와인품질 예측 결과보고서

권이태, 김예원, 김태희

2021-06-12

---

1. 서론

최근 MZ세대(1981년~2010년 출생한 이들을 이르는 신조어)가 주요 소비층으로 급부상한 가운데, 주류 산업의 지형 역시 이에 맞춰 급변하고 있다. 이마트24가 공개한 2021년 1~2월 주류 구매 실적에 따르면 와인 매출은 전년 동기 대비 209% 증가하였으며, 해당 성장세에는 MZ세대의 5만원 이하의 중저가 와인 소비량 증가가 주요한 것으로 알려졌다[1-3]. 가성비(가격에 대한 성능 비)를 중시하는 MZ세대의 특성상, 소주와 맥주로 대표되던 기존 주류들과는 달리 고급스러운 분위기를 낼 수 있는 와인이 인기를 끌었다는 것이다. 이처럼 요즈음엔 레스토랑에서 소믈리에와 함께 와인을 즐기는 것이 아니라 집에서 간편하게 와인을 즐기는 '홈술' 문화가 정착하고 있다. 그러나 와인에 대한 전문적인 지식 없이 와인의 품질을 판단하는 것은 불가능하기에, 가정이나 매장에서 와인 품질을 간단하게 평가할 수 있는 기술의 필요성은 더욱 커지고 있다.

이 프로젝트에서는 이러한 수요에 맞추어 와인의 다양한 화학적/물리적 특성을 이용해 와인의 품질을 예측하는 모델을 구축하고자 하였다. 휘발산, 시트르산, 이산화황, 알코올 등은 그 양을 간단히 측정할 수 있을 뿐더러 와인의 맛과 향, 숙성에 중요하다고 알려져 있다. 또한, 우리가 와인을 평가할 때 중량감을 그 중요한 기준으로 여긴다는 점에서 밀도도 중요한 기준이 될 것이다. 이때 화합물들이 너무 많거나 밀도가 너무 높을 경우와 그 반대의 경우 모두 음용자에게 거부감을 줄 수 있다는 점에서, 각 수치들이 각각의 중앙값에 가까울수록 품질이 높을 것이라는 가설을 기반으로 분석을 진행하였다. 탐색적 자료 분석을 통해 와인 품질에 따른 각 변수들의 중간값과의 차이를 시각화하고 이를 상관분석해 가설이 예측에 적절한 수단임을 보였으며, 레드 와인과 화이트 와인을 분리하여 변수들의 산점도를 그림으로써 각각을 분리해 분석하는 것이 합당함을 확인하였다. 이후 랜덤포레스트에서 적절한 트리 수와 분기 수를 이용해 와인 품질 예측을 수행하였으며, 평가 자료에서 약 65%의 정확도를 얻어냈다. 범위를 넓힐 경우 약 95% 이상의 정확도를 얻어냈기에, 해당 예측 방법은 충분히 사용될 가능성이 있을 것으로 생각된다.

와인 소비와 품질 평가에 대한 수요가 늘어남과 동시에, 가정이나 영세상인들의 수제 와인 양조 역시 꾸준히 진행되고 있다. 특히 기획재정부가 '2018 세법 후속 시행령 개정안 마련'에서 소규모주류제조면허 대상에 과실주를 포함하기로 결정한 것이 결정적이었다[5]. 이에 따라 기존 와인 양조업계는 경쟁력을 얻기 위해, 개인들은 자신의 입맛에 맞는 와인을 양조하기 위해 와인 양조 시 첨가해야 하는 물질들에 대한 연구를 계속 진행하고 있다. 그 중 대표적인 것이 적절한 이산화황의 농도에 대한 논의다. 프레스크 아이슬 와인협회(Presque Isle Wine Cellars)는 이산화황이 와인 저장에 도움을 주며 몇몇 효모와 아세트산 박테리아, 젖산 박테리아의 생장을 방해한다고 언급하고 있다[6]. 이산화황은 효모와 같은 혐기성 생물들의 대사 과정에서 자주 발견되는 물질로, 그 농도는 이들의 활성에 주로 억제하는 방향으로 큰 영향을 미친다[7]. 발효 과정에서 효모는 와인의 품질에 관여하는 모든 화학물질의 대사를 담당하며, 이는 곧 효모와 다른 박테리아들의 활성 정도 조절이 와인의 맛을 결정하는 아주 중요한 요인이라는 것이다. 예를 들어, 만약 산미가 풍부한 와인을 양조하고자 한다면, 효모의 시트르산 대사량에 이산화황이 어떤 영향을 미치는지 확인한 이후, 와인의 시트르산 양이 가장 많게 되는 농도의 이산화황을 넣음으로써 산미를 임의로 조절할 수 있을 것이다[8-9]. 이처럼 이산화황의 농도를 조절하면 미생물들의 활성을 조절함으로써 와인으로 하여금 원하는 맛과 향을 가지도록 할 수 있다.

이 보고서의 기타 연구 가설은 이러한 예측을 직접 확인하고 이산화황 농도에 따른 각 성질들의 차이를 분석하기 위하여, 교락 요인인 '효모의 활성'을 조절하는 주요 변수가 이산화황이라는 가설을 세웠다. 그 다음, 탐색적 자료 분석을 통해 효모의 활성과 관련되어 있다고 알려진 변수들과 이산화황에 대하여 상관관계를 맺고 있는 변수인 시트르산, 휘발산, 잔당, 알코올을 발굴하였다. 이들에 대하여 이산화황의 양에 따라 각각의 양, 그리고

비율이 어떻게 달라지는지 확인하였고, 이를 통해 각각의 양을 조절하기 위해 이산화황의 양을 어떻게 해야 하는지를 분석하였다. 이러한 결과는 바디감, 중량감 등의 다양한 와인의 특성과 이산화황의 양이 어떤 관계를 맺고 있는지에 대한 기초적인 식견을 제공함으로써 다양한 와인의 양조에 도움이 될 것이다.

2. 분석의 주요목적

2-1. 와인 품질 가설

와인 품질 가설 : 변수의 값이 각 변수의 중앙값에 가까울수록 quality가 높을 것이다.

