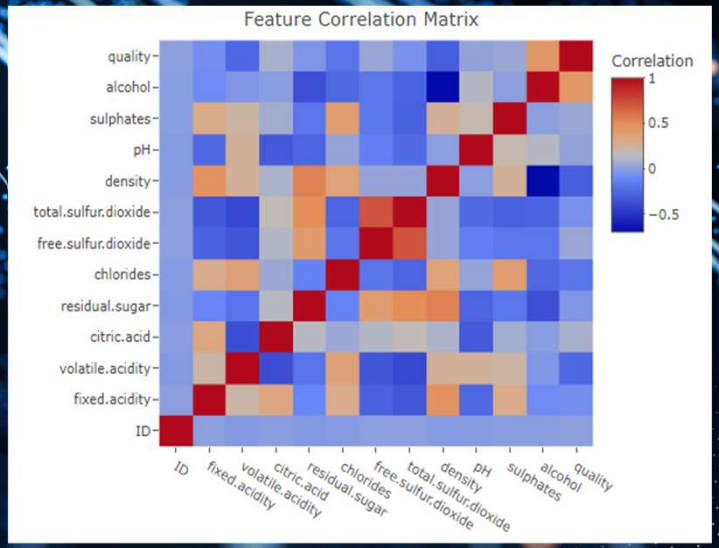
# Wine Quality

A predictive analysis of factors contributing to high-quality wines.

Gerden Clark
Solomon Mathew
Vanessa Nkongolo

# EXPLORATORY DATA ANALYSIS

- Three datasets : training, testing and supplementary wine types and locations.
- Training set contains 13 columns and 5463 rows; Test set has 12 columns, 1034 rows; Supplementary set has 3 columns, 6497 rows.
- 152 Missing values for type in the train data and 23 missing values for type in the test data were imputed using the most frequent type found in the data.
- Data was split using a 75/25 train-test split for model evaluation.



# EXECUTIVE SUMMARY

01

**Model Design**

Polynomial Regression & Random Forest Models

02

**Evaluation**

Model Performance assessed using Mean Absolute Error (MAE)

03

**Key findings**

Top 3 features related to Wine Quality

04

**Conclusion**

Recommendations and Conclusion.

# Analysis

## Regression Model Evaluation

Model assessed using Mean Absolute Error (MAE).

- Developed a polynomial linear regression model, that performed well for predicting wine quality with a mean absolute error (MAE) of 0.4944648.
- The average predicted quality grade for the 1034 wines in the test set is 5.815.

```
# Predictions and MAE
predictions = predict(fit, newdata = test_data)
actual = test_data$quality
mae = mean(abs(actual - predictions))
cat("Polynomial Regression Mean Absolute Error on Training Data:", mae, "\n")
```

```
## Polynomial Regression Mean Absolute Error on Training Data: 0.4944648
```

# Analysis

## Random Forest Model Evaluation

Model assessed using Mean Absolute Error (MAE).

- The random forest model  provided an effective tool for predicting wine quality with low error rates. With an MAE 0.1773722
- The average predicted quality grade for the 1034 wines in the test set is 5.805.

```
# Evaluate model on training data
train_preds = predict(rf_model, x_train)
mae_train = mae(as.numeric(y_train), as.numeric(train_preds))
cat("Random Forest Mean Absolute Error on Training Data:", mae_train, "\n")
```

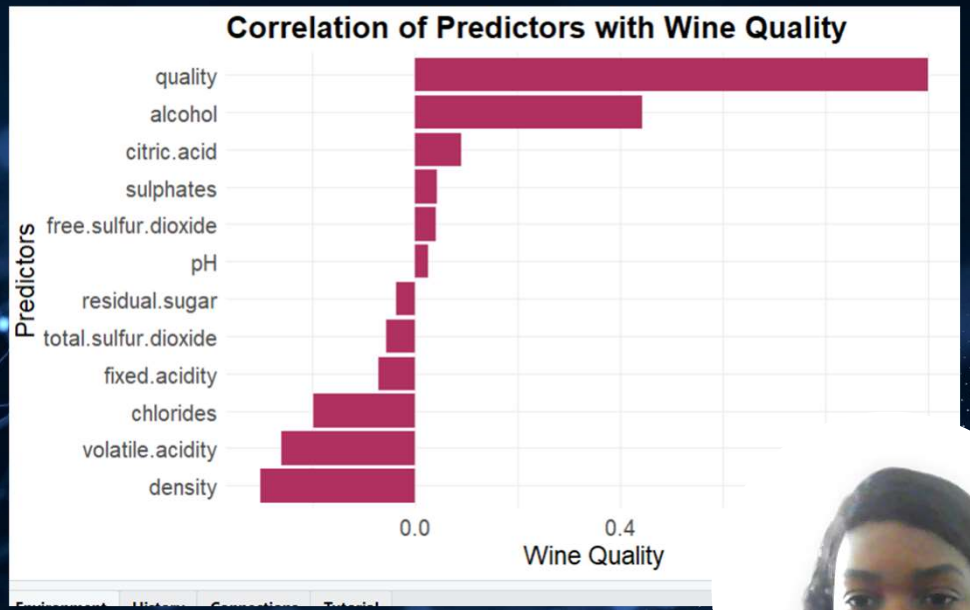
```
## Random Forest Mean Absolute Error on Training Data: 0.1773722
```

