August 10th, 2021

The University of Texas at Austin
McCombs School of Business

Wine Quality Data Analysis Project

What makes a wine "high quality"?

PRESENTED BY:
Rohan Garg, Yanqi Liang, Junsu Kim, & Christian Alfonso

OUR PURPOSE

- Red Wine Quality dataset
  - 11 predictors & wine quality
- Determine the chemical components of a high-quality red wine

WHAT WE DID

- Exploratory Data Analysis
  - Categorization of the response variable
  - Boxplots for predictors
  - Pivot tables
  - Bar Charts
- Model Building
  - Multiple Linear Regression
  - Logistic Regression

EXPLORATORY DATA ANALYSIS

CATEGORIZING QUALITY

- Divide "Quality" into three groups
  1. Low (46.5%)
  2. Medium (40%)
  3. High (13.5%)

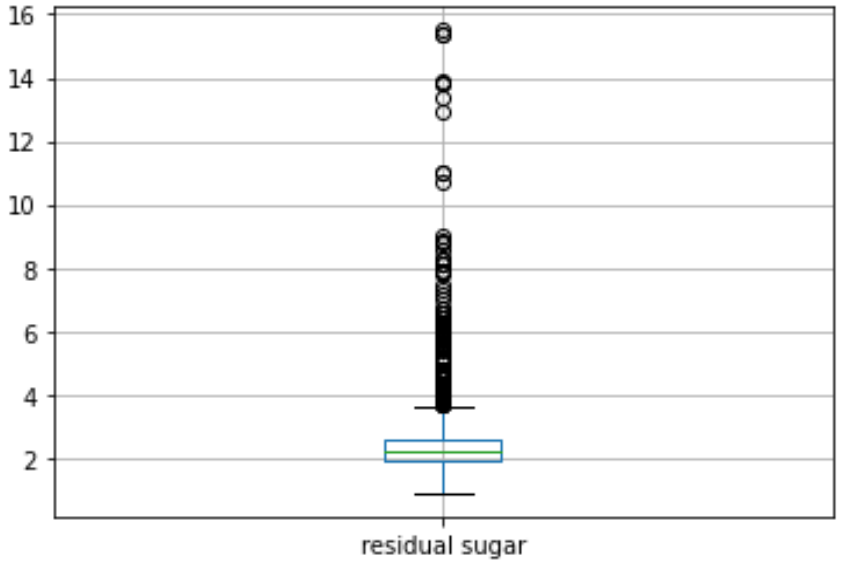
Original Count for Each Wine Quality

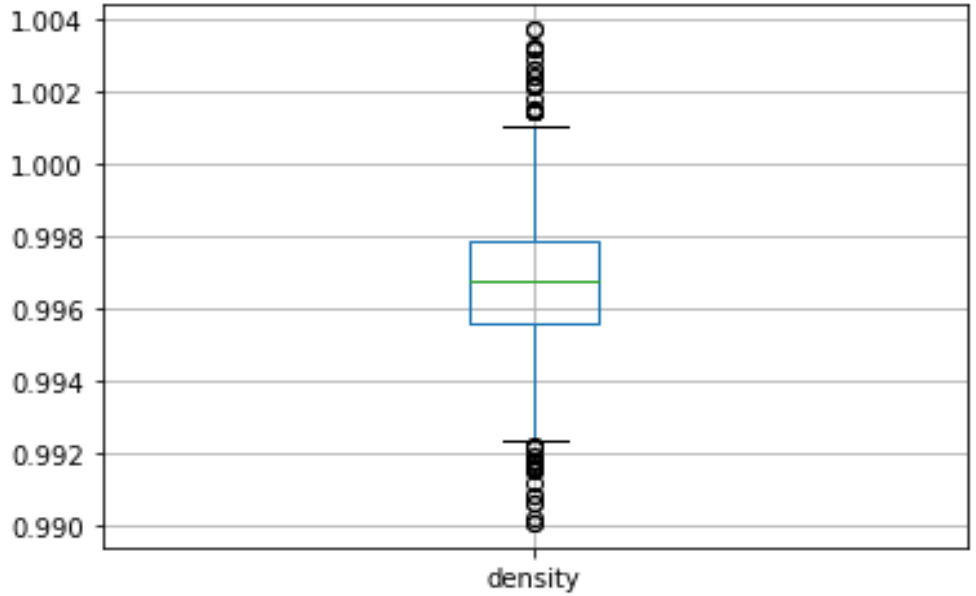
```
5    681
6    638
7    199
4     53
8     18
3     10
Name: quality, dtype: int64
```

