탐색적 데이터 분석(EDA)

**Wine Quality - White**

과목: 데이터기반 통계분석 시스템구축

제출일: 7월 24일 (화)

이름: 임 정

1. 개요

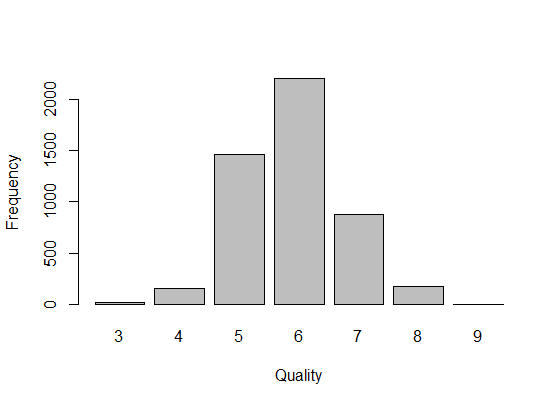
|  |
| --- |
| 'data.frame': 1599 obs. of 12 variables:  $ fixed.acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...  $ volatile.acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...  $ citric.acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...  $ residual.sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...  $ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...  $ free.sulfur.dioxide : num 11 25 15 17 11 13 15 15 9 17 ...  $ total.sulfur.dioxide: num 34 67 54 60 34 40 59 21 18 102 ...  $ density : num 0.998 0.997 0.997 0.998 0.998 ...  $ pH : num 3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...  $ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...  $ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...  $ quality : int 5 5 5 6 5 5 5 7 7 5 ... |

분석결과 Red wine 데이터에는 총 12개의 변수 1599개의 관측치가 있으며 12개의 변수 중에 11개는 연속형 1개는 범주형 변수로 나누어지는 것을 확인할 수 있었다. 따라서 다음과 같이 표현될 수 있다.

|  |  |
| --- | --- |
| 종속변수 | ‘quality’ |
| 설명변수(11개) | ‘fixed.acidity’, ‘volatile.acidity’, ‘citric.acid’, ‘residual.sugar’, ‘chlorides’, ‘free.sulfur.dioxide’, ‘total.sulfur.dioxide’, ‘density’, ‘pH’, ‘sulphates’, ‘alcohol’ |

2. 변수탐색

1. Quality



Group을 나눠주는 변수인 quality에 따른 데이터의 분포를 살펴본 결과는 위와 같다. 가장 낮은 등급인 3의 값에서부터 가장 높은 등급 9의 값을 가지는데, 위의 그래프를 살펴보면 중간 등급인 5~7 등급에 대다수의 와인이 분포하고 있는 것을 알 수 있다.

1. Fixed.quality

|  |  |
| --- | --- |
|  | Fixed\_acidity 변수를 등급별로 나누어서 분포를 살펴 보았을 때 위와 같이 결과가 나왔다. 등급 간에 평균이나 중앙값의 큰 차이가 없어 보인다. 하지만 산포는 등급이 올라갈수록 조금 줄어드는 경향이 있다. 하지만 이는 큰 차이가 아니라고 판단되고 그렇기 때문에 fixed\_acidity 변수는 등급의 차이를 확인하는데에 유의한 변수라고 보기 힘들다. |
|  |

.

1. Volatile.acidity

|  |  |
| --- | --- |
|  | Volatile.acidity 변수를 등급별로 나누어서 분포를 살펴 보았을 때, 등급에 따라 평균, 중앙값의 큰 차이가 보이지 않는다고 판단할 수 있다. 5,6 등급의 경우에 다른 등급에 비해서 아웃라이어가 많이 분포했지만 전체적으로 보았을 때에 등급별로 이 변수가 차이를 확인하는데에 유의한 변수라고 보기 힘들다. |
|  |

