경영데이터분석기초

- SPSS, Excel을 활용한 통계분석 -

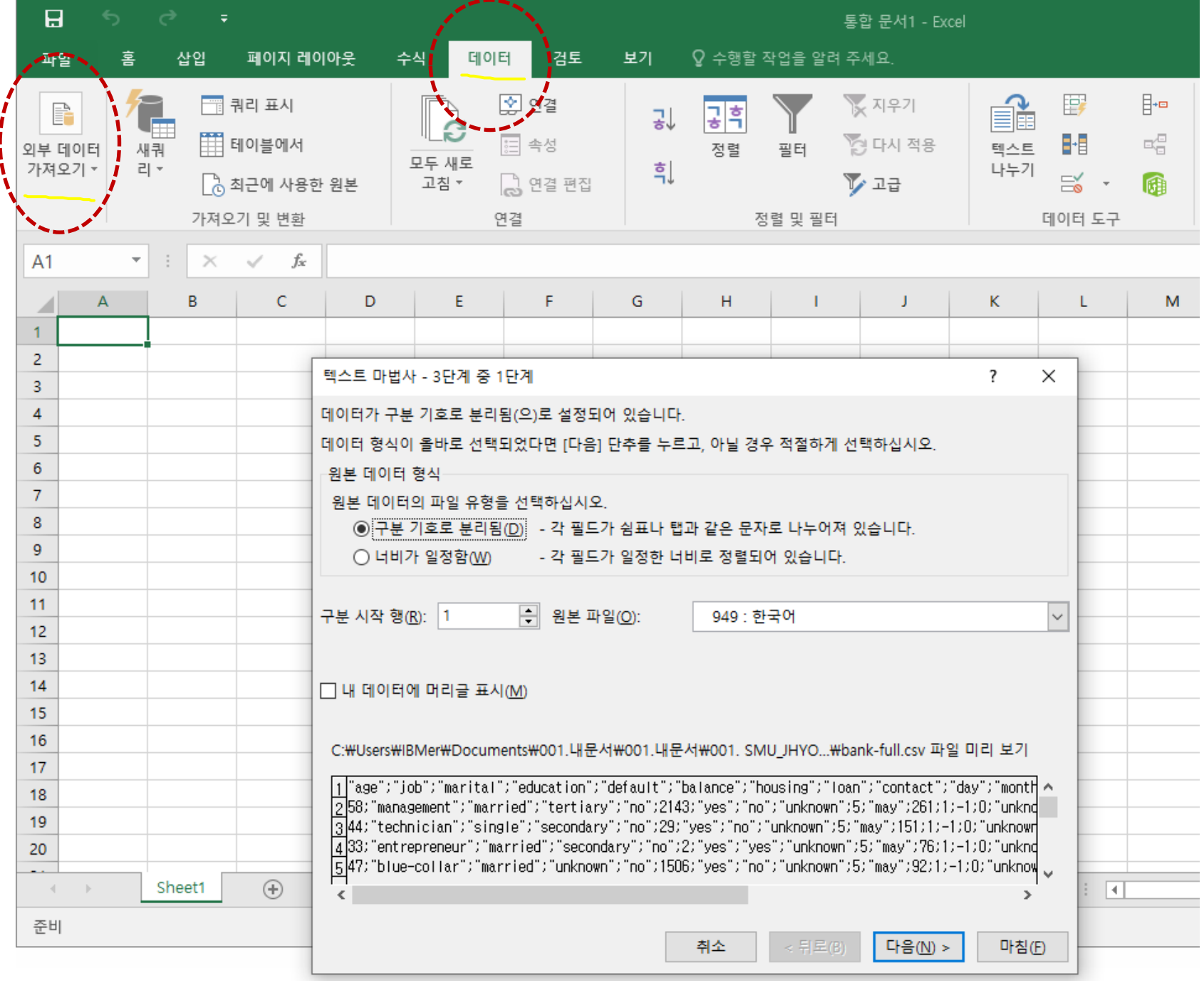
유 진 호

jhyoo@smu.ac.kr

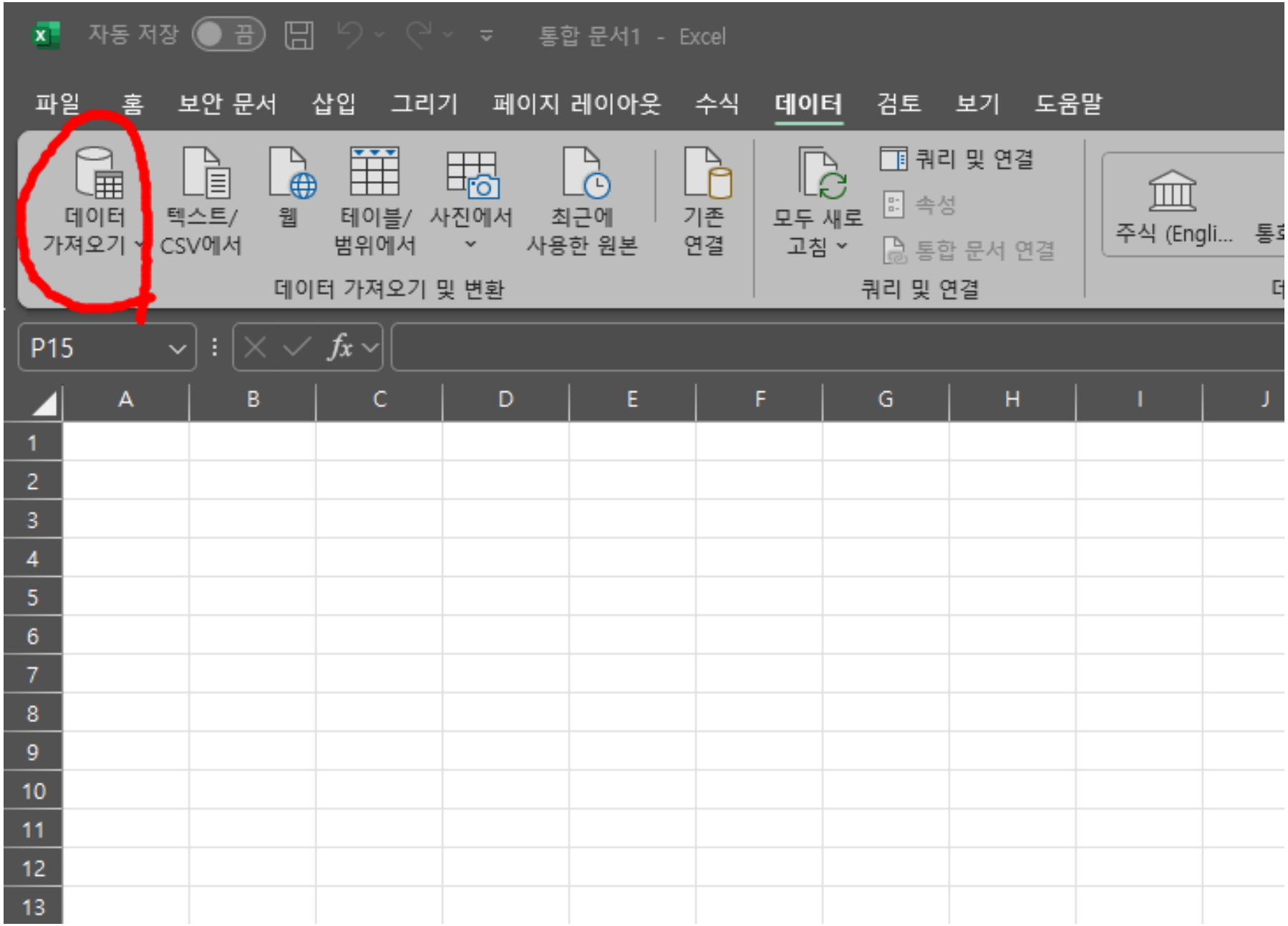
bank-full.csv - Windows 메모장

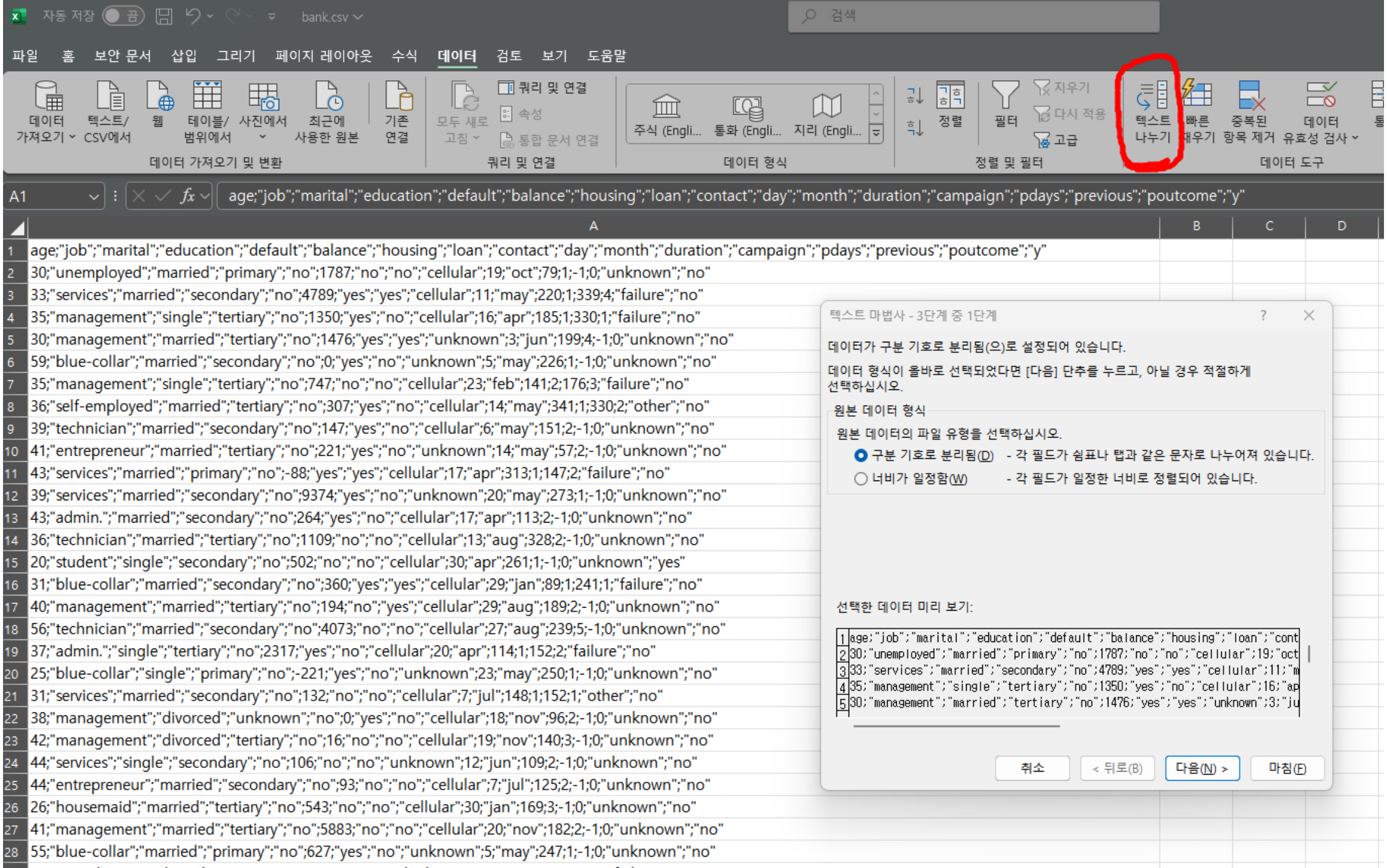
파일(F) 편집(E) 서식(O) 보기(V) 도움말

```
"age";"job";"marital";"education";"default";"balance";"housing";"loan";"contact";"day";"month";"duration";"campaign";"pdays";"previous";"poutcome";"y"
58;"management";"married";"tertiary";"no";2143;"yes";"no";"unknown";5;"may";261;1;-1;0;"unknown";"no"
44;"technician";"single";"secondary";"no";29;"yes";"no";"unknown";5;"may";151;1;-1;0;"unknown";"no"
33;"entrepreneur";"married";"secondary";"no";2;"yes";"yes";"unknown";5;"may";76;1;-1;0;"unknown";"no"
47;"blue-collar";"married";"unknown";"no";1506;"yes";"no";"unknown";5;"may";92;1;-1;0;"unknown";"no"
33;"unknown";"single";"unknown";"no";1;"no";"no";"unknown";5;"may";198;1;-1;0;"unknown";"no"
35;"management";"married";"tertiary";"no";231;"yes";"no";"unknown";5;"may";139;1;-1;0;"unknown";"no"
28;"management";"single";"tertiary";"no";447;"yes";"yes";"unknown";5;"may";217;1;-1;0;"unknown";"no"
42;"entrepreneur";"divorced";"tertiary";"yes";2;"yes";"no";"unknown";5;"may";380;1;-1;0;"unknown";"no"
