Wine Quality

Linear Regression

Noah Love, Frederick Cordova, Giovanni Sanchez, Matthew Chen

Business Problem

- Utilize the tools and Linear Regression techniques learned in this class to determine which factors contribute most to the quality of wine.
- In addition, we can possibly better understand if the human quality of tasting can be related to wine's chemical properties.
- Chemical properties of interest:

```
"alcohol"              "chlorides"         "citric.acid"
"density"              "fixed.acidity"     "free.sulfur.dioxide"
"pH"                   "residual.sugar"    "sulphates"
"total.sulfur.dioxide" "volatile.acidity"
```

The Dataset

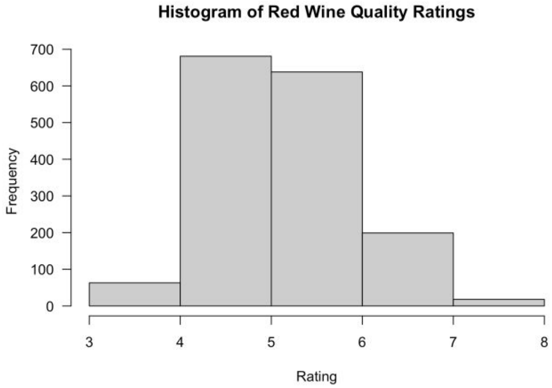
Details of the dataset

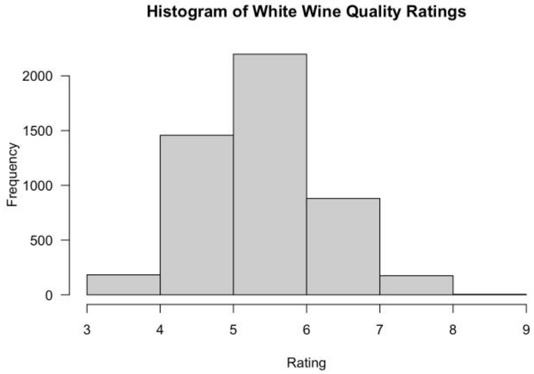
- Source: UCI - https://archive.ics.uci.edu/ml/datasets/Wine+Quality
- 2 datasets: One for white, another for red wine
- # samples: 1599 samples
- # variables: 12 total variables: 11 Predictor Variables, 1 Response Variable

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
1	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
2	7.8	0.88	0.00	2.6	0.098	25	67	0.9968	3.20	0.68	9.8	5
3	7.8	0.76	0.04	2.3	0.092	15	54	0.9970	3.26	0.65	9.8	5
4	11.2	0.28	0.56	1.9	0.075	17	60	0.9980	3.16	0.58	9.8	6
5	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
6	7.4	0.66	0.00	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

