When Mood Meets Market

A TIME SERIES ANALYSIS OF DEPRESSION, WEATHER, AND STOCK MARKET BEHAVIOR

APAN 5400

Group Project Proposal

Group 1: Ji Qiu, Wenda Zhang, Boa Kim, Liuyang Li, Ziyi Zhong

Columbia University

November 2025

01 BACKGROUND & BUSINESS USE CASE

Currently, US-based investment funds hold more than $50 trillion in assets under management. However, only approximately $440 billion, or less than 1% of the market capitalization, relies on models backed by behavioral or sentiment-based data. Behavioral and sentiment data are often overlooked in investment models despite existing literatures highlighting the connections between weather, mental health metrics, and stock market performances.

- Weathers and mental health conditions tend to influence investors' sentiment, resulting in changes in risk-taking behaviour.
- These changes in behaviour further influence their trading decisions, volume, and hence returns/volatility.

Our project transforms weather, sentiment, and market data into actionable insights for investors by uncovering behavioral patterns and environmental influences not visible in traditional financial analysis.

As the demand for more sophisticated model rises, we believe that more investment funds would adopt behavioral and sentiment based models in their valuation models.

$ 50 Trillion

U.S.-domiciled investment funds asset under management

$440 Billion

Asset under management for behavioral or sentiment-based funds.

Source: Chen, Y. (2024). *Extreme weather, investor mood, and stock market performance: Evidence from China. Sustainability, 16*(16), 6802.
Janzen, B. (2022). *Temperature and mental health: Evidence from helpline calls.* arXiv preprint arXiv:2207.04992.

02 METHODOLOGY & IMPACT

Bridging behavioral finance and data analytics to uncover the emotional and environmental drivers behind market trends.

Methodology

- Integrate multi-source datasets — combining financial (Yahoo Finance), weather (Open-Meteo), substance (OpenFDA), and mood data (CC News & Google Trends) into one consistent framework.
- Develop a time series model (2016–2022) to analyze how weather and depression indicators correlate with market movements.
- Apply Python-based and SQL-based data processing for cleaning, merging, and aligning variables across multiple data types.
- Quantify correlations and lagged effects between emotional and environmental factors and stock market volatility.

Impact

- Provides empirical evidence on how psychological and environmental conditions influence financial behavior.
- Helps investors and researchers recognize mood-dependent and weather-driven market patterns.
- Enhances forecasting accuracy by incorporating behavioral and environmental dimensions into quantitative finance.
- Contributes to behavioral finance research by expanding prior studies with broader data coverage and more variables.

Source: Tarczyński, W. (2021). The influence of investors' mood on the stock prices. Energies, 14(21), 7396. https://doi.org/10.3390/en14217396

03 DATA SOURCE & PROCUREMENT DETAILS

Data Source1 : Yahoo Finance *(mainly S&P 500 performance)*	**Data Source 2:Open Metro** *(mainly weather information)*	**Data Source 3: Open FDA** *(mainly substance data implying moods)*	**Data source 4: CC News & Google Pytrends** *(mainly depression text data of news)*
Data source Description： Yahoo Finance provides comprehensive financial market data including stock prices, indices and securities — ideal for analyzing market trends, volatility, and performance	**Data source Description：** An open-source weather API that provides historical and forecast data on temperature, precipitation—ideal for studying the relationship between weather patterns and stock market behaviour	**Data source Description:** OpenFDA provides publicly accessible data on drugs and is valuable for identifying mental health trends	**Data source Description:** Use Python to scrape historical text data from CC News, a large-scale English news dataset, and extract articles containing the keyword “depression.” We also use Google Pytrends to get the data on depression text data in the US.
Available Attributes: • Past S&P 500 performance: Date;Open, High, Low, Close,Volume, Return, and 7-Day Volatility	**Available Attributes:** • Daily temperature: maximum, minimum and mean daily temperature of all American States; Date • Rainfall: rainfall of all American states;Date	**Available Attributes:** • Substance abuse : Fentanyl, oxycodone, methadone and alprazolam abuses; Date	**Available Attributes：** • CC News Depression Text: Date, title, text, domain

Format:all data are csv files

Year Range:2016/01/01 to 2022/12/31(all data)

Data limitation:

• All datasets are static CSV files without live API connections, which limits real-time updates or automatic data refresh.

• The coverage and time range depend on the original source export, so new records or recent data may not be included.