와인의 밀도, 이산화황, 시트르산 등은 너무 적으면 와인의 풍미와 식감을 저해하여 맛이 없게 만들지만, 너무 많을 경우에는 거부감을 일으킬 수도 있다. 이 보고서에서는 위와 같은 일반적인 상식을 통계적으로 보이고자 했다. 특히, 우리는 그 적정치를 중앙값이라고 둔 후 각 변수가 그들의 중앙값으로부터 얼마나 떨어져 있는지가 품질에 큰 영향을 미치며, 그 절댓값이 클수록 품질은 낮아짐을 보이고자 했다. 이후 이를 기반으로 다양한 예측 방법을 통해 와인의 품질을 높은 정확도로 예측하고자 했으며, 오차가 나더라도 큰 문제가 없도록 RMSE 역시 낮은 모형을 추구하였다. 와인의 품질은 음용자의 취향에도 기반하며 이러한 기술이 사용될 주 영역은 고급 레스토랑이 아니라 온라인 쇼핑몰 등이라 예상하였기에, 다양한 방법의 적용을 통해 0.6 이상의 정확도를 얻고자 하였다.

그러나 분석 과정에서 데이터 처리를 하면 중앙값으로부터 해당 값이 얼마나 떨어져 있는지를 기반으로 분석하게 되는데, 이 경우 대소에 대한 정보가 사라진다는 큰 단점이 있었다. 즉 중앙값이 50인 자료에 대해서 0과 100이 동일하게 처리되는 것이다. 또한, 최적의 와인이 가지고 있는 변수들의 양이 과연 중앙값인지에 대해서는 잘 알려져 있지 않기에, 중앙값에 대하여 자료를 처리하는 것은 비합리적일 수 있다. 이와 더불어 학습에 이용될 자료 크기가 제한되어 있으며 높은 복잡도의 분석 프로그램을 돌릴 만한 하드웨어적 여건이 되지 않기에 높은 정확도로 와인 품질을 추측하기에는 무리가 있다. 이는 3이나 9와 같이 그 품질의 와인 수가 적어 특징을 파악하기 어려운 경우에 더욱 두드러진다.

2-2. 기타 연구 가설

기타 연구 가설 : 변수 간의 관계에서 중요한 교락 요인은 '효모의 활성'이며, 이산화황은 이를 조절한다.

기존 연구와 와인 관련 사이트들로부터 이산화황이 와인 양조에 주요한 요인임을 파악하였다. 여기서 보이고자 한 것은 크게 두 개로, 첫째는 이산화황 그 자체가 다양한 변수의 양에 영향을 미친다는 것이다. 와인에 포함된 물질들은 원래 포도에 포함된 물질, 첨가되는 물질, 발효 결과 형성되는 물질로 구분할 수 있을 것이다. 이 중 조절하기 가장 쉬운 것은 첨가되는 물질인 이산화황으로, 이에 따라 효모의 활성이 달라지며 반응물들과 생성물들의 양을 보일것이다. 여기서는 상관분석과 산점도의 확인을 통해 이를 확인하고자 했다. 둘째는 이산화황이 발효에 미치는 영향이다. 발효는 혐기성 상태에서 효모에 의해 주로 매개되며, 반응물과 생성물의 비를 조절하는 것은 보통 이들이 가진 생화학적 경로의 활성이다. 따라서 이산화황이 효모의 활성에 영향을 미친다면 생성물과 반응물 간의 비율이 달라질 것이다. 따라서 교락 요인인 이산화황을 통제하고 시트르산/휘발산, 알코올/당의 관계를 봄으로써 와인 양조 중의 효모도 실험실에서의 그것과 동일한 이산화황에의 감수성을 가지고 있는지, 혹은 다른 요인 때문인지 확인하고자 했다.

그러나 앞서 언급된 문제와 같이 이 경우에도 학습에 이용될 자료 크기가 제한되어 있으며 분석 지식과 하드웨어의 부족이 있었다. 또한 이와 같이 반응물과 생성물의 관계를 보기 위해서는 초기 상태 역시 모두 같다는 가정이 있어야 한다. 예를 들어, 모두 같은 양의 시트르산을 포함한 포도를 이산화황의 양을 달리해 가며 양조하였다면, 양조된 와인의 시트르산/휘발산 비는 그 활성에 대한 좋은 근거이다. 허나, 시트르산이 원래 많은 포도를 발효시킬 경우에는 시트르산이 적은 포도를 사용하였을 때보다 양조된 와인의 시트르산 양과 휘발산 양이 모두 많을 것이기에 이산화황의 영향이 가려진다. 여기서는 그 비율을 확인함으로써 그 효과를 최대한 줄여보고자 하였지만, 와인 양조장 내부의 환경은 실험실과 달리 다양한 물질이 혼재해 있고 그 반응식을 정확히 알 수 없기에 이것이 완전히 해결될 수는 없을 것이다.

3. 분석방법

3-1. 데이터 코드북

Variable	Description
quality	품질; 와인의 종합적인 품질
fixed acidity	고정산; 와인에 포함된 비휘발성 산의 총량
volatile acidity	휘발산; 와인에 포함된 휘발성 산의 총량
citric acid	시트르산; 와인에 포함된 시트르산의 총량
residual sugar	잔당; 발효 이후 와인 속에 포함된 잔당
chlorides	염화물; 염화물의 양
free sulfur dioxide	자유 이산화황; 분자 혹은 이온 상태로 존재하는 이산화황의 양
total sulfur dioxide	총 이산화황; 자유 이산화황과 다른 분자에 결합한 부착 이산화황의 합
density	밀도; 부피에 대한 질량의 비
pH	pH; 수소이온 농도에 음의 로그를 취한 값
sulphates	황산염; 황산염의 양
alcohol	알코올 도수
type	와인 종류(red = 적색/레드 와인, white = 백색/화이트 와인)

3-2. 와인 품질 예측

와인의 품질을 예측하기 위해 세운 가설인 '각 변수의 중앙값에 가까울수록 품질이 높을 것이다'는 과도한 산미나 부족한 중량감이 와인의 맛을 해칠 수 있다는 일반적 상식으로부터 비롯한 것이었다. 따라서 먼저 탐색적 자료 분석을 통해 데이터가 이러한 예측을 따르는지 확인하고자 하였다. 중앙값과의 차이의 절댓값을 새로운 변수에 할당한 후, 품질에 따른 그 분포를 상자 그림으로써 시각화하고 상관분석을 통해 가설이 합당한지 사전 검증하였다. 이때 중앙값을 이용한 이유는 각 변수들로부터 차이 절댓값의 합이 가장 작은 값이 중앙값이기 때문이다. 또한, 후의 분석을 위하여 주어진 `train.csv`의 데이터를 예측 모델의 설립에 사용할 학습용 자료와 평가용 자료로 8:2 비율 하에 분리하였다.

