기초 데이터 분석 및 실습 세션 프로젝트 안내

1. 프로젝트 목표

세션 기간 동안 배운 내용을 종합하여 데이터 수집부터 해석까지의 데이터 분석 전단계를 체계적으로 수행하는 것을 경험하기

2. 프로젝트 포함 항목

2.1. 필수 항목

2.1.1. 데이터 준비: 데이터 크롤링 및 소개

데이터 크롤링 및 소개

- 윤리 및 사회적 규범을 준수하는 크롤링 코드 또는 API 활용 코드 작성
- 크롤링한 데이터에 대한 간단한 설명 (예: 기상청 제공 2024년도 월별 및 지역별 강수량, 네이버 기사 등)
- 다음 형태의 데이터를 열(column)로 포함

  i. ID 또는 Key 등 샘플의 고유 값을 나타내는 데이터 (예: ID, 핸드폰 번호 등)

  ii. 명목형 데이터: 카테고리를 분류할 수 있는 데이터 (예: 성별, 지역, 등)

  iii. 수치형 데이터: 최대/최소, 평균, 표준편차 등 기초 통계량을 구할 수 있는 데이터 (예: 나이, 기온, 강수량 등)

- 크롤링한 데이터 CSV 파일로 저장

2.1.2 Numpy, Pandas 및 기타 라이브러리를 이용한 데이터 전처리

데이터 전처리

- 결측치 데이터 처리: 결측치가 있다면 제거 또는 다른 값으로 대체
- 중복 데이터 처리: 중복된 데이터가 있다면 중복되는 것 중 하나의 샘플만 남기고 나머지 제거
- 분석에 필요없는 열(column) 제거: 분석 대상이 되는 열들만 남기고 나머지 제거
- 정규표현식을 통한 단어 추출
- etc.

2.1.3 전처리 후 데이터 분석

데이터 분석

- 데이터 통계량 및 카테고리 분석: 수치형 데이터의 기초 통계량과 카테고리 데이터의 항목 별 샘플 개수 등에 대한 기본적인 분석
- 여러 개의 열을 함께 고려한 분석 (예: 성별 평균 연령, 지역별 강수량 평균 등)
- 시각화: 의미 있는 내용을 시각화 (예: 기상청 데이터 중 지역별 평균 강수량의 막대 그래프, 네이버 영화 리뷰 중 가장 많이 나온 단어 WordCloud 작성)
- 데이터 분석을 통해 확인한 사실 명시
- etc.

2.2. 추가(자유) 항목

- 해당 데이터를 이용하여 자유롭게 시도해보고 싶은 것들 적용 후 결과 도출

- 세션 시간에 배우지 않은 것도 가능

- 필수 사항은 아니며 최대한 자유롭게 수행

- 창의성 및 참신성에 따라 가산점

- 예) 상관관계 분석, 지도 시각화, 실시간 어플리케이션, etc.

3. 프로젝트 제출물

1) PPT 작성 후 PDF로 저장한 파일 (최소 15장, 제목-목차-마지막 표지 등 모두 포함)

2) 코드 (Notebook 파일은 필수이며, 필요시 function 등 import해서 사용하는 코드들을 정리한 py파일 추가 제출 가능)

3) 크롤링한 데이터를 CSV 파일로 저장한 것 4) 모든 프로젝트물 명 통일: [기데분프로젝트]이름.pdf/csv/ipynb

4. 프로젝트 수행 가이드라인

분석 대상 데이터를 정할 때, 단순한 데이터가 아닌 추후 여러 가지 분석 시도 혹은 해석이 가능하며, 실생활과 관련이 밀접한 데이터를 선택할 것

정제되지 않은 데이터를 정제하는 연습 또한 데이터 분석에서 매우 중요한 절차이니, 아주 깔끔하게 정제된 데이터만 구하려고 할 필요 없음

통계적 가설 검정은 아직 배우지 않았으므로 정량적으로 수행할 필요는 없으나, 대상 데이터에 대해서 여러 가지 가설을 세우고 이에 대한 참/거짓을 판단할 수 있는 여러 가지 정성적 도구를 시도하는 것을 권장(단순 통계량 비교하기, 다양한 형태의 그래프/그림 그려보기 등)

- 예시: 보건복지부의 질병 발생률 데이터를 활용
- 가설: 대한민국의 지역과 연령, 그리고 성별에 따른 미세먼지로 인한 폐질환 발병률이 다를 것이다
- 분석 예시: 대한민국 시도별 발병률, 남/녀별 발병률, 연령대별 발병율에 대한 그래프/그림을 그려 유의미한 차이가 있는지를 해석

수업 시간에 배우지 않은 내용들이라도 스스로 여러 자료를 통해서 시도할 수 있는 다양한 분석 기법 최대한 시도하는 것을 권장(설사 이론을 잘못 이해하여 틀리게 적용하더라도 그에 대한 감점/페널티 부여하지 않음, 다양한 시도 유무가 제 1 기준임)

코딩 자체가 어려울 경우 ChatGPT 등의 도움을 받아서 코드 작성해도 됨

5. 프로젝트 주요 일정

공지 일정: 11월 19일(수) 7주차 세션

제출 기한: 12월 2일(화) 23:59

발표 일정: 12월 3일(수) 8주차 세션

우수 프로젝트 발표: 12월 21일