INFO 6105 Final Project

Shaofan Wei

002815198

1. Introduction

The quality assessment of wine has always been a significant topic in wine research and production. The quality of wine is influenced not only by the grape variety, cultivation environment, and winemaking process but also closely related to various physical and chemical properties. These properties include but are not limited to acidity, sugar content, alcohol content, and the presence of various compounds. Understanding how these factors affect the sensory evaluation and quality rating of wine is key to improving wine production and tasting.

This study aims to explore the relationship between the chemical components of red and white wines and their quality ratings. By analyzing the wine quality dataset from the UCI Machine Learning Repository[1], this study focuses on identifying those physical and chemical characteristics that significantly impact wine quality assessment. Additionally, this research also aims to compare whether there are differences in the quality evaluation standards between red and white wines. Through an in-depth analysis of the physical and chemical characteristics of wine, this study hopes to provide valuable insights for wine producers, consumers, and researchers.

2. Method

The dataset used in this study includes samples of red and white wines from the UCI Machine Learning Repository. These datasets detailed a variety of physical and chemical characteristics of wine, such as fixed acidity, volatile acidity, citric acid, residual sugar, and quality ratings based on sensory tests. The sample size was 1,599 for red wine and 4,898 for white wine.

In the preliminary stage of data processing, special attention was paid to the treatment of outliers in characteristics such as residual sugar, chlorides, and free sulfur dioxide. After careful consideration, we chose to retain these outliers because they, although deviating from the norm, might represent natural variations rather than erroneous data. This retention strategy helps ensure the completeness and authenticity of the analysis.

