WINE ANALYSIS

Capstone Project-1

Huseyin YILMAZ

1. INTRODUCTION

1.a. Why we need to do this analysis

Everyone has an idea about the concept of wine quality. But when comes to defining precisely what that means there is often silence. For most wine critics, quality refers to what they personally consider 'good' versus 'bad' wine, and correspondingly desirable versus aversive. This is usually framed within the context of conformity relative to established, learned norms for the wines concerned. This indicates, and rightly so, that quality is not only subjective, but also involves both intrinsic (sensory) and extrinsic (contextual) components.

For the purist, only sensory inputs should be involved. Because we are primarily creatures of sight, visual characteristics of the wine can, and usually do, almost subversively bias our perception of a wine's fragrance and taste. Thus, true aficionados prefer to sample their wines knowing nothing about its origin and in officially standardized (ISO) black, wine tasting glasses. Quality is assessed only on those attributes that the wine communicates to our senses of smell and taste – just what is in the glass.

When to Ignore Wine Tasting Notes

There are specific reasons why it can better to ignore wine tasting notes, or at least read them with a grain of salt. They don't apply 100% of the time with 100% of wine drinkers, but they are at least issues you should take into account whenever you see wine tasting notes.

Tasting Notes can Prejudice or Influence Your Perception

It's easy to prejudice or influence people's perception of almost anything, and even without trying very hard. You may not realize it consciously, but commentary and reviews you've seen about movies, food, books, and even other people will become part of an unconscious filter through which all new experiences and information flow. You can't help it and you can't avoid it, so descriptions you read about a wine will cause you to experience that wine differently than you would have had you read nothing at all. This isn't always bad, since a reference to "strawberry notes" might help you recognize the very subtle strawberry flavors in the background, but it's not something you want let get out of hand either.

Tasting Notes Can Set You Up for Failure

Most tasting notes in wine reviews are written by people with a lot of knowledge about and experience with wines — or at least probably more knowledge and experience than you have. These wine drinkers have developed the ability to pick out subtle flavors and aromas in wines which may not be readily apparent to the average wine drinker. So if you read about "strawberry notes" yet are unable to detect anything remotely like strawberries — and consistently have similar experiences when drinking a wine after reading an expert's tasting notes — you might start thinking of yourself as a "failure" when it comes to drinking wine. You aren't, though, and don't need to such discouragement.

Tasting Notes Can Be Overwritten and Too Complicated

Wines may not all taste alike, but coming up with different words and phrases to express subtle differences in flavor or aroma can be difficult. Add to this the fact that the average wine reviewer or writer might be trying to write scores of wine tasting notes in any given year, and you

can just imagine how difficult it will get for even an excellent writer. As a result, 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.

1.b. The problem

As it can be concluded from above reasons, the problem is dependability of notes or scores of wine quality. We may not end the discussion on this topic but we can contribute. In order to make this contribution we will try to predict wine quality with using its sensory inputs.

1.c. The Dataset

In order to solve our problem we will try to establish a model with using a set of observations on a number of white wine varieties involving their chemical properties and ranking by tasters. Wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability. Pricing of wine depends on such a volatile factor to some extent. Another key factor in wine certification and quality assessment is physicochemical tests which are laboratory-based and takes into account factors like acidity, pH level, presence of sugar and other chemical properties. For the wine market, it would be of interest if human quality of tasting can be related to the chemical properties of wine so that certification and quality assessment and assurance process is more controlled.

Two datasets are available of which one dataset is on red wine and have 1599 different varieties and the other is on white wine and have 4898

varieties. Only white wine data is analysed. All wines are produced in a particular area of Portugal. Data are collected on 12 different properties of the wines one of which is Quality, based on sensory data, and the rest are on chemical properties of the wines including density, acidity, alcohol content etc. All chemical properties of wines are continuous variables. 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.

2. Data Understanding and Wrangling

2.a Obtaining

Data set obtained as plain text files from UC Irvine Machine Learning Repository (link: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/)

After downloading the text files, they were converted to csv files, and loaded with pandas.

2.b. Understanding and Wrangling

There are 12 features in the data set. So, except the response feature (quality), we have 11 predictor features.

We changed the feature names with using "_" in order to make coding easier.

```
In [20]: wine.columns = ['fixed_acidity', 'volatile_acidity', 'citric_acid', 'residual_sugar',
                         'chlorides', 'free_sulfur_dioxide', 'total_sulfur_dioxide', 'density',
                         'pH', 'sulphates', 'alcohol', 'quality']

In [108]: wine.head()
```

Out[108]:

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

According to statistics summary (below table), 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 model is performed outliers must be taken care of.

