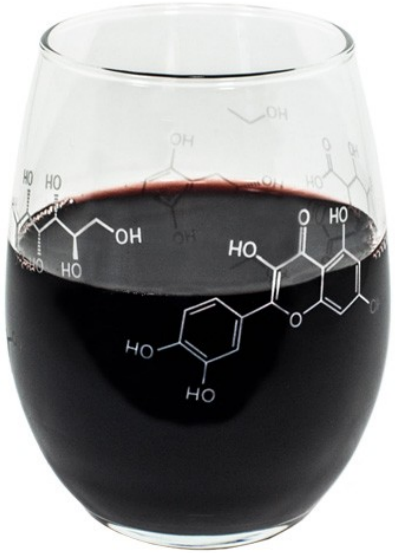
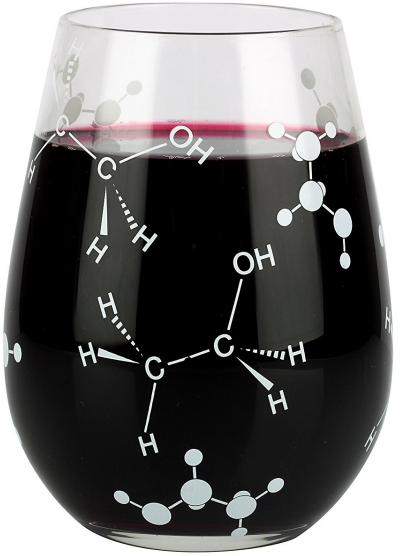
Quality and Chemistry of Wines



Aaron Levine
February 2017



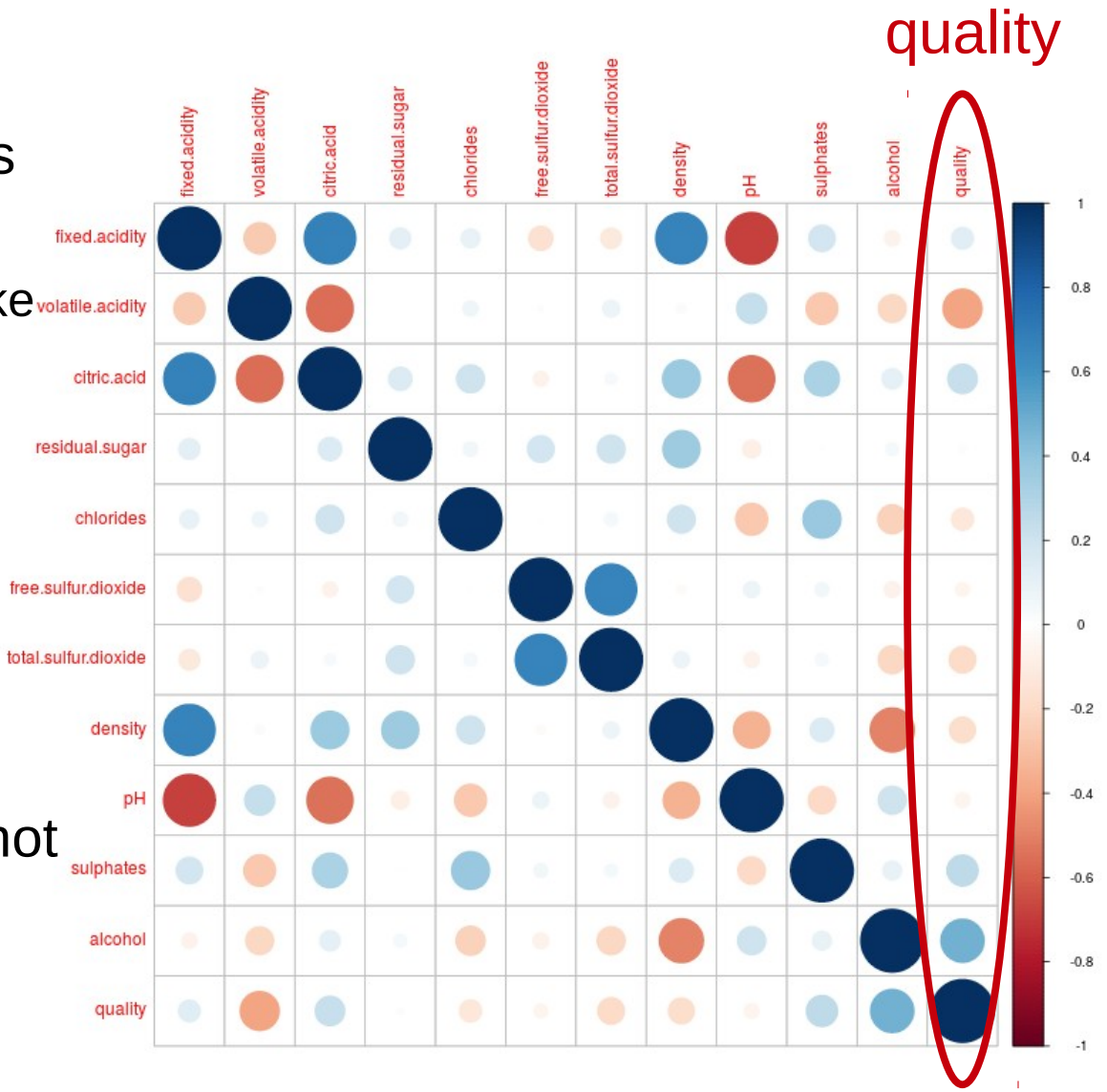
Overview

- Is it possible to predict the quality of a wine from its chemical composition?
- Restaurants can simply inquire about the chemistry of a wine before deciding whether or not to request it
  - Useful for new wines that haven't been rated yet
- Public data available on the UCI machine learning repository
  - Contains 11 chemical attributes of red and white wines
    - Density, pH, citric acid content, etc.
  - Contains quality ratings for each wine as determined by professional critics
- Let's develop a model to do this!

Red Wines

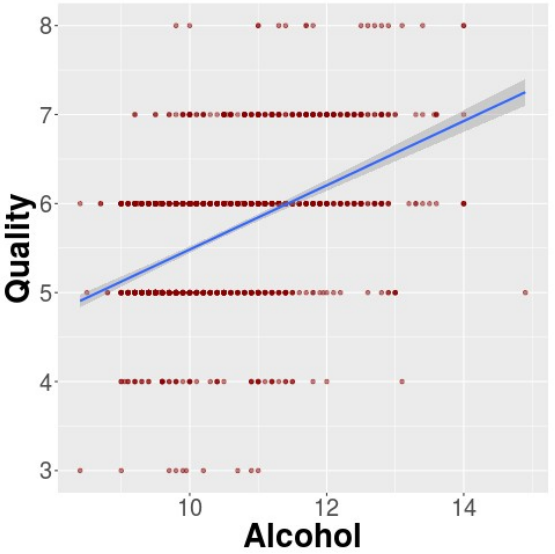
Correlation Matrix

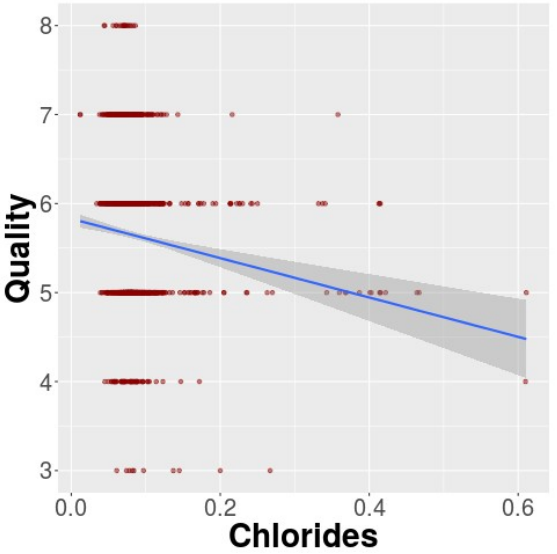
- Which of the 11 features are important for identifying wines by their quality?
  - With only 11 features, can make a matrix of correlation coefficients to visualize relationships
- Strongest correlations to quality
  - Volatile acidity
  - Alcohol content
- Correlation coefficients may not tell the full story

