Gnosiom - an open source end-to-end solution to factor investment research in US stocks

Yuhuang Chen

Feb 26 2025

Disclaimer

- This is not investment advice—just an open-source research tool for educational purposes.
- Wharton Research Data Services (WRDS) was used in preparing this analysis. The data and tools provided constitute valuable intellectual property and trade secrets of WRDS and/or its third-party suppliers.
- To protect proprietary information and comply with data usage agreements, in this slide, all figures have been aggregated and anonymized. No raw or sensitive data is revealed in these slides, and the displayed results are intended solely for illustrative and educational purposes.

What an idealized backtest engine should be

- High-quality cleaned data with daily resolution
- No survivorship bias, corporate actions such as delisting, dividend, split, spin-off should be handled
- Being able to do the universe stock screening, i.e. not being constrained as fixed several symbols
- Raw data can be downloaded for further large-scale investigation

CRSP and Compustat

- Data quality requirement satisfied
- In most literature, only the monthly data are used for academic research
- In reality the freedom to do daily rebalancing is needed

Gnosiom: What it is, What it can do

- An open-source investment research engine based on CRSP and Compustat data
- Support two common tasks in factor investment: Factor Evaluation and Backtest
- Factor Evaluation: focus on the predictability of future returns and the trading costs is neglected
- Backtest: given a good factor, display the P&L of top ranking stocks portfolio with all trading costs (tax, slippage, commission fee) considered.
- GitHub link: https://github.com/Quaizz/Backtest

Workflow

To run this engine, here is the workflow, before we get the results we need to load and preprocess the data in the first few steps

Step 1: Data Import. We start by running a notebook (e.g. create dsf v2.ipynb) that pulls raw financial data from an online source (using the WRDS Python API) and saves it into our local database (DuckDB).

Step 2: Define Your Factor. Next, you open a Python file (FactorCalculator.py) where you write a custom SQL query. This query calculates a factor (e.g. ROE, PB ratio) that will be used later in the analysis.

Workflow

Step 3: Calculate Daily Factor Data. Then, you run another notebook (create factor parquet.ipynb) that applies your custom factor calculation to generate daily data. For example, if your factor is ROE, you' ll end up with a daily snapshot of ROE values across different companies. (the file path is like /factor data/roe/roe-1988-01-04.parquet) For each day and each factor you will get one parquet file.

Step 4: Factor Evaluation. After that, you analyze the computed factor by grouping the data (using factor analysis.ipynb). You can find more sample results from the slides

