Dang Thien Nguyen

Dxn2100021

CS 4372.501 – Computational Methods for Data Scientists

RED WINE QUALITY DATA ANALYSIS



1. Background

This case study examines a dataset of red wines from Portugal. The goal is to predict the quality rating of a wine (integer scores of 38) from its chemical properties. The ability to predict quality from objective factors like acidity and alcohol content is helpful for winemakers, sommeliers, and consumers.

The dataset contains 1599 red wines with 11 measured attributes, including alcohol content, pH, sulfates, volatile acidity, and density. There is also a categorical “wine type” and expert quality rating.

# Load the dataset.

The wine quality dataset was sourced from the UCI Machine Learning Repository. It contains physicochemical properties and quality ratings for 1599 red wines from the Vinho Verde region in Portugal. In this project, we will focus on Red Wine only.

The data was available as a CSV file with the following attributes:

* Fixed acidity
* Volatile acidity
* Citric acid
* Residual sugar
* Chlorides
* Free sulfur dioxide
* Total sulfur dioxide
* Density
* pH
* Sulphates
* Alcohol
* Quality (score between 0 and 10)



# Exploratory Data Analysis

## 3.1 Checking Null and Missing Values

A screen shot of a computer

Description automatically generated The data consisted of 1599 rows and 12 columns, representing the number of wine samples and features, respectively.

A screen shot of a computer

Description automatically generatedThe red wine data was checked for null values using df.isnull().sum(). This identified no null values.

The red wine data was checked for null values using df.isnull().sum(). This identified no null A computer screen with white text

Description automatically generatedvalues.

And it appeared that the data also didn’t have any null value or any missing data.

The feature data types were deemed appropriate a floating point for the chemical properties and an integer for the quality.

