Wine Quality Classification

Nirajan Bekoju
Bishal Adhikari
Rijan Pokhrel
Manoj Khatri

Topics of discussion

- Problem Statement
- Feature Engineering and EDA
- Model development and Evaluation
- Demo

Problem Statement

- The dataset describes the amount of various chemicals present in wine and their effect on it's quality.
- The datasets can be viewed as classification or regression tasks.
- The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones)
- The complexity arises due to the fact that the dataset has fewer samples, & is highly imbalanced

Feature Engineering and EDA

Dataset

Out[3]:

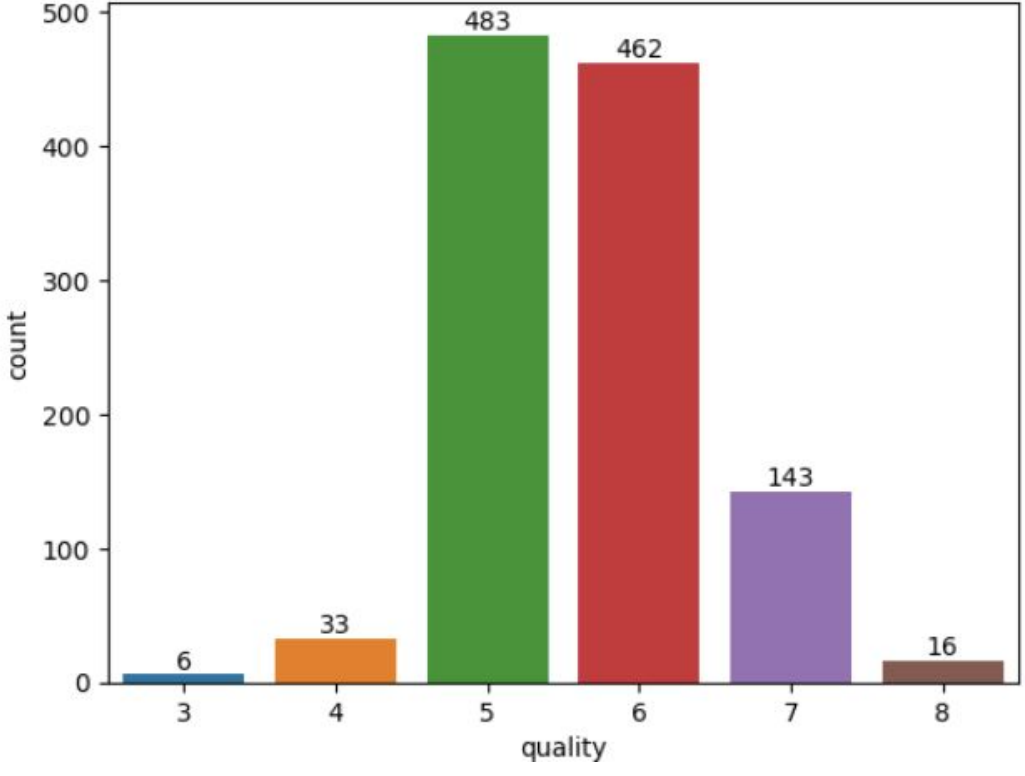
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	2
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	3
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	4

