차차차!!!

외제차 왜 이렇게 비싸냐

김석준
김성표
임진혁
이세환

목차

개요

조원 소개

김석준

PCA 활용

김성표

클러스터링 실행

임진혁

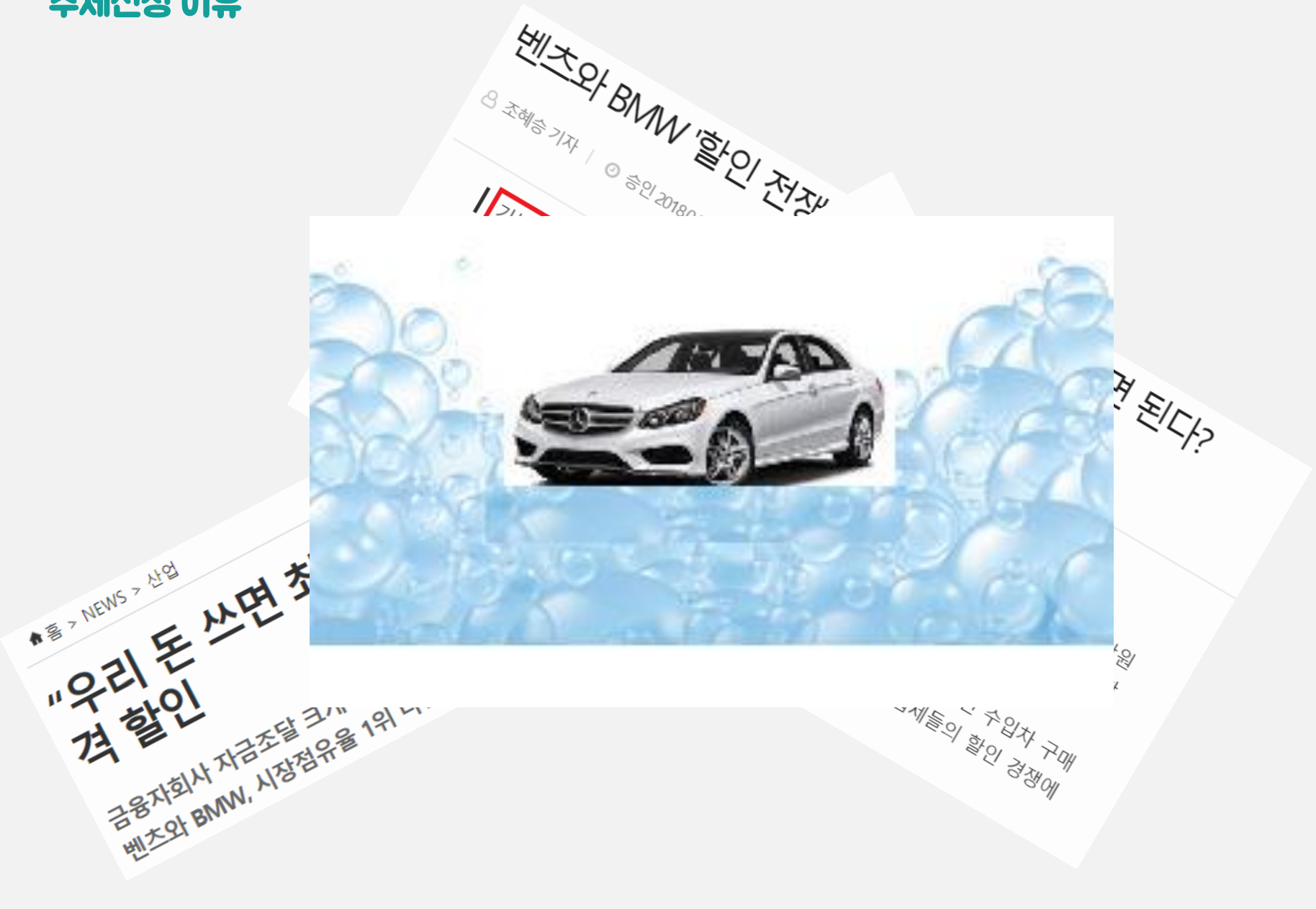
국산차와 외제차의 가격예측 모델

이세환

데이터 전처리 및 ppt제작

개요

주제선정 이유


벤츠와 BMW '할인 전쟁
조혜승 기자
홈 > NEWS > 산업
"우리 돈 쓰면 최
격 할인
금융자회사 자금조달 크
벤츠와 BMW, 시장점유율 1위
면 된다?
수입차 구매
제들의 할인 경쟁에

개요

주제선정 이유

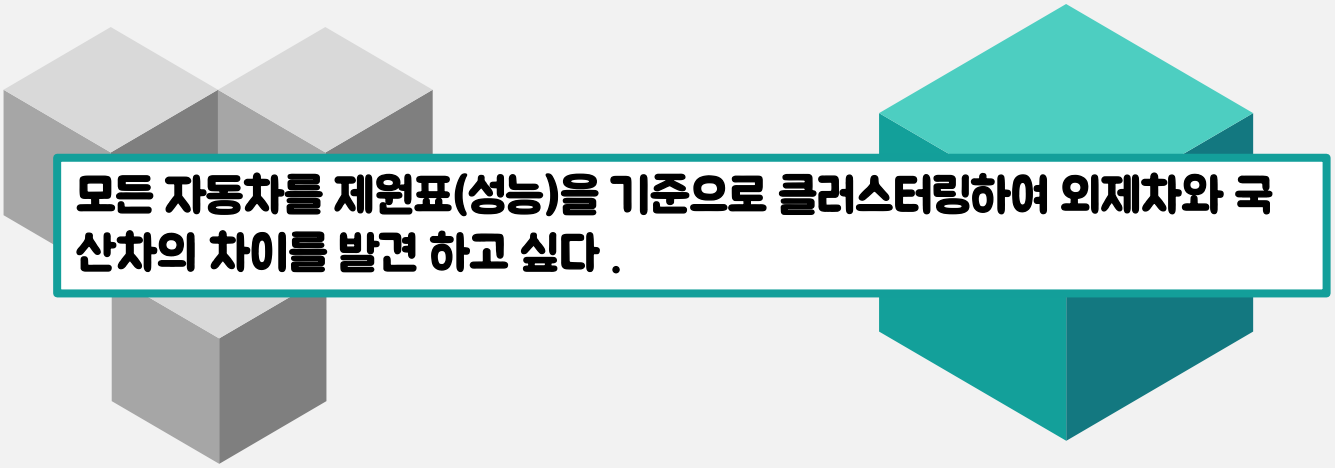

이렇게 할인을??
01
02
이렇게 할인해준다는건
그래도 남는 계산이라는거
아니냐?
왜 이렇게 비싼거지
외제차는?
03
04
외제차가 국산차 보다 뭐
가 잘나서 이리 비싼거냐?
두 그룹이 차이가 있는지
있다면 무엇인지!!
05

개요

초기 분석 목표

문제 정의

해결 방안



모든 자동차를 제원표(성능)을 기준으로 클러스터링하여 외제차와 국산차의 차이를 발견 하고 싶다 .

