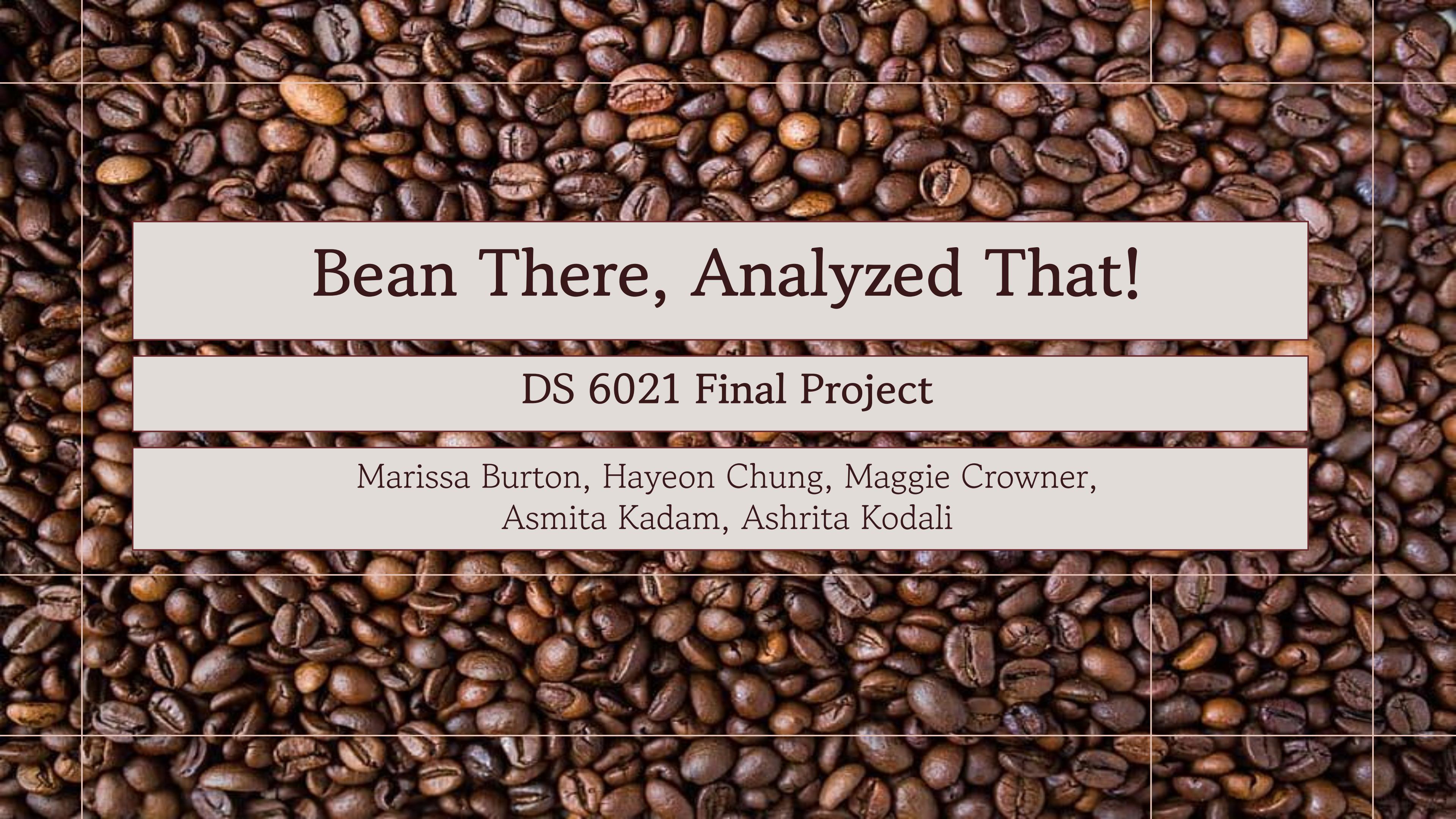


Bean There, Analyzed That!

