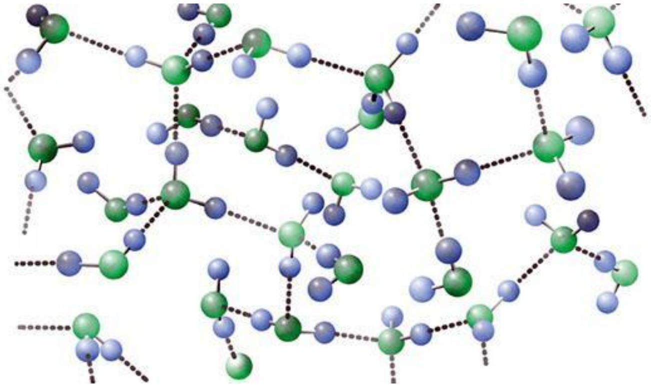
Modeling Wine Quality with Physicochemical Properties

Jason Witry, Jerome Doe, Armand Heydarian, Hang Zhao

Taste

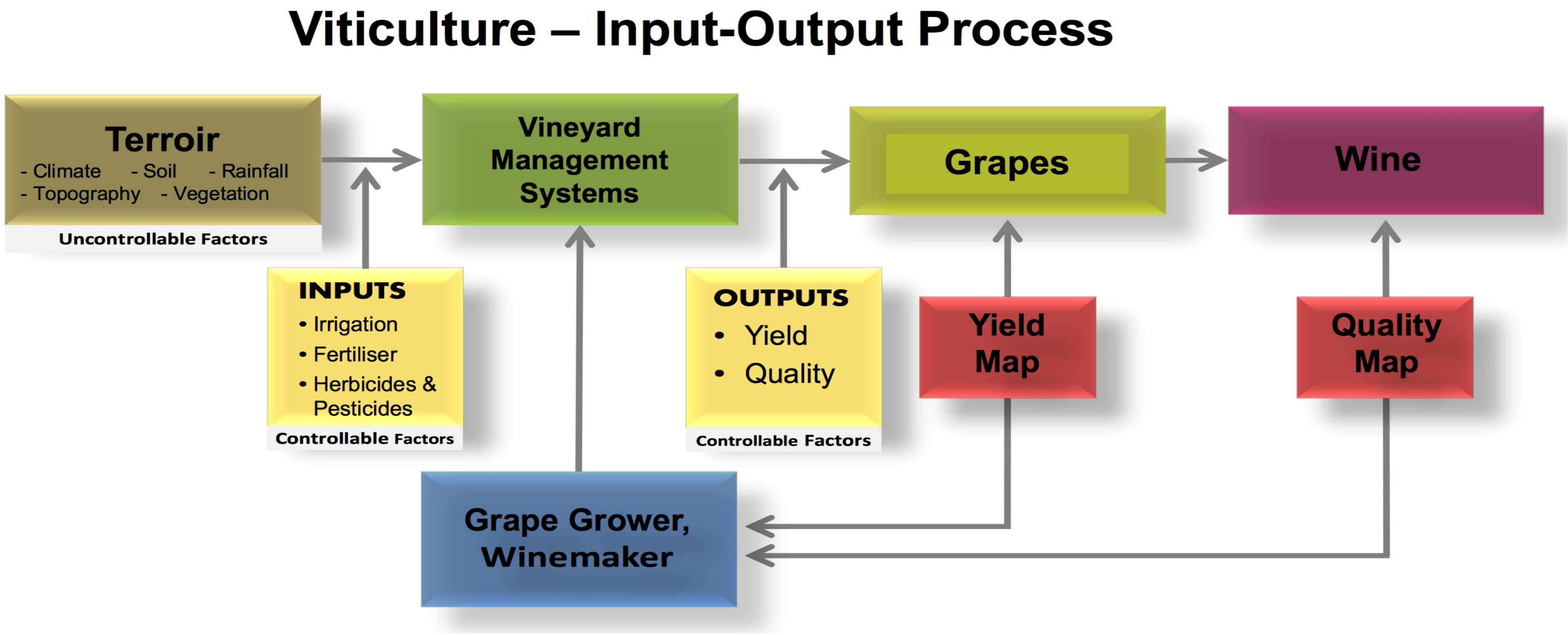

BUMBLE BEE
QUALITY YOU CAN SEE
Premium Quality
LIFT RING TO RIM, PULL BACK
SARDINES
IN HOT SAUCE
NET WT 3.75 oz (106 g)
INGREDIENTS: SARDINES, WATER, MALTODEXTRIN, ACETIC ACID, SALT, PAPRIKA, MODIFIED CORNSTARCH, XANTHAN GUM, POLYSORBATE 80, EXTRACT OF PAPRIKA, CARAMEL COLOR, SPICES AND EXTRACTS OF SPICES
DISTRIBUTED BY:
©BUMBLE BEE FOODS, LLC
SAN DIEGO, CA 92186
FOR NUTRITION FACTS
PLEASE WRITE TO
THE ABOVE ADDRESS
PRODUCT OF POLAND
5440

Taste



? ? ?

Factors Affecting Wine Quality



Methods to Ensure Wine Quality

- Have a team of sommelier come taste the wine and rate it
- Cons: expensive, time-consuming and may not always be able to get enough sample size
- DATA SCIENCE APPROACH!
- Cortez 2009

Modeling wine preferences by data mining from physicochemical properties

Paulo Cortez [a], António Cerdeira [b], Fernando Almeida [b], Telmo Matos [b], José Reis [a, b]

Show more

https://doi.org/10.1016/j.dss.2009.05.016 Get rights and content

Abstract

We propose a data mining approach to predict human wine taste preferences that is based on easily available analytical tests at the certification step. A large dataset (when compared to other studies in this domain) is considered, with white and red *vinho verde* samples (from Portugal). Three regression techniques were applied, under a computationally efficient procedure that performs simultaneous variable and model selection. The support vector machine achieved promising results, outperforming the multiple regression and neural network methods. Such model is useful to support the oenologist wine tasting evaluations and improve wine production. Furthermore, similar techniques can help in target marketing by modeling consumer tastes from niche markets.

“Modeling wine preferences by data mining from physicochemical properties”

- Predicting wine quality using SVM, Neural Networks, and Multiple Regression
- SVM performed the best
- Feature Selection: Sensitivity Analysis
- Results:

White wine predictions				
4	5	6	7	8
0	2	17	0	0
19	55	88	1	0
7	**833**	598	19	0
4	235	**1812**	144	3
0	18	414	**441**	7
0	3	71	43	**59**
0	1	3	2	0
63.3	72.6	60.3	67.8	85.5
90.0	93.3	81.9	90.3	96.2

Support Vector Machines

	White wine		
	MR	NN	SVM
MAD	0.59 ± 0.00	0.58 ± 0.00	**0.45** ± 0.00[a]
Accuracy$_{T=0.25}$ (%)	25.6 ± 0.1	26.5 ± 0.3	**50.3** ± 1.1[a]
Accuracy$_{T=0.50}$ (%)	51.7 ± 0.1	52.6 ± 0.3	**64.6** ± 0.4[a]
Accuracy$_{T=1.00}$ (%)	84.3 ± 0.1	84.7 ± 0.1	**86.8** ± 0.2[a]
Kappa$_{T=0.5}$ (%)	20.9 ± 0.1	23.5 ± 0.6	**43.9** ± 0.4[a]
Inputs ($\overline{I}$)	9.6	9.3	10.1
Model	–	$\overline{H} = 2.1$	$\overline{\gamma} = 2^{1.55}$
Time (s)	**551**	1339	30674

How can Wine Quality be predicted by physicochemical properties?