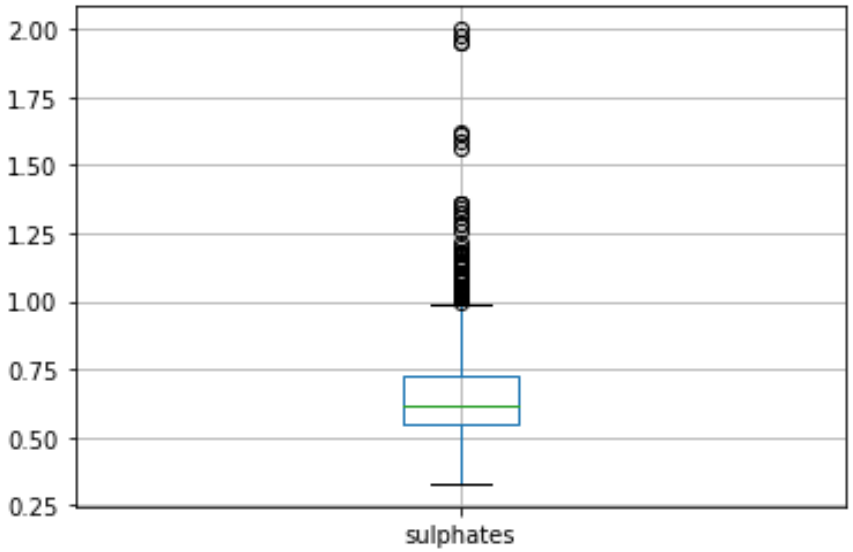
Count for the 3 Groups of Wine Quality

```
Low       744
Medium    638
High      217
Name: Quality Quartile, dtype: int64
```

PREDICTOR BOXPLOTS

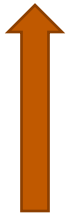


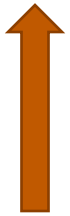




MEDIAN PREDICTOR VALUES FOR LOW/MEDIUM/HIGH QUALITIES

Quality Quartile	alcohol	chlorides	citric acid	density	fixed acidity	free sulfur dioxide	pH	quality	residual sugar	sulphates	total sulfur dioxide	volatile acidity
Low	9.7	0.081	0.22	0.996935	7.8	14.0	3.31	5	2.2	0.58	45.0	0.59
Medium	10.5	0.078	0.26	0.996560	7.9	14.0	3.32	6	2.2	0.64	35.0	0.49
High	11.6	0.073	0.40	0.995720	8.7	11.0	3.27	7	2.3	0.74	27.0	0.37

