Assignment 1 Report: Linear Regression Analysis

Created by: Shraddha Gangaram and Tanvi Nair

Wine Quality Dataset: https://archive.ics.uci.edu/dataset/186/wine+quality

Objective: The objective of this project is to analyze the dataset, perform preprocessing and exploratory data analysis (EDA), and then build two different models for regression:

1. Stochastic Gradient Descent using the SGDRegressor library of Scikit-learn.
2. Ordinary Linear Regression using the statsmodels library.

Dataset Introduction: The dataset used in this study is the Red Wine Quality Dataset from the UCI Machine Learning Repository. It contains 1,599 observations of red Portuguese *Vinho Verde* wine, each described by 11 chemical attributes and wine quality score (the target variable).

Attributes:

- Fixed acidity: non-volatile acids in wine.
- Volatile acidity: acetic acid.
- Citric acid: adds freshness, can improve wine quality.
- Residual sugar: sugar left after fermentation.
- Chlorides: salt content.
- Free sulfur dioxide / Total sulfur dioxide: preservatives.
- Density: density of wine (linked with sugar & alcohol).
- pH: acidity (low pH = more acidic).
- Sulphates: sulfur dioxide compounds (stability, bitterness).
- Alcohol: % of alcohol/ethanol content.

Target Variable:

- Quality: integer value from 0–10.

Data Cleansing and Pre-processing: Before analyzing the dataset, some preprocessing steps were carried out to ensure the data was consistent and reliable. The changes made are as follows:

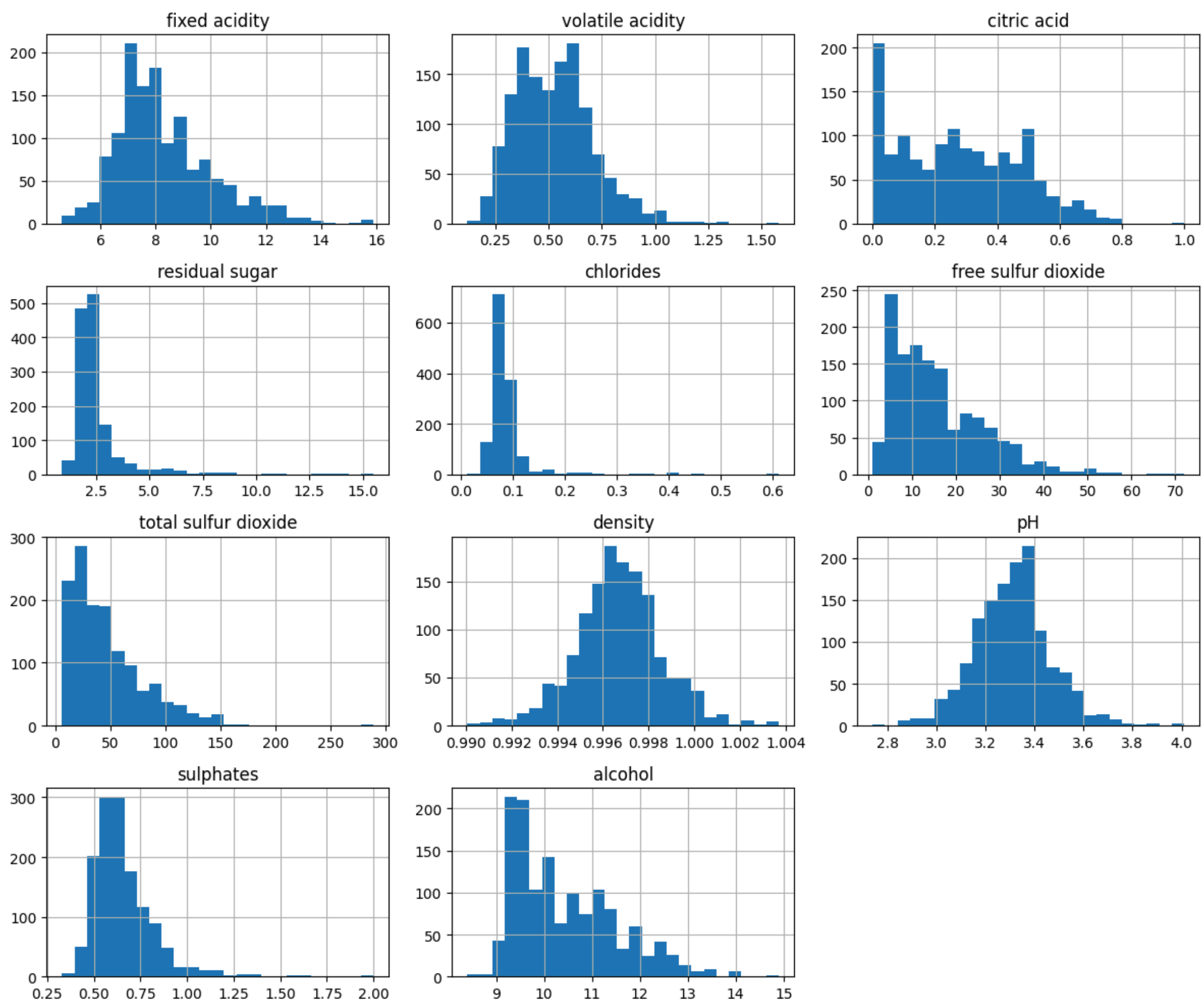
- Removed null or NA values.
- Dropped duplicate rows.
- Checked for irrelevant variables (none found).
- Converted any categorical variables to numeric (none found).

