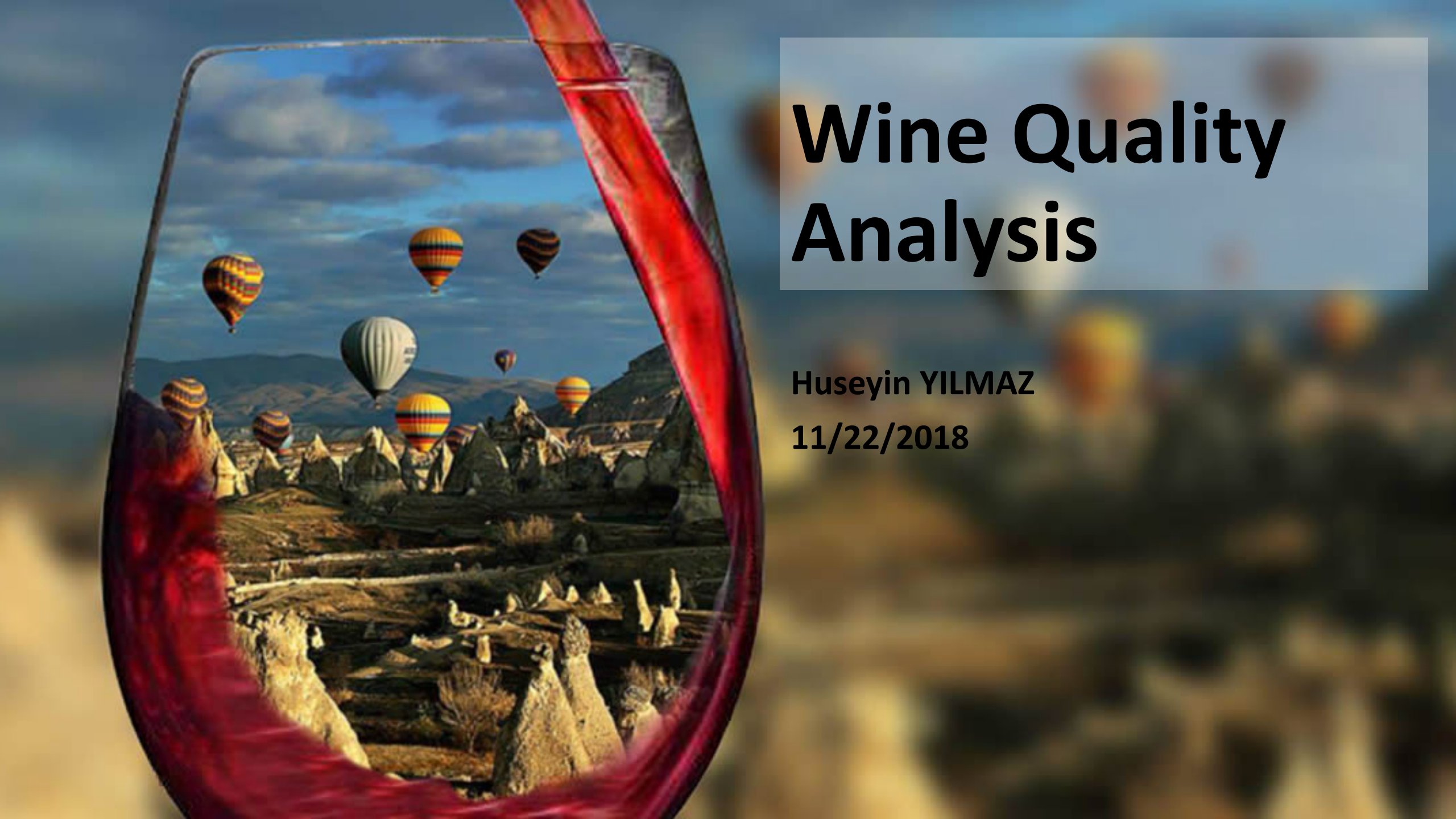


Wine Quality Analysis

Huseyin YILMAZ

11/22/2018

AGENDA

- Introduction
- Objective
- Data Analysis & Modeling
- Findings & Conclusion

Introduction

- Everywhere you look and see discussions about wine, you are sure to find wine tasting notes and wine scores.
- The descriptions wine reviewers write end up using all sorts of flowery, complicated language just to avoid writing the exact same descriptions over and over. This isn't very helpful to newer and less experienced wine drinkers.

Introduction

- We have a dataset which is on white wine and have 4898 observations. Data are collected on 12 different attributes of the wines and Quality is an ordinal variable with possible ranking from 1 (worst) to 10 (best). Each variety of wine is tasted by three independent tasters and the final rank assigned is the median rank given by the tasters.

Objective

- We want to predict the wine quality with using its attributes and to be helpful to newer and less experienced wine drinkers.

