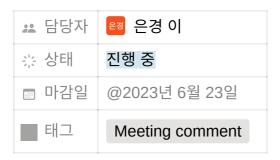
20230623_Meeting_comment_정



Meeting Comment 정리)

- Data EDA 재진행 필요
- 1) 취득일(ECNY_DT) , 상실일(OUT_DT) 변수 값 형태가 이상한 관측치를 가지는 객체의 기록 전부 제거

(해당 Data Example)

: "ECNY_DT", "OUT_DT" 변수 길이가 8이 아닌 경우



- 2) 재정의한 "DURATION" 변수값이 음수인 관측치를 가지는 객체의 기록 전부 제거
- 3) 사업장 정보 변수 중 하나라도 결측이 있는 관측치를 가지는 객체의 기록 전부 제거

4) 개인 별 근무일 중첩이 있는 객체(즉, two-job인 객체 의미)의 기록 전부 제거

(Data Example)



- 업종, 직종, 직종차수, 우편번호, 종사자수 column이 없는 개인 (또는 row)를 배제하는 게 분석인 편하나, 우선 <u>사업장 정보 변수 missing에 패턴이 있는지 살펴볼 필요는 있</u>음.
- 해당 관측치 전부 추후에는 포함할 필요 있음

5) 추후 보고를 위해 아래 사항 파악할 필요가 있음.

- Raw data의 distinct INDI ID: xx명
- Data quality check한 후 이상치 존재(사업장 변수 missing도 포함)하는 객체 제외한 후 파악한distinct INDI_ID: xx명
- 근무기록 중 중첩이 있는 개인 제거한 후 파악한 distinct INDI_ID: xx명
- 최초입사 전 암발병자 제거한 후 파악한 distinct INDI ID: xx명

📋 Outcome 변수 / 공변량 재정의

1) "개인 암 이력 변수" 재정의 필요

: Outcome이 폐암인 경우, <u>백혈병도 기타 암 이력에 포함</u>해야 함. (백혈병에 대해서도 마찬 가지)

→ Outcome 별 개인 암 이력 변수 따로 생성하는 것이 better

2) "Entry" covariate는 Dummy variable로 재생성

: 즉, "1995~1999년 첫 입사" / "2000 ~ 2004년 첫 입사" ... "2015~2018년 첫 입사" 각각 변수 따로 생성해 1 아니면 0의 값을 가지는 이변량 변수로 재생성하는 것이 better