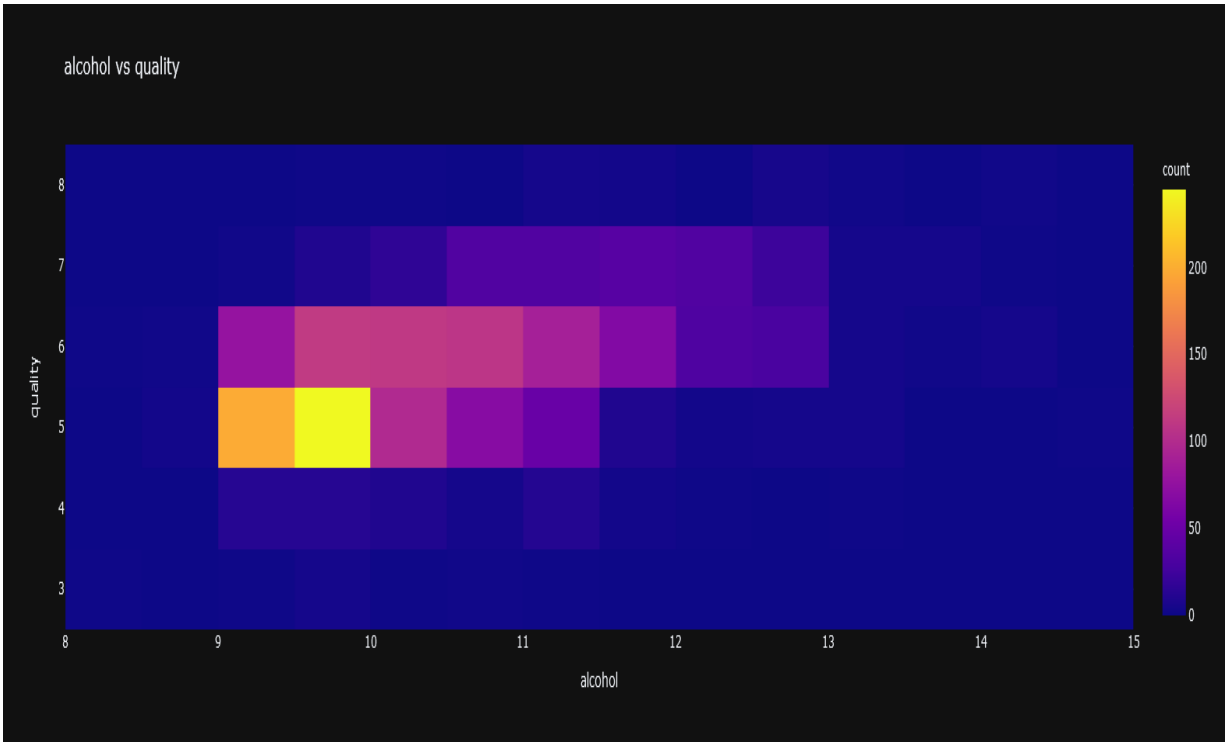

residual sugar in the wine
300
250
200
150
100
50
0
count
2
4
6
8
10
12
14
residual sugar


