Modeling Wine Preferences via Data Mining

Dev Misra

Agenda


01
Introduction
02
Data Description
03
Techniques
04
Results
05
Conclusion

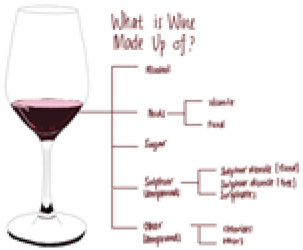
Background on Wine

- Wine was once viewed as a luxury good, but is now enjoyed by a wide range of consumers

- Portugal is a top ten wine exporting country
    - Exports of *Vinho Verde* have increased by 36% from 1997 to 2007

- Wine certifications and quality assessments are essential in the wine industry towards enhancing growth

Data Description



- Two large datasets, one for **red wine** and one for **white wine**

	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free sulfur Dioxide	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol	Quality
1	7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5
2	7.8	0.880	0.00	2.60	0.098	25	67	0.9968	3.20	0.68	9.8	5
3	7.8	0.760	0.04	2.30	0.092	15	54	0.9970	3.26	0.65	9.8	5
4	11.2	0.280	0.56	1.90	0.075	17	60	0.9980	3.16	0.58	9.8	6
5	7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5
6	7.4	0.660	0.00	1.80	0.075	13	40	0.9978	3.51	0.56	9.4	5
7	7.9	0.600	0.06	1.60	0.069	15	59	0.9964	3.30	0.46	9.4	5
8	7.3	0.650	0.00	1.20	0.065	15	21	0.9946	3.39	0.47	10.0	7
9	7.8	0.580	0.02	2.00	0.073	9	18	0.9968	3.36	0.57	9.5	7
10	7.5	0.500	0.36	6.10	0.071	17	102	0.9978	3.35	0.80	10.5	5
11	6.7	0.580	0.08	1.80	0.097	15	65	0.9959	3.28	0.54	9.2	5
12	7.5	0.500	0.36	6.10	0.071	17	102	0.9978	3.35	0.80	10.5	5
13	5.6	0.615	0.00	1.60	0.089	16	59	0.9943	3.58	0.52	9.9	5
14	7.8	0.610	0.29	1.60	0.114	9	29	0.9974	3.26	1.56	9.1	5
15	8.9	0.620	0.18	3.80	0.176	52	145	0.9986	3.16	0.88	9.2	5
16	8.9	0.620	0.19	3.90	0.170	51	148	0.9986	3.17	0.93	9.2	5

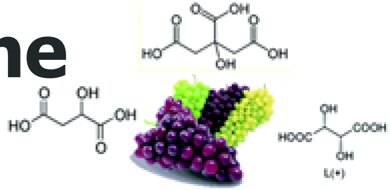
Red Wine

	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free sulfur Dioxide	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol	Quality
1	7.0	0.270	0.36	20.70	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
2	6.3	0.300	0.34	1.60	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
3	8.1	0.280	0.40	6.90	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
4	7.2	0.230	0.32	8.50	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
5	7.2	0.230	0.32	8.50	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
6	8.1	0.280	0.40	6.90	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
7	6.2	0.320	0.16	7.00	0.045	30.0	136.0	0.9949	3.18	0.47	9.6	6
8	7.0	0.270	0.36	20.70	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
9	6.3	0.300	0.34	1.60	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
10	8.1	0.220	0.43	1.50	0.044	28.0	129.0	0.9938	3.22	0.45	11.0	6
11	8.1	0.270	0.41	1.45	0.033	11.0	63.0	0.9908	2.99	0.56	12.0	5
12	8.6	0.230	0.40	4.20	0.035	17.0	109.0	0.9947	3.14	0.53	9.7	5
13	7.9	0.180	0.37	1.20	0.040	16.0	75.0	0.9920	3.18	0.63	10.8	5
14	6.6	0.160	0.40	1.50	0.044	48.0	143.0	0.9912	3.54	0.52	12.4	7
15	8.3	0.420	0.62	19.25	0.040	41.0	172.0	1.0002	2.98	0.67	9.7	5