Workflow

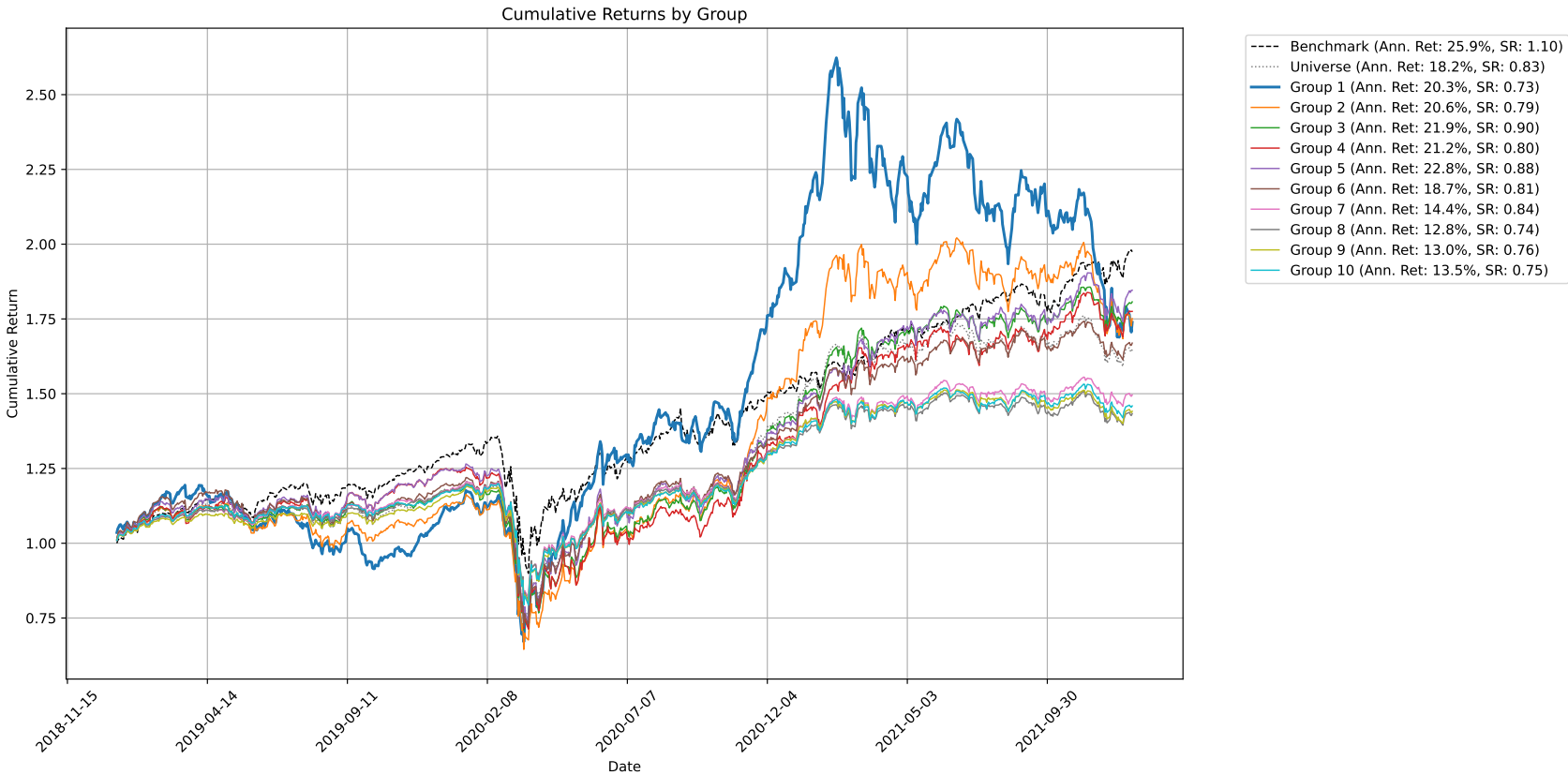
Step 5: Backtest the Trading Strategy Finally, you run a notebook (backtest portfolio.ipynb) that simulates a trading strategy based on your factor. This backtesting step includes realistic details like taxes, commissions, slippage, and other market factors. You can also find more results from the slides

For the rest of the slide, you will see the figures that illustrate the functionality of step 4 and step 5.

