KaxaNuk

Sharing knowledge

Data Curation & Feature Engineering

Session 2

Content

- ➔ The Data Challenge in Investment Research
- ➔ Introduction to Data Curator
- ➔ Outlier Detection & Correction
- ➔ Feature Engineering
- ➔ Hands-On: Your First Data Pipeline

The Data Challenge in Investment Research

Bottlenecks

The Data Challenge

60–80%

of a quant researcher’s time may be wasted on repetitive data wrangling tasks.

Reproducibility Crisis

- Undocumented workflows prevent reliable validation of results.

Scattered Sources

- Inconsistent provider formats require heavy normalization.

Manual Errors

- Human intervention introduces silent data failures.

Lost Velocity

- Time spent fixing pipelines reduces research productivity.

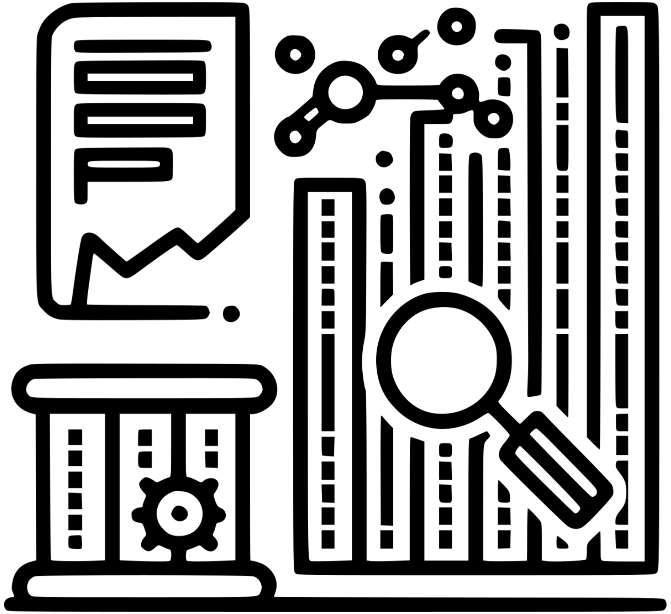
The Messy Middle Problem

Major bottlenecks:

- Data quality
- Pipeline consistency
- Governance gaps

In quant investing, this looks like:

- 40 Jupyter notebooks
- 15 slightly different datasets
- No reproducibility
- No audit trail



Professional Research Demands

Structured, versioned, reproducible data systems.

The Tooling To Speed Up Feature Engineering

Data Curator

Data Weaver

Data Analyzer

Introduction to Data Curator

The Solution to Improve Data Wrangling Tasks

Data Curator

A Python-based tool designed to **build structured databases** for market, fundamental, and alternative data, **transforming chaotic wrangling into a systematic workflow**.

Unified Data Access

Single interface for various data providers and custom in-house datasets, with homogenized tags.

Vendor Data Verification

Assess whether the provider's data is reliable or affected by serious issues (e.g., negative values, restated data).

Prebuilt Calculations

Ready-to-use functions for outlier adjustments, ratios, momentum, trend, volatility, volume, etc.

Custom Features

Easily define your own custom feature functions to extend the capabilities of the Data Curator.

Excel-Based Configuration: Specify tickers, dates, and features in a spreadsheet, **no complex coding required for setup**.

