[숙명DSS센터] 통계 데이터 분석을 위한 범주형 데이터 전처리 방법

SPSS, R, 파이썬 등의 프로그램을 활용해 통계 분석을 하기 위해선 데이터 내 분석을 위해 사용할 변수가 모두 문자가 아닌 **숫자**로 이루어져 있어야 합니다. 따라, 문자로 이루어진 범주형 데이터 변수를 수치형 변수로 변환하는 방법에 대해 안내해 드립니다. 용이한 분석 서비스 전달을 위해 해당 방법에 맞춰 데이터를 정제한 뒤 파일을 업로드 해주시기 바랍니다.

1. 라벨링(Labelling)

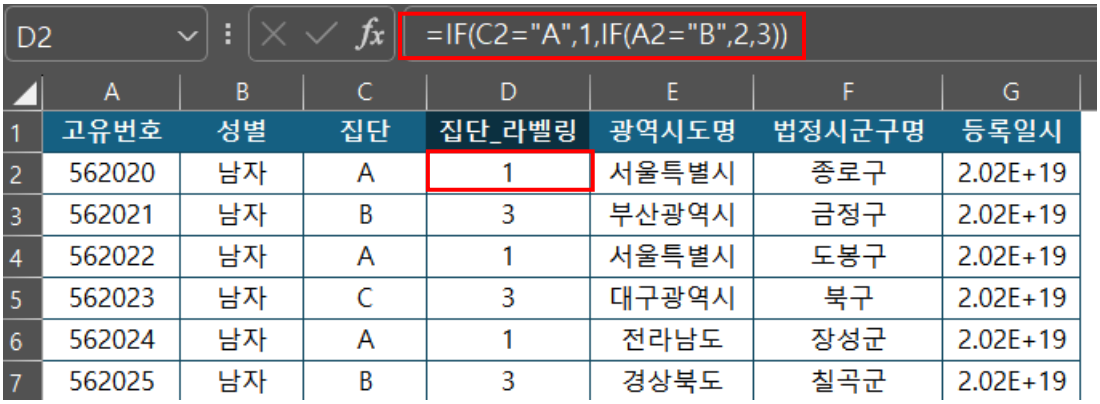
SPSS를 사용한 T 검정, ANOVA 등 집단간 비교 / 트리 계열 모델을 사용한 분석의 경우 변수를 해당 방식으로 변환 해주시기 바랍니다. R을 사용한 집단 간 비교는 라벨링 없이 factor 지정만으로도 분석이 가능합니다.

집단	집단_라벨링
A	1
B	2
C	3

집단을 다음과 같이 겹치지 않는 숫자로 변환 해주시기 바랍니다.

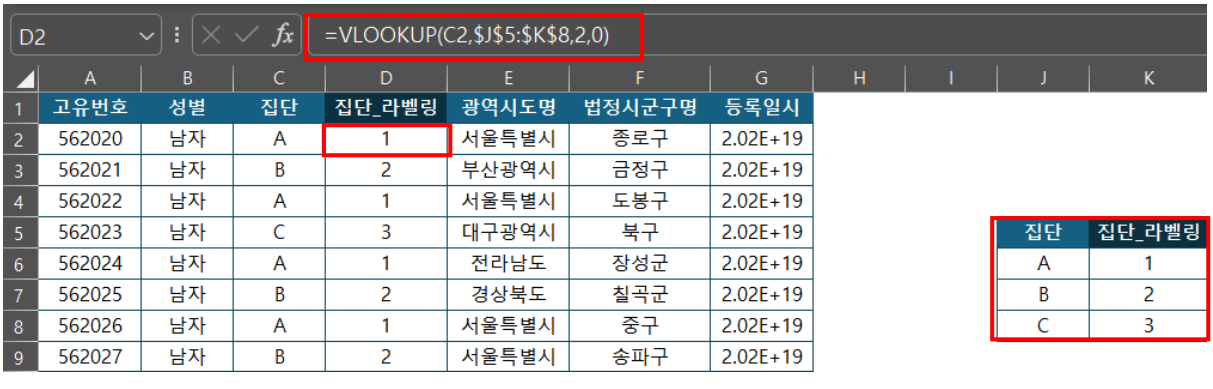
1-1. 엑셀을 사용한 라벨링

- **IF구문**을 사용한 라벨링



	A	B	C	D	E	F	G
1	고유번호	성별	집단	집단_라벨링	광역시도명	법정시군구명	등록일시
2	562020	남자	A	1	서울특별시	종로구	2.02E+19
3	562021	남자	B	3	부산광역시	금정구	2.02E+19
4	562022	남자	A	1	서울특별시	도봉구	2.02E+19
5	562023	남자	C	3	대구광역시	북구	2.02E+19
6	562024	남자	A	1	전라남도	장성군	2.02E+19
7	562025	남자	B	3	경상북도	칠곡군	2.02E+19

