# Wine Classification

- by Mugdha Paithankar

Beer is made by men, wine is the drink of gods.

# DATA

THE SCIENCE BEHIND WINES

- UCI Machine Learning Repository
- The dataset contains red and white variants of the Portuguese “Vinho Verde” wine.
- physicochemical input variables
- sensory output variable

# Project Procedure


Get the data!

Data Wrangling

Exploratory Data Analysis

Statistical Testing

Machine Learning

Insights/ Conclusion

Future Directions

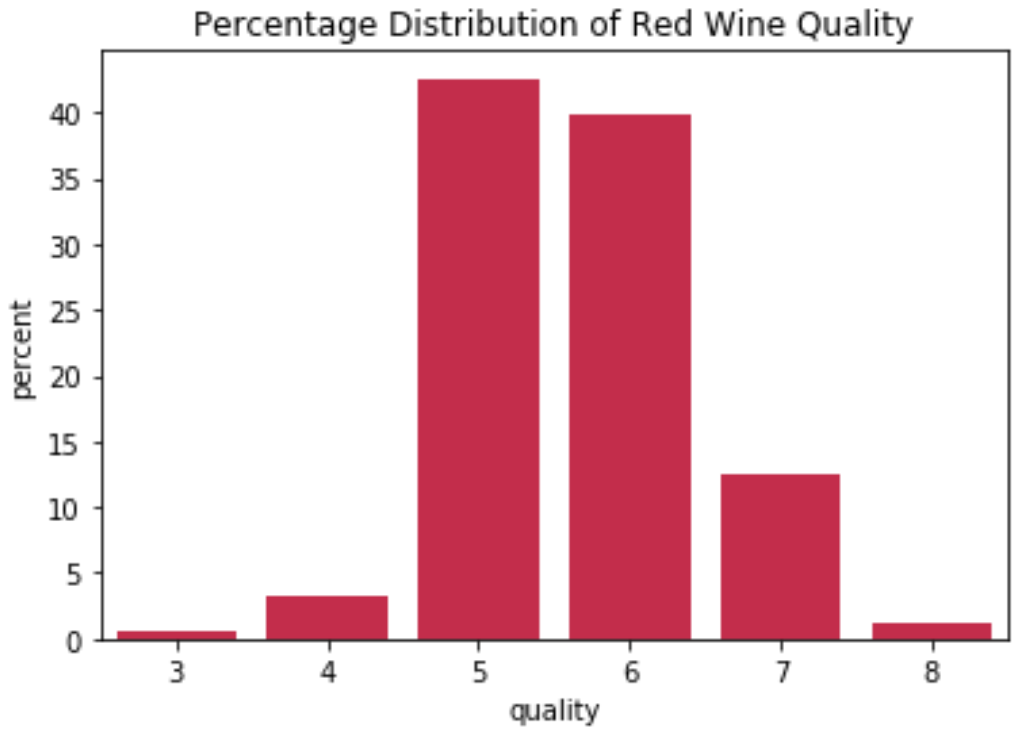
# Obtain the data/ Get to know the data!

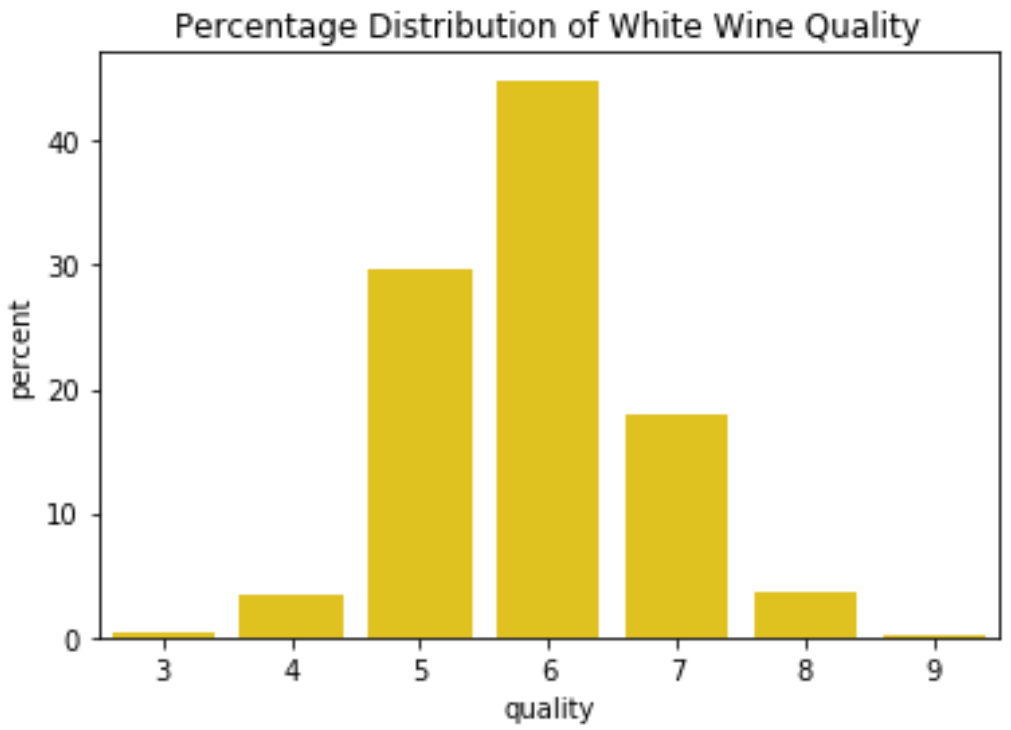
- fixed acidity
- volatile acidity
- citric acid
- residual sugars
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulfates
- alcohol

All features are measured in g/dm^3 except for total + free sulfur dioxide and alcohol which is in mg/dm^3 and % by volume respectively.