Data Analysis

Variables of the dataset

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulfates
- Alcohol
- Quality (score between 0 and 10)

Data Analysis

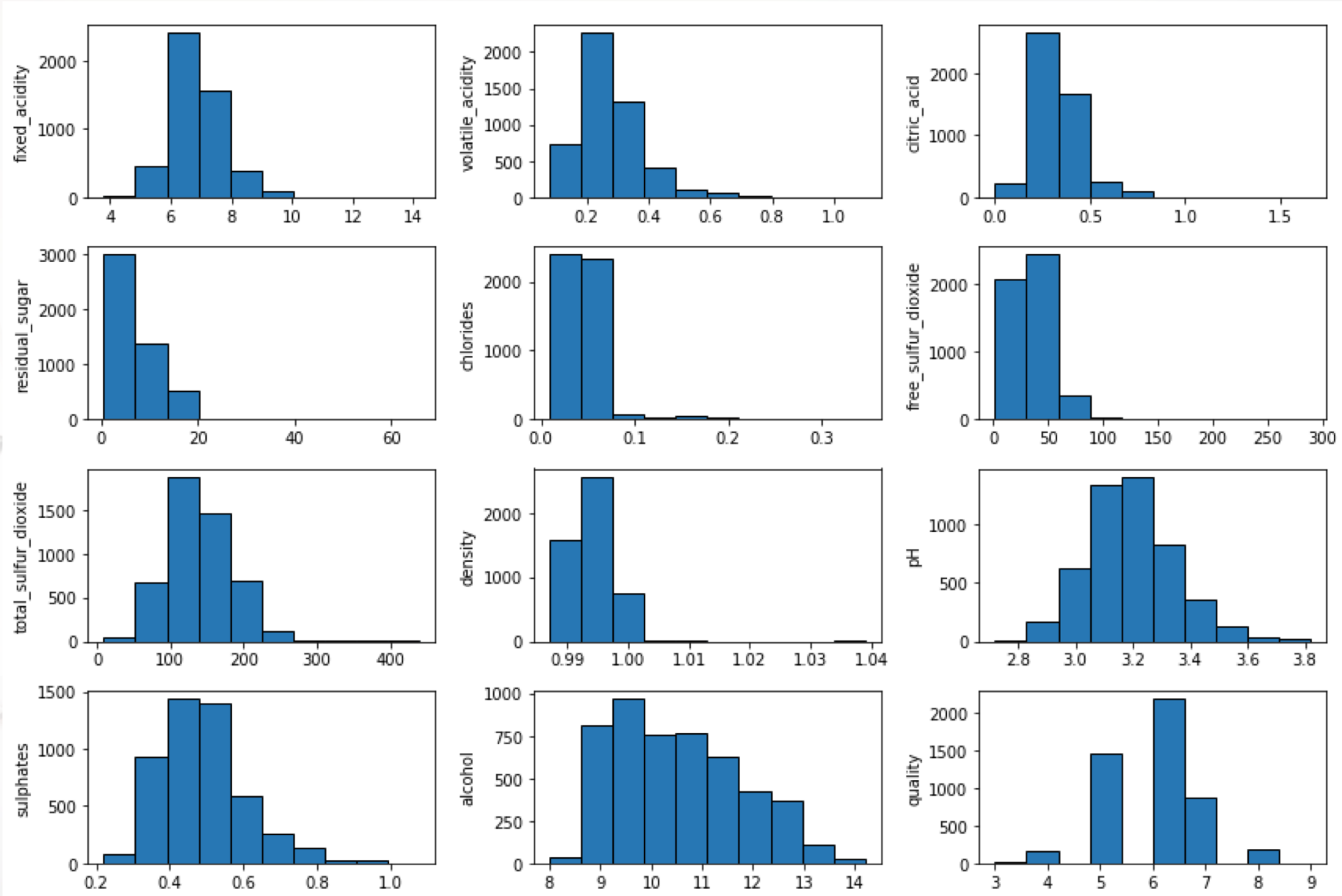
Statistical summary

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000

- Range is much larger compared to the IQR. Mean is usually greater than the median. These observations indicate that there are outliers in the data set and before any analysis is performed outliers must be taken care of.

