데이터랭글링 평가 과제

2018112462 유근태

1. XML 파일 수집 및 설명

-1. 데이터 출처: 공공데이터포털

-2. 데이터 링크: https://www.data.go.kr/data/15052782/fileData.do

-3. 데이터 설명: 동북아역사재단에서 제공하는 고구려 성곽에 대한 정보가 담겨 있는 데이터입니다. 이 데이터셋은 고구려성 개관, 환인, 집안 지역, 소자하 일대, 요하 중류 일대, 동요하~송화강 일대, 압록강 일대, 초하, 애하~대양하 일대, 요동반도 연안, 흔하 일대 등의 고구려 성의 위치와 특징에 대한 설명을 포함하고 있습니다.

2. XML 파싱

먼저 루트 요소에서 'level1' 태그를 가진 모든 요소를 찾고, 찾은 데이터를 level1_list에 저장합니다. 이후 각 'level1' 요소에서 'front/biblioData/title/mainTitle'과 'front/biblioData/title/alternative' 태그를 가진 요소들을 찾습니다. 그런 다음, 각 'level1' 요소 내에서 'level2' 요소와 그 하위의 'front/biblioData/title/mainTitle' 요소를 찾습니다. 두 리스트를 동시에 순회하면서 각 'level2' 요소의 'id' 속성과 'level2_maintitle' 요소의 'volume' 속성을 추출합니다. 이렇게 파싱한 3개의 텍스트 값과 2개의 속성값을 리스트로 묶어 list1에 추가합니다. list1을 pandas의 DataFrame 함수를 활용하여 'maintitle', 'alternative', 'level2_maintitle', 'volume', 'id' 컬럼으로 구성된 데이터프레임. df로 구성하여 출력을 진행했습니다.

3. 누락데이터 평가 및 구현

우선 isnull() 함수를 활용해서 누락 데이터이면 True 반환하고, 유효한 데이터이면 False를 반환했습니다. 그 결과, alternative 컬럼에서 index 9, 10에서 True가 반환된 것을 확인할 수 있습니다. 다음으로 sum() 함수를 통해 컬럼별 null 값의 개수를 구해봤는데, alternative 열에서 누락 데이터가 2개인 것을 확인할 수 있었습니다. 그래서 이를 null 값 percent로 산출해본 결과, 전체 42개의 데이터 중 2개의 데이터가 결측값이므로 alternative 열은 2 / 42 * 100 = 4.761905%가 이상치로 구성되어 있다는 결과를 도출할 수 있었습니다.