White Wine

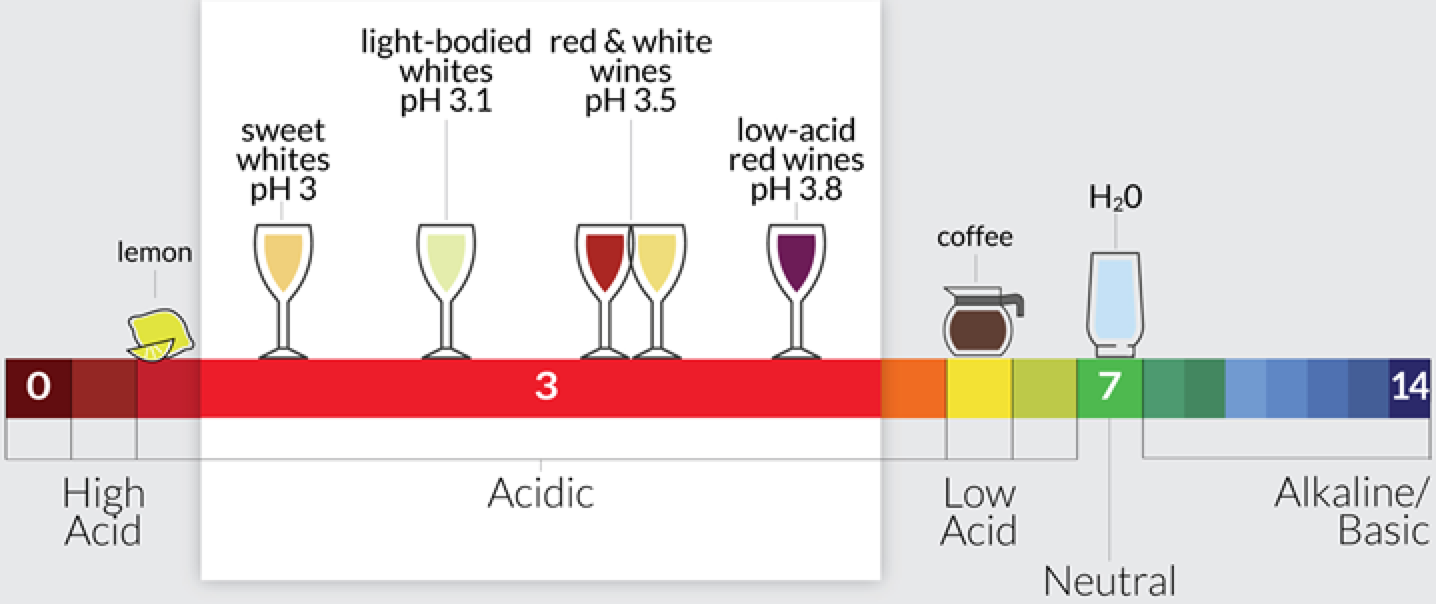
- **Objective**
  - Build a **model to predict wine taste preferences** based on distinct variables

Importance of Specific Variables in Wine



- Total **acidity** tells us the concentration of acids present in wine
- **Potential of hydrogen (pH)** level tells us how intense these acids taste
    - Measures the degree of relative alkalinity of a liquid on a scale of 0 to 14, with 7 being neutral
- Winemakers use **pH** as a way to measure ripeness in relation to **acidity**
    - Low pH wines will taste tart and crisp
    - Higher pH wines will taste flat and lack freshness
    - Most importantly, higher pH wines are more susceptible to bacterial growth, as bacteria thrive in higher pH environments

pH Scale



light-bodied whites pH 3.1
red & white wines pH 3.5
sweet whites pH 3
low-acid red wines pH 3.8
H_2O
coffee
lemon
0
3
7
14
High Acid
Acidic
Low Acid
Neutral
Alkaline/ Basic

Ideal Acidity for Wine

- Ideal pH range for red wine is **3.3 - 3.6**

- Ideal pH range for white wine is **3.0 - 3.4**

- Warmer climates result in higher sugar and lower acidity, whereas cooler climates result in lower sugar and higher acidity

