**지하철 역별 특징 데이터 정제 과정**

|  |  |  |
| --- | --- | --- |
| **STEP** | **설명** | **주요 처리 내용** |
| **STEP 1** | 기본 정보 추출 | - 서울 교통빅데이터플랫폼에서 지하철 역 정보 수집  - 해당 데이터를 기본 데이터프레임으로 사용 |
| **STEP 2** | 호선 ID → 호선명 매핑 | - 동일 플랫폼에서 호선 정보 수집  - ‘지하철 호선 ID’를 기준으로 ‘호선명’ 열 추가 |
| **STEP 3** | 환승역 정보 추가 및 누락값 보완 | - 공공데이터포털에서 환승역 정보 확보  - ‘역명’ 기준으로 ‘환승역명’ 매핑  - 누락 시 ‘지하철역ID’ 중복 확인  - 검색엔진 활용하여 보완 |
| **STEP 4** | 파생변수 생성 | - 주소 → 시/군/구/동으로 분리  - 이전/다음 역명은 지하철역ID 기준으로 매핑 |

**STEP 1) 기본 정보 데이터 추출ㄴ**

* **이용 사이트 : 서울 교통빅데이터플랫폼 TData**
* **데이터 원본 url :** [**https://t-data.seoul.go.kr/dataprovide/trafficdataviewfile.do?data\_id=10054**](https://t-data.seoul.go.kr/dataprovide/trafficdataviewfile.do?data_id=10054)
* **해당 데이터를 기본 데이터프레임으로 사용**

**STEP 2) 지하철 호선 ID 와 실제 호선명 매핑**

* **이용 사이트 : 서울 교통빅데이터플랫폼 TData**
* **데이터 원본 url :** <https://t-data.seoul.go.kr/dataprovide/trafficdataviewfile.do?data_id=10044>
* **해당 데이터의 ‘지하철호선명’ 열을 기본 데이터프레임에 추가 (기준 : 지하철 호선 ID)**

**STEP 3) 지하철 호선 ID 와 실제 호선명 매핑**

* **이용 사이트 : 공공데이터포털**
* **데이터 원본 url :**
  + <https://data.seoul.go.kr/dataList/OA-12033/S/1/datasetView.do>
  + https://www.data.go.kr/data/15098252/fileData.do
* **해당 데이터의 ‘환승역명’ 열을 기본 데이터프레임에 추가 (기준 : 역명)**
* **해당 제공데이터로 확인하지 못한 경우, ‘지하철역ID’열(고유값)을 기준으로 중복값 체킹**
* **이후, 동일역명 다른호선 혹은 신설 호선 등의 누락 데이터는 검색엔진 이용하여 매핑 진행**
* **1차 매핑 이후, 2차 확인으로 검색엔진 활용 2차 매핑 후 전체 체킹 완료**

**STEP 4) 파생변수 생성**

**① 시/군/구/동 등 : ‘기본주소’ 열 공백 기준 분리 후 각각 파생변수 생성**

**② 이전지하철역명 / 다음지하철역명 : 고유값인 지하철역ID 를 기준으로 매핑진행**