RED WINE

Assignment

Course: ALY6015 Intermediate Analytics

Instructor: Fidel Rodriguez

By,

Pragati Koladiya | NUID: 001029445

AGENDA

- Introduction
- Exploratory Data Analysis (EDA)
- Pre-processing
- Models Implemented
- Comparisons of models
- Testing sample data
- Implementation of Professor's Feedback
- Conclusion

GOALS

- To compare several classification algorithms to predict wine quality
- Show some red wine information details with Exploratory Data Analysis(EDA)
- Find best model to predict quality of wine
- The goal of predictive analysis is to avoid overfitting and find the model that gives the highest accuracy.

DATASET INFORMATION

The data is about red wine samples (vinho verde) from Portugal.

The data contains 1599 observations (wine samples) and 12 attributes related to the wine.

The data was collected from May 2004 to February 2007

Vinho Verde is not a grape variety, it is a DOC for production of wine. The name means "green wine," but translates as "young wine", with wine being released three to six months after the grapes are harvested. They may be red, white or rose and they are usually consumed soon after bottling.

DATASET INFORMATION

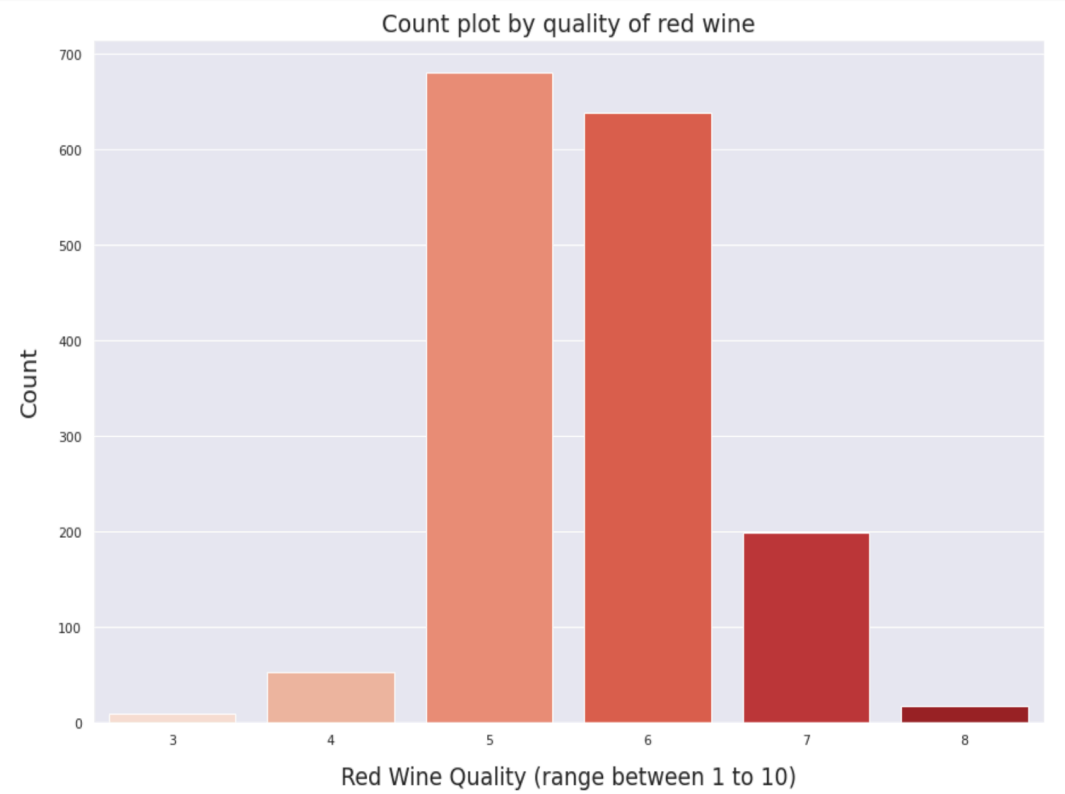
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

1599 rows × 12 columns

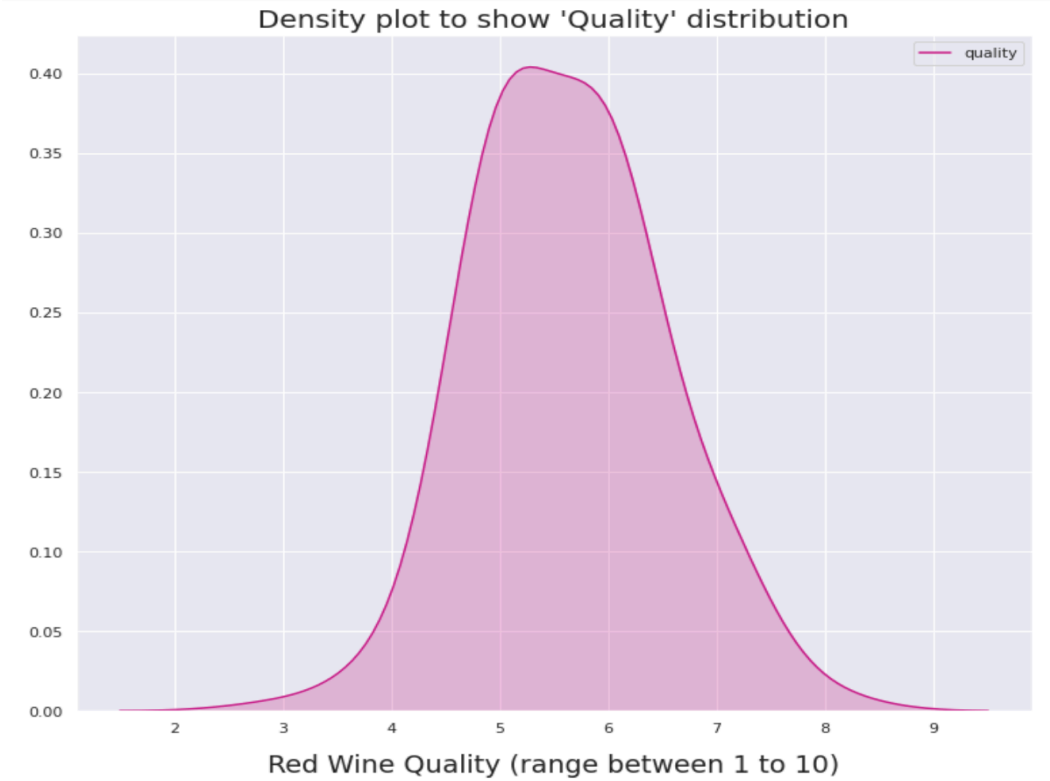
EXPLORATORY DATA ANALYSIS (EDA)

1. Red wine quality count?
2. Alcohol Vs Wine Quality?
3. Alcohol Vs Density?
4. Sulphates Vs Reviews?
5. pH Vs Reviews?
6. Citric Acid Vs Reviews?
7. Acidity in Wine Vs Quality?
8. Residual sugar Vs Red wine reviews?
9. Chlorides Vs Quality?
10. Total Sulfur dioxide Vs Reviews?