DS 6021 Final Project

Marissa Burton, Hayeon Chung, Maggie Crowner,
Asmita Kadam, Ashrita Kodali

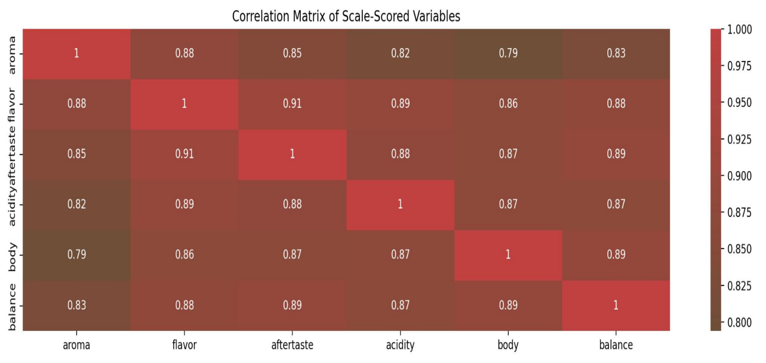
Data

- Collected from the Coffee Quality Institute
- Current data scraped using code from https://github.com/jldbc/coffee-quality-database/tree/master
- **24 Columns:** Country of Origin, Number of Bags, Bag Weight, Harvest Year, Grading Date, Processing Method, Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup Score, Sweetness, Moisture, Category One Defects, Quakers, Color, Category Two Defects, Expiration, Altitude, Species, Total Quality; ~400 rows

Research Questions

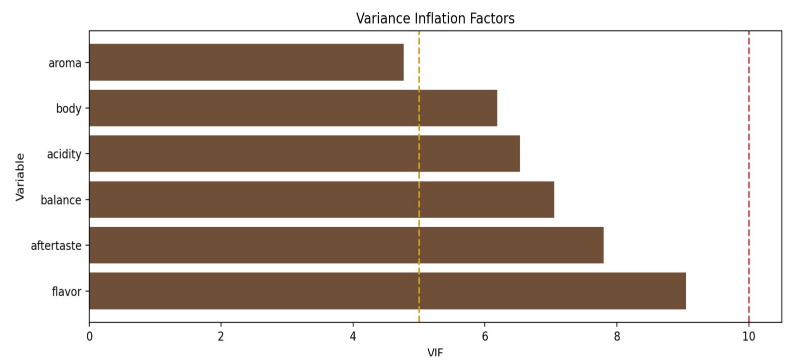
- What distinct profiles of arabica/robusta coffee beans can we identify using K-Means Clustering?
- Can we predict total coffee quality scores based on certain characteristics using Linear Regression?
- How well can we classify coffee beans as arabica or robusta based on their characteristics?
- How well can we predict the altitude of the coffee bean farms using K-Nearest Neighbors Regression?
- Are we able to effectively use Multilayer Perceptrons to predict the market grade of coffee beans based on farming and physical attributes?

Exploratory Data Analysis

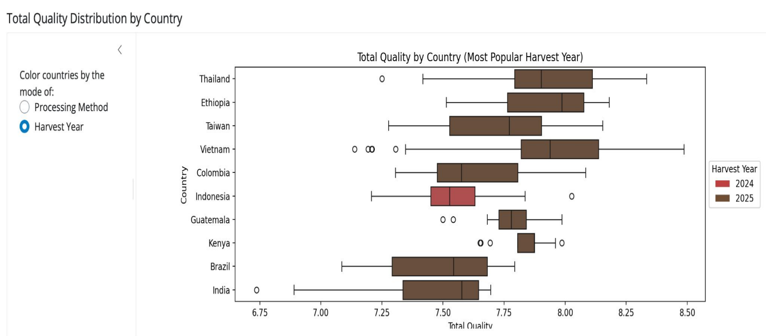


Correlation Matrix of Scale-Scored Variables

	aroma	flavor	aftertaste	acidity	body	balance
aroma	1	0.88	0.85	0.82	0.79	0.83
flavor	0.88	1	0.91	0.89	0.86	0.88
aftertaste	0.85	0.91	1	0.88	0.87	0.89
acidity	0.82	0.89	0.88	1	0.87	0.87
body	0.79	0.86	0.87	0.87	1	0.89
balance	0.83	0.88	0.89	0.87	0.89	1