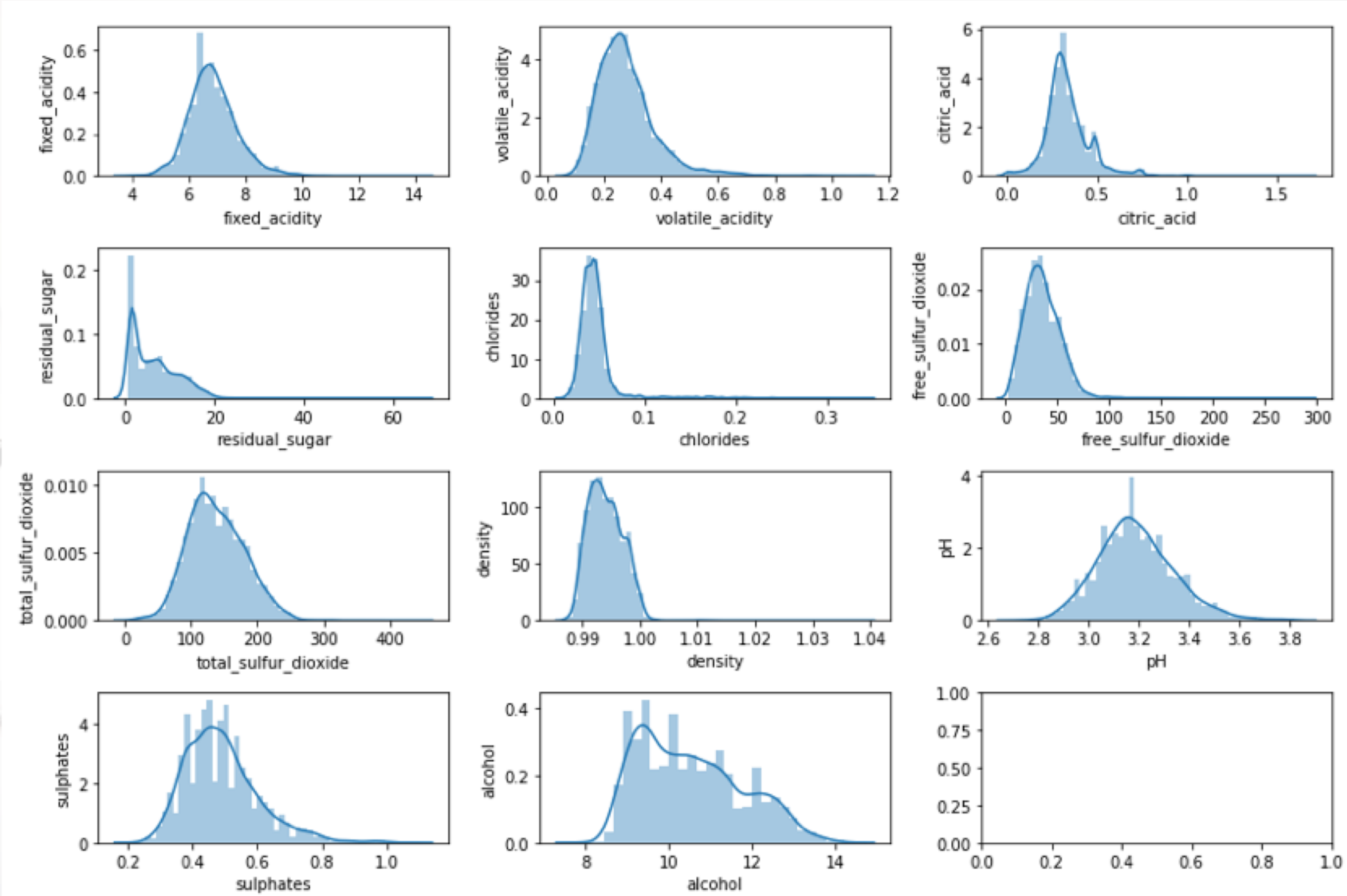
Exploratory Data Analysis

- Quality has most values concentrated in the categories 5, 6 and 7. Only a small proportion is in the categories [3, 4] and [8, 9] and none in the categories [1, 2] and 10.