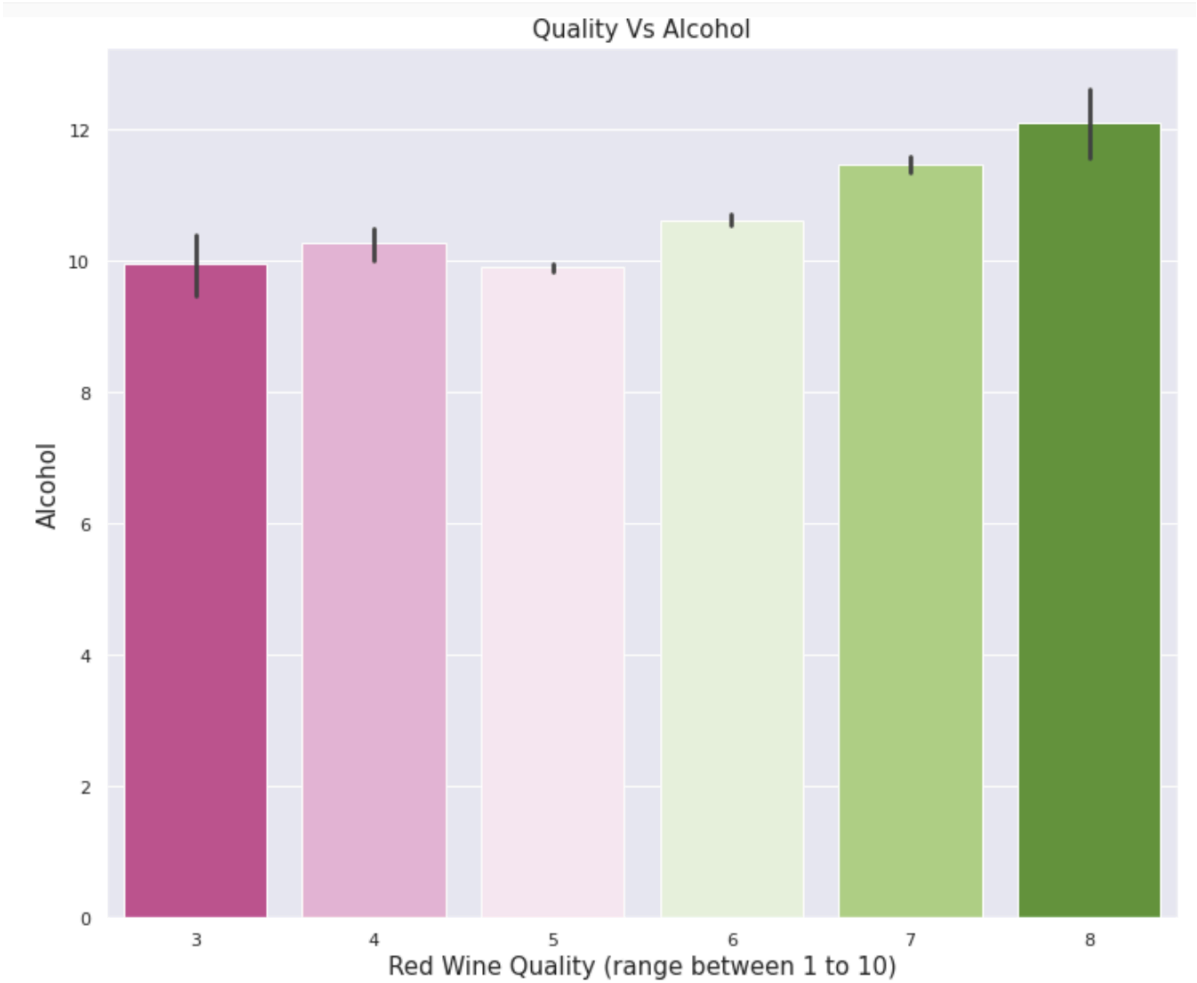
RED WINE QUALITY COUNT?



- Maximum quality counts are dense around 5 and 6 while quality count 3,4,7 and 8 are relatively low.

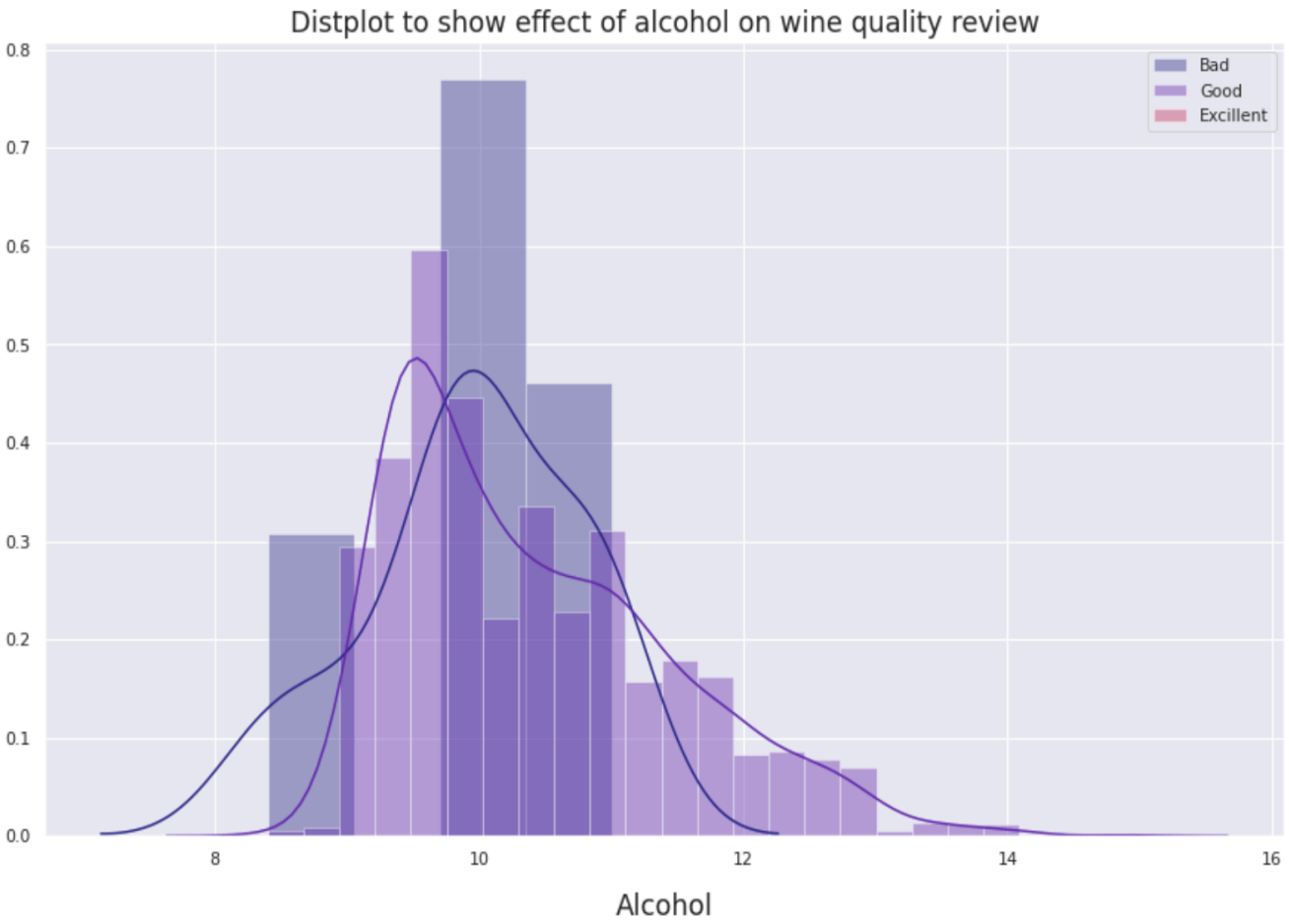


- The data is not equally distributed, so we will consider some actions in future so that results may not bias to majority.