	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	4898
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267	5
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621	0
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000	3
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000	5
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000	6
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000	6
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000	9

Some question to trigger exploratory analysis;

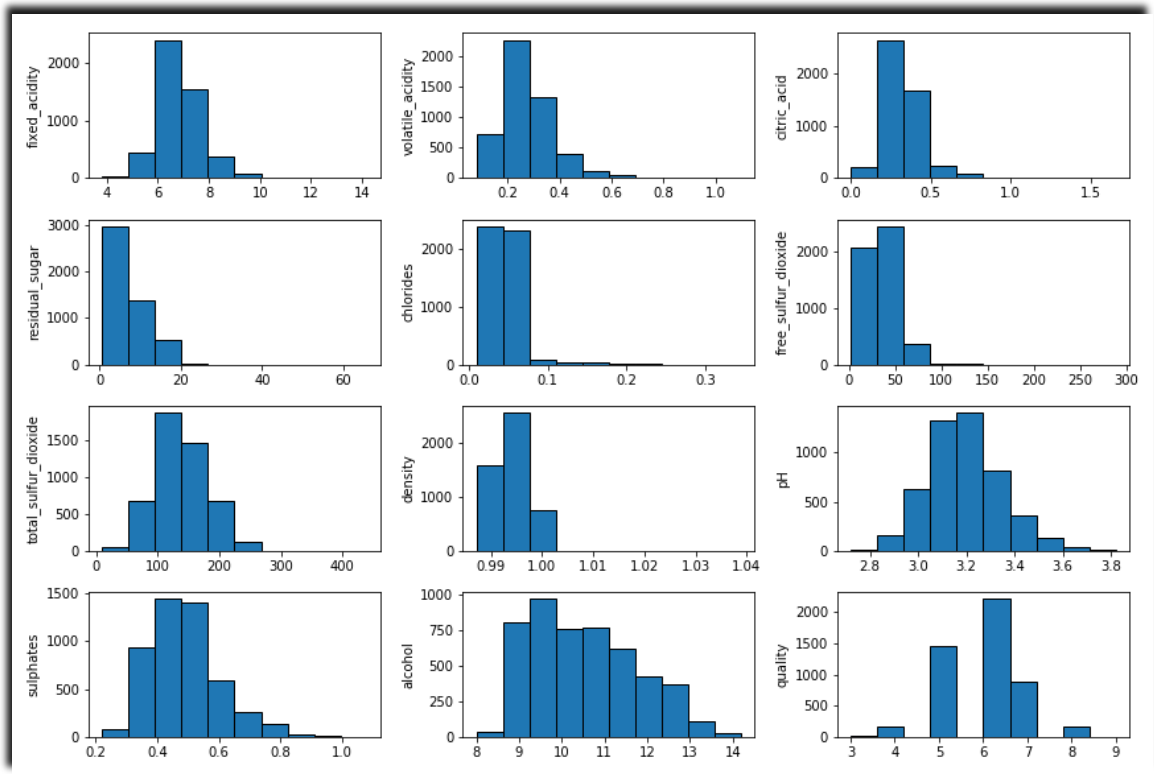
- *What is the relationship between each feature and our response value?*
- *Are those relationships linear or nonlinear, positive or negative?*
- *Which features have outliers?*

