# Data Collection (Dataset Card Analog)

This section describes the source and method used to acquire the raw data for the BTC/QQQ Cross-Market Extension.

| Component | Description |
|---|---|
| **Data Source** | Financial Data API (Twelve Data) |
| **Assets Acquired** | **BTC/USD** (Cryptocurrency) and **QQQ** (Invesco QQQ Trust ETF) |
| **Collection Method** | Data was acquired via REST API calls, requesting **daily** time-series data using a personalized API key. The request spanned from **2005-11-01 to 2025-11-14**. |
| **Compliance/License** | The collection process adhered strictly to the principle of respecting the API provider's **Terms of Service (ToS)**, **rate limits**, and **copyright/license requirements**. The non-commercial nature of this academic project minimizes the license burden. |
| **Raw Features** | datetime, open, high, low, close for both assets, with volume included for QQQ. |

# Data Preprocessing

This describes the cleaning, handling of missing values, and initial alignment necessary before feature engineering.

## 1. Initial Cleaning and Alignment

The data requires careful synchronization due to the disparate trading schedules of the two markets:

- **Data Type Conversion:** The datetime column was converted to a pandas datetime index and used to sort the data chronologically. All price/volume columns were converted to a numeric format.
- **Missing Value Handling:** Rows with missing closing prices were dropped as they are essential for calculating returns.
- **Time Alignment (Inner Join):** BTC trades 24/7 (daily entries), while QQQ trades only on weekdays. An **inner join** was performed on the datetime index to align the combined dataset, synchronizing it to the **QQQ trading calendar** (weekdays only).

## 2. Target Variable Preprocessing

The primary goal of preprocessing was to construct the required target variable, an analog to the competition's market_forward_excess_returns.

- **Risk-Free Rate:** The {Risk_Free_Rate} was defined as a constant daily rate (e.g., 2% annualized)
- **Forward Returns:** The daily return for each asset was calculated and then shifted backward one day to represent the future return the model aims to predict.
- **Expected Return:** A **5-year (1260-day) rolling mean** of the forward returns was computed to establish a long-term expectation baseline.

- **Winsorization:** The final target was obtained by subtracting the expected return from the forward return and then **winsorizing** the result using a **Median Absolute Deviation (MAD) criterion of 4**. This process minimizes the influence of extreme outliers, promoting a more stable target for the machine learning models.

# Feature Engineering Process

The feature engineering process was designed to convert the raw daily OHLCV data for BTC and QQQ into a comprehensive set of predictive signals. This feature set strategically mirrors the categories found in the main S&P 500 competition (MOM*, V*, P*, E*, D*) to build a robust model for financial time-series forecasting.

## 1. Core Technical and Momentum Indicators (MOM*, P*)

These features capture the intrinsic movement and trend of each asset:

- **Momentum Indicators (MOM*)**: **Daily Return** and **Momentum Ratios** (current price relative to N days prior) were computed for lookback windows (N=5, 10, 20, 60) to capture short- and long-term trend strength.
- **Trend Following (P*)**: **Simple Moving Averages (SMA)** were calculated across the same lookback windows. These define the asset's overall trend and act as a price-level proxy.
- **Advanced Trend (M*)**: The **Relative Strength Index (RSI)** was included to measure the speed and change of price movements, providing signals on overbought/oversold conditions.

## 2. Volatility and Risk Metrics (V*)

Risk features are essential for managing the project's 120% Volatility Constraint:

- **Rolling Volatility (V*)**: The **Standard Deviation of daily returns** was computed over windows (e.g., 20, 60 days) to directly measure historical risk.
- **EWMA Volatility (V*)**: The **Exponentially Weighted Moving Average (EWMA) Volatility** was generated, which places greater emphasis on recent data, providing a more responsive estimate of next-day risk compared to simple rolling standard deviation.
- **Price and Range Proxies (V*)**:
    - **Intraday Volatility Proxy** (ln(High/Low)) measures volatility within a single day.
    - The **OC Range Ratio** and **BTC Overnight Gap** capture volatility arising from price movement relative to the day's total range or gaps between trading sessions.

## 3. Cross-Market and Macroeconomic Proxies (E*, I*)

The combination of BTC (crypto risk) and QQQ (US growth equity) data is leveraged to create powerful cross-market signals, serving as proxies for the original **Macro Economic (E*)** and **Interest Rate (I*)** features:

- **Spillover Effects (E*)**: The **previous day's return of the cross-asset** (QQQ $\text{Return}_{t-1}$ predicts BTC $\text{Return}_t$ and vice versa) was used to model the immediate transmission of global risk-on/risk-off sentiment across markets.
- **Rolling Correlation (E*)**: The **30-day Rolling Correlation between BTC and QQQ** returns was calculated. This serves as a regime indicator: high correlation implies markets are unified by macro factors, while low correlation suggests decoupled, asset-specific trading.
- **Volume Indicators (M*)**: For QQQ, **Rolling Average Volume** and the **Volume Ratio** (Current Volume / Avg $\text{Volume}_N$) were included to measure liquidity and trend conviction.

## 4. Regime and Structural Features (D*)

- **Calendar Indicators (D*)**: Categorical features for **Day of Week (DOW)**, **Month**, and **Quarter** were one-hot encoded to capture structural trading patterns, such as the volatility change between weekdays and weekends in the BTC market.

This detailed and justified feature set provides the necessary input for training machine learning models to predict the target variable, which is the final winsorized excess return.