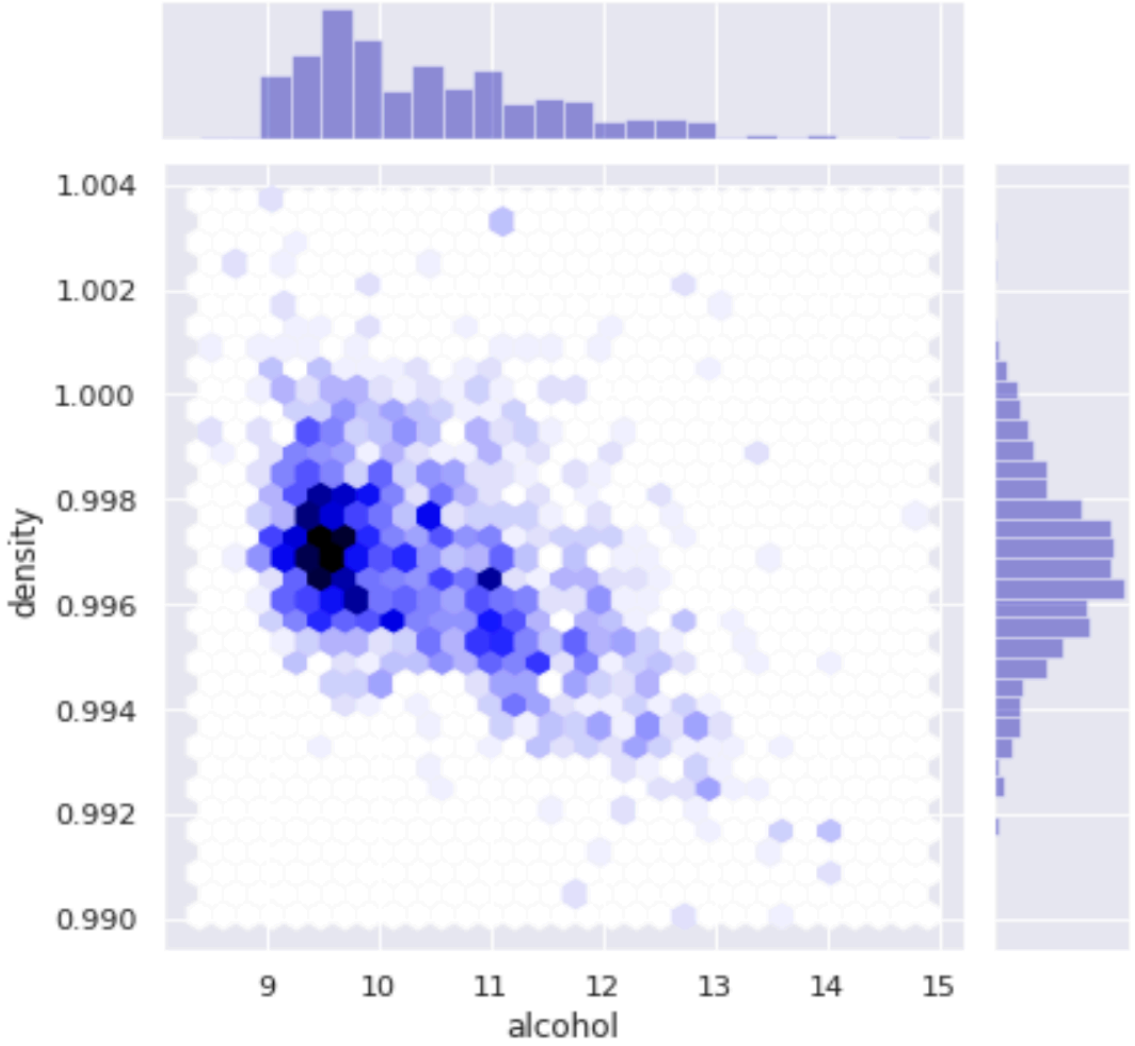
ALCOHOL VS WINE QUALITY?

- Wine with the lower than 10% alcohol is considered as bad wine quality.
- Wine with higher rating tend to have higher amount of alcohol in it.



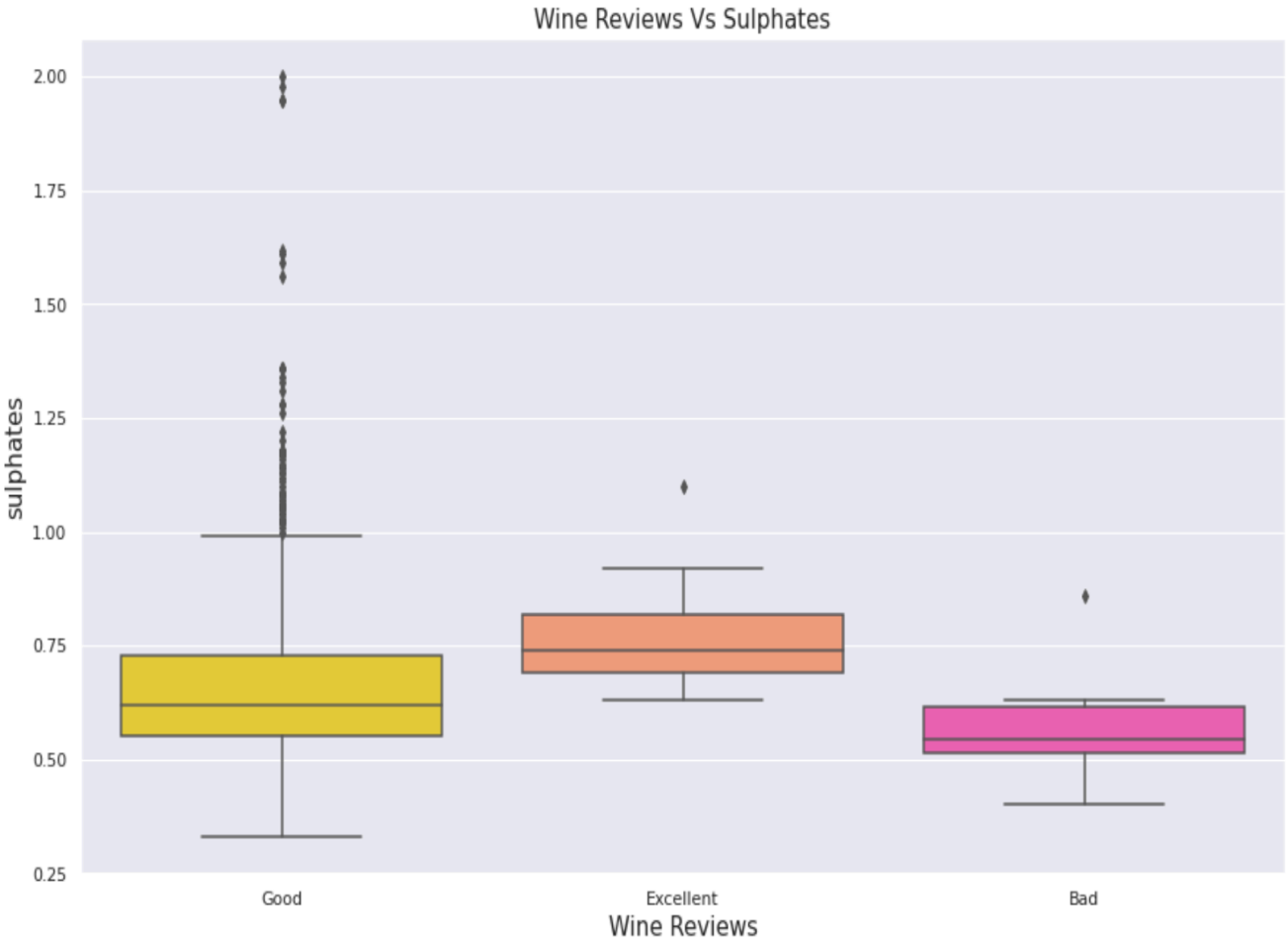
ALCOHOL VS WINE QUALITY?