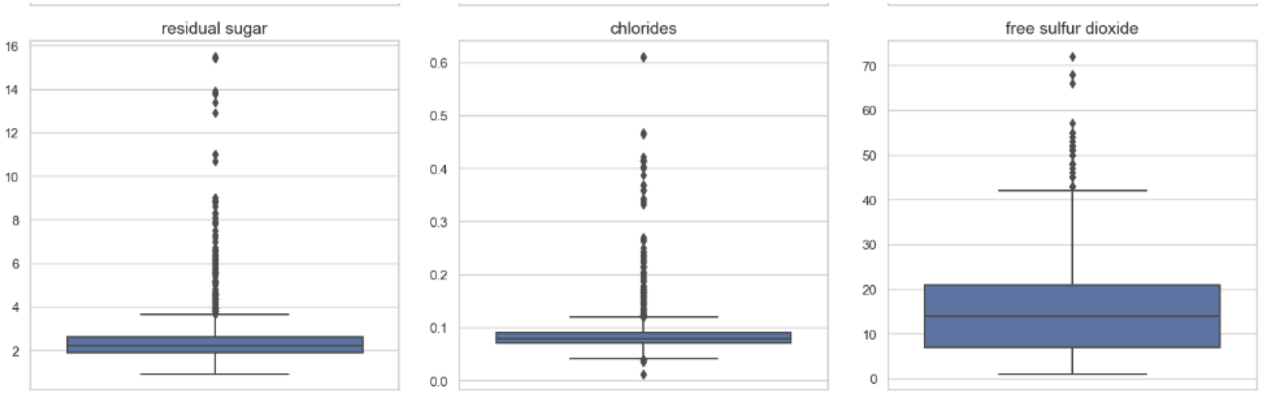
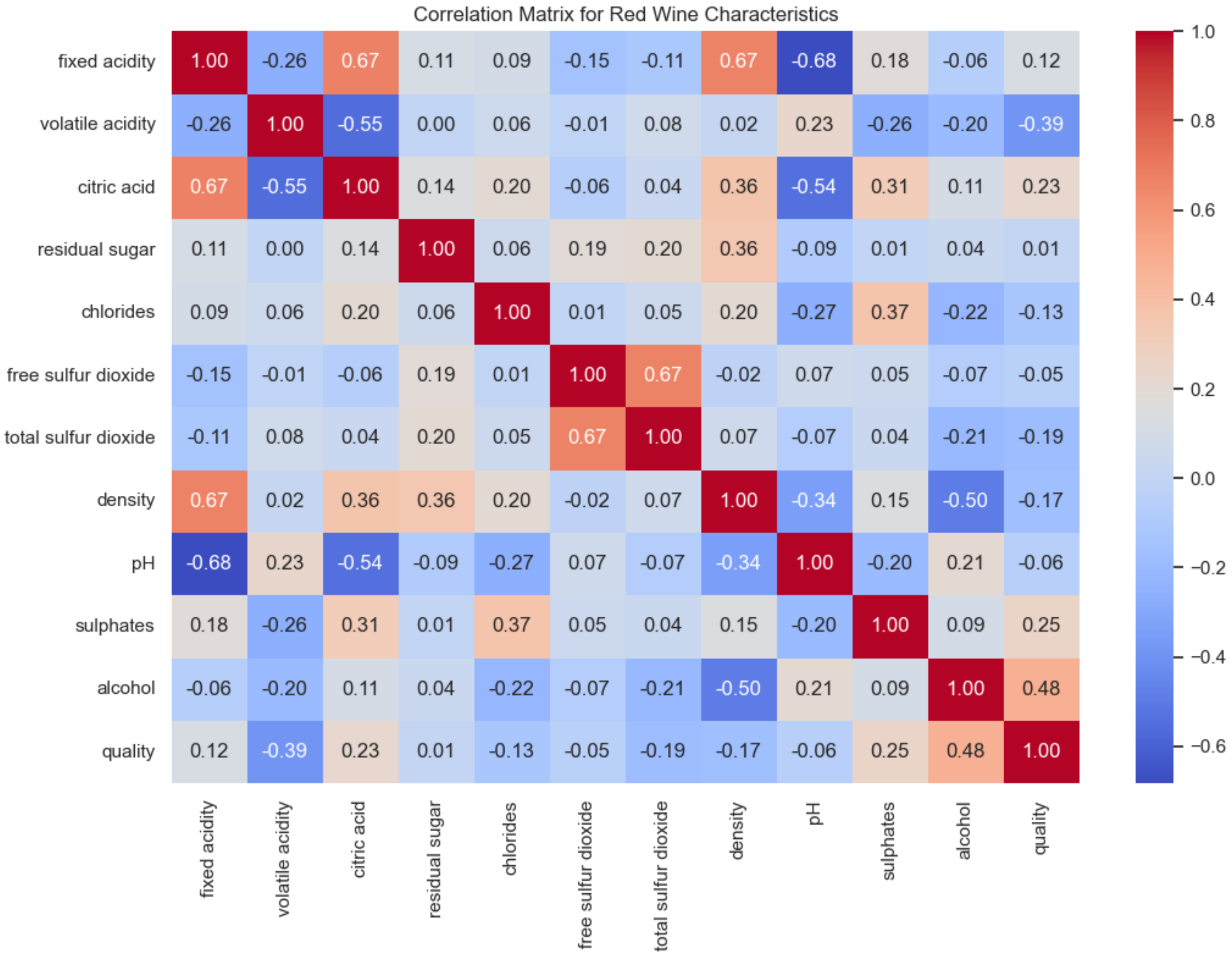


Figure 1, Outliers in the Box Plot

Then, feature selection was conducted by exploratory data analysis to determine the features most relevant to wine quality rating. To reduce the dimensionality of the data and identify the most important features, PCA was performed on the datasets. Four principal components were chosen for the red wine dataset, and five for the white wine dataset. This decision was based on the variance contribution rate of each principal component to ensure that most information was retained. Regression models were established to analyze the impact of different chemical characteristics on wine quality rating. The performance of the regression models was evaluated using indicators such as Mean Squared Error and Coefficient of Determination.

3. Results

Through detailed visualization analysis using heatmaps, we observed some apparent trends. For example, a significant positive correlation was found between alcohol content and wine quality rating, suggesting that wines with higher alcohol content generally have higher quality ratings. Conversely, a negative correlation was found between density and quality rating, indicating that heavier wines might be of lower quality. This visualization method provides an intuitive way to understand how different characteristics interact and how they affect the overall quality of the wine.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12
volatile acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.20	-0.39
citric acid	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	-0.09	0.01	0.04	0.01
chlorides	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	-0.02	0.07	0.05	-0.07	-0.05
total sulfur dioxide	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	-0.07	0.04	-0.21	-0.19
density	0.67	0.02	0.36	0.36	0.20	-0.02	0.07	1.00	-0.34	0.15	-0.50	-0.17
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1.00	-0.20	0.21	-0.06
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.20	1.00	0.09	0.25
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00	0.48
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05	-0.19	-0.17	-0.06	0.25	0.48	1.00

Figure 2, Heatmap of Red Wine

In the red wine dataset, we analyzed four principal components. These principal components captured most of the variance in the data (about 70%), representing the key information in the dataset.