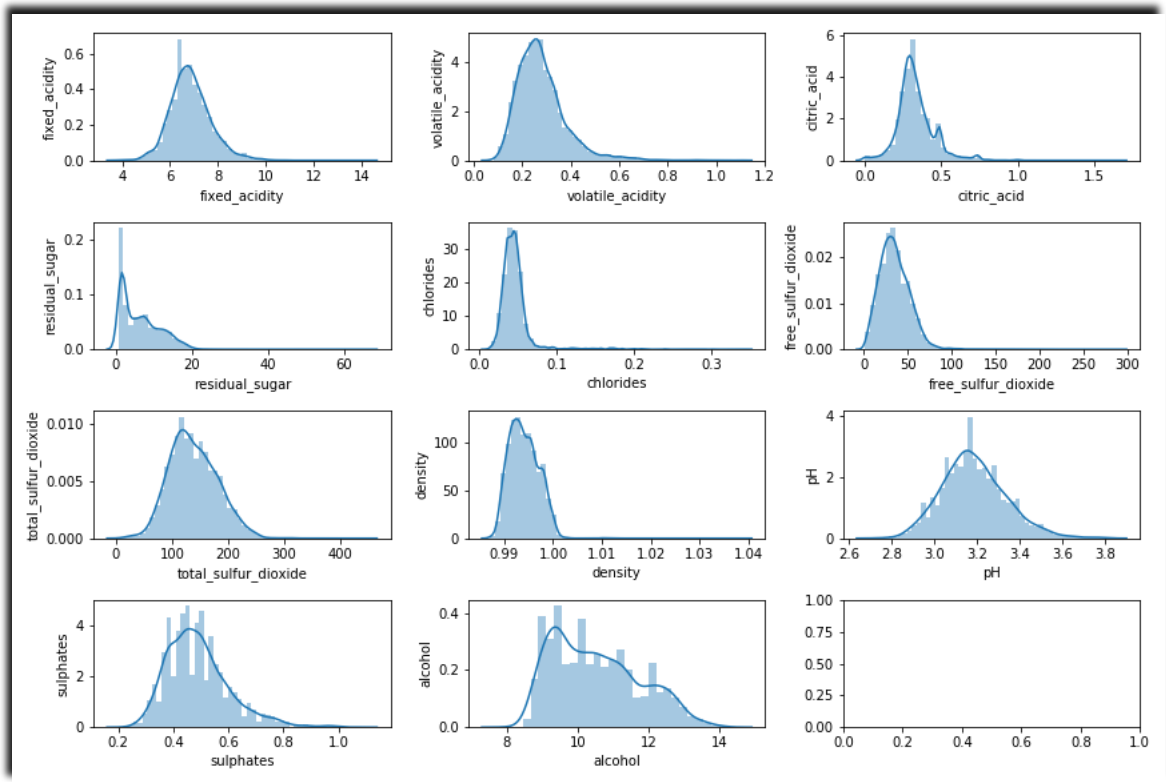
3. Exploratory Data Analysis

In order to understand the characteristics of the features and relationships between these features, we will apply exploratory data analysis.

Distribution histograms and curves is very helpful to get an initial sense.

