제 8회 KCB 시각화 경진대회

슈퍼콘 팀

임정우 손동희 오세인 이현재 최인서

목차

1 전국민 카드 및 대출 이용통계 데이터 (credit.csv) 분석

Jeju 지역 거주지(제주도) 금융라이프 데이터 분석 2

1. 전국민 카드 및 대출 이용통계 데이터 분석 (credit.csv)

DACON
DATA TO VALUE

단순히 변수 별로 시각화를 진행하였을 경우 변수의 개수도 많고 개별 변수의 영향력만 고려하게 되므로 인사이트 도출이 쉽지 않다고 판단하였다. 이에 우선적으로 **고객들의 금융 스타일을 파악**하고 규명하기 위해 **기존 변수들과 파생변수들을 이용하여 클러스터링을 진행**하였다.

```
## ************************************************************************
## * Among all indices:
## * 1 proposed 2 as the best number of clusters
## * 13 proposed 3 as the best number of clusters
## * 4 proposed 4 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 3 proposed 8 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
```

<사용된 변수>
1) 은행업종 총 대출금액
2) 카드업종 총 대출금액
3) 할부금융업종 총 대출금액
4) 보험업종 총 대출금액
5) 저축은행업종 총 대출금액
6) 담보대출 대출금액 총합
7) 신용카드 일시불 이용금액 합
8) 신용카드 할부 이용금액 합
9) 실카드 사용수
10) 총 대출 약정 금액 - 총 대출금액
11) 분할상환비율

먼저 기존의 변수와 파생된 변수들로 클러스터링을 해 보았을 때, 3개의 군집이 가장 적절하다고 나왔다.(kmeans)
각 군집별 특성을 살펴보면 다음과 같다.

- **클러스터1:** 은행업종 총 대출금액, 담보대출 대출금액 총합 변수의 값이 높았다. 즉, 제1금융권 대출을 주로 하며 집이나, 차 같은 담보대출이 있는 스타일이다.
- **클러스터2:** 저축은행은행업종 총 대출금액 변수의 값이 높았다. 즉, 은행대출과 담보대출 등 금리가 낮은 대출과는 관련이 상대적으로 떨어지는 반면에 저축은행의 대출과 같이 상대적으로 금리가 높은 대출을 가지고 있는 스타일이다.
- **클러스터3:** 카드업종 총 대출금액, 할부금융업종 총 대출금액, 보험업종 총 대출금액, 신용카드 총 이용금액(할부+일시), 분할상환비 변수에서 높은 값이 나왔다. 즉, 제2금융권에서 주로 대출을 하면서 분할 상환비율이 높은 금융스타일이다.

1. 전국민 카드 및 대출 이용통계 데이터 분석 (credit.csv)

클러스터링을 통해 사람들의 금융스타일을 구분하고 파악할 수 있지만 한 클러스터 내에서의 변화나 차이는 규명하지 못한다. 이를 더욱 엄밀하게 파악하고 해석하기 위하여 Factor analysis를 진행하였고 해석가능한 잠재변수를 찾고자 했다. 사용된 변수는 clustering에서 사용한 변수와 동일하다.

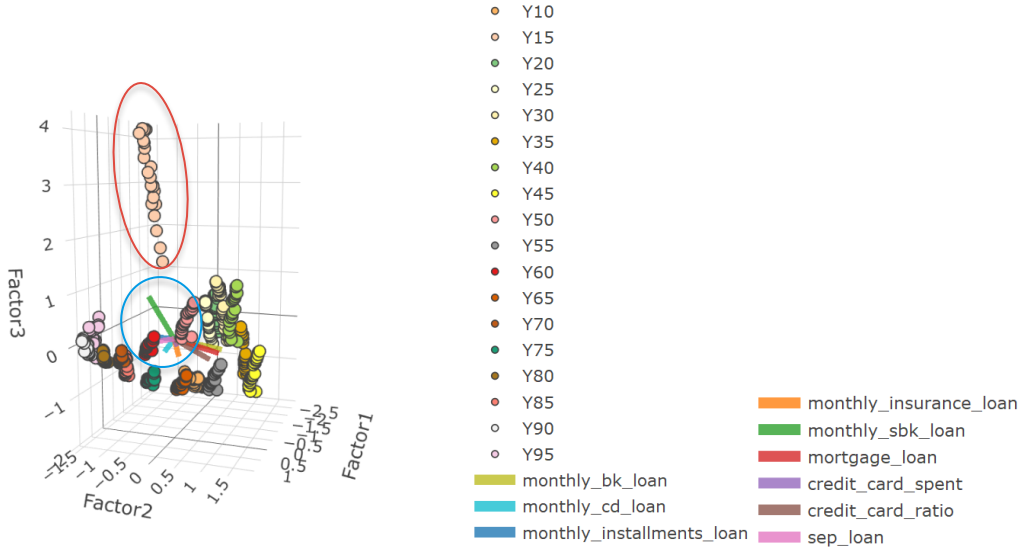
Factor analysis 결과 모든 변수를 설명할 수 있는 잠재변수 3개를 만들어낼 수 있었고 각각의 변수는 다음과 같은 의미를 가진다.