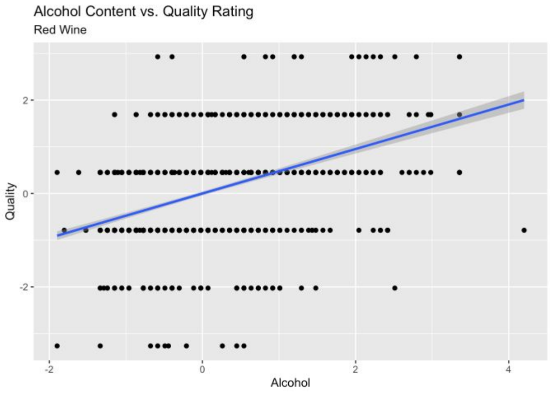
Cursory examination

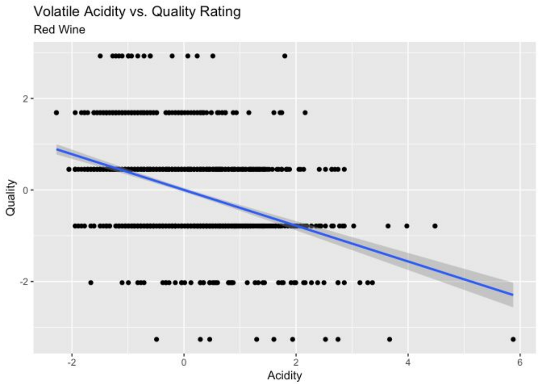


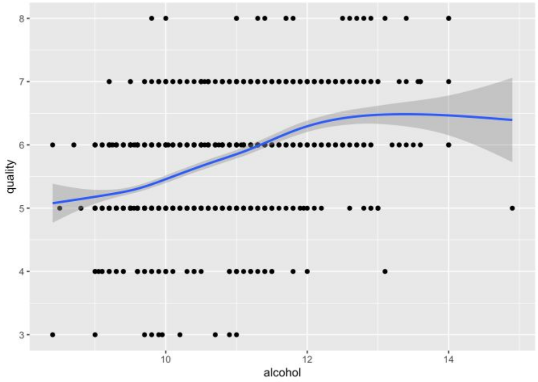


- The white wine ratings are **distributed similarly**, although a bit more symmetrically. Our model will probably not perform well when predicting white wine ratings below 4 or above 8.
- We may **consider winsorizing** or truncating our dataset to only consider observations where the **rating was above 4 and below 7** (or 8, for white wine)

Cursory examination







- After standardizing, there are a few variables that appear to be predictive of quality
- Anticipate alcohol and volatile acidity to be influential predictor variables
- However further inspection with geom_smooth reveals alcohol may not be helpful across all values

Train & Test

A 70:30 Train:Test data split on red wine left us with the following dimensions

```
dim(train)
```

```
## [1] 1119   12
```

```
dim(test)
```

```
## [1] 480  12
```

Initial Results: Full Model

Our initial model included all variables.

R_{adj}^2 = 0.3561 for full model

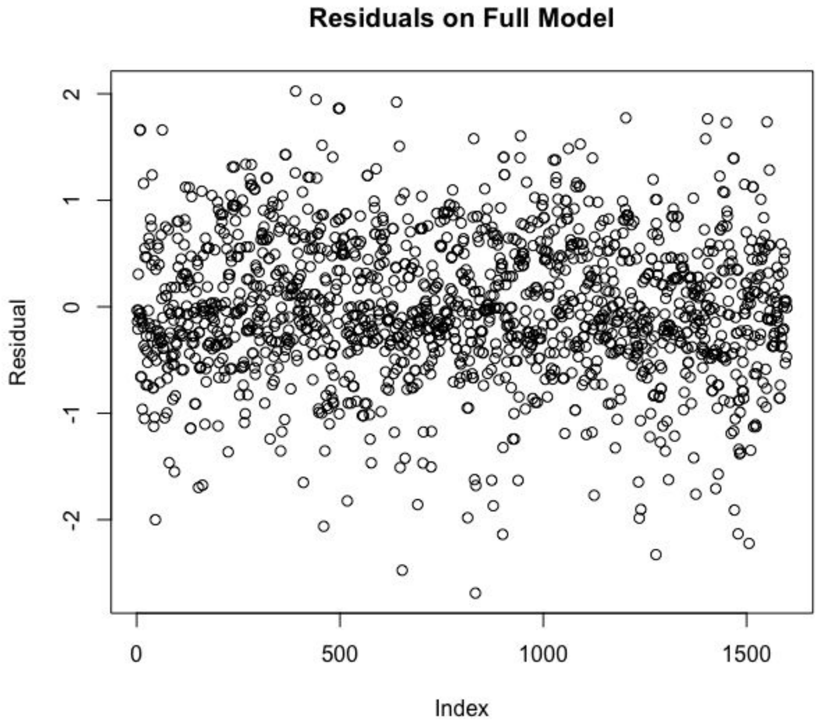
Important variables:

- Volatile acidity
- Chlorides
- Total sulfur dioxide
- Sulphates
- Alcohol

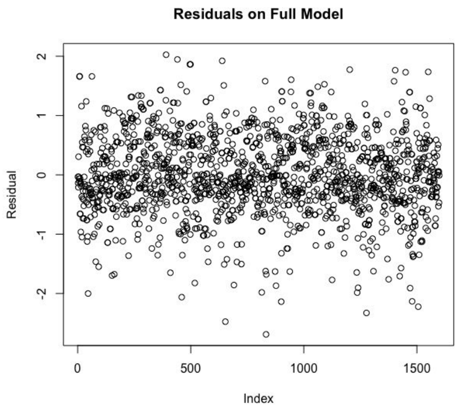
We can try to do better

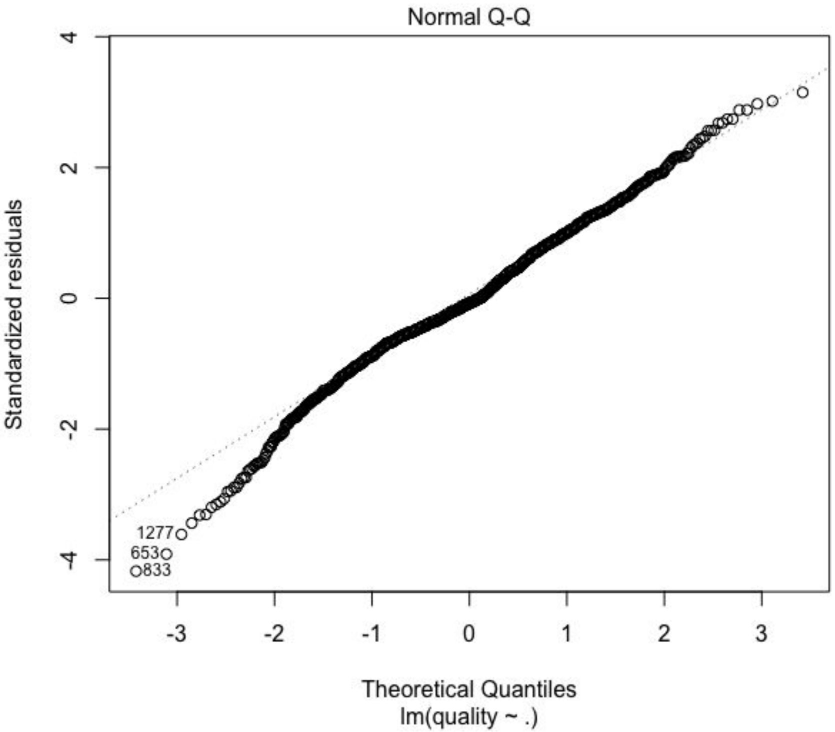
```
## Call:
## lm(formula = quality ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.197e+01  2.119e+01   1.036   0.3002
## fixed.acidity        2.499e-02  2.595e-02   0.963   0.3357
## volatile.acidity    -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## citric.acid         -1.826e-01  1.472e-01  -1.240   0.2150
## residual.sugar       1.633e-02  1.500e-02   1.089   0.2765
## chlorides           -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## free.sulfur.dioxide  4.361e-03  2.171e-03   2.009   0.0447 *
## total.sulfur.dioxide -3.265e-03 7.287e-04  -4.480 8.00e-06 ***
## density             -1.788e+01  2.163e+01  -0.827   0.4086
## pH                  -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates            9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol              2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Initial Results: Full Model

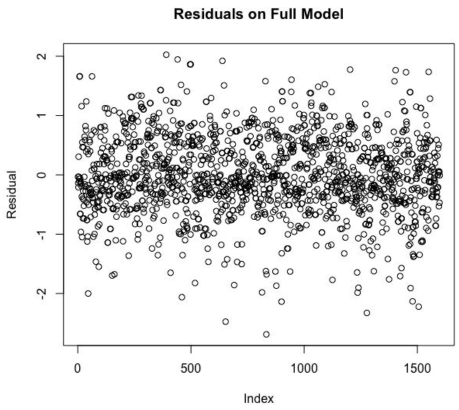


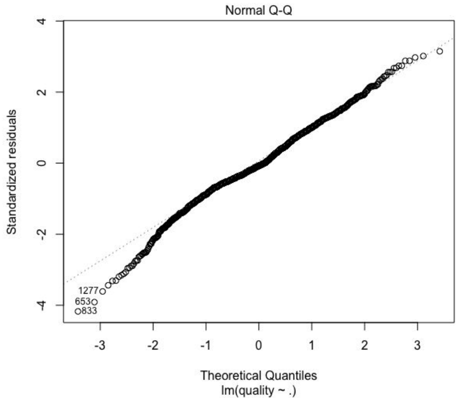
Initial Results: Full Model

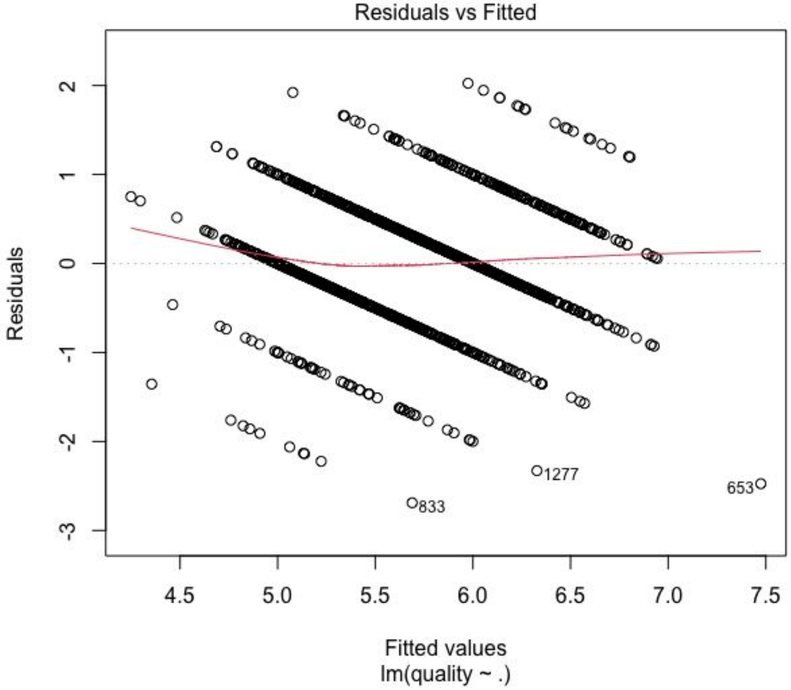




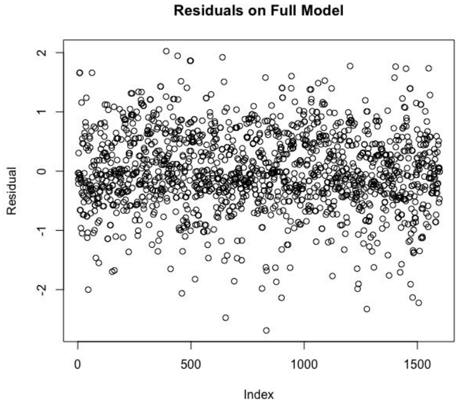
Initial Results: Full Model