Portugal “Vinho Verde” Region

- White wines
- Data collected 2004-2007
- Characterized by small growers
- Wine released 3-6 months after harvesting

Variable Description

1. ***Fixed acidity:*** most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
2. ***Volatile acidity:*** the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. ***Citric acid:*** found in small quantities, citric acid can add 'freshness' and flavor to wines
4. ***pH:*** describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)
5. ***Residual sugar:*** the amount of sugar remaining after fermentation stops、
6. ***Chlorides:*** the amount of salt in the wine
7. ***Alcohol:*** the percent alcohol content of the wine

acid → vinegar → less sweet

Residual sugar → sugar → more sweet

Variable Description

8. Density: the density of water is close to that of water depending on the percent alcohol and sugar content

9. Free sulfur dioxide: the free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

10. Total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO_2 is mostly undetectable in wine, but at free SO_2 concentrations over 50 ppm, SO_2 becomes evident in the nose and taste of wine

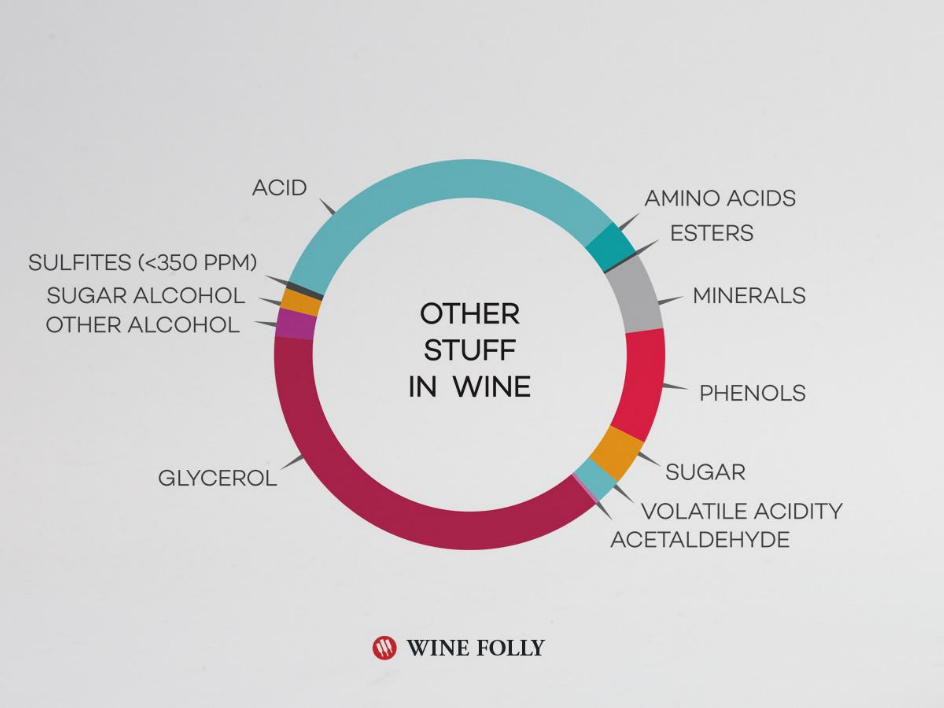
11. Sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

12. Quality: output variable (based on sensory data, score between 0 and 10)

just-struck match → pungent smell → SO_2 → worse taste

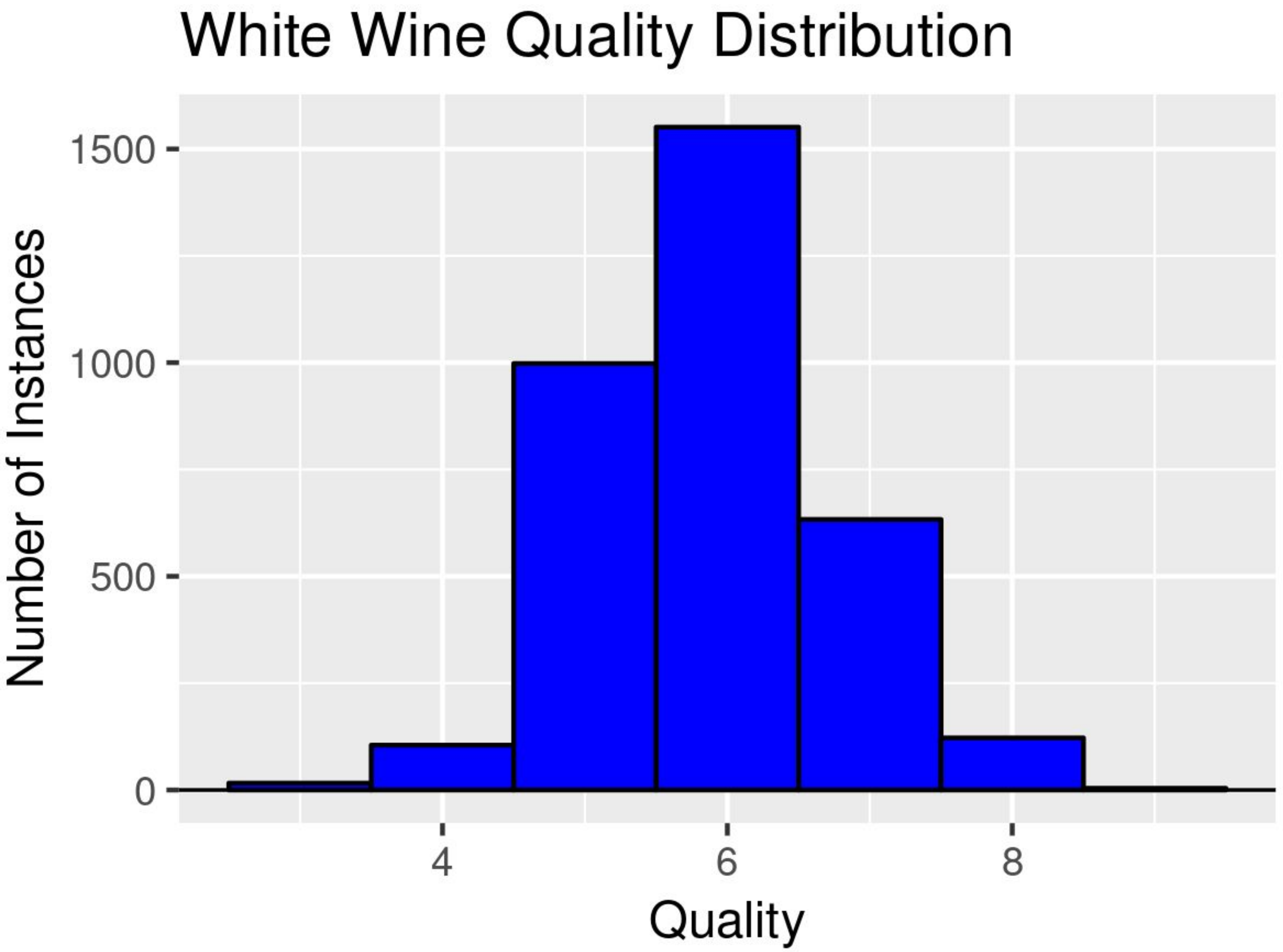
Limitations

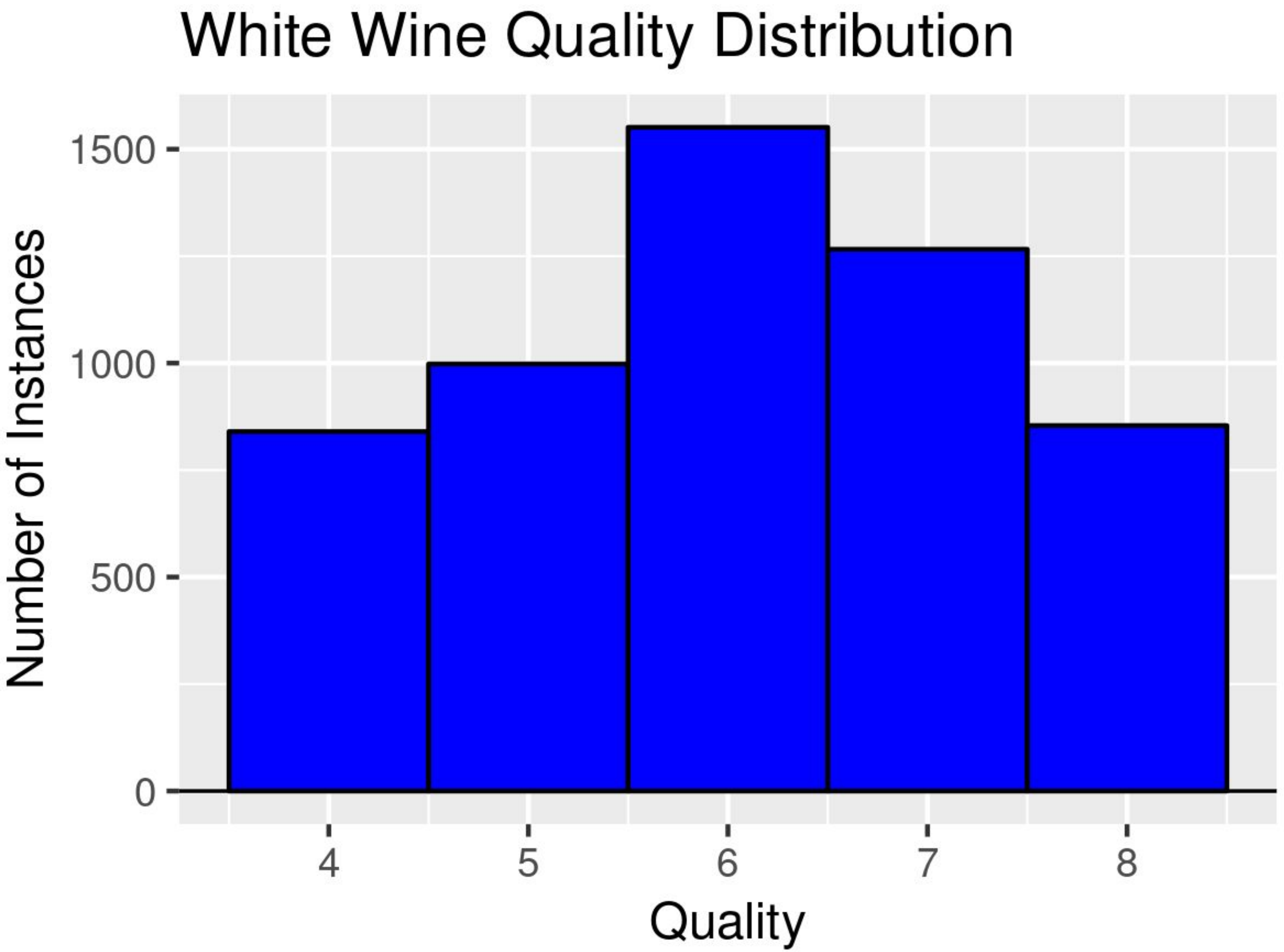
- In some cases, only 3 people assessed wine quality
- Although the data set may capture the majority of contributors to quality, there are hundreds of other compounds in wine that may have an effect



