Factor-Based Stock Recommendation Tool

DS 5110

Team Members: Chen Yang, Amseen Shaik

1 Project Kickoff

1.1 Project Scope

The scope of this project focuses on building a web-based application of data pipeline and analytics platform that integrates ETL workflows, financial data processing, and multi-factor modeling for stock portfolio management.

- Compute and visualize factor-based scores (momentum, value, volatility).
- Construct and backtest portfolios with rebalancing and evaluation.
- Support advanced analytics through customizable SQL queries.
- Future work: natural-language strategy definition via LLM.

1.2 Expected Outcomes

1. Functional Streamlit prototype for stock exploration and backtesting.
2. Multi-factor selection model with ranking, weighting, and portfolio construction.

1.3 Milestones and Timeline

- Week 1: Database schema design and data ingestion setup
- Week 2: ETL pipeline implementation and data validation
- Week 3–4: Factor computation and initial backtesting framework
- Week 5: Streamlit integration and analytical visualization
- Week 6: Final integration, reproducibility validation, and presentation

2 Team Roles

Chen Yang is responsible for the core data infrastructure, backtesting framework, and user interface.

- **Data (OHLCV layer):** Designing and maintaining the DuckDB schema, collecting and validating OHLCV data (open, high, low, close, volume) for S&P companies through the IBKR API, and ensuring data consistency for backtests.
- **Backtesting Engine:** Building the backtest framework to support multi-factor portfolio strategies, periodic rebalancing, and performance evaluation metrics.
- **User Interface:** Developing the Streamlit application that enables interactive factor selection, visualization, and portfolio analytics.

Ameen Shaik focuses on complementary data enrichment and integration.

- **Financial Indicators:** Scraping and preprocessing company-level fundamentals and ratios (e.g., P/E, P/B, earnings growth) using sources such as Yahoo Finance (`yfinance`).
- **Corporate Actions:** Collecting and adjusting for events such as stock splits, dividends, and mergers to ensure accurate adjusted-price series and factor computations.
- **Data Support:** Collaborating on the ingestion pipeline to align financial indicators and corporate actions with OHLCV data for seamless integration into the factor model.

3 Data Ingestion and Processing Workflow

3.0.1 Overview

- **Sources:** Yahoo Finance, IBKR, optional Kaggle/Alpha Vantage.
- **Universe:** S&P500 constituents in the past 8 years.
- **Storage:** Parquet snapshots + DuckDB tables.
- **Modes:** Batch (historical backfill), Incremental (daily refresh).
- **Reproducibility:** Immutable raw snapshots, run_id, logging.

3.0.2 Workflow

1. Extract: Pull OHLCV + fundamentals, store raw Parquet.
2. Validate: Schema + quality checks.
3. Transform: Normalize, compute returns, standardize ratios.
4. Load: Upsert into prices/fundamentals/securities.
5. Factors: Momentum, value, volatility computation.
6. Predictions: ML-based expected return estimation from factor features. (Optional)
7. Reporting: Expose to Streamlit UI.
8. Logging: Metadata for reproducibility.

4 Skills and Tools

Python, SQL/DuckDB, pandas/numpy, matplotlib/plotly, Streamlit, vectorbt, GitHub, Overleaf, Excel.

5 Initial Setup

- Progress Tracker: Google Sheets Link
- GitHub: `https://github.com/cyang0227/ds5110.git`