Predicting Wine Quality Based on Physicochemical Properties

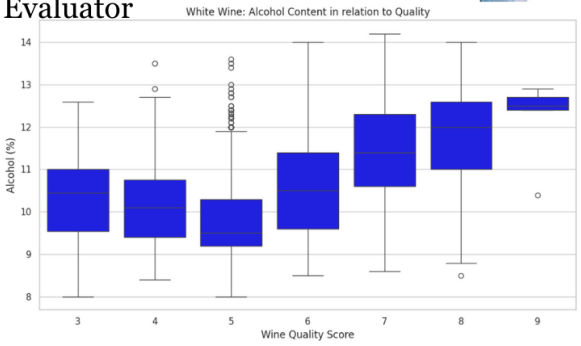
Group 5, Arnav Chawla
Guadalupe Ramirez Lara

Learn More: GitHub

Used pySpark for Vector Assembler, Linear Regression, and Regression Evaluator

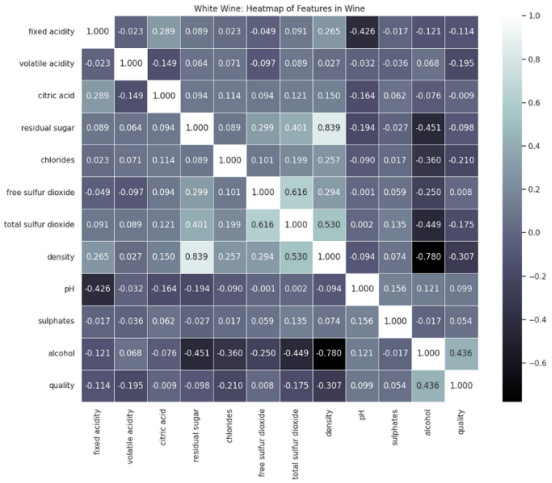
Linear Regression Model

- Feature Coefficients used to interpret variable's affect on wine quality. Positive coefficients increase the predicted value, while negative one's decrease it.



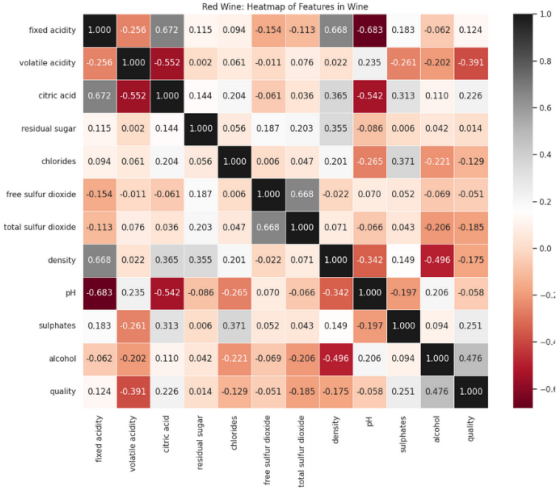
Analyzing the Results:

- Density(-) & Volatile acidity(-)are the most influential features
- Alcohol & sulphates(+) are correlated with quality
- Chlorides(-) have an affect in red wine but very little in white wine



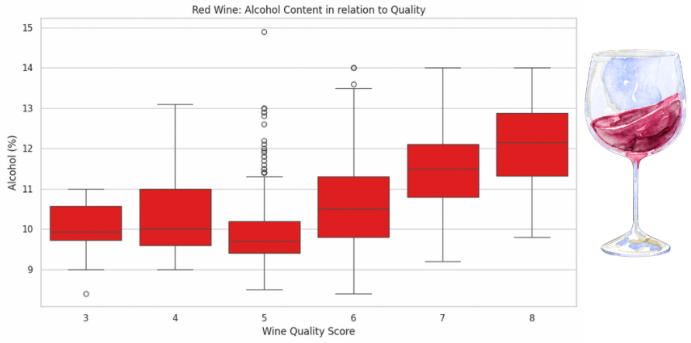
White Wine:

- Correlation between alcohol and quality: 0.436
- The RMSE value is = 0.7581



Red Wine:

- Correlation between alcohol and quality: 0.476
- The RMSE value is = 0.6747



- The heatmaps show how two different variables affect the wine quality (shows how linearly related two variables are)
- The box plots explore how alcohol content varies with wine quality (shows how strong of a predictor it is)

Moving Forward:

Since our RMSE value is significantly high for both red and white wine, we could consider altering the data quality and quantity, or trying different models to predict wine quality.