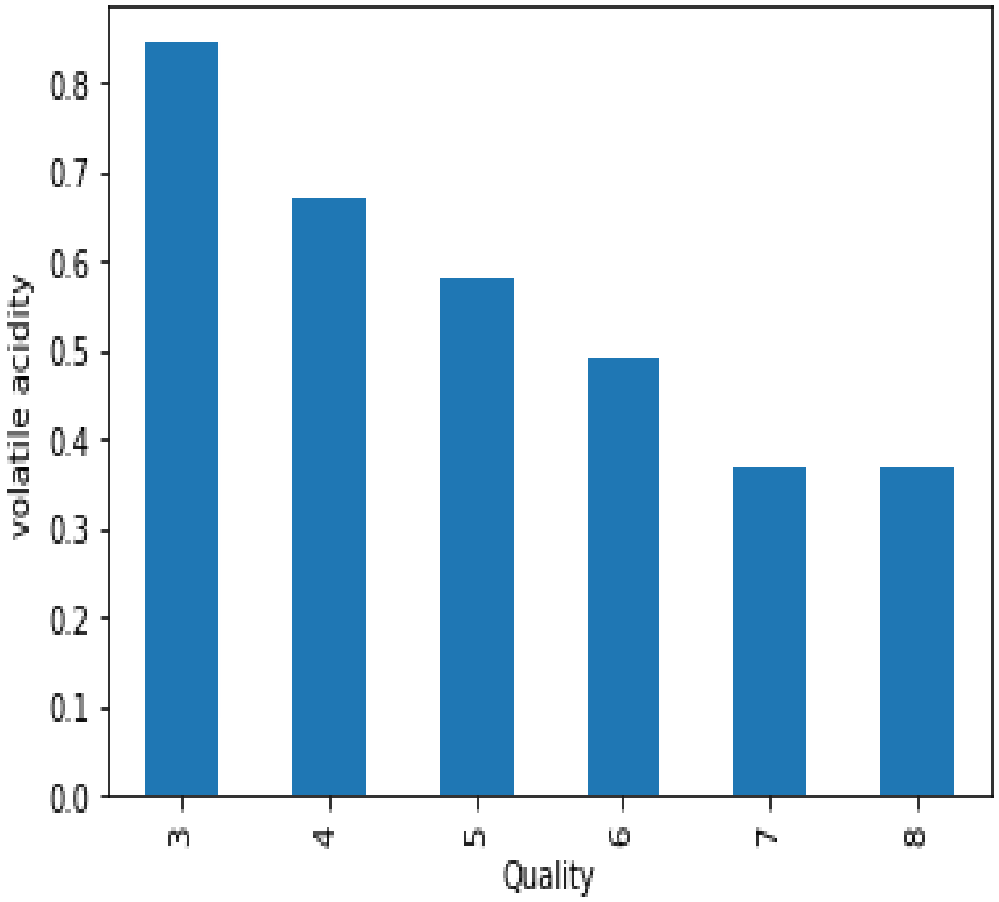




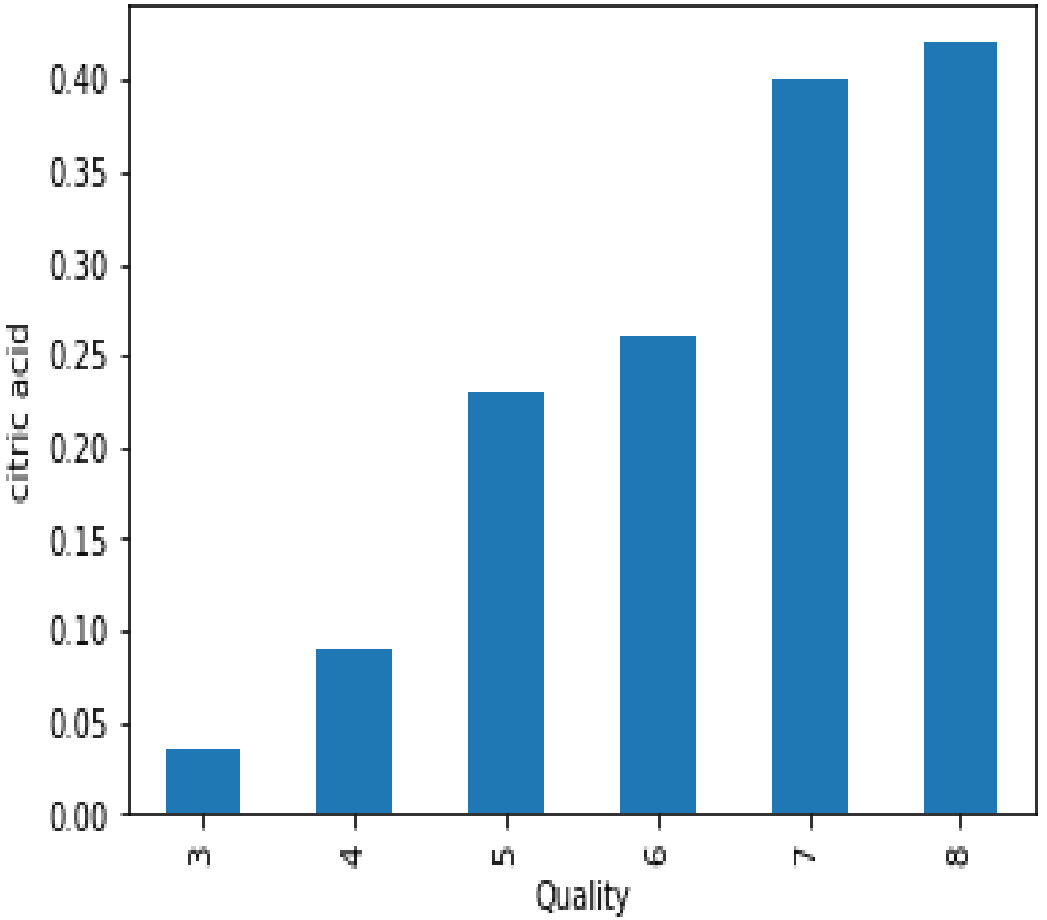
MEDIAN PREDICTOR VALUES FOR EACH NUMERICAL QUALITY