랜덤 포레스트는 와인 품질과 같은 범주형 자료의 값를 예측하는 데 많이 사용되는 분석 기법으로, 여러 개의 의사결정나무를 무작위로 형성하여 분류를 진행한다. 이 보고서에서는 학습용 자료를 바탕으로 구성한 랜덤포레스트를 이용하여 평가용 자료에서 품질을 예측하고, 이를 실제 품질과 비교하여 정확도와 RMSE(Root Mean Square Error)를 도출하였다. 또한 해당 과정에서 만드는 의사결정나무의 개수와 분기의 개수를 바꾸어가며 이들을 평가하였으며, 그 결과 가장 높은 효율을 보여주는 모델을 선택하고자 하였다. 이와 더불어, 와인 품질에 큰 영향을 미치는 변수가 무엇인지 역시 확인하였다. 이러한 분석을 바탕으로 최적의 의사결정나무 개수와 분기수를 선택하였다. 위의 랜덤포레스트 분석은 패키지 `randomForest`를 이용해 진행하였으며, 패키지 `Metrics` 등이 사용되었다.

인공신경망은 실제 동물의 신경망을 본따 만든 학습 알고리즘으로, 관측값들을 이용해 각 층 간의 연결을 조정하고 이를 바탕으로 평가자료의 값을 예측한다. 여기서는 앞서 전처리한 자료를 이용해 패키지 `nnet`에서 제공하는 함수로써 모형 구축과 예측을 수행하였다. 이때, 랜덤포레스트에서 의사결정나무를 여러 개 만들어 예측하듯 신경망을 여러 개 만든 이후 그 최빈값을 예측값으로 설정함으로써 오차율을 줄이고자 했다.

3-3. 기타 연구 가설: 와인 양조에서의 이산화황의 역할

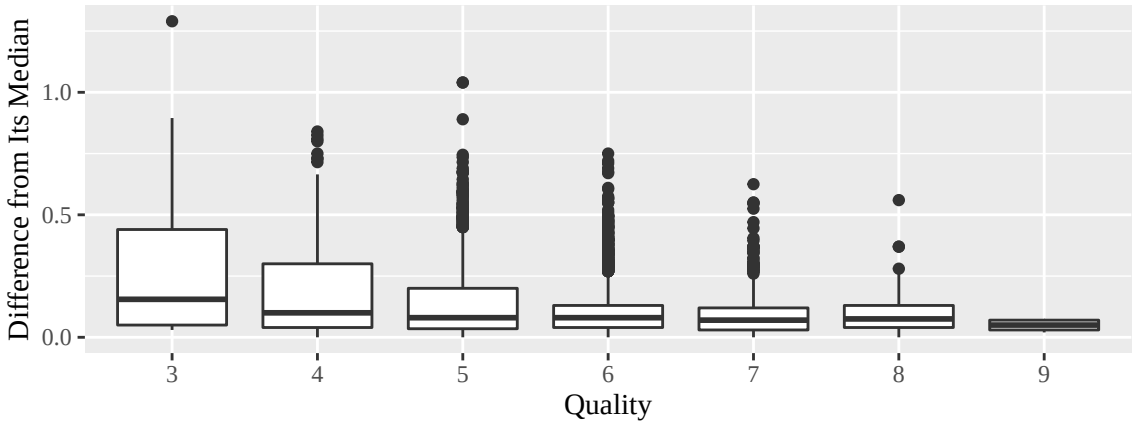
와인 종류를 제외한 수치형 변수들에 대해 히트맵을 그려 각각의 상관계수가 얼마나 되는지와 그것이 유의한지를 확인하고, 매커니즘 증거와 일치하는지를 검증하였다. 또한, 이산화황이 각 변수들의 관계에 어떤 영향을 미치는지 확인하기 위해서 상관분석을 이용하여 그 양에 따른 시트르산, 휘발산, 잔당, 알코올의 양을 분석하였다. 이때, 와인의 종류에 따라 그 경향성이 다를 것이라 예측하여 `type`을 기준으로 나누어 분석했으며, whisker range를 벗어나는 것들은 데이터에서 제거한 후 총 이산화황과 자유 이산화황 모두에 대해 시행하였다. 이후 분석하고자 하는 화학 반응이 반응물의 주 사용처이자 생성물의 주 생성 원인임을 보기 위하여 이산화황의 농도에 따라 4개 블록으로 블록화한 후 ANOVA를 시행하였다. 또한, 다중회귀를 통해 이산화황 농도가 일정할 때 당과 알코올, 시트르산과 휘발산의 비율 변화를 비교하였다.

4. 분석결과

4-1. 와인 품질 예측

와인 품질 예측을 위한 탐색적 자료 분석

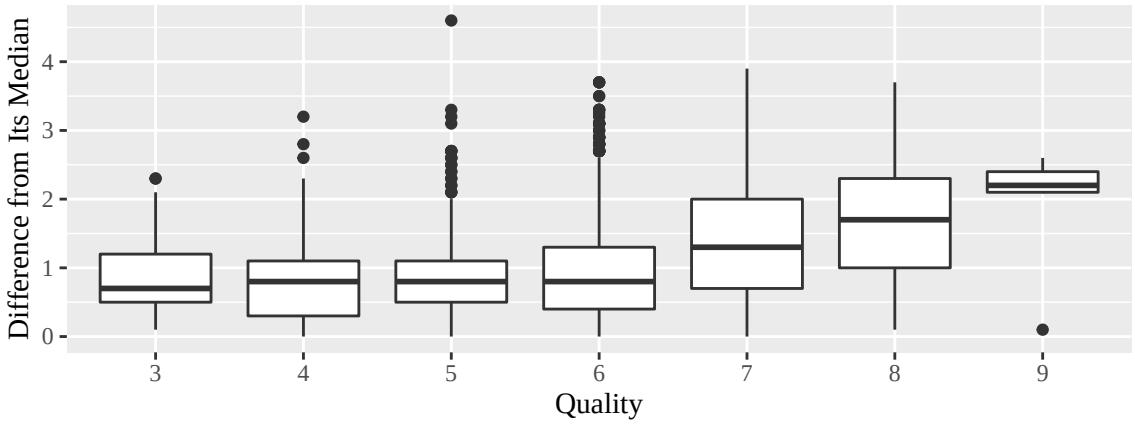
본격적인 분석을 진행하기에 앞서, 와인 품질에 따른 변수의 분포를 상자 그림을 통해 시각화하였다(Figure 1).



