ESC 20-1 Final Project

Week1

Team 2

김윤환 조인식 신혜연 백채빈 이상완 손지우

Our Goal

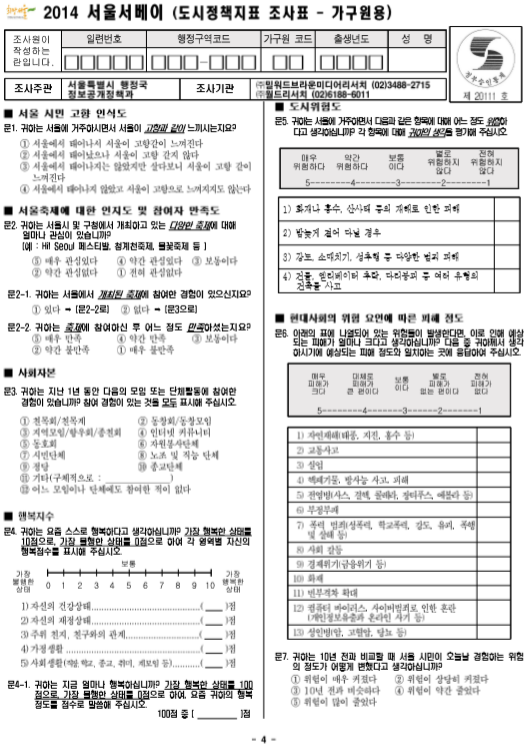
서울시 가구원의 행복도
예측 모델 만들기

Contents

Pre-processing
Visualization

Feature
Extraction

Seoul Happiness 2014



2014 서울서베이 (도시정책지표 조사표 - 가구원용)

조사원이 작성하는 란입니다.	일련번호	행정구역코드	가구원 코드	출생년도	성 명
	□□□□□	□□□-□□□	□□	□□□□	

조사주관	서울특별시 행정국 정보공개정책과	조사기관	㈜힐워드브라운미디어리서치 (02)3488-2715 ㈜컬트리서치 (02)6188-6011

제 20111 호

■ 서울 시민 고향 인식도

문1. 귀하는 서울에 거주하시면서 서울이 *고향과 같이* 느껴지는지요?

① 서울에서 태어나서 서울이 고향같이 느껴진다
② 서울에서 태어났으나 서울이 고향 같지 않다
③ 서울에서 태어나지는 않았지만 살다보니 서울이 고향 같이 느껴진다
④ 서울에서 태어나지 않았고 서울이 고향으로 느껴지지도 않는다

■ 서울축제에 대한 인지도 및 참여자 만족도

문2. 귀하는 서울시 및 구청에서 개최하고 있는 *다양한 축제*에 대해 얼마나 관심이 있습니까?
(예 : Hi Seoul 페스티벌, 청계천축제, 불꽃축제 등)

③ 매우 관심있다 ④ 약간 관심있다 ③ 보통이다
② 약간 관심없다 ① 전혀 관심없다

문2-1. 귀하는 서울에서 *개최된 축제*에 참여한 경험이 있으신지요?

① 있다 → (문2-2로) ② 없다 → (문3으로)

문2-2. 귀하는 *축제*에 참여하신 후 어느 정도 *만족*하셨는지요?

⑤ 매우 만족 ④ 약간 만족 ③ 보통이다
② 약간 불만족 ① 매우 불만족

■ 사회자본

문3. 귀하는 지난 1년 동안 다음의 모임 또는 단체활동에 참여한 경험이 있습니까? 참여 경험이 있는 것을 모두 표시해 주십시오.

① 친목회/친목계 ② 동창회/동창모임
③ 지역모임/향우회/종친회 ④ 인터넷 커뮤니티
⑤ 동호회 ⑥ 자원봉사단체
⑦ 시민단체 ⑧ 노조 및 직능 단체
⑨ 정당 ⑩ 종교단체
⑪ 기타(구체적으로 :)
⑫ 어느 모임이나 단체에도 참여한 적이 없다

■ 행복지수

문4. 귀하는 요즘 스스로 행복하다고 생각하십니까? *가장 행복한 상태*를 10점으로, *가장 불행한 상태*를 0점으로 하여 각 영역별 자신의 행복점수를 표시해 주십시오.

보통
가장 불행한 상태 0 1 2 3 4 5 6 7 8 9 10 가장 행복한 상태

1) 자신의 건강상태.....(___)점
2) 자신의 재정상태.....(___)점
3) 주위 친지, 친구와의 관계.....(___)점
4) 가정 생활.....(___)점
5) 사회 생활(학업 학교, 종교, 취미, 레포츠 등).....(___)점

문4-1. 귀하는 지금 얼마나 행복하십니까? *가장 행복한 상태*를 100점으로, *가장 불행한 상태*를 0점으로 하여, 요즘 귀하의 행복 정도를 점수로 말씀해 주십시오.

100점 중 [_____]점

■ 도시위험도

문5. 귀하는 서울에 거주하면서 다음과 같은 항목에 대해 어느 정도 *위험*하다고 생각하십니까? 각 항목에 대해 *귀하의 생각*을 평가해 주십시오.

	매우 위험하다	약간 위험하다	보통이다	별로 위험하지 않다	전혀 위험하지 않다
	5	4	3	2	1
1) 화재나 홍수, 산사태 등의 재해로 인한 피해					
2) 방射능 물질 누출 경우					
3) 강도, 소매치기, 성추행 등 다양한 범죄 피해					
4) 건물 엘리베이터 추락, 다리붕괴 등 여러 유형의 건축물 사고					

■ 현대사회의 위험 요인에 따른 피해 정도

문6. 아래의 표에 나열되어 있는 위험들이 발생한다면, 이로 인해 예상되는 피해가 얼마나 크다고 생각하십니까? 다음 중 귀하께서 생각하시기에 예상되는 피해 정도와 일치하는 곳에 응답해 주십시오.

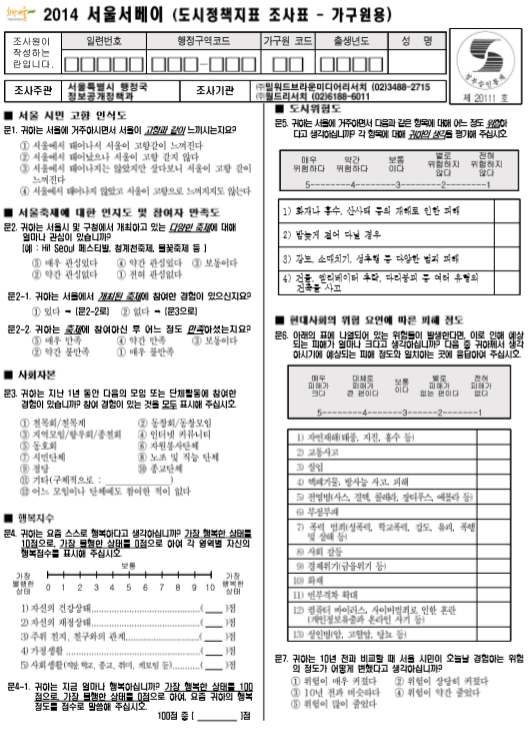
	매우 피해가 크다	다소 피해가 클 것이다	보통이다	별로 피해가 없을 것이다	전혀 피해가 없다
	5	4	3	2	1
1) 자연재해(태풍, 지진, 홍수 등)					
2) 교통사고					
3) 실업					
4) 백혈기물, 방사능 사고, 피해					
5) 전염병(사스, 결핵, 콜레라, 장티푸스, 에볼라 등)					
6) 부정부패					
7) 폭력 범죄(성폭력, 학교폭력, 갈도, 유괴, 폭행 및 살해 등)					
8) 사회 갈등					
9) 경제위기(금융위기 등)					
10) 화재					
11) 빈부격차 확대					
12) 컴퓨터 바이러스, 사이버범죄로 인한 혼란 (개인정보유출과 온라인 사기 등)					
13) 성인범죄, 고령화, 알코 등)					

문7. 귀하는 10년 전과 비교할 때 서울 시민이 오늘날 경험하는 위험의 정도가 어떻게 변했다고 생각하십니까?

① 위험이 매우 커졌다 ② 위험이 상당히 커졌다
③ 10년 전과 비슷하다 ④ 위험이 약간 줄었다
⑤ 위험이 많이 줄었다

- 4 -

Seoul Happiness 2014



2014 서울서베이 (도시정책지표 조사표 - 가구원용)

조사원이 작성하는 란입니다.

일련번호	행정구역코드	가구원 코드	출생년도	성 명
□□□□□	□□□-□□□	□□	□□□□	

조사주관	서울특별시 행정국 정보공개정책과	조사기관	㈜힐워드브라운미디어리서치 (02)3488-2715 ㈜힐드리서치 (02)6188-6011

제 20111 호

■ 서울 시민 고향 인식도

문1. 귀하는 서울에 거주하시면서 서울이 고향과 같이 느끼시는지요?

① 서울에서 태어나서 서울이 고향같이 느껴진다
② 서울에서 태어났으나 서울이 고향 같지 않다
③ 서울에서 태어나지는 않았지만 살다보니 서울이 고향 같이 느껴진다
④ 서울에서 태어나지 않았고 서울이 고향으로 느껴지지도 않는다

■ 서울속세에 대한 인지도 및 참여자 만족도

문2. 귀하는 서울시 및 구청에서 개최하고 있는 다양한 축제에 대해 얼마나 관심이 있습니까?
(예 : Hi Seoul 페스티벌, 청계천축제, 월보축제 등)

⑤ 매우 관심있다 ④ 약간 관심있다 ③ 보통이다
② 약간 관심없다 ① 전혀 관심없다

문2-1. 귀하는 서울에서 개최된 축제에 참여한 경험이 있으신지요?

① 있다 → [문2-2로] ② 없다 → [문3으로]

문2-2 귀하는 축제에 참여하신 후 어느 정도 만족하셨는지요?

⑤ 매우 만족 ④ 약간 만족 ③ 보통이다
② 약간 불만족 ① 매우 불만족

■ 사회자본

문3. 귀하는 지난 1년 동안 다음의 모임 또는 단체활동에 참여한 경험이 있습니까? 참여 경험이 있는 것을 모두 표시해 주십시오.

① 친목회/동창회 ② 동창회/동창모임
③ 계역모임/향우회/종친회 ④ 인터넷 커뮤니티
⑤ 동호회 ⑥ 자원봉사단체
⑦ 시민단체 ⑧ 노조 및 직능 단체
⑨ 정당 ⑩ 종교단체
⑪ 기타(구체적으로 :) ⑫ 어느 모임이나 단체에도 참여한 적이 없다

■ 행복지수

문4. 귀하는 요즘 스스로 행복하다고 생각하십니까? 가장 행복한 상태를 10점으로, 가장 불행한 상태를 0점으로 하여 각 영역별 자신의 행복점수를 표시해 주십시오.

보통
가장 불행한 상태 0 1 2 3 4 5 6 7 8 9 10 가장 행복한 상태

1) 자신의 건강상태.....(___)점
2) 자신의 재정상태.....(___)점
3) 주위 친지, 친구와의 관계.....(___)점
4) 가정생활 .....(___)점
5) 사회생활(직장, 학교, 종교, 취미, 레포츠 등).....(___)점

문4-1. 귀하는 지금 얼마나 행복하십니까? 가장 행복한 상태를 100점으로, 가장 불행한 상태를 0점으로 하여, 요즘 귀하의 행복 정도를 점수로 말씀해 주십시오.

100점 중 [_____]점

■ 도시위험도

문5. 귀하는 서울에 거주하면서 다음과 같은 항목에 대해 어느 정도 위험하다고 생각하십니까? 각 항목에 대해 귀하의 생각을 평가해 주십시오.

매우 위험하다 5 | 약간 위험하다 4 | 보통이다 3 | 별로 위험하지 않다 2 | 전혀 위험하지 않다 1

1) 홍수나 호우, 산사태 등의 재해로 인한 피해	
2) 밤늦게 걸어 다닐 경우	
3) 장보 소매치기, 성추행 등 다양한 범죄 피해	
4) 전염, 멜리베이비 추락, 다리붕괴 등 여러 유형의 건축물 사고	

■ 현대사회의 위험 요인에 따른 피해 정도

문6. 아래의 표에 나열되어 있는 위험들이 발생한다면, 이로 인해 예상되는 피해가 얼마나 크다고 생각하십니까? 다음 중 귀하께서 생각하시기에 예상되는 피해 정도와 일치하는 곳에 응답해 주십시오.

매우 피해가 크다 5 | 다소 피해가 클 것이다 4 | 보통이다 3 | 별로 피해가 없을 것이다 2 | 전혀 피해가 없다 1

1) 자연재해(태풍, 지진, 홍수 등)	
2) 교통사고	
3) 실업	
4) 핵폐기물, 방사능 사고, 피해	
5) 전염병(사스, 결핵, 콜레라, 장티푸스, 에볼라 등)	
6) 부정부패	
7) 폭력, 범죄(살옥력, 학교폭력, 갈도, 유괴, 폭행 및 살해 등)	
8) 사회 갈등	
9) 경제위기(금융위기 등)	
10) 화재	
11) 빈부격차 확대	
12) 컴퓨터 바이러스, 사이버범죄로 인한 혼란 (개인정보유출과 온라인 사기 등)	
13) 실업(방업, 고령화, 탈노 등)	

문7. 귀하는 10년 전과 비교할 때 서울 시민이 오늘날 경험하는 위험의 정도가 어떻게 변했다고 생각하십니까?

