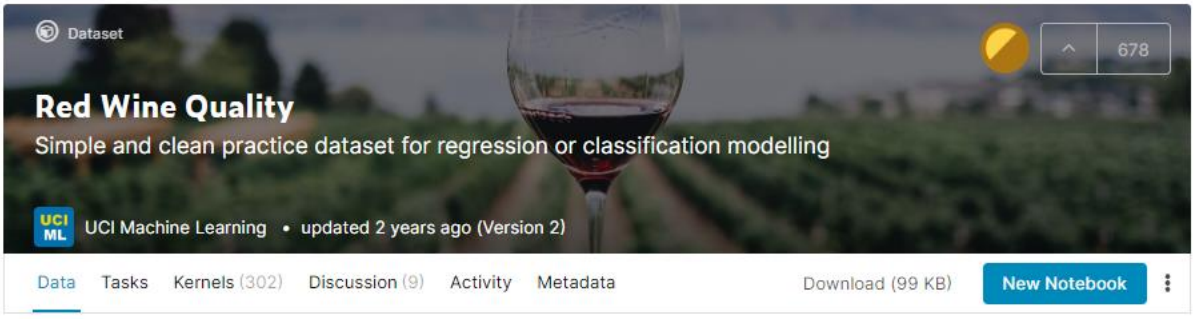
레드 와인 품질 데이터분석 보고서

수학과
이은주

관찰 방법 및 데이터 설명

어떤 와인이 좋은 와인이라고 할 수 있을까? 데이터를 쉽게 구할 수 있는 사이트인 Kaggle 데이터 중 레드 와인 품질 (Red-Wine Quality)을 이용하여 좋은 와인에 영향을 주는 요인을 찾아보자.



< 그림0. Kaggle-Red Wine Quality >

사용 데이터 : https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009#winequality-red.csv

https://archive.ics.uci.edu/ml/datasets/wine+quality 에도 공개되어 있는 데이터로 그룹화를 처음 시작하는 사람들에게 추천하는 데이터이다. 2009년 10월 27일에 등록되었으며, 판매량과 판매위치 등에 대한 정보 없이 화학적 데이터만을 가지고 있다. 자세한 데이터 출처는 아래 참고문헌에 남긴다.

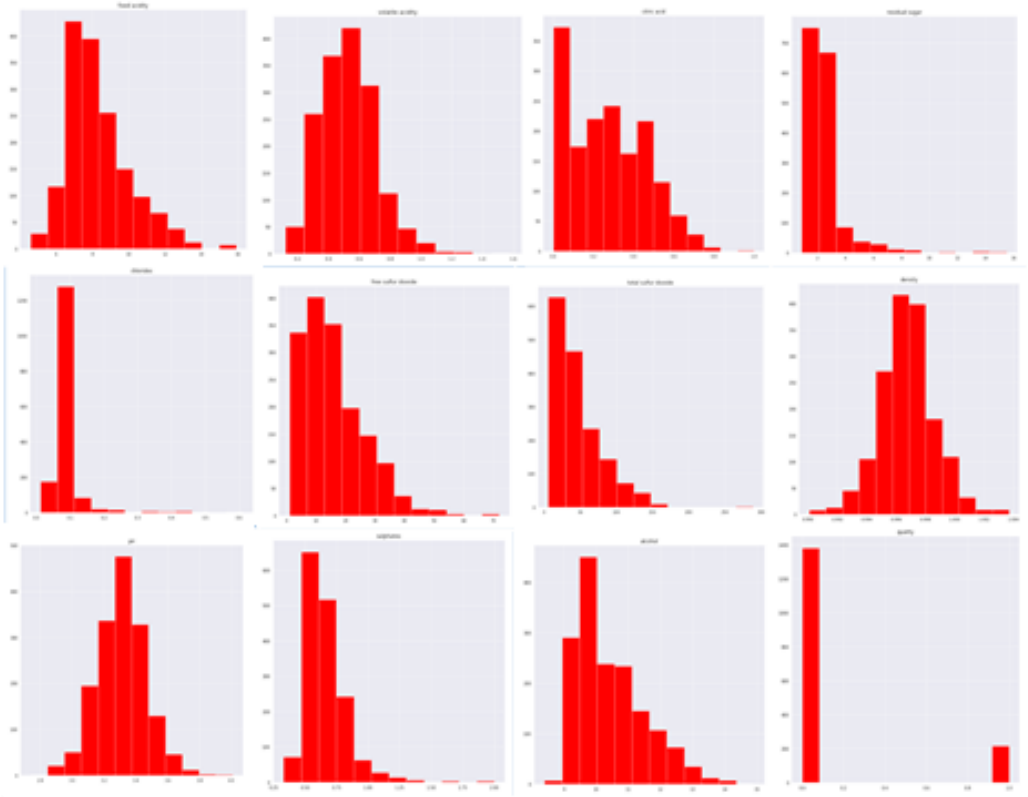
데이터 사이즈는 1599 rows x 12 columns으로 11개의 변수와 1개의 label를 갖는다. 원본데이터의 label은 score(0 – 10) 값으로 되어 있지만, 이후에 분류 문제로 생각하여 logistic regression 모델을 돌려볼 예정이므로 score 0 ~ score 6.5 는 label 0 (bad)로 score 6.5 ~ score 10 은 label 1 (good) 으로 변경하였다.

