WINE QUALITY PREDICTION USING MACHINE LEARNING

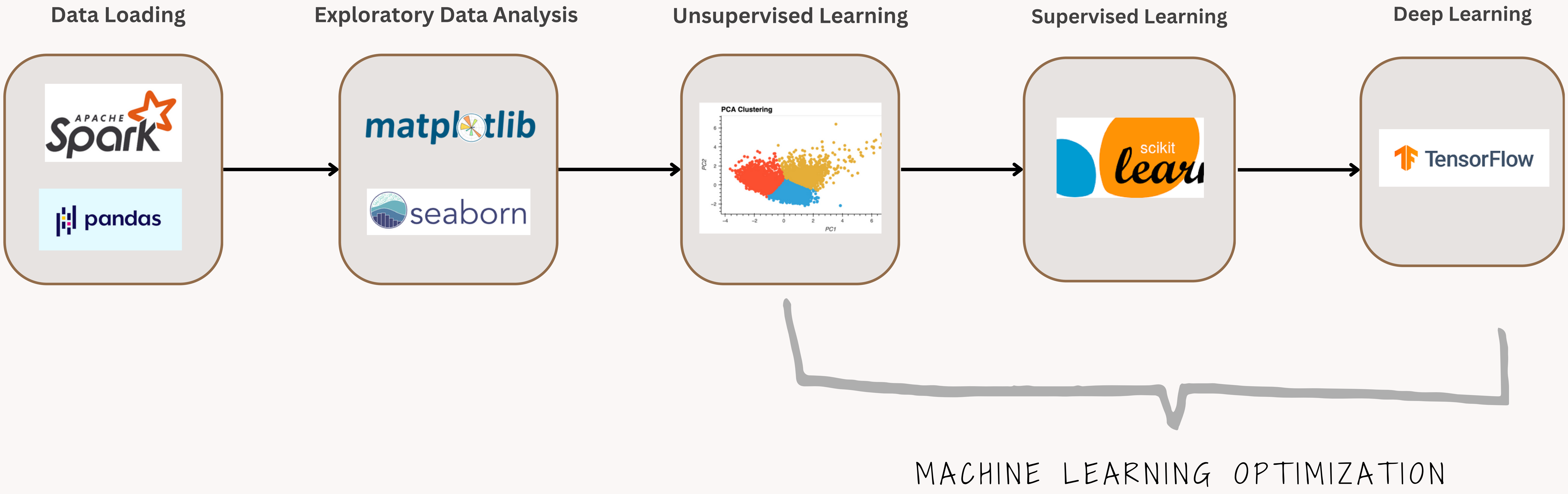
GROUP MEMBERS
KATHERINE HONG
FIZA TARIQ
NIVETHA SUNDAR
SHIPRA GUPTA
MIMI GEORGE

INTRODUCTION

Our goal was to utilize two separate datasets dedicated to red and white wine and analyze them against various components that make wine great, like acidity, citric acid, pH, etc. With this information, we were keen to explore,

- How does each component or ingredient relate to the quality of wine? Does it differ between red and white wine?
- What are some important factors to consider when measuring the quality of both types of wine? Are there significant differences or outliers?
- Can we group/cluster wines based on their chemical compositions?
- How accurate can certain machine learning models be in predicting the type of wine over quality?

