<산업 안전 보건 연구원 - Data EDA Check 사항 정리>

- 6월 25일 Version
- 6월 26일 Update
- 작성자 : 이은경

<Data EDA 항목>

1) 해당 조건을 만족하는 "객체"의 기록 전부 제거

1)-① Data quality check

* 취득일(ECNY_DT), 상실일(OUT_DT), 암 진단일자(fdx1-fdx6) 변수 값의 길이가 8이 아닌 경우
  ↳ 각 변수 값의 길이가 8이긴 하지만, 실제 달력 상에 존재하지 않는 날짜의 경우도 포함해서 제외
  ex) 1999.13.31 / 1999.04.31. 등

* 사망 일자(DTH_DATE1-DTH_DATE3) 변수가 결측이 아닌데 각각 변수 값의 길이가 4자, 2자, 2자가 아닌 경우
  ↳ 각 변수 값의 길이가 4자, 2자, 2자이긴 하지만, 실제 달력 상에 존재하지 않는 날짜도 포함해서 제외
  ex) 1999.13.31 / 1999.04.31. 등

* 취득일(ECNY_DT) ≥ 상실일(OUT_DT)(↔ "취득일 - 상실일" 값이 0 이하)인 경우

* 생년월일 값이 1940년 ~ 1999년 사이 + 1월 ~ 12월 사이 + 1일 ~ 31일 사이가 아닌 경우
  : 해당 조건 통해 AGE(입사 시 연령) 변수 값이 100 초과이거나 음수인 객체 제거 가능
  + INDI_ID가 '1000000000001'인 객체도 제외 가능
  ↳ 각 변수 값의 길이가 4자, 2자, 2자이긴 하지만, 실제 달력 상에 존재하지 않는 날짜도 포함해서 제외
  ex) 1999.13.31 / 1999.04.31. 등

1)-② 특정 제외조건

* 최초 고용보험 등록일 전에 암 이력이 존재하는 경우
* 고려하는 사업장 정보 변수 중 하나라도 missing이 있는 경우
* 근무 일자 중 겹치는 (즉, two-job) 기간이 존재하는 경우

2) 특정 조건 만족하는 "관측치"만 제거

2)-① 고용보험 취득일(ECNY_DT)가 2018.12.31.이후인 관측치 제거
↳ 암DB 기록이 2018.12.31.까지만 존재하기 때문

3) 특정 값 "대체"

3)-① OUT_DT(상실일)이 2018.12.31.이후이거나 결측인 관측치의 경우 모두 "2018.12.31."로 변경
↳ 추적 종료 시점이 모두 2018.12.31.이므로

3)-② 사망일(DTH_DATE1 ~ DTH_DATE3)이 고용보험 상실일(OUT_DT)보다 과거 시점인 경우 상실일(OUT_DT) 값을 사망일로 변경
↳ 행정 처리 상의 문제로 보임.