외제차와 국산차의 가격 차이를 결정짓는 요인은 무엇일까?

성능이 비슷하다면 가격의 차이는 디자인으로 결정되는 것이 아닐까?

제원이 차이가 확연히 난다. -> 비싼 값의 원인: 이유가 있었다.

제원이 차이가 별로 안난다. -> 1)제원에 없는 것이 영향을 준다.
2)괜히 비싸다.

데이터 소개 및 분석 기법

데이터 설명

- 데이터 수집 방법 : www.enuri.com 직접 수집
- 데이터 종류 : 국산 52종 외제 105종 총 157대, 17개사의 차량 데이터
- 변수 설명 (총 22개의 변수)
  - 차종 : 차량 모델 이름
  - 브랜드 : 제조사
  - 판매가 : 국내 판매가 (부가세 및 옵션 미포함)
  - 공식 연비 : 제조사 제공 공식 연비 / 1리터 또는 1갤런으로 갈 수 있는 거리
  - 배기량 : 제조사 제공 공식 배기량 / 실린더의 부피
  - 최고출력 : 제조사 제공 최고 마력
  - 최대토크 : 엔진의 회전력을 나타내는 수치
  - 엔진형식 : 엔진의 종류
  - 미션형식1 : 자동 or 수동
  - 미션형식2 : 기어 단수
  - 연료탱크 : 총 연료 탱크 량
  - 구동방식 : 가솔린 or 경유
  - 공차중량 :
  - 길이, 폭, 높이, 부피, 면적, 승차인원, 도어수
  - 타이어 : 타이어의 크기
  - 나라 : 국산 or 외제

데이터 소개 및 분석 기법

데이터 설명

-헷갈릴 수 있는 변수 개념

마력

주어진 '시간' 안에 할 수 있는 일의 총량

토크

순간적으로 발휘할 수 있는 '힘'의 총량

엔진 형식(좌 v자 / 우 L자 형식)

데이터 소개 및 분석 기법

데이터 설명

- 헷갈릴 수 있는 변수 개념

배기량 -> 엔진 실린더 내부의 체적
쉽게 말해 엔진의 크기. -> 높은 배기량->엔진의 폭력/출력이 강함

배기량이 높으면 그만큼 환경오염도가 높음 따라서 배기량에 따라 세금을 차등 부여

<1,000cc 경차
<1,600cc 소형차
<2,000cc 중형차
그 이상 – 대형차.

폭/너비/부피 등

데이터 소개 및 분석 기법

분석기법- 뭘 어떻게 할 것인가!!

PCA : PCAMixdata 패키지를 사용
factor변수를 포함한 PCA를 실행할 수 있게 해주는 패키지

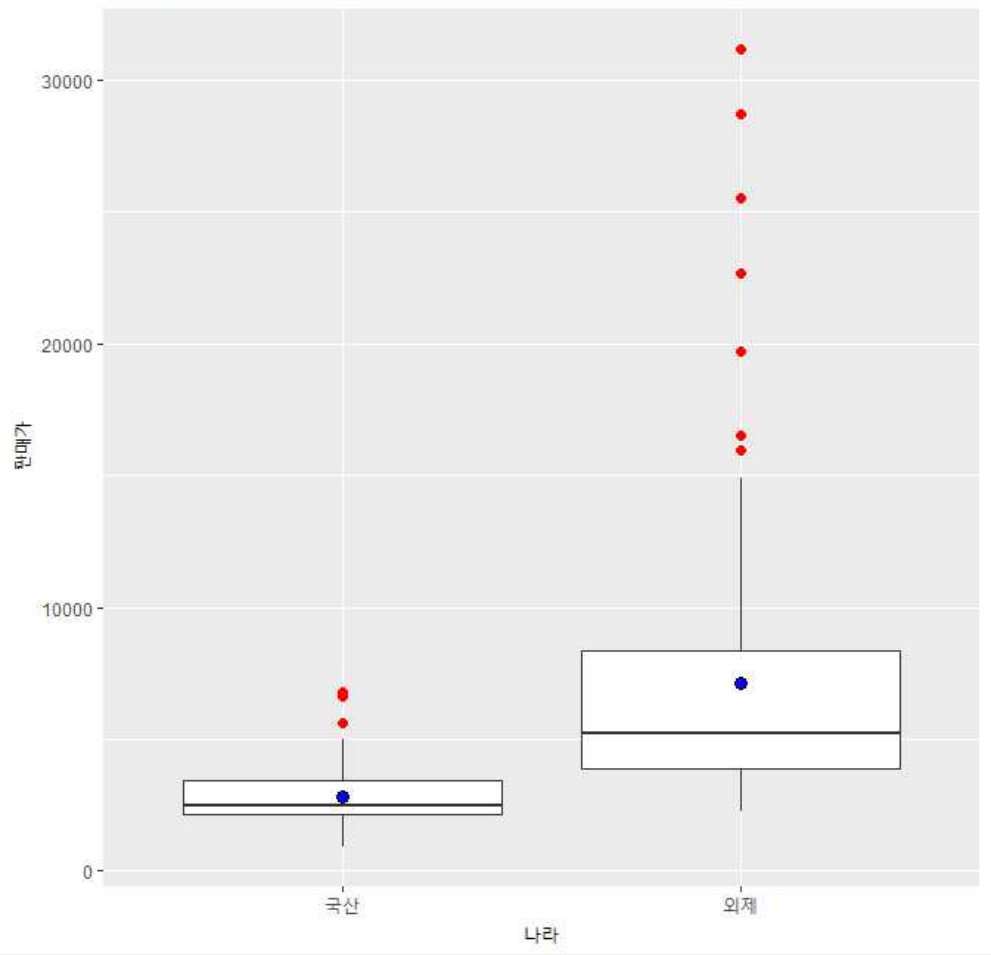
제원을 pca로 표현하여 데이터를 평면에 시각화하고 외제차와 국산차에 큰 차이가 있는지 확인

클러스터링 : K-means

가격 예측 모델 : 국산차의 가격을 예측해주는 모델을 만들고 그 모델에
외제차의 스펙을 넣어주면 어느 정도의 차이가 날까?