① 위험이 매우 커졌다 ② 위험이 상당히 커졌다
③ 10년 전과 비슷하다 ④ 위험이 약간 줄었다
⑤ 위험이 많이 줄었다

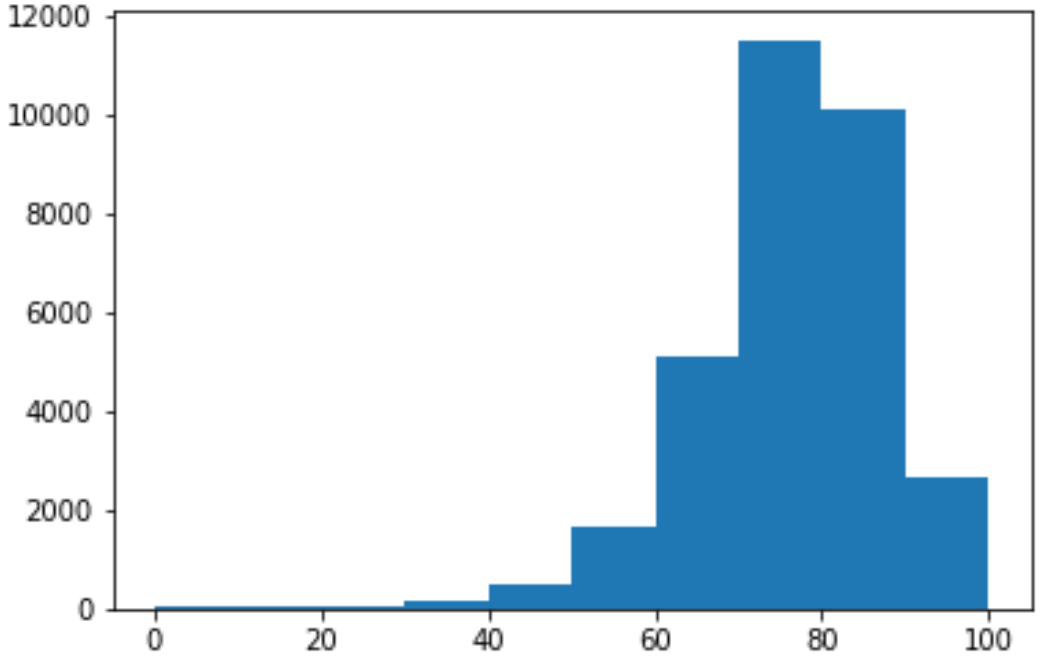
- 4 -

- 198 variables
- 31848 rows

```
df.shape
```

```
(31848, 198)
```

Y

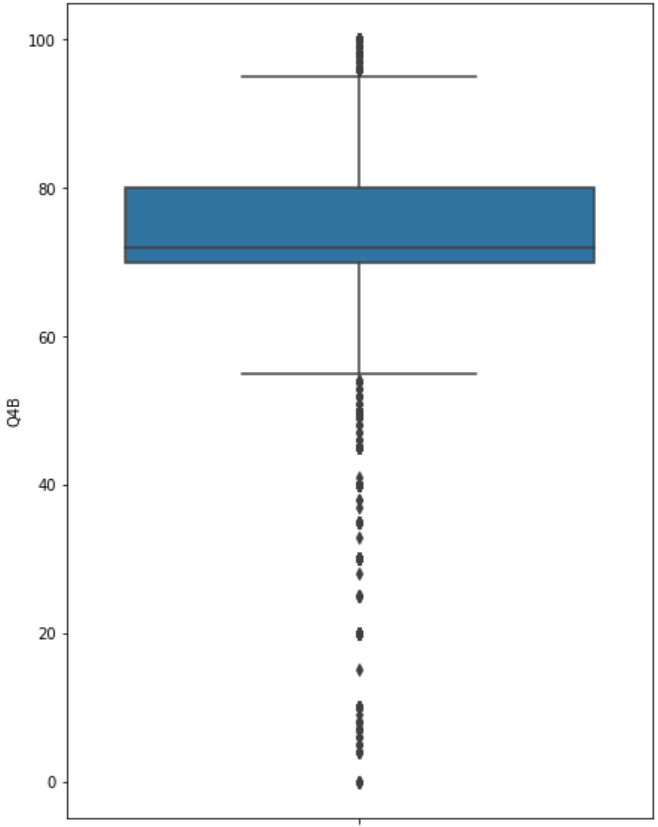


- 0-100 주관적 행복점수
- Mean 72.68

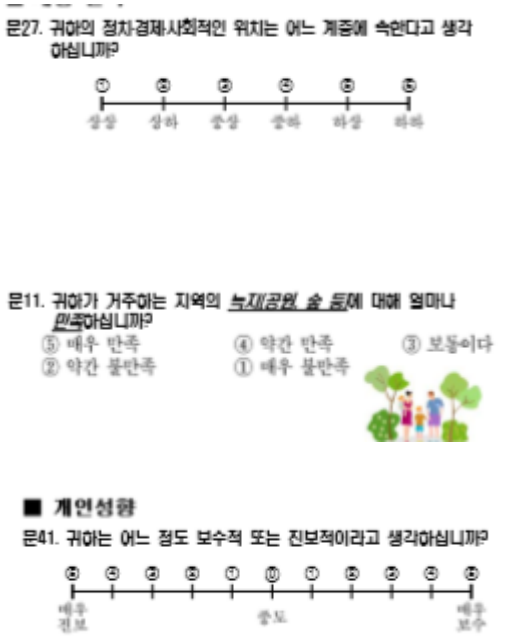
Y

```
y.describe()
```

```
count    31848.000000
mean        72.680482
std         11.822053
min          0.000000
25%         70.000000
50%         72.000000
75%         80.000000
max        100.000000
Name: Q4B, dtype: float64
```



X variables


문27. 귀하의 정치·경제·사회적인 위치는 어느 계층에 속한다고 생각하십니까?

⑦ ⑥ ⑤ ④ ③ ②

상상 상하 중상 중하 하상 하하

문11. 귀하가 거주하는 지역의 *녹지공원, 숲 등*에 대해 얼마나 *만족*하십니까?

⑤ 매우 만족 ④ 약간 만족 ③ 보통이다

② 약간 불만족 ① 매우 불만족

■ 교통 이용 만족도

문45. 귀하는 서울의 *교통수단*에 대해 어느 정도 *만족*하십니까? 교통수단의 *쾌적성, 정시성, 친절함* 등을 *전반적*으로 생각하여 다음 각 항목에 대한 *귀하의 생각*을 말씀해 주십시오.

매우 만족	약간 만족	보통 이다	약간 불만족	매우 불만족	이용하지 않음
5	4	3	2	1	9

1) 버스(시내버스, 마을버스, 좌석버스 등)	
2) 지하철	
3) 택시	

■ 개인성향

문41. 귀하는 어느 정도 보수적 또는 진보적이라고 생각하십니까?

⑪ ⑩ ⑨ ⑧ ⑦ ⑥ ⑤ ④ ③ ② ①

매우 진보 중도 매우 보수

서울시민 고향 인식도
서울축제에 대한 인지도
사회자본
도시위험도
문화활동
환경
여가활동형태
운동
평생교육
자원봉사활동
계층인식
사회적 신뢰
교통이용만족도
은퇴시기
…

변수삭제

1)정보 없음
Ex)FC 가구원코드

2)겹치는 정보
Ex) 출생년도-나이
지역분류 변수 다수 존재

3)행복지수와 관련 없는 변수
Ex) 여가활동의 종류/ 공공시설 불편이유/ 마을공동체관련 질문

변수변환

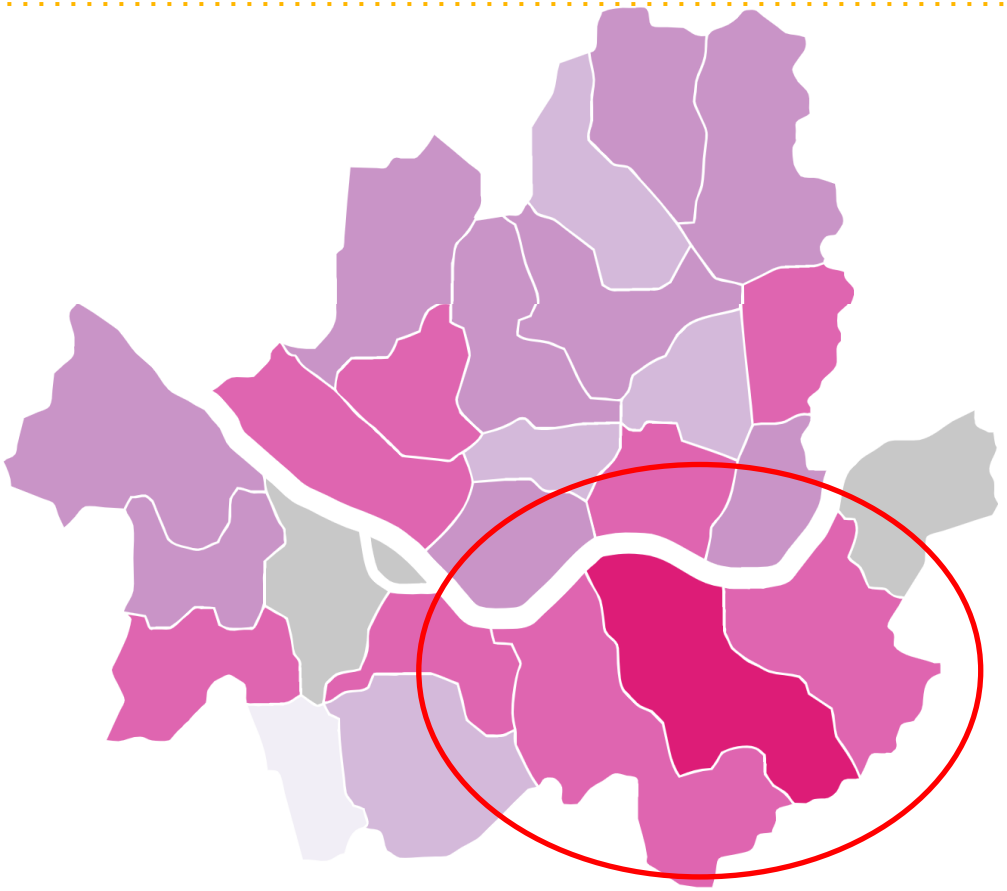
1)카테고리가 너무 많다

자가/전세/반전세/월세/상속 -> 자가/전세/기타
단체활동경험 -> 단체활동 개수
박물관,전시회 방문횟수/비용 -> 문화체험 연간 방문횟수
1순위 /2순위 -> 1순위만 선택
응답내 의미가 겹친다 -> 묶자

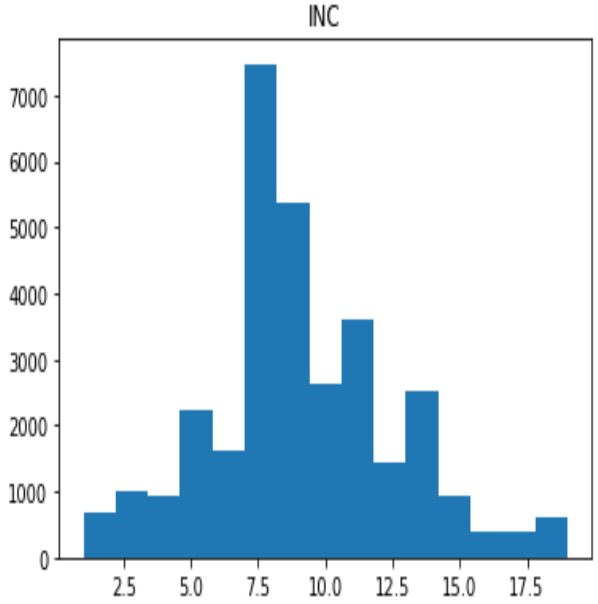
2) 이어지는 문항

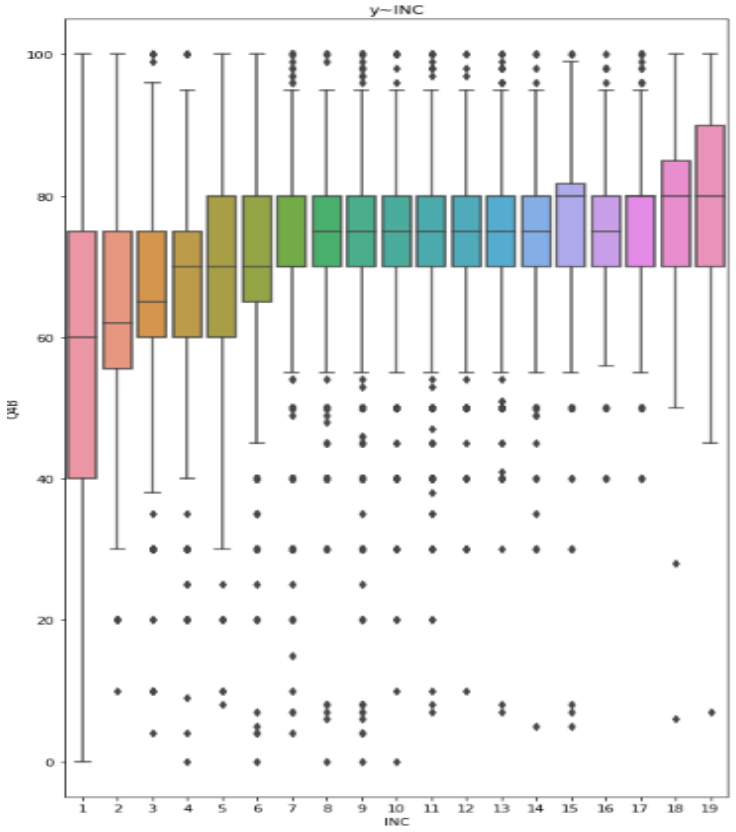
축제경험있다-> 3번문항으로 가시오
-> 관련변수 합쳐서 카테고리화

행정구역

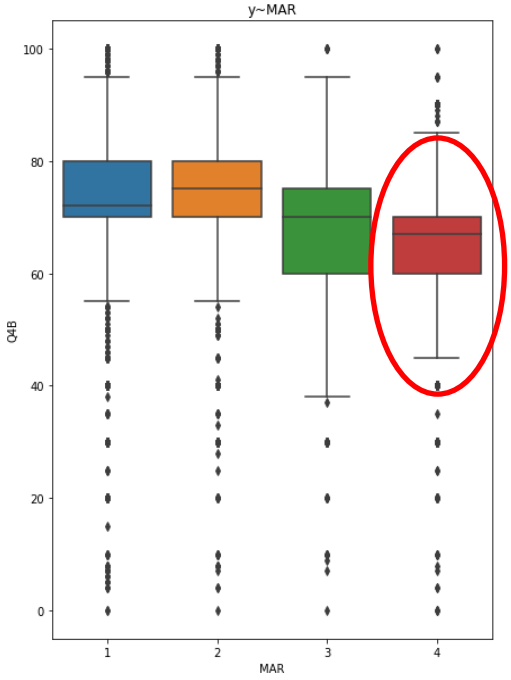


소득

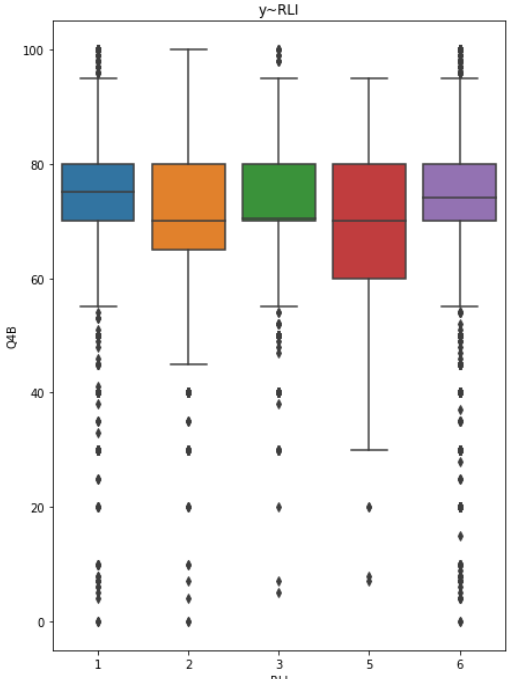

INC
7000
6000
5000
4000
3000
2000
1000
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