- The amount of alcohol in Bad wine quality is normally distributed from 0 to 12 whereas the good quality of wine contains more alcohol and the data is right-skewed.



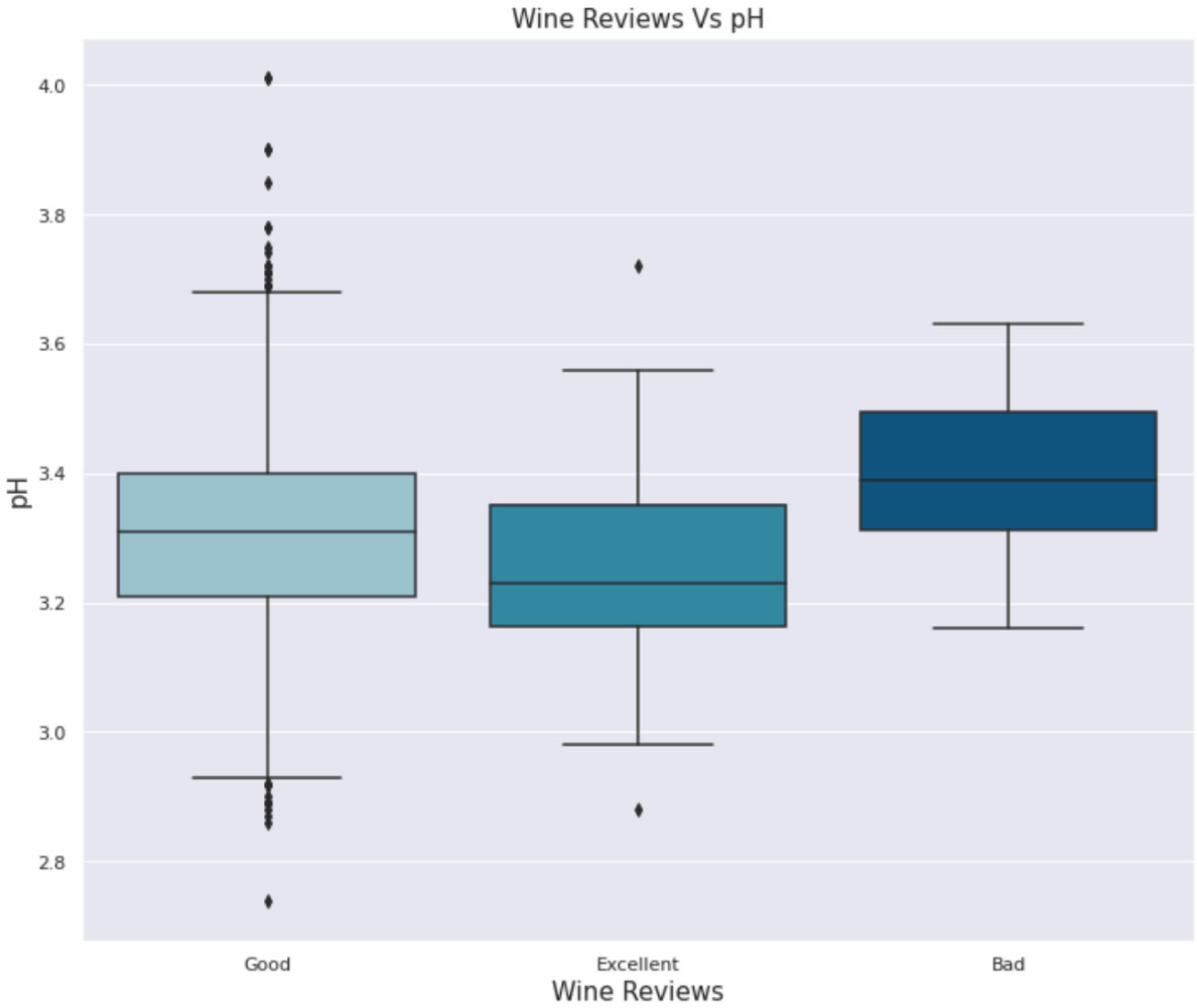
ALCOHOL VS DENSITY?

- The alcohol is denser between 9 and 10 while the density is denser between 1 and 0.996.
- The density column has equally distributed data while alcohol has right-skewed data.



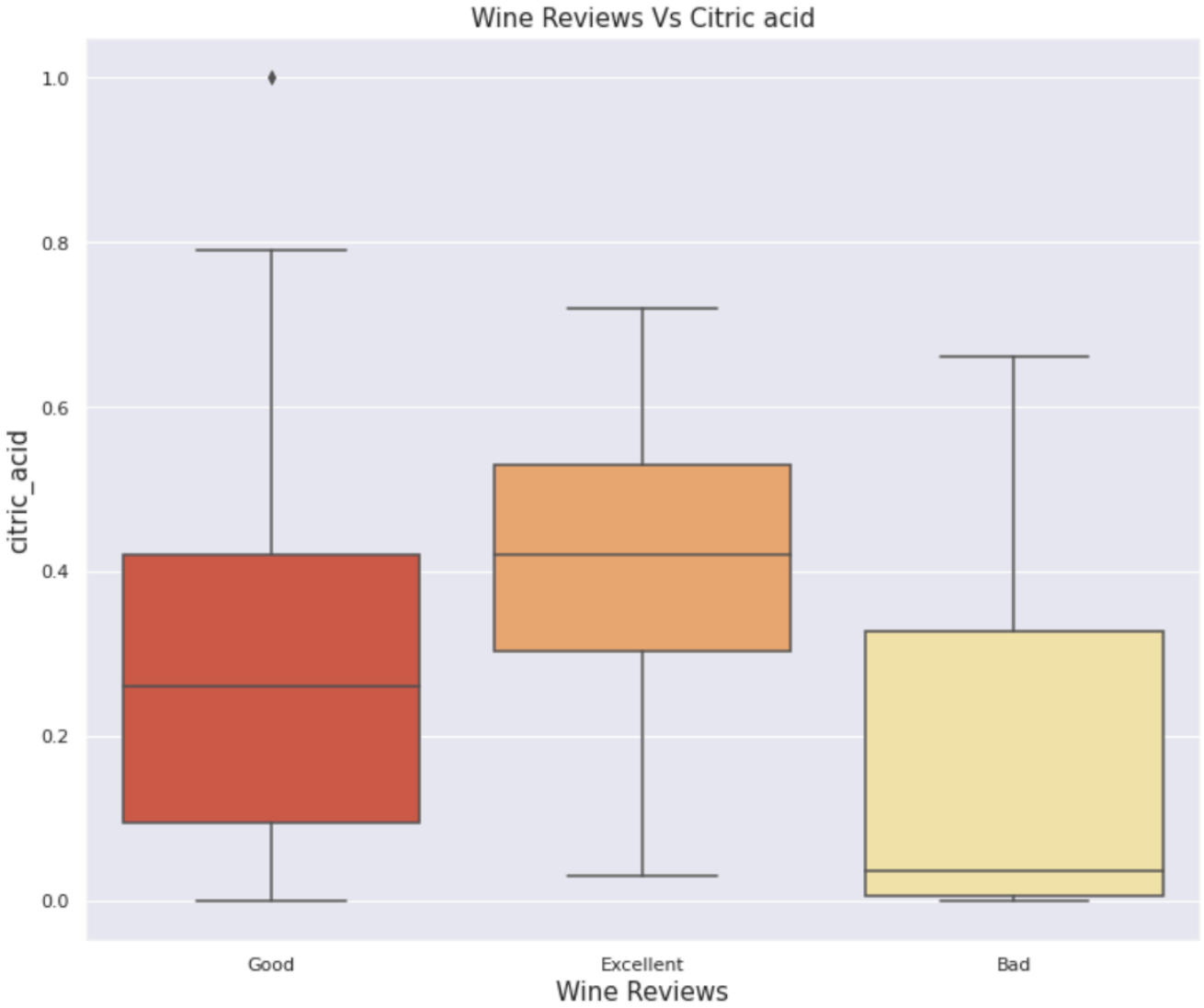
SULPHATES VS REVIEWS?