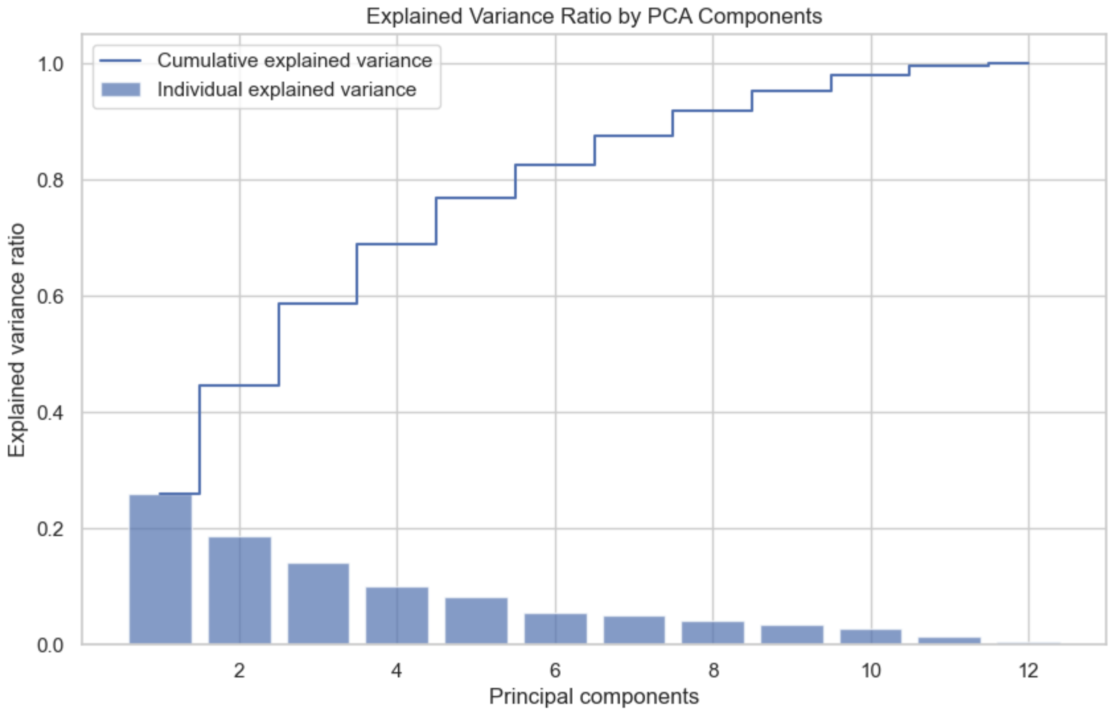


Figure 3, Explained Variance Ratio of Red Wine

In the white wine dataset, chose five principal components, which also captured most of the variance (about 70%).

