**SNU 빅데이터 아카데미**

**-데이터 기반 통계분석 시스템구축-**

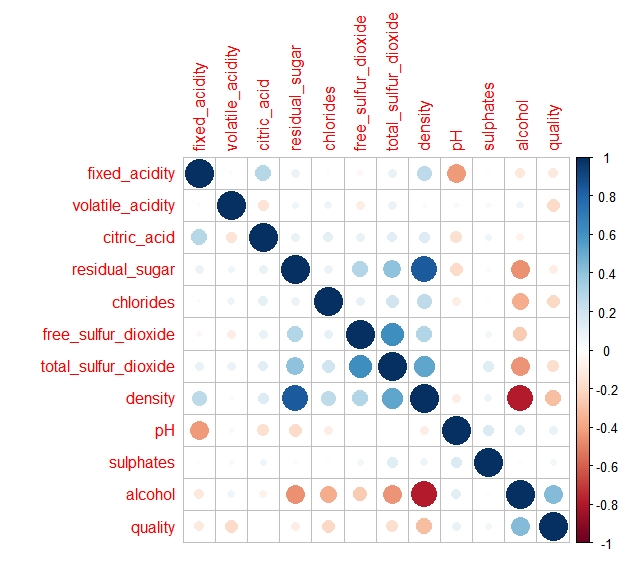
학생명: 허현

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.

먼저 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.