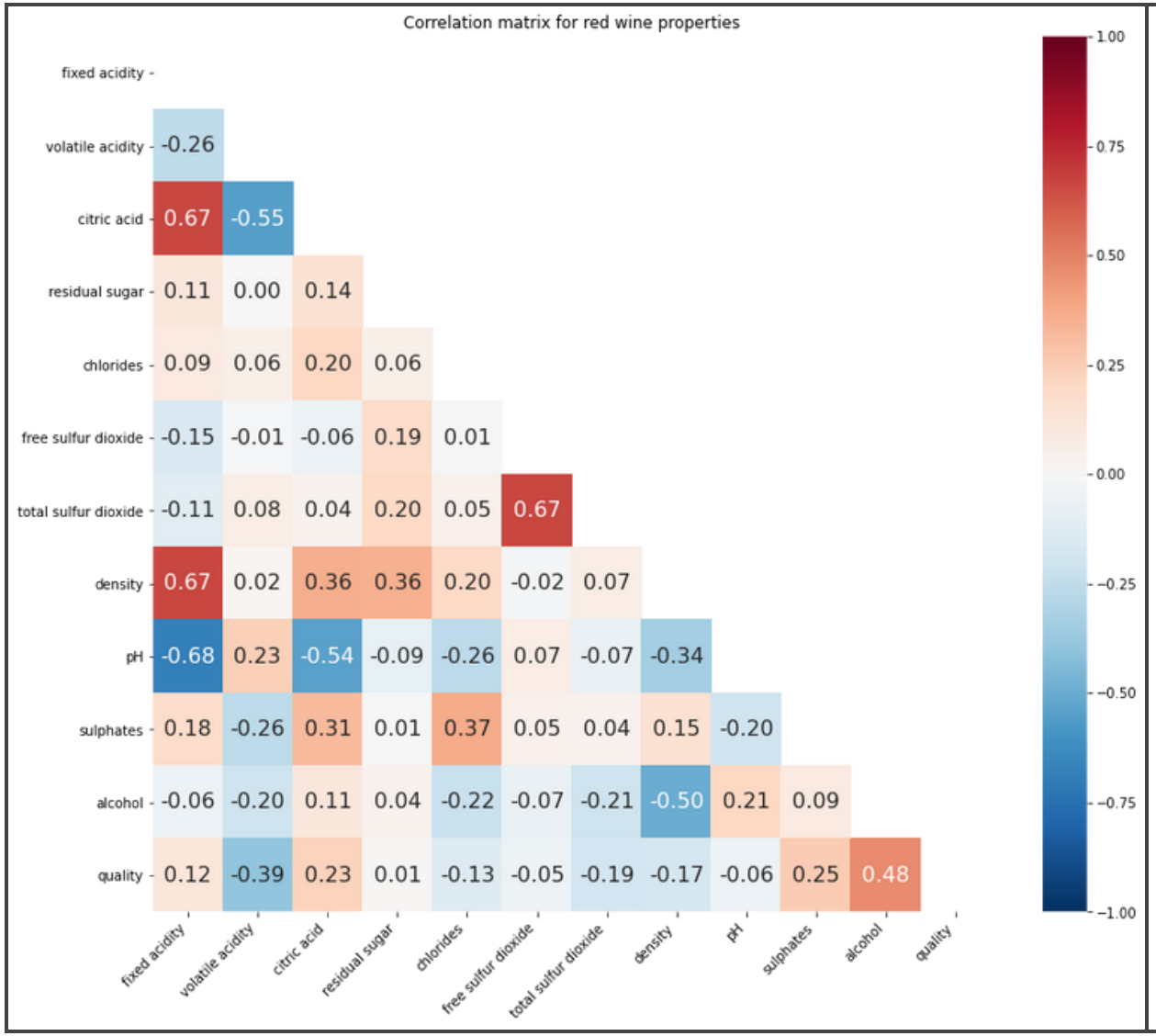
# Let's start with the EDA!

# Quality distribution of red and white wines

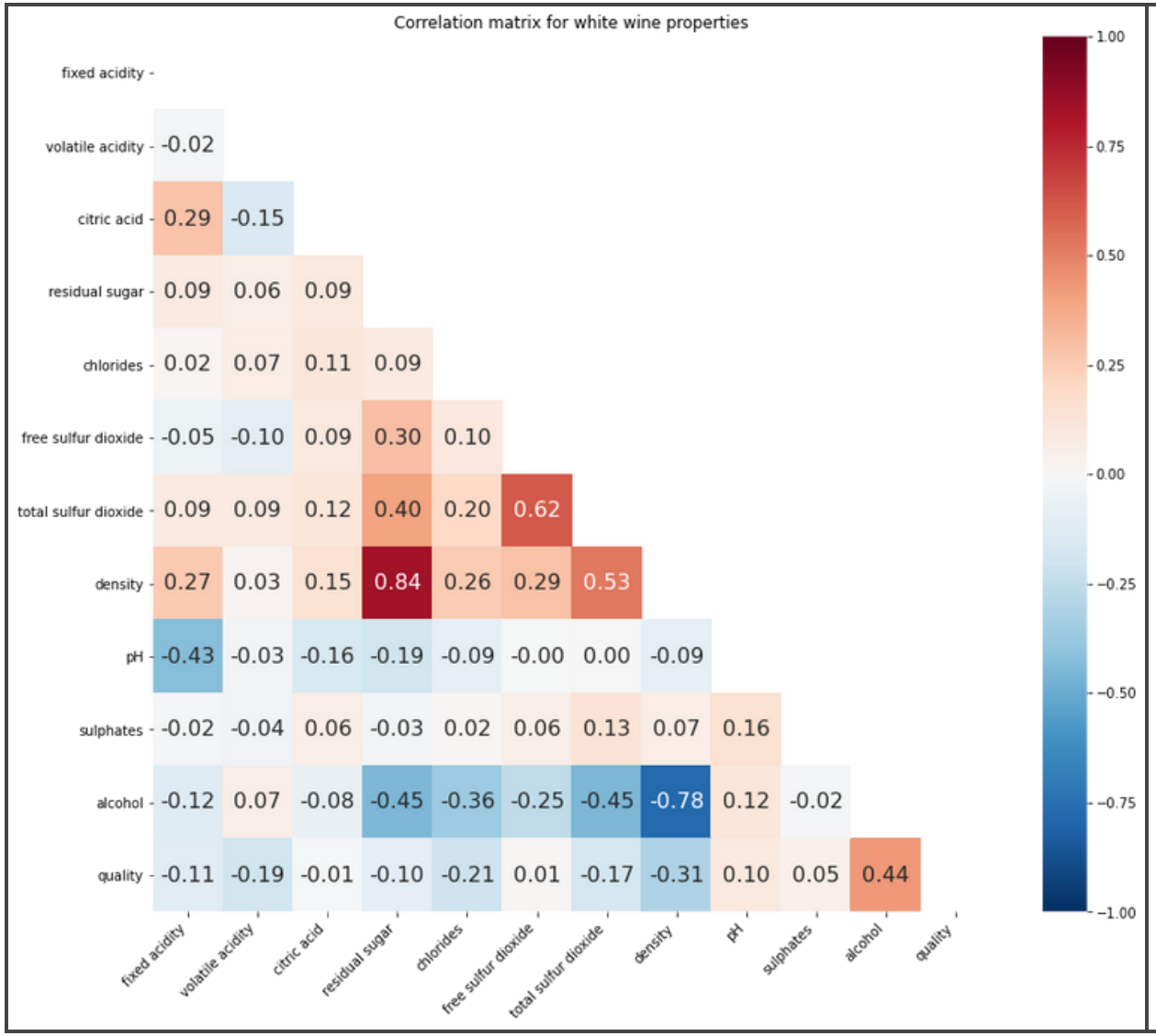



A majority of red and white wines (about 85%) have a quality rating of 5 and 6. There are no 9s for red wine, very few wines (less than 1%) with a quality rating of 3, and only 2% (for red wines) with a rating of 8.




- fixed acidity has moderate correlations with pH, density and citric acid.
- Free sulfur dioxide has strong correlation with total sulfur dioxide
- Alcohol has a correlation of 0.48 with quality.
- pH has the least correlation of -0.06 with quality!



