# 러시아의 알코올 소비(Alcohol Consumption in Russia)에 대한 다변량 분석

20171038 김유경
20171063 정소희

Abstract

**Purpose** : This analysis is a graduation paper that examines and summarizes the relationship between alcohol types in Russia's per capita consumption (1998-2016).

**Method** : Kaggle.com's MM (UIISS)-Intergrated Departmental Information and Statistics System derived data. All multivariate analysis was performed with the package 'R 3.6.1'.

**Result** : Beer consumption was the largest and fluctuated the most (Of the five variables). The strongest correlation between [Brandy, Champagne]. The number of principal components is 2. and the second principal component is approximately 72%, The contrast between [Vodka, Brandy]. Cumulative explanation after Varimax rotation with 2 factors is 58.1%. Clustering with 5 clusters

## 1. 서론

알코올은 오늘날 현대 사회에서 기쁜 날 좋은 사람들과의 자리에 항상 함께하며 스트레스 배출구로써의 역할까지 많은 성인들에게 없어서는 안될 존재가 되었다. 전 세계 어느 나라를 가도 그 나라만의 음주 문화가 존재하고 즐겨 마시는 술이 다양하다. 우리나라에서는 평일 간 열심히 일한 자, 주말 전 마지막 금요일은 뜨겁게 즐기자는 심리가 하나의 문화처럼 자리 잡으며 “불타는 금요일”이라는 은어가 국어사전에까지 등재되었다. 여기에 코로나 시국까지 겹쳐 집에서 혼술을 즐기는 사람들이 늘어나면서 연간 알코올 판매량이 급증했다는 뉴스 보도도 있었다. 농림축산식품부와 한국 농수산식품유통공사의 ‘2019 주류시장 트렌드 보고서’에 따르면 우리나라 월평균 주종별 음주 비중은 맥주(36.9%), 소주(32.9%), 전통주(20.1%)로 맥주와 소주가 비슷한 퍼센트로 주류 소비 시장의 거의 70%가량을 차지하고 있다. 세계인이 모두 즐기는 알코올 하면 단연코 처음으로 생각나는 나라는 바로 러시아이다. 알코올 섭취로 많은 사회적 문제들이 나타나면서 러시아 정부의 최근 10년간 알코올 규제로 인해 알코올 소비량이 많이 줄어든 추세지만, 여전히 많은 사람에게 알코올의 나라로 강렬하게 인식되고 있는 러시아에서는 어떤 종류의 술을 많이 소비할까 문득 궁금증이 생겼다. 본 논문에서는 1998년도부터 2016년도까지의 러시아 알코올 소비에 대한 데이터에 있는 2개의 범주형 변수와 5개의 연속형 변수들로 기초통계량들을 구해보고, 상관분석, 주성분 분석, 인자분석, 군집 분석 등 다변량 분석을 통해 각 요인들 간 관계들을 파악하고자 한다.

## 2. 러시아의 알코올 소비 데이터

### 2.1 데이터 출처 및 변수 설명

분석에 사용된 러시아의 1인당 연간 알코올 소비량 데이터(1998~2016년도)는 kaggle(https://www.kaggle.com/) 에서 얻은 데이터이다. 본 데이터는 year, region, wine, beer,

vodka, champagne, brandy 총 7개의 변수를 갖는데, 이 중 wine, beer, vodka, champagne, brandy 5개 변수는 연속형, year, region은 범주형 변수이며 모든 변수를(결측값 제거) 사용하여 분석을 진행하였다. 아래 표 2.1 는 각 변수에 대한 설명이다.

| 변수명 | 변수 설명 | 변수 단위 |
|---|---|---|
| year | 1998-2016 사이의 연도 | . |
| region | 러시아 연방 지역들의 이름 | . |
| wine | 1인당 연간 와인 판매량 | *l* (리터, 연속형) |
| beer | 1인당 연간 맥주 판매량 | *l* (리터, 연속형) |
| vodka | 1인당 연간 보드카 판매량 | *l* (리터, 연속형) |
| champagne | 1인당 연간 샴페인 판매량 | *l* (리터, 연속형) |
| brandy | 1인당 연간 브랜디 판매량 | *l* (리터, 연속형) |

표 2.1 변수명 및 변수 설명

## 2.2 기초통계량

표 2.2는 러시아 알코올 소비에 대한 데이터에서 연속형 변수들에 대한 기초통계량으로 각각 변수에 대한 평균, 표준편차, 최댓값, 최솟값을 나타낸다. 단위는 리터로 모두 같으며, 평균은 beer(맥주)가 가장 높고 brandy(브랜디)가 가장 낮다. 즉 러시아에서 가장 많이 소비되고 있는 알코올의 종류는 beer(맥주)라는 것을 알 수 있다.