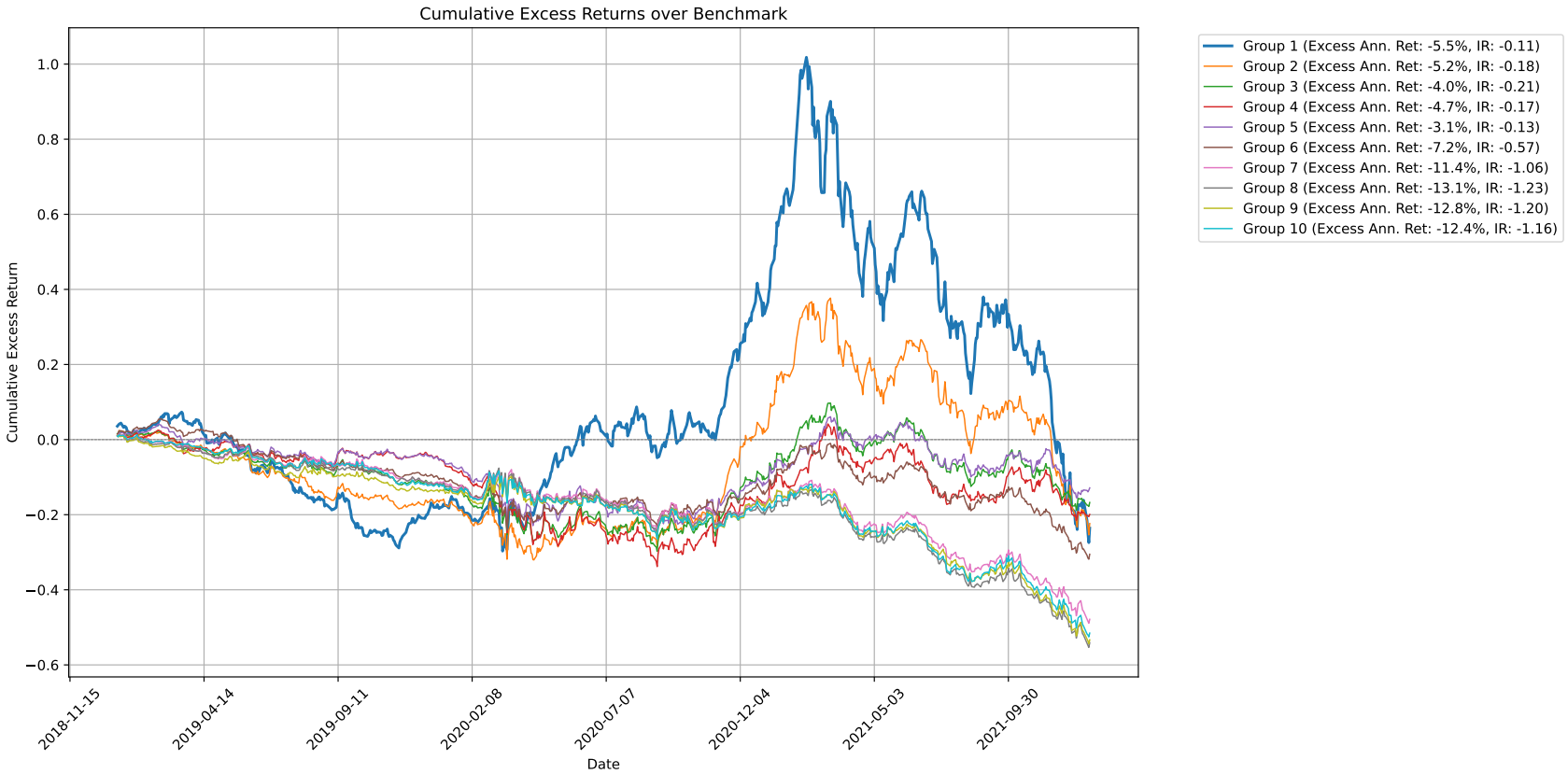
Factor Evaluation

- Separate the investment universe into groups based on the percentile of factor values
- Examine the the relationship between factor value and return