alcohol in the wine
250
200
150
100
50
0
count
9
10
11
12
13
14
15
alcohol

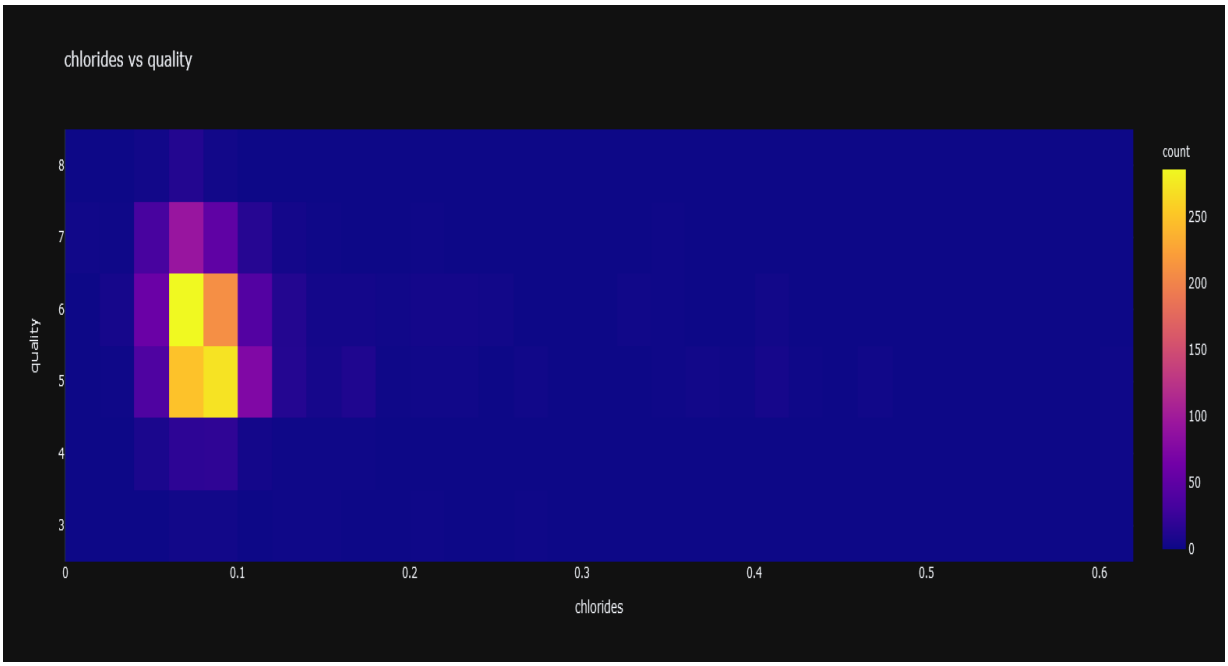


- Bivariate analysis

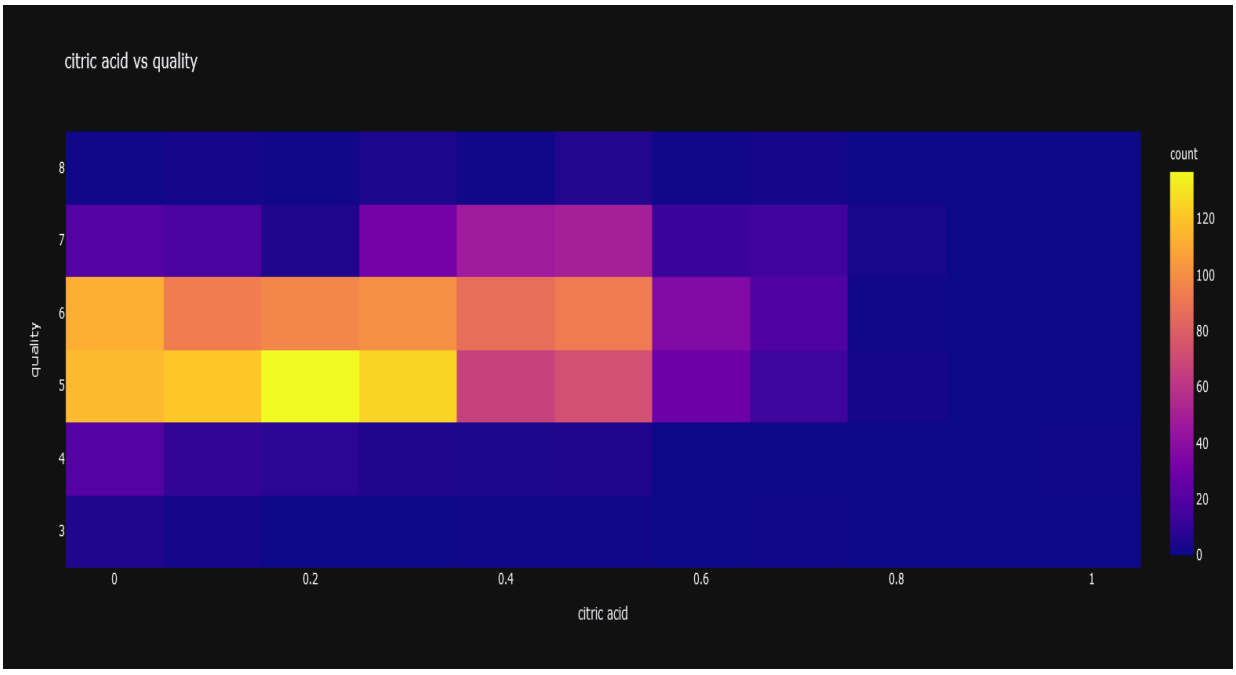
Alcohol vs quality



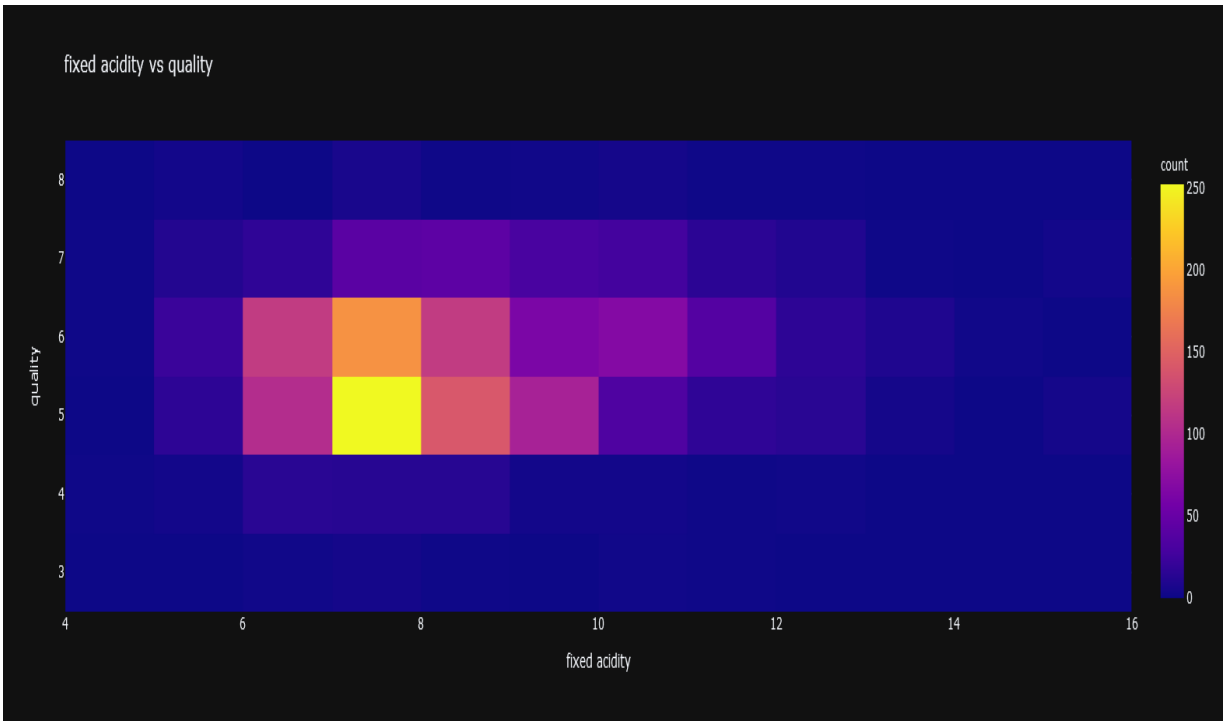
Chloride vs quality



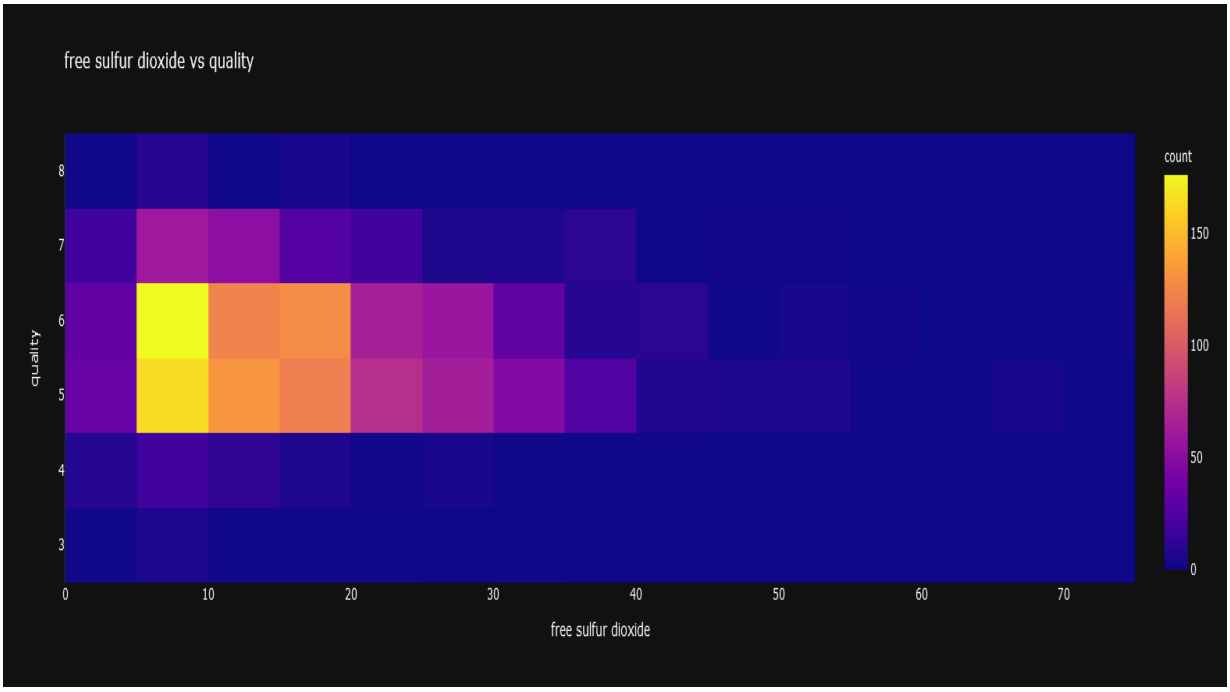
Citric acid vs quality



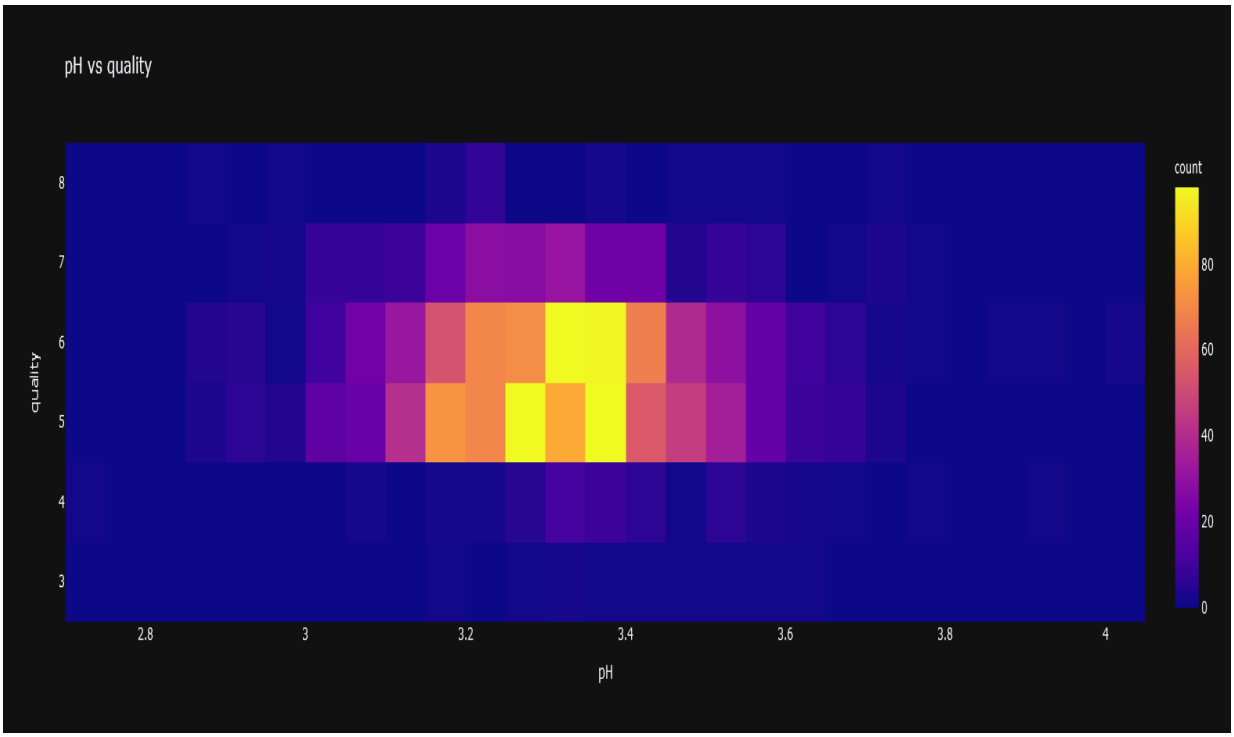