y~INC
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
INC

혼인상태/ 종교

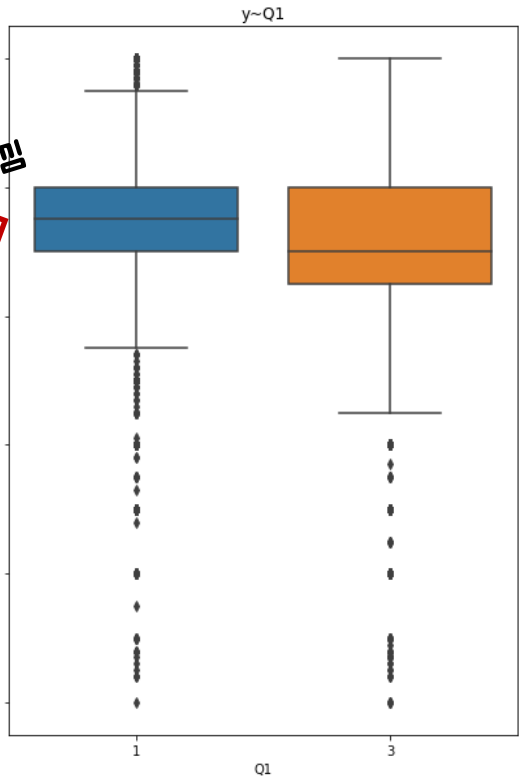


기혼 미혼 이혼 사별

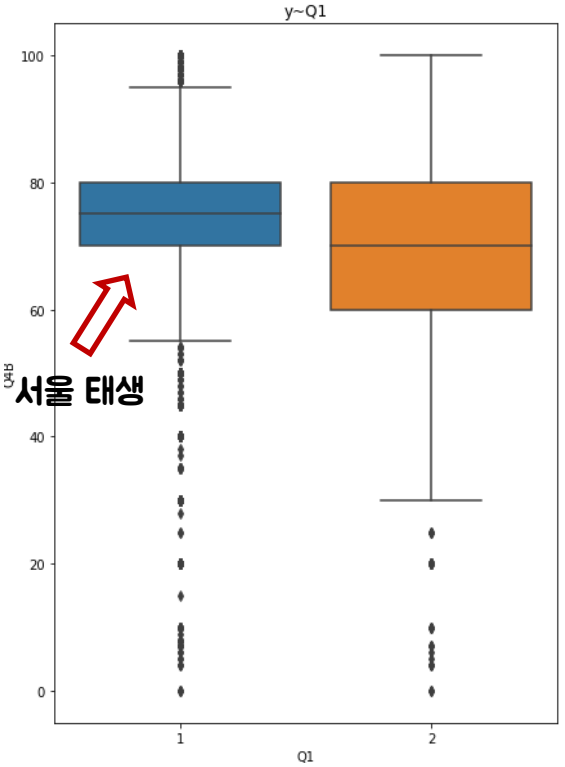


개신교 불교 천주교 기타 무교

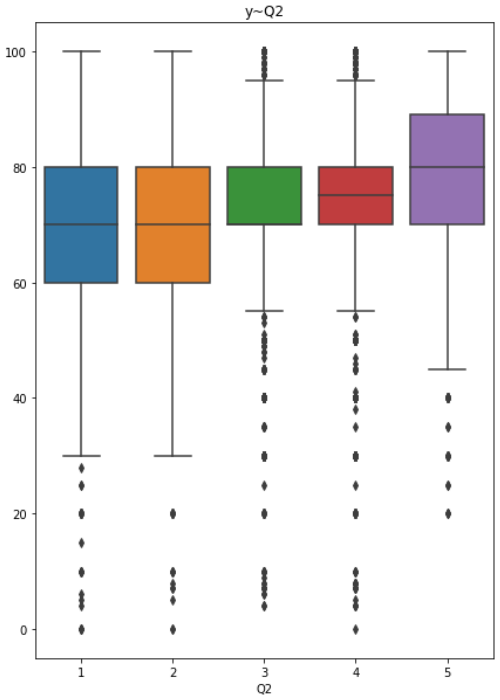
서울을 고향이라 느낌



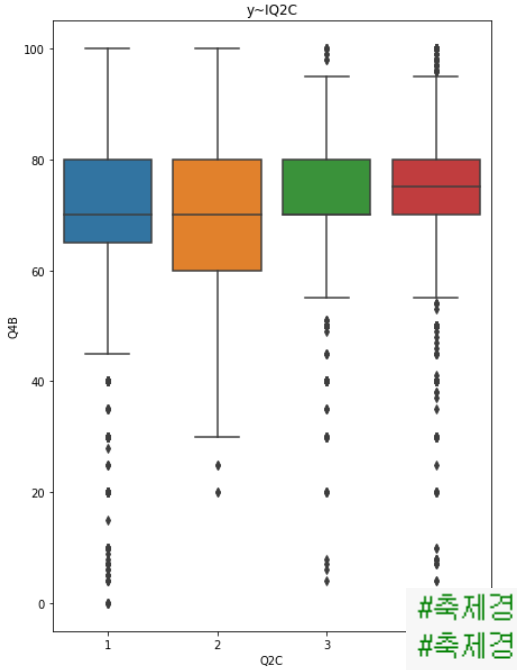
서울 태생



축제관련

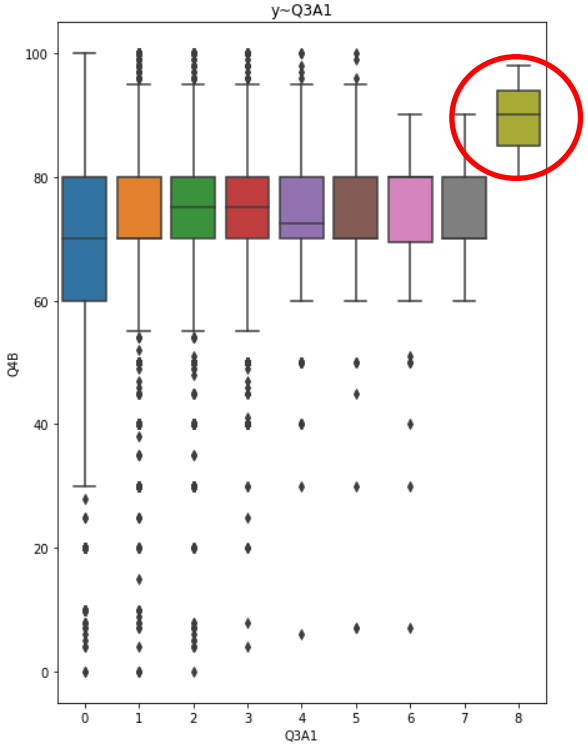


불만족 만족

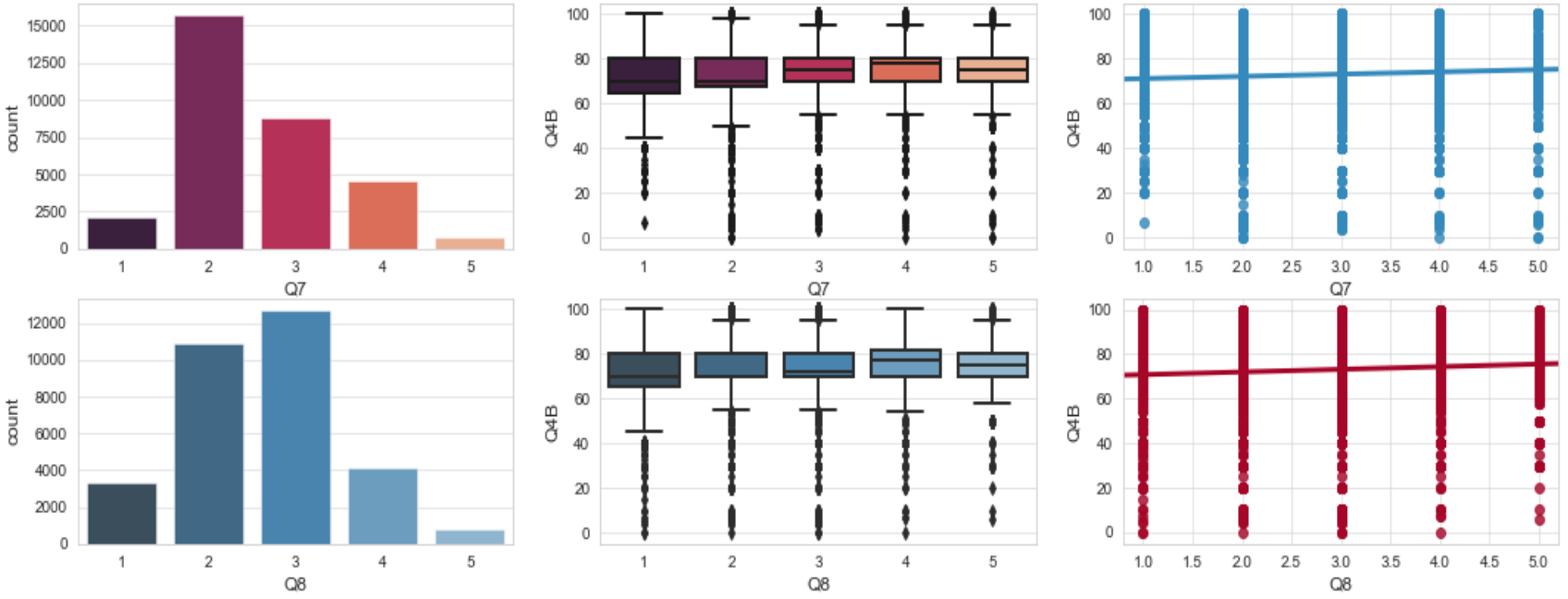


#축제경험있다(1)-만족(4)->4
#축제경험있다(1)-보통(3)->3
#축제경험있다(1)-불만족(2)->2
#축제경험없다(2)->1

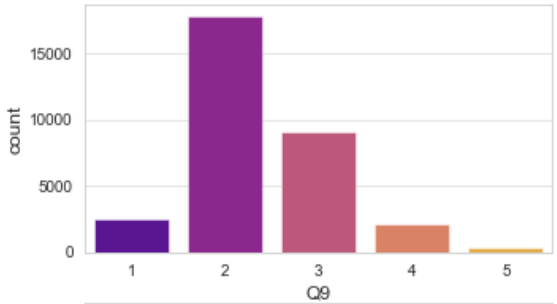
단체활동 개수

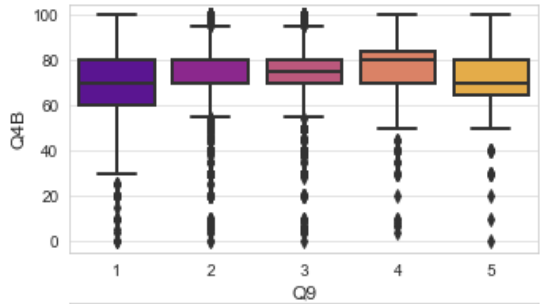


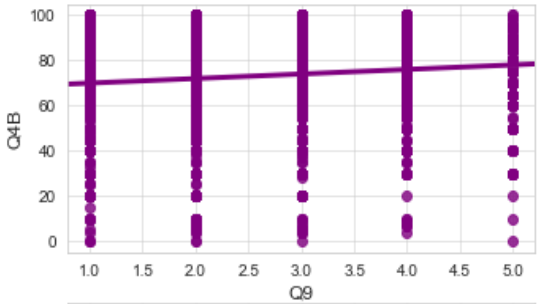
위험사회에 관한 인식

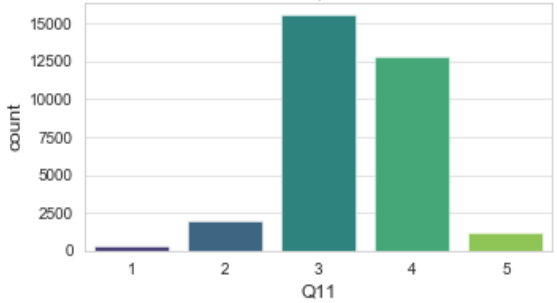


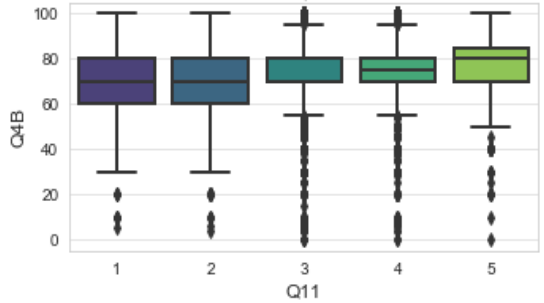
스트레스와 녹지만족도

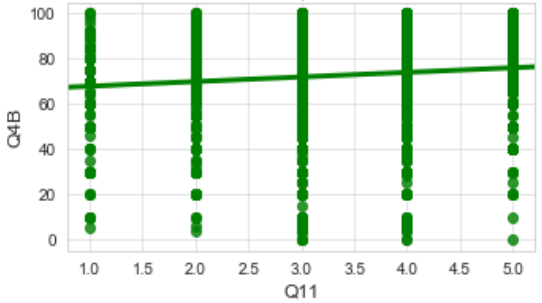






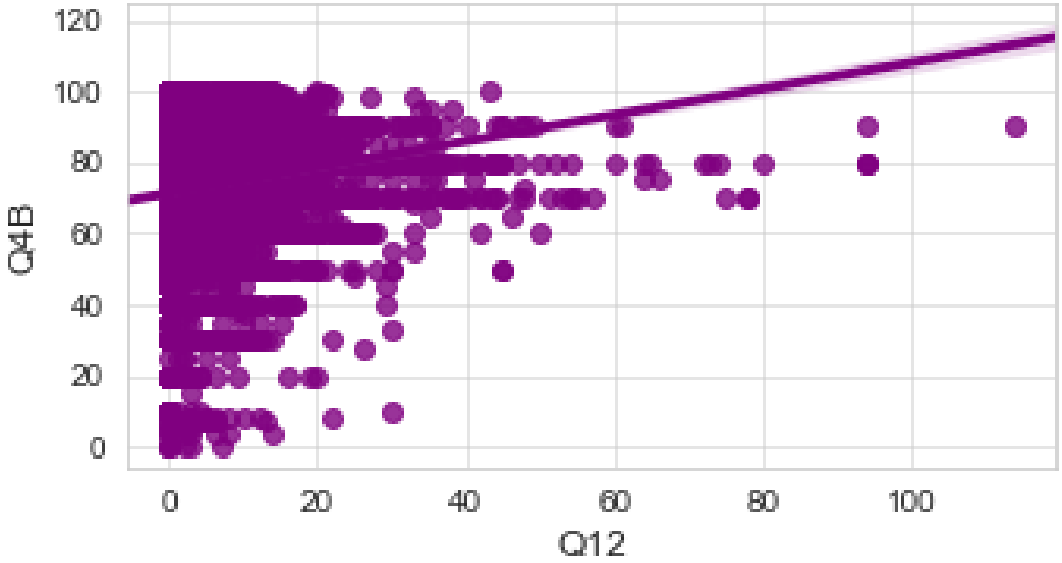






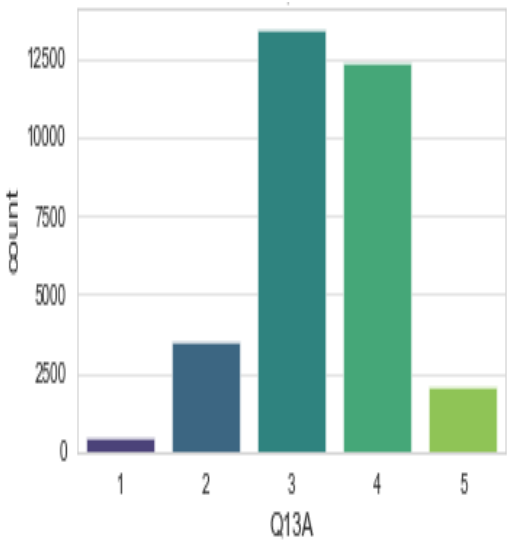
스트레스가 없고
녹지만족함

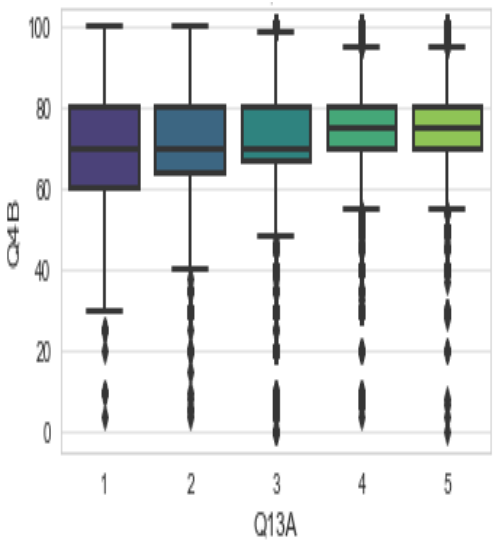
문화경험수