## 

## 3.2 Let's do some Statistics

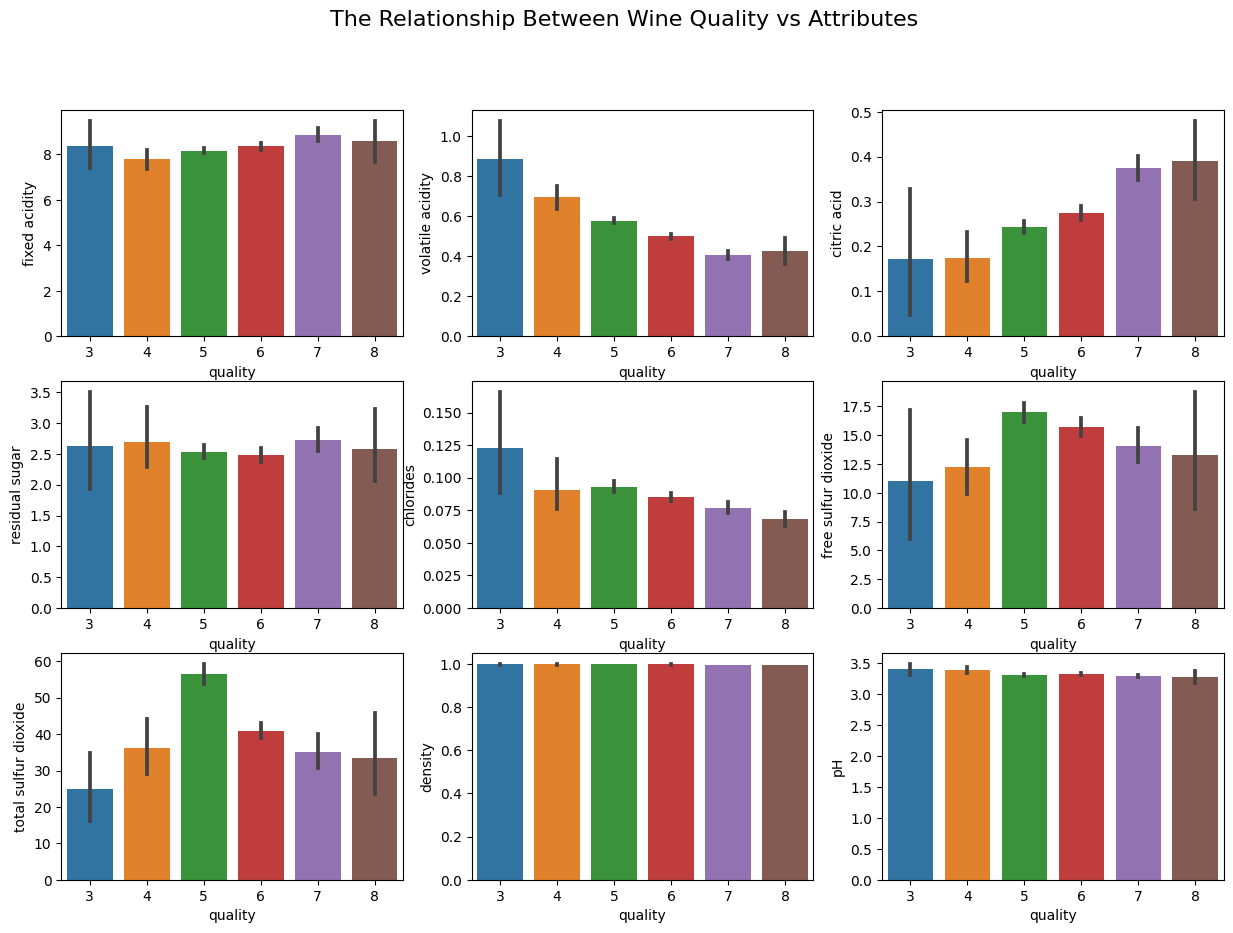
A screenshot of a computer screen

Description automatically generated

* The average value of fixed acidity is 8.31, the highest value is 15.9, and the lowest value is 8.32
* The average value of volatile acidity is 0.53, the highest value is 1.58, and the lowest value is 0.12
* The average value of citric acid is 0.27, the highest value is 1, and the lowest value is 0
* The average value of residual sugar is 2.54, the highest value is 15.5, and the lowest value is 0.9
* The average value of chlorides is 0.08, the highest value is 0.61, and the lowest value is 0.01
* The average value of free sulfur dioxide is 15.87, the highest value is 72, and the lowest value is 1
* The average value of total sulfur dioxide is 46.46, the highest value is 289, and the lowest value is 6
* The average value of density is 0.99, the highest value is 1, and the lowest value is 0.99
* The average value of pH is 3.31, the highest value is 4.01, and the lowest value is 2.74
* The average value of sulphates is 0.66, the highest value is 2, and the lowest value is 0.33
* The average value of alcohol is 10.42, the highest value is 14.9, and the lowest value is 8.4
* The average value of quality is 5.64, the highest value is 8, and the lowest value is 3