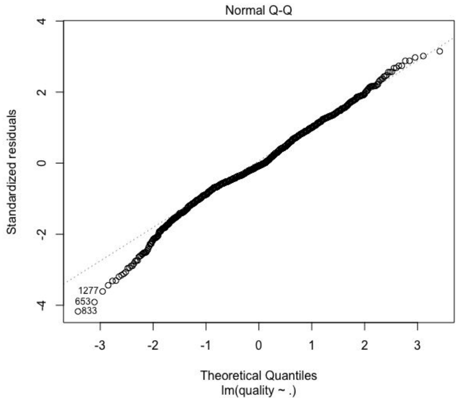


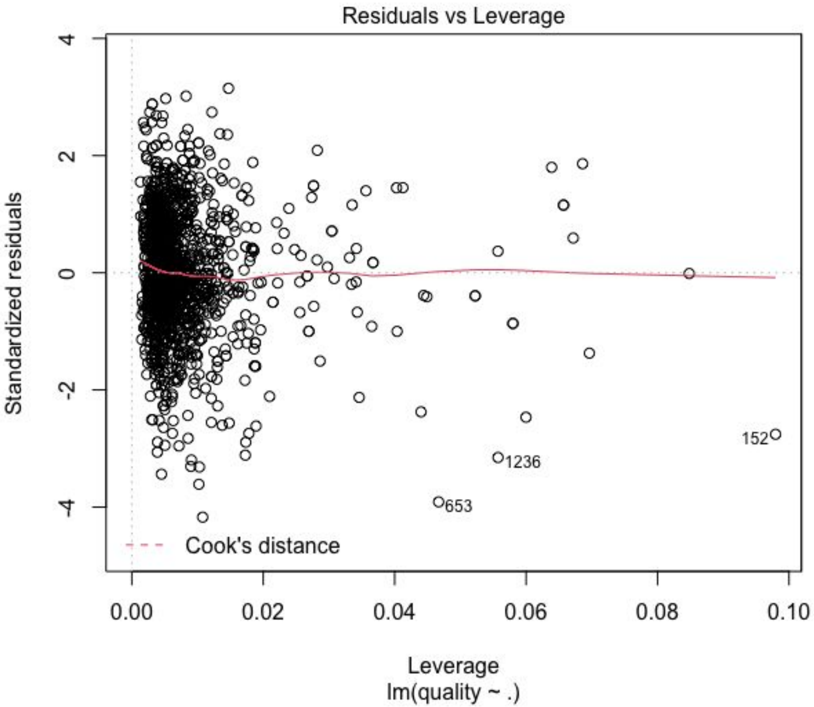


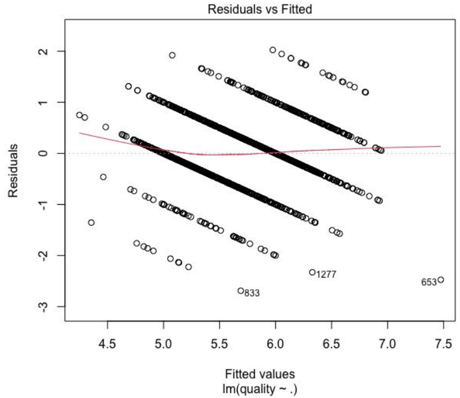


Initial Results: Full Model









Stepwise Regression

Attempted to pare down the variables using stepwise

Made several attempts at different subsets of variables with similar effect

We find inclusion of similar variables we thought important from preliminary results

$R_{adj}^2 = 0.3567$

Not great results, let’s check out

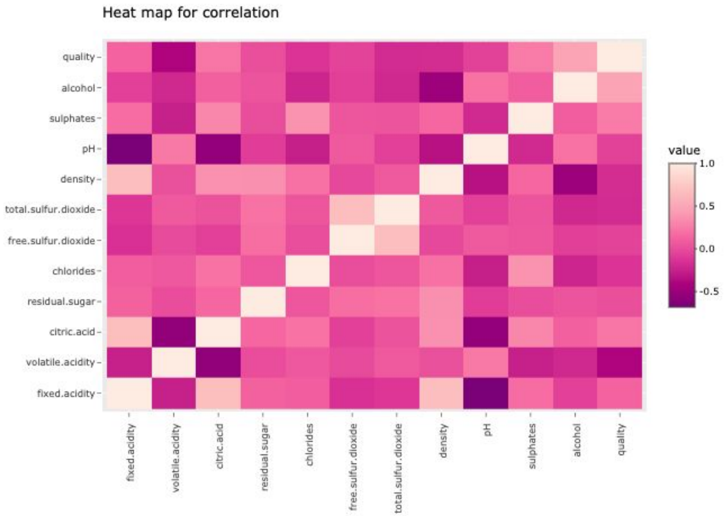
correlative values

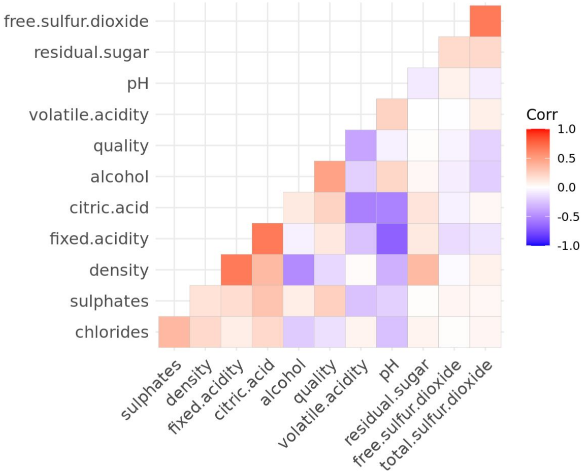
```
## Step:  AIC=3158.98
## .outcome ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                        Df Deviance    AIC
## <none>                      667.54 3159.0
## - free.sulfur.dioxide   1   669.93 3162.7
## - pH                    1   674.61 3173.8
## - total.sulfur.dioxide  1   678.32 3182.6
## - chlorides             1   678.35 3182.7
## - sulphates             1   694.60 3220.5
## - volatile.acidity      1   709.85 3255.3
## - alcohol               1   792.02 3430.4
```

Correlation

At this point, we thought multicollinearity may be a problem

Also, realizing that perhaps there isn't much to garner from the dataset





We see there isn't that much correlation among features. Importantly, quality doesn't exhibit much correlation with any specific variable, except maybe alcohol.

