알 프로젝트

정효인

궁금증

데이터 분석에 앞서 …
도대체 우리는 데이터 분석을 왜 해야하는 걸까

우리는 이 과정의 첫 수업부터 지금까지 약 12주 간..
5번의 시험을 쳤고, 6번의 프로젝트를 진행했으며 약 7권의 책을 끝냈다

하지만, 기초 데이터 베이스의 개념부터 프로그래밍, 통계학, 분석 툴을 배우고
프로젝트 주제 선정을 위해 데이터를 찾고 또 찾으며.. 분석을 위해 또다시 데이터와 씨름을
하면서도, 정작 데이터 분석을 해야하는 **본질적인 이유**에 대해 고민한 적은 없었다

데이터 분석 : 아이데이션 -> 주제 선정 -> 데이터 찾기 -> 데이터 전처리 -> 데이터 검증 -> 시각화

Why?

통찰

데이터 분석의 본질, 미래에 대한 답은 ~~과거에~~ 데이터에 있다

누가 조국의 미래를 묻거든
고개를 들어 관악을 보게 하라 - 서울대학교

과거는 미래를 보는 창
- 진관타오

미래에 대한 최고의 예언자는 과거이다
- 조지 고든 바이런

역사를 잊은 민족에게 미래는 없다
- 윈스턴 처칠

-

데이터 분석을 통해 미래를 들여다 볼 수 있다면 어떤 미래를 가장 보고싶을까?

올해 경제 성장률 -1.1% 전망, 코로나, n포 세대, 갈수록 암담한 취업 전망..

우리는 행복해질 수 있을까?

제가 선정한 이슈는 우리의 ‘행복’ 입니다.

우리의 해석

모두에게 다르게 적용되는 행복의 기준,
행복의 사전적 의미를 들여다 보니 …

-

행·복 幸福

: 사람이 생활 속에서 기쁘고 즐겁고
만족을 느끼는 상태에 있는 것

인간에게 있어서 인생의 궁극적인 목표는 행복이다.
- 플라톤 -

우리의 해석

모두에게 다르게 적용되는 행복의 기준,
행복의 사전적 의미를 들여다 보니 …

-

행·복 幸福

: 사람이 생활 속에서 기쁘고 즐겁고 -----▸ ①
만족을 느끼는 상태에 있는 것 -----▸ ②

} 행복의 두가지 조건

인간에게 있어서 인생의 궁극적인 목표는 행복이다.
- 플라톤 -

행복의 두가지 조건을
어떻게 수치화 할까

/

여가 지수

휴식을 겸한 다양한
취미활동이 포함되는
경제 활동 이외의 시간으로
개인이 처분할 수 있는
자유로운 시간 *여기서는
가계의 소비지출 총액에서
오락·문화 비용이
차지하는 비율로 정의했다

엥겔 지수

엥겔지수는 가계의 소비
지출 총액에서 식료품비가
차지하는 비율이다
식료품의 경우 일반적으로
소득의 크기와 상관없이
반드시 일정 부분 소비해야
하기 때문에 가계의 생활
수준을 가늠할 수 있다.

‘ 사람이 생활 속에서 기쁘고 즐겁고 , 만족을 느끼는 상태에 있는 것 ’

여가지수 = 사람이 생활 속에서 기쁘고 즐겁고

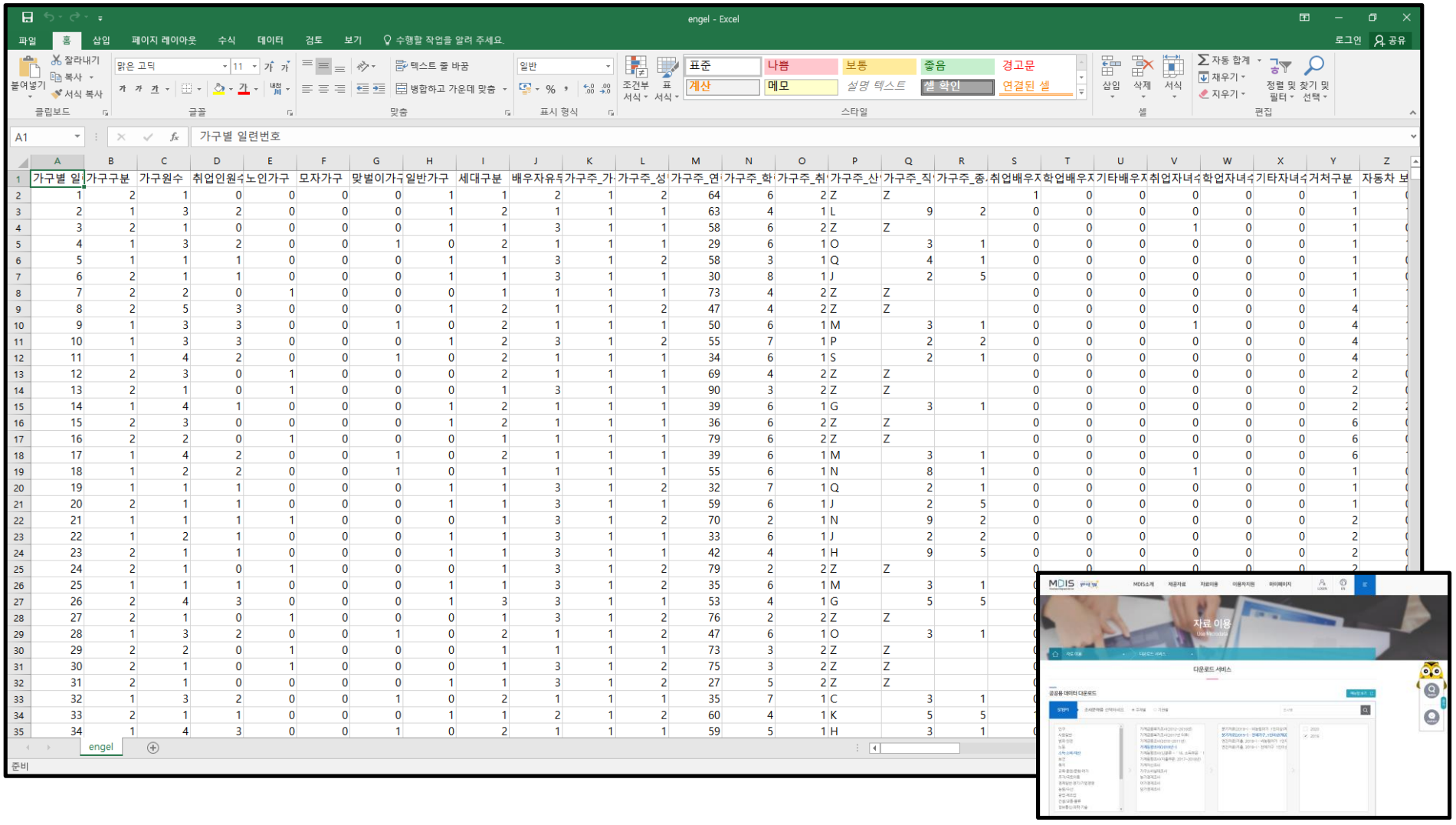
엥겔지수 = 만족을 느끼는 상태에 있는 것

$$\frac{\text{여가지수}}{\text{엥겔지수}} = \frac{\text{사람이 생활 속에서 기쁘고 즐겁고}}{\text{만족을 느끼는 상태에 있는 것}}$$

01 데이터 구축

가계동향조사 연간자료 – 지출부문(2019~)

데이터 크기: 9510x162

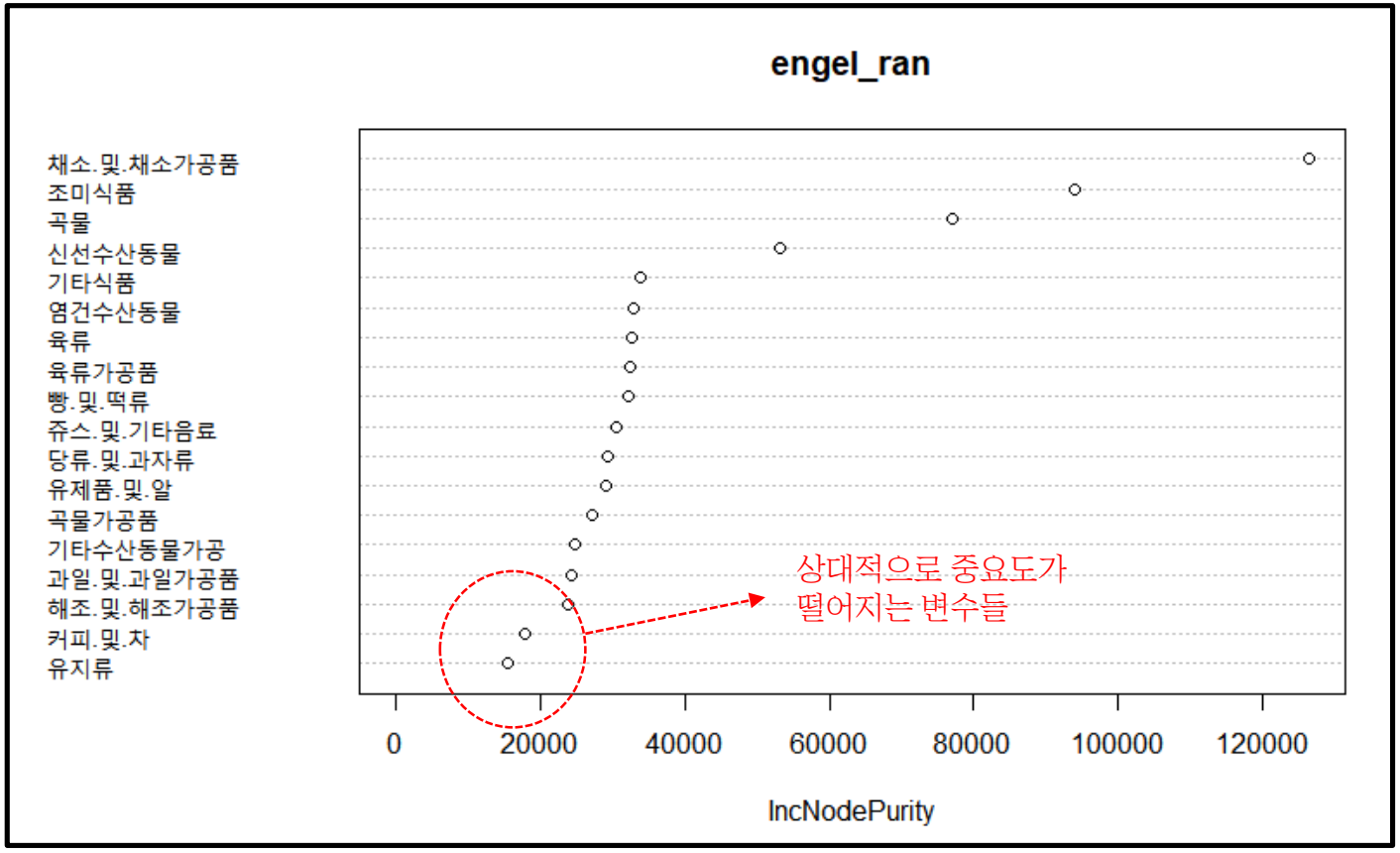


출처: 마이크로데이터 통합서비스 포털(MDIS)