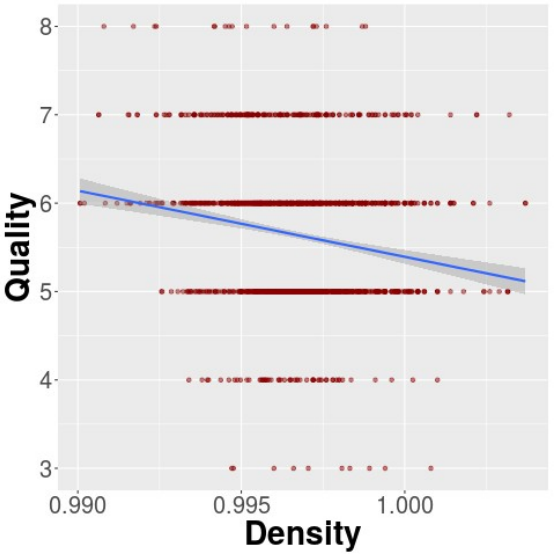


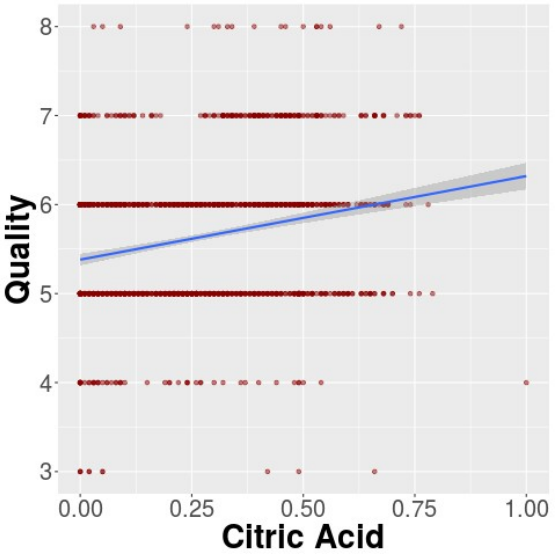
Visualize Relationships

- Plot quality vs features
  - Add linear regression for each plot



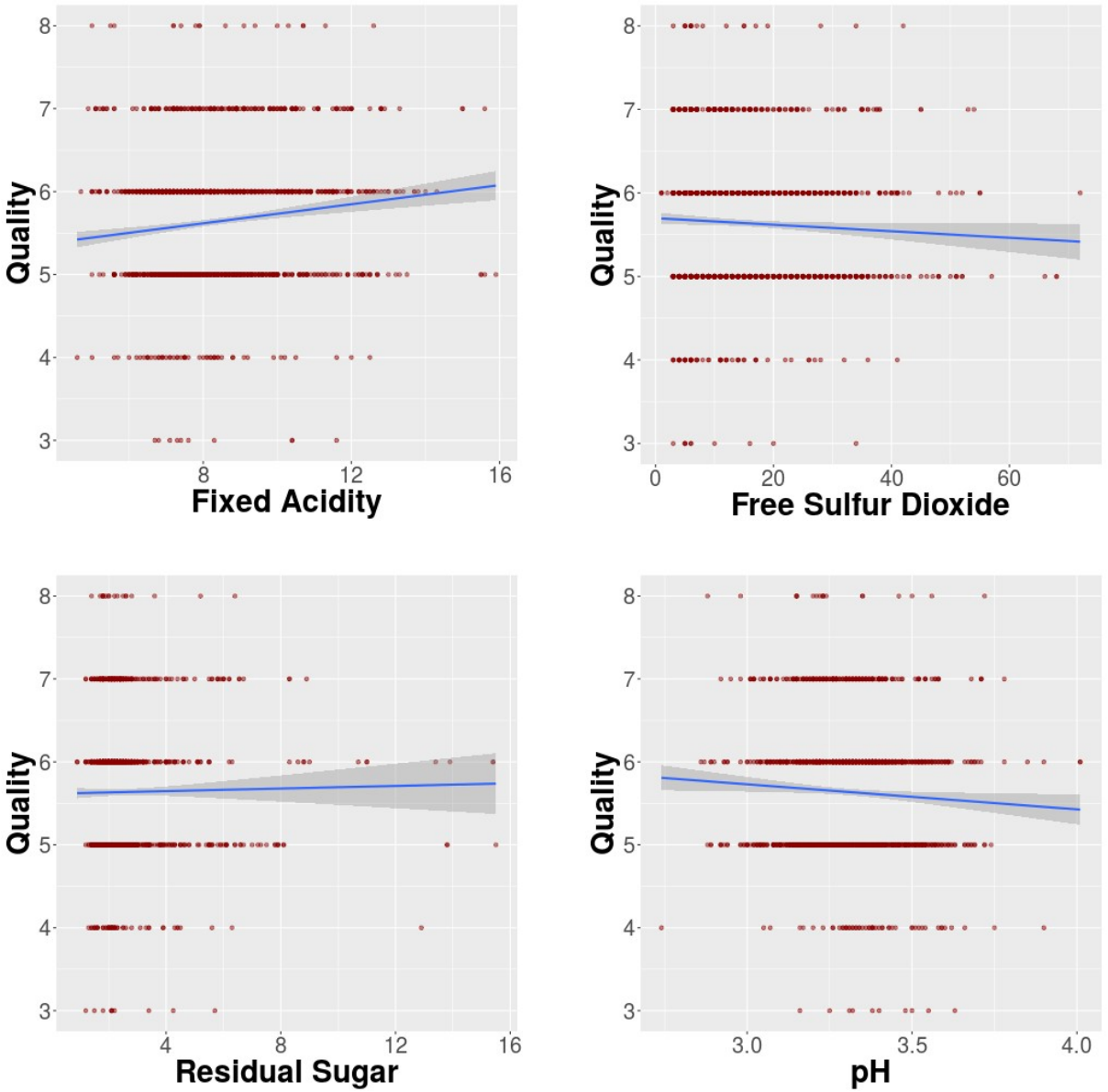




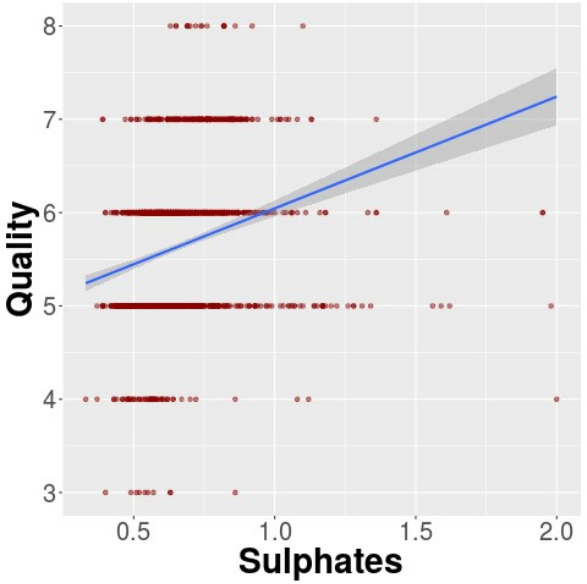


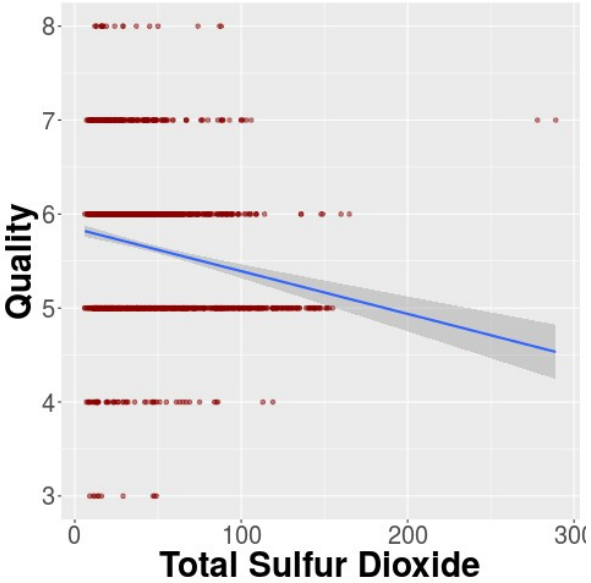
Visualization Continued

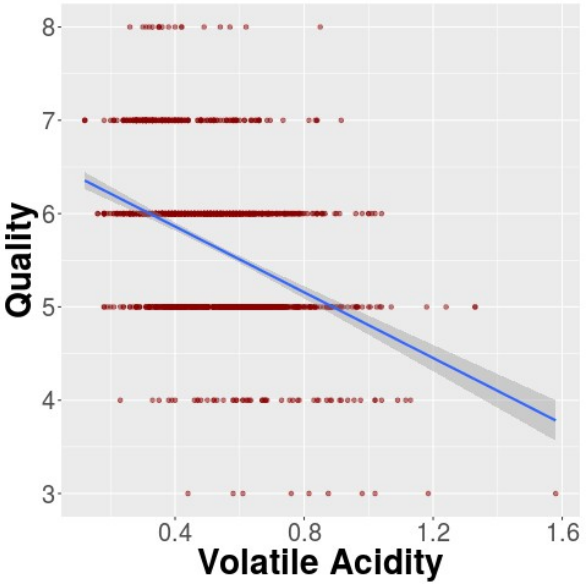
Weak Correlations



Visualization Part 3







Building a Model

- Now it's time to build a machine learning model that predicts the quality of a wine from its chemistry
- Quality of wines are known
  - Supervised learning
  - Multiclass classification
- Use a support vector machine (SVM) with a radial basis function (RBF) kernel

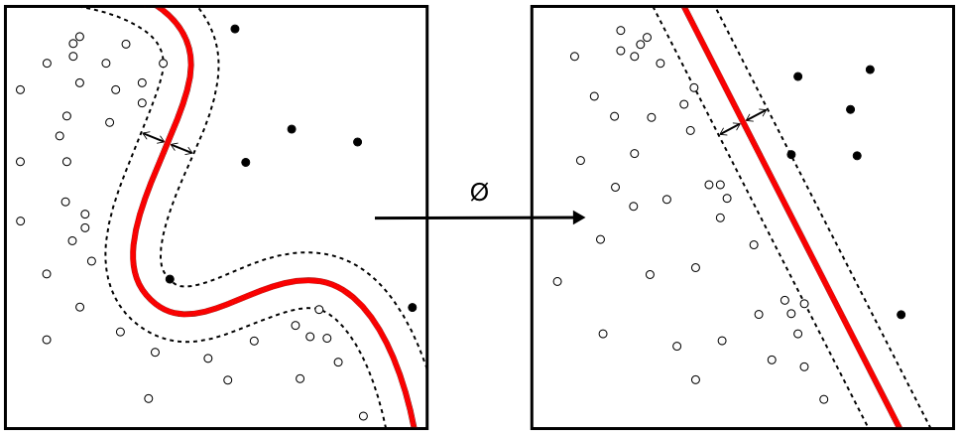


Image source: wikipedia

Training the SVM

- Randomly divide the data into training, cross validation, and testing
  - 60% train, 20% cross validation, 20% testing
  - Use same random seed in R for reproducibility
- Use cross validation set to select optimal values of C and γ
  - Higher values of C increase penalty for missclassification, results in tighter fit to training data (high variance)
  - Lower values of γ: smoother variance of the RBF kernel near data points, results in looser fit to training data (high bias)

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$$

