CSCI 4360 Data Science II Project I

Ayush Kumar, Faisal Hoissain, Brandon Amirouche

February 23, 2021

Contents

1 Introduction and Methodology

1.1 Code Credits

2 Auto-MPG

3 Concrete Compressive Strength

4 Crime

5 Wine Quality

5.1 The Dataset

The wine quality data is actually two datasets, one for red wine and the other for white wine. Both of them have the same 11 features, but there may be differences in the features that matter, as well as the overall results of regression. A dummy variable could be created to meld the datasets into one overall set, but if there are substantial differences between the two it may substantially add to model complexity. The dataset has the following variables:

1. fixed.acidity
2. volatile.acidity
3. citric.acid
4. residual.sugar
5. chlorides
6. free.sulfur.dioxide

7. total.sulfur.dioxide
8. density
9. pH
10. sulfates
11. alcohol
12. quality (the response variable)

The red wine dataset has 1,599 observations, and the white wine dataset has 4,898 observations. The two files can be found in the data/WineQuality folder, and were orignally downloaded from the UCI Machine Learning Repository. [insert link here]

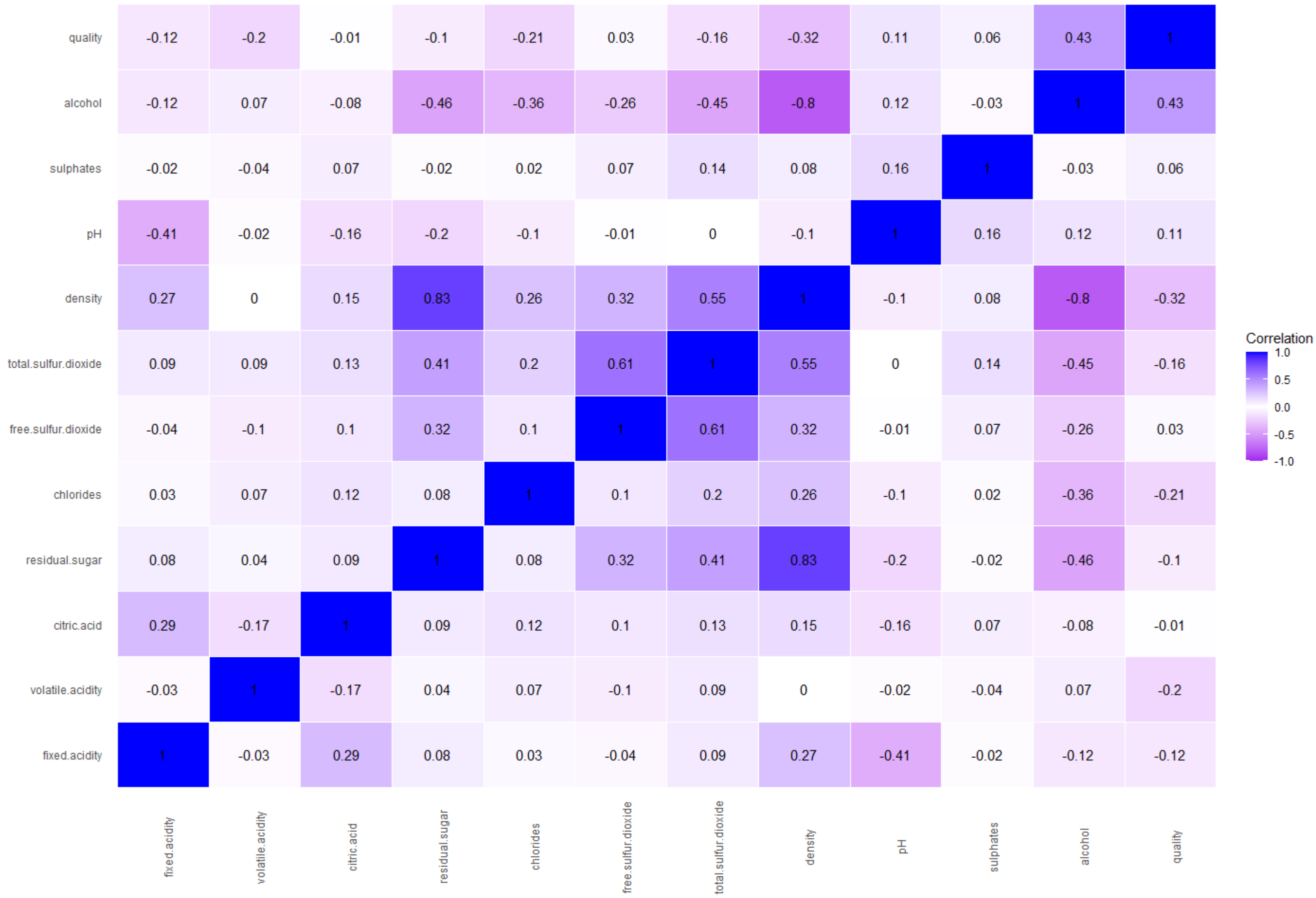