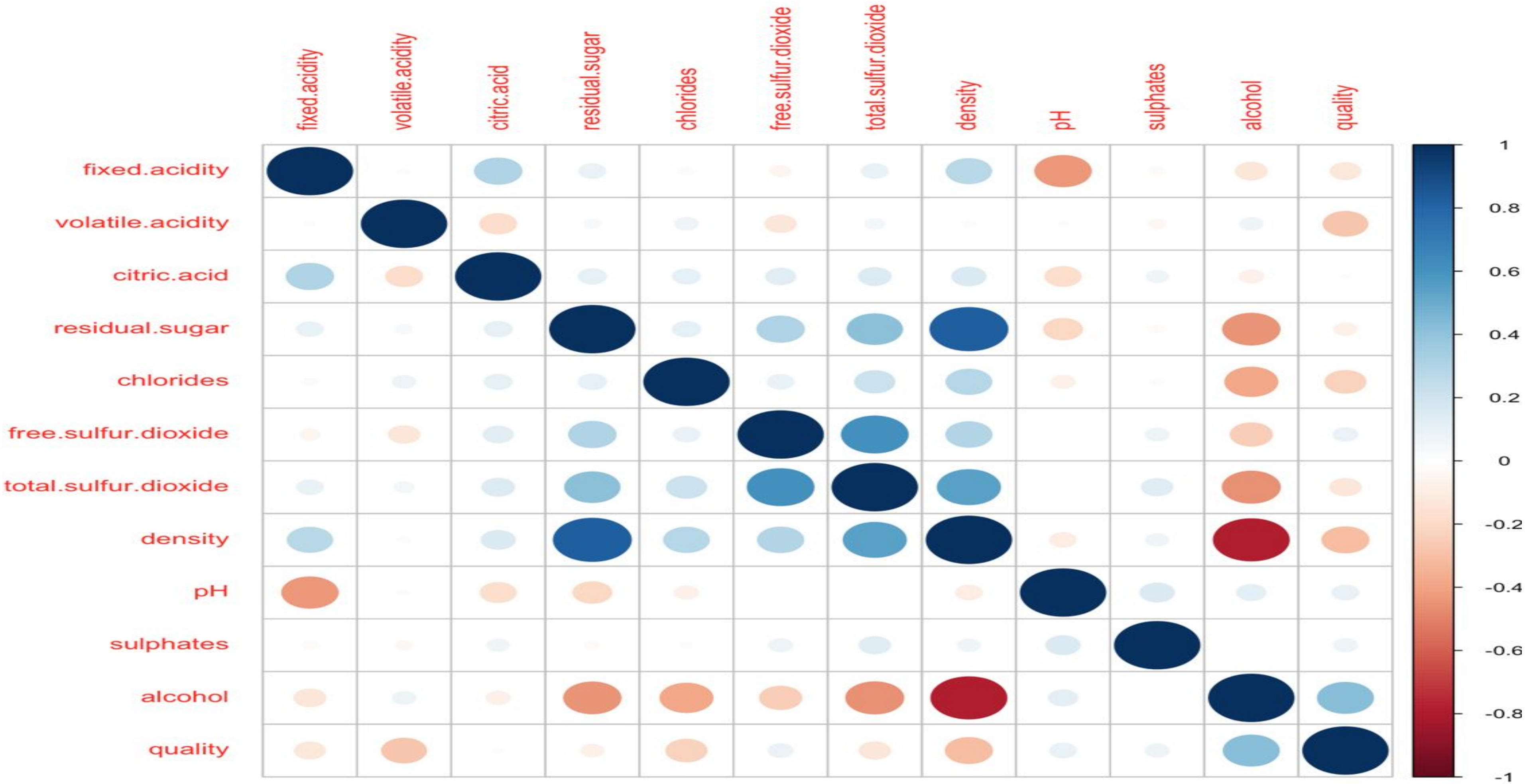
Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
fixed.acidity	4,898	6.855	0.844	4	6.3	7.3	14
volatile.acidity	4,898	0.278	0.101	0.080	0.210	0.320	1.100
citric.acid	4,898	0.334	0.121	0.000	0.270	0.390	1.660
residual.sugar	4,898	6.391	5.072	0.600	1.700	9.900	65.800
chlorides	4,898	0.046	0.022	0.009	0.036	0.050	0.346
free.sulfur.dioxide	4,898	35.308	17.007	2	23	46	289
total.sulfur.dioxide	4,898	138.361	42.498	9	108	167	440
density	4,898	0.994	0.003	0.987	0.992	0.996	1.039
pH	4,898	3.188	0.151	3	3.1	3.3	4
sulphates	4,898	0.490	0.114	0.220	0.410	0.550	1.080
alcohol	4,898	10.514	1.231	8.000	9.500	11.400	14.200
quality	4,898	5.878	0.886	3	5	6	9


White Wine Quality Distribution
1500
1000
500
0
4
6
8
Number of Instances
Quality