In [4]:
```
# Dimension of dataset
df.shape
```

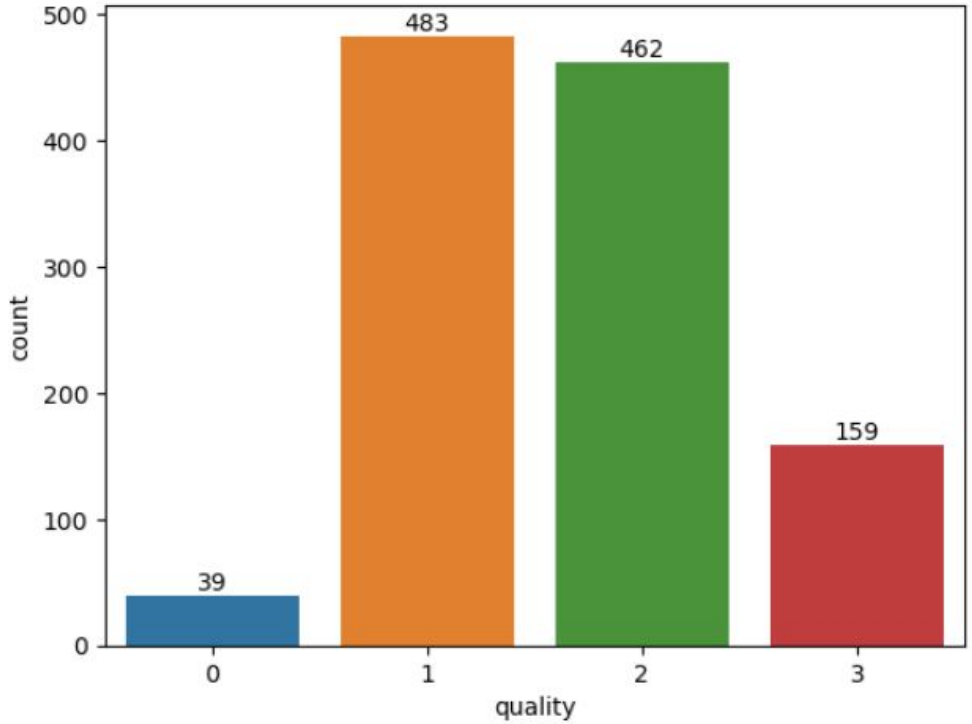
Out[4]:
```
(1143, 13)
```

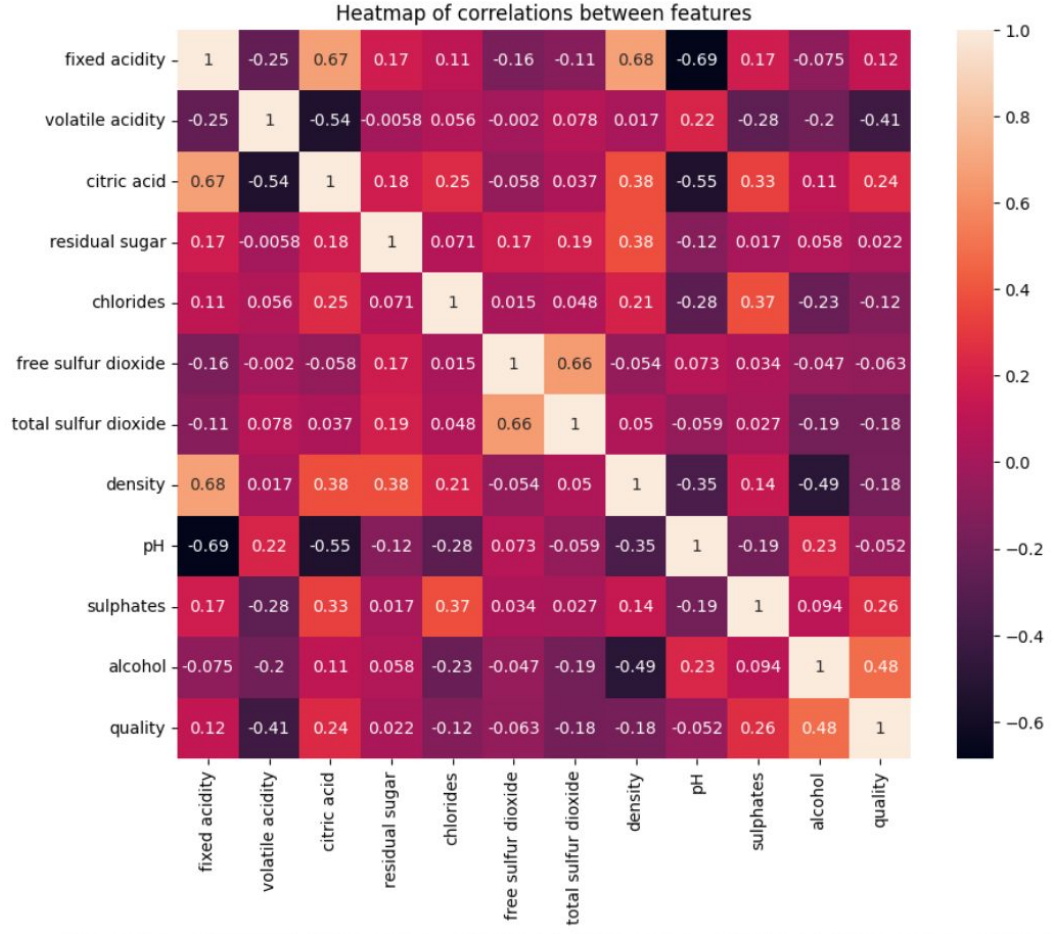
```
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1143 non-null   float64
 1   volatile acidity      1143 non-null   float64
 2   citric acid           1143 non-null   float64
 3   residual sugar        1143 non-null   float64
 4   chlorides             1143 non-null   float64
 5   free sulfur dioxide   1143 non-null   float64
 6   total sulfur dioxide  1143 non-null   float64
 7   density               1143 non-null   float64
 8   pH                    1143 non-null   float64
 9   sulphates             1143 non-null   float64
 10  alcohol               1143 non-null   float64
 11  quality               1143 non-null   int64
 12  Id                    1143 non-null   int64
dtypes: float64(11), int64(2)
memory usage: 116.2 KB
```