02 데이터 정제

여가·엥겔 지수의 최적의 변수는 (1)

식료품비에 해당하는 하위 변수가 19가지이기 때문에 최적의 변수만을 찾기 위한 분류 진행



| 분류 방법 |

랜덤포레스트

02 데이터 정제

여가·엥겔 지수의 최적의 변수는 (1)

식료품비에 해당하는 하위 변수가 19가지이기 때문에 최적의 변수만을 찾기 위한 분류 진행

```
> engel_ran <- randomForest(formula = engel ~ 곡물 +
+                                 곡물가공품 +
+                                 빵.및.떡류 +
+                                 육류 +
+                                 육류가공품 +
+                                 신선수산동물 +
+                                 염건수산동물 +
+                                 기타수산동물가공 +
+                                 유제품.및.알 +
+                                 해조.및.해조가공품 +
+                                 과일.및.과일가공품 +
+                                 채소.및.채소가공품 +
+                                 해조.및.해조가공품 +
+                                 당류.및.과자류 +
+                                 조미식품 +
+                                 기타식품 +
+                                 유지류 +
+                                 쥬스.및.기타음료 +
+                                 커피.및.차 ,
+                                    data = engel, na.action = na.omit,
+                       importance = FALSE, ntree = 500 ,proximity = TRUE)
```

```
                Type of random forest: regression
                      Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 57.7317
                    % Var explained: 44.16
```

모델의 설명력: 44.16

나무 수가 너무 작으면 설명력↓
너무 크면 시간이 오래 걸려 비효율적

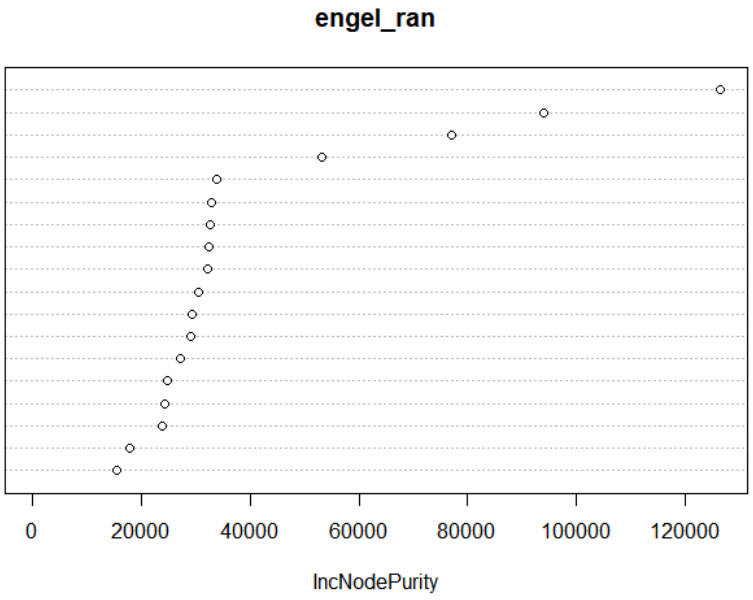
나무수 100~1000까지중,
최적의 나무수를 500으로 설정

최적의 나무 수를 찾아
모델의 설명력을 높이는 과정 진행

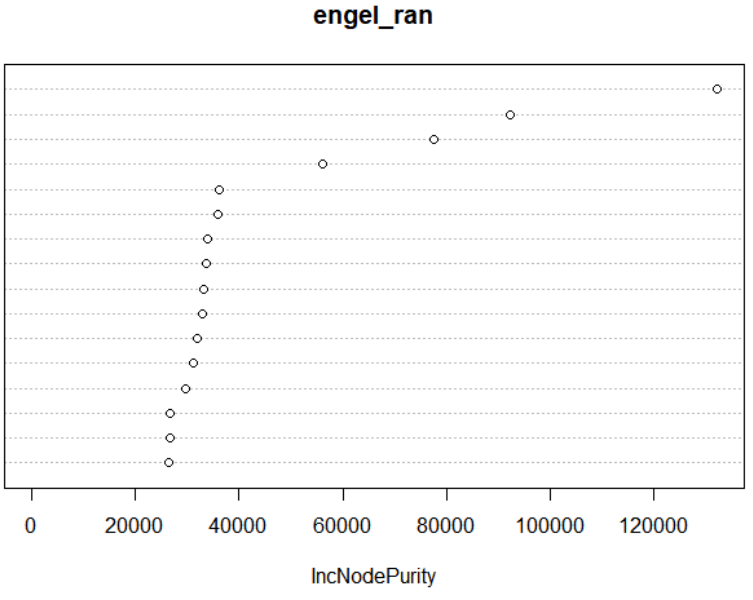
02 데이터 정제

여가·엥겔 지수의 최적의 변수는 (2)

식료품비에 해당하는 하위 변수가 19가지이기 때문에 최적의 변수만을 찾기 위한 분류 진행




```
                Type of random forest: regression
                      Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 57.7317
                    % Var explained: 44.16
```

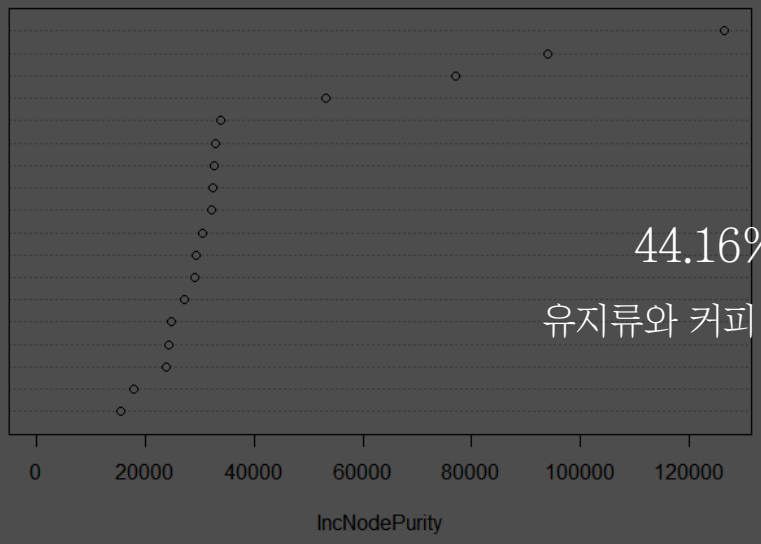



```
                Type of random forest: regression
                      Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 57.42584
                    % Var explained: 44.46
```

02 데이터 정제

여가·엥겔 지수의 최적의 변수는 (2)

식료품비에 해당하는 하위 변수가 19가지이기 때문에 최적의 변수만을 찾기 위한 분류 진행

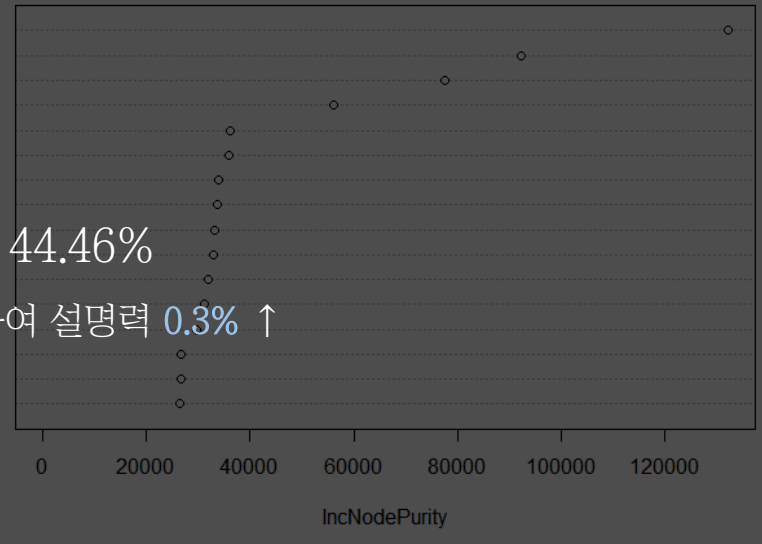



```
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 57.7317
                    % Var explained: 44.16
```

44.16%

44.46%

유지류와 커피 및 차 변수를 제외하여 설명력 0.3% ↑

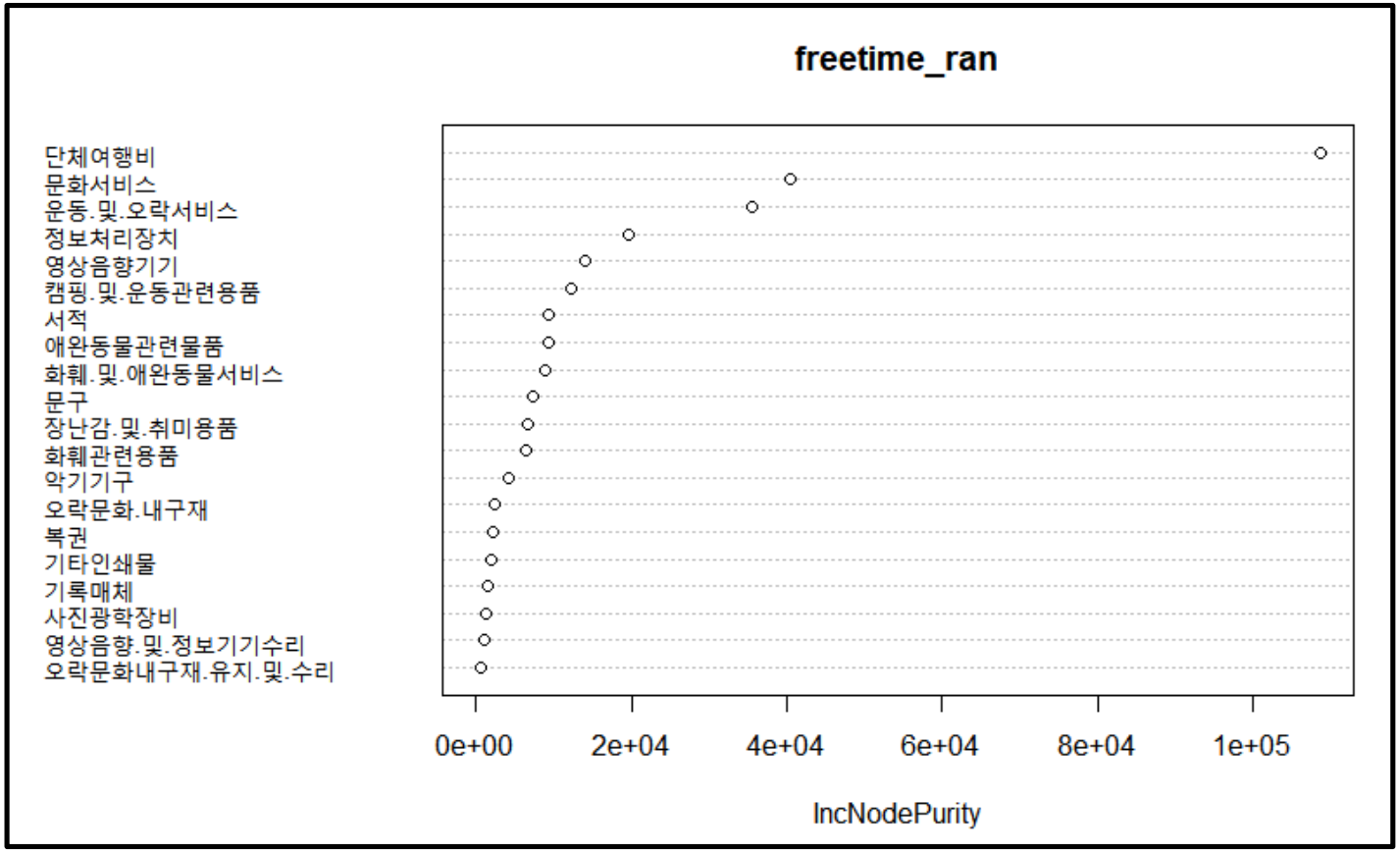



```
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 57.42584
                    % Var explained: 44.46
```