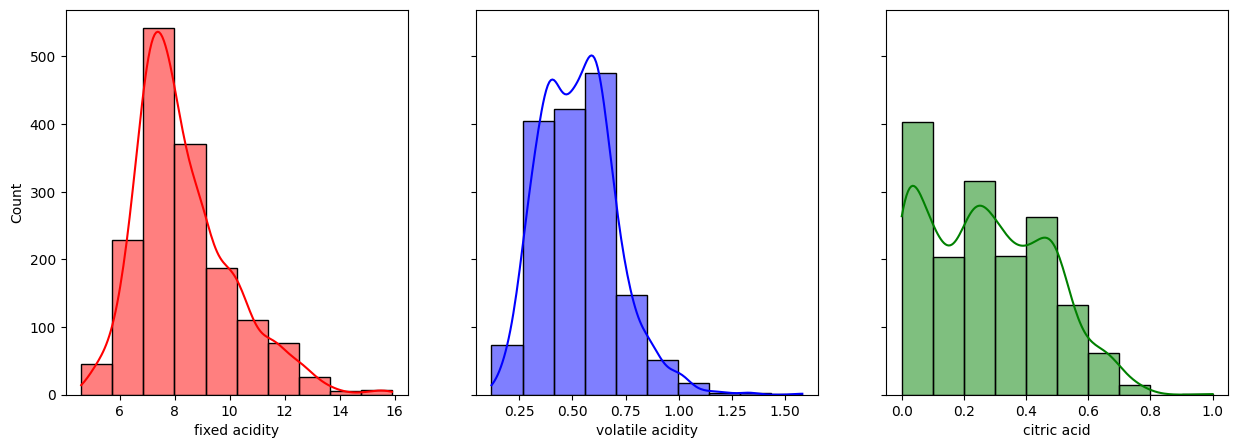
## 3.3 Data Visualization

### 3.3.1 Distribution

Here is an analysis of the bar plots showing the relationship between different wine quality attributes and the target variable quality:

* Fixed Acidity: No clear trend, acidity levels similar across quality ratings
* Volatile Acidity: Clear downward trend, higher quality wines have lower volatile acidity
* Citric Acid: No clear trend, some fluctuation but no correlation
* Residual Sugar: No clear trend, residual sugar fairly constant across qualities
* Chlorides: No clear trend, chlorides level unaffected by quality
* Free Sulfur Dioxide: Higher quality wines tend to have slightly higher free SO2
* Total Sulfur Dioxide: Higher quality wines tend to have noticeably higher total SO2
* Density: Slight upward trend, with higher densities at higher qualities
* pH: No clear trend, pH level does not seem correlated with quality

### Hisplot

The histograms offer more clarity on the distribution and attributes of the chemical characteristics.

* Fixed Acidity:
  + Unimodal distribution
  + Approximately normal / Gaussian shape but slightly left-skewed.
  + Peak frequency values clustered around 8-9
  + Most wines have fixed acidity values in the range of 6-11
* Volatile Acidity:
  + Bimodal distribution with two peaks
  + First peak around 0.2-0.4 (lower volatile acidity)
  + Second peak around 0.6-0.8 (higher volatile acidity)
  + Two clusterings indicate two common levels of volatile acidity.
* Citric Acid:
  + Bimodal distribution
  + First peak around 0.1-0.2 (lower citric acid)
  + Second peak around 0.4-0.6 (higher citric acid)
  + Two peaks indicate two common levels of citric acid content.

A graph of a graph of a graph

Description automatically generated with medium confidence

* Residual Sugar:
  + Unimodal distribution, skewed right with a long tail.
  + Most values clustered on left side between 0-10, with decreasing frequency towards the right.
  + Indicates most wines have low-moderate residual sugar content.
* Chlorides:
  + Unimodal distribution, approximately normal / Gaussian shape
  + Peak frequency values centered around 0.05-0.1
  + The symmetric bell curve indicates most values clustered around the mean.
* Free Sulfur Dioxide:
  + Bimodal distribution with two peaks
  + First peak between 0-50 (lower sulfur dioxide)
  + Second peak around 125-150 (higher sulfur dioxide)
  + Bimodality indicates two common levels of sulfur dioxide content.

A graph of a normal distribution

Description automatically generated

* Total Sulfur Dioxide:
  + Unimodal, right-skewed distribution
  + Most values clustered between 50-200 ppm.
  + Long tail indicates some wines have much higher total SO2.
* Density:
  + Unimodal, approximately normal distribution
  + Symmetric bell curve shape clustered around 0.99-1.0 g/cm^3.
  + Indicates most wines have similar density values.
* pH:
  + Unimodal distribution, slightly left-skewed.
  + Most values centered around 3.2-3.4
  + Wider spread indicates broader range of pH values.