데이터 소개 및 분석 기법

국산 및 외제차의 가격 비교



단위 : 만 원

외제차 가격 요약

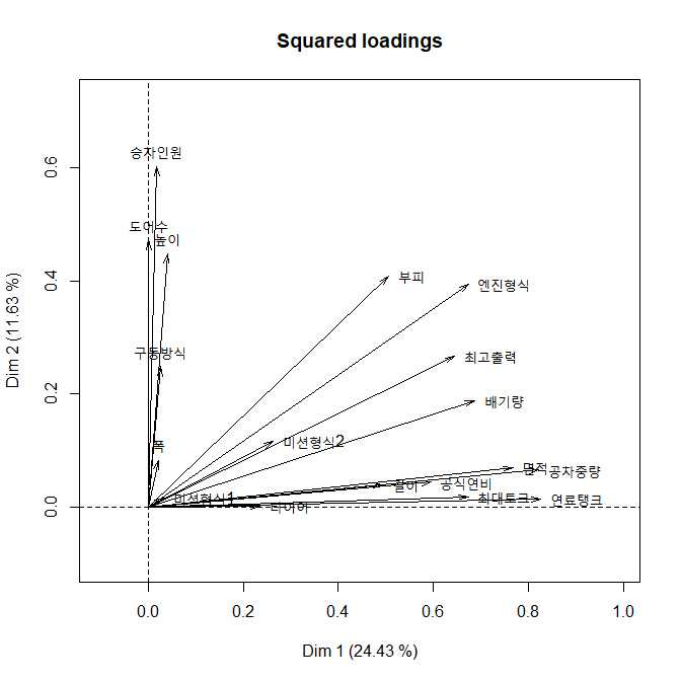
최대값 : 31,200
평균값 : 7,102
최소값 : 2,260

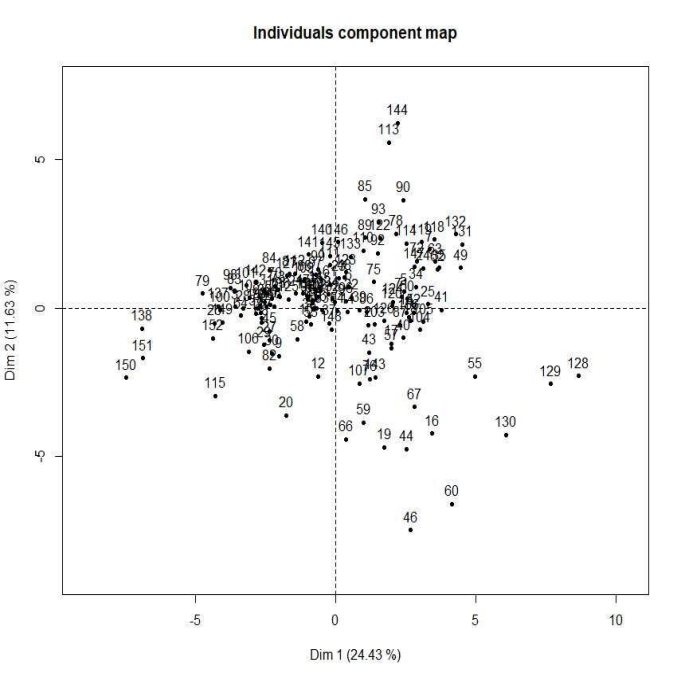
국산차 가격 요약

최대값 : 6,798
평균값 : 2,819
최소값 : 908

분석 결과 및 해석

분석 결과 시각화(PCA)



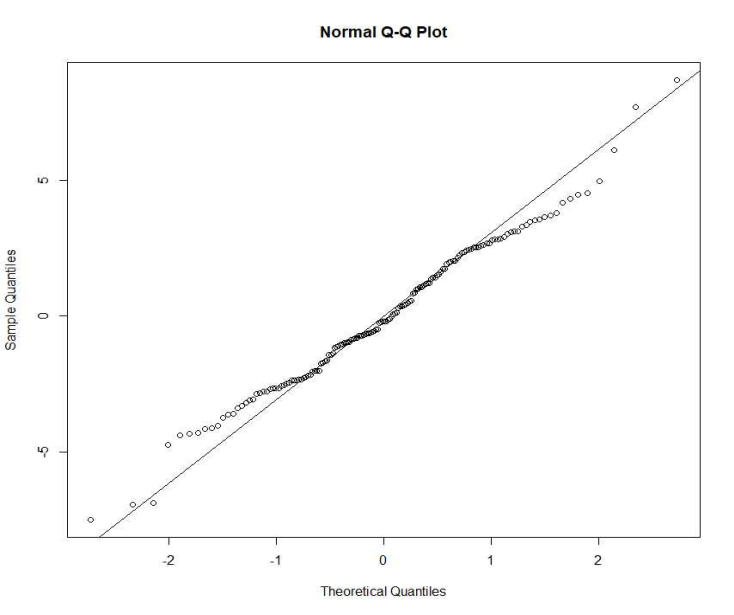


각 변수가 축소된 변수에 미친 영향 확인

2개의 변수로 차원축소 된 공간 내에서 데이터 분포

분석 결과 및 해석

분석 결과 검정 (정규성 검정)



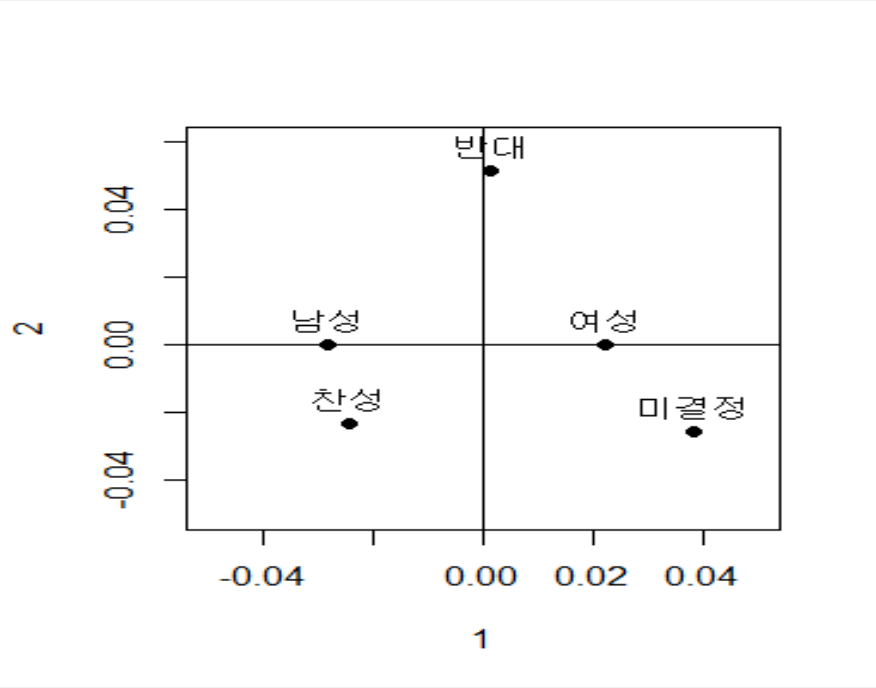

```
        Shapiro-Wilk normality test

data:  pca$ind$coord[, 1]
W = 0.98866, p-value = 0.2401
```

데이터들이 정규성을
적절히 따라가는 것을 알 수 있다.

귀무가설 : 데이터는 정규분포이다
대립가설 : 데이터는 정규분포가 아니다.
P-value가 0.05보다 크므로 귀무가설을 기각할 수 없다.