| variable | mean | sd | max | 50% | min |
|---|---|---|---|---|---|
| wine | 5.64 | 2.81 | 18.1 | 5.40 | 0.1 |
| beer | 51.52 | 25.18 | 207.3 | 49.97 | 1 |
| vodka | 11.85 | 5.1 | 40.6 | 11.50 | 0.4 |
| champagne | 1.32 | 0.8 | 5.56 | 1.20 | 0.1 |
| brandy | 0.53 | 0.40 | 2.3 | 0.40 | 0 |

표 2.2 연속형 변수 기초통계량

그림 2.1은 범주형 변수 년도 별로 각 연속형 변수의 평균을 구해 히스토그램으로 나타낸 것으로 알코올 종류별 소비량 추이를 한눈에 볼 수 있다. 전체적으로 맥주 소비량이 가장 높았다는 것을 확인할 수 있고, 2008년 이후로 서서히 전체적인 알코올 소비량이 줄어들고 있는 것을 확인할 수 있다.
그림 2.2는 (년도 + 지역) 별 각 연속형 변수의 평균을 구해 정리한 표이다.

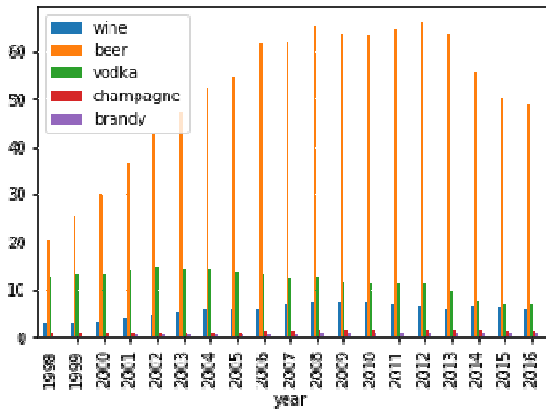


그림 2.1 년도 별 변수 평균 히스토그램

| year | region | wine | beer | vodka | champagne | brandy |
|---|---|---|---|---|---|---|
| 1998 | Altai Krai | 3.3 | 19.2 | 11.3 | 1.1 | 0.1 |
| | Altai Republic | 3.4 | 7.6 | 9.0 | 0.5 | 0.1 |
| | Amur Oblast | 2.1 | 21.2 | 17.3 | 0.7 | 0.4 |
| | Arkhangelsk Oblast | 4.3 | 10.6 | 11.7 | 0.4 | 0.3 |
| | Astrakhan Oblast | 2.9 | 18.0 | 9.5 | 0.8 | 0.2 |
| ... | ... | ... | ... | ... | ... | ... |
| 2016 | Vologda Oblast | 8.6 | 37.3 | 9.9 | 1.2 | 0.6 |
| | Voronezh Oblast | 5.5 | 43.4 | 4.3 | 1.3 | 0.6 |
| | Yamalo-Nenets Autonomous Okrug | 4.5 | 75.8 | 8.2 | 1.7 | 1.3 |
| | Yaroslavl Oblast | 10.2 | 38.0 | 8.9 | 1.4 | 1.0 |
| | Zabaykalsky Krai | 6.4 | 30.8 | 6.8 | 0.9 | 0.3 |

표 2.3 년도 + 지역별 변수 평균

## 3. 다변량 통계분석

### 3.1 상관분석

러시아에서 많이 소비하는 술 종류들을 대상으로 상관분석을 실시하여 변수 간 관련성을 알아보고자 한다. 상관분석을 통해 상관계수를 구하고 해석하여 상관관계의 정도를 파악할 수 있다. 상관계수는 변수 간 관계의 정도와 방향을 하나의 수치로 요약해주는 지수로 이 지수를 해석하여 상관 정도를 파악할 수 있다. 상관계수는 -1.00에서 +1.00 사이의 값을 가지며, +1.00에 가까울수록 강한 양의 상관관계를 나타내고 -1.00에 가까울수록 강한 음의 상관관계를 나타낸다. 보통 상관계수가 -0.2 사이에서 +0.2 사이이면 상관관계가 거의 없다고 간주한다. 양의 상관관계는 한 변수가 증가함에 따라 다른 변수도 증가함을 의미하고, 음의 상관관계는 한 변수가 증가함에 따라 다른 변수는 감소함을 의미한다. 표 3.1은 변수 간의 상관계수를 표로 나타낸 상관행렬 표이다. 모든 변수는 양의 상관관계를 가지고 있음을 확인할 수 있는데, 초록색으로 나타낸 vodka와 beer은 +0.2보다 작은 수로 상관관계가 거의 없다고 볼 수 있다. 빨간색으로 나타낸 변수들의 상관계수는 +0.4 이상의 비교적 높은 양의 상관계수를 가지고 있는 것을 볼 수 있는데 이 중 brandy와 champagne은 +0.78로 강한 양의 상관관계에 있음을 알 수 있다.

