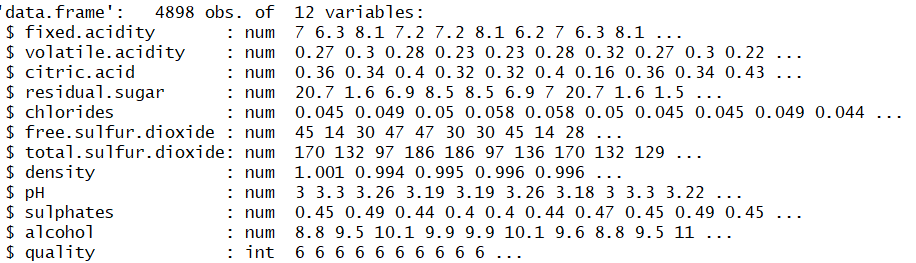
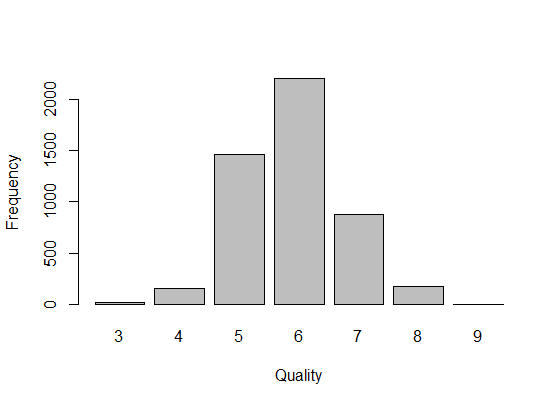
1. **Wine(White) Data**의 등급별 탐색적 자료 분석
2. 데이터 구성



Winequality-white data는 총 4898개의 관측치와 12개의 변수로 이루어져 있음을 확인할 수 있다.

Quality 변수를 제외한 모든 변수가 numeric의 형태를 가지고 있고 quality 변수만 integer 형태이다.



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. 변수에 따른 등급별 기초통계량 및 Plot
   1. fixed\_acidity

|  |  |
| --- | --- |
|  |  |
| 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 (단위 부피당 아세트산의 양)

|  |  |
| --- | --- |
|  |  |
| 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 (구연산)

|  |  |
| --- | --- |
|  |  |
| 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 (단위 부피당 당분의 양)

|  |  |
| --- | --- |
|  |  |
| 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 (단위 부피당 염화나트륨 양)

|  |  |
| --- | --- |
|  |  |
| 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 (단위 부피당 (free) 상태에 있는 무수아황산의 양)

|  |  |
| --- | --- |
|  |  |
| 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 (총 무수아황산의 양)

|  |  |
| --- | --- |
|  |  |
| 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 (밀도)

|  |  |
| --- | --- |
|  |  |
| 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 (산도)

|  |  |
| --- | --- |
|  |  |
| 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. | |

* 1. Sulphates (단위 부피당 아황산칼륨의 양)

|  |  |
| --- | --- |
|  |  |
| 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. | |

* 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. | |