DATA PIPELINE



Data Loading

Exploratory Data Analysis

Unsupervised Learning

Supervised Learning

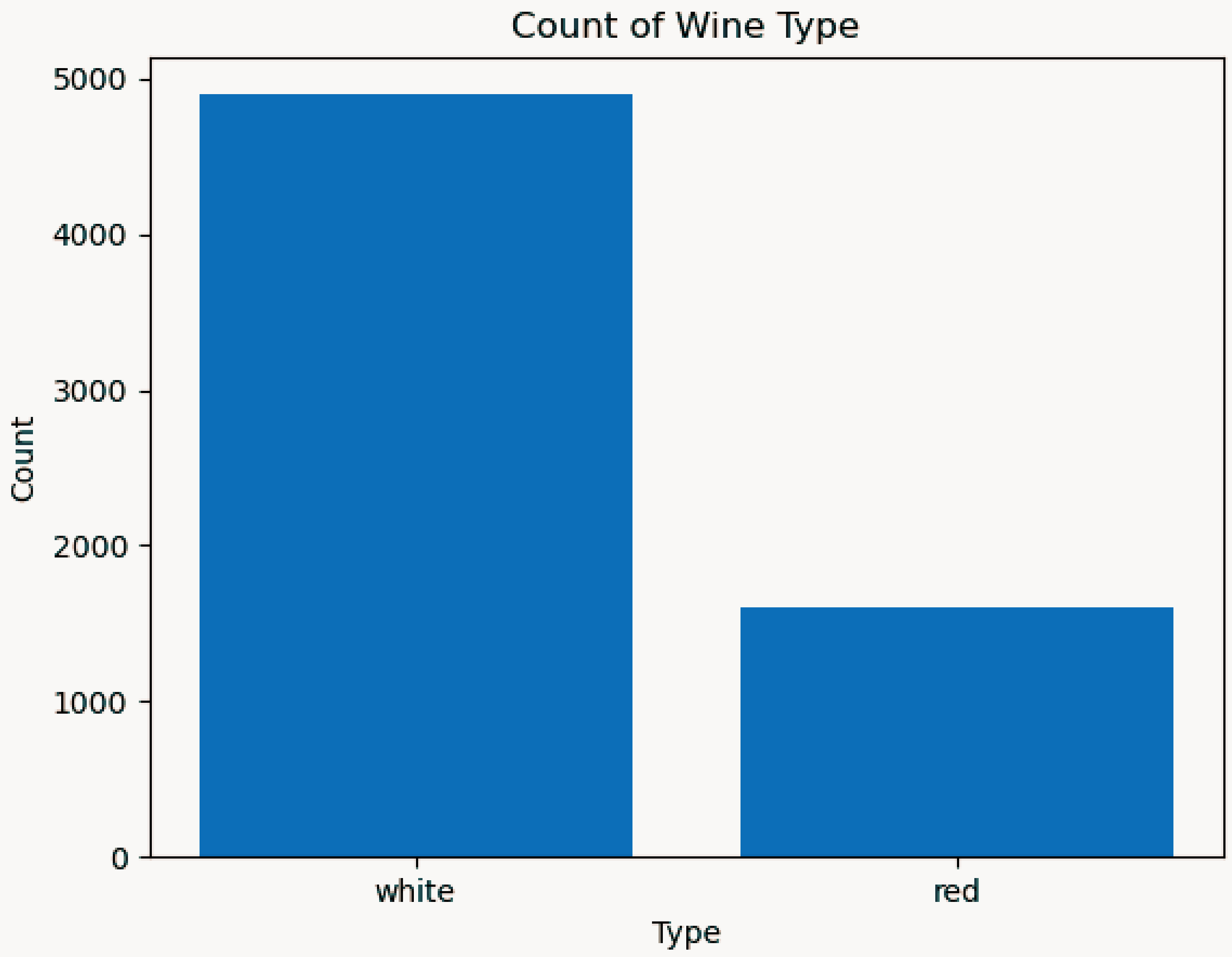
Deep Learning

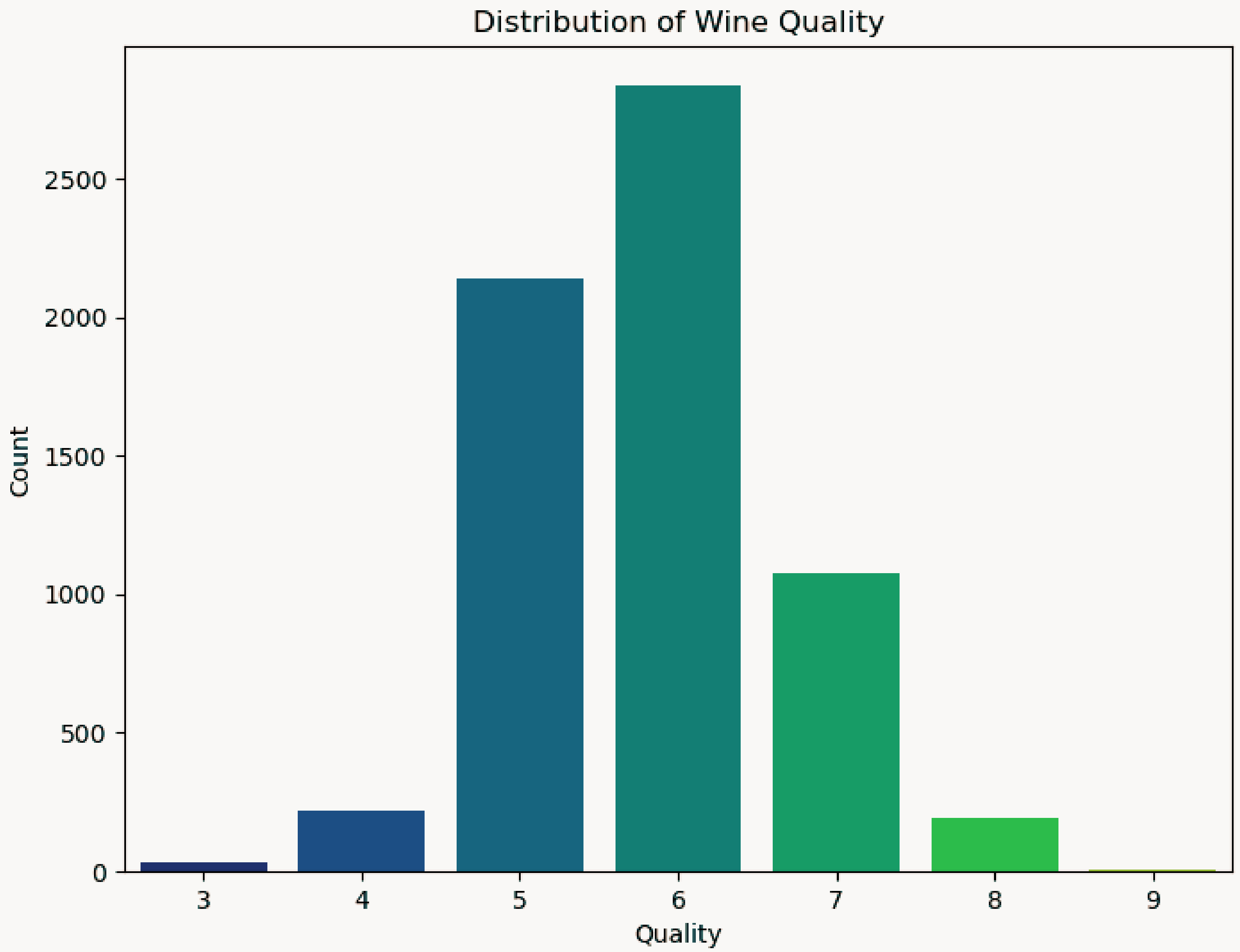
MACHINE LEARNING OPTIMIZATION

EXPLORATORY DATA ANALYSIS

After zipping both datasets for red and white wine, we began by looking into the distribution of our data and we found the following:

- Our dataset included more white wine than red wine instances
- The majority of our wines were of average quality

