data preprocessing load map

Date	@2023년 4월 10일
Type	📌 과제
Status	In progress
Subject	캡스톤 디자인

사용한 데이터:

▼ 서울특별시 건축물대장 정보

- shape: (924620, 24)
- new shape: (924620, 14) → 현재 row는 아무 곳에도 안건드림!
- original columns:

  '관리_건축물대장_PK', '관리_상위_건축물대장_PK', '대장_구분_코드', '대장_종류_코드', '시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지', '특수지_명', '블록', '로트', '건물_명', '위반_건축물_여부', '대장_일련번호', '총괄표제부_일련번호', '표제부_일련번호', '전유부_일련번호', '새주소_도로_코드', '새주소_법정동_코드', '새주소_지상지하_코드', '새주소_본_번', '새주소_부_번', '작업_일자'

- new columns:

  '관리_건축물대장_PK', '관리_상위_건축물대장_PK', '대장_구분_코드', '대장_종류_코드', '시군구_코드', '법정동_코드', '번', '지', '건물_명', '위반_건축물_여부', '새주소_도로_코드', '새주소_지상지하_코드', '새주소_본_번', '새주소_부_번'

- 제거한 columns:

  1. 대지 구분 코드: 원래는 0이면 대지, 1이면 산, 2면 블록에 있는 건물임을 나타냄. 건물의 추천과정에서 건물이 위치한 대지가 무엇인지는 상관 없기 때문에 제거
  2. 특수지 명: 개발 지역에 대한 정보를 담고 있으나 구획이나 블록에대한 정보로 큰 의미 없다고 판단하여 제거
  3. 블록: 특정한 구역에 개발 계획상의 블록 번호를 적어놓은듯 보이나 값이 2064개를 제외하고는 NaN값이기 때문에 제거
  4. 로트: 위의 블록과 연관된 데이터. 블록보다 심함,, 962개 값 제외하고 NaN이라 제거.

5. 각종 일련번호: 테이블 간의 PK관계는 건축물 대장 PK로 연결되기 때문에 각 문서에 대한 일련번호는 필요 없음.
6. 작업일자: 그냥 이 데이터 언제 만든지에 대한 정보라 제거

- 남은 columns
  1. 대장 구분 코드: 일반 건축물과 집합 건축물을 구분하는 코드. 1: 일반, 2 : 집합
  2. 대장 종류 코드: 해당 건축물 대장과 관련있는 타 대장 정보를 담음.

     1: 총괄표제부, 2: 일반건축물, 3: 표제부, 4: 전유부
  3. 시군구, 법정동, 번, 지, 새주소 도로코드, 새주소 본.부번 → 건물의 위치를 파악하기 위해 남김
  4. 위반 건축물 여부: 입점하려는 건물이 위반 건축물일 경우 행정상의 불이익이 크기 때문에 추천에서 배제하기 위해 남김 → 추후 Full table에서 위반 건축물의 경우 제거할 예정

▼ 서울시 건축물대장 표제부 정보

- shape: (606665, 30)
- new shape: (, 15) row 수는 현재 변화 없음
- original columns:

  '건축물대장 관리번호', '주용도 구분', '기타 용도', '세대수', '가구수', '구조 구분', '기타 구조', '지붕 구분', '기타 지붕', '건축 면적', '연 면적', '용적율 산정 연면적', '높이', '대지 면적', '건폐율', '용적율', '지상층수', '지하층수', '승용 승강기수', '비상용 승강기수', '부속 건축물수', '부속 건축물 면적', '옥내 자주식 대수', '옥내 자주식 면적', '옥외 자주식 대수', '옥외 자주식 면적', '옥내 기계식 대수', '옥내 기계식 면적', '옥외 기계식 대수', '옥외 기계식 면적'
- new columns:

  '건축물대장 관리번호', '주용도 구분', '기타 용도', '세대수', '가구수', '건폐율', '용적율', '지상층수', '지하층수', '승용 승강기수', '비상용 승강기수', '옥내 자주식 대수', '옥외 자주식 대수', '옥내 기계식 대수', '옥외 기계식 대수'
- 제거한 columns:
  1. 구조 구분, 기타 구조, 지붕 구분, 기타 지붕: 해당 건물이 지어진 방법에 대한 내용으로 추천과는 연관성이 없어서 제거

2. 연면적, 용적률 산정 연면적: 건물 바닥 넓이의 총 합. 용적률과 큰 연관이 있기 때문에 제거
3. 높이: 건축물의 높이는 추천과 큰 상관이 없어서 제거
4. 부속 건축물수, 면적: 부속 건축물은 입점하는 상가들과는 관련이 없다고 판단하여 제거 → 추후 더 면밀히 조사해본 후 수정 필요시 수정하겠음
5. 옥내외 자주,기계식 면적: 주차장의 면적 정보. 주차 대수가 더욱 중요한 요소라고 판단하여 제거