Fixed acidity vs quality



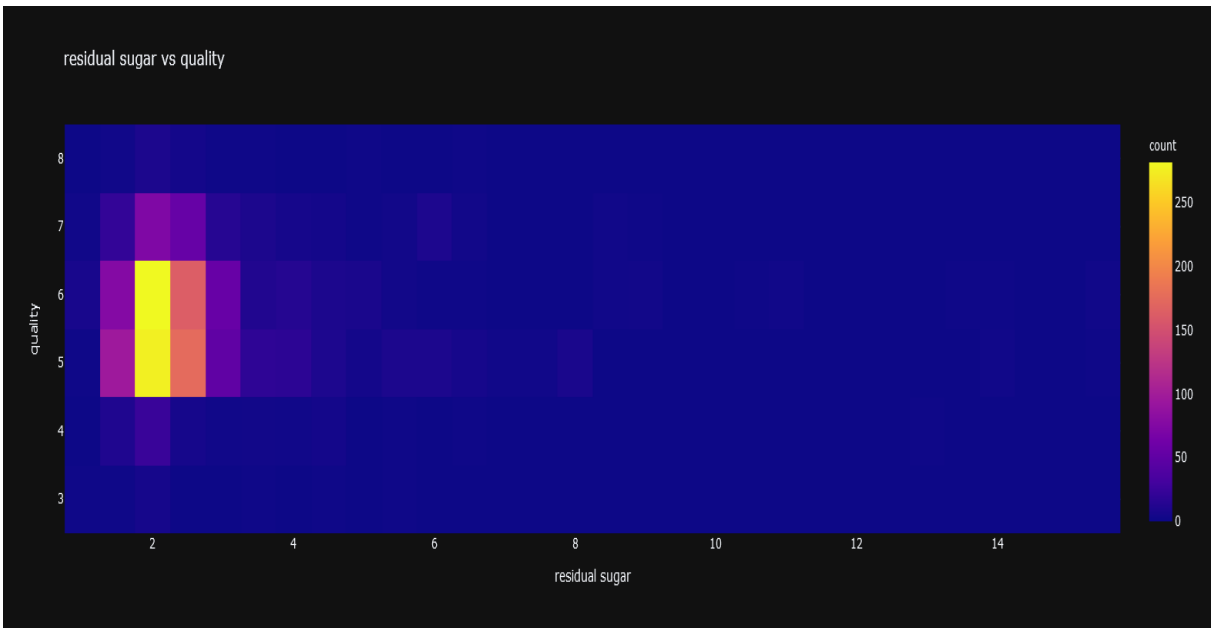
Free sulfur dioxide vs quality



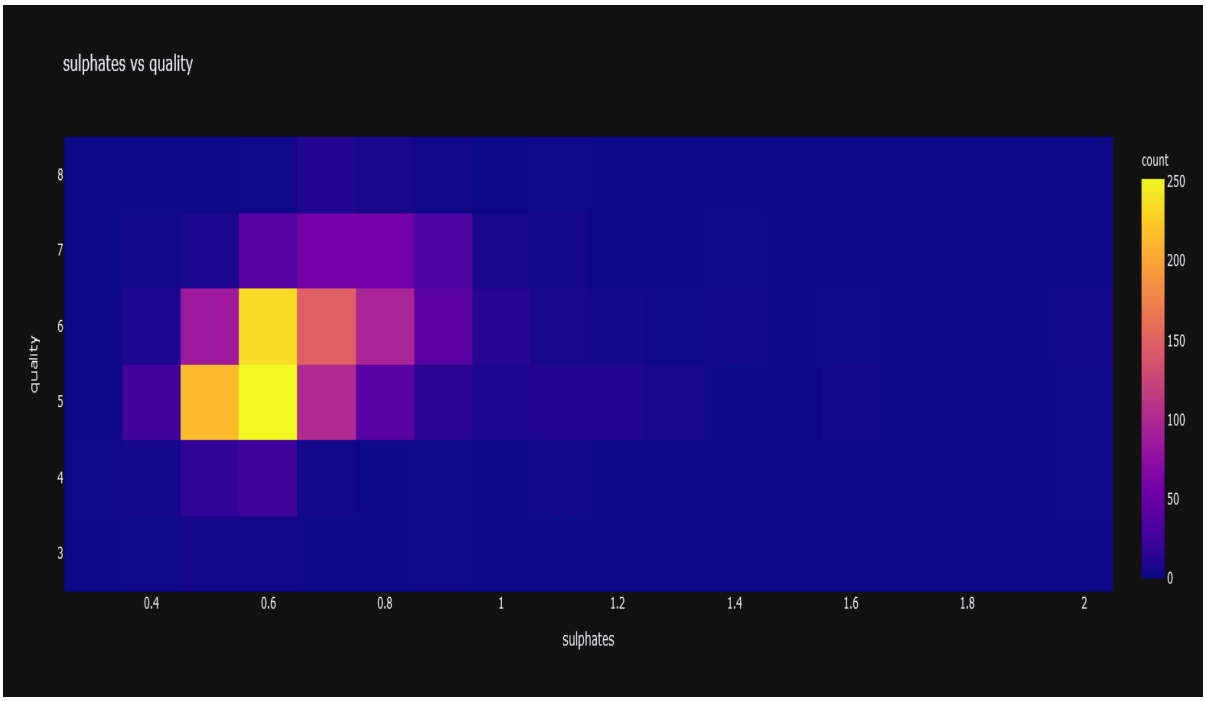
pH vs quality



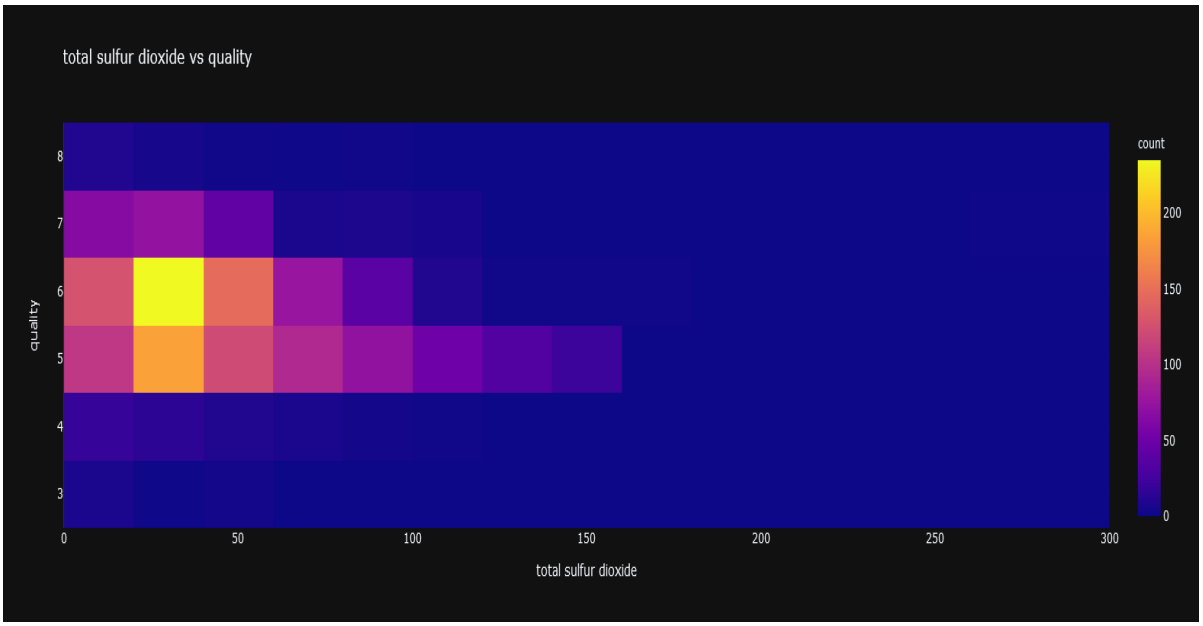
Residual sugar vs quality



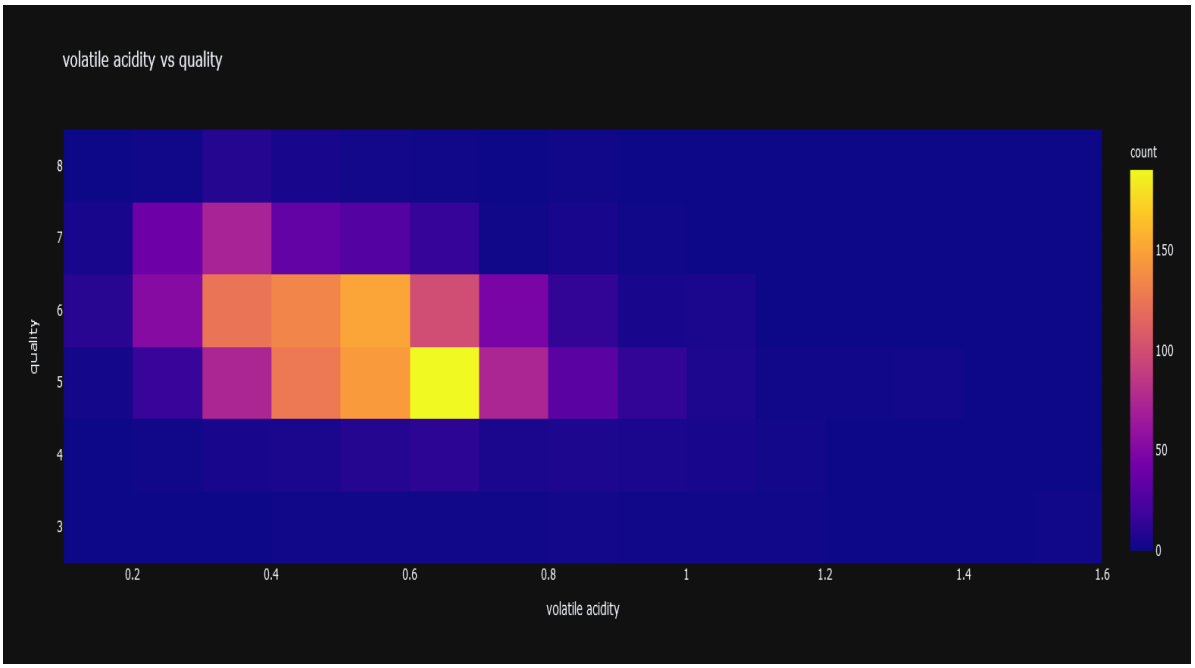
Sulphates vs quality



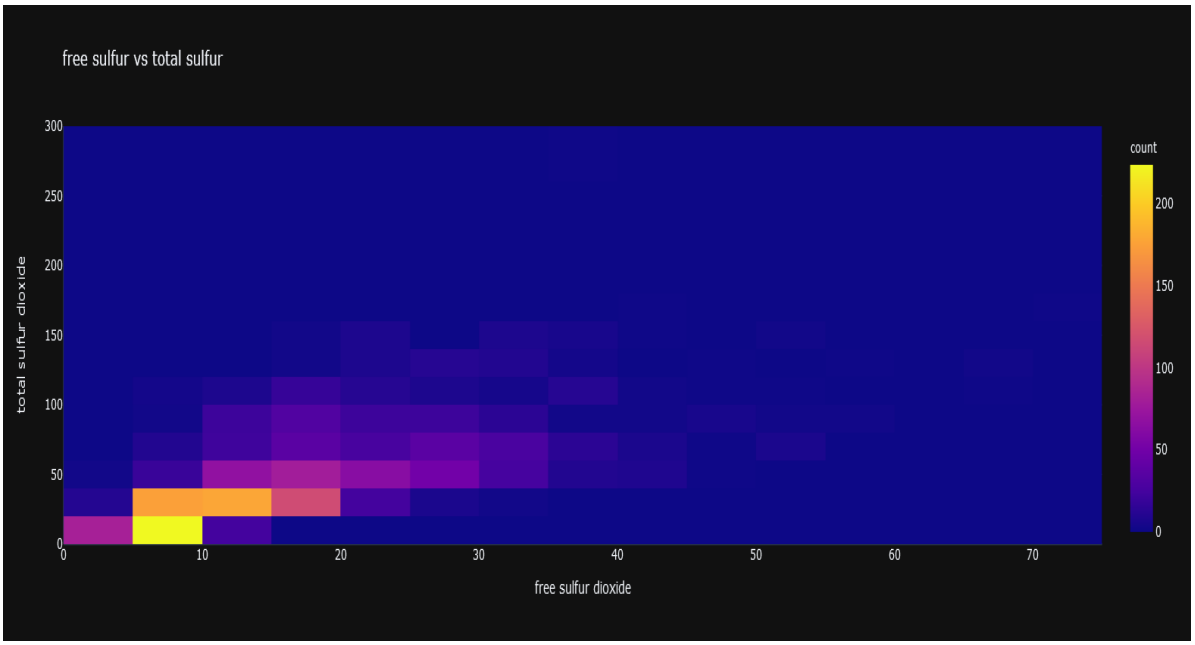
Total sulfur dioxide vs quality