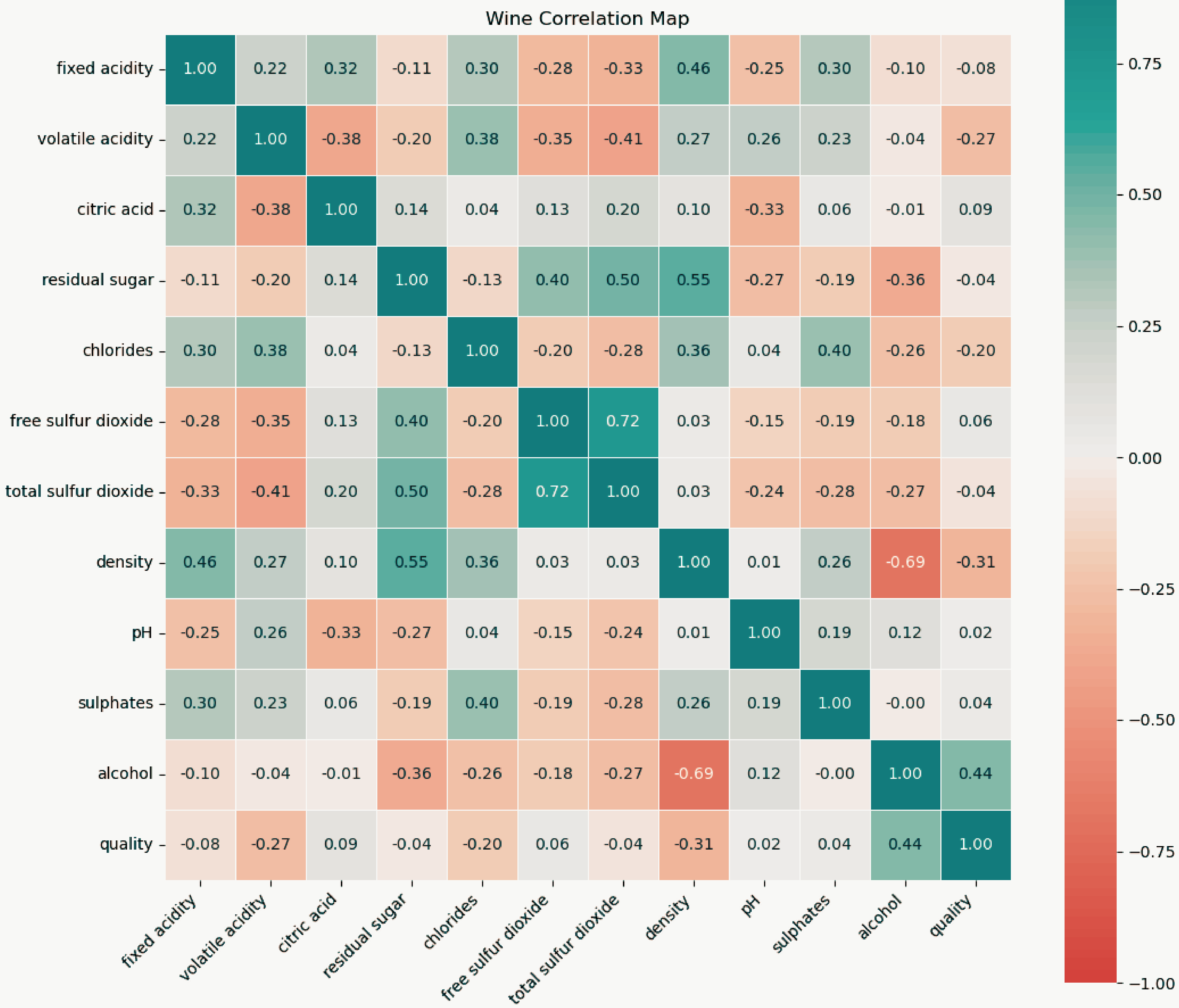




CORRELATION PLOT

In our correlation plot, we dig deeper into how each ingredient correlates with the other, and how all of this contributes to the quality of wine.

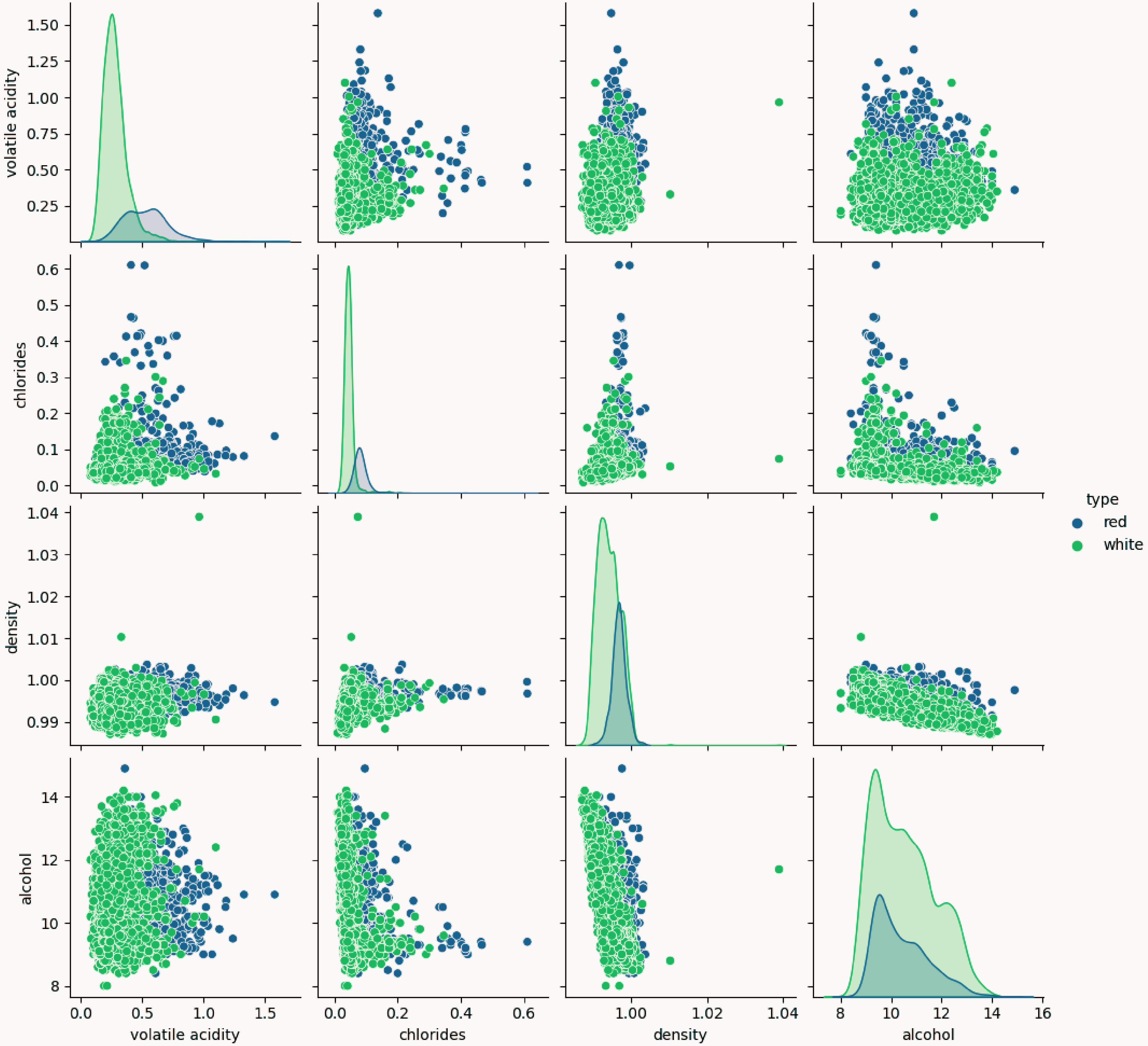
- Immediately we notice that volatile acidity, chlorides, density, and alcohol content have relatively extreme relationships with quality.
- Other components like sulfur dioxide (which are preservatives ensuring the longevity of wine) positively correlate with residual sugars, as this means that the sugar naturally occurring in grapes is preserved longer. However, this is irrelevant when it comes to wine quality.
- It's also observed with wine of any type, the more alcohol there is in wine, the less density it will have. And higher alcohol content means better quality.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	0.22	0.32	-0.11	0.30	-0.28	-0.33	0.46	-0.25	0.30	-0.10	-0.08
volatile acidity	0.22	1.00	-0.38	-0.20	0.38	-0.35	-0.41	0.27	0.26	0.23	-0.04	-0.27
citric acid	0.32	-0.38	1.00	0.14	0.04	0.13	0.20	0.10	-0.33	0.06	-0.01	0.09
residual sugar	-0.11	-0.20	0.14	1.00	-0.13	0.40	0.50	0.55	-0.27	-0.19	-0.36	-0.04
chlorides	0.30	0.38	0.04	-0.13	1.00	-0.20	-0.28	0.36	0.04	0.40	-0.26	-0.20
free sulfur dioxide	-0.28	-0.35	0.13	0.40	-0.20	1.00	0.72	0.03	-0.15	-0.19	-0.18	0.06
total sulfur dioxide	-0.33	-0.41	0.20	0.50	-0.28	0.72	1.00	0.03	-0.24	-0.28	-0.27	-0.04
density	0.46	0.27	0.10	0.55	0.36	0.03	0.03	1.00	0.01	0.26	-0.69	-0.31
pH	-0.25	0.26	-0.33	-0.27	0.04	-0.15	-0.24	0.01	1.00	0.19	0.12	0.02
sulphates	0.30	0.23	0.06	-0.19	0.40	-0.19	-0.28	0.26	0.19	1.00	-0.00	0.04
alcohol	-0.10	-0.04	-0.01	-0.36	-0.26	-0.18	-0.27	-0.69	0.12	-0.00	1.00	0.44
quality	-0.08	-0.27	0.09	-0.04	-0.20	0.06	-0.04	-0.31	0.02	0.04	0.44	1.00