https://kaxanuk-data-curator.readthedocs.io

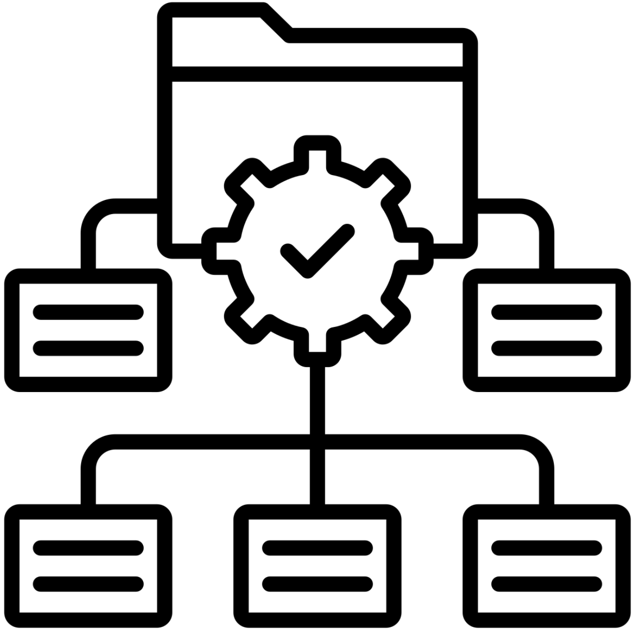
Structured Financial Data

Data is messy:

- Missing earnings dates
- Inconsistent ticker mapping
- Outliers
- Corporate actions
- Different frequencies

Structured data means:

- Aligned timestamps
- Point-in-time data
- Cleaned values
- Standardized schema
- Version-controlled datasets



Time Matters: Avoiding Look-Ahead Bias

Every data point must respect:

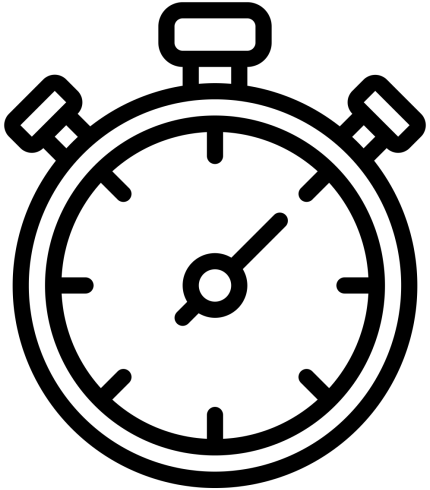
❑ **Release date**
❑ **Availability date**
❑ **Revision history**

Example:

- Q1 earnings released on May 5
- You cannot use them in March backtests.

Feature Engineering must be:

Chronologically honest.



Getting Structured Financial Data

The 3-Step Workflow

01 Install

Set up the environment. Requires Python 3.12+.

```
pip install kaxanuk.data_curator
```

# Optional extension

```
pip install kaxanuk.data_curator_extensions.yahoo_finance
```

02 Configure

Initialize and edit settings in Excel.

```
kaxanuk.data_curator init excel
```

Config/parameters_datacurator.xlsx
Define Tickers, Dates, Providers

Config/.env Store API Keys securely

03 Execute

Run the pipeline to fetch and process data.

```
kaxanuk.data_curator run
```

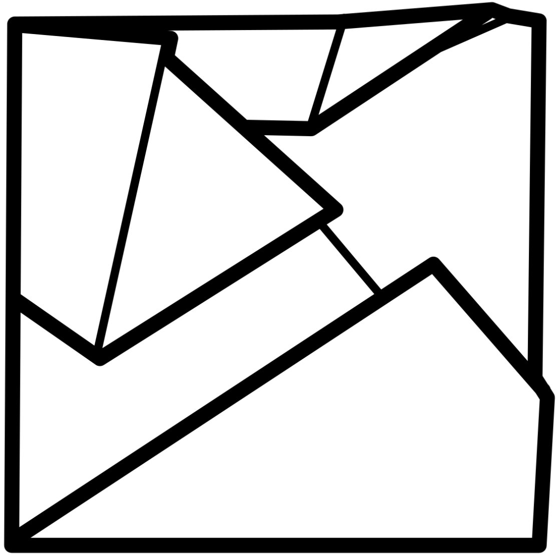
Output/ Structured CSV/Parquet files ready for analysis

https://kaxanuk-data-curator.readthedocs.io/en/stable/user_guide/quick_start.html