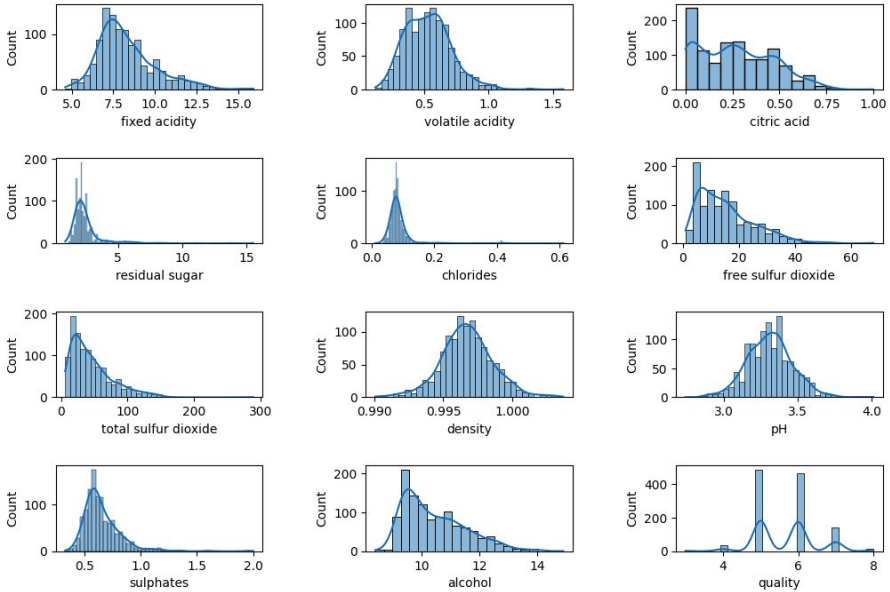
The count plot of the whole dataset on the basis of the quality of wine is shown aside.



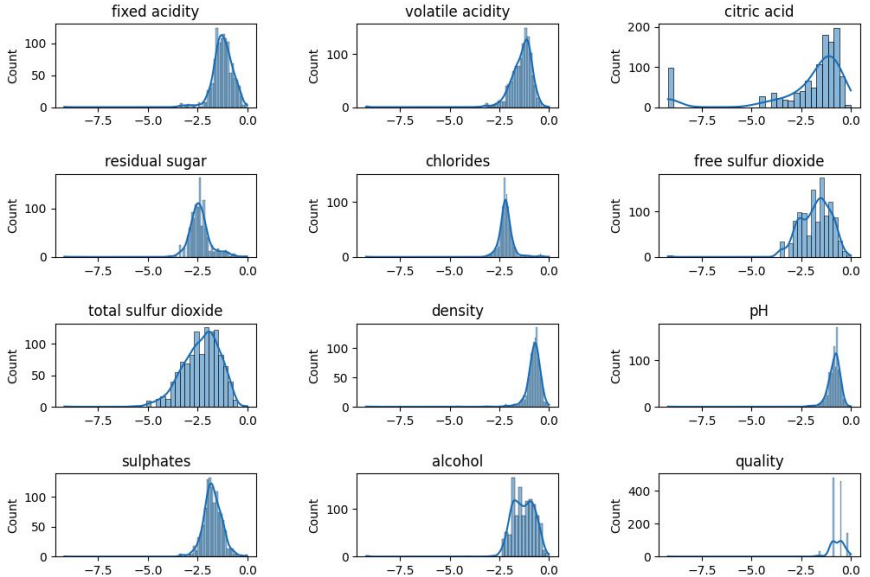
- label 3 and label 4 => label 0
- label 5 => label 1
- label 6 => label 2
- label 7 and label 8 => label 3