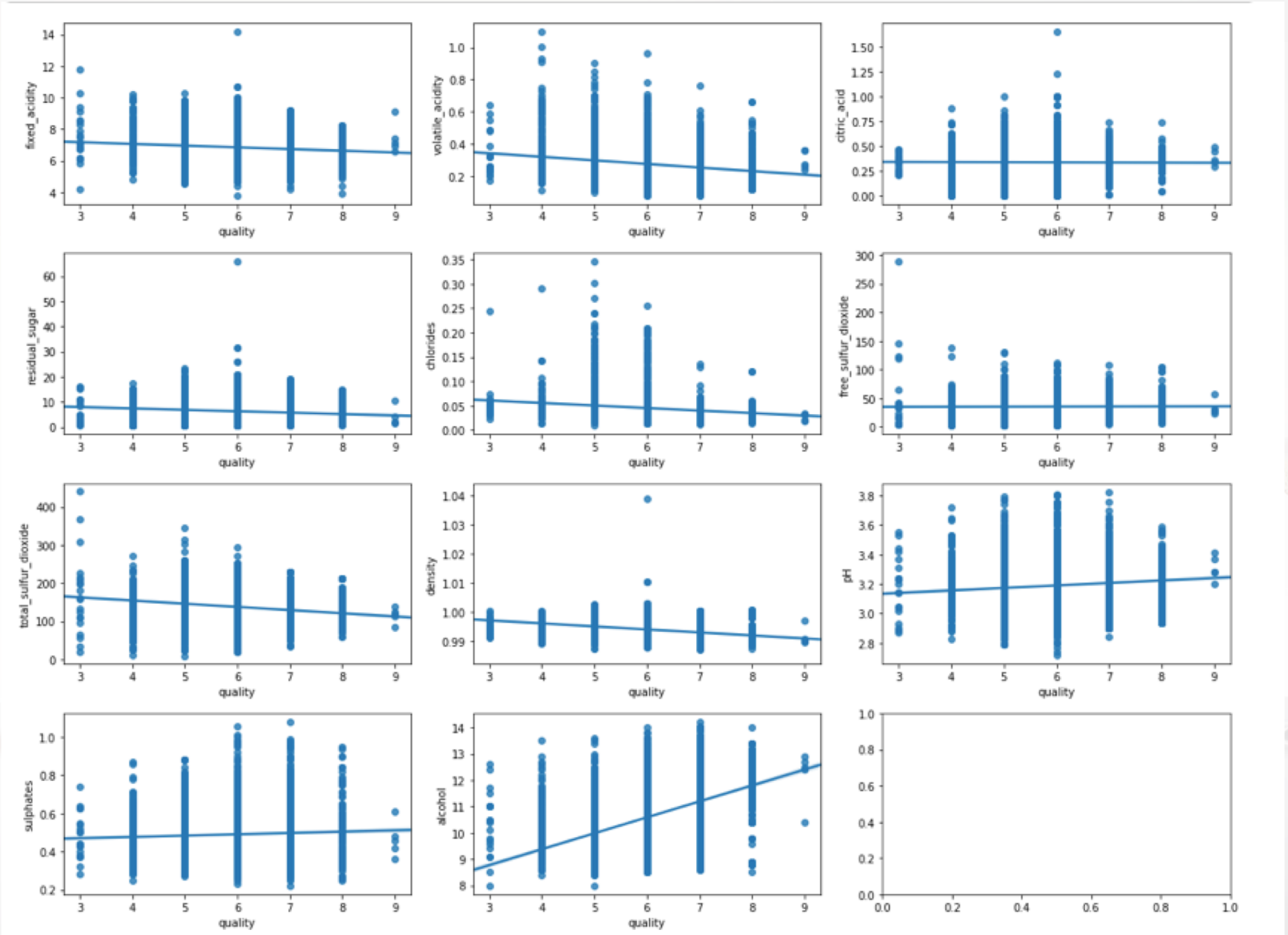
Exploratory Data Analysis

- All variables have outliers.
- Residual sugar has a positively skewed distribution; even after eliminating the outliers distribution will remain skewed.