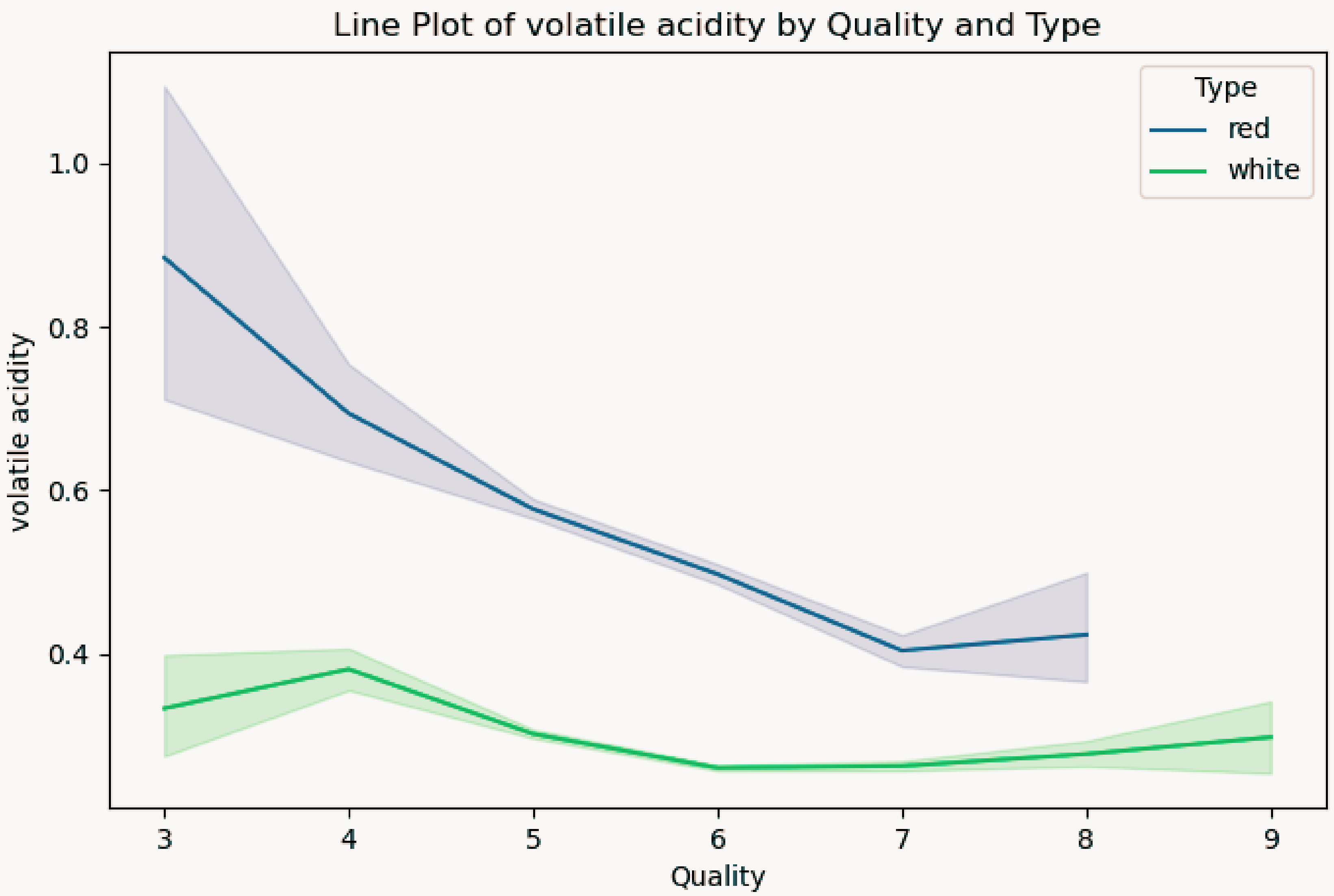
PAIR PLOT

- We performed a pair plot to compare each attribute that determines wine quality.
- The blue represents the red wine and the green represents the white wine.
- In our pair plots, we see obvious relationships between volatile acidity and alcohol.
- While alcohol is equally distributed for red and white types, red wine tends to have more acidic properties.