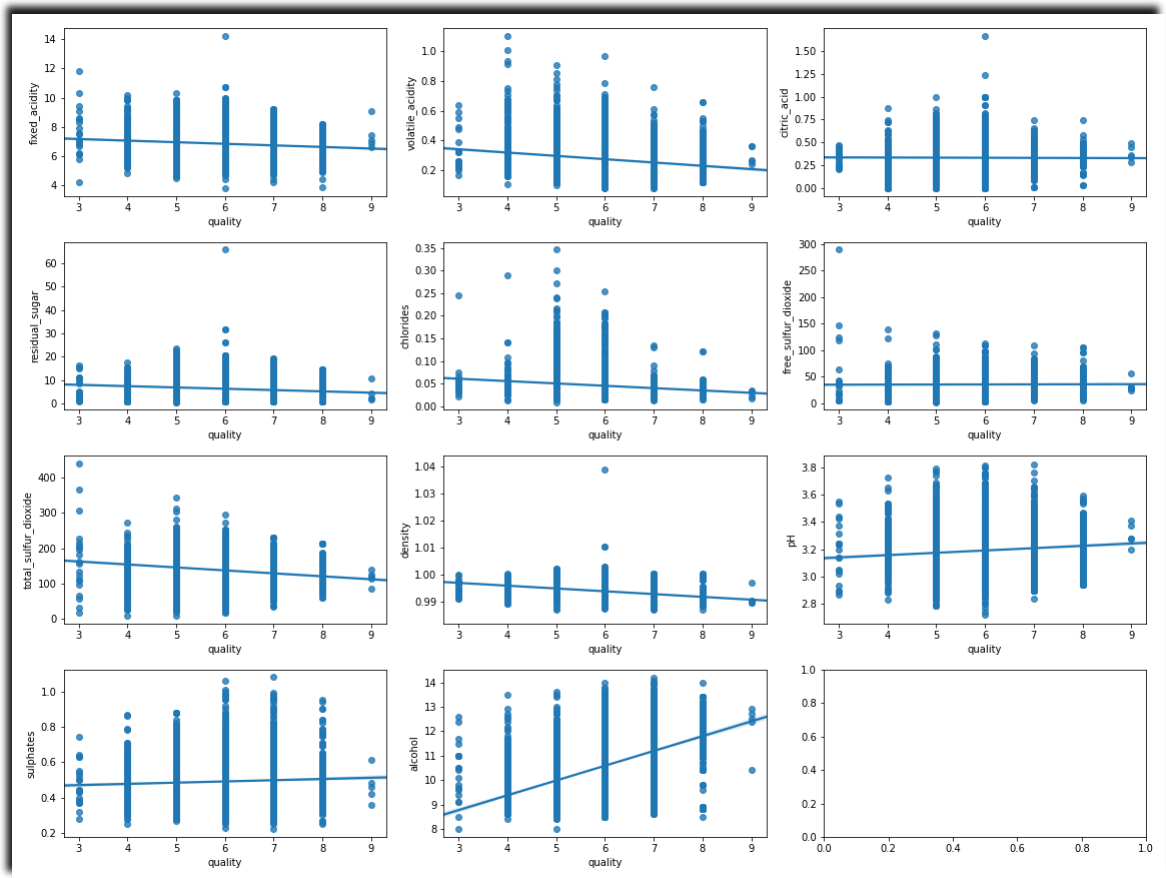
3.a. Distribution Histogram





3.b. Scatter Plots

Scatter plots will increase our understanding of features



Scatter plot is one of the way to make a comparison and to understand the distribution of the data. But it has some drawbacks here. For instance, in this data set, our target value is integer and it is used for each plot in order to make a comparison between the target variable and rest of the other variables. The problem with the integer variables is that one of the axes of the plot has constant values and this makes you unable to understand the density of the specific values. In other words, when you see a solid line, you cannot be sure that whether there are 50 points or 500 points under that solid line. So it is a problem.

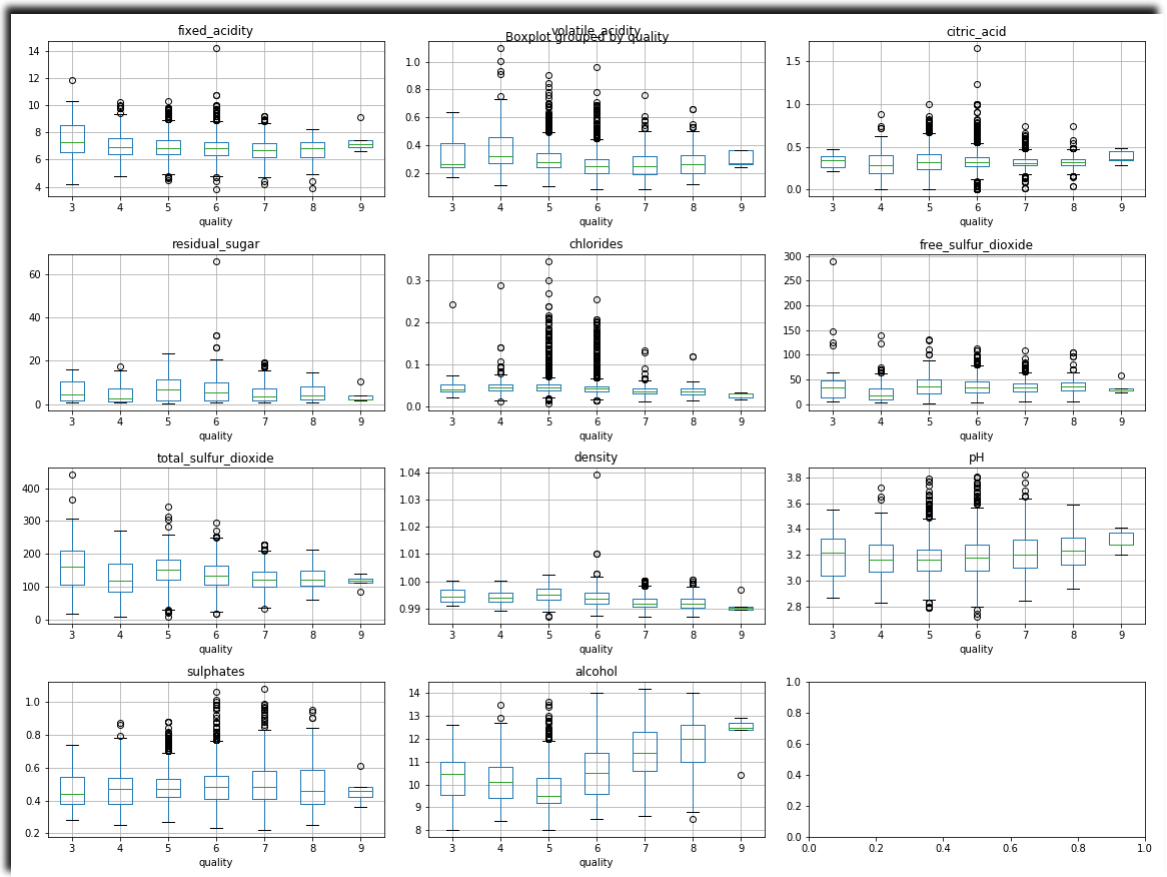
Seaborn Scatter plot has same drawbacks with the matplotlib one, but it compensates some of these drawbacks with using a linear line which gives a critical clue about the relationship between the compared variables. For example, let's compare the scatter plots of "alcohol-quality" and "pH-quality", in order to understand the power of linear