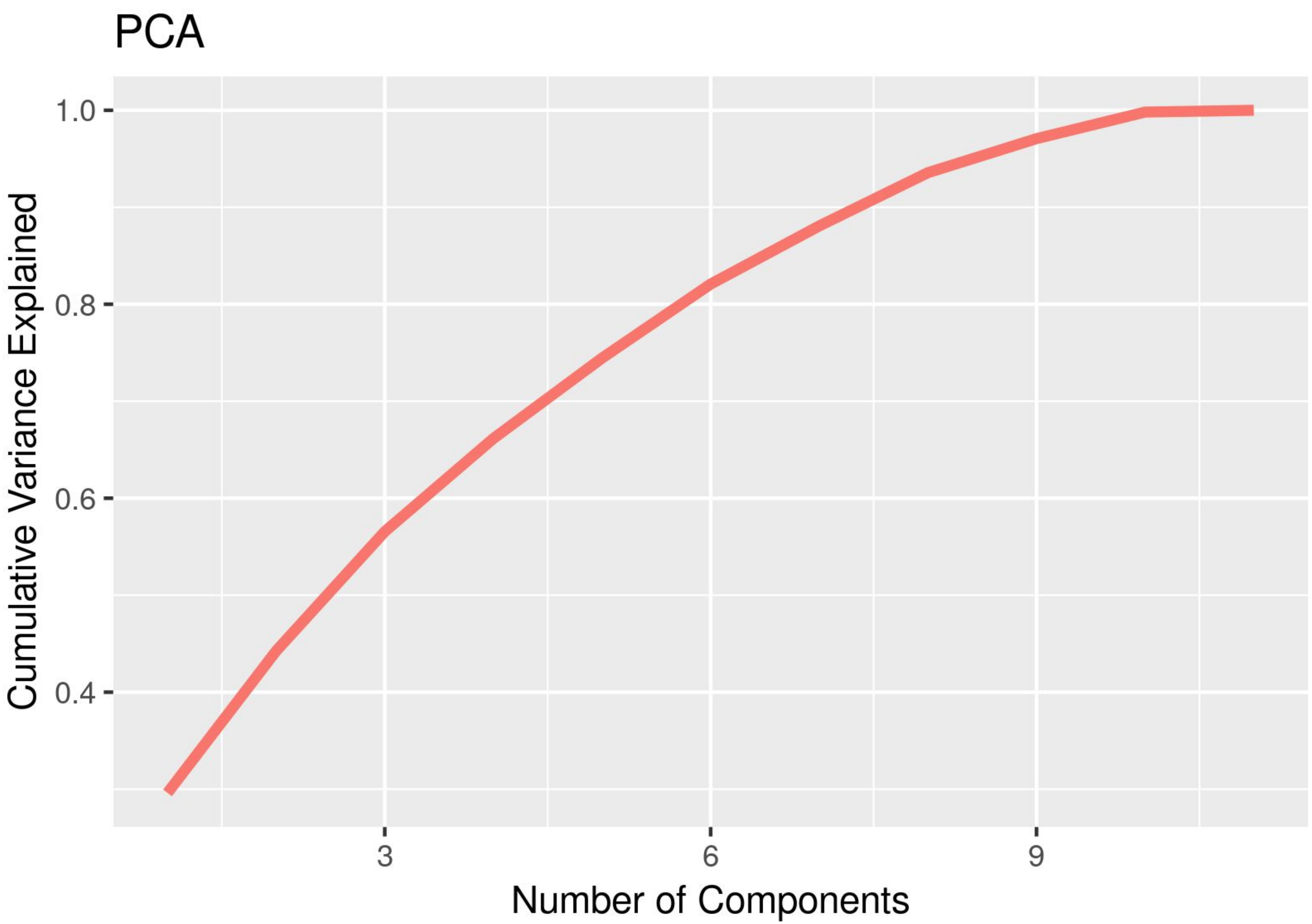

White Wine Quality Distribution
Number of Instances
1500
1000
500
0
4
5
6
7
8
Quality