[Figure 1] Box plot of quality and differences of volatile acid content from its median

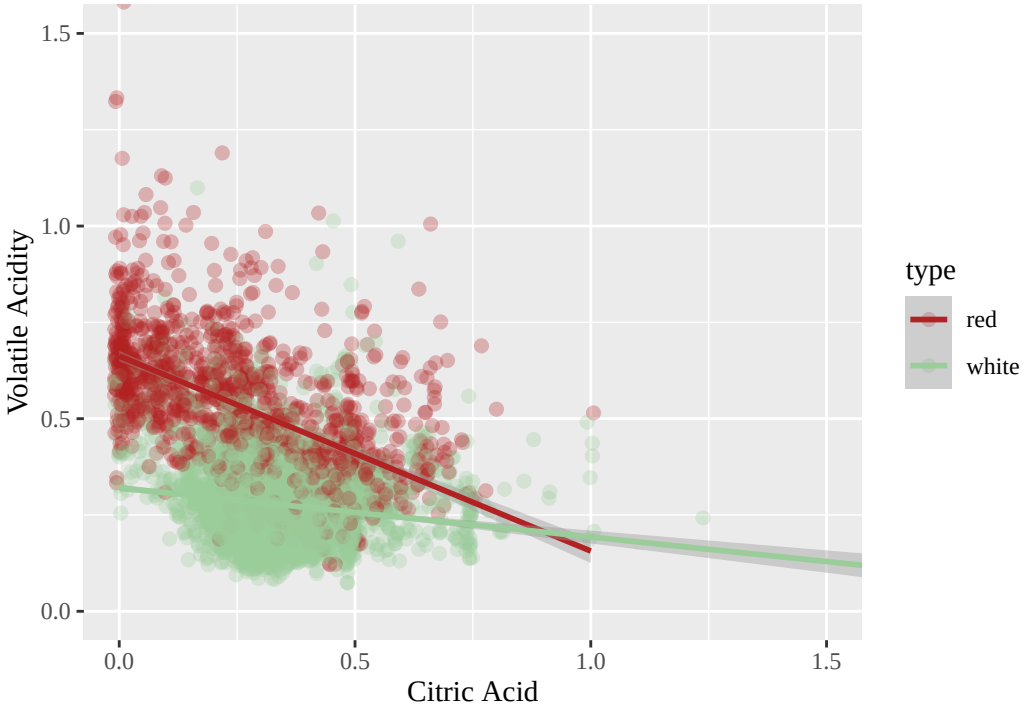
휘발산을 비롯한 다양한 변수들에서는 와인 품질이 높아질수록 그 중앙값과의 차이들이 전체적으로 낮아지는 경향성이 관찰되었으며, IQR 역시 감소해 특정 값에 가까울수록 품질이 높다는 가설과 어느 정도 일치하는 경향성이 나타났다. 즉, 중앙값에 가까울수록 와인 품질이 높다는 가설이 몇몇 변수에 대해서는 옳음을 시각적으로 확인할 수 있었다. 이를 다르게 말하면, 와인 품질과 중앙값과의 차이는 음의 상관관계를 이루어야 한다. 휘발산에 대해 둘 사이의 상관분석을 시행한 결과 상관계수는 -0.205로 음이었고 해당 값은 유의하였으며, 이는 탐색적 자료 분석을 통한 추측을 뒷받침한다.

와인 종류를 제외한 11개의 수치형 자료에 대해 모두 이와 같은 분석을 진행한 경우에는 시트르산, 고정산을 비롯한 9개의 변수에 대해서는 음 혹은 0에 가까운 상관계수 값이 등장하지만 pH와 알코올에 대해서는 양의 상관관계가 나타났으며, 알코올의 경우에는 상관계수 0.316이 유의하였다(Figure 2). 레드 와인과 화이트 와인에 대해 이런 추세는 유지되었으며, 이는 알코올이 적은 와인은 품질 5~6 가량의 일반적인 평가를 받지만 알코올이 많은 와인은 다른 요인들에 따라 3~4의 매우 낮은 품질 혹은 7~9의 높은 품질로 평가되기 때문으로 생각된다. 그러나 전체적으로 증가하는 경향성이 나타났다는 것은 가설과는 반대되지만 중앙값이 품질 예측에 여전히 중요한 기준으로 기능할 수 있음을 의미하기에, 후의 분석에 있어서도 이를 포함하여 관찰하였다(Supplementary 1-3).



[Figure 2] Box plot of quality and differences of alcohol content from its median

중앙값을 기반으로 분석을 진행할 때, 유일하게 전처리되지 못하는 변수는 이산형 변수인 와인의 종류다. 자료에는 적포도를 기반으로 하는 레드 와인과 청포도나 껍질을 까낸 적포도를 발효시킨 화이트 와인이 포함되어 있으며, 이들의 맛과 향, 평가 기준, 그리고 첨가물의 양은 서로 다른 관점에서 접근하여야 한다. 아래는 이를 확인하기 위하여 주요 변수인 휘발산과 시트르산의 관계를 와인 종류에 따라 표시한 산점도이다.



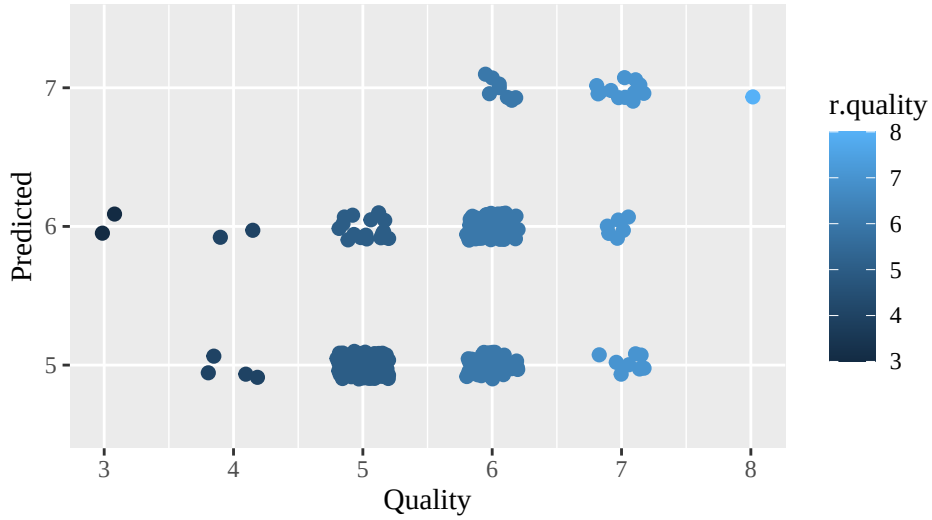
[Figure 3] Scatter plot of citric acid content and volatile acid content

위와 같이 레드 와인과 화이트 와인에는 그 주원료의 차이로부터 비롯된 변수의 양과 그 사이 관계의 차이가 있다. 따라서 이들에 같은 잣대를 들이대는 것은 불가능하기에, 추후의 분석에서 적포도주와 백포도주를 나누어 모형을 구축하는 것은 합리적이라는 결론을 내릴 수 있었다.

