Red Wine Quality Analysis

Red Wine Break Down - Agenda

- Fixed Acidity, Volatile Acidity, Citric Acid – Rebekkah Alexander
- Residual Sugar & Alcohol – Danita Charles
- Free Sulfur Dioxide & Total Sulfur Dioxide – Bryan Thomas
- Density & pH – Aaron Wood
- Sulphates & Chlorides – Barry Maldonado

Fixed Acidity, Volatile Acidity, Citric Acid

Rebekkah Alexander

Acidity in Wine

Fixed Acidity

- Tartaric Acid
- Malic Acid
- Citric Acid
- Adds crispness, balance and refreshing character

Volatile Acidity

- Acetic Acid
- Ethyl Acetate
- Small amounts add complexity and bright, tangy notes
- Excessive amounts lead to sour, vinegary wine that smells like nail polish remover

Citric Acid

- Found naturally in grapes
- Adds tangy and slightly citrus flavors

Measured in Grams per Liter (g/L)

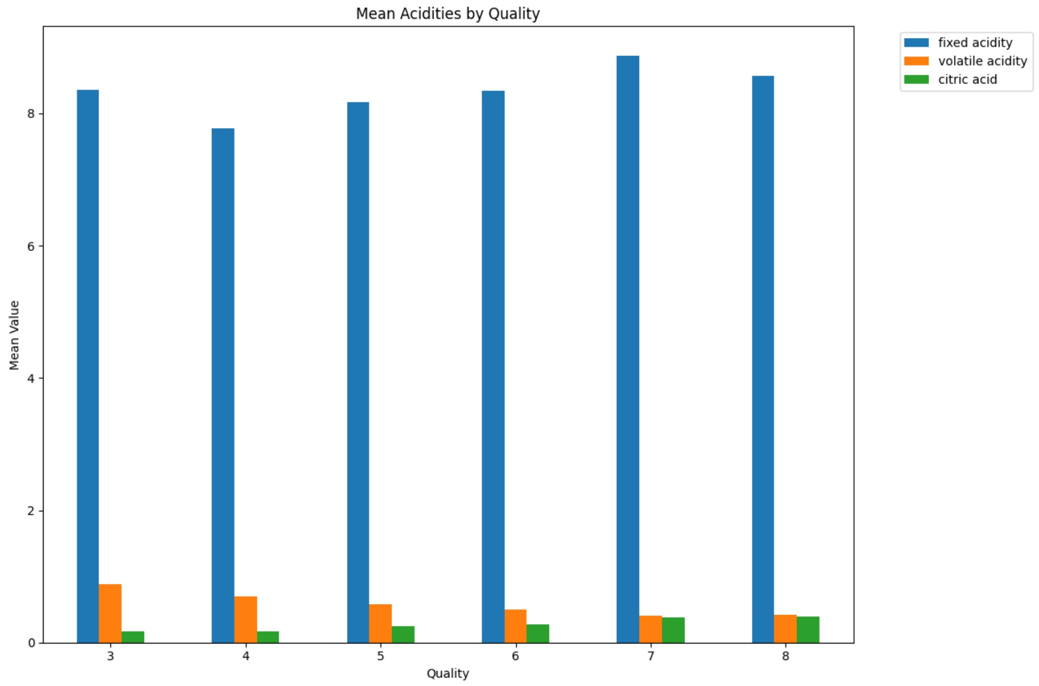
Analysis

Logistic Regression: 41% Accuracy

Decision Tree: 72% Accuracy

Random Forest: 78% Accuracy

SMOTE
Creates synthetic data to balance out class imbalances




```
# Initialize the Logistic Regression model
lr = LogisticRegression()

# Train the Logistic Regression model
lr.fit(X_train_scaled, y_train)

# Make predictions on the test set
y_pred_lr = lr.predict(X_test_scaled)

# Calculate the accuracy score
accuracy_lr = accuracy_score(y_test, y_pred_lr)
print(f"Logistic Regression Accuracy: {accuracy_lr:.2f}")
```

```
Logistic Regression Accuracy: 0.41
```

```
# Initialize the Random Forest Classifier
rf = RandomForestClassifier(random_state=42)

# Train the Random Forest Classifier
rf.fit(X_train_scaled, y_train)

# Make predictions on the test set
y_pred_rf = rf.predict(X_test_scaled)

# Calculate the accuracy score
accuracy_rf = accuracy_score(y_test, y_pred_rf)
print(f"Random Forest Accuracy: {accuracy_rf:.2f}")
```

```
Random Forest Accuracy: 0.78
```

```
# Initialize the Decision Tree Classifier
dt = DecisionTreeClassifier()

# Train the Decision Tree Classifier
dt.fit(X_train_scaled, y_train)

# Make predictions on the test set
y_pred_dt = dt.predict(X_test_scaled)

# Calculate the accuracy score
accuracy_dt = accuracy_score(y_test, y_pred_dt)
print(f"Decision Tree Accuracy: {accuracy_dt:.2f}")
```

```
Decision Tree Accuracy: 0.72
```

```
Optimal Acid Levels for Low-Quality Wine:
fixed acidity       8.011342
volatile acidity    0.815877
citric acid         0.144134
```

```
Optimal Acid Levels for Medium-Quality Wine:
fixed acidity       8.357173
volatile acidity    0.492967
citric acid         0.296870
```

```
Optimal Acid Levels for High-Quality Wine:
fixed acidity       8.295652
volatile acidity    0.416990
citric acid         0.371611
```

Synthetic Testing

- Creates 1000 points of synthetic data based off the data provided
- Tests that data with the trained Random Forest model

```
# Generate synthetic acid combinations
fixed_acidity = np.linspace(X['fixed acidity'].min(), X['fixed acidity'].max(), 100)
volatile_acidity = np.linspace(X['volatile acidity'].min(), X['volatile acidity'].max(), 100)
citric_acid = np.linspace(X['citric acid'].min(), X['citric acid'].max(), 100)

# Create a grid of synthetic data
synthetic_data = pd.DataFrame({
    'fixed acidity': np.random.choice(fixed_acidity, 1000),
    'volatile acidity': np.random.choice(volatile_acidity, 1000),
    'citric acid': np.random.choice(citric_acid, 1000)
})

# Predict quality for synthetic data
synthetic_data_scaled = scaler.transform(synthetic_data)
synthetic_predictions = rf.predict(synthetic_data_scaled)

# Filter high-quality predictions
high_quality_synthetic = synthetic_data[synthetic_predictions == 8]

# Find the mean acid levels for "perfect" high-quality wine
print("Perfect Acid Blend:")
print(high_quality_synthetic.mean())
```

```
Perfect Acid Blend:
fixed acidity       8.575926
volatile acidity    0.547677
citric acid         0.531145
dtype: float64
```

Every time the code is ran; it will output slightly different results because it works with synthetic data based off the data it is provided.