Attribute Summary: A descriptive summary was generated for all attributes, showing their ranges, means, and standard deviations.

Summary statistics:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000
mean	8.310596	0.529478	0.272333	2.523400	0.088124	15.893304	46.825975	0.996709	3.309787	0.658705	10.432315
std	1.736990	0.183031	0.195537	1.352314	0.049377	10.447270	33.408946	0.001869	0.155036	0.170667	1.082065
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996700	3.310000	0.620000	10.200000
75%	9.200000	0.640000	0.430000	2.600000	0.091000	21.000000	63.000000	0.997820	3.400000	0.730000	11.100000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000

Attribute Distributions: Histograms were created for all predictor variables to visualize their distributions.



As seen above, most attributes are not normally distributed:

- Residual sugar, chlorides, sulphur dioxide, sulphates, and citric acid are strongly right-skewed.
- Alcohol and volatile acidity show moderate skewness.
- Fixed acidity is slightly skewed.
- Density and pH are the closest to normal, but still not perfectly symmetric.

Overall, this is expected because the chemical properties of wine are naturally bound, influenced by winemaking practices, and often cluster around typical production levels.

Data Transformation and Exploration: After cleaning, the dataset was standardized (mean = 0, standard deviation = 1) and normalized to ensure comparability across variables. These transformations prevent attributes with larger scales from dominating the model and help make correlation patterns more interpretable.

Correlation Matrix: A table was created to show linear relationships among all attributes.