랜덤포레스트를 이용한 와인 품질 예측

랜덤포레스트를 적용하기에 앞서 적절한 의사결정나무의 개수를 찾기 위하여 이를 바꾸어가면서 레드 와인의 학습용 자료를 이용해 오차율을 계산하였다(Supplementary 4). 일반적으로는 그 개수가 증가할수록 오차율 역시 감소하였으며, 변수 `ntree`가 300 이상이 될 경우 큰 차이는 없는 것으로 관찰되었다. 이것이 증가하면 알고리즘의 시공간 복잡도가 증가하므로, 적절한 효율을 얻기 위하여 이 값은 300, 400, 500을 선택하여 분석을 진행하였다. 또한 여기에서 변수들의 중요도를 확인한 결과 알코올의 중요도가 가장 높았고, 뒤를 이어 휘발산, 총 이산화황, 황산염 등이 나타났다(Supplementary 5). 너무 많은 변수를 포함할 경우 과적합이 우려되기에, 분기의 수는 2, 3, 4를 선택하여 분석을 진행하기로 결정하였다.

레드 와인과 화이트 와인에 대하여 중앙값으로부터 차이의 절댓값으로 변수들을 새롭게 정의한 후 의사결정나무의 개수와 가지의 개수를 달리해가며 모델을 얻은 후, 평가용 자료에 이를 도입해 이를 평가하였다(Figure 4, Supplementary 6-7).



[Figure 4] Confusion matrix between predicted quality and red wine quality

Figure 4는 의사결정나무의 개수의 개수가 500개, 변수의 개수가 4개일 때의 랜덤포레스트 모형을 구축하고 이를 이용해 레드 와인의 평가용 자료에 대한 혼동행렬을 표현한 것이다. 실제 품질와 와인 품질이 정확하고 정밀하게 예측될수록 예측 방식의 효율이 좋다는 점을 고려하였을 때, 우상향의 대각선에 많은 점이 분포하며 그 주변에 대부분의 점이 있음은 예측 방식이 효과적임을 의미한다. 그 정확도는 레드 와인에 대해 0.632, RMSE는 0.734로 나타났으며, 다른 트리 개수나 분기 개수에 대해서도 정확도는 대략 0.6~0.65 정도로 나타났다. 따라서 랜덤 포레스트는 일반적으로 60%의 레드 와인에 대해서 그 품질을 정확히 맞출 수 있으며, 기준을 넓혀 원 품질과 예측 품질이 1 이하로 차이나는 것까지 허용한다면 95% 이상의 높은 정확도를 보였다. 동일한 방법으로 화이트 와인의 평가자료에 대한 예측을 했을 때는정확도가 0.624, RMSE가 0.701이었다. 최종적으로, 주어진 `test.csv`의 와인들 역시 변수 `type`을 기준으로 나눈 이후 전처리하여 와인 품질을 예측하였다(Appendix 1).

인공신경망을 이용한 와인 품질 예측

	3	4	5	6	7	8	9
3	1	0	0	0	0	0	0
5	2	4	86	34	4	0	0
6	0	2	19	72	21	1	0
7	0	0	2	9	13	0	0

	3	4	5	6	7	8	9
4	1	4	2	2	0	0	0
5	1	22	149	74	3	0	0
6	0	8	99	279	91	14	1
7	0	0	1	30	44	13	0

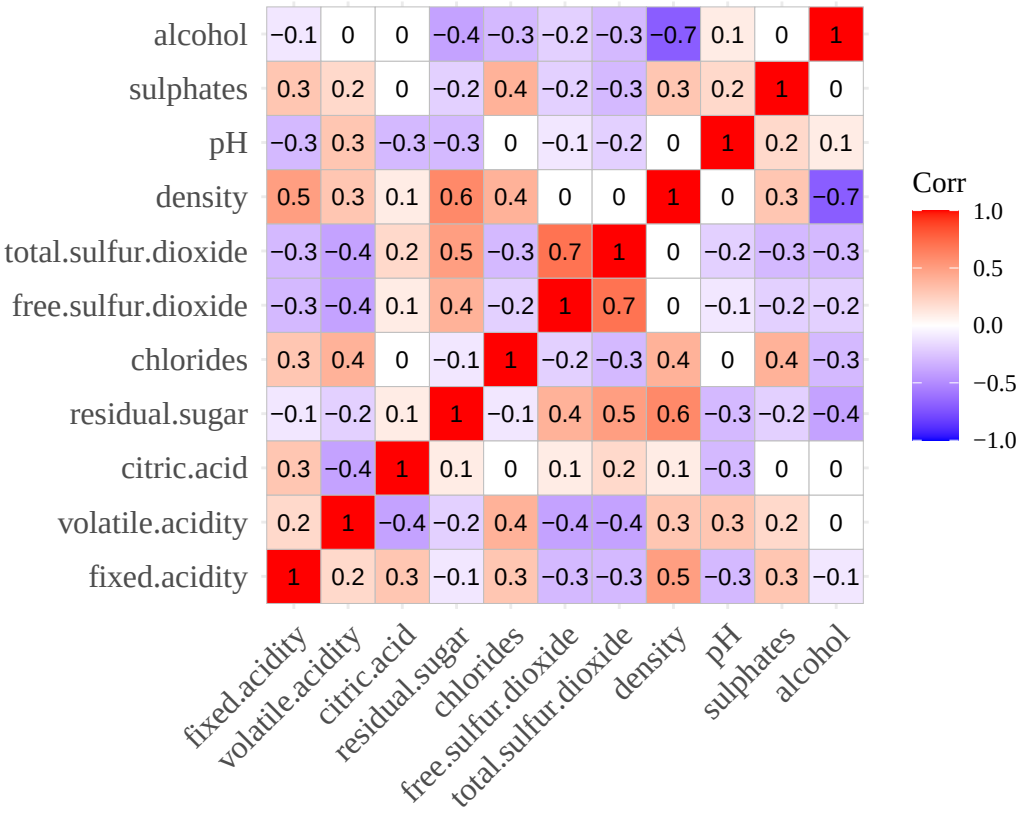
앞서 언급하였듯이, 위에서와 같이 데이터를 처리할 경우 중앙값으로부터의 차이를 기준으로 데이터를 새롭게 구축할 수 있지만, 그 과정에서 그 대소에 대한 정보가 일부 사라진다. 따라서 이를 피하기 위해, 비선형 데이터의 예측에도 용이한 인공신경망 과정에서는 데이터 전처리 없이 진행하였다. 분석 결과 그 정확도는 레드 와인과 화이트 와인 모두에 대해 랜덤포레스트와 유사하게 0.6 가량(위;레드 와인: 0.637, 아래;화이트 와인: 0,568)이 나왔다. 그 RMSE값은 레드 와인과 화이트 와인에 대해 각각 0.697, 0.738로 역시 랜덤포레스트에서와 같았다. 그러나 인공신경망을 적용하기에는 자료의 크기가 비교적 작고, 많은 시간을 요한다는 그 효율성은 낮다고 생각되었다. 따라서, 추후 발전은 랜덤포레스트를 이용해 진행하는 편이 더욱 좋다는 결론을 내릴 수 있었다.