Residual Sugar & Alcohol

Danita Charles

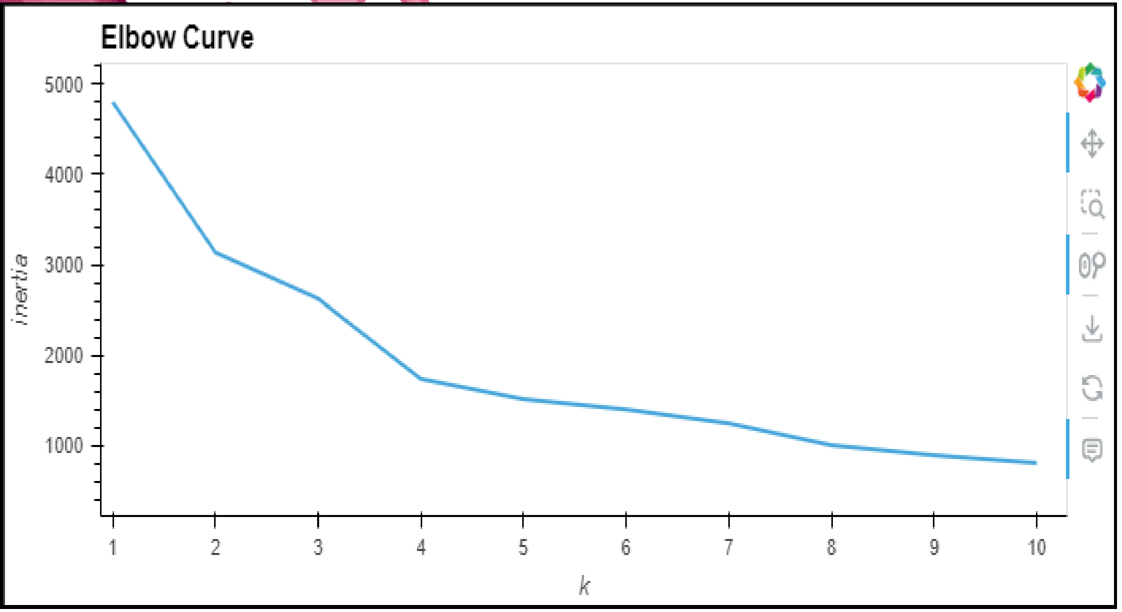
Wine Residual Sugar & Alcohol Content

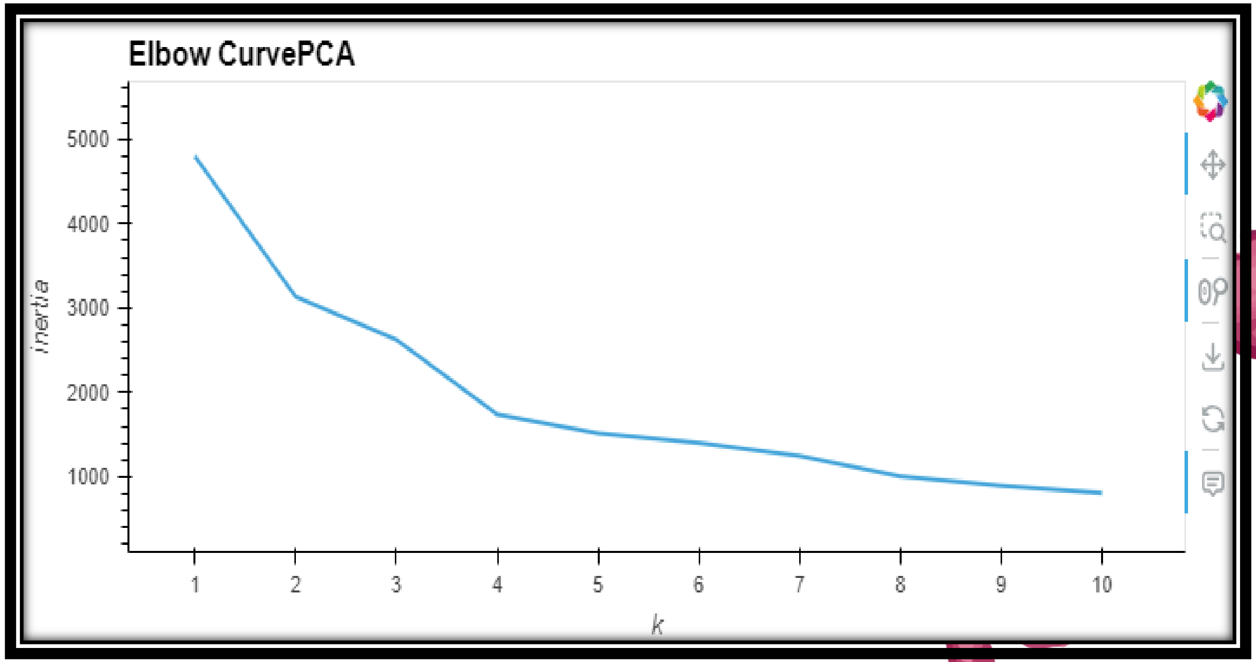
Do the physicochemical properties residual sugar and alcohol level contribute to specific quality scores of red wine?

- Residual sugar levels range 0.9(low)-15.5(high)
- Alcohol level range 8.4(low)- 14.9(high)
- Quality level range 3(low)-8(high)