- **RC1**는 카드업종 총 대출금액, 할부금융업종 총 대출금액, 보험업종 총 대출금액, 신용카드 총 이용금액(할부+일시), 분할상환비 변수에서 높은 로딩값을 가졌다. 즉, RC1은 **제2금융권에서 주로 대출을 하면서 분할 상환비율이 높은 직장인일 가능성**을 나타내는 변수이다. (Cluster 3의 금융스타일 정도를 나타내는 지수)
- **RC2**는 은행업종 총 대출금액, 실카드 사용비율, 담보대출 대출금액 총합 변수에서 높은 로딩값을 가졌다. **제1금융권 대출을 주로 하며 집이나, 차 같은 담보대출이 있을 가능성**을 나타내는 변수이다. (Cluster 1의 금융스타일 정도를 나타내는 지수)
- **RC3**은 저축은행업종 총 대출금액 변수에서 높은 로딩값을 가졌다. 은행대출과 담보대출 등 금리가 낮은 대출과는 관련이 상대적으로 떨어지는 반면 **저축은행의 대출과 같이 상대적으로 금리가 높은 대출을 가지고있을 가능성**을 나타내는 변수이다. (Cluster 2의 금융스타일 정도를 나타내는 지수)

즉, 클러스터 1에 속하는 금융스타일을 가졌다면 RC2 변수의 값이 높을 것이고 클러스터 2에 속하는 금융스타일을 가졌다면 RC3 변수의 값이 높을 것이다. 더 나아가 같은 클러스터를 가졌다 해도 RC1변수와 RC2변수 값의 고저에 따라 **사람들의 변화하는 금융스타일을 더욱 엄밀하게 파악**할 수 있을 것이다.

1. 전국민 카드 및 대출 이용통계 데이터 분석 (credit.csv)

각 pop_cd 별로 factor analysis결과를 3D-plot으로 시각화 하였다.

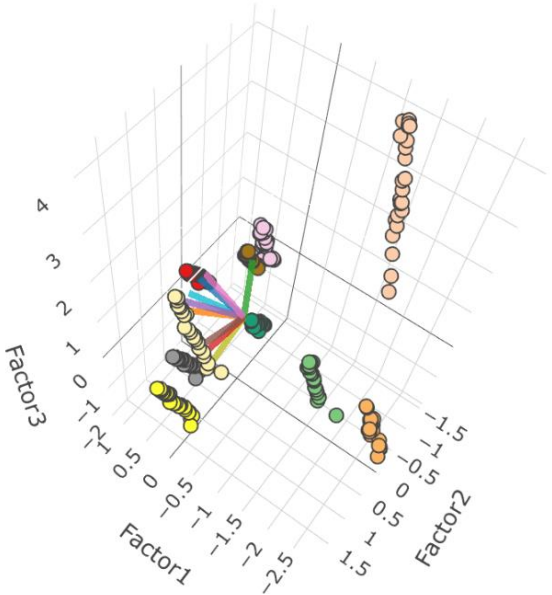


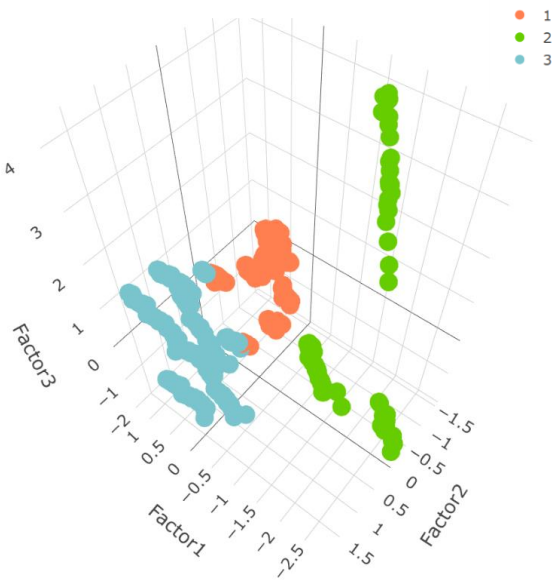
해당 시각화는 다음과 같이 해석할 수 있다.