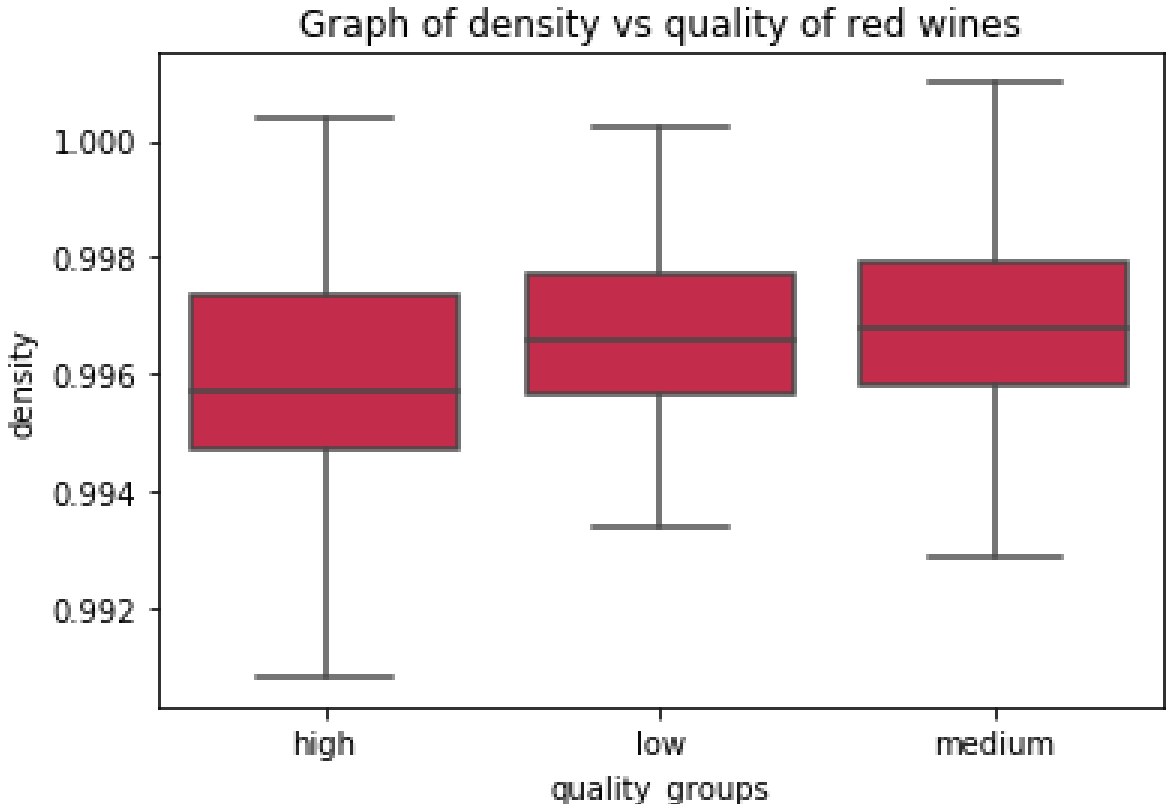

- density has moderate correlations with residual sugar and alcohol
- Free sulfur dioxide has a moderate correlation with total sulfur dioxide
- Alcohol has a correlation of 0.44 with quality.
- citric acid has the least correlation of -0.01 with quality!

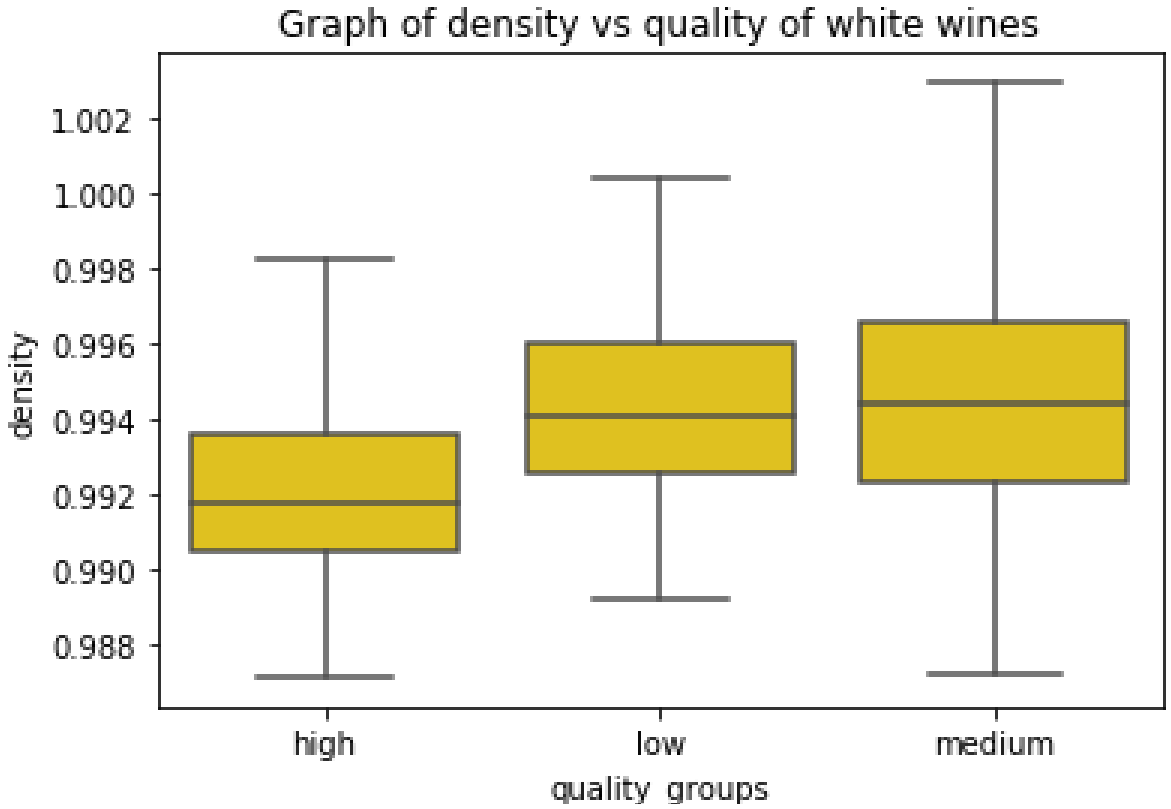
# What next in EDA?

Answer: Box Plots!

- Visualize the change in features with respect to quality of the wine
- I divided the quality ratings of red and white wines into 3 categories:
  - Low (1-4)
  - Medium (5-6)
  - High (7-10)
- do highly rated wines have greater alcohol content? Do low rated wines have higher density?