| | wine | beer | vodka | champagne | brandy |
|---|---|---|---|---|---|
| wine | 1.00 | 0.49 | 0.27 | 0.47 | 0.56 |
| beer | 0.49 | 1.00 | 0.19 | 0.45 | 0.46 |
| vodka | 0.27 | 0.19 | 1.00 | 0.27 | 0.21 |
| champagne | 0.47 | 0.45 | 0.27 | 1.00 | 0.78 |
| brandy | 0.56 | 0.46 | 0.21 | 0.78 | 1.00 |

표 3.1 변수 간 상관행렬

아래 그림 3.1은 위 상관행렬을 여러 가지 그래프로 시각화한 것으로 비교적 한눈에 변수 간 상관관계를 파악할 수 있다. 모든 상관관계가 양의 상관관계였기 때문에 붉은색은 없고 모두 푸른색으로 나타나 있는 것을 볼 수 있다. 강한 양의 상관관계를 가지고 있었던 brandy와 champagne은 진한 파란색으로, 비교적 강한 관계를 가지고 있던 변수 쌍인 [champagne과 wine], [champagne과 beer], [brandy와 wine], [brandy와 beer은 중간 진하기의 파란색으로, 상관관계가 거의 없다고 보여지는 (+0.2보다 작은 상관계수를 갖는) 변수 쌍인 [vodka와 beer] 그리고 이에 근접한 상관계수를 가지고 있는 변수 쌍인 [vodka와 wine], [champagne과 vodka], [brandy와 vodka]는 매우 연한 파란색으로 나타나는 것을 볼 수 있다. 즉 색이 진할수록 강한 양의 상관관계에 가까우며, 색이 연할수록 상관관계가 없다고 볼 수 있다.

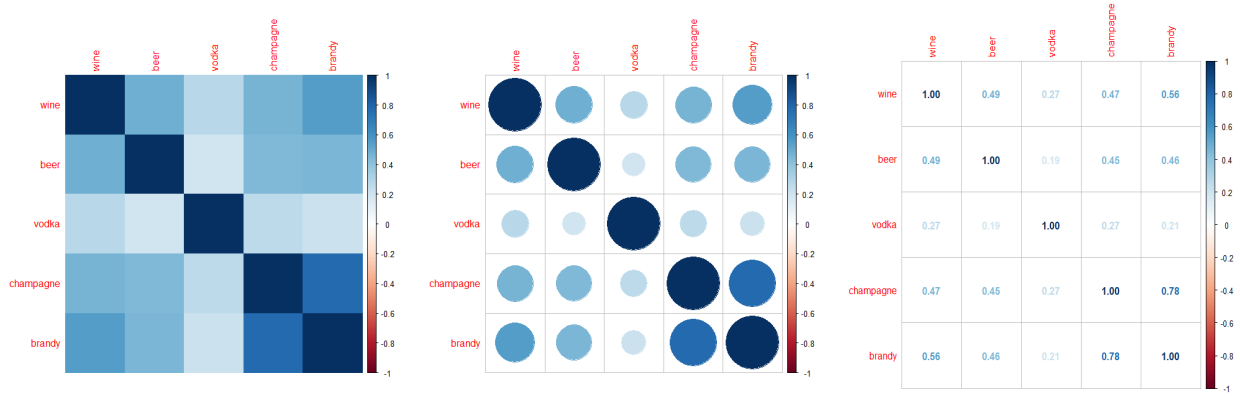


그림 3.1 상관행렬의 시각화

### 3.2 주성분 분석

주성분 분석은 여러 개의 반응변수로 얻어진 다변량 데이터에 대해 분산-공분산 구조를 변수들의 선형결합식으로 설명하고자 하는 방법이다. 주성분 분석의 목적은 차원 축소, 변동이 큰 축 탐색, 주성분을 통한 데이터의 해석 등에 있다. 주성분은 서로 독립적인 새로운 변수이며 p개의 변수에 포함된 전체 변동을 p개보다 작거나 같은 m개의 주성분으로 대신하여 설명한다. 일반적으로 주성분 분석은 표본 공분산행렬을 이용하지만, 변수의 단위가 다르거나 분산의 차이가 큰 경우 표본 상관행렬을 이용한다. 이 데이터에서는 변수의 단위는 같지만 분산의 차이가 크기 때문에 표본 공분산행렬을 사용할 경우 분산이 큰 변수가 주성분의 압도적인 비중을 차지할 수 있으므로 분석의 균형을 유지하기 위해서 표본 상관행렬을 이용해 주성분 분석을 시행하였다.

| | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 |
|---|---|---|---|---|---|
| 표준편차 | 1.65 | 0.94 | 0.82 | 0.71 | 0.46 |
| 분산 비율 | 0.55 | 0.18 | 0.13 | 0.10 | 0.04 |
| 누적 비율 | 0.55 | 0.72 | 0.86 | 0.96 | 1.00 |

표 3.2 상관행렬을 이용한 분산 설명량

주성분 분석 시 원래 데이터에 대한 정보의 손실을 최소화하기 위해서는 적절한 개수의 주성분을 선택해야 한다. 주로 누적 비율이 전체 주성분의 70%~90%가 되도록 주성분의 수를 결정한다, 위의 표 3.2를 보면 첫 번째 주성분은 전체의 약 55% 두 번째 주성분은 전체의 약 72%를 보여주고 있다. 따라서 이 데이터에서 적절한 주성분의 개수는 2개라고 할 수 있다.