- In a less acidic environment, a winemaker needs to compensate with higher doses of sulfur dioxide (SO_2) to keep bacteria away
  - E.g. a red wine with a pH of 3.9 would require about 60 mg/L of free SO_2 to inhibit bacteria whereas a similar wine but with a pH of 3.2 would only require about 13 mg/L

How Winemakers Control for Acidity

- Wines of **different acidity levels** can be **blended** to increase or lower the pH
- **Acid reduction** using potassium bicarbonate ($KHCO_3$) or agents such as ACIDEX to remove acidity and raise the pH
- **Cold stabilization** of wine can be used to increase or decrease pH
- **H_2O** can be added to wine to **dilute** its acidity and increase the pH
- **Malolactic fermentation** can raise the pH and alter the acidity of wine

Multiple Linear Regression

- Response/Dependent variable(s)
  - Wine Quality
- Regressor/Independent variable(s):
  - Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates, Alcohol
- Training and Test sets
  - Red wine
    - Red training set: [1:800,]
    - Red test set: [801:nrow(red),]
  - White wine
    - White training set: [1:2400,]
    - White test set: [2401:nrow(white),]

```
> dim(red)
[1] 1599   12
> dim(white)
[1] 4898   12
```

Assumptions for Regression

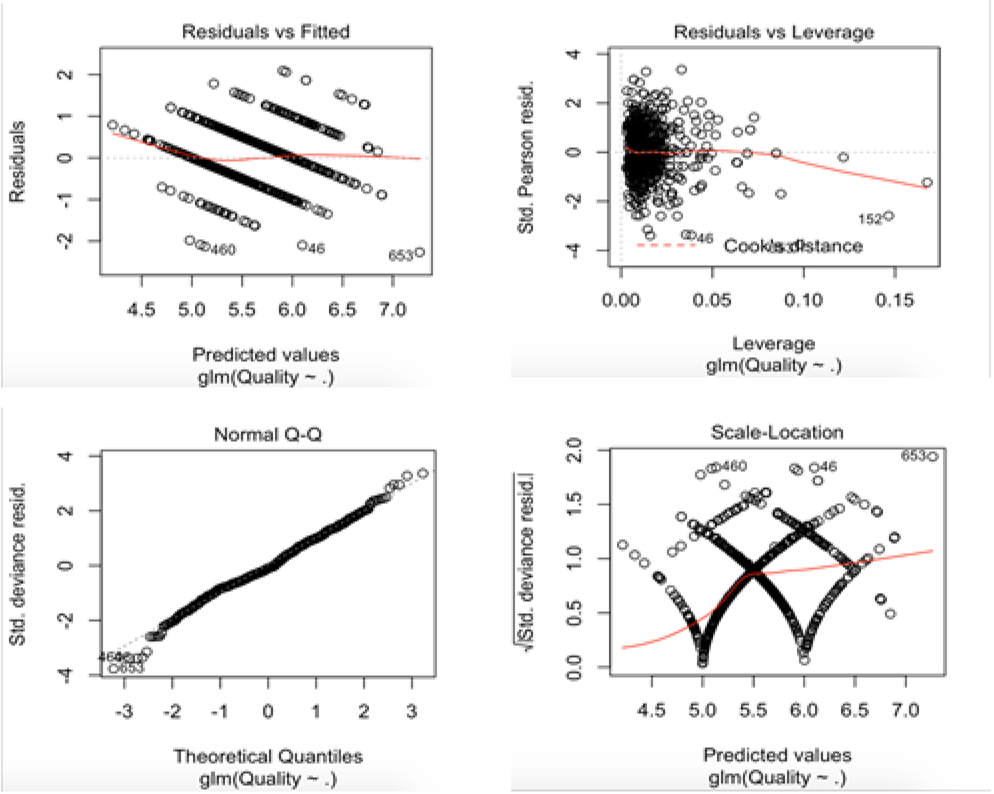
- **L.I.N.E.** assumptions/conditions must be met within both datasets to draw inferences from or make predictions from the model
  - **L**inearity
    - Relationship between dependent and independent variables is linear
  - **I**ndependence of Errors
    - No correlation between consecutive residuals
    - Each independent variable can be tested using VIF values
  - **N**ormality of Error
    - Residuals are normally distributed
  - **E**qual Variance
    - Residuals have a constant variance at every level of x

```
#assumptions for multiple linear regression - red
Rm<-lm(Quality~., data=Rtrain)
plot(Rm)
```

```
#assumptions for multiple linear regression - white
Wm<-lm(Quality~., data=Wtrain)
plot(Wm)
```

