# [숙명DSS센터] 통계 데이터 분석을 위한 범주형 데이터 전처리 방법

SPSS, R, 파이썬 등의 프로그램을 활용해 통계 분석을 하기 위해선 데이터 내 분석을 위해 사용할 변 수가 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 분석 이 가능합니다.

집단	집단_라벨링
А	1
В	2
С	3

집단을 다음과 같이 겹치지 않는 숫자로 변환 해주시기 바랍니다.

#### 1-1. 엑셀을 사용한 라벨링

## ■ IF구문을 사용한 라벨링

D2	2	<u>~]:[×</u>	$\checkmark f_x$	=IF(C2="A",1	,IF(A2="B",2,	.3))	
	А	B	с	D	E	F	G
1	고유번호	성별	집단	집단_라벨링	광역시도명	법정시군구명	등록일시
2	562020	남자	А	1	서울특별시	종로구	2.02E+19
3	562021	남자	В	3	부산광역시	금정구	2.02E+19
4	562022	남자	А	1	서울특별시	도봉구	2.02E+19
5	562023	남자	C	3	대구광역시	북구	2.02E+19
6	562024	남자	А	1	전라남도	장성군	2.02E+19
7	562025	남자	В	3	경상북도	칠곡군	2.02E+19

VLOOKUP, HLOOKUP을 통한 라벨링

D	2	~) <b>:</b> [×	$\sqrt{f_x}$	=VLOOKUP(	C2,\$J\$5:\$K\$8	8,2,0)				
	A	В	с	D	E	F	G	н		к
1	고유번호	성별	집단	집단_라벨링	광역시도명	법정시군구명	등록일시			
2	562020	남자	Α	1	서울특별시	종로구	2.02E+19			
3	562021	남자	В	2	부산광역시	금정구	2.02E+19			
4	562022	남자	Α	1	서울특별시	도봉구	2.02E+19			
5	562023	남자	С	3	대구광역시	북구	2.02E+19		집단	집단_라벨링
6	562024	남자	Α	1	전라남도	장성군	2.02E+19		А	1
7	562025	남자	В	2	경상북도	칠곡군	2.02E+19		В	2
8	562026	남자	А	1	서울특별시	중구	2.02E+19		С	3
9	562027	남자	В	2	서울특별시	송파구	2.02E+19			

#### 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

보기(火) 데이터(D)	변환(T) 분석(A) 그래프(G) 유틸리(	(U) 확	🔁 다른 변수로 코딩변경				×	
) 🛄 🗠 🤉	금 변수계산(C)				문자변수 -> 출력변수	- 출력변수	🍓 다른 변수로 코딩변경: 기존값 및 새로운 값	х
고유번호 🚜 562020 남자	값이동(E)	<b>a</b> 8	/ 🔗 고유번호		집단> ?	이름(N):	- 기존값	세로운 값
562021 남자 562022 남자	·····································		🚜 경렬 🔒 광역시도명			집단_레이블	C	○ 시스템 걸속값(!)
562023 급사 562024 남자 562025 남자	[웹 자동 코딩변경(A)		♣ 법정시군구명 ★ 등록익시			레이블(L):	● 시스템 결측값(S) ● 시스템 또는 사용자 결측값(U)	◎ 기존값 복사만
562026 남자 562027 남자	▶ 월 시각적 구간화(B) 않 최적 구간화(I)		V 012/1	*		<b>増</b> 落(H	● 범위(M):	기존값 -> 새로운 값 <u>(D):</u> 'X -> 1
562028 남자 562029 남자	모형화를 위한데이터 준비(P) 위변수 생성(K)	•					HAD	( <b>‡7</b> /A)
562030 남자 562031 남자	. 🛗 날짜 및 시간 마법사(D) 🥅 시계열 변수 생성(M)		-				● 최저값배서 다음 값까지 범위(G):	H 3C
562032 급사 562033 여자 562034 날자	💱 결측값 대체(V) 🎯 난수 생성기(G)		-					4.7.00
562035 남자 562036 남자	변환 중지 Ctri+C A 서울특별	A	-		기존값 및 세도꾼 값(인)		● 다음 값에서 최고값까지 범위(E): -	출력변수가문자열임(B) 너비(W): 8
562037 남자 562038 남자	B         서울특별           A         부산광역	시 시			조건() (선택적 케이스 선택 조건)		○ 기타 모든 값( <u>0</u> )	🚺 숫자형 문자를 숫자로 변환(5:~5)Ш)
562039 남자 562040 남자	C 서울특별 A 경기도	AI	확인	l l	(여넣기(P) 재설정(R) 취소	도움말	月金〇	취소 도움말

