빅데이터 분석 경진 대회 보고서

와인 입문자를 위한 필터링 이후 와인 추천 시스템

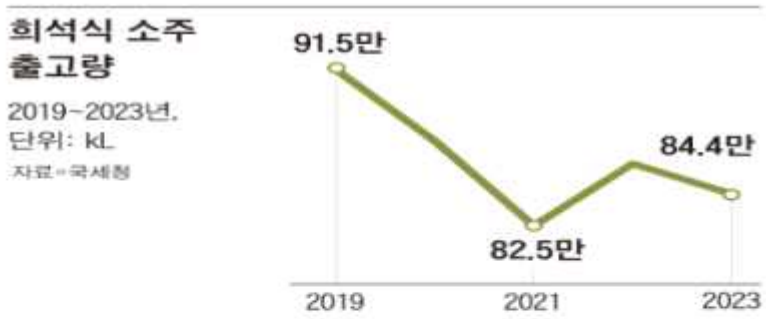
팀명 : 215 (김시현, 배현준, 오정우)

목차

서론 1. 데이터 분석 배경과 필요성

2020년부터 코로나19로 인해 사회적 거리두기와 방역 지침이 강화되면서, 전통적인 술자리와 회식 문화가 축소되었다. 이러한 변화는 소주와 같은 기존 주류의 소비 감소와 함께, 와인을 포함한 다양한 와인, 위스키 등 주류 소비 증가라는 새로운 트렌드를 불러왔다. 특히 와인은 고급스러운 이미지와 다채로운 특성으로 인해 소비자들의 관심을 끌고 있다.

그러나 와인을 처음 접하는 소비자들은 방대한 종류와 복잡한 특성에 대한 지식 부족으로 인해 자신에게 적합한 와인을 선택하는 데 어려움을 겪고 있다. 이는 와인 시장의 접근성을 낮추는 요인으로 작용하며, 소비자들의 구매 만족도를 저하시키는 문제로 이어지고 있다. 이러한 문제를 해결하기 위해 데이터 분석을 통해 소비자의 조건에 맞춰 와인 추천 시스템을 제공할 필요성이 있다고 판단되어 분석을 하게 되었다.



그래픽=백형선

국세청에 따르면, 국내 전체 주류 출고량은 코로나 이전이었던 2019년 384만1000kL에서 작년 361만9000kL로 약 6% 줄었다. 20세 이상 국민의 1인당 연간 알코올 소비량도 2015년 9.813L에서 2021년 8.071L로 18% 감소했다. 올해 초부터 지난 7월까지의 주세 수입 또한 작년 같은 기간보다 6.6% 줄어들었다.

특히 한때 '국민 술'로 불렸던 희석식 소주를 외면하는 현상이 두드러진다. 과거 주머니 가벼운 대학생들이 싼값으로 금방 취할 수 있었던 소주를 즐겨 찾았고, MT나 학과 행사 등에서 '사발식' 등으로 소주를 폭음했던 문화가 코로나를 지나면서 '멸종' 수준이라는 것이다. 코로나 이전 2019년 91만5596kL였던 희석식 소주 출고량은 작년 84만4250kL로 약 8% 감소했다. 주류 업계에선 "젊은 대학생과 직장인의 소주 소비가 줄어든 탓"이라고 분석한다.

《 출처 : 김병권 기자, 조선 일보(2024.09.21), "크~~" 가 사라졌다... 소주 안 마시는 2030, https://www.chosun.com/national/national_general/2024/09/21/EQGAX3MCIFHMNM66INJ6QL37FY/ 》

서론 2. 분석 목적 및 목표

본 분석의 목적은 와인 평점 데이터를 활용하여 소비자에게 적합한 와인을 추천할 수 있는 데이터 기반 모델을 개발하는 데 있다. 이를 통해 소비자들에게 맞춤형 정보를 제공하여 와인 선택 과정을 보다 쉽게 하고, 와인 시장의 접근성을 향상시키는 것을 목표로 한다.

1. k-means clustering을 이용한 모델링
와인 데이터를 기반으로 k-means clustering을 적용하여 와인의 특성을 그룹화하고, 소비자의 선호에 맞는 군집을 타겟으로 설정하였다.

2. 소비자의 조건과 모델링을 이용한 와인 추천 시스템
소비자가 제공하는 조건(가격, 평점 등)을 바탕으로 필터링한 뒤, 필터링된 데이터 중에서 학습된 모델을 기반으로 추천하는 시스템을 만든다.

3. 데이터 기반 추천을 통해 와인 소비의 효율성과 만족도 증대
데이터 기반의 추천 시스템은 소비자들에게 최적화된 와인을 제안하여 구매 결정을 돕고, 와인 선택 과정의 편리성을 증대시킨다. 이를 통해 와인 소비의 효율성을 높이고, 궁극적으로 소비자 만족도를 극대화하는 것을 목표로 한다.

본론 1. 데이터 설명 - 데이터 출처 및 수집

데이터 출처 :

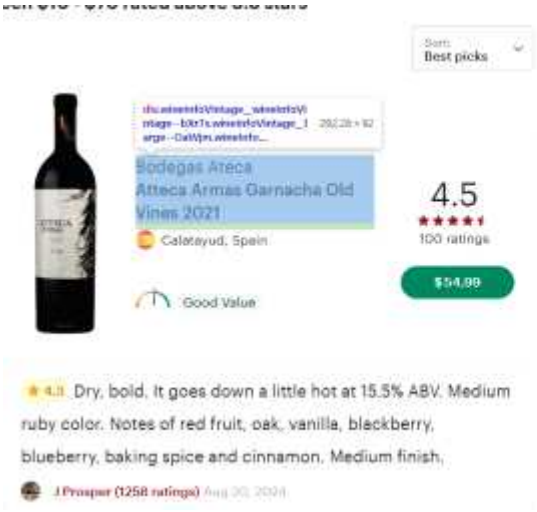
Vivino : 2011년에 시작되어 현재 전 세계적으로 수백만 명의 사용자들이 이용하는 플랫폼이며 주요 기능으로는

- 와인 리뷰 및 평가
- 와인 구매
- 와인 추천
- 와인 검색 및 정보
- 와인 커뮤니티

Vivino는 와인 초보자부터 전문가까지 다양한 사람들이 와인을 더 쉽게 탐색하고, 구매할 수 있는 와인 전문 사이트이다.

데이터 수집 :

이번 분석에서는 Python의 Selenium 패키지를 활용하여 웹 자동화를 수행했다. Selenium의 WebDriver를 사용하여 지정한 HTML 요소에서 필요한 정보를 추출하고, 이를 기반으로 데이터를 수집하였다. 해당 웹 페이지의 특정 텍스트를 검색하고, 동적 컨텐츠도 자동으로 로딩되도록 처리하여 정확한 데이터를 확보했다.



본론 1. 데이터 설명 -데이터 주요 변수 및 관측치

1. 관측치 설명

와인 종류(wine)	관측치(observation)
red_wine	2028
Rose_wine	52
sparkling_wine	164
white_wine	826
fortified_wine	74
dessert_wine	144

와인 종류를 따로 수집한 이유는 와인 별 추천 시스템을 만들고자 하였다. 관측치가 적은 와인들도 있지만 와인의 특성을 제대로 알 수 없는 상황이라 생각했고, 2~3개의 와인을 '기타 와인'으로 묶는 것은 하고자 하는 분석의 방향성과 어떤 영향이 있을지 알 수 없기에 하지 않았다.

