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
- 8월 31일 일요일 20시 미팅
- 각자 브랜치 생성 및 Create Pull Request 수행
- main에 개인별 폴더 및 파일 업로드
- 리드미 파일 생성기 (다양한 템플릿)
  - https://readme.so/editor
- 마크다운 작성법 예시
  - https://stackedit.io/app#
  - https://www.markdownguide.org/cheat-sheet/

1. 데이터 정제(영혜님)
   - 총 21개 장르인데, 여기서 더 데이터가 축소될 수 있음.
   - 의미 없는 value는 빼고, 집중해서 보는 것도 의미있을 듯 (20개 이하는 제외)
2. 각 시기별 상승, 하락에 대한 이유 (재원,호성,시훈 미션)
   - 상승, 하락 요인 1개씩 정리 / 시각화 가시성 향상
   - 현재는 하락만 간단히 고려된 상황(경제위기, 영화산업 스튜디오 붕괴 등 이유)
   - 하락 및 상승 이유에 대한 심화 분석 필요

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
        1.
    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(청중), 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. 0826 이후 할 일 선정
                1. 데이터 정제(영혜님) - 총 21개 장르인데, 여기서 더 데이터가 축소될 수 있음.
                    1. 의미 없는 value는 빼고, 집중해서 보는 것도 의미있을 듯 (20개 이하는 제외)
                2. 장르에 얽매이지 않고, 시기별 Rate, Vote, Gross 평균 - 시각화(완료)
                3. 각 시기별 상승, 하락에 대한 이유
                    1. (재원,호성,시훈 미션) 상승, 하락 요인 1개씩 정리 / 시각화 예쁘게...
                    - 현재는 하락만 생각됨(경제위기, 영화산업 스튜디오 붕괴 등 이유)
                    - 추후에 상승 이유도 넣어보면 좋을 듯
        3. 시각화 제대로 하기
            1. 시기별 Rate, 관객 수, 흥행수익 평균 추이 시각화
                1. 약간 예쁘게... 그래프 종류... 등 시도해보기.
            2. 시기별 Rate, 관객 수, 흥행수익 평균 추이 상승/하락 요인 분석

        1. 보고서에 바로 넣어도 될만큼 예쁘게 만들기
3. 프로젝트 정리 보고 양식 결정 (다음 미팅(0831)에서 결정)
    1. Step 1~5. 호성님 영혜님 가산점
    2. 심화에서는 1개씩 상승/하락 요인 범위를 최종적으로 합치는 방식
    3. 재원 코드를 가져와서 저 양식에 추가해서
    4. SEABorn 등.. 예쁜 라이브러리로 시각화 하기.