ETL strategy

Step	Tool	Purpose
extract	csv/JSON	Manual or automated downloads
Transform	Python	Clean, normalize and join datasets
Load(raw)	MongoDB	Store original structured records
Load(Cleaned)	Postgre SQL	Create analytics-ready relational schema

04 RATIONALE FOR SELECTED TECHNOLOGIES

API Layer — Python

- Python is used to connect and retrieve data from multiple public APIs, including financial market, weather, and sentiment sources.
- Using libraries such as yfinance, pytrends, and requests, the project collects raw data efficiently without building a separate Flask server.
- This layer focuses on secure and direct data ingestion, minimizing infrastructure complexity.

Back-End Data Store — PostgreSQL, Mongo DB

- PostgreSQL provides great support for complex queries, joins, and aggregations, making it ideal for relational and structured data analysis.
- MongoDB offers a schema-flexible document store that efficiently handles varied and nested content such as parsed documents and extracted metadata,
  enabling a hybrid storage architecture that supports both structured and unstructured data sources.

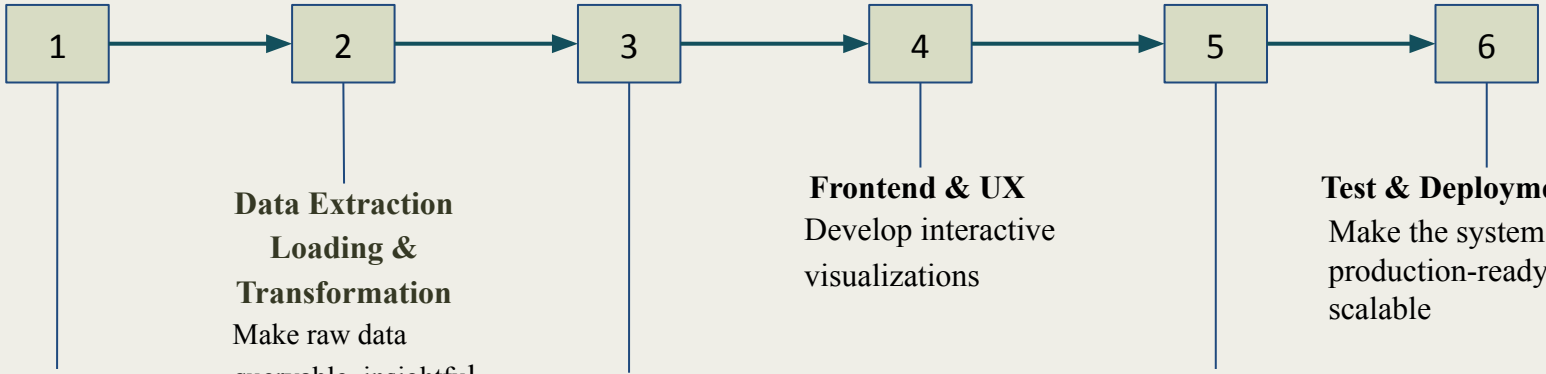
Data Ingestion & Processing - Python & Apache Spark

- This layer performs data transformation, feature engineering, and model-based analysis on the collected datasets.
- Python (pandas, scikit-learn, pyod) handles most of the data cleaning.
- For scalability, Apache Spark can be integrated to process large-scale or time-series data, such as historical prices, weather logs, or sentiment text.
- This architecture enables flexible expansion from local prototypes to distributed analytics environments.