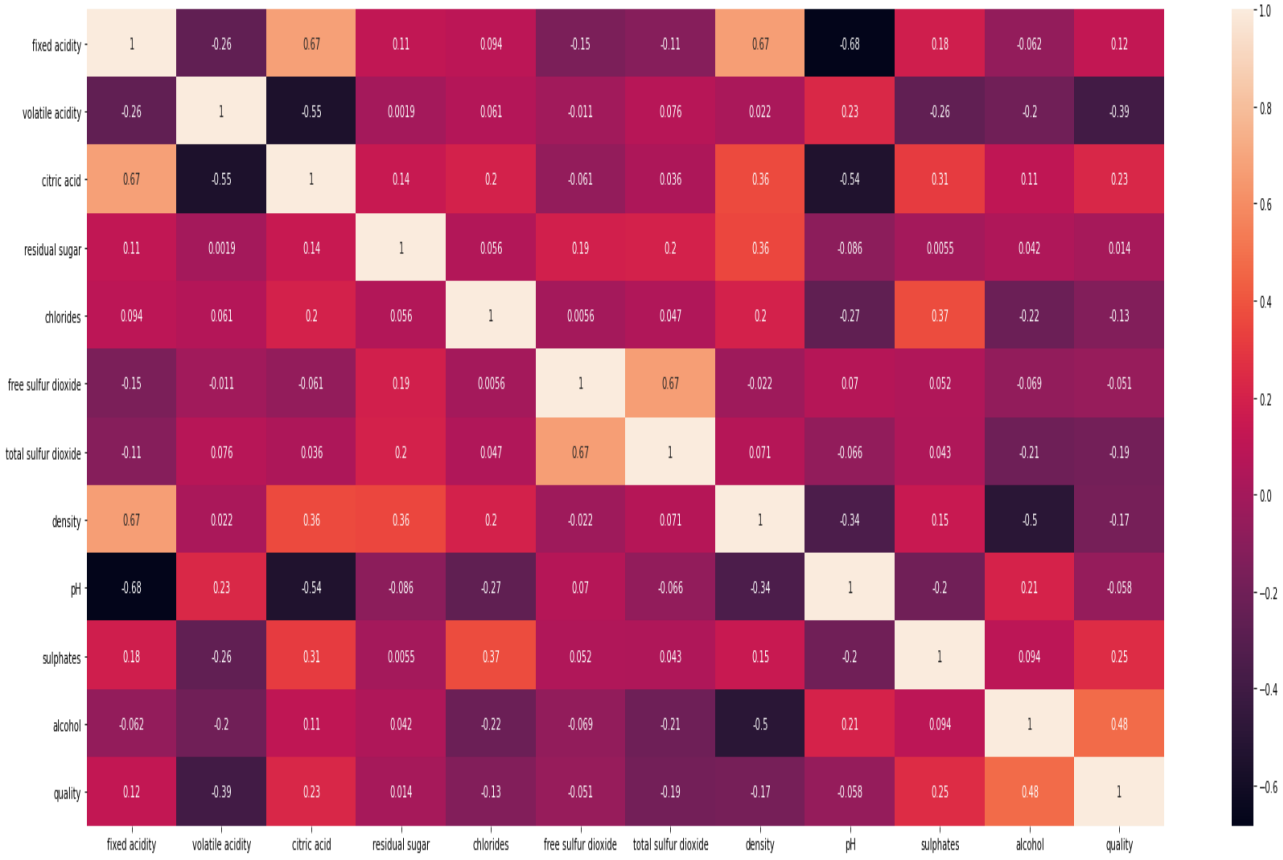
Volatile acidity vs quality



Free sulfur vs total sulfur



- Correlation heat map



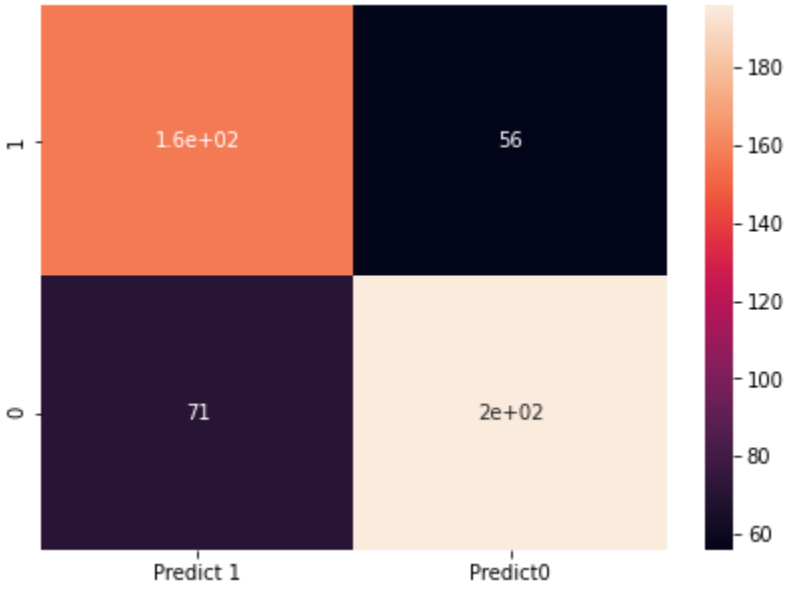
- Summary of training at least three classification models which should be variations that cover using a simple logistic regression as a baseline and other classification methods. Preferably, all use the same training and test splits, or the same cross-validation method.
  - I used the following classification method for choosing the best model suitable for this problem: -
    - Logistic regression
    - KNN classifier
    - Decision tree
    - Pruned decision tree
    - Random forest
    - Support vector machine with 4 different kernels
    - AdaBoost
    - Gradient boost
    - Bagging
    - Naïve Bayes