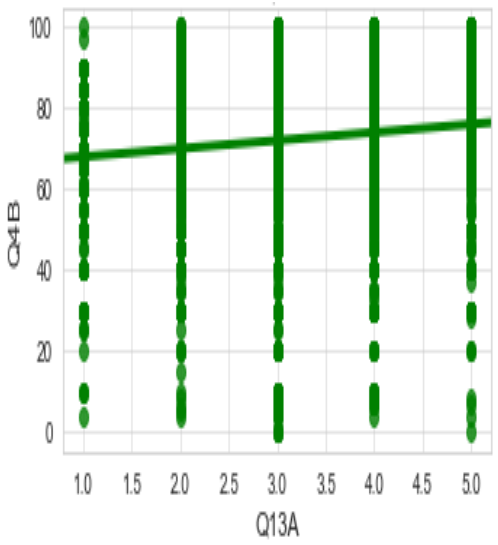


1년간 문화경험 총 횟수

문화환경 만족도

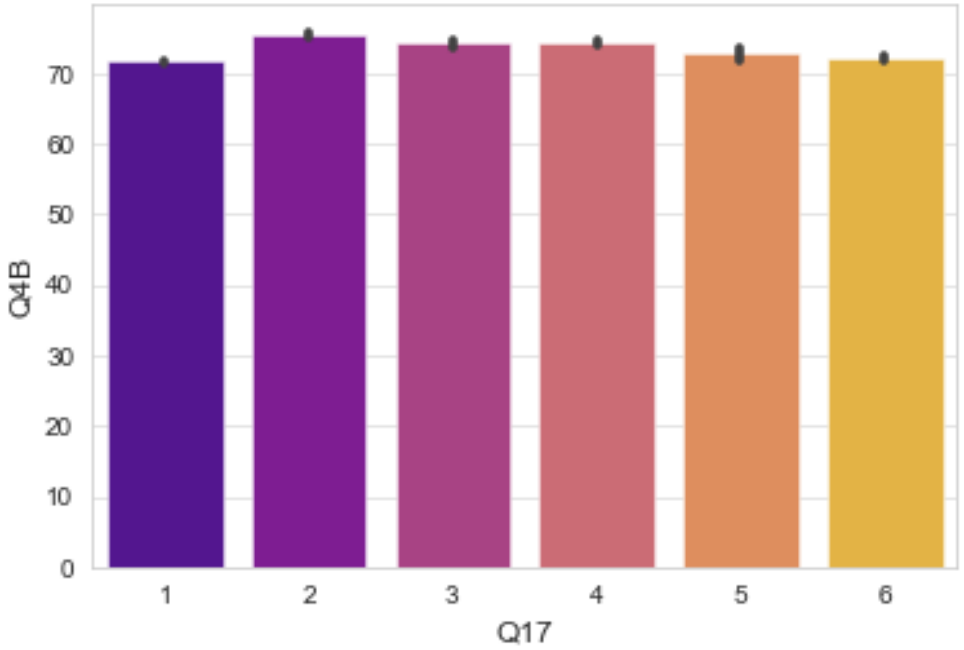






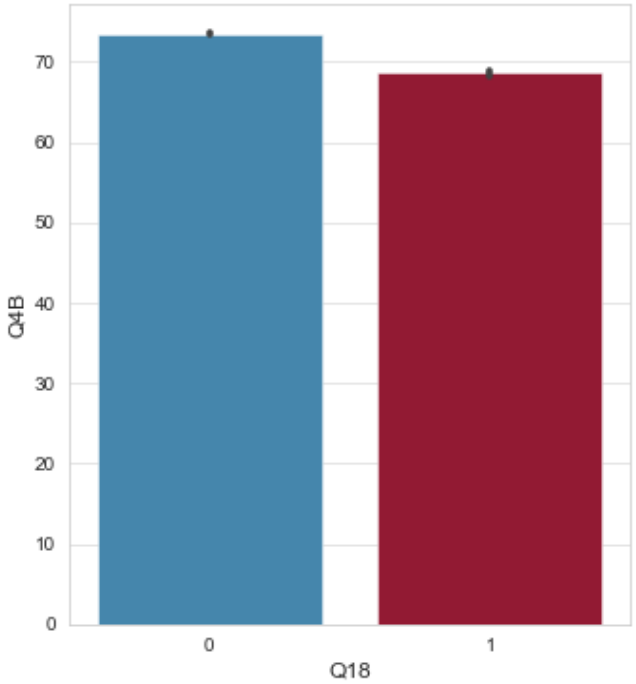
문화 시설 만족

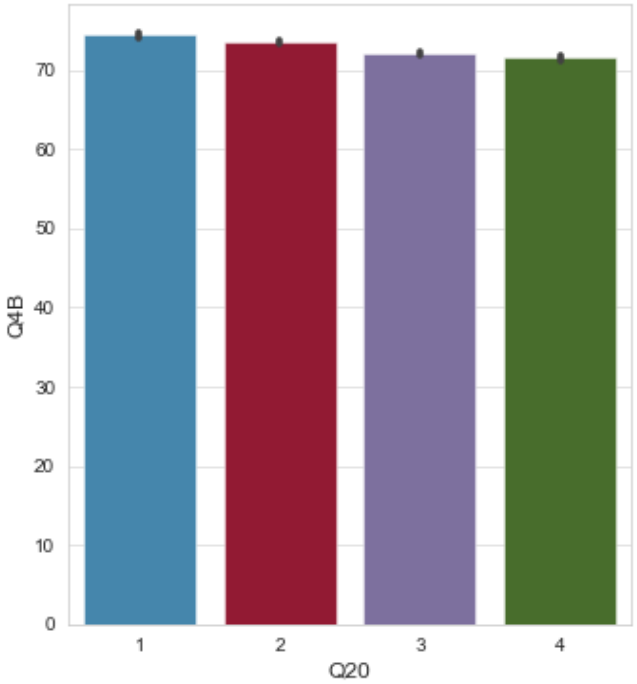
취미



1 : TV, 컴퓨터
2 : 문화,창작 활동
3 : 운동
4: 여행
5: 사회봉사
6 : 기타

기타

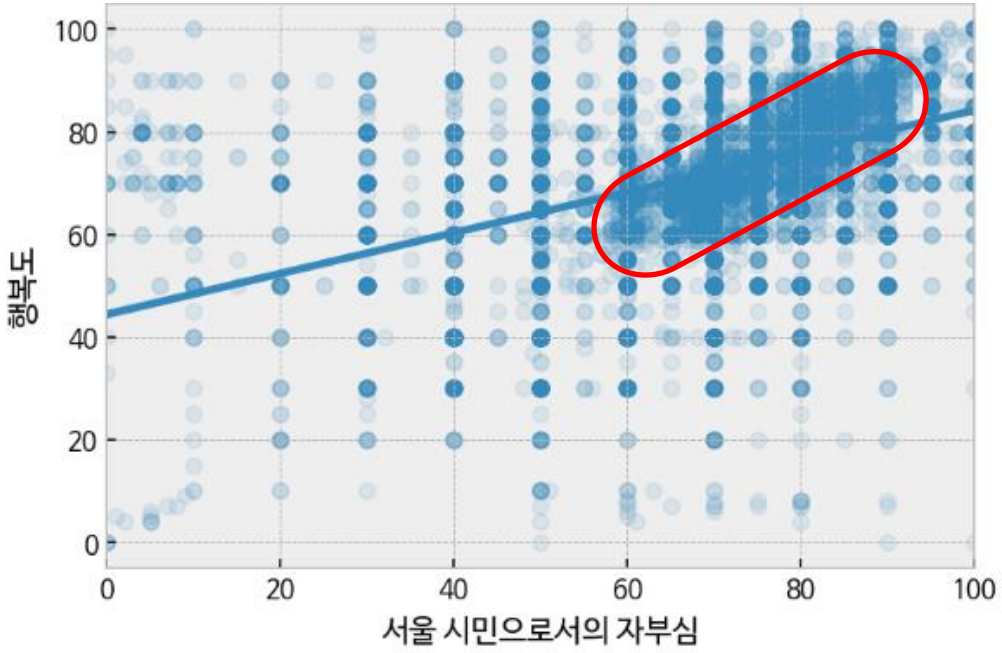




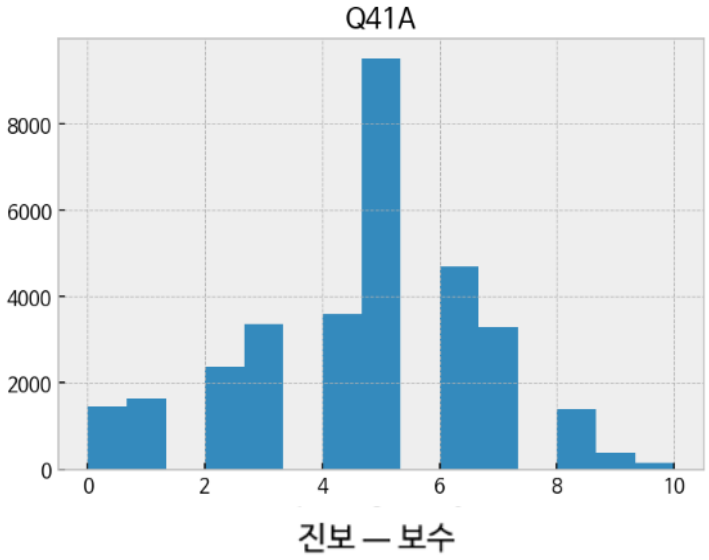
1 : 혼자 취미활동
0 : 혼자가 아님

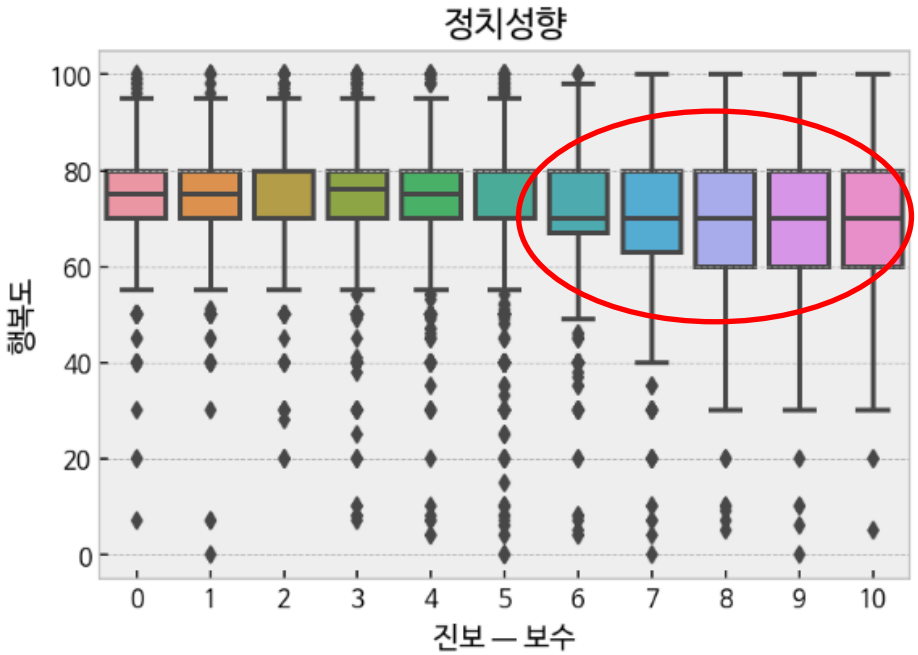
→ 4로 갈수록 운동량 적음

서울 시민 자부심

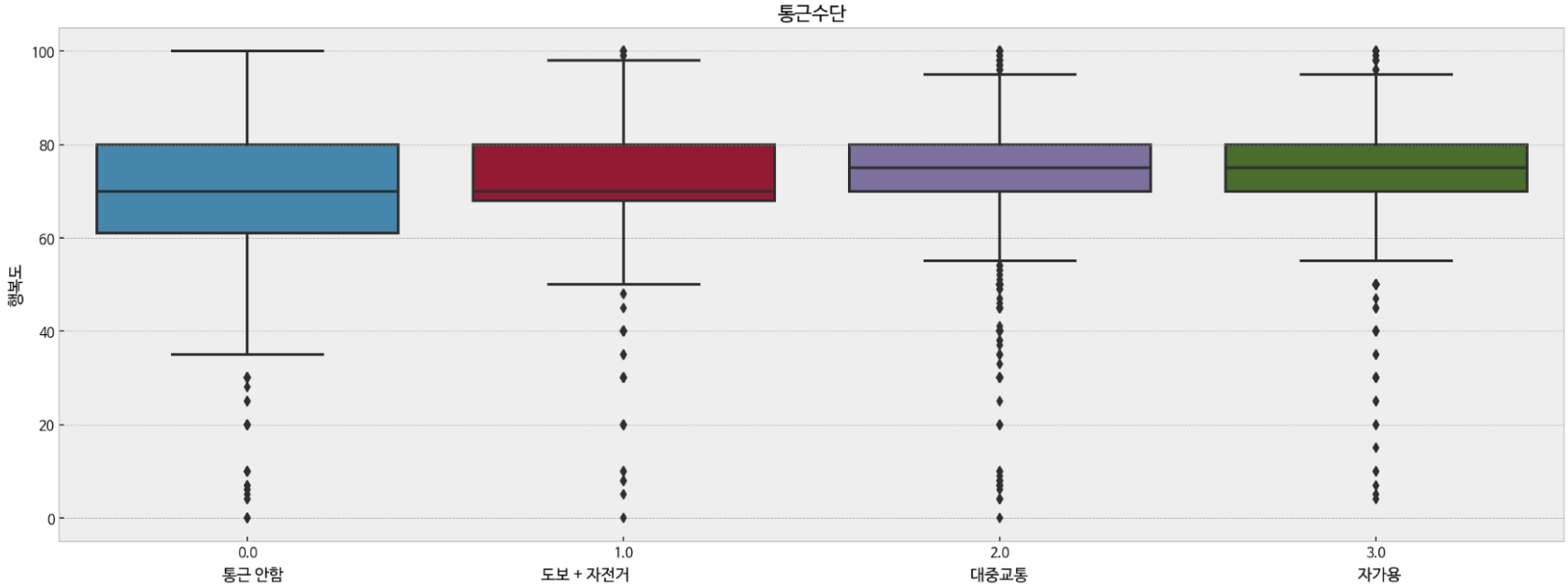


정치 성향

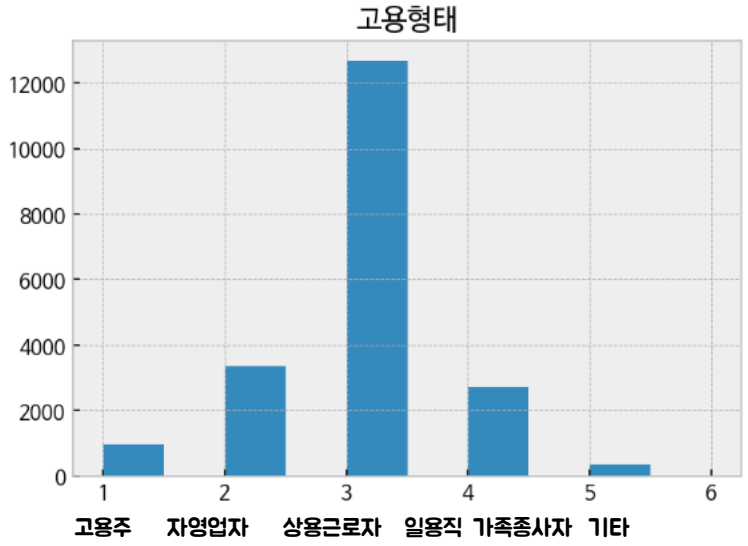

Q41A
8000
6000
4000
2000
0
0
2
4
6
8
10
진보 — 보수

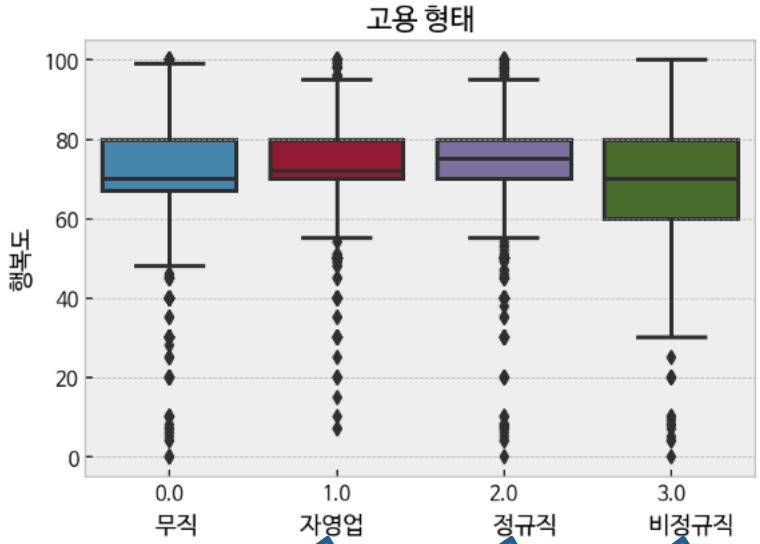

정치성향
100
80
60
40
20
0
보복행
0
1
2
3
4
5
6
7
8
9
10
진보 — 보수

통근 수단

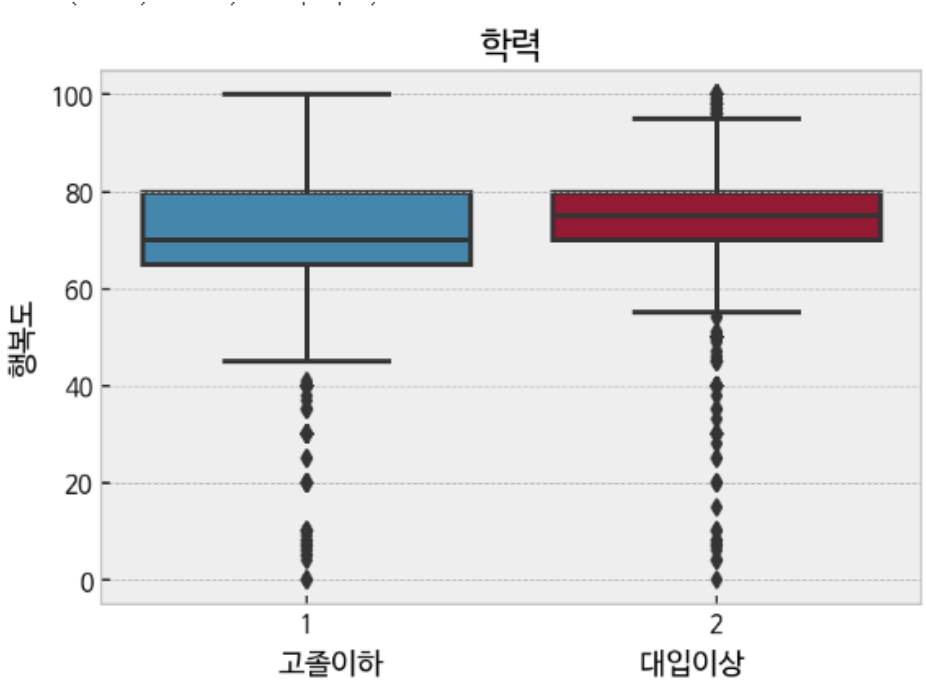


고용 형태


고용형태
12000
10000
8000
6000
4000
2000
0
1
2
3
4
5
6
고용주
자영업자
상용근로자
일용직
가족종사자
기타


고용 형태
행복도
100
80
60
40
20
0
0.0
1.0
2.0
3.0
무직
자영업
정규직
비정규직

학력



자전거 이용

문44. 귀하는 현재 *자전거를 이용* 하고 계십니까?