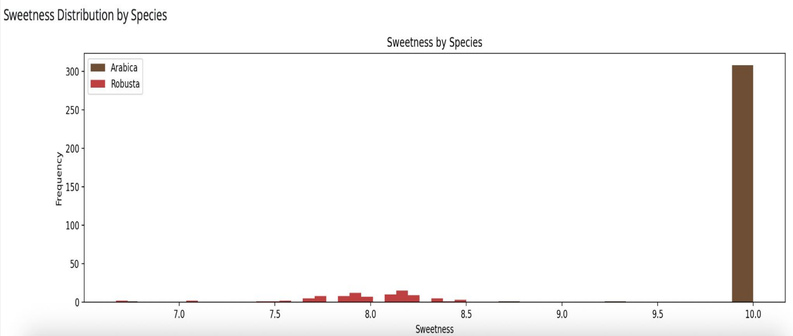


Variance Inflation Factors



Total Quality Distribution by Country

Color countries by the mode of:
○ Processing Method
● Harvest Year

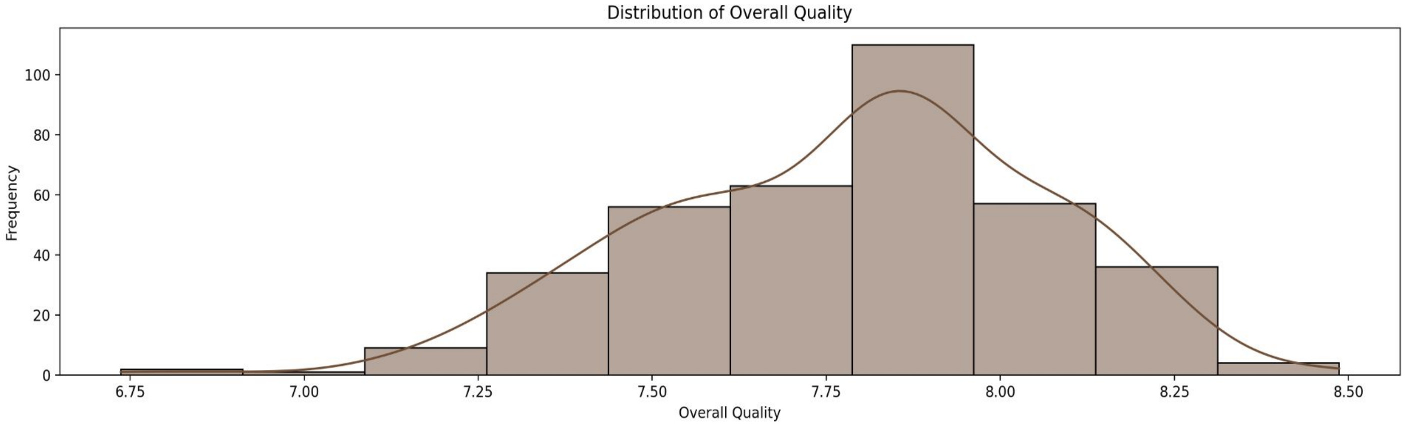


Sweetness Distribution by Species