- Wine to being a prestigious quality requires an average sulphates level to be around 0.75 grams.
- The bad wine contains sulphate level approximately below 0.55 grams.
- The box plot also helps to identify the outliers.



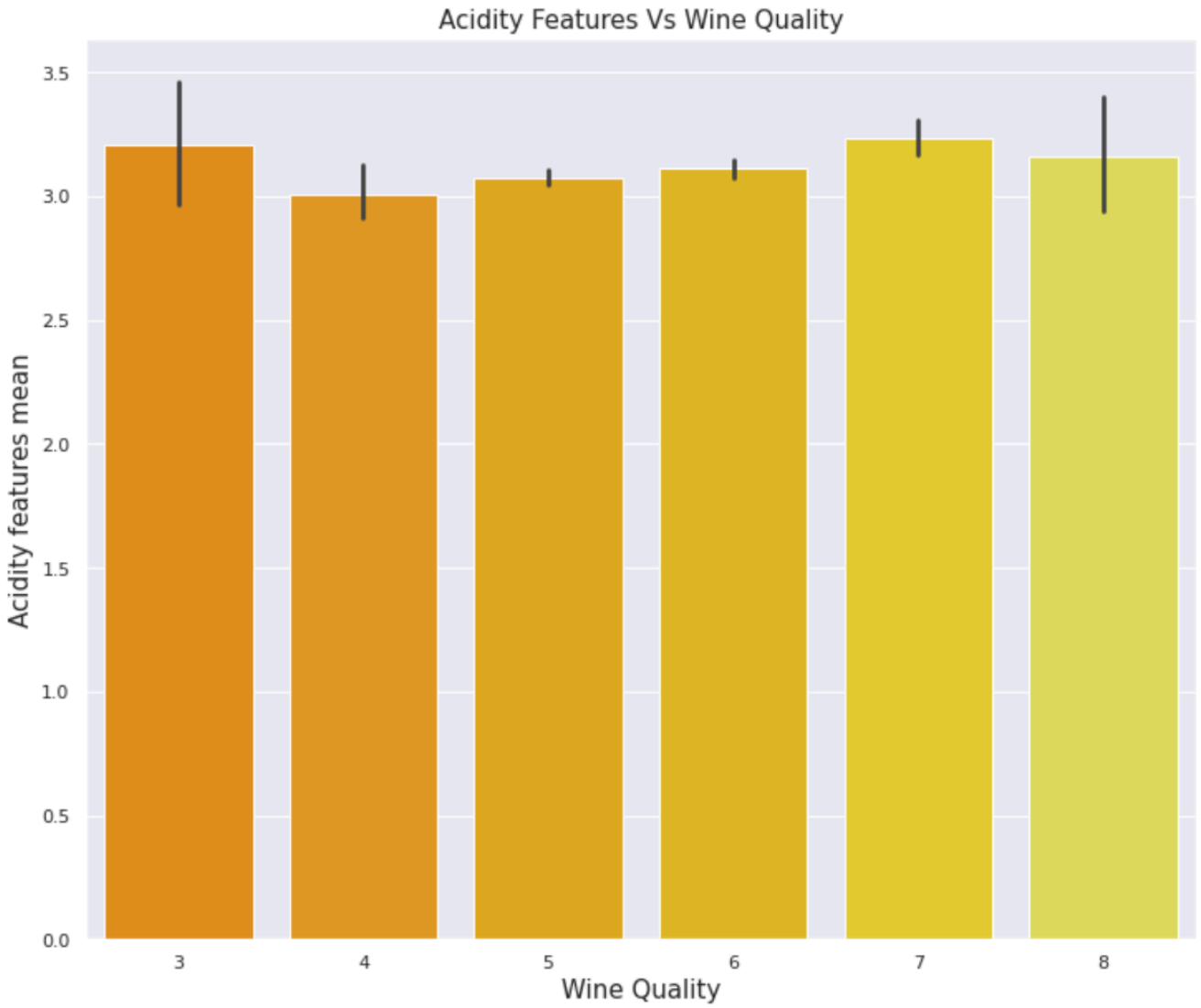
PH VS REVIEWS?

- The pH shows how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)
- Most wines are between 3-4 on the pH scale
- The excellent wine contains pH, which is approximately an average of 3.2 grams per decimetre whereas bad wine contains pH approximately 3.4 grams per decimetre.