Correlation Plot



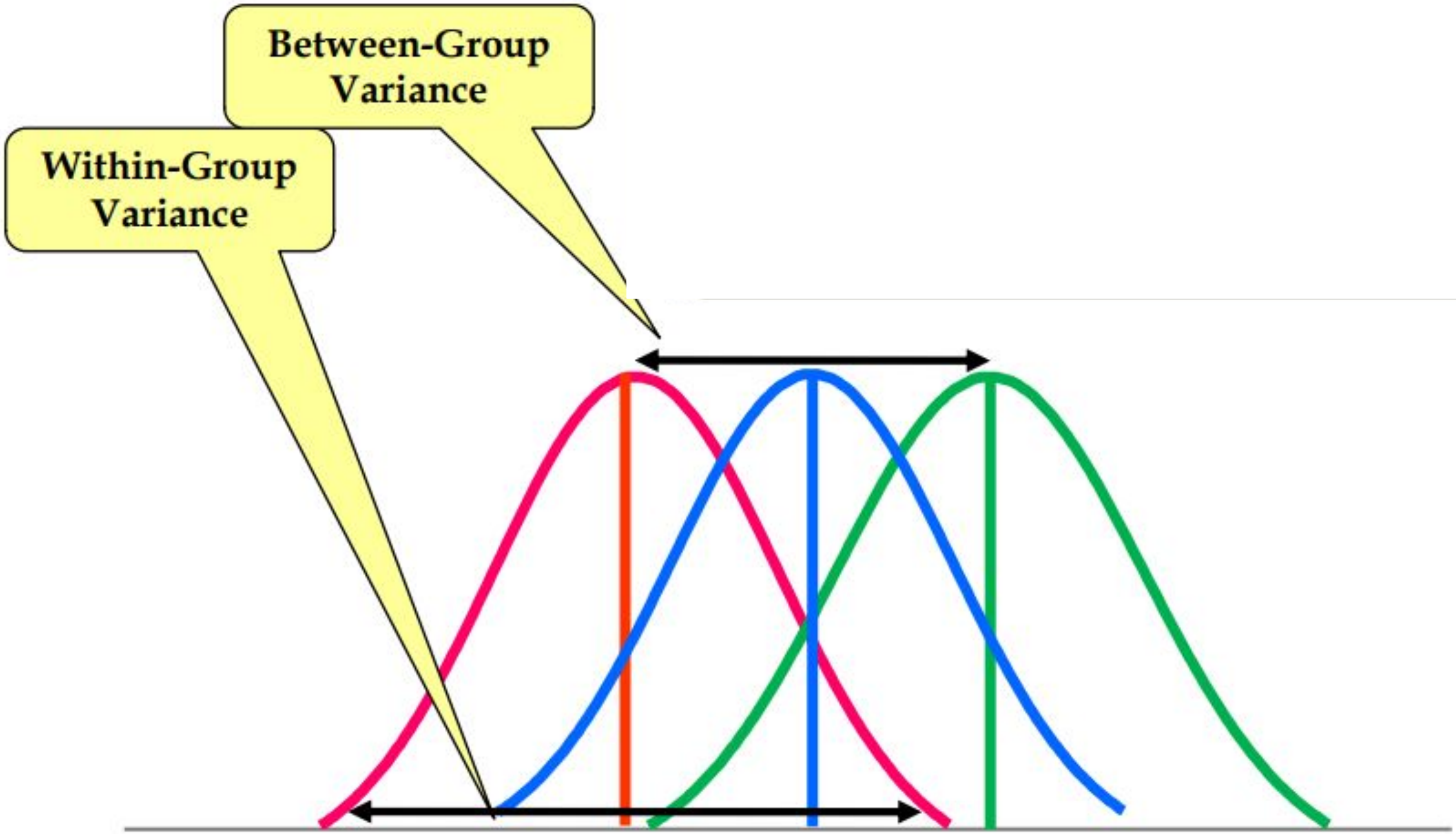
Principal Components Analysis

- Steady increase
- Exclude PC11



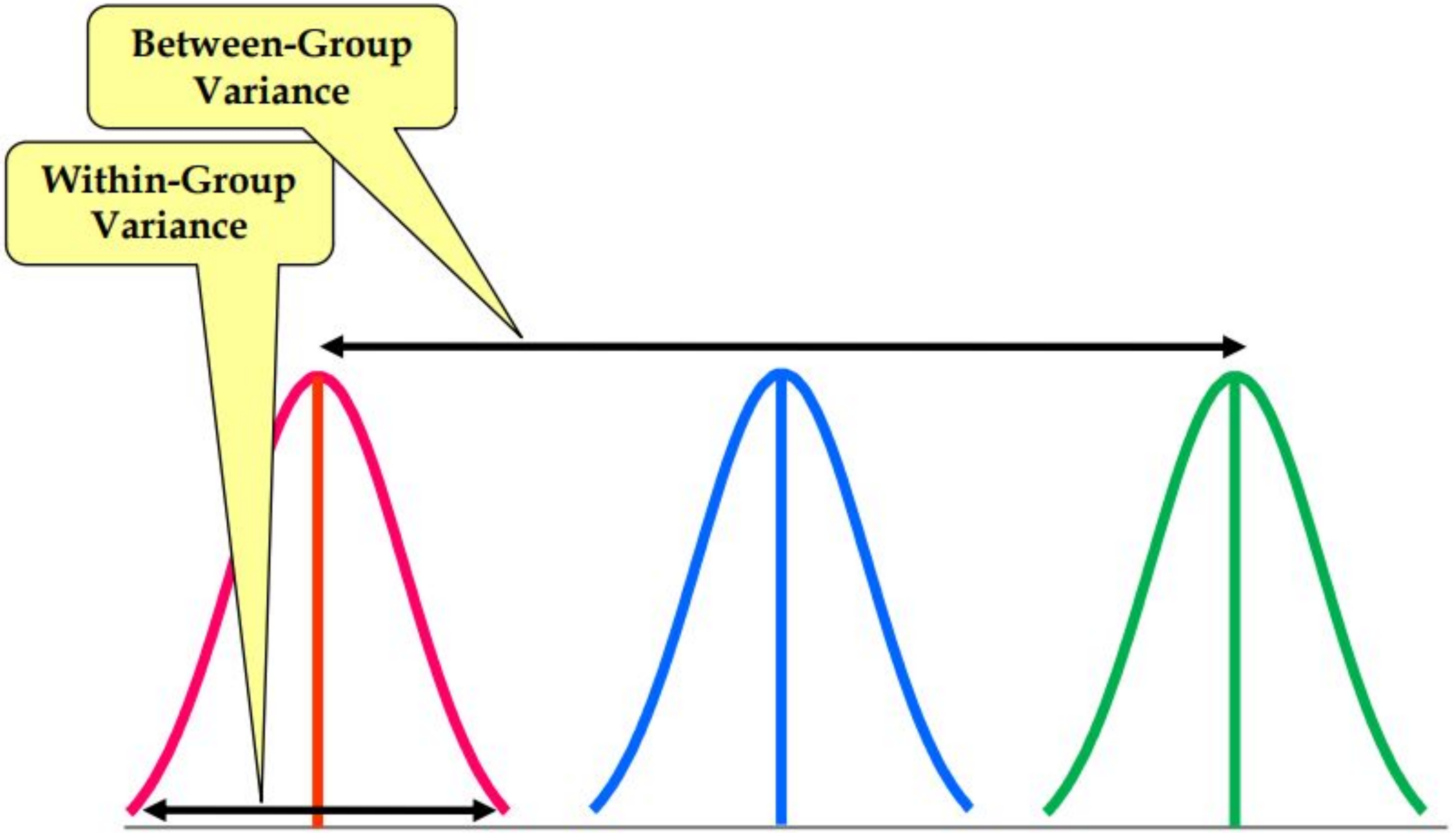
Anova Analysis

- Bad predictor
- Low F-stat
- Higher p-value



Anova Analysis

- Good predictor!
- High F-stat
- Low p-value



Anova Results

```
                               fvals
 [1,] "alcohol"                "232.254312769435"
 [2,] "density"                "103.474858788422"
 [3,] "volatile.acidity"       "63.1971816854159"
 [4,] "chlorides"              "42.1557508453574"
 [5,] "total.sulfur.dioxide"   "40.7707009126757"
 [6,] "residual.sugar"         "22.1156588646865"
 [7,] "free.sulfur.dioxide"    "20.6019121498521"
 [8,] "fixed.acidity"          "11.6229269392978"
 [9,] "pH"                     "11.546538492975"
[10,] "sulphates"              "5.2186585535878"
[11,] "citric.acid"            "3.0520898260995"
                               pvals
 [1,] "fixed.acidity"          "2.24784040834199e-09"
 [2,] "volatile.acidity"       "1.33208434454332e-51"
 [3,] "citric.acid"            "0.0159922417751507"
 [4,] "residual.sugar"         "4.79173802063143e-18"
 [5,] "chlorides"              "1.4868966352724e-34"
 [6,] "free.sulfur.dioxide"    "8.60466932069265e-17"
 [7,] "total.sulfur.dioxide"   "2.02619064828755e-33"
 [8,] "density"                "2.8139783530395e-83"
 [9,] "pH"                     "2.59727367807412e-09"
[10,] "sulphates"              "0.000344250587062572"
[11,] "alcohol"                "1.55706345424093e-176"
```

```
              fvals
 [1,] "PC3"   "248.13841872865l"
 [2,] "PC1"   "178.482206774759"
 [3,] "PC9"   "137.95145977849"
 [4,] "PC4"   "97.527290615611"
 [5,] "PC2"   "75.5393997791381"
 [6,] "PC8"   "40.5774142523372"
 [7,] "PC5"   "27.8038732730827"
 [8,] "PC11"  "16.4546941675465"
 [9,] "PC6"   "6.34280705371403"
[10,] "PC7"   "2.95603104028173"
[11,] "PC10"  "0.773369508254746"
              pvals
 [1,] "PC1"   "6.40408044743834e-144"
 [2,] "PC2"   "1.98585792591803e-62"
 [3,] "PC3"   "3.96739779676727e-196"
 [4,] "PC4"   "2.76592789698053e-80"
 [5,] "PC5"   "6.86895868182001e-23"
 [6,] "PC6"   "4.34380917209942e-05"
 [7,] "PC7"   "0.0187933899001202"
 [8,] "PC8"   "1.4820548605322e-33"
 [9,] "PC9"   "1.79896117894 84e-112"
[10,] "PC10"  "0.542356783962971"
[11,] "PC11"  "2.081579154564e-13"
```

Models

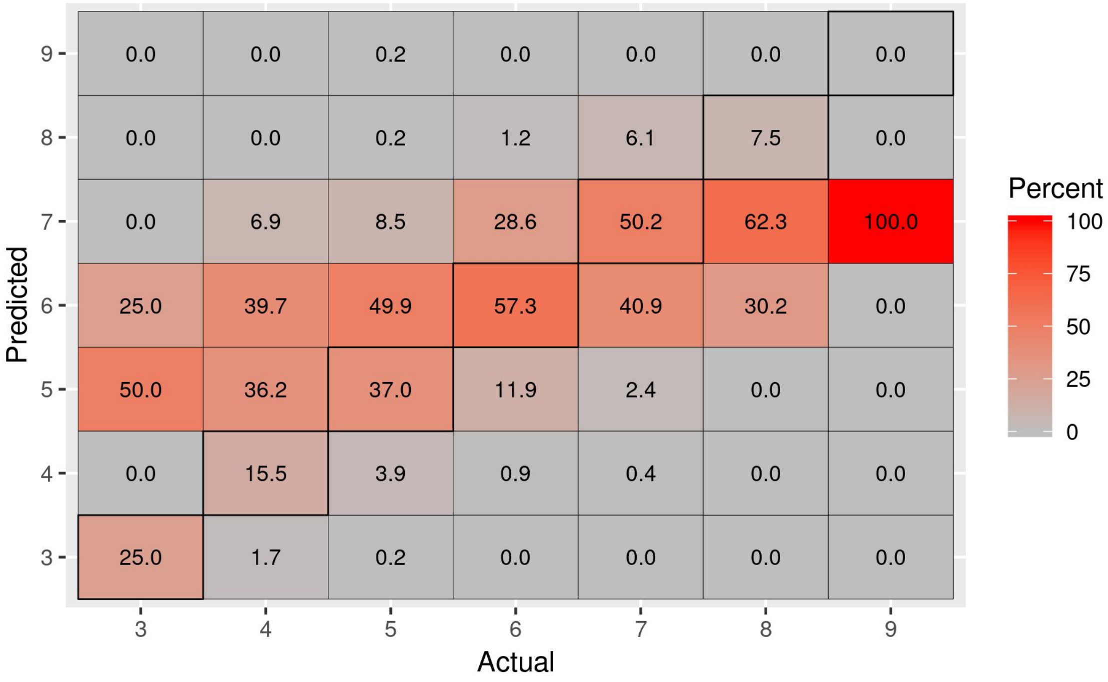
- Linear Regression
- Principal Components Regression
- Ordered Logistic Regression

