1차 프로젝트 - 영화 데이터 분석

일정

- 8월 22일 금요일 20시 미팅
  - 각자 Step1~6 코드 설명
  - 다음 미팅까지 해올 할일 정하기
    - 영화 데이터(IMDB) 사용 이유 결정
    - 데이터 분석으로 할 수 있는 솔루션 범위 정의
    - 협업 방식 결정(Github 활용)
    - 데이터 분석 주제 결정
  - 코드 리뷰 도구 공유(https://www.coderabbit.ai/)
- 8월 26일 화요일 20시 미팅
  - 할 일 1: 3.1 Github 사용법 알아보기
  - 할 일 2: 3.2.1 기초 통계/기술 통계 해보기
  - 할 일 3: 3.2.3.1 시기별 Rate, 시기별 관객 수, 시기별 흥행수익 시각화

진행

1. 영화 데이터 쓰는 이유:
   1. LLM 파인 튜닝 시 IMDB를 테스트용으로 많이 사용 - 데이터셋에 익숙
   2. 확장이므로, 기존 스탭1~6과 연결을 위해.
2. 데이터 분석으로 할 수 있는 솔루션 범위:
   1. 특정 데이터를 뽑아서, 비교해서 해석까지 하는 것
   2. 데이터 분석 -> 통계, 어떤 장르가 지금 인기다! 요즘 유행은 이거다!
   3. 통계 -> 실제 숫자로 해석에 대한 증거 제공
   4. 기초 통계/기술통계 -> 탐색적 데이터 분석(EDA)
3. 어떤 주제로 어떻게 할지 결정:
   1. Github 사용해서 협업
      - 4명이 각자 써놓은 코드 분리
      - 어떻게 해야 꼬이지 않고 잘 쓸 수 있을지 고민
      - 일단은, 각자 코드 준비,
      - 각자 폴더가 있는게 나중에 정리할 때 용이할 것
      - 각자 알아본 내용 다음 미팅에서 공유 -> 실습
   2. 데이터 분석 주제 결정
      1. 기초 통계/기술 통계
         - EDA에 필요한 정보
         - 1. 연도(Title 내 존재), 2. 장르(Genre), 3. 흥행수입(Gross), 4. 평점(Rate(청

중), 5. Metascore(전문가, 평론가)),6. 관객수(Votes in Info)

- 시기 = 10년
- 시기별 각 장르 개수 => 결과물: 10년 간 각 장르의 개수
- 시기별 각 장르에 대한 Rate, Metascore, 흥행수입, 관객수
    1. 시기별 장르 Rate => 결과물: 10년 간 각 장르의 Rate 평균
    2. 시기별 장르 관객 수 => 결과물: 10년 간 각 장르의 관객 수 평균
    3. 시기별 장르 흥행수익 => 결과물: 10년 간 각 장르의 흥행수익 평균
        - Metascore는 EDA 단계에서 인기 항목에서 제외할 예정으로 조사 X

2. 탐색적 데이터 분석(EDA) - 시기는 10년
    1. 흥행수익 기준 TOP 10 영화/장르 추출 (연도 상관X)
    2. 인기의 기준 3개 - Rate, 관객 수, 흥행성(흥행수익/관객수)
        - 흥행 수익은 흥행성으로 대체
        - Metascore는 전문가의 입장이므로 Rate로 대체
    3. 시기별 장르 흥행성 => 10년 간 각 장르의 (흥행수익/관객수) 평균 -> 시각화 1번 가능
    4. 각 인기 종류를 기준으로 시기별 인기가 높은 장르 뽑기
        - "10년 간" 정의: 0~9까지의 년도, 예: 1920~1929
        - 10년 간 Rate, 관객 수, 흥행성이 가장 좋은 장르가 나옴
        - 10년 동안 Rate가 가장 높은 장르, 관객 수가 가장 높은 장르,... 이런식
        - 다음 미팅에서 시각화할지 결정
    - 상관 관계는 다음 미팅에서 어느정도 진행된 후 생각해보기
3. 시각화 제대로 하기
    1. 시기별 Rate, 시기별 관객 수, 시기별 흥행수익 시각화
        1. 약간 예쁘게... 그래프 종류... 등 시도해보기.
    2. 시기별 Rate와 시기별 흥행성 상관관계 분석:
        - 1920년대 Rate / 흥행성 / 관객 수
        - 1930년대 Rate / 흥행성 / 관객 수
        - 1940년대 Rate / 흥행성 / 관객 수
        - 1950년대 Rate / 흥행성 / 관객 수

3. 프로젝트 정리 보고 양식 결정
    1. 다음 미팅에서 결정