[feature 설명]

변수 이름	자료형	변수 설명
1. fixed acidity	연속	결합 산도 (쉽게 증발하지 않는 산), 와인의 산도 결정
2. volatile acidity	연속	휘발성 산도, 와인의 향과 연관이 많음
3. citric acid	연속	시트르산, 와인의 신선함과 관련
4. residual sugar.	연속	잔여 설탕, 와인의 단맛을 결정
5. chlorides	연속	와인의 소금 양, 와인의 짠맛의 원인
6. free sulfur dioxide	연속	free 이산화항
7. total sulfur dioxide	연속	총 이산화황
8. density	연속	밀도, 와인의 무게감을 의미
9. pH	연속	Ph, 산성 염기성 정도
10. sulphates	연속	이산화황 가스
11. alcohol	연속	알코올, 와인의 단맛과 연관
12. Quality (label)	이산	0(Bad) or 1(Good)

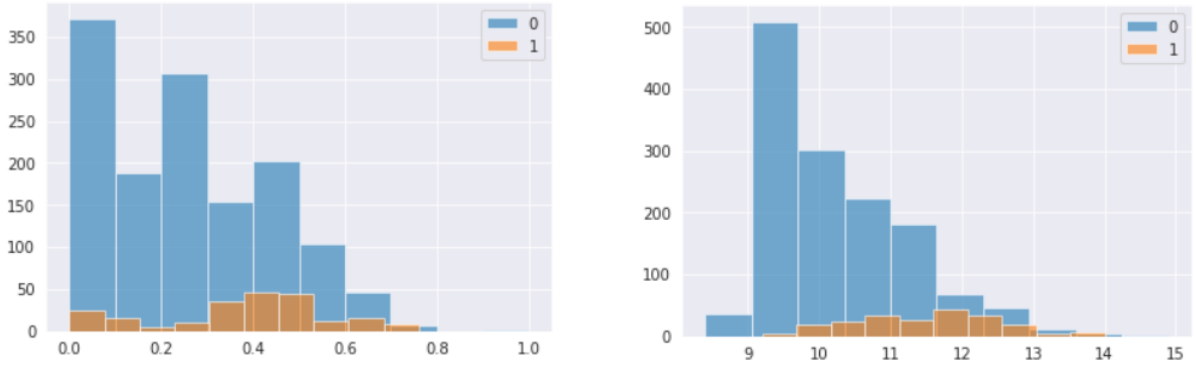
데이터분석

- 데이터분포 확인



< 그림 1. 변수들의 히스토그램 >

마지막 히스토그램은 label 데이터의 히스토그램으로 0 데이터가 1데이터 보다 많음을 알 수 있다. 4번 변수(residual sugar)와 5번 변수(chlorides) 그래프에서 한쪽으로 치우쳐 있는데 이를 통해 잔여 설탕과 소금은 대부분 비슷한 정도의 비율로 와인이 만들어짐을 확인 할 수 있다. label별로 히스토그램을 겹쳐 그려본 결과 분포가 눈에 띄게 다른 부분은 2가지 변수이다 .



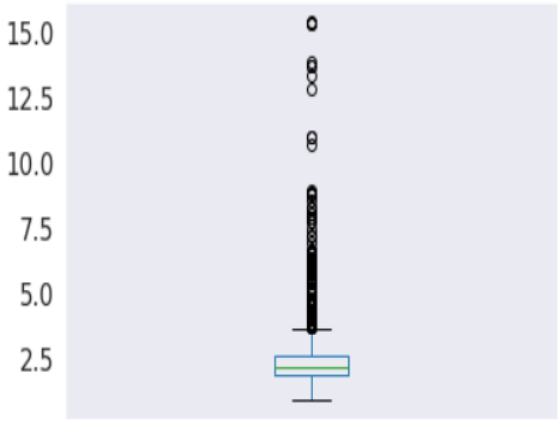
< 그림 2. citric acid 히스토그램 과 Alcohol 히스토그램 >

citric acid 변수에서 label 0인 부분은 0.0과 가까운 부분이 많은 반면, label 1인 부분은 0.4 근처이고 Alcohol 변수에서 label 0인 부분은 대부분 10 이하이면 label 1인 부분은 고루 퍼져있고, 12근처에 가장 많음을 알 수 있다.