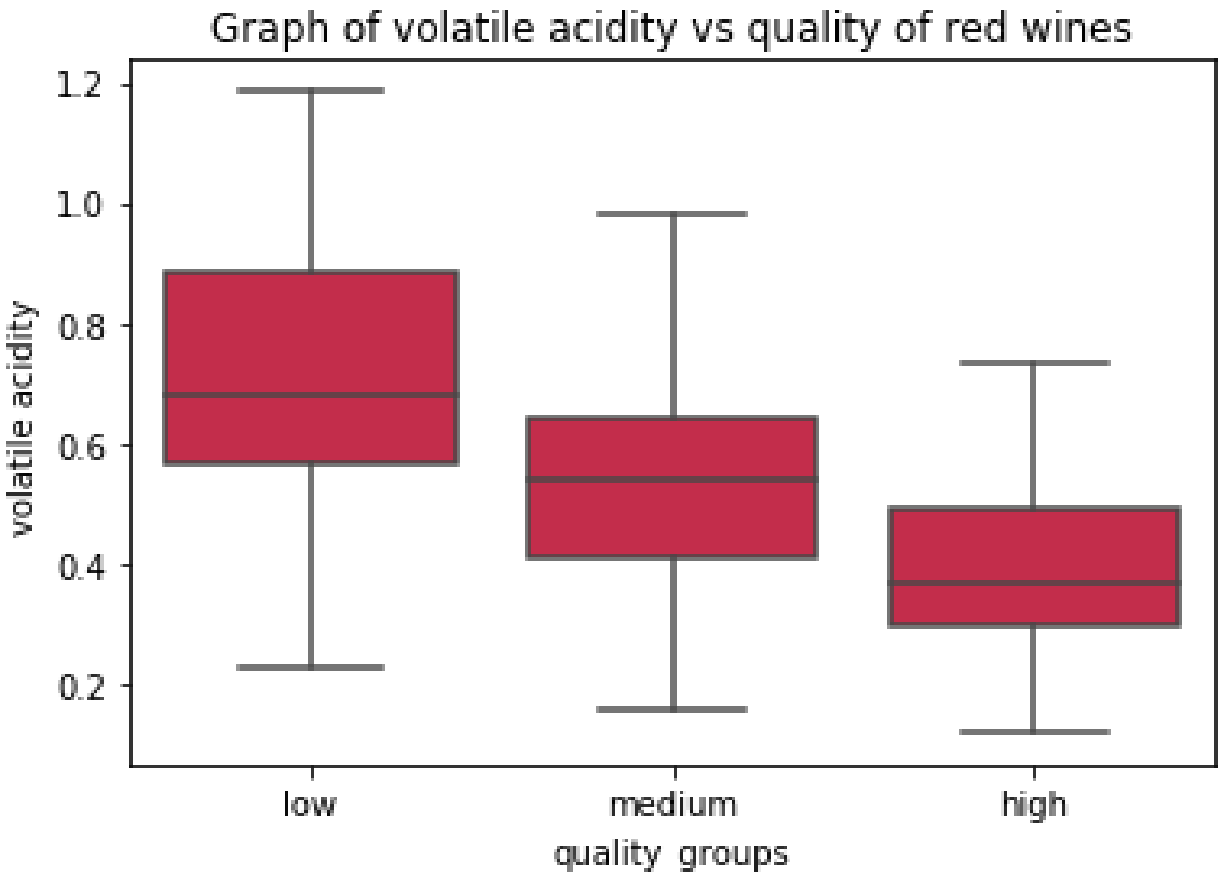
# Box plot 1: Density VS Quality

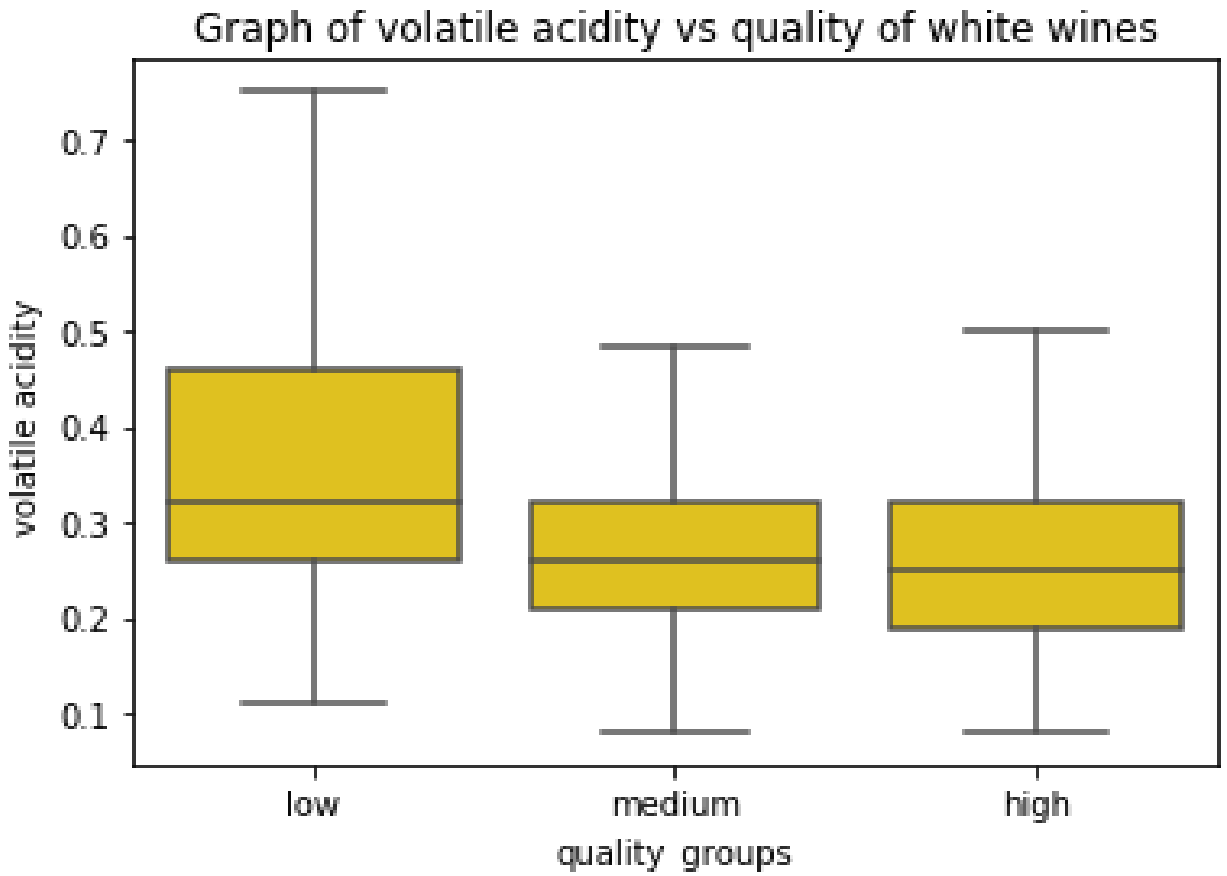



High quality red and white wines seem to have slightly lesser densities compared to low and medium quality ones.