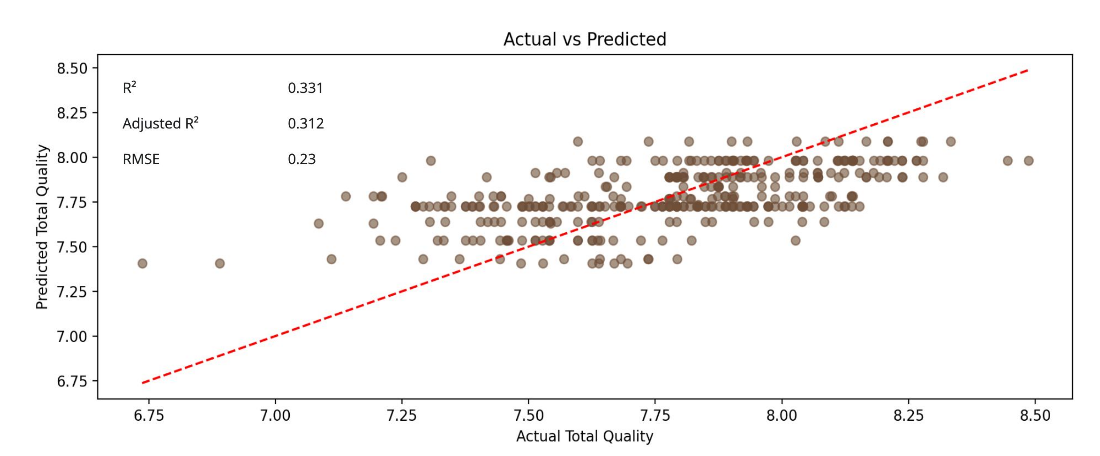
Model 1: Linear Regression

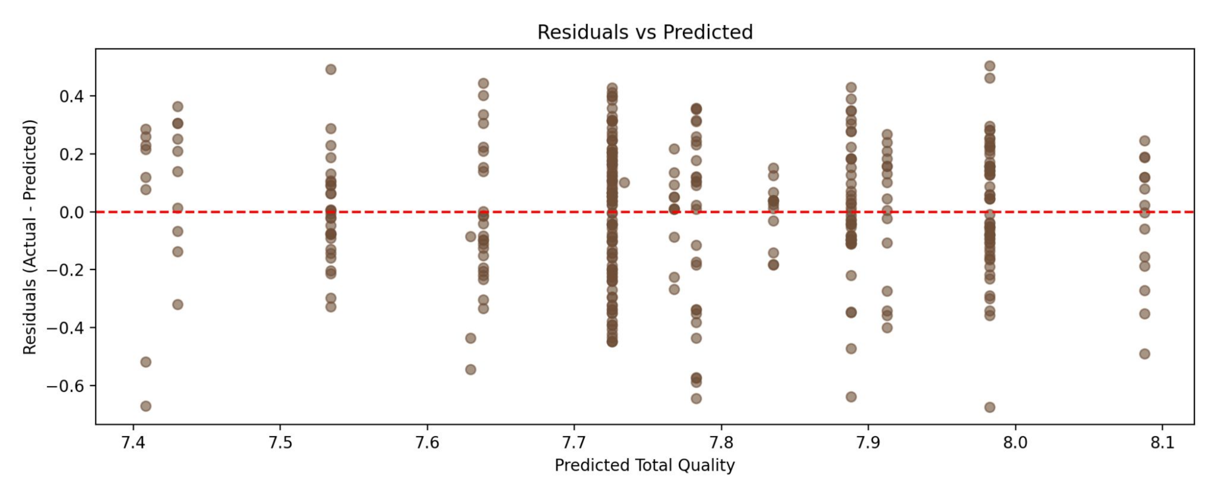
- **Predictor Variables:** Country of Origin and Species
- **Target Variable:** Total Quality (combination of 6 scale-scored variables)