# Analysis

## Key Findings

Top 3 variables with the highest correlation to wine quality are:

- Alcohol
- Density
- Volatile acidity




# Rshiny App

## Wine Quality Analysis Dashboard

**Upload CSV File:**

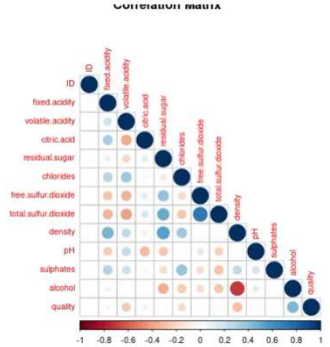
Browse... Wine Train.csv

Upload complete

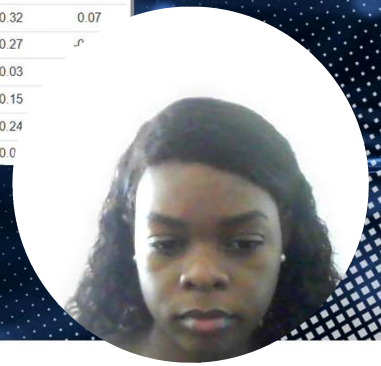
**Select Visualization Type:**

Correlation Plot

Refresh Plot




| ID | fixed.acidity | volatile.acidity | citric.acid | residual.sugar | chlorides | free.sulfur.dioxide | total.sulfur.dioxide | density | pH | sulphates |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.00 | 0.00 | -0.02 | -0.00 | -0.02 | -0.01 | 0.01 | 0.00 | -0.01 | -0.01 | -0.00 |
| 0.00 | 1.00 | 0.22 | 0.33 | -0.11 | 0.30 | -0.28 | -0.33 | 0.47 | -0.25 | 0.30 |
| -0.02 | 0.22 | 1.00 | -0.37 | -0.20 | 0.37 | -0.35 | -0.40 | 0.26 | 0.26 | 0.22 |
| -0.00 | 0.33 | -0.37 | 1.00 | 0.13 | 0.05 | 0.12 | 0.18 | 0.10 | -0.32 | 0.07 |
| -0.02 | -0.11 | -0.20 | 0.13 | 1.00 | -0.13 | 0.41 | 0.50 | 0.55 | -0.27 | -0 |
| -0.01 | 0.30 | 0.37 | 0.05 | -0.13 | 1.00 | -0.19 | -0.27 | 0.36 | 0.03 | |
| 0.01 | -0.28 | -0.35 | 0.12 | 0.41 | -0.19 | 1.00 | 0.72 | 0.03 | -0.15 | |
| 0.00 | -0.33 | -0.40 | 0.18 | 0.50 | -0.27 | 0.72 | 1.00 | 0.03 | -0.24 | |
| -0.01 | 0.47 | 0.26 | 0.10 | 0.55 | 0.36 | 0.03 | 0.03 | 1.00 | 0.0 | |



# Model Output

- Mean absolute error values of models:
  Linear Regression = 0.494 MAE
  Random Forest = 0.177 MAE
- Random forest is the higher performing model and will be used for the predictive model.
- The data output requested is a CSV file called “New_Wine_Test_Predictions” and is housed in the GitHub repository.

| | ID | Random_Forest_Predictions | Polynomial_Linear_Regression_Predictions |
|---|---|---|---|
| 2 | 5464 | 6.5567 | 5.88460677112841 |
| 3 | 5465 | 5.33906666666667 | 5.38255355366818 |
| 4 | 5466 | 5.11376666666667 | 5.05423755567365 |
| 5 | 5467 | 6.17806666666667 | 6.07182030592836 |
| 6 | 5468 | 5.44253333333333 | 5.25800685864326 |
| 7 | 5469 | 6.09726666666667 | 6.13186668971442 |
| 8 | 5470 | 6.3221 | 6.69539817777087 |
| 9 | 5471 | 6.19196666666667 | 6.52039525325389 |
| 10 | 5472 | 5.98883333333334 | 6.2830171403907 |
| 11 | 5473 | 6.08776666666667 | 6.010515166125 |
| 12 | 5474 | 5.9252 | 6.11834833432368 |
| 13 | 5475 | 6.34556666666667 | 6.37502439406757 |
| 14 | 5476 | 5.12806666666667 | 5.08621051861781 |
| 15 | 5477 | 6.8523 | 6.67585460360067 |
| 16 | 5478 | 5.2301 | 5.32614737987987 |
| 17 | 5479 | 5.74933333333333 | 5.94542797702426 |

# Conclusion

- Based on the importance of alcohol content and citric acid in wine in relation to quality, a primary goal for Robert Renzoni Vineyards would be to aim for higher alcohol content and more citric acid.
- Wine type and location are not major determining factors for quality.
- The random forest model is more accurate than a polynomial regression model, with a measured mean absolute error (MAE) of 0.177, and is the recommended model for predicting wine quality grades.

# THANK YOU

GERDEN CLARK: gerdenc@mail.smu.edu
SOLOMON MATHEW: stmatthew@mail.smu.edu
VANESSA NKONGOLO : vnkongolo@mail.smu.edu

# Appendix

Published Rmarkdown
GitHub Repository
Rshiny App