4-2. 기타 연구 가설: 이산화황이 와인 양조에 미치는 영향의 분석

이산화황과 변수 사이 관계에 대한 상관분석

효모의 활성은 포도와 더불어 와인의 맛과 향을 조절하는 매우 중요한 요소임이 틀림없다. 특히, 선행 연구들은 당이 알코올로 분해되는 과정, 그리고 시트르산이 휘발산으로 분해되는 과정이 와인의 발효에 사용되는 효모인 S.cerevisiae과 C.lipolytica 등에 의해 수행된다고 밝히고 있다[7-8]. 이는 곧 처음에 투입된 포도의 수준이 유사하다고 가정하였을 때 당과 알코올, 시트르산과 휘발산 사이에는 음의 상관관계가 있어야 함을 의미한다. 이를 히트맵을 통해 확인한 결과가 아래와 같다(Figure 5).

[Figure 5] Correlation Matrix for Variables

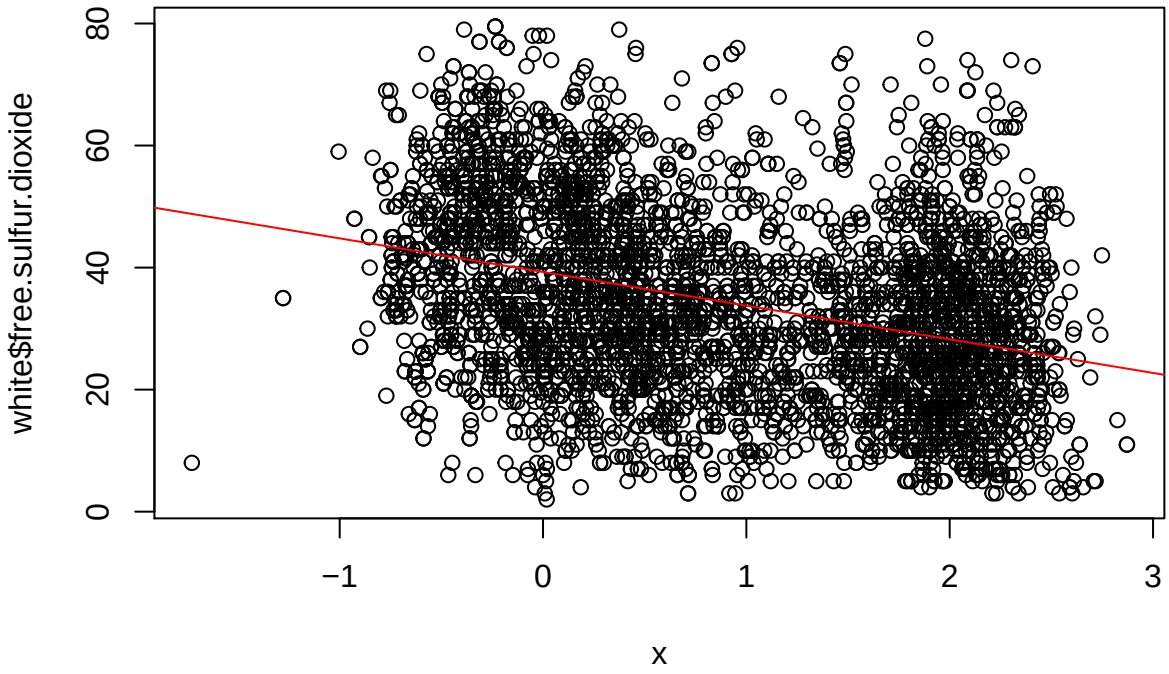


예상한 바와 같이 휘발산과 시트르산 사이의 상관계수는 -0.38로 음이었으며, 이는 상관분석 결과 유의한 값으로 나타났다. 또한 음의 상관관계는 잔당과 알코올에 대해서도 비슷한 상관계수 -0.36을 가지고 나타났으며, 이 역시 유의했다. 즉 시트르산으로부터 휘발산, 당으로부터 알코올을 얻어내는 생화학적 반응은 와인 발효 과정에서 일어남을 확인할 수 있었다. 이러한 경향성은 레드 와인과 화이트 와인에서 일부 다르게 나타났으며, 이에 따라 추후의 분석에서도 둘을 나누어 시행하였다(Figure 3, Supplementary 8).

이때 특이한 것이, 이산화황은 그 형태와 상관 없이 휘발산의 양과 알코올의 양에 대해서는 유의미한 음의 상관관계를 가지나, 시트르산과 잔에 대해서는 모두 유의미한 양의 상관관계를 가지고 있었다. 이산화황이 미생물 생장을 억제함으로써 혐기 상태 반응조 내 미생물 생태를 조절하는 데 효과적임을 고려하였을 때, 이산화황에 의해 이러한 물질들의 농도가 변하는 것은 자연스럽다. 양의 상관관계를 가지는 시트르산과 잔당이 모두 반응물이며 음의 상관관계를 가지는 휘발산과 알코올은 생성물임을 고려해 보았을 때, 효모의 대사 과정은 이산화황에 억제됨을 예상할 수 있었다(Supplementary 9).

