빅데이터를 활용한 와인 품질 및 종류 분석

7조 조원

유형기 심우주
유상훈 유병승
강정화 이용선
김건 최규영
김현석 홍동연

목차
Contents

발표 배경

“소믈리에의 평가는 항상 옳을까?”

발표 배경

Orley Ashenfelter의 보르도와인 가격 예측

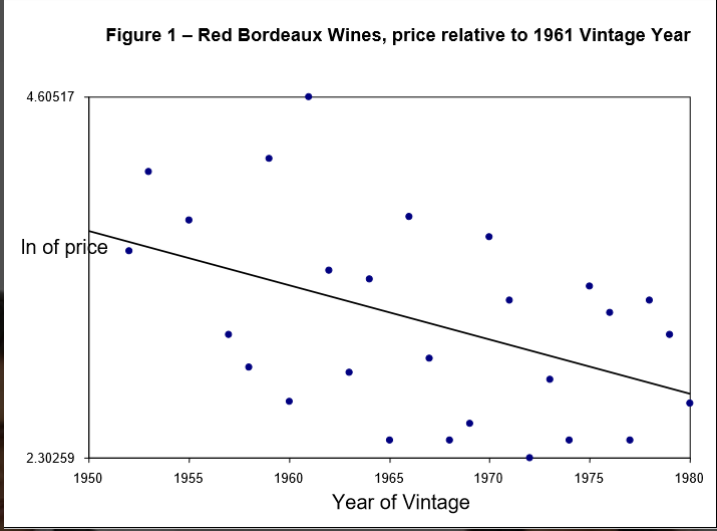


Table 1 – London Auction Prices for Select Mature Red Bordeaux Wines, 1990-1991 (per dozen bottles in $US)

	Chateaux (Vineyards)						
Vintage	Lafite	Latour	Cheval Blanc	Cos d'Estournel	Montrose	Pichon Lalande	Average
1960	494	464	486				479
1961*	4335	5432	3534	1170	1125	1579	4884
1962*	889	1064	821	521	456	281	977
1963	340	471		251			406
1964*	649	1114	1125	315	350	410	882
1965	190	424				258	307
1966*	1274	1537	1260	546	482	734	1406
1967*	374	530	441	213	236	243	452
1968	223	365	274				294
1969	251	319		123	84	152	285
Average	1504	1935	1436	553	530	649	

소개 – 변수 정하기

결합산(fixed acid) : 주로 타르타르산(tartaric),사과산(malic)으로 이루어져 있고 와인의 산도를 제어 한다.

휘발산(volatile acidity) : 와인의 향에 연관이 많다.

구연산(citric acid) : 와인의 신선함을 올려주는 역할, 산성화에 연관을 미친다

잔여 설탕(residual sugar) : 와인의 단맛을 올려준다.

염화물(chlorides) : 와인의 짠맛의 원인이며 와인의 신맛을 좌우하는 성분

황 화합물 : 황 화합물은 원하지 않는 박테리아와 효모를 죽여서 와인을 오래 보관하는 역할 (free sulfur dioxide, total sulfur, dioxide sulphates)

밀도(density) : 바디의 높고 낮음을 표현하는 와인의 무게감을 의미한다.

산성도(pH) : 와인의 신맛의 정도를 나타낸다.

알코올(alcohol) : 와인의 과 단맛을 주며 와인의 바디감에 영향을 준다.

