**데이터 수집 미니 프로젝트**

**산업구조 변화대응특화훈련**

**조해민**

**#01. 프로젝트 개요**

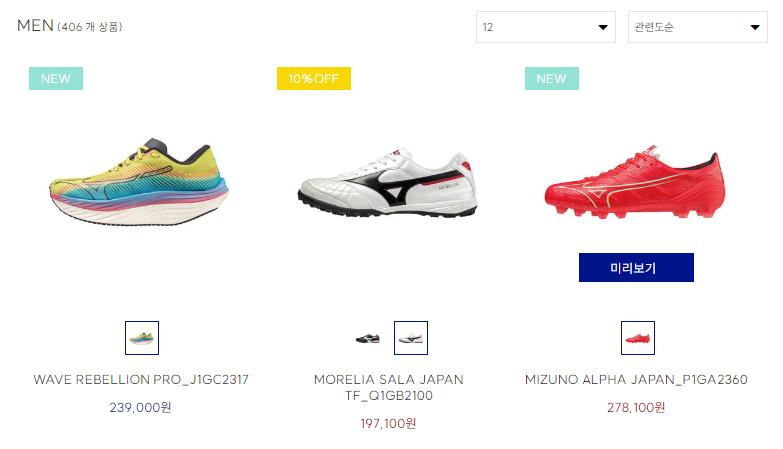
* **데이터 수집 주제**

**: 남성 스포츠 용품이라는 관심사를 정하여 인지도가 있으면서 가성비가 좋은 브랜드인 미즈노와 아디다스의 스포츠 용품에 대한 정보를 수집하고자 했다.**

**#02. 작업 개요**

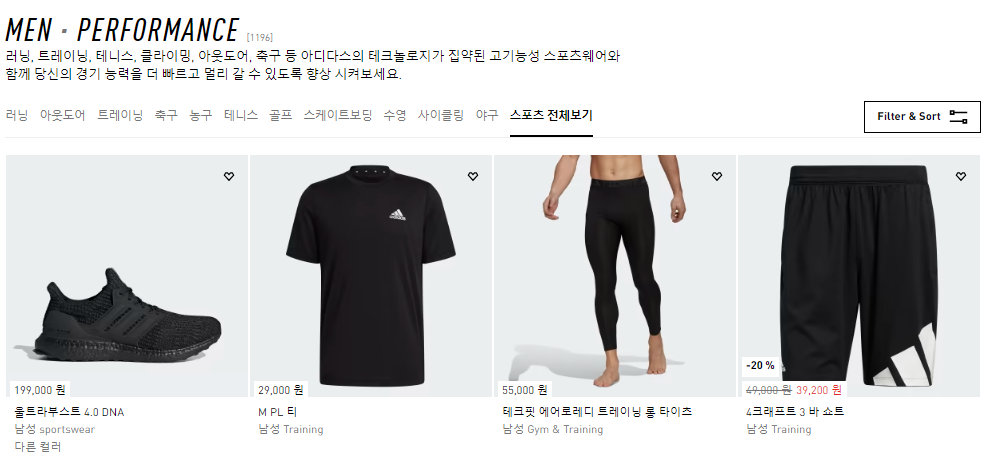
1. **데이터 수집 웹사이트 결정**
2. **미즈노 남성 스포츠 용품 웹사이트**