Before

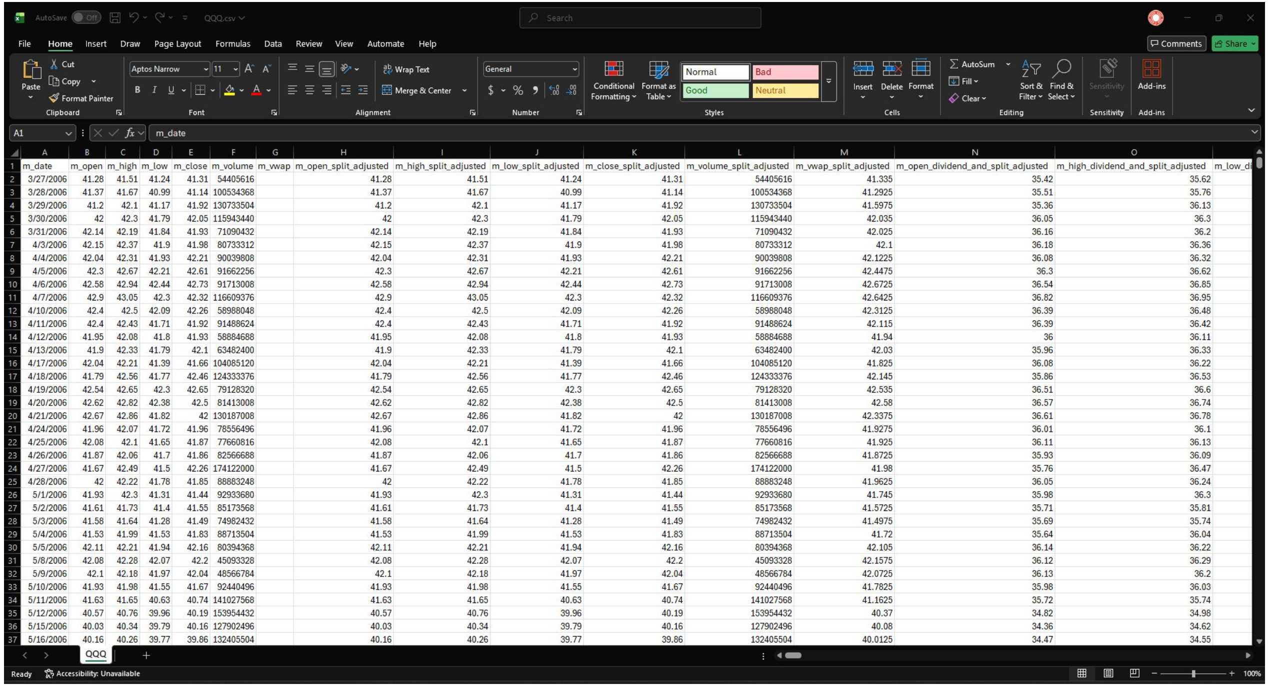
Fragmented infrastructure:

- Multiple APIs with heterogeneous formats
- Inconsistent field naming across providers
- Missing or partially available data
- Manual joins and ad-hoc transformations
- Point-in-time misalignment
- Non-reproducible research notebooks
- Difficult feature sharing across teams
- High debugging overhead
- ...