Vietoris – Rips complex

Statistical process for estimating the relationships among variables

Reconstruction ontology

Orthogonal Basis

PCA

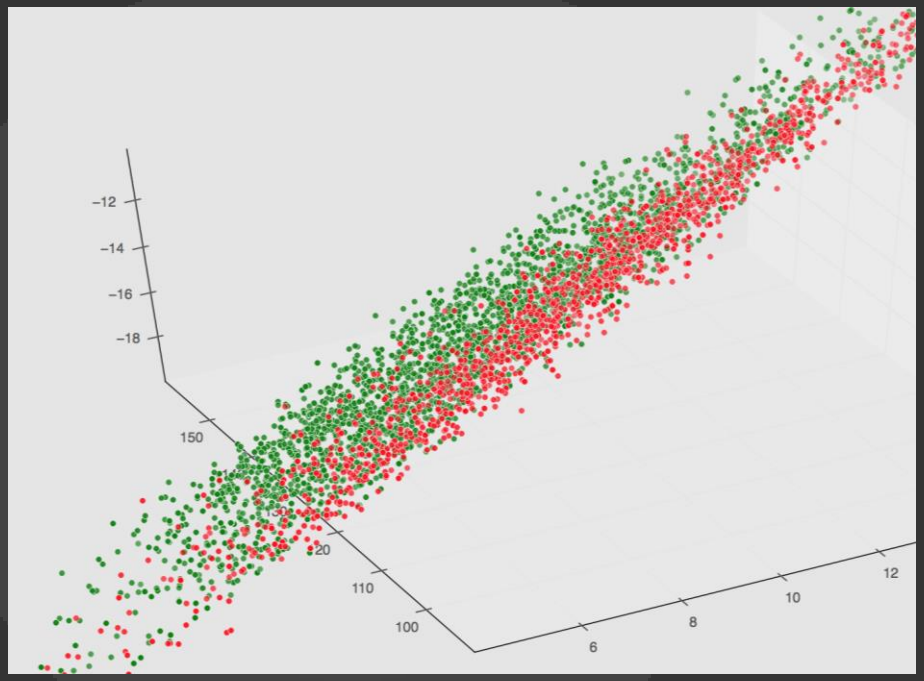
X′X 의 eigenvalue $x_1, x_2, \cdots, x_{11}$
(단, x_i = var($V_i^T X$))

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_{11}} \geq 0.7$$

$$\frac{\lambda_1+\lambda_2+\lambda_3}{\lambda_1+\lambda_2+\cdots+\lambda_{11}} \geq 0.7$$

실제 결과

PCA



: 고품질 와인
(quality =6,7,8)

: 저품질 와인
(quality =3,4,5)

- 데이터들의 분리가 어느 정도 이루어짐을 알 수 있다.

PCA

한계점

- 3차원 이상으로의 차원축소를 고려하지 못했다.

- 주요인자를 구하지 못하였다.