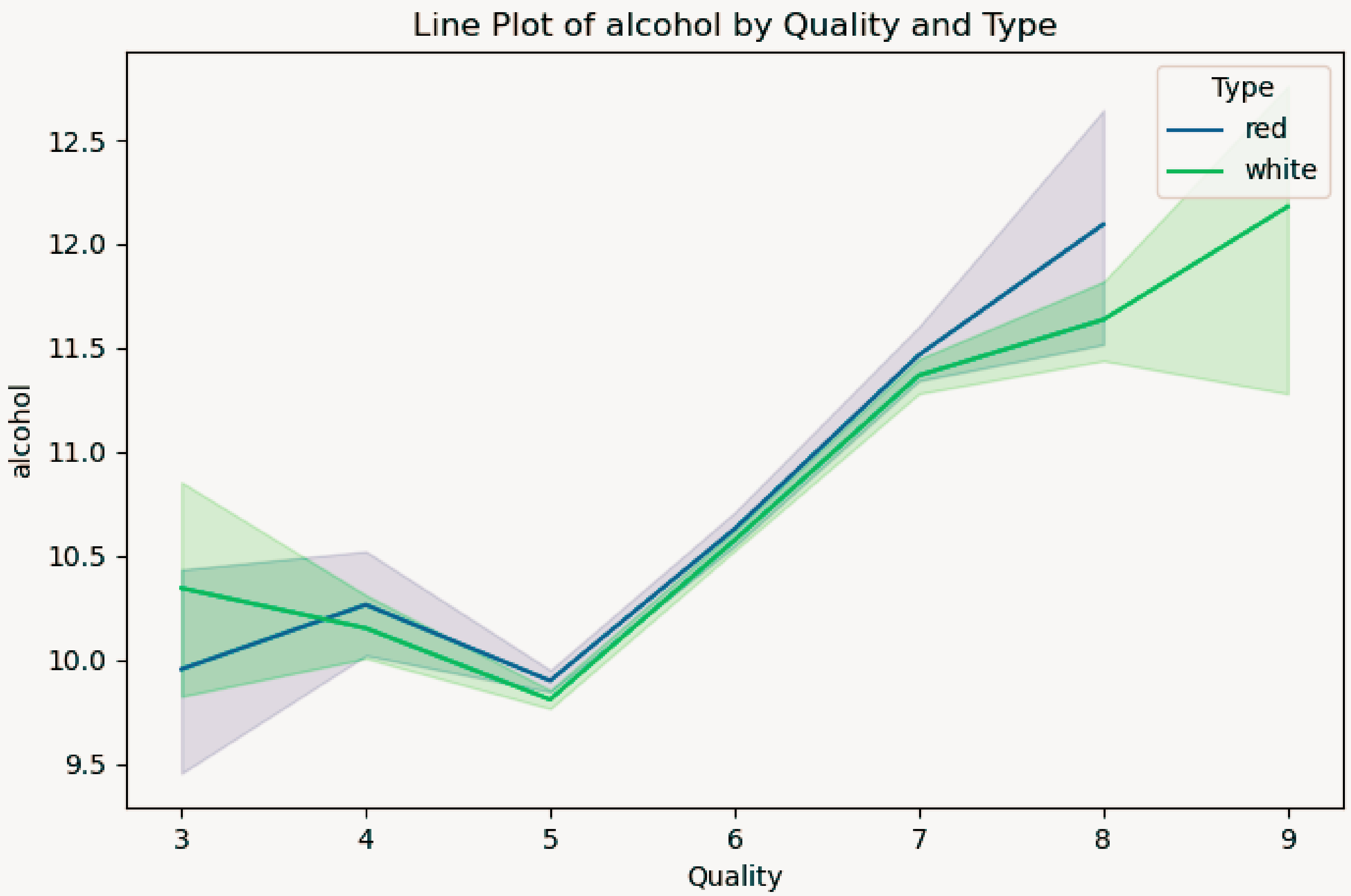


COMPARATIVE LINE PLOTS

- Volatile acidity has a relatively negative relationship with quality, and this is more prevalent with red wine. It does not seem to affect the quality of white wine.
- Both red and white wine have a positive relationship when we compare alcohol against quality.
  - It would seem that anything under 10.0% alcohol content would lower the quality score of both red and white wine.