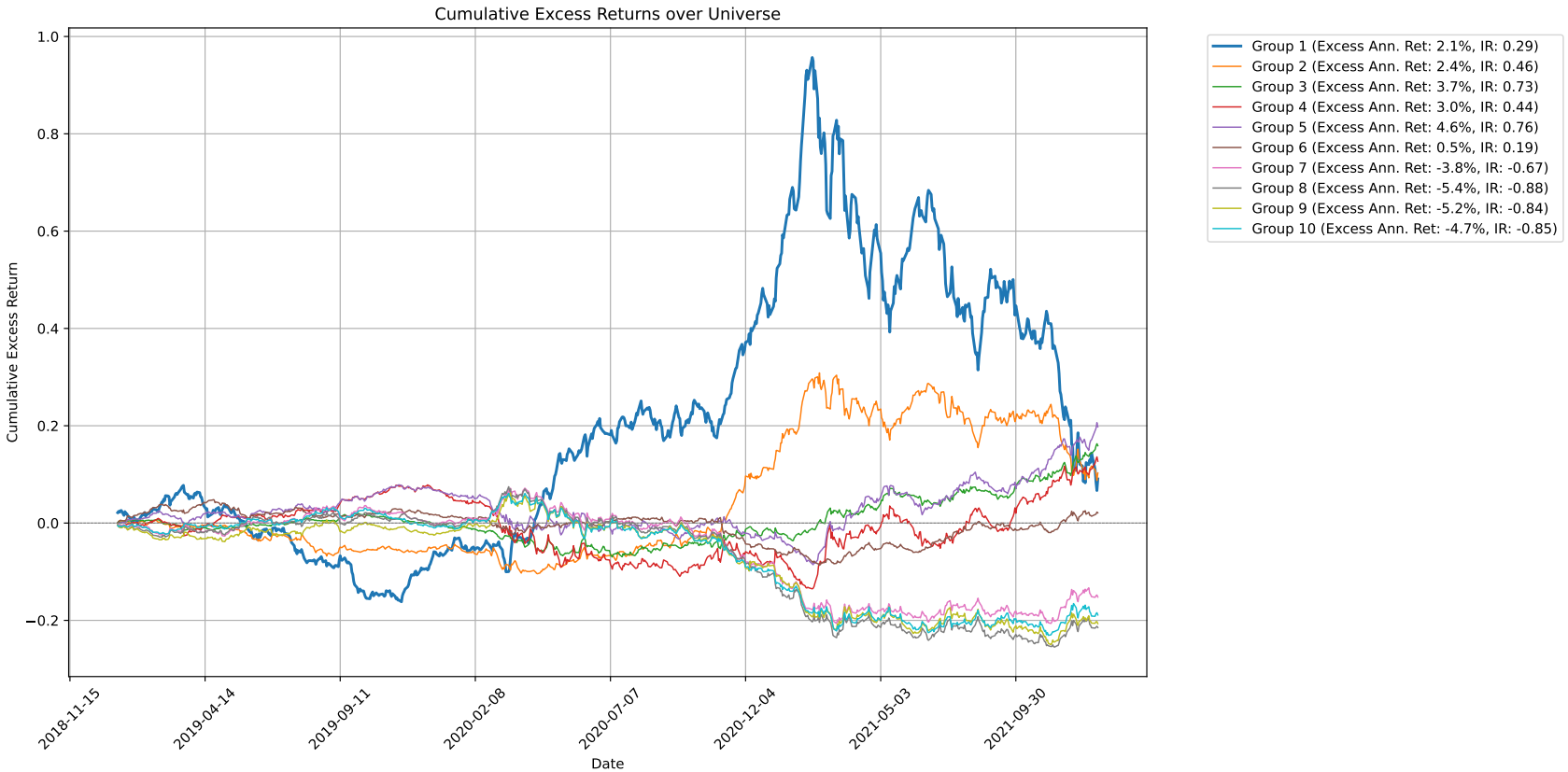
Visualization Tools: Return By Groups