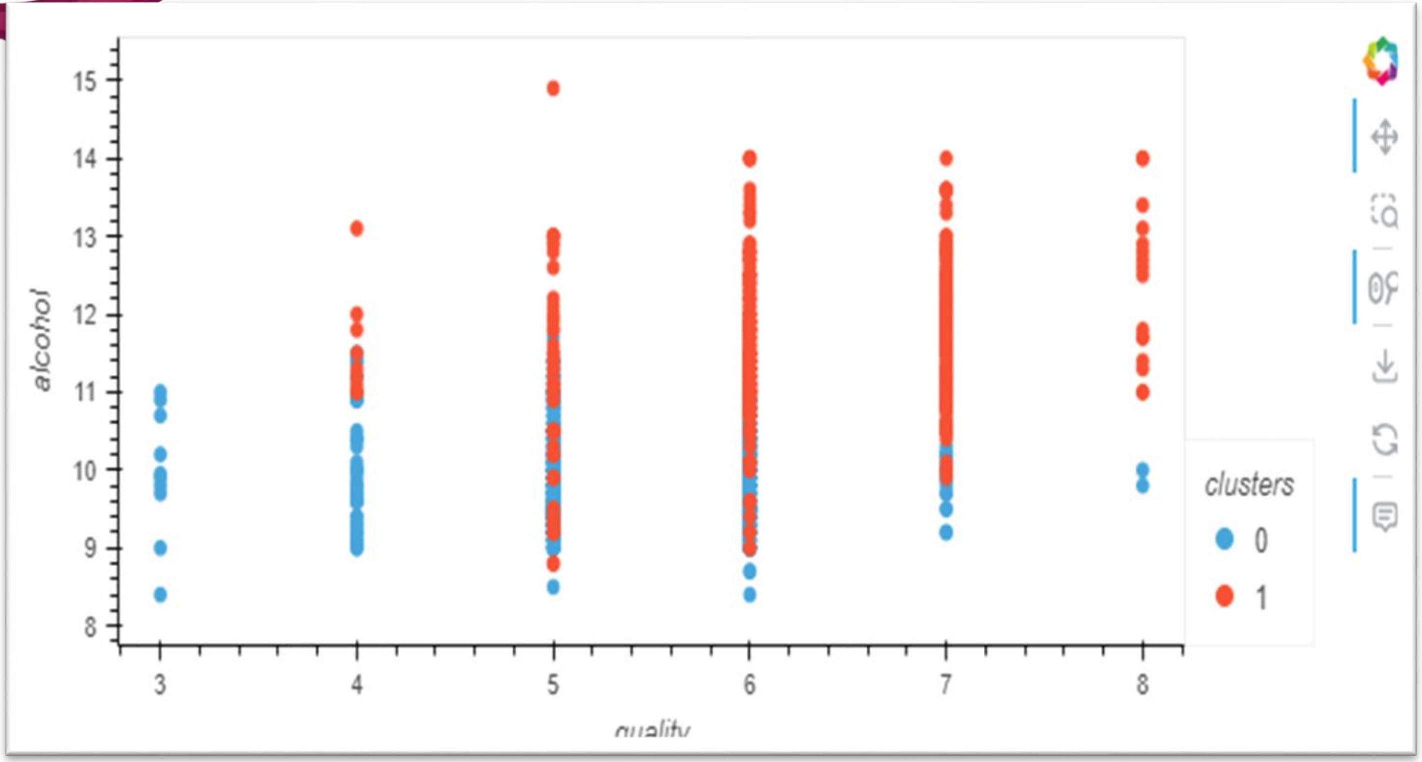
Unsupervised Cluster Analysis

- **Kmeans**
- **Optimized with PCA**



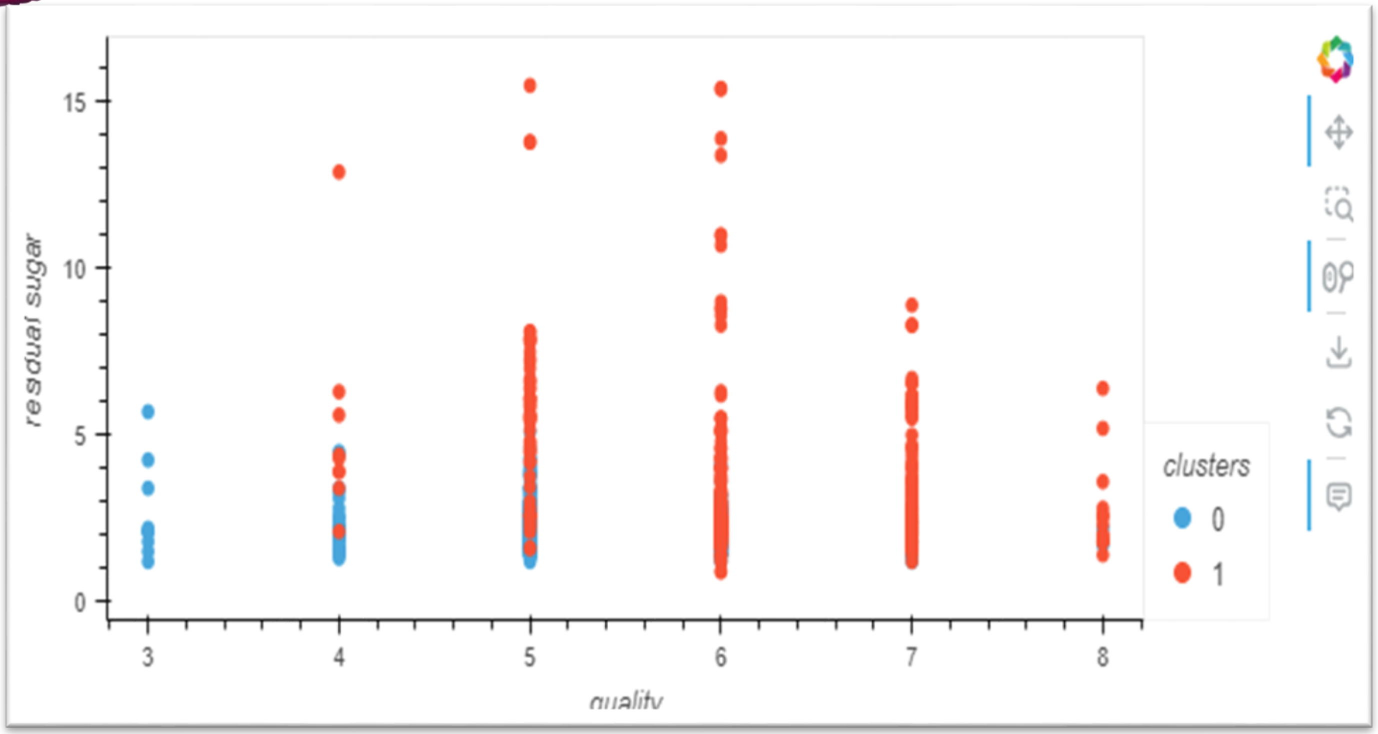


Alcohol Content



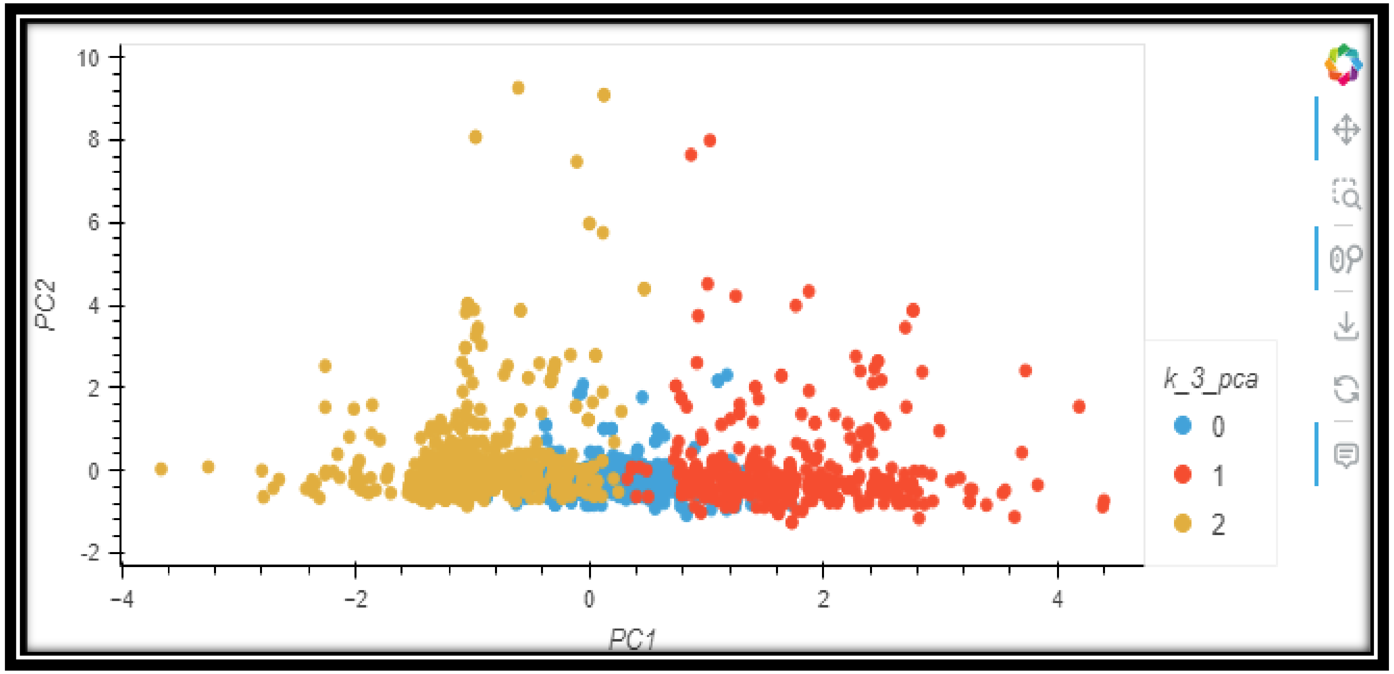
Alcohol content levels 9% -12% are more associated with red wine quality level 6.

