기타 아이디어

1. 응답률이 높은 사람 vs. 낮은 사람
2. 응답의 일관성이 높은 사람 vs. 낮은 사람

전처리 완료

1. GU: 그대로
2. FAM: FAM1이 있어서 FAM2는 제거. 실제로도 변수 연관성이 너무 높았다.
3. MAR: 그냥 놔두는 걸로
4. JEN: 그대로
5. INC: 빈곤층, 중산층, 상류층으로 나누었다.
6. REL: 1~2(가구주, 배우자) vs. 그외

사실 그렇게 의미있는 변수는 아닐 거라고 생각된다.

9(기타) 제거? 15개

1. RLI: 1~6까지 밖에 없어서 5(기타) 제거 / 24개밖에 없음
2. FC 제거
3. Q1: 결측치 x, 무응답 x
4. Q2: Q2A 결측치 x, 무응답 x / Q2B
5. Q5(도시 위험도): 평균 값 (A1~A4 결측치 x)
6. Q6: A1~A13 평균값 / 결측치 x
7. Q7: Q8과 연관성 높긴 하지만 일단 놔두자.
8. Q9: 결측치 x 그대로 놔두기.
9. Q10: 평균 값 / 결측치 많아서 그냥 제거하거나 평균값? / 일단 제거함
10. Q11: 그대로 두기
11. Q12I: Q12와 연관성 비교해보고 제거하기 / 일단 제거함
12. Q13: 그대로 두기
13. Q14: 요인의 원인으로 나누기

1, 2, 7, 9, 10, 52: 개인적인 이유

3, 4, 5, 6, 11: 외부적인 이유

8, 98, 99: 기타 또는 없음

// 실제로 데이터 보니까 1~8밖에 없었고, 8(기타)가 18개 밖에 없으니 제거해도 되지 않을까?

// Q14A2는 결측치 662개라서 그냥 지우고 Q14A1만 보자!!

1. Q16: Q16A는 fillna(0)
2. Q17:

Q17A1만 고려 / Q17A2 결측치 2851개, Q17A3 결측치 6884개이다.

정적 vs. 동적으로 나누고 어떤 활동을 가장 선호하는지

1,2,4,7,10,16 : 정적

3,5,6,8,9,12,13,14,15,17,51,52,54 : 동적

// 실제로 데이터 보니까 1~11밖에 없었고, 11(기타)가 22개 밖에 없으니 제거해도 되지 않을까?

1. Q18 1(개인) vs. 나머지(여럿)으로 나누기
2. Q19:

Q19A1만 고려 / Q19A2 결측치가 3260이나 된다.

정적 vs. 동적으로 나누고 어떤 활동을 가장 선호하는지

1,2,4,7,10,16 : 정적

3,5,6,8,9,12,13,14,15,17,51,52,54 : 동적

// 11(기타)가 2개 밖에 없으니 제거해도 되지 않을까?

1. Q20: Q20A 결측치가 9125 이므로 제거한다.

Q20B 결측치 많아서 제거한다.

전처리 미완료

1. HTYP: 단독주택 vs. 2~4(아파트, 다세대주택, 연립주택/빌라)

~~5(기타)는 131개밖에 안 되는데 제거할까? 제거함~~

1. HLIV: 1(자가) vs 2~4(전세, 보증부월세, 월세)

~~5(무상 및 기타)랑 9(무응답) 결측치가 각각 43, 38이니까 없애도 되지 않을까? 제거함~~

1. BIR: 나이로 계산할까? 2014년도 기준으로? … 잘 못하겠다
2. Q3(지난 1년간 모임 또는 단체활동 참여 경험): A1~A11 참여종류 개수로 통합할 예정
3. Q12(A~H): Q12\*1 총합 & Q12\*2 총합