**SNU 빅데이터 아카데미**

**-데이터 기반 통계분석 시스템구축-**

학생명: 허현

프로젝트명: 화이트 와인 데이터 EDA

EDA(Exploratory data analysis)의 목적은 데이터를 다양하게 살펴봄으로써 앞으로 분석의 방향성을 구상하고, 통계적 수치가 주는 함정에 빠지지 않게 하는 것에 있다. EDA 과정을 충분히 거치지 않으면 후에 분석 전체를 다시 시작해야 하는 상황에 처할 수도 있다. 이 보고서는 앞으로 어떻게 분석을 할지도 염두에 두고 EDA를 진행했다.

화이트 와인 데이터는 총 4898개의 관측 데이터와 12개의 특성 데이터로 구성되어 있다. 이 중 quality 특성은 총 3점에서 9점으로 구성되어 있으며 높을수록 와인의 품질이 좋았음을 나타낸다. 나머지 11개의 특성은 quality와 연관성이 있는 것으로 추정되는 특성들이다.

우선 반응변수인 quality 특성의 빈도를 알아보았다. 반응변수는 불균형성이 심했다.

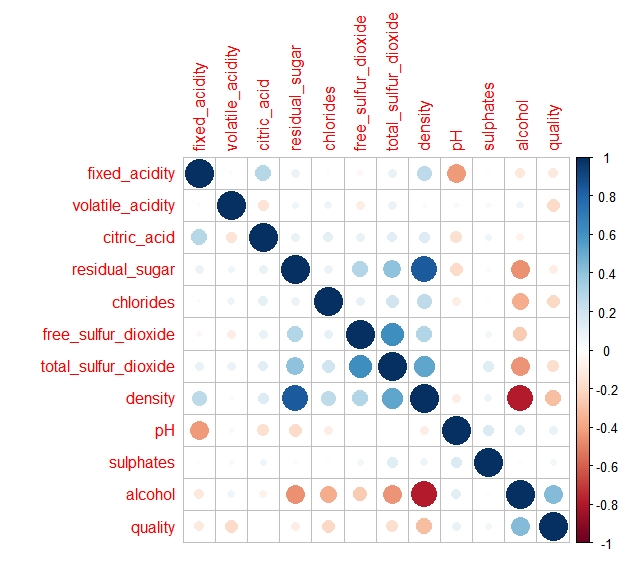
quality

3 4 5 6 7 8 9

20 163 1457 2198 880 175 5

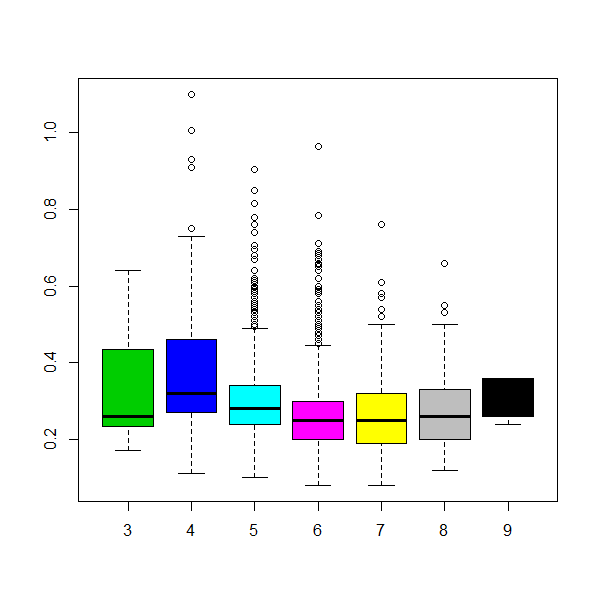
점수가 5, 6, 7점인 데이터는 일정 수준 이상으로 분석에 문제가 없을 것으로 보이지만 그 외의 점수 데이터, 특히 3점과 9점 데이터의 경우 추후 분석을 위해서는 데이터를 더 수집하거나 가지고 있는 데이터를 의도적으로 재활용하여야 제대로 분석이 가능할 것으로 보인다. 이런 불균형성 때문에 후에 박스플랏으로 시각화를 할 때 대부분의 아웃라이어가 5, 6, 7점에서 발생했다.