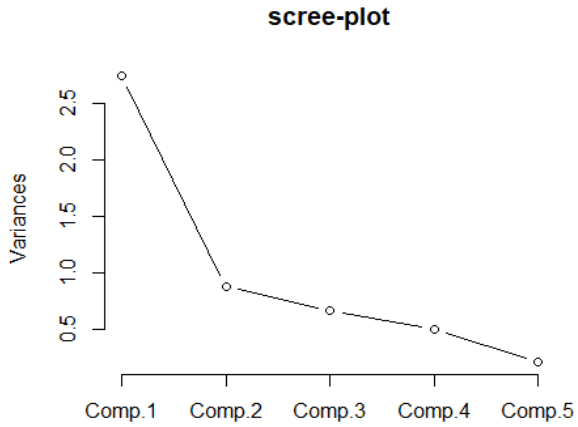


그림 3.2 스크리 그래프

위의 그림 3.2는 스크리 그래프를 통해 적절한 주성분 개수를 선택하는 방법이다. 스크리 그래프를 통한 방법은 그래프의 가파른 정도를 보고 큰 고유값과 작은 고유값을 구분하여 자연스럽게 적절한 개수를 정하는 방법이다. 그림을 보면, Comp.2인 두 번째 이후로 경사가 완만해지는 것을 볼 수 있다. 따라서 스크리 그래프를 통한 방법에서도 이 데이터에서 적절한 주성분의 개수는 2개라는 것을 알 수 있다.

| | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 |
|---|---|---|---|---|---|
| wine | 0.465 | | 0.366 | 0.788 | 0.172 |
| beer | 0.429 | -0.145 | 0.690 | -0.564 | |
| vodka | 0.262 | 0.954 | | -0.113 | |
| champagne | 0.509 | -0.144 | -0.485 | -0.218 | 0.661 |
| brandy | 0.522 | -0.221 | -0.389 | | -0.725 |

표 3.3 주성분 계수

위의 표 3.3은 러시아의 알코올 소비 데이터의 주성분 계수를 나타낸 것으로, 주성분 계수를 선형결합식의 형태로 나타내면 다음과 같다.

첫 번째 주성분: $Y_1 = 0.465 \times wine + 0.429 \times beer + 0.262 \times vodka + 0.509 \times champagne + 0.522 \times brandy$

두 번째 주성분: $Y_2 = -0.145 \times beer + 0.954 \times vodka - 0.144 \times champagne - 0.221 \times brandy$

첫 번째 주성분에서는 각 계수의 부호가 모두 같으므로 이 데이터의 가중평균을 나타낸다고 할 수 있다. vodka는 가중평균이 0.262로 다른 변수들에 비해 주성분 $Y_1$을 설명하는 비중이 작다.

두 번째 주성분에서는 vodka의 값이 커질수록 주성분의 값이 커지고, brandy의 값이 커질수록 주성분의 값이 작아지기 때문에 vodka와 brandy의 대비성분이라고 할 수 있다.

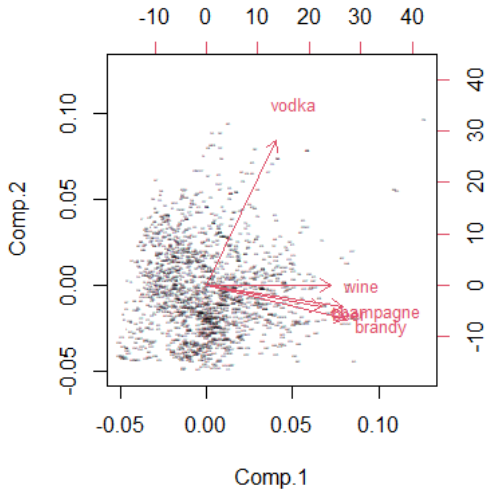


그림 3.3 주성분 그래프

그림 3.3은 첫 번째 주성분과 두 번째 주성분에 대해 그린 주성분 그래프이다. 화살표가 이루는 각도가 작을수록, 화살표가 같은 방향을 가리킬수록 변수들의 상관성이 높아지게 되는데, 그래프에서 vodka를 제외한 wine, beer, champagne, brandy가 서로 가까운 곳에 위치하고, 같은 곳을 가리키고 있으므로 vodka를 제외한 이들의 상관성이 매우 높다는 것을 알 수 있다.