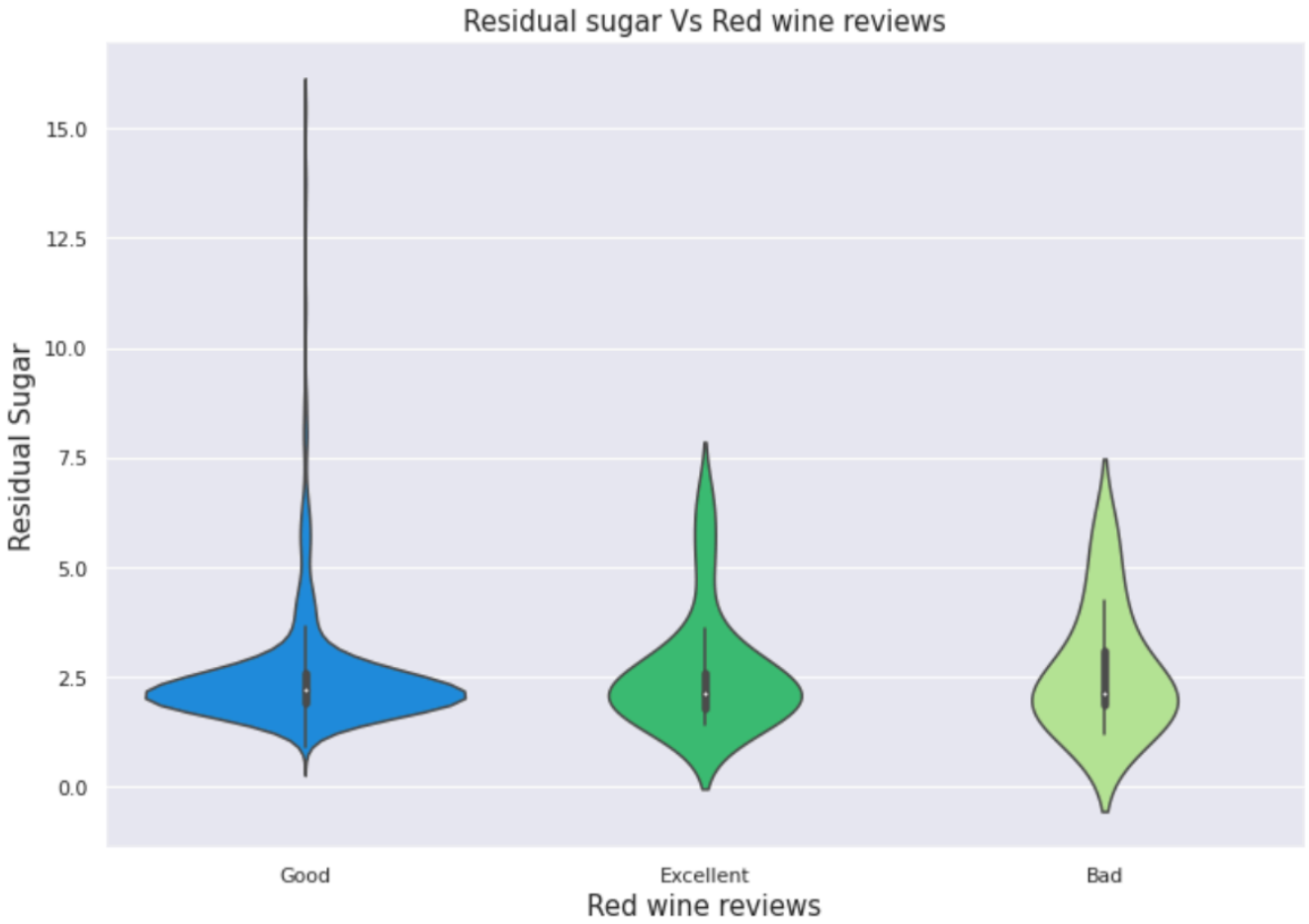
CITRIC ACID VS REVIEWS?

- It found in small quantities, citric acid can add 'freshness' and flavour to wines
- In order to be qualifies as better wine, it should have higher amount of citric acid in it.



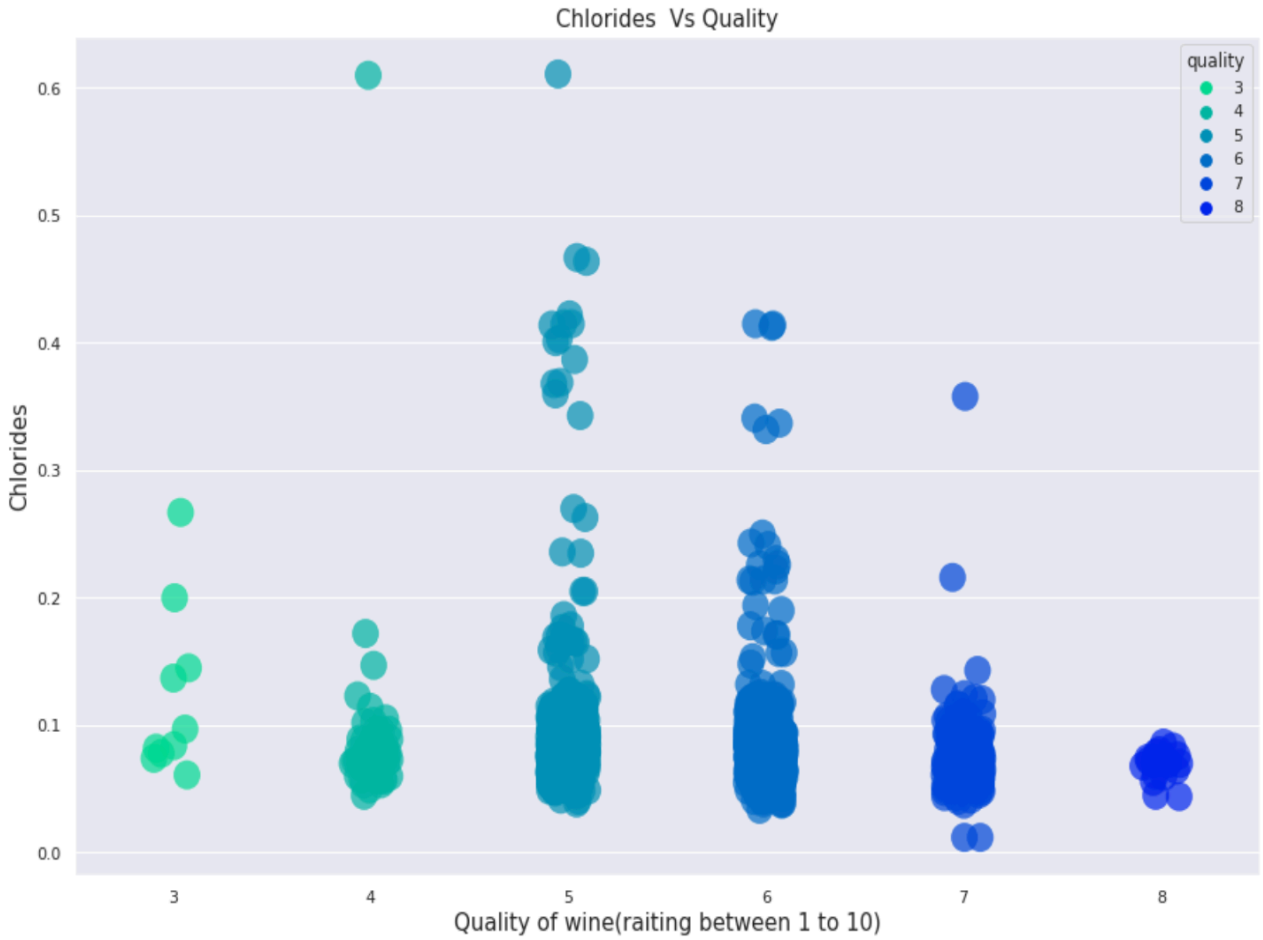
ACIDITY IN WINE VS QUALITY?

- To plot this graph I have created new column of “acidic_features” ' to group the acidic features and visualize the data with 'quality' column which includes ('fixed_acidity', 'volatile_acidity’, 'citric_acid’, 'pH') with mean.
- Almost all kinds of wine have an average of 3 to 3.5 grams per decimetre of acidic content in it. The wine without acidic content is almost impossible to make.