quality	alcohol	chlorides	citric acid	density	fixed acidity	free sulfur dioxide	pH	residual sugar	sulphates	total sulfur dioxide	volatile acidity
3	9.925	0.0905	0.035	0.997565	7.50	6.0	3.39	2.1	0.545	15.0	0.845
4	10.000	0.0800	0.090	0.996500	7.50	11.0	3.37	2.1	0.560	26.0	0.670
5	9.700	0.0810	0.230	0.997000	7.80	15.0	3.30	2.2	0.580	47.0	0.580
6	10.500	0.0780	0.260	0.996560	7.90	14.0	3.32	2.2	0.640	35.0	0.490
7	11.500	0.0730	0.400	0.995770	8.80	11.0	3.28	2.3	0.740	27.0	0.370
8	12.150	0.0705	0.420	0.994940	8.25	7.5	3.23	2.1	0.740	21.5	0.370

BAR CHARTS: STRONG CORRELATION

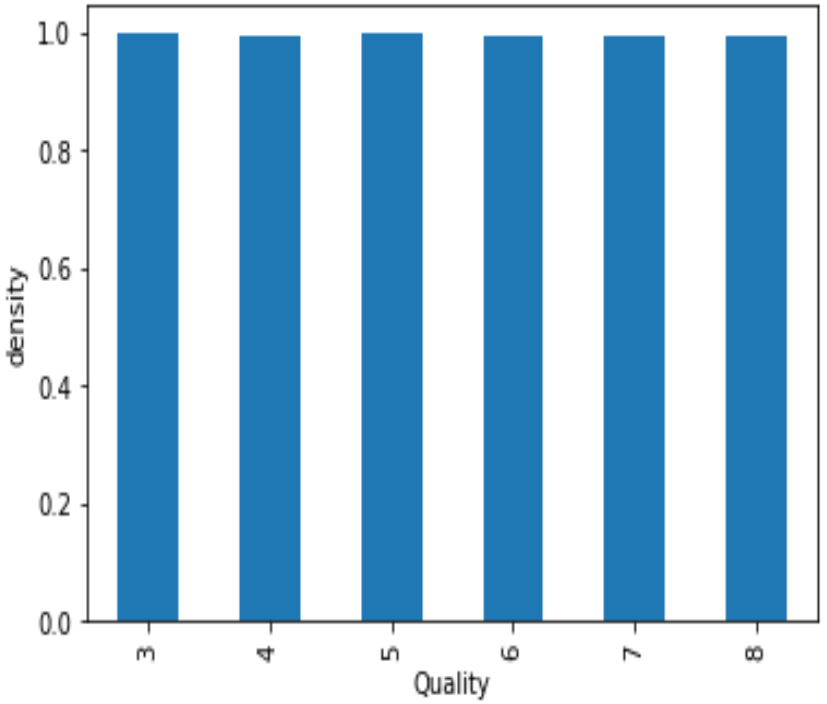


Volatile Acidity vs Quality

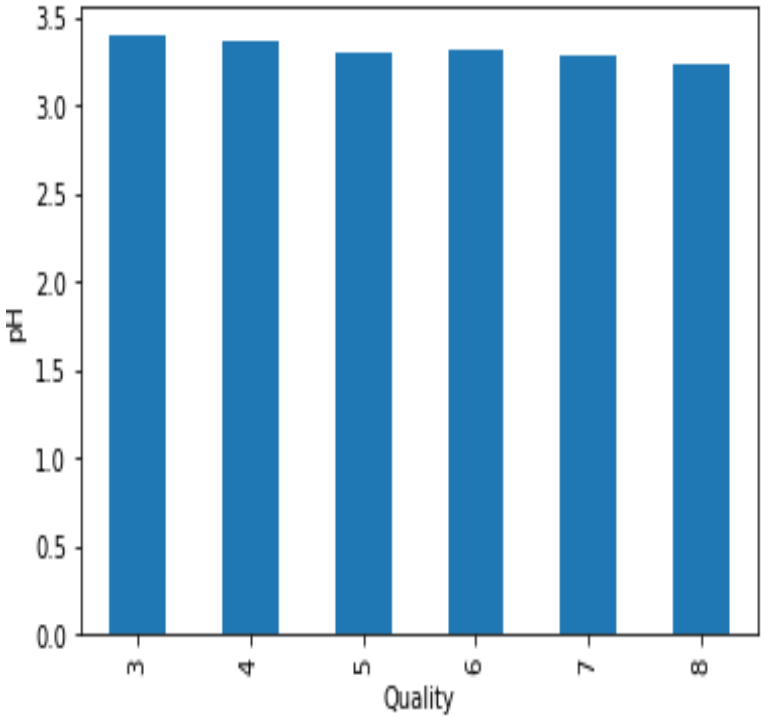


Citric Acid vs Quality

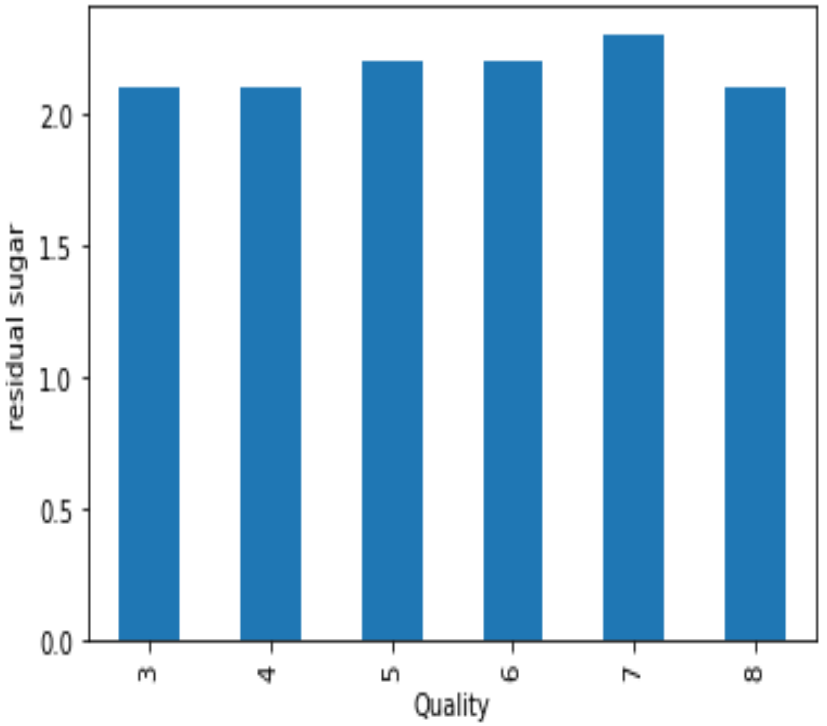
BAR CHARTS: WEAK CORRELATION



Density vs Quality



pH vs Quality



Residual Sugar vs Quality

MODEL BUILDING

MULTIPLE LINEAR REGRESSION MODEL

quality =

- 0.2762 * 'alcohol'
- + 0.0250 * 'fixed acidity'
- - 1.0836 * 'volatile acidity'
- - 0.1826 * 'citric acid'
- + 0.0163 * 'residual sugar'
- - 1.8742 * 'chlorides'
- + 0.0044 * 'free sulfur dioxide'
- - 0.0033 * 'total sulfur dioxide'