02 데이터 정제

여가·엥겔 지수의 최적의 변수는 (3)

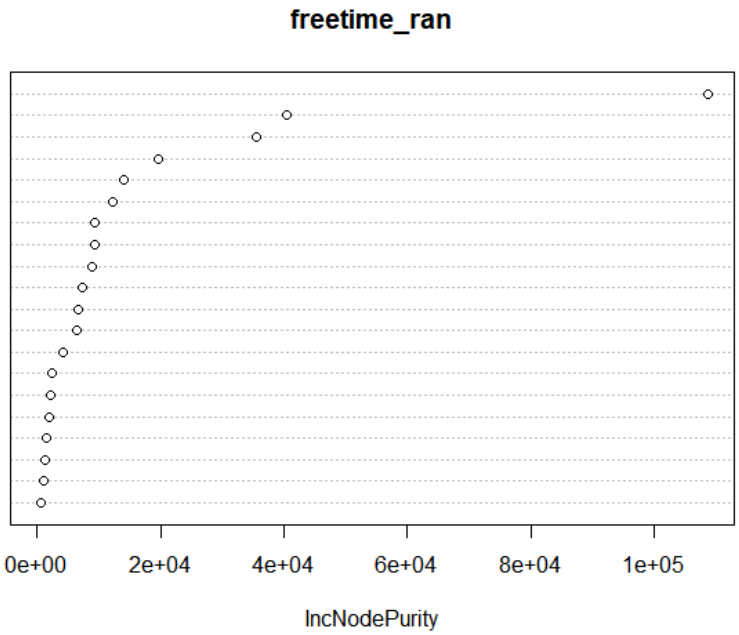
오락 및 문화 비용에 해당하는 하위 변수 역시 20가지이기 때문에 엥겔지수와 동일한 과정 반복



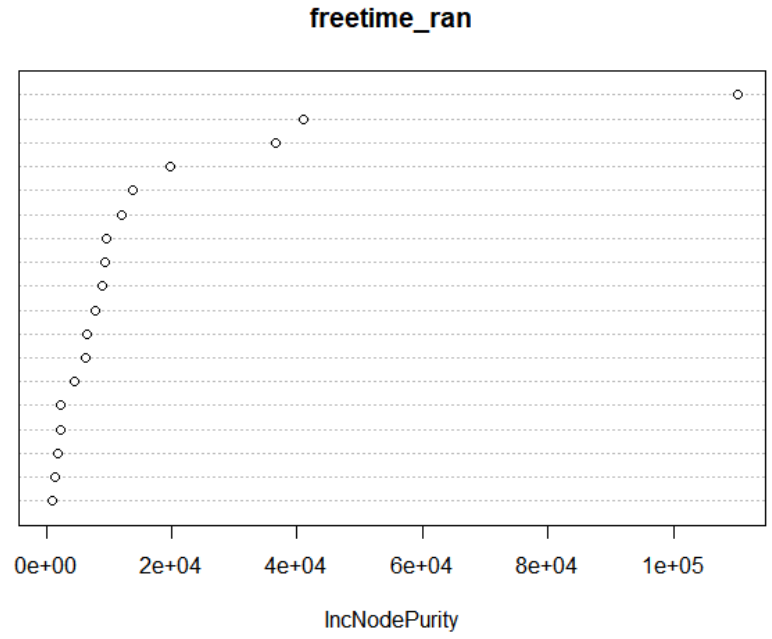
02 데이터 정제

여가·엥겔 지수의 최적의 변수는 (3)

오락 및 문화 비용에 해당하는 하위 변수 역시 20가지이기 때문에 엥겔지수와 동일한 과정 반복




```
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 11.75116
                    % Var explained: 71.63
```

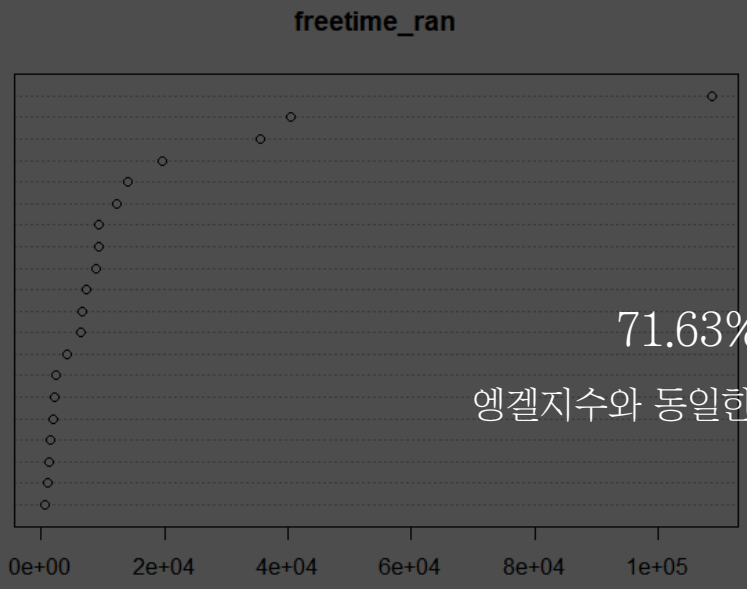



```
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 11.6664
                    % Var explained: 71.83
```

02 데이터 정제

여가·엥겔 지수의 최적의 변수는 (3)

오락 및 문화 비용에 해당하는 하위 변수 역시 20가지이기 때문에 엥겔지수와 동일한 과정 반복

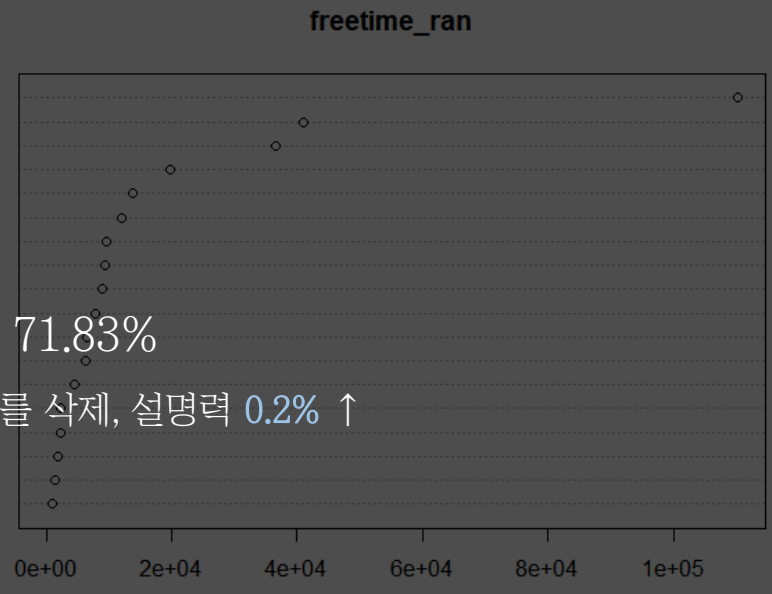



```
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 11.75116
                    % Var explained: 71.63
```

71.63%

71.83%

엥겔지수와 동일한 방법으로 변수 3개를 삭제, 설명력 0.2% ↑



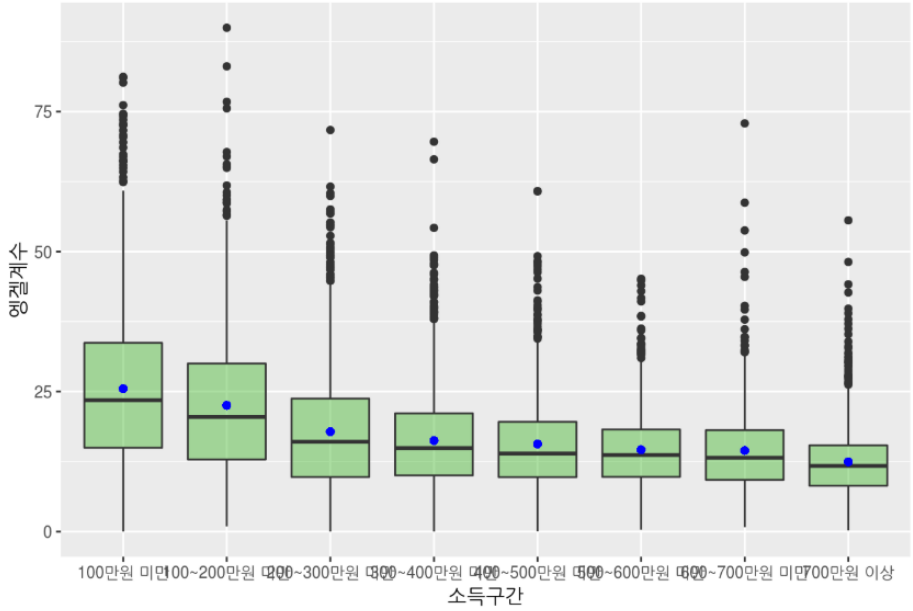

```
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 11.6664
                    % Var explained: 71.83
```