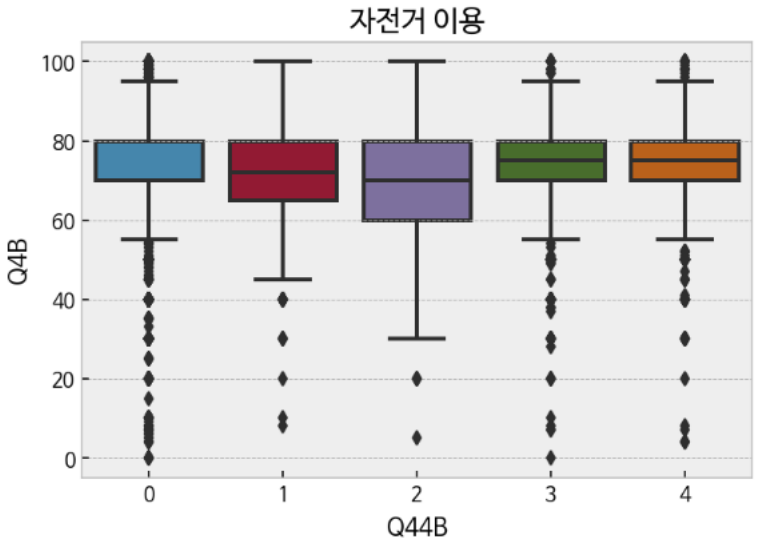
① 이동(출퇴근, 장보기, 업무 등) 수단으로 이용 ➡ **[문44-1로]**

② 운동(레저 등) 수단으로 이용 ➡ **[문44-1로]**

③ 이용하지 않음 ➡ **[문45로]**

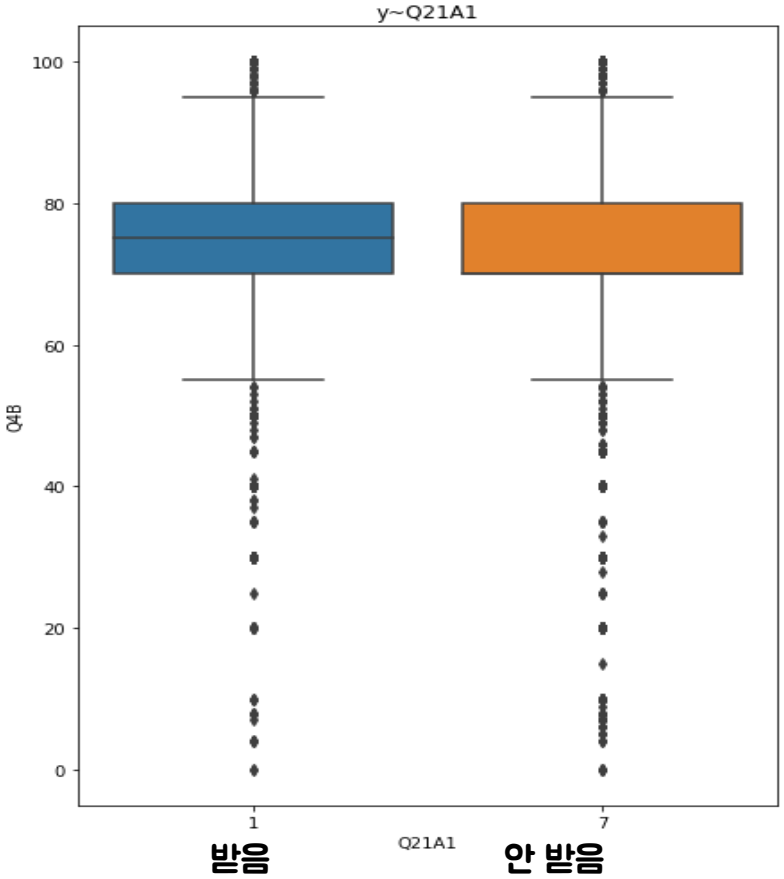
문44-1. 귀하는 *자전거 이용 환경*에 대해 전반적으로 *얼마나 만족* 하십니까?

⑤ 매우 만족 ④ 약간 만족 ③ 보통이다
② 약간 불만족 ① 매우 불만족

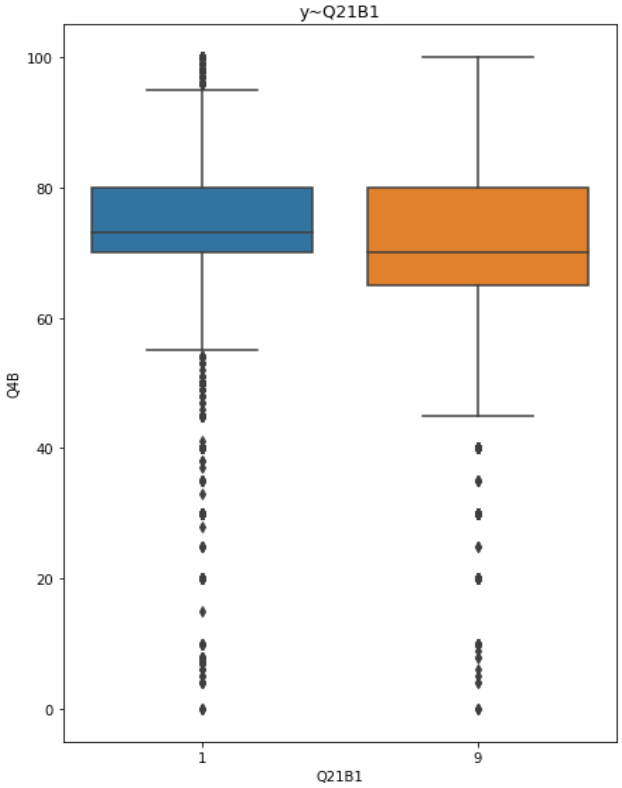


0. 자전거 이용 안함
1. 자전거 통근 + 환경만족
2. 자전거 통근 + 환경 불만족
3. 자전거 운동 + 환경 만족
4. 자전거 운동 + 환경 불만족

교육

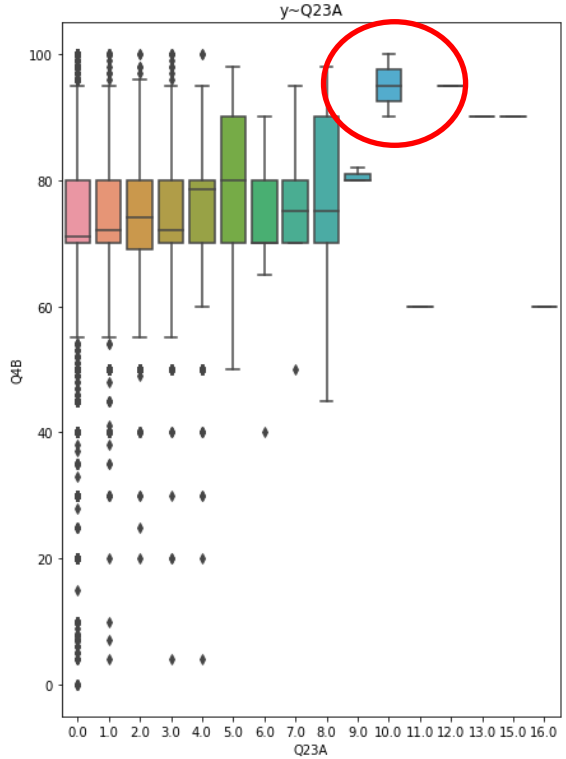


교육



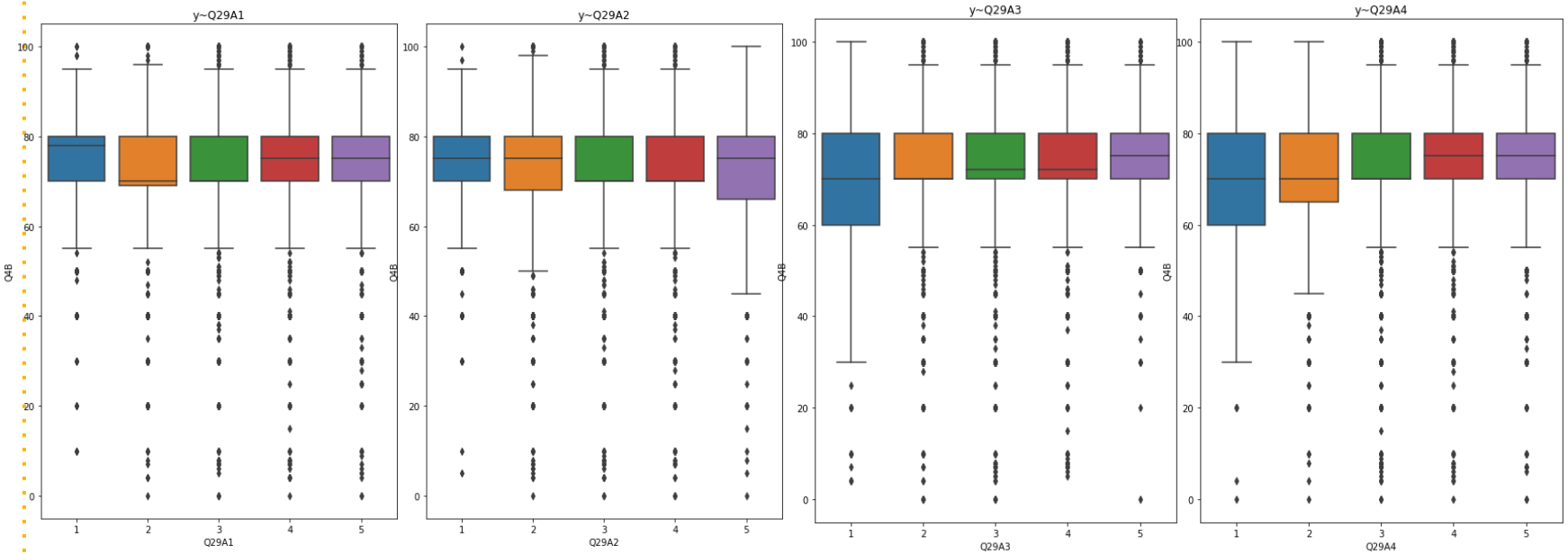
받을 의향　　받을 의향 X

자원봉사활동 관련


y~Q23A
Q4B
Q23A
0
20
40
60
80
100
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 15.0 16.0