Unified Time Series Output Across All Tools

Data Curator



m_date	m_open	m_high	m_low	m_close	m_volume	m_vwap	m_open_split_adjusted	m_high_split_adjusted	m_low_split_adjusted	m_close_split_adjusted	m_volume_split_adjusted	m_vwap_split_adjusted	m_open_dividend_and_split_adjusted	m_high_dividend_and_split_adjusted	m_low_di
3/27/2006	41.28	41.51	41.24	41.31	54405616		41.28	41.51	41.24	41.31	54405616	41.335	35.42	35.62	
3/28/2006	41.37	41.67	40.99	41.14	100534368		41.37	41.67	40.99	41.14	100534368	41.2925	35.51	35.76	
3/29/2006	41.2	42.1	41.17	41.92	130733504		41.2	42.1	41.17	41.92	130733504	41.5975	35.36	36.13	
3/30/2006	42	42.3	41.79	42.05	115943440		42	42.3	41.79	42.05	115943440	42.035	36.05	36.3	
3/31/2006	42.14	42.19	41.84	41.93	71090432		42.14	42.19	41.84	41.93	71090432	42.025	36.16	36.2	
4/3/2006	42.15	42.37	41.9	41.98	80733312		42.15	42.37	41.9	41.98	80733312	42.1	36.18	36.36	
4/4/2006	42.04	42.31	41.93	42.21	90039808		42.04	42.31	41.93	42.21	90039808	42.1225	36.08	36.32	
4/5/2006	42.3	42.67	42.21	42.61	91662256		42.3	42.67	42.21	42.61	91662256	42.4475	36.3	36.62	
4/6/2006	42.58	42.94	42.44	42.73	91713008		42.58	42.94	42.44	42.73	91713008	42.6725	36.54	36.85	
4/7/2006	42.9	43.05	42.3	42.32	116609376		42.9	43.05	42.3	42.32	116609376	42.6425	36.82	36.95	
4/10/2006	42.4	42.5	42.09	42.26	58988048		42.4	42.5	42.09	42.26	58988048	42.3125	36.39	36.48	
4/11/2006	42.4	42.43	41.71	41.92	91488624		42.4	42.43	41.71	41.92	91488624	42.115	36.39	36.42	
4/12/2006	41.95	42.08	41.8	41.93	58884688		41.95	42.08	41.8	41.93	58884688	41.94	36	36.11	
4/13/2006	41.9	42.33	41.79	42.1	63482400		41.9	42.33	41.79	42.1	63482400	42.03	35.96	36.33	
4/17/2006	42.04	42.21	41.39	41.66	104085120		42.04	42.21	41.39	41.66	104085120	41.825	36.08	36.22	
4/18/2006	41.79	42.56	41.77	42.46	124333376		41.79	42.56	41.77	42.46	124333376	42.145	35.86	36.53	
4/19/2006	42.54	42.65	42.3	42.65	79128320		42.54	42.65	42.3	42.65	79128320	42.535	36.51	36.6	
4/20/2006	42.62	42.82	42.38	42.5	81413008		42.62	42.82	42.38	42.5	81413008	42.58	36.57	36.74	
4/21/2006	42.67	42.86	41.82	42	130187008		42.67	42.86	41.82	42	130187008	42.3375	36.61	36.78	
4/24/2006	41.96	42.07	41.72	41.96	78556496		41.96	42.07	41.72	41.96	78556496	41.9275	36.01	36.1	
4/25/2006	42.08	42.1	41.65	41.87	77660816		42.08	42.1	41.65	41.87	77660816	41.925	36.11	36.13	
4/26/2006	41.87	42.06	41.7	41.86	82566688		41.87	42.06	41.7	41.86	82566688	41.8725	35.93	36.09	
4/27/2006	41.67	42.49	41.5	42.26	174122000		41.67	42.49	41.5	42.26	174122000	41.98	35.76	36.47	
4/28/2006	42	42.22	41.78	41.85	88883248		42	42.22	41.78	41.85	88883248	41.9625	36.05	36.24	
5/1/2006	41.93	42.3	41.31	41.44	92933680		41.93	42.3	41.31	41.44	92933680	41.745	35.98	36.3	
5/2/2006	41.61	41.73	41.4	41.55	85173568		41.61	41.73	41.4	41.55	85173568	41.5725	35.71	35.81	
5/3/2006	41.58	41.64	41.28	41.49	74982432		41.58	41.64	41.28	41.49	74982432	41.4975	35.69	35.74	
5/4/2006	41.53	41.99	41.53	41.83	88713504		41.53	41.99	41.53	41.83	88713504	41.72	35.64	36.04	
5/5/2006	42.11	42.21	41.94	42.16	80394368		42.11	42.21	41.94	42.16	80394368	42.105	36.14	36.22	
5/8/2006	42.08	42.28	42.07	42.2	45093328		42.08	42.28	42.07	42.2	45093328	42.1575	36.12	36.29	
5/9/2006	42.1	42.18	41.97	42.04	48566784		42.1	42.18	41.97	42.04	48566784	42.0725	36.13	36.2	
5/10/2006	41.93	41.98	41.55	41.67	92440496		41.93	41.98	41.55	41.67	92440496	41.7825	35.98	36.03	
5/11/2006	41.63	41.65	40.63	40.74	141027568		41.63	41.65	40.63	40.74	141027568	41.1625	35.72	35.74	
5/12/2006	40.57	40.76	39.96	40.19	153954432		40.57	40.76	39.96	40.19	153954432	40.37	34.82	34.98	
5/15/2006	40.03	40.34	39.79	40.16	127902496		40.03	40.34	39.79	40.16	127902496	40.08	34.36	34.62	
5/16/2006	40.16	40.26	39.77	39.86	132405504		40.16	40.26	39.77	39.86	132405504	40.0125	34.47	34.55	