2. 변수 설명

변수(feature) 개수 : 6개

리뷰 평점	Rating	수치형
리뷰 수	Rating.Count	수치형
가격	Price	수치형
제조 국가	Country	범주형
제조사(혹은 제조인)	Brand	범주형
빈티지(포도 생산 년도)	Vintage	수치형

	Rating	Rating.Count	Price	Country	Brand	Vintage
1	4.9	66	6110491	France	Château Pétrus	1960
2	4.8	1540	3192869	France	Château Haut-Brion	1989
3	4.8	1446	2715774	France	Château Latour	1982
4	4.8	1246	6752735	France	Château Pétrus	1990
5	4.8	103	6679336	France	Château Pétrus	2018
6	4.8	69	1461864	France	E. Guigal	1990
7	4.8	39	470974	France	Jean-Michel Gerin	1999
8	4.8	30	3009371	France	Henri Bonneau	1990
9	4.7	35071	8434798	France	Château Pétrus	1951

본론 2. 데이터 구조 이해 - 수치화 및 시각화

(1) Summary() 함수 이용하여 수치화

- red wine

```
     Rating           Name           Rating.Count          Price            Country              Brand              Wine.Name            Vintage         
 Min.   :3.600   Length:2028        Min.   :   25.0   Min.   :   9542   Length:2028        Length:2028        Length:2028        Length:2028       
 1st Qu.:3.900   Class :character   1st Qu.:   82.0   1st Qu.:  51367   Class :character   Class :character   Class :character   Class :character  
 Median :4.100   Mode  :character   Median :  251.5   Median : 102018   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :4.113                      Mean   : 1838.0   Mean   : 287521                                                                            
 3rd Qu.:4.300                      3rd Qu.:  912.2   3rd Qu.: 214076                                                                            
 Max.   :4.900                      Max.   :81950.0   Max.   :8683137                                                                            
```

- white wine

```
     Rating           Name           Rating.Count          Price            Country              Brand              Wine.Name            Vintage         
 Min.   :3.000   Length:826         Min.   :   25.0   Min.   :  10203   Length:826         Length:826         Length:826         Length:826        
 1st Qu.:3.700   Class :character   1st Qu.:   84.0   1st Qu.:  27742   Class :character   Class :character   Class :character   Class :character  
 Median :3.900   Mode  :character   Median :  195.0   Median :  37883   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :3.856                      Mean   :  973.8   Mean   :  90813                                                                            
 3rd Qu.:4.100                      3rd Qu.:  482.8   3rd Qu.:  58730                                                                            
 Max.   :5.000                      Max.   :29980.0   Max.   :7034104                                                                            
```

- Rose wine

```
     Rating          Name           Rating.Count         Price            Country            Brand              Wine.Name           Vintage         
 Min.   :3.000   Length:52          Min.   :  25.0   Min.   : 10203   Length:52          Length:52          Length:52          Length:52         
 1st Qu.:3.600   Class :character   1st Qu.:  45.5   1st Qu.: 22393   Class :character   Class :character   Class :character   Class :character  
 Median :3.700   Mode  :character   Median : 132.0   Median : 28192   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :3.742                      Mean   : 627.6   Mean   : 29282                                                                            
 3rd Qu.:4.000                      3rd Qu.: 351.5   3rd Qu.: 32560                                                                            
 Max.   :4.800                      Max.   :10636.0   Max.   :104875                                                                           
```

- fortified wine

```
     Rating          Name           Rating.Count         Price            Country            Brand              Wine.Name           Vintage         
 Min.   :3.200   Length:74          Min.   :  28.0   Min.   : 23341   Length:74          Length:74          Length:74          Length:74         
 1st Qu.:3.725   Class :character   1st Qu.: 150.0   1st Qu.: 37049   Class :character   Class :character   Class :character   Class :character  
 Median :3.900   Mode  :character   Median : 556.5   Median : 57111   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :3.977                      Mean   :1485.4   Mean   : 188282                                                                           
 3rd Qu.:4.200                      3rd Qu.:2087.0   3rd Qu.: 176470                                                                           
 Max.   :4.800                      Max.   :13138.0   Max.   :2232557                                                                          
```

- sparkling wine

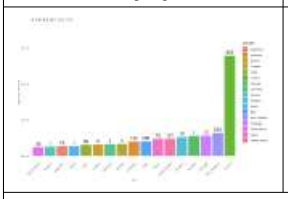
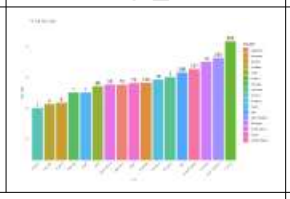
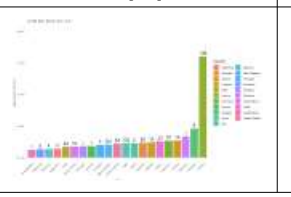
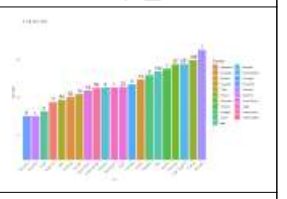
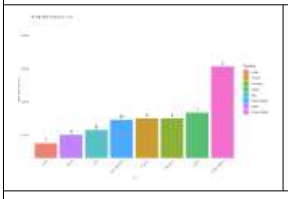
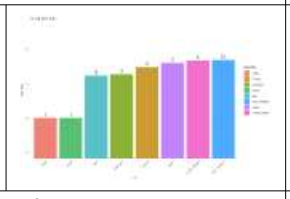
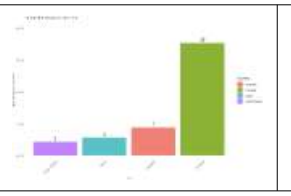
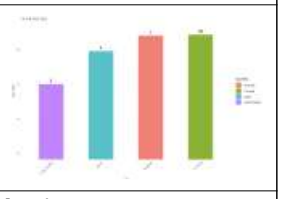
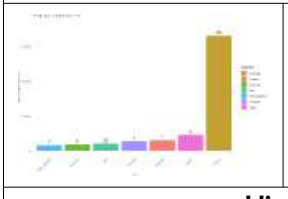
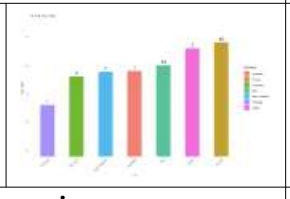
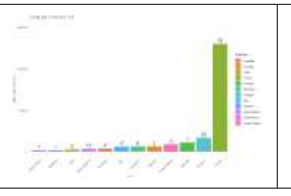
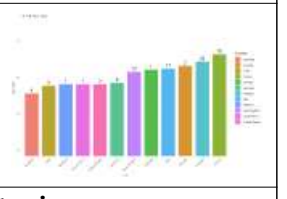
```
     Rating          Name           Rating.Count         Price            Country            Brand              Wine.Name           Vintage         
 Min.   :3.300   Length:164         Min.   :  26.0   Min.   :  7810   Length:164         Length:164         Length:164         Length:164        
 1st Qu.:3.900   Class :character   1st Qu.: 162.8   1st Qu.: 34408   Class :character   Class :character   Class :character   Class :character  
 Median :4.100   Mode  :character   Median : 625.5   Median : 65564   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :4.076                      Mean   :6793.4   Mean   : 274783                                                                           
 3rd Qu.:4.300                      3rd Qu.:3418.0   3rd Qu.: 170734                                                                           
 Max.   :4.800                      Max.   :135578.0   Max.   :9682586                                                                         
```

- dessert wine

```
     Rating          Name           Rating.Count         Price            Country            Brand              Wine.Name           Vintage         
 Min.   :3.600   Length:144         Min.   :   25.00   Min.   :  17979   Length:144         Length:144         Length:144         Length:144        
 1st Qu.:4.000   Class :character   1st Qu.:   57.75   1st Qu.:  54996   Class :character   Class :character   Class :character   Class :character  
 Median :4.200   Mode  :character   Median :  168.50   Median : 104426   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :4.185                      Mean   : 4015.88   Mean   : 677506                                                                           
 3rd Qu.:4.325                      3rd Qu.: 1991.50   3rd Qu.: 503086                                                                           
 Max.   :4.800                      Max.   :50034.00   Max.   :13028375                                                                          
```

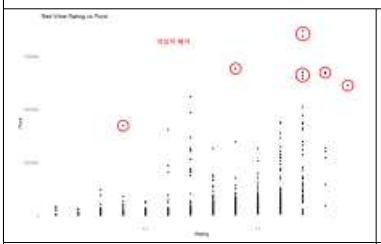
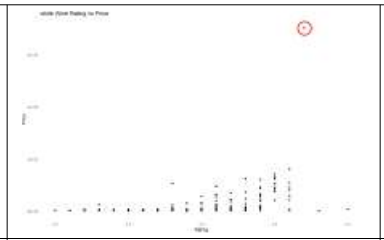
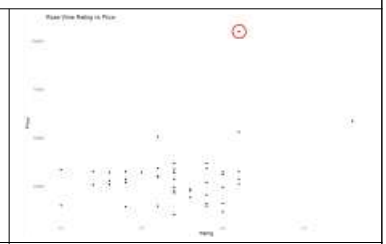
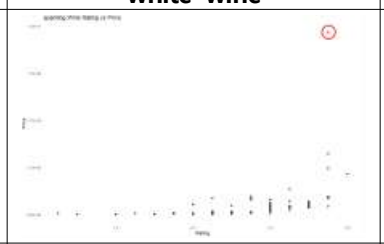
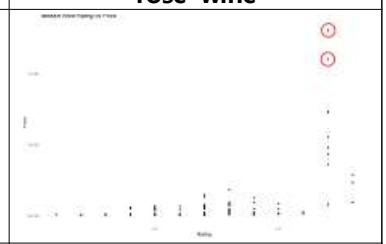
(2) 데이터 시각화

(뒤에 더 큰 이미지가 있으니 참고할 수 있습니다.)