### Colleration

After using Pair plot, We notice that there is correlation between some variables, and this is called multicollinearity

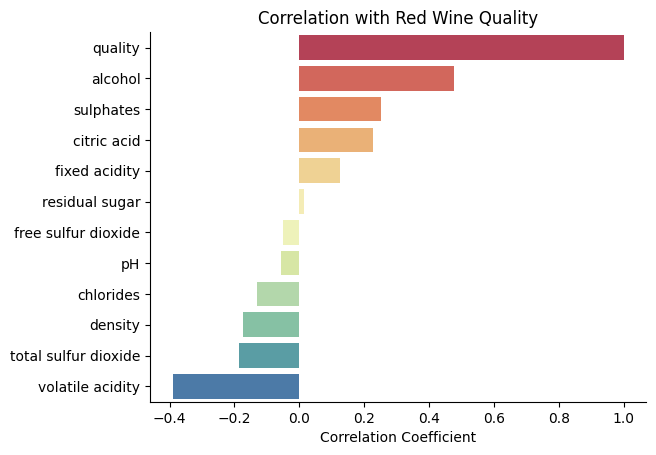
A screenshot of a graph

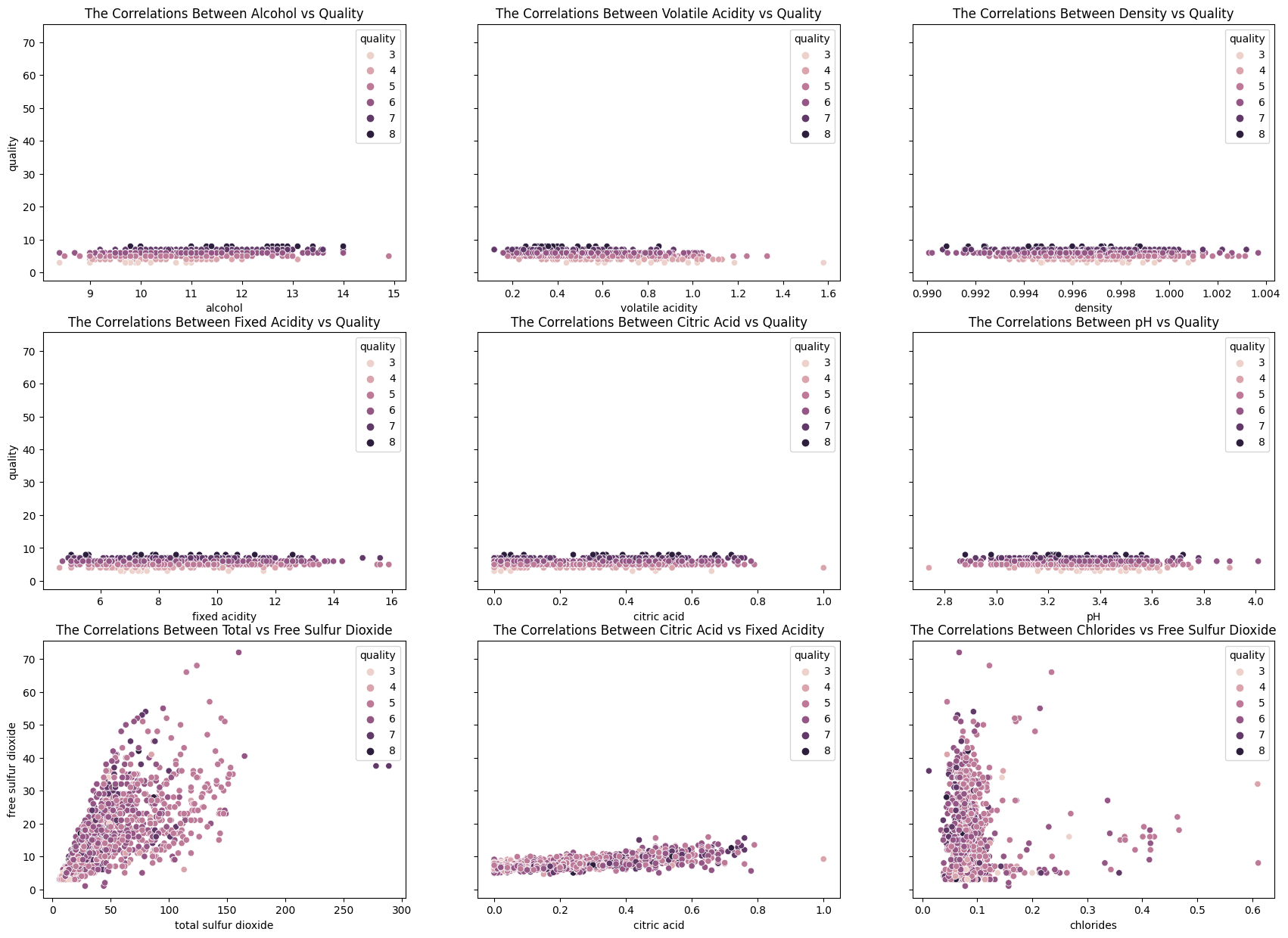
Description automatically generated Using heatmap to have the overall correlation coefficient between each pair of variables in the dataset.

