Base Table 만들기

## 0. 라이브러리 설치 및 데이터 불러오기

# 필요 라이브러리 설치  
import numpy as np  
import pandas as pd

# 데이터 load  
data = pd.read\_csv('/Users/limdonggeon/Desktop/KPMG/Work/Law/data/main\_data.csv')

/var/folders/fs/zfypqyv96hs22x794hfx0ycm0000gn/T/ipykernel\_13528/235597709.py:2: DtypeWarning: Columns (2,4,5,8,13,16,17,18,19,24) have mixed types. Specify dtype option on import or set low\_memory=False.  
 data = pd.read\_csv('/Users/limdonggeon/Desktop/KPMG/Work/Law/data/main\_data.csv')

data.info()

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 861719 entries, 0 to 861718  
Data columns (total 25 columns):  
 # Column Non-Null Count Dtype   
--- ------ -------------- -----   
 0 소관부처명 861666 non-null object   
 1 법령명 861702 non-null object   
 2 법령구분 861704 non-null object   
 3 조번호 861129 non-null object   
 4 항번호 666590 non-null object   
 5 호번호 504904 non-null object   
 6 조문제목 805869 non-null object   
 7 조문 848856 non-null object   
 8 사무판단 767124 non-null object   
 9 사무판단근거 700871 non-null object   
 10 사무명 60113 non-null object   
 11 수행주체 60116 non-null object   
 12 사무유형 60071 non-null object   
 13 위임사무판단 761139 non-null object   
 14 위임근거규정 5311 non-null object   
 15 수임기관 4416 non-null object   
 16 특행기관 93089 non-null object   
 17 재위임사무판단 702502 non-null object   
 18 재위임근거규정 54 non-null object   
 19 재수임기관 13 non-null object   
 20 위탁사무판단 758703 non-null float64  
 21 위탁근거규정 3975 non-null object   
 22 수탁기관 3952 non-null object   
 23 사무유형(소분류) 60114 non-null object   
 24 기타 3 non-null object   
dtypes: float64(1), object(24)  
memory usage: 164.4+ MB

# 1. 데이터 전처리

## 1) 설명변수(소관부처명, 법령명, 조번호, 항번호, 호번호, 조문제목, 조문)가 모두 결측치인 행 삭제

def x\_null\_drop(df):   
 select\_column = ['소관부처명', '법령명', '조번호', '항번호', '호번호', '조문제목', '조문']  
 delete\_row\_idx = list(df[df[select\_column].isnull().all(axis = 1)].index)  
 delete\_row\_idx.sort(reverse = True)  
 for i in delete\_row\_idx:  
 df = df.drop([i],axis = 0)  
 return df

data = x\_null\_drop(data)

## 2) 소관부처명 결측치 처리

* 소관부처명 결측치: 45개
* 동일한 법령에 대해서는 동일한 소관부처를 가짐
* 이에, 다른 행 중 동일한 법령을 지닌 소관부처 파악 후 결측치 채워줌

def dep\_law\_preprocessing(df):  
 # department\_idx: '법령명'은 채워져있는데 '소관부처명'은 채워져있지 않은 행의 index  
 department\_idx = df[df['소관부처명'].isnull() & df['법령명'].notnull()].index  
 # department\_name\_list: '소관부처명'이 채워져야할 법령명  
 department\_name\_list = df[df['소관부처명'].isnull() & df['법령명'].notnull()]['법령명'].unique()  
  
 department\_dic = {}  
 department\_dic['건설산업기본법'] = '국토교통부'  
 department\_dic['보건범죄단속에관한특별조치법시행령'] = '보건복지부'  
 department\_dic['항로표지법'] = '해양수산부'  
 department\_dic['수산자원관리법'] = '해양수산부'  
 department\_dic['연안관리법'] = '해양수산부'  
 department\_dic['야생생물 보호 및 관리에 관한 법률'] = '환경부'  
  
 for i in range(len(department\_idx)):  
 for j in range(len(department\_name\_list)):  
 df.loc[department\_idx[i],'소관부처명'] = department\_dic[department\_name\_list[j]]  
  
 # '소관부처명', '법령명' 모두 채워져있지 않은 행  
   
 return df

data = dep\_law\_preprocessing(data)

## 3) 법령명 결측치 처리

* 조, 항, 조문 통해 법령명 찾아 삽입

def law\_name\_preprocessing(df):  
 idx = 14168  
 df.loc[idx,'소관부처명'] = '고용노동부'  
 df.loc[idx,'법령명'] = '근로자퇴직급여 보장법'  
 df.loc[idx,'법령구분'] = 1  
  
 idx = 198519  
 df.loc[idx,'소관부처명'] = '국토교통부'  
 df.loc[idx,'법령명'] = '택수운송사업의 발전에 관한 법률'  
 df.loc[idx,'법령구분'] = 1  
 df.loc[idx,'조번호'] = 11  
 df.loc[idx,'항번호'] = 1  
 df.loc[idx,'조문제목'] = '감차계획의 수립 및 시행 등'  
 idx = 686791  
 df.loc[idx,'소관부처명'] = '해양수산부'  
 df.loc[idx,'법령명'] = '수산업ㆍ어촌 공익기능 증진을 위한 직접지불제도 운영에 관한 법률'  
  
 idx = 708300  
 df.loc[idx,'소관부처명'] = '해양수산부'  
 df.loc[idx,'법령명'] = '해양공간계획 및 관리에 관한 법률'  
  