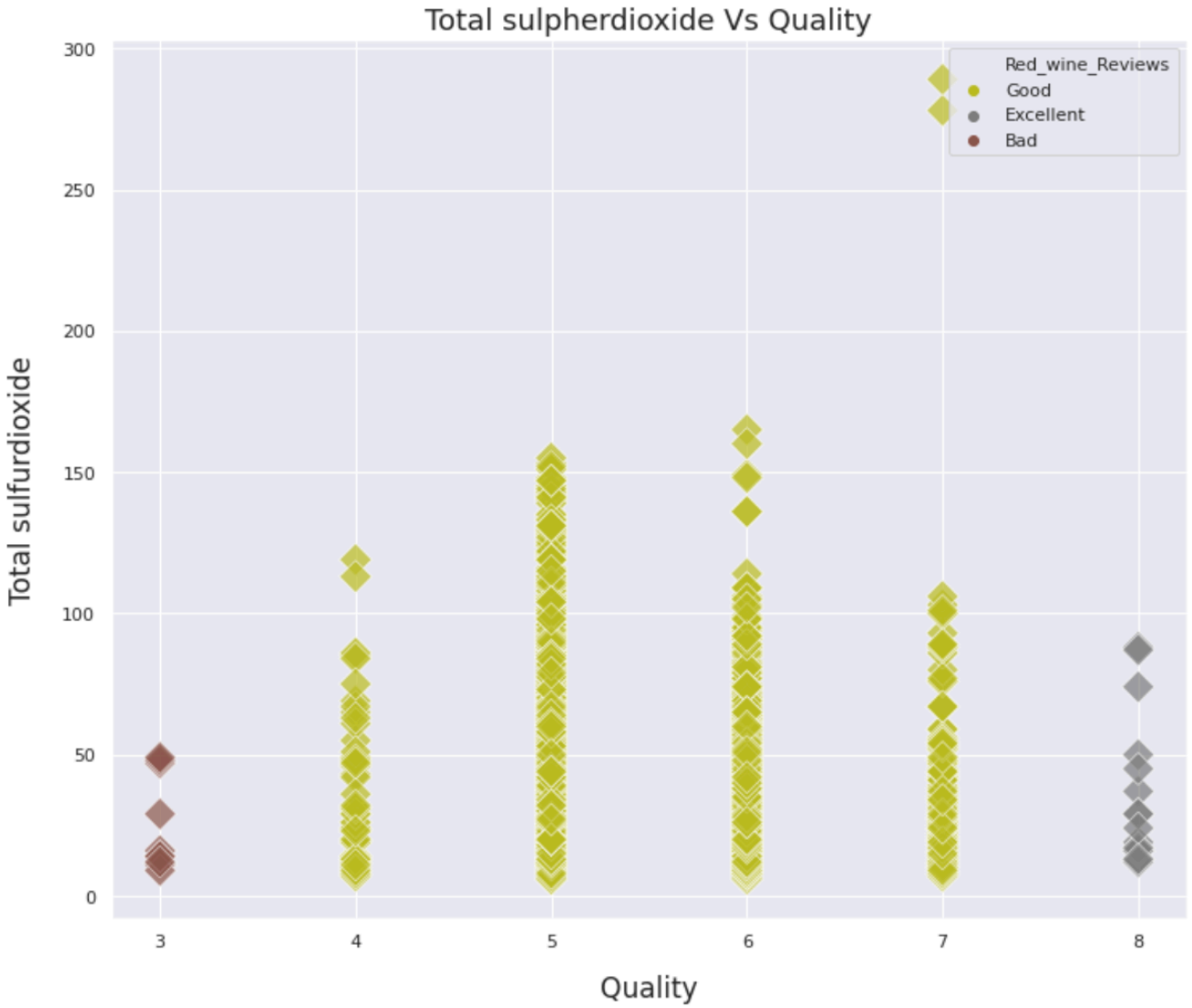
RESIDUAL SUGAR VS RED WINE REVIEWS?

- It is difficult to find a wine with sugar level less than 1 gram/litter .
- Wines with greater than 45grams/litter considered to be a sweet wine.
- It is apparent that wine quality does not affect by the sugar level as far as the sugar level is close around 2.5 grams per decimetre



CHLORIDES VS QUALITY?

- Chlorides is the amount of salt in the wine.
- Chlorides and quality appear a negative relationship for red wines. which can be interpreted as salt is not much required for a good quality.



TOTAL SULFUR DIOXIDE VS REVIEWS?

- Total sulfur dioxide shows a negative relationship to quality in red wines.
- low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

PREPROCESSING

- **Checking missing values:**
  The dataset does not contain any missing values

- **Dropping columns:**
  To leverage data for better efficiency and accuracy, dropped columns which are redundant.(dropped 1 columns "Red_wine_review" which was created for visualization purpose)

- **Removing Outliers:**
  Important to remove outliers since they would most likely affect the performance of machine learning models.
- I have used zscore() function to remove outliers which is defined in SciPy library and set the threshold = 3.
- Z-score identifies how many standard deviations away a data point is from the mean. The data points which are too far from the mean are considered as outliers.

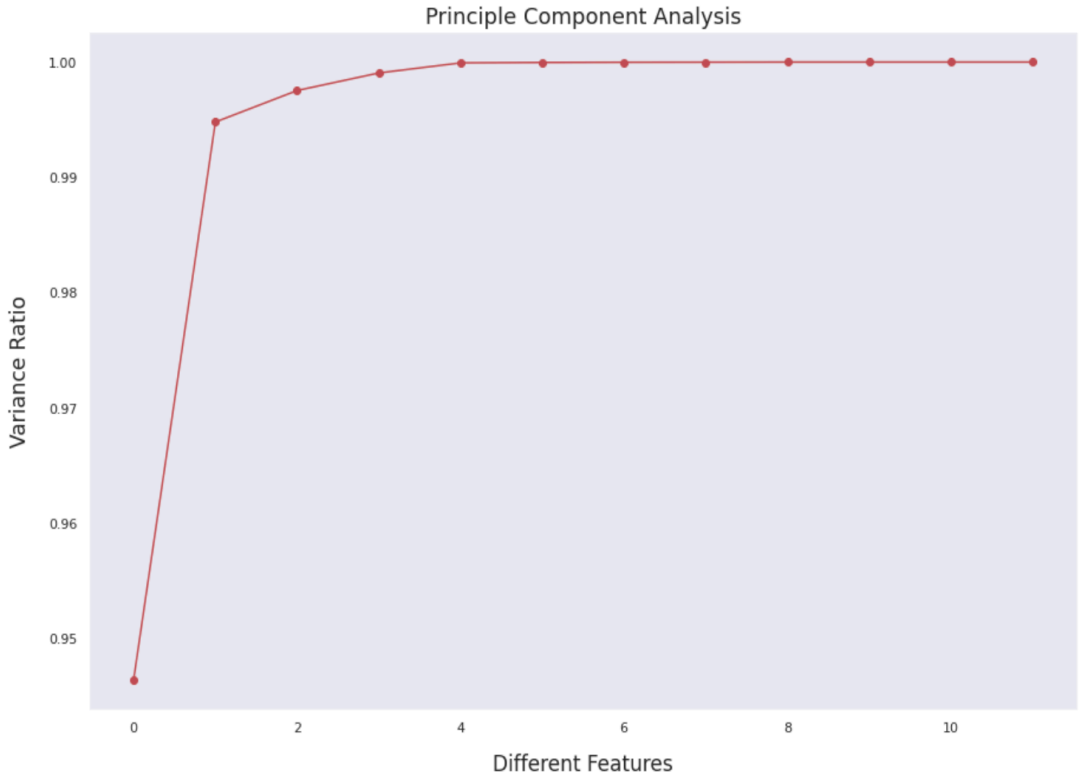
- **Principle Component Analysis (PCA):**
  To speed up a machine learning algorithm PCA is used.

- **Final arrangements before model comparison:**
  Sorting important features.

Need of PCA: High dimensional data is extremely complex to process due to inconsistence in the features which increases the computation time and make data processing more complex.

What is PCA?

It's a dimension reduction techniques that enables you to identify the correlations and patterns in a data set so that it can be transformed into a dataset of significantly lower dimensions without loss of any important information.