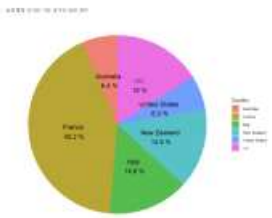
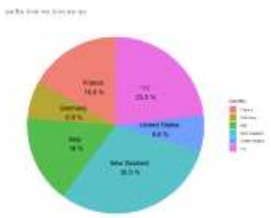
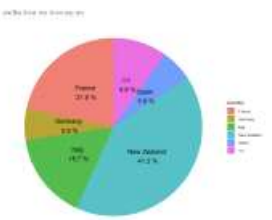
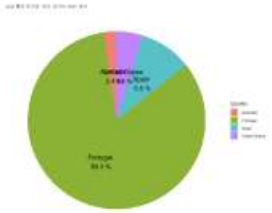
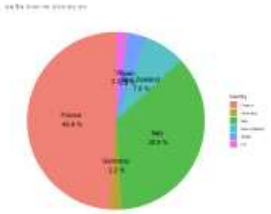
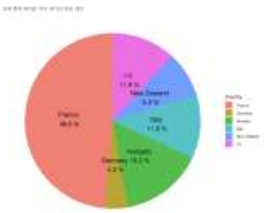
와인 종류 별 평균 가격과 평점			
가격	평점	가격	평점
			
red wine		white wine	
			
Rose wine		fortified wine	
			
sparkling wine		dessert wine	

가격과 평점 역시 프랑스에서 생산된 와인이 높은 것으로 나타난다.

※ 와인의 본고장인 프랑스산 와인이 대중들에게 인기가 많고 평가가 좋다고 할 수 있다.

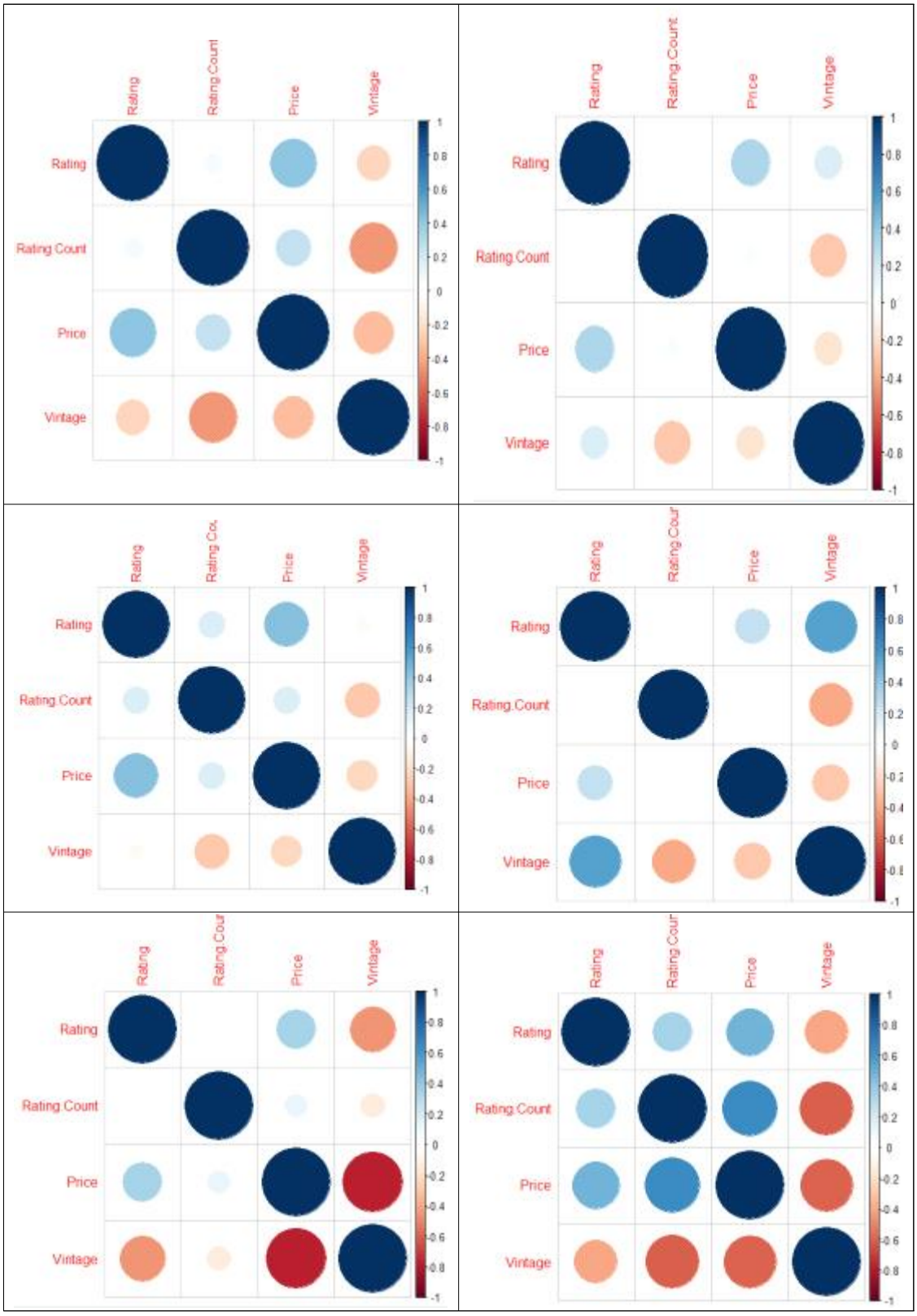
평점과 가격에 대한 산점도		
		
red wine	white wine	rose wine
		
fortified wine	sparkling wine	dessert wine

가격이 높다고 무조건 높은 평점을 얻는 것은 아니였지만 같은 평점이여도 너무 높은 가격이 있으면 분석에 좋지 않은 영향이 있을 수 있다고 생각하여 데이터 전처리 과정에서 이상치로 다룰 예정이다.

와인 종류 별 생산 국가 비중		
		
red wine	**white wine**	**Rose wine**
		
fortified wine	**sparkling wine**	**dessert wine**

6가지 와인 중 레드 와인(Red Wine), 스파클링 와인(Sparkling Wine), 디저트 와인(Dessert Wine)의 최대 생산 국가는 프랑스입니다. 또한, 포티파이드 와인(Fortified Wine)을 제외한 나머지 2가지 와인에서도 프랑스가 높은 비중을 차지하고 있습니다.

3) 수치형 변 간의 연관성 파악



상관행렬과 VIF분석을 통하여 확인한 결과, 일부 변수들 간에 높은 상관관계가 있다는 것이 관찰되었다. 이는 다중공선성의 가능성이 있을 수 있으므로, 분석 및 모델링 과정에서 문제가 될 수 있다. 특히, 다중공선성은 변수 간 상호 연관성을 증폭시켜 회귀계수의 불안정성을 초래하거나, 모델이 복잡해지며, 해석력이 저하될 수 있으므로 후에 데이터 전처리 과정에서 수치형 변수들을 정규화, 표준화하는 방법을 고려할 것이다.

추가로 VIF 분석도 모두 진행하였을 때 상관행렬과 같이 특정 변수들에서 연관성이 있다고 할 수 있는 값이 나왔기 때문에 연관성이 있다고 판단하였다. (사진이 너무 많아 생략하였음.)

