<산업보건연구원 – Meeting 자료>

- 6월 2일 Version
- 작성자 : 이은경

<What TO DO>

: "BYEAR" 변수의 unique element 확인 + 분포 파악
(생년이 어느 연도부터 시작인지 확인 / 이상치 존재하는지 확인)

: "BMONTH", "BDAY" 변수의 unique element 재확인 (이상치 존재하는지 확인)

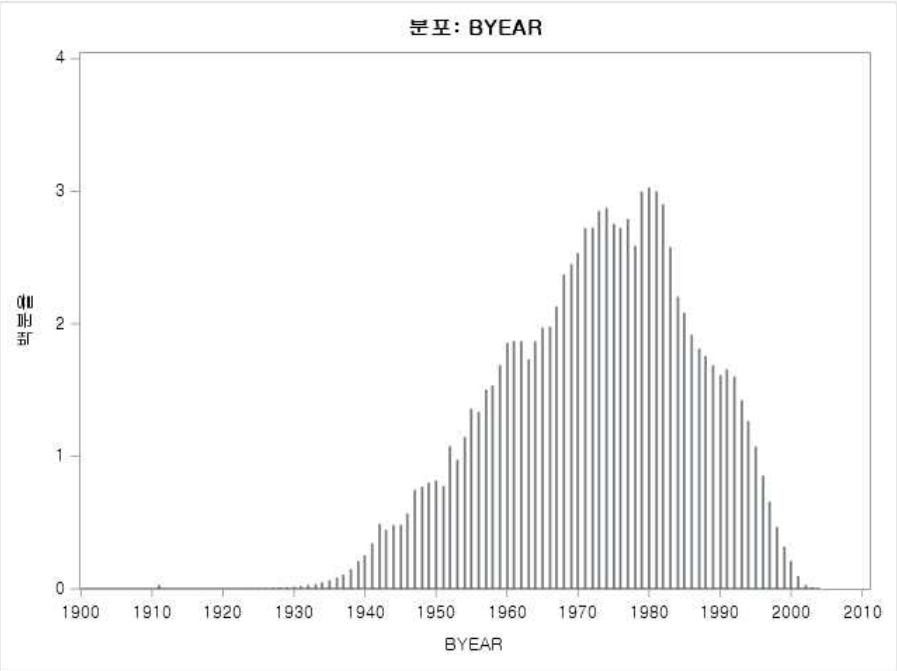
: Personal Information table 생성하기 전, 데이터 정제 + Personal Information table 생성

<Share & Result>

1) "BYEAR"

: 최솟값이 '1900'년도 / 이상치 제외 최댓값은 '2010'년도 / 결측치도 존재

["BYEAR" 변수 히스토그램]



[BYEAR 분포표]

	Min	Q1	Mean	Q2	Q3	Max
BYEAR	1900	1963	1972.64	1974	1983	2004

: 생년이 1960년대 ~ 1980년대에 몰려 있음을 알 수 있다.

2) "BMONTH"
: 결측치, '00' 값도 존재 / 이상치라 여겨지는 값은 '13' ~ '99'

3) "BDAY"
: 결측치, '00' 값도 존재 / 이상치라 여겨지는 값은 '32' ~ '99'

4) Personal Information table 생성하기 전, 이상치 제외 필요한 변수만 가져옴.
: "INDI_ID" 변수 값이 '1000000000001'이 아님 + "BYEAR"가 2004년 이하
+ "BMONTH"가 12 이하 + "BDAY"가 31 이하 조건 부여

: "INDI_ID", "SEX", "BYEAR", "BMONTH", "BDAY" 변수만 KEEP
: "BYEAR", "BMONTH", "BDAY" 변수 수치형으로 모두 변환
--- "data" tbl

("data" table example)

	성별	INDI_ID	BYEAR	BMONTH	BDAY
15	남	1000001846316	1946	10	6
16	남	1000001846316	1946	10	6
17	남	1000001846316	1946	10	6
18	남	1000001846317	1946	10	6
19	남	1000001846317	1946	10	6
20	남	1000001846317	1946	10	6
21	남	1000001846317	1946	10	6
22	남	1000001846317	1946	10	6

⇒ INDI_ID가 같으면 "성별", "BYEAR", "BMONTH", "BDAY" 변수 값 모두 동일하므로, "PROC SORT NODUPKEY" option 사용해 unique row만 가져옴.

(Final Personal Information table example)

	성별	INDI_ID	BYEAR	BMONTH	BDAY
	남	1000000000002	2000	1	1
	남	1000000000003	2000	1	1
	남	1000000000004	2000	1	1
	남	1000000000005	2000	1	1
	남	1000000000006	2000	1	1
	남	1000000000007	2000	1	1
	남	1000000000008	2000	1	1
	남	1000000000009	2000	1	1
	남	1000000000010	2000	1	1
	남	1000000000011	2000	1	1
	남	1000000000012	2000	1	1
	남	1000000000013	2000	1	1
	남	1000000000014	2000	1	1
	남	1000000000015	2000	1	1
	남	1000000000016	2000	1	1
	남	1000000000017	2000	1	1
	남	1000000000018	2000	1	1
	남	1000000000019	2000	1	1
	남	1000000000020	2000	1	1
	남	1000000000021	2000	1	1

총 관측치 수 = "31360814"