03 데이터 분석

소득 구간 별 여가·엥겔 지수

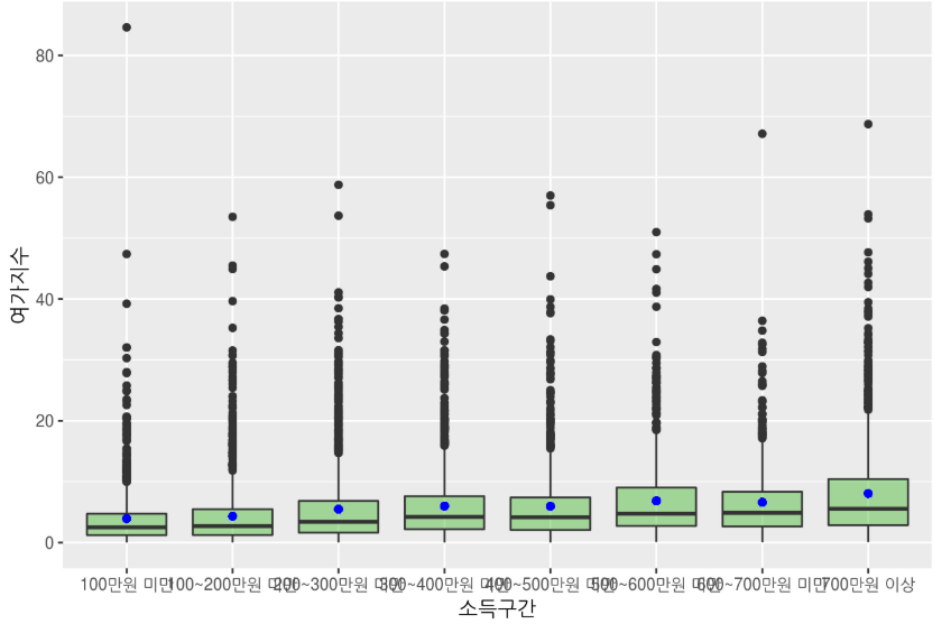
소득 구간 별로 엥겔 및 여가 지수의 차이가 있는지 보기 위해 ANOVA-TEST 진행

‘소득구간 ↗, 엥겔지수 ↘’




```
> oneway.test(engel ~ 소득구간, data=engel, var.equal=FALSE)

        One-way analysis of means (not assuming equal variances)

data:  engel and 소득구간
F = 99.814, num df = 7.0, denom df = 2498.6, p-value < 2.2e-16
```

‘소득구간 ↗, 여가지수 ↗’




```
> oneway.test(freetime ~ 소득구간, data=engel, var.equal=FALSE)

        One-way analysis of means (not assuming equal variances)

data:  freetime and 소득구간
F = 32.378, num df = 7.0, denom df = 2534.2, p-value < 2.2e-16
```

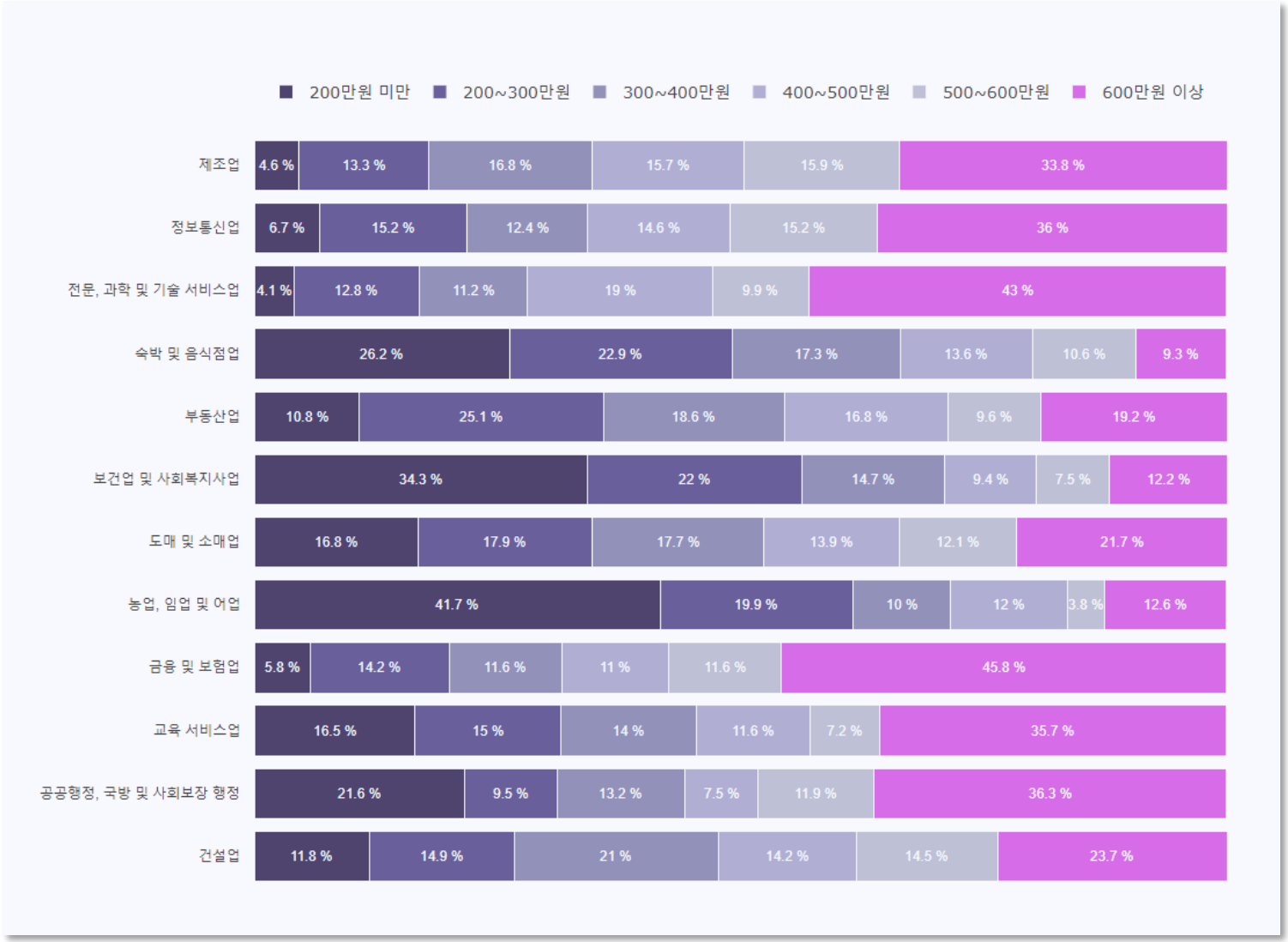
03 데이터 분석

산업별 소득 구간의 비율 차이

산업 군 별로
소득 구간의 차이가
있는지 카이제곱 독립성
검증 진행 및 시각화

| 검증방법 |
카이제곱 독립성 검정
(P-value < 2.2e-16)

```
> chisq.test(engel$inderstry, engel$소득구간)

        Pearson's Chi-squared test

data:  engel$inderstry and engel$소득구간
X-squared = 1243.6, df = 112, p-value < 2.2e-16
```