PCAMIX 설명



다중 대응 일치분석(MCA)

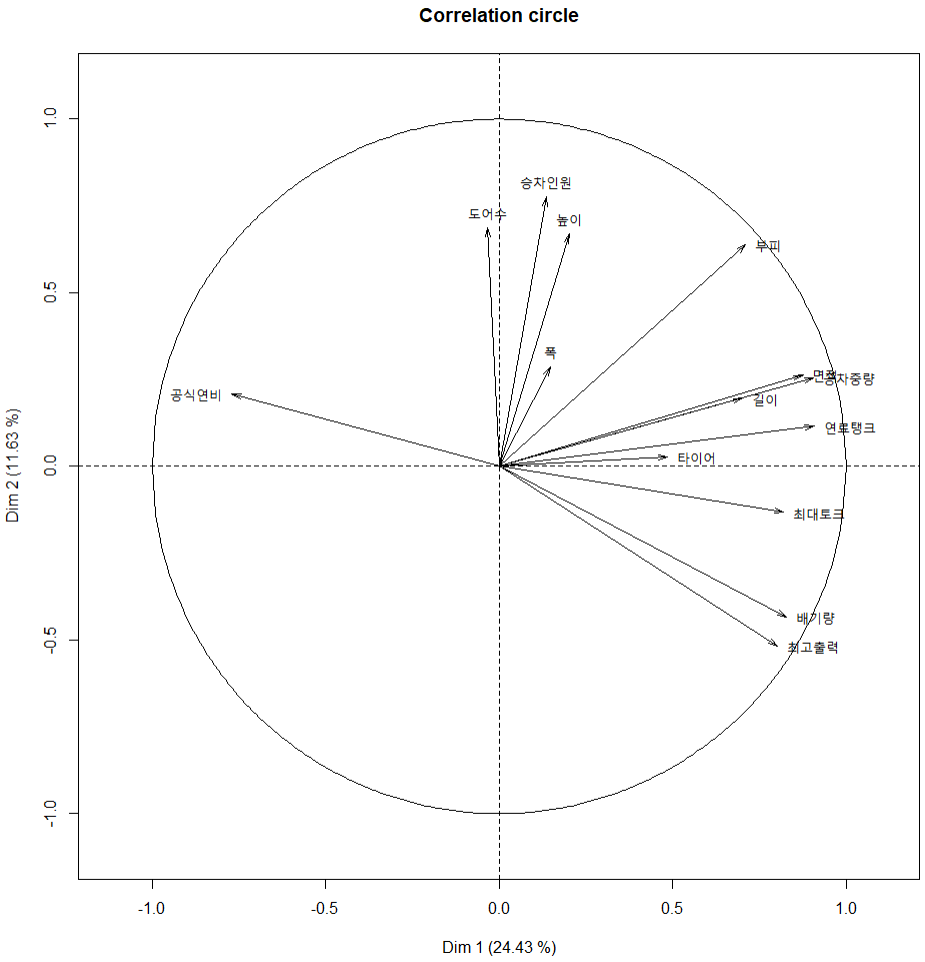
범주형 변수를 저차원에 뿌려주어 변수 간의 연관을 봄.

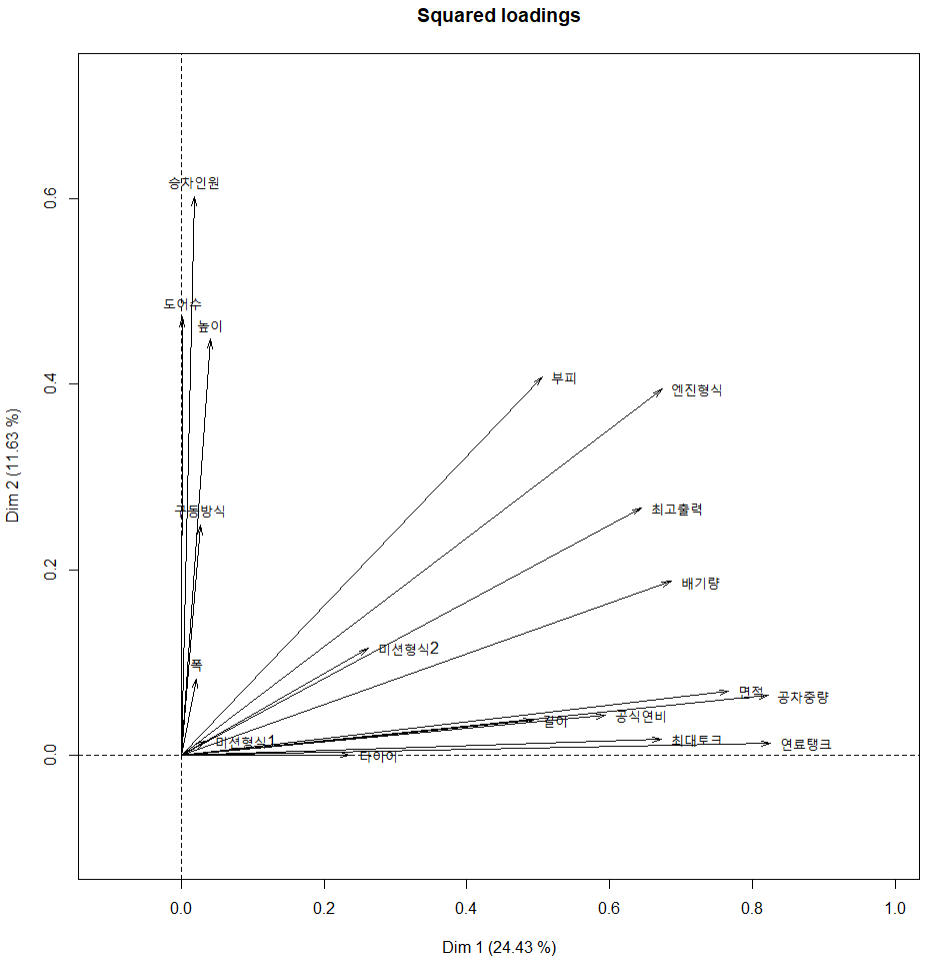
질적 및 양적 변수가 혼합 된 개인 또는 집합에 대한 주성분 분석을 수행한다.
PCAmix는 특수한 경우로서 일반적인 주성분 분석 (PCA)과 다중 대응 분석 (MCA)을 포함한다.

-> pc1의 전체변수에 대한 설명력이 떨어짐..