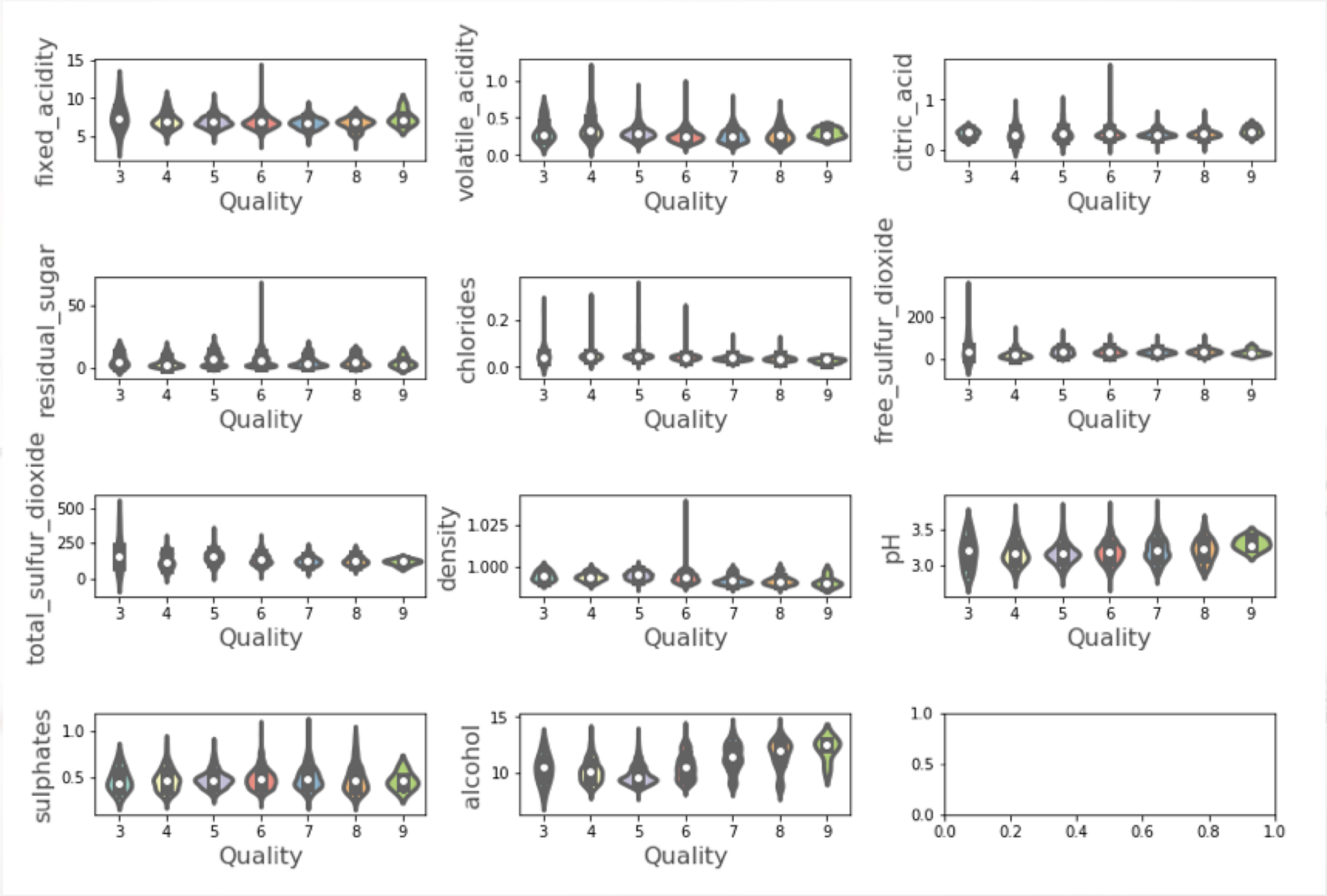
Exploratory Data Analysis

- Alcohol has a significant linear relationship but other variables have weak or no linear relationship with quality.



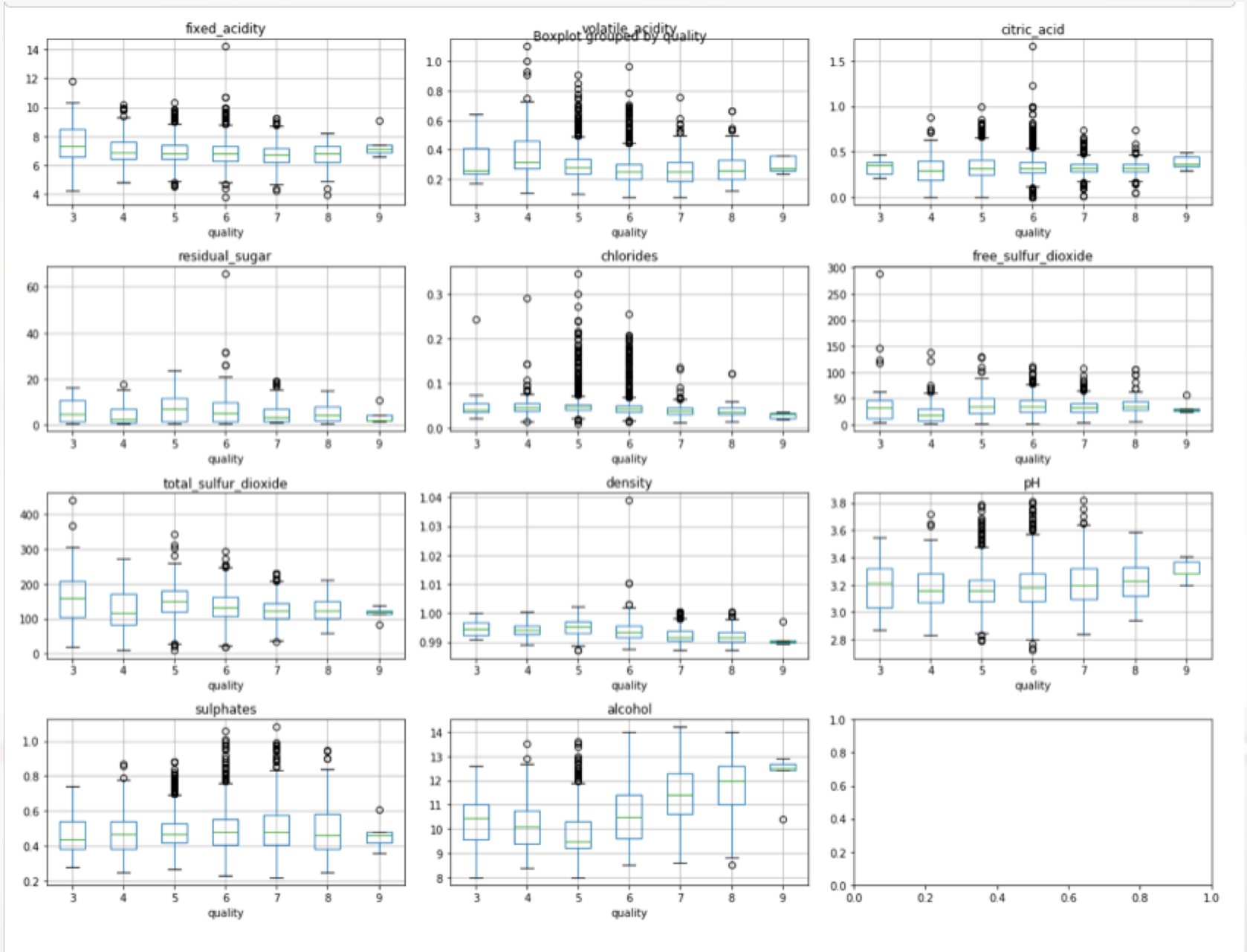
Exploratory Data Analysis

- Fixed acidity, volatile acidity and citric acid have outliers. If those outliers are eliminated distribution of the variables may be taken to be symmetric.