03 데이터 분석

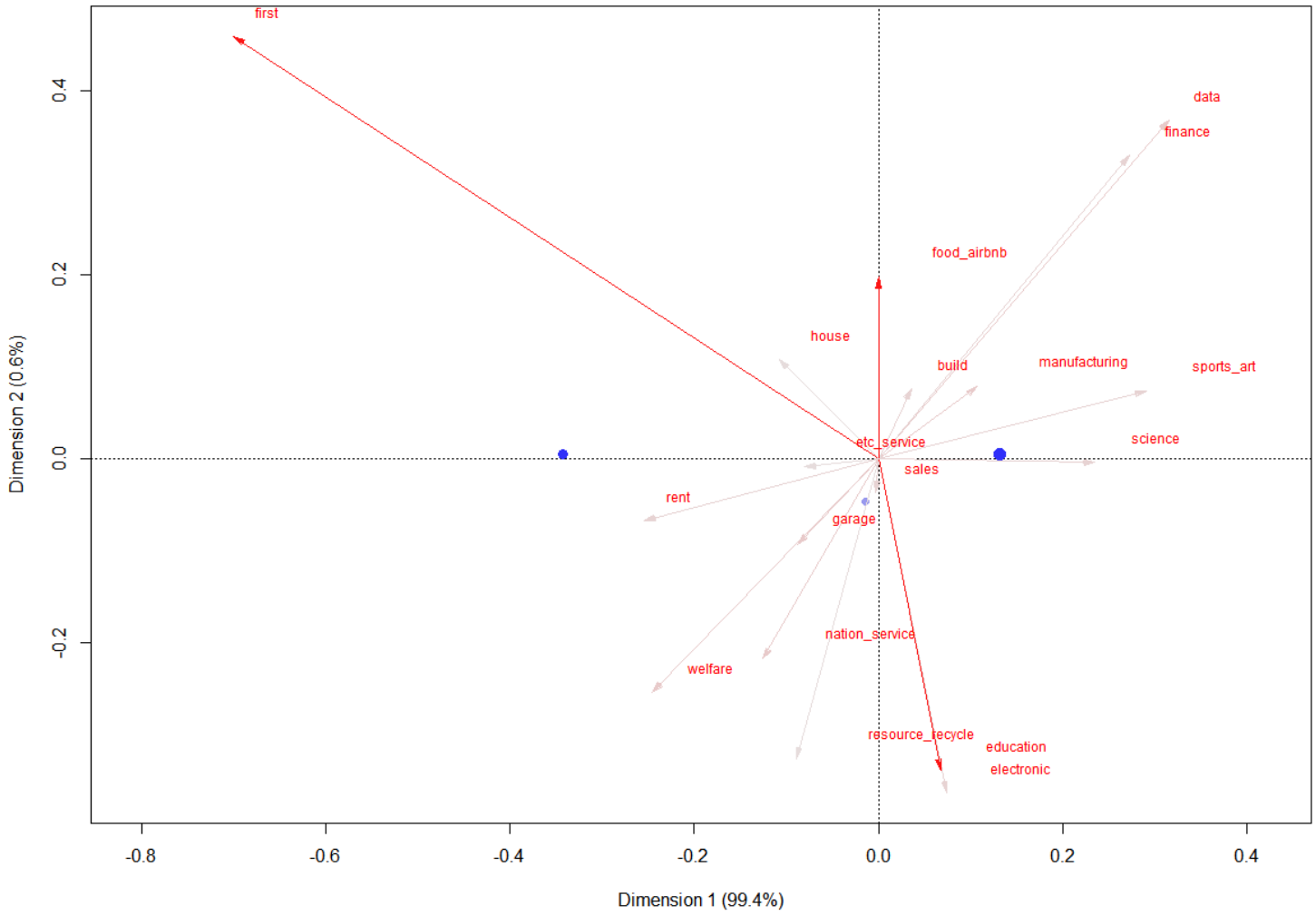
PCA

산업별 각 지수의 주성분 분석

임의로 정의한 3가지 지수와 산업별로 어떠한 관계가 있는지 설명하기 위해 PCA를 활용한 시각화

| 사용된 변수(차원) |

1) 엥겔지수
2) 여가지수
3) 행복지수



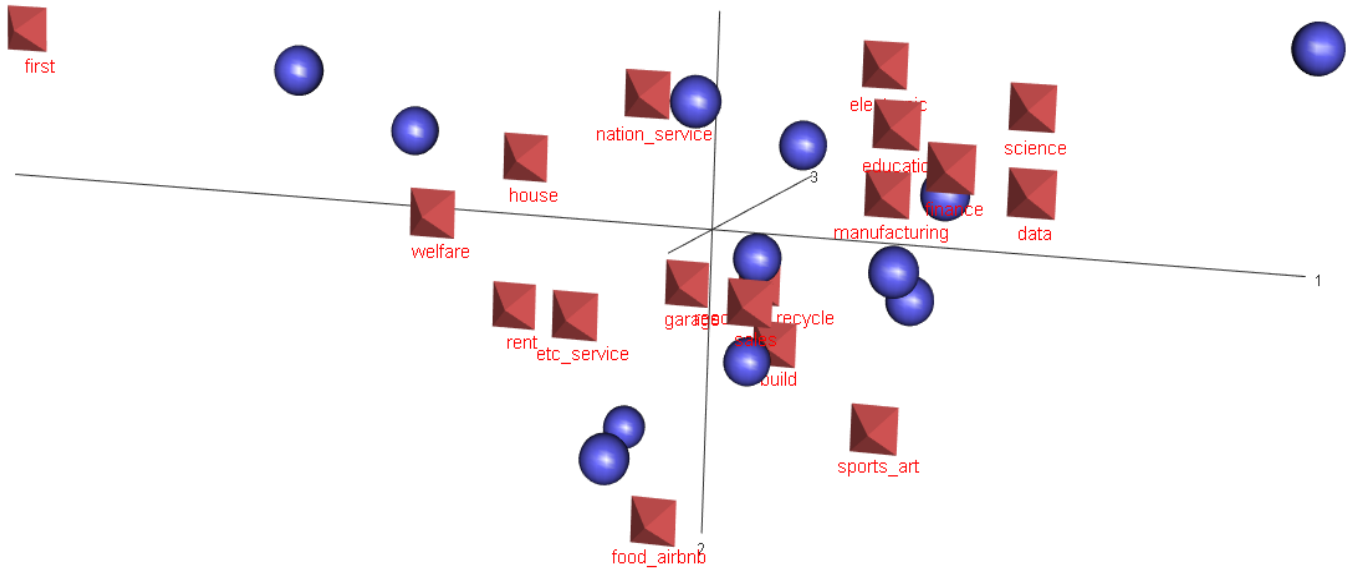
PCA

산업별
각 지수의
주성분 분석

엔겔지수와 같이 모든
항목을 소비지출로 나눈
지수를 도출(12개)

소비 행태를 통한 산업별
이미지를 알아보기 위해
PCA 3D를 활용한 시각화

| 사용된 변수(차원) |
12개



03 데이터 분석

전체가 아닌 '정보통신업'의 모습은

보다 세밀한 분석을 위해, 산업군 == 정보통신업 으로 표본 범위 축소 및
'연차 구분' 컬럼 추가 등 데이터 2차 정제

| 데이터 정제 과정 |

1) 표본 단위 : 가구 => 개인
기존 데이터의 기준이 '가구'였기 때문에
소비 변수들을 가구 원 수로 나누어 '개인'으로 변환

2)산업군을 '정보통신업'으로 한정하여 데이터 추출

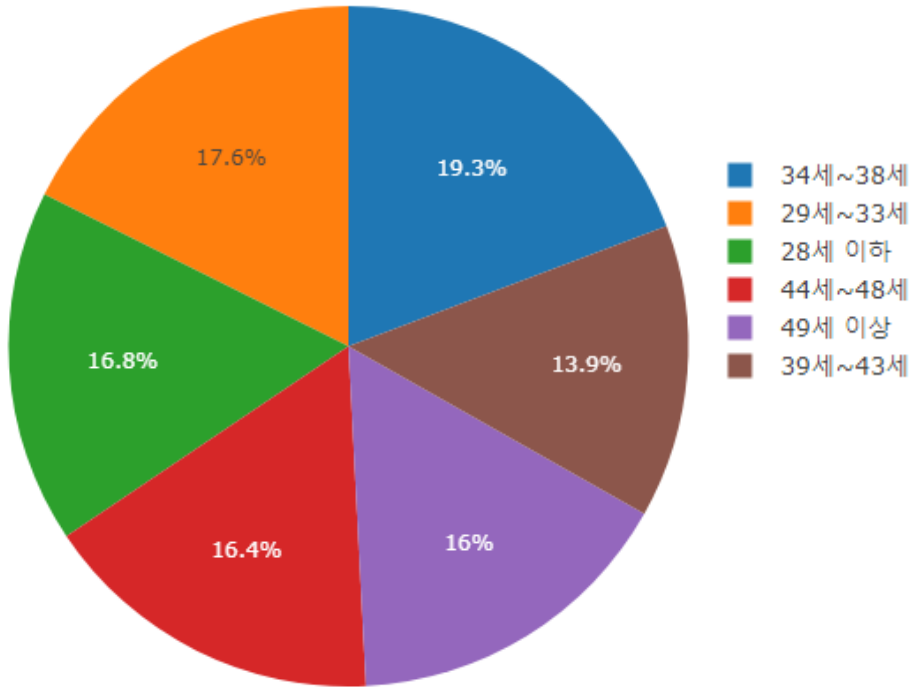
3) 5년 단위의 연차 컬럼 추가
우리 반의 평균 연령(28세)을 취업 나이로 기준 삼아
5년 단위의 연차 컬럼을 추가

```
> chisq.test(table(years$연차구분))

        Chi-squared test for given probabilities

data:  table(years$연차구분)
X-squared = 2.2951, df = 5, p-value = 0.807
```

| 검증방법 |
카이제곱 적합도 검정
(P-value : 0.807)

*P값이 0.05보다 크므로, 연차 별로 표본의 수에는 차이가 없다

' 연차별 비율 그래프 '



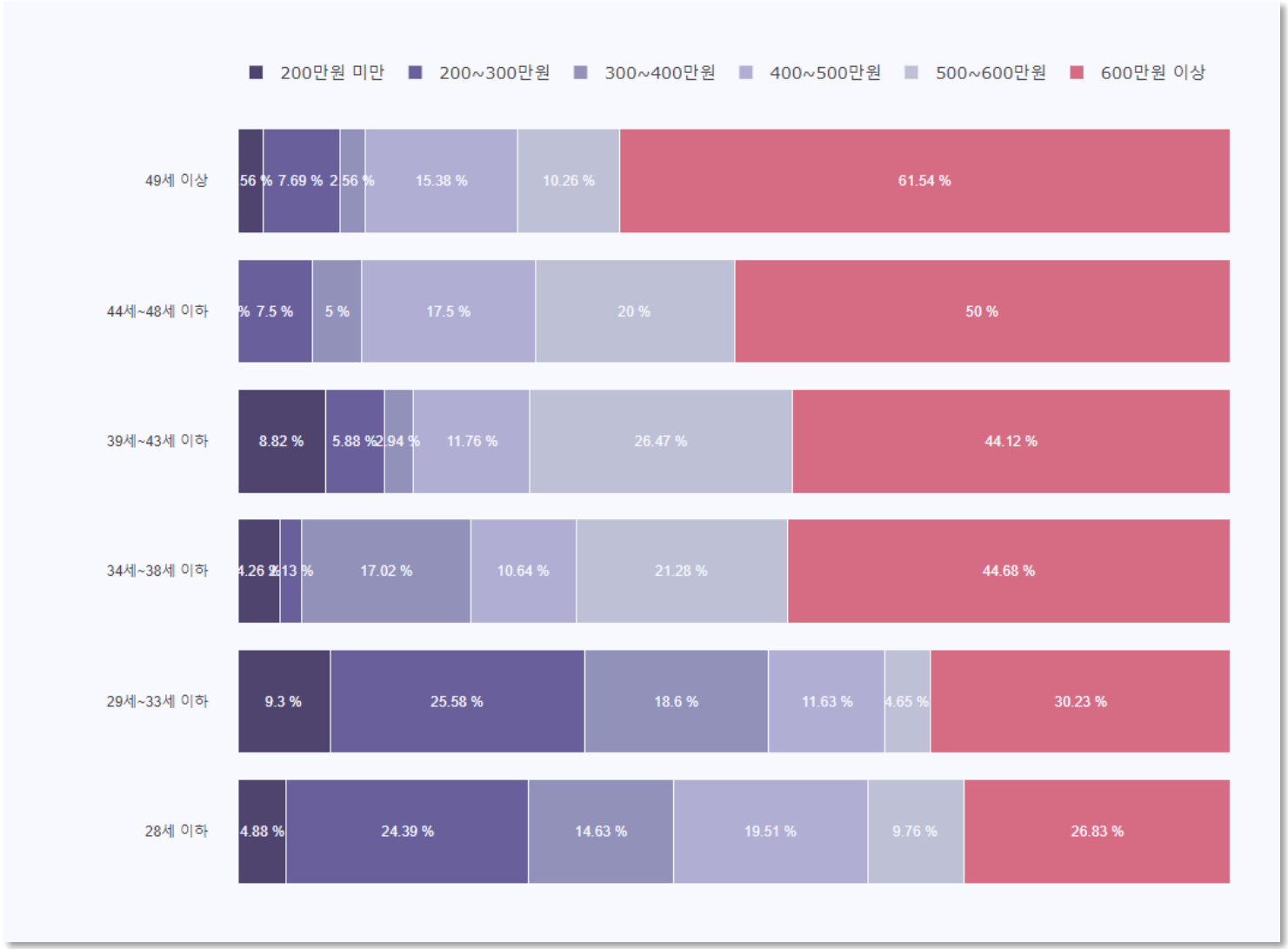
03 데이터 분석

연차별 소득 구간의 비율 차이

연차 별로
소득 구간의 차이가
있는지 카이제곱 독립성
검증 진행 및 시각화

| 검증방법 |
카이제곱 독립성 검증
(P-value: 0.001476)

```
> chisq.test(years$연차구분, years$소득구간)

        Pearson's Chi-squared test

data:  years$연차구분 and years$소득구간
X-squared = 65.134, df = 35, p-value = 0.001476
```



