집행부원 대상의 설문은 일반부원의 응답에서 6개의 질문이 추가된 것.

일시적으로 이 6개의 질문을 제거한 후 두 데이터프레임을 합치고, 제거했던 응답들은 따로 동아리 컬럼에 다시 merge할 예정 (대표 정보를 동아리별로 생성 후 합칠 것임.)

동아리에 열심히 참여한 정도는 매우 아니다만 아닐 시, 같은 가중치를 같도록 함 (사실 처리를 안했어 넘 주관적일거 같아서,, 안하는게 좋지 않을까)

1. time stamp, 기프티콘 번호 등 사용하지 않는 열 제거

2. column의 길이가 너무 길어 축약한 이름으로 변경

---------------------------------------------------------------------------------------------------------------------------

진행순서 : 5점 scale이 가능했던 column들 수치화하기 →

나머지 5점 scale이 아닌 경우(e.x 아님, 보통, 좋음 or 회비) 수치화하기 →

범주형 변수 더미변수화하기 (추천하는 학년 1,2,3,4)

3. 전혀 그렇지 않다 - 매우 그렇다의 설문 항목들을 어떻게 변경할지 생각.

3.1. 일반적인 환산 척도

마케팅조사에서 특정대상 혹은 사안에 대한 소비자 인식 혹은 태도를 측정하기 위해 가장 널리 쓰이는 척도 중 하나가 바로 리커드척도(Likert Scale)와 어의차이척도(Semantic Differenital Scale)이다. 리커트 척도는 척도항목별로 해당항목의 의미를 부여하는 것이며, 어의차이척도는 척도구성항목의 가장 큰 값과 작은 값에만 의미를 부여하는 것이다. 예를 들어 브랜드 호감도를 측정한다고 할 때 리커트 척도는 '1. 전혀 호감이 가지 않는다 2. 별로 호감이 가지 않는 편이다. 3. 보통이다 4. 호감이 가는 편이다 5. 매우 호감이 간다'로 구성이 되지만, 어의차이 척도에서는 1과 5에만 의미를 부여해 주고 2, 3, 4에는 의미를 부여해 주지 않는다. 실무에서는 리커트척도나 어의차이척도를 100점 환산해서 분석하는 경우가 많으며, 이러한 척도항목의 100점 환산공식은 아래와 같다.

**\* 척도항목의 100점 환산공식 - [(응답치-1)/(척도항목수-1)]\*100**

예를 들어보면, 5점 척도에서 4로 응답한 경우 공식을 적용해 보면 [(4-1)/(5-1)]\*100 = 75가 된다. 7점척도에서 4로 응답한 경우에는 [(4-1)/(7-1)]\*100 = 50이 된다. 이 공식에서는 척도항목에서 가장 낮은 값이 0, 가장 높은 값이 100이 되도록 환산하게 된다.



→ 위의 환산 점수는 사람들이 100점 만점으로 알아보기 쉽게 환산하는 목적이기 때문에 우리의 분류 목적을 위한 점수로는 걸맞지 않다고 생각 ( 어차피 scaling도 하기 때문), 또한 1점을 0점으로 환산할 때 적절한 분석이 어려울 것 같아 하지 않기로 함…

3.2. scaling을 하기 때문에 그대로 1~5점으로 진행할 예정

3.3. 5점 스케일을 하는 척도는 전혀 그렇지 않다 ~ 매우 그렇다의 5개를 각각 1,2,3,4,5로 변환하는 것.

3.4. 동아리 활동 소요시간인 2시간~4시간, 4시간~6시간 등의 수치는 중간값을 활용. 마지막 값은 등차수열로 이용 (8시간 초과인 경우 그동안의 중간값이 3,5,7이었으니 9로 대체), 동아리 평균 활동비도 마찬가지.

3.5. 활동 강제 정도에 대해 고민함. 자유, 반드시 한번 필참, 필참 세 개의 분류가 된 변수. 123의 스케일을 줄지, 범주형 변수로 바꿀지 고민. 수치형으로 바꾸기엔 척도가 정확한 역할을 하지 않을 것 같아 카테고리화하기로 결정 (다른 의견 있음 말해죵)

3.6. 열정적으로 활동하는 동아리원 (5명 내외, 10명 내외, 15명 내외 … ) 의 변수는 각각 나타낸 숫자로 변환 (범위를 나타낸게 아니므로 중간값 사용 X)

3.7. 그 외 ‘소모임, 스터디’ 변수 같이 (알수 없음, 있음, 비교적 많음) 5점 척도는 아니더라도 적절한 ordinary number를 줄 수 있다면 0,1,2의 식으로 줌.

3.8. 활동 빈도는 category로 변환.

3.9. Na값은 (=알수없음) 각 동아리 내에서의 중앙값or최빈값으로 대체할 예정. (각 동아리마다 관측치가 적으면 힘들거같으면,...전체 column에서의 평균값으로 대체?)

3.10. 범주형 변수로 바꿀 때는 Na값을 제거 (하나의 범주에 할당하기엔 리스크가 크다고 생각) ,, 수치형은 그냥 중간값 대체가 가능하지만 이건 못할거같은데

3.11. 추천하는 학년은 중복선택이 이뤄짐. 모두를 선택할 경우 1부터 4의 변수를 1로 반환