Line Plot of volatile acidity by Quality and Type

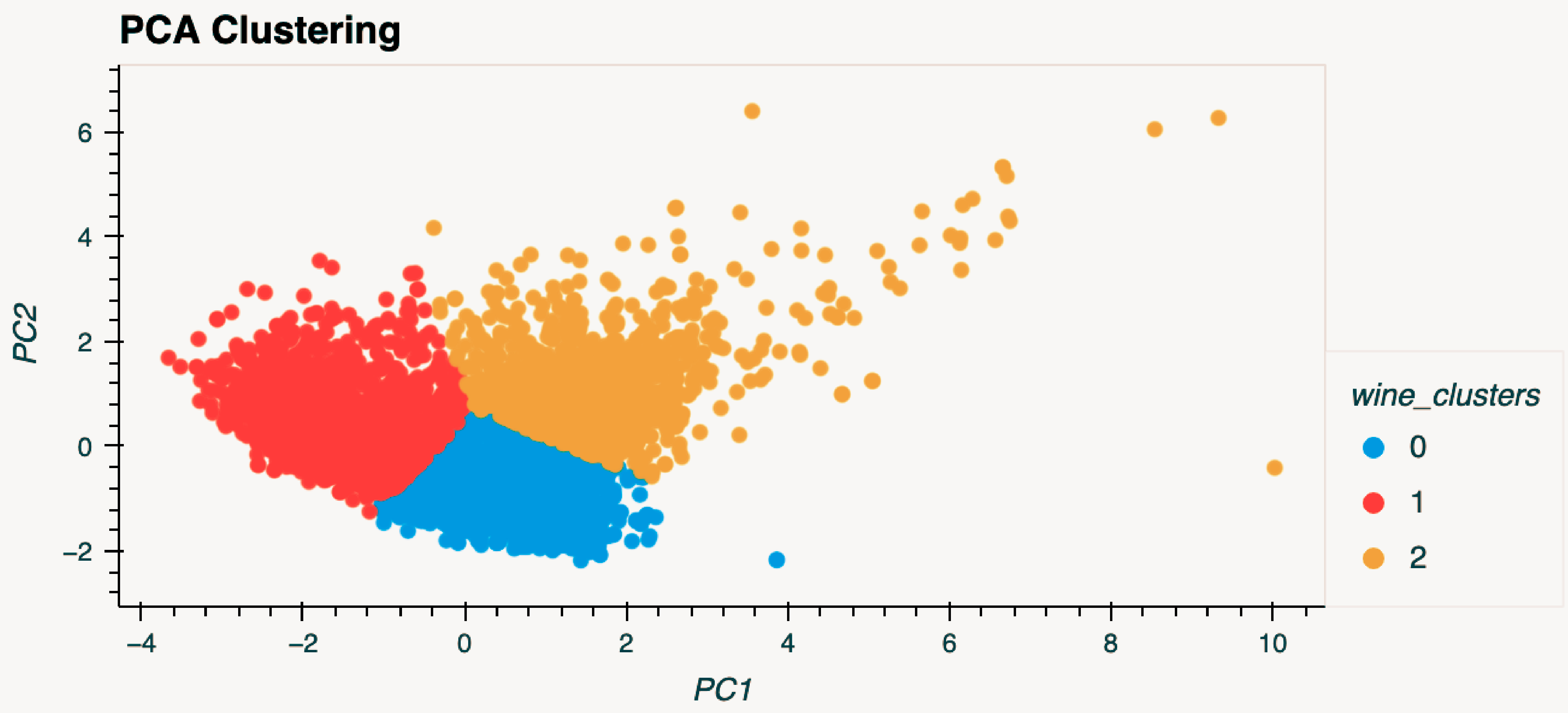


Line Plot of alcohol by Quality and Type

UNSUPERVISED LEARNING

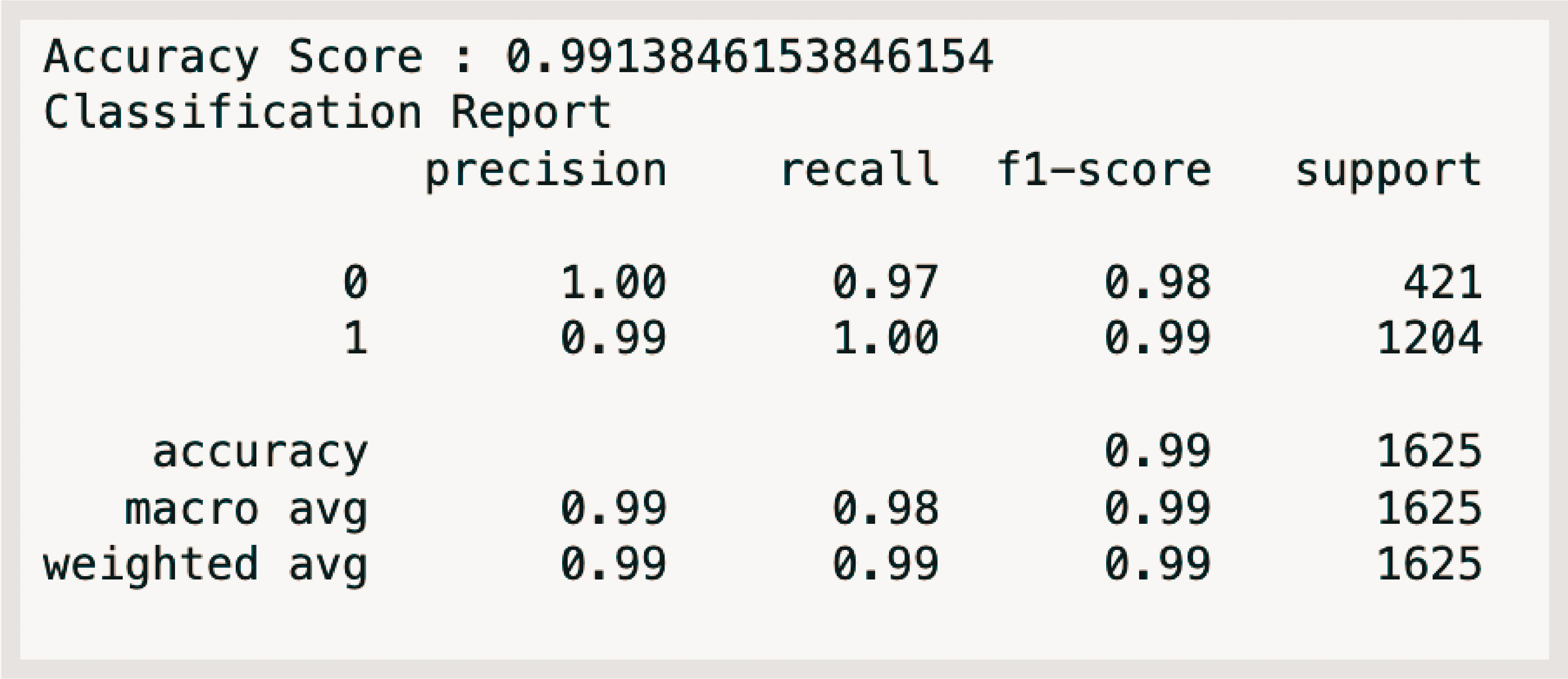
To begin with, we had 15 features that contributed to much noise in our dataset so based on our heat map and pair plots, we reduced our data frame to 4 features that heavily affect the quality of wine. Since our data was primarily a numerical dataset, we found it best to start by clustering our data given the various components in the wine. By doing so, we

- see clear distinctions between each cluster in a reduced dimensional space allowing us to identify unique groups of wine features,
- achieve a cumulative variance of 78% which suggests that the selected features or dimensions (principal components) are informative and collectively account for a large proportion of the data's patterns and distinctions.