58;"retired";"married";"primary";"no";121;"yes";"no";"unknown";5;"may";50;1;-1;0;"unknown";"no"
43;"technician";"single";"secondary";"no";593;"yes";"no";"unknown";5;"may";55;1;-1;0;"unknown";"no"
41;"admin.";"divorced";"secondary";"no";270;"yes";"no";"unknown";5;"may";222;1;-1;0;"unknown";"no"
29;"admin.";"single";"secondary";"no";390;"yes";"no";"unknown";5;"may";137;1;-1;0;"unknown";"no"
53;"technician";"married";"secondary";"no";6;"yes";"no";"unknown";5;"may";517;1;-1;0;"unknown";"no"
58;"technician";"married";"unknown";"no";71;"yes";"no";"unknown";5;"may";71;1;-1;0;"unknown";"no"
57;"services";"married";"secondary";"no";162;"yes";"no";"unknown";5;"may";174;1;-1;0;"unknown";"no"
51;"retired";"married";"primary";"no";229;"yes";"no";"unknown";5;"may";353;1;-1;0;"unknown";"no"
45;"admin.";"single";"unknown";"no";13;"yes";"no";"unknown";5;"may";98;1;-1;0;"unknown";"no"
57;"blue-collar";"married";"primary";"no";52;"yes";"no";"unknown";5;"may";38;1;-1;0;"unknown";"no"
```



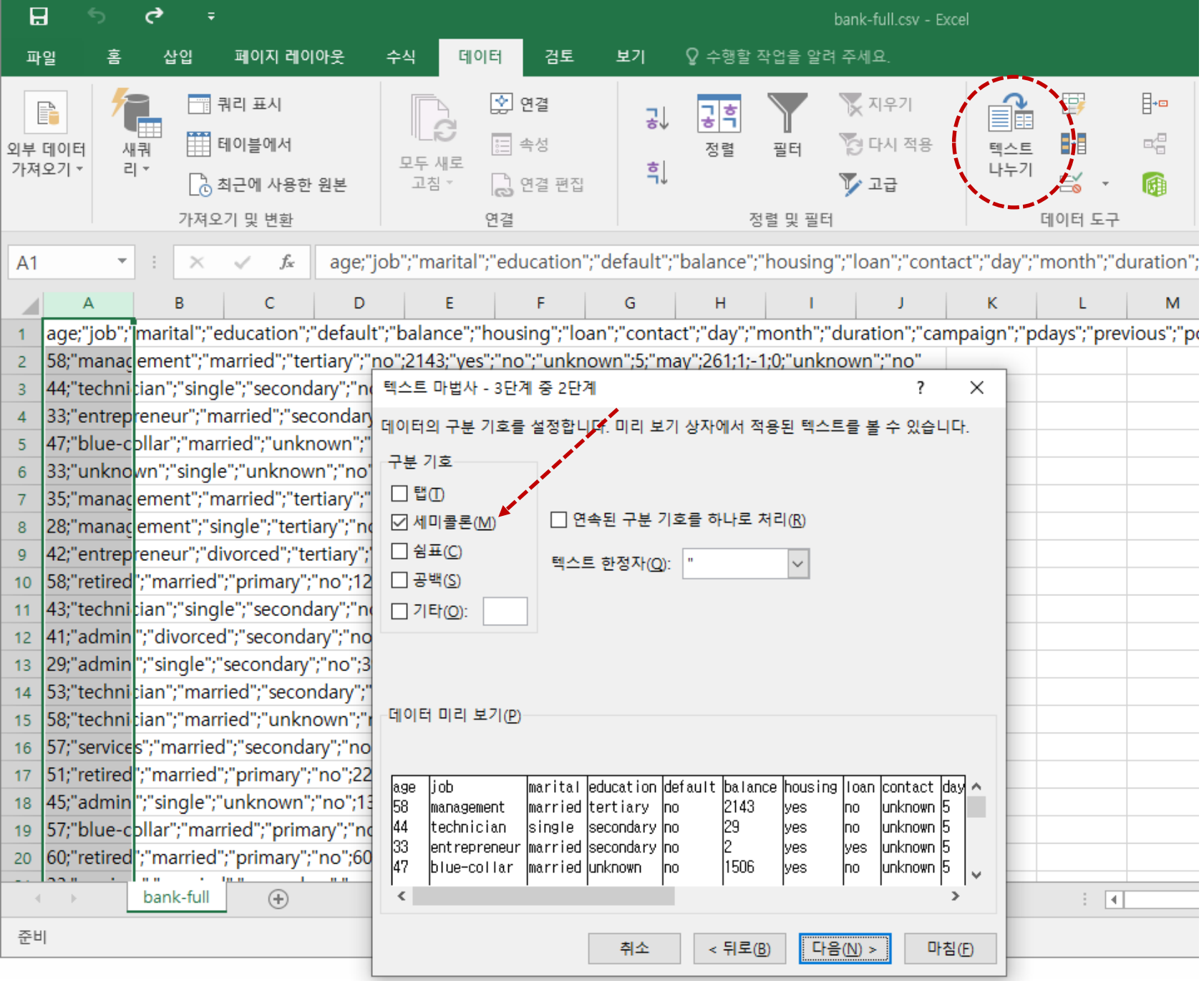
- CSV 파일을 엑셀에서 불러오기
- CSV[comma separated value]
  - 쉼표(comma)를 기준으로 항목을 구분하여 저장한 데이터,
  - 수많은 애플리케이션에서 취급하는 범용 형식,
  - CSV 형식의 파일은 텍스트 파일로 보존하여 문서 처리기나 편집기에서 열람·편집할 수 있음.