- 남은 columns:
  1. 주용도 구분: 해당 건축물이 위치하는 부지에 대한 행정정보(근린생활시설 등등)
  2. 기타 용도: 해당 건축물이 사용되는 용도
  3. 건폐율: 건설 부지에서 건물이 차지하는 땅의 비율
  4. 용적률: 전체 대지면적에 대한 건물 연면적의 비율

▼ 서울시 건축물대장 총괄표제부 정보

- shape: (19840, 21)
- new shape: (, 10)
- original columns:

  '건축물대장 관립번호', '대지 면적', '건축 면적', '건폐율', '연면적', '용적율 산정 연면적', '용적율', '주건축물수', '부속 건축물수', '부속 건축물 면적', '세대수', '가구수', '총주차수', '옥내 자주식 대수', '옥내 자주식 면적', '옥외 자주식 대수', '옥외 자주식 면적', '옥내 기계식 대수', '옥내 기계식 면적', '옥외 기계식 대수', '옥외 기계식 면적'

- new columns:

  '건축물대장 관립번호', '건폐율', '용적율', '주건축물수', '세대수', '가구수', '옥내 자주식 대수', '옥외 자주식 대수', '옥내 기계식 대수', '옥외 기계식 대수'

- 제거한 columns:
  1. 대지, 건축면적, 용적율산정 연면적: 건폐율과 용적률로 유추 가능
  2. 나머지는 표제부랑 같은 이유!
- 남긴 columns:
  1. 세대수, 가구수: 사실 지울까 말까 고민 엄청 하다가 애매해서 일단 남겨뒀다,, 상가 건물 위에 있는 가정집이라면 분명 긍정적이던 부정적이던 영향을 끼칠

수 도 있다고 판단했음. 추후에 상관관계 등 통계적 기법으로 확인하고 제거 여부 다시 선택하도록 하겠음.

▼ 서울특별시_건축물대장_층별개요정보(2019년)
→ 19년 이후 데이터도 있지만.. 사이즈가 너무 커서 안열림.. 나중에 해결해보겠습니다!

- shape: (845040, 13)
- new shape: (, 10)
- original columns:

  '관리_층별_개요_PK', '관리_건축물대장_PK', '관리_주_건축물대장_PK', '주_부속_구분_코드', '주_부속_일련번호', '층_구분_코드', '층_번호', '층_번호_명', '구조_코드', '기타_구조', '주_용도_코드', '기타_용도', '면적'

- new columns:

  '관리_층별_개요_PK', '관리_건축물대장_PK', '관리_주_건축물대장_PK', '주_부속_구분_코드', '층_구분_코드', '층_번호', '층_번호_명', '주_용도_코드', '기타_용도', '면적'

▼ 서울시 건축물대장 지역지구구역 정보

- shape: (775035, 6)
- new shape: (, 5)
- original columns:

  '관리_지역지구구역_pk', '관리_건축물대장_pk', '지역지구구역_구분_코드', '지역지구구역_코드', '대표_여부', '기타_지역지구구역'

- new columns:

  '관리_건축물대장_pk', '지역지구구역_구분_코드', '지역지구구역_코드', '대표_여부', '기타_지역지구구역’

▼ 서울시 건축물대장 전유부 정보

- shape = new shape: (3626813, 4)
- original columns = new columns:

  '관리_건축물대장_PK', '동명칭', '호_명', '층_구분_코드’

- 전유부는 column의 개수가 적고 전부 유의미하다고 판단하여 일단 건들지 않은 상태

▼ 서울시 건축물대장 전유공유면적표 정보

- shape: (512878, 12)
- new shape: (, 9)
- original columns:

  '관리_전유_공용_면적_pk', '호별명세_pk', '평형_구분_명', '전유_공용_구분_코드', '주_부속_구분_코드', '층_구분_코드', '층_번호', '구조_코드', '주_용도_코드', '기타_용도', '면적', '작업_일자'

- new columns

  '관리_전유_공용_면적_pk', '호별명세_pk', '전유_공용_구분_코드', '주_부속_구분_코드', '층_구분_코드', '층_번호', '주_용도_코드', '기타_용도', '면적'

- 제거한 columns:
  1. 평형구분 명: 84B처럼 각 건물에서 평형을 구분하는 코드. 면적과 중복이기 때문에 제거
  2. 구조코드: 건물을 지을 때 건물을 지은 방식에 대한 정보
- 남긴 columns:
  1. 호별 명세: 있어야 하는건지 확실하지 않지만 없애도 되는지 확실하지 않아서 남겨둠. 추후에 재조정 예정