5.2 Exploratory Data Analysis

The first step in EDA was to checkout the collinearity of features, and how well they correlate with our response variables.

Red Wine Correlation Matrix (Plotted with R)

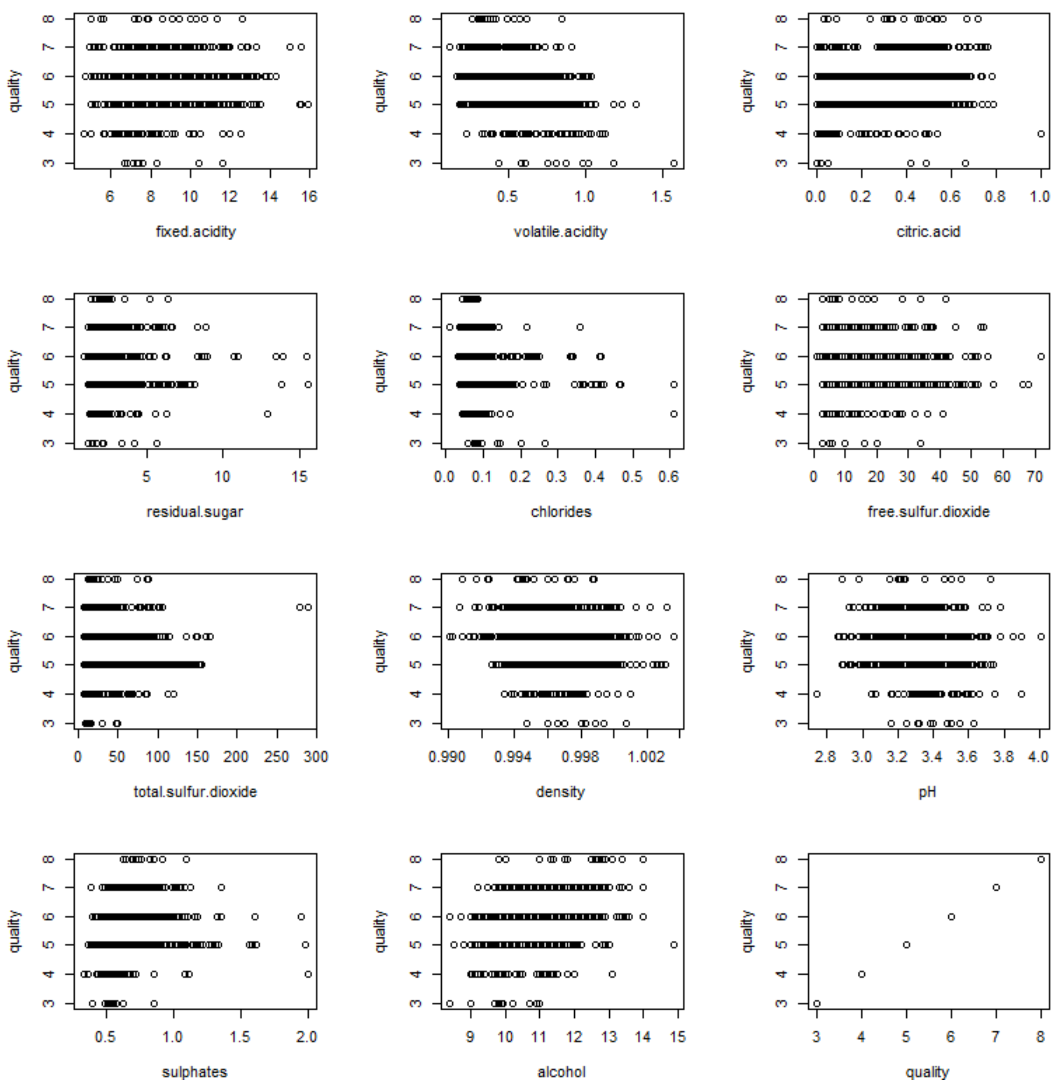
From the Red Wine Correlation Matrix (Figure 1) we can see that there are a few variables that have fairly strong correlations with each other. The citric acid content and the fixed acidity seem to be highly correlated with each other, and this pattern holds true for many of the features that depend on acidity. We can see that pH is highly correlated with fixed acidity and citric acid as well. This will be important to consider as it may create a problem in our regression. Based on the results of our variable screening methods, we may consider dropping one or more these variables, or using them as an instrumental variable to reduce the endogeneity issue with this dataset. Another highly correlated issue may be density and fixed acidity, as well as free.sulfur.dioxide and total sulfur dioxide. It may be worth our time to standardize some of these variables to limit collinearity, and this will be considered during variable selection.