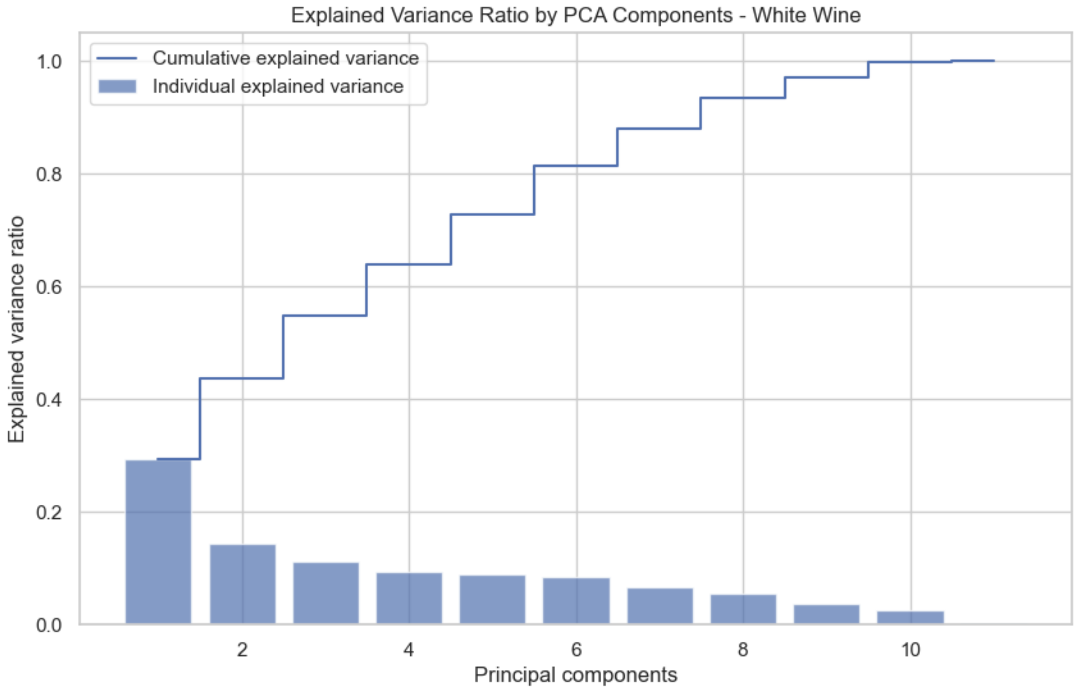
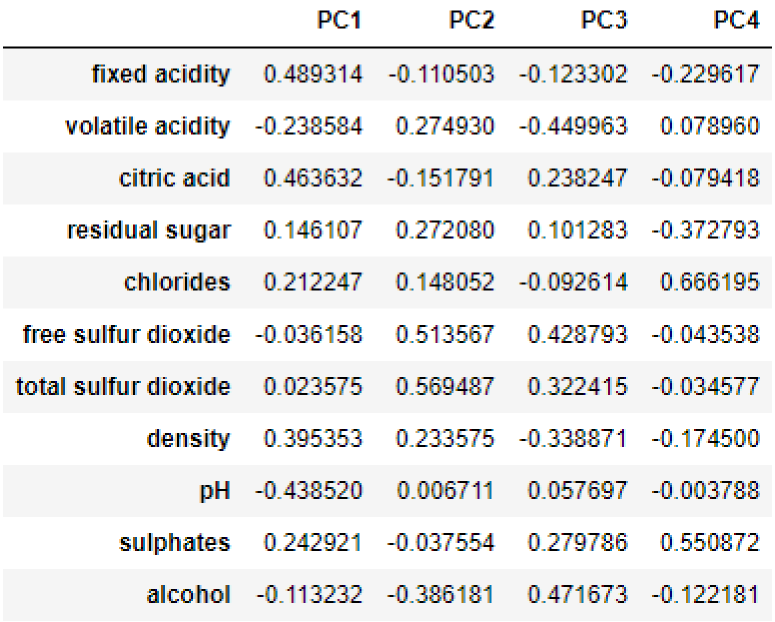


Figure 4, Explained Variance Ratio of White Wine

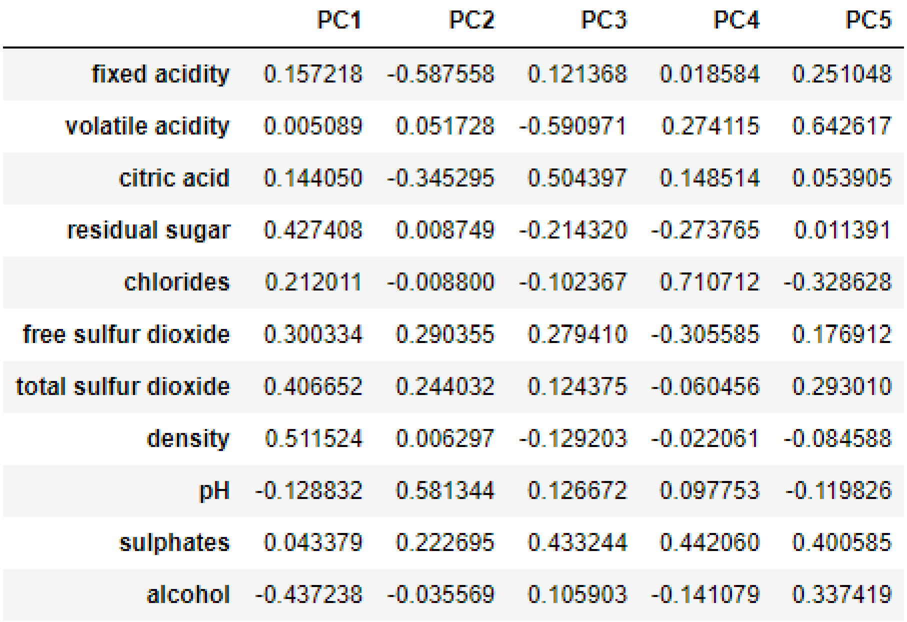
In the red wine dataset, the four principal components were influenced by several original data characteristics. For instance, one principal component might be mainly influenced by alcohol content and total acidity. This indicates that these characteristics play an important role in wine quality rating.