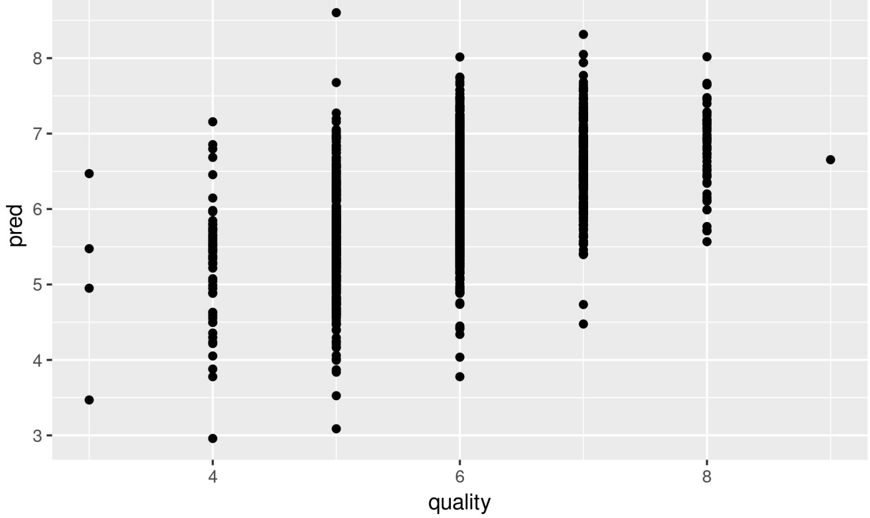
Linear Regression

- Call Selected Variables
- All coefficients significant

```
Call:
lm(formula = quality ~ volatile.acidity + citric.acid + residual.sugar +
    free.sulfur.dioxide + pH + alcohol, data = winequality_bal_sc)

Residuals:
     Min       1Q   Median       3Q      Max
-3.11075 -0.66950 -0.01801  0.67349  2.80580

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.04873    0.01352 447.467  < 2e-16 ***
volatile.acidity    -0.36875    0.01397 -26.393  < 2e-16 ***
citric.acid         -0.05050    0.01416  -3.565 0.000366 ***
residual.sugar       0.24623    0.01579  15.596  < 2e-16 ***
free.sulfur.dioxide  0.25197    0.01470  17.136  < 2e-16 ***
pH                   0.10114    0.01421   7.116 1.25e-12 ***
alcohol              0.73377    0.01494  49.101  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1 on 5466 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared:  0.3881,    Adjusted R-squared:  0.3874
F-statistic: 577.7 on 6 and 5466 DF,  p-value: < 2.2e-16
```

Linear Regression Results



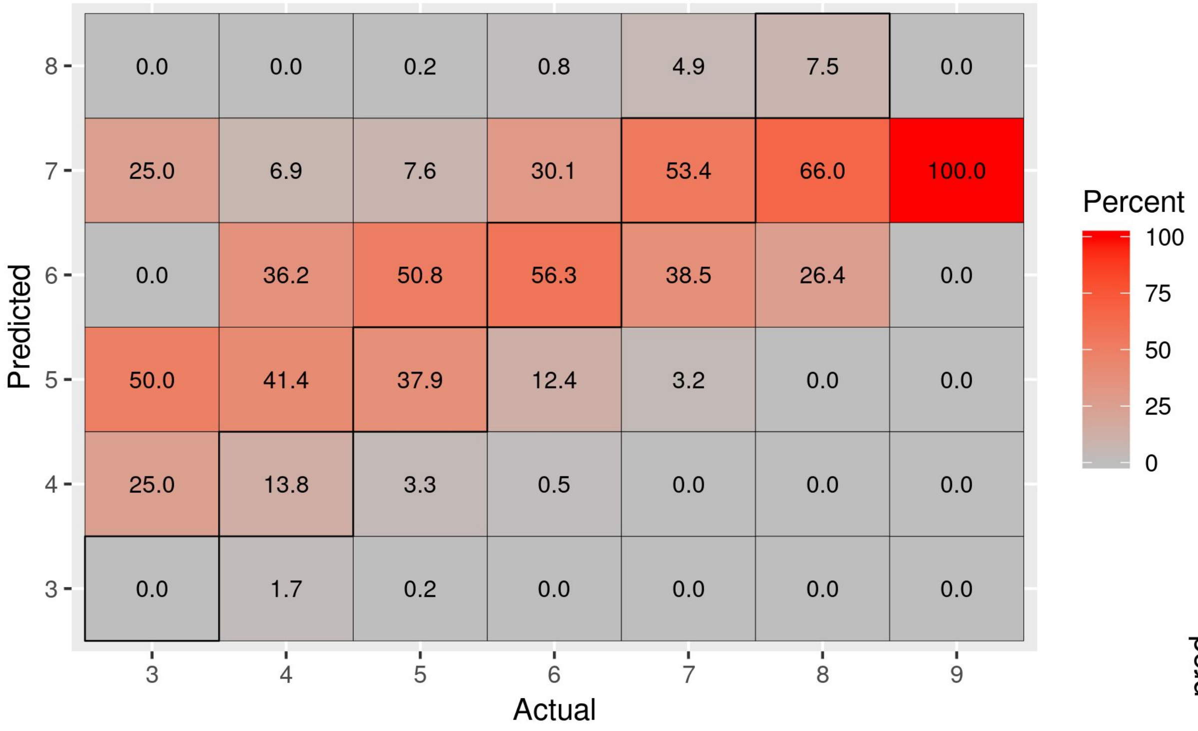


Principal Components Regression

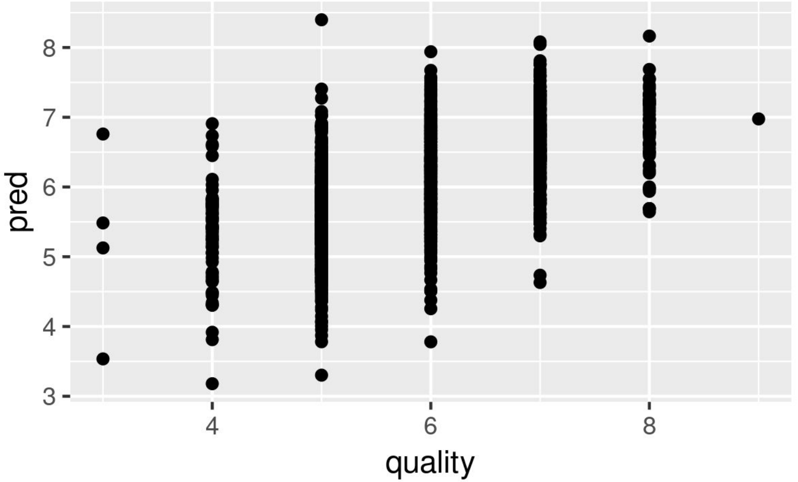
- Call Selected PC’s
- All coefficients significant

```
Call:
lm(formula = df1 ~ PC1 + PC2 + PC3 + PC4 + PC8 + PC9 + PC11,
    data = PCA_Wine)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.2589 -0.6611 -0.0230  0.6664  3.9961 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.050140   0.013359 452.901   <2e-16 ***
PC1          0.165881   0.007404  22.405   <2e-16 ***
PC2         -0.195104   0.010559 -18.478   <2e-16 ***
PC3         -0.396257   0.011501 -34.455   <2e-16 ***
PC4         -0.293908   0.013042 -22.536   <2e-16 ***
PC8          0.274815   0.017222  15.957   <2e-16 ***
PC9          0.609005   0.021529  28.287   <2e-16 ***
PC11         0.815146   0.094375   8.637   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9873 on 5455 degrees of freedom
  (46 observations deleted due to missingness)
Multiple R-squared:  0.4025,    Adjusted R-squared:  0.4017 
F-statistic: 524.9 on 7 and 5455 DF,  p-value: < 2.2e-16
```

PCR Results



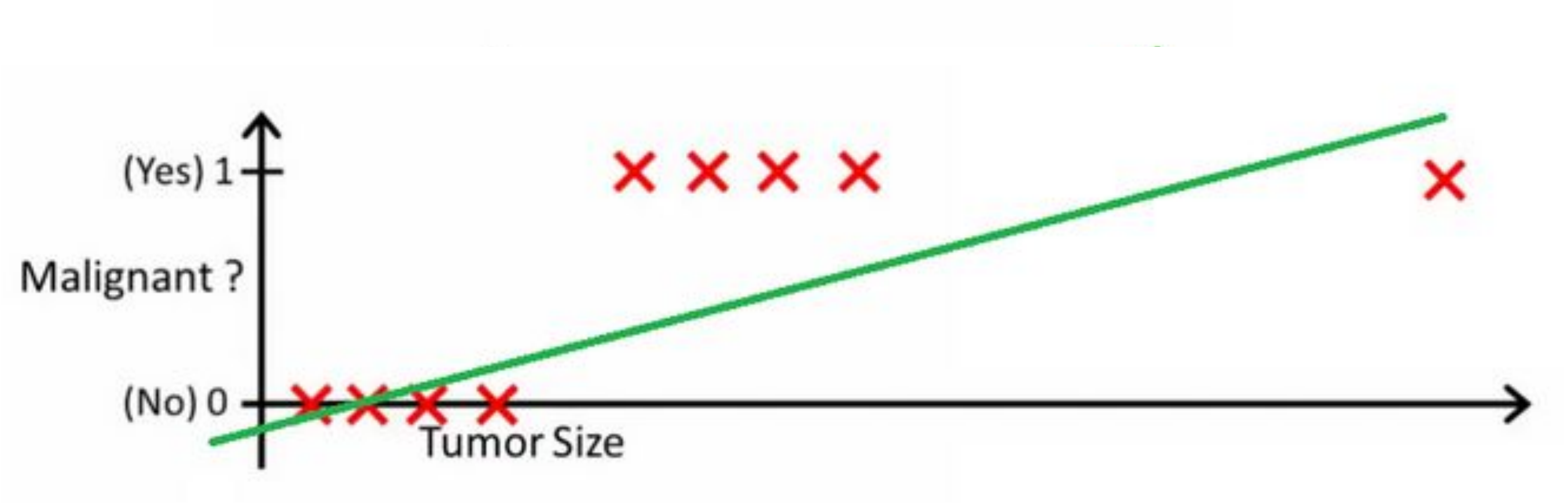
Predicted
Actual
Percent
100
75
50
25
0
3
4
5
6
7
8
9
0.0 0.0 0.2 0.8 4.9 7.5 0.0
25.0 6.9 7.6 30.1 53.4 66.0 100.0
0.0 36.2 50.8 56.3 38.5 26.4 0.0
50.0 41.4 37.9 12.4 3.2 0.0 0.0
25.0 13.8 3.3 0.5 0.0 0.0 0.0
0.0 1.7 0.2 0.0 0.0 0.0 0.0



