Data Description

1. 데이터 파일 종류

```
data/
├── preprocessed/           # 가공된 데이터 저장
│   ├── all_data.csv        # 아래 데이터 전체 통합(매치가 안되는 날짜 nan으로 채움, 2015/01/02 - 2025/03/23)
│   ├── fertilizer.csv      # 비료 가격 데이터(month 기준, 2015/01 - 2024/11)
│   ├── macroeconomic.csv   # 거시경제지표 데이터(Year 기준, 2015 - 2023)
│   ├── market.csv          # 환율, 커피 가격 등 시장 데이터(Date 기준, 2015/01/01 - 2025/04/09)
│   ├── weather_with_lag.csv  # 기후 데이터(lag feature 포함, Date 기준, 2015/01/01 - 2025/03/23)
│   ├── weather.csv         # 기후 데이터(Date 기준, 2015/01/01 - 2025/03/23)
├── raw/                    # 가공되지 않는 데이터 저장
```

2. 컬럼 설명

> 줄이 그어져 있는 컬럼은 전처리 과정에서 삭제된 컬럼

2-1. 기후 데이터

컬럼명	설명
~~YEAR~~	~~연도 정보~~
~~MO~~	~~월 정보~~
~~DY~~	~~일 정보~~
Date	날짜 (YYYY-MM-DD 형식)
ALLSKY_SFC_UV_INDEX	전체 하늘 자외선 지수
ALLSKY_SFC_SW_DWN	전체 하늘 지표면 단파 복사량 (W/m^2)
WS2M	지표면(2m) 풍속 (m/s)
~~CLRSKY_SFC_SW_DWN~~	~~맑은 하늘 지표면 단파 복사량 (W/m^2)~~
T2M	2m 기온 (°C)
~~T2M_MAX~~	~~최고 기온 (°C)~~
~~T2M_MIN~~	~~최저 기온 (°C)~~
RH2M	상대 습도 (%)
PRECTOTCORR	강수량 (mm)

컬럼명	설명
PS	기압 (Pa)
~~CLRSKY_SFC_PAR_TOT~~	~~맑은 하늘 지표면 광합성 유효복사 (mol/m²/day)~~
~~WS10M_MAX~~	~~10m 최대 풍속 (m/s)~~
~~WS10M_MIN~~	~~10m 최소 풍속 (m/s)~~
~~ALLSKY_KT~~	~~전체 하늘 투과율 지수~~
~~T2M_RANGE~~	~~일교차 (T2M_MAX - T2M_MIN)~~
~~WS10M~~	~~10m 평균 풍속 (m/s)~~
~~TS~~	~~지면 온도 (°C)~~
~~WSC~~	~~풍향 코드 혹은 기타 기상 요소~~
locationName	지역 정보 (예: brazil_varginha)
season_tag	수확 시기 구분 (harvest, pre-harvest, off-season 등)
days_until_harvest	해당 날짜로부터 다음 수확까지 남은 일수

+ Lag Features

컬럼명	설명
T2M_lag_1m ~ T2M_lag_6m	2m 기온의 이전 1~6개월 값
WS2M_lag_1m ~ WS2M_lag_6m	지표면 풍속의 이전 1~6개월 값
ALLSKY_SFC_SW_DWN_lag_1m ~ _6m	지표면 단파 복사량의 이전 1~6개월 값
ALLSKY_SFC_UV_INDEX_lag_1m~_6m	자외선 지수의 이전 1~6개월 값
PRECTOTCORR_lag_1m ~ _6m	강수량의 이전 1~6개월 값
RH2M_lag_1m ~ RH2M_lag_6m	상대 습도의 이전 1~6개월 값
PS_lag_1m ~ PS_lag_6m	기압의 이전 1~6개월 값

2-2. 시장 데이터

컬럼명	설명
~~Date(조인 기준, 기후 Date를 살림)~~	~~날짜 (YYYY-MM-DD 형식)~~
Coffee_Price	커피 가격 (국제 시세 기준)
Crude_Oil_Price	원유 가격 (배럴당 가격, 보통 Brent 기준)
USD_KRW	미국 달러 대비 원화 환율
USD_BRL	미국 달러 대비 브라질 헤알 환율

컬럼명	설명
USD_COP	미국 달러 대비 콜롬비아 페소 환율
USD_ETB	미국 달러 대비 에티오피아 비르 환율
Coffee_Price_Return	커피 가격의 전일 대비 수익률 (%)
Crude_Oil_Price_Return	원유 가격의 전일 대비 수익률 (%)
USD_KRW_Return	원-달러 환율의 전일 대비 수익률 (%)
USD_BRL_Return	브라질 환율 수익률
USD_COP_Return	콜롬비아 환율 수익률
USD_ETB_Return	에티오피아 환율 수익률
~~month~~	~~월 단위 기준일 (YYYY-MM-01 형식)~~
Urea_price	요소비료 가격 (단위: USD/톤 등)
Dap_price	인산디암모늄(DAP) 비료 가격 (USD/톤 등)