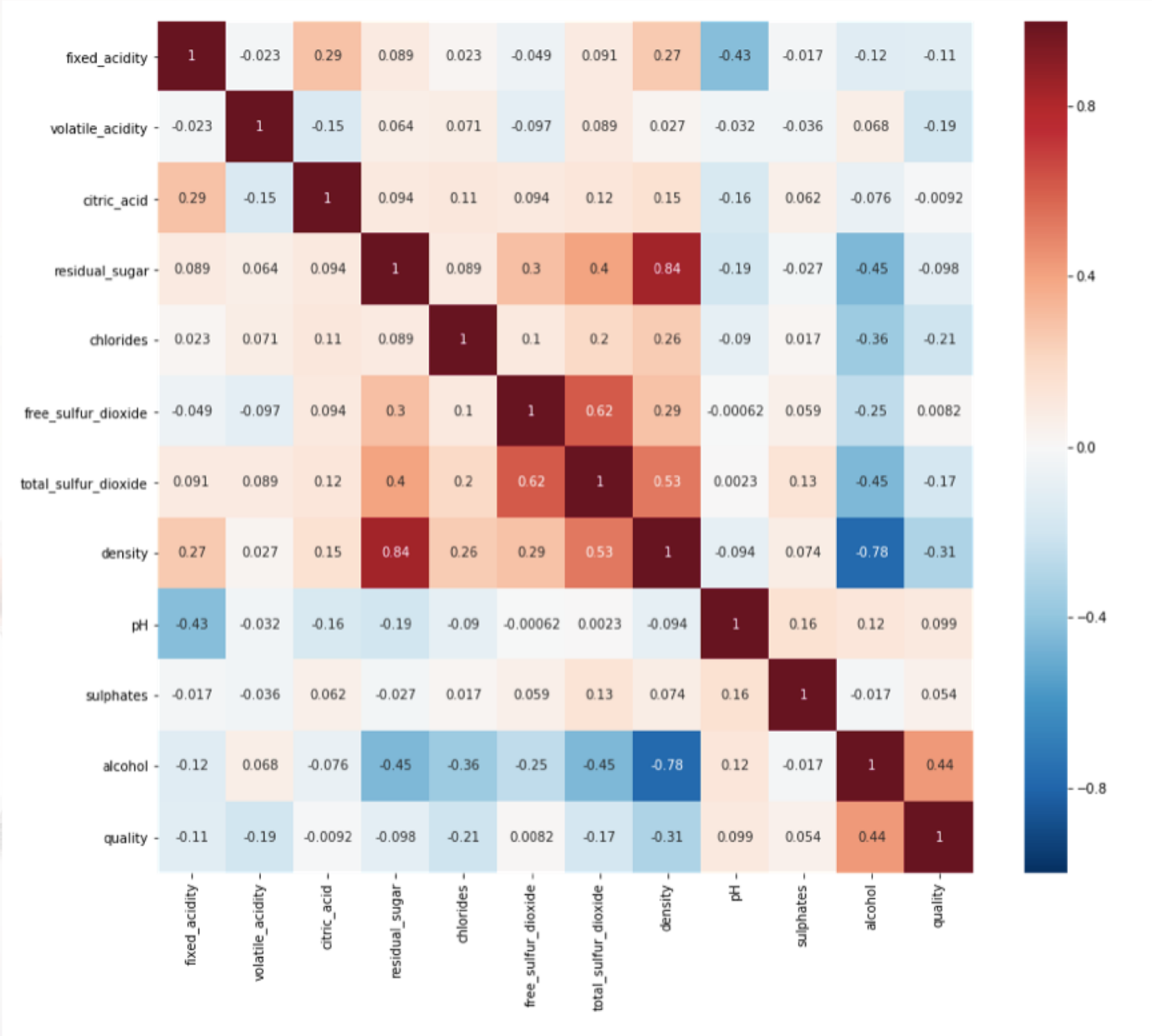
Exploratory Data Analysis

- Some of the variables, e.g . free sulphur dioxide, density, have a few outliers but these are very different from the rest.
- Mostly outliers are on the larger side.



Exploratory Data Analysis

- No varibales Some of the variables, e.g . free sulphur dioxide, density, have a few outliers but these are very different from the rest.
- Mostly outliers are on the larger side.



EDA Findings

- All variables have outliers
- Quality has most values concentrated in the categories 5, 6 and 7.

Only a small proportion is in the categories [3, 4] and [8, 9] and none in the categories [1, 2] and 10.

- Fixed acidity, volatile acidity and citric acid have outliers. If those outliers are eliminated distribution of the variables may be taken to be symmetric.

EDA Findings

- Residual sugar has a positively skewed distribution; even after eliminating the outliers distribution will remain skewed.
- Some of the variables, e.g . free sulphur dioxide, density, have a few outliers but these are very different from the rest.
- Mostly outliers are on the larger side.
- Alcohol has an irregular shaped distribution but it does not have pronounced outliers.

