Predicting Wine Quality

EASTWOOD LOFTUS

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

Summary of Methodologies

- Data Collection
- EDA
- Model Building

Summary of Results

- RFC highest accuracy
- Minor-major increases in accuracy after parameter fitting.

Introduction

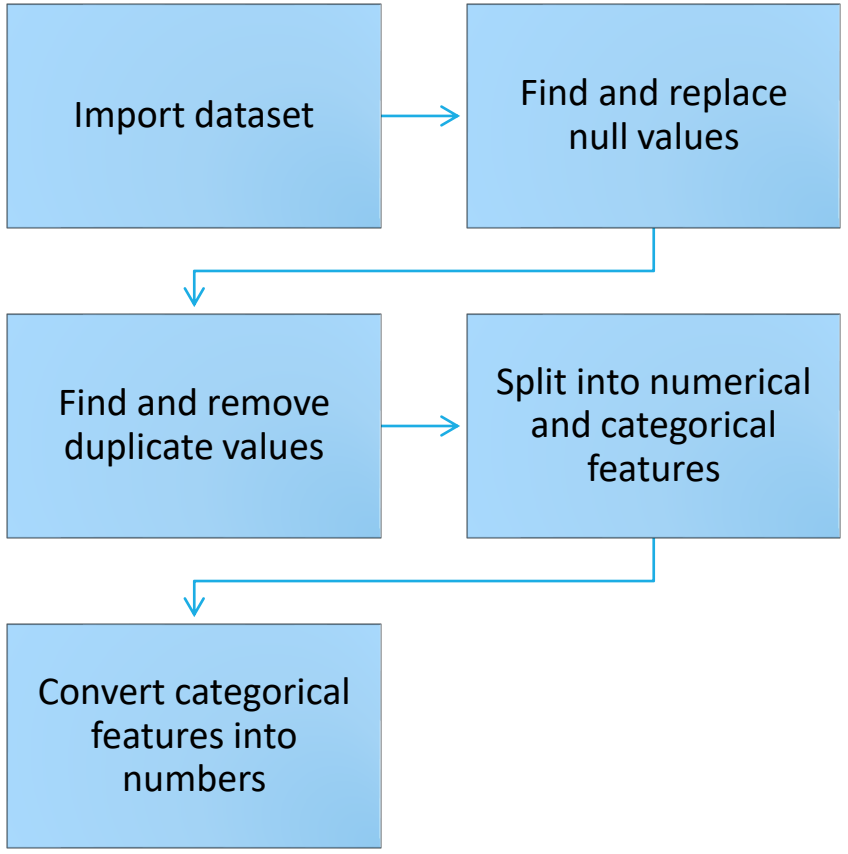
Project Background and Context

- For this project, I will attempt to predict wine quality by training various machine learning models with data from a given dataset, and showcasing which has the highest accuracy.

Methodology

Methodology – Data Collection

- Import dataset from link
- Find the total null values, and replace with specific mean values for each column.
- Find and remove all duplicate entries for the dataset, so the model will be less biased.
- Split numerical and categorical values, and shift categorical values into numerical for later analysis and model building