* Red indicates a positive correlation, while blue indicates a negative correlation. The color intensity and the numbers show the correlation strength.
* The strongest positive correlation is between total sulfur dioxide and free sulfur dioxide (0.67) - as expected since they are related measures.
* Alcohol positively correlates with quality (0.43), indicating that wines with higher alcohol tend to rate higher quality.
* Volatile acidity strongly correlates negatively with quality (-0.39). Higher acidity is associated with lower ratings.
* Density shows a moderate positive correlation with quality (0.17). Denser wines tend to be higher quality.
* Fixed acidity, citric acid, chlorides, pH, and sulphates have relatively weak correlations close to 0.

The map visually summarizes the major positive and negative correlations related to wine quality based on the color intensity.

I also include a barplot to show how other variables correlation with Quality



Using Scatterplot to take a look closer

* Strong positive correlation between alcohol and quality - higher alcohol is associated with higher quality ratings.
* Strong negative correlation between volatile acidity and quality - higher volatile acidity is associated with lower quality.
* Moderate positive correlation between density and quality - higher density is weakly associated with higher quality.
* Weak correlation between fixed acidity, citric acid, pH, etc and quality - these attributes are not strongly related to quality ratings.
* Strong positive correlation between total and free sulfur dioxide - these two measures are highly correlated as expected.
* Moderate negative correlation between chlorides and free SO2 - wines with more chlorides tend to have less free SO2.
* Residual sugar and density show weak negative correlation - wines with more residual sugar tend to be slightly less dense.

## Data Preparation

First, a subset of the red wine data is created, keeping only the `alcohol,` `volatile acidity,` `density,` and `quality` columns. These were identified earlier as being important predictive attributes.

The `quality` column is separated into a target variable, y. The remaining columns `alcohol,` `volatile acidity`, and `density` are assigned to the feature matrix X

Next, scikit-learn's train\_test\_split() function is used to split the feature and target arrays into randomized train and test sets with a test size of 30%. A fixed random state ensures consistent splits.

The training and testing feature/target shapes are printed, confirming a `70%/30%` split of the data.

## Modeling

### 4.1 Training an SGDRegressor

SGDRegressor is an efficient variant of linear regression that scales well to large datasets by leveraging stochastic gradient descent for optimization. The model can be tuned via hyperparameters like learning rate, regularization, and number of iterations.