전진선택법(Forward Selection method)을 이용하여 회귀식 추정하기

회귀분석
(Regression Analysis

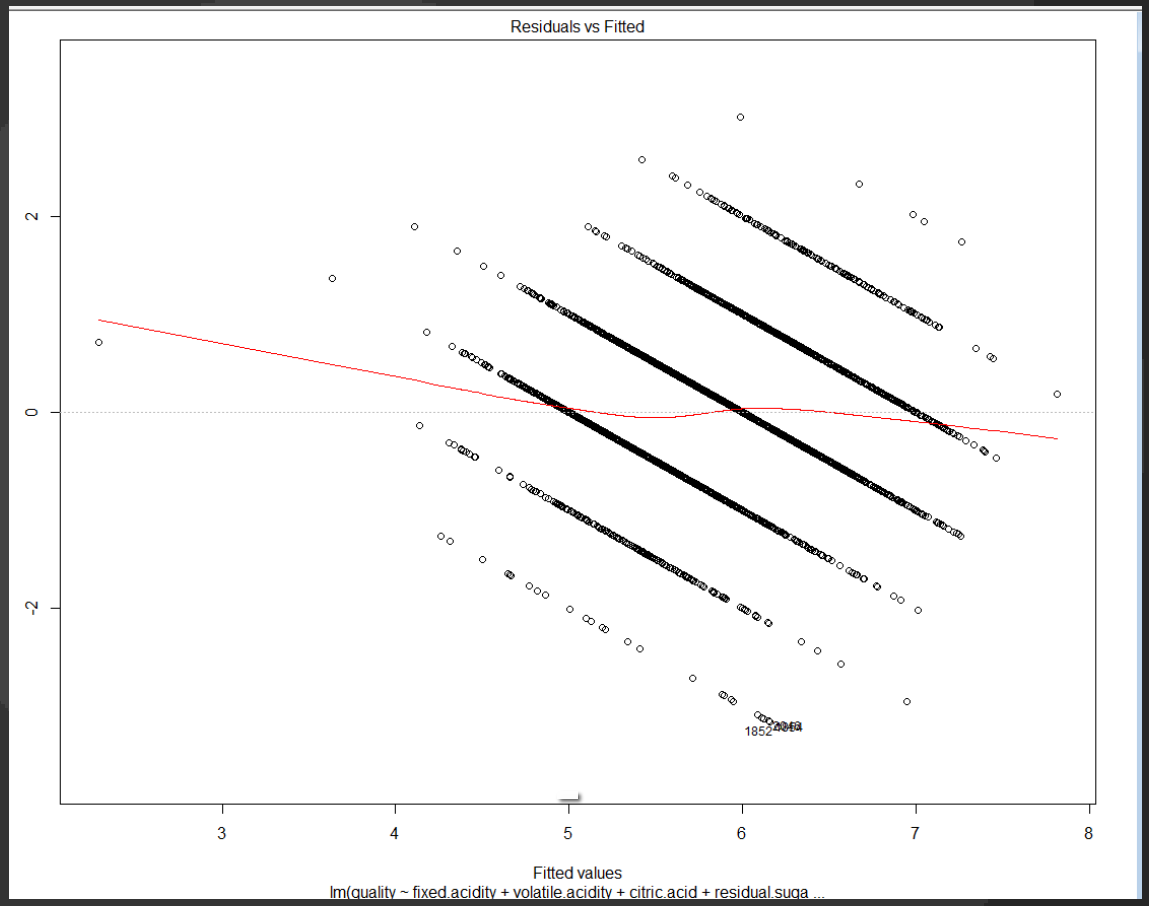
Quality = 4.4300987+(-1.0127527)× (휘발산)+(-2.0178138) ×(염화물) + (0.0050774) ×(free.무수아황산)+(-0.0034822) ×(total.무수아황산)+(-0.4826614) ×(pH)+(0.8826651) ×(황산염)+(0.2893028) ×(알코올)

Residual standard error: 0.6477 on 1591 degrees of freedom Multiple R-squared: 0.3595, Adjusted R-squared: 0.3567 F-statistic: 127.6 on 7 and 1591 DF, p-value: < 2.2e-16

Residual standard error: 0.7094 on 6430 degrees of freedom(1 observation deleted due to missingness) Multiple R-squared: 0.3467, Adjusted R-squared: 0.34 F-statistic: 51.71 on 66 and 6430 DF, p-value: < 2.2e-16