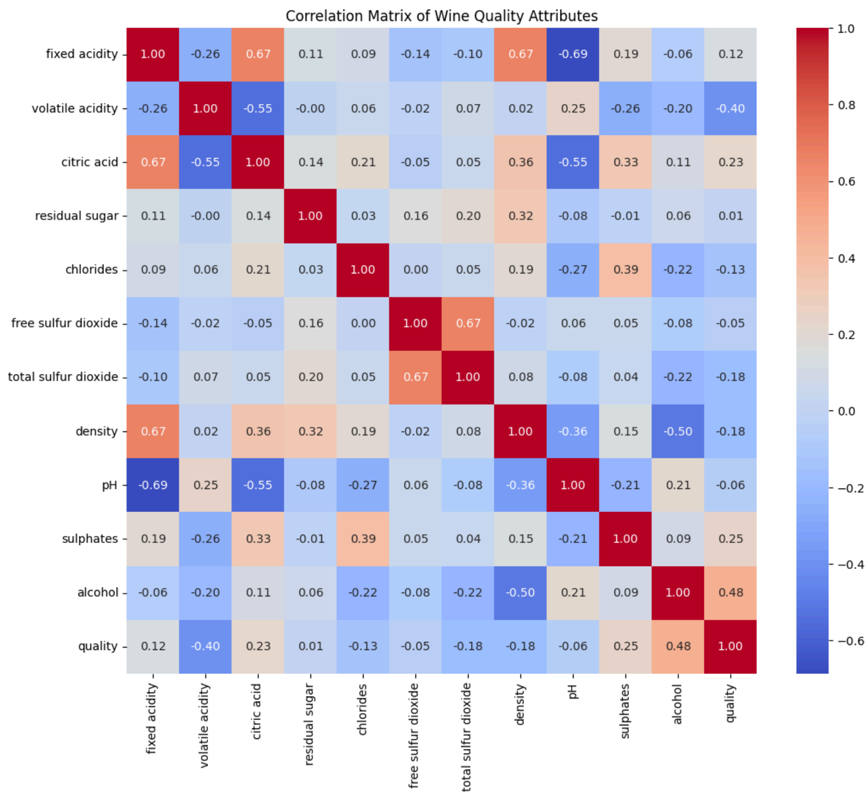
Correlation matrix:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.000000	-0.255124	0.667437	0.111025	0.085886	-0.140580	-0.103777	0.670195	-0.686685	0.190269	-0.061596	0.119024
volatile acidity	-0.255124	1.000000	-0.551248	-0.002449	0.055154	-0.020945	0.071701	0.023943	0.247111	-0.256948	-0.197812	-0.395214
citric acid	0.667437	-0.551248	1.000000	0.143892	0.210195	-0.048004	0.047358	0.357962	-0.550310	0.326062	0.105108	0.228057
residual sugar	0.111025	-0.002449	0.143892	1.000000	0.026656	0.160527	0.201038	0.324522	-0.083143	-0.011837	0.063281	0.013640
chlorides	0.085886	0.055154	0.210195	0.026656	1.000000	0.000749	0.045773	0.193592	-0.270893	0.394557	-0.223824	-0.130988
free sulfur dioxide	-0.140580	-0.020945	-0.048004	0.160527	0.000749	1.000000	0.667246	-0.018071	0.056631	0.054126	-0.080125	-0.050463
total sulfur dioxide	-0.103777	0.071701	0.047358	0.201038	0.045773	0.667246	1.000000	0.078141	-0.079257	0.035291	-0.217829	-0.177855
density	0.670195	0.023943	0.357962	0.324522	0.193592	-0.018071	0.078141	1.000000	-0.355617	0.146036	-0.504995	-0.184252
pH	-0.686685	0.247111	-0.550310	-0.083143	-0.270893	0.056631	-0.079257	-0.355617	1.000000	-0.214134	0.213418	-0.055245
sulphates	0.190269	-0.256948	0.326062	-0.011837	0.394557	0.054126	0.035291	0.146036	-0.214134	1.000000	0.091621	0.248835
alcohol	-0.061596	-0.197812	0.105108	0.063281	-0.223824	-0.080125	-0.217829	-0.504995	0.213418	0.091621	1.000000	0.480343
quality	0.119024	-0.395214	0.228057	0.013640	-0.130988	-0.050463	-0.177855	-0.184252	-0.055245	0.248835	0.480343	1.000000