2-3. 거시경제지표 데이터

컬럼명	설명
~~Area(조인 기준, 기후 locationName을 살림)~~	~~국가명 (거시경제 데이터 기준)~~
~~Year(조인 기준, 기후 Date을 살림)~~	~~연도 (거시경제 데이터 기준)~~
Production	농업 생산지표 (단위에 따라 다름)
~~index~~	~~인덱스 구분자 또는 레코드 식별자~~
Agricultural raw materials exports (% of merchandise exports)	농업 원자재 수출 비율 (% 기준)
Merchandise trade (% of GDP)	상품 무역 비율 (GDP 대비 %)
Unemployment, male (% of male labor force)	남성 실업률 (ILO 추정 기준)
GDP per capita (current US$)	1인당 GDP (현재 미국 달러 기준)
IMF repurchases and charges (TDS, current US$)	IMF 차입 상환 및 수수료 (달러 기준)
Food production index (2014-2016 = 100)	식량 생산 지수 (기준 연도: 2014–2016 = 100)
Political Stability and Absence of Violence/Terrorism...	정치 안정성 및 폭력/테러 부재 지표 (상위 90% 신뢰 구간 백분위)
GDP per capita growth (annual %)	1인당 GDP 연간 성장률 (%)
Merchandise exports to low- and middle-income...	저/중소득 국가로의 상품 수출 비율 (%)
Export unit value index (2015 = 100)	수출 단가 지수 (2015년 기준 = 100)
Rural population	농촌 인구 수

컬럼명	설명
Permanent cropland (% of land area)	영구 농지 비율 (국토 면적 대비 %)
Cereal yield (kg per hectare)	곡물 수확량 (헥타르당 킬로그램)

3. 데이터 가공 과정

3-1. 기후 데이터 가공

1. NASA에서 WSC(풍향 코드) 데이터 같은 경우는 특정 지역들(네 지역 정도)만 제공함. 따라서 컬럼 통일을 위해 삭제하였음.
2. 최근 날짜의 기후 데이터는 전부 -999로 설정되어 있었고, 그 비중은 전체 데이터에 0.3% 정도밖에 되지 않아 삭제하였음.(2025/03/24-2025/04/09 삭제)
3. ALLSKY_SFC_UV_INDEX 컬럼의 결측치는 8% 정도로 삭제하기에는 너무 많았음. 따라서 월평균으로 보간하였음.
4. YEAR, MO, DY를 하나의 컬럼으로 "Date"로 통합하였음.
5. 불필요한 컬럼들을 삭제하였음.(`2. 컬럼 설명`에서 확인 가능 & `raw/`에서 찾을 수 있음.)
6. 파생 컬럼(수확시기, 수확 시기까지 남은 기간, 재배지역 이름)을 추가하였음.
7. locationName 별로 기후 컬럼에 1~6개월의 lag feature를 생성하였음.

3-2. 시장 데이터 가공

> market.csv에 해당

1. 환율, 유가, 커피 가격의 결측치는 전날로 보간하였음.(시장이 안열렸기에 결측치 발생으로 판단)
2. 가격은 이전 값으로 대체했기 때문에, 변화율 결측치는 0.0으로 처리하였음.

3-3. all_data.csv 데이터 가공

1. `weather.csv`에는 최근 데이터에 결측치가 발생하였고, `3-1` 설명과 같이 결측치 보정함.
2. `weather_with_lag.csv`는 lag feature 생성 이후, 초반 일부 lag 결측치를 제외하면 결측치가 없음.
3. `weather_with_lag.csv` + `marker.csv` 조인 후 **첫 번째 행(2015/01/01) 삭제하여 결측치 제거**하였음. → `market.csv`의 변화율은 첫날 기준 데이터가 없어 결측치가 발생했기 때문
4. `fertilizer.csv` 까지 조인 후에는 1,005 개의 결측치가 발생하였고 nan 상태로 놔둠. 이는 전체 행에서 2.97%만 해당하는 크기임. `fertilizer.csv`의 데이터는 2015/01부터 2024/11까지만 제공해서 발생한 결측치임.(기존 데이터는 2015/01/02 - 2025/03/23) → `fertilizer.csv`는 월 단위 데이터로, 2015/01 ~ 2024/11까지만 제공되며, 일별로 기존 월 데이터를 확장해 사용
5. `all_data.csv`는 기존 데이터에 `macroeconomic.csv` 거시경제 지표 데이터를 조인 한 데이터임. 이 데이터에는 다음과 같이 결측치가 발생하였음.

> "총 33,593개의 샘플에서 15,485개의 샘플에 결측치가 존재함." 결측치가 있는 컬럼 요약 (lag feature 제외)

컬럼명	결측치 개수
Urea_price	1005
Dap_price	1005

컬럼명	결측치 개수
Production	4019
Agricultural raw materials exports (% of merchandise exports)	4019
Merchandise trade (% of GDP)	4019
Unemployment, male (% of male labor force) (modeled ILO estimate)	4019
GDP per capita (current US$)	4019
IMF repurchases and charges (TDS, current US$)	4019
Food production index (2014-2016 = 100)	7304
Political Stability and Absence of Violence/Terrorism: Percentile Rank, Upper Bound of 90% Confidence Interval	4019
GDP per capita growth (annual %)	4019
Merchandise exports to low- and middle-income economies within region (% of total merchandise exports)	13874
Export unit value index (2015 = 100)	4019
Rural population	4019
Permanent cropland (% of land area)	7304
Cereal yield (kg per hectare)	7304

거시경제지표 데이터는 2015 - 2023 기간밖에 제공을 하지 않아 많은 결측치가 발생하였음. 또한 거시경제지표 데이터는 연 단위이며, 일 단위 데이터와 병합을 위해 별도 보간 없이 그대로 일별로 확장 처리하였음.