pred
quality
3
4
5
6
7
8

Limitations of Regression

- Interpretability of coefficients
- Where should the threshold between classes be? 4.5? 4.8?



Ordered Logistic Regression

- Ordinal Target with Numeric Predictors
- Log odds measured differently

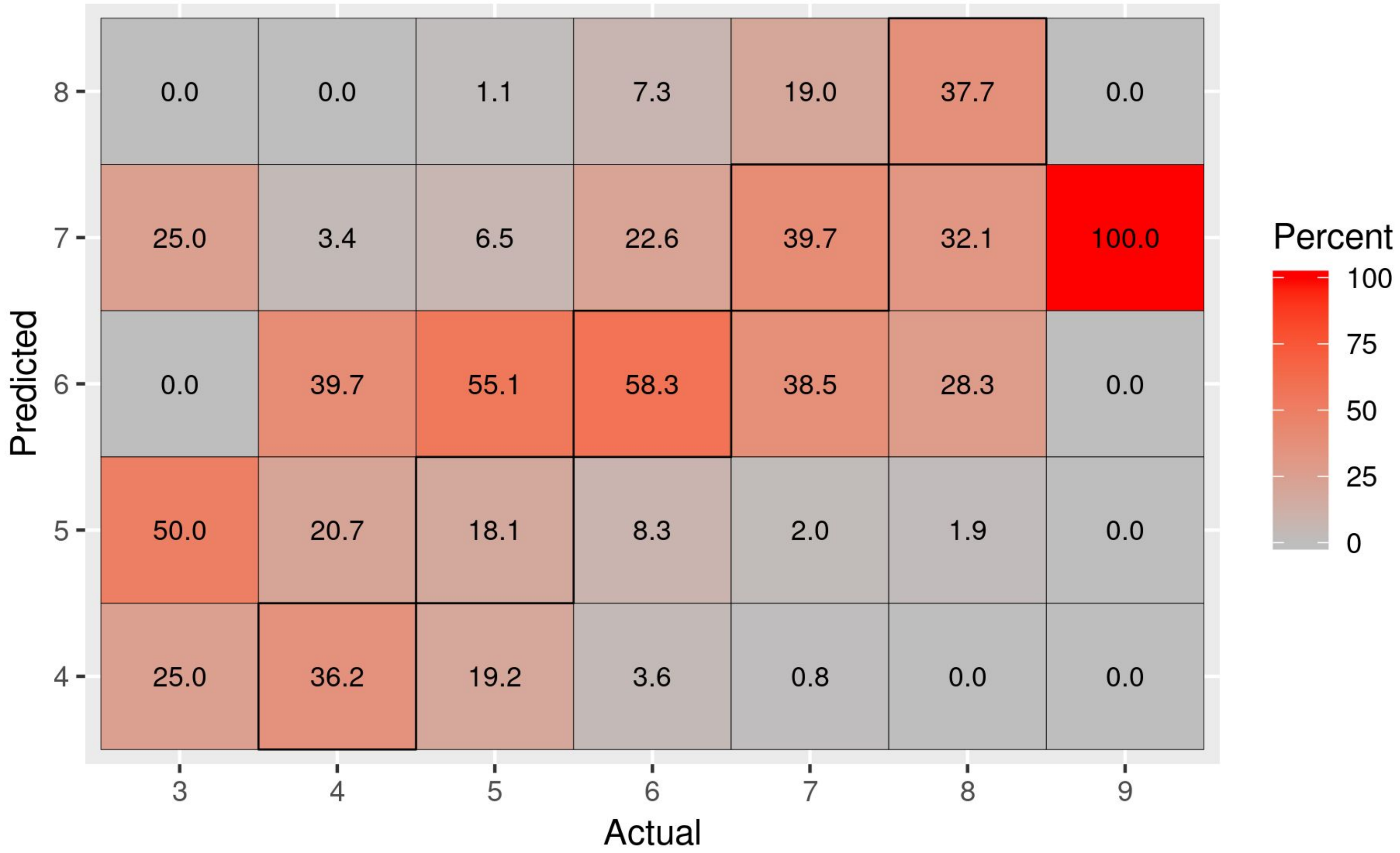
$$
\begin{aligned}
\text{poor,} \quad & \log \frac{p_1}{p_2+p_3+p_4+p_5}, & 0 \\
\text{poor or fair,} \quad & \log \frac{p_1+p_2}{p_3+p_4+p_5}, & 1 \\
\text{poor, fair, or good,} \quad & \log \frac{p_1+p_2+p_3}{p_4+p_5}, & 2 \\
\text{poor, fair, good, or very good,} \quad & \log \frac{p_1+p_2+p_3+p_4}{p_5}, & 3
\end{aligned}
$$

Ordered Logistic Regression

- All errors less than coefficients
- Variables Selected

```
Call:
polr(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar
+
    free.sulfur.dioxide + total.sulfur.dioxide + pH + density,
    data = winequality_bal, Hess = TRUE)

Coefficients:
                          Value Std. Error   t value
fixed.acidity         5.164e-01  0.0341135    15.138
volatile.acidity     -5.519e+00  0.2503980   -22.040
residual.sugar        3.848e-01  0.0060713    63.388
free.sulfur.dioxide   2.313e-02  0.0021182    10.920
total.sulfur.dioxide  1.895e-03  0.0008408     2.254
pH                    3.868e+00  0.1897306    20.386
density              -8.307e+02  0.3675593 -2259.972

Intercepts:
    Value      Std. Error t value
4|5  -810.2089     0.3777 -2144.8887
5|6  -808.7376     0.3782 -2138.4250
6|7  -807.0399     0.3797 -2125.5306
7|8  -805.4540     0.3813 -2112.3660

Residual Deviance: 14463.34
AIC: 14485.34
(46 observations deleted due to missingness)
```

Ordered Logistic Regression



Conclusions

- ANOVA Analysis does not yield much information for feature selection
- PCA with ANOVA allows us to drop one variable, not a significant difference however.
- Linear Regression fits the data fairly well, with around 46.2% accuracy. This is better than random classification.
- Principal Components Regression does not improve the accuracy much, at 46.4%.
- Ordinal Logistic Regression leads to an accuracy of 40.1%
- Ordinal Logistic Regression fits the ends better, while Regression fits the middle classes better
- We did not achieve the same accuracy as Cortez et al, but we can predict quality at a higher rate than random prediction.
- Future steps will include incorporating k-fold cross validation with Ordinal Logistic Regression. We believe this is the best model for this dataset, in spite of the accuracy differences because it is a more suited prediction for future data.