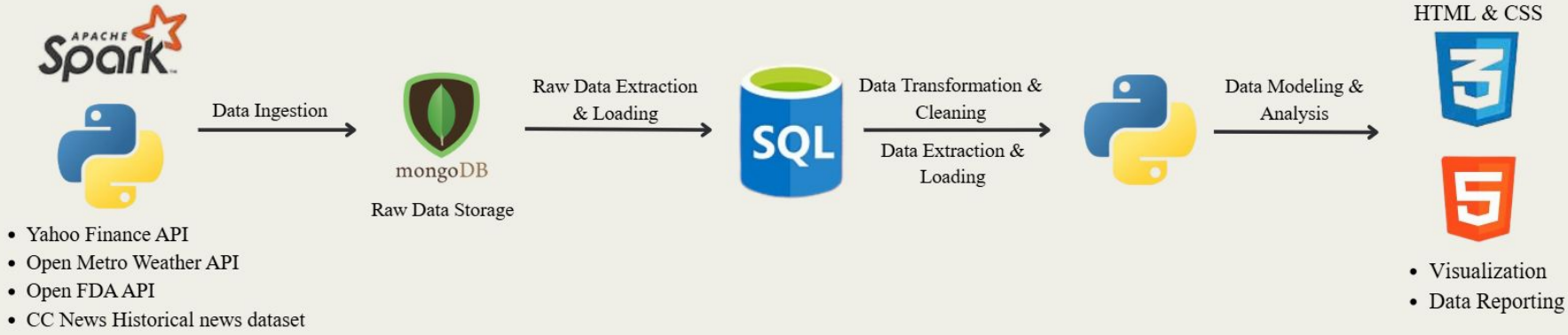
Front-End - HTML & CSS, Streamlit

- The front-end interface is built primarily with Streamlit, providing an interactive dashboard for visualizing model performance.
- HTML and CSS are used to enhance the layout and styling of Streamlit components, enabling a more customized and polished user experience.
- This simple web-based design allows users to access results directly through any browser without additional setup.

05 IMPLEMENT PLAN



06 DATA FLOW



Data Collection Layer → **Data Storage Layer** → **Application Layer**

07 TEAM MEMBER PLANNED CONTRIBUTION

Based on our agreement, we have decided to concentrate on our respective tasks to integrate all components of the data platform and achieve our common goal. This will ensure a seamless alignment of the overall data flow.

Ji Qiu	Wenda Zhang	Boa Kim	Liuyang Li	Ziyi Zhong
Proposal Writing & Data Source Searching	*Data Processing & System Preparation*	*Database & Data Pipeline Development (Leader)*	*Technical Design & Rationale*	*Presentation Design & Content Integration*
• Drafted and refined the proposal content based on team discussions and dataset findings. • Designed the Data Source & Procurement slides using materials provided by Boa and Wenda. • Assisted with Boa and Wenda in collecting data sources and processing codes for the project's data analysis.	• Worked on data extraction and cleaning, transforming weather and sentiment datasets into unified formats. • Data validation and technical preparation for pipeline testing. • Proposed and explored UI design ideas (Flask or Streamlit) from the user perspective.	• Collected datasets from Yahoo Finance, Google Trends, and Open-Meteo for integration. • Built the SQL database, creating and joining tables for time-series analysis. • Developed the Python–SQL data connection pipeline, ensuring accurate and efficient data flow.	• Led the Design Choices & Technology Rationale section explaining system architecture and tool comparison. • Assisted in slide revision and refinement, integrating technical details with clear visuals. • En logical flow between data and design sections.	• Structured the overall presentation flow, connecting each section smoothly from background to conclusion. • Created and refined the Background & Business Use Case, Team Roles, and final summary slides. • Ensured visual consistency and narrative coherence by integrating all members' work into a unified deck.

08 SCALABILITY

Data Infrastructure & Cloud Resources

Current Architecture (MVP Stage)

- **Data Source:** Multiple CSV datasets (weather, stock market, and sentiment data), all downloaded manually from public portals.
- **Process:** Manual or semi-automated ETL pipeline using Python scripts; data updated periodically instead of real-time ingestion.
- **Storage:** MongoDB (raw data) and PostgreSQL (cleaned analytical data) hosted via AWS RDS / MongoDB Atlas Free Tier.

Expected Cost

Step	Tool / Platform	Est. Cost (month)
CSV Data Storage	CSV Download	Free
MongoDB (Raw Data)	MongoDB Atlas (Free / Shared Tier)	Free – $15
PostgreSQL(cleaned data)	AWS RDS / GCP Cloud SQL	$25 – $40
Cloud Compute (ETL Scripts) and Storage	RDS , EC2, S3	Free

Future Scalability Plan

- **Data Source Expansion / Automated Integration**
  Upgrade to automated data collection pipelines using APIs (e.g., Open-Meteo, Yahoo Finance, or Google Trends) for frequent and scalable ingestion. This transition will enable near real-time updates and reduce manual maintenance.
- **Database & Architecture Scalability**
  Implement database partitioning and indexing in PostgreSQL for faster analytical queries, and enable MongoDB sharding to handle large raw datasets efficiently. Both databases can be scaled independently, depending on data volume growth.
- **Cloud Infrastructure Scalability**
  Deploy the ETL workflow and databases on scalable AWS services (EC2,RDS, S3) or equivalent GCP infrastructure. Cloud auto-scaling can be enabled to accommodate higher data volumes, concurrent users, and heavier analytical workloads.

Thank you!

APAN 5400

Group Project Proposal

Group One: Ji Qiu, Wenda Zhang, Boa Kim, Liuyang Li, Ziyi Zhong

Columbia University

November 2025