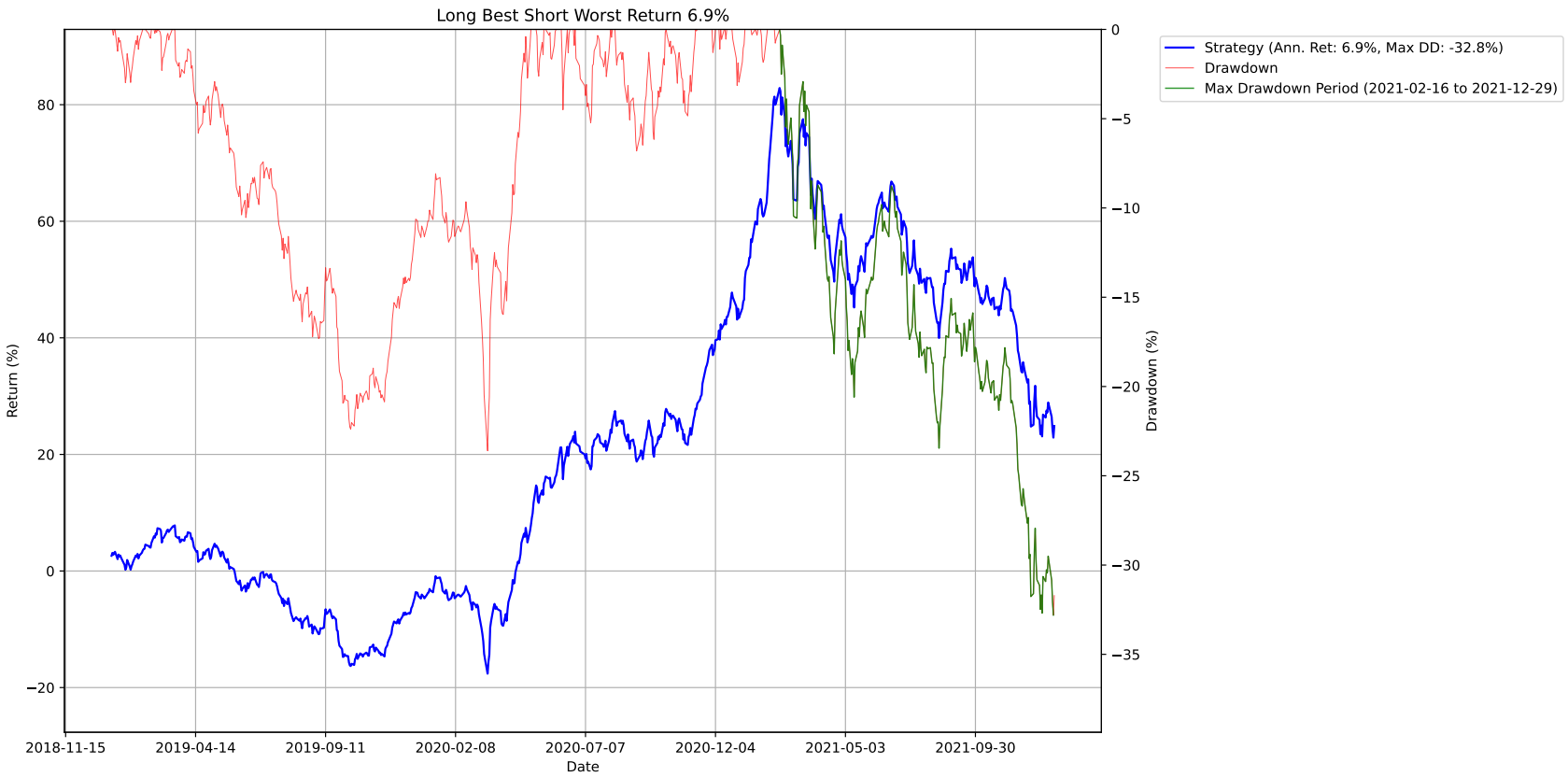
Visualization Tools: Excess Return over Benchmark