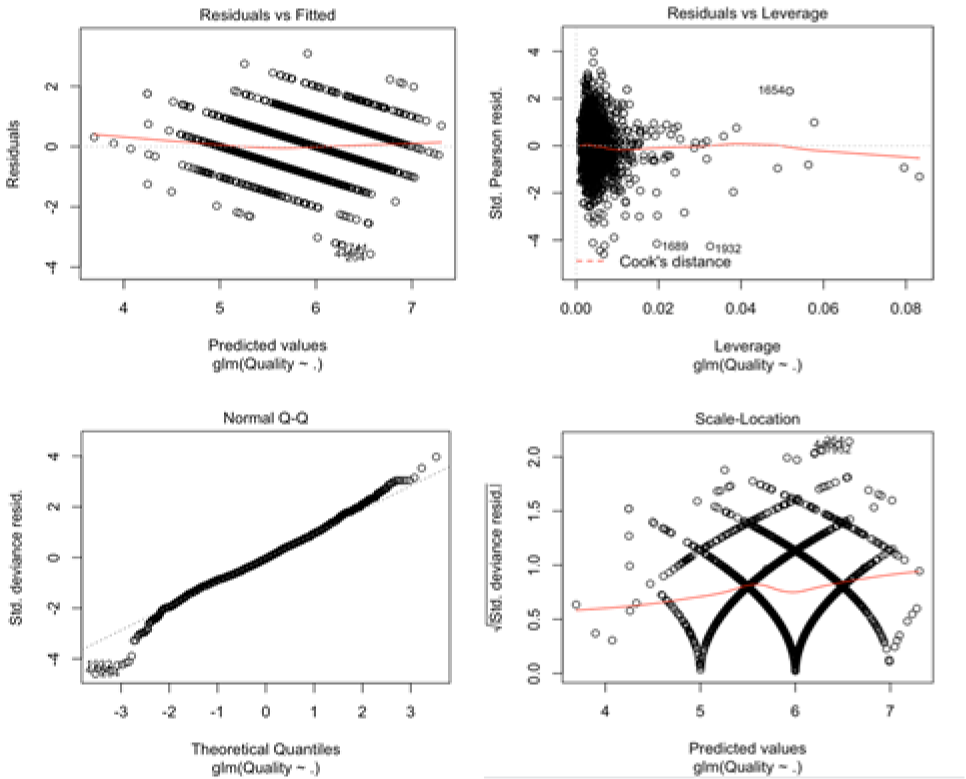
Red Wine Assumptions

- Linearity
  - No significant U-shape in “Residuals vs Fitted”
- Independence of Errors
  - No cyclical patterns in “Residuals vs Leverage”
- Normality of Error
  - Residuals are normally distributed in “Normal Q-Q”
- Equal Variance
  - Inconsistent variance in “Scale-Location”



White Wine Assumptions

- Linearity
  - No significant U-shape in "Residuals vs Fitted"
- Independence of Errors
  - No cyclical patterns in "Residuals vs Leverage"
- Normality of Error
  - Residuals are normally distributed in "Normal Q-Q"
- Equal Variance
  - Inconsistent variance in "Scale-Location"