본론3. 데이터 전처리

(1) 결측값 처리

와인 정보를 가져오는 과정에서 빈티지(Vintage : 와인에 사용된 포도가 수확된 년도)에 대한 정보가 없는 경우가 있었기에 N/A로 처리하고 삭제하였다.
빈티지를 다른 값으로 대체하는 선택을 하기에는 좋은 품질의 포도가 생산된 년도가 특정한 패턴이 있다고 보기 힘들기 때문에 다른 값으로 대체하는 것이 힘들다고 생각하였다.

(2) 이상치 처리

데이터 시각화 과정에서 산점도를 통해서 같은 평점이여도 너무 높은 가격을 가진 와인들이 있었고, 그 와인들이 충분히 분석을 함에 있어서 안좋은 영향, 특히, 군집화를 할 예정인데 이상치는 그룹으로 묶기 힘들다고 판단하여 제거하였다.

(3) 범주형 변수 처리

1. 제조 국가에 대한 인코딩

Country 변수가 France, Italy 등처럼 텍스트로 되어있고, 후에 분석을 할 때 범주형 데이터를 수치형 데이터로 변환하여야 분석을 진행할 수 있어서 수치형 데이터로 변환하였다.
처음에는 빈도 인코딩을 통하여 변환을 하였지만 France → 855, Chile → 96 처럼 변환된 값들끼리의 차이가 심해 빈도가 높은 순서대로 순서 인코딩을 하였다.

Country	Frequency	Crank
France	855	1
Italy	296	2
New Zealand	293	3
Australia	130	4
United States	127	5
Chile	96	6
Spain	79	7
Argentina	75	8
South Africa	28	9
Hungary	18	10
Portugal	12	11
Austria	6	12
Canada	4	13
Germany	3	14
Georgia	1	15
Greece	1	16
Israel	1	17

2. 제조사에 대한 인코딩

제조사에 대해서 라벨 인코딩을 진행하였지만 종류가 너무 많아 분석을 제대로 진행할 수 없다고 판단하였고,
Sum(와인$Rating * 와인$Rating.Count) / sum(와인$Rating.Count)로 평균 평점을 구하여 높은 순서대로 순위 인코딩을 하였다.
Ex) red_wine은 794개의 제조사(혹은 제조인)이 있었다.
1~100 = 1
101~200 = 2
.
.
.
701~794 = 8

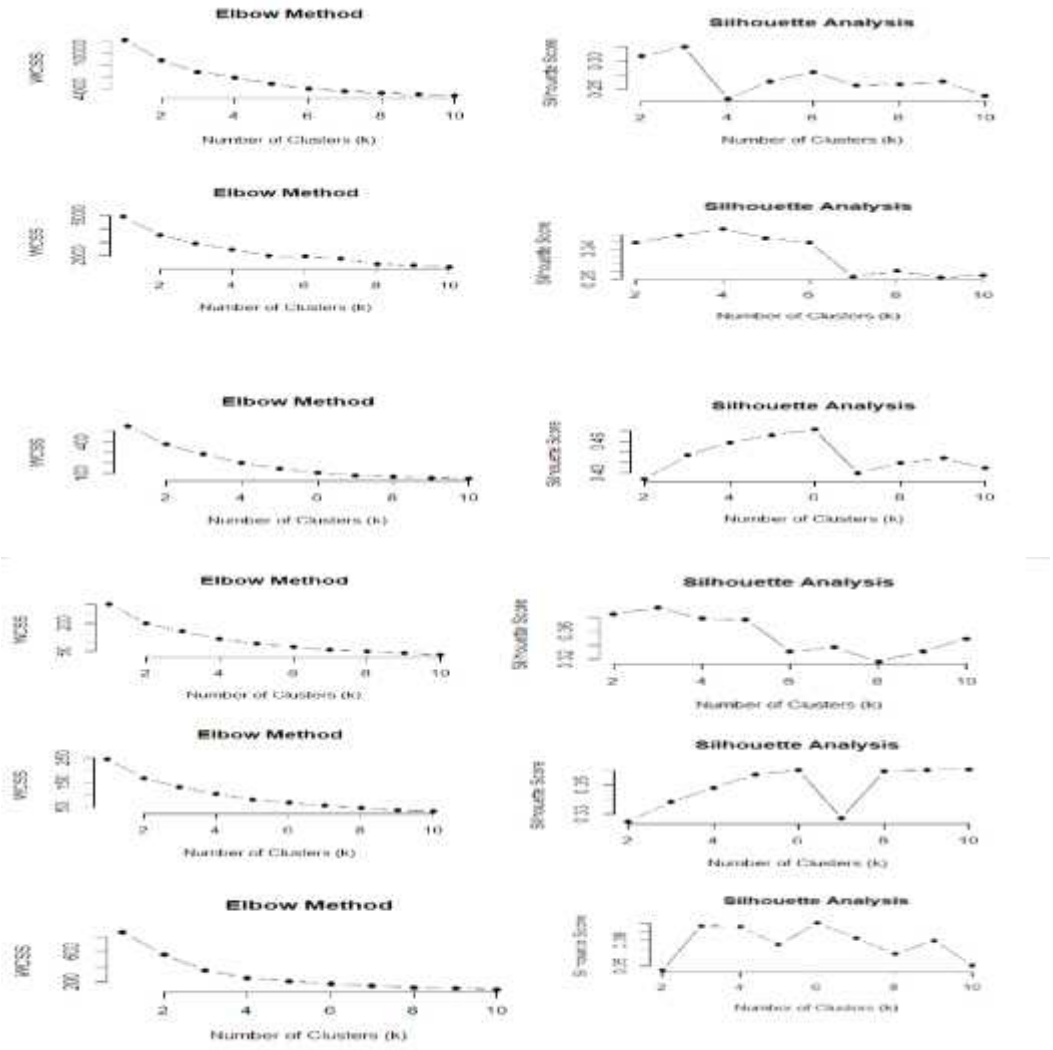
3. 수치형 데이터 표준화

데이터 분석을 하는 과정에서 비지도학습을 선택하였고 k - means clustering 기법을 사용하였고, 이 방법은 그룹화가 중요한데 범위가 넓으면 힘들다. 수치형 데이터들 리뷰 평점과 리뷰 수와 가격에 대해서 범위가 너무 넓다고 생각되어 표준화를 진행하였다.
추가로 앞에서 변수 간의 연관성이 있다고 판단되었기 때문에 수치형 데이터에 대해서 표준화를 진행하였다. 하지만 빈티지 데이터는 표준화하지 않았는데 빈티지(포도 생산 년도) 라는 정보를 표준화했을 때 모델에서 충분히 설명될 수 없다고 생각을 하였기에 표준화를 하지 않았다.

Rating	Rating.Count	Price
1.0000000	5.004577e-04	0.703393345
0.9230769	1.849252e-02	0.367013562
0.9230769	1.734513e-02	0.312008112
0.9230769	1.490388e-02	0.777439228
0.9230769	9.520903e-04	0.768976878

본론 4. 데이터 분석

현재 데이터에는 '와인 추천 점수' 같은 특정한 종속 변수가 없으므로 비지도 학습 기반으로 모델링을 실시하였으며, 데이터 내의 잠재적인 패턴과 그룹 구조를 탐색하기 위해 k-means 클러스터링 기법을 적용하였고, 엘보드 메소드와 실루엣 방법을 이용하여 최적의 k를 찾았다.