변수 이름	평균	표준편차	최소값	25%	50%	75%	최대값
fixed acidity	8.319	1.741	4.600	7.100	7.900	9.200	15.900
volatile acidity	0.528	0.179	0.120	0.390	0.520	0.640	1.580
citric acid	0.271	0.194	0.000	0.090	0.260	0.420	1.000
residual sugar.	2.539	1.409	0.090	1.900	2.200	2.600	15.500
chlorides	0.088	0.0470	0.012	0.070	0.079	0.090	0.6110
free sulfur dioxide	15.875	10.460	1.000	7.000	14.000	21.000	72.000
total sulfur dioxide	46.468	32.895	6.000	22.000	38.000	62.000	289.000
density	0.997	0.001	0.990	0.995	0.996	0.997	1.004
pH	3.312	0.154	2.740	3.210	3.310	3.400	4.010
sulphates	0.658	0.169	0.3300	0.550	0.620	0.730	2.000
alcohol	10.423	1.065	8.400	9.500	10.200	11.100	14.900

< 그림 3. 자료 요약 >

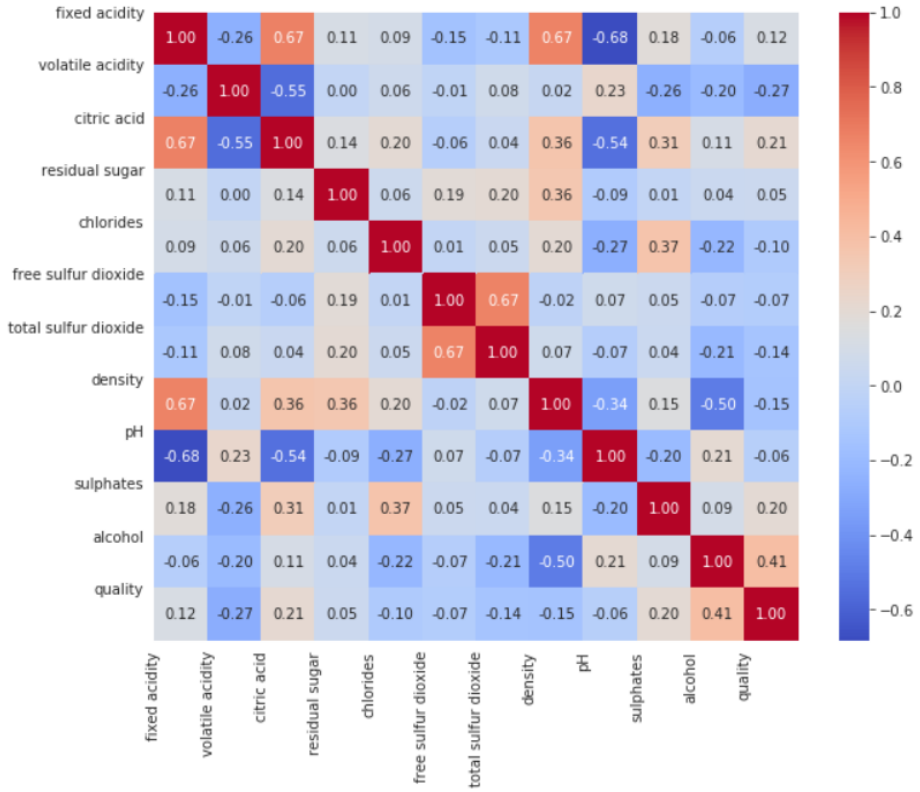
요약된 자료 중 (residual sugar)자료의 평균, 최소값, 최대값과 표준편차를 비교했을 때 이상값(outlier)이 있을 것이란 판단이 들어 boxplot을 그려 보았다.



< 그림 4. Residual sugar boxplot >

실제로 많은 이상치가 존재함을 확인 할 수 있다. 이러한 이상값은 이후 모델 성능을 낮아지게 하므로 미리 처리를 해주어야 한다. 예를 들면, 이상치를 삭제하거나, log값을 씌워주어 변수를 scale을 할 때 분포를 치우치지 않게 만든다.

- 변수간 상관관계 분석



< 그림 5. correlation matrix >

<그림 5.>에서 free sulfur dioxide, total sulfur dioxide 의 상관계수가 0.67로 약한 양의 상관관계가 있음을 알 수 있다. 두 변수 모두 이산화황에 대한 자료로 상관이 있는 변수이다.

fixed acidity 변수와 pH 변수 사이 상관계수가 -0.68로 약한 음의 상관관계가 있음을 알 수 있다. fixed acidity 변수는 결합 산도 (쉽게 증발하지 않는 산)을 나타내며, 산도가 증가 할수록 산성이 증가하므로 pH가 낮아지기 때문에 약한 음의 상관관계를 보인다.

- Logistic Regression 결과

Good = 217개, bad = 1382개

Train 1279개 데이터, test 320개 데이터로 나눈다. 11개의 데이터를 standardscaler를 이용해 scale 해준 후 모델을 돌린 결과

		Pred	
		0	1
TRUE	0	264	9
	1	34	13

< 그림6. Confusion matrix >

34개의 1종오류와 9개의 2종오류가 발생했다.

관찰 후 소감 및 참고문헌

우선 데이터를 선택하는데 어려움이 있었다. 이전에 해보았던 데이터 요약과 정리를 위주로 보고서를 작성하였는데, 통계적 지식이 부족해 그래프 속 의미를 도출하지 못한 부분이 많았다. 다변량분석 수업을 듣고 같은 데이터를 분석하여 비교해 볼 것 이다.

참고문헌 :

Kaggle,2020.03.25, https://www.kaggle.com/tolgahancepel/red-wine-quality-classification-analysis-eda

데이터 출처 : Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal
@2009