20230619_Data_EDA_Plan(공통처리)

담당자	은경 이
상태	진행 중
마감일	@2023년 6월 20일

📋 참고 문서

사업장 관련 변수 전처리 및 전체 전처리 계획

Goal)

: 공통적으로 처리해야 하는 Data EDA 과정 거친 뒤, 각자 필요한 변수만 KEEP한 raw data 구축

What TO DO)

: 포함조건, 제외조건 파악 / 기존에 진행했던 Data quality check 공유 / Data split / Code 구상

Share & Result

0. 사용하지 않는 변수 제거

: 사망DB column + 정의 방식이 잘못되어 있는 “AGE”, “DURATION” 변수 삭제

(사용하지 않은 변수 목록)

AGE	DURATION	DREGDATE	REAL_AUTH_CODE1
REAL_AUTH_CODE2	DTH_POS	MRR_STATUS	DTH_EDU
DTH_AGE	NATIONALITY_CLASS_F	NATIONALITY_F	CODE_103
CODE_56	DTH_JOB	NEW_DTH_CODE1	NEW_DTH_CODE2

1. 생년월일 value Limitation

: BYEAR(생년)은 1940년 ~ 1999년 / BMONTH(월)은 1월 ~ 12월 / BDAY(일)은 1일 ~ 31일로 값 제한

— 값 제한한 후, distinct INDI_ID 얼마나 삭제되었는지 N수 count 필요함.

```
/* 0번, 1번 동시 처리 code */
DATA temp;
SET dir.Db3;
```

```
IF (input(BYEAR, 12.) between 1940 and 1999) and (input(BMONTH, 12.) between 1 and 12)
and (input(BDAY, 12.) between 1 and 31);
DROP AGE DURATION DREGDATE REAL_AUTH_CODE1 REAL_AUTH_CODE2 DTH_POS MRR_STATUS
     DTH_EDH DTH_AGE NATIONALITY_CLASS_F NATIONALITY_F CODE_103 CODE_56 DTH_JOB
     NEW_DTH_CODE1 NEW_DTH_CODE2;
run;

/* N수 확인 -- 기존 N수 - 정제한 data N수 */
proc sql;
create total_N as select distinct INDI_ID from dir.Db3;
quit;

proc sql;
create table count_N as select distinct INDI_ID from temp;
quit;
```

2. “ECNY_DT”(취득일) = “OUT_DT”(상실일) 혹은 ECNY_DT가 2019년 이후인 관측치 삭제

: “ECNY_DT” = “OUT_DT” 인 것은 기록 상의 오류이므로 해당 관측치 삭제 필요

+ 암 추적 기간이 2018년까지이므로 고용보험 취득일이 2019년 이후인 관측치 또한 삭제 필요

— 삭제되는 관측치 몇 개 인지 파악하기

```
DATA temp2;
SET temp;
IF ECNY_DT = OUT_DT or ECNY_DT >'20190101' then delete;
run;
```

3. Data quality check

: 데이터 입력 방식이 맞지 않는 관측치 삭제

3-1) “ECNY_DT” 변수 길이가 8이 아니거나 OUT_DT가 결측이 아니면서 변수 길이가 8이 아닌 관측치 제외

: “ECNY_DT”, “OUT_DT” 변수 형태가 기본적으로 YYYYMMDD 이므로 길이가 8이어야 함.

3-2) “OUT_DT”(상실일) 변수가 결측인 경우는 아직 근무 지속 중임을 의미

: 마지막 follow-up 기간이 2018년까지로 정했으므로 해당 변수 값을 **‘20181231’로 대체**

```
/* 2번, 3번 동시 처리 code */
DATA temp3;
SET temp2;
IF length(ECNY_DT) != 8 or (not missing(OUT_DT) and length(OUT_DT)!=8) then delete;
IF OUT_DT='' then OUT_DT='20181231';
run;
```

4. “AGE”(연령), “DURATION”(근속연수) 변수 재생성

: “AGE”, “DURATION” 변수 값을 살펴보니 잘못된 부분이 많음을 확인

— 연령이 음수 혹은 100 초과 / 근속연수가 결측인 경우 존재

: 변수 생성 후, “DURATION”(근속연수)가 음수인 관측치는 삭제

— (ECNY_DT > OUT_DT로 인해 발생한 error)

```
DATA temp4;
SET temp3;
AGE = round((input(ECNY_DT, yymmdd8.)-input(catx(BMONTH, BDAY, BYEAR), yymmdd8.))/365.25,
            .01);
DURATION = round((input(OUT_DT, yymmdd8.) - input(ECNY_DT, yymmdd8.))/365.25, .01);
IF DURATION<0 then delete;
run;
```

5. 최초 고용보험 등록 이전 암 발생한 객체 제외

: 순수하게 노동으로 인한 폐암 / 백혈병 발생 위험률 추정하고자 하는 것이므로 최초 고용보험 등록 이전 암 발생한 객체는 제외 필요

5-1) 중복 행 제거

: “NO”, “INDI_ID”, “ECNY_DT”, “OUT_DT”, “AGE”, “DURATION” 변수 값이 모두 같은 관측치가 여러 행이 존재하는 것을 이전에 발견함.