References

https://stats.stackexchange.com/questions/22381/why-not-approach-classification-through-regression

https://ragrawal.wordpress.com/2011/05/16/visualizing-confusion-matrix-in-r/

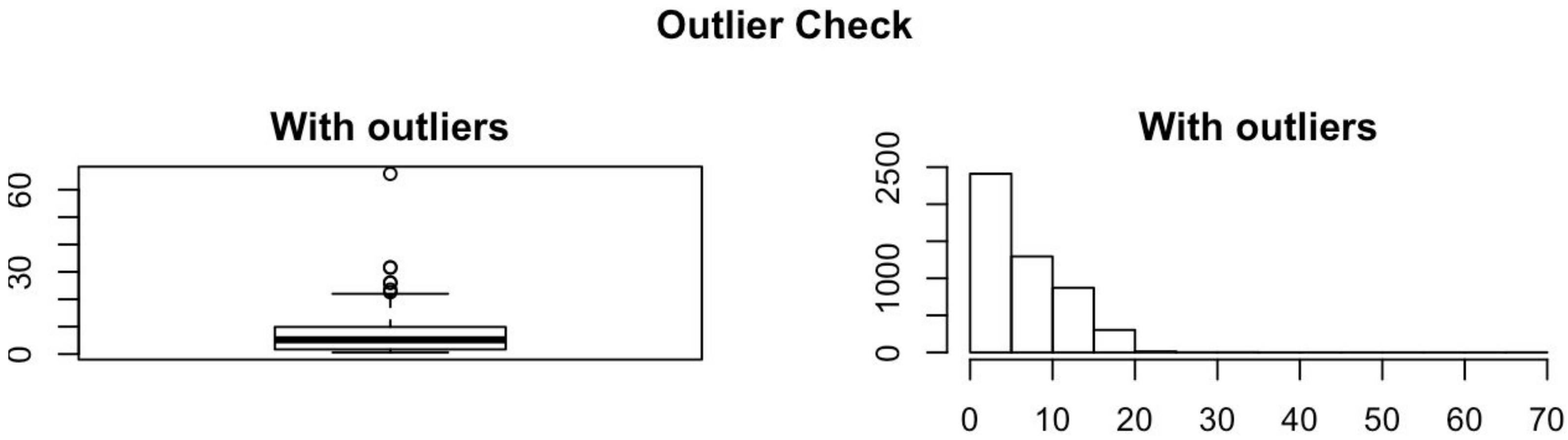
Cortez, P. et al., Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems 47 (2009) 547-553. https://www.sciencedirect.com/science/article/pii/S0167923609001377?via%3Dihub

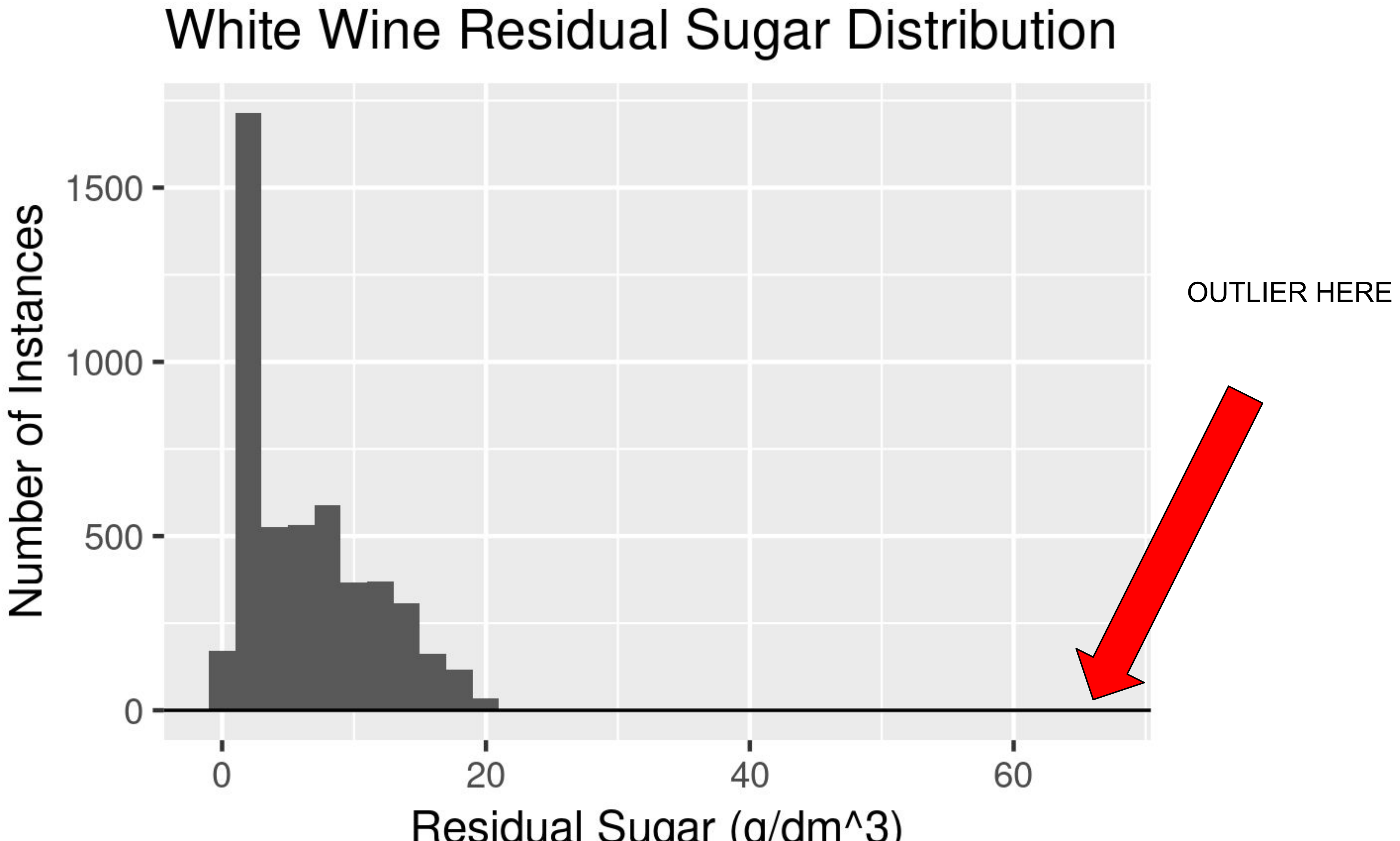
https://en.wikipedia.org/wiki/Vinho_Verde

Questions?

BACKUP SLIDES BELOW

Outlier Handling




White Wine Residual Sugar Distribution
Number of Instances
1500
1000
500
0
0
20
40
60
Residual Sugar (g/dm^3)
OUTLIER HERE

Wine Quality Dataset

```
'data.frame':   4898 obs. of  12 variables:
 $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
 $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
 $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
 $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
 $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
 $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
 $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
 $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
 $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
 $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
 $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
 $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

Anova Results

	diff	lwr	upr	p adj
5-4	0.344	0.573	0.115	0.001
6-4	0.423	0.198	0.648	0.00001
7-4	1.215	0.979	1.452	0.00000
6-5	0.767	0.673	0.860	0.00000
7-5	1.559	1.441	1.678	0.00000
7-6	0.793	0.682	0.903	0.00000

Alcohol, F = 397

	diff	lwr	upr	p adj
5-4	-0.195	-0.372	-0.019	0.023
6-4	-0.292	-0.466	-0.118	0.0001
7-4	-0.395	-0.577	-0.212	0.00000
6-5	-0.096	-0.169	-0.024	0.003
7-5	-0.199	-0.291	-0.108	0.00000
7-6	-0.103	-0.188	-0.018	0.011

Fixed Acidity, F = 16.68

	diff	lwr	upr	p adj
5-4	-0.079	-0.100	-0.059	0.00000
6-4	-0.121	-0.141	-0.101	0.00000
7-4	-0.118	-0.140	-0.097	0.00000
6-5	-0.041	-0.050	-0.033	0.00000
7-5	-0.039	-0.050	-0.029	0.00000
7-6	0.002	-0.008	0.012	0.941

Volatile Acidity, F = 123.2

	diff	lwr	upr	p adj
5-4	25.625	16.880	34.370	0.00000
6-4	11.768	3.173	20.364	0.002
7-4	-0.164	-9.193	8.865	1.000
6-5	-13.857	-17.434	-10.280	0.00000
7-5	-25.790	-30.310	-21.269	0.00000
7-6	-11.933	-16.156	-7.709	0.00000

Total Sulfur Dioxide, F = 81.45

	diff	lwr	upr	p adj
5-4	0.001	-0.003	0.006	0.845
6-4	-0.005	-0.009	-0.0004	0.026
7-4	-0.012	-0.017	-0.007	0.00000
6-5	-0.006	-0.008	-0.004	0.00000
7-5	-0.013	-0.016	-0.011	0.00000
7-6	-0.007	-0.009	-0.005	0.00000

Chlorides, F = 74.63