ANOVA를 이용한 이산화황의 효과 분석

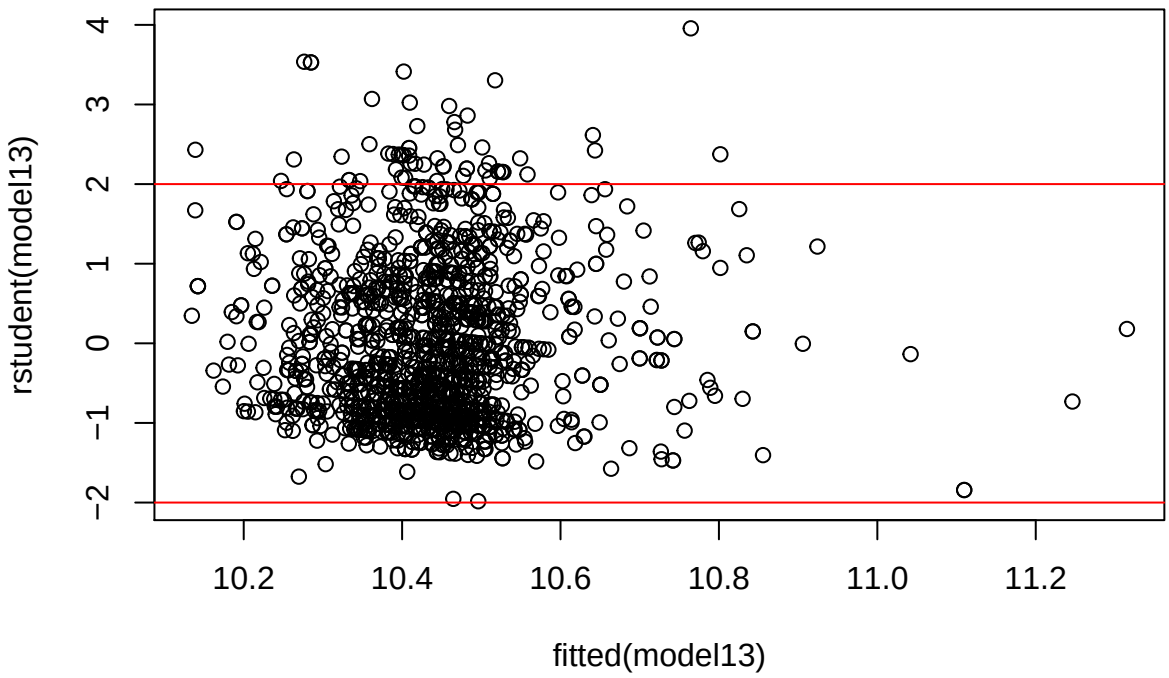
먼저, 자유 이산화황의 양을 기반으로 전체 데이터를 네 개로 나누어 블록화하였다. 그 다음, 각각 자료에 대해 $\log(Product/Reactant)$을 계산한 후 이산화황의 양에 따른 그 변화를 ANOVA와 산점도를 이용하여 분석했다. 이때 $\log(Product/Reactant)$를 계산한 이유는, 일반적인 화학 반응에서 이 값이 평형의 위치를 표현하기 때문이다. 알코올과 잔당에 대해 이를 분석한 결과 자유 이산화황 양에 따른 통계적인 유의미한 차이가 있는 것으로 관찰되었으며, 산점도로부터 이는 음의 상관관계를 가짐을 알 수 있었다. 즉, 이산화황의 농도가 높아질수록 생성물의 양이 줄어든다는 것이다. 이는 앞서 예상한 것처럼 이산화황이 효소의 활성을 억제하여 당이 알코올로 변하는 과정을 억제함을 의미한다(Figure 6; 아래 그림, Supplementary 10).



다중회귀를 통한 이산화황의 효과 분석

레드 와인과 화이트 와인으로 나누어 설명변수를 자유 이산화황과 잔당으로 하고 반응변수를 알코올로 하는 다중회귀분석과 설명변수를 자유 이산화황과 시트르산으로 하고 반응변수를 휘발산으로 하는 다중회귀분석을 시행하고자 했다. 이를 위해, 먼저 잔차도를 보아 다중선형회귀의 가정들을 만족하는지 확인하였다. 그 결과 아래 코드를 이용해 독립성, 선형성, 등분산성, 정규성의 가정들을 어느 정도 만족함을 관찰할 수 있었다(Figure 7; 아래 그림).

```
model13<-lm(red$alcohol~red$residual.sugar+red$free.sulfur.dioxide,red)
plot(fitted(model13),rstudent(model13))
abline(h=c(-2,2),col="red")
```



레드 와인과 화이트 와인으로 나누어 설명변수를 자유 이산화황과 잔당으로 하고 반응변수를 알코올로 하는 다중회귀분석과 설명변수를 자유 이산화황과 시트르산으로 하고 반응변수를 휘발산으로 하는 다중회귀분석을 시행하였다. 화이트 와인에서는 잔당의 회귀계수가 −0.099로 음수이고 이산화황의 회귀계수는 −0.01로 음수였다. 이들은 모두 작은 값이기는 했지만 p값은 매우 작아 통계적으로 유의함이 밝혀졌으며, 이는 이산화황은 효모의 활성을 막음으로써 알코올의 생성을 막는다는 사실과 일치한다. 반면 레드 와인에서는 이산화황의 회귀계수가 −0.008로 음수였으나, 잔당에 대해서는 그 계수가 0.0766 가량으로 양수였다. 또한, 이들은 유의수준 1%에서 의미있었다. 따라서 이산화황은 효모의 활성을 억제하지만, 레드 와인에서는 다른 매커니즘의 존재로 인해 오히려 잔당과 알코올은 양의 상관관계를 맺음을 알 수 있다. 이는 포도 종류의 차이로 인하여 생기는 결과로 보인다.