PRINCIPLE COMPONENT ANALYSIS(PCA)

- Explained variance ratio tells us how much information(variance) can be attributed to each of the principal components
- The initial 5 features have a high variance rest of the other are the almost at same level
- The below is the list of 5 features is being use for machine learning models

	0	1
0	PC0	total_sulfur_dioxide
1	PC1	free_sulfur_dioxide
2	PC2	fixed_acidity
3	PC3	residual_sugar
4	PC4	alcohol

MODELS IMPLEMENTED

- **Logistic Regression:**

Logistic Regression is used when the dependent variable(target) is categorical.

- **Random Forest Classifier:**

The random forest algorithm can be used for both regression and classification tasks. It mainly joins different types of algorithms or the same algorithm multiple times to form a more powerful prediction model.

- **Decision Tree:**

Decision Tree is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables

- **KNN Classifier:**

KNN is k nearest neighbours algorithm. It is a supervised learning technique which can be used for classification and regression problems but mostly for classification problems.

COMPARISON OF MODELS

Model	Accuracy Scores	Cross Validation	ROC/AUC	f1-score(with weighted average)
Random Forest	0.88	0.86	0.67	0.88
KNN Classifier with optimum k value	0.87	0.00	0.55	0.83
Decision Tress	0.86	0.82	0.72	0.86
Logitic Regression	0.86	0.85	0.57	0.83
KNN	0.84	0.00	0.00	0.83

- The table contains information of model with model evaluation techniques accuracy score, cross validation, ROC/AUC and f1-score
- All the model are preforming with good by accuracy, but the best model is Random Forest which we will use it for predicting the wine quality.

TESTING SAMPLE DATA

	fixed_acidity	residual_sugar	free_sulfur_dioxide	total_sulfur_dioxide	alcohol	Prediction	0	1
0	7.3	2.0	17	106	9.5	0	0.72	0.28
1	6.7	2.0	8	30	10.5	0	0.72	0.28
2	5.6	6.1	8	121	9.2	0	0.70	0.30
3	7.5	1.8	9	95	10.5	0	0.74	0.26
4	5.6	1.6	50	37	9.3	0	0.74	0.26
5	5.6	1.6	11	29	9.3	0	0.74	0.26
6	7.4	1.7	13	29	9.3	0	0.74	0.26
7	8.9	1.8	22	112	9.2	0	0.74	0.26
8	7.9	6.1	17	44	9.0	0	0.70	0.30
9	7.6	1.7	28	40	9.3	0	0.74	0.26
10	6.9	2.4	8	129	9.2	0	0.72	0.28
11	8.9	1.8	4	52	9.7	0	0.68	0.32
12	7.9	1.7	17	43	9.5	0	0.74	0.26
13	7.6	2.3	8	14	9.5	0	0.72	0.28
14	7.9	1.8	12	49	9.4	0	0.74	0.26
15	7.9	1.8	9	40	9.5	0	0.74	0.26
16	8.5	1.6	8	38	9.5	0	0.74	0.26
17	6.9	2.3	8	72	9.4	0	0.72	0.28
18	7.9	2.3	17	58	9.5	0	0.72	0.28

- **Using Dummy Excel File:**
  - Created an excel file with 5 columns (which we have used during modelling).
  - The dummy excel file is generated using function '=INDEX(X1:XN,RANDBETWEEN(x1,xn)'
  - To predict the wine quality here, I have used a model with best accuracy(88%) name Random Forest

Results:

By observing above values we can interpret that for given inputs the quality of wine will be bad.

More detail view, low free sulfur dioxide, residual sugar and high alcohol shows that 72% that wine being bad.

TESTING SAMPLE DATA

- **Using single row dummy data:**

To show the performance of different models for same single line dummy data

Input:

```
# The attribute names are in following sequence,
# 'fixed_acidity', 'residual_sugar', 'free_sulfur_dioxide', 'total_sulfur_dioxide', 'alcohol'
Xnew = [[7, 2, 9, 18, 12]]
```

Results:

```
Result of classification:
Logistic Regression: Label = [1] --> good wine
Random Forest Classifier: Label = [0] --> bad wine
Decision Tree: Label = [0] --> bad wine
K-Nearest Neighbors Classifier: Label = [0] --> bad wine
K-Nearest Neighbors Classifier: Label = [0] --> bad wine
```

Observation:

While most models seem to agree with Random Forest model prediction, the logistic regression is the only outlier (As can be seen from the above result)

Hence, we can prove that the wine quality can be predicated accurately by using good accuracy score model (Random Forest)

IMPLEMENTATION OF PROFESSOR'S FEEDBACK

We would like to expand our implementation to find out consensus among the models:

- If all models agree on the same outcome then that should be the final outcome.
- If there is a tie then the model with higher accuracy will determine the final outcome.
- Attribute sequence - 'fixed_acidity', 'residual_sugar', 'free_sulfur_dioxide', 'total_sulfur_dioxide', 'alcohol'

Conditions

Case 1: The count of **bad wine is greater than good wine**, then the *majority* models **predicts quality of wine will be Bad wine**

Case 2: The count of **good wine is grater than bad wine**, then the *majority* models **predicts quality of wine will be Good wine**

Case 3: The count of **good wine is equal to count of bad wine**, then the ***model with highest accuracy*** will display with its prediction(which might be good or bad)

Input1:

```
Xnew = [[7, 2, 9, 18, 12]]
```

```
txt = [label_LR,label_RF,label_DT,label_KNN1]
```

```
txt
```

```
['Good', 'Bad', 'Bad', 'Bad']
```

Results:

```
Majority models predicts quality of wine will be Bad wine
```

Input2:

```
Xnew = [[3.5, 2.5, 9, 52, 12]]
```

```
txt = [label_LR,label_RF,label_DT,label_KNN1]
```

```
txt
```

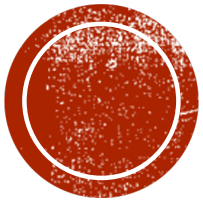
```
['Good', 'Bad', 'Bad', 'Good']
```

Results:

```
Random Forest
0.88
Bad
```

CONCLUSION

- In this assignment, I used K-Nearest Neighbours, Logistic Regression with polynomial features, Decision Tree, and Random Forest. With the accuracy score, cross-validation, classification report and ROC/AUC score for each model.



- To speed up a machine learning algorithm, I have used Principal Component Analysis (PCA) since input dimension is too high.

- Even though the quality of a wine is a subjective matter, prediction can still help tremendously to make a better choice. We were able to predict the wine quality using the best model (i.e. random forest), with accuracy score 88%.

REFERENCES AND LINKS

1] Understanding PCA (Principal Component Analysis) with Python
https://towardsdatascience.com/dive-into-pca-principal-component-analysis-with-python-43ded13ead21

2] The python graph gallery
https://python-graph-gallery.com/110-basic-correlation-matrix-with-seaborn/

3] Styling plots with Seaborn
http://jose-coto.com/styling-with-seaborn

4] What is Acidity in Wine?
https://www.winemag.com/2019/06/19/what-is-acidity-in-wine/

5] PCA using Python (scikit-learn)
https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

6] Ways to Detect and Remove the Outliers
https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
https://en.wikipedia.org/wiki/Vinho_Verde
https://www.w3schools.com/python/python_lambda.asp

Link to notebook:
https://colab.research.google.com/drive/1_Qwob1tFwH7EMnSXDB_FKPeq8lpSO8Wh