**:** [**https://kor.mizuno.com/kr-kr/044/?start=0&sz=12**](https://kor.mizuno.com/kr-kr/044/?start=0&sz=12)



1. **아디다스 남성 스포츠 용품 웹사이트**

**:** [**https://www.adidas.co.kr/men-performance?start=0**](https://www.adidas.co.kr/men-performance?start=0)



1. **적용 기술 종류**

* **데이터 수집: Selenium 사용 크롤링**
* **데이터 저장: docker를 통해 MySQL에 수집한 데이터를 저장**

1. **수집 계획**
2. **Selenium 패키지를 사용하여 웹 페이지 코드 수집**

**: chrome driver를 통해 Chrome 브라우저 가동**

**: 전체 남성 스포츠 용품에 대한 정보를 수집하기 위해 각 페이지 별로 상품 정보 수집 예정 -> 페이지 별 상품에 접근하기 위해 URL을 분석하여 변수 값 취득 필요**

**: 상세한 상품정보를 수집하기 위해 상품 리스트 웹페이지에서 상품 상세보기 링크를 통해 각 상품의 상세보기 페이지에서 상품에 대한 상세정보 수집 예정.**

1. **응답결과로부터 내용 추출하기**
2. **상품 웹사이트**

**: 상품 리스트가 나온 페이지에서 HTML 코드를 분석하여 상품 이름, 이미지 링크, 상세보기 링크 추출**

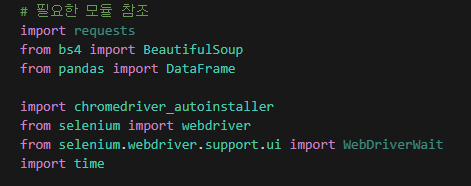
**: 추출된 상품 상세보기 링크와 chrome driver를 통해 Chrome 브라우저 재가동하여 상품 상세보기 페이지에 접속**

**: 접속된 상품 상세보기 페이지에서 상품의 정가, 최종 금액, 리뷰 수 정보를 추출**

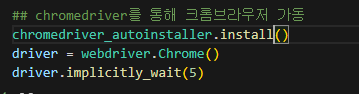
1. **xlsx파일 형태로 저장하고 추출된 데이터를 병합 후 DB에 저장**

**#03. 수집 프로그램 구현**

1. **미즈노 웹사이트**
2. **필요한 모듈 참조**



1. **Chrome driver을 통해 크롬브라우저 가동**

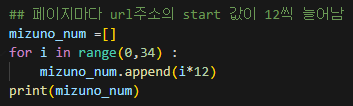


1. **상품 웹 페이지 주소를 페이지 별로 분석**

**: 1페이지 주소:** [**https://kor.mizuno.com/kr-kr/044/?start=0&sz=12**](https://kor.mizuno.com/kr-kr/044/?start=0&sz=12)

**: 2페이지 주소:https://kor.mizuno.com/kr-kr/044/?start=12&sz=12**

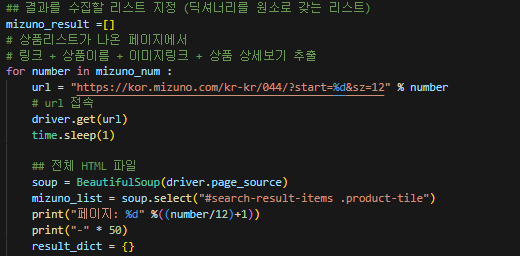
* + - * + **start=숫자가 나오는데 12씩 증가함.**



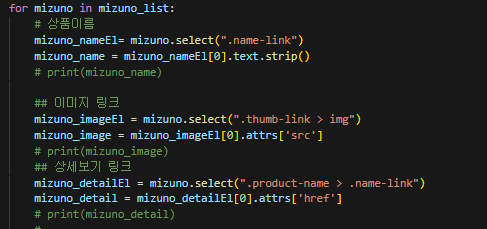
**(34페이지까지 있으나 시간관계상 5페이지 까지 진행)**

1. **응답 결과로부터 내용 추출하기**

* **(4) -1 상품 리스트 각 페이지 별로 접근 + HTML 소스코드 select**



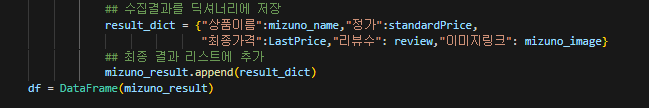
* **(4) - 2 상품 리스트 웹 페이지 에서 상품정보 수집**

****

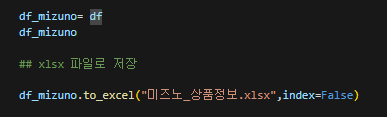
* **(4) - 3 상품 상세정보 링크 접속하여 상세정보 수집**

****

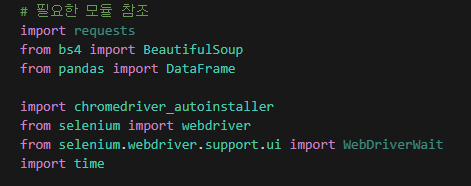
* **(4) – 4 추출 결과를 결과 리스트에 넣고 데이터프레임 형태로 저장**