- Following where the classification report and confusion matrix of the models: -

Logistic regression

Classification Report

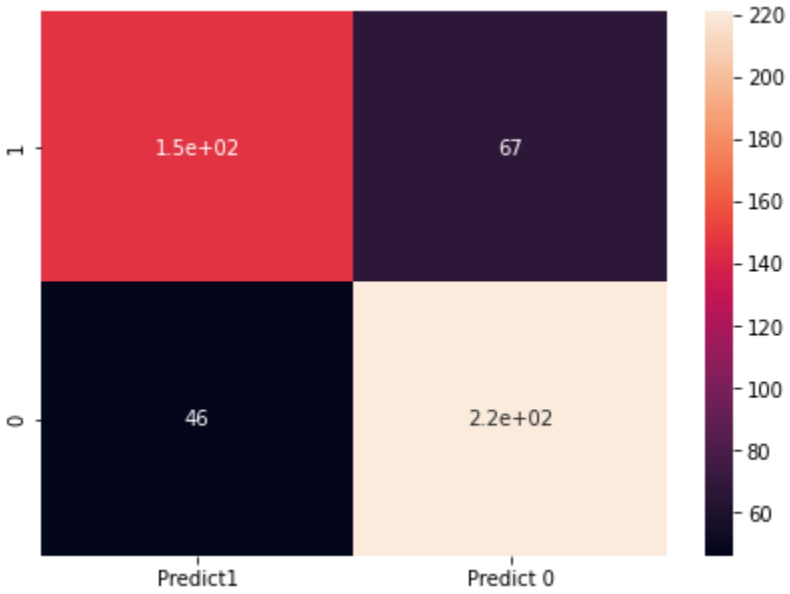
	precision	recall	f1-score	support
1	0.78	0.73	0.76	267
0	0.69	0.74	0.71	213
accuracy			0.74	480
macro avg	0.73	0.74	0.73	480
weighted avg	0.74	0.74	0.74	480



KNN classifier

Classification Report

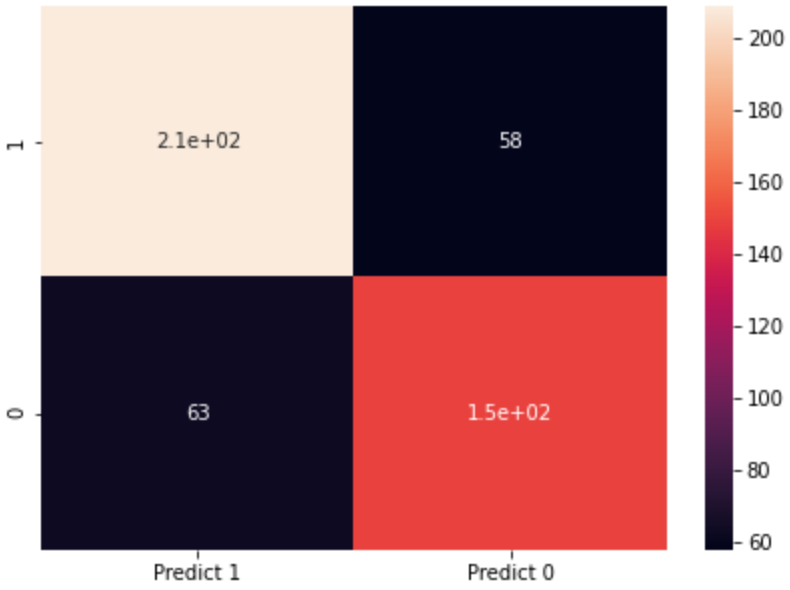
	precision	recall	f1-score	support
1	0.77	0.83	0.80	267
0	0.76	0.69	0.72	213
accuracy			0.76	480
macro avg	0.76	0.76	0.76	480
weighted avg	0.76	0.76	0.76	480



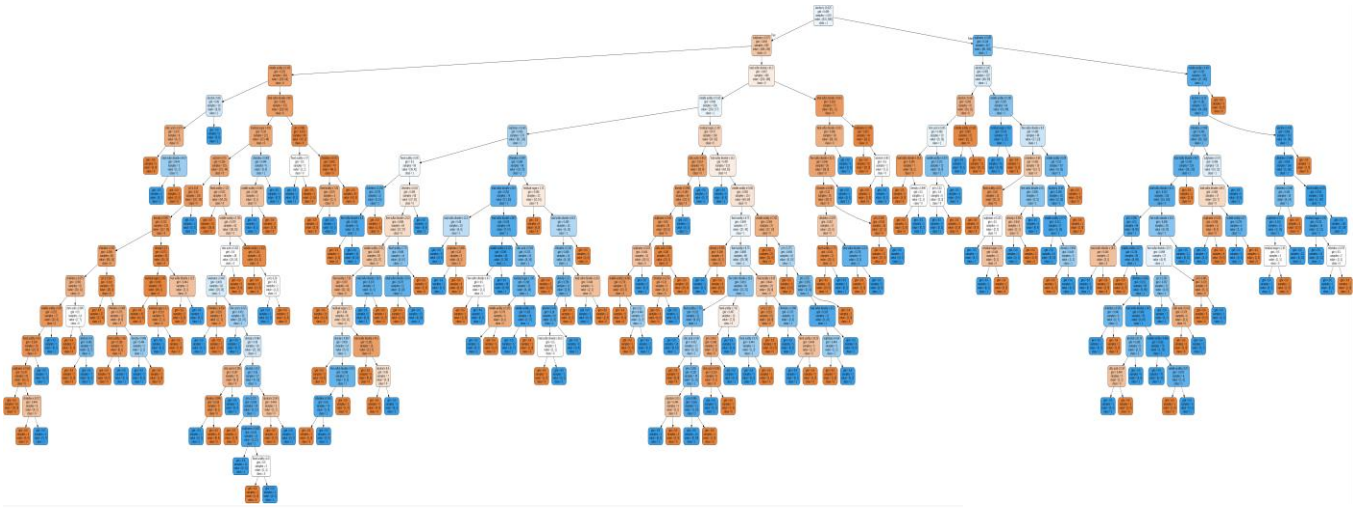
Naïve Bayes Classifier

Classification Report

	precision	recall	f1-score	support
1	0.77	0.78	0.78	267
0	0.72	0.70	0.71	213
accuracy			0.75	480
macro avg	0.74	0.74	0.74	480
weighted avg	0.75	0.75	0.75	480