Residual Sugar



Residual sugar levels 1-5 are more associated with quality level 6

Residual Sugar & Alcohol Content Combined



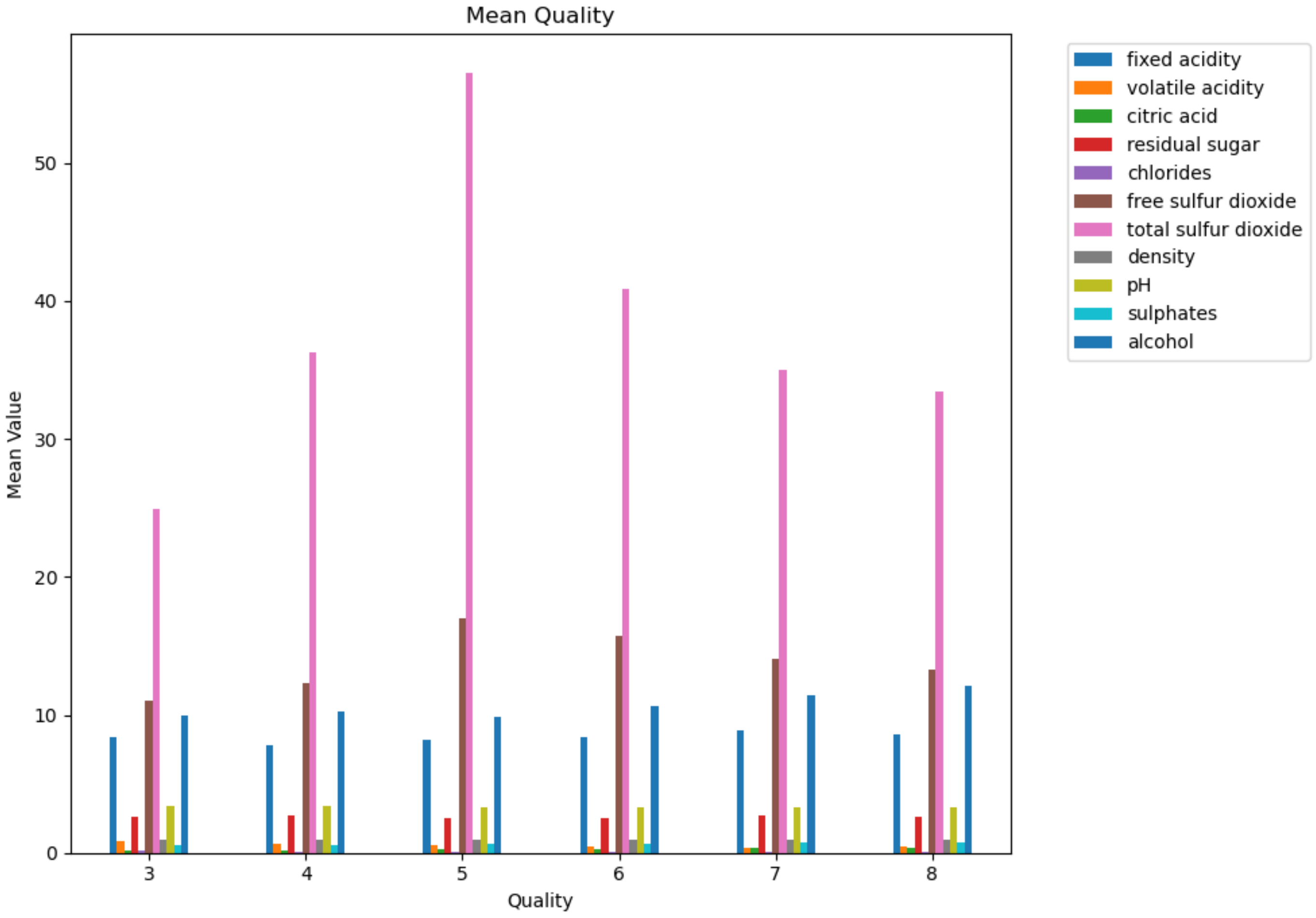
Optimized Cluster Analysis using Principal Component Analysis (PCA)

Mid levels residual sugar 6-7 combined with alcohol content (10-12) associated with high quality 6-7.

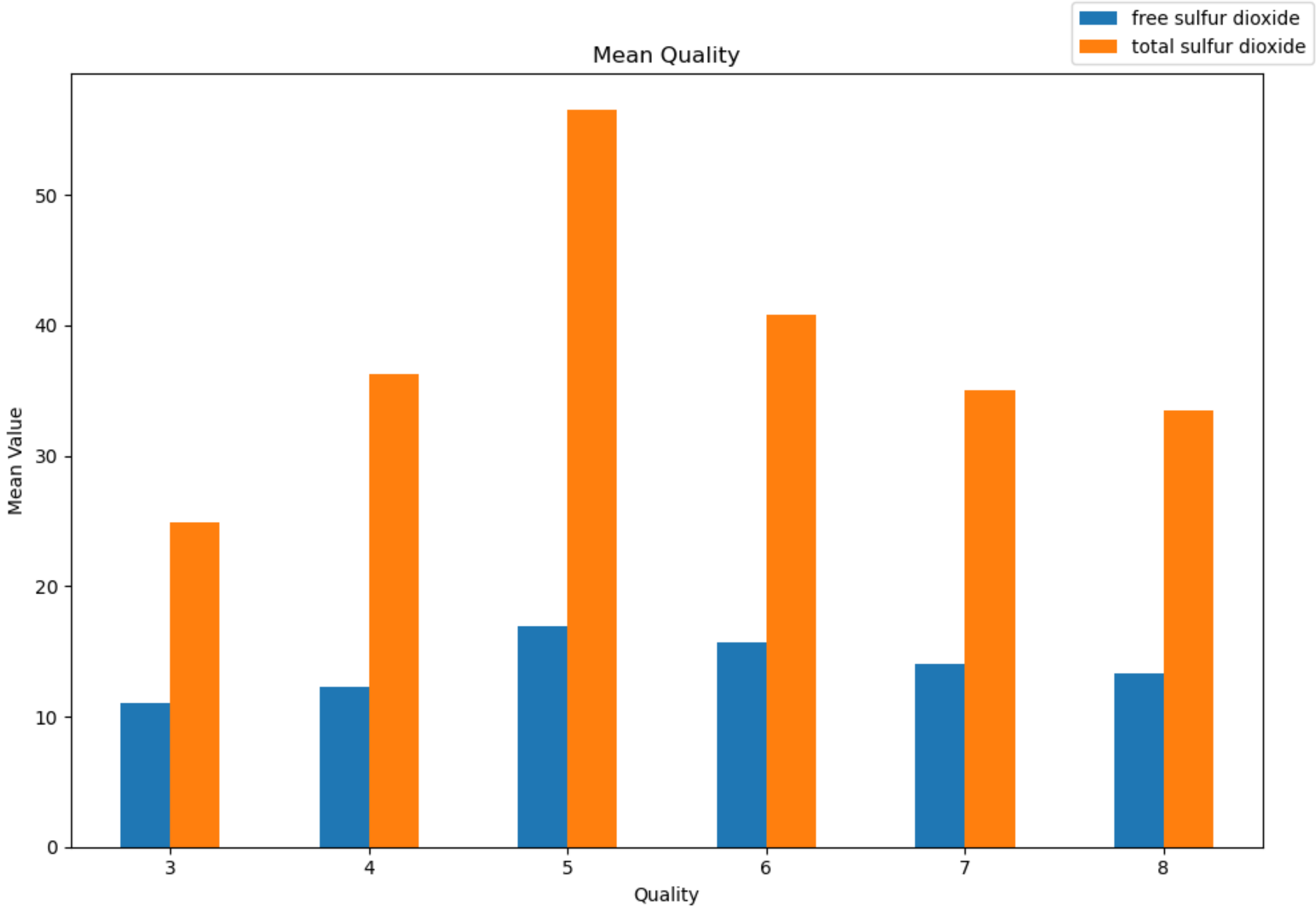
Free Sulfur Dioxide & Total Sulfur Dioxide