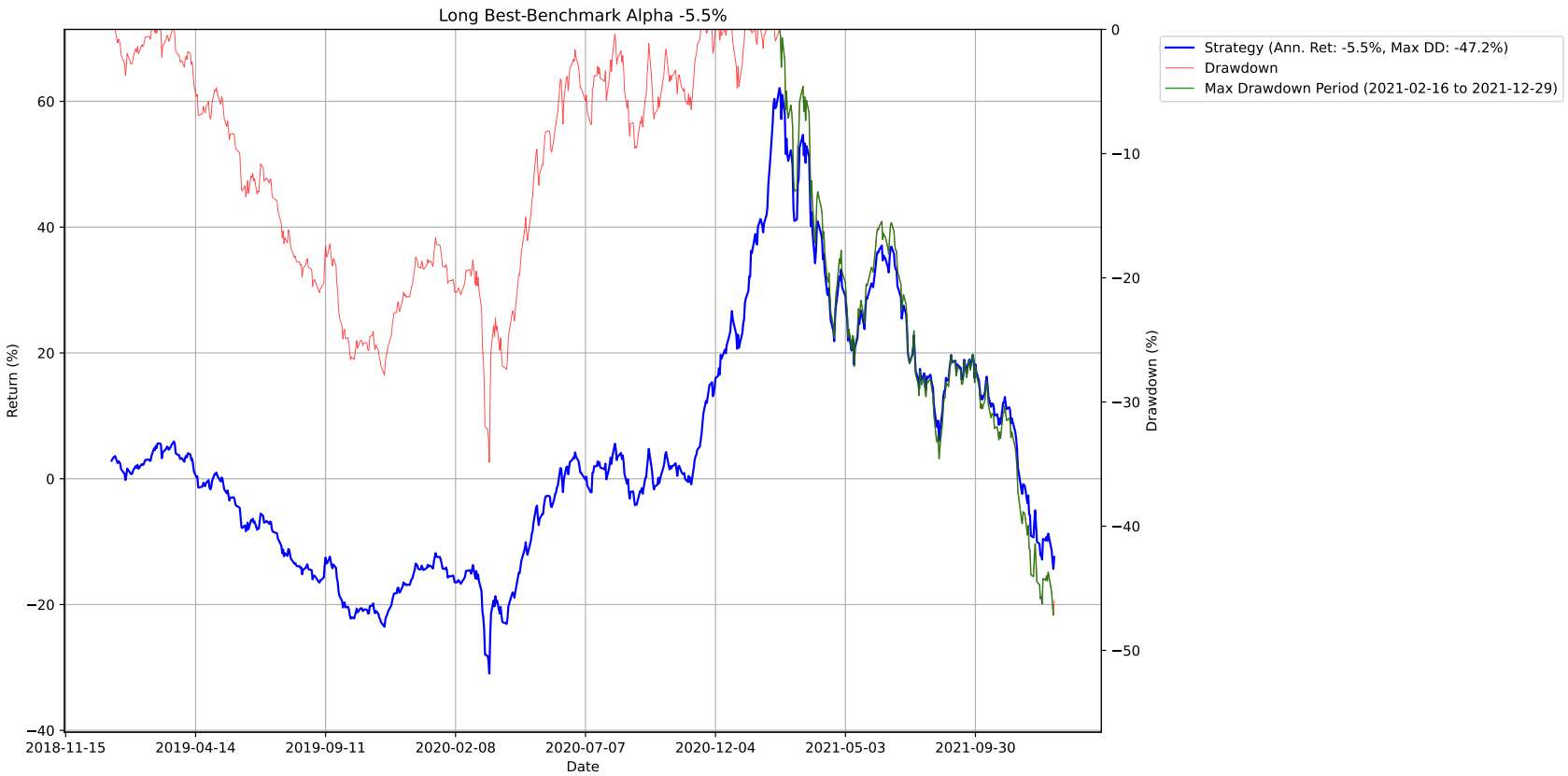
Visualization Tools: Excess Return over universe