Distribution of Total Quality



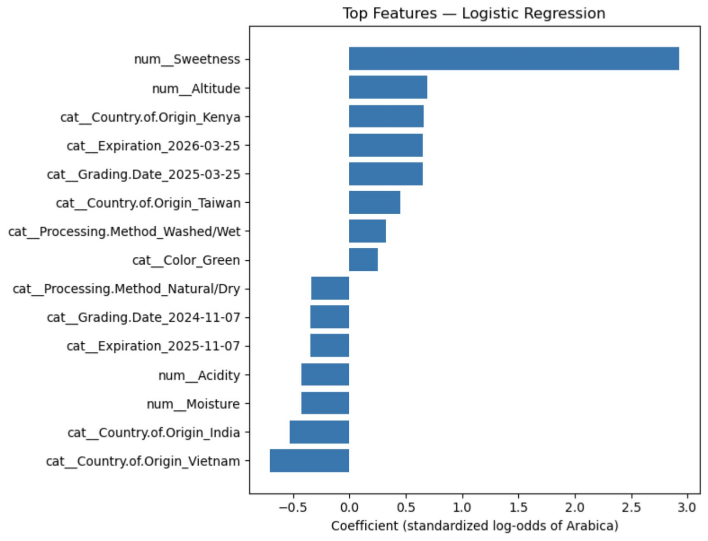
Linear Regression Plots





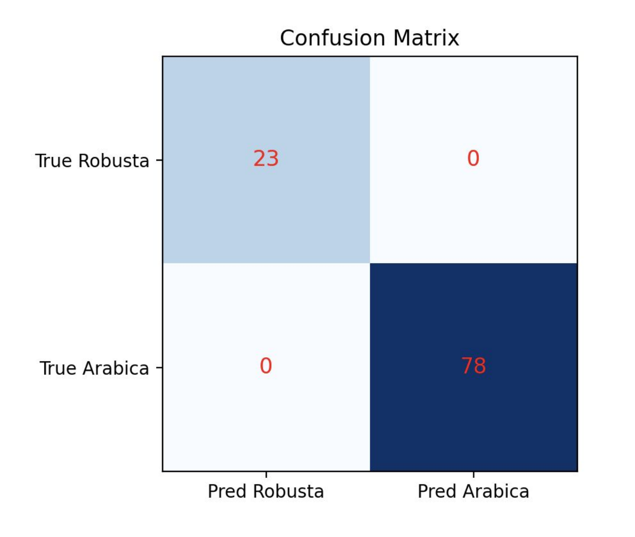
Model 2: Logistic Regression

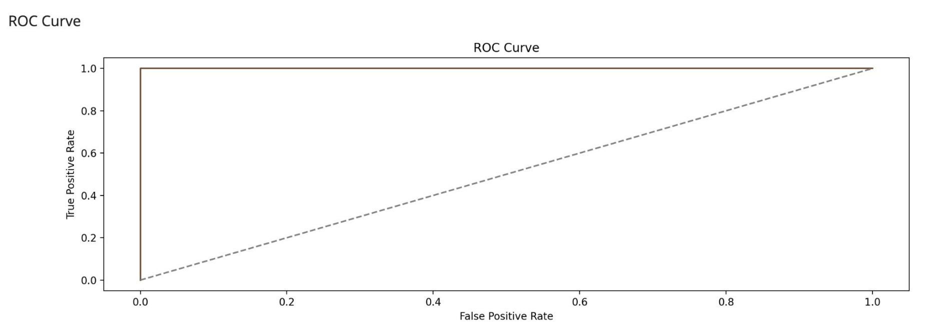
- The goal of **Logistic Regression** is to predict coffee bean species (Arabica, Robusta) based on different attributes.
- To prepare the data for logistic regression, we used a **scikit learn pipeline** that automates all preprocessing (SimpleImputer, StandardScaler, OneHotEncoder, ColumnTransformer, Pipeline)
- Used a **train/test split** with 75% training data and 25% testing data
- Fit the model and found the **most influential features**, as shown in the plot below:



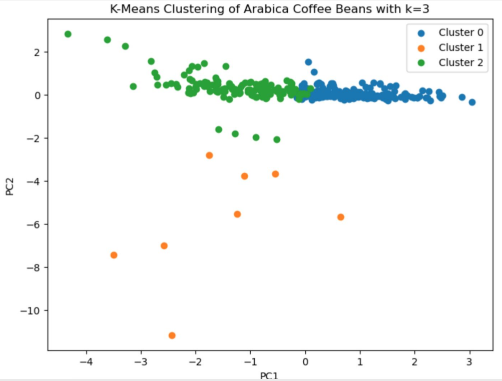
Model 2: Logistic Regression

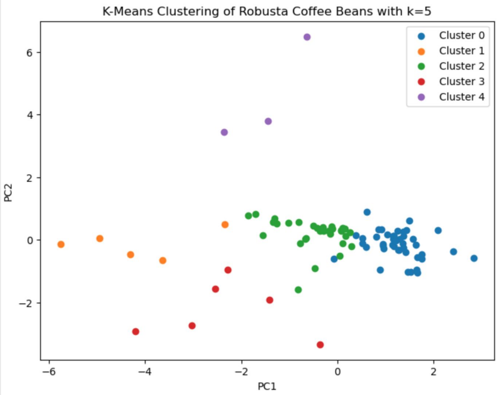
- Our model performed extremely well on test data as seen in the **Confusion Matrix**
- **ROC Curve** plots True Positive rate vs False Positive Rate.
- The shape shows the model is **consistently good** at distinguishing from Arabica from Robusta.



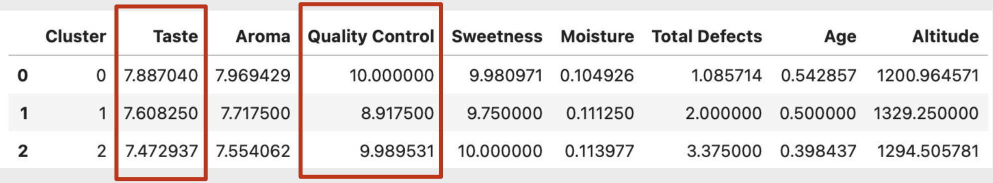


Model 3: K-Means



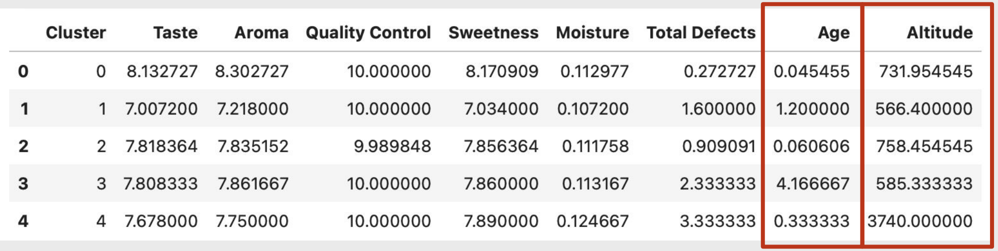


Arabica:



	Cluster	Taste	Aroma	Quality Control	Sweetness	Moisture	Total Defects	Age	Altitude
0	0	7.887040	7.969429	10.000000	9.980971	0.104926	1.085714	0.542857	1200.964571
1	1	7.608250	7.717500	8.917500	9.750000	0.111250	2.000000	0.500000	1329.250000
2	2	7.472937	7.554062	9.989531	10.000000	0.113977	3.375000	0.398437	1294.505781

Robusta:

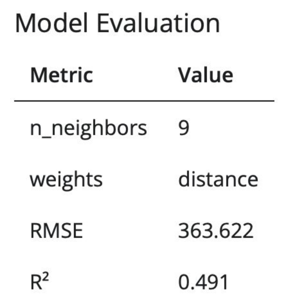


	Cluster	Taste	Aroma	Quality Control	Sweetness	Moisture	Total Defects	Age	Altitude
0	0	8.132727	8.302727	10.000000	8.170909	0.112977	0.272727	0.045455	731.954545
1	1	7.007200	7.218000	10.000000	7.034000	0.107200	1.600000	1.200000	566.400000
2	2	7.818364	7.835152	9.989848	7.856364	0.111758	0.909091	0.060606	758.454545
3	3	7.808333	7.861667	10.000000	7.860000	0.113167	2.333333	4.166667	585.333333
4	4	7.678000	7.750000	10.000000	7.890000	0.124667	3.333333	0.333333	3740.000000

*Taste, Quality Control, Total Defects, Age are aggregates/calculations from our original variables

Model 4: KNN

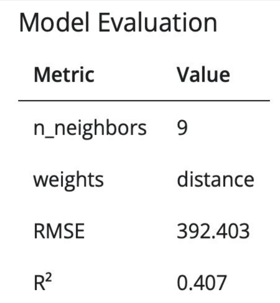
1) Without PCA



Model Evaluation

Metric	Value
n_neighbors	9
weights	distance
RMSE	363.622
R^2	0.491

2) With PCA

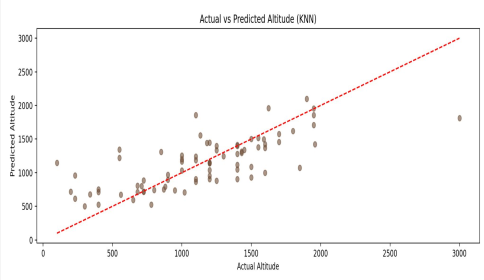


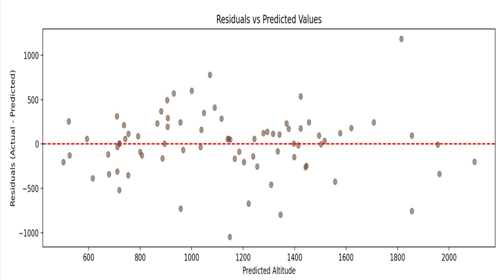
Model Evaluation

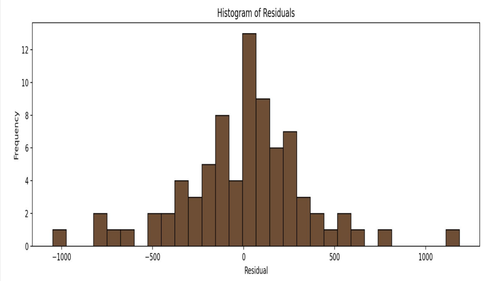
Metric	Value
n_neighbors	9
weights	distance
RMSE	392.403
R^2	0.407

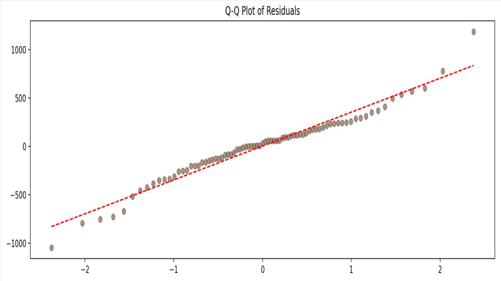
Without PCA yielded better results!