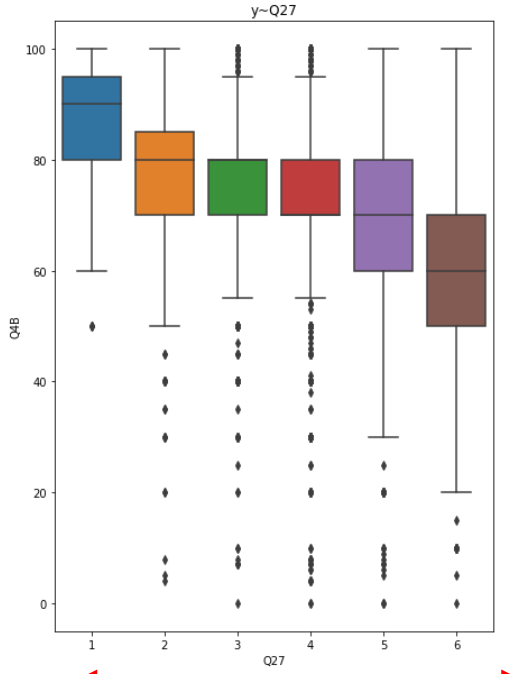

Q4B
Q23A
150
100
50
0
−50
−10
−5
0
5
10
15
20
25

사회적 약자에 대한 태도

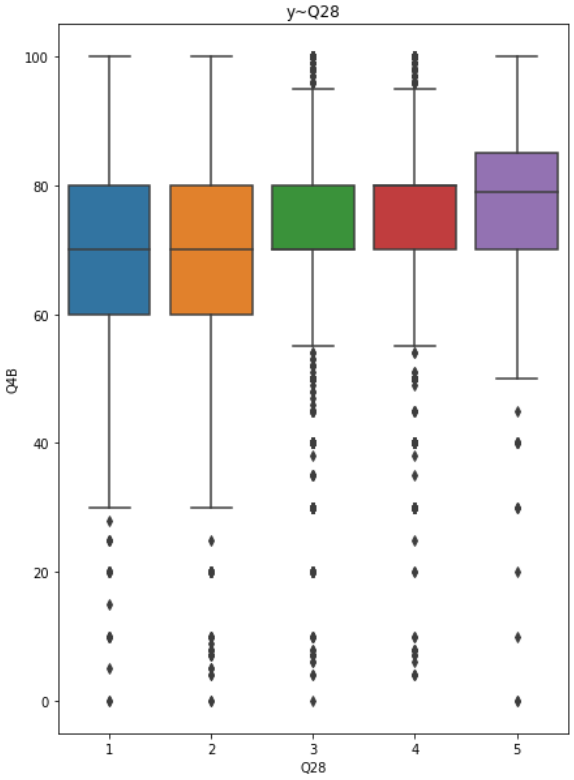


엄격 ← → 관대

계층 인식



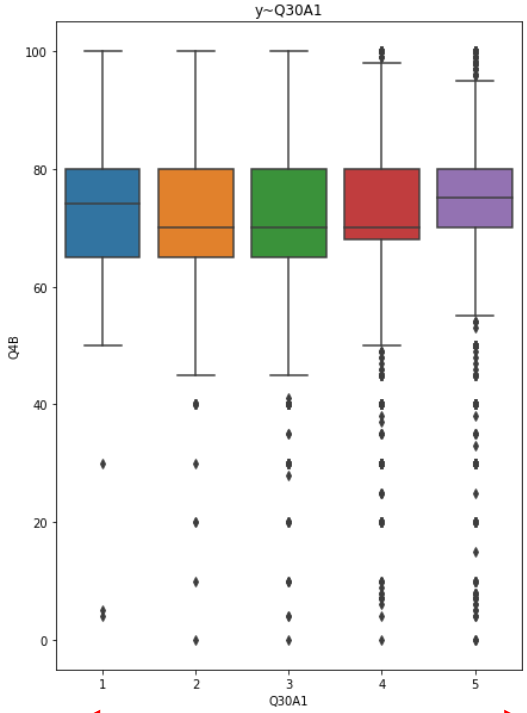
상상 하하



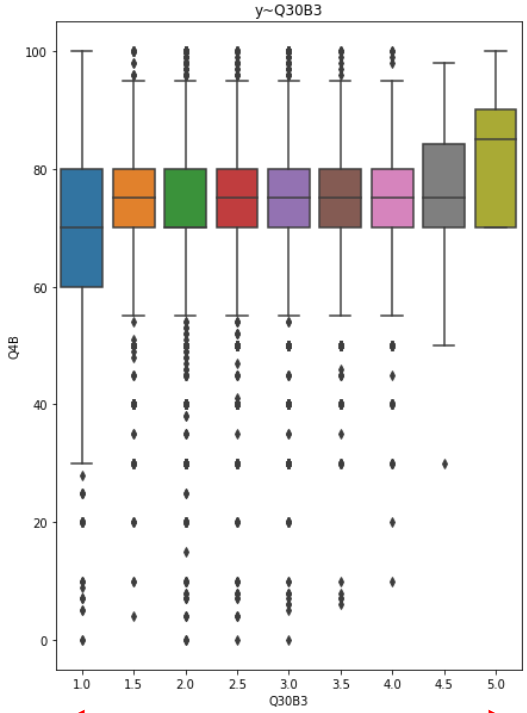
낮음 높음

사회적 신뢰

가족



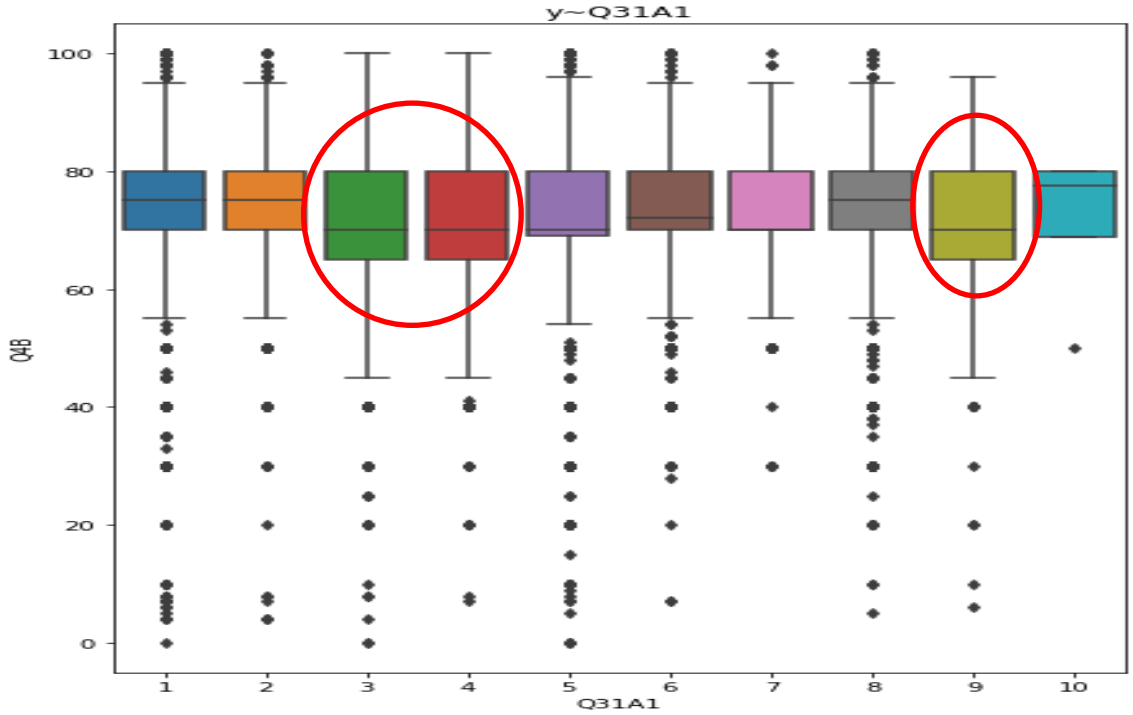
신뢰 X → 신뢰



신뢰 X → 신뢰

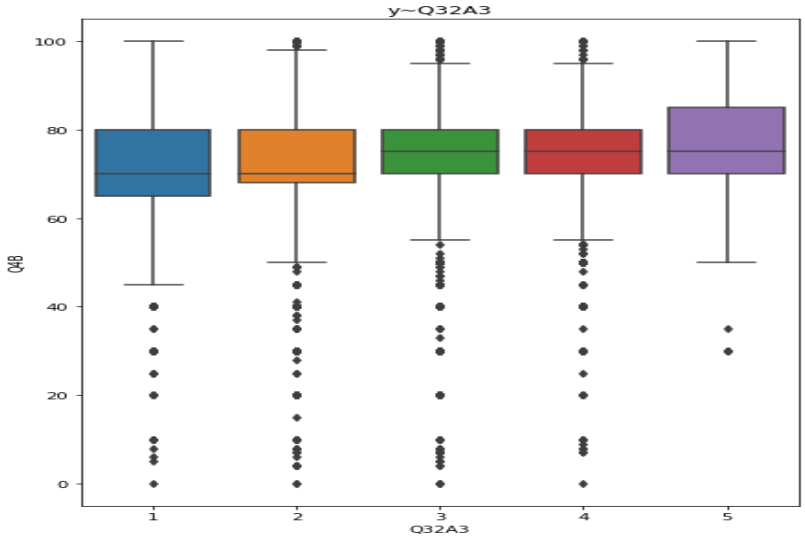
외부 사람
= 처음 만난 + 다른 나라 사람

사회적 차별

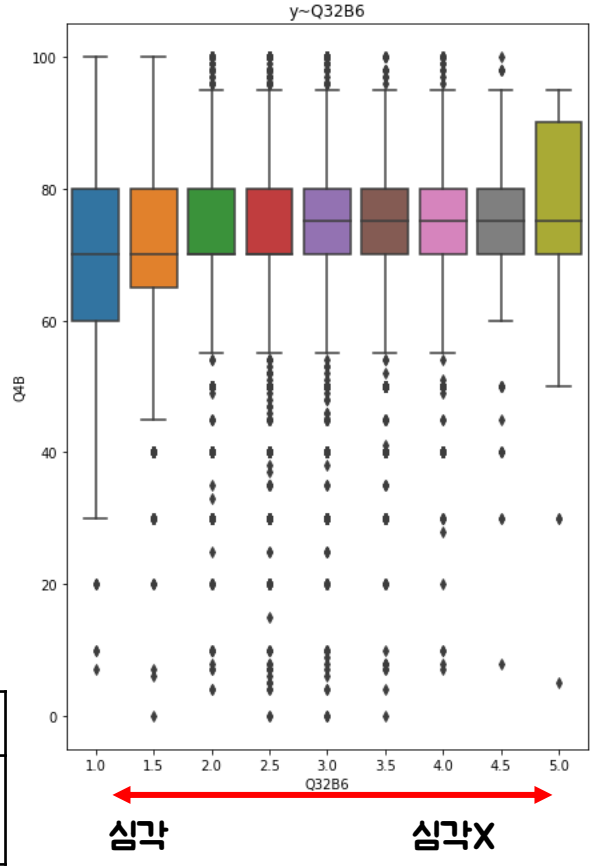


교육 국적 나이 성별 소득 외모 종교 직업 출신 기타

가족 관련 위험

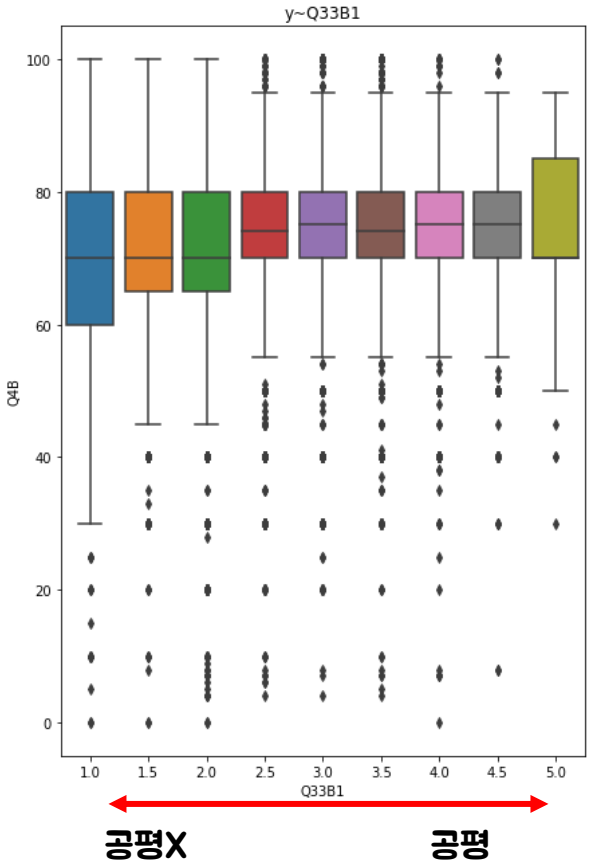


이혼 + 무자녀 또는 저출산	가족 형태
가족의 상부상조 기능 감소 + 재산 분배에 대한 갈등	가족 기능 (Q32B6)

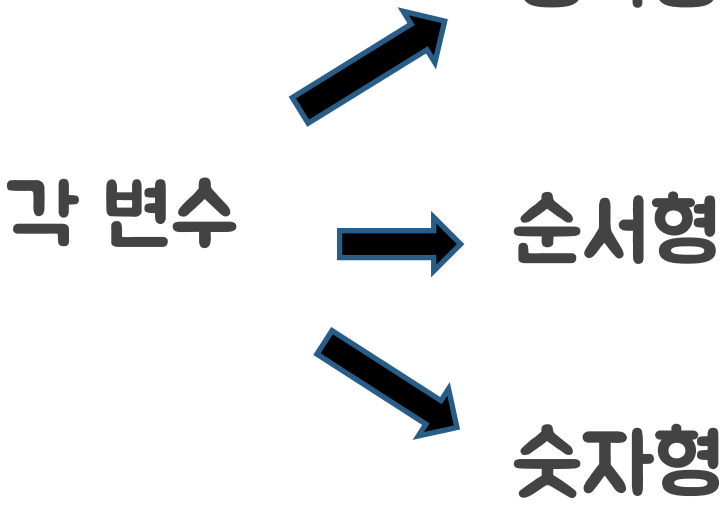


사회정의

수입과 소득 + 일자리/취업 기회	소득 (Q33B1)
대학교육의 기회	대학교육의 기회
수도권,지방의 발전 + 도시,농촌의 발전	지역 간 격차
사회복지 + 조세(세금) 정책	사회복지
남녀 평등 + 소수자의 권리	인권



각 변수 간의 독립성



카이제곱 검정
~ Cramer's V

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

스피어만 상관계수

Cramer's V란?

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

2개 이상의 '범주'로 나눈 집단
간의 상관계수

Cramer's V란?

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

$$V = \sqrt{\frac{\chi^2}{n\,(q-1)}}$$

χ^2 : 카이자승 공식에 의해 구함

n : 총사례수

q : 줄 또는 칸의 유목수 중에 적은 숫자

특징1) 측정치들의 분포에 상관없이 적용
특징2) 항상 양수
=> 카이제곱검정의 효과크기

Cramer's V란?

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

χ^2: 카이자승 공식에 의해 구함
n: 총 사례 수
q: 줄 또는 칸의 수 중 적은 숫자

Cramer's V란?

EFFECT SIZE: CRAMER'S V

Contingency tables larger than the 2 × 2 table (i.e., that have more than two rows and/or two categories) typically use another measure of effect size, Cramer's V. Like phi values, **Cramer's V** calculates effect size values that range between 0.000 and 1.000.

The judgment for the magnitude of Cramer's V does not always use the guidelines we discussed for the other effect size measures (0.100 for small, 0.300 for medium, and 0.500 for large) that pertain to the other tests. This is because there are adjustments made to the formula caused by the shape of the table. Cohen (1988) provides an adjusted set of guidelines in a series of tables. For the 2 × 3 table in our example, we can use the same 0.100, 0.300, and 0.500 guidelines for judging the effect size for Cramer's V as we did for C. Larger tables will have reduced magnitude effect size criteria that determine small, medium, and large effects.[4] (For example, if our table had been 3 × 4, the effect size criteria for a large effect would be 0.354, not 0.500.)

[4] We do not have the space in this book to develop this matter given its technical nature. However, SPSS reports the actual significance appropriate to any size of table, so in a practical sense, we do not need to delve more deeply into the subject at this point.

0.00–0.10	Ne
0.10–0.20	We
0.20–0.40	Mo
0.40–0.60	Re
0.60–0.80	Strong
0.80–1.00	Very strong

Interpretation of Φ in Chi-statistics or Cramér's V

ér's V is an effect size measu
gly two categorical fields are

ffect size is calculated in the

etermine which field has the
ıbtract 1 from the number o
ultiply the result by the total
vide the chi-square value by
uare test of independence
ke the square root.

1. Interpretation of effect siz

t size	Interpretation
ES ≤ 0.2	The result is weak / weakly associated
0.2 < ES ≤ 0.6	The result is mod
ES > 0.6	The result is strong. The fields are strongly associated.

Cramer's V Coefficient (V)

Useful for comparing multiple X^2 test statistics and is generalizable across contingency tables of varying sizes. It is not affected by sample size and therefore is very useful in situations where you suspect a statistically significant chi-square was the result of large sample size instead of any substantive relationship between the variables. It is interpreted as a measure of the relative (strength) of an association between two variables. The coefficient ranges from 0 to 1 (perfect association). In practice, you may find that a Cramer's V of .10 provides a good minimum threshold for suggesting there is a substantive relationship between two variables.

$$V = \sqrt{\frac{X^2}{n(q-1)}}$$ where q = smaller # of rows or columns

Describing Strength of Association

Characterizations

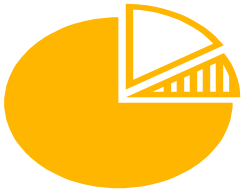
>.5	high association
.3 to .5	moderate association
.1 to .3	low association
0 to .1	little if any association

-of-PH-in-Chi-statistics-or-Cramers-V_tbl2_311335682

g=PA98&lpg=PA98&dq=cramer%27s+v+interpretation+large+size&source=bl&ots=ytqopG9M
ko&sa=X&ved=2ahUKEwjUgKaDiNTpAhWlHqYKHUkKDfMQ6AEwF3oECAgQAQ#v=onepage&q=c
false

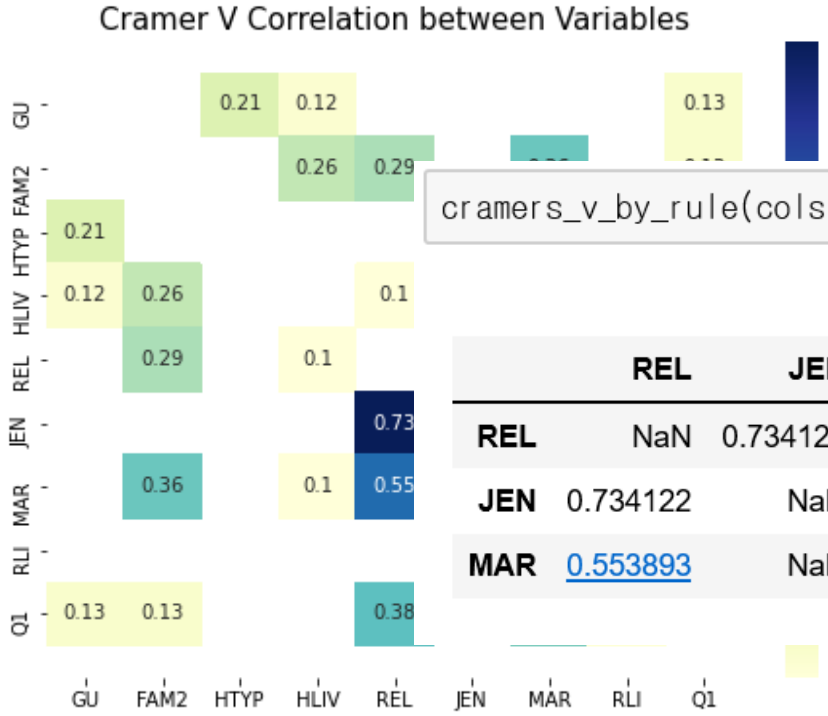
https://www.researchgate.net/figure/Interpretation-of-Phi-and-Cramers-V_tbl2_326885374

Cramer's V란?



```
def direct_cramer_v_interpret(X, Y):
    table = make_contingency_table(X, Y)
    cramer = cramers_stat(table)
    if cramer < 0.1:
        print(X, ',', Y, ':', '연관성이 없다.')
    elif cramer < 0.3:
        print(X, ',', Y, ':', '약한 연관성이 있다.')
    elif cramer < 0.5:
        print(X, ',', Y, ':', '보통의 연관성이 있다.')
    else:
        print(X, ',', Y, ':', '강한 연관성이 있다.')
```