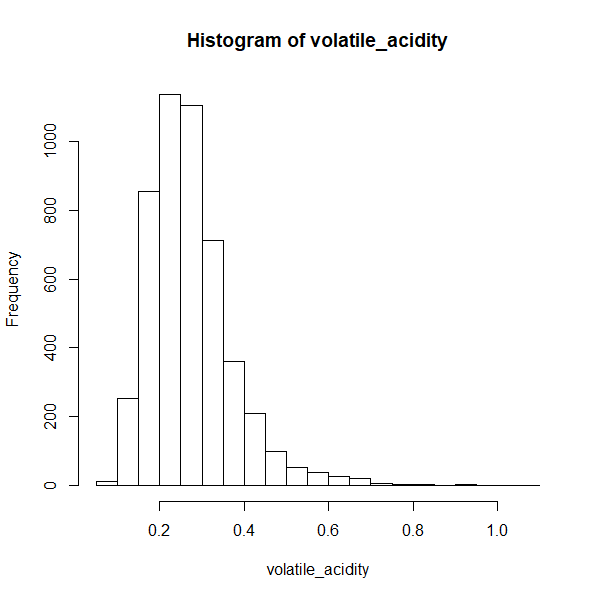
다음으로 모든 특성간 연관성을 알기 위해 correlation을 구하고 이를 시각화했다. 대각선 방향으로는 같은 데이터를 비교했기 때문에 모두 1이 나타나고, 연관성 분석은 A에서 B로 가는 방향성이 없기 때문에 대칭행렬의 속성도 가지고 있다. 아래의 그래프는 가장 양의 상관관계가 강한 1일 때 파란색이 짙고, 가장 음의 상관관계가 강한 -1일 때 붉은색이 짙게 나타난다. 이 수치를 볼 때 주의해야 할 점은 여기서 나타내는 상관관계라는 것은 선형성에 대해서만 설명력을 가지고, y = x^2 이나 원형 관계와 같은 관계에 대해서는 설명력이 떨어진다는 점이다. 이런 관계를 보이는 것으로 추정된다면 가지고 있는 데이터를 수정하거나 선형관계 외에도 분석이 가능한 다른 분석기법을 사용해야 한다.

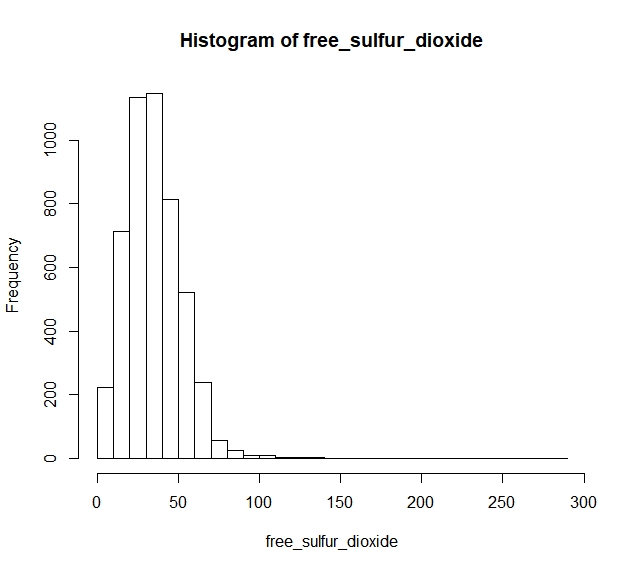
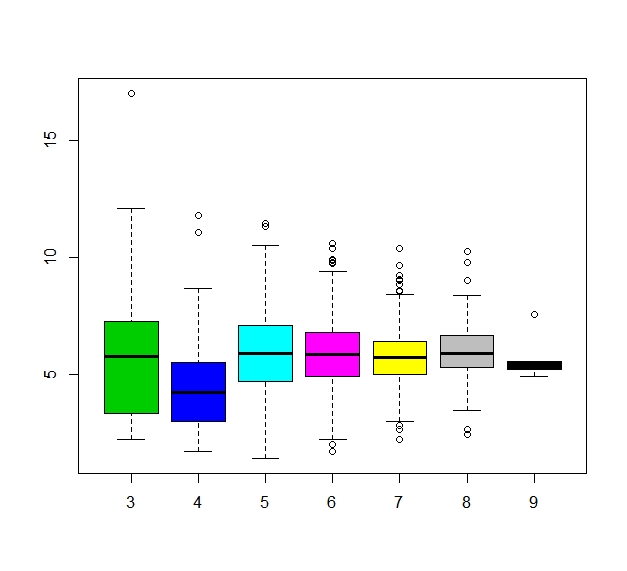