	PC1	PC2	PC3	PC4
fixed acidity	0.489314	-0.110503	-0.123302	-0.229617
volatile acidity	-0.238584	0.274930	-0.449963	0.078960
citric acid	0.463632	-0.151791	0.238247	-0.079418
residual sugar	0.146107	0.272080	0.101283	-0.372793
chlorides	0.212247	0.148052	-0.092614	0.666195
free sulfur dioxide	-0.036158	0.513567	0.428793	-0.043538
total sulfur dioxide	0.023575	0.569487	0.322415	-0.034577
density	0.395353	0.233575	-0.338871	-0.174500
pH	-0.438520	0.006711	0.057697	-0.003788
sulphates	0.242921	-0.037554	0.279786	0.550872
alcohol	-0.113232	-0.386181	0.471673	-0.122181

Figure 5, Correlation Matrix of Red Wine

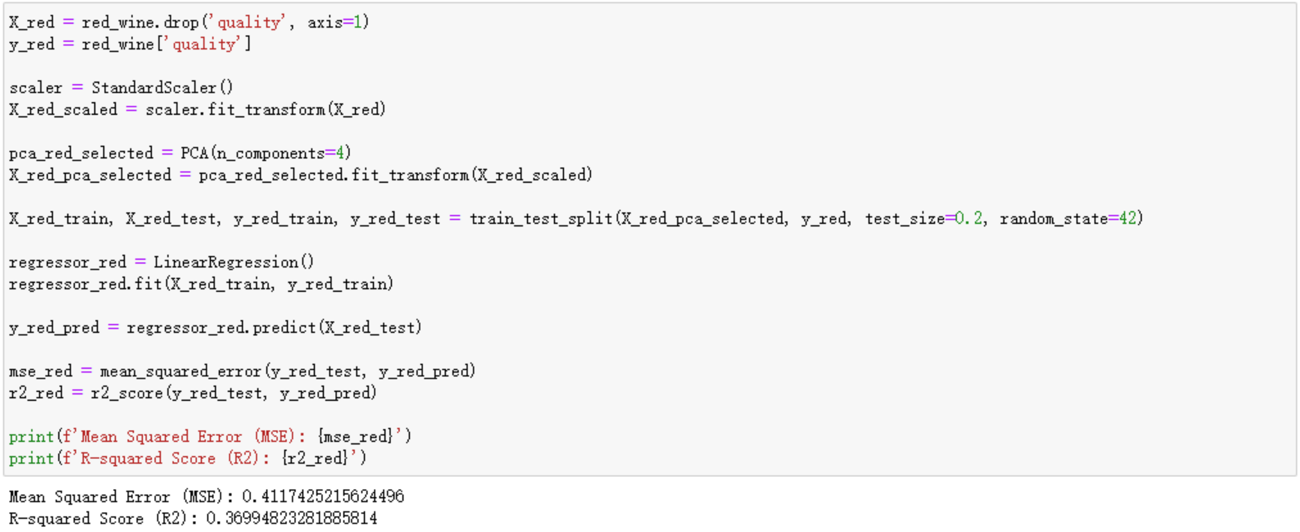
For white wine, the five principal components were also influenced by a range of characteristics, such as volatile acidity and citric acid content, which together define the unique quality of white wine.