03 데이터 분석

연차별 각 지수의 차이

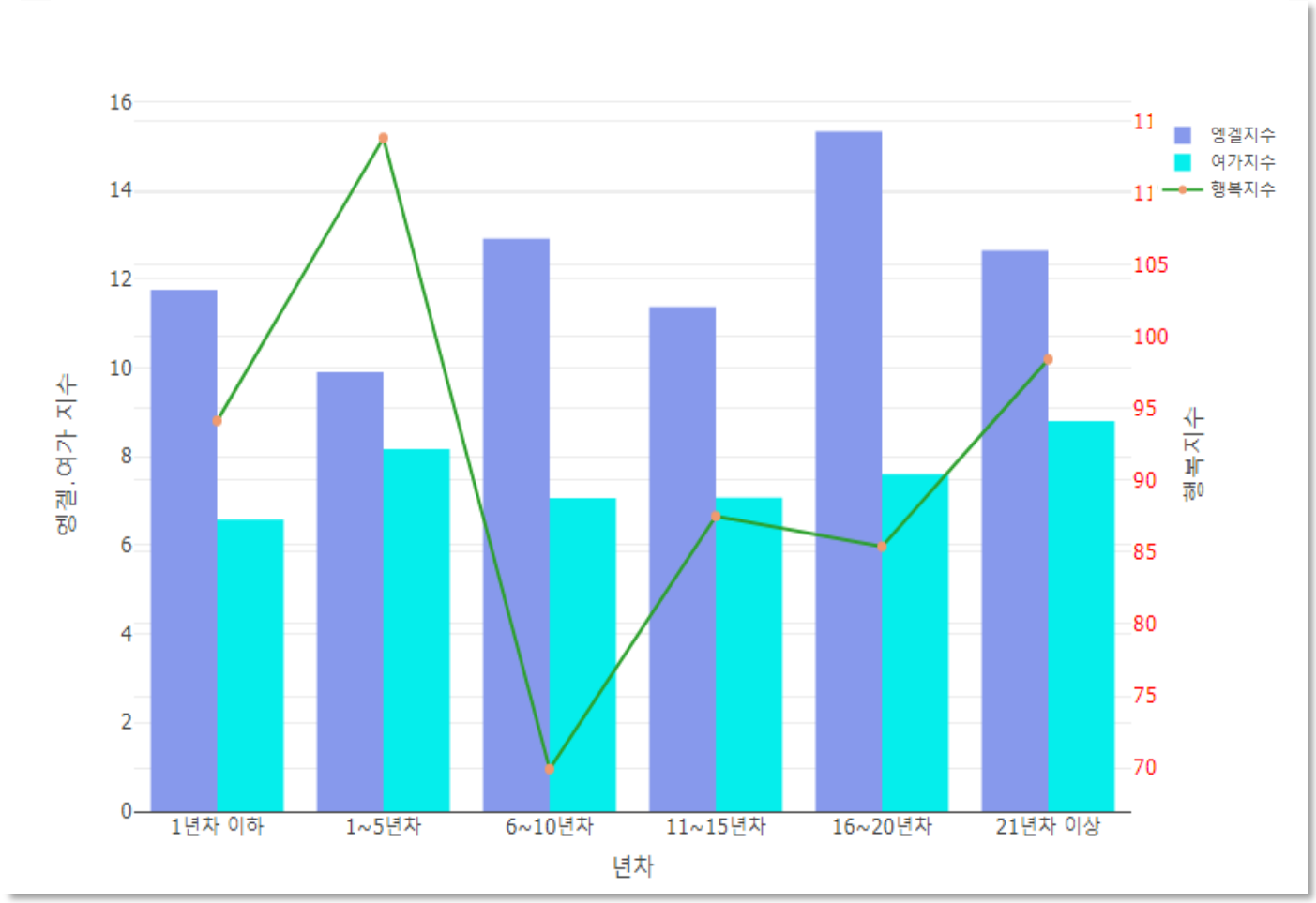
연차 별로
여가/엥겔/행복 지수의
차이를 통계적으로
검정하고 시각화

| 검증방법 |
지수 별로 ANOVA-TEST 3번 진행

1) 엥겔지수: P-value: 0.00947

2) 여가지수: P-value: 0.786(차이X)

3) 행복지수: P-value: 0.5058(차이X)



03 데이터 분석

연차 별로 엥겔 지수의 차이가 나는 원인은

엥겔/여가/행복 지수 중, 유일하게 통계적으로 유의미한 차이를 보이는 엥겔 지수
추가적인 분석을 위해 사후 검정을 진행하고 연차 별 엥겔 지수 차이의 원인을 추측

|사후검정|

```
$groups
       engel groups
5 15.327500      a
3 12.912766     ab
6 12.643590     ab
1 11.751220     ab
4 11.373529     ab
2  9.902326      b

attr(,"class")
[1] "group"
```

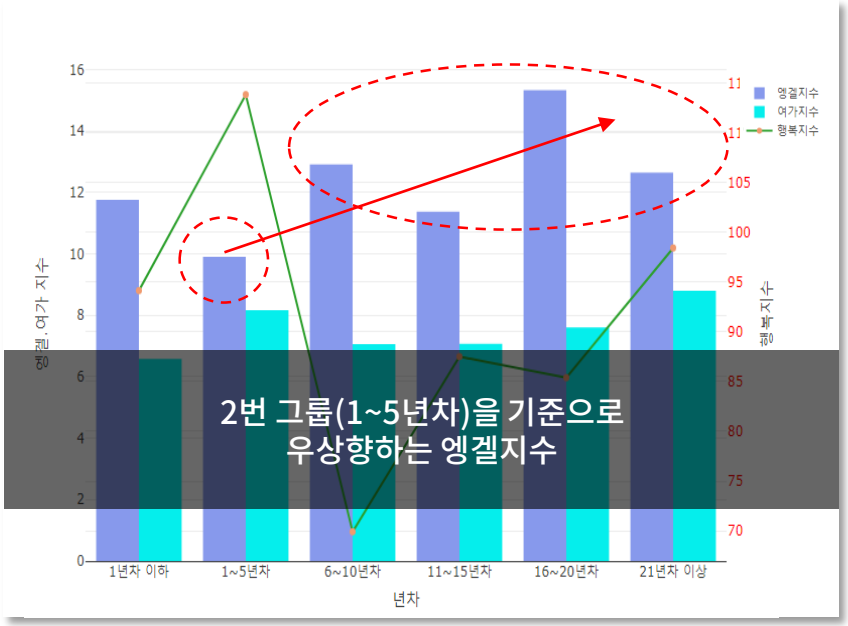
*검정 방법으로 bonferroni 사용

“

명확한 차이를 보이는 그룹 2,5

분류가 애매한 그룹들의 추가적인
평균 비교 필요성 有

”



03 데이터 분석

연차 별로 엔겔 지수의 차이가 나는 원인은

엔겔/여가/행복 지수 중, 유일하게 통계적으로 유의미한 차이를 보이는 엔겔 지수
추가적인 분석을 위해 사후 검정을 진행하고 연차 별 엔겔 지수 차이의 원인을 추측

|사후검정|

```
$groups
      engel groups
5 15.327500      a
3 12.912766     ab
6 12.643590     ab
1 11.751220     ab
4 11.373529     ab
2  9.902326      b

attr(,"class")
[1] "group"
```

*검정 방법으로 bonferroni 사용

“

명확한 차이를 보이는 그룹 2,5

분류가 애매한 그룹들의 추가적인
평균 비교 필요성 有

”

“

**2번 그룹과 3번 그룹의
평균 차이를 검정해보니**

1) 정규성 검정(P-value: 0.8061 / 0.2018)
2) 등분산성 검정(P-value : 0.4649)
3) T-TEST(P-value:0.004339)

통계적으로 유의미한 차이가 있다.

”

**1~5년 차를 기준으로 엔겔지수의
유의미한 평균차이의 원인은 ‘결혼’일 가능성 有**

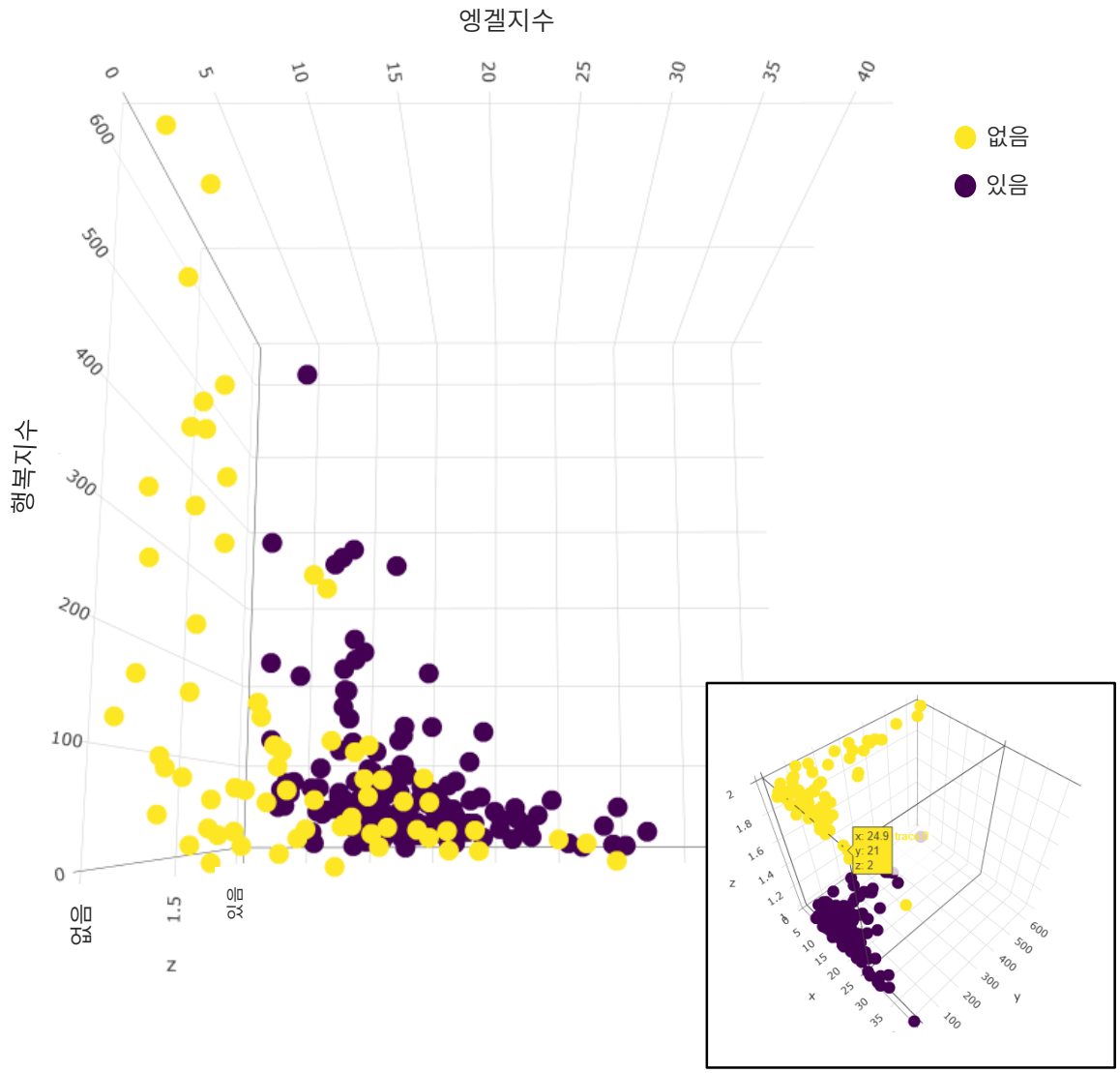
03 데이터 분석

결혼 유무에 따른 행복과 엔겔 지수

결혼 유무에 따른
행복 및 엔겔 지수의
차이를 알아보기 위해
3D그래프로 시각화

| 검증방법 |
지수 별로 ANOVA-TEST 3번 진행

1) **엔겔지수: P-value: 0.004288**
2) 여가지수: P-value: 0.6416(차이X)
3) **행복지수: P-value: 0.01453**