이와 같은 방법으로 시트르산과 휘발산에 대해서 보았을 때에도 유사한 관계가 관찰되었다. 화이트 와인에 대해 다중회귀분석을 수행한 결과 시트르산의 회귀계수는 −0.11로 음수였다. 레드 와인의 경우에는 이 값이 −0.51로 절댓값이 매우 컸다. 반면, 이들 각각의 경우에서 자유 이산화황에 대한 회귀계수를 보았을 때에는 0.001보다 그 계수가 작았으며, 레드 와인의 경우에는 통계적으로 유의하지 않았다. 또, 결정계수 값도 매우 낮았다. 이는 시트르산과 휘발산에 대해서는 이산화황이 아닌 다른 교락 요인이 더 중요하게 작용함을 암시한다. 즉, 시트르산이 많은 포도일수록 휘발산의 함량은 적은 등의 다른 관계가 음의 상관관계에 대한 주요 원인이지, 이산화황은 큰 영향을 미치지 않는다고 결론내릴 수 있었다.

5. 결론

각 변수들이 적당한 값을 취할수록 높은 품질을 가질 것이라는 일반적인 예측을 근거로 한 와인 품질 가설은 탐색적 자료 분석을 통해 많은 변수에 대해 성립함을 확인할 수 있었다. 그러나 알코올과 같은 몇몇 변수에 대해서는 이것이 적용되지 않았고, 이 처리는 데이터의 대소 관련 정보를 축소한다는 단점이 있었다. 레드 와인과 화이트 와인을 나누어 예측을 수행하였을 때 랜덤 포레스트는 짧은 처리 시간과 높은 정확도, 낮은 RMSE를 보여주었으며 가장 좋은 알고리즘이라 평가되었다. 가지의 개수와 분기의 개수를 달리해 가면서 이를 시행한 결과 가지의 개수가 증가할수록 오차율이 낮아지는 것을 보여주었으며, 결과적으로 약 0.6 정도의 정확도를 얻을

수 있었다. 다만, 매커니즘 증거와 자료 개수의 부족으로 인해 더 높은 정확도를 얻는 데에는 실패하였다. 또한, 알코올이 여기서 중요한 변수로서 작용함이 밝혀졌는데 알코올은 중앙값으로부터의 거리와 품질 간의 관계가 예상한 바와 같지 않았기에 RMSE는 더욱 올라갔을 것으로 추정된다. 중앙값은 단지 추정된 값이었기에, 더욱 많은 부석을 통해 특정 값을 잡은 이후 그 점과의 차이를 이용하여 분석을 시행한다면 더욱 좋은 결과를 얻을 수 있을 것이다. 또한 ANN 역시 실시하였으나 처리 시간이 길고 만족스럽지 못한 결과를 보였고, 더욱의 개선이 필요할 것으로 보인다. 이처럼 와인 품질 예측은 약 60%의 예측 정확도를 보였으며, 이를 더욱 개선할 경우 제안한 바처럼 가정이나 소규모 양조장에서도 손쉽게 품질을 예측할 수 있을 것이라 기대된다.

기타 연구 가설에서는 이산화황의 양에 따른 다른 변수들의 변화를 효모의 활성 관점에서 분석하였다. 특히, 효모의 생화학 경로에 포함되어 있는 당에서 알코올로의 발효와 시트르산에서의 휘발산 변화를 집중적으로 분석하였다. ANOVA와 상관분석, 다중회귀 등을 통해 이산화황의 영향을 분석한 결과, 당과 알코올에 대해서는 그 비율에 이산화황이 큰 영향을 미치는 것으로 확인되었고, 이는 우리의 예상과 같이 효모의 활성을 억제함으로써 알코올을 줄이는 방향으로 작용했다. 반면, 시트르산과 휘발산의 관계는 비교적 약하게 나타났고 통계적으로 그리 유의하지 않았다. 그러나 이산화황이 알코올과 당의 비율에 어떤 영향을 미치는지를 확인할 수 있었기에, 이는 와인의 바디감과 향을 자신이 원하는 대로 조절하는 과정에 큰 도움을 줄 것으로 보인다.

6. 각 조원이 보고서 및 발표에 기여한 바

권이태 : 보고서 R markdown 작성, 자료 정리 및 데이터 시각화, 와인 품질 가설의 데이터 분석 도움(KNN, ANN)

김예원 : 변수에 대한 탐색적 자료 분석, 와인 품질 가설 분석 코드 작성 및 분석(의사결정나무, Random Forest), 예측 모델 신뢰성 분석, 와인 품질 예측

김태희 : 기타 연구 가설 관련 탐색적 자료 분석, 기타 연구 가설 코드 작성 및 분석, 기타 연구 가설 검증, 이산화황 농도 결정을 통한 와인 양조의 개인화 가능성 분석

7. 참고 문헌

1. 이마트24 (2021). 이마트24 와인, 올해도 3배 성장 중! 3월에도 와인 커뮤니티 달군다!. https://www.shinsegaegroupnewsroom.com/54624/ (검색일: 2020.06.11)
2. 신세계그룹 (2021). 이마트, 역대 최대 규모 와인장터 개최. https://www.shinsegaegroupnewsroom.com/60282/ (검색일: 2020.06.11)
3. 이마트 (2021). 2021 1분기 보고서. http://www.emartcompany.com/ko/investor/irfinance_list.do (검색일: 2020.06.11)
4. Andrew L. Waterhouse et al (2016). Understanding Wine Chemistry. Wiley.
5. 기획재정부 (2019). 2018년도 세법 후속 시행령 개정안 수정사항. https://www.moef.go.kr/nw/nes/detailNesDtaView.do?searchBbsId1=MOSFBBS_000000000028&searchNttId1=MOSF_000000000026762&menuNo=4010100 (검색일: 2020.06.11)
6. Presque Isle Wine Cellars. Use and Measurement of Sulfur Dioxide in Wine. https://www.piwine.com/media/home-wine-making-basics/using_sulfur_dioxide.pdf (검색일: 2020.06.11)
7. Guy N. B. (1951). Basic Effects of Sulfur Dioxide on Yeast Growth. Am J Enol Vitic. 2. 43-53.
8. Farouk A.H. et al. (1981). Fermentative Production of Citric Acid by Yeasts. Agricultural Wastes. 3(1). 21-33.
9. 엄경자 (2006). [엄경자의 와인이야기] 타닌 · 산 · 알콜 · 당 … 성분의 조화가 맛 결정. https://www.hankyung.com/society/article/2006102051881 (검색일: 2020.06.11)

8. 부록

이 보고서의 사용 코드 및 보조 자료는 재현 가능성의 보존을 위해 GitHub에 r markdown 파일 형식으로도 게시하였다. 접근 가능한 주소는 아래와 같다:

https://github.com/Yitae-Kwon/2021-Spring-StatLab-Project