Ex. Red와인 k - means clustering(k = 3)

Vintage	Rating	Rating.Count	Price	rank	Crank	cluster
1960	1.0000000	5.004577e-04	0.703393345	1	1	1
1989	0.9230769	1.849252e-02	0.367013562	1	1	3
1982	0.9230769	1.734513e-02	0.312008112	1	1	3
1990	0.9230769	1.490388e-02	0.777439228	1	1	1
2018	0.9230769	9.520903e-04	0.768976878	1	1	3
1990	0.9230769	5.370766e-04	0.167441759	1	1	3

1) Cross Validation

Red, White 와인은 관측치가 충분하기 때문에 train = 0.8 / test = 0.2 로 분할한 이후 k - fold cv (k = 5)를 이용하여 교차 검증을 실시한다.
Sparkling, Rose, Fortified, Dessert 와인은 관측치가 적기 때문에 LOOCV(Leave-One-Out Cross Validation 를 이용하여 관측치만큼 교차 검증을 실시함

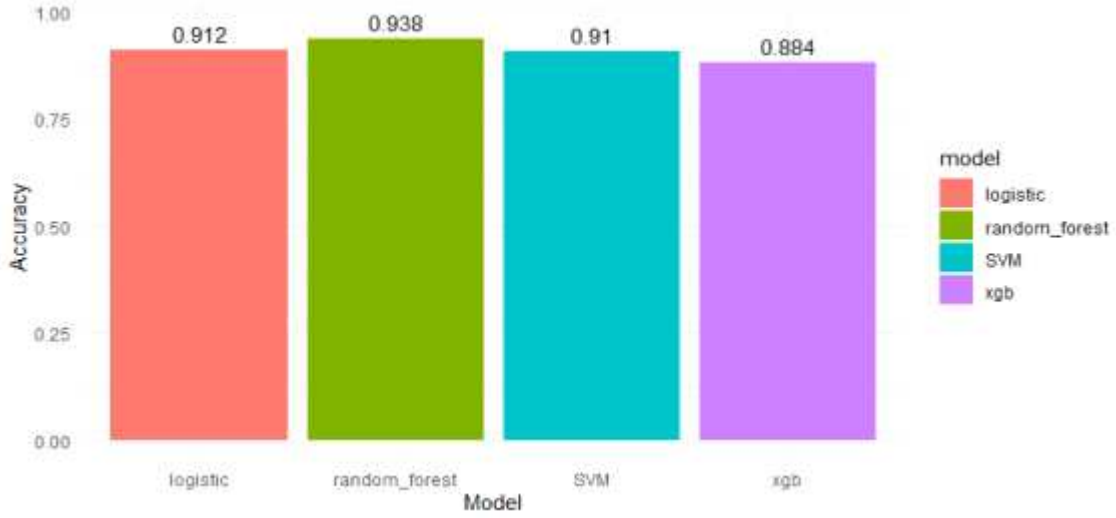
2) 모델링

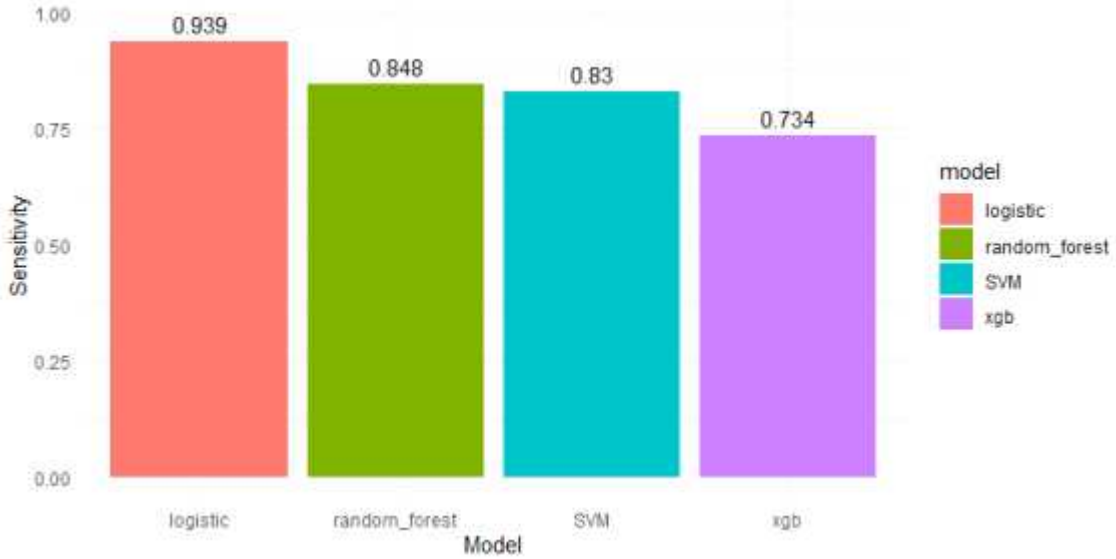
로지스틱 회귀 (Logistic Regression, LR)
단순하면서도 해석 가능성이 높아, 클러스터와 독립 변수 간의 관계를 파악하는 데 유리 클러스터 간 차이가 선형적으로 구분 가능하다면, 로지스틱 회귀는 효과적인 분석 도구이다.
랜덤 포레스트 (Random Forest, RF)
비선형 데이터에도 강하며, 변수 중요도(feature importance)를 제공해 클러스터에 기여하는 주요 변수를 파악할 수 있고, 복잡한 데이터 구조에서 클러스터와 독립 변수 간의 관계를 비선형적으로 모델링할 수 있다.
XGBoost (XGB)
고성능 부스팅 알고리즘으로, 과적합 방지와 속도가 뛰어나다. 데이터가 크거나 복잡할 때, 클러스터 분류 문제에 강력한 성능을 발휘한다.
서포트 벡터 머신 (SVM)
높은 차원의 데이터를 효과적으로 처리하며, 비선형 데이터를 분리하기 위해 커널 함수를 활용하고, 클러스터 간 경계가 선형적으로 분리되지 않는 경우 강력한 성능을 발휘한다.

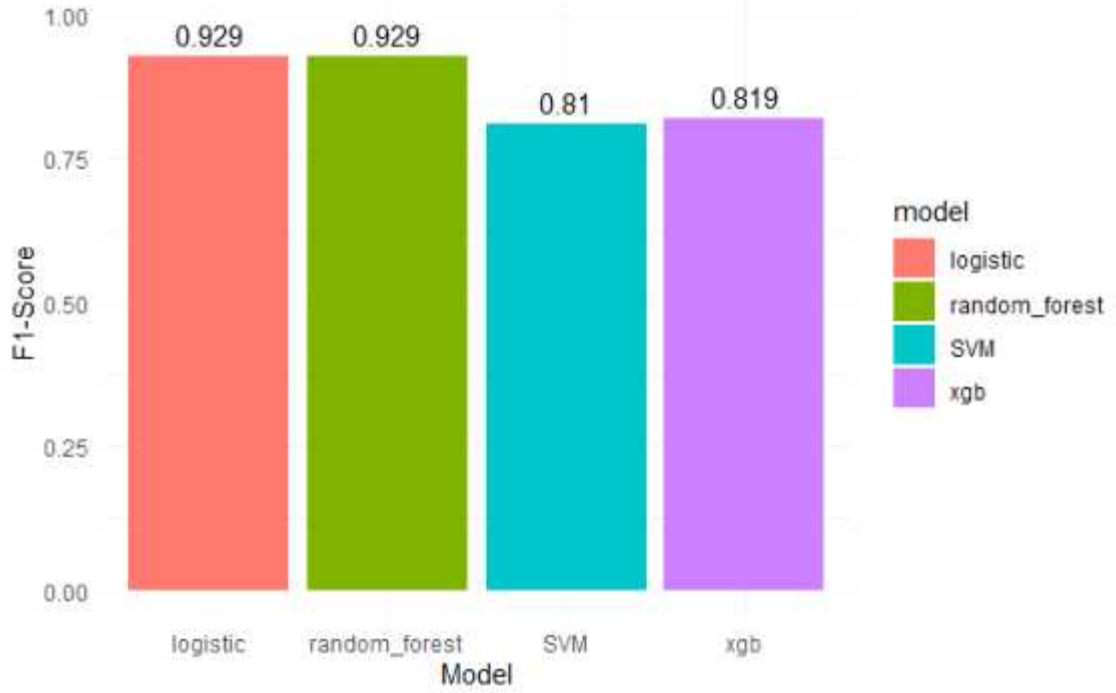
3) 성능 평가

K - means clustering을 사용하였기 때문에 정확도, 정밀도, F1-Score를 중심으로 모델을 평가하였다.
6종류 와인에 분석을 진행하여 나온 정확도, 정밀도, F1-Score 값 6개에 평균을 내었을 때

구분	Logistic Regression	Random Forest	XGB	SVM
Accuracy	0.912	0.938	0.91	0.884
Sensitivity	0.939	0.848	0.83	0.734
F1-Score	0.929	0.929	0.81	0.819


1.00
0.75
0.50
0.25
0.00
Accuracy
0.912
0.938
0.91
0.884
logistic
random_forest
SVM
xgb
Model
model
logistic
random_forest
SVM
xgb


1.00
0.75
0.50
0.25
0.00
Sensitivity
0.939
0.848
0.83
0.734
logistic
random_forest
SVM
xgb
Model
model
logistic
random_forest
SVM
xgb


1.00
0.75
0.50
0.25
0.00
F1-Score
0.929
0.929
0.81
0.819
logistic
random_forest
SVM
xgb
Model
model
logistic
random_forest
SVM
xgb

결론 1. 분석 결과 및 와인 추천 시스템

로지스틱 회귀분석은 다양한 모델 중에서 가장 우수한 성능을 보였기 때문에 최종적으로 모델링에 선택되었다. 로지스틱 회귀는 변수 간의 관계를 명확하게 해석할 수 있는 장점이 있으며, 이로 인해 예측 결과에 대한 해석을 쉽게 할 수 있다는 장점이 있다. 분석의 목표에 맞춰, 유저가 필요로 하는 정보를 효율적으로 처리하고 제공하기 위해, 주요 프로세스를 함수 형태로 구현하였다.