위 그래프에서 아주 강한 상관성을 보이는 관계는 alcohol과 density, 그리고 density와 residual\_sugar의 관계였다. 각 특성에 대해 자세히 알아보고자 검색을 해본 결과, alcohol 즉 알코올은 와인에서 물을 제외하고 가장 많은 성분으로 와인의 강도를 결정한다. 당분의 발효 과정을 통해 알코올이 생긴다. Density는 와인의 농도로서 알코올과 설탕 함량에 따라 결정된다. Residual sugar는 발효가 끝난 후 잔류하는 설탕의 양으로 1g/liter 이하인 와인은 드물고, 45g/liter 이상의 와인은 달콤하다고 여겨진다. 또한 위에서 언급한 density와 alcohol은 quality와의 상관성이 높은 것으로 보아 alcohol이 높을수록, density가 낮을수록 quality는 좋아질 것이라고 예측해 볼 수 있다.

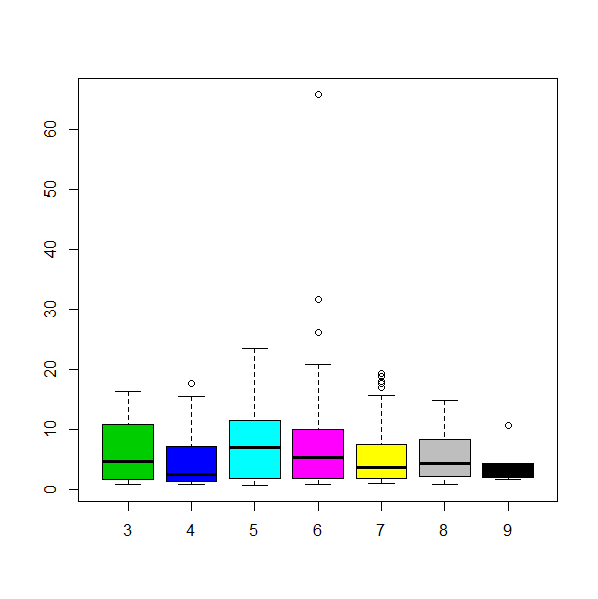
개별 변수에 대해 EDA 할 때는 그 변수를 앞으로 사용할 때 그대로 사용할지 아웃라이어를 제거해야할지 log나 루트 등을 씌워 형태를 변경시켜야 할지 등에 대해 고민했다. 사용하고자 하는 모델에 따라서는 어떠한 변형도 요구하지 않는 경우도 있지만, 형태를 변경했을 때 관계가 적어보였던 변수가 관계가 높아지기도 하기 때문이다. 일례로 자동차의 속도와 정지 시 제동 거리의 관계에서는 제동 거리가 속도의 제곱에 비례한다고 한다. 또한 되도록이면 전체적으로 보기보다는 quality별로 나눠서 시각화 하는 방법을 지향했다. Quality 속성이 이산형이기 때문에 전체적으로 정규분포 꼴을 따른다 해도 개별 속성으로 보면 정규분포 꼴을 따르지 않을수도 있기 때문이다. 다만 이 때 정규분포는 예시일 뿐 꼭 정규분포 꼴을 따라야 한다는 조건이 아니다.