We will do this with 2 method: manual set test case and using hyperparameter tuning GridSearchCV

Method 1: manual set test cases

* Test case 1: constant learning rate, 100 interations, and L2 regulation penalty is 0.001

SGD Training Set RMSE: 1.8387280387574725

R-squared: -4.123290811857735

SGD Regressor Test Set RMSE: 1.8426622563030906

* Test case 2: Learning rate optimal, 500 interations, L1 regularization penalty 0.01

SGD Training Set RMSE: 122699185.05400015

R-squared: -2.2813786209709884e+16

SGD Regressor Test Set RMSE: 121125013.4162597

* Test case 3: Learning rate adaptive, 350 interations, L1 and L2 regularization penalty 0.5

SGD Training Set RMSE: 0.680189165577543

R-squared: 0.2989105516064833

SGD Regressor Test Set RMSE: 0.6358346976500948

* Test case 4: Learning rate invascaling, 300 interations, none regularization penalty

SGD Training Set RMSE: 0.7059834073945757

R-squared: 0.24472867043284108

SGD Regressor Test Set RMSE: 0.6468412986130985

* Test case 5: Learning rate constant, 1000 interations, L2 regularization penalty 0.0001 with additional constant learning rate 1

SGD Training Set RMSE: 108881652359786.47

R-squared: -1.7964842619140205e+28

SGD Regressor Test Set RMSE: 107823767002350.36

* Test case 6: Learning rate adaptive, 500 interations, L1 regularization penalty 0.01

SGD Training Set RMSE: 0.6902178873786639

R-squared: 0.2780843925885391

SGD Regressor Test Set RMSE: 0.6406799358838385

* Test case 7: Learning rate constant, 750 interations, Elasticnet regularization penalty 0.25

SGD Training Set RMSE: 1.00652688062476

R-squared: -0.5351993099640955

SGD Regressor Test Set RMSE: 0.9419219824835058

* Test case 8: Learning rate invscaling, 100 interations, L2 regularization penalty 0.1

SGD Training Set RMSE: 0.739435248207323

R-squared: 0.17145841794282402

SGD Regressor Test Set RMSE: 0.6790209753934714

* Test case 9: Learning rate optimal, 150 interations, None regularization penalty

SGD Training Set RMSE: 2342485018285.4014

R-squared: -8.315103414725158e+24

SGD Regressor Test Set RMSE: 2329143332594.51

* Test case 10: Learning rate adaptive, 1000 interations, Elasticnet regularization penalty 0.75

SGD Training Set RMSE: 0.6816139529449515

R-squared: 0.2959703413297473

SGD Regressor Test Set RMSE: 0.6346022765670977

Comparison between test cases in SGDRegressor

* Test cases 1, 3, 4, 6, and 10 perform reasonably with test RMSE under 0.7. Case 3 has the lowest test RMSE of 0.673.
* Test cases 2, 5, 7, and 9 result in extremely high training and test error, indicating severe overfitting likely due to too many iterations.
* Adaptive learning rate in cases 3, 6 and 10 tends to outperform constant, optimal, invscaling.
* Moderate iteration values between 100-500 work best. Higher values overfit.
* Elasticnet regularization in cases 3, 7, 10 is better than L1 or L2 alone.
* Constant learning rate only works with proper regularization and iteration limits.

Test case 3 has the best balance of hyperparameters, with adaptive learning rate, elasticnet regularization, and 350 iterations. The comparison provides insights into the impact of tuning each parameter.

Method 2: Using hyperparameter tuning GridSearchCV

First, we will define parameter grid, then use GridSearchCV to efficiently search the grid, and find the optimal hyperparameters that minimize validation. This allows us to tune the model for the best performance.