White Wine Correlation Matrix (Plotted with R)



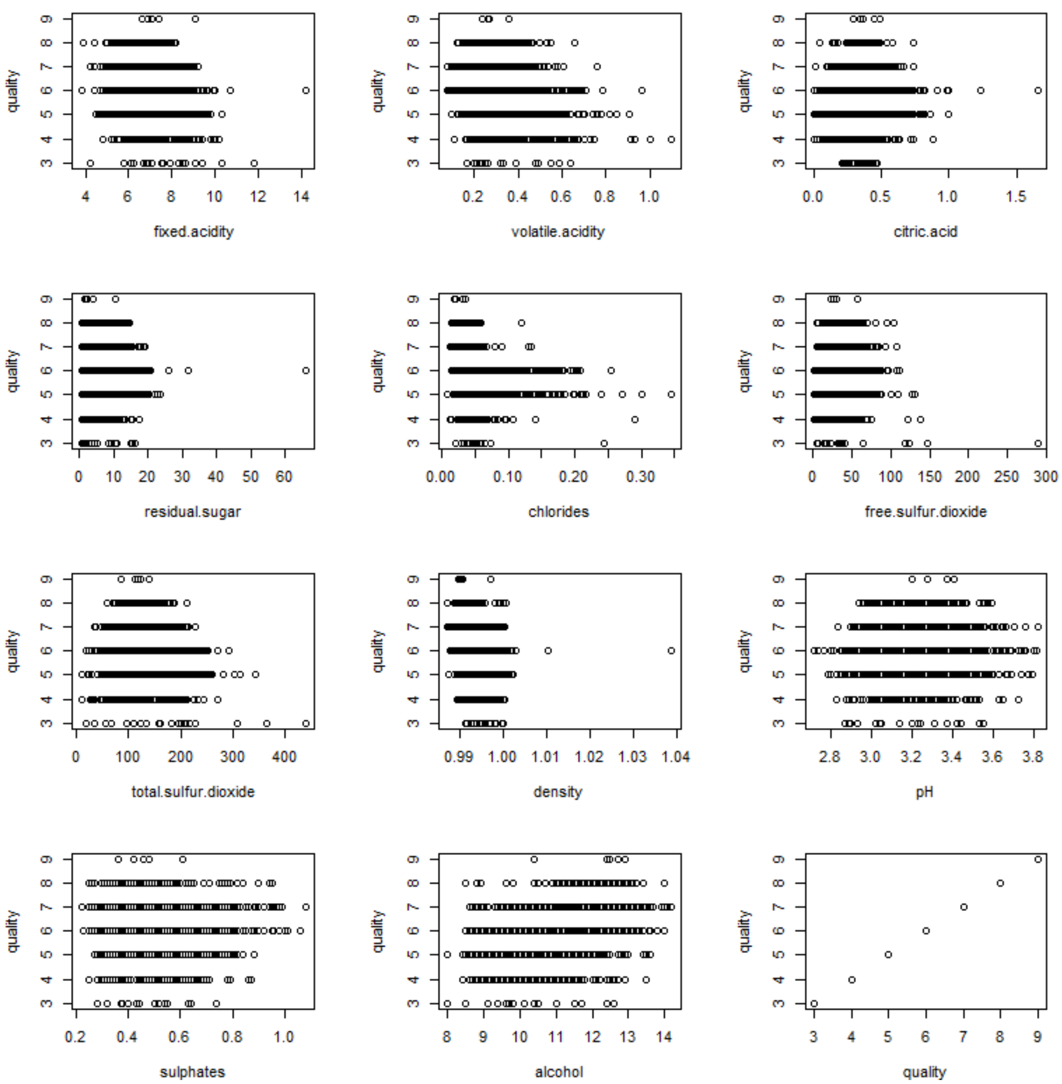
Moving on to the white wine dataset, we can see very clearly that the behavior is very different from the collinearity between variables is quite high for two interactions in specific. The first is between density and residual sugar, the second is between alcohol and density. Both of these are higher than 0.8, but residual sugar seems to have very little correlation with our response variables, so it may not end up being an issue after variable selection. The high correlation between alcohol and density seems to pose a much more substantial issue because both have some correlation with quality of the wine. We may standardize one or both of these variables during variable screening, or remove one of them from our regression on the basis of our 4 criterion.

