230419 시계열팀 주분 회의; EDA 역할 분담

To do: EDA 역할분담 (종목별로? 아니면 역할별로?), 같이 만나서 할 날짜 정하기, EDA 완료 일정 공유

진행상황

: 데이터 수집 완, 공통지표 데이터셋 병합 중

같이 만나서 작업해요... 2주만...

: 수요일 6시 이후~10시 이전, 금요일 3시~6시 어떤가여?

EDA 완료 목표일자

: (데이터 수집이 이미 끝난 관계로...) 4/27(목)!!!!

[EDA]

최종 데이터셋 구성

- 종목 별로 하나의 데이터셋으로 합치기 (총 5개의 데이터셋)
- 날짜를 기준으로 병합, 연/월/일 변수 생성
- 일단 기사 변수는 제외하고 합치기

X변수

- 각 X변수의 결측치, 이상치 확인 후 처리
- 각 X변수에 대한 분포 확인, 추세 및 계절성 확인 (시각화)

- X변수끼리 상관분석
- (선택) X변수들만 클러스터링 (비슷하게 묶이는 변수들 구분)
- 파생변수 생성 (ex. 전일 대비 주가 상승률)

Y변수

- Y변수의 분포, 추세, 계절성 확인
- Y변수만 가지고 과거의 값들로 현재의 값 예측했을 때, 몇 시점까지의 데이터가 유의미한 영향이 있는지 분석 (시계열 분석)
- 각 X변수와 Y변수에 대한 상관분석 (어떤 변수가 Y변수와 높은 상관관계를 갖는가?) (~t-1 시점까지의 과거의 X변수의 t시점의 Y값에 대한 상관관계 파악)
- 토론방, 기사 데이터 등 주요 X변수들도 몇 시점 전까지의 데이터가 Y변수에 유의미한 영향이 있는지 분석 (시계열 분석)
- (가능하다면) 어떤 X변수과 Y변수에 주로 후반영되는지 / 어떤 X변수가 Y변수에 주로 선반영되는지

기사 데이터 (생각보다 쉬울 것 같음)

- 변수 2개: 일별 긍부정도, 일별 기사 개수
- 변수 생성 후 다른 X변수, Y변수와의 상관분석 필요
- https://medium.com/naver-cloud-platform/%EC%9D%B4%EB%A0%87%EA%B2%8C-%EC%82%AC%EC%9A%A9%ED%95%98%EC%84%B8%EC%9A%94-%ED%85%8D%EC%8A%A4%ED%8A%B8-%EA%B0%90%EC%A0%95-%EB%B6%84%EC%84%9D-%EC%84%9C%EB%B9%84%EC%8A%A4-%EA%B5%AC%ED%98%84%ED%95%98%EA%B8%B0-clova-sentiment-%ED%99%9C%EC%9A%A9%EA%B8%B0-5d9db7b0209b
- https://techblog-history-younghunjo1.tistory.com/111
- 문제는 보통 감성분석이 긍정/중립/부정인데 중립이 너무 많이 나올 것 같음. 어떻게 긍

부정도를 수치화할 것이냐? 아니면 긍정, 부정, 중립 세 개를 다 변수로 만드는 방법도 있음

이후의 계획...

- 데이콘 팀 결성 하는 거 까먹으면 안됨 (시험 끝난 이후에 합시다)
- 프로젝트 방향성 확정 (아마 매매(매도매수유지) 추천으로 갈 듯? 몰라요)
- 몇 년치 데이터를 모델링에 활용할 것이냐
- 변수선택법, 사용 모델 등.....