### 3.3 인자분석

인자분석은 변수 간 내재하고 있는 공통적인 구조를 파악하고, 데이터의 특성을 몇 개의 인자로 축약하여 설명하고자 하는 분석법이다. 인자분석을 여러 개 변수의 상관성 구조를 나타내는 몇 개의 인자로 분석하는 것으로 인자분석을 통해 인자가 형성하는 차원과 그 차원에서의 변수 위치 및 의미를 파악할 수 있다. 주성분 분석에서 주성분은 관측된 변수들의 선형 결합 식으로 정의되는 반면, 인자분석에서는 변수들이 공통인자들의 선형 결합 식으로 정의된다는 측면에서 구분할 수 있다. 인자분석은 연속형 변수에 적용할 수 있으므로 분석데이터의 모든 변수(wine, vodka, beer, champagne, brandy)를 사용하여 분석하였다.

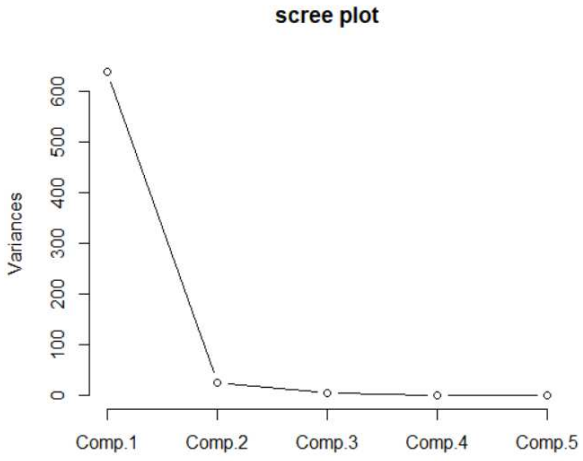


그림 3.4 인자 개수에 대한 스크리 그래프

그림 3.4은 인자분석 결과를 해석하기 위한 인자 개수를 결정할 수 있는 스크리 그래프로(scree graph) 값이 큰 고유값부터 크기순으로 점이 찍히며 값의 차이가 크면 가파른 경사를, 고유값의 변화가 작으면 완만한 경사로 그려진다. 큰 고유값과 작은 고유값을 구분하는 방법으로 가파른 정도를 보고 가파른 부분에 해당하는 고유값까지를 인자의 개수로 결정하는데, 두 번째 고유값에서 세 번째 고유값의 기울기가 완만해진 것으로 보아 적절한 인자의 개수는 2개임을 알 수 있다.

| 변수 | 회전 전 | | Varimax 회전 후 | | Promax 회전 후 | |
|---|---|---|---|---|---|---|
| | Factor1 | Factor2 | Factor1 | Factor2 | Factor1 | Factor2 |
| wine | 0.477 | 0.753 | 0.233 | 0.860 | -0.190 | 1.020 |
| beer | 0.456 | 0.364 | 0.328 | 0.482 | 0.157 | 0.458 |
| vodka | 0.268 | 0.182 | 0.202 | 0.252 | 0.123 | 0.223 |
| champagne | 0.997 | | 0.957 | 0.282 | 1.115 | -0.170 |
| brandy | 0.784 | 0.241 | 0.678 | 0.462 | 0.637 | 0.231 |

표 3.4 인자분석 결과 인자 적재 값 비교

인자분석 해석의 편리를 위해 인자회전을 시도하여 단순한 구조가 되도록 변환할 수 있는데, 인자 적재 값들의 제곱을 취해 이들의 분산을 최대화하여 인자적재값들의 차이가 많이 나도록 하는 Varimax 직교 회전 방법을 사용하였다. 표 3.4 는 Varimax 회전 전후 인자적재값을 비교하여 소수 셋째 자리까지 나타낸 것이다. 인자적재값이 0.5를 넘어서는 경우, 각 인자에서 변수 비중이 크다는 의미이므로 빨간색으로 특이값을 표시하였다. 전체 누적 설명력은 약 58.1% 갸량으로 나타났다. 요인1의 고유값은 1.578, 설명력은

31.6%로 요인2의 고유값은 1.329, 설명력은 약 26.6%로 파악되었다.

Varimax 회전 후 인자 모형은 다음과 같다.

$$Wine = 0.233 \times factor1 + 0.860 \times factor2 + \epsilon_1$$
$$Beer = 0.328 \times factor1 + 0.482 \times factor2 + \epsilon_2$$
$$Vodka = 0.202 \times factor1 + 0.252 \times factor2 + \epsilon_3$$
$$Champane = 0.957 \times factor1 + 0.282 \times factor2 + \epsilon_4$$
$$Brandy = 0.678 \times factor1 + 0.462 \times factor2 + \epsilon_5$$