Training Results

- Use svm with library(e1071) in R
- Train with alcohol, volatile acidity
  - Highest correlation to quality
  - Start small to avoid overtraining
- Use cross validation set to determine C and γ
  - Optimal value: C = 3, γ= 1
  - 57.7% accuracy on CV set
- Results in 52.8% accuracy on test set
  - Can this be improved?

C

γ	0.01	0.03	0.1	0.3	1	3	10	30	100
0.01	0.429	0.429	0.508	0.53	0.539	0.542	0.542	0.542	0.542
0.03	0.429	0.505	0.533	0.539	0.542	0.542	0.542	0.539	0.533
0.1	0.483	0.53	0.542	0.542	0.542	0.539	0.542	0.545	0.545
0.3	0.527	0.539	0.539	0.542	0.549	0.542	0.552	0.549	0.561
1	0.517	0.555	0.552	0.542	0.549	0.577	0.567	0.552	0.561
3	0.429	0.552	0.561	0.558	0.574	0.567	0.533	0.524	0.53
10	0.429	0.486	0.545	0.555	0.542	0.527	0.52	0.524	0.533
30	0.429	0.429	0.527	0.52	0.545	0.517	0.53	0.52	0.517
100	0.429	0.429	0.429	0.502	0.53	0.542	0.53	0.527	0.53

Cross Validation Results

- Train with all 11 features
- See Cross Validation table
  - Optimal value: C = 1, γ= 1
  - 63.3% accuracy on CV set
- 59.1% accuracy on test set
  - 58.8% chance 5 prediction is correct
  - 76.9% of 5's correctly identified
  - 55.7% chance 6 prediction is correct
  - 59.3% of 6's correctly identified
  - 84.2% chance 7 prediction is correct
  - 38.1% of 7's correctly identified
  - No 3, 4, 8, predictions made