- 각 시점의 pop_cd가 해당 변수에 중요한 요인에 영향을 많이 받았다
- Y15(10대 남성)이 monthly_sbk_loan(저축은행업종 총 대출금액) 변수에 중요한 factor3에 영향을 많이 받았다.
- Y30 35(30대 남여) / Y40 45(40대 남여) – bank_loan , mortgage_loan에 영향을 많이 받았다.

* Rmd 파일로 보시면 그래프를 움직이면서 데이터를 확인해보실 수 있습니다.

1. 전국민 카드 및 대출 이용통계 데이터 분석 (credit.csv)

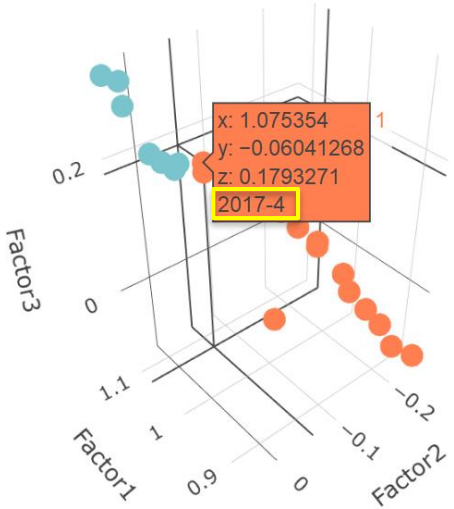


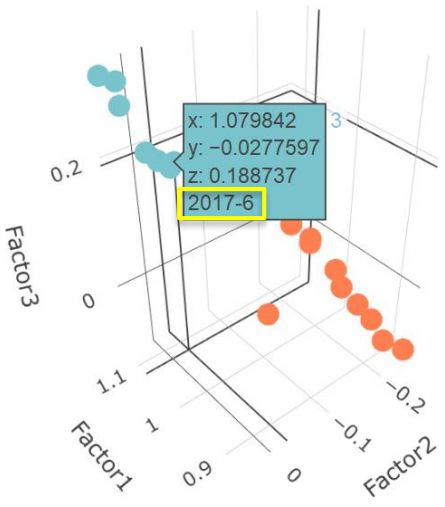


pop_cd의 금융 스타일(클러스터)를 factor analysis 결과에 중첩해보았는데,
시간대에 따라 클러스터가 변하는 pop_cd를 찾을 수 있었다.

* Rmd 파일로 보시면 그래프를 움직이면서 데이터를 확인해보실 수 있습니다.

1. 전국민 카드 및 대출 이용통계 데이터 분석 (credit.csv)

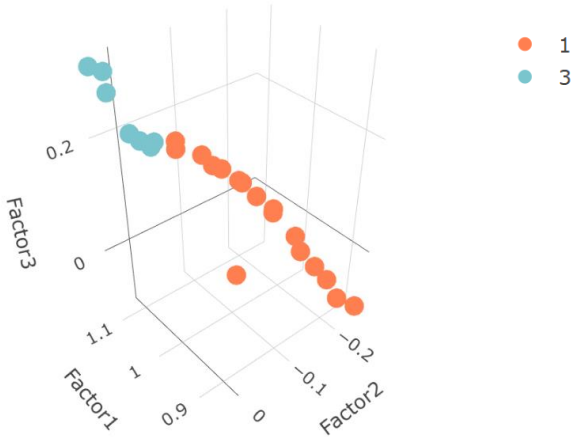




그 중, **60대 여성.남성의 경우, 시간이 지남에 따라서 클러스터1(제1금융권 사용 및 담보대출이 존재하는 스타일)에서 클러스터3(제2금융권 주로 사용 및 분할상환비율이 높은 스타일)로 변화**한다는 사실을 알 수 있었다.

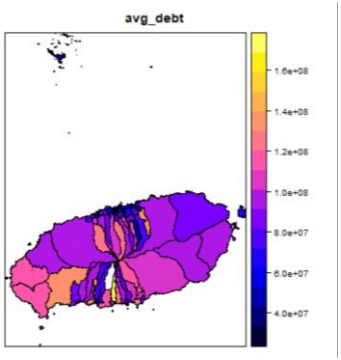
* Rmd 파일로 보시면 그래프를 움직이면서 데이터를 확인해보실 수 있습니다.