Cramer's v

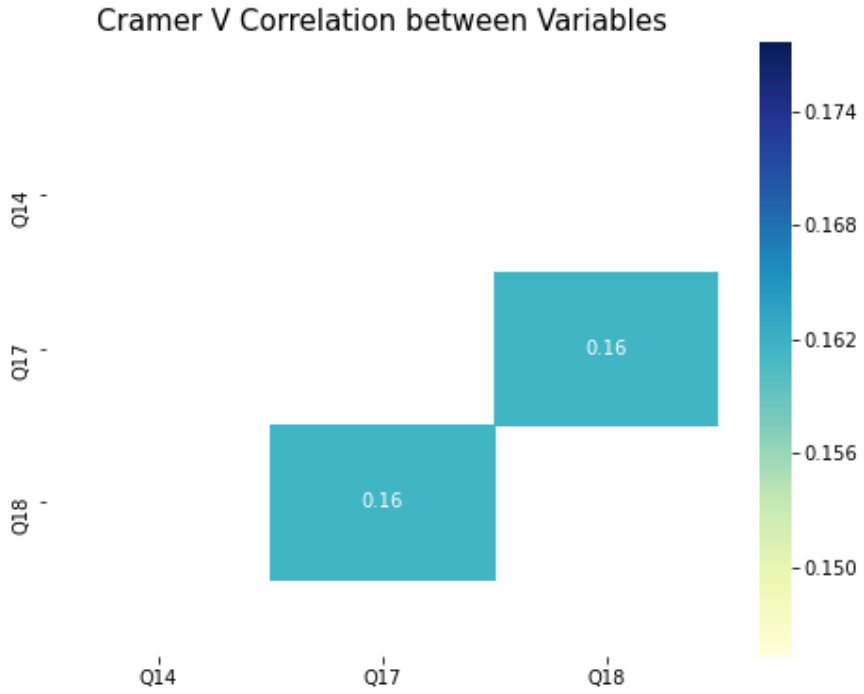



```
cramers_v_by_rule(cols)
```

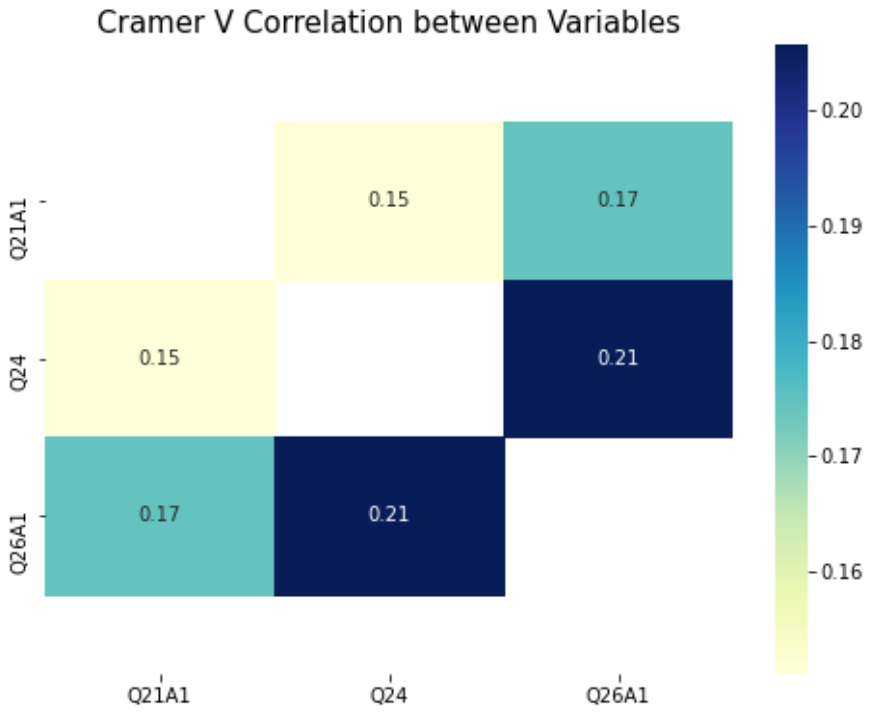
	REL	JEN	MAR
REL	NaN	0.734122	0.553893
JEN	0.734122	NaN	NaN
MAR	0.553893	NaN	NaN

	FAM2	REL	MAR	Q1
FAM2	NaN	NaN	0.363610	NaN
REL	NaN	NaN	NaN	0.381702
MAR	0.36361	NaN	NaN	0.344918
Q1	NaN	0.381702	0.344918	NaN

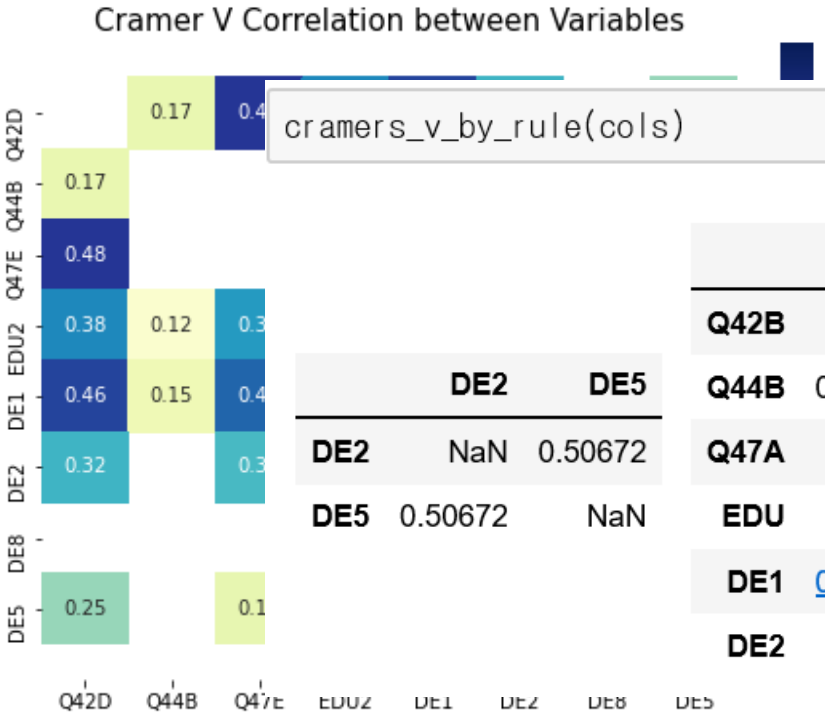
Cramer's v


Cramer V Correlation between Variables
Q14
Q17
Q18
0.16
0.16
0.174
0.168
0.162
0.156
0.150
Q14
Q17
Q18

Cramer's v


Cramer V Correlation between Variables
Q21A1
Q24
Q26A1
0.15
0.17
0.15
0.21
0.17
0.21
Q21A1
Q24
Q26A1
0.20
0.19
0.18
0.17
0.16

Cramer's v




```
cramers_v_by_rule(cols)
```

	DE2	DE5
DE2	NaN	0.50672
DE5	0.50672	NaN

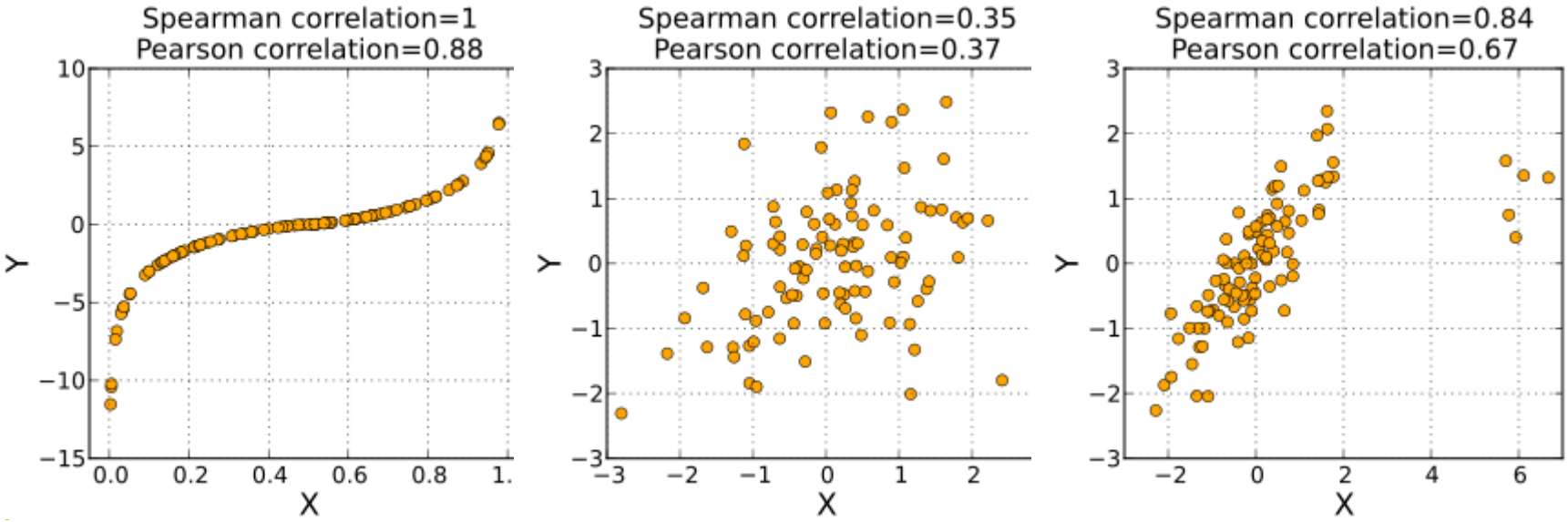
	Q42B	Q44B	Q47A	EDU	DE1	DE2
Q42B	NaN	0.349537	NaN	NaN	0.326015	NaN
Q44B	0.349537	NaN	NaN	NaN	NaN	NaN
Q47A	NaN	NaN	NaN	NaN	0.316307	NaN
EDU	NaN	NaN	NaN	NaN	NaN	0.311649
DE1	0.326015	NaN	0.316307	NaN	NaN	NaN
DE2	NaN	NaN	NaN	0.311649	NaN	NaN

Spearman Rank Correlation

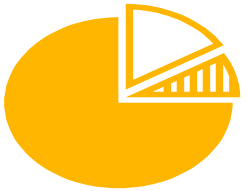
$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}}$$

스피어만 상관 계수는 두 변수의 순위 사이의 통계적 의존성을 측정하는 비모수적인 척도이다. 이는 두 변수의 관계가 단조 함수를 사용하여 얼마나 잘 설명될 수 있는지를 평가한다. 스피어만 상관 계수는 순위가 매겨진 변수 간의 피어슨 상관 계수 로 정의된다.

Spearman Rank Correlation

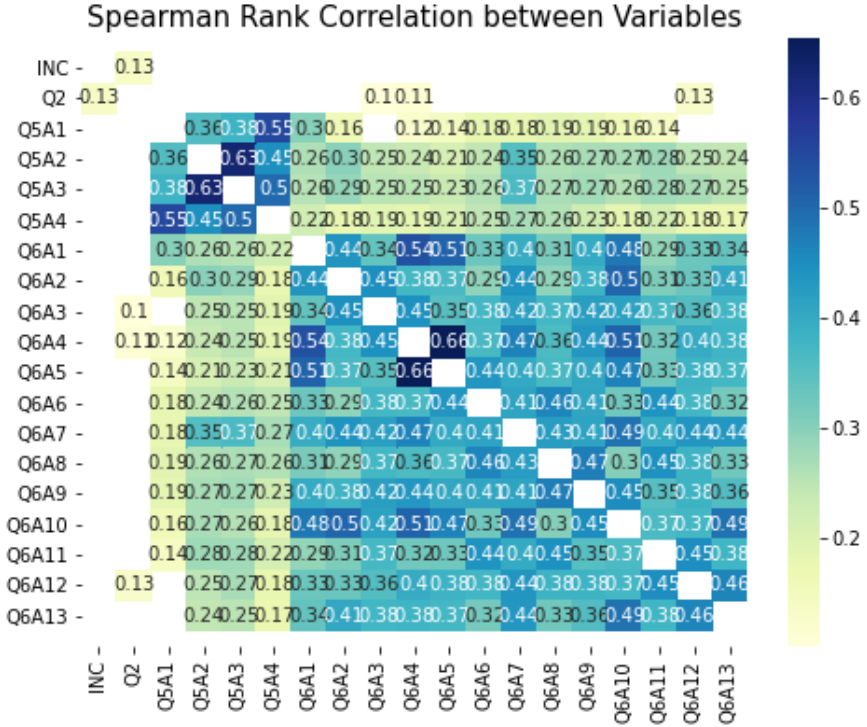


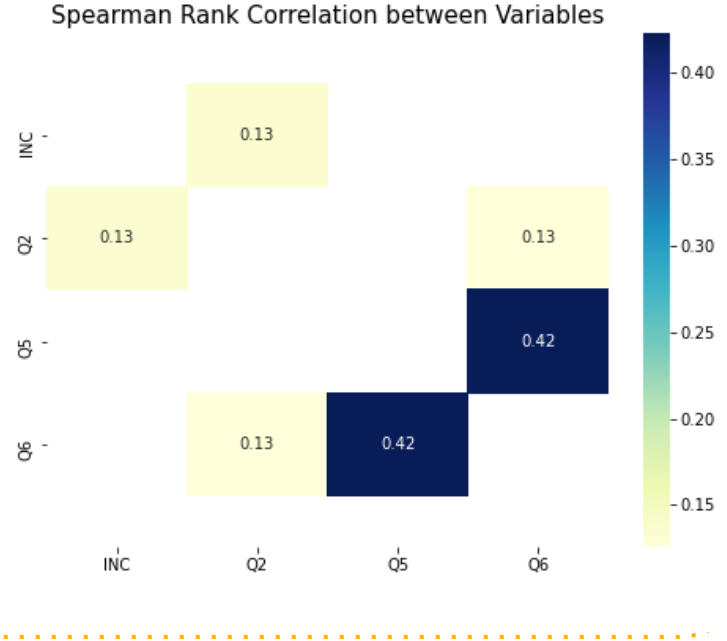
Spearman Rank Correlation



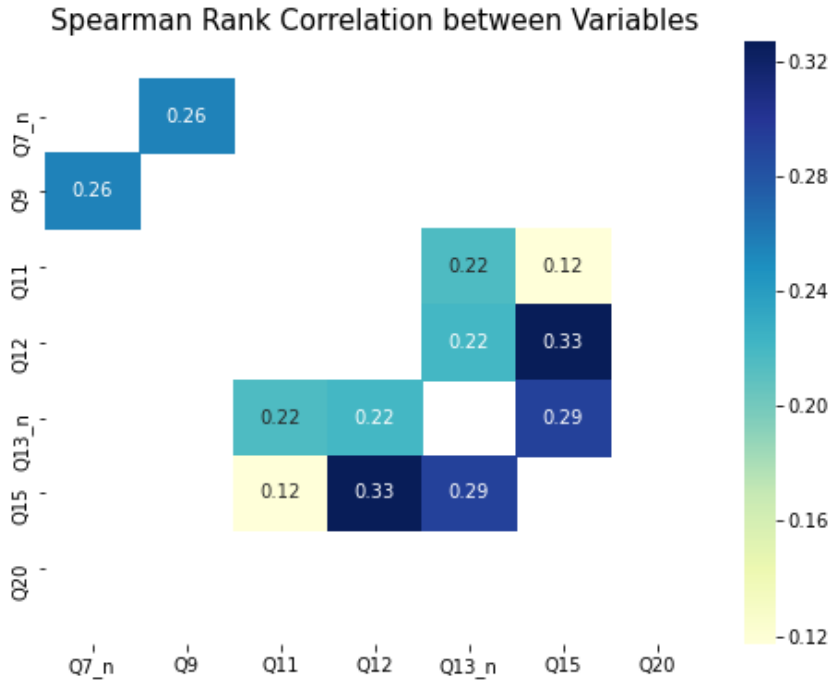
```
def spearmanr_interpret(X, Y):
    spearman_rho = calculate_spearmanr(X, Y)
    if abs(spearman_rho) < 0.1:
        print(X, ',', Y, ':', '연관성이 없다.')
    elif abs(spearman_rho) < 0.3:
        print(X, ',', Y, ':', '약한 연관성이 있다.')
    elif abs(spearman_rho) < 0.5:
        print(X, ',', Y, ':', '보통의 연관성이 있다.')
    else:
        print(X, ',', Y, ':', '강한 연관성이 있다.')
```