Cross Validation

C

γ	0.01	0.03	0.1	0.3	1	3	10	30	100
0.01	0.429	0.429	0.524	0.555	0.555	0.567	0.58	0.596	0.577
0.03	0.429	0.52	0.558	0.564	0.564	0.592	0.602	0.58	0.602
0.1	0.429	0.536	0.561	0.561	0.618	0.605	0.621	0.602	0.564
0.3	0.429	0.436	0.542	0.583	0.624	0.605	0.574	0.586	0.58
1	0.429	0.429	0.429	0.514	0.633	0.621	0.608	0.614	0.614
3	0.429	0.429	0.429	0.433	0.577	0.571	0.571	0.571	0.571
10	0.429	0.429	0.429	0.436	0.552	0.552	0.552	0.552	0.552
30	0.429	0.429	0.429	0.436	0.542	0.542	0.542	0.542	0.542
100	0.429	0.429	0.429	0.436	0.536	0.536	0.536	0.536	0.536

Test

True

Pred	3	4	5	6	7	8
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	3	13	100	49	5	0
6	1	5	30	73	21	1
7	0	0	0	1	16	2
8	0	0	0	0	0	0

Identifying Best Red Wines: Overview

- The previous model was successful in identifying average wines, but unsuccessful in identifying outliers
- Restaurants and consumers don't want to identify average with high accuracy, they want to identify the best and/or avoid the worst
- Apply an SVM to two classes of wines: best(7-8) and not best(3-6)
  - Highly skewed classes
  - Precision (P) = $\frac{True\ Positives}{True\ Positives + False\ Positives}$
  - Recall (R) = $\frac{True\ Positives}{True\ Positives + False\ Negatives}$
- Use F-Score metric = $\frac{2 * P * R}{P + R}$

Identifying Best Red Wines: Results

- Train SVM with all 11 features
  - Class 0: Not best wines
  - Class 1: Best wines
- CV results
  - Optimal value: C = 100, γ= 0.3
  - 0.645 F score
- Test results
  - 0.568 F score
  - 5 5's, 13 6's, 18 7's, 3 8's identified as top wines (Class 1)
    - Over half of wines are in fact top quality
    - No terrible wines (3,4) were false positives
  - 10 7's, 4 8's missed identification (Class 0)
    - Over half of top wines in the test set were correctly identified
    - Room for improvement: over half of 8's are not correctly identified

Cross Validation

γ \ C	0.01	0.03	0.1	0.3	1	3	10	30	100
0.01	NaN	NaN	NaN	NaN	NaN	NaN	0.138	0.394	0.456
0.03	NaN	NaN	NaN	NaN	0.038	0.412	0.45	0.539	0.559
0.1	NaN	NaN	NaN	0.04	0.432	0.53	0.6	0.577	0.588
0.3	NaN	NaN	NaN	0.218	0.557	0.598	0.581	0.622	0.645
1	NaN	NaN	NaN	NaN	0.575	0.61	0.602	0.602	0.602
3	NaN	NaN	NaN	NaN	0.393	0.469	0.469	0.469	0.469
10	NaN	NaN	NaN	NaN	0.339	0.393	0.393	0.393	0.393
30	NaN	NaN	NaN	NaN	0.339	0.367	0.367	0.367	0.367
100	NaN	NaN	NaN	NaN	0.339	0.339	0.339	0.339	0.339

Test

Pred \ True	0	1
0	267	14
1	18	21

Identifying Worst Red Wines: Results

- An average consumer might want to avoid bad wines (3,4 quality)
- Train SVM
  - Class 0: Not bad wines
  - Class 1: Bad wines
- CV results
  - Optimal value: C = 30, γ= 0.3
  - 0.316 F score
  - NaN score when no true positives
- Test results
  - No terrible wines sucessfully identified

Cross Validation

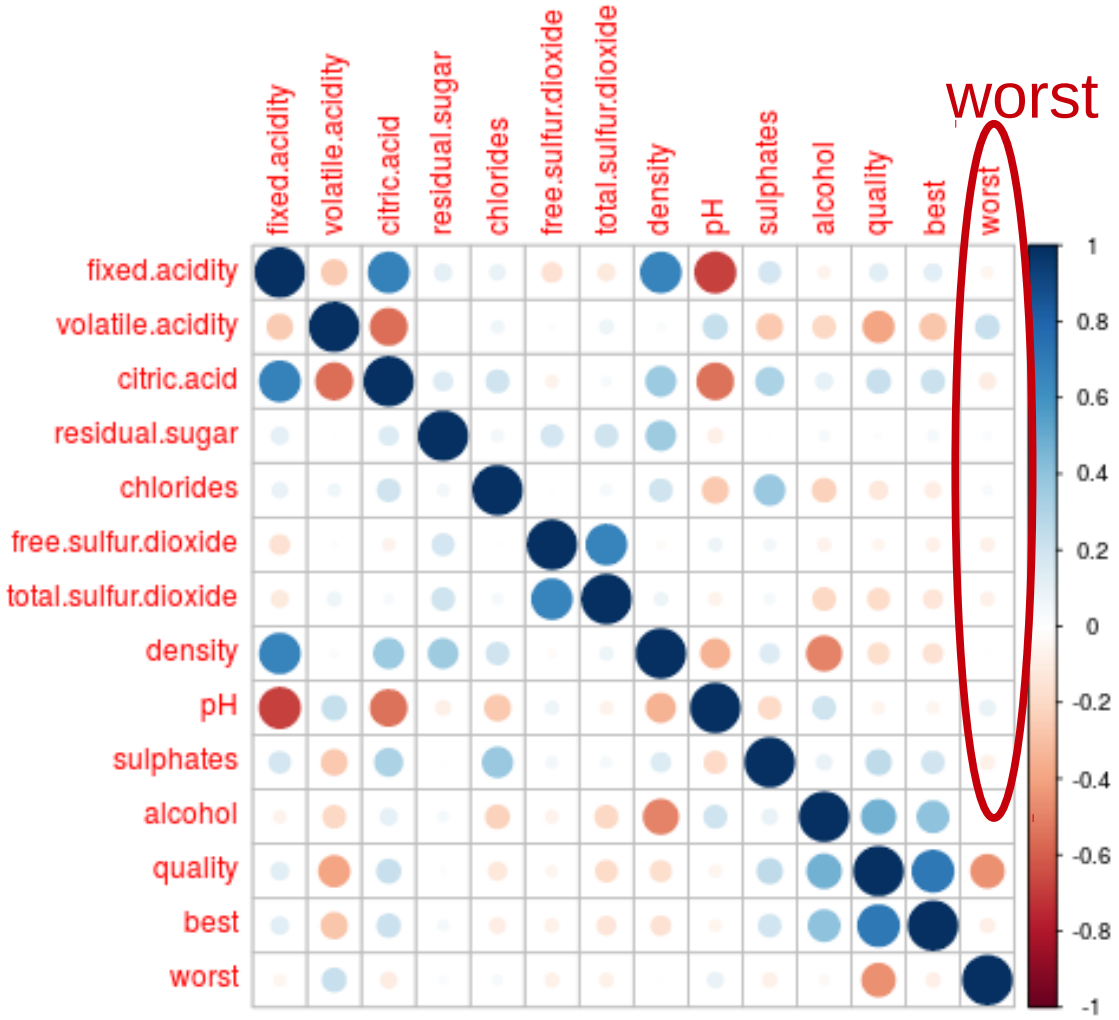
C

γ	0.01	0.03	0.1	0.3	1	3	10	30	100
0.01	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.03	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.143	NaN
0.1	NaN	NaN	NaN	NaN	NaN	NaN	0.143	0.267	0.261
0.3	NaN	NaN	NaN	NaN	NaN	0.154	0.235	0.316	0.316
1	NaN	NaN	NaN	NaN	NaN	0.154	0.154	0.154	0.154
3	NaN	NaN	NaN	NaN	NaN	0.154	0.154	0.154	0.154
10	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
30	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Test