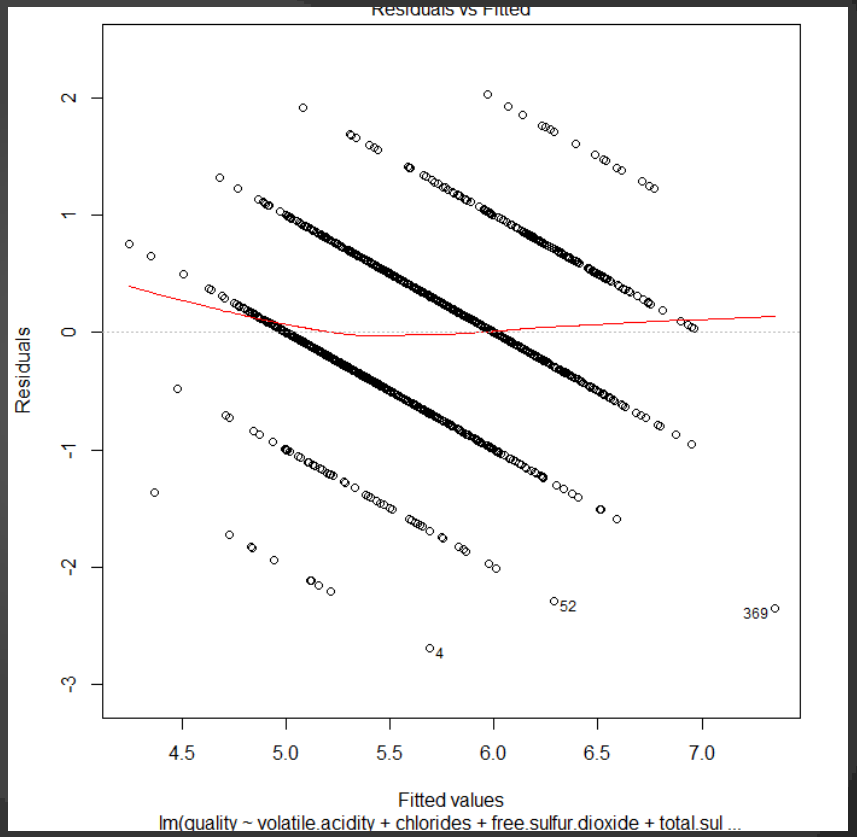
- 교호작용이 없을 때

회귀분석
(Regression Analysis



- 교호작용이 있을 때

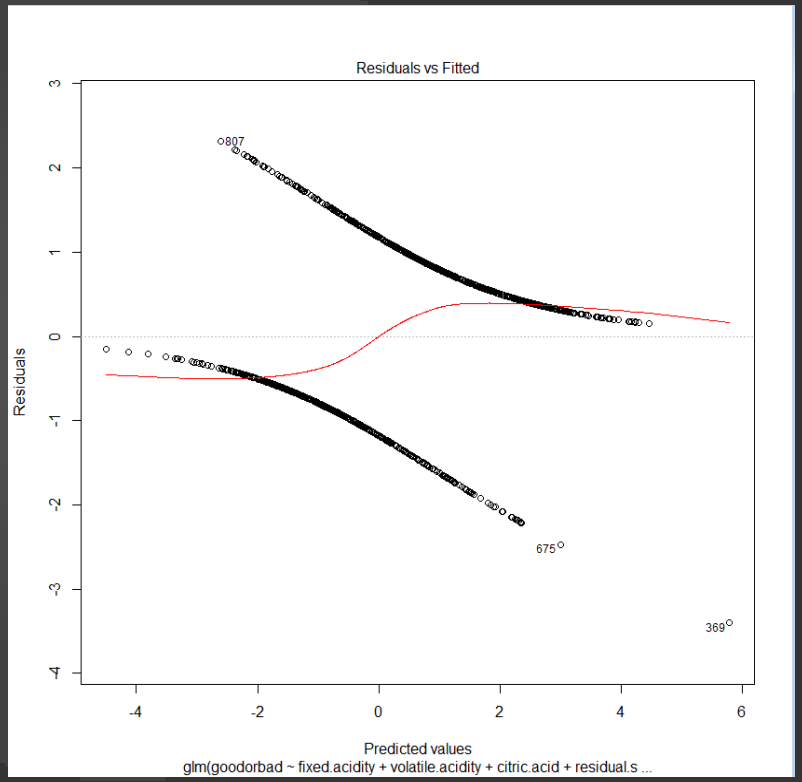
회귀분석
(Regression Analysis



분석 기법

- 로지스틱 모형

회귀분석