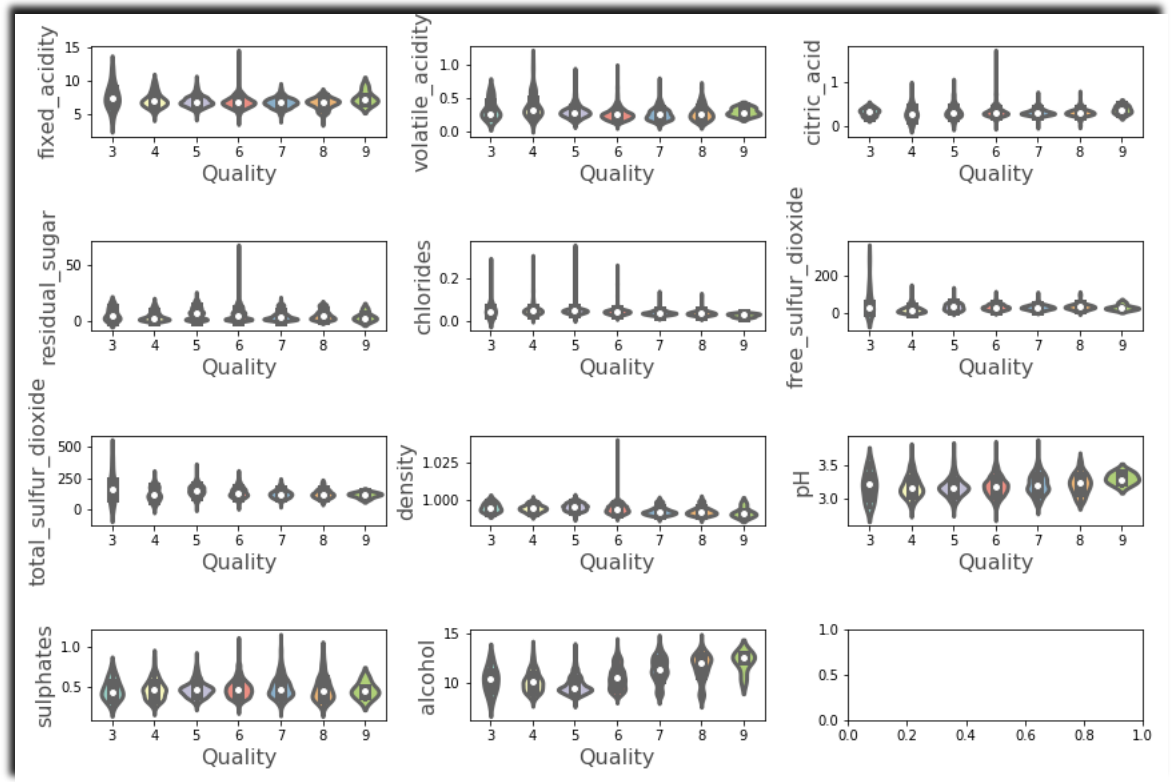
model. When you just look at the points both plots look very close to each other. But linear lines show that how they have different relationships with “quality”. While alcohol has a significant positive relationship between quality, "pH" does not.

So, we have now a better understanding of the influences of the features on the quality by means of seaborn scatter plots. Positive relationship between the "alcohol" and the quality is clearly visible.