분석 결과 및 해석

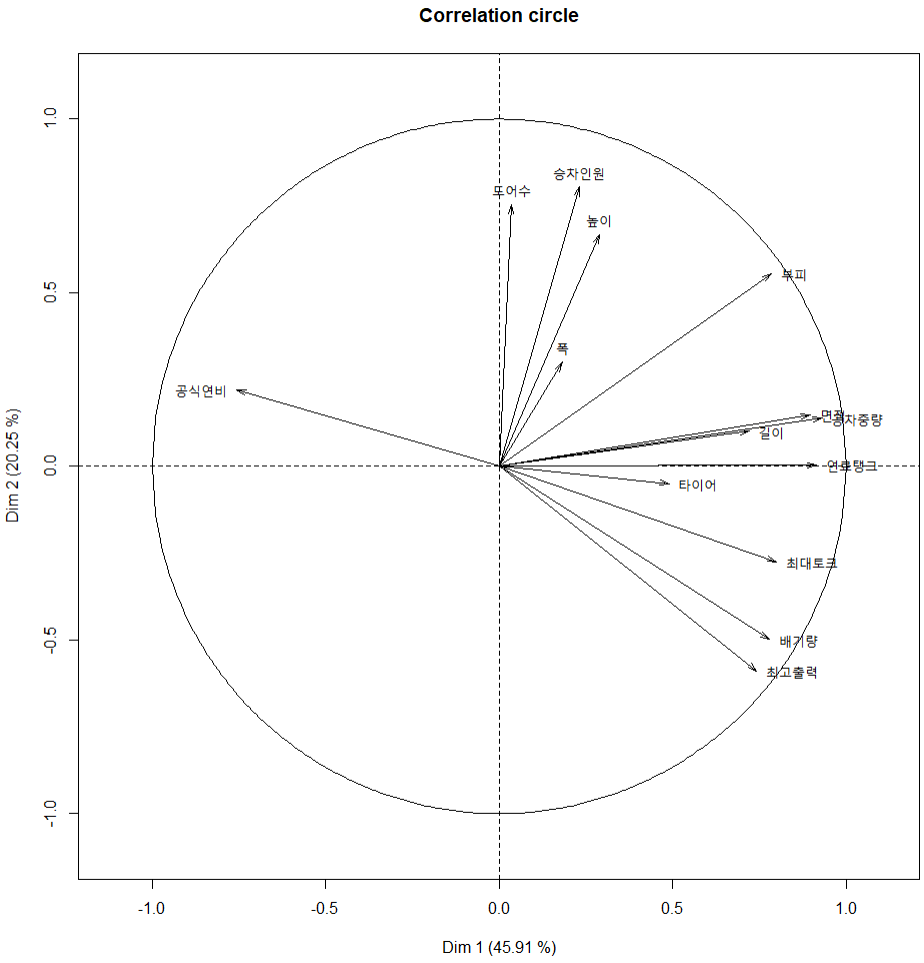
분석 결과 시각화(PCA)

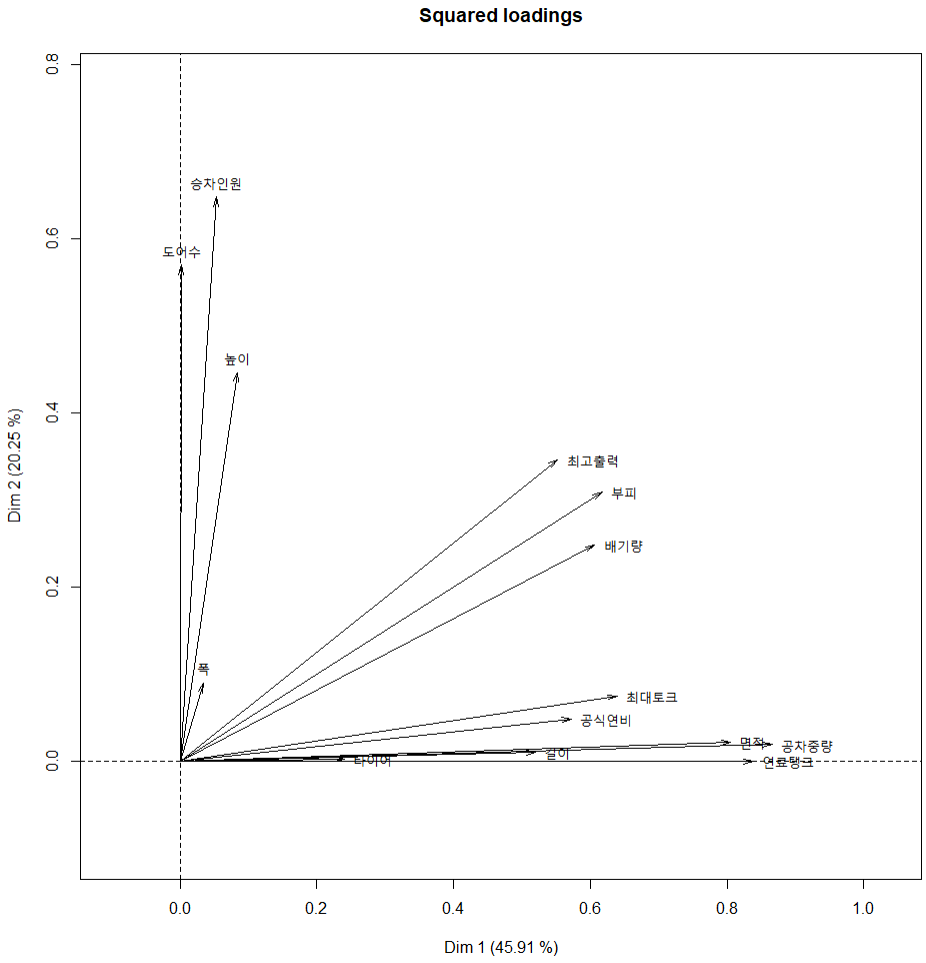




분석 결과 및 해석

분석 결과 시각화(PCA)