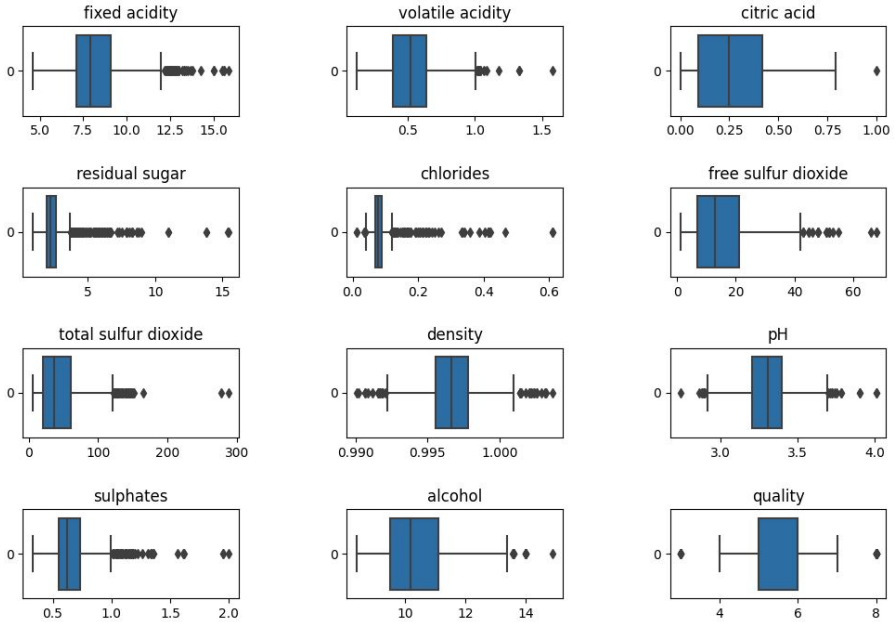

Heatmap of correlations between features
fixed acidity 1 -0.25 0.67 0.17 0.11 -0.16 -0.11 0.68 -0.69 0.17 -0.075 0.12
volatile acidity -0.25 1 -0.54 -0.0058 0.056 -0.002 0.078 0.017 0.22 -0.28 -0.2 -0.41
citric acid 0.67 -0.54 1 0.18 0.25 -0.058 0.037 0.38 -0.55 0.33 0.11 0.24
residual sugar 0.17 -0.0058 0.18 1 0.071 0.17 0.19 0.38 -0.12 0.017 0.058 0.022
chlorides 0.11 0.056 0.25 0.071 1 0.015 0.048 0.21 -0.28 0.37 -0.23 -0.12
free sulfur dioxide -0.16 -0.002 -0.058 0.17 0.015 1 0.66 -0.054 0.073 0.034 -0.047 -0.063
total sulfur dioxide -0.11 0.078 0.037 0.19 0.048 0.66 1 0.05 -0.059 0.027 -0.19 -0.18
density 0.68 0.017 0.38 0.38 0.21 -0.054 0.05 1 -0.35 0.14 -0.49 -0.18
pH -0.69 0.22 -0.55 -0.12 -0.28 0.073 -0.059 -0.35 1 -0.19 0.23 -0.052
sulphates 0.17 -0.28 0.33 0.017 0.37 0.034 0.027 0.14 -0.19 1 0.094 0.26
alcohol -0.075 -0.2 0.11 0.058 -0.23 -0.047 -0.19 -0.49 0.23 0.094 1 0.48
quality 0.12 -0.41 0.24 0.022 -0.12 -0.063 -0.18 -0.18 -0.052 0.26 0.48 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6