와인 추천 시스템 예시

1) 아무 조건이 없는 와인 추천

모델링 이후 추천된 레드 와인

Vintage	Rating	Rating.Count	Price	rank	Crank	cluster
1960	1.0000000	0.0005004577	0.7033933	1	1	3
1990	0.9230769	0.0149038755	0.7774392	1	1	3
2018	0.9230769	0.0009520903	0.7689769	1	1	3
1951	0.8461538	0.4277815075	0.9713684	1	1	3
1963	0.8461538	0.4277815075	0.7774392	1	1	3
1968	0.8461538	0.4277815075	0.7428840	1	1	3

원본 레드 와인 데이터로 변환

	Rating	Rating.Count	Price	Country	Brand	Wine.Name	Vintage
1	4.9	66	6110491	France	Château Pétrus	Pomerol 1960	1960
4	4.8	1246	6752735	France	Château Pétrus	Pomerol 1990	1990
5	4.8	103	6679336	France	Château Pétrus	Pomerol 2018	2018
9	4.7	35071	8434798	France	Château Pétrus	Pomerol 1951	1951
10	4.7	35071	6752735	France	Château Pétrus	Pomerol 1963	1963
11	4.7	35071	6453017	France	Château Pétrus	Pomerol 1968	1968

2) 필터링 이후 와인 추천, 제조 국가가 이탈리아(인코딩 = 2)인 와인 중에서 추천

필터링, 모델링 이후 추천된 레드 와인

	Vintage	Rating	Rating.Count	Price	rank	Crank	cluster
1187	1955	0.3076923	5.772719e-01	0.0324668145	3	2	3
457	2018	0.5384615	2.221544e-02	0.0118929925	6	2	2
533	2017	0.5384615	2.477876e-03	0.0057036327	5	2	2
551	2000	0.5384615	1.696674e-03	0.0166665610	6	2	2
666	2000	0.4615385	3.748551e-02	0.0118548307	6	2	2
673	2006	0.4615385	2.652426e-02	0.0112379008	5	2	2

원본 레드 와인 데이터로 변환

	Rating	Rating.Count	Price	Country	Brand	Wine.Name	Vintage
1187	4.0	47318	291146	Italy	Antinori	Vila Antinori Chianti Classico Riserva 1955	1955
457	4.3	1845	112697	Italy	Donnafugata	Mille E Una Notte 2018	2018
533	4.3	228	59013	Italy	Di Majo Norante	Don Luigi Riserva 2017	2017
551	4.3	164	154101	Italy	Ruffino	Romitorio di Santedame Toscana 2000	2000
666	4.2	3096	112366	Italy	Aldo Rainoldi	Sfursat di Valtellina 2000	2000
673	4.2	2198	107015	Italy	Cantina Zaccagnini	San Clemente Montepulciano d'Abruzzo Terre di Casauria Ri...	2006

다른 와인들에 대한 추천

빈티지가 1960~1980년 사이인 스파클링 와인 추천

	Rating	Rating.Count	Price	Country	Brand	Wine.Name	Vintage
15	4.5	23791	484435	France	Perrier-Jouët	Belle Epoque Brut Champagne 1975	1975
16	4.5	19746	1357887	France	Ruinart	Dom Ruinart Blanc de Blancs Brut Champagne 1961	1961
34	4.3	6553	972541	France	Charles Heidsieck	Brut Millésimé 1961	1961
58	4.1	26	880792	France	Ruinart	Dom Ruinart Blanc de Blancs Brut Champagne 1976	1976
69	3.9	6182	42939	Italy	Guido Berlucchi	'61 Franciacorta Satén N.V.	1963
82	3.8	5079	35070	Italy	Guido Berlucchi	'61 Franciacorta Brut N.V.	1963

평균 평점 300위 안의 제조사 중 화이트 와인 추천

	Rating	Rating.Count	Price	Country	Brand	Wine.Name	Vintage
1	5.0	151	80739	Germany	Carl Loewen	1896 Riesling 2023	2023
2	4.8	152	23414	Germany	Jakob Schneider	Niederhäuser Felsensteyer Riesling Trocken 2022	2022
5	4.6	237	34498	Germany	Fritz Haag	Brauneberger Juffer Sonnenuhr Riesling Spätlese 2019	2019
8	4.6	142	425716	France	Vincent Dauvissat	Les Preuses Chablis Grand Cru 2015	2015
21	4.4	3631	81331	Germany	Egon Müller - Scharzhof	Scharzhofberger Riesling Kabinett 2002	2002
24	4.4	698	48444	France	Gustave Lorentz	Vieilles Vignes Riesling Alsace Grand Cru Altenberg de Berg...	2017

프랑스산이고 빈티지가 1990~2010년 사이인 디저트 와인 추천

	Rating	Rating.Count	Price	Country	Brand	Wine.Name	Vintage
125	3.8	9644	21403	France	JP. Chenet	Delicious Medium Sweet Moelleux Red 2006	2006
126	3.8	9644	21403	France	JP. Chenet	Delicious Medium Sweet Moelleux Red 2004	2004
132	3.7	1589	53507	France	Dourthe	Grands Terroirs Sauternes 2005	2005
139	3.7	1589	53507	France	Dourthe	Grands Terroirs Sauternes 2005	2005
9	4.7	562	763353	France	Château d'Yquem	Sauternes 2000	2000
14	4.6	224	269126	France	Château Climens	Barsac (Premier Grand Cru Classé) 2001	2001

결론 2. 개선점 및 한계점

개선점 :

1. 와인 추천 알고리즘을 도입하면, 기존의 와인 필터링 함수 기반 접근 방식보다 더 정교하고 개인화된 추천이 가능해질 것이다. 이를 통해 추천 정확도를 높일 수 있으며, 사용자의 선호를 더 잘 반영한 결과를 얻을 수 있다.

2. 변수 개수가 적어 모델에 큰 영향을 미치는 변수들이 존재하는데, 특히 리뷰 평점이 과도한 영향력을 미칠 수 있다. 이로 인해 특정 변수에 대한 의존도가 높아지며, 데이터의 다양성과 변화를 반영하는 데 어려움이 있을 수 있다.

3. 레드와 화이트 와인을 제외한 나머지 와인 종류에 대한 데이터가 적어, 각 관측치가 모델에 미치는 영향이 매우 크다. 특히 관측치가 적은 경우, 성능 평가에서 하나의 예측 실패가 모델 지표에 큰 영향을 미칠 수 있어, 모델의 안정성과 신뢰성에 대한 문제가 발생할 수 있다.

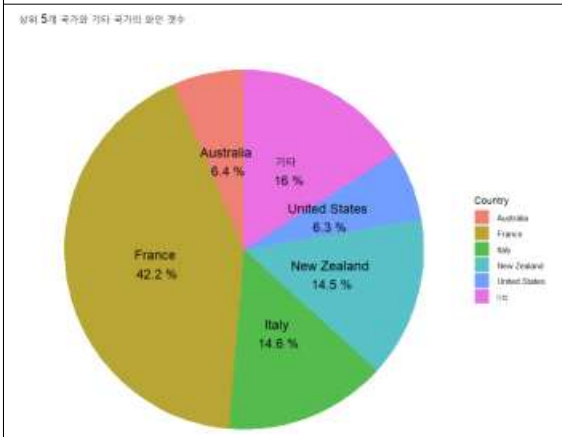
한계점:

1. K-means 클러스터링을 사용하여 유사한 속성을 가진 관측치들을 군집화했지만, 각 군집이 데이터의 대표성을 잘 반영할 수 있을지에 대한 의문이 존재한다. 클러스터의 수와 초기 중심점 선택에 따라 군집화 결과가 달라질 수 있으며, 군집의 성능을 보장하는 것이 어렵다.