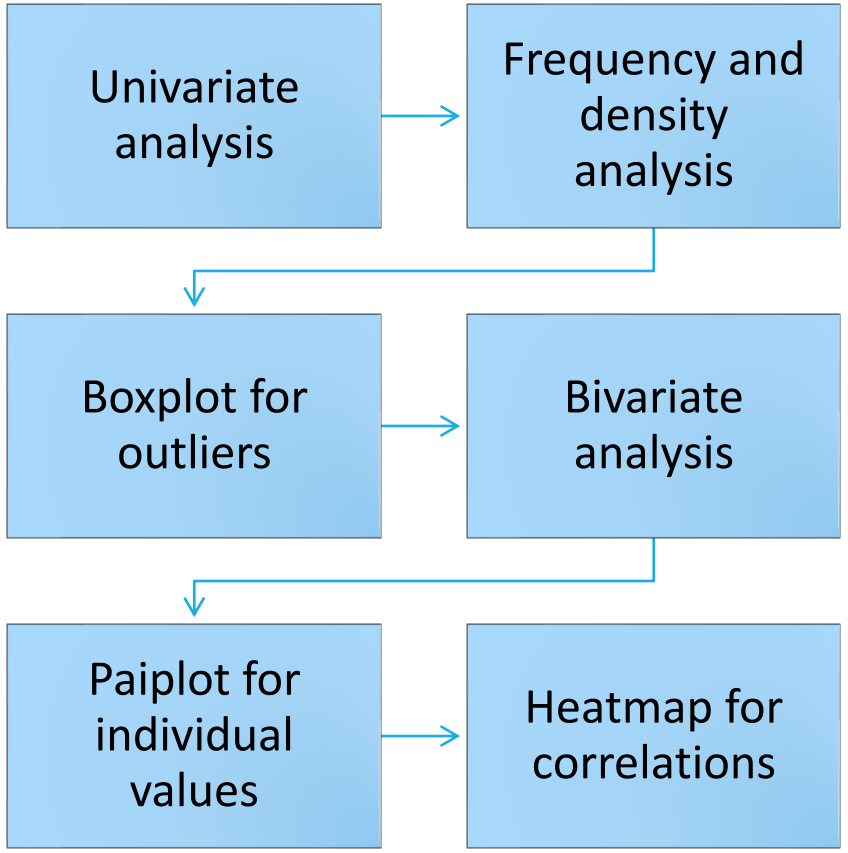
Methodology – EDA

Univariate Analysis

- Frequency and density analysis reveals the curve of the values, while a boxplot makes it easy to identify outliers and the mean values.

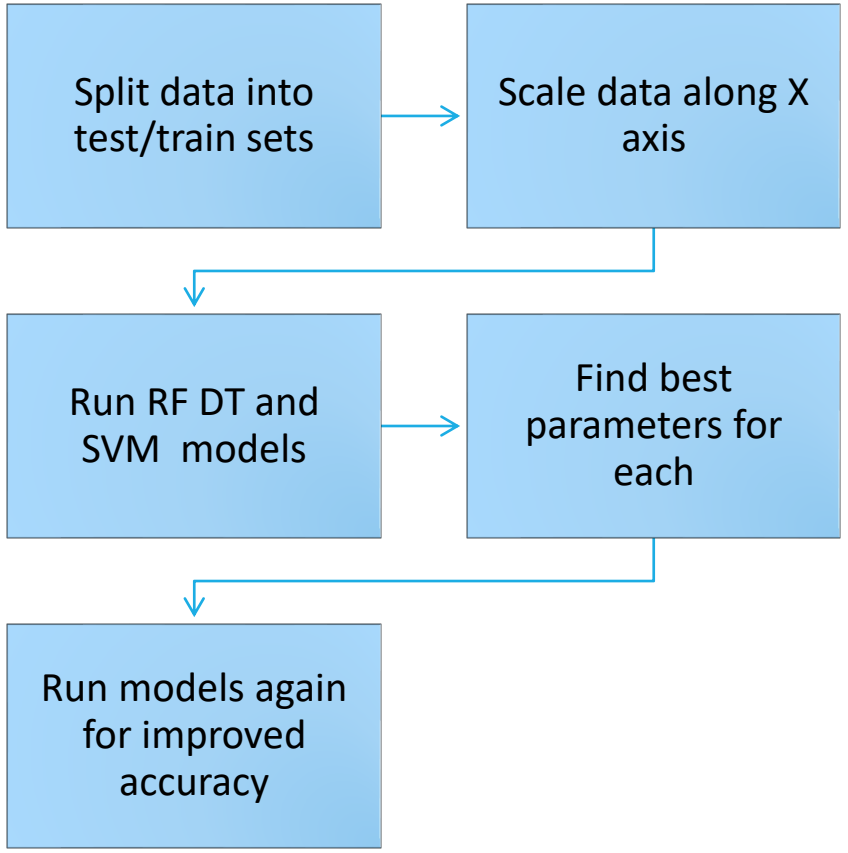
Bivariate Analysis

- Using pairplot and a heatmap of the data, we can see the correlations in the data and the clusters formed in our values.



Methodology – Model Building

- Run evaluation on scaled data to determine accuracy, recall and f1-score. Generate a confusion matrix and determine overall accuracy.
- Using a Gridsearch, find the best parameters for the model and determine the new accuracy.