Outlier Detection & Correction

Ensuring Data Integrity

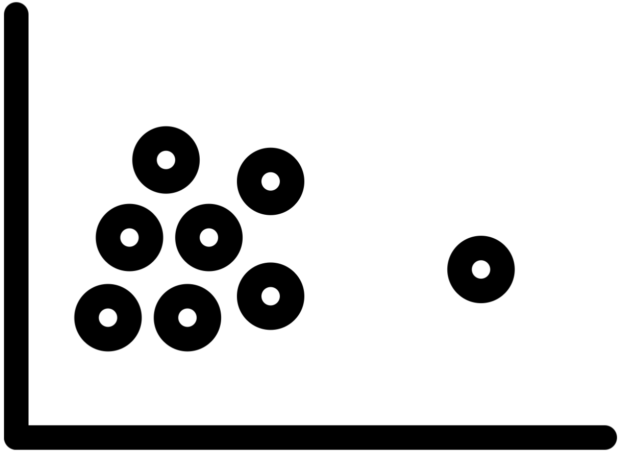
Why Outliers Destroy Signals

Outliers may come from:

- Data entry errors
- API glitches
- Stock splits not adjusted
- Extreme corporate events

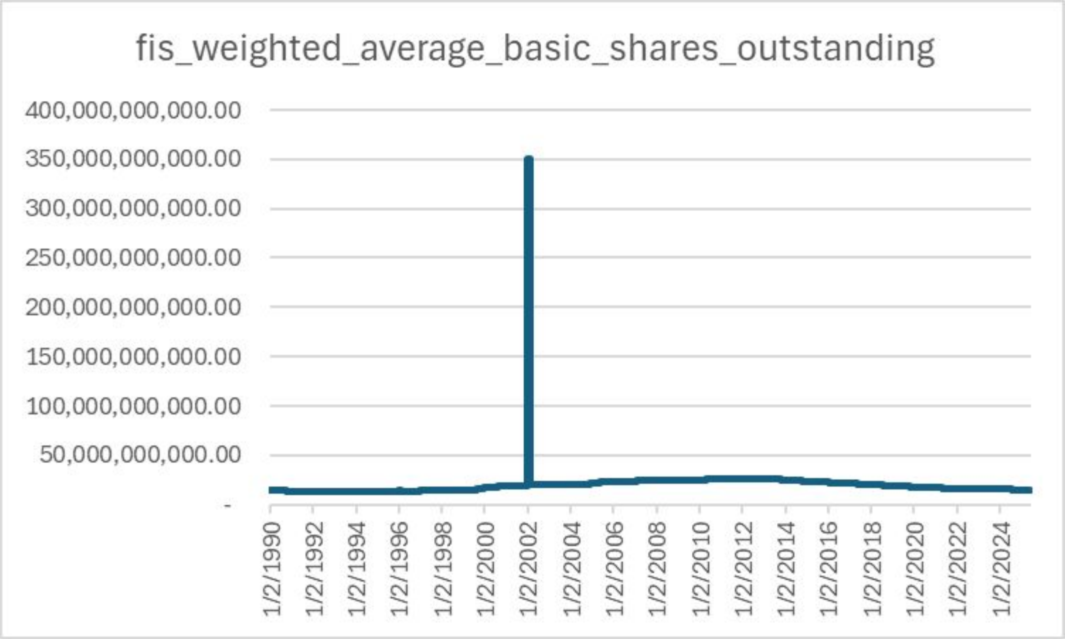
Consequences:

- Inflated z-scores
- False momentum signals
- Volatility distortion



Data Entry Errors

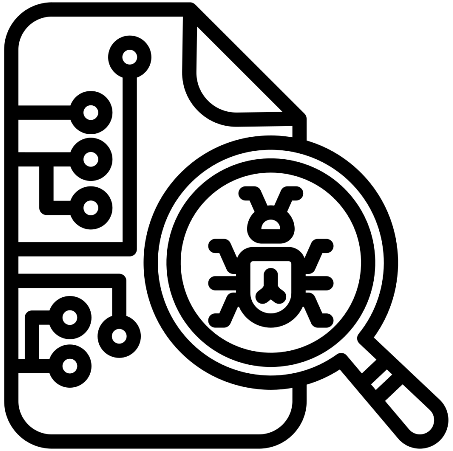
Data
Analyzer



Methods for Outlier Detection

Some ideas we can test:

- Rolling Z-Score Thresholds
- Median Absolute Deviation (MAD)
- Volatility-Scaled Return Filters
- Abnormal price gaps vs rolling volatility
- Structural break detection
- Stale value detection
- Near-zero denominator flags
- Corporate action validation (splits, dividends)
- Ratio sanity checks (negative or impossible values)
- Extreme growth without event justification
- Accounting restatement detection



Methods for Outlier Correction

We can implement:

- Z-score filtering
- Rolling median deviation
- Winsorization
- Robust scaling
- Percentile clipping

Important:

Outlier correction must be rule-based, reproducible, and grounded in financial logic (e.g., corporate actions and fundamental events).

Correction vs Removal

Never “manually delete”.

Choose:

- Winsorize
- Cap values
- Replace with rolling median
- Flag for exclusion
- …

The pipeline must explain itself.

Auditability > Convenience.

Outlier Detection and Correction

Data Curator enables **custom calculations** to diagnose, validate, and transform data into research-ready inputs.

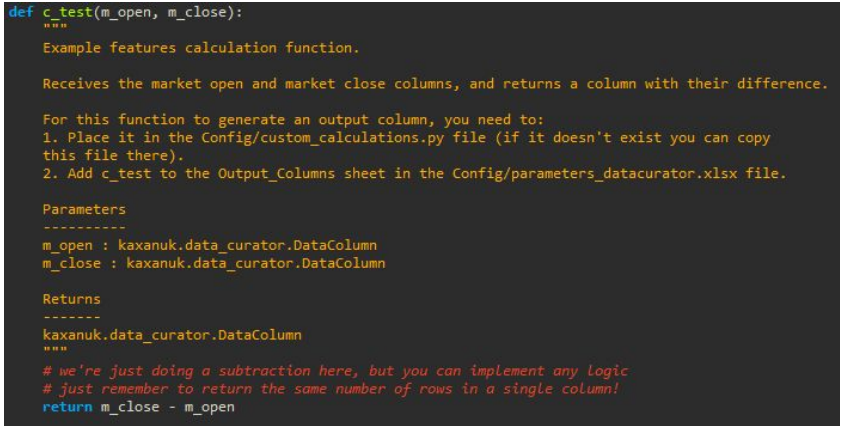
We define explicit rules.

Those rules:

- Detect anomalies
- Correct distortions
- Standardize scaling
- Preserve chronology
- Remain reproducible

Outlier handling becomes part of the research system, **not a notebook hack**.

