**Final Report: Wine Quality Regression Analysis**

**1. Introduction**

The objective of this project was to build and evaluate multiple regression models to predict wine quality based on various physicochemical properties. The models used in this analysis include Linear Regression, Lasso Regression, Polynomial Regression, and Random Forest Regression.

**2. Data Preparation**

**2.1. Data Cleaning**

* **Dataset**: winequality-red.csv
* **Steps**:
  + Removed rows where citric acid was 0.0.
  + Removed duplicate rows.
  + Addressed outliers using z-scores, removing data points beyond 4 standard deviations.

**2.2. Data Standardization**

* Standardized features to have mean 0 and standard deviation 1.

**3. Model Training and Evaluation**

**3.1. Linear Regression**

* **Model**: Linear Regression
* **Performance**:
  + **R-squared**: 0.4252
  + **Mean Squared Error (MSE)**: 0.3087

**3.2. Lasso Regression**

* **Model**: LassoCV with cross-validation (cv=5)
* **Performance**:
  + **Optimal alpha**: (determined by LassoCV)
  + **R-squared**: 0.4372
  + **Mean Squared Error (MSE)**: 0.3023

**3.3. Polynomial Regression**

* **Model**: Polynomial Regression (degree=2)
* **Performance**:
  + **R-squared**: 0.3593
  + **Mean Squared Error (MSE)**: 0.3441

**3.4. Random Forest Regression**

* **Model**: Random Forest with GridSearchCV for hyperparameter tuning
* **Performance**:
  + **R-squared**: 0.4158
  + **Mean Squared Error (MSE)**: 0.3137

**4. Model Comparison**

| **Model** | **R-squared** | **Mean Squared Error** |
| --- | --- | --- |
| Linear Regression | 0.4252 | 0.3087 |
| Lasso Regression | 0.4372 | 0.3023 |
| Polynomial Regression | 0.3593 | 0.3441 |
| Random Forest | 0.4158 | 0.3137 |

**Insights:**

* **Best Performer**: Lasso Regression with the highest R-squared (0.4372) and lowest MSE (0.3023).
* **Polynomial Regression**: Added complexity did not improve the performance significantly.
* **Random Forest**: Slightly lower performance compared to Lasso Regression.

**5. Residual Analysis**

* **Residual Plot**: Residuals were randomly scattered around zero, indicating a good model fit without systematic errors.
* **Histogram of Residuals**: Residuals followed a normal distribution, confirming that the assumptions of linear regression were met.

**6. Conclusion**

Lasso Regression was identified as the best model for predicting wine quality based on the given dataset. This model not only provided the highest R-squared and lowest MSE but also offered the benefit of feature selection by shrinking some coefficients to zero. The residual analysis confirmed that the model fit was appropriate, with no evident patterns in the residuals.

Future steps could include further fine-tuning of hyperparameters, experimenting with other advanced models, or integrating ensemble methods for potential performance improvements.