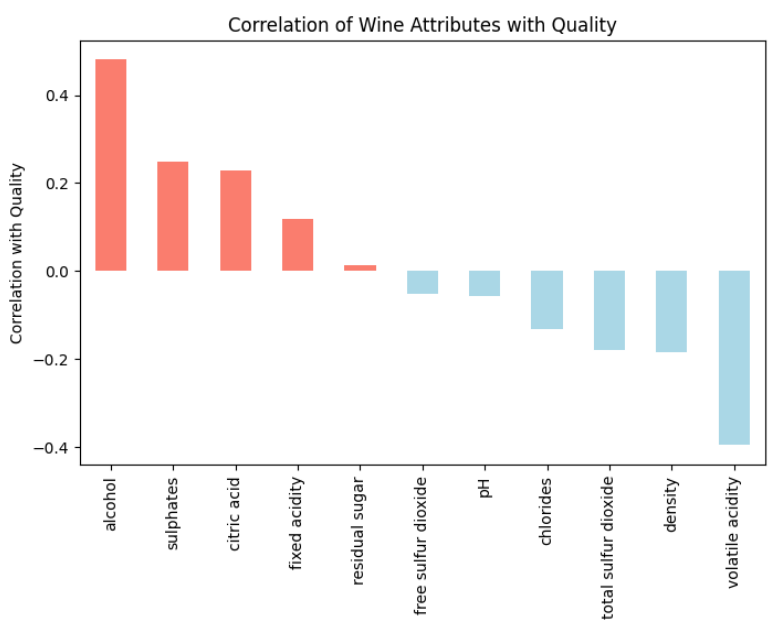
Correlation with Target: A summary showing how each attribute correlates with the target variable (wine quality)

```
Correlation of each attribute with target (quality):
quality                 1.000000
alcohol                 0.480343
sulphates               0.248835
citric acid             0.228057
fixed acidity           0.119024
residual sugar          0.013640
free sulfur dioxide    -0.050463
pH                     -0.055245
chlorides              -0.130988
total sulfur dioxide   -0.177855
density                -0.184252
volatile acidity       -0.395214
```

Heatmap: A color-coded visualization of the correlation matrix, highlighting strong positive and negative relationships.



Bar Plot: A visualization displaying the strength of correlation of each attribute with the target variable, making it easier to identify the most influential predictors.



Feature Selection: After examining the correlations of each attribute with wine quality through the above numerical outputs and visualizations, the following predictors seem to have the strongest linear relationships with the target variable.

```
Attributes with strong positive correlation with quality:
alcohol        0.480343
sulphates      0.248835
citric acid    0.228057
Name: quality, dtype: float64

Attributes with strong negative correlation with quality:
volatile acidity   -0.395214
Name: quality, dtype: float64
```

Therefore, these features were selected as the important predictors to be used in model building:

```
Selected important features for modeling:
['alcohol', 'sulphates', 'citric acid', 'volatile acidity']
```

Training and Testing Sets: The dataset was split into training (80%) and testing (20%) subsets to allow for model evaluation on unseen data. The predictor variables were then standardized to ensure all features contributed equally to the models.

Assumptions

- Regression Task: The wine quality dataset is treated as a regression problem, assuming quality can be modeled as a continuous numeric outcome.
- Linearity: The relationship between selected predictors (alcohol, sulphates, citric acid, volatile acidity) and wine quality is approximately linear.
- Independence: Each wine sample is independent, which means that no autocorrelation exists among observations.

Model Construction with SGDRegressor (with Hyperparameter Tuning): A Stochastic Gradient Descent (SGD) Regressor was trained using the four selected features: alcohol, sulphates, citric acid, and volatile acidity. To optimize performance, a grid search was conducted over multiple hyperparameters, including loss functions, penalties, alpha values, learning rates, and initial step sizes.