3.c. Violin and Box Plots

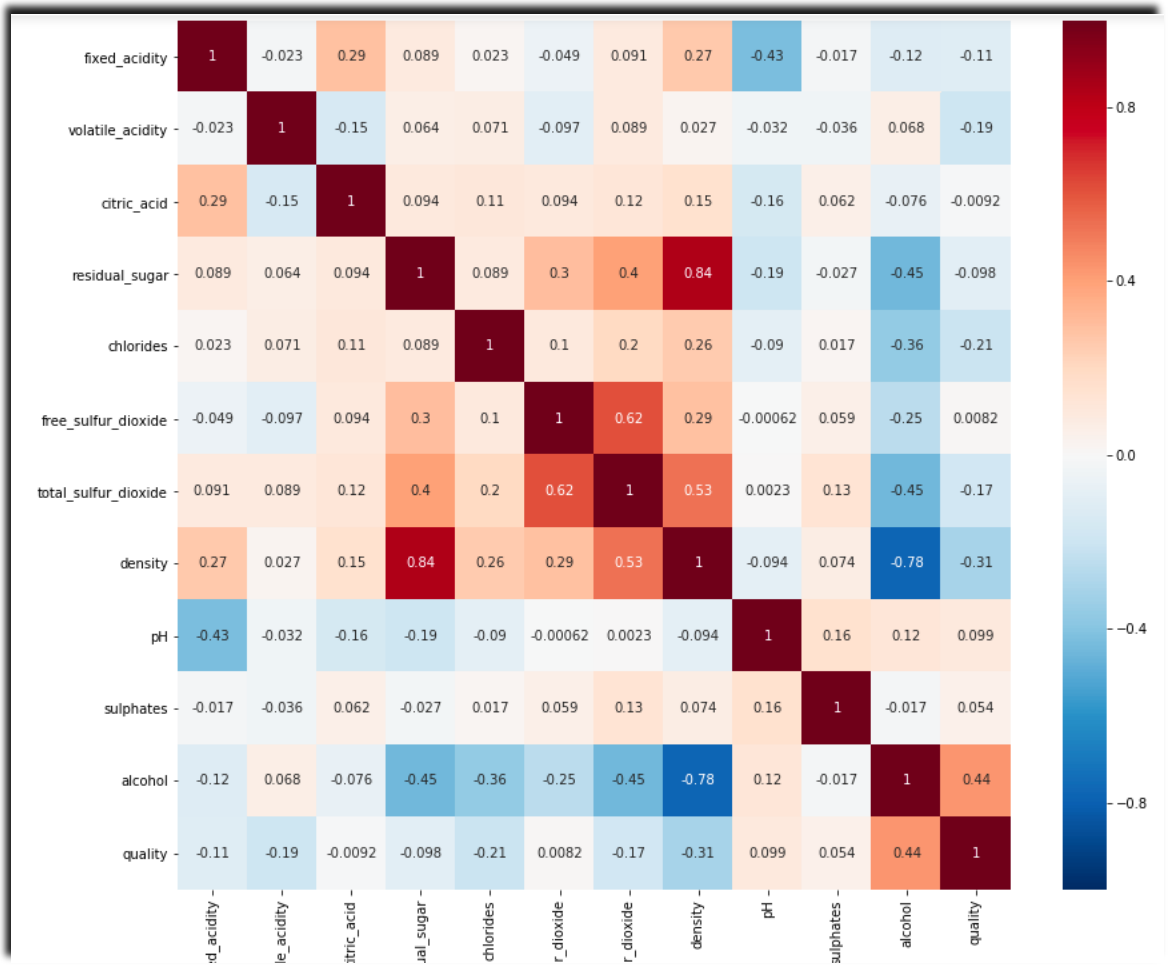
Outliers are more visible with violin and box plot graphs. They make the outlier detection easier and we will be able to figure out that which features need to be touched.





3.d. Correlation Matrix

Correlation matrix help us to validate our findings until here. We will combine all results for selecting and improving our features.



3.e. 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.*
- *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.*