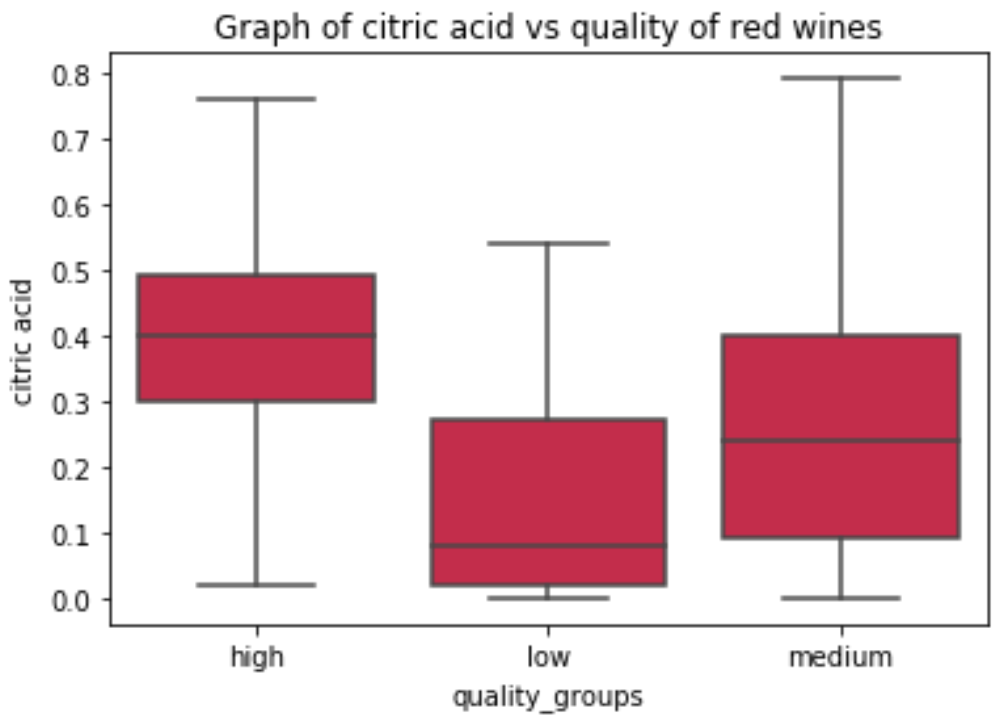
# Box plot 2: Volatile Acidity VS Quality

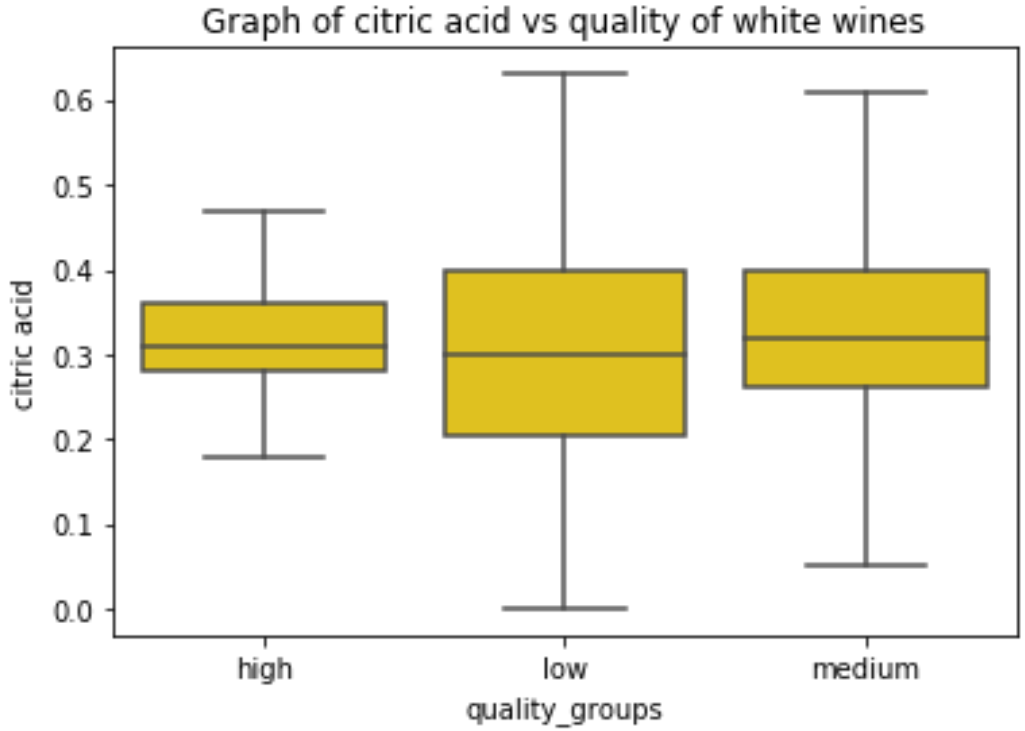



Volatile acidity content is lesser in high quality red and white wines.