****

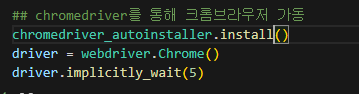
* **(4) – 5 데이터프레임을 xlsx 파일 형태로 저장**



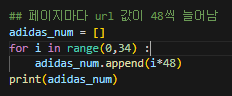
1. **아디다스 웹사이트**
2. **필요한 모듈 참조**



1. **Chrome driver을 통해 크롬브라우저 가동**

****

1. **상품 웹 페이지 주소를 페이지 별로 분석**



**(25페이지까지 있으나 시간관계상 5페이지까지 진행)**

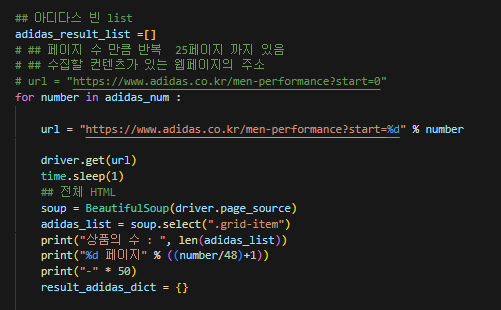
**: 1페이지 주소:** **https://www.adidas.co.kr/men-performance?start=0**

**: 2페이지 주소:** [**https://www.adidas.co.kr/men-performance?start=48**](https://www.adidas.co.kr/men-performance?start=48)

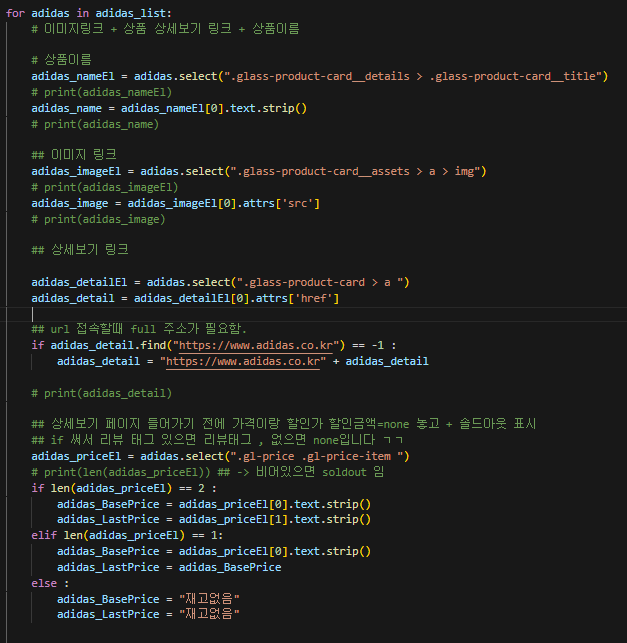
* + - * + **start=숫자가 나오는데 48씩 증가함.**

1. **응답 결과로부터 내용 추출하기**

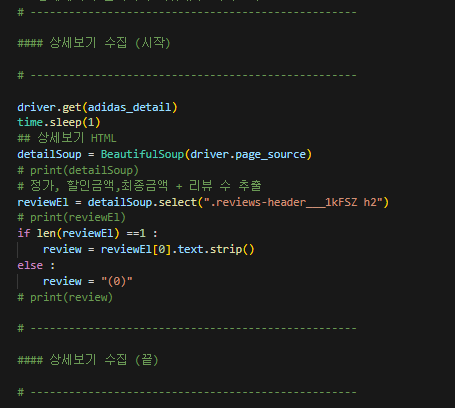
* **(4) -1 상품 리스트 각 페이지 별로 접근 + HTML 소스코드 selcet**