자동 저장
끔
통합 문서1 - Excel
파일
홈
보안 문서
삽입
그리기
페이지 레이아웃
수식
데이터
검토
보기
도움말
데이터 가져오기
텍스트/ CSV에서
웹
테이블/ 범위에서
사진에서
최근에 사용한 원본
기존 연결
모두 새로 고침
쿼리 및 연결
속성
통합 문서 연결
주식 (Engli...
데이터 가져오기 및 변환
쿼리 및 연결
P15
A
B
C
D
E
F
G
H
I
J
1
2
3
4
5
6
7
8
9
10
11
12
13



	A
1	age;"job";"marital";"education";"default";"balance";"housing";"loan";"contact";"day";"month";"duration";"campaign";"pdays";"previous";"poutcome";"y"
2	30;"unemployed";"married";"primary";"no";1787;"no";"no";"cellular";19;"oct";79;1;-1;0;"unknown";"no"
3	33;"services";"married";"secondary";"no";4789;"yes";"yes";"cellular";11;"may";220;1;339;4;"failure";"no"
4	35;"management";"single";"tertiary";"no";1350;"yes";"no";"cellular";16;"apr";185;1;330;1;"failure";"no"
5	30;"management";"married";"tertiary";"no";1476;"yes";"yes";"unknown";3;"jun";199;4;-1;0;"unknown";"no"
6	59;"blue-collar";"married";"secondary";"no";0;"yes";"no";"unknown";5;"may";226;1;-1;0;"unknown";"no"
7	35;"management";"single";"tertiary";"no";747;"no";"no";"cellular";23;"feb";141;2;176;3;"failure";"no"
8	36;"self-employed";"married";"tertiary";"no";307;"yes";"no";"cellular";14;"may";341;1;330;2;"other";"no"
9	39;"technician";"married";"secondary";"no";147;"yes";"no";"cellular";6;"may";151;2;-1;0;"unknown";"no"
10	41;"entrepreneur";"married";"tertiary";"no";221;"yes";"no";"unknown";14;"may";57;2;-1;0;"unknown";"no"
11	43;"services";"married";"primary";"no";-88;"yes";"yes";"cellular";17;"apr";313;1;147;2;"failure";"no"
12	39;"services";"married";"secondary";"no";9374;"yes";"no";"unknown";20;"may";273;1;-1;0;"unknown";"no"
13	43;"admin.";"married";"secondary";"no";264;"yes";"no";"cellular";17;"apr";113;2;-1;0;"unknown";"no"
14	36;"technician";"married";"tertiary";"no";1109;"no";"no";"cellular";13;"aug";328;2;-1;0;"unknown";"no"
15	20;"student";"single";"secondary";"no";502;"no";"no";"cellular";30;"apr";261;1;-1;0;"unknown";"yes"
16	31;"blue-collar";"married";"secondary";"no";360;"yes";"yes";"cellular";29;"jan";89;1;241;1;"failure";"no"
17	40;"management";"married";"tertiary";"no";194;"no";"yes";"cellular";29;"aug";189;2;-1;0;"unknown";"no"
18	56;"technician";"married";"secondary";"no";4073;"no";"no";"cellular";27;"aug";239;5;-1;0;"unknown";"no"
19	37;"admin.";"single";"tertiary";"no";2317;"yes";"no";"cellular";20;"apr";114;1;152;2;"failure";"no"
20	25;"blue-collar";"single";"primary";"no";-221;"yes";"no";"unknown";23;"may";250;1;-1;0;"unknown";"no"
21	31;"services";"married";"secondary";"no";132;"no";"no";"cellular";7;"jul";148;1;152;1;"other";"no"
22	38;"management";"divorced";"unknown";"no";0;"yes";"no";"cellular";18;"nov";96;2;-1;0;"unknown";"no"
23	42;"management";"divorced";"tertiary";"no";16;"no";"no";"cellular";19;"nov";140;3;-1;0;"unknown";"no"
24	44;"services";"single";"secondary";"no";106;"no";"no";"unknown";12;"jun";109;2;-1;0;"unknown";"no"
25	44;"entrepreneur";"married";"secondary";"no";93;"no";"no";"cellular";7;"jul";125;2;-1;0;"unknown";"no"
26	26;"housemaid";"married";"tertiary";"no";543;"no";"no";"cellular";30;"jan";169;3;-1;0;"unknown";"no"
27	41;"management";"married";"tertiary";"no";5883;"no";"no";"cellular";20;"nov";182;2;-1;0;"unknown";"no"
28	55;"blue-collar";"married";"primary";"no";627;"yes";"no";"unknown";5;"may";247;1;-1;0;"unknown";"no"

데이터 구분기호


bank-full.csv - Excel
파일 홈 삽입 페이지 레이아웃 수식 데이터 검토 보기 수행할 작업을 알려 주세요.
외부 데이터 가져오기
새로 고침
쿼리 표시
테이블에서
최근에 사용한 원본
가져오기 및 변환
모두 새로 고침
연결
속성
연결 편집
연결
정렬
필터
지우기
다시 적용
고급
정렬 및 필터
텍스트 나누기
데이터 도구
A1
age;"job";"marital";"education";"default";"balance";"housing";"loan";"contact";"day";"month";"duration";
텍스트 마법사 - 3단계 중 2단계
데이터의 구분 기호를 설정합니다. 미리 보기 상자에서 적용된 텍스트를 볼 수 있습니다.
구분 기호
탭(T)
세미콜론(M)
쉼표(C)
공백(S)
기타(O):
연속된 구분 기호를 하나로 처리(R)
텍스트 한정자(Q):
데이터 미리 보기(P)
age job marital education default balance housing loan contact day
58 management married tertiary no 2143 yes no unknown 5
44 technician single secondary no 29 yes no unknown 5
33 entrepreneur married secondary no 2 yes yes unknown 5
47 blue-collar married unknown no 1506 yes no unknown 5
취소
< 뒤로(B)
다음(N) >
마침(F)
bank-full
준비

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
2	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
3	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
4	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
5	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
6	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
7	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
8	28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
9	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
10	58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
11	43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
12	41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
13	29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
14	53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
15	58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
16	57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
17	51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
18	45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no
19	57	blue-collar	married	primary	no	52	yes	no	unknown	5	may	38	1	-1	0	unknown	no
20	60	retired	married	primary	no	60	yes	no	unknown	5	may	219	1	-1	0	unknown	no
21	33	services	married	secondary	no	0	yes	no	unknown	5	may	54	1	-1	0	unknown	no
22	28	blue-collar	married	secondary	no	723	yes	yes	unknown	5	may	262	1	-1	0	unknown	no
23	56	management	married	tertiary	no	779	yes	no	unknown	5	may	164	1	-1	0	unknown	no
24	32	blue-collar	single	primary	no	23	yes	yes	unknown	5	may	160	1	-1	0	unknown	no
25	25	services	married	secondary	no	50	yes	no	unknown	5	may	342	1	-1	0	unknown	no
26	40	retired	married	primary	no	0	yes	yes	unknown	5	may	181	1	-1	0	unknown	no

데이터 불러오면
행의 개수는?

