

데이터 분석 체크리스트

1. 주제 선정

- 핵심 키워드가 무엇이며, 선정한 이유가 무엇인가요?
- 키워드를 통해 어떤 메시지를 전달하고 싶나요?
- 기대하는 결말이 있나요?

2. 데이터 수집

1. 파일 내려받기

- 파일을 다운로드하는 방법으로 데이터를 가져올 때 단점이 무엇일까요?

2. OPEN API

- 요청 주소가 무엇인가요?
- 요청 파라미터가 무엇인가요?
- 출력 파라미터가 무엇인가요?
- 출력 결과에서 내가 원하는 데이터는 어디에 있나요?

3. 정적 크롤링(BeautifulSoup)

- 어떤 데이터를 수집했나요?
- 기능마다 함수로 만들 수 있나요?
- (심화) 기능을 모아 클래스로 만들 수 있나요?

4. 동적 크롤링(Selenium)

- 어떤 데이터를 수집했나요?
- 기능마다 함수로 만들 수 있나요?
- (심화) 기능을 모아 클래스로 만들 수 있나요?

3. 데이터 구조 이해 (EDA)

1) 데이터 불러오기

- 다양한 데이터를 불러올 수 있나요?(csv, tsv, xlsx)
- FileNotFoundError 를 만나면 어떻게 대처해야 하나요?
- ModuleNotFoundError 를 만나면 어떻게 대처해야 하나요?
- UnicodeDecodeError 를 만나면 어떻게 대처해야 하나요?

2) 데이터 정보 파악하기

- 데이터의 행, 열 개수를 알고 싶을 때는 어떻게 하나요?
- 데이터의 열과 dtype이 일치하지 않을 때에는 어떻게 해야 하나요?
- 시계열 데이터를 만났을 때에는 어떻게 해야 하나요?
- 범주형 데이터를 만났을 때 어떤 종류들이 있는지 알고 싶을 땐 어떻게 해야 하나요?

3) 결측치 처리하기

- 결측치 개수를 알고 싶을 때에는 어떻게 하나요?
- 결측치 처리하는 방법은 무엇이 있나요?

4) 데이터 요약하기

- 데이터의 요약통계량을 한 번에 보고 싶을 때에는 어떻게 하나요?
- 그룹별로 해당 열의 요약통계량을 출력할 수 있나요?
- 인덱스와 컬럼, 값을 새롭게 설정해서 새로운 데이터프레임을 만들 수 있나요?

5) 데이터 시각화하기

- 데이터의 유형은 어떤 것들이 있나요?(연속형, 수치형, 범주형, 순서형, 시계열 등)
- 데이터의 유형에 따라 그릴 수 있는 시각화 그래프는 무엇인가요?
- 시각화 그래프의 title, 크기 등을 조절할 수 있나요?
- 컬럼 간의 관계를 확인하기 위해서는 어떤 시각화를 해야 하나요?
- 시각화하면서 마주했던 어려움은 무엇이었나요?

4. 데이터 검증(CDA)

- 왜 데이터 검증이 필요한가요?
- 가설검정결과를 어떻게 해석해야 할까요?(통계량, 유의확률)
- 정규성 검정을 하는 방법
- 독립된 두 집단의 평균을 검정하는 방법
- 한 집단의 전과 후 차이를 검정하는 방법
- 독립된 세 집단 이상의 평균을 검정하는 방법
- 범주형 데이터 두 집단의 독립성을 검정하는 방법

5. 인사이트 도출

- 주제를 선정한 이유와 연관지어서 결론을 도출했나요?
- 주제 선정 시 기대했던 결론과 비슷한가요?
- 어떤 서비스 제안을 할 수 있을까요?
- 서비스에 인공지능을 넣는다면 어떻게 할 수 있을까요?

기타

- uv를 통해 새 프로젝트를 만들면 어떤 파일이 생기나요?
- 새로운 가상환경을 만들기 위해서는 어떻게 해야 하나요?
- pyproject.toml과 uv.lock이 있을 때 다시 가상환경을 만들어서 프로젝트에서 설치한 라이브러리를 한번에 설치하려면 어떻게 해야 하나요?
- 가상환경 내에서 라이브러리의 충돌이 있었습니다. 어떻게 해결해야 할까요?