- - 17.8812 * 'density'
- - 0.4137 * 'pH'
- + 0.9163 * 'sulphates'

Multiple Linear Regression

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.361
Model:	OLS	Adj. R-squared:	0.356
Method:	Least Squares	F-statistic:	81.35
Date:	Sun, 08 Aug 2021	Prob (F-statistic):	1.79e-145
Time:	15:36:11	Log-Likelihood:	-1569.1
No. Observations:	1599	AIC:	3162.
Df Residuals:	1587	BIC:	3227.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	21.9652	21.195	1.036	0.300	-19.607	63.538
alcohol	0.2762	0.026	10.429	0.000	0.224	0.328
Q("fixed acidity")	0.0250	0.026	0.963	0.336	-0.026	0.076
Q("volatile acidity")	-1.0836	0.121	-8.948	0.000	-1.321	-0.846
Q("citric acid")	-0.1826	0.147	-1.240	0.215	-0.471	0.106
Q("residual sugar")	0.0163	0.015	1.089	0.276	-0.013	0.046
chlorides	-1.8742	0.419	-4.470	0.000	-2.697	-1.052
Q("free sulfur dioxide")	0.0044	0.002	2.009	0.045	0.000	0.009
Q("total sulfur dioxide")	-0.0033	0.001	-4.480	0.000	-0.005	-0.002
density	-17.8812	21.633	-0.827	0.409	-60.314	24.551
pH	-0.4137	0.192	-2.159	0.031	-0.789	-0.038
sulphates	0.9163	0.114	8.014	0.000	0.692	1.141

Omnibus:	27.376	Durbin-Watson:	0.585