엑셀파일로 저장

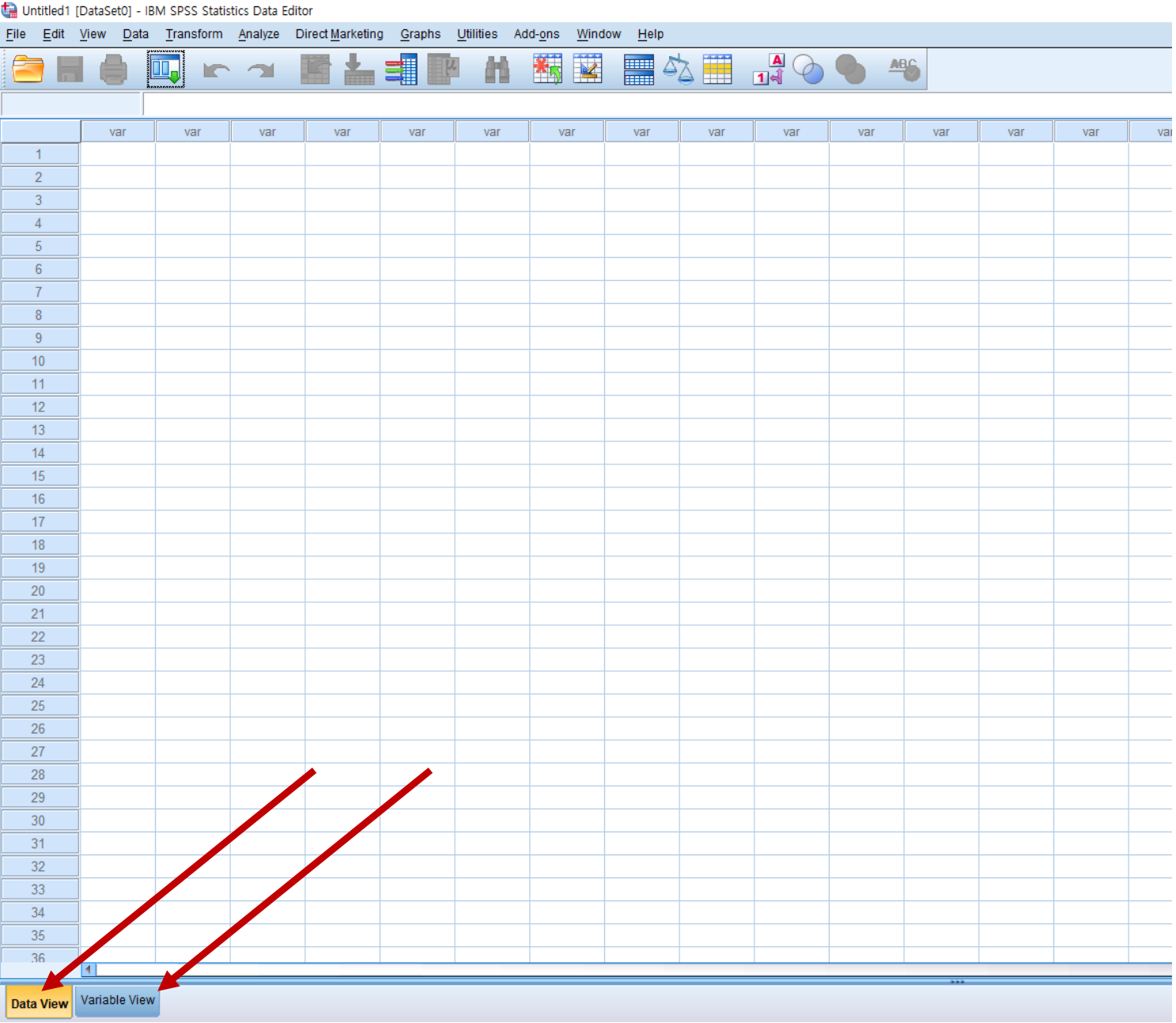
Input variables:

bank client data:
1 - age (numeric)
2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student","blue-collar","selfemployed","retired","technician","services")
3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown", "primary", "secondary", "tertiary")
5 - default: has credit in default? (binary: "yes","no"), 채무불이행
6 - balance: average yearly balance, in euros (numeric), 연간 평균 잔액
7 - housing: has housing loan? (binary: "yes","no"), 주택담보대출
8 - loan: has personal loan? (binary: "yes","no"), 개인신용대출

related with the last contact of the current campaign:
9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12 - duration: last contact duration, in seconds (numeric), 통화지속시간

other attributes:
13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):
17 - y - has the client subscribed a term deposit? (binary: "yes","no"), 정기예금에 가입하였나?


Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
var
Data View
Variable View

SPSS 첫 화면

bank.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
1	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
2	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
3	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
4	30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
5	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
6	35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
7	36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
8	39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
9	41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
10	43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
11	39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
12	43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
13	36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
14	20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes
15	31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no
16	40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	-1	0	unknown	no
17	56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0	unknown	no
18	37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	152	2	failure	no
19	25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	-1	0	unknown	no
20	31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	152	1	other	no
21	38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	-1	0	unknown	no
22	42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	-1	0	unknown	no
23	44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	-1	0	unknown	no
24	44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	-1	0	unknown	no
25	26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	169	3	-1	0	unknown	no
26	41	management	married	tertiary	no	5883	no	no	cellular	20	nov	182	2	-1	0	unknown	no
27	55	blue-collar	married	primary	no	627	yes	no	unknown	5	may	247	1	-1	0	unknown	no
28	67	retired	married	unknown	no	696	no	no	telephone	17	aug	119	1	105	2	failure	no
29	56	self-employed	married	secondary	no	784	no	yes	cellular	30	jul	149	2	-1	0	unknown	no
30	53	admin.	married	secondary	no	105	no	yes	cellular	21	aug	74	2	-1	0	unknown	no
31	68	retired	divorced	secondary	no	4189	no	no	telephone	14	jul	897	2	-1	0	unknown	yes
32	31	technician	married	secondary	no	171	no	no	cellular	27	aug	81	3	-1	0	unknown	no
33	59	management	married	secondary	no	42	no	no	cellular	21	nov	40	1	-1	0	unknown	no
34	32	management	single	tertiary	no	2536	yes	no	cellular	26	aug	958	6	-1	0	unknown	yes
35	49	technician	married	tertiary	no	1235	no	no	cellular	13	aug	354	3	-1	0	unknown	yes

Data View Variable View

Data 가져오기

1. *.csv
2. *.xlsx
3. *.sav

- CSV 파일을 엑셀에서 불러오기
- CSV[comma separated value]
  - 쉼표(comma)를 기준으로 항목을 구분하여 저장한 데이터,
  - 수많은 애플리케이션에서 취급하는 범용 형식,
  - CSV 형식의 파일은 텍스트 파일로 보존하여 문서 처리기나 편집기에서 열람·편집할 수 있음.
- 엑셀데이터 불러오기(bank.xlsx)

*제목없음2 [데이터집합1] - IBM SPSS Statistics Data Editor

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 창(W) 도움말(H)