(Regression Analysis



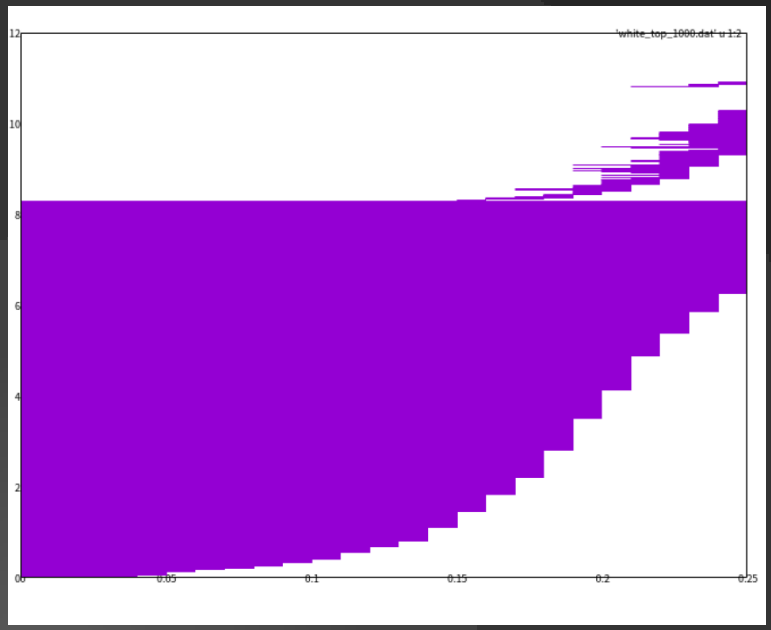
회귀분석
(Regression Analysis

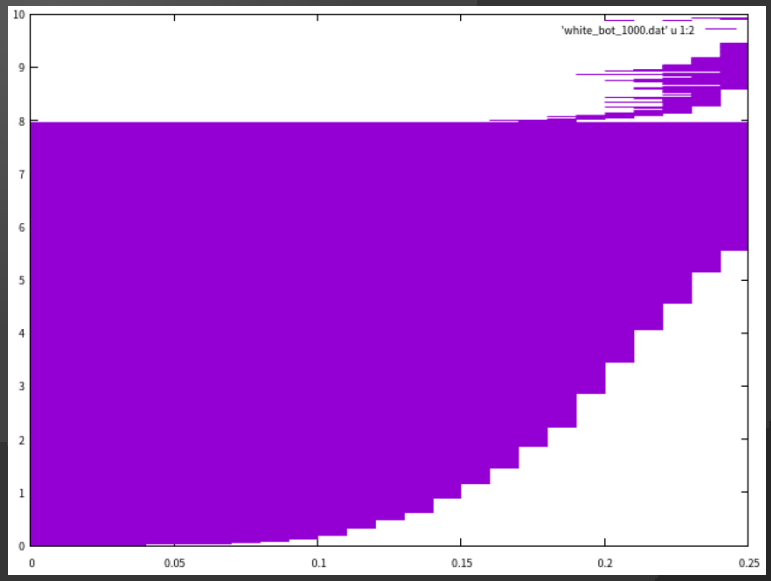
결론

-다양한 모델을 이용하여
회귀선을 추정하였지만,
11개의 변수와 품질간에는
상관관계를 발견하기 힘들었다.