#plotly_scatter3d

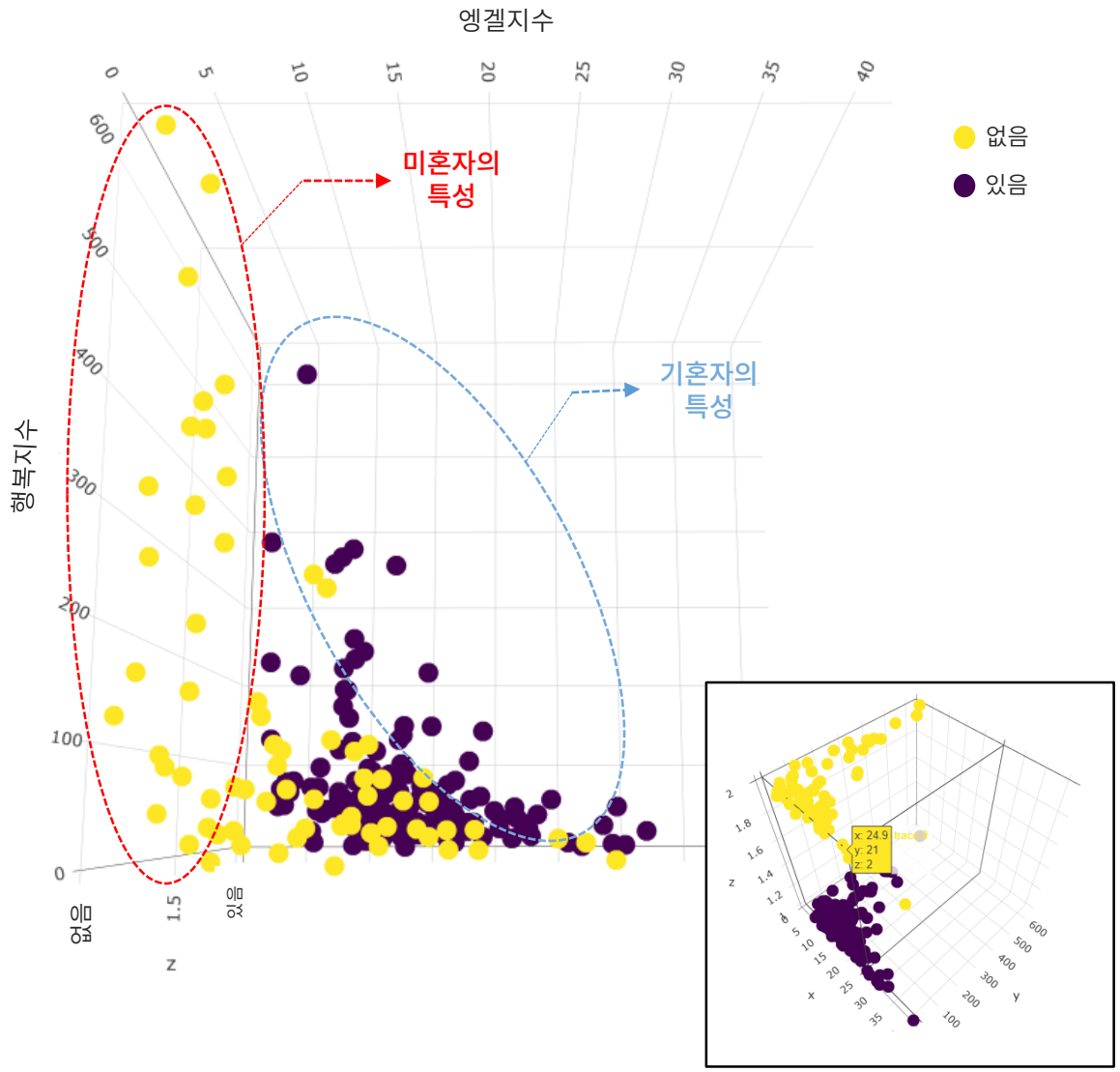
03 데이터 분석

결혼 유무에 따른 행복과 엔겔 지수

결혼 유무에 따른
행복 및 엔겔 지수의
차이를 알아보기 위해
3D그래프로 시각화

| 검증방법 |
지수 별로 ANOVA-TEST 3번 진행

1) **엔겔지수: P-value: 0.004288**
2) 여가지수: P-value: 0.6416(차이X)
3) **행복지수: P-value: 0.01453**



#plotly_scatter3d

호기심

if… 우리가 결혼을 해서 서울에 산다면,
어느 동네에 살아야 가장 행복할 수 있을까?

데이터: 서울시 문화 공간 현황 (2019~)

데이터 내용: 영화관, 공연장, 도서관, 체육센터, 예술극장, 박물관, 미술관, 갤러리, 아이스링크장 등 작은 구별 도서관부터 DDP (동대문 디자인 플라자)와 같은 대규모 문화 복합시설을 포함한 통계 자료

시설 개수: 10313개

데이터: 서울시 생필품 농축 수산물 가격정보 (2019~)