For each combination, the model was trained and evaluated on the training and test sets, and the results were recorded and saved into a CSV file (sgd_results.csv) for comparison:

sgd_results.csv

loss	penalty	alpha	learning_rate	eta0	train_r2	test_r2	mse	mae	explained_variance	r2
squared_error	l2	0.0001	constant	0.001	0.3212216328809123	0.41986190369788456	0.4109442205464015	0.49606083781378607	0.4227369120188662	0.41986190369788456
squared_error	l2	0.0001	constant	0.01	0.3130205866579171	0.4029053669100262	0.4229554827577079	0.4998021232298458	0.4038081474967046	0.4029053669100262
squared_error	l2	0.0001	optimal	0.001	0.25856934452350566	0.3598929541133936	0.4534235774461962	0.5126069436247223	0.3601075561739212	0.3598929541133936
squared_error	l2	0.0001	optimal	0.01	0.25856934452350566	0.3598929541133936	0.4534235774461962	0.5126069436247223	0.3601075561739212	0.3598929541133936
squared_error	l2	0.001	constant	0.001	0.3212293393527279	0.4197837207849676	0.41099960187097484	0.49612023378146486	0.4226603625237516	0.4197837207849676
squared_error	l2	0.001	constant	0.01	0.31304789928646604	0.4028432467503934	0.42299948593685305	0.4998588825949046	0.40374485704601437	0.4028432467503934
squared_error	l2	0.001	optimal	0.001	0.26832697483410706	0.32082937986937354	0.48109448920287823	0.535785055478289	0.33345259798133065	0.32082937986937354
squared_error	l2	0.001	optimal	0.01	0.26832697483410706	0.32082937986937354	0.48109448920287823	0.535785055478289	0.33345259798133065	0.32082937986937354
squared_error	l2	0.01	constant	0.001	0.32178307301906217	0.41584966076798435	0.4137863163404553	0.49849172830385247	0.4192308067577035	0.41584966076798435
squared_error	l2	0.01	constant	0.01	0.3132882580767661	0.402201438432908	0.42345411462000687	0.5004255798764904	0.40309160005872735	0.402201438432908
squared_error	l2	0.01	optimal	0.001	0.3125900311651082	0.399164626501901	0.42560525814926037	0.5035668398506605	0.4131797070519444	0.399164626501901
squared_error	l2	0.01	optimal	0.01	0.3125900311651082	0.399164626501901	0.42560525814926037	0.5035668398506605	0.4131797070519444	0.399164626501901
squared_error	l1	0.0001	constant	0.001	0.32121632830485025	0.4198325963265135	0.4109649805946747	0.4960735807962307	0.42270847059407735	0.4198325963265135
squared_error	l1	0.0001	constant	0.01	0.3131094026517227	0.40268657335923763	0.42311046645169814	0.49988533487749137	0.40356475796483116	0.40268657335923763
squared_error	l1	0.0001	optimal	0.001	0.09714560816773798	0.12129162763989287	0.6224382254308518	0.6437182525335519	0.13184833036913302	0.12129162763989287
squared_error	l1	0.0001	optimal	0.01	0.09714560816773798	0.12129162763989287	0.6224382254308518	0.6437182525335519	0.13184833036913302	0.12129162763989287
squared_error	l1	0.001	constant	0.001	0.3213418475547257	0.419211230044571 4	0.41140512904214627	0.49644418652937655	0.4221079902567092	0.4192112300445714
squared_error	l1	0.001	constant	0.01	0.31755440280932234	0.42208053794925915	0.4093726379716314	0.4949502513420555	0.42224268439673285	0.42208053794925915
squared_error	l1	0.001	optimal	0.001	0.2906432079721911	0.3579168719131436	0.4548233468540209	0.5224193775712065	0.3735051112757155	0.3579168719131436
squared_error	l1	0.001	optimal	0.01	0.2906432079721911	0.3579168719131436	0.4548233468540209	0.5224193775712065	0.3735051112757155	0.3579168719131436
squared_error	l1	0.01	constant	0.001	0.32147259397117367	0.4117836761791501	0.4166664803535803	0.5009045557430926	0.41525760653153465	0.4117836761791501
squared_error	l1	0.01	constant	0.01	0.31289672034326954	0.39662883788020764	0.4274014988810007	0.5029943623037952	0.3972214758801128	0.39662883788020764