		True 0	True 1
Pred	*0*	269	12
	1	39	0

Full Red Wine Correlation Matrix
With Best, Worst

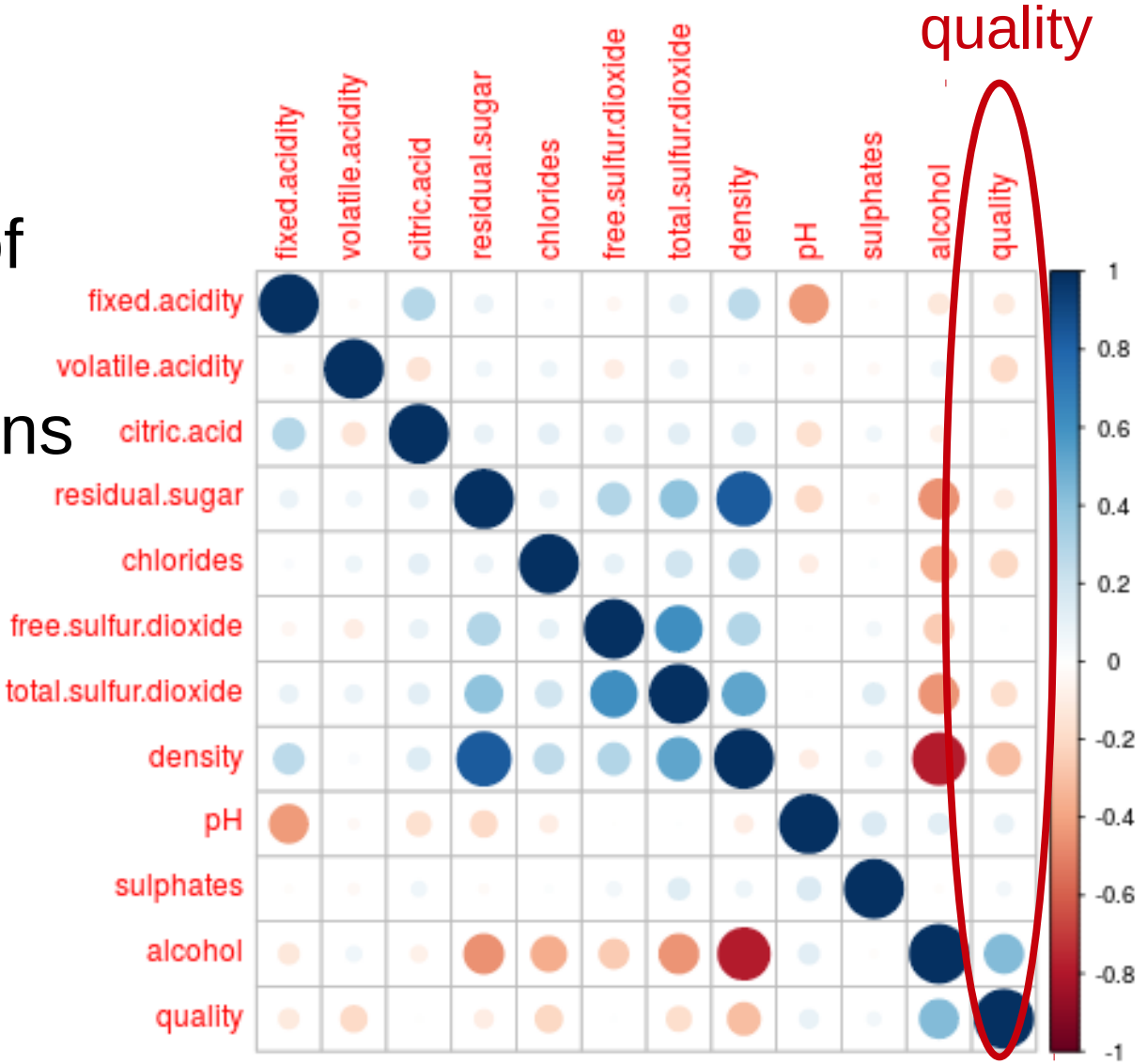


Reason for poor worst wines identification: no features strongly correlated to poor quality.

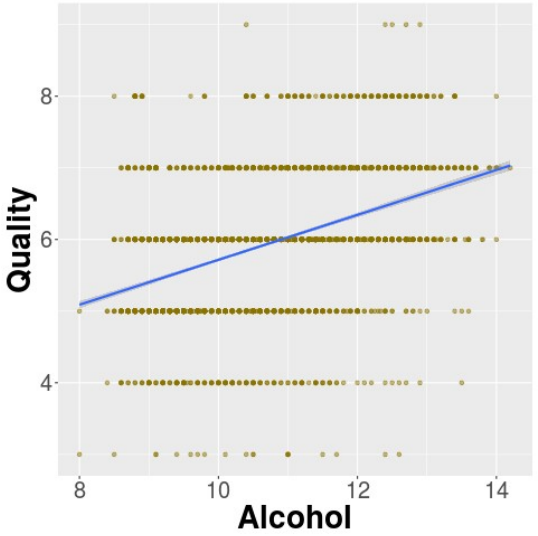
White Wines

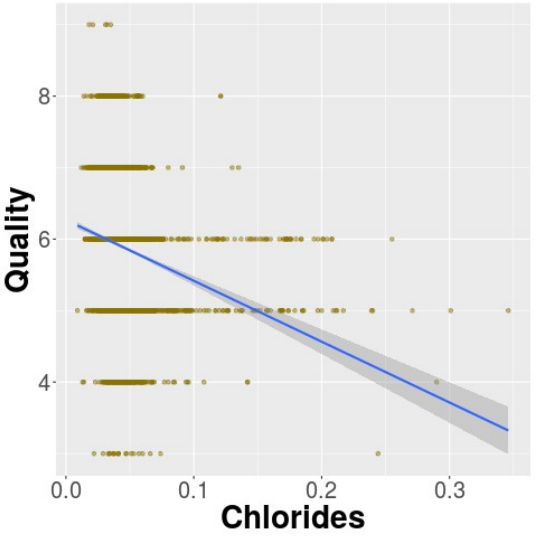
Correlation Matrix

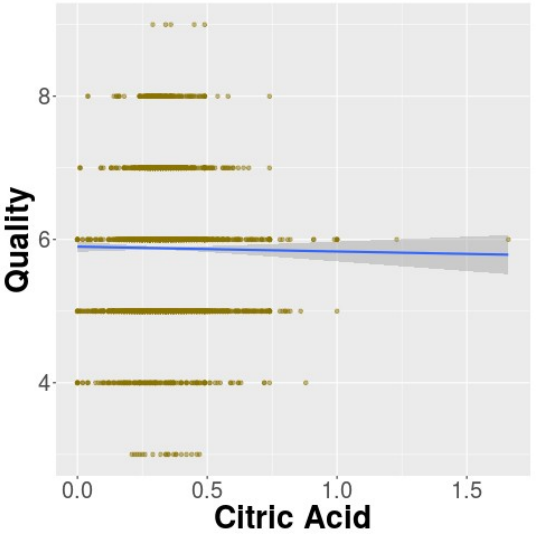
- Alcohol, density strongest indicators of quality
- Many weak correlations to quality