Feature Engineering

Feature elimination with using OLS results

- OLS summary shows us the most unrelated 3 features are X2 (citric acid), X4 (cholorides) and X6(total sulfur dioxide). these features will be removed before the modeling in order to increase our accuracies.

	coef	std err	t	P>|t|	[0.025	0.975]
const	150.1928	18.804	7.987	0.000	113.328	187.057
x1	0.0655	0.021	3.139	0.002	0.025	0.106
x2	-1.8632	0.114	-16.373	0.000	-2.086	-1.640
x3	0.0221	0.096	0.231	0.818	-0.166	0.210
x4	0.0815	0.008	10.825	0.000	0.067	0.096
x5	-0.2473	0.547	-0.452	0.651	-1.319	0.824
x6	0.0037	0.001	4.422	0.000	0.002	0.005
x7	-0.0003	0.000	-0.756	0.450	-0.001	0.000
x8	-150.2842	19.075	-7.879	0.000	-187.679	-112.890
x9	0.6863	0.105	6.513	0.000	0.480	0.893
x10	0.6315	0.100	6.291	0.000	0.435	0.828
x11	0.1935	0.024	7.988	0.000	0.146	0.241

Feature Engineering

Cleaning of outliers

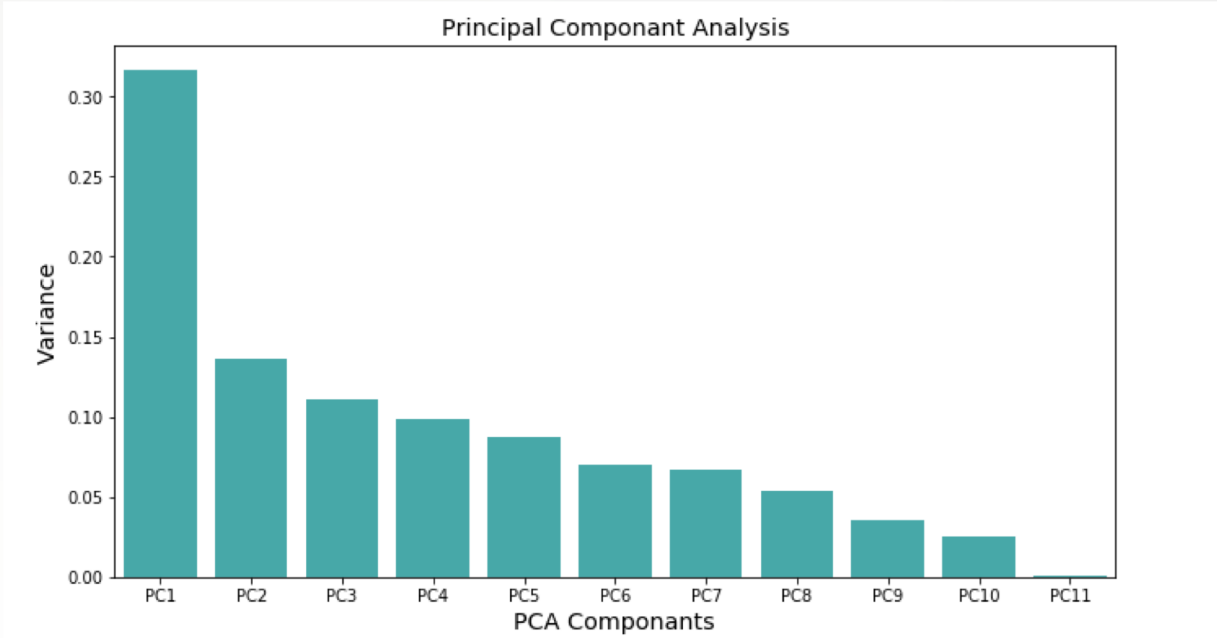
- In order to improve our scores, outliers have been removed from the data set

```
data=pd.read_csv('winequality-white.csv', delimiter=';')
for feature in data.keys():
    step=1.5*((np.percentile(data[feature], q=75))-(np.percentile(data[feature], q=25)))
    lower_bound=(np.percentile(data[feature], q=25))-step
    upper_bound=(np.percentile(data[feature], q=75))+step
    data.drop(list(data.loc[data[feature]<lower_bound].index)+
              list(data.loc[data[feature]>upper_bound].index), inplace=True)
```

Feature Engineering

Principal Component Analysis (PCA)

Linear Discriminant Analsis (LDA)