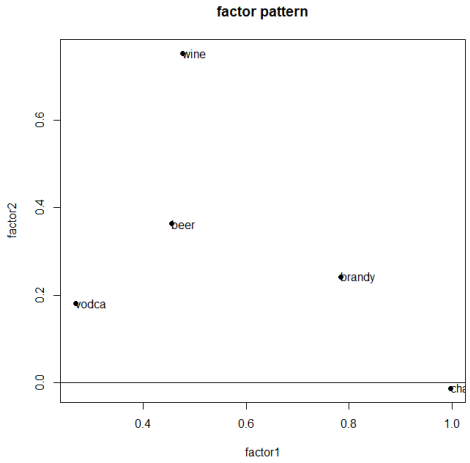


그림 3.5 회전 전 인자 패턴

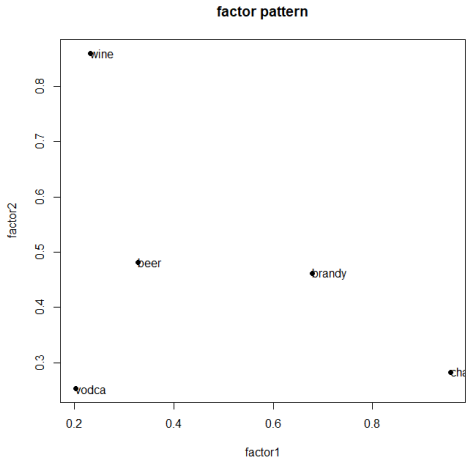


그림 3.6 Varimax 회전 후 인자 패턴

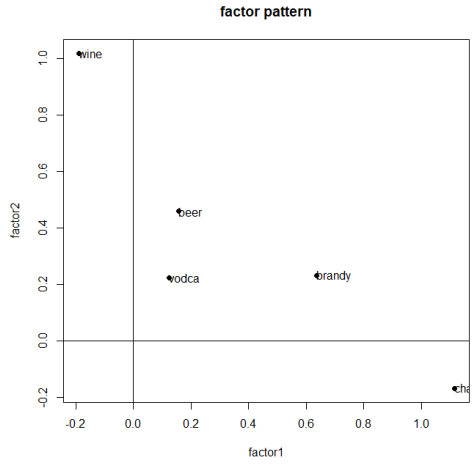


그림 3.7 Promax 회전 후 인자 패턴

그림 3.5는 회전 전 인자 패턴, 3.6은 Varimax 회전된 인자 패턴, 3.7은 Promax 회전된 인자 패턴을 나타내는 그림으로 각 변수에 대한 인자 적재값을 알려주므로 인자의 의미에 대한 정보를 얻을 수 있다. Varimax 회전 후 인자의 누적 설명력은 58.1%, Promax 회전 후 누적 설명력은 62.1%이다.

### 3.4 군집 분석

군집 분석은 시스템을 표현하는 데이터로부터 구조를 찾아내고 통계적 특성이 서로 다른 군집으로 분리 가능한지를 알아내는 것으로, 구체적인 군집 분석 방법에 따라 군집화 결과에 차이가 생길 수 있다. 군집 분석에서는 군집의 개수나 구조에 대한 가정 없이 다변량 데이터로부터 거리 기준에 의해 자발적인 군집화를 유도한다. 군집 분석의 목적은 적절한 군집으로 데이터를 나누고 각 군집의 특성, 군집 간 차이에 관해 연구하는 것이다. 각 관측 벡터에 대해 어떻게 군집으로 나눌 것인지를 판단하기 위해 관측벡터간의 거리를 이용하여 유사성을 측정하거나 또는 근접성을 측정하는 방법도 있다. 군집 분석은 계층적, 비계층적 군집 방법으로 나뉘는데 계층적 군집 분석은 처음 n개 군집부터 시작하여 점차 군집 개수를 줄여나가는 방법으로 최단연결법, 최장연결법, 평균연결법, Ward방법이 여기 속한다. 비계층적 군집 분석에는 전체 데이터를 K개의 군집으로 나누는 K-평균법이 있다.

계층적 군집 방법으로 분석데이터의 변수들을(모두 연속형 변수) 분석하려 하였으나, 데이터 크기가 너무 커서 관측 벡터 간 거리로부터 정상적으로 실행할 수 없었다. 따라서 비 계층적 군집 방법인 K-means clustering 으로 군집 분석을 시행하였다.

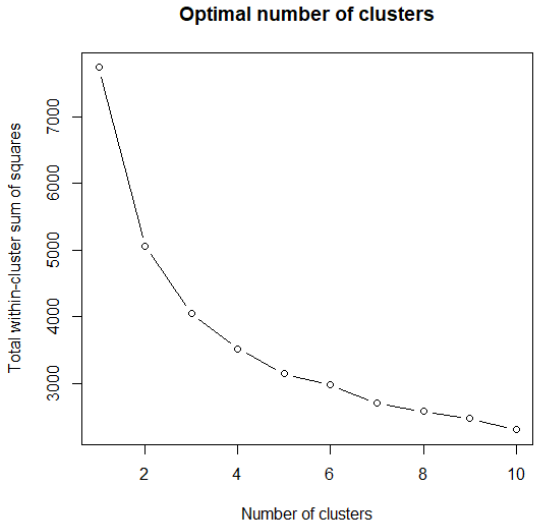


그림 3.8 군집 개수 결정 그래프

```
*******************************************************************
* Among all indices:
* 8 proposed 2 as the best number of clusters
* 1 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 9 proposed 5 as the best number of clusters
* 3 proposed 6 as the best number of clusters
* 1 proposed 8 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  5

*******************************************************************
```