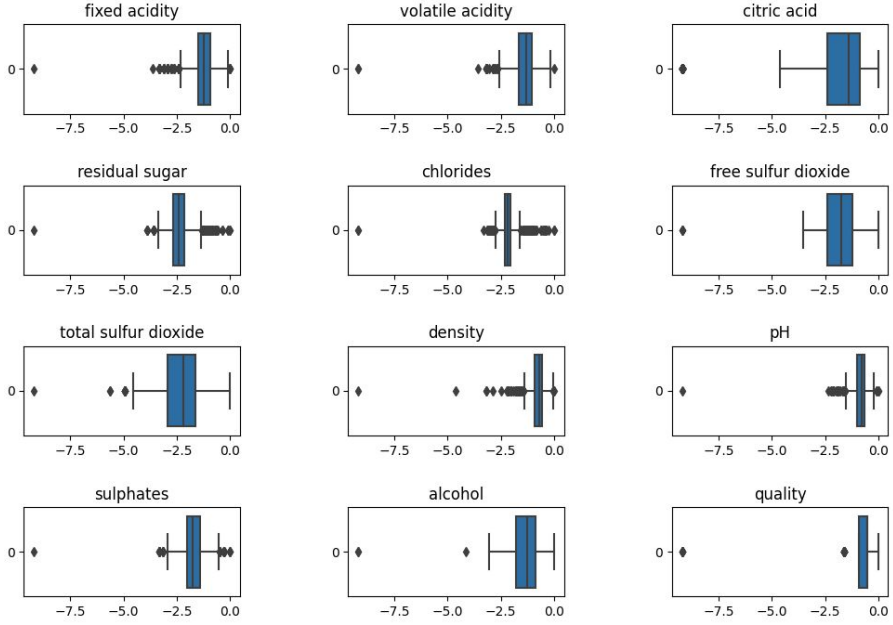
Data Distribution



Data Distribution after min max scaling and log transformation



Box plot of features



Box plot after min max scaling and log transformation

Model Development and Evaluation

multi_class	solver	micro avg	macro avg	weighted avg
ovr	newton-cg	0.64	0.49	0.62
multinomial	newton-cg	0.66	0.53	0.65

Table 5.1: Logistic regression hyperparameter and evaluation

id	learning_rate	max_depth	n_estimators	subsample	micro	macro	weighted
xgb_v1	0.0824	16	113	0.5394	0.64	0.49	0.65
xgb_v2	0.0903	19	127	0.6520	0.63	0.48	0.63

Table 5.2: XGBClassifier hyperparameter and evaluation metrics

model	micro	macro	weighted
AdaBoost	0.51	0.46	0.55
Gradient Boosting	0.61	0.50	0.61
xgb_v1	0.64	0.49	0.65

Table 5.3: Boosting algorithms evaluation metrics

id	C	kernel	micro	macro	weighted
1	1	ovr	0.67	0.46	0.65
2	1	linear	0.62	0.44	0.61

Table 5.5: SVM hyperparameters and evaluation metrics

id	n_estimators	max_depth	min_samples_split	min_samples_leaf	micro	macro	weighted
1	100	None	None	1	0.73	0.54	0.72
2	200	20	2	1	0.69	0.51	0.68

Table 5.4: Random Forest hyperparameters and evaluation metrics

Wine Quality Prediction

Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulfur Dioxide
7.3	1.07	0.09	1.7	0.178	10

Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol
89	0.9962	3.3	0.57	9

Predict

Logistic Regression

0	0.07
1	0.87
2	0.06
3	0.00

XGBClassifier

0	0.42
1	0.35
2	0.22
3	0.01

Random Forest

0	0.19
1	0.46
2	0.34
3	0.01

SVM

0	0.04
1	0.86
2	0.09
3	0.00

Thank You