	PC1	PC2	PC3	PC4	PC5
fixed acidity	0.157218	-0.587558	0.121368	0.018584	0.251048
volatile acidity	0.005089	0.051728	-0.590971	0.274115	0.642617
citric acid	0.144050	-0.345295	0.504397	0.148514	0.053905
residual sugar	0.427408	0.008749	-0.214320	-0.273765	0.011391
chlorides	0.212011	-0.008800	-0.102367	0.710712	-0.328628
free sulfur dioxide	0.300334	0.290355	0.279410	-0.305585	0.176912
total sulfur dioxide	0.406652	0.244032	0.124375	-0.060456	0.293010
density	0.511524	0.006297	-0.129203	-0.022061	-0.084588
pH	-0.128832	0.581344	0.126672	0.097753	-0.119826
sulphates	0.043379	0.222695	0.433244	0.442060	0.400585
alcohol	-0.437238	-0.035569	0.105903	-0.141079	0.337419

Figure 6, Correlation Matrix of White Wine

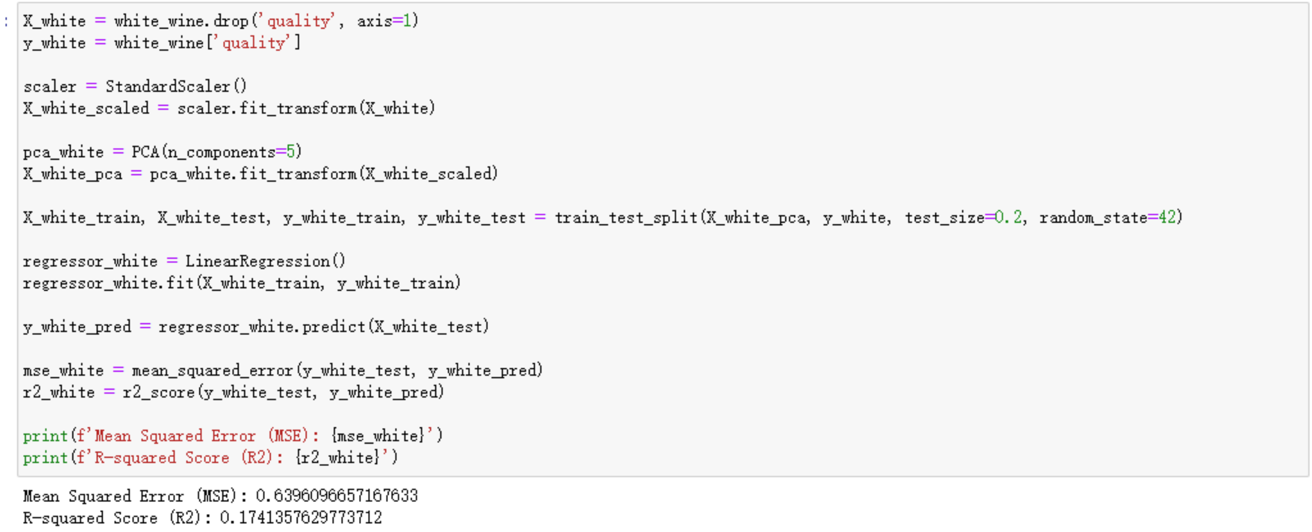
In terms of performance evaluation, linear regression models were constructed for both red and white wine datasets. For the red wine model, the Mean Squared Error was 0.41, and the Coefficient of Determination was 0.37, indicating that the model could predict wine quality to a certain extent. In contrast, the white wine model had a Mean Squared Error of 0.64 and a Coefficient of Determination of 0.17, showing that the model's accuracy in predicting the quality of white wine was relatively low and required further optimization and adjustment.



```
X_red = red_wine.drop('quality', axis=1)
y_red = red_wine['quality']

scaler = StandardScaler()
X_red_scaled = scaler.fit_transform(X_red)

pca_red_selected = PCA(n_components=4)
X_red_pca_selected = pca_red_selected.fit_transform(X_red_scaled)

X_red_train, X_red_test, y_red_train, y_red_test = train_test_split(X_red_pca_selected, y_red, test_size=0.2, random_state=42)

regressor_red = LinearRegression()
regressor_red.fit(X_red_train, y_red_train)

y_red_pred = regressor_red.predict(X_red_test)

mse_red = mean_squared_error(y_red_test, y_red_pred)
r2_red = r2_score(y_red_test, y_red_pred)

print(f'Mean Squared Error (MSE): {mse_red}')
print(f'R-squared Score (R2): {r2_red}')
```

```
Mean Squared Error (MSE): 0.4117425215624496
R-squared Score (R2): 0.36994823281885814
```