분석 기법

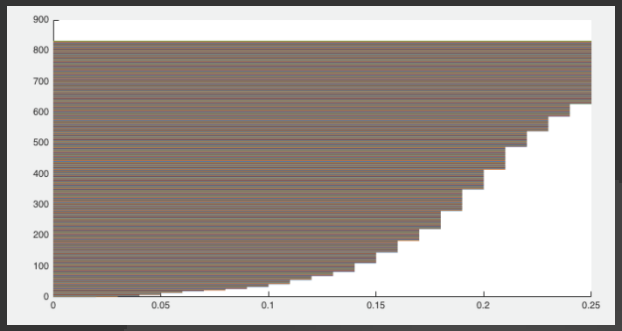
TDA

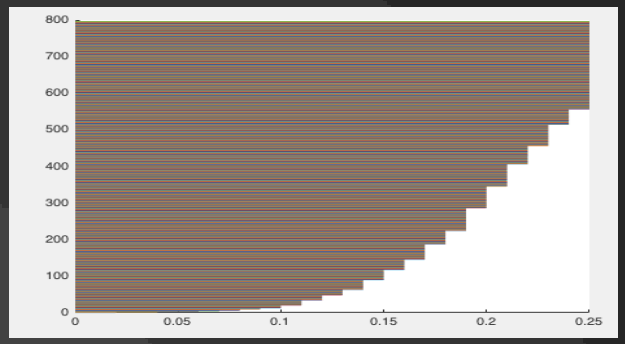

'white_top_1000.dat' u 1:2
12
10
8
6
4
2
0
0.05
0.1
0.15
0.2
0.25

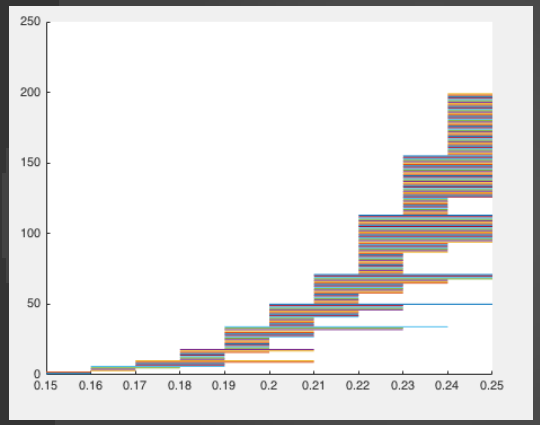

'white_bot_1000.dat' u 1:2
10
9
8
7
6
5
4
3
2
1
0
0
0.05
0.1
0.15
0.2
0.25

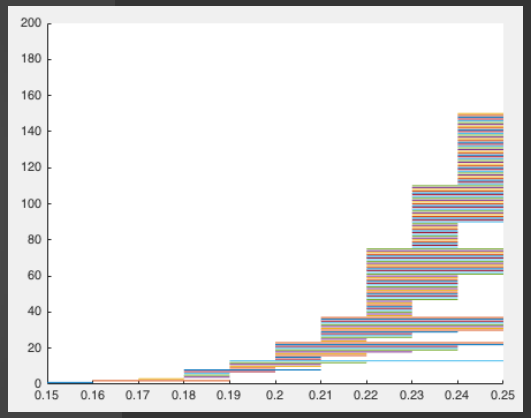
분석 기법

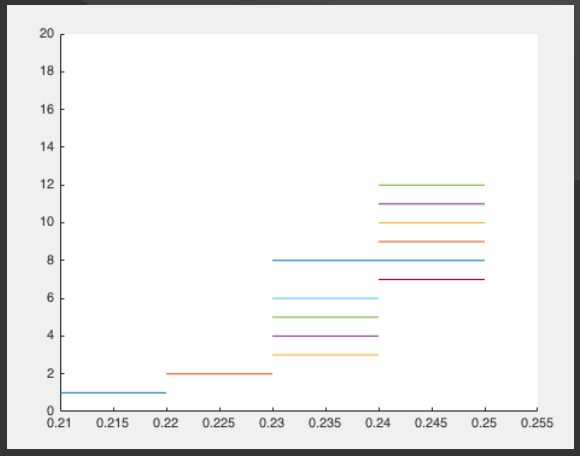
TDA

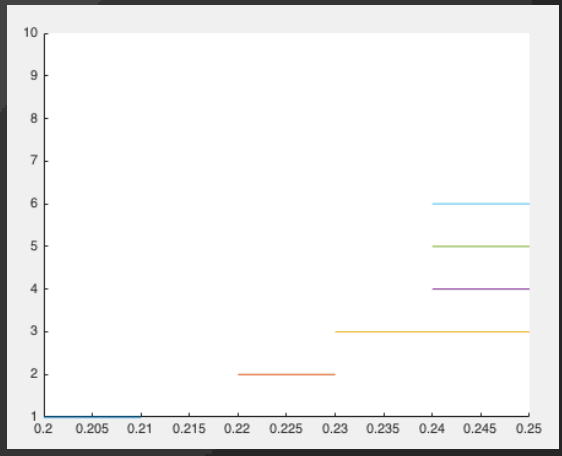

900
800
700
600
500
400
300
200
100
0
0
0.05
0.1
0.15
0.2
0.25


800
700
600
500
400
300
200
100
0
0
0.05
0.1
0.15
0.2
0.25


250
200
150
100
50
0
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25


200
180
160
140
120
100
80
60
40
20
0
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25