Results

Results of EDA – Univariate Analysis

Nothing truly unique can be spotted during this analysis, with most graphs having a decent bell curve and some being slightly biased towards lower values.

Results of EDA – Bivariate Analysis

From viewing the heatmap, we can see that there are strong correlations between:

- The free sulphur dioxide and total sulphur dioxide (0.72)
- Residual sugar and Density (0.52)
- Density and Alcohol (-0.67)

and medium:

- Residual sugar and total sulfur dioxide (0.49)
- Fixed acidity and density (0.48)
- Alcohol and quality (0.47)

and more.

Results of Data Models – RFC

On inspection of the generated report, we can see that the overall accuracy score of the Random Forest is ~55.8%. However, upon finding and fitting the best parameters, the accuracy increases to ~56.4%, which can be considered a minor increase.

```
rfc_gs.best_score_
```

```
np.float64(0.5639137606682003)
```

```
rfc_gs.best_params_
```

```
{'rfc__max_depth': 17, 'rfc__max_features': 'sqrt', 'rfc__n_estimators': 33}
```

```
Classification report of the Model:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         9
           4       0.00      0.00      0.00        32
           5       0.61      0.62      0.61       356
           6       0.54      0.68      0.60       469
           7       0.50      0.31      0.39       172
           8       0.25      0.04      0.07        26

    accuracy                           0.56      1064
   macro avg       0.32      0.28      0.28      1064
weighted avg       0.53      0.56      0.53      1064

Confusion Matrix of the given Model:
 [[  0   1   4   3   1   0]
 [  0   0  21  11   0   0]
 [  0   0 220 134   2   0]
 [  0   0 109 319  39   2]
 [  0   0   8 109  54   1]
 [  0   0   1  12  12   1]]
Accuracy score of the Model:
0.5582706766917294
```

Results of Data Models – DTC

On inspection of the generated report, we can see that the overall accuracy score of the Decision Tree is ~46.8%. However, upon finding and fitting the best parameters, the accuracy increases to ~54%, which can be considered a major increase.

```
dtc_gs.best_params_
```

```
{'dtc__criterion': 'gini',
 'dtc__max_depth': 5,
 'dtc__max_features': 8,
 'dtc__min_samples_leaf': 2}
```

```
dtc_gs.best_score_
```

```
np.float64(0.5390052023854841)
```

```
Classification report of the Model:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         9
           4       0.10      0.12      0.11        32
           5       0.52      0.48      0.50       356
           6       0.52      0.54      0.53       469
           7       0.40      0.39      0.40       172
           8       0.05      0.08      0.06        26

    accuracy                           0.47      1064
   macro avg       0.27      0.27      0.27      1064
weighted avg       0.47      0.47      0.47      1064

Confusion Matrix of the given Model:
 [[  0   0   4   3   2   0]
 [  0   4  12  14   1   1]
 [  1  20 170 137  25   3]
 [  0  16 120 255  62  16]
 [  0   1  22  66  67  16]
 [  0   0   1  13  10   2]]
Accuracy score of the Model:
0.4680451127819549
```

Results of Data Models – SVC

On inspection of the generated report, we can see that the overall accuracy score of the Random Forest is ~55.8%. However, upon finding and fitting the best parameters, the accuracy increases to ~56.1%, which can be considered a minor increase.

```
svc_gs.best_params_
```

```
{'svc__C': 1, 'svc__kernel': 'rbf'}
```

```
svc_gs.best_score_
```

```
np.float64(0.5613260841980989)
```

```
Classification report of the Model:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         9
           4       0.00      0.00      0.00        32
           5       0.60      0.61      0.61       356
           6       0.54      0.73      0.62       469
           7       0.53      0.20      0.29       172
           8       0.00      0.00      0.00        26

    accuracy                           0.56      1064
   macro avg       0.28      0.26      0.25      1064
weighted avg       0.52      0.56      0.52      1064

Confusion Matrix of the given Model:
 [[  0   0   5   4   0   0]
 [  0   0  23   9   0   0]
 [  0   0 218 138   0   0]
 [  0   0 106 342  21   0]
 [  0   0   9 129  34   0]
 [  0   0   0  17   9   0]]
Accuracy score of the Model:
0.5582706766917294
```

