Red Wine Quality

Data

- “Red Wine Quality” From Keggle
- 11 Independent Variables -> Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates,Alcohol
- 1 Dependent Variable -> Quality
    - Scale From 0 (Worst) to 10 (Best)
- 1599 Samples

Data Example

fixed acidit	volatile acic	citric acid	residual su	chlorides	free sulfur	total sulfur	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6

Multiple Linear Regression

- Find a Linear Equation to Predict Red Wine's quality

```
df <- read.csv("winequality-red.csv", sep=",", header = T)

FixedAcidity <- df[,1]
VolatileAcidity <- df[,2]
CitricAcid <- df[,3]
ResidualSugar <- df[,4]
Chlorides <- df[,5]
FreeSulfurDioxide <- df[,6]
TotalSulfurDioxide <- df[,7]
Density <- df[,8]
PH <- df[,9]
Sulphates <- df[,10]
Alcohol <- df[,11]
Quality <- df[,12]

model <- lm(Quality ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar
+ Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density + PH + Sulphates + Alcohol)
summary(model)
```

Multiple Linear Regression : Result

```
Call:
lm(formula = Quality ~ FixedAcidity + VolatileAcidity + CitricAcid +
    ResidualSugar + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
    Density + PH + Sulphates + Alcohol)

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
FixedAcidity         2.499e-02  2.595e-02   0.963   0.3357
VolatileAcidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
CitricAcid          -1.826e-01  1.472e-01  -1.240   0.2150
ResidualSugar        1.633e-02  1.500e-02   1.089   0.2765
Chlorides           -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
FreeSulfurDioxide    4.361e-03  2.171e-03   2.009   0.0447 *
TotalSulfurDioxide -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
```

$$
\begin{aligned}
Quality = 21.97 &+ (0.02499 * FixedAcidity) \\
&+ (-1.084 * VolatileAcidity) \\
&+ (-0.1826 * CitricAcid) \\
&+ (0.01633 * ResidualSugar) \\
&+ (-1.874 * Chlorides) \\
&+ (0.004361 * FreeSulfurDioxide) \\
&+ (-0.003265 * TotalSulfurDioxide) \\
&+ (-17.88 * Density) + (-0.4137 * PH) \\
&+ (0.9163 * Sulphates) + (0.2762 * Alcohol)
\end{aligned}
$$

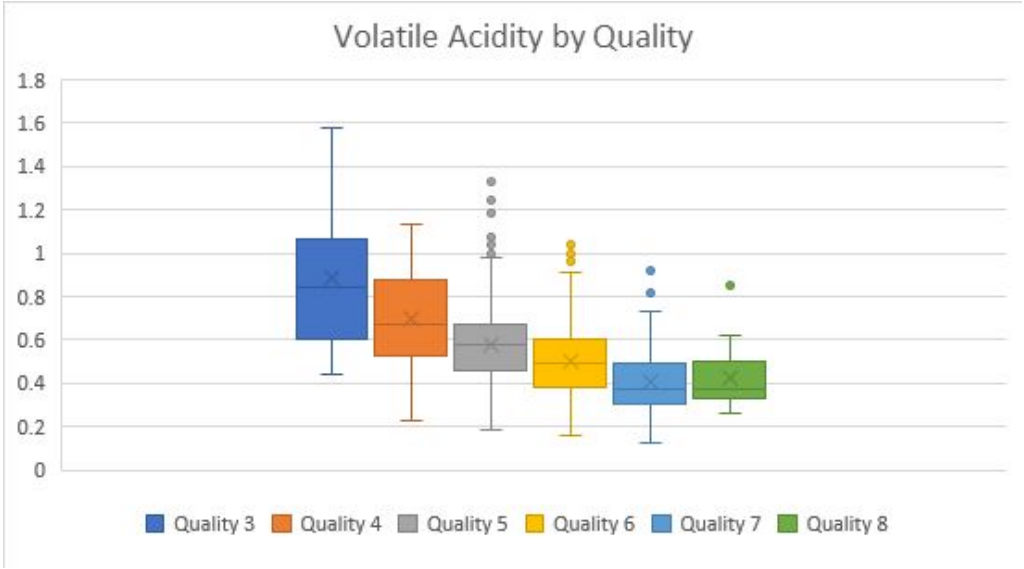
- Significant Variables
  - Volatile Acidity, Chlorides, Total Sulfur Dioxide, Sulphates, Alcohol
- R-squared = 0.3606 (Quite Low)

Multiple Linear Regression : R-squared

- Why R-squared is quite low
  - Quality
    - Discrete
    - Little Sample in some Quality score