그림 3.9 NbClust 이용한 최적 군집 개수 결정

데이터를 가장 잘 설명할 수 있는 군집의 개수를 찾기 위해 Elbow methods를 사용하여 최적의 군집 개수를 찾을 수 있다. 그림 3.8에서 기울기가 가장 완만해지는 때인 5가 군집 개수가 5였을 때 최적의 군집 개수임을 알려준다. 그림 3.9에서 NbClust를 사용하여 최적의 군집 개수를 계산한 결과가 5로 동일하다.

| cluster num | wine | beer | vodka | champagne | brandy |
|---|---|---|---|---|---|
| 1 | 6.526 | 55.655 | 11 | 1.462 | 0.598 |
| 2 | 7.584 | 103.046 | 15.546 | 1.984 | 0.866 |
| 3 | 2.945 | 19.157 | 10.693 | 0.785 | 0.261 |
| 4 | 6.872 | 76.151 | 12.426 | 1.685 | 0.724 |
| 5 | 5.298 | 38.255 | 12.124 | 1.109 | 0.416 |

표 3.5 k-평균법을 사용한 군집 별 평균 (군집 5개)

다음 표 3.5는 k-평균법으로 5개의 군집으로 나눈 후, 군집 별 평균을 소수점 셋째 자리까지 반올림하여 구한 표이다. 각각 군집 1에서는 417개, 군집 2에서는 116개, 군집 3에서는 303개, 군집 4에서는 294, 군집 5에서는 419개의 관측치가 존재하며 클러스터별 제곱합 비중은 (cluster sum of squares by cluster) 86.4%를 차지한다. 군집 2번에서는 beer의 비중이 특히 크고 군집 3번에서는 wine의 비중이 특히 작은 것을 확인할 수 있다.

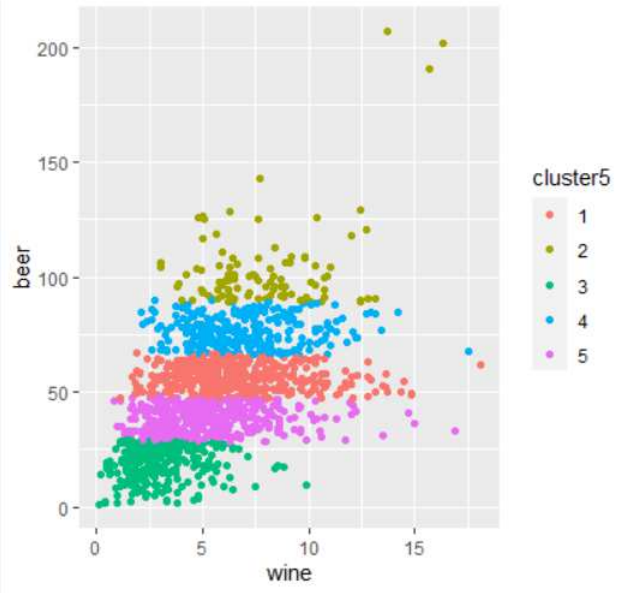


그림 3.10 클러스터 5개 실행 후 그래프

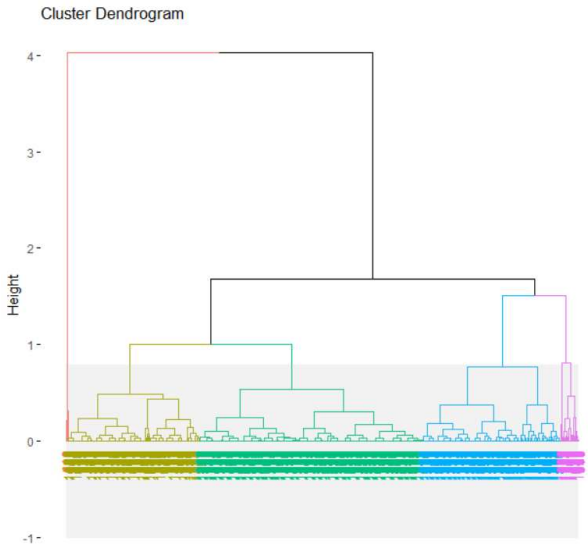


그림 3.11 클러스터 덴드로그램

그림 3.10은 5개 군집으로 나눠 분석했을 때 변수 중 한 예로 wine과 beer의 그래프를 그린 것인데 살짝 겹치는 부분이 있지만, 전체적으로 잘 분리되어 있다. 그림 3.11 카테고리 군집 별 wine 소비량 덴드로그램으로 빨강, 노랑, 초록, 파랑, 보라 다섯 가지 색으로 분리되어 5개 군집으로 잘 나뉘어 있는 것을 볼 수 있다. table()로 클러스터 된 값과 실제 값이 맞는지 확인했을 때 결과가 [417, 116, 303, 294, 419]로 앞 k-평균법(군집 5개)으로 군집 분석한 값과 정확히 일치하는 것을 확인할 수 있었다.

