4/30 건축물 데이터 전처리 상태

진행도: 95% (6 / 7)

1. 건축물대장 - 완료

2. 표제부 - 완료

3. 총괄표제부 - 완료

4. 층별개요 – 완료

5. 지역지구구역 – 진행중(90%)

6. 전유부 - 완료

7. 전유공유면적표 – 완료

\*\*통합 – 진행중(40%)

이번 주는 크게 바뀐게 없어서 일단 4월 간 진행된 내용에 대한 요약을 하겠습니다.

<세부사항>

**건축물대장**

* Shape: (948397, 24) 🡪 (948397, 9)
* 건축물의 주소를 가진 테이블 🡪 주소와 다른 대장과의 관련이 있는 칼럼 제외 제거
* 주소를 도로명 주소가 존재하는 경우 사용하고 도로명이 없는 경우 지번 주소 사용

**표제부**

* Shape: (606892, 30) 🡪 (123535, 9)
* 건물의 골조와 용적률, 승강기, 주차 대수 등의 정보를 담은 테이블
* 건물의 골조와 같은 내용은 우리가 알고자 하는 내용 아니니 제거
* 상가는 근린생활시설에만 입점할 수 있기 때문에 근린생활시설만 남김
* 승강기 수와 주차 대수가 여러 칼럼으로 나눠져 있음 🡪 통합

**총괄표제부**

* Shape: (19840, 21) 🡪 (19840, 7)
* 표제부와 거의 동일한 내용을 담은 테이블.
* 차이점은 한 필지에 건물이 여러 개 있는 경우 총괄 표제부가 발급.
* 표제부와 같은 작업 실행

**층별개요정보**

* Shape: (845040, 13) 🡪 (170016, 10)
* 건축물 층별 구조, 용도, 면적 등을 담은 테이블
* 건물의 구조적 특성은 제거.
* 주용도코드가 5자리의 실수이거나 Z로 시작하는 총 5자리 코드
* 5자리 실수는 1000단위로 대분류, Z로 시작하는 코드는 Z이하 100단위로 대분류
* 코드 분류 해서 근린생활시설만 남김

**지역지구구역**

* Shape: (775537, 6) 🡪 (775537, 5)
* 해당 건축물의 위치가 어떤 구역인지에 대한 정보를 담은 테이블
* 건축물 대장에서 제일가는 문제아
* 코드 질의해놓고 답변 기다리는 중
* 이것만 하면 다 끝나는데,,,

**전유부**

* Shape: (3631394, 4)🡪 (3631394, 4)
* 각 호수에 대한 정보를 담은 테이블
* 전유공유면적표와 Join해서 사용

**전유공유면적표**

* Shape: (521029, 12) 🡪 (76491, 9)
* 각 호수에 대한 자세한 정보를 담은 테이블
* 전유부와 Join해서 사용
* 호수 별 용도를 알 수 있음. 층별개요와 마찬가지로 근린생활시설만 남김.

**통합**

* 현재 있는 데이터끼리 묶어보는 중.
* Inner Join 시행한다면 3개의 큰 테이블로 나올 예정
* Outer Join 시행하면 1개의 큰 테이블로 나옴. 그러나 옆으로 커짐. 그리고 대장 별 칼럼이 달라서 대장 별 One-Hot Encoding 수행 필요.
* 위도와 경도 값을 넣는 과정에서 좀 시간이 오래 걸려서 코드 최적화 해서 해결해볼 예정. 정 안되면 며칠동안 켜두죠 뭐…ㅎㅎㅎ

**Clustering**

* 일단 저번에 해놨던 코드 깃에 올려뒀음.
* 데이터 통합 해보면서 상황에 맞게 바꿀 예정
* Clustering 추가로 조사 해보는 중.
* 많은 클러스터링 기법이 있는데 일단 Simulation은 K-means로 일단 하고 실제 서비스 모델은 클러스터링 해서 나온 데이터가 Regression 통과 했을 때 성능 가장 좋은 걸로 쓰는게 가장 좋지 않을까 하는게 제 생각입니당.