 idx = 708831  
 df.loc[idx,'소관부처명'] = '해양수산부'  
 df.loc[idx,'법령명'] = '해양폐기물 및 해양오염퇴적물 관리법'  
  
 idx = 766079  
 df.loc[idx,'소관부처명'] = '행정안전부'  
 df.loc[idx,'법령명'] = '새마을금고법'  
  
 idx = 859679  
 df.loc[idx,'소관부처명'] = '환경부'  
 df.loc[idx,'법령명'] = '미세먼지 저감 및 관리에 관한 특별법'  
  
 idx = 859692  
 df.loc[idx,'소관부처명'] = '환경부'  
 df.loc[idx,'법령명'] = '미세먼지 저감 및 관리에 관한 특별법'  
  
 idx = 859755  
 df.loc[idx,'소관부처명'] = '환경부'  
 df.loc[idx,'법령명'] = '미세먼지 저감 및 관리에 관한 특별법'  
   
 return df

data = law\_name\_preprocessing(data)

## 4) 법령구분 처리

* 법령구분 결측치 처리
* 법령구분 자료형 int로 통일

def law\_category\_preprocessing(df):  
 # '법령구분'이 결측치인 행들의 index  
 null\_idx = df[df['법령구분'].isnull()].index  
 # 국가법령정보센터 확인 결과, 결측치인 모든 행들은 법률, 즉 '1'에 해당  
 for i in null\_idx:  
 df.loc[i,'법령구분'] = 1  
   
 # 국가법령정보센터 확인 결과, '법령구분'이 공백으로 되어있는 행은 시행령, 즉 '2'에 해당  
 df.loc[df['법령구분']==' ', "법령구분"] = 2  
  
 # 법령구분 자료형 통일  
 df['법령구분'] = df['법령구분'].astype('int64')  
  
 return df

data = law\_category\_preprocessing(data)

## 5) 사무판단 처리

* ’ ’ -> nan, ‘0’ -> 0 , ‘1’ -> 1, ‘0 1’ -> 2 float 형태로 변환

def decision\_preprocessing(df):  
 # 표기방식 통일  
 idx\_nan = df[(df['사무판단'] == ' ')].index #idx\_nan: '사무판단'이 nan인 행의 index  
 for i in idx\_nan:  
 df.loc[i,'사무판단'] = np.nan  
 idx\_0 = df[(df['사무판단'] == '0')].index #idx\_0: '사무판단'이 '0'인 행의 index  
 for i in idx\_0:  
 df.loc[i,'사무판단'] = 0  
 idx\_1 = df[(df['사무판단'] == '1')].index #idx\_1: '사무판단'이 '1'인 행의 index  
 for i in idx\_1:  
 df.loc[i,'사무판단'] = 1  
 idx\_2 = df[(df['사무판단'] == '0 1')].index #idx\_2: '사무판단'이 '0 1'인 행의 index  
 for i in idx\_2:  
 df.loc[i,'사무판단'] = 2  
  
 # 오류 행 삭제  
 ## 경우1: 사무가 아님에도 사무 유형이 분류된 경우  
 delete\_0\_idx = list((df[(df['사무판단'] == 0) & (df['사무유형(소분류)'].notna())]).index)  
 df = df.drop(delete\_0\_idx, axis = 0)  
   
 ## 경우2: 사무임에도 사무 유형이 분류되지 않은 경우  
 delete\_1\_idx1 = list((df[(df['사무판단'] == 1) & (df['사무유형'].isnull())]).index)  
 df = df.drop(delete\_1\_idx1, axis = 0)  
   
 delete\_1\_idx2 = list((df[(df['사무판단'] == 1) & (df['사무유형(소분류)'].isnull())]).index)  
 df = df.drop(delete\_1\_idx2, axis = 0)  
  
 # 결측행 처리  
 ## 경우1: 사무 유형이 분류된 경우 => '1'로 채움  
 change\_1\_idx = df[(df['사무판단'].isnull()) & (df['사무유형'].notna()) & (df['사무유형(소분류)'].notna())].index  
 df.loc[change\_1\_idx, '사무판단'] = 1  
   
 # 경우2: 사무 유형이 분류되지 않은 경우 => '0'으로 채움  
 change\_0\_idx = df[(df['사무판단'].isnull()) & (df['사무유형'].isnull()) & (df['사무유형(소분류)'].isnull())].index  
 df.loc[change\_0\_idx,'사무판단'] = 0  
   
 # 자료형 통일  
 df['사무판단'] = df['사무판단'].astype('int64')  
  
 return df

data = decision\_preprocessing(data)

## 6) 소관부처명 공백 처리

def blank\_preprocessing(df):  
 df.loc[df['소관부처명']=="교육부,\n고용노동부", "소관부처명"] = '고용노동부,교육부'  
 df.loc[df['소관부처명']=="과학기술정보통신부, \n교육부", "소관부처명"] = '과학기술정보통신부,교육부'  
 df.loc[df['소관부처명']=="교육부,\n과학기술정보통신부", "소관부처명"] = '과학기술정보통신부,교육부'  
 return df

data = blank\_preprocessing(data)

## 7) 조문, 조문 제목 결측치 처리

* 조문, 조문 제목 null값이면 ’0’으로 채움

def law\_preprocessing(df):  
 df.loc[df['조문제목'].isna(), '조문제목'] = '0'  
 df.loc[df['조문'].isna(), '조문'] = '0'  
 return df

data = law\_preprocessing(data)