The best-performing configuration is shown in the results below:

```
SGD Best hyperparameter combination based on Test R²:
            loss penalty  alpha learning_rate  eta0  train_r2   test_r2  \
17  squared_error      l1  0.001      constant  0.01  0.317554  0.422081

        mse      mae  explained_variance        r2
17  0.409373  0.49495            0.422243  0.422081
```

SGD Regressor Model Interpretations:

- R-squared: 0.422 – This suggests that approximately 42.2% of the variance in wine quality can be explained by the model. While moderate, it indicates that other factors not included in the dataset may also influence wine quality.
- Best Hyperparameters: loss = squared_error, penalty = l1, alpha = 0.001, learning rate = constant, eta0 = 0.01.

Error Metrics:

- Mean Squared Error (MSE): 0.409 – Reflects the average squared difference between predicted and actual values. Since MSE penalizes larger errors more heavily, this value suggests that most predictions are reasonably close to the true quality ratings.
- Mean Absolute Error (MAE): 0.495 – Represents the average absolute difference between predictions and actual values. Given that wine quality scores in the dataset range from 3 to 8, an average error of less than 0.5 is relatively small, indicating that predictions are often within half a point of the true score.
- Explained Variance Score: 0.422 – Very close to the R^2 score, further supporting that the model captures about 42% of the variation in wine quality.

SGD Regressor Coefficients: To further interpret the contribution of each feature, we examined the coefficients of the SGD Regressor. These coefficients indicate the magnitude and direction of each predictor's effect on the predicted wine quality (positive values indicate that higher feature values increase the predicted wine quality, whereas negative values indicate a decrease):

```
SGD Regressor Coefficients:
alcohol              0.324497
sulphates            0.099736
citric acid          0.000000
volatile acidity    -0.250316
dtype: float64
```

Coefficients:

- Alcohol: 0.3245 – For a one-unit increase in alcohol content (while holding other variables constant), the predicted wine quality increases by 0.3245, indicating a positive contribution of alcohol.
- Sulphates: 0.0997 – For a one-unit increase in sulphates (while holding other variables constant), the predicted wine quality increases by 0.0997, showing a small but positive impact.
- Citric acid: 0.0000 – The coefficient for citric acid is zero, meaning it does not contribute to the model's predictions.
- Volatile acidity: -0.2503 – For a one-unit increase in volatile acidity (while holding other variables constant), the predicted wine quality decreases by 0.2503, indicating a negative impact.

Overall Interpretation: The SGD model suggests that alcohol, sulphates, and volatile acidity are meaningful predictors of wine quality. As seen/explained above, alcohol and sulphates have a positive impact on predicted wine quality, while volatile acidity has a negative impact. Lastly, citric acid does not appear to contribute to the model, as its coefficient is zero. Overall, while the model demonstrates a moderate ability to predict wine quality, there is still substantial unexplained variance, suggesting that additional features or more complex models could improve performance.

Model Construction with OLS Regression: An Ordinary Least Squares (OLS) regression model was built using the four selected features: alcohol, sulphates, citric acid, and volatile acidity. The model was constructed with the statsmodels library to obtain detailed statistical insights, including R^2 values, F-statistics, coefficient estimates, and diagnostic tests.