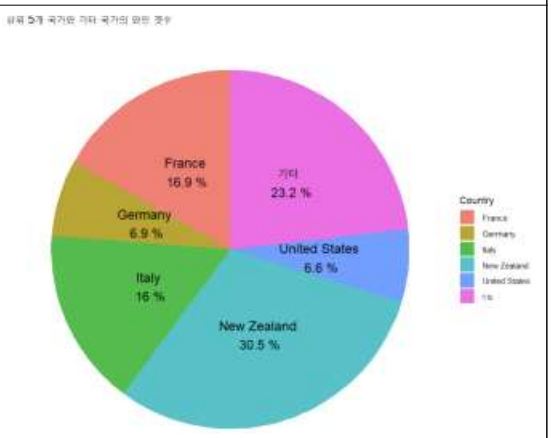
2. 수치형 변수를 표준화하는 과정에서 필터링이 어려운 상황이 발생할 수 있다. 표준화를 통해 값의 범위가 비슷해지지만, 이는 변수 간 상대적 중요도를 왜곡할 수 있고, 필터링이 의도한 대로 잘 작동하지 않을 수 있다.

3. 범주형 변수를 인코딩하는 과정에서 제조사를 평균 평점 기준으로 임의로 인코딩했으나, 이 방법이 실제 데이터와의 연관성을 반영하는 데 한계가 있을 수 있다. 인코딩의 근거가 부족하고, 평균 평점이 특정 제조사의 품질을 제대로 나타내지 못할 가능성이 있어, 모델에 잘못된 정보를 제공할 수 있다.

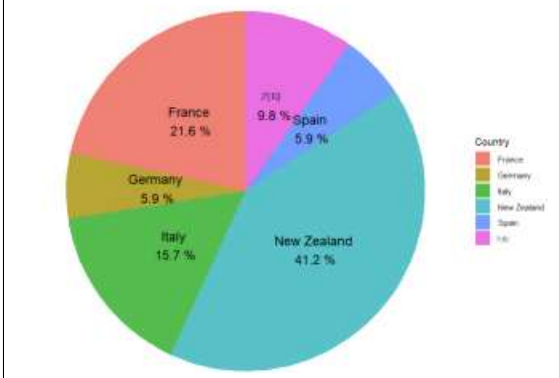
와인 종류 별 생산 국가 비중



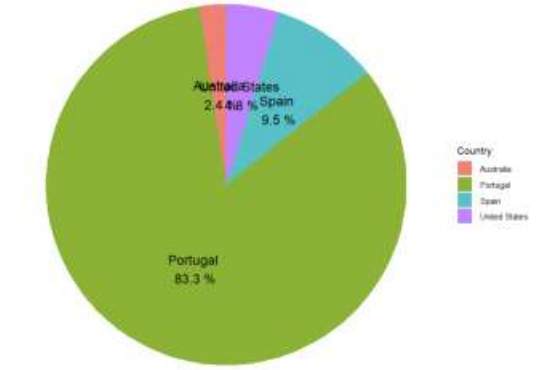
red wine



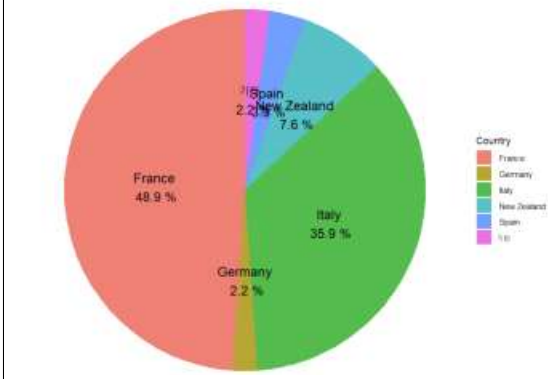
white wine



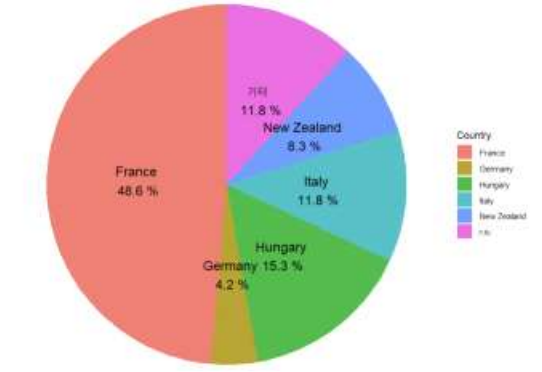
Rose wine



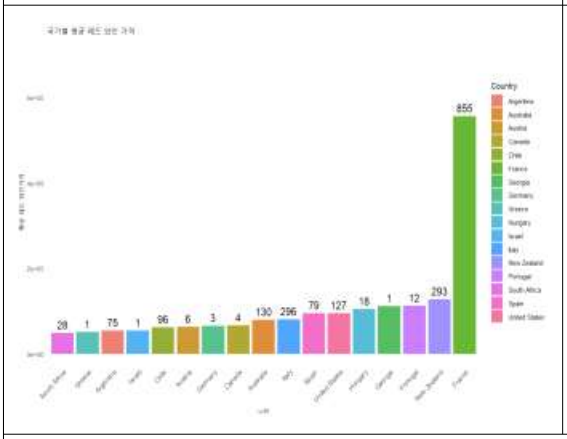
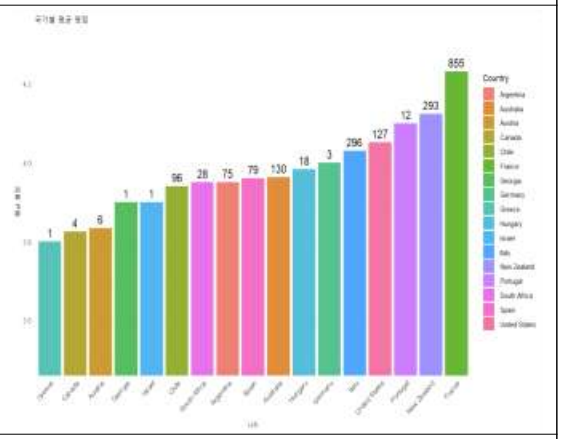
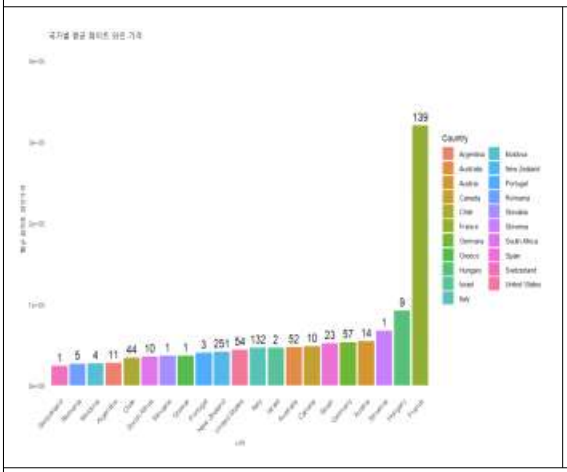
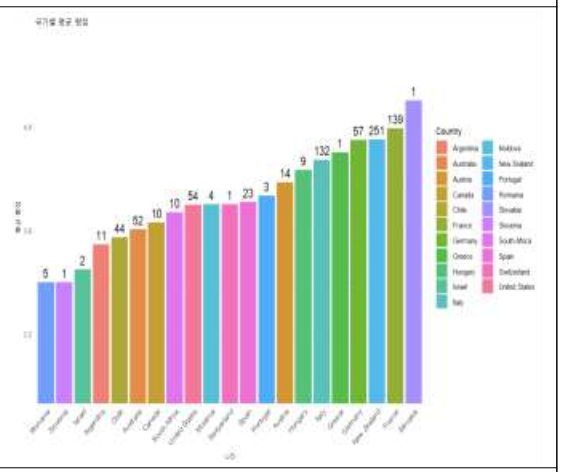
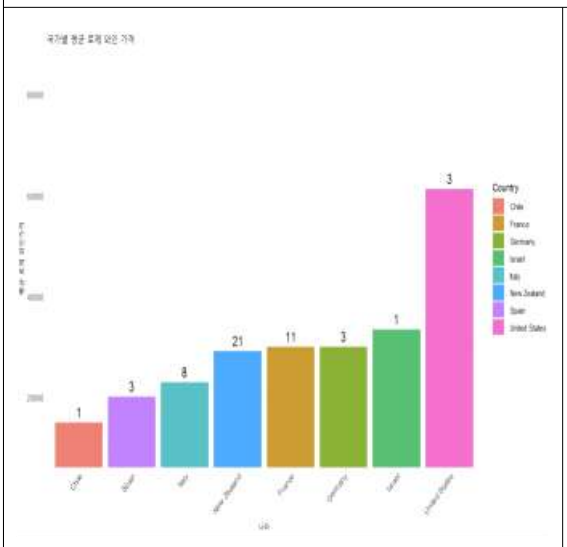
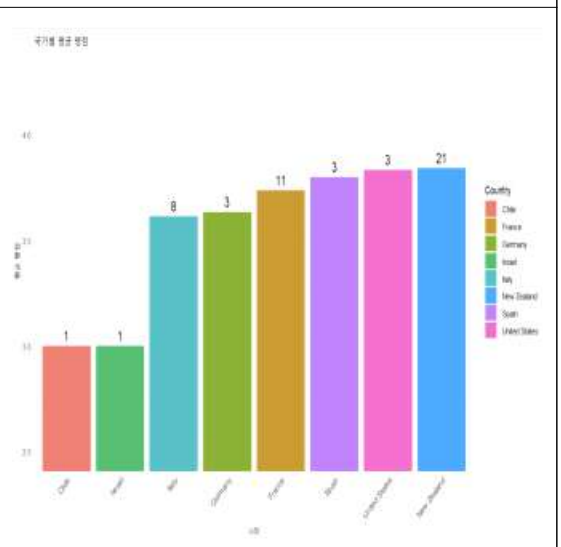
fortified wine