Red Wine Scatter Plots (R)



Quality seems to be a multi-categorical variable, based on this information maybe regression is not the best way to predict on this dataset, but it may give a great deal of information regarding feature selection. Some features exhibit very peculiar behavior of having similar characteristics for both low and high quality wines, but not for middle quality wines. This can be illustrated when we take a look at total sulfur dioxide, chlorides, and residual sugar. This pattern may mislead us during variable selection, so it may be a good idea to add in quadratic terms for these variables to account for their curving behavior. If these variables fail to be selected initially then I will rerun variable selection methods taking into account their quadratic terms. There may be other transformations needed, but we can determine those after variable selection bases on partial residual plots.

White Wine Scatter Plots (R)



We once again see very similar patterns to the first dataset in terms of residual sugar, chlorides, and sulfur dioxide. There does not seem to be anything else out of the ordinary in these scatterplots. Taking this into account the obvious next step is variable screening and selection.

5.3 Variable Screening and Selection

Variable Selection for Red Wine

Output has been modified for space. Run code for full output.

```
Start:  AIC=-682.5
quality ~ 1

                        Df Sum of Sq     RSS      AIC
+ alcohol                1   236.295  805.87 -1091.65
+ volatile.acidity       1   158.967  883.20  -945.14
+ sulphates              1    65.865  976.30  -784.89
+ citric.acid            1    53.405  988.76  -764.61
+ total.sulfur.dioxide   1    35.707 1006.46  -736.24
+ density                1    31.887 1010.28  -730.19
+ chlorides              1    17.318 1024.85  -707.29
+ fixed.acidity          1    16.038 1026.13  -705.29
+ pH                     1     3.473 1038.69  -685.84
+ free.sulfur.dioxide    1     2.674 1039.49  -684.61
<none>                               1042.17  -682.50
```

```
+ residual.sugar        1     0.197 1041.97  -680.80

Step:  AIC=-1091.65
quality ~ alcohol

                      Df Sum of Sq    RSS     AIC
+ volatile.acidity     1    94.074 711.80 -1288.1


Step:  AIC=-1288.14
quality ~ alcohol + volatile.acidity

                      Df Sum of Sq    RSS     AIC
+ sulphates            1   19.6916 692.10 -1331.0

Step:  AIC=-1331
quality ~ alcohol + volatile.acidity + sulphates

                      Df Sum of Sq    RSS     AIC
+ total.sulfur.dioxide  1    8.2176 683.89 -1348.1


Step:  AIC=-1348.1
quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide

                     Df Sum of Sq    RSS     AIC
+ chlorides           1    8.0370 675.85 -1365.0

Step:  AIC=-1365
quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
    chlorides

                     Df Sum of Sq    RSS     AIC
+ pH                  1    5.9189 669.93 -1377.1


Step:  AIC=-1377.06
quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
    chlorides + pH

                     Df Sum of Sq    RSS     AIC
+ free.sulfur.dioxide  1   2.39413 667.54 -1380.8
<none>                             669.93 -1377.1


Step:  AIC=-1380.79
quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
    chlorides + pH + free.sulfur.dioxide

                Df Sum of Sq    RSS     AIC
<none>                       667.54 -1380.8


Call:
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
    total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide,
    data = red_wine)
```

One of the variables that was omitted here was residual sugar, one of the variables that exhibited very strong quadratic behavior. I will be rerunning forward selection while including the quadratic term for residual sugar and it will be included for all of the following variable selection methods as well. This is the final call when I include the quadratic term in the forward step.

```
Call:
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide,
data = red_wine)
```

As can be seen even the inclusion of the quadratic term did not change the results of the forward selection. Below are listed the final conclusions of backwards elimination and step wise regression.