53597415	373412	19980618	20000401	1000001273396	55	2
53597416	373412	19980618	20000401	1000001273396	55	2
53597417	373412	19980618	20000401	1000001273396	55	2
53597418	373412	19980618	20000401	1000001273396	55	2
53597419	373412	19980618	20000401	1000001273396	55	2
53597420	373412	19980618	20000401	1000001273396	55	2
53597421	373412	19980618	20000401	1000001273396	55	2
53597422	373412	19980618	20000401	1000001273396	55	2
53597423	373412	19980618	20000401	1000001273396	55	2
53597424	373412	19980618	20000401	1000001273396	55	2
53597425	373412	19980618	20000401	1000001273396	55	2
53597426	373412	19980618	20000401	1000001273396	55	2
53597427	3814311	19950701	19961231	1000001273397	52	1
53597428	3814311	19950701	19961231	1000001273397	52	1
53597429	3814311	19950701	19961231	1000001273397	52	1
53597430	3814311	19950701	19961231	1000001273397	52	1
53597431	3814311	19950701	19961231	1000001273397	52	1
53597432	3814311	19950701	19961231	1000001273397	52	1
53597433	3814311	19950701	19961231	1000001273397	52	1
53597434	3814311	19950701	19961231	1000001273397	52	1
53597435	3814311	19950701	19961231	1000001273397	52	1
53597436	3814311	19950701	19961231	1000001273397	52	1
53597437	3814311	19950701	19961231	1000001273397	52	1
53597438	3814311	19950701	19961231	1000001273397	52	1
53597439	3814311	19950701	19961231	1000001273397	52	1

→ 암 DB column 값 까지 모두 동일하다면, 필요없는 관측치임.

```
/* 전체 column 값이 동일한 경우 제거하는 option 사용 */
proc sort data=temp4 noduprecs;
by INDI_ID;
run;
```

5-2) 최초 고용보험 등록일과 가장 과거에 발병한 암 진단일자 비교

: 최초 고용보험 등록일 이전에 암 이력이 있는 사람은 제외

```
proc sort data = temp4;
by INDI_ID ECNY_DT;
run;

DATA temp5;
SET temp4;
BY INDI_ID ECNY_DT;
IF first.INDI_ID;
IF ECNY_DT < min(fdx1-fdx6) then do;
  output exclude;
  delete;
end;
run;
```

→ “exclude” table N수 count 확인 필요

6. Data Split

: 서로가 필요한 변수는 다르므로, 정제한 데이터 기준으로 Data split 진행

(윤서)

(취득일, 상실일)	최초 고용보험 등록 일자	근속 연수	업종정보 (대분류, 중분류, 소분류)	직종 코드	직종 차수	상시 근로자 수	사업장 주소 정보

(은경)

생년월일	최초 고용보험 일자 + entry (최초 등록 연도 범주화 변수)	성별	AGE (입사 당시 나이)	근속 연수	(취득일, 상실일)	암 이력 + 진단 일자	사망 여부 + 사망 일자	Outcome (폐암 / 백혈병 진단 여부 + 시점)

✔️ TO DO LIST

1) Demographic Information table 생성

: “BYEAR” + “BMONTH” + “BDAY” 변수 합쳐서 “BIRTH_DATE” 변수 새로 생성

: 다른 table과의 joint key는 “INDI_ID”

INDI_ID	BIRTH_DATE	SEX	first_ECNT_DT	ENTRY

2) 입사 당시 연령 / 근속연수 / (취득일, 상실일) 정보 담긴 table 생성

: 모두 time-varying covariate이며, 다른 table과의 joint key는 “INDI_ID”, “NO”

INDI_ID	NO	ECNY_DT	OUT_DT	AGE	DURATION

→ 해당 table 정의한 뒤, 모두 값이 같은 관측치가 여러 개 존재하는 경우 있는지 재확인

(Comment)

: 각 반복 측정치의 (취득일, 상실일)을 (취득일 - 첫 고용보험 등록일, 상실일 - 첫 고용보험 등록일)형태로 바꾸어줄 필요가 있음.

→ Duration 형태로 바꾸어줘야 함을 의미.