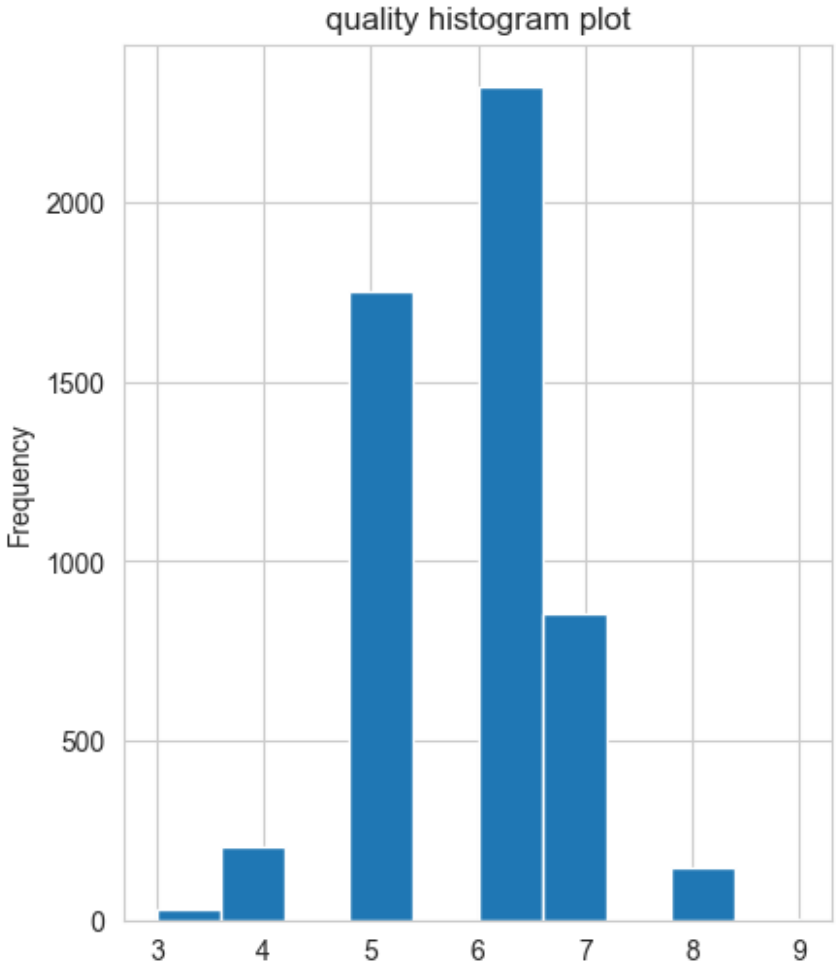
Conclusion

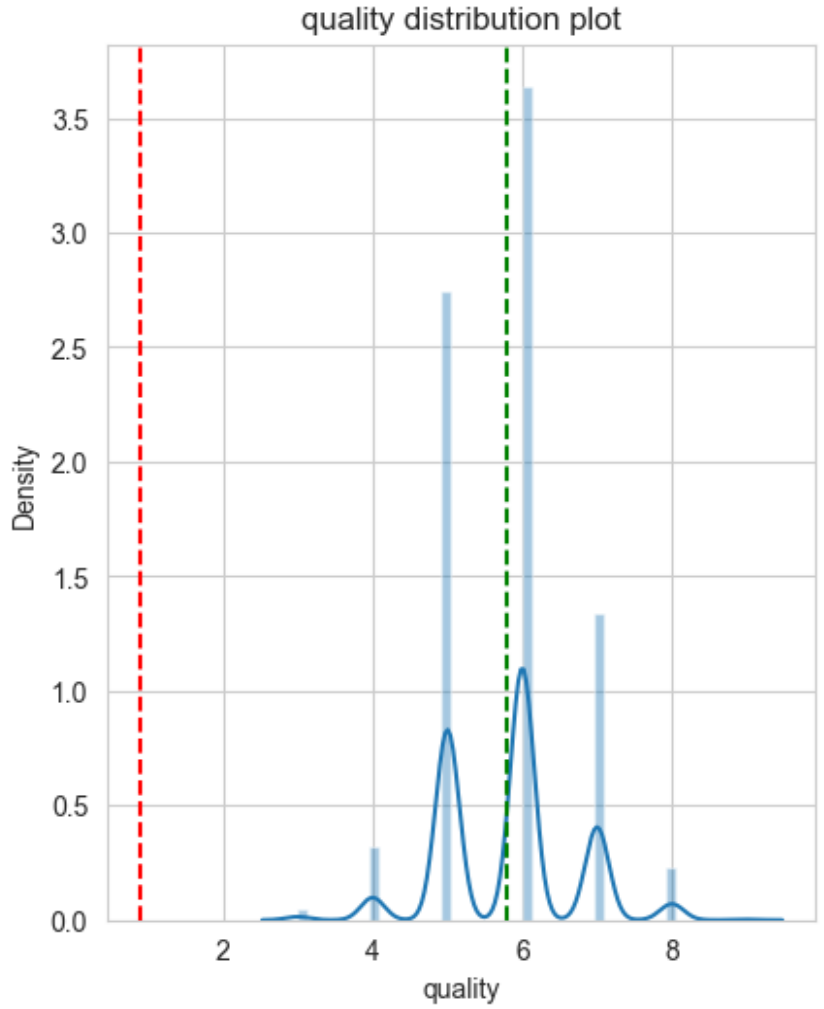
Conclusion

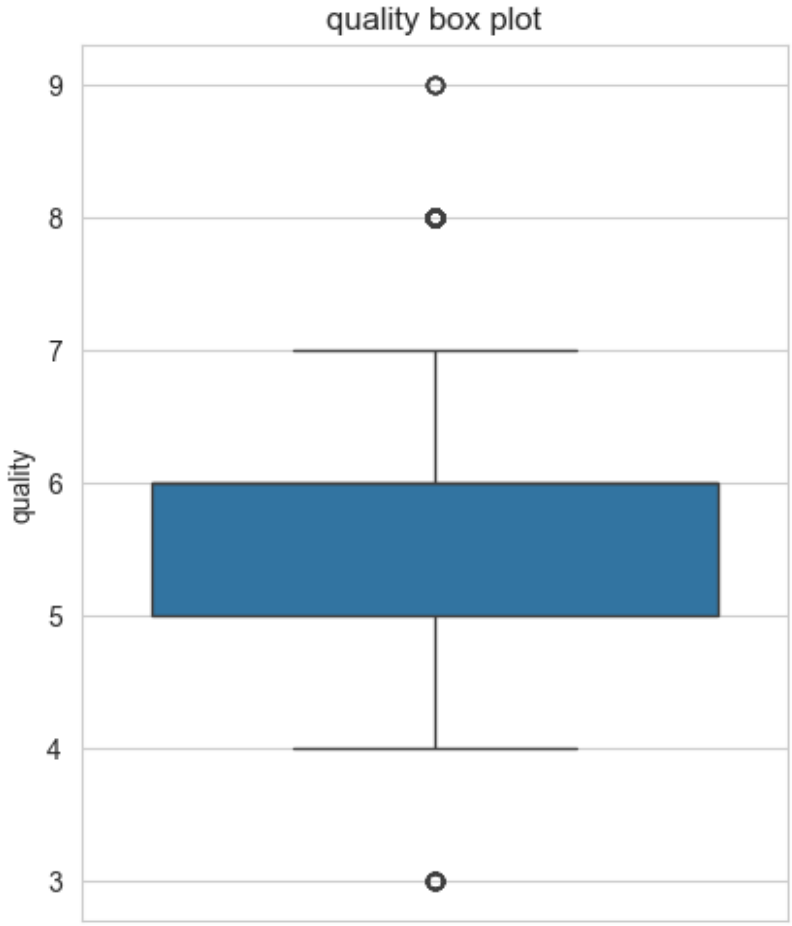
In conclusion, we can determine that should a model be used for prediction, the Random Forest and the SVM can be recommended from our testing, as they have really similar scores before and after the parameter fitting.