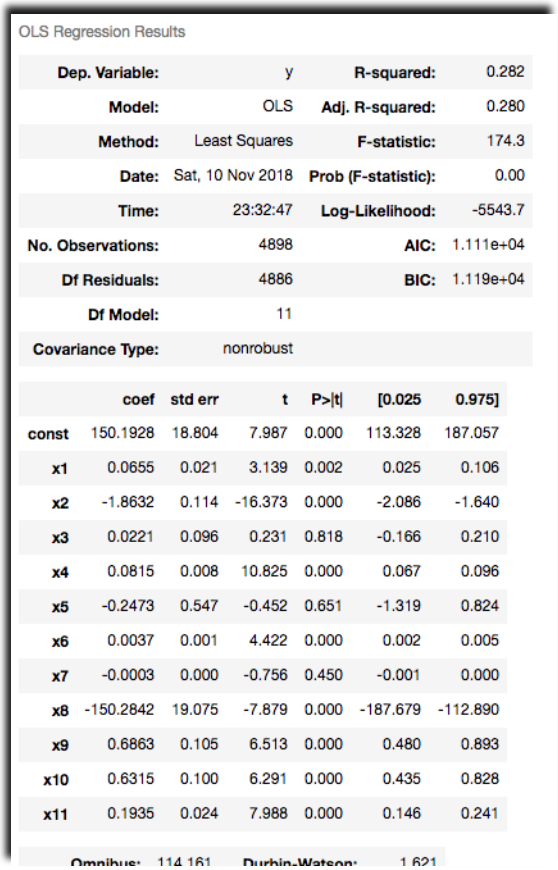
4. Feature Engineering

According to the above EDA findings we don't have so many choices as the relationship between the predictor features and target feature are not so strong. Correlations values also confirm these findings.

Along with the EDA findings we will use an elimination technique in order to support our findings and select the correct features.

4.a. 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. We will also use the above EDA findings and remove additional 3 features which are clearly have almost even relationship with “quality”; "free sulfur dioxide", "fixed acidity" and "residual sugar". Eventually, we will keep just 5 features.



OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.282
Model:	OLS	**Adj. R-squared:**	0.280
Method:	Least Squares	**F-statistic:**	174.3
Date:	Sat, 10 Nov 2018	**Prob (F-statistic):**	0.00
Time:	23:32:47	**Log-Likelihood:**	-5543.7
No. Observations:	4898	**AIC:**	1.111e+04
Df Residuals:	4886	**BIC:**	1.119e+04
Df Model:	11		
Covariance Type:	nonrobust		

	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

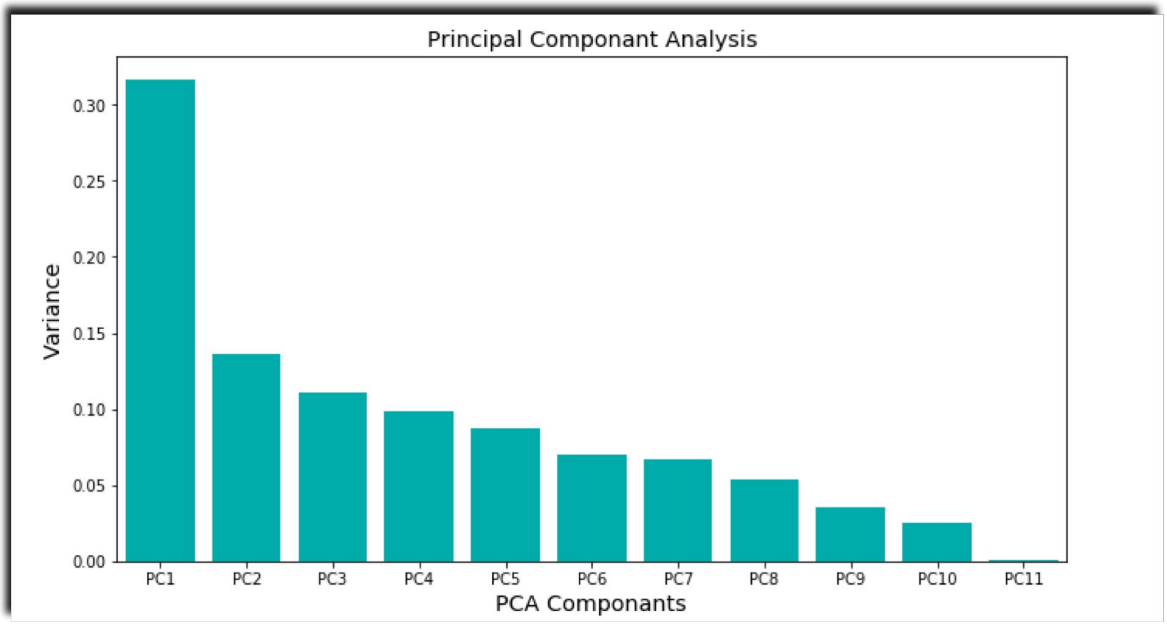
Omnibus:	114.161	**Durbin-Watson:**	1.621

4.b. Outlier Cleaning

In order to improve our model performances all the outliers have been cleaned from the data set.

4.c. Dimension Reduction

Two different approach have been used for this analysis and both of them compared at the end.



From the Principle Component Analysis (PCA) scree plot for our data set, 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.

Since the number of the components is not small and the linear relationships among them are not strong, it is hard to interpret the principal components.