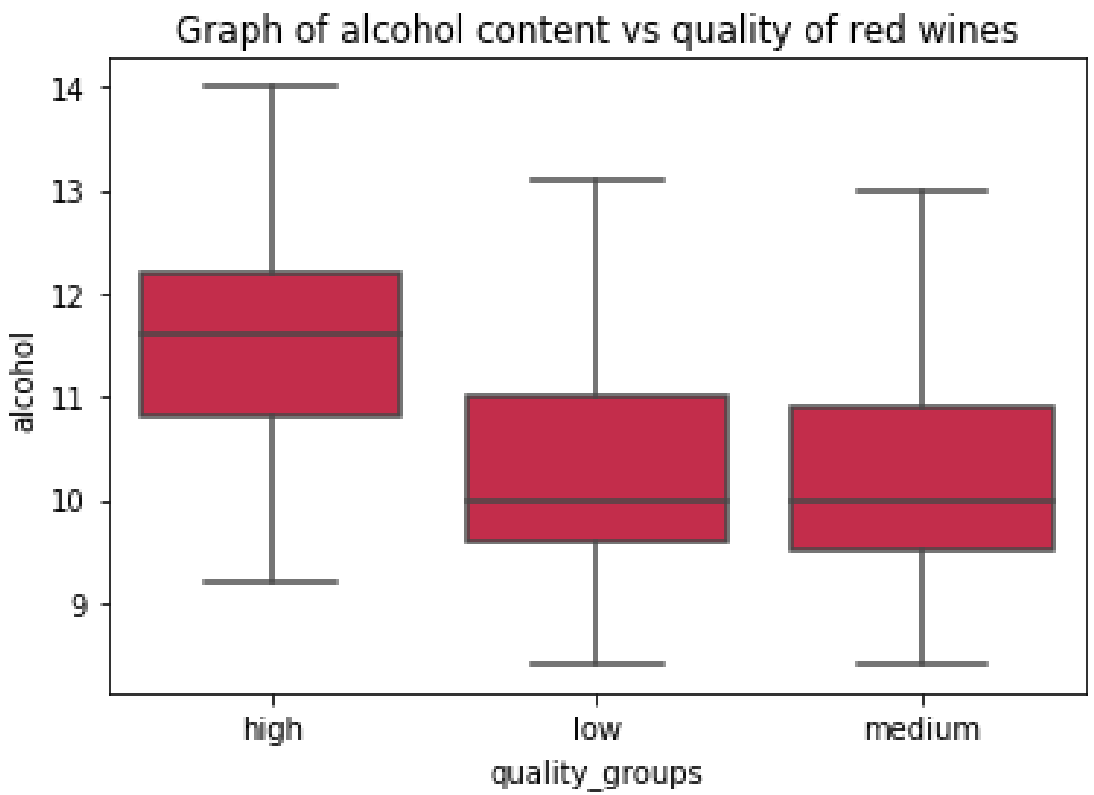
# Box plot 3: Citric Acid VS Quality

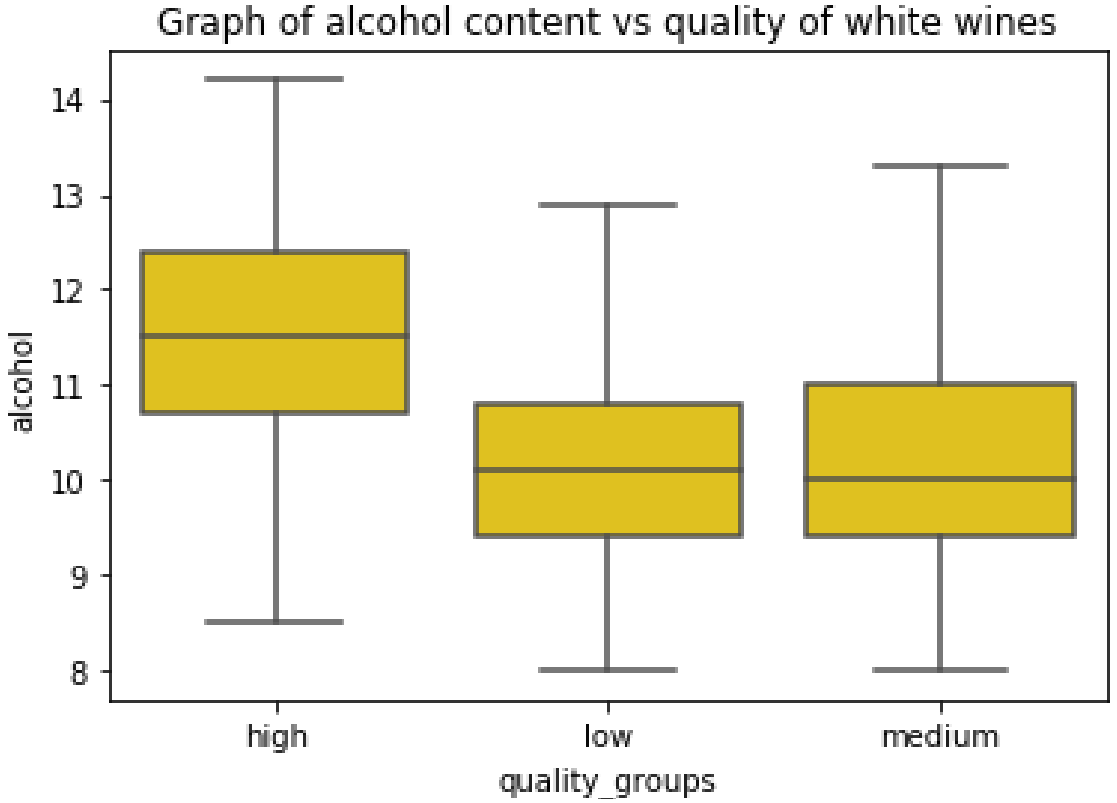



Citric acid content does not vary a lot according to quality groups for white wines. For red wine, however, citric acid seems to have a strong positive relationship with wine quality.

# Box plot 4: Alcohol content VS Quality




The alcohol content for high quality red and white wines seems to vary significantly when compared to other quality groups.

# Tukey's Test:

## OVERVIEW

- This method tests at P<0.05 (correcting for the fact that multiple comparisons are being made which would normally increase the probability of a significant difference being identified).
- A results of 'reject = True' means that a significant difference has been observed.
- Tukey's test compares the means of all treatments to the mean of every other treatment

# Results of Tukey Test

## Statistically significant difference in means: Yes or No.

| Quality Groups | Density | Volatile Acidity | Citric Acid | Alcohol |
|---|---|---|---|---|
| **For Red Wines** | | | | |
| **High-Low** | No | Yes | Yes | Yes |
| **High-Medium** | Yes | Yes | Yes | Yes |
| **Low-Medium** | No | Yes | Yes | No |
| **For White Wines** | | | | |
| **High-Low** | Yes | Yes | Yes | Yes |
| **High-Medium** | Yes | Yes | No | Yes |
| **Low-Medium** | No | Yes | Yes | No |