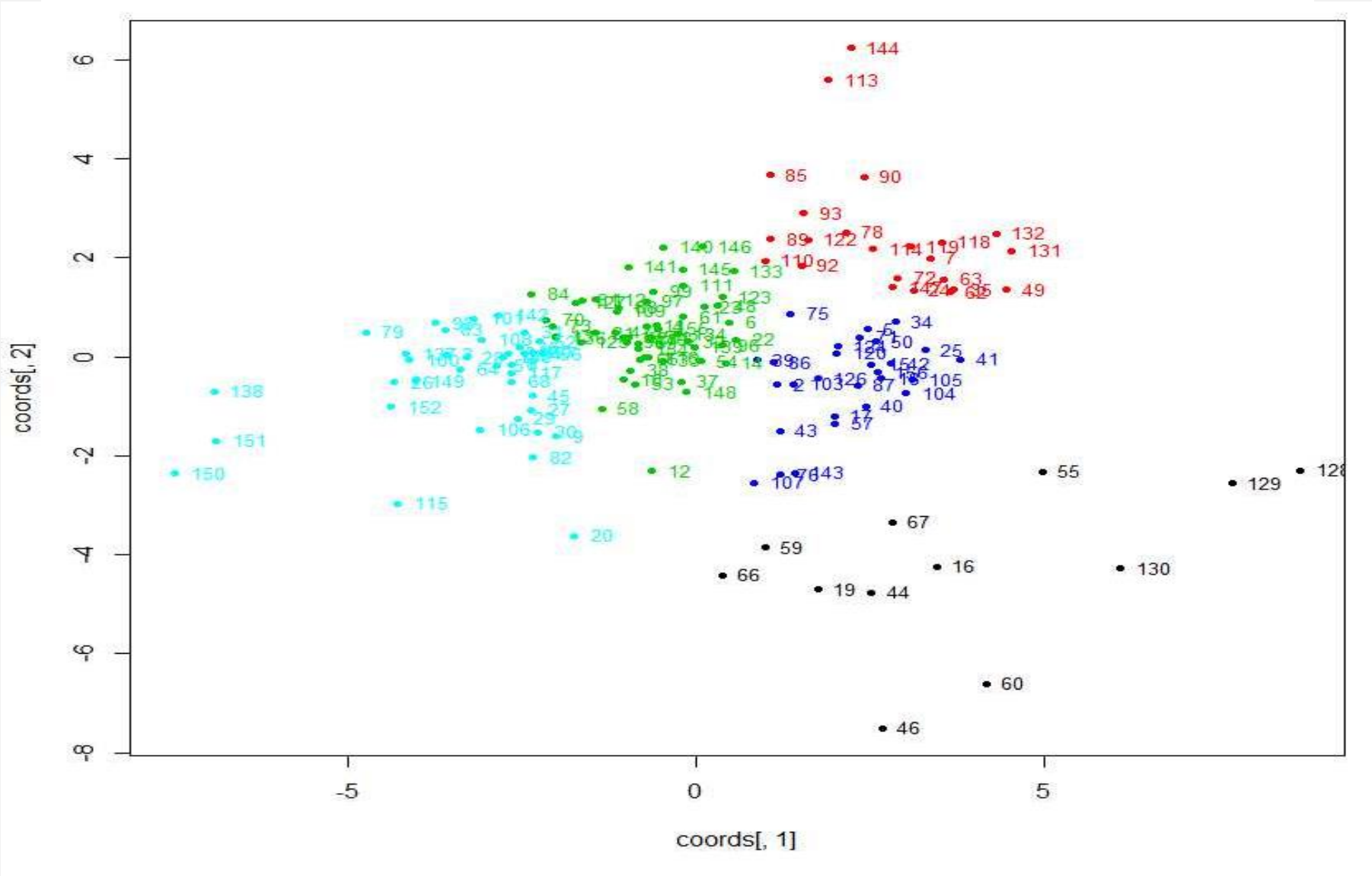




확실히 수치값만 넣은 pca가 좀 더 분산 설명력이 높음
하지만 장점이 있으니..!! 이렇게 굳이 한 의미가 있으니!!

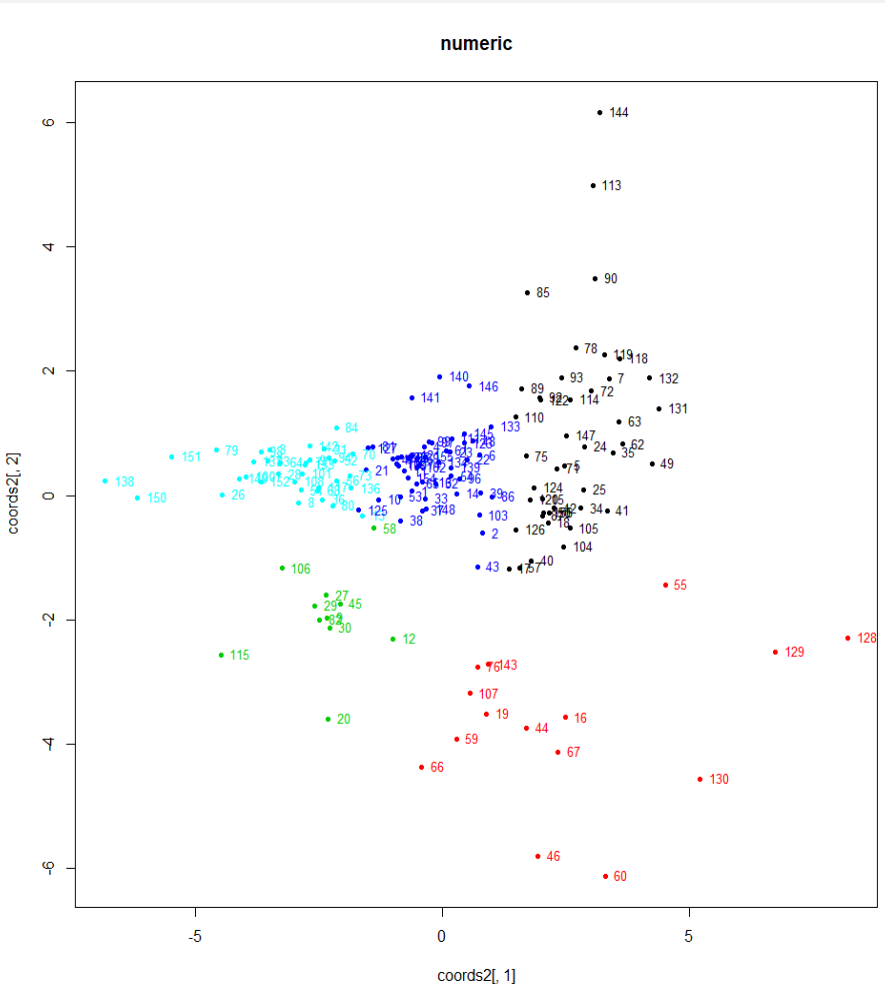
분석 결과 및 해석

분석 결과 시각화 (각 클러스터 특징 설명)



