데이터 전처리 과정

해야 하는 것 : 오류 사항 점검 및 조치 / 데이터 구조 및 특성 변경

ex) 중복값 제거, 이상치 및 결측치 보정, 데이터 연계/통합, 데이터 구조 변경

우리가 조사했던 데이터들 :

사이버범죄율 :

- 경찰청 공공데이터 개방 → 2021 시도청별 사이버범죄 발생 및 검거 데이터
- 전자기기 보급률
- 온라인 상거래 이용률
- 디지털 교육 : 공공데이터 - 지역별 디지털 배움터 이용자 수

고령화

- 주민등록 인구통계 → 행정구역별 평균 연령 데이터 → 고령화 정도 수치화
- 독거노인 가구비율 : KOSIS 통계청 e-지방지표 독거노인가구비율, 공공데이터포털
- 평균 학력 : 교육통계서비스 KESS

데이터	출처	비고
사이버범죄율 관련 데이터		
시도청별 사이버범죄 발생 및 검거	경찰청 공공데이터 개방	2021년 데이터
전자기기 보급률	미표기	추가 출처 확인 필요
온라인 상거래 이용률	미표기	추가 출처 확인 필요
지역별 디지털 배움터 이용자 수	공공데이터포털	디지털 교육 관련 지표
고령화 관련 데이터		
행정구역별 평균 연령	주민등록 인구통계	고령화 정도 수치화
독거노인 가구비율	KOSIS 통계청 e-지방지표, 공공데이터포털	-

데이터	출처	비고
평균 학력	교육통계서비스(KESS)	-

헐 공공데이터포털이 그 기관?에 불나서 잠시 서비스가 중단되었음.

이전에 드라이브에 올려뒀던 데이터 : 주민등록인구현황(2025년 7월)(고령인구현황)

지역별 평생교육기관 개황

지역별 디지털배움터 이용자 수

행정구역별 위경도 좌표

내 컴퓨터에 있었던 데이터 : (위 데이터 제외) 보건복지부 독거노인 수 연령별, 시도별

컴퓨터 내 폴더에 정리함

그러면 지금 나한테 없는 게 사이버범죄 발생 및 검거 데이터, 전자기기 보급률, 온라인 상거래 이용률, 평균 학력?

그냥 인터넷에서 찾아보기로 함

전자기기 보급률 :

악 다 공공데이터포털에 연결되어있어서 힘듦... 파일이 올 때까지 있는 걸 통합하기로.

→ 컴퓨터 내에서 새로운 가상환경을 만들고, 라이브러리를 설치함. 이 방식은 다음 글을 참고 : (개발환경 만드는 과정 자체는 동일, 라이브러리만 다름)

사용한 라이브러리 종류는 requirements 내에 들어있기 때문에 파악 가능

lidar 3d point cloud 만들기 과정

토큰명 : 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

지금 '고령화' 관련 데이터들만 있는 것 같다. 공통점은 모두 '지역'에 따라 엮을 수 있다는 것.
근데 원래 우리의 생각은 '시간에 따라' 고령화가 심화됨에 따라 사이버 범죄율이 어떻게 변화했는가, 였는데

지역에 따라 엮는 것을 시도함

하나의 큰 표로 만드는 코드를 chatgpt와 만들기를 시도하다가 claude와 만듦...

→ n번째 시도에서 어느정도 성공

github에서 preprocess5.py가 해당 코드임

다른 자료들은 아직 반영 x.