Bryan Thomas

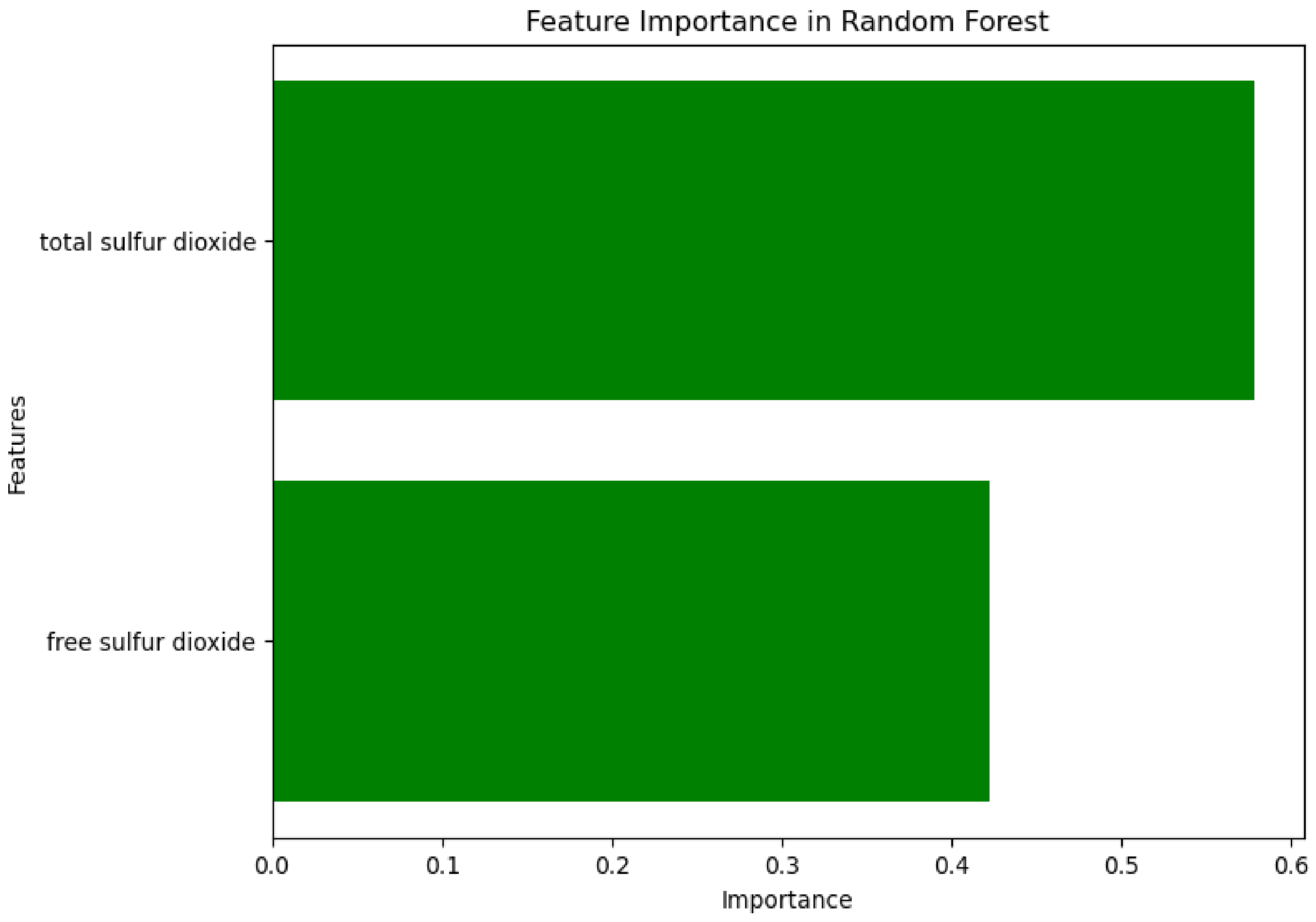
Wine Free & Total Sulfur Dioxides



Wine Sulfur Dioxide Content



Sulfur Dioxide Importance



Density & pH

Aaron Wood

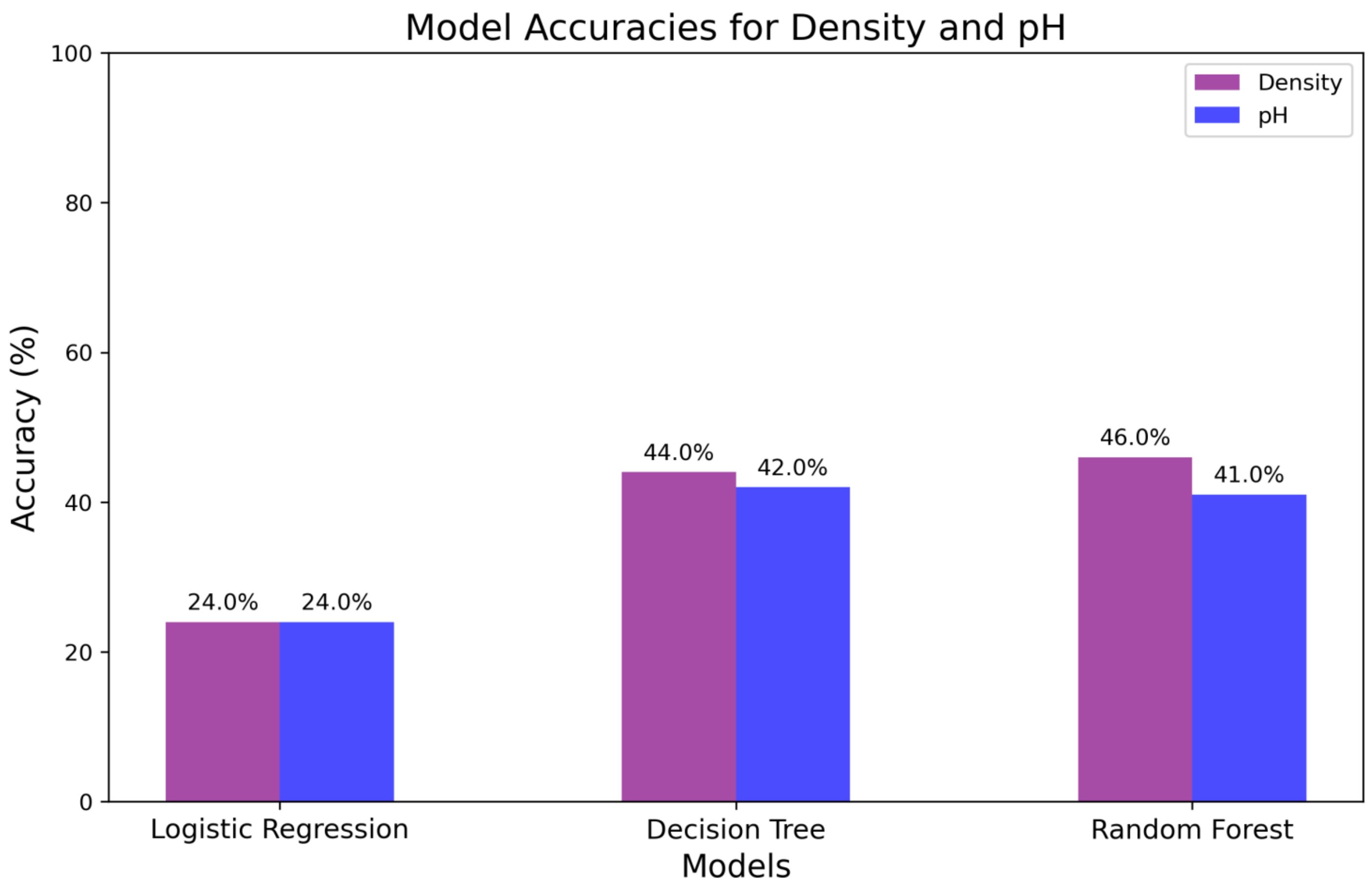
Wine Density & pH

Aaron Wood

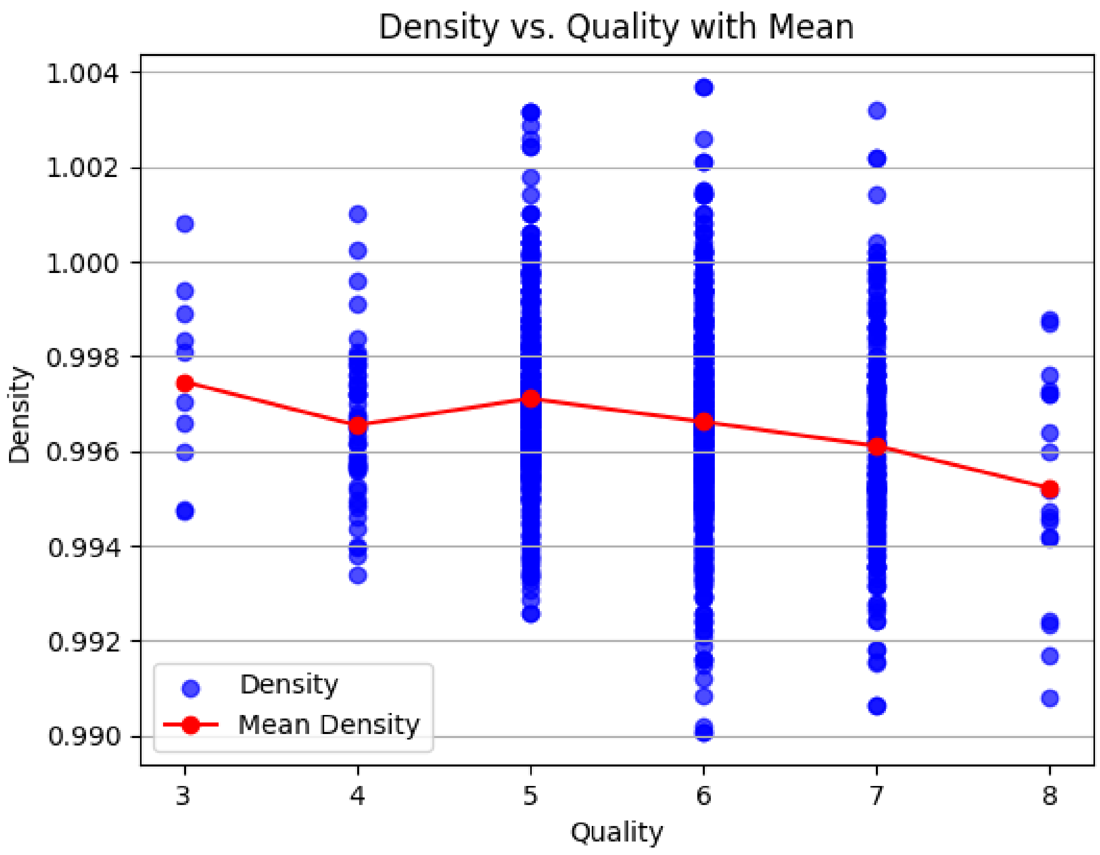
Wine density refers to the amount of dissolved solids (e.g. – sugars) dissolved in the wine. Dryer wines have lower densities than sweeter wines.

Wine pH refers to the acidity of the wine. Wine pH will impact the taste, color, and aging of the wine.