Model Diagnostic Plots



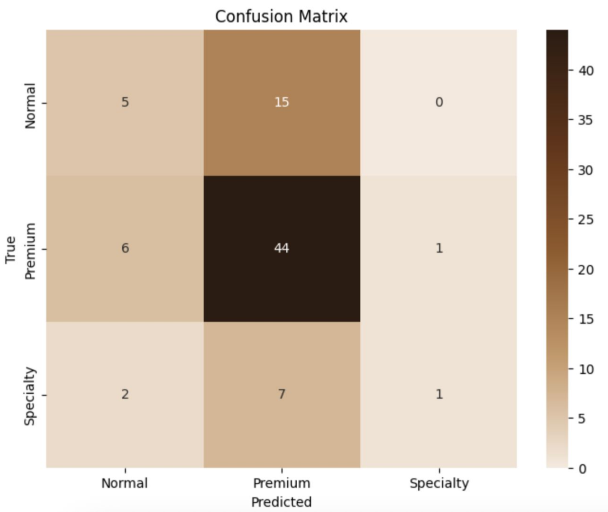




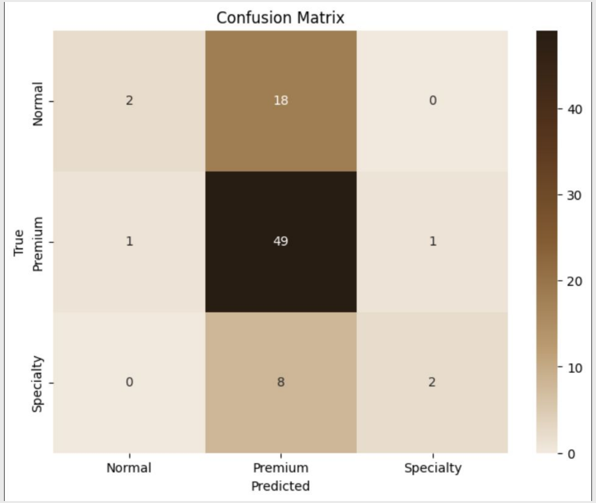


Model 5: MLP

Simple MLP (fewer layers):



Complex MLP (more layers):



Best Accuracy: 0.617

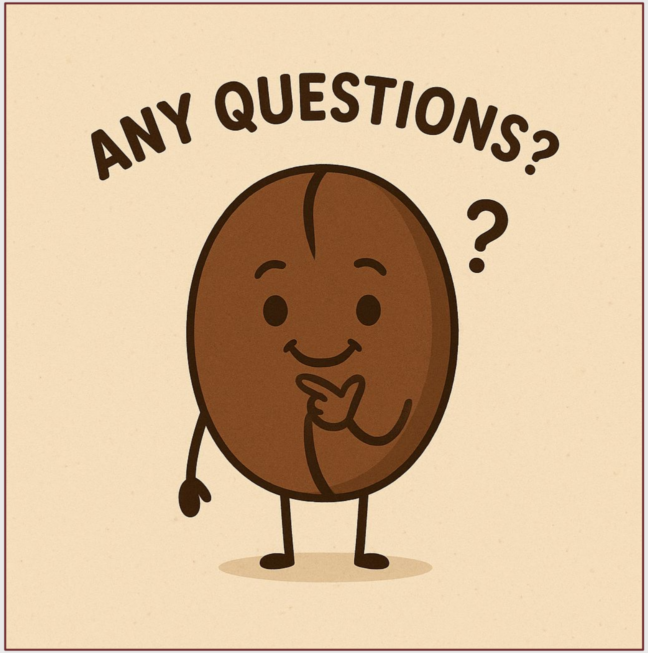
Best Accuracy: 0.653

Conclusions

Shiny App Link: https://maggiecrowner.shinyapps.io/coffee_quality_app/

- ➢ **Linear Regression:** Country of Origin and Species are the best predictors of Total Quality in a linear model.
- ➢ **Logistic Regression:** Sweetness is the strongest predictor of coffee species, with higher sweetness levels increasing the likelihood that a bean is Arabica.
- ➢ **K-Means:** Clustering highlights key differences between arabica and robusta beans in how various factors interact in different ways, including the opposing effects of age and altitude.
- ➢ **KNN:** KNN provides limited predictive power for estimating coffee farm altitude.
- ➢ **MLP:** MLP provides also provides limited predictive power for assessing market grade as a result of class imbalance.

THANK YOU


ANY QUESTIONS?
?