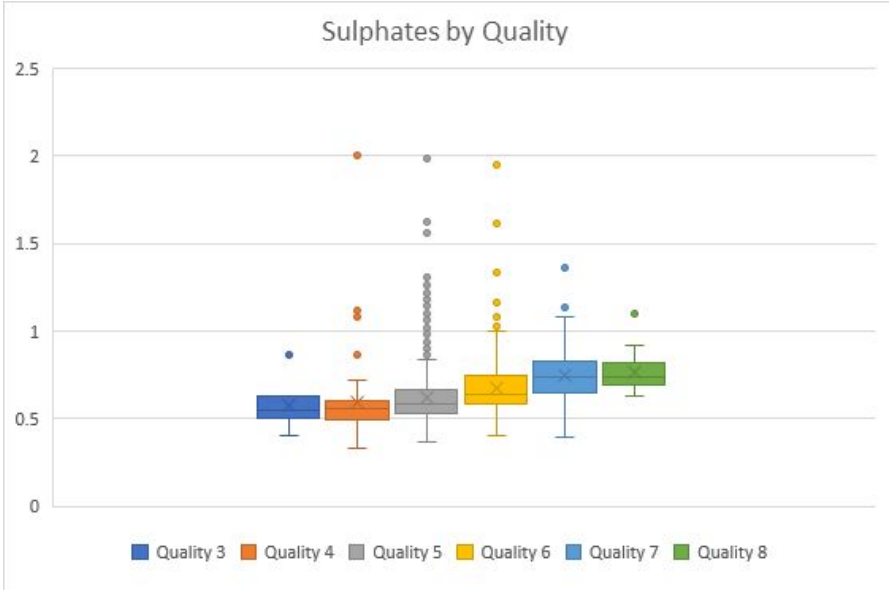
Multiple Linear Regression : Correlation

- Volatile Acidity
  - Correlation coefficient = -0.3905578



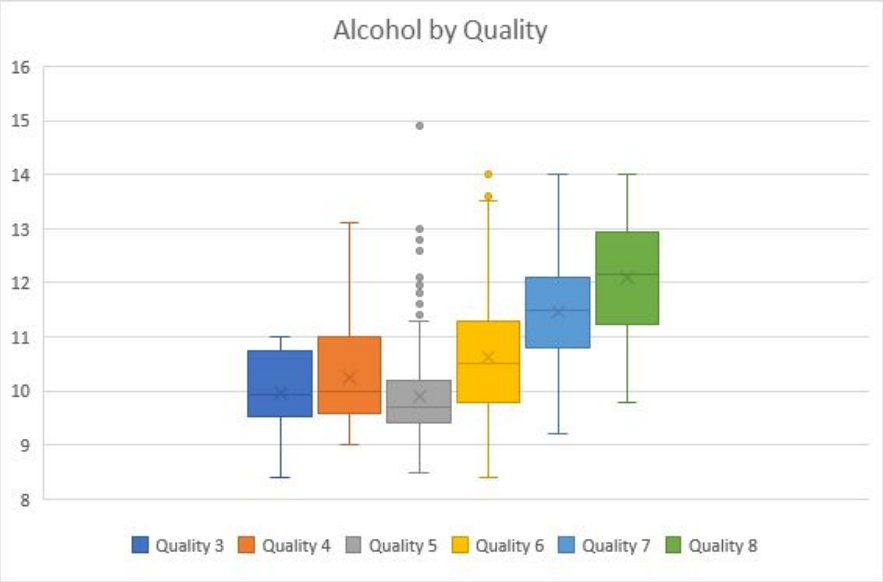
Multiple Linear Regression : Correlation

- Sulphates
  - Correlation coefficient = 0.2513971



Multiple Linear Regression : Correlation

- Alcohol
  - Correlation coefficient = 0.4761663



Anova-2-Factors (Fixed Acidity , Residual Sugar)

fixed acidity
7.4
7.8
7.8
11.2
7.4
7.4
7.9
7.3
7.8
7.5
6.7
7.5

residual sugar
1.9
2.6
2.3
1.9
1.9
1.8
1.6
1.2
2
6.1
1.8
6.1
1.6
1.6

quality
5
5
5
6
5
5
5
7
7
5

ResSugar				fixed_acidity		
0	0.5	0		0	6	58
0.5	1	2		6	7	254
0	1.5	50		7	8	504
0	2	464		8	9	316
2	2.5	616		9	10	191
2.5	3	279		10	11	138
3	25	240		11	12	72
		0		12	16	66

Anova-2-Factors (Fixed Acidity , Residual Sugar)

ตารางแสดง Wine Quality ในระดับ Fixed Acidity และ Residual Sugar ต่างๆ

		residual sugar			
		[0,2)	[2,2.5)	[2.5,3)	>=3
fixed_acidity	[0,6)	5.78125	6.09090	5	6
	[6,7)	5.393258426	5.66346	5.57894	5.428571428571
	[7,8)	5.479768786	5.60696	5.57894	5.314814814814
	[8,9)	5.670588235	5.56557	5.61904	5.565217391304
	[9,10)	5.923076923	5.57142	5.53333	5.7
	[10,11)	5.925925925	5.84782	6	5.923076923076
	[11,12)	6.083333333	5.83333	5.72222	6.166666666666
	>=12	5.571428571	5.85	5.88888	5.904761904761

Anova-2-Factors (Fixed Acidity , Residual Sugar)

SST	1.988783830906	df	31
SSA	0.20855349200	df	7
SSB	0.01286175858	df	3
SSE	1.76736858031	df	21