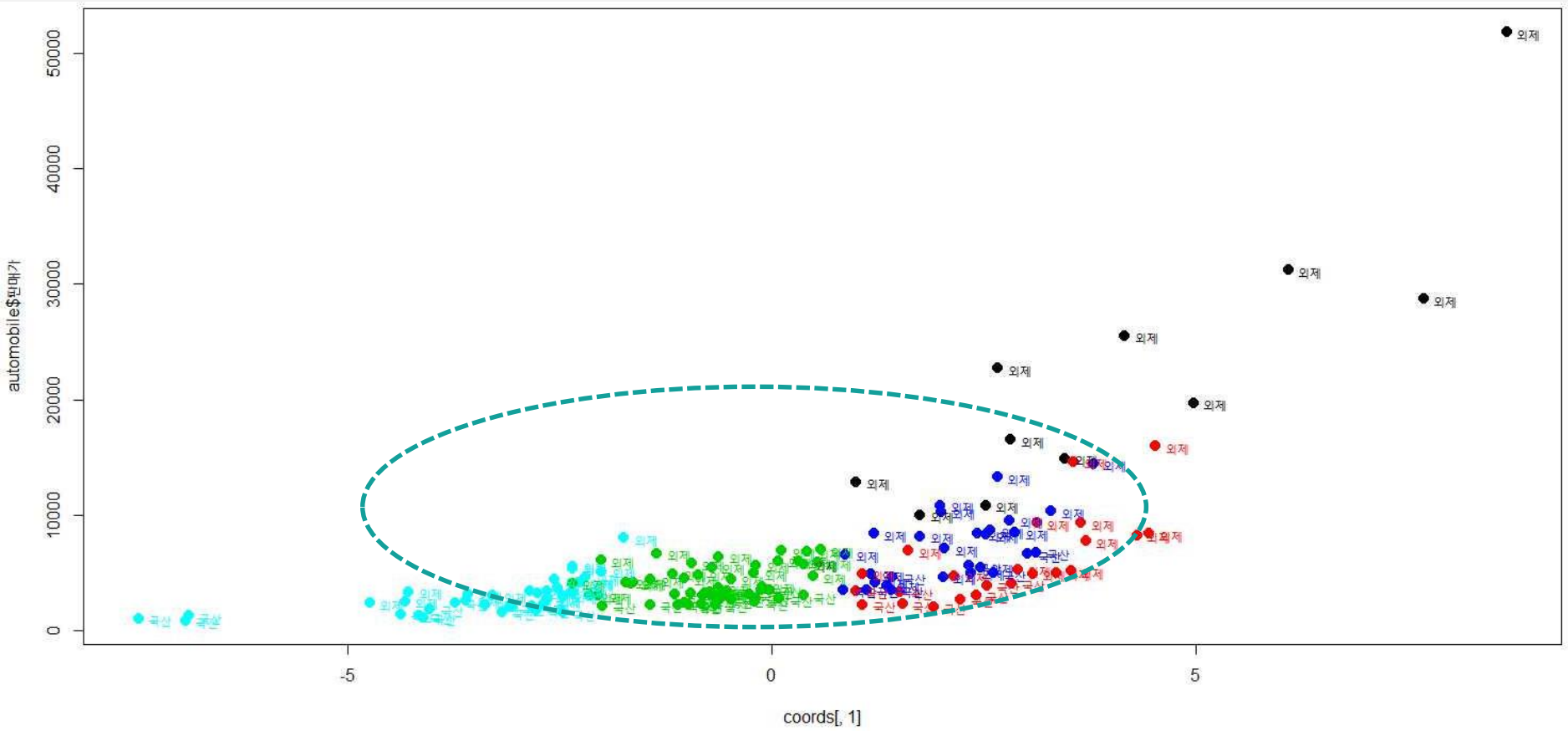
분석 결과 및 해석

분석 결과 시각화 (각 클러스터)



분석 결과 및 해석

분석 결과 시각화 (군집내 가격 비교)



제원을 차원축소한 축을 기준으로 클러스터 내에서의 국산과 외제차 간의 판매가 비교
같은 제원(클러스터일 경우) 외제차가 가격이 더 높이 형성되어있음

분석 결과 및 해석

활용 방안 (기본적인 자동차 추천 모델 구축)

다항 로지스틱 회귀 활용

- Multinom 함수를 통해 다항로지스틱 회귀로 추천 시스템 개발
- 반응변수를 클러스터로 나머지 제원을 설명변수로 설정하였다.

```
m <- multinom(cluster ~ ., data = train)

predict(m, newdata = test, type = "class") # 예측데이터가 속하는 군집
predict(m, newdata = test, type = "probs") # 예측데이터가 해당되는 군집에 속할 확률

predicted <- predict(m, newdata = test)
sum(predicted == test$cluster)/NROW(test) # 정확도 측정
```

- 성능 검정

```
> sum(predicted == test$cluster)/NROW(test) # 정확도 측정
[1] 0.7954545
> xtabs(~predicted + test$cluster) # 혼돈행렬
          test$cluster
predicted  1  2  3  4  5
        1  2  0  0  1  0
        2  0  6  0  0  1
        3  0  0 11  1  0
        4  0  0  3  6  0
        5  1  0  2  0 10
>
```

추천 시스템

다항 로지스틱 회귀를 활용한 자동차 추천 시스템

원하는 차량의 제원값을 넣으면, 가장 비슷한 차종들이 분류되어있는 군집단 번호를 제시

```
> # 가장 싸고, 평균적으로 성능을 내는, 그리고 브랜드는 기아(국산)꺼를 선호합니다 이에 맞는 차를 추천해주세요
> predict(m, newdata = data.frame(브랜드='기아',판매가=min(car$판매가),공식연비=mean(car$공식연비),배기량=mean(car$배기량),
+                                 최고출력=mean(car$최고출력),최대토크=mean(car$최대토크),엔진형식='직렬4',미션형식1='자동',미션형식2='5단',
+                                 연료랭크=mean(car$연료랭크),구동방식='가솔린',공차중량=mean(car$공차중량),길이=mean(car$길이),폭=mean(car$폭),높이=mean(car$높이),
 부피=mean(car$부피),면적=mean(car$면적),
+                                 승차인원=mean(car$승차인원),도어수=mean(car$도어수),
+                                 타이어=20,나라='국산'), type = "class")
[1] 1
Levels: 1 2 3 4 5
```

군집 내에서 가장 적절한 차량 몇 대를 추천

```
          차종 브랜드 판매가
149     스팅어    기아   3500
150     스토닉    기아   1910
151        모닝    기아   1075
152        레이    기아   1281
153  프라이드    기아   1456
>
```

가격 예측 모델

먼저 회귀식으로 !!

```
Call:
lm(formula = 판매가 ~ . - 차종, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-3400.4  -562.1   -53.5   577.5  5095.9

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.694e+03  4.951e+03   0.342  0.73288
브랜드기아     -2.528e+03  9.908e+02  -2.552  0.01210 *
브랜드닛산     -7.274e+02  1.220e+03  -0.596  0.55210
브랜드도요타   -1.371e+03  1.126e+03  -1.217  0.22605
브랜드랜드로버  2.321e+03  1.148e+03   2.021  0.04573 *
브랜드벤츠      2.219e+03  9.776e+02   2.270  0.02516 *
브랜드벤틀리    3.031e+04  2.847e+03  10.648  < 2e-16 ***
브랜드볼보     -7.748e+02  9.284e+02  -0.835  0.40578
브랜드삼성     -9.755e+02  1.046e+03  -0.932  0.35325
브랜드쉐보레K  -1.859e+03  1.315e+03  -1.413  0.16036
브랜드쌍용     -1.785e+03  9.064e+02  -1.969  0.05149 .
브랜드아우디    1.293e+02  7.102e+02   0.182  0.85588
브랜드포드     -1.927e+03  1.061e+03  -1.816  0.07205 .
브랜드폭스바겐 -3.163e+02  8.388e+02  -0.377  0.70685
브랜드푸조      1.470e+02  9.394e+02   0.156  0.87598
브랜드현대     -1.779e+03  7.903e+02  -2.252  0.02634 *
브랜드혼다     -1.140e+03  1.125e+03  -1.013  0.31316
공식연비        3.506e+02  1.086e+02   3.227  0.00165 **
배기량          4.152e-02  6.691e-01   0.062  0.95063
최고출력        1.741e+01  7.102e+00   2.452  0.01579 *
최대토크        8.402e+01  4.396e+01   1.911  0.05857 .
엔진형식V10     8.718e+03  2.726e+03   3.199  0.00180 **
엔진형식v6     -1.235e+03  2.070e+03  -0.597  0.55205
엔진형식V6     -1.084e+03  1.241e+03  -0.874  0.38421
엔진형식V8      2.638e+03  1.928e+03   1.369  0.17390
엔진형식W12    -1.862e+04  3.192e+03  -5.833 5.56e-08 ***
엔진형식직렬3  -1.131e+03  2.287e+03  -0.495  0.62185
엔진형식직렬4  -4.889e+02  1.070e+03  -0.457  0.64868
엔진형식직렬5  -6.300e+02  1.549e+03  -0.407  0.68502
엔진형식직렬6   3.038e+02  1.352e+03   0.225  0.82264
미션형식1자동   5.378e+02  5.891e+02   0.913  0.36330
미션형식25단   -2.421e+03  1.897e+03  -1.276  0.20451
미션형식26단   -2.253e+03  1.783e+03  -1.263  0.20911
미션형식27단   -2.645e+03  1.852e+03  -1.428  0.15617
미션형식28단   -2.330e+03  1.870e+03  -1.246  0.21543
미션형식2무단  -3.021e+03  1.954e+03  -1.546  0.12488
연료탱크        6.872e+01  2.961e+01   2.321  0.02213 *
구동방식경유   -2.390e+03  9.134e+02  -2.620  0.01003 *
```