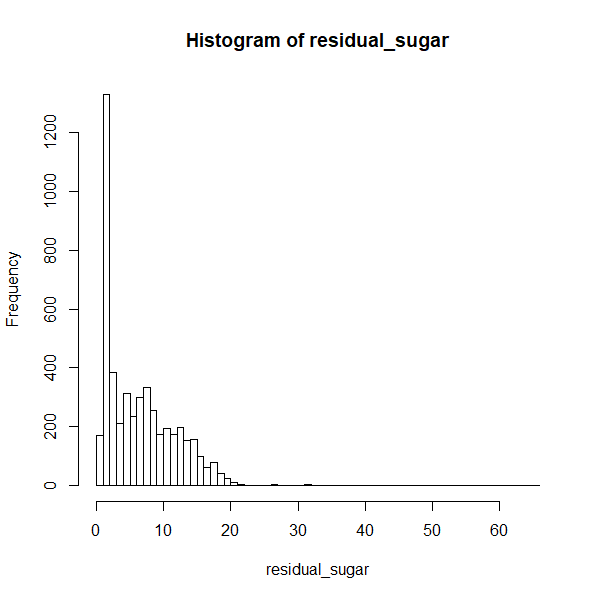
11개의 속성 중에는 후에 분석 시 고려해야 할 명확한 특징이 있거나 quality별로 차이가 명확한 속성도 있었고 그렇지 않은 것도 있었다. 여기서는 언급할 가치가 있는 특성에 대해서만 시각화 결과를 보이고 설명할 것이다.



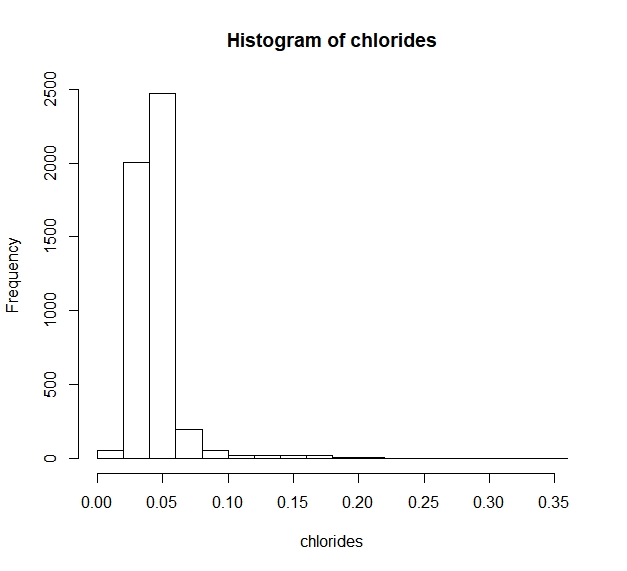
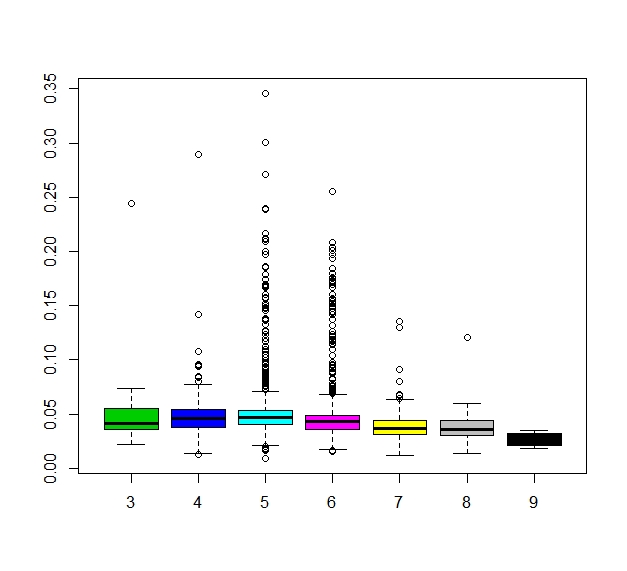
이 변수는 volatile\_acidity으로 휘발성 산이라서 와인에 함량이 너무 높으면 불쾌한 식초 맛이 날 수 있다고 한다. 박스플랏에서 4점부터 8점 사이의 아웃라이어들이 점점 밑으로 내려오는(산의 함량이 낮아지는) 형태를 보인다. 물론 아웃라이어는 대표성을 가지지 못하기 때문에 참고로만 봐야 할 것이다. 비슷한 변수로는 미생물 성장과 포도주의 산화 방지를 위해 들어가는 free\_sulfur\_dioxide가 있다. 산화 방지를 위해 들어가는 것이기 때문에 당연히 산화가 방지될 수준만 필요로 할 뿐, 그 이상은 quality에 역효과를 줄 수 있다.

