<산업보건연구원 Data 파악 결과 정리>

- 2월 14일 Version

<What TO DO>

: USB로 받은 파일 잘 열리나 확인 / Data 구성하는 변수 간단하게 확인

<Share & Result>

1) SAS 상에서 데이터는 잘 열림을 확인

2) 간단한 Procedure 결과 확인하는 부분임에도 소요시간이 꽤 걸림.

2)-⓪. Data label과 실제 변수명이 다르므로 변수명 먼저 확인하는 과정이 필요해 보임.

-- 변수명이 의미하는 것이 무엇인지도 알아야 함.

2)-①. 연령 / 근속 연수 / 등록 일자 / 종료 일자 / 사망 연월 등 수치형일 것이라 예상되는 변수들이 모두 문자형임을 확인 → 수치형으로 변환하는 과정이 필요할 수도 있을 것 같음.

2)-②. 근속 연수(Duration) 변수값은 무조건 양수이어야 할 것이라 기대되나, 음수인 객체가 소수 존재함을 확인. → Duration이 -1 / -2 / -7 / -8인 record가 여럿 존재함.

근속년수

DURATION	빈도	백분율	누적 빈도	누적 백분율
-1	6	0.00	6	0.00
-2	1	0.00	7	0.00
-7	1	0.00	8	0.00
-8	7	0.00	15	0.00

2)-③. "BYEAR" 변수

: 2023년을 초과한 연도 값이 존재

2043	1
2047	1
2049	2
2050	1
2058	1
2059	1
2060	1
2066	1
2068	3
2069	1
2071	3
2073	1
2074	2
2076	3
2077	1
2078	2
2079	1
2080	2
2083	1
2088	1

2)-④. "BMONTH" 변수

: 0 ~ 12 값이 대부분이나, 13 ~ 99 범주 내 값을 가지는 관측치가 여럿 존재함을 확인.

2)-⑤. "BDAY" 변수

: 0 ~ 31 값이 99.15% 비중을 차지하나, 32 ~ 99 범주 내 값을 가지는 관측치가 여러 개 있음.

2)-⑥. “AGE” 변수

: 변수값이 음수, 0인 관측치,
100 혹은 200이 넘는 값을 가지는 관측치도 존재함을 확인함.

AGE	빈도
-1	1
-40	1
-42	1
-46	2
-47	1
-49	1
-53	1
-57	1
-64	1
-65	2
-66	2
-67	1
-68	2
-69	2
-70	2
-71	1
-73	1
-74	2
-75	4
-79	1
-81	1
0	1

→ 해당 객체들의 취득일 / 상실일 확인

나이	취득일	상실일
	20020807	20110419
	20020807	
	20170216	20180401
	20180402	20180815
-40	20040116	
-42	20050516	20060601
-46	20040105	20040328
-46	20040105	20040327
-47	20040402	20040701
-49	20100101	
-53	20070201	20130601
-57	20031229	
-64	20031117	20040713
-65	20040401	
-74	19950801	19960314
-69	20001101	20010519
-65	20040302	20040621
-67	20040301	20040401
	20090501	20100618
-68	20040219	20040731
-68	20040201	20050101
-70	20040614	20041025
-71	20031201	20051011
-70	20041013	20061130
-73	20040101	20040229
-75	20010301	20011101
-75	20010301	20020618
-69	20090807	
-75	20040101	20100101
-75	20040501	20040725
-74	20050214	
-66	20150302	20150627
-66	20150706	20151031
-81	20030506	20070101
-79	20091124	
0	19950701	19950816
-1	19980101	19980501

나이	취득일	상실일
108	20081013	20101130
108	20081103	20100603
109	20090512	20100120
109	20090811	20100719
109	20090907	20100430
110	20100101	
110	20100101	20101231
110	20100624	20100724
110	20100629	20101031
110	20100709	
110	20100809	
110	20101201	
110	20101230	20101231
110	20101230	20101231
110	20101230	20101231
110	20101230	20101231

: 취득일 > 상실일인 경우가 없음에도 불구하고 AGE가 음수 / 이유 파악 필요해 보임.
: AGE 값이 100 이상인 객체들도 100 이상인 이유를 짐작할 수 없음.

---

2)-⑦. “INDI_ID” 변수

: ‘사업장 주소’ / ‘연번’이 다름에도 불구하고 “INDI_ID” 변수가 같은 객체가 여럿 존재함.