Accuracy Issues

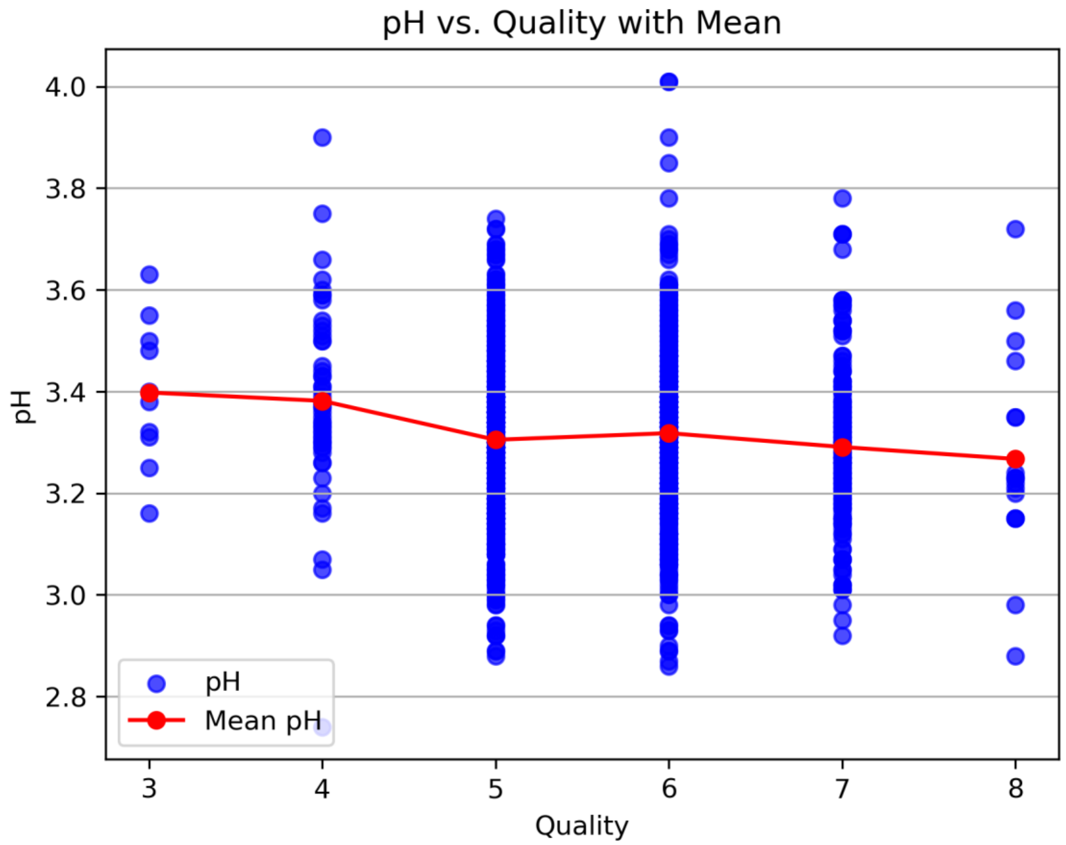


Wine Density



- Modeled density for quality wine is .9948 g/mL.
- Logistic Regression accuracy is 24%.
- Decision Tree accuracy is 44%.
- Random Forest accuracy is 46%.
- Wine density is NOT an appropriate measure for predicting wine quality. Sufficient accuracy could not be obtained by the models targeting standalone physiochemical density and the mean may be just as useful within the density data

Wine pH

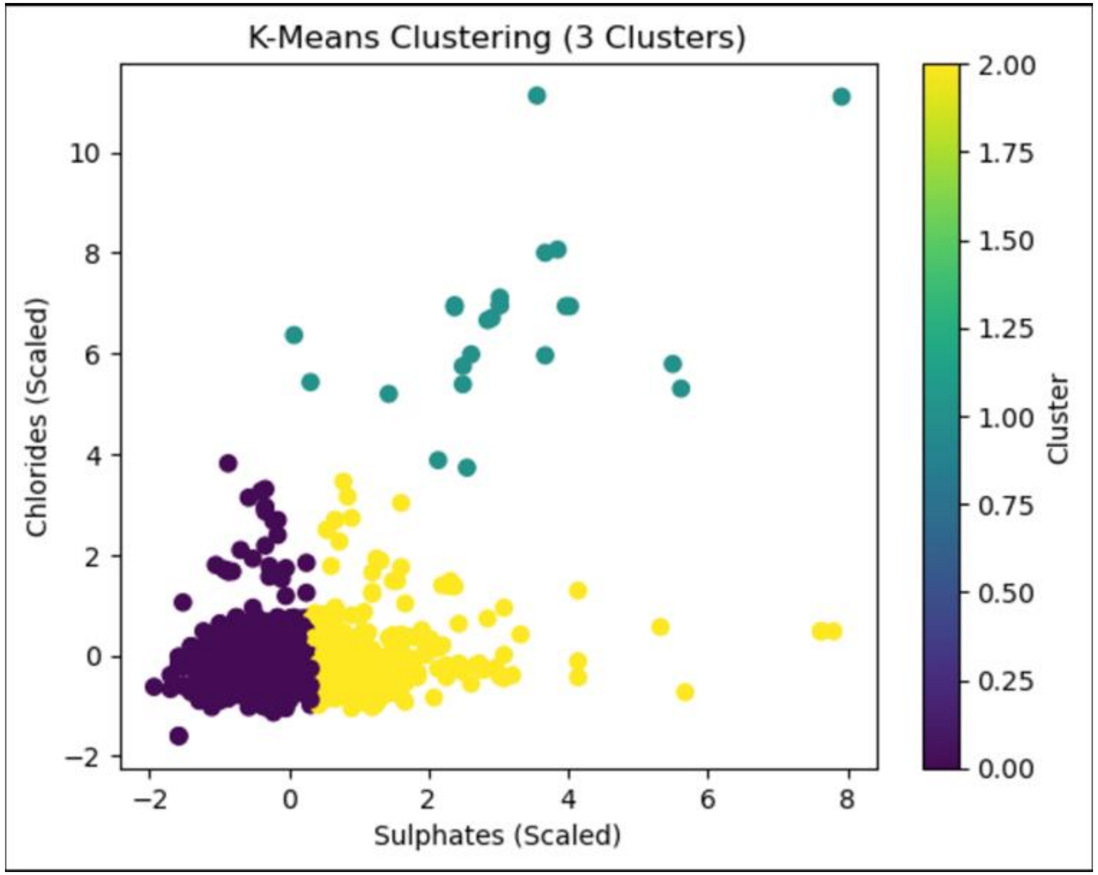


- Modeled pH for quality wine is 3.25.
- Logistic Regression accuracy is 24%.
- Decision Tree accuracy is 42%.
- Random Forest accuracy is 41%.
- Wine pH is NOT an appropriate measure for predicting wine quality. Sufficient accuracy could not be obtained by the models targeting standalone physiochemical density and the mean may be just as useful within the pH data.

Sulfates & Chlorides

Barry Maldonado

Wine pH



Key Findings:

1. Clustering Results:

- Cluster 1: Highest average quality (6.06), associated with moderate Sulphates and Chlorides.
- Cluster 0: Intermediate quality (5.48), balanced feature values.
- Cluster 2: Lowest average quality (5.29), driven by high Sulphates and very high Chlorides.

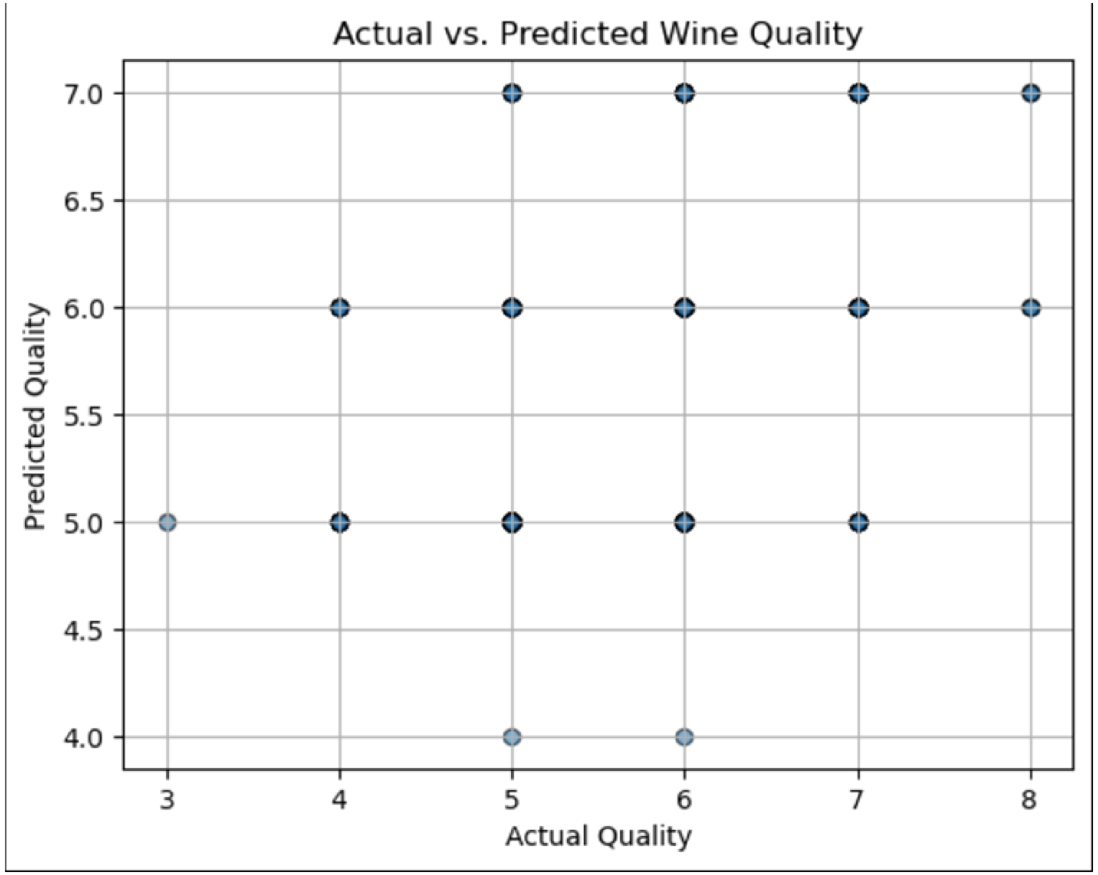
2. Feature Importance:

- Sulphates contributed significantly to the prediction of wine quality, positively impacting quality in moderate ranges.
- Chlorides had a smaller influence, negatively impacting quality when values were high.

3. Supervised Model Performance:

- Initial Model Accuracy: Approximately X% (replace with actual score).
- Tuned Model Accuracy: Improved to Y% (replace with actual score).
- Classification Report: Showed consistent performance across wine quality categories, with room for improvement in extreme cases.

Wine pH

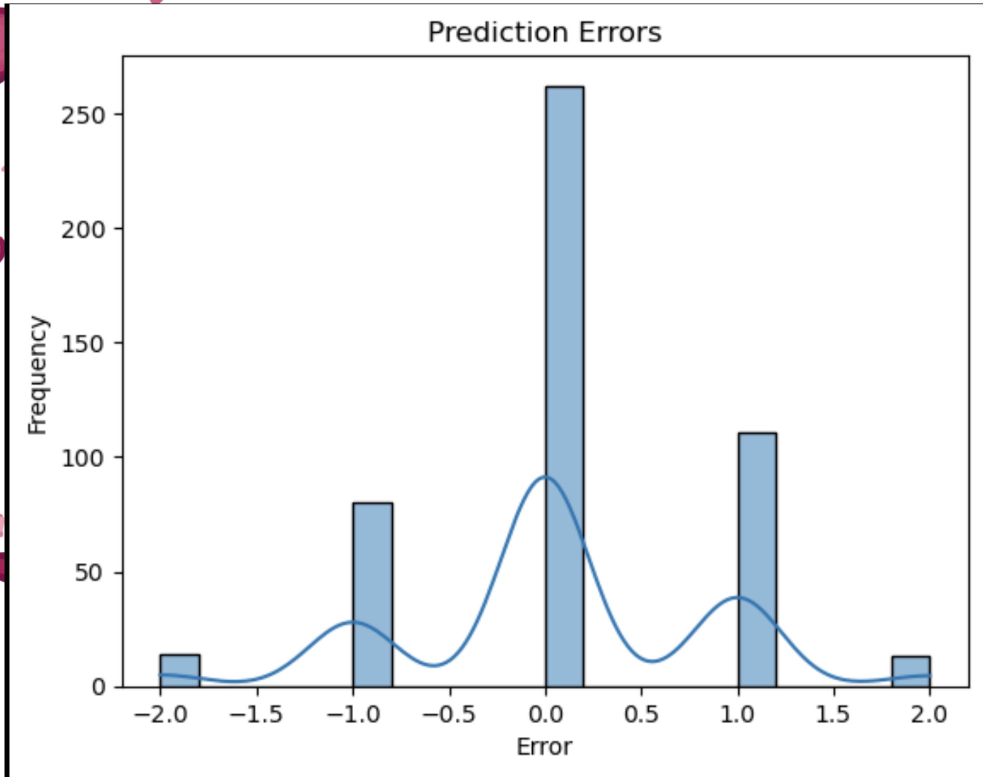


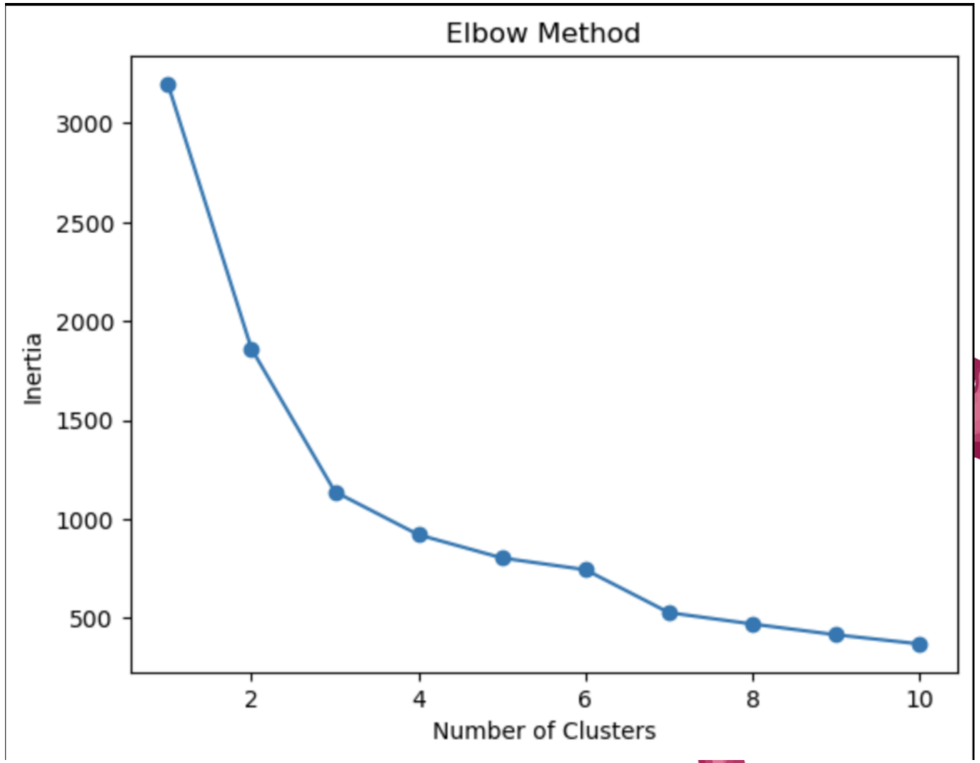
Recommendations For Winemakers:
- Maintain moderate Sulphates and minimize Chlorides to enhance wine quality.

Future Work:
- Incorporate additional physicochemical features (e.g., acidity, alcohol content) for a comprehensive analysis.
- Explore DBSCAN for more flexible clustering to detect noise or outliers.

Model Refinement:
- Continue hyperparameter tuning and consider ensemble methods for improved prediction accuracy

Wine Sulfates & Chlorides





Questions?

Thank you!