Correlation & Multicollinearity

We thought it would be helpful to remove the most highly correlated variables

Taking a look at VIF, we confirm what we see visually in the correlation matrix

- Fixed acidity has highest VIF of 7.7675 followed by density at 6.343760
- These two variables don't have high correlation to quality, so let's try to remove them
- Mean VIF of 3.1049 > 1 - tells us multicollinearity is an issue

Removing these two variables and running lm, we see a slight improvement with:

R_{adj}^2 = 0.3565 - 0.0004 higher than full model, but worse than stepwise

We'll attempt Ridge Regression as remediation for multicollinearity issue

Ridge Regression + Cross Validation

Hoped to reduce effects of multicollinearity among features by using ridge regression

Found optimal lambda of 0.06310

Performance on test data:

$R^2 = 0.3809$

RMSE = 0.6563

Not the best results, but better

LASSO

Continuing search for simpler model

Prioritize variable selection with Lasso

over ridge

Optimal lambda value of 0.005012

Performance on test data:

R^2 = 0.3892

RMSE = 0.6519

Most influential variables:

- Alcohol: 0.3147
- Volatile acidity: -0.1838
- Sulphates: 0.1481

$$\hat{Y} = 5.6247 + 0.1289X_1 - 0.1838X_2 + 0.009215X_3 - 0.09065X_4 + 0.02873X_5 - 0.08711X_6 - 0.04197X_7 + 0.1481X_8 + 0.3147X_9$$