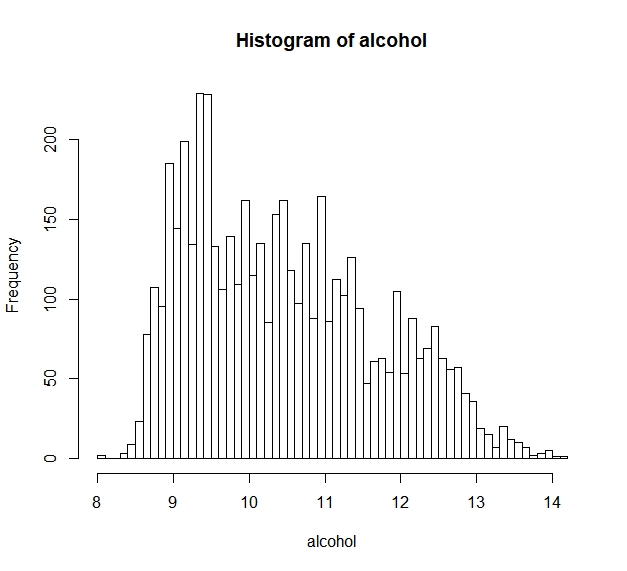
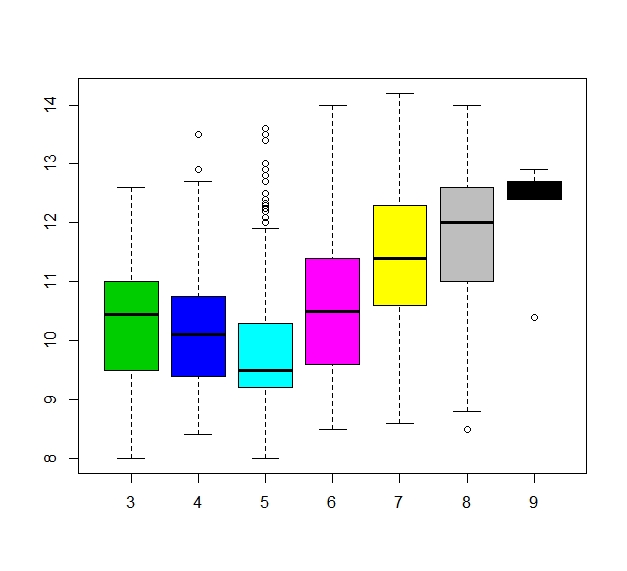
다음은 앞서 언급한 residual\_sugar이다. 전체적 빈도로 보면 3~5 정도쯤에서 굉장히 많은 양이 나오는 것을 알 수 있는데 박스플랏으로 그려본 결과 이 현상은 모든 quality에 대해 동일한 양상인 것을 알 수 있다.





Chlorides 변수는 residual\_sugar 변수와 비슷하게 나왔는데 chlorides는 소금의 양을 의미한다고 한다. 와인에는 일정 수준의 당의 양, 소금의 양 같은 것이 고착화된 것으로 보인다. 다만 이 변수는 상대적으로 아웃라이어가 더 많았기 때문에 추후 분석을 위해서는 아웃라이어에 대한 조치를 취할지 아닐지 고민해 봐야 할 것이다.



마지막으로 alcohol은 와인의 알코올 함량에 대한 변수이다.

우선 히스토그램 상으로는 알코올 함량이 다양했음을 알 수 있고 그 중에서도 3번 정도의 꼭지점을 볼 수 있다. 이를 고려하고 박스플랏을 보면 전체적으로 quality 3에서 5로 가면서 알코올 함량이 낮아지다가 5부터는 다시 높아지는 경향성을 볼 수 있다. 이를 통해 quality와 알코올 함량은 상관성이 있는 것으로 보이며, 이를 잘 활용하면 좋은 변수로 쓰게 될 것이다.