	age	job	marital	education	default	balance	housing	loan	contact	day	mo...	duration	campaign	pdays	previous	poutcome	y
4494	28.0	technician	single	tertiary	no	0	yes	no	unknown	4	jun	205	6	-1	0	unknown	no
4495	26.0	technician	single	secondary	no	668	yes	no	unknown	28	may	576	3	-1	0	unknown	yes
4496	48.0	management	married	tertiary	no	1175	yes	no	telephone	18	nov	1476	3	-1	0	unknown	no
4497	30.0	blue-collar	single	secondary	no	363	no	no	cellular	28	jul	171	3	-1	0	unknown	no
4498	31.0	entrepreneur	single	tertiary	no	38	no	no	cellular	20	nov	185	2	-1	0	unknown	no
4499	31.0	management	married	tertiary	no	1183	yes	no	unknown	27	may	676	6	-1	0	unknown	no
4500	45.0	blue-collar	divorced	primary	no	942	no	no	cellular	21	nov	362	1	-1	0	unknown	no
4501	38.0	admin.	married	secondary	no	4196	yes	no	cellular	12	may	193	2	-1	0	unknown	no
4502	34.0	management	married	tertiary	no	297	yes	no	cellular	26	aug	63	4	-1	0	unknown	no
4503	42.0	services	married	secondary	no	-91	yes	yes	cellular	5	feb	43	1	-1	0	unknown	no
4504	60.0	self-employed	married	primary	no	362	no	yes	cellular	29	jul	816	6	-1	0	unknown	yes
4505	42.0	blue-collar	single	secondary	no	1080	yes	yes	cellular	13	may	951	3	370	4	failure	yes
4506	32.0	admin.	single	secondary	no	620	yes	no	unknown	26	may	1234	3	-1	0	unknown	yes
4507	42.0	unemployed	divorced	tertiary	no	-166	no	no	cellular	29	aug	85	4	-1	0	unknown	no
4508	33.0	services	married	secondary	no	288	yes	no	cellular	17	apr	306	1	-1	0	unknown	no
4509	42.0	admin.	married	unknown	no	642	yes	yes	unknown	16	may	509	2	-1	0	unknown	no
4510	51.0	technician	married	tertiary	no	2506	no	no	cellular	30	nov	210	3	-1	0	unknown	no
4511	36.0	technician	divorced	secondary	no	566	yes	no	unknown	20	may	129	2	-1	0	unknown	no
4512	46.0	blue-collar	married	secondary	no	668	yes	no	unknown	15	may	1263	2	-1	0	unknown	yes
4513	40.0	blue-collar	married	secondary	no	1100	yes	no	unknown	29	may	660	2	-1	0	unknown	no
4514	49.0	blue-collar	married	secondary	no	322	no	no	cellular	14	aug	356	2	-1	0	unknown	no
4515	38.0	blue-collar	married	secondary	no	1205	yes	no	cellular	20	apr	45	4	153	1	failure	no
4516	32.0	services	single	secondary	no	473	yes	no	cellular	7	jul	624	5	-1	0	unknown	no
4517	33.0	services	married	secondary	no	-333	yes	no	cellular	30	jul	329	5	-1	0	unknown	no
4518	57.0	self-employed	married	tertiary	yes	-3313	yes	yes	unknown	9	may	153	1	-1	0	unknown	no
4519	57.0	technician	married	secondary	no	295	no	no	cellular	19	aug	151	11	-1	0	unknown	no
4520	28.0	blue-collar	married	secondary	no	1137	no	no	cellular	6	feb	129	4	211	3	other	no
4521	44.0	entrepreneur	single	tertiary	no	1136	yes	yes	cellular	3	apr	345	2	249	7	other	no
4522																	

*제목없음2 [데이터집합1] - IBM SPSS Statistics Data Editor

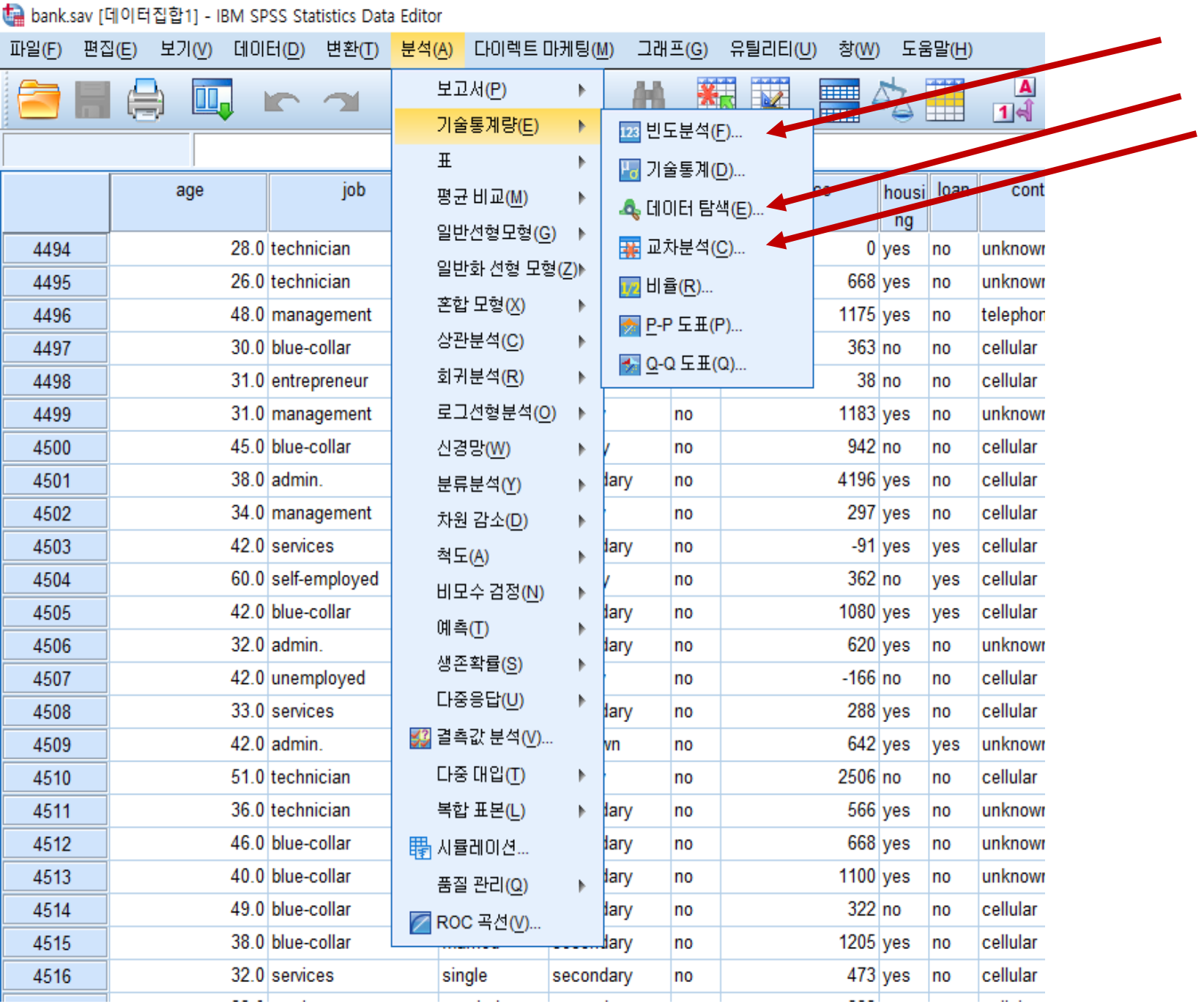
파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 창(W) 도움말(H)

	이름	유형	너비	소수점이...	설명	값	결측값	열	맞춤	측도	역할
1	age	숫자	12	1		없음	없음	12	오른쪽	척도(S)	입력
2	job	문자	13	0		없음	없음	13	왼쪽	명목(N)	입력
3	marital	문자	8	0		없음	없음	8	왼쪽	명목(N)	입력
4	education	문자	9	0		없음	없음	9	왼쪽	명목(N)	입력
5	default	문자	3	0		없음	없음	3	왼쪽	명목(N)	입력
6	balance	숫자	12	0		없음	없음	12	오른쪽	척도(S)	입력
7	housing	문자	3	0		없음	없음	3	왼쪽	명목(N)	입력
8	loan	문자	3	0		없음	없음	3	왼쪽	명목(N)	입력
9	contact	문자	9	0		없음	없음	9	왼쪽	명목(N)	입력
10	day	숫자	12	0		없음	없음	12	오른쪽	척도(S)	입력
11	month	문자	3	0		없음	없음	3	왼쪽	명목(N)	입력
12	duration	숫자	12	0		없음	없음	12	오른쪽	척도(S)	입력
13	campaign	숫자	12	0		없음	없음	12	오른쪽	척도(S)	입력
14	pdays	숫자	12	0		없음	없음	12	오른쪽	척도(S)	입력
15	previous	숫자	12	0		없음	없음	12	오른쪽	척도(S)	입력
16	poutcome	문자	7	0		없음	없음	7	왼쪽	명목(N)	입력
17	y	문자	3	0		없음	없음	3	왼쪽	명목(N)	입력

Data Type 정하기,
- 숫자
- 문자

척도 정하기

- 연속형(수치형)
  - Scale
  - Numeric
- 범주형(이산형)
  - Nominal
  - Categorical
- 순서형(Ordinal)



Data 탐색하기(**빈도분석**)
.평균
.표준편차
.첨도/왜도
.분포 모양(흩어진 정도)
.히스토그램
.Box-Plot 등

연속형 변수와 타겟변수간 관련성 탐색
(타겟변수에 영향을 줄 수 있는 연속형 변수 찾기)
.Explore(**데이터탐색**)

이산형 변수와 타겟변수간 연관성 분석
.crosstabs(**교차분석**)