Appendix

REPOSITORY FOR IMAGES AND FILES

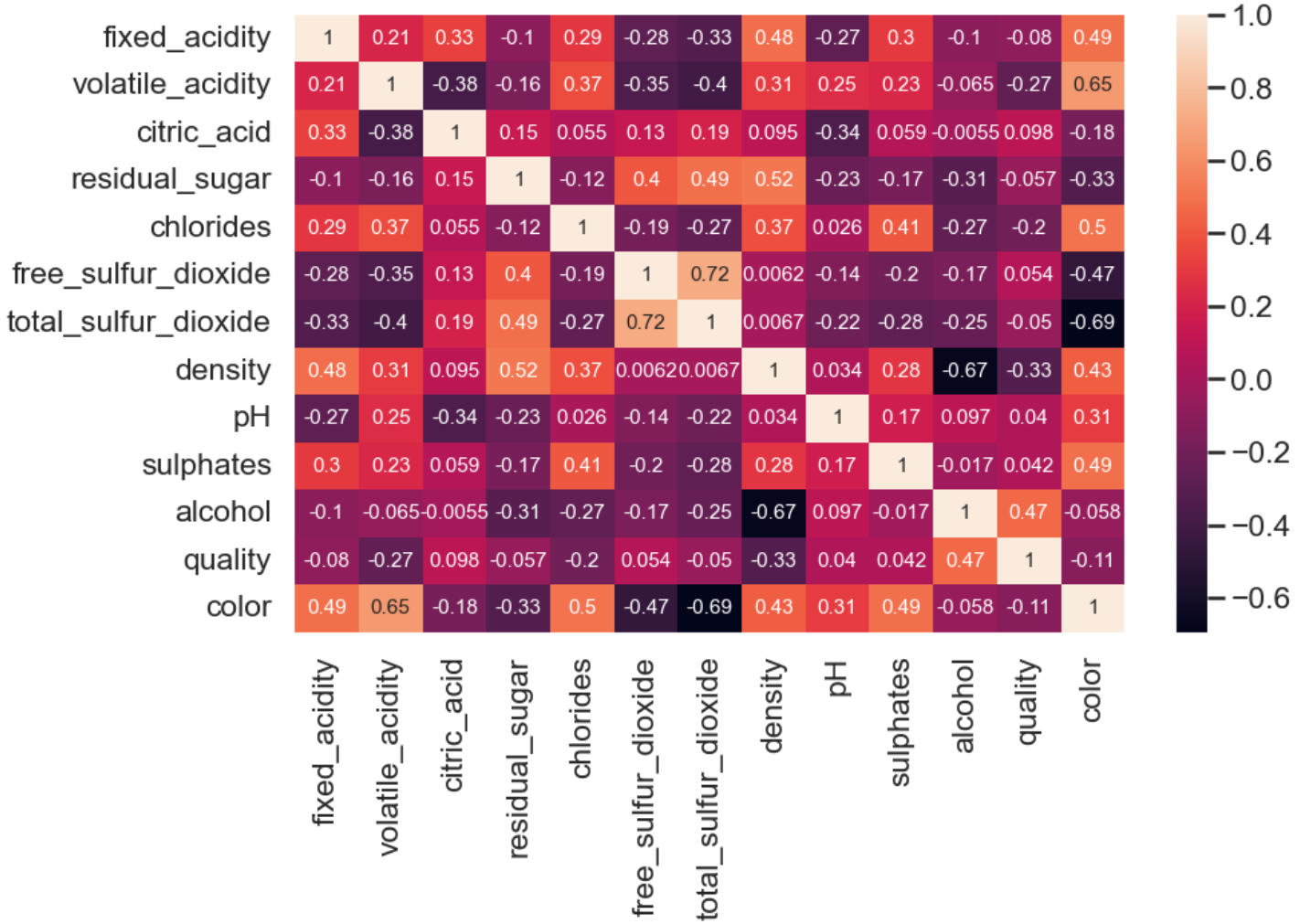






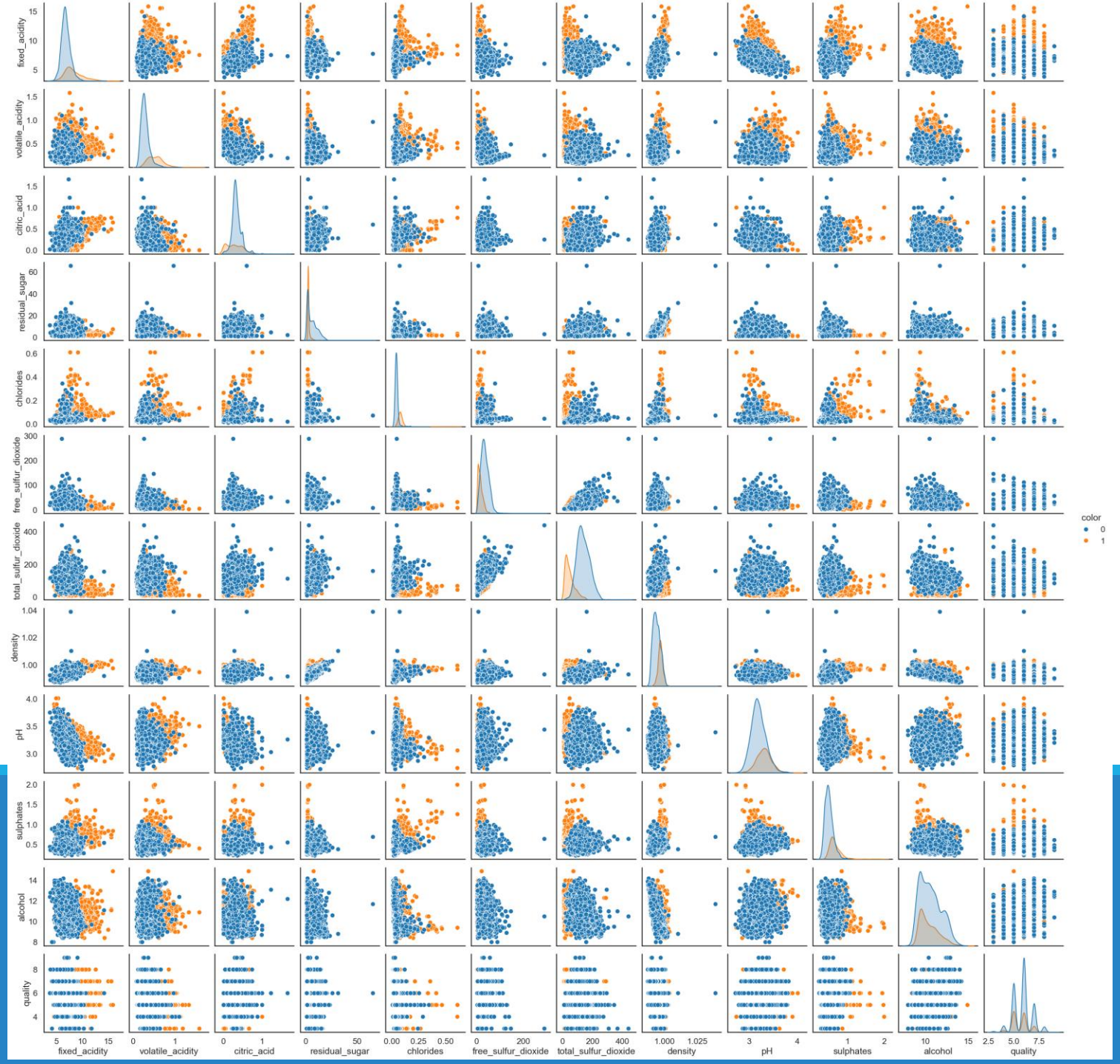
Univariate Analysis of Quality

Sample image of data analysis



Heatmap of Correlations between Columns

Pairplot of Values



```
rfc_gs.best_score_
```

```
np.float64(0.5639137606682003)
```

```
rfc_gs.best_params_
```

```
{'rfc__max_depth': 17, 'rfc__max_features': 'sqrt', 'rfc__n_estimators': 33}
```

```
dtc_gs.best_params_
```

```
{'dtc__criterion': 'gini',
 'dtc__max_depth': 5,
 'dtc__max_features': 8,
 'dtc__min_samples_leaf': 2}
```

```
dtc_gs.best_score_
```

```
np.float64(0.5390052023854841)
```

```
svc_gs.best_params_
```

```
{'svc__C': 1, 'svc__kernel': 'rbf'}
```

```
svc_gs.best_score_
```

```
np.float64(0.5613260841980989)
```

Results of Parameter Search