****

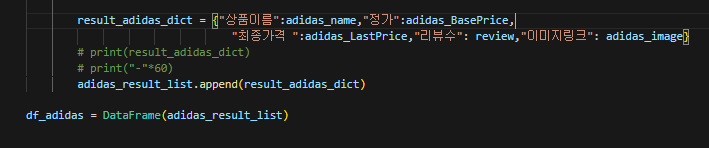
* **(4) – 2 상품 리스트 웹 페이지 에서 상품정보 수집**



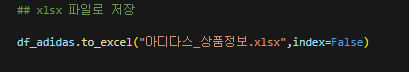
* **(4) - 3 상품 상세정보 링크 접속하여 상세정보 수집**

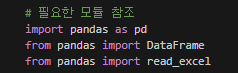
****

* **(4) – 4 추출 결과를 결과 리스트에 넣고 데이터프레임 형태로 저장**



* **(4) – 5 xlsx파일 형태로 저장**

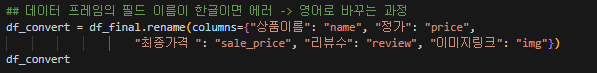
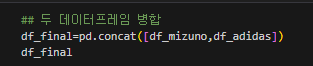


1. **데이터 병합**
2. **각 Excel 파일 불러오기**

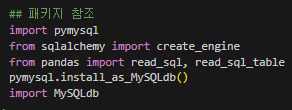


**C:\Users\user\AppData\Local\Microsoft\Windows\INetCache\Content.Word\아디다스 불러오기.png**

1. **두 데이터프레임 병합**



1. **MYSQL에 접속 및 저장**



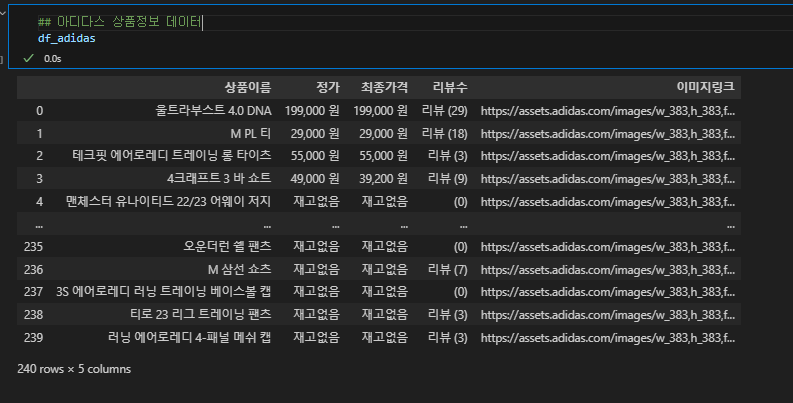
****

**#04. 수집 결과물**

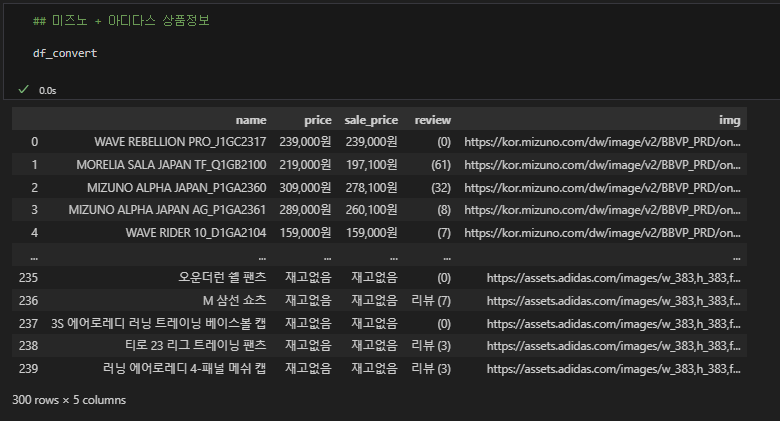
1. **미즈노 상품정보 데이터 …**



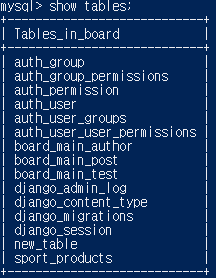
1. **아디다스 상품정보 데이터**



1. **병합된 데이터**



1. **데이터가 DB에 저장이 되었는지 확인**
2. **DB에서 테이블 삽입 확인**

****

1. **DB에서 테이블 조회**