- **VLOOKUP, HLOOKUP**을 통한 라벨링



	A	B	C	D	E	F	G
1	고유번호	성별	집단	집단_라벨링	광역시도명	법정시군구명	등록일시
2	562020	남자	A	1	서울특별시	종로구	2.02E+19
3	562021	남자	B	2	부산광역시	금정구	2.02E+19
4	562022	남자	A	1	서울특별시	도봉구	2.02E+19
5	562023	남자	C	3	대구광역시	북구	2.02E+19
6	562024	남자	A	1	전라남도	장성군	2.02E+19
7	562025	남자	B	2	경상북도	칠곡군	2.02E+19
8	562026	남자	A	1	서울특별시	중구	2.02E+19
9	562027	남자	B	2	서울특별시	송파구	2.02E+19

집단	집단_라벨링
A	1
B	2
C	3

1-2. R을 사용한 라벨링

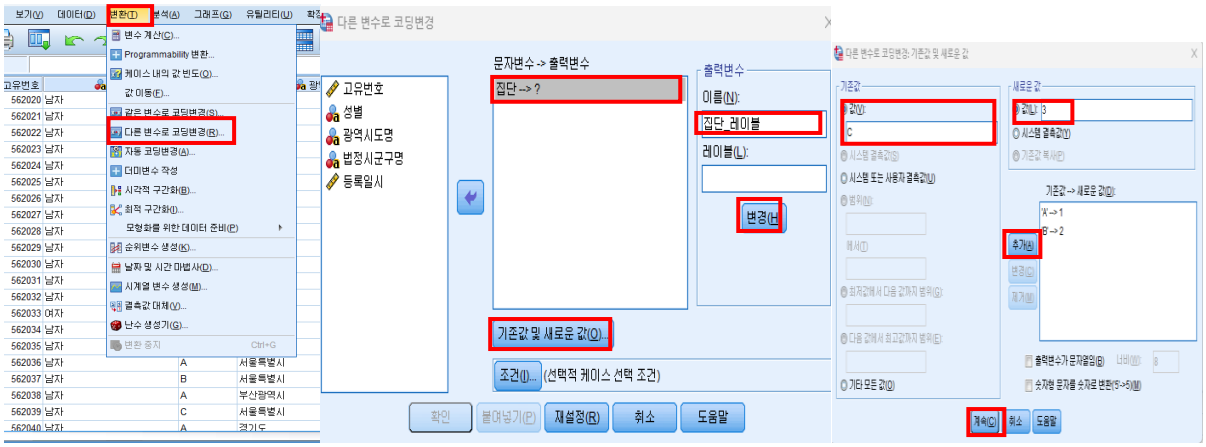
다음과 같은 코드를 사용해 변수를 수치로 변경해주시기 바랍니다.

```
# R에서 라벨 인코딩 실시하기

# 1. 집단 변수를 factor로 변환
data$집단 <- factor(data$집단, level = c("A", "B", "C"))
# as.numeric()을 통해 숫자형 변수로 변환
data$집단 <- as.numeric(data$집단)
```