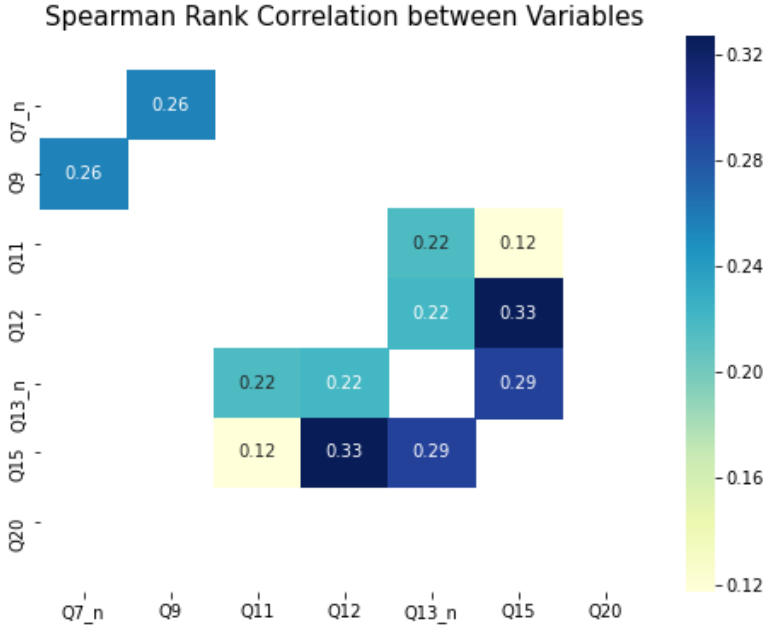
Spearman 상관계수



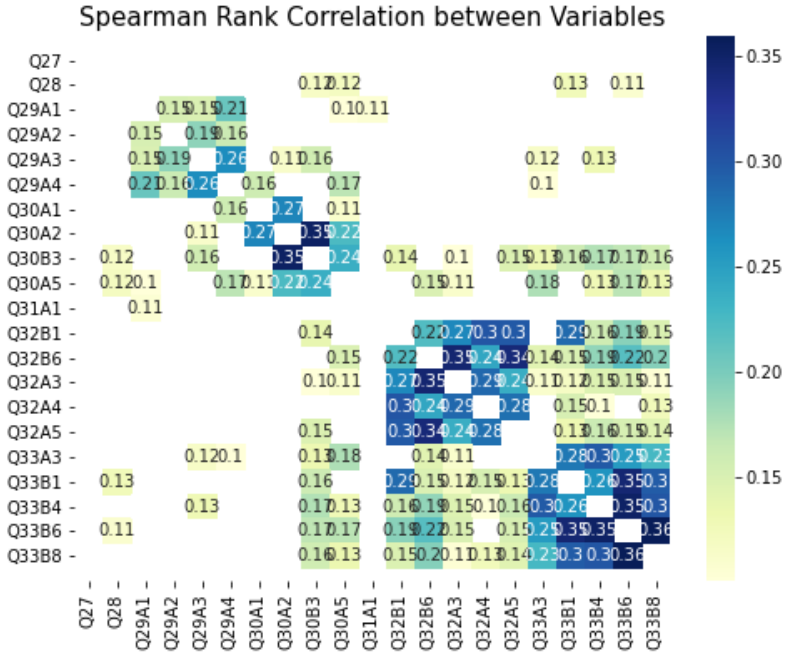


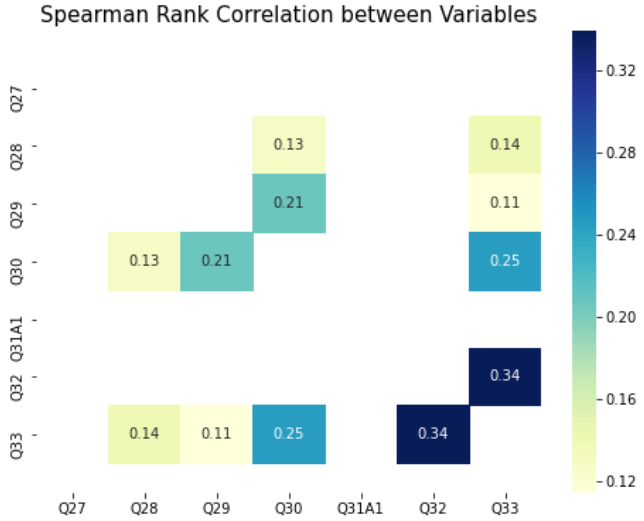
Spearman 상관계수



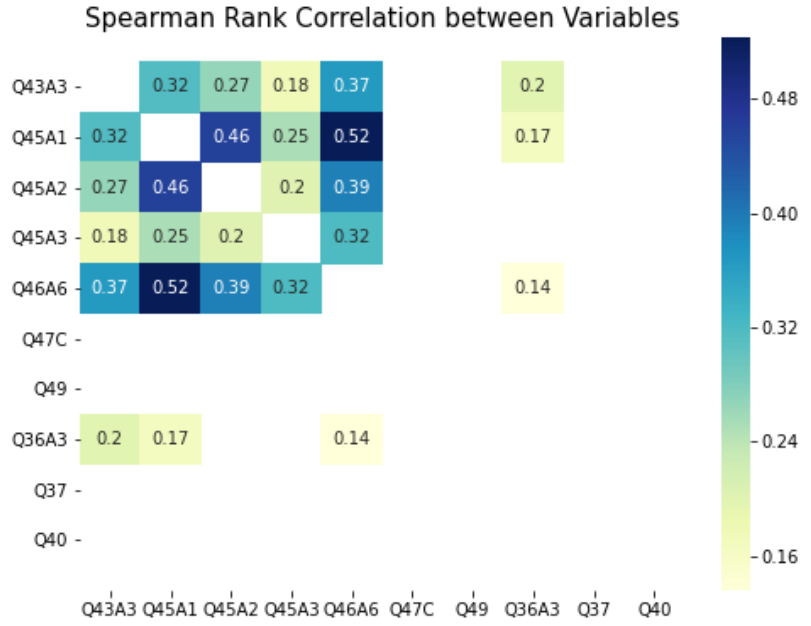


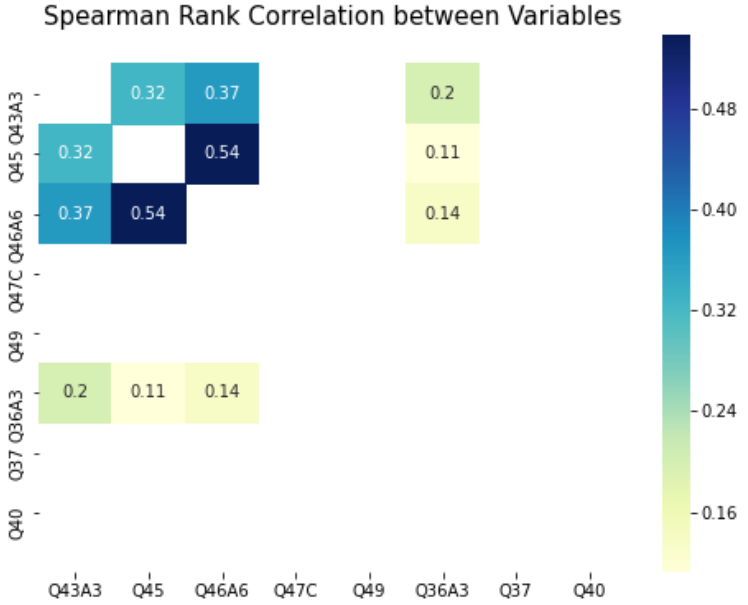
Spearman 상관계수





Spearman 상관계수

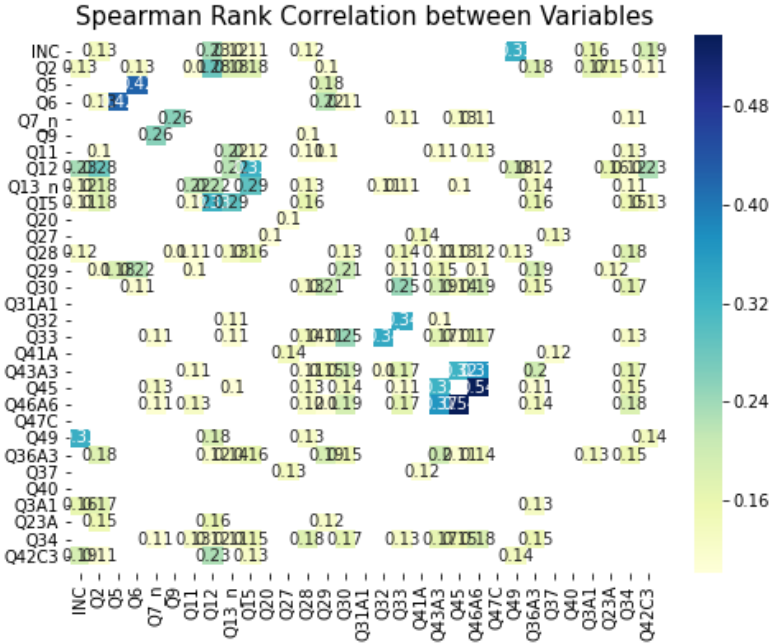




Spearman 상관계수

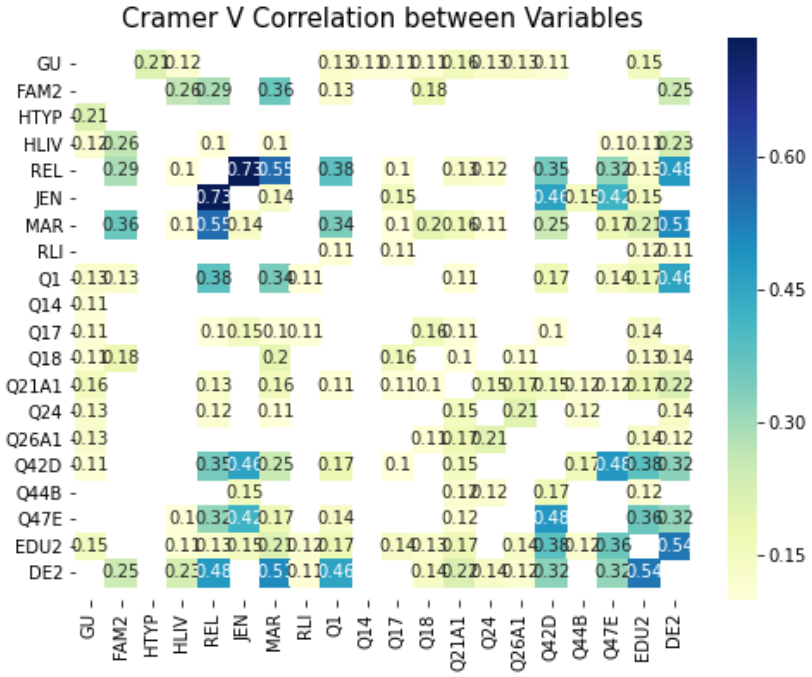
```
#Q5
data = data.assign(Q5 = np.mean(data.loc[:,'Q5A1':'Q5A4'], axis=1)) # A1~A4 평균값
data.drop(['Q5A1','Q5A2','Q5A3','Q5A4'], axis=1) #기존 변수를 제거 #inplcae=True는

#Q6.
data = data.assign(Q6 = np.mean(data.loc[:,'Q6A1':'Q6A13'], axis=1)) # A1~A13평균값
data.drop(['Q6A1','Q6A2','Q6A3','Q6A4','Q6A5','Q6A6','Q6A7','Q6A8','Q6A9','Q6A10','

#Q7,8
data = data.assign(Q7_n = np.mean(data.loc[:,'Q7':'Q8'], axis=1)) # Q7~Q8평균값으로
data.drop(['Q7','Q8'], axis=1) #기존 변수를 제거 #inplcae=True는 아직 안 넣음

#Q13,Q13A
data = data.assign(Q13_n = np.mean(data.loc[:,'Q13':'Q13A'], axis=1)) # Q13~Q13A평
data.drop(['Q13','Q13A'], axis=1) #기존 변수를 제거 #inplcae=True는 아직 안 넣음

#Q29
data = data.assign(Q29 = np.mean(data.loc[:,'Q29A1':'Q29A4'], axis=1)) # Q13~Q13A평
data.drop(['Q29A1','Q29A2','Q29A3','Q29A4'], axis=1) #기존 변수를 제거 #inplcae=Tru

#Q30
data = data.assign(Q30 = np.mean(data.loc[:,'Q30A1':'Q30A5'], axis=1)) # Q13~Q13A평
data.drop(['Q30A1','Q30A2','Q30A3','Q30A4','Q30A5'], axis=1) #기존 변수를 제거 #inp

#Q32
data = data.assign(Q32 = np.mean(data.loc[:,'Q32A1':'Q32A7'], axis=1)) # Q13~Q13A평
data.drop(['Q32A1','Q32A2','Q32A3','Q32A4','Q32A5','Q32A6','Q32A7'], axis=1) #기존

#Q33
data = data.assign(Q33 = np.mean(data.loc[:,'Q33A1':'Q33A9'], axis=1)) # Q13~Q13A평
data.drop(['Q33A1','Q33A2','Q33A3','Q33A4','Q33A5','Q33A6','Q33A7','Q33A8','Q33A9']


#Q45
data = data.assign(Q45 = np.mean(data.loc[:,'Q45A1':'Q45A3'], axis=1)) # A1~A3평균
data.drop(['Q45A1','Q45A2','Q45A3'], axis=1) #기존 변수를 제거 #inplcae=True는 아직
```

순서형 변수 Total



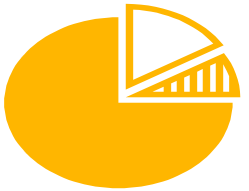
	Q5	Q6	Q43A3	Q45	Q46A6
Q5	NaN	0.422838	NaN	NaN	NaN
Q6	0.422838	NaN	NaN	NaN	NaN
Q43A3	NaN	NaN	NaN	NaN	0.366373
Q45	NaN	NaN	NaN	NaN	0.538366
Q46A6	NaN	NaN	0.366373	0.538366	NaN

명목형 변수 Total



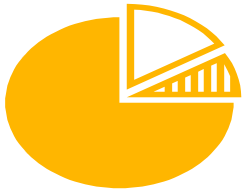
	REL	JEN	MAR	EDU2	DE2
REL	NaN	0.734122	0.553893	NaN	NaN
JEN	0.734122	NaN	NaN	NaN	NaN
MAR	0.553893	NaN	NaN	NaN	0.506720
EDU2	NaN	NaN	NaN	NaN	0.538172
DE2	NaN	NaN	0.506720	0.538172	NaN

전체 Total / Strong Corr



	Q42C3	REL	JEN	MAR	Q42D	EDU2	DE2
Q42C3	NaN	NaN	NaN	NaN	0.689181	NaN	NaN
REL	NaN	NaN	0.734122	0.553893	NaN	NaN	NaN
JEN	NaN	0.734122	NaN	NaN	NaN	NaN	NaN
MAR	NaN	0.553893	NaN	NaN	NaN	NaN	0.506720
Q42D	0.689181	NaN	NaN	NaN	NaN	NaN	NaN
EDU2	NaN	NaN	NaN	NaN	NaN	NaN	0.538172
DE2	NaN	NaN	NaN	0.506720	NaN	0.538172	NaN

전체 Total / Moderate Corr



	Q12	Q23A	Q42C3	FAM2	REL	JEN	MAR	Q1	Q24	Q42D	Q47E	EDU2	DE2
Q12	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.317194	NaN
Q23A	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.312008	NaN	NaN	NaN	NaN
Q42C3	NaN	NaN	NaN	NaN	0.306921	0.402452	NaN	NaN	NaN	NaN	0.461518	0.389835	NaN
FAM2	NaN	NaN	NaN	NaN	NaN	NaN	0.363610	NaN	NaN	NaN	NaN	NaN	NaN
REL	NaN	NaN	0.306921	NaN	NaN	NaN	NaN	0.381702	NaN	0.350983	0.316603	NaN	0.476977
JEN	NaN	NaN	0.402452	NaN	NaN	NaN	NaN	NaN	NaN	0.463026	0.422675	NaN	NaN
MAR	NaN	NaN	NaN	0.36361	NaN	NaN	NaN	0.344918	NaN	NaN	NaN	NaN	NaN
Q1	NaN	NaN	NaN	NaN	0.381702	NaN	0.344918	NaN	NaN	NaN	NaN	NaN	0.456939
Q24	NaN	0.312008	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Q42D	NaN	NaN	NaN	NaN	0.350983	0.463026	NaN	NaN	NaN	NaN	0.481647	0.384876	0.324049
Q47E	NaN	NaN	0.461518	NaN	0.316603	0.422675	NaN	NaN	NaN	0.481647	NaN	0.362904	0.316871
EDU2	0.317194	NaN	0.389835	NaN	NaN	NaN	NaN	NaN	NaN	0.384876	0.362904	NaN	NaN
DE2	NaN	NaN	NaN	NaN	0.476977	NaN	NaN	0.456939	NaN	0.324049	0.316871	NaN	NaN

앞으로...

1) Feature Extraction
PCA < FAMD

2)Analysis
SVM
XGboost
Group Lasso
Random Forest
*stacking

<Feature Extraction 관련 생각>

생각1. 잠재변수를 3개 또는 5개로 설정하고 싶다.
=> FA를 했을 때 3개 또는 5개로 나오면 좋겠다.
WHY?
=> 1) 3개: 행복의 구성요소 3개(LS, PA, NA)
⇒ 2) 5개: 성격 5요인(외향성, 개방성, 신경증, 성실성, 원만성)

생각1-1.
위의 생각에 따라, 미리 각 요인에 해당하는 값들에 적합할 항목들을 나눠보고
그것들의 밸런스를 위해 추가적인 변수를 제작해보자. ex) 부정적 감정을 측정하기 위해 어떤 것들 ratio로 제시해본다던가

Q&A

감사합니다^^