Elastic Net

Let’s see if we can do any better

Hyperparameter values:

Alpha: 0.03193

Lambda: 0.01560

Performance on test data:

$R^2 = 0.3871$

RMSE = 0.6530

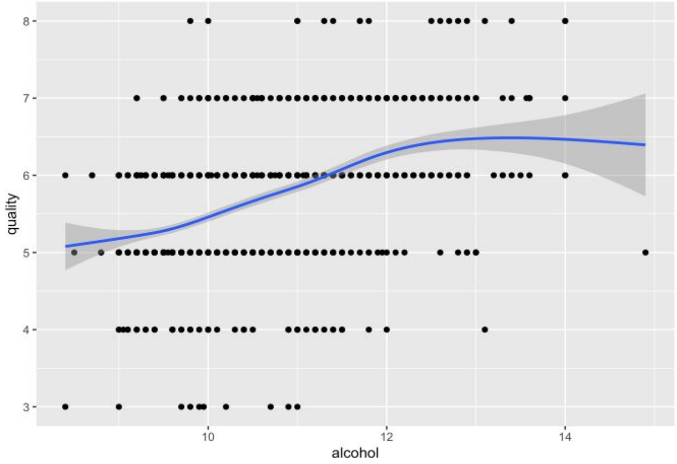
Elastic Net

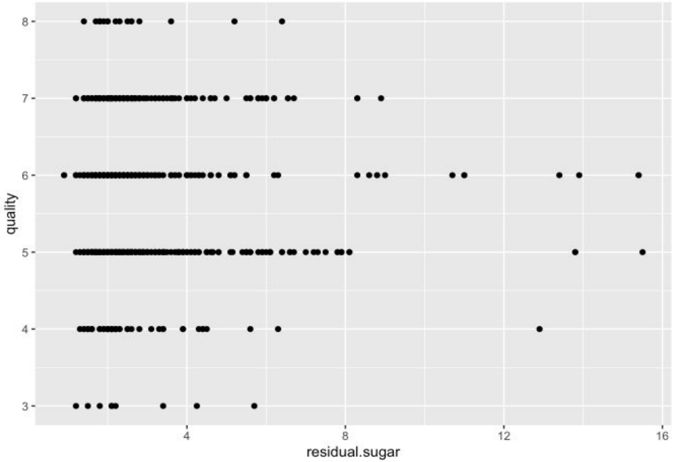
Clear that there's not much to go off of here

Perhaps somewhat of a relationship with alcohol, as we saw in the correlation matrices

But even that falls off after about 12%

We see also that there are several outliers, residual sugar being one example of features with dispersed points





Takeaways

- Best performing model achieved with Lasso - R^2 = 0.3892
- Possibly looking at a subset of important features
    - Age
    - Region
- Perhaps a problem better suited for clustering or decision trees
- At the end of the day, maybe winemaking is more art than science
- Subjectivity of quality indicator may mean that most important features are not quantifiable
    - Psychology of wine quality & price
- Github repo: https://github.com/noahlove/linear-regression-final-project/tree/main

Improvements

1. Random Forest Classifier works well (86% predictive accuracy)
2. Stochastic Gradient Descent Classifier (84% accuracy)
3. Grid Search Vector Classification with CV (90%)
4. Multivariate analysis with expert insight
5. PCA (98.5 %) or maybe ICA