평균, 편차, 분산, 표준편차

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

$$x_i - \bar{x}$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

	국어	수학
점수	90	70
평균	70	50
표준편차	10	5

평균, 분산, 표준편차

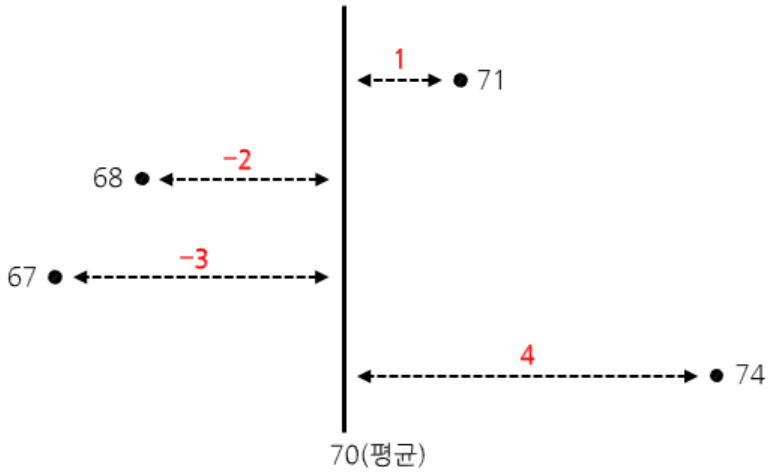
$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

X: 키

X1: 1번째 사람의 키
X2: 2번째 사람의 키
…



$(x_i - \bar{x})$ ← 편차

$(x_i - \bar{x})^2$ ← 편차 제곱

$\sum(x_i - \bar{x})^2$ ← 편차 제곱합

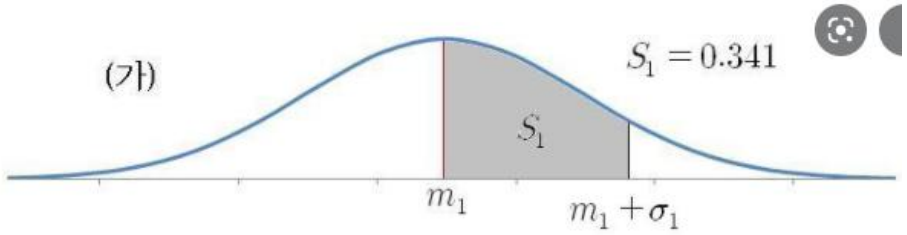
제곱합

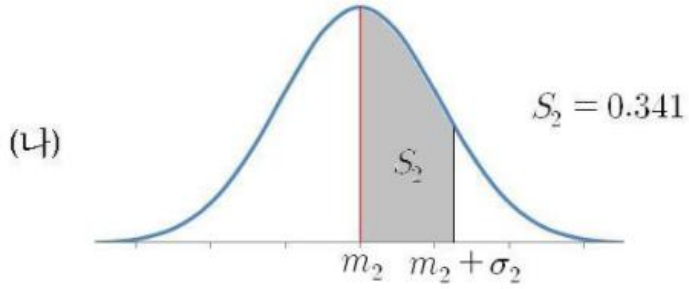
$$\frac{\sum(x_i - \bar{x})^2}{(n-1)}$$ ← 편차 제곱의 평균(**분산**)

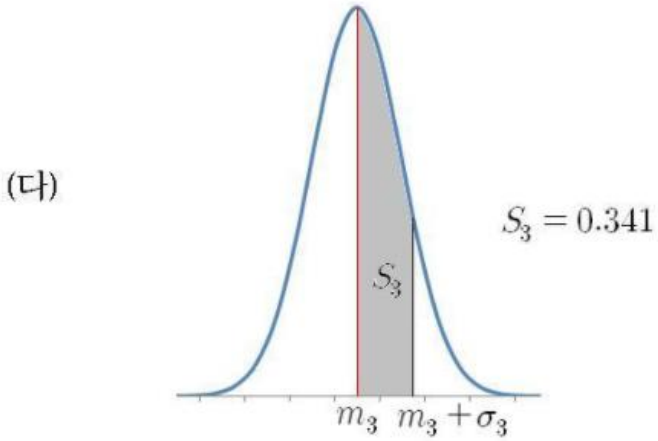
Start	1번	2번	3번	4번	5번
변량	175	177	179	181	183
평균 $\left(\frac{\text{변량의 합}}{\text{인원수}}\right)$	$= \frac{175+177+179+181+183}{5} = 179$				
편차 (변량-평균)	175-179 =-4	177-179 =-2	179-179 =0	181-179 =2	183-179 =4
편차제곱 (편차*편차)	(-4)*(-4) =16	(-2)*(-2) =4	(0)*(0) =0	(2)*(2) =4	(4)*(4) =16
분산 (편차제곱의 평균)	$= \frac{16+4+0+4+16}{(5-1)}$				
표준편차 $(\sqrt{\text{분산}})$					

평균, 분산, 표준편차

Start	1번	2번	3번	4번	5번
변량	175	175	175	175	175
평균 (변량의 합 / 인원수)	$= \frac{175+175+175+175+175}{5} = 175$				
편차 (변량-평균)	175-175 =0	175-175 =0	175-175 =0	175-175 =0	175-175 =0
편차제곱 (편차*편차)	(0)*(0) =0	(0)*(0) =0	(0)*(0) =0	(0)*(0) =0	(0)*(0) =0
분산 (편차제곱의 평균)	$= \frac{0+0+0+0+0}{(5-1)}$				
표준편차 ($\sqrt{분산}$)	$= \sqrt{0} = 0$				

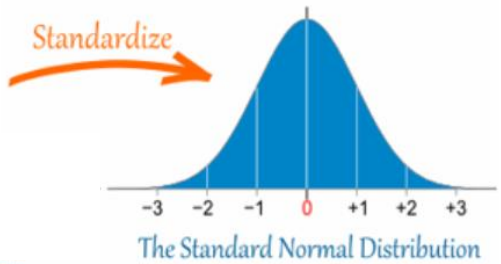






표준화

$$Z = \frac{X - \mu}{\sigma}$$



	국어	수학
점수	90	70
평균	70	50
표준편차	10	5
표준점수	2.0	4.0

$$\frac{\sum (X_i - \bar{X})^3 / N}{S^3}$$

왜도(skewness)

Definition of Skewness

For univariate data $Y_1, Y_2, \ldots, Y_N$ the formula for skewness is:

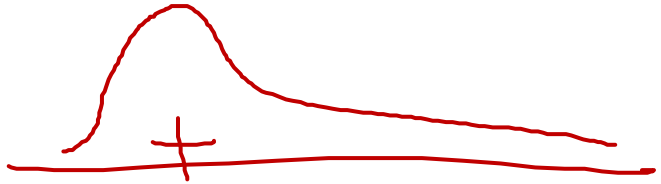
$$g_1 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^3 / N}{s^3}$$

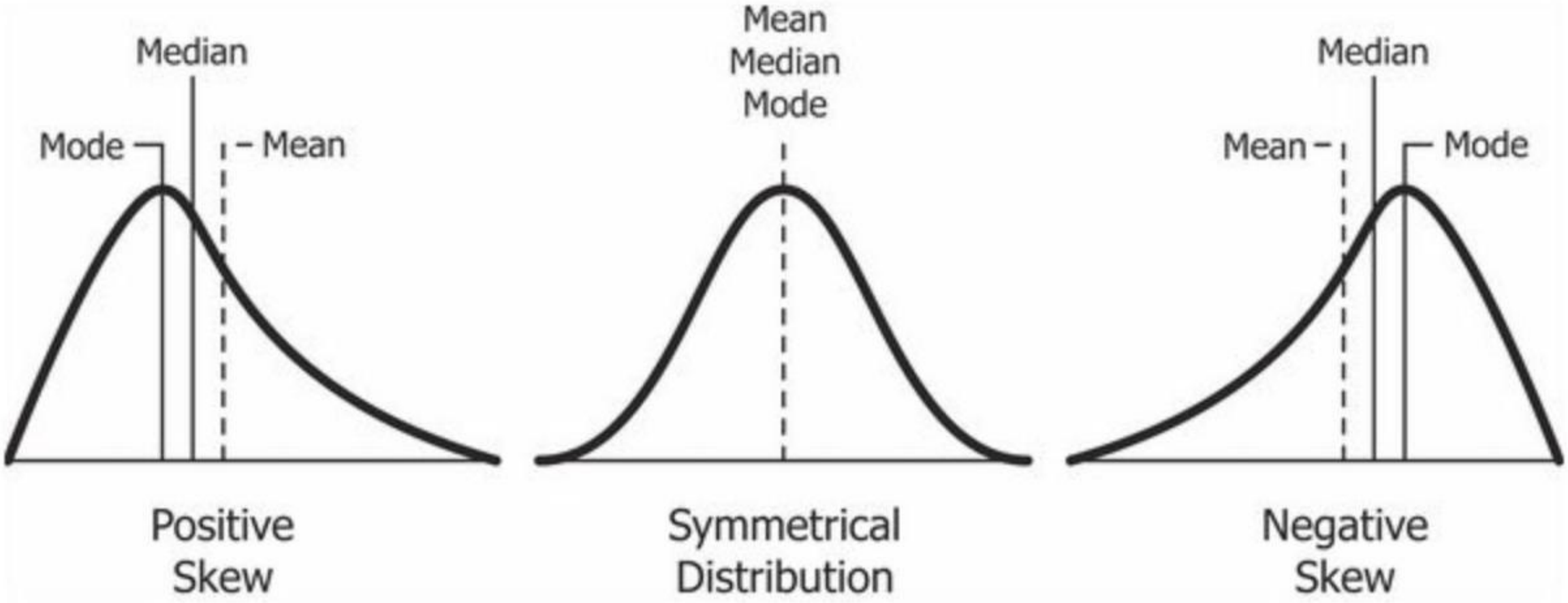
where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that in computing the skewness, the s is computed with N in the denominator rather than $N - 1$.