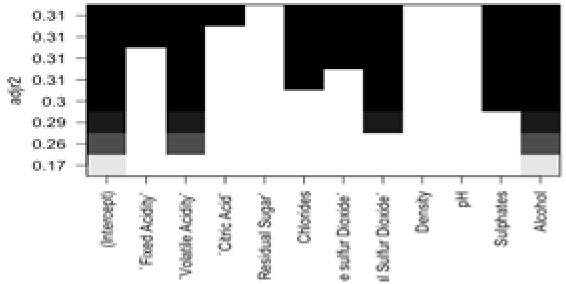
Model Fitness – Red Wine

- Adjusted R-Squared: 0.3098
    - A low adjusted R-squared indicates that the additional input variables are not adding value to the model
    - Currently, the red wine model is a bad fit
- Significance at α = 0.05
    - The model is not statistically significant at α = 0.05, as the p-value from the model is 0.6358, which is greater than 0.05

```
> summary(Rm)

Call:
lm(formula = Quality ~ ., data = Rtrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.26520 -0.39961 -0.06639  0.44318  2.09402 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            13.814270  29.160616   0.474   0.6358    
`Fixed Acidity`         0.037628   0.035297   1.066   0.2867    
`Volatile Acidity`     -1.023671   0.158995  -6.438 2.10e-10 ***
`Citric Acid`          -0.264088   0.193707  -1.363   0.1732    
`Residual Sugar`        0.002101   0.022651   0.093   0.9261    
Chlorides              -1.194417   0.508944  -2.347   0.0192 *  
`Free sulfur Dioxide`   0.005029   0.003468   1.450   0.1474    
`Total Sulfur Dioxide` -0.004895   0.001044  -4.690 3.22e-06 ***
Density               -10.588858  29.763640  -0.356   0.7221    
pH                     -0.086864   0.261459  -0.332   0.7398    
Sulphates               0.680155   0.138437   4.913 1.09e-06 ***
Alcohol                 0.267089   0.033492   7.975 5.36e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6335 on 788 degrees of freedom
Multiple R-squared:  0.3193,	Adjusted R-squared:  0.3098 
F-statistic: 33.61 on 11 and 788 DF,  p-value: < 2.2e-16
```

Improved Red Wine Model

- Best regressor(s) to incorporate into model:
  - Used *regsubsets* {leaps}
  - Fixed Acidity, Volatile Acidity, Citric Acid, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates, and Alcohol
- Adjusted R-Squared: 0.312
  - Compared to the original values, the relatively higher Adjusted R-Squared indicates the regressors can add more value to the model
- Significance at α = 0.05
  - The model is statistically significant at α = 0.05, as the p-value from the model is 2e-16, which is less than 0.05




```
> summary(fitR)

Call:
lm(formula = Quality ~ `Fixed Acidity` + `Volatile Acidity` +
    `Citric Acid` + Chlorides + `Free sulfur Dioxide` + `Total Sulfur Dioxide` +
    Sulphates + Alcohol, data = Rtrain)

Residuals:
     Min       1Q   Median       3Q      Max
-2.22837 -0.40410 -0.06757  0.44475  2.10794

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.962163   0.303470   9.761  < 2e-16 ***
`Fixed Acidity`         0.035104   0.017966   1.954   0.0511 .
`Volatile Acidity`     -1.033679   0.157581  -6.560 9.74e-11 ***
`Citric Acid`          -0.257201   0.193081  -1.332   0.1832
Chlorides              -1.160423   0.490300  -2.367   0.0182 *
`Free sulfur Dioxide`   0.004677   0.003417   1.369   0.1715
`Total Sulfur Dioxide` -0.004787   0.001016  -4.713 2.88e-06 ***
Sulphates               0.682293   0.134398   5.077 4.79e-07 ***
Alcohol                 0.269975   0.024477  11.030  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6325 on 791 degrees of freedom
Multiple R-squared:  0.3189,	Adjusted R-squared:  0.312
F-statistic: 46.29 on 8 and 791 DF,  p-value: < 2.2e-16
```

