

# 서울시 매물 데이터

1. 목적 및 배경
2. 데이터 개요
  - 2.1 데이터 소스
  - 2.2 주요 컬럼
  - 2.3 대상 데이터
3. 전처리 목표
4. 전처리 단계
  - 4.1 데이터 수집
  - 4.2 데이터 필터링
  - 4.3 데이터 정제
  - 4.4 데이터 저장
5. 결과 및 요약
6. 향후 작업
7. 부록

## 1. 목적 및 배경 [↗](#)

서울시 부동산 매물 데이터를 분석 가능하도록 정제하여 SQLite 데이터베이스에 저장.

## 2. 데이터 개요 [↗](#)

### 2.1 데이터 소스 [↗](#)

- <<https://www.rter2.com>> API

### 2.2 주요 컬럼 [↗](#)

- **property\_id**: 매물의 고유 ID
- **price1**: 매매가 또는 보증금 (임대 매물의 경우)
- **price2**: 월세 (임대 매물의 경우)
- **useApproveDay**: 건물 준공일
- **facilities**: 매물의 편의 시설 정보 (예: 주차, 엘리베이터)

### 2.3 대상 데이터 [↗](#)

서울시 25개 구의 매물 데이터를 수집하여 매매가, 임대가, 준공일, 편의 시설 등을 기반으로 지역별 매물 트렌드를 분석

## 3. 전처리 목표 [↗](#)

- 결측값 처리
  - 일부 필드에 누락된 값(None)이 있는 경우 기본값 설정
- 데이터 변환
  - API 응답 데이터를 데이터베이스 형식에 맞게 변환
  - 날짜 포맷 변환 (useApproveDay)
- 중복 제거

- 중복된 매물 데이터 제거
- 데이터 검증
  - API에서 가져온 데이터를 enum 타입과 매핑하여 유효성 검사
- 데이터 저장
  - SQLite 데이터베이스에 적재 (충돌 시 업데이트)

## 4. 전처리 단계 [↗](#)

### 4.1 데이터 수집 [↗](#)

- 작업
  - process\_all\_districts() 함수에서 서울시의 각 구에 대해 매물 데이터 수집
  - 비동기 요청(aiohttp)을 사용하여 API 데이터 수집

### 4.2 데이터 필터링 [↗](#)

- 작업
  - 필수 필드가 누락된 매물 데이터 필터링
  - 기존 매물 정보와 비교하여 중복 매물 업데이트

### 4.3 데이터 정제 [↗](#)

- 작업
  - 시설 및 사진 데이터를 문자열로 변환
  - 필드 값 유효성 검사 및 변환 (enum 매핑)

### 4.4 데이터 저장 [↗](#)

- 작업
  - 정제된 데이터를 SQLite 데이터베이스 테이블에 저장

## 5. 결과 및 요약 [↗](#)

전처리 작업 완료 후 데이터의 상태 요약

- 처리 전 데이터
  - API에서 수집한 JSON 데이터 (총 100,000개 매물)
- 처리 후 데이터:
  - 중복 제거 및 필터링 후 약 95,000개
  - SQLite 데이터베이스 테이블
    - property\_locations: 매물 위치 정보
    - property\_info: 매물 기본 정보
    - sales: 매매 정보
    - rentals: 임대 정보

## 6. 향후 작업 [↗](#)

- 추가적으로 필요한 전처리 작업 또는 개선 사항
  - 매물 사진의 실제 유효성 검증
  - 네트워크 속도 최적화를 위한 요청 스케줄링

- 고정 필드 값(enum) 업데이트 시 유연성 개선

## 7. 부록 [↗](#)

- PropertyInfo 테이블

컬럼명	설명
property_id	매물 ID (고유값)
property_type	매물 타입 (예: 아파트, 빌라 등)
building_name	건물 이름
address	매물 주소

- PropertyLocations 테이블

컬럼명	설명
property_id	매물 ID (외래키)
administrative_dong	행정동 정보
latitude	위도
longitude	경도