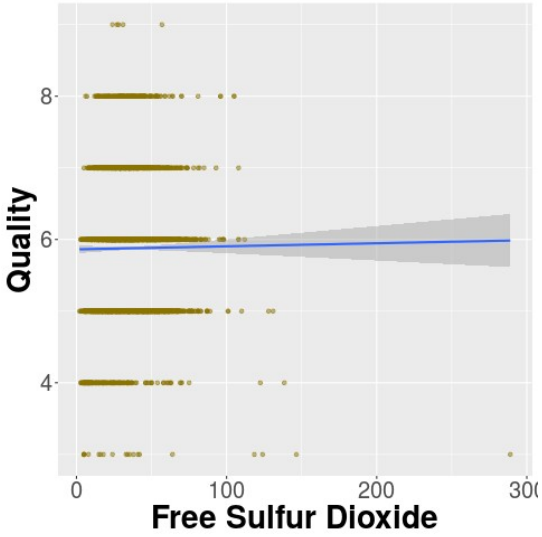


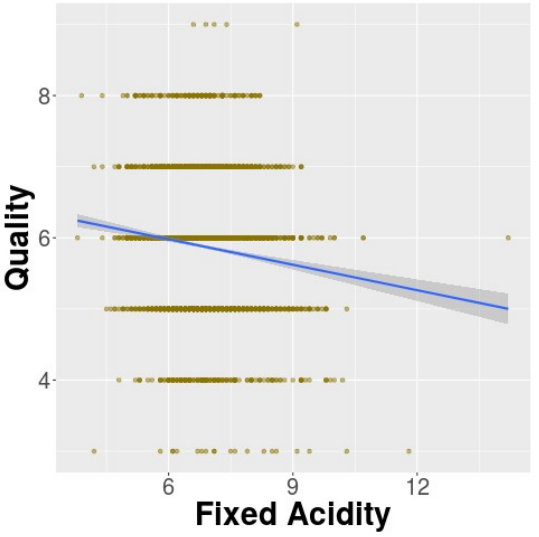
Visualize Relationships

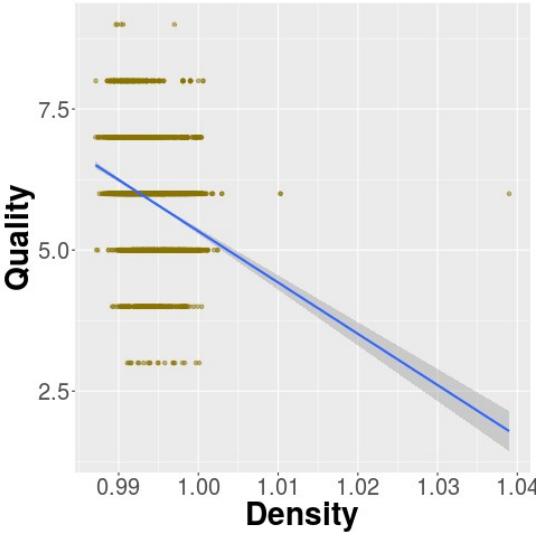




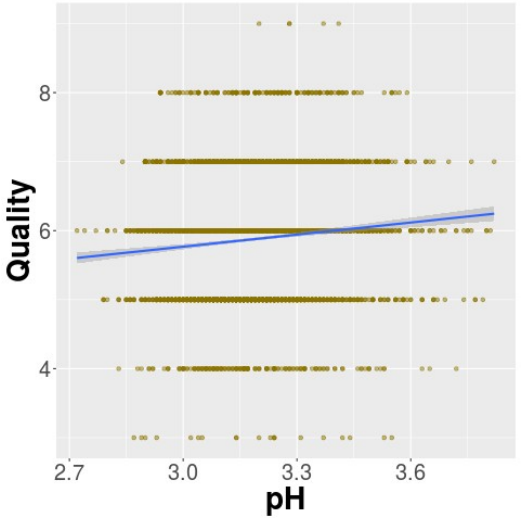


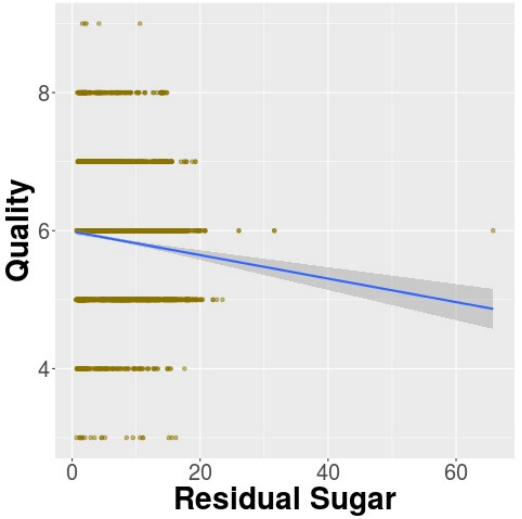


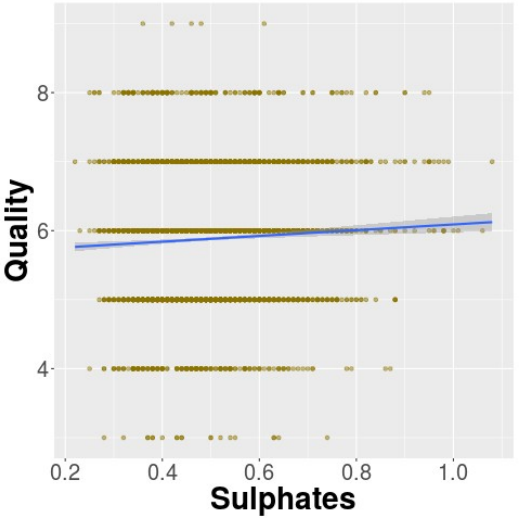


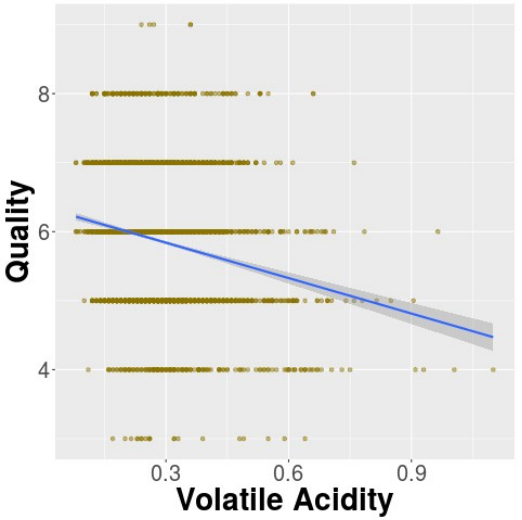


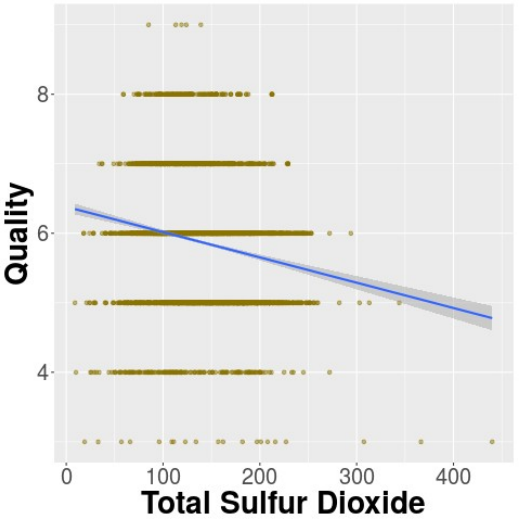
Visualization Continued











Identifying White Wines By Quality

- Train SVM with all 11 features
- CV results
  - Optimal value: C = 1, γ= 1
  - Results in 63.2% accuracy on CV set
- 63.2% accuracy on test set
  - 100% chance 4 prediction is correct
  - 4% of 4's correctly identified
  - 74.6% chance 5 prediction is correct
  - 44.8% of 5's correctly identified
  - 73.8% chance 6 prediction is correct
  - 89.3% of 6's correctly identified
  - 100% chance 8 prediction is correct
  - 30% of 8's correctly identified
  - No 3 or 9 predictions made

Cross Validation

C

γ	0.01	0.03	0.1	0.3	1	3	10	30	100
0.01	0.442	0.442	0.461	0.528	0.546	0.552	0.559	0.572	0.573
0.03	0.442	0.442	0.525	0.55	0.566	0.567	0.58	0.596	0.59
0.1	0.442	0.444	0.54	0.569	0.577	0.599	0.6	0.6	0.587
0.3	0.442	0.442	0.509	0.569	0.605	0.61	0.614	0.615	0.618
1	0.442	0.442	0.444	0.491	0.632	0.626	0.63	0.629	0.63
3	0.442	0.442	0.442	0.454	0.614	0.621	0.623	0.623	0.623
10	0.442	0.442	0.442	0.45	0.599	0.607	0.607	0.607	0.607
30	0.442	0.442	0.442	0.449	0.585	0.589	0.589	0.589	0.589
100	0.442	0.442	0.442	0.449	0.578	0.578	0.578	0.578	0.578

Test

True

Pred	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0
5	0	5	129	39	1	0	0
6	5	19	156	399	102	16	0
7	0	0	3	9	81	5	1
8	0	0	0	0	0	9	0
9	0	0	0	0	0	0	0

Identifying Best White Wines

- Train SVM with all 11 features
  - Class 0: Not best wines
  - Class 1: best wines (7, 8, or 9)
- CV results
  - Optimal value: C = 10, γ= 1
  - Results in 0.637 F score
- Test results:
  - 0.580 F score
  - 5 5's, 28 6's, 88 7's, 13 8's, 1 9 identified as top wines (Class 1)
    - Only 33 out of 102 predicted top wines were not of top quality
    - No terrible wines (3,4) were false positives
    - Sole 9 in test set identified as top wine
  - 98 7's, 17 8's, 0 9's missed identification (Class 0)
    - Approximately half of top wines in the test set were successfully identified

Cross Validation

γ \ C	0.01	0.03	0.1	0.3	1	3	10	30	100
0.01	NaN	NaN	NaN	NaN	0.128	0.375	0.45	0.487	0.521
0.03	NaN	NaN	NaN	0.289	0.414	0.447	0.538	0.547	0.567
0.1	NaN	NaN	0.253	0.428	0.517	0.536	0.582	0.602	0.619
0.3	NaN	NaN	0.226	0.404	0.536	0.615	0.595	0.615	0.632
1	NaN	NaN	NaN	0.24	0.605	0.624	0.637	0.637	0.637
3	NaN	NaN	NaN	0.058	0.505	0.535	0.535	0.535	0.535
10	NaN	NaN	NaN	0.058	0.446	0.472	0.472	0.472	0.472
30	NaN	NaN	NaN	0.048	0.428	0.44	0.44	0.44	0.44
100	NaN	NaN	NaN	0.048	0.428	0.428	0.428	0.428	0.428

Test

Pred \ True	0	1
0	730	115
1	33	102

Identifying Worst Wines

- An average consumer might want to avoid bad wines (3,4 quality)
- Train SVM with all 11 features
  - Class 0: Not bad wines
  - Class 1: Bad wines
- CV results
  - Optimal value: C = 100, γ= 0.01
  - 0.261 F score on CV set
- Test results
  - 0.065 F score
  - 1 out of 28 bad wines successfully flagged
  - 2 False positives were 5's

Cross Validation

C

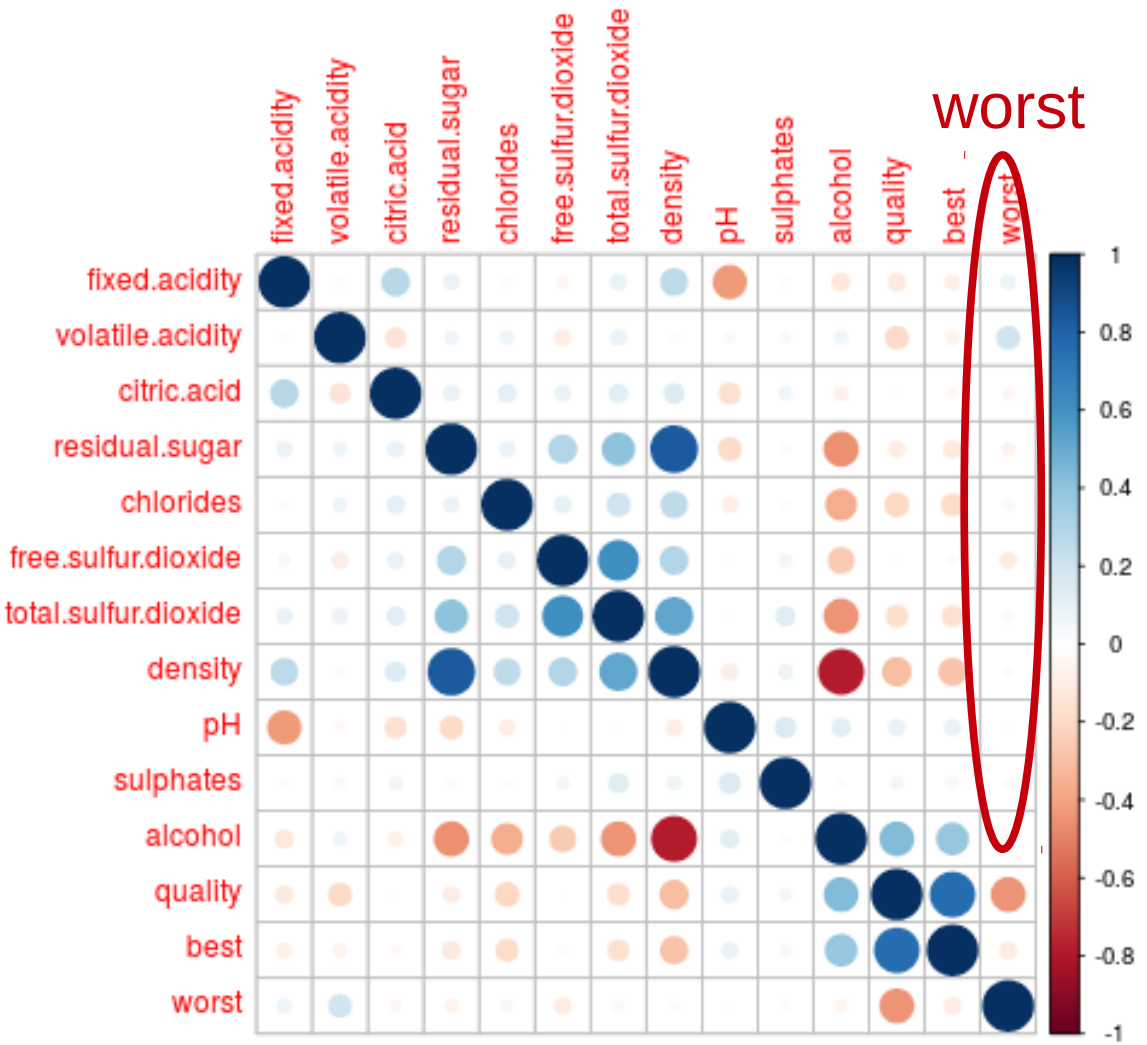
γ	0.01	0.03	0.1	0.3	1	3	10	30	100
0.01	NaN	NaN	NaN	NaN	NaN	NaN	0.056	0.154	0.261
0.03	NaN	NaN	NaN	NaN	NaN	0.053	0.213	0.213	0.208
0.1	NaN	NaN	NaN	NaN	0.15	0.178	0.213	0.214	0.194
0.3	NaN	NaN	NaN	NaN	0.108	0.217	0.2	0.235	0.231
1	NaN	NaN	NaN	NaN	NaN	0.105	0.105	0.105	0.105
3	NaN	NaN	NaN	NaN	0.108	0.108	0.108	0.108	0.108
10	NaN	NaN	NaN	NaN	0.056	0.108	0.108	0.108	0.108
30	NaN	NaN	NaN	NaN	0.056	0.108	0.108	0.108	0.108
100	NaN	NaN	NaN	NaN	0.056	0.056	0.056	0.056	0.056