데이터 내용: 시장, 마트, 백화점 등에서 판매하는 식료품을 지역/품목/판매규격/가격 등으로 정리해놓은 통계 자료

품목 개수: 800519개

문화공간의수

\+

식료품의 수

=

사람이 생활 속에서 기쁘고 즐겁고

\+

만족을 느끼는 상태에 있는 것

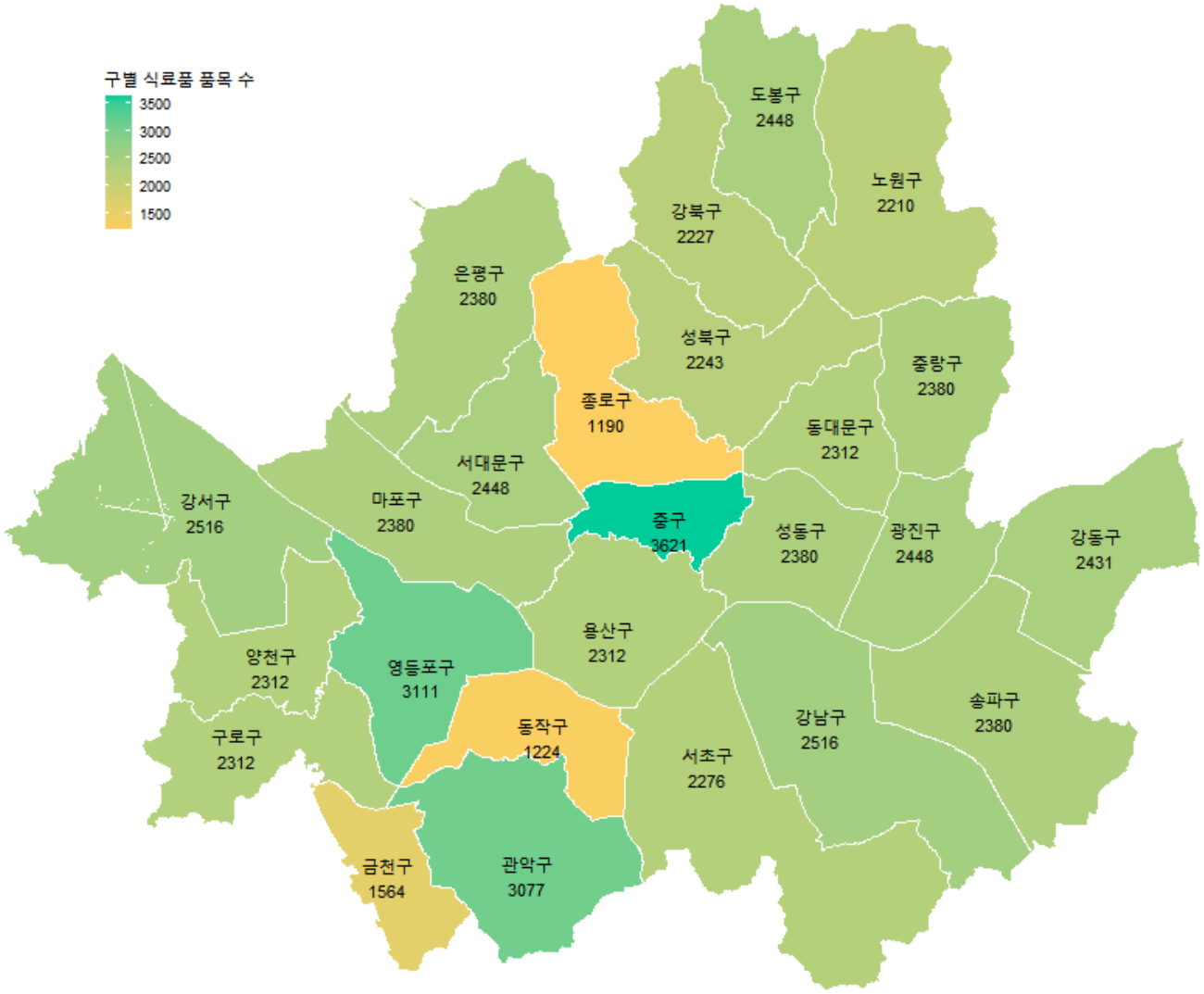
03 데이터 분석

서울시 구별 식료품 품목 개수

서울시25개 자치구별
위도와 경도를 활용,
식료품 품목의 수에 따라
자치구의 색을 다르게 표현

| 식료품 품목 순위 |

1) 중구
2) 영등포구
3) 관악구



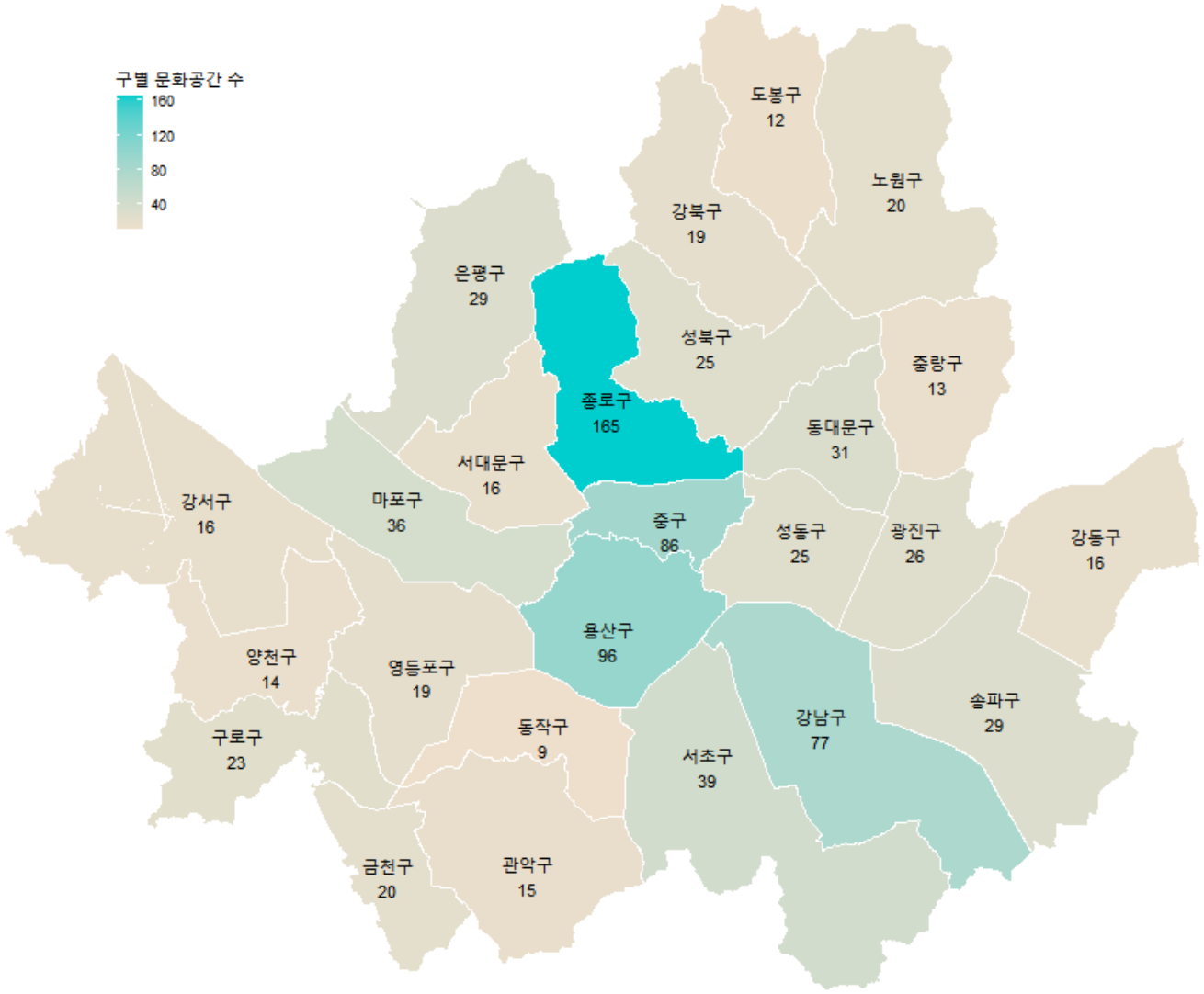
03 데이터 분석

서울시 구별 문화공간 수

서울시25개 자치구별
위도와 경도를 활용,
문화공간의 수에 따라
자치구의 색을 다르게 표현

| 문화공간 순위 |

1) 종로구
2) 용산구
3) 중구



03 데이터 분석

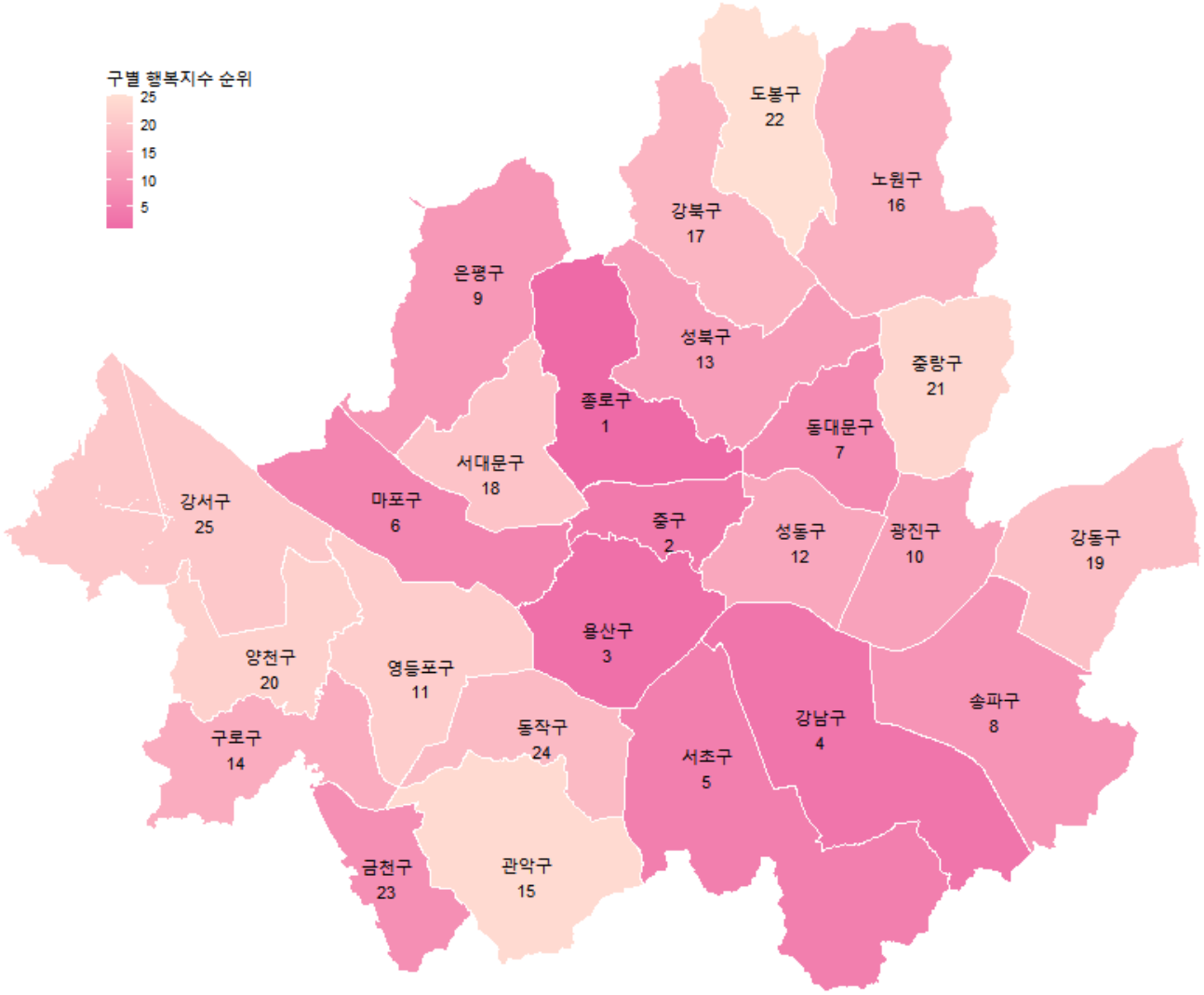
서울시 구별 행복지수

문화공간과 식료품 품목의
표본 비율이 1:80

행복지수 표현을 위해
식료품 품목개수/80으로
표본의 비율을 맞춘 후 더하여
행복지수 도출

| 행복지수 순위 |

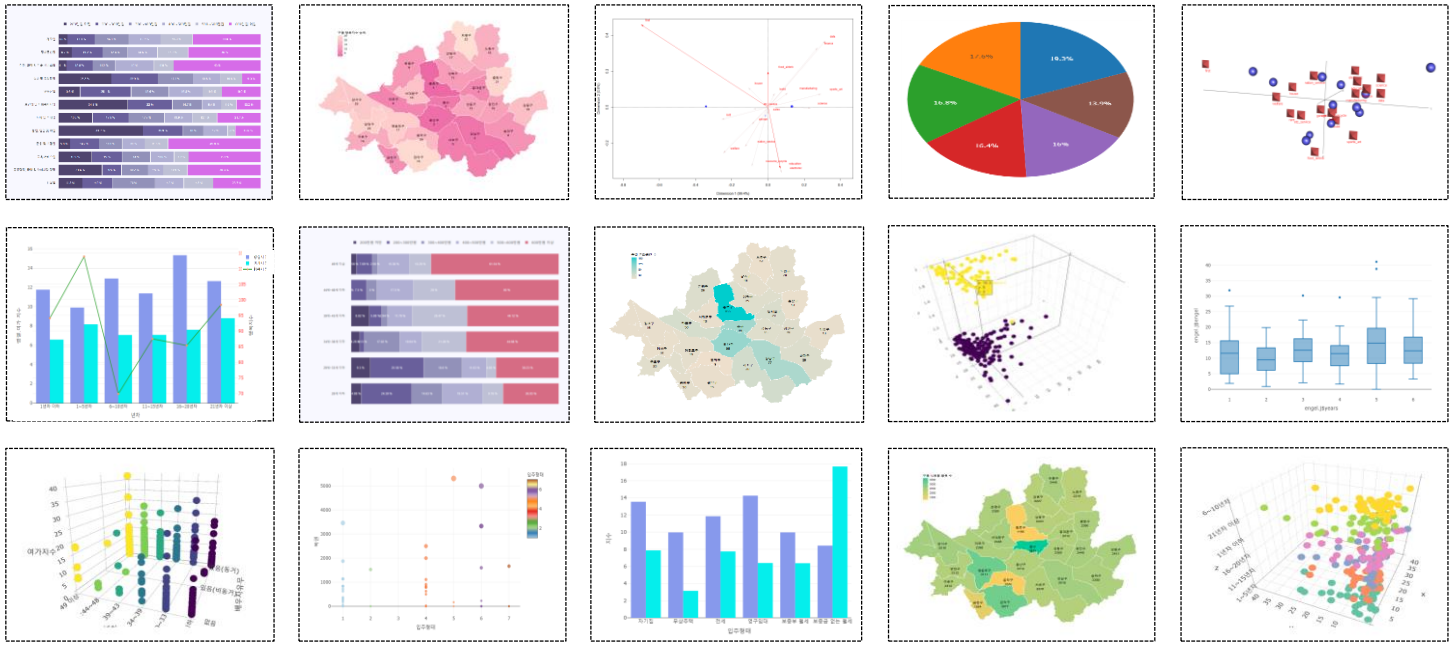
1) 종로구
2) 중구
3) 용산구



R shiny

가계지출 통계 자료를 이용하여 시각화한 자료를 R shiny를 통해 웹과 연동 및 업로드

#ggmap #ggplot2 #plotly #3d_plot #stack_barchart #pca



@https://junghi.shinyapps.io/project_rshiny/

감사합니다

정효인