KOSPI Dataset Description

1. Overview of Data Schema

This dataset was constructed to build a model that predicts the
next-day excess return of the KOSPI index.
The final dataset includes both raw market information and derived features designed to capture return patterns, trends, volatility, and downside risk.

Feature List

date, close, high, low, open, volume,
ret_1d, ret_5d, ret_22d,
ma_5, ma_20, ma_60, ma_120,
vol_22d, rsi_14,
macd, macd_signal, macd_hist,
atr_14,
bb_middle, bb_upper, bb_lower,
momentum_10, momentum_20, momentum_60,
drawdown_60, max_drawdown_60
vix, usdkrw, gold_price, NSI(News Sentimental Index)

2. Data Collection and Processing

① Price-Based Raw Data

- Columns: date, open, high, low, close, volume
- Source: Yahoo Finance (^KS11)
- Tool: Python yfinance
- Purpose:
  Raw price information forms the foundation for all other engineered features and is essential for any return prediction task.

② Return & Momentum Features

- Columns: ret_1d, ret_5d, ret_22d, momentum_10, momentum_20, momentum_60
- Source: Derived from the **close** price series.
- Method
  - **Returns:** Computed as percentage changes in closing prices over 1, 5, and 22 trading days.
  - **Momentum:** Calculated as the difference between today's closing price and the closing price 10, 20, or 60 days prior.
- Rationale:
  - Past return patterns are among the strongest predictors of future excess returns
  - Momentum captures short- and medium-term directional movements
  - These features enable the model to learn how price dynamics evolve across time horizons

③ **Trend Indicators**

- Columns: ma_5, ma_20, ma_60, ma_120, macd, macd_signal, macd_hist
- Source: All trend indicators are computed using the **close** price series.
- Method:
  - **Moving Averages (MA):** Calculated as simple rolling averages over windows of 5, 20, 60, and 120 days.
  - **MACD:**
    - 12-day and 26-day exponential moving averages (EMA) of close
    - MACD = EMA12 – EMA26
    - Signal = 9-day EMA of MACD
    - Histogram = MACD – Signal
- Rationale:
  - Moving averages extract long- and short-term trend information
  - MACD features help identify potential trend reversals
  - As returns are heavily influenced by market regimes, trend indicators provide essential predictive context

④ **Volatility & Risk Indicators**

- Columns: vol_22d, atr_14, bb_middle, bb_upper, bb_lower
- Source: Derived using combinations of **close**, **high**, and **low** price series.
- Method:
  - Rolling Volatility (vol_22d):
    - Standard deviation of 1-day returns over a 22-day window
  - ATR (atr_14):
    - Based on True Range: max(high-low, |high-prev_close|, |low-prev_close|)
    - Averaged over 14 days
  - Bollinger Bands:
    - Middle band: 20-day moving average of close
    - Upper band: middle + (2 × 20-day rolling standard deviation)
    - Lower band: middle – (2 × 20-day rolling standard deviation)
  - RSI (14):
    - Based on 14-day smoothed average gains and losses
- Rationale:
  - Volatility reflects uncertainty and market stress, both closely linked to excess return behavior
  - ATR measures absolute price movement
  - Bollinger Bands capture both volatility and relative price position
  - These features help the model recognize high-risk or high-uncertainty environments

⑤ **Downside Risk Measures**

- Columns: drawdown_60, max_drawdown_60
- Source: Derived from **close** price history.
- Method:
  - Compute rolling 60-day maximum price
  - Drawdown = (close – rolling_max) / rolling_max
  - Max drawdown = minimum drawdown within the past 60 days
- Rationale:

    - Drawdown features quantify how far the index has fallen from recent highs
    - They indicate market weakness or oversold conditions
    - Downside risk signals provide supplementary information relevant to future return shifts

⑥ **External Market Indicators**

- Columns: vix, usdkrw, gold_price, NSI(News Sentimental Index)
- Source: Yahoo Finance
- Tool: Python yfinance
- Rationale:
    - External macro-financial variables provide information that is not contained within domestic KOSPI price movements.
    - VIX captures **global volatility shocks**, USD/KRW exchange rate reflects **foreign capital sensitivity and risk-on/off flows**, and gold price represents **global safe-haven sentiment**.
    - Including these exogenous indicators helps the model detect broader market conditions that influence short-term excess returns in the Korean equity market, particularly during periods of heightened uncertainty or macro-driven movements.

3. Summary

The KOSPI dataset was designed explicitly for **excess return prediction**,
organized into a structured hierarchy:

Price → Returns/Momentum → Trend → Volatility → Downside Risk

Each feature group captures a different aspect of market behavior,
and together they form a comprehensive set of inputs
that enhance the model's ability to forecast next-day returns.