Test

True

Pred	0	1
0	950	27
1	2	1

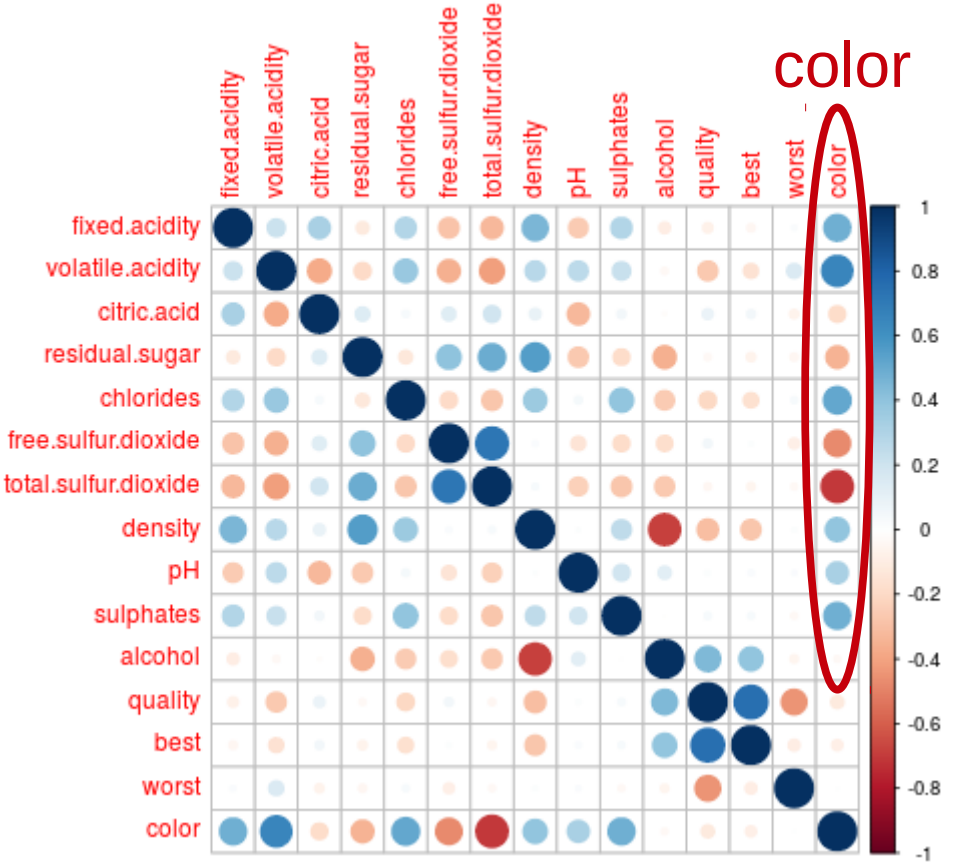
Full White Wine Correlation Matrix
With Best, Worst



Reason for poor worst wines identification: no features strongly correlated to poor quality.

Red vs White Wine

- Goal: If wine chemistry known, sort wine into red or white, and then apply appropriate model to determine quality
- Correlations between wine color and all features except alcohol content
  - White wines: class 0
  - Red wines: class 1
  - White wines have more total sulfur dioxide, red wines have more volatile acidity



Wine Color SVM Results

- Available features are extremely helpful for distinguishing between wine color
  - White wines: class 0
  - Red wines: class 1
- CV results
  - Optimal value: C = 10, γ= 0.03
  - Results in 0.989 F score
- Test results
  - Only 0.04% of wines incorrectly identified

Cross Validation

C

γ	0.01	0.03	0.1	0.3	1	3	10	30	100
0.01	0.911	0.967	0.978	0.982	0.982	0.983	0.986	0.986	0.986
0.03	0.962	0.974	0.98	0.983	0.983	0.988	0.989	0.983	0.983
0.1	0.942	0.956	0.972	0.985	0.983	0.986	0.986	0.982	0.98
0.3	0.607	0.871	0.929	0.964	0.98	0.983	0.982	0.982	0.982
1	NaN	0.042	0.532	0.812	0.907	0.916	0.916	0.916	0.916
3	NaN	NaN	NaN	0.126	0.532	0.58	0.58	0.58	0.58
10	NaN	NaN	NaN	0.006	0.292	0.305	0.305	0.305	0.305
30	NaN	NaN	NaN	0.006	0.26	0.287	0.287	0.287	0.287
100	NaN	NaN	NaN	0.006	0.251	0.251	0.251	0.251	0.251

Test

True

Pred	0	1
0	978	4
1	1	317

Conclusions

- Using only basic wine chemistry, the color of a wine can be identified with about 99% accuracy
- After dividing by color, wines in the top 10% can be selected with 50-70% accuracy
  - Restaurants can order new top wines using chemistry data before experts have tasted the wine
- Need more relevant features to prevent bad wines from being identified as average wines