20
18
16
14
12
10
8
6
4
2
0
0.21
0.215
0.22
0.225
0.23
0.235
0.24
0.245
0.25
0.255


10
9
8
7
6
5
4
3
2
1
0.2
0.205
0.21
0.215
0.22
0.225
0.23
0.235
0.24
0.245
0.25

요약

PCA

3차원으로 차원을 축소를 시켰을때 유의한 결과를 얻었지만
그 이상으로의 축소결과와 주요 인자를 구하지는 못하였다.

회귀분석
(Regression Analysis

Forward selection(전진선택법)을 이용하여 적합회귀식을 구하였지만
결정계수가 매우 낮았고 관측값역시 유의미한 결과를 얻을 수 없었다.

TDA

Persistent Homology를 Barcode로 시각화 하여 두 집단 간 데이터
구조의 차이를 관찰해보려 했지만 의미있는 결과를 얻을 수 없었다.

“11가지 변수로 와인의 종류를 구분할 수 있을까 ?”

후진제거법(Backward elimination method) 을 이용하여 회귀식 추정하기

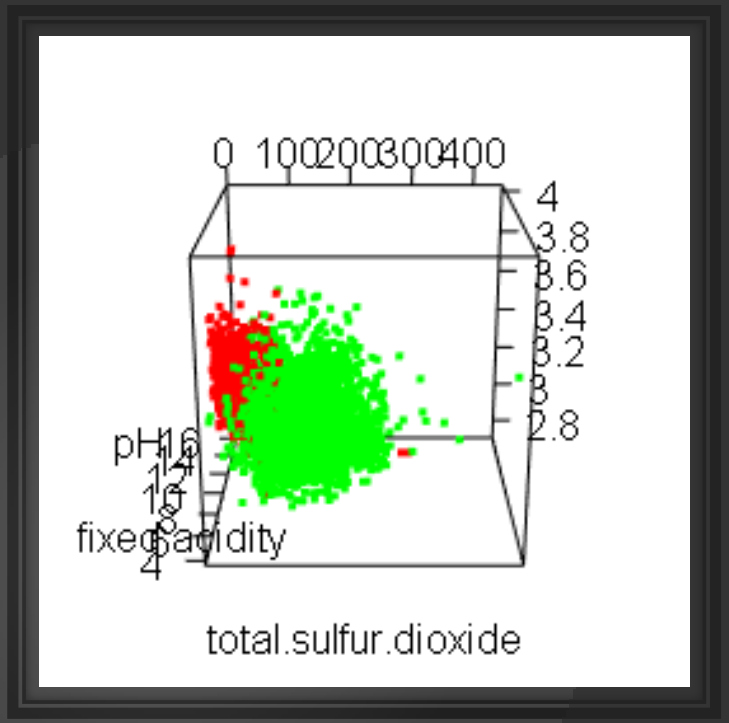
Regression Analysis

Quality ~ 2.5000675+(-1.3242155)×(fixed acidity)+(-0.7656644)×(total.sulfur.dioxide)+(-0.22314675)×(pH)

Residual standard error: 0.3345 on 1591 degrees of freedom
Multiple R-squared: 0.7696, Adjusted R-squared: 0.7856
F-statistic: 127.6 on 7 and 1591 DF, p-value: < 2.2e-16

적용 - 회귀분석

Regression Analysis



: 화이트 와인

: 레드 와인

데이터들의 분리가 잘 이루어졌다.

Regression Analysis

->주요 요소들에 의해 와인의 종류를 구분할 수 있다.

결론

1. PCA기법을 제외하고 나머지 분석기법을 사용했을 땐 11가지 요소들이 와인의 품질에 영향을 준다는 결론을 내릴 수 없다.

2. 회귀분석을 통하여 11가지 요소들에 따라 와인의 종류를 구별할 수 있다는 결과를 얻었다.