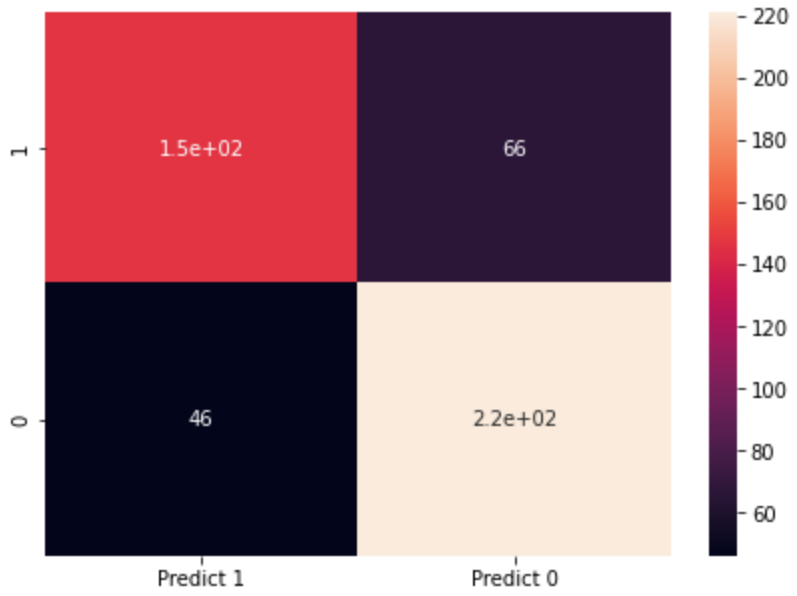


Decision tree

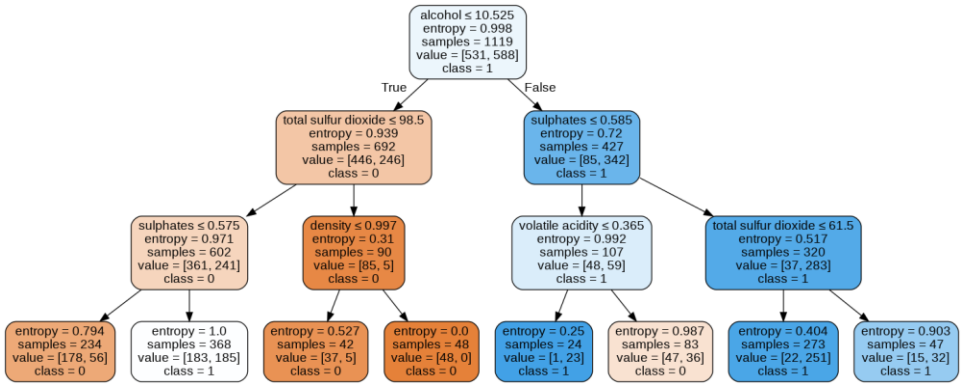


Classification Report

	precision	recall	f1-score	support
1	0.77	0.83	0.80	267
0	0.76	0.69	0.72	213
accuracy			0.77	480
macro avg	0.77	0.76	0.76	480
weighted avg	0.77	0.77	0.77	480



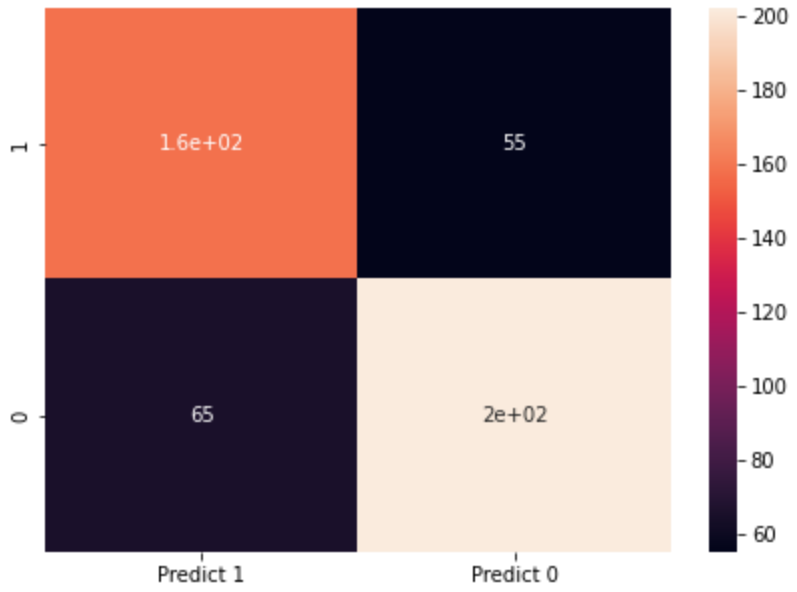
Pruned Decision tree



AdaBoost Classifier

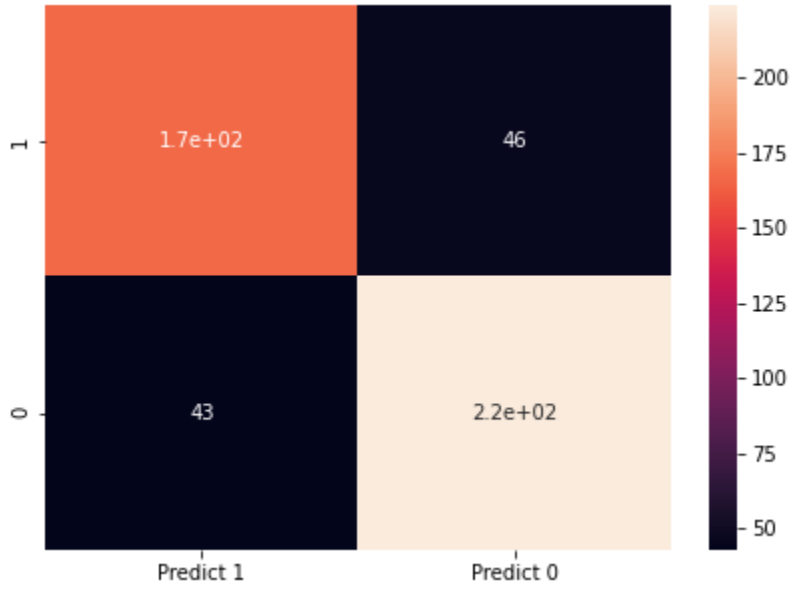
Classification Report

	precision	recall	f1-score	support
1	0.79	0.76	0.77	267
0	0.71	0.74	0.72	213
accuracy			0.75	480
macro avg	0.75	0.75	0.75	480
weighted avg	0.75	0.75	0.75	480



Random Forest

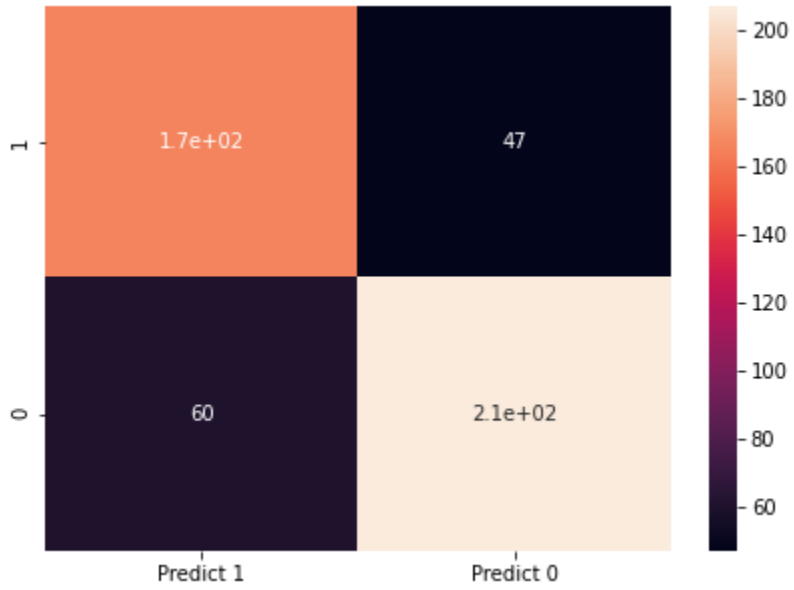
```
Classification Report
              precision    recall  f1-score   support

           1       0.83      0.84      0.83       267
           0       0.80      0.78      0.79       213

    accuracy                           0.81       480
   macro avg       0.81      0.81      0.81       480
weighted avg       0.81      0.81      0.81       480
```