왜도가 음수일 경우에는 확률밀도함수의 왼쪽 부분에 긴 꼬리를 가지며 중앙값을 포함한 자료가 오른쪽에 더 많이 분포해 있다.

왜도가 양수일 때는 확률밀도함수의 오른쪽 부분에 긴 꼬리를 가지며 자료가 왼쪽에 더 많이 분포해 있다는 것을 나타낸다.

평균과 중앙값이 같으면 왜도는 0이 된다.




Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

$\Sigma (X_i - \bar{X})^4 / N$

첨도(kurtosis)

Definition of Kurtosis

For univariate data $Y_1, Y_2, ..., Y_N$ the formula for kurtosis is:

$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that in computing the kurtosis, the standard deviation is computed using N in the denominator rather than $N - 1$.

Alternative Definition of Kurtosis

The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis (often referred to as "excess kurtosis"):

SPSS

$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4} - 3 = 0$$

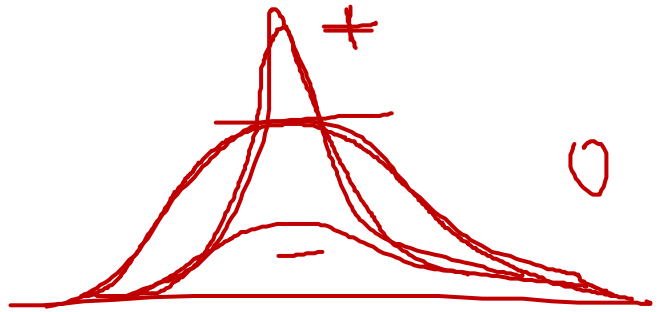
This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with the second definition positive kurtosis indicates a "heavy-tailed" distribution and negative kurtosis indicates a "light tailed" distribution.

3

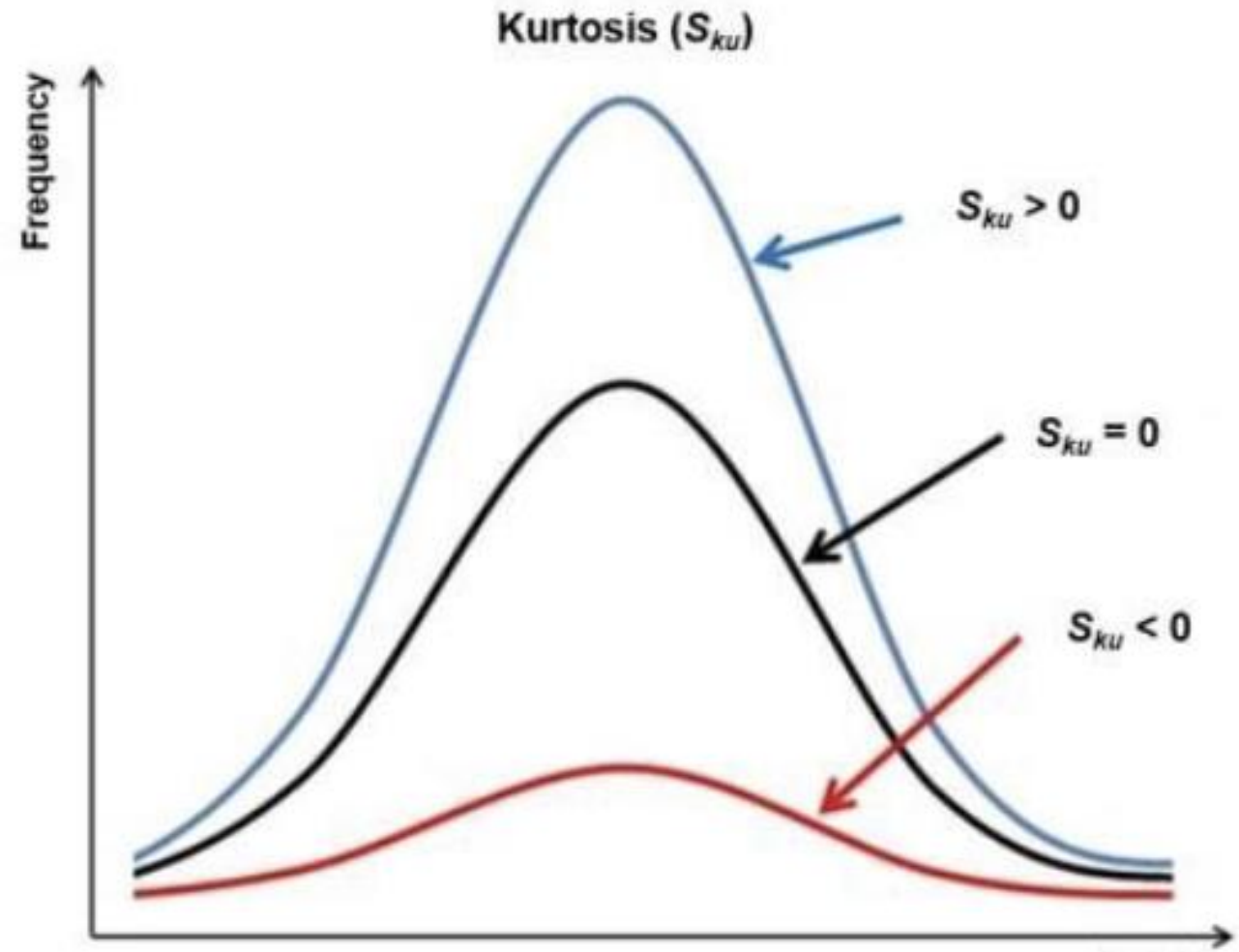
첨도값(K)이 3에 가까우면 산포도가 정규분포에 가깝다.

3보다 작을 경우에는(K<3) 정규분포보다 더 완만하게 납작한 분포로 판단할 수 있으며,

첨도값이 3보다 큰 양수이면(K>3) 산포는 정규분포보다 더 뾰족한 분포로 생각할 수 있다.