Custom Calculations



```python
def c_test(m_open, m_close):
    """
    Example features calculation function.

    Receives the market open and market close columns, and returns a column with their difference.

    For this function to generate an output column, you need to:
    1. Place it in the Config/custom_calculations.py file (if it doesn't exist you can copy
    this file there).
    2. Add c_test to the Output_Columns sheet in the Config/parameters_datacurator.xlsx file.

    Parameters
    ----------
    m_open : kaxanuk.data_curator.DataColumn
    m_close : kaxanuk.data_curator.DataColumn

    Returns
    -------
    kaxanuk.data_curator.DataColumn
    """
    # we're just doing a subtraction here, but you can implement any logic
    # just remember to return the same number of rows in a single column!
    return m_close - m_open
```

https://kaxanuk-data-curator.readthedocs.io/en/stable/api_reference/custom_calcs.html

Some Examples

Outlier Adjusted Data

- **Price Spike Filter** — Flag returns exceeding a multiple of rolling volatility and cap or replace when no corporate action justifies the move.
- **Corporate Action Adjustment** — Detect split- or dividend-driven price jumps and adjust the historical series to preserve economic continuity.
- **Ratio Explosion Control** — Nullify or redesign ratios when denominators approach zero to prevent artificial signal distortions.
- **Abnormal Growth Check** — Identify extreme revenue or earnings changes without event support and winsorize or smooth accordingly.
- **Extreme Z-Score Cap** — Bound cross-sectional z-scores at predefined limits to stabilize rankings and reduce signal instability.
- **Missing Data Guardrail** — Apply exclusion or forward-fill only under predefined, financially justified rules.
- **Negative or Impossible Values Filter** — Flag and correct financially inconsistent observations (e.g., negative shares outstanding).
- **Stale Price Detection** — Detect abnormal price stagnation and classify as illiquidity or potential data error.

Feature Engineering

From Data to Alpha Signals and Risk Factors

What Is a Feature?

A **feature is a structured variable** derived from data that captures information relevant for **prediction** or **risk modeling**.

Hypothesis-Driven Feature Engineering

Avoid: Data mining 200 features and picking the best.

Instead: Start from hypothesis.

Example

“**Markets underreact to earnings surprises.**”

Then engineer:

- Standardized earnings surprise
- Post-event drift
- Volume confirmation

Feature engineering follows theory.

Types of Features

Univariate (1×N) Features

- Derived from each asset's own history.
- Examples: Trend, rolling volatility, growth rates, historical z-scores.