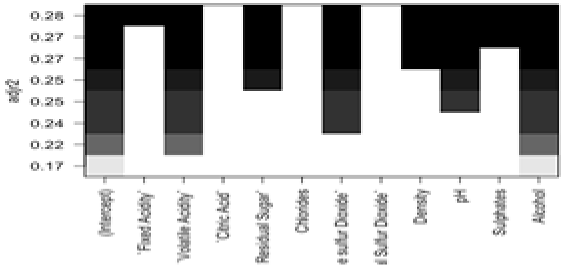
Model Fitness – White Wine

- Adjusted R-Squared: 0.2787
  - A low adjusted R-squared indicates that the additional input variables are not adding value to the model
  - Currently, the white wine model is a bad fit

- Significance at α = 0.05
  - The model is statistically significant at α = 0.05, as the p-value from the model is 2.54e-15, which is less than 0.05

```
> summary(Wm)

Call:
lm(formula = Quality ~ ., data = Wtrain)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5669 -0.5154 -0.0374  0.4796  3.0864 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            2.529e+02  3.176e+01   7.964 2.54e-15 ***
`Fixed Acidity`        1.348e-01  3.295e-02   4.090 4.45e-05 ***
`Volatile Acidity`    -1.743e+00  1.652e-01 -10.547  < 2e-16 ***
`Citric Acid`          7.870e-02  1.305e-01   0.603   0.5464    
`Residual Sugar`       1.072e-01  1.206e-02   8.886  < 2e-16 ***
Chlorides             -2.367e-01  7.427e-01  -0.319   0.7500    
`Free sulfur Dioxide`  6.113e-03  1.336e-03   4.575 5.01e-06 ***
`Total Sulfur Dioxide` 1.077e-04  5.467e-04   0.197   0.8439    
Density               -2.551e+02  3.220e+01  -7.923 3.51e-15 ***
pH                     1.187e+00  1.624e-01   7.312 3.58e-13 ***
Sulphates              8.936e-01  1.530e-01   5.839 5.95e-09 ***
Alcohol                9.932e-02  4.013e-02   2.475   0.0134 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7791 on 2388 degrees of freedom
Multiple R-squared:  0.282,	Adjusted R-squared:  0.2787 
F-statistic: 85.27 on 11 and 2388 DF,  p-value: < 2.2e-16
```

Improved White Wine Model



- Best regressor(s) to incorporate into model:
  - Used *regsubsets* {leaps}
  - Fixed Acidity, Volatile Acidity, Residual Sugar, Free Sulfur Dioxide, Density, pH, Sulphates, and Alcohol

- Adjusted R-Squared: 0.2795
  - Compared to the original values, the relatively higher Adjusted R-Squared indicates the regressors can add more value to the model

- Significance at α = 0.05
  - The model is statistically significant at α = 0.05, as the p-value from the model is 2e-16, which is less than 0.05

```
> summary(fitW)

Call:
lm(formula = Quality ~ `Fixed Acidity` + `Volatile Acidity` +
    `Residual Sugar` + `Free sulfur Dioxide` + Density + pH +
    Sulphates + Alcohol, data = Wtrain)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5485 -0.5163 -0.0338  0.4800  3.0869

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.519e+02  3.005e+01   8.383  < 2e-16 ***
`Fixed Acidity`       1.375e-01  3.180e-02   4.324 1.59e-05 ***
`Volatile Acidity`   -1.758e+00  1.607e-01 -10.934  < 2e-16 ***
`Residual Sugar`      1.073e-01  1.153e-02   9.310  < 2e-16 ***
`Free sulfur Dioxide` 6.282e-03  1.077e-03   5.834 6.16e-09 ***
Density              -2.541e+02  3.043e+01  -8.350  < 2e-16 ***
pH                    1.179e+00  1.561e-01   7.551 6.12e-14 ***
Sulphates             9.008e-01  1.519e-01   5.931 3.46e-09 ***
Alcohol               1.021e-01  3.950e-02   2.584  0.00982 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7787 on 2391 degrees of freedom
Multiple R-squared:  0.2819,    Adjusted R-squared:  0.2795
F-statistic: 117.3 on 8 and 2391 DF,  p-value: < 2.2e-16
```

Model Evaluation

- Mean Absolute Percentage Error (MAPE) is used to determine the prediction error of the models