But we will use first 5 components to be able to explain at least 80% of the variance.

5. Modeling

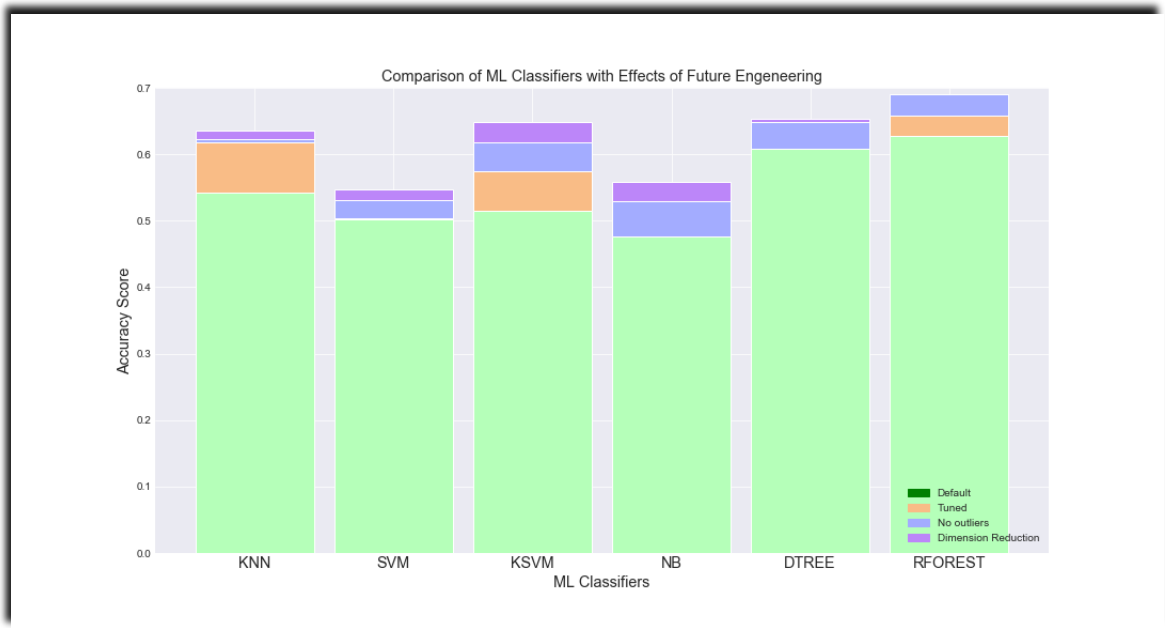
Six different machine learning algorithms have been used to predict the wine quality. Algorithms:

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

All models have five different accuracy scores. We got the first score (default) without applying hyper parameter tuning and out outlier cleaning; after hyper parameter tuning we got the second score; after outlier cleaning third one, after dimension reduction with PCA we got the forth one and fifth one after dimension reduction with LDA. And also contribution of the each step has been calculated in percentage.

	default	tuned	no_outliers	pca	lda	total_improvement_(%)	tuning_cont_(%)	no_outliers_cont_(%)	pca_cont_(%)	lda_cont_(%)
KNN	0.5418	0.6173	0.6225	0.6356	0.6055	0.1731	0.0755	0.0052	0.0131	0.0000
SVM	0.5030	0.5040	0.5307	0.5465	0.5307	0.0865	0.0010	0.0267	0.0158	0.0000
KSVM	0.5153	0.5744	0.6186	0.6487	0.5439	0.2589	0.0591	0.0442	0.0301	0.0000
NB	0.4765	0.4765	0.5294	0.5583	0.5386	0.1717	0.0000	0.0529	0.0289	0.0092
DTREE	0.6081	0.5709	0.6120	0.6159	0.6304	0.0367	0.0000	0.0411	0.0039	0.0184
RFOREST	0.6275	0.6591	0.6907	0.6736	0.6408	0.1007	0.0316	0.0316	0.0000	0.0000

Below graph shows the general comparison of the algorithms over our data set and also effects of each step.



6. Conclusion

In this study, wine quality is tried to modeled with wine attributes (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. 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.

Why our highest score is not so high;

• The first thing the distribution of the target variable. 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. Small proportions decrease the chance of high accuracy scores of these levels.

• Relationships between the predictor variables and the target variable. There is no strong relationship between quality and anyone of the predictors. So, this makes difficult to predict.

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

7. Next Step

Without full knowledge of each wine's we can't get high prediction scores. 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.