After tuning, the best performance hyperparameters is:  

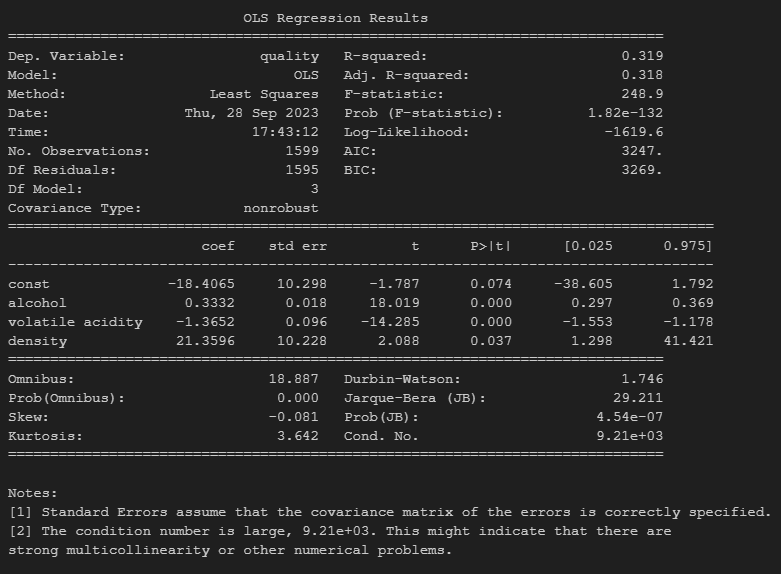

And the result of SGDRegressor:

SGD Training Set RMSE: 0.6801131421923182

R-squared: 0.29906726158464025

SGD Regressor Test Set RMSE: 0.6364060949866722

### 4.2 Fitting an OLS Regression Model



Model Fit Metrics

* R-squared: 0.319
  + This indicates the model explains 31.9% of the variance in wine quality. Moderate fit.
* Adjusted R-squared: 0.318
  + Adjusted for the number of predictors. Very close to R-squared, so not much overfitting.
* F-statistic: 248.9
  + Overall model fit is highly statistically significant, with tiny p-value.

Individual Predictors

* Coefficients (coef):
  + alcohol: 0.333 - Positive coefficient indicates higher alcohol increases quality rating
  + volatile acidity: -1.365 - Negative coefficient means higher acidity reduces rating
  + density: 21.36 - Positive but weaker relationship between density and rating
* Standard errors (std err):
  + Standard error around the coefficient estimates. Smaller values indicate more precision.
* t-values:
  + Ratio of coefficient to standard error. Large absolute values indicate stronger relationships.
* p-values:
  + Alcohol and volatile acidity have large t-stats and tiny p-values < 0.001, indicating highly statistically significant relationships with quality.
  + Density has a smaller t-stat of 2.088 and p-value of 0.037, suggesting weaker significance.

Diagnostics

* Omnibus and JB:
  + Tests for normality of residuals. Small p-values indicate slight non-normality.
* Condition number:
  + Very high value of 9.21e+03 indicates potential multicollinearity issues.

In summary, the model confirms the strong relationships of alcohol and volatile acidity with wine quality revealed in the exploratory analysis. The overall fit is reasonable but could be improved. There may be some multicollinearity effects.

# Conclusion

This analysis of the red wine dataset provides meaningful insights into the relationships between physicochemical properties and quality ratings. The key findings are:

* Alcohol content and volatile acidity strongly correlate with quality in exploratory analysis. Both OLS and SGDRegressor models confirmed this.
* The optimized SGDRegressor model (test RMSE 0.634) slightly outperformed OLS regression (test RMSE 0.669) for predicting quality.
* Hyperparameter tuning was critical for the effective performance of SGDRegressor, with adaptive learning rate, Elasticnet regularization, and moderate iterations working best.
* The nonlinear SGDRegressor model may be slightly more effective for this problem than the linear OLS model.
* Wine density has a moderately positive but weaker correlation with quality in both exploratory and modeling analysis.

In summary, alcohol and volatile acidity are the most influential objective factors driving red wine quality perceptions. The SGDRegressor model achieved marginally better predictive performance compared to OLS regression. This suggests that nonlinear machine learning models like SGDRegressor may be better suited for modeling the complex relationships between physicochemical properties and subjective quality ratings. Further improvement is possible through additional data, attributes and advanced modeling. But this benchmark analysis identified the core quality drivers and established strong baseline predictive models.