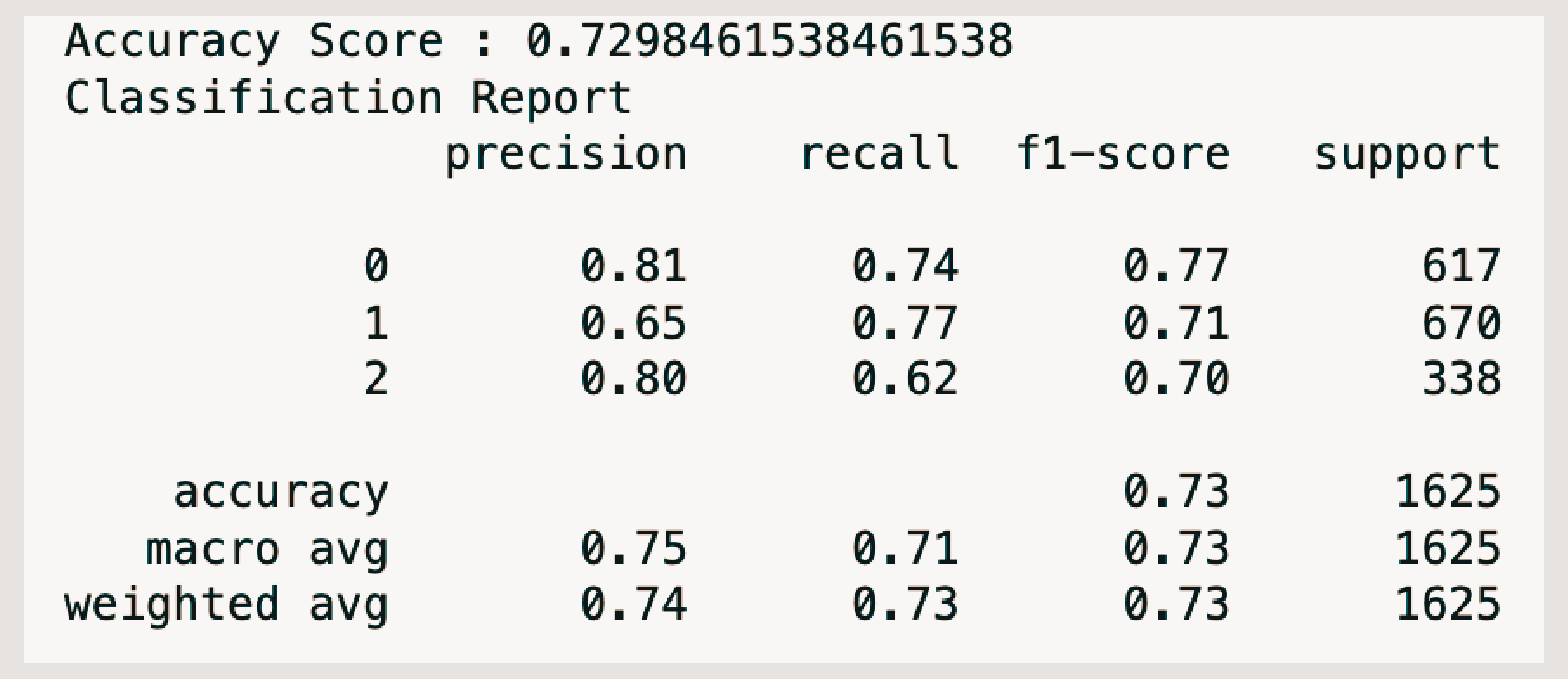


SUPERVISED LEARNING(WINE QUALITY)

- We began by using a Random Forest Classifier model using two classes to analyze the accuracy of predictions based on the type of wine. We found our model had very high accuracy. A source published by the University of California, Davis speaks extensively on the accuracy of the Random Forest model when training our dataset. This can confirm that there is no leakage in the dataset.
- Upon this, we decided to pivot into using the same model to predict the quality of wine instead.
  - We binned our quality into 3 classes (below average, average, and above average)
  - Ran predictions again to find that our model had a pretty low accuracy rate of 73%.



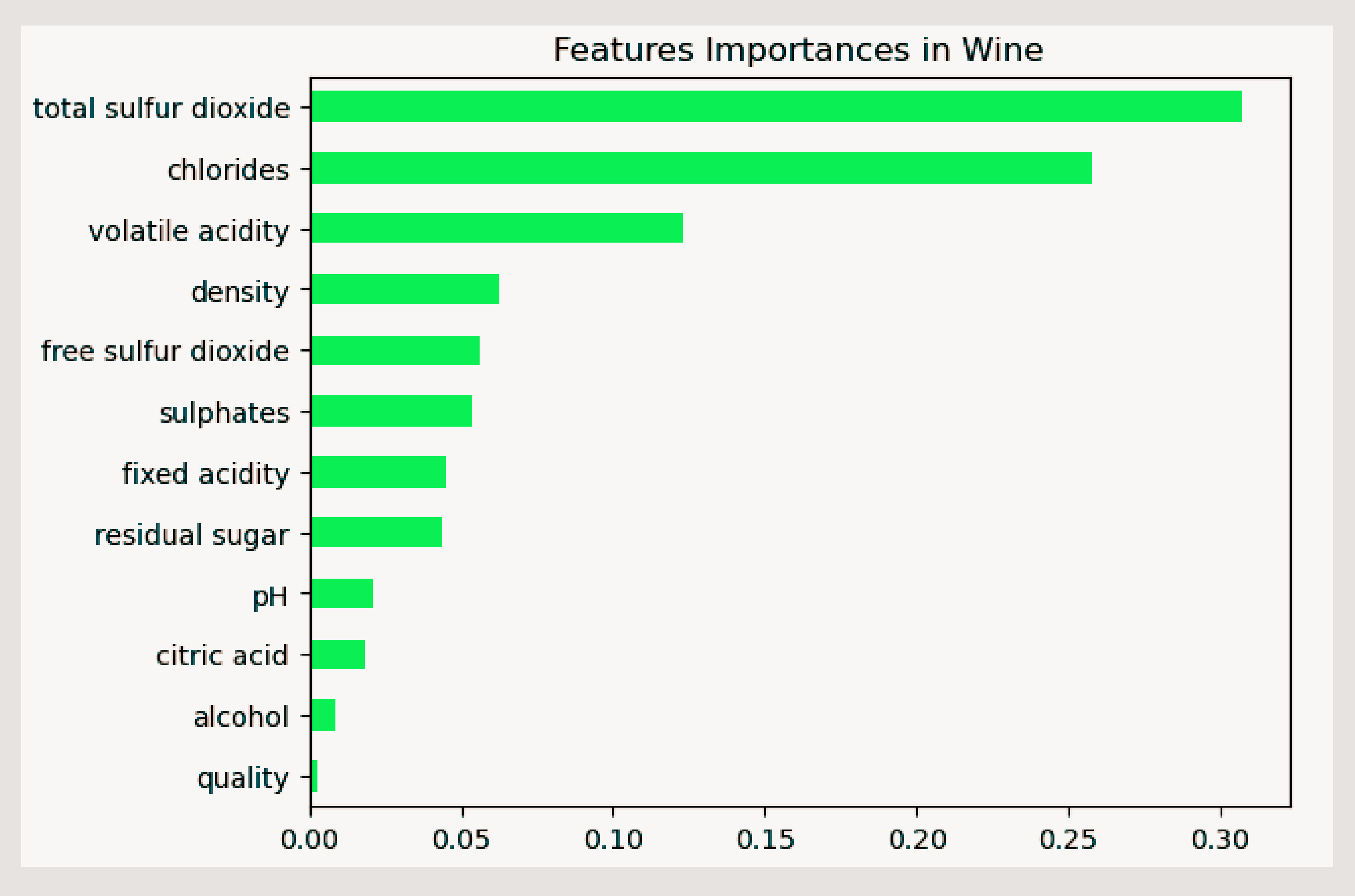
```
Accuracy Score : 0.9913846153846154
Classification Report
              precision    recall  f1-score   support

           0       1.00      0.97      0.98       421
           1       0.99      1.00      0.99      1204

    accuracy                           0.99      1625
   macro avg       0.99      0.98      0.99      1625
weighted avg       0.99      0.99      0.99      1625
```