1. Citric.acid

|  |  |
| --- | --- |
|  | Citric.acid 변수를 등급별로 나누어서 분포를 살펴 보았을 때, 등급에 따라 평균, 중앙값의 큰 차이가 보이지 않는다고 판단할 수 있다. 이 변수 역시 5,6 등급의 경우에 다른 등급에 비해서 아웃라이어가 많이 분포했지만 전체적으로 보았을 때에 등급별로 이 변수가 차이를 확인하는데에 유의한 변수라고 판단하기 힘들다. |
|  |

1. Residual.sugar

|  |  |
| --- | --- |
|  | Residual.sugar 변수를 등급별로 나누어서 살펴본 결과, 평균, 중앙값이 모든 등급에서 비슷하고 boxplot을 보아도 같은 결과를 얻을 수 있었다. 따라서 등급별로 큰 차이가 있다고 판단하기 어렵다. |
|  |

1. Chlorides

|  |  |
| --- | --- |
|  | Chlorides 변수를 등급별로 살펴본 결과, 평균과 중앙값이 등급이 올라갈수록 조금씩 감소하는 것을 확인할 수 있었다. 또한 등급이 올라갈수록 산포도 역시 감소하고 있는 것으로 보아 이 변수는 등급 간의 차이를 확인할 수 있는 유의미한 변수라고 판단할 수 있다 |
| . |

1. Free.sulfur.dioxide

|  |  |
| --- | --- |
|  | Free.sulfur.dioxide 변수를 등급별로 살펴본 결과 평균이나 중앙값, 분산이 등급이 올라감에 따라 어떠한 패턴을 가진다고 볼수가 없고 그 값들 역시 모든 등급에서 비슷한 값으로 볼 수 있었다. 따라서 이 변수는 등급간의 차이를 나타내기에 어려운 변수라고 판단할 수 있다. |
|  |

1. Total.sulfur.dioxide

|  |  |
| --- | --- |
|  | Total.sulfur.dioxide 변수를 등급별로 살펴 보았을 떄, 등급이 올라갈수록 평균적으로 무수아 황산의 양이 작아지는 것을 확인할 수 있었다. 또한 높은 등급일수록 산포도 역시 줄어드는 것을 확인할 수 있기 때문에 이 변수는 등급 간의 차이를 확인하는데에 필요한 변수라고 판단할 수 있다. |
|  |

1. Density

|  |  |
| --- | --- |
|  | Density 변수를 등급별로 나누어서 살펴본 결과, 평균, 중앙값이 모든 등급에서 비슷하고 boxplot을 보아도 같은 결과를 얻을 수 있었다. 따라서 등급별로 큰 차이가 있다고 판단하기 어렵다. |
|  |

1. pH

|  |  |
| --- | --- |
|  | pH 변수를 등급별로 나누어서 살펴본 결과 등급이 올라갈수록 pH의 평균과 중앙값이 미세하게 커지는 것을 확인할 수 있었다. 높은 등급의 와인일수록 낮은 등급의 와인보다 pH가 평균적으로 높은 것이라고 판단할 수 있으므로 pH는 등급별로 차이를 보여줄 수 있는 변수라고 판단된다. |
| C:\Users\jayjunglim\wine_trash\plot.png |

1. sulphates

|  |  |
| --- | --- |
|  | Sulphates 변수를 등급별로 나누어서 살펴본 결과, 평균, 중앙값이 모든 등급에서 비슷하고 boxplot을 보아도 같은 결과를 얻을 수 있었다. 따라서 등급별로 큰 차이가 있다고 판단하기 어렵다. |
| C:\Users\jayjunglim\wine_trash\plot.png |

1. Alcohol

|  |  |
| --- | --- |
|  | Alcohol 변수를 등급별로 살펴본 결과, 3~6 등급의 와인의 평균, 중앙값은 비슷한 수치를 나타내는 것을 확인할 수 있었고 이에 비해 7,8 등급의 와인의 경우에는 이보다 조금 더 큰 수치를 가지는 것을 확인할 수 있었다. |
| C:\Users\jayjunglim\wine_trash\plot.png |