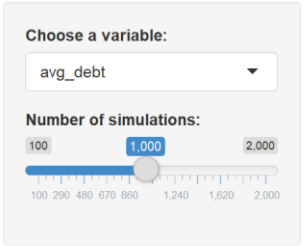
1. 전국민 카드 및 대출 이용통계 데이터 분석 (credit.csv)



- 이를 해석해보자면 정년을 맞이하는 고객들이 직장을 그만두게 되면서 금융스타일의 형태가 주로 큰 금액을 대출하는(클러스터 1) 금융스타일에서 비교적 적은 금액을 다양한 곳에서 대출하는(클러스터 3) 금융스타일로 변화했다고 할 수 있다.
- 이렇게 고객층 별로 각각 시각화를 해보면 그 **고객층의 금융스타일 변화를 Factor analysis의 잠재변수를 통해서** 알 수 있다.
- 60대 고객층 외에도 시간이 지남에 따라 클러스터의 카테고리는 달라지지 않지만 그 안에서 금융스타일의 변화가 일어나는 고객층을 위의 분석을 통해 파악할 수 있을 것이다. 또한, **이를 개별 고객 데이터에 적용한다면 새로운 고객군을 타겟팅한 마케팅을 진행하거나 맞춤 상품 제안**을 할 수 있을 것이다.

2. 제주도 금융라이프 데이터 분석(jeju_financial_life_data.csv)






```
        Monte-Carlo simulation of Moran I

data:  jeju_no_na_emd_sp_without_no_neighbor_region@data[, input$va
riable] %>%
weights: nb2listw(jeju_nb_without_no_neighbor_region)
number of simulations + 1: 1000
    as.numeric()
weights: nb2listw(jeju_nb_without_no_neighbor_region)
number of simulations + 1: 1000

statistic = 0.36389, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater
```

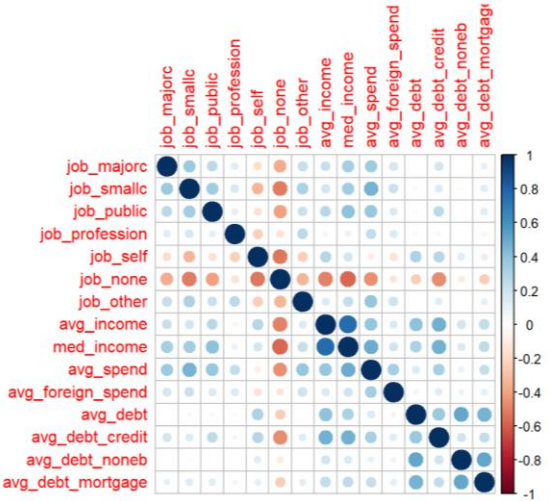
지역적으로 수치적인 차이는 있었지만, **통계적으로 분석을 진행하는 것이 유의미한지 파악**하기 위하여 **지역적 자기상관성 테스트를 진행**하였다.

p-value가 0.05보다 작으면, 지역적 자기 상관성이 존재한다는 의미이고, 0.05보다 크면 지역적 자기 상관성이 존재하지 않는다는 의미이다. 분석 결과, **모든 변수에서 지역적 자기 상관성이 대부분 존재**하였다. 이는 즉, 특정 지역에 각 변수들의 값이 몰려 있거나 특정 지역에 한정 되어 있다고 할 수 있다. 이를 통해 시각화를 통한 수치적 차이가 의미 있는 결과라는 정당성을 얻을 수 있었다.

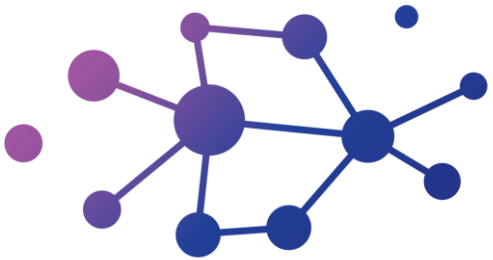
실제로 위에서 각 변수들을 지도 위에 시각화 하였을 때 뚜렷한 차이를 보였다. 이를 통해 각 변수들이 어느 지역에서 높은지, 어느 지역에 몰려있는 지 파악할 수 있다.
또한, **변수 별로 지도 시각화를 해본 결과 몇몇 변수들이 유사한 패턴**을 보이는 것을 볼 수 있었다.
이를 통계적으로 규명하기 위하여 상관관계를 볼 수 있는 correlation plot을 그린 후, 확인해 보았다.

* Rmd 파일로 보시면 변수를 선택하여 그래프를 보실 수 있습니다.

2. 제주도 금융라이프 데이터 분석(jeju_financial_life_data.csv)



평균 연소득과 평균 채무 보유액이 양의 상관관계를 보이는 것으로 보아, 두 변수는 비슷한 지역에서 유사한 패턴을 보인다고 할 수 있다. 두 변수의 관계를 해석해 본다면 소득이 많다고 해서 채무가 없는 것이 아니라, 오히려 소득이 있는 사람들이 빚을 더 많이 지는 경향이 있다고 해석할 수 있다. 이와 같은 과정을 통하여 **유사한 패턴을 보이는 변수들을 규명**할 수 있고 이 변수들 사이의 **상관성이 높은 지역을 탐구하여 지역적 특징과 이들의 금융 스타일을 파악**할 수 있다.



감사합니다