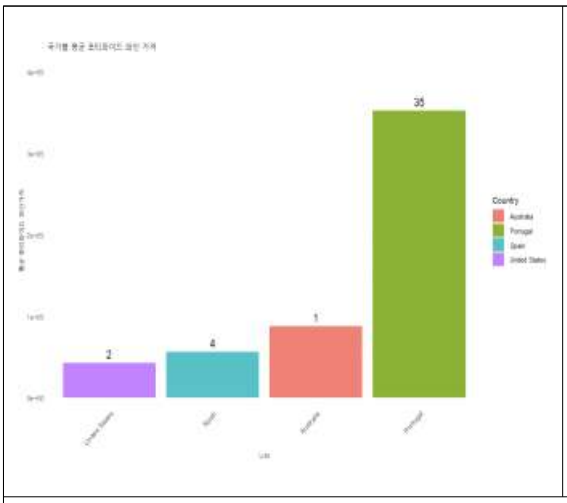


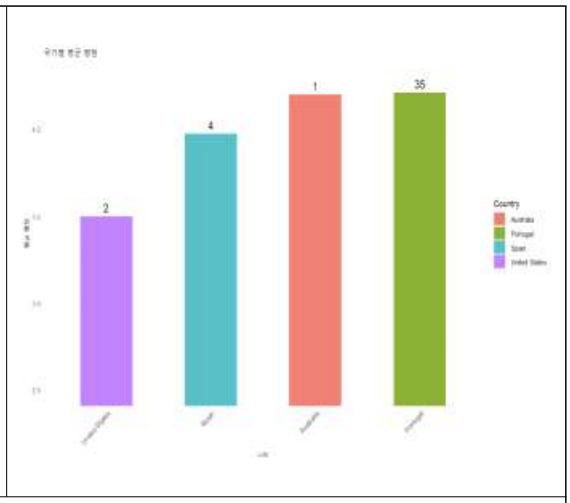
sparkling wine



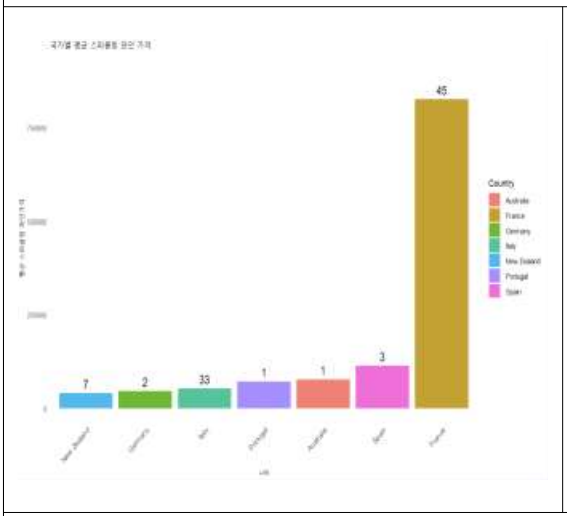
dessert wine

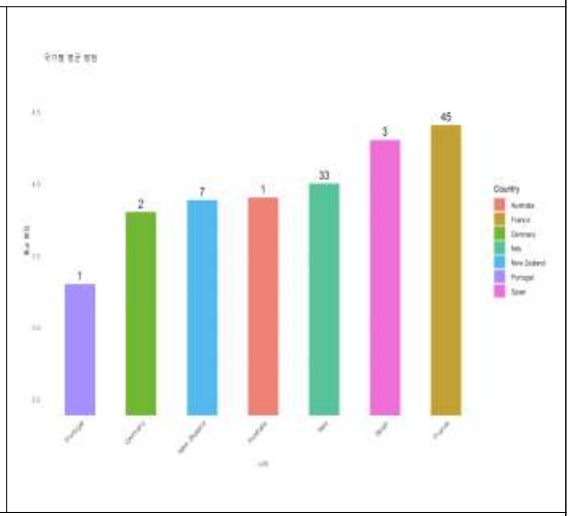
와인 종류 별 평균 가격과 평균 평점	
가격	평점
	
red wine	
	
white wine	
	
Rose wine	



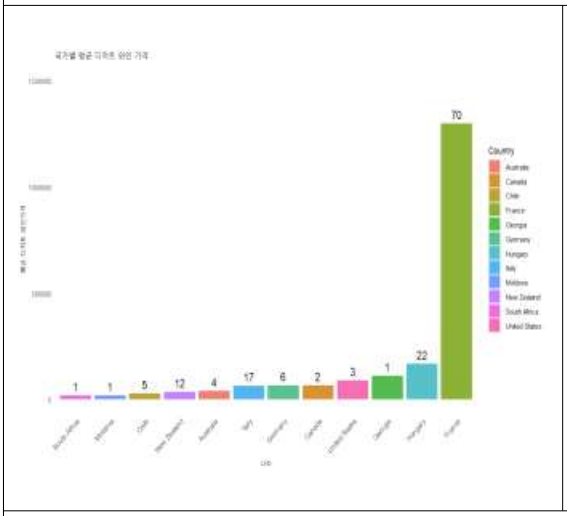


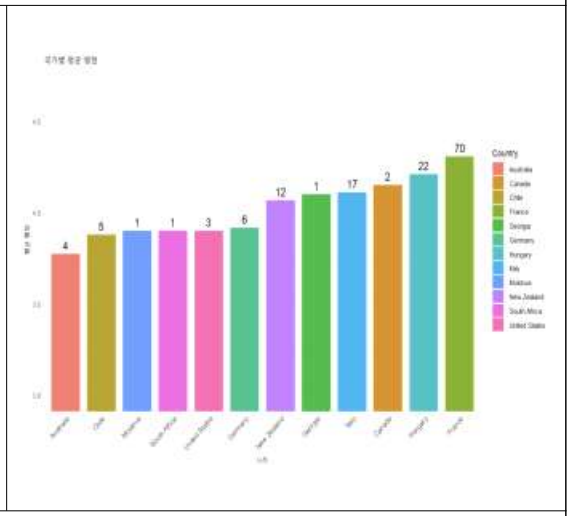
fortified wine





sparkling wine





dessert wine