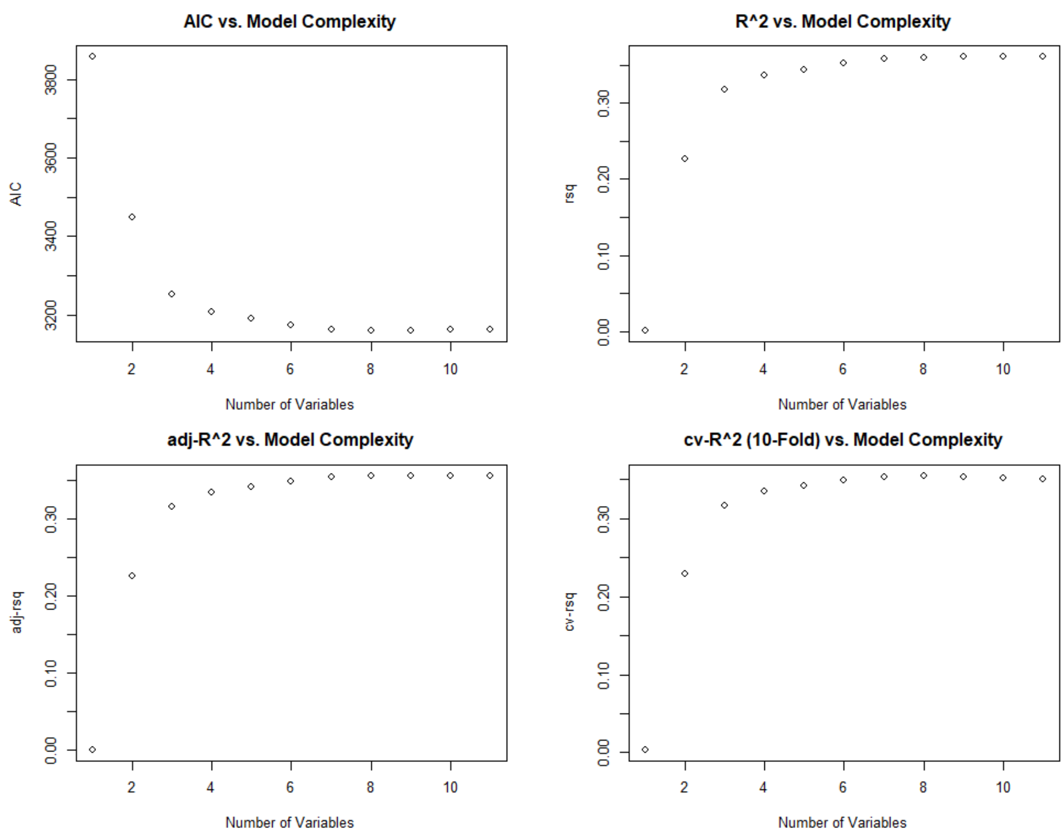
Backward Elimination

```
Call:
lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
total.sulfur.dioxide + pH + sulphates + alcohol, data = red_wine)
```

Step Regression

```
Call:
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
total.sulfur.dioxide + chlorides + pH + free.sulfur.dioxide,
data = red_wine)
```

While the methods may be different every single one of our variable screening methods chose the same seven variables to include: alcohol, volatile.acidity, sulphates, total.sulfur.dioxide, chlorides, pH, and free.sulfur.dioxide. These are the variables that will be included in all our models for red wine moving forward.



Here we can see that as we increase the number of variables, we see diminishing marginal improvement in quality of fit. For measures that penalize model complexity such as R^2 and AIC we can even see a light bend towards the quality of fit worsening as model complexity increases. The diminishing returns can in part be explained by variables that have weak correlation

Variable Selection for White Wine

6 Seoul Bike Rentals

7 Air Quality