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

	이름	유형	너비	소수점이.	. 레이블	값	결측값	열	맞춤	
1	고유번호	숫자	6	0		지정않음	지정않음	8	📰 오른쪽	
2	성별	문자	18	0		지정않음	지정않음	20	틀 왼쪽	
3	집단	문자	3	0		지정않음	지정않음	8	■ 왼쪽	
4	광역시도명	문자	75	0		지정않음	지정않음	26	≣ 왼쪽	
5	법정시군구명	문자	72	0		지정않음	지정않음	26	≣ 왼쪽	
6	등록일시	지수표기	8	1		지정않음	지정않음	8	▓ 오른쪽	
7	집단_레이블	숫자	8	2		지정않음	지정않음	18	遭 오른쪽	- 🐫 값 데 이글 🛛 💦 🗙
8										
9				1	🔒 값 레이블				×	┌값레이블
10										
11					값레이블					시군값(U), [3] 맞춤법(S),
12	1				기준값(U):				맞춤법(S)	
13	]				레이블(L):					데이크(도), 동세집단
14								_		1.00 - "A 신청지단"
15					李7H(A)					
16					#7(C)					추가( <u>A</u> ) 2.00 = "B 실험집단"
17					1000					(HIZIO)
18					M M(M)					
19										TH THOM
20						확인	취소 도움말			
21										
22										
	4	-		_						화이 최소 도운막
데이터 보기(	(D) 변수 보기(\	n								
							A			

# 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>B</b>	집단	집단_A	집단_ <b>B</b>	집단_ <b>C</b>
А	1	0	Α	1	0	0
В	0	1	В	0	1	0
С	0	0	С	0	0	1

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

## 2.1. 엑셀을 사용한 더미 변수 생성

D	<u>-</u> 2	~) : [×	$\sqrt{f_x}$	=IF(C2="B",1	,0)	E2		~) <b>:</b> [X	$\checkmark f_x$	=IF(C2="B",1	,0)
	A	B	c	D	E		А	В	С	D	E
1	고유번호	성별	집단	집단_A	집단_ <b>B</b>	1	고유번호	성별	집단	집단_A	집단_ <b>B</b>
2	562020	남자	А	0	0	2	562020	남자	А	0	0
3	562021	남자	В	0	1	3	562021	남자	В	0	1
4	562022	남자	Α	1	0	4	562022	남자	А	1	0
5	562023	남자	С	0	0	5	562023	남자	С	0	0
6	562024	남자	Α	1	0	6	562024	남자	А	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]

↓(E) 보기(V)	데이터(D)	변환( <u>T</u> )	분석( <u>A</u> )	그래프( <u>G</u> )	유틸리티( <u>U</u> )	1	🔒 더미변수 작성		×			
		📑 변수	계산 <mark>(C</mark> )				변수()):		다음에 대한 더미변수 작성( <u>C</u> ):			
	•	🛨 Prog	rammability	(변환			🛷 고유변호	]	🔏 집단			
		🛃 भाषा.	스 내의 값 빈	민도( <u>O</u> )		L	응 성별 - 카이니드며					
🛷 고유번호	Ja	2) DI	토(F)			5	👍 형국시도장 🚑 법정시군구명					
562020	남자		a( <u>_</u> /				🛷 등록일시					
562021	남자	🔤 같은	변수로 코딩	변경( <u>S</u> )								
562022	남자	🏧 다른	변수로 코딩	변경( <u>R</u> )					수효과 너비면수			
562023	남자	🛐 자동	코딩변경(A)	)				]	♥ 두보과 더미 작정(M) 로토 미르(세명하나비 A도나하 레VO)			
562024	남자	E GUU	벼스 작성				더미변수레이블		두드 이용(전력한 번부당 번 개)( <u>0</u> ). 집단 데미			
562025	남자					-	◎ 값 레이클 사용(⊻) ◎ 가 사용(?)					
562026	남자	P: 시작·	의 구간와(법	)								
562027	남자	🔀 최적	구간화([)			-	- 강 순서					
562028	남자	모형	화를 위한 더	I이터 준비( <u>P</u> )	•	1	◎ 오를차순(A)		이원 영모역용			
562029	남자	🛃 순위!	변수 생성( <u>K</u>	)		F.	◎ 내림차순(D)		르티 이르(F):			
562030	남자	<u>⇔</u> ⊎ль	및 시가미바	14(D)		1			+= 018( <u>c</u> ).			
562031	날자			a / ( <u>C</u> )		F.	애크로		매크로 이름(C)			
562032	는 · 남자	🔤 Л Л.	걸 면수 생성	( <u>M</u> )		F	🥅 매크로 정의에서 첫 번째 더미 범주 생략(E)					
562033	여자	🐫 결측	값대체(\_)			-	참고: 매크로 이름은 일반적으로 I로 시작됩니다.		. 사외 사호자용			
562034	날자	🝘 난수	생성기 <mark>(G)</mark>			-			B 전 8 포 1 8 모든 삼원 상호작용에 대한 더미 작성(H)			
562035	나자	- 변화	중지		Ctrl+G	F.	측정 수준 사용		₹E 01#:			
562036	나자		Δ		서운트병시	-	◎ 작도면수 값에 대한 더미를 작성하지 않음(U) ○ 모든 벼스에 대한 더미 장성(C)					
562030					개등금융기 1186여기		O TEE MALEAN 10(E)		매크로 이름:			
562020	다니		0		비운국들시		이 대화 상자에는 Python Essentials가 필요합니다.					
562030			~		파요트웨기 							
562039	리지				지골국립시 거리도	_	확인 물여넣기(만) 지	#열성( <u>R</u> )	쥐소 노동말			
EF-MAN												