Data Curator

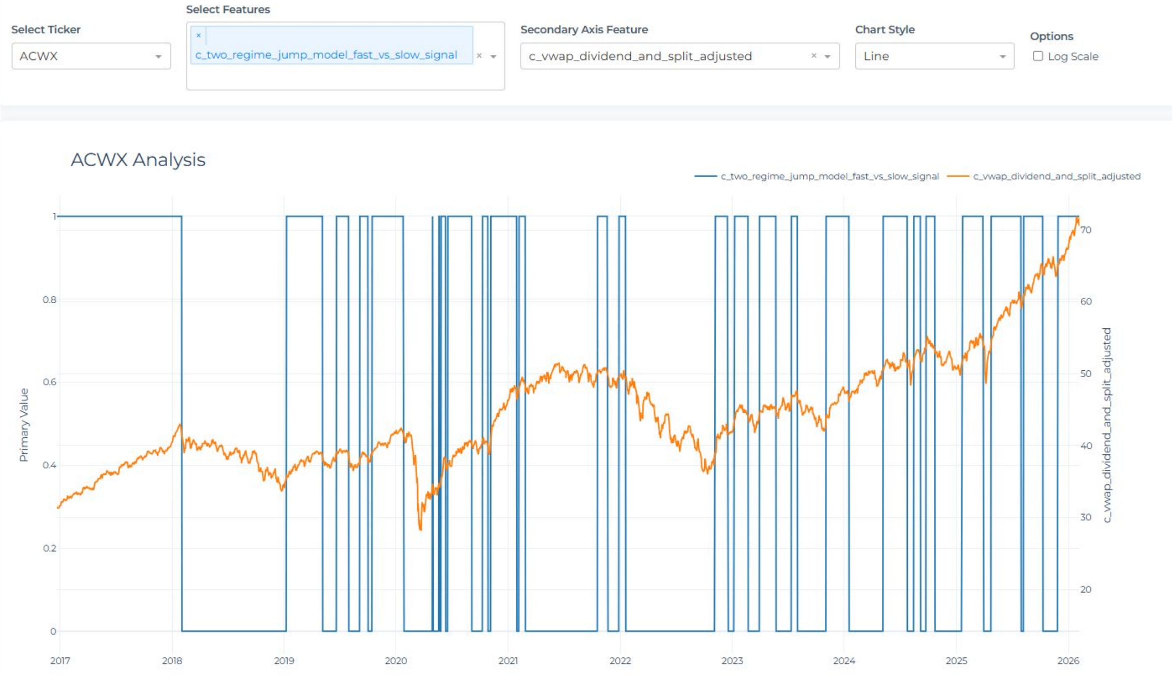
Multivariate (N×N) Features

- Relative measures across assets at a given time.
- Examples: Sector-neutral ranks, cross-sectional z-scores, relative strength.

Data Weaver

Feature Analysis

Data Analyzer



Select Ticker: ACWX

Select Features: c_two_regime_jump_model_fast_vs_slow_signal

Secondary Axis Feature: c_vwap_dividend_and_split_adjusted

Chart Style: Line

Options: ☐ Log Scale

ACWX Analysis
c_two_regime_jump_model_fast_vs_slow_signal
c_vwap_dividend_and_split_adjusted
Primary Value
1
0.8
0.6
0.4
0.2
0
70
60
50
40
30
20
c_vwap_dividend_and_split_adjusted
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026

Features Can Be Used as Alpha Signals or Risk Factors

Alpha Signals

Idiosyncratic predictive features designed to capture pricing inefficiencies and generate **excess returns**.

Risk Factors

Systematic drivers of returns that are common across assets and **explain risk premia**.

Some Examples

Alpha Signals

- **Earnings Surprise** — Abnormal returns following unexpected earnings results.
- **Short-Term Reversal** — Mean reversion after extreme short-term moves.
- **Residual Momentum** — Trend persistence unexplained by major risk factors.
- **Quality** — Excess returns linked to profitability and balance sheet strength.
- **Accruals** — Mispricing related to low earnings quality.
- **Sentiment** — Predictive signals derived from news or textual tone.
- **Dispersion** — Return opportunities arising from cross-sectional disagreement.
- **Event-Driven** — Temporary mispricing around corporate events (M&A, buybacks, guidance).

Risk Factors

- **Market** — Exposure to overall asset class return movements.
- **Sectors** — Exposure to industry-specific economic dynamics.
- **Beta** — Sensitivity of an asset’s returns to the market benchmark.
- **Size** — Exposure linked to company market capitalization.
- **Value** — Exposure to relatively undervalued companies.
- **Momentum** — Exposure to persistent price trends.
- **Volatility** — Exposure to return variability and risk level.
- **Growth** — Exposure to companies with strong fundamental expansion.

Let’s code!

Hands-On: Your First Data Pipeline

Disclaimers

The content of this document is strictly informative and does not constitute an offer or recommendation of KaxaNuk S.C. to buy, sell or subscribe any kind of securities, or to perform specific transactions. KaxaNuk S.C. is not responsible for the interpretation given to the information and /or content of this document. KaxaNuk S.C. does not accept or will accept any liability for losses or damages resulting from investment decisions that would have been based on this document. The persons responsible for the preparation of this content certify that the opinions stated reflect their own point of view and do not represent the view of KaxaNuk S.C. nor of its officials. This document is based on publicly available information which are considered reliable, however KaxaNuk S.C. makes no warranty regarding its accuracy or completeness.

Contact

research@kaxanuk.mx