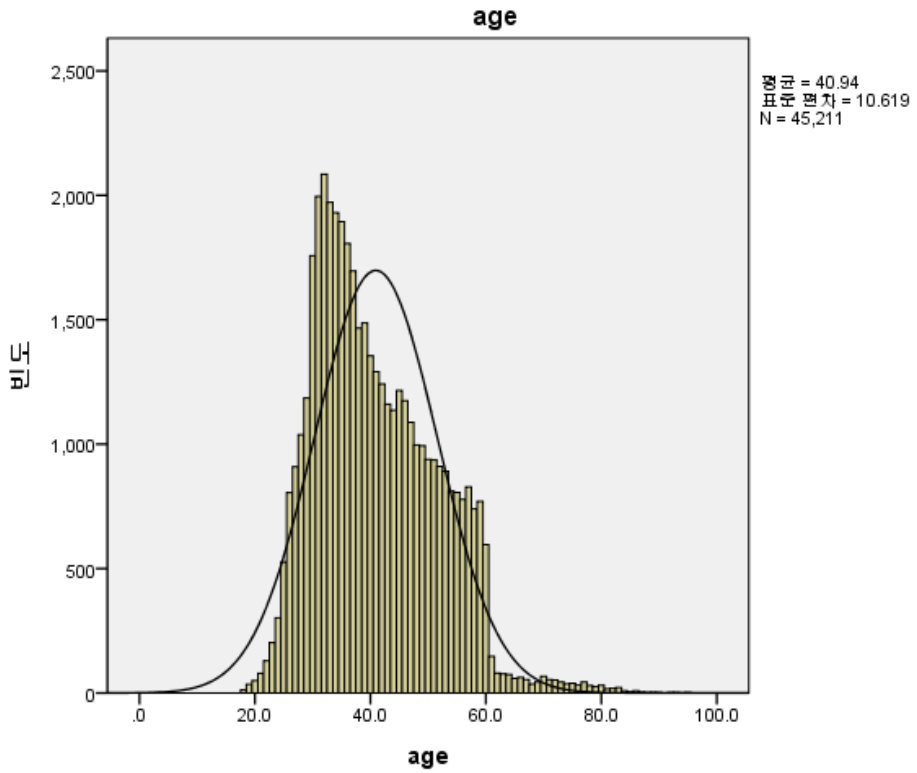


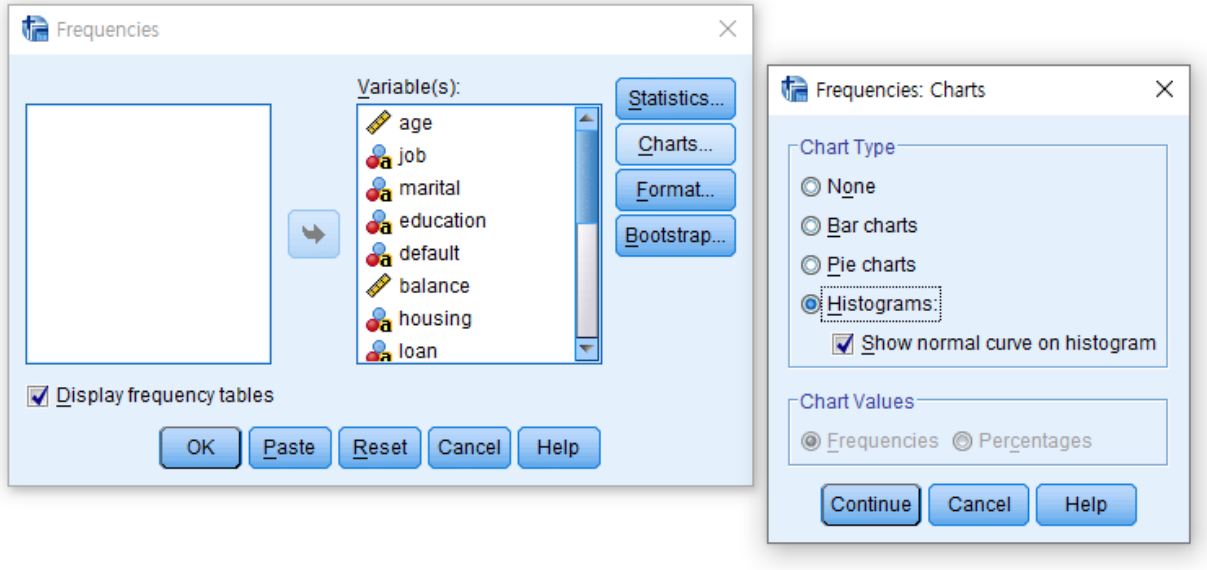
SPSS에서는 0 값이 기준


Kurtosis (S_{ku})
Frequency
$S_{ku} > 0$
$S_{ku} = 0$
$S_{ku} < 0$

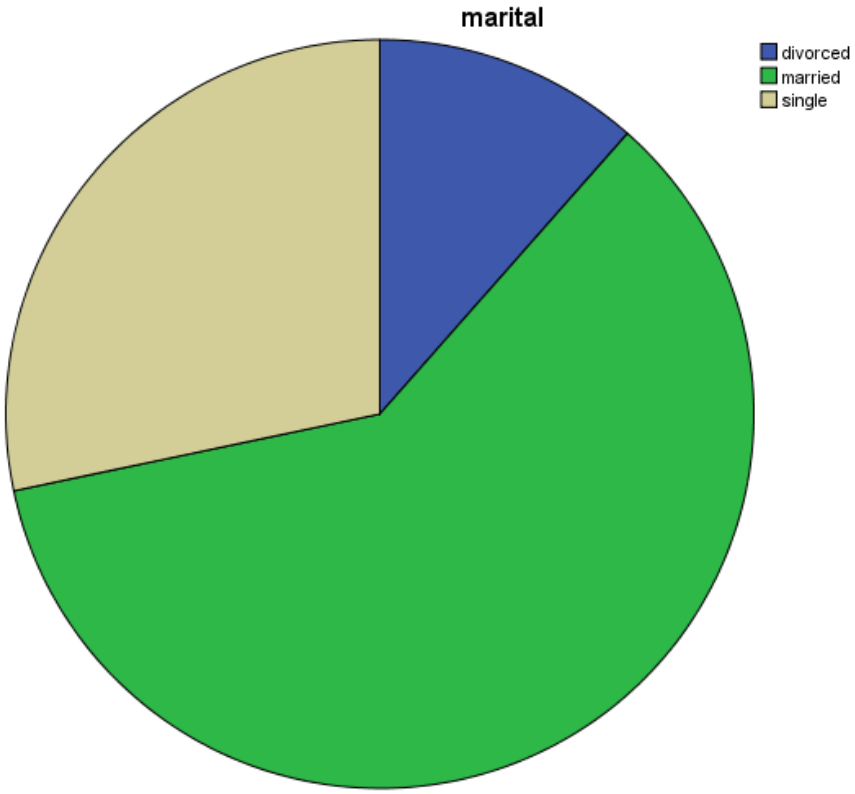
Histograms

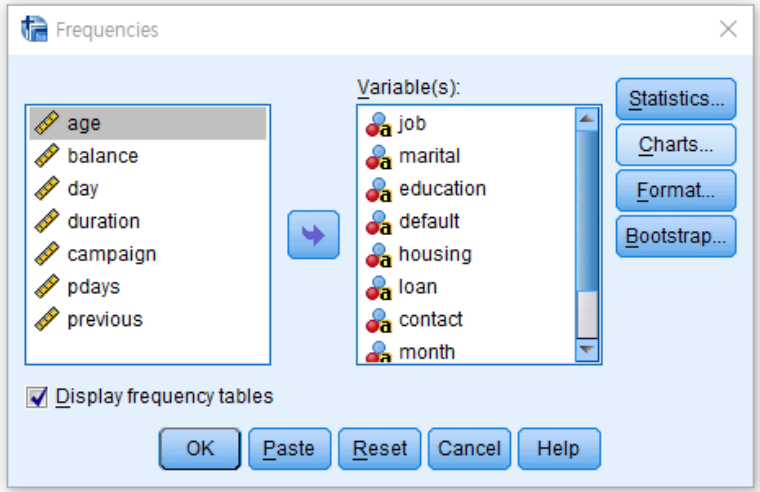
히스토그램

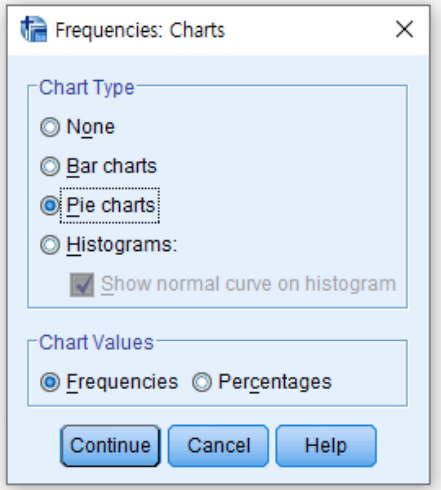




Pie Charts


marital
divorced
married
single


Frequencies
Variable(s):
age
balance
day
duration
campaign
pdays
previous
job
marital
education
default
housing
loan
contact
month
Statistics...
Charts...
Format...
Bootstrap...
Display frequency tables
OK
Paste
Reset
Cancel
Help


Frequencies: Charts
Chart Type
None
Bar charts
Pie charts
Histograms:
Show normal curve on histogram
Chart Values
Frequencies
Percentages
Continue
Cancel
Help

엑셀에서

- **Age, balance, duration 변수에 대한**
  - 각각의 총합(sum), 평균(Average), 최대값(Max), 최소값(Min)을 구하기
  - 분산, 표준편차 엑셀에서 구하기(되도록 함수 사용하지 않기)
  - 교재(P.188~195) 참고하기

=sum(A2:A4521)