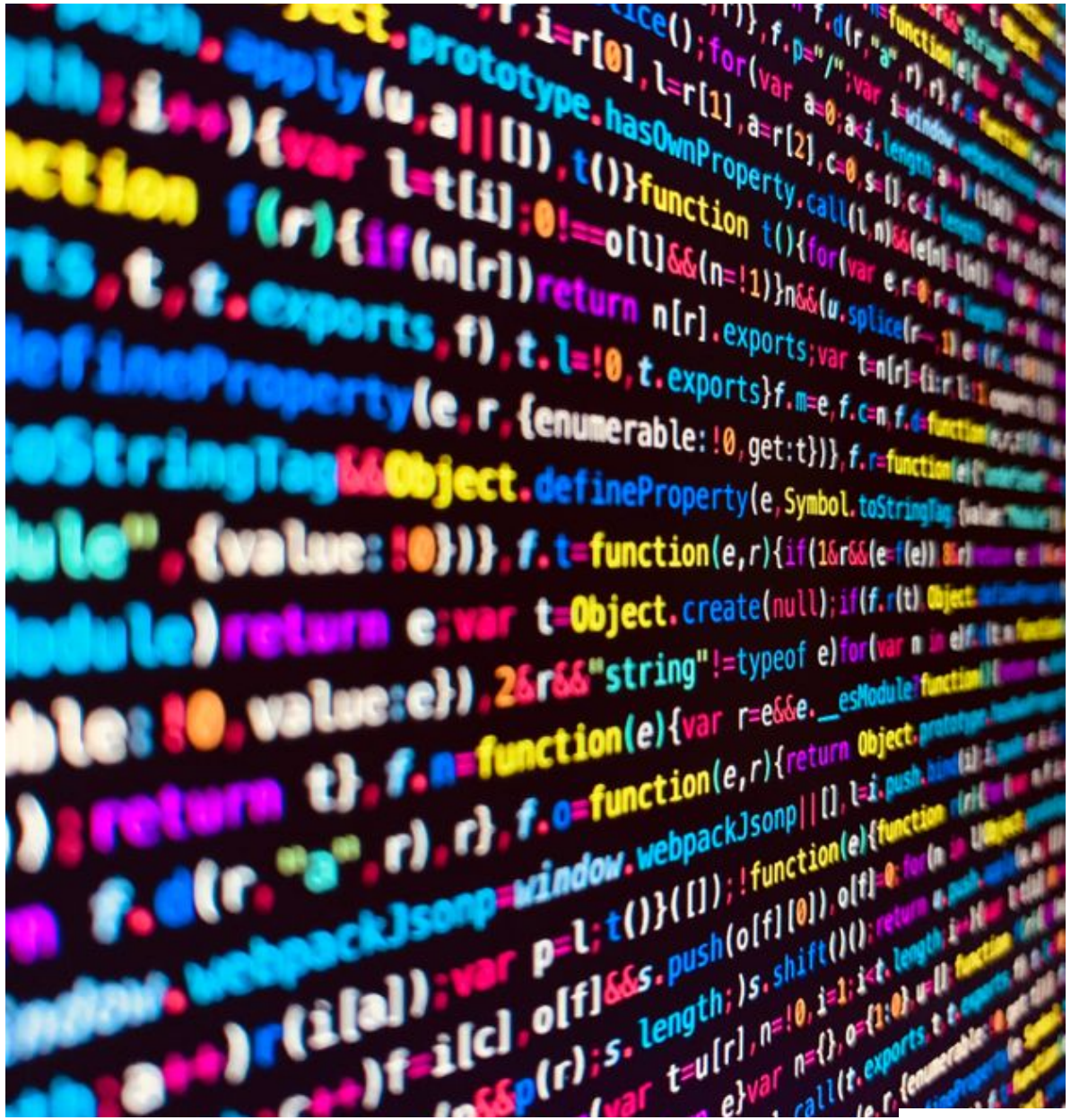
# Further Statistical Analysis

- **VIF Score check for multi collinearity:** A VIF above 10 indicates high correlation and might be a cause for concern. But the VIF scores for the red and white wine dataset were below 5!
- **Recursive Feature Elimination (RFE) using logistic regression:** After scaling the dataset, I implemented RFE to check if any of the wine features are unnecessary or redundant. Features were ranked using RFE's ranking_ and support_ attribute. RFE returned all the 11 features as important.
- **Logistic Regression model:** I built a Logistic Regression model using statsmodels, with all the 11 features.

# Result of Logistic Regression and Odds Ratio

- **Results of Logistic Regression:** After scaling the data, all red wine features except for fixed acidity, citric acid, free sulfur dioxide and density were found to have a statistically significant effect on quality. For white wines, after scaling the data, except for pH and sulfates, all other features had a significant effect on quality.
- **Odds ratio:** highest for sulfates in case of red wines and white wines. It was the least for chlorides in case of red wines and white wines.
- 31.40 odds ratio for red wine sulfates and 2.64 for alcohol.
- 3.67, 3.60, 2.36 odds ratio for sulfates, pH and alcohol in white wines.

# Onto Machine Learning. (where the magic happens!)



# A brief outline of procedure

- In the wines dataset, there were 1382 red wines of “poor quality”, rated below and 7 and 217 of high quality. 3838 white wines were ”poor quality” and 1060 of high quality.
- I used sklearn’s train_test_split to split the already standardized dataset into train and test sets.
- The models were trained using sklearn’s Logistic Regression, XGBoost, Decision Tree and Random Forest.
- All models were tuned for hyper parameters using GridSearchCV.
- AUC score was used to evaluate the model performance.

# Result of grid searching for red and white wines

| Model Type | Parameters with grid search for red wines | Parameters with grid search for white wines | AUC Score :red wines | AUC score:white wines |
|---|---|---|---|---|
| **Logistic Regression** | C: 5.179474679231202, penalty: l2 | C: 268.2695795279727, penalty: l2 | 0.8485 | 0.8033 |
| **Decision Tree** | max_depth=4, max_leaf_nodes=6, min_samples_leaf=1, min_samples_split=2 | max_depth=13, max_leaf_nodes=41, min_samples_leaf=1, min_samples_split=2 | 0.75 | 0.74 |
| **Random Forest** | max_depth : 20, max_features: 0.25, min_samples_split: 2, n_estimators: 150 | max_features: 0.25, min_samples_split: 2, n_estimators: 250 | 0.9200 | 0.9190 |
| **XGBoost** | learning_rate=0.05, max_depth=4, n_estimators=180 | learning_rate=0.1, max_depth=3, n_estimators=100 | 0.9041 | 0.8900 |

# Result of random forest classification

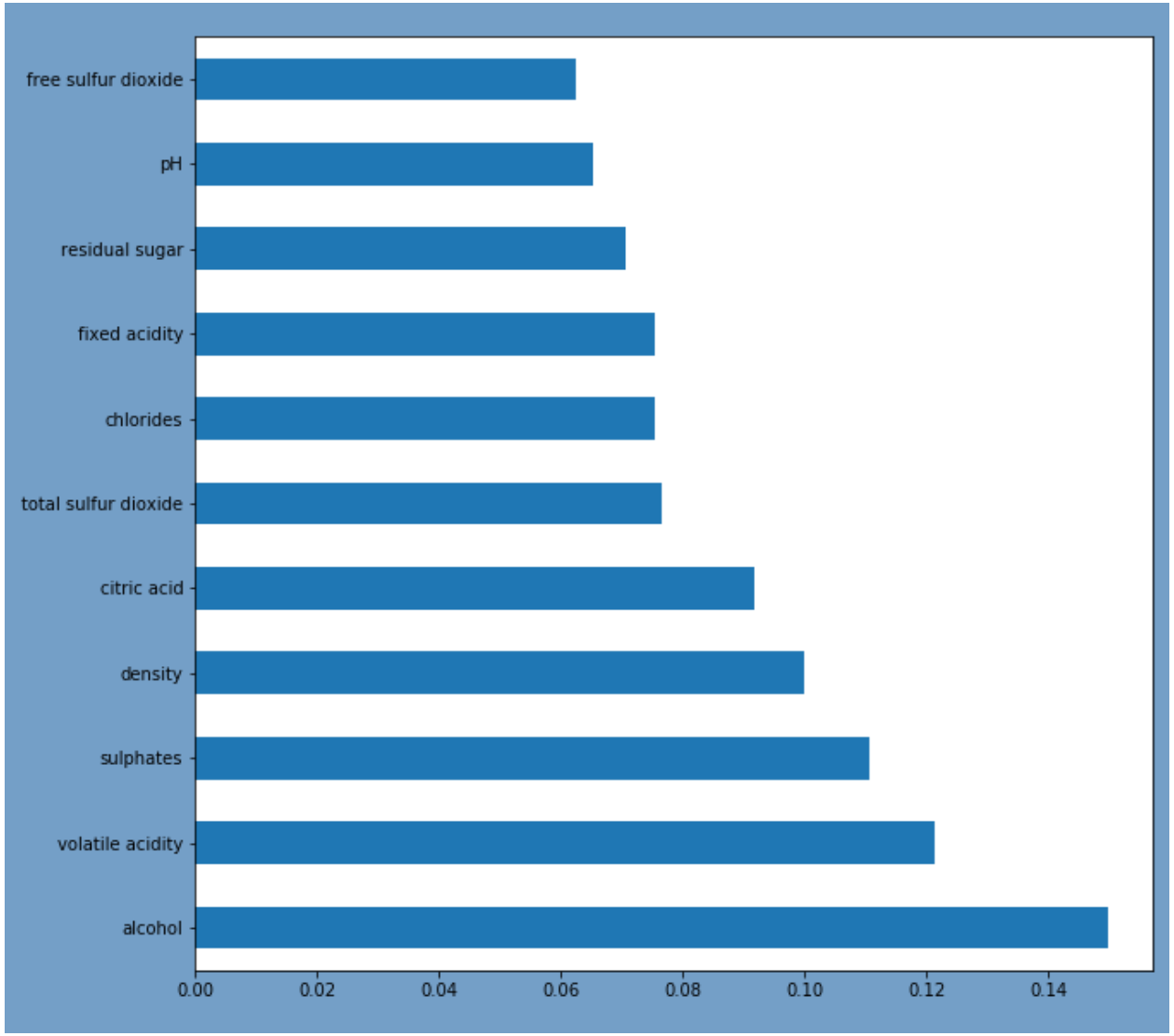
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.93 | 0.98 | 0.96 | 420 |
| 1.0 | 0.82 | 0.52 | 0.63 | 60 |
| accuracy | | | 0.93 | 480 |
| macro avg | 0.88 | 0.75 | 0.80 | 480 |
| weighted avg | 0.92 | 0.93 | 0.92 | 480 |