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.965
Skew:	-0.168	Prob(JB):	1.27e-09
Kurtosis:	3.708	Cond. No.	1.13e+05

LOGISTIC REGRESSION MODELS

Weights for Model 1

- (All Predictors)

Q("volatile acidity")	-3.185168
pH	-1.702287
chlorides	-1.082183
density	-1.043145
Q("citric acid")	-0.737660
Q("fixed acidity")	-0.027132
Q("total sulfur dioxide")	-0.020082
Q("free sulfur dioxide")	0.028765
Q("residual sugar")	0.119569
alcohol	0.914453
sulphates	1.859851
dtype: float64	

Accuracy Rate: 71.46%

Weights for Model 2

- (All except PH, density, ...)

Q("volatile acidity")	-2.862298
chlorides	-1.205371
Q("citric acid")	-0.441335
Q("total sulfur dioxide")	-0.017711
Q("free sulfur dioxide")	0.031386
Q("fixed acidity")	0.095689
alcohol	0.956457
sulphates	2.050538
dtype: float64	

Accuracy Rate: 72.08%

Weights for Model 3

- (Only Volatile acidity and citric acid)

Q("volatile acidity")	-3.825206
Q("citric acid")	-0.167636
dtype: float64	

Accuracy Rate: 61.67%

CONCLUSION

- Most Important variables:
  - Volatile acidity
  - Sulphate
  - Chloride
  - Alcohol
- Consider other aspects of red wine making


THANK YOU