Gradient Boost

Classification Report

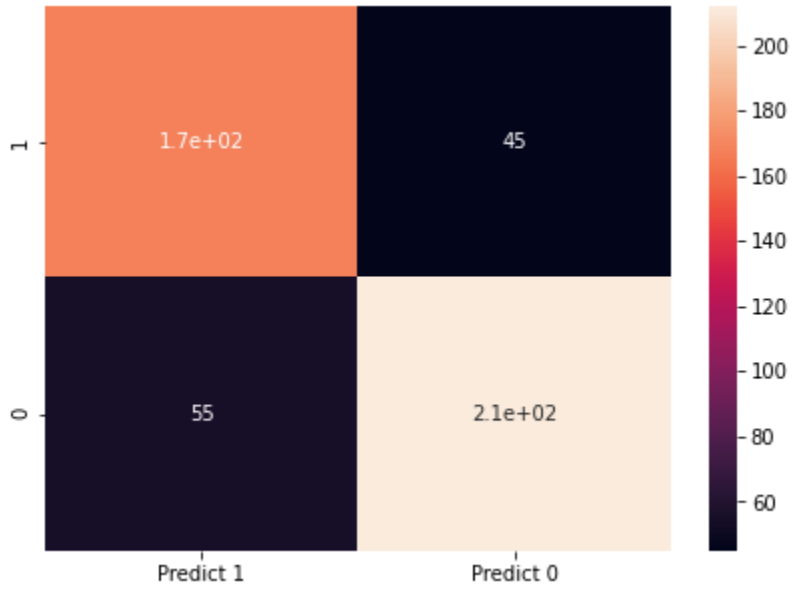
	precision	recall	f1-score	support
1	0.81	0.78	0.79	267
0	0.73	0.78	0.76	213
accuracy			0.78	480
macro avg	0.77	0.78	0.78	480
weighted avg	0.78	0.78	0.78	480



Bagging Classifier

Classification Report

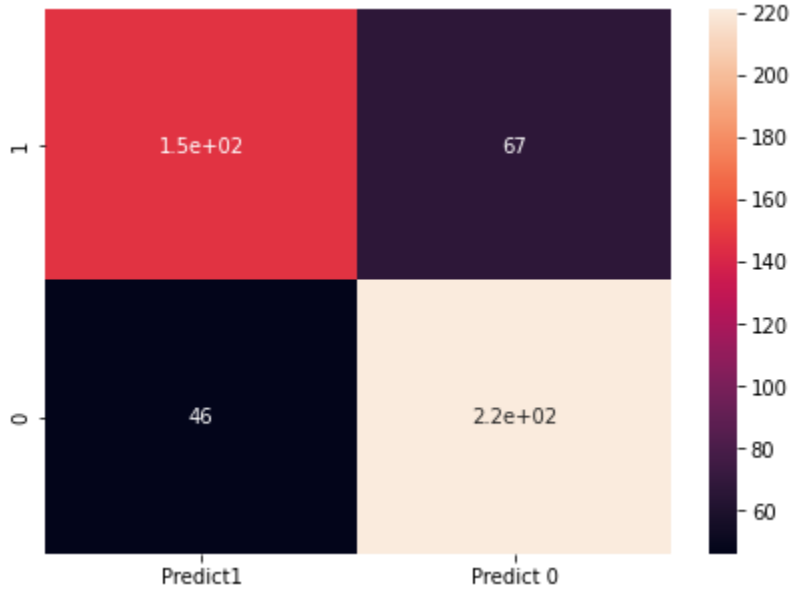
	precision	recall	f1-score	support
1	0.82	0.79	0.81	267
0	0.75	0.79	0.77	213
accuracy			0.79	480
macro avg	0.79	0.79	0.79	480
weighted avg	0.79	0.79	0.79	480



SVM (Linear kernel)

Classification Report

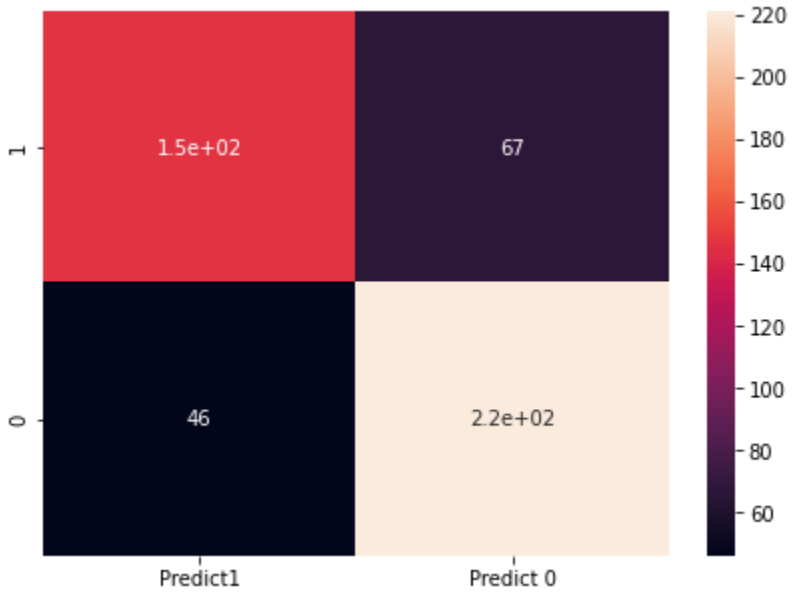
	precision	recall	f1-score	support
1	0.78	0.67	0.72	267
0	0.65	0.76	0.70	213
accuracy			0.71	480
macro avg	0.71	0.71	0.71	480
weighted avg	0.72	0.71	0.71	480



SVM(rbf Kernel)

Classification Report

	precision	recall	f1-score	support
1	0.63	0.88	0.73	267
0	0.70	0.35	0.47	213
accuracy			0.65	480
macro avg	0.67	0.62	0.60	480
weighted avg	0.66	0.65	0.62	480



- Insights derived from these models whereas follows:
  - The best model for this dataset was random forest because it has most accuracy of 80 compared to other model at 95% confidence level
  - The classification report also shows that it has higher F1 score and good recall and precision compared to other models

- Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model adding specific data features to achieve a better explanation or a better prediction.
    - This model can be further improved by Hyperparameter tuning
    - The PCA approach can also help in improve the accuracy further
    - This model can also be compared with other algorithms like Gridsearch CV or Randomsearch CV etc.