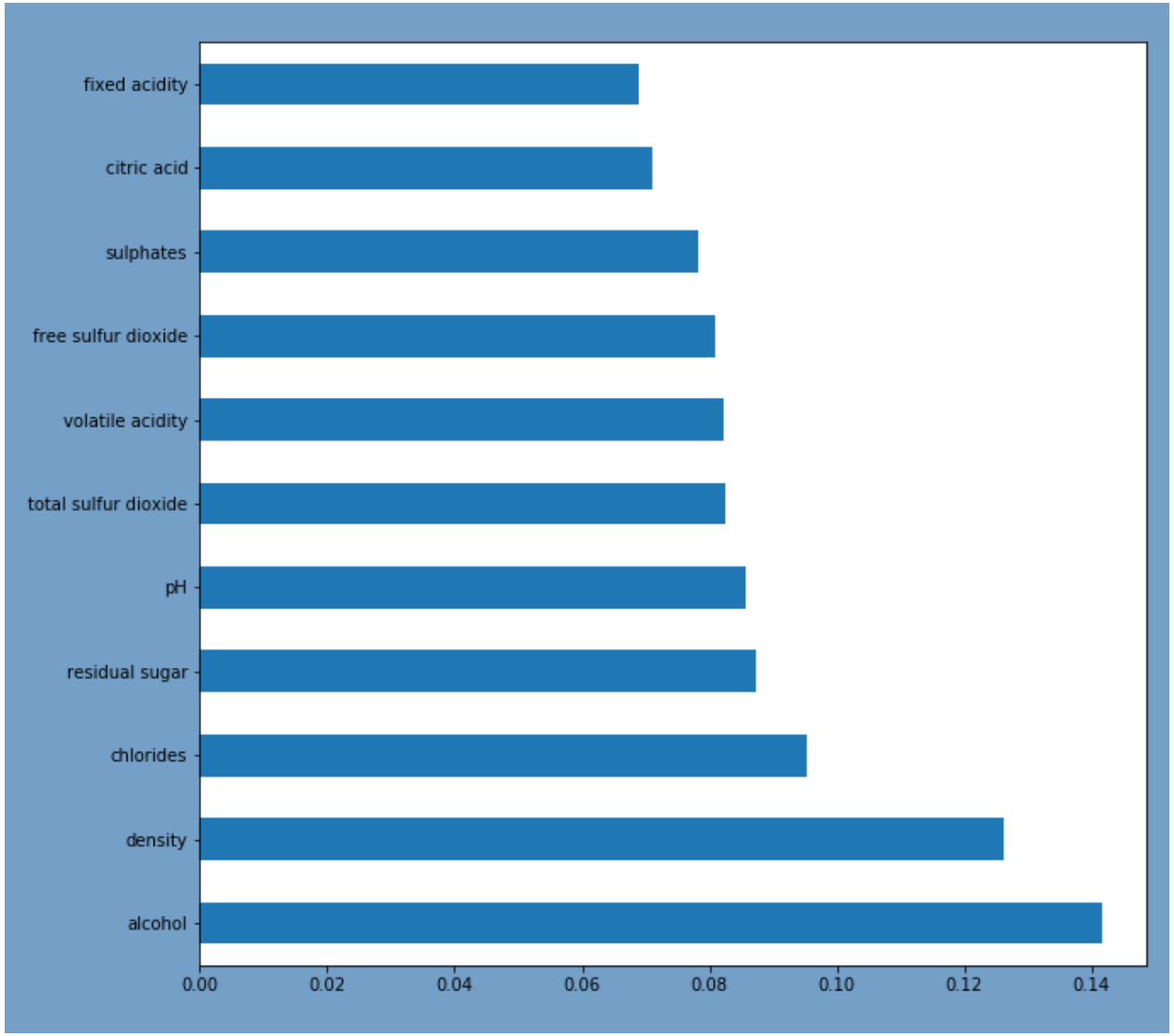
```
[[413   7]
 [ 29  31]]
```

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.88 | 0.97 | 0.93 | 1143 |
| 1.0 | 0.86 | 0.55 | 0.67 | 327 |
| accuracy | | | 0.88 | 1470 |
| macro avg | 0.87 | 0.76 | 0.80 | 1470 |
| weighted avg | 0.88 | 0.88 | 0.87 | 1470 |

```
[[1114   29]
 [ 147  180]]
```

Classification and confusion matrix report for red and white wines respectively




## Feature importances for red and white wines, using random forest's feature_importances_ attribute

# Conclusion

- From the correlation matrix, box plots, Tukey test, logistic regression and the machine learning analysis, it is clear that alcohol content has the largest effect on the quality of red and white wines!
- Sulfate content is another important deciding factor for red wines whereas chlorides and residual sugars could play an important role in white wine quality.
- The odds ratio was high for pH and sulphates but the feature_importances_ attribute of random forest differed from that.

# Future Directions

- Get more data! The dataset had a total of 6497 wine ratings. More ratings could help improve the mode.
- The breadth of the dataset could be improved. This dataset has only red and white wines. Wines like Rose could be added
- The number of features could be increased, we are focusing here on a few physicochemical properties but there may be others that are relevant as well or perhaps even confounding that would lead us to different results
- For this project I focused on binary classification (either low or high rating), but implementing multiclass classification for machine learning with high, low and medium quality groups (like I did in boxplots) could also be informative

# The End!
# Questions?