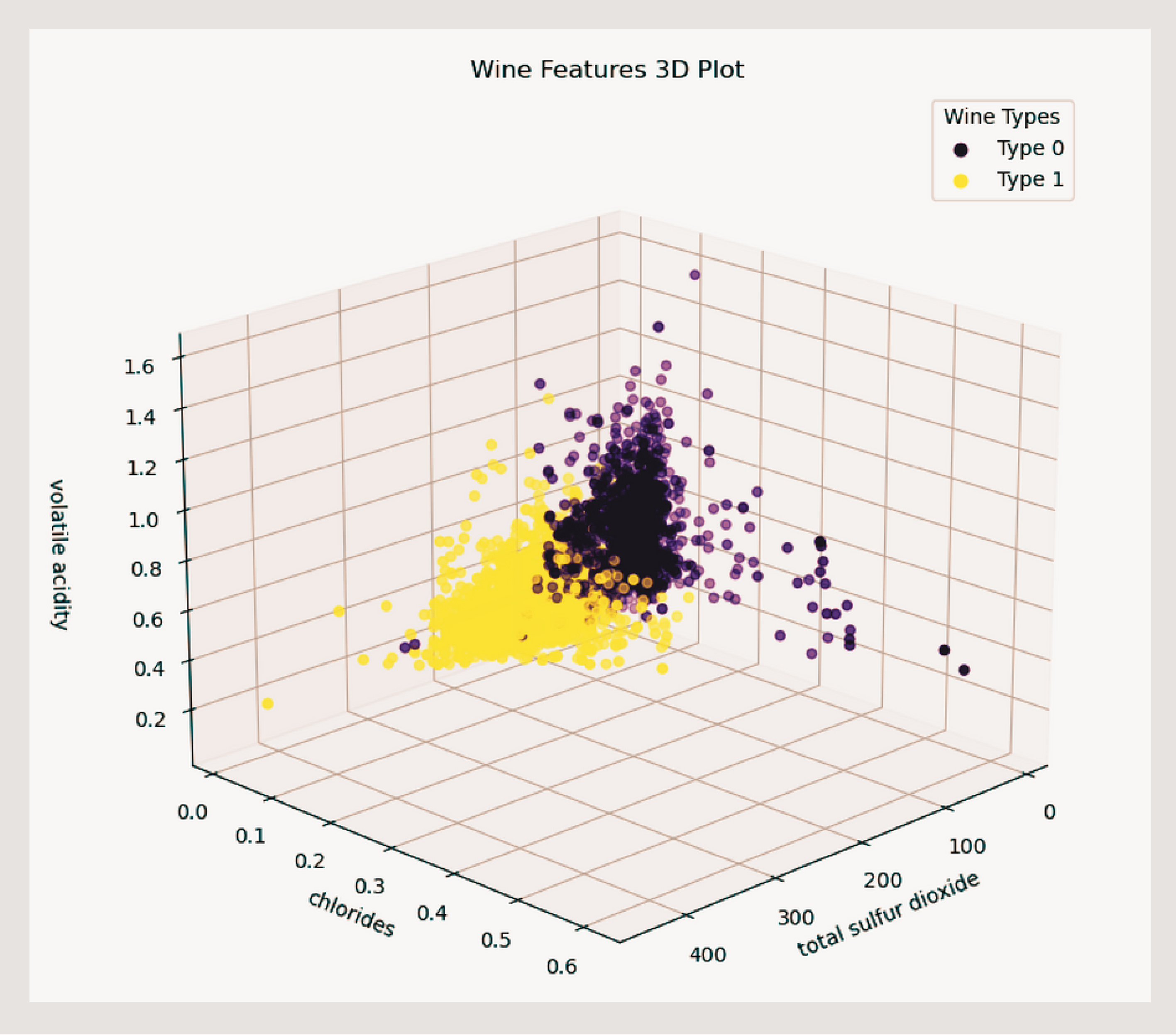


```
Accuracy Score : 0.7298461538461538
Classification Report
              precision    recall  f1-score   support

           0       0.81      0.74      0.77       617
           1       0.65      0.77      0.71       670
           2       0.80      0.62      0.70       338

    accuracy                           0.73      1625
   macro avg       0.75      0.71      0.73      1625
weighted avg       0.74      0.73      0.73      1625
```

FEATURE IMPORTANCE (WINE-TYPE)

- Our decision tree helped us understand which features heavily contributed to our wine-type predictions. After digging deeper, we notice that sulfur dioxide, chlorides, and volatile acidity play a major role in predicting the type of wine.
- We also plotted a 3D scatter plot to show how these features relate to each other depending on the wine type 0 for red and 1 for white wines.





DECISION TREE DEMO

NEURAL NETWORK OPTIMIZATION

	Attempt 1	OPTIMIZED MODEL Attempt 2	Attempt 3
Hidden Layers	3	4	3
Nodes/ Neurons	input - 20 1 - 10 2- 8	input - 100 1 - 60 2 - 30 3 - 15	input - 60 1 - 30 2 - 15
Activation	input - relu 1 - relu 2 - sigmoid	input - relu 1 - relu 2 - relu 3 - sigmoid	input - relu 1 - relu 2 - tanh
Accuracy Score	77% (Loss : 0.48)	80.6% (Loss: 1.7)	75% (Loss: 0.52)

- **Attempt 1**
  - Dropped the “type” column and utilized all features present in our original data set.
  - Began with 3 hidden layers and 1 input as a start.
- **Attempt 2 (OPTIMIZED MODEL)**
  - We didn’t change our data frame but instead increased our hidden layers and the amount of nodes in each layer.
- **Attempt 3**
  - We retained columns that were seen to be important features contributing to wine quality predictions in our Random Forest Classifier model.
  - We also reduced our hidden layers once more and used a “tanh” activation function.

CONCLUSION

- Our Neural Network models and optimization yielded higher accuracy when predicting quality, specifically, attempt 2 with an accuracy rating of ~81%
- PCA clustering results in distinct clusters and a healthy variance ratio
- Strong relationships were identified in our correlation maps and pair plots.
    - For red wine
        - strong positive correlation with quality - alcohol content, sulfates, citric acid.
        - strong negative correlation with quality - volatile acidity, total sulfur dioxide.
    - For white wine
        - strong positive correlation with quality - alcohol
        - strong negative correlation with quality - density, volatile acidity, and chlorides.

FUTURE WORK

- We could expand the data set from just wine to different types of alcohol such as beer, ciders, and distilled drinks and can use this expanded data set to better train our model.
- It would also be interesting to see if we could have found other data sets with more features to see if that would have changed our results.
- With more time, it would have been nice to be able to create a dashboard with an interface displaying our findings and having users able to see the wine predictions from our model.
- We can also try retaining 3 bins of quality categories when training our neural network model.