The OLS Regression Results are shown below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                quality   R-squared:                       0.322
Model:                            OLS   Adj. R-squared:                  0.319
Method:                 Least Squares   F-statistic:                     128.4
Date:                Sun, 21 Sep 2025   Prob (F-statistic):           9.78e-90
Time:                        23:21:21   Log-Likelihood:                -1113.0
No. Observations:                1087   AIC:                             2236.
Df Residuals:                    1082   BIC:                             2261.
Df Model:                           4
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const                2.6629      0.246     10.814      0.000       2.180       3.146
alcohol              0.3077      0.019     16.002      0.000       0.270       0.345
sulphates            0.6352      0.126      5.047      0.000       0.388       0.882
citric acid         -0.0616      0.129     -0.477      0.633      -0.315       0.192
volatile acidity    -1.2120      0.139     -8.710      0.000      -1.485      -0.939
==============================================================================
Omnibus:                       18.295   Durbin-Watson:                   1.952
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.534
Skew:                          -0.150   Prob(JB):                     1.05e-06
Kurtosis:                       3.720   Cond. No.                         137.
==============================================================================
```

OLS Regression Model Interpretations:

- Our R-squared of 0.322 suggests that approximately 32.2% of the variance in wine quality can be explained by the selected features (alcohol, sulphates, citric acid, and volatile acidity). This is a relatively low R-squared, indicating that these features alone do not fully capture the factors influencing wine quality.
- Adjusted R-squared: 0.319 - This adjusted R-squared accounts for the number of predictors in the model and is slightly lower than the R-squared, as expected.
- F-statistic: 128.4 - This is a test of the overall significance of the regression model. The large F-statistic and the very small p-value (9.78e-90) indicate that the model is statistically significant, meaning that at least one of the predictors is related to wine quality.

Coefficients and P-values:

- Constant: 2.6629 - This is the intercept of the model. It represents the expected wine quality when all predictor variables are zero.
- Alcohol: 0.3077 - For a one-unit increase in alcohol content (while holding other variables constant), the expected wine quality increases by 0.3077. The p-value (0.000) is less than 0.05, indicating that alcohol content is a statistically significant predictor of wine quality.

- Sulphates: 0.6352 - For a one-unit increase in sulphates (while holding other variables constant), the expected wine quality increases by 0.6352. The p-value (0.000) is less than 0.05, indicating that sulphates are also a statistically significant predictor of wine quality.
- Citric acid: -0.0616 - For a one-unit increase in citric acid (while holding other variables constant), the expected wine quality decreases by 0.0616. The p-value (0.633) is greater than 0.05, indicating that citric acid is not a statistically significant predictor of wine quality in this model.
- Volatile acidity: -1.2120 - For a one-unit increase in volatile acidity (while holding other variables constant), the expected wine quality decreases by 1.2120. The p-value (0.000) is less than 0.05, indicating that volatile acidity is a statistically significant predictor of wine quality.

Other Diagnostics:

- Omnibus, Jarque-Bera, Skew, and Kurtosis are tests for the normality of the residuals. The significant p-values for Omnibus and Jarque-Bera, along with the non-zero skew and kurtosis different from 3, suggest that the residuals are not normally distributed.
- Durbin-Watson had a statistic of 1.952, which tests for autocorrelation in the residuals. Such a value close to 2 suggests no significant autocorrelation.
- Condition Number: 137 - A high condition number (usually above 30) indicates potential multicollinearity, which can make the coefficient estimates unstable. Therefore, the value of 137 suggests that there might be some multicollinearity among the predictor variables.

Overall Interpretation: The OLS model suggests that alcohol, sulphates, and volatile acidity are statistically significant predictors of wine quality. Alcohol and sulphates have a positive impact, while volatile acidity has a negative impact. Citric acid does not appear to be a significant predictor in this model. The model explains a moderate amount of the variance in wine quality, but the non-normal residuals and potential multicollinearity suggest that there might be limitations or areas for further investigation in the model.

Conclusion: In this study, we analyzed the Red Wine Quality dataset and applied both Stochastic Gradient Descent (SGD) and Ordinary Least Squares (OLS) regression models to predict wine quality based on selected chemical features. Both models consistently highlighted alcohol, sulphates, and volatile acidity as key predictors, with citric acid having negligible influence. While the SGD model achieved slightly better predictive performance, the OLS model offered detailed statistical insights into the significance of each feature. Overall, these findings demonstrate the importance of chemical composition in determining wine quality and provide a foundation for further predictive modeling.