







# 데이터 수집 및 전처리: 태은, 지호

## 1. 데이터 수집

-  리서치팀 자료 수신
  - 금융시장 리서치팀(후민)으로부터 종목별 건제 지표 및 시장 환경 데이터 수신
  - 예: 금리, 환율, 인플레이션, 각종 뉴스 요약 등
-  공식 API 또는 웹 크롤링 활용
  - Yahoo Finance, Alpha Vantage, Quandl 등에서 다음 데이터 수집:
    - 종가(Close), 시가(Open), 고가(High), 저가(Low), 거래량(Volume)
    - 일별, 주별, 월별 주가 데이터
    - 배당 정보(회사가 주주에게 이익을 나눠주는 금액), PER(주가수익비율,  $\text{주가} \div \text{주당순이익}$ ), PBR(주가순자산비율,  $\text{주가} \div \text{주당순자산}$ ), 시가총액(기업의 전체 가치를 의미하는  $\text{주가} \times \text{총 주식 수}$ ) 등 기업의 재무 상태를 나타내는 주요 지표 예시를 포함함
-  기간 설정 및 일부 결측치 허용
  - 모든 종목의 주가 데이터를 같은 기간(예: 2018년 1월 1일부터 2025년 5월 9일까지) 동안 빠짐없이 모으기
  - 일부 날짜에 결측치(데이터가 없는 경우)가 있더라도 전체 데이터를 분석에 활용할 수 있도록 허용 있는 결측치가 있는 날짜는 제거하거나 적절한 값으로 채워 넣어도 됩니다.

---

## 2. 데이터 정리 및 통합

-  컬럼명 통일 및 필요 컬럼 추출
  - 컬럼명 영문화 및 표준화 (종가 → Close, 날짜 → Date)
  - 불필요한 여러 열 제거 및 필요한 열만 유지
-  형식 통일 및 인코딩 처리
  - 날짜 형식 통일 (예: 2025-05-09 형식의 YYYY-MM-DD)
  - 문자열-숫자 변환 (예: '100'이라는 글자를 숫자 100으로 변환)
-  데이터 통합 및 정렬
  - 종목별 데이터를 하나로 병합하여 CSV 파일로 저장 (파일 형식은 .csv, 인코딩은 UTF-8)
  - 날짜 순서(오름차순)로 정렬하고, 모든 종목에 같은 날짜가 존재하는지 확인
  - CSV파일에는 다음 항목들이 포함되어야 합니다:
    - 종목명(Ticker): 예를 들어 Apple은 'AAPL', Microsoft는 'MSFT'처럼 각 회사의 고유한 약어입니다.
    - 날짜(Date): 데이터가 기록된 날짜. 예: '2025-05-09'
    - 시가(Open): 해당 날짜에 시장이 열릴 때의 주가. 예: 145.00
    - 종가(Close): 해당 날짜에 시장이 마감될 때의 주가. 예: 150.20
    - 거래량(Volume): 하루 동안 거래된 주식 수. 예: 3,000,000 이런 형식으로 구성된 CSV는 모델 학습에 필요한 주요 데이터를 깔끔하게 정리할 수 있도록 도와줍니다.

### 3. 결승치 및 이상치 처리

- ✕ 결승치 처리

- 단순 제거 (dropna기법) 또는 평균/이전 값으로 채워주기 (fillna기법)
- 금융 특성에 따라 forward fill, backward fill 적용 (아래서 설명)
- **Forward fill**: 결측값을 **이전 값으로 채우는 방법**입니다. 예를 들어 주가 데이터에서 5월 1일 값이 100이고, 5월 2일 값이 비어 있다면, 5월 2일을 100으로 채웁니다.
- **Backward fill**: 결측값을 **다음 값으로 채우는 방법**입니다. 예를 들어 5월 2일이 비어 있고, 5월 3일 값이 102라면, 5월 2일을 102로 채웁니다.

- 🕯 이상치 제거

- Z-score, IQR 방식으로 이상치 탐지 (이상치는 일반적인 값 범위에서 벗어나는 값으로, 주가 데이터에서는 갑자기 급등하거나 급락한 비정상적인 값을 말함)
- **Z-score**: 각 데이터가 평균에서 얼마나 떨어져 있는지를 나타내는 지표로, 일반적으로 Z-score가 3보다 크거나 -3보다 작으면 이상치로 간주하여 해당 지표는 평균값으로 맞추거나 삭제 혹은 이상적인 값을 적어줌

Z 점수 구하는 공식까진... 안 알려줘도 되겠죠..?

- **IQR (Interquartile Range)**: 데이터의 중간값을 기준으로 사분위 범위를 계산해, 1사분위(Q1)와 3사분위(Q3)를 기준으로  $Q1 - 1.5 \times IQR$ 보다 작거나,  $Q3 + 1.5 \times IQR$ 보다 큰 값들을 이상치로 판단
- 갑작스러운 급등이나 급락과 같은 비정상적인 주가 변화(스파이크)는 모델의 예측에 방해가 되기 때문에, 이상치로 간주해 분석에서 제외하거나 보정하여 처리

---

### 4. 정규화 및 스케일링

- ⊗ 모델 학습을 위한 정규화

- 주가나 거래량처럼 숫자의 크기가 크게 차이 나는 데이터를 같은 범위로 맞추는 작업
- 예를 들어, 어떤 회사의 주가는 1000원인데 다른 회사는 1만원이라면, 이 숫자들을 0~1 사이로 바꿔서 모델이 잘 비교할 수 있게 만들 > 이렇게 하는 이유가 10억 이랑 10만원 비교하는 것도 컴퓨터에서는 제법 많은 시간을 잡아먹어서 최적화를 위한 과정이라고 생각해주세요.
- 보통 Min-Max Scaling(최솟값과 최댓값 기준) 또는 Standard Scaling(평균과 표준편차 기준) 방법을 사용함

---

### 5. 최종 저장 및 공유

- ⊗ 하나의 CSV 파일로 저장

- 여러 종목의 데이터를 하나의 파일로 정리해서 저장
- 팀원들과 쉽게 공유할 수 있도록 형식을 통일하고, 뒤에 있는 `processed/` 폴더에 저장
- 예: `processed_data_2024_05_09.csv`, `processed_data_2023_12_31.csv`, `processed_data_2025_01_15.csv`

- 📁 디렉터리 구조 정리

- **raw/**: 원본 데이터를 저장하는 폴더 (리서치팀 또는 외부 API에서 받은 처음 데이터)
- **processed/**: 전처리를 마친 데이터를 저장하는 폴더 (결측치 처리, 정규화 등 수행 후)
- 파일 이름에 날짜와 종목명을 포함해 나중에 찾기 쉽게 관리함

---

☒ 협업 체크리스트

- output 폴더에

**항목 완료 유무 [O / X]**

리서치팀 자료 수신	O / X
종목별 시계열 데이터 수집	O / X
결측치 및 이상치 처리	O / X
정규화 및 저장	O / X
전처리 완료 데이터 팀 공유	O / X