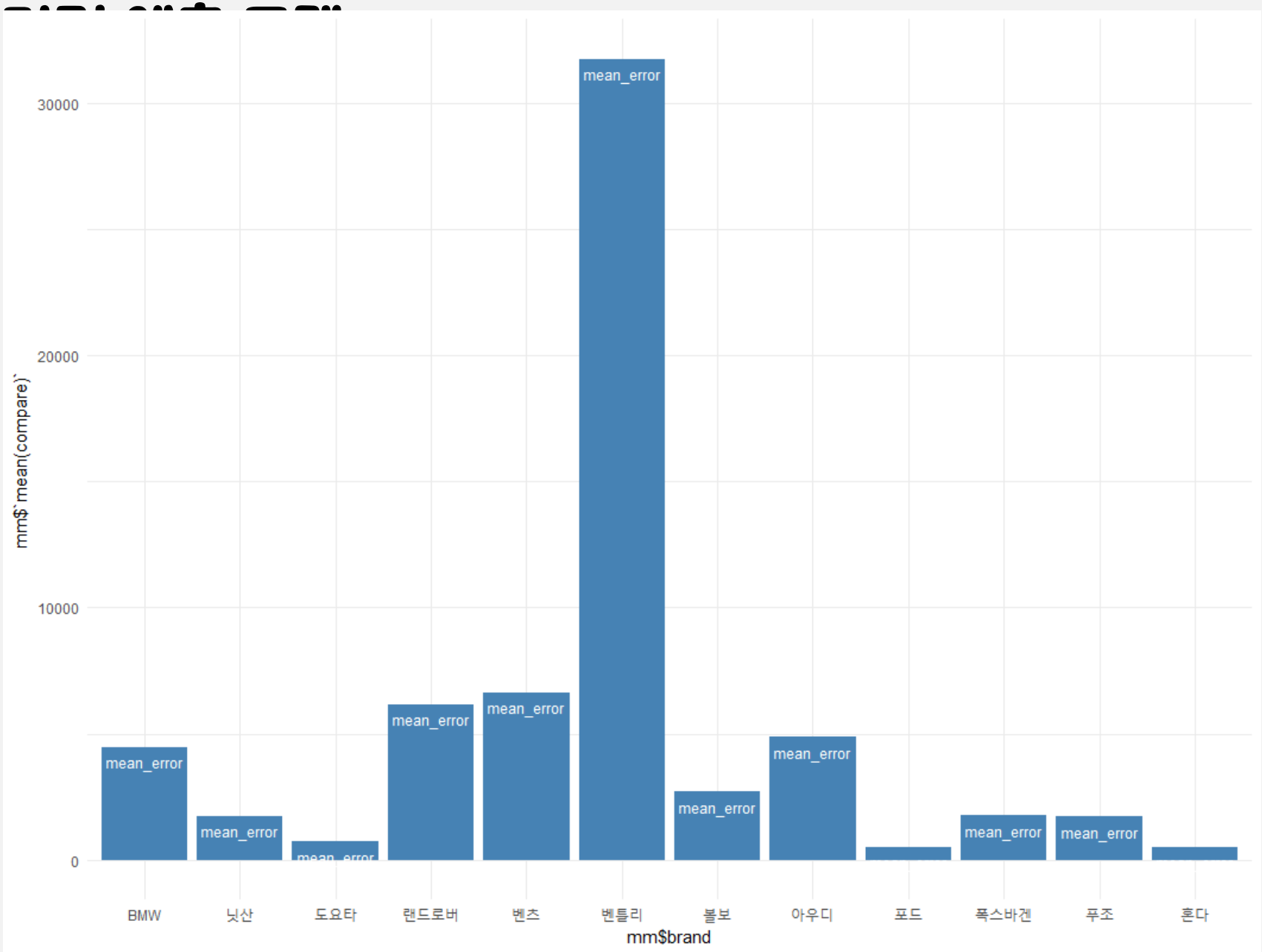
```
> summary(fit3)

Call:
lm(formula = 판매가 ~ . - 차종 - 브랜드 - 엔진형식 - 미션형식1 -
    미션형식2 - 길이 - 높이, data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-6040.6 -1293.6   228.7  1322.2 14247.2

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.065e+03  3.919e+03  -2.058 0.041446 *
공식연비       6.689e+02  1.493e+02   4.480 1.52e-05 ***
배기량         2.597e+00  6.036e-01   4.303 3.12e-05 ***
최고출력      -1.172e+01  8.062e+00  -1.453 0.148348
최대토크       3.576e+02  5.137e+01   6.961 1.15e-10 ***
연료탱크       4.823e+01  4.123e+01   1.170 0.244082
구동방식경유 -6.395e+03  1.012e+03  -6.317 3.23e-09 ***
공차중량       3.777e+00  2.099e+00   1.799 0.074080 .
폭            -4.292e-02  3.091e-01  -0.139 0.889768
부피           8.064e+01  2.545e+02   0.317 0.751819
면적          -2.112e+03  5.994e+02  -3.523 0.000575 ***
승차인원      -6.388e+01  2.856e+02  -0.224 0.823313
도어수        -3.266e+01  4.021e+02  -0.081 0.935389
타이어        -1.202e+01  8.101e+01  -0.148 0.882217
나라외제       6.722e+02  5.845e+02   1.150 0.252058
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2700 on 142 degrees of freedom
Multiple R-squared:  0.8207,    Adjusted R-squared:  0.803
F-statistic: 46.42 on 14 and 142 DF,  p-value: < 2.2e-16
```


mean_error
30000
20000
10000
0
mm$`mean(compare)`
BMW
닛산
도요타
랜드로버
벤츠
벤틀리
볼보
아우디
포드
폭스바겐
푸조
혼다
mm$brand

>

분석 결과 및 해석

분석 결과 해석

✓ 성능이 비슷한 차라도 외제차가 국산차에 비해 비싼 경향을 보인다.

✓ 단순히 정형 데이터로 확인할 수 있는 성능(제원)을 넘어서 디자인과 인지도 같은 비정형 요소가 가격에 영향을 미친다고 생각할 수 있다.

Thank you