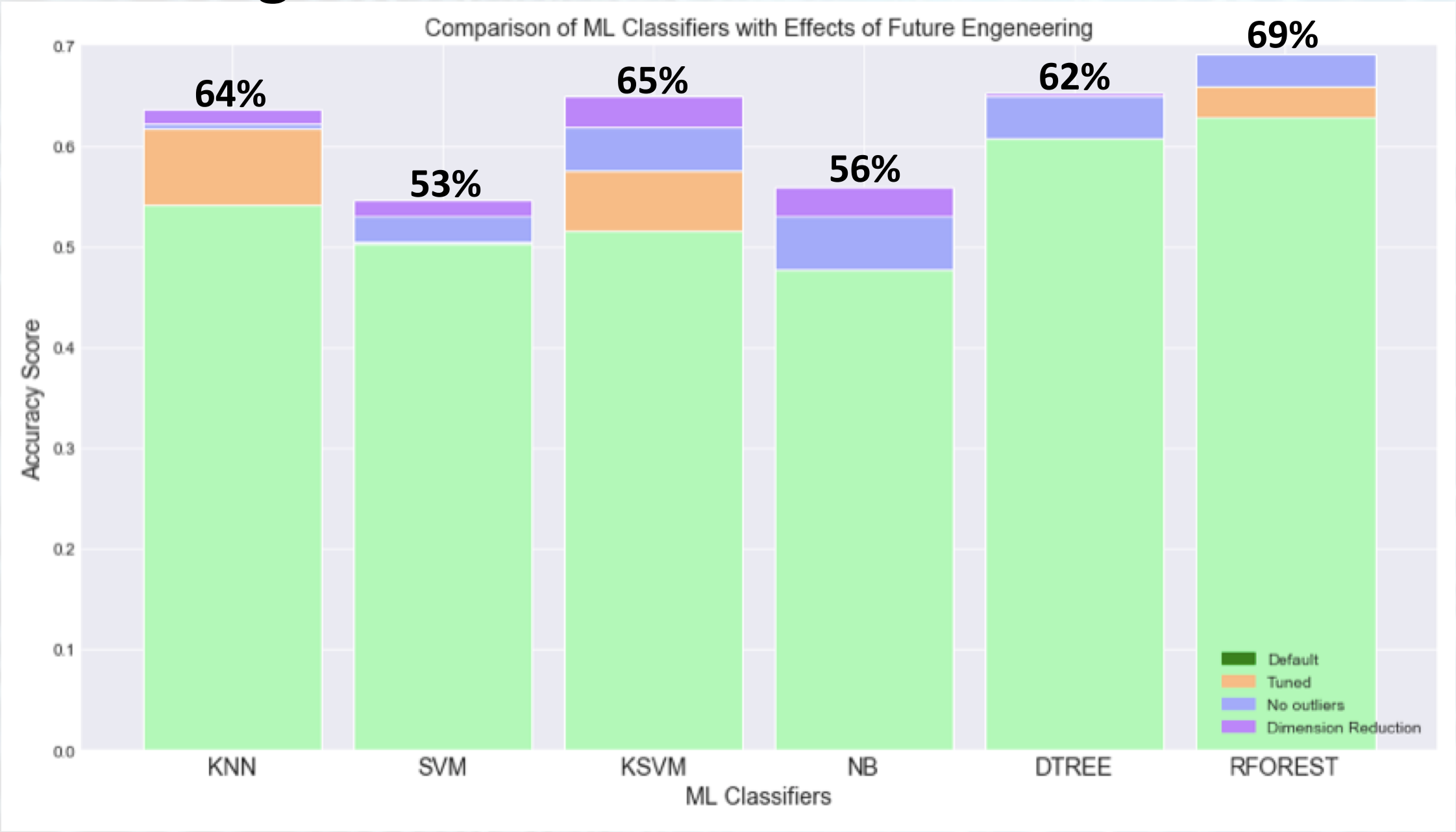
- We found that the first dimension explains about 31.6% of total variance, the second dimension explains about 13.6%, and the third dimension explains about 11%. This means the first 3 dimensions together explains only about 55% of the total variance, which is quite low. Additionally, we don't see any sharp drop off in the percentage of variance explained from this scree plot, suggesting no natural cut off point in keeping certain dimensions and discarding others.

Modeling

Algorithms

- K-Nearest Neighbors (K-NN)
- Support Vector Machine (SVM)
- Kernel SVM
- Naïve Bayes
- Decision Tree
- Random Forest

Modeling


Comparison of ML Classifiers with Effects of Future Engeneering
Accuracy Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
64%
53%
65%
56%
62%
69%
KNN
SVM
KSVM
NB
DTREE
RFOREST
ML Classifiers
Default
Tuned
No outliers
Dimension Reduction

Conclusion

- In this study, wine quality is tried to modeled with wine attributes of (i) Fixed acidity (ii) Volatile acidity (iii) Citric acid (iv) Residual sugar (v) Chlorides (vi) Free sulfur dioxide (vii) Total sulfur dioxide (viii) Density (ix) pH (x) Sulfates (xi) Alcohol.

- With the exception of alcohol, none of the attributes showed significant correlations to wine quality, indicating that they in fact cover different aspects of wine quality. However, as expected, no single predictor variable is able to fully describe all aspects of wine quality.

- Even after cleaning the outliers and applied dimensional reduction, the best prediction was 69% which is with random forest.

Conclusion

- We also reached below findings:
  - We observed that hyperparameter tuning has an improving effect on scores up to 13%.
  - Cleaning Outliers made positive effects over all models.
  - Principal Component Analysis generally made positive effect but did't affect Random Forest Model.

- In conclusion, however the best score is not so high, we reached the best score in regards of the data in hand.

Next Step

- Without full knowledge of each wine's, getting high prediction score is difficult. The results of this study can only serve as an initial look at quality. It is therefore necessary to validate the reported findings on a more extensive sample set, including, but not limited to, retail price and the year the wine was made.