## 4. 결론

현대사회에서 알코올은 현대인들에게 위로가 되어주고 하나의 문화가 되어가고 있다. 흔히 술을 가장 많이 그리고 잘 마신다고 알려져 있는 나라인 러시아에서는 어떤 술을 즐겨마시고 그 소비량을 어느정도인지 알아보기 위하여 본 분석를 실시하였다. 본 분석에서는 러시아의 1인당 연간 알코올 소비량 데이터(1998~2016년도)에서 범주형 변수인 year, region과 연속형 변수인 wine, beer, vodka, champagne, brandy를 이용하여 기초통계량을 구하고, 상관분석, 주성분분석, 인자분석, 군집분석 등 여러 다변량 분석을 시행하였다.

첫 번째로 기초통계량 조사 결과, 러시아에서 즐겨 마시는 주종인 wine, beer, vodka, champagne, brandy 다섯가지 주종의 1인당 소비량에 대한 각 변수들의 평균, 표준편차, 최소값과 최대값을 확인할 수 있었다. 다섯가지 주종 중 beer에 대한 평균이 압도적으로 맥주 소비량이 굉장히 크다는 것을 알 수 있었고 그와 동시에 표준편차 역시 매우 커서 변동량도 매우 크다는 것을 알 수 있었다.

두 번째로 변수들간의 상관관계를 파악하기 위해 상관분석을 실시하였다. 모든 상관계수가 양수 값을 가져 대부분 적절한 양의 상관성을 가지고 있는 것으로 보였지만, 그 중 상관계수가 압도적으로 큰 brandy와 champagne의 강한 양의 상관관계와, 이와 대비되어 vodka와 beer은 상관성이 없음을 발견하였다.

세 번째로는 표본 상관행렬을 이용해 주성분분석을 실시하였다. 데이터에 대한 정보의 손실을 최소화하기 위해서 스크리 그래프로 적절한 주성분의 개수를 2개로 결정한 후에 분석을 실시하였다. 그 결과 첫 번째 주성분에서는 각 계수의 부호가 모두 같으므로 이 데이터의 가중평균을 나타내는 것을 알았다. 두 번째 주성분에서는 vodka의 값이 커질수록 주성분의 값이 커지고, brandy의 값이 커질수록 주성분의 값이 작아지므로 vodka와 brandy는 서로 대비관계에 있음을 알 수 있었다. 두 번째 주성분은 누적비율 약 72%를 설명하였다.

네 번째로는 인자분석을 시행하였다. 우선 스크리 그래프를 통해 적절한 인자의 개수를 2개로 결정 후 분석을 실시하였다. 해석의 편리를 위해 Varimax 회전 후 요인1은 약 31.6%를, 요인2는 약 26.6%의 설명력을 가졌고, 누적 설명력은 58.1% 였다.

마지막으로 연속형 변수만을 사용하여 군집분석을 시행하였다. 역시 Elbow methods (스크리 그래프)와 NbClust를 사용하여 최적의 군집 개수를 계산한 결과대로 군집 개수를 5개로 결정 후 비계층적 군집방법인 K-means clustering 분석을 실시하였다. 각각 [417, 116, 303, 294, 419] 의 관측치가 존재했고, 클러스터별 제곱합 비중 86.4%였다. 특히 군집 2번에서는 beer의 비중이 크고 군집 3번에서는 wine의 비중이 특히 작은 것을 발견할 수 있었다.

## 참고문헌


1. 김재희 저 | 교우사 | 2015, <R 다변량 통계분석(개정판)>
2. 덕성여자대학교 정보통계학과(2019) 다변량 통계분석 모음집, 덕성여자대학교
3. 강현철, 연규필 외 1명 저 | 자유아카데미 | 2021. <R을 활용한 다변량 자료분석 방법론>
4. 시모야마 테루마사 외 2명, 손민규 저 | 손민규 역 | 위키북스 | 2020,
   <파이썬 데이터 분석 실무 테크닉 100>
5. [R, Python 분석과 프로그래밍의 친구 (by R Friend)]
https://rfriend.tistory.com/586?category=706119
6. https://www.khan.co.kr/world/europe-russia/article/201910021608001
7. 우리나라 선호 주종 https://www.yna.co.kr/view/AKR20200918125000030
8. 상관분석 http://www.6025.co.kr/bbs/board.php?bo_table=cust_in&wr_id=13
9. K-Means Clustering in R: Algorithm and Practical Examples ...
https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/
10. Factor Analysis with the Principal Factor Method - RPubs
https://rpubs.com/aaronsc32/factor-analysis-principal-factor-method