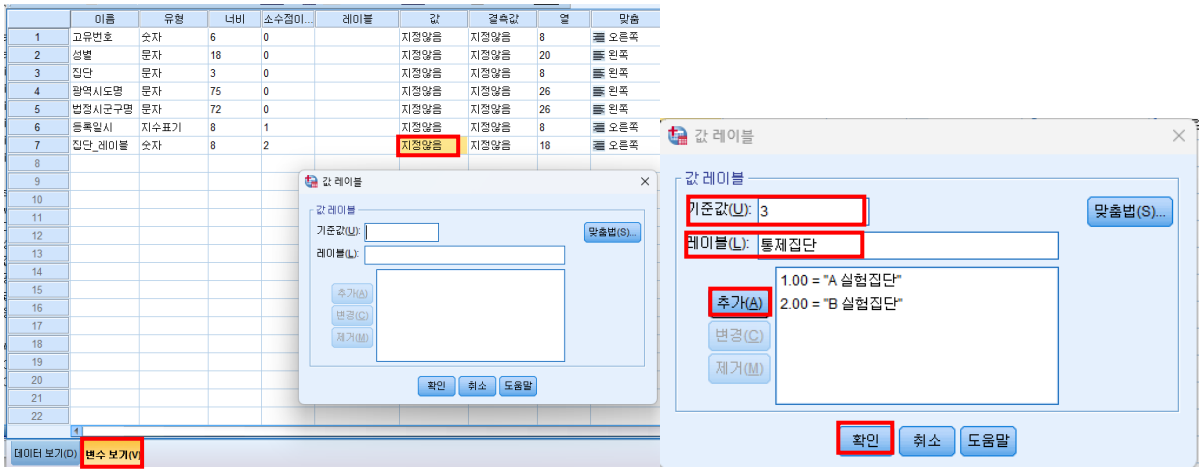
1-3. SPSS를 사용한 라벨링

변환 > 다른 변수로 코딩 변경 > [문자변수 -> 출력변수]에 변환할 변수 추가 > [출력 변수]에 새로운 변수명 기입 후 [기존 값 및 새로운 값 클릭] > [기존 값]에 문자형 변수를, [새로운 값]에 변환할 수치 기재 후 추가



SPSS 데이터 파일인 .sav 형식의 파일을 보내주실 경우 각 라벨링에 따른 값을 추가 부탁드립니다.

좌측 하단의 [변수 보기] > 변수의 [값] 클릭 > [값 레이블]에서 값에 따른 레이블을 추가



2. 더미 변수(Dummy Variable) 생성

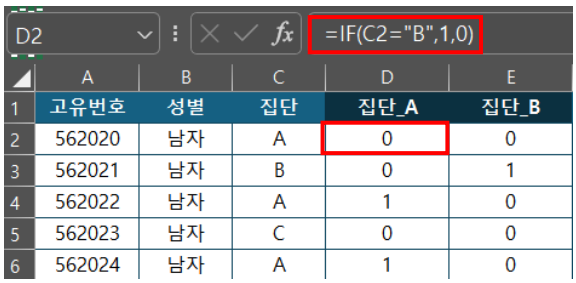
다중 회귀 분석, ANCOVA 등 집단 변수를 통제 변수로 사용하는 분석의 경우 변수를 해당 방식으로 변환해주시기 바랍니다.

집단	집단_A	집단_B
A	1	0
B	0	1
C	0	0

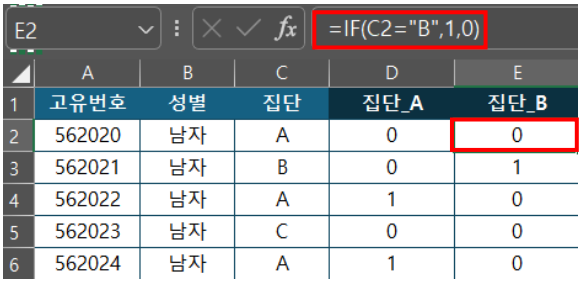
집단	집단_A	집단_B	집단_C
A	1	0	0
B	0	1	0
C	0	0	1

집단을 범주의 포함 여부에 따라 0, 1의 수치를 부여해주시기 바랍니다. 좌측 그림과 같이 범주 수보다 하나 적은 개수의 변수 생성이 기본적이지만, 우측과 같이 범주 수만큼 변수를 생성하셔도 괜찮습니다.

2.1. 엑셀을 사용한 더미 변수 생성



	A	B	C	D	E
1	고유번호	성별	집단	집단_A	집단_B
2	562020	남자	A	0	0
3	562021	남자	B	0	1
4	562022	남자	A	1	0
5	562023	남자	C	0	0
6	562024	남자	A	1	0



	A	B	C	D	E
1	고유번호	성별	집단	집단_A	집단_B
2	562020	남자	A	0	0
3	562021	남자	B	0	1
4	562022	남자	A	1	0
5	562023	남자	C	0	0
6	562024	남자	A	1	0

2.2. R을 사용한 더미 변수 생성

```
# R에서 더미 변수 생성하기

#1. fastDummies 패키지 설치
install.packages("fastDummies")
library(fastDummies)
# 더미변수로 변경할 변수 선택 후 생성
dummy_data <- dummy_cols(data, select_columns = "집단")
```

2.3. SPSS를 사용한 더미 변수 생성

변환 > 더미변수 작성 > [다음에 대한 더미변수 작성]에 변환할 변수 추가 > [주효과 더미변수]에 생성할 변수명 작성 > [확인]