- MAPE of red wine model = 9.5976%

- MAPE of white wine model = 10.3711%

```
MAPE<-function(pred, true)
{
  return(100*mean(abs((pred-true)/true), na.rm=T))
}

MAPE(Rpred, Rtest$Quality)
MAPE(Wpred, Wtest$Quality)
```

```
> MAPE(Rpred, Rtest$Quality)
[1] 9.597551
> MAPE(Wpred, Wtest$Quality)
[1] 10.37114
```

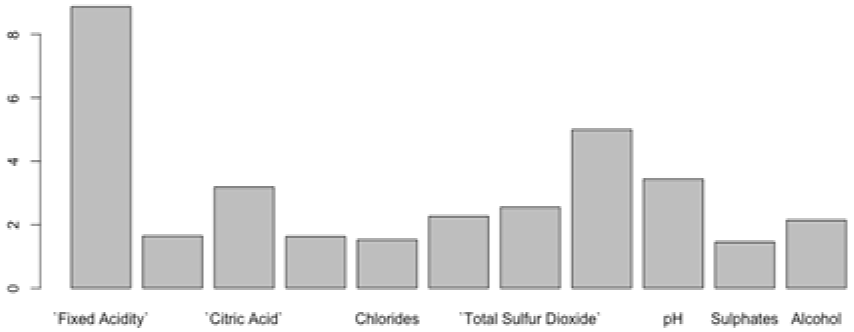
- Both models have a MAPE of < 10.5%, indicating that the average unsigned percentage error for each model is very low (models are good fits)

Variance Inflation Factor – Red Wine

```
> RVIF
     `Fixed Acidity`   `Volatile Acidity`          `Citric Acid`
            8.866974             1.645710               3.184926
    `Residual Sugar`            Chlorides  `Free sulfur Dioxide`
            1.627857             1.528297               2.265589
`Total Sulfur Dioxide`            Density                     pH
            2.551407             4.996512               3.438727
           Sulphates              Alcohol
            1.455214             2.144835

> which(RVIF>5)
`Fixed Acidity`
              1
```

- Threshold for VIF: 5
- A variable with a higher VIF contributes more to the standard error of a regression

- `which(RVIF>5)` highlights regressors that exhibit multicollinearity
  - ∴ Fixed Acidity is highly collinear with the other regressors in the model

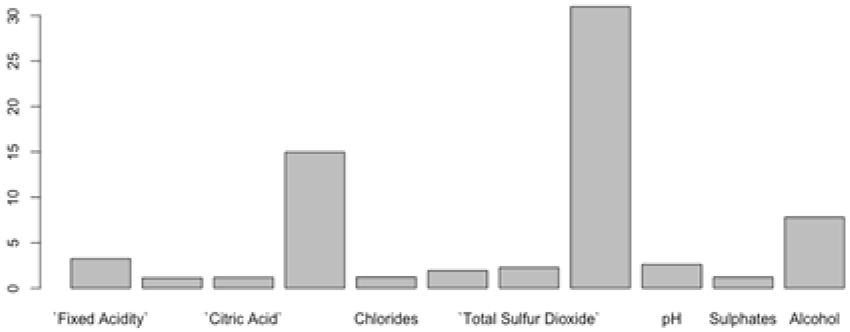


Variance Inflation Factor – White Wine

```
> WVIF
      `Fixed Acidity`    `Volatile Acidity`         `Citric Acid`
             3.241740              1.128829              1.182005
     `Residual Sugar`             Chlorides `Free sulfur Dioxide`
            14.969678              1.213800              1.928162
`Total Sulfur Dioxide`              Density                    pH
             2.271571             30.975080              2.611740
            Sulphates               Alcohol
             1.204943              7.799363

> which(WVIF>5)
`Residual Sugar`          Density          Alcohol
               4                8               11
```

- Threshold for VIF: 5
- A variable with a higher VIF contributes more to the standard error of a regression

- *which(RVIF>5)* highlights regressors that exhibit multicollinearity
  - ∴ Residual Sugar, Density, and Alcohol are highly collinear with the other regressors in the model