Figure 7, MSE and R2 Values for the Linear Regression Model of Red Wine



```
X_white = white_wine.drop('quality', axis=1)
y_white = white_wine['quality']

scaler = StandardScaler()
X_white_scaled = scaler.fit_transform(X_white)

pca_white = PCA(n_components=5)
X_white_pca = pca_white.fit_transform(X_white_scaled)

X_white_train, X_white_test, y_white_train, y_white_test = train_test_split(X_white_pca, y_white, test_size=0.2, random_state=42)

regressor_white = LinearRegression()
regressor_white.fit(X_white_train, y_white_train)

y_white_pred = regressor_white.predict(X_white_test)

mse_white = mean_squared_error(y_white_test, y_white_pred)
r2_white = r2_score(y_white_test, y_white_pred)

print(f'Mean Squared Error (MSE): {mse_white}')
print(f'R-squared Score (R2): {r2_white}')
```

```
Mean Squared Error (MSE): 0.6396096657167633
R-squared Score (R2): 0.1741357629773712
```

Figure 8, MSE and R2 Values for the Linear Regression Model of White Wine

These results showcase the main findings in analyzing the red and white wine datasets. Through further analysis and model optimization, a more accurate understanding of the key factors affecting wine quality can be achieved.

4. Discussion

The purpose of this study is to explore the key chemical characteristics affecting the quality of red and white wines and to analyze these characteristics through principal component analysis and linear regression models.

Heatmap analysis showed that certain chemical characteristics, such as alcohol content, positively correlate with wine quality, while others, like density, show a negative correlation. This difference might reflect the chemical changes in different types of wines during the production process. PCA revealed the most important variables in the datasets. In the red wine dataset, four principal components effectively represented the key aspects of the data. However, in the white wine dataset, due to its complexity, five principal components were chosen, yet they still did not adequately analyze high-quality white wines. This suggests that the quality of white wine might be influenced by a broader range of variables. Although the linear regression models could predict wine quality to some extent, the performance of the models indicates room for improvement. Especially in the white wine dataset, the model's accuracy was relatively low, possibly due to complex non-linear relationships among variables in the dataset.

Future research could consider using more complex models, such as neural networks, to capture the non-linear relationships in the data. Neural networks are particularly suitable for handling complex patterns that linear models find difficult to capture, as they can learn complex mappings from input to output through their multi-layer structure and non-linear activation functions.

5. Conclusion

This study provides a deep understanding of the factors affecting wine quality by analyzing the relationship between the chemical components of red and white wines and their quality ratings. Using the wine quality dataset from the UCI Machine Learning Repository, we demonstrated how to identify and analyze the key physical and chemical characteristics affecting wine quality through data processing, feature selection, PCA analysis, and linear regression models.

Heatmap analysis revealed the correlations between different chemical characteristics and wine quality, where some characteristics like alcohol content positively correlate with wine quality, while others like density show a negative correlation. PCA further revealed the key variables in the red and white wine datasets. The red wine dataset showed a better data representation with four principal components, while the white wine required five principal components to capture its more complex characteristic patterns.

The linear regression model showed better predictive ability in the red wine dataset but performed less satisfactorily in the white wine dataset, pointing out the model's limitations in handling complex non-linear relationships. This suggests that future research needs to consider using more advanced models, such as neural networks, to better capture and interpret the non-linear relationships in the data. Neural networks, with their multi-layered structure and complex functions, can reveal hidden patterns and relationships in the data, providing more accurate tools for predicting and improving wine quality.

6. Reference

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” in Decision Support Systems, Elsevier, vol. 47, no. 4, pp. 547–553, 2009, ISSN: 0167-9236. [online] http://dx.doi.org/10.1016/j.dss.2009.05.016