Hypothesis	Mean Square (MS)	Test Statistic (f)	Rejection region(α = 0.05)	
$H_{0A} vs\ H_{0A}$	0.02979335600	0.35400678895	2.4876	Do not Reject
$H_{0B} vs\ H_{0B}$	0.00428725286	0.05094144544	3.0725	Do not Reject
Error	0.08416040858			

Analysis of Categorical Data

		residual sugar			
		[0,2)	[2,2.5)	[2.5,3)	>=3
fixed_acidity	[0,6)	32	22	1	3
	[6,7)	89	104	19	42
	[7,8)	173	201	76	54
	[8,9)	85	122	63	46
	[9,10)	39	77	45	30
	[10,11)	27	46	39	26
	[11,12)	12	24	18	18
	>=12	7	20	18	21

Independence test

Hypothesis Test:

Null Hypothesis (Ho)=> Pij = Pi*Pj

Alternative hypothesis (Ha)=> Ho is not true

ทดสอบค่า residual sugar Independence กับ fixed_acidity หรือไม่

Analysis of Categorical Data

		residual sugar			
		[0,2)	[2,2.5)	[2.5,3)	>=3
fixed_acidity	[0,6)	16.8305	22.3439	10.1200	8.70544
	[6,7)	73.7060	97.8511	44.3189	38.1238
	[7,8)	146.251	194.161	87.9399	75.6472
	[8,9)	91.6973	121.736	55.1369	47.4296
	[9,10)	55.4246	73.5809	33.3264	28.6679
	[10,11)	40.0450	53.1632	24.0787	20.7129
	[11,12)	20.8930	27.7373	12.5628	10.8067
	>=12	19.1519	25.4258	11.5159	9.90619

โดยจะได้ค่าต่างๆดังนี้ (กำหนด α = 0.001)

- Test Statistic = 120.017
- Degree of Freedom = 21
- α = 0.001
- Chisquare test = 46.797

Note: แม้ว่า α = 0.001 ก็ยัง Reject

สรุป: ค่า Teststatistic > ค่า Chisquare test ดังนั้นเราจะปฏิเสธสมมติฐานนี้และสรุปได้ว่าค่า Residual Sugar ที่ระดับต่างๆ นั้น Dependence ต่อ Fixd_Acidity

Analysis of variance

ตารางแสดงถึงค่าของน้ำตาลที่มีอยู่ในไวน์ โดยแยกกลุ่มออกตามคุณภาพของไวน์

Group of quality	Residual Sugar										Sample Mean	Sample SD
3	2.2	2.1	4.25	1.5	3.4	2.1	1.2	2.1	5.7	1.8	2.635	1.401596
4	4.4	1.5	2.8	2.1	2.1	1.5	1.4	3.4	1.3	1.6	2.21	1.024641
5	2	1.5	2	2.5	2.4	2.4	2	2.5	1.8	2	2.11	0.331495
6	1.9	2	1.6	2	2.8	2.1	2.4	2	2	2.6	2.14	0.356526
7	2.2	2.2	2.6	1.8	5.6	3.5	5.6	2.5	2.5	3.2	3.17	1.370361
8	1.4	2.2	5.2	2.8	2.6	2.6	2	2.3	1.8	1.9	2.48	1.045413
										Xbar	2.4575	

ภาพจาก :
https://www.kinlakestars.com/master-art-%E0%B8%82%E0%B8%AD%E0%B8%87%E0%B9%84%E0%B8%A7%E0%B8%99%E0%B9%8C-%E0%B8%93-wine-club-%E0%B9%82%E0%B8%A3%E0%B8%87%E0%B9%81%E0%B8%A3%E0%B8%A1%E0%B9%82%E0%B8%9F%E0%B8%A3%E0%B9%8C%E0%B8%8B/

Analysis of variance

สมมติฐานคั้งต้นให้ค่าเฉลี่ยของปริมาณน้ำตาลที่อยู่ในกลุ่มของคุณภาพต่างๆมีปริมาณที่เท่ากัน

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$

สมมติฐานรองคือค่าเฉลี่ยของปริมาณน้ำตาลที่อยู่ในกลุ่มของคุณภาพต่างๆมีปริมาณที่ไม่เท่ากัน

$H_a: Not\ all\ \mu_i's\ are\ equal$ (at least two of the μi 's are different)

Analysis of variance

	df	SS	MS	f	Rejection region	
Treatment	5	301.4660625	60.2932125	58.1406621	3.37691159	REJECT
Error	54	55.99925	1.037023148			
Total	59	357.4653125				

ทำการทดสอบ Test Statistic ที่ค่า $\alpha = 0.01$

H_0 is reject . ค่าเฉลี่ยของปริมาณน้ำตาลที่อยู่ในกลุ่มของคุณภาพต่างๆมีปริมาณที่ไม่เท่ากัน

สรุป ค่าเฉลี่ยของปริมาณน้ำตาลในไวน์ที่อยู่ในกลุ่มของคุณภาพไวน์ที่ระดับต่างๆมีปริมาณที่ไม่เท่ากัน