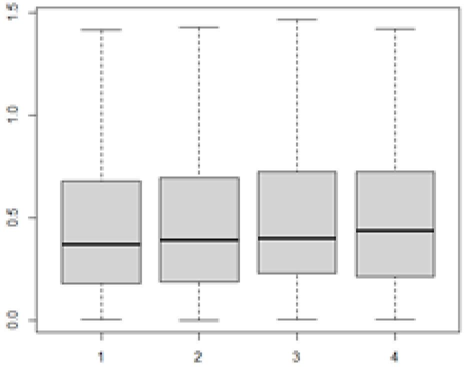
Red Wine Model Robustness (K-fold CV Test)

- K-fold cross validation is a procedure used to estimate the skill of the model on new data
- Four fold Cross Validation
  - Because the performance metrics across all four folds are similar, the red wine model can be described as robust
  - In other words, the model performance stays stable when the data (in both the training and test sets) changes, thus, it is robust

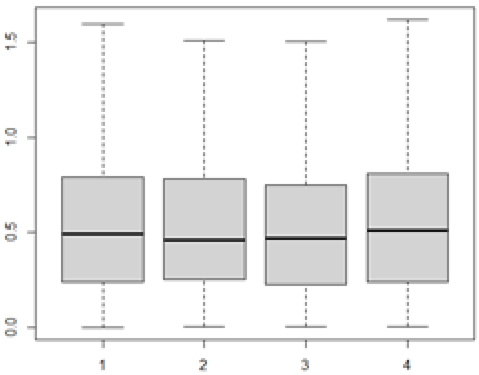
```
K<-4
dt1<-red[sample(1:nrow(red), nrow(red)), ]
len<-floor(nrow(dt1)/K) # number of obs. in testing set
pred<-matrix(, len, K)
test.all<-matrix(, len, K)
pred.err<-matrix(, len, K)


for(k in 1:K)
{
  index <- ((k-1)*len+1):(k*len)

  te<-dt1[index, ]
  tr<-dt1[-index, ]

  fit <- lm(Quality~., data=red)

  pre<- predict(fit, newdata=te)
  test.all[1:len, k]<-te$Quality
  pred[1:len, k]<- pre
  pred.err[1:len, k]<- abs(pre-te$Quality)

}

boxplot(pred.err, outline = FALSE)
```

White Wine Model Robustness (K-fold CV Test)

- Because the performance metrics across all four folds are similar, the white wine model can be described as robust
- In other words, the model performance stays stable when the data (in both the training and test sets) changes, thus, it is robust



Red Wine



White Wine

Red Wine Model Conclusion

- Best model (eight regressors):

```
> summary(fitR)

Call:
lm(formula = Quality ~ `Fixed Acidity` + `Volatile Acidity` +
    `Citric Acid` + Chlorides + `Free sulfur Dioxide` + `Total Sulfur Dioxide` +
    Sulphates + Alcohol, data = Rtrain)
```

- (RVIF>5):

```
> which(RVIF>5)
`Fixed Acidity`
              1
```

- Variable(s) most important towards determining the quality of red wine:
    - Volatile Acidity, Citric Acid, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates, and Alcohol

White Wine Model Conclusion

- Best model (eight regressors):

```
> summary(fitW)

Call:
lm(formula = Quality ~ `Fixed Acidity` + `Volatile Acidity` +
    `Residual Sugar` + `Free sulfur Dioxide` + Density + pH +
    Sulphates + Alcohol, data = Wtrain)
```

- (RVIF>5):

```
> which(WVIF>5)
`Residual Sugar`         Density          Alcohol
               4               8               11
```

- Variable(s) most important towards determining the quality of white wine:
  - Fixed Acidity, Volatile Acidity, Free Sulfur Dioxide, pH, and Sulphates

References

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

https://www.sciencedirect.com/science/article/pii/S0167923609001377?via%3Dihub

https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod3/3/index.html

https://www.wineperspective.com/wine-acidity/

https://winemakermag.com/technique/1650-monitoring-adjusting-ph

https://winemakermag.com/article/547-phiguring-out-ph

https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/

THANK YOU