Visualization Tools: Performance of Long Short Portfolio



Visualization Tools: Performance of Long Only Over Benchmark



Visualization Tools: Stats of Long Short Portfolio

Long-Short Monthly Statistics

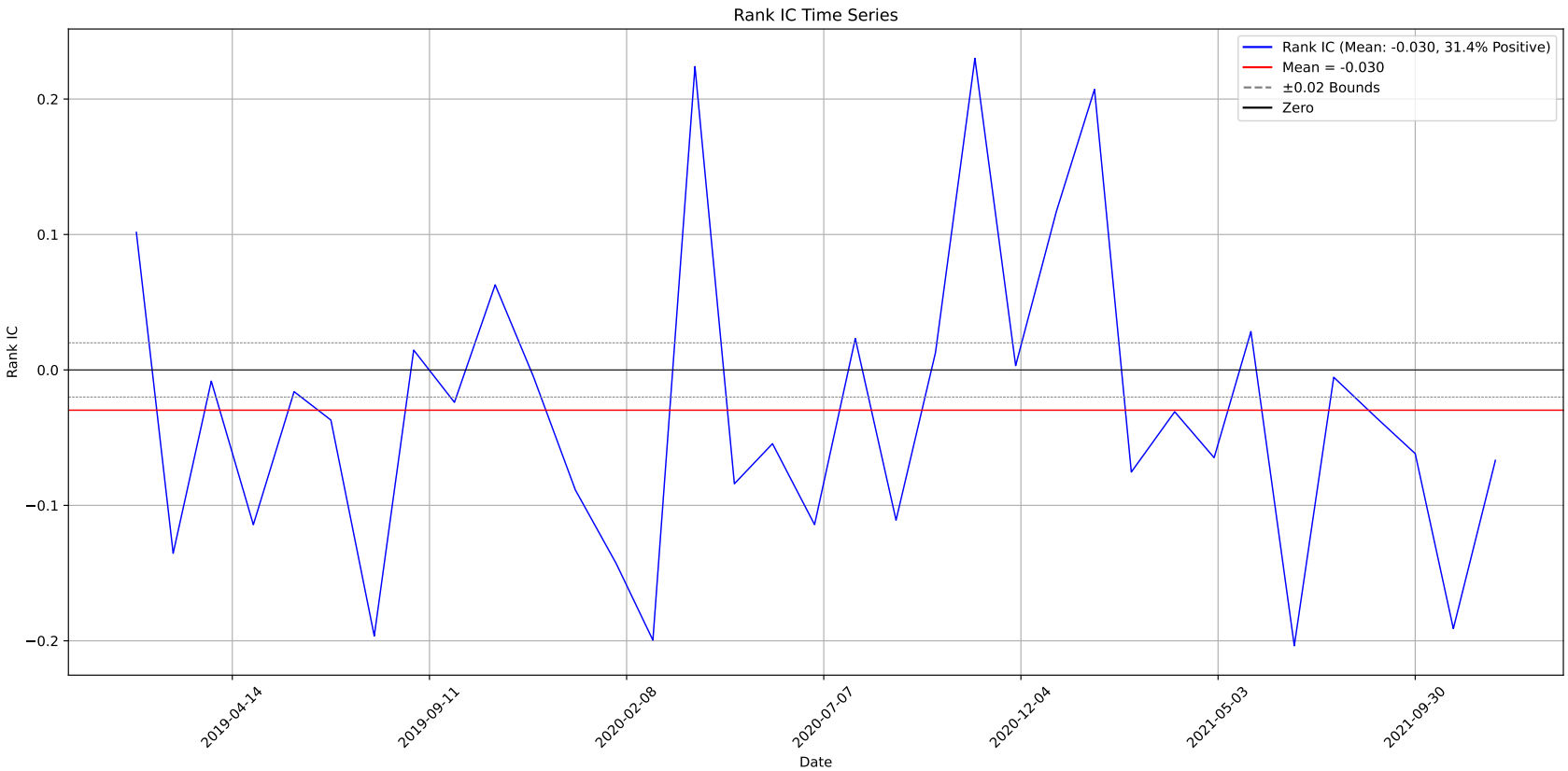
Year	Jan%	Feb%	Mar%	Apr%	May%	Jun%	Jul%	Aug%	Sep%	Oct%	Nov%	Dec%	Total%	YMDD%	Start_date	End_date	Month_odds%
2019	1.16	5.03	-0.8	-2.61	-4.24	1.34	-5.07	-4.5	-4.09	-1.31	6.86	5.06	-4.04	-22.39	20190321	20191010	41.67
2020	-1.02	-0.37	-1.59	14.33	7.32	6.22	-0.63	1.6	0.04	-0.53	13.84	3.29	49.44	-16.89	20200116	20200316	58.33
2021	12.32	6.24	-3.41	-4.1	-1.72	5.04	-8.67	2.52	-1.89	-3.62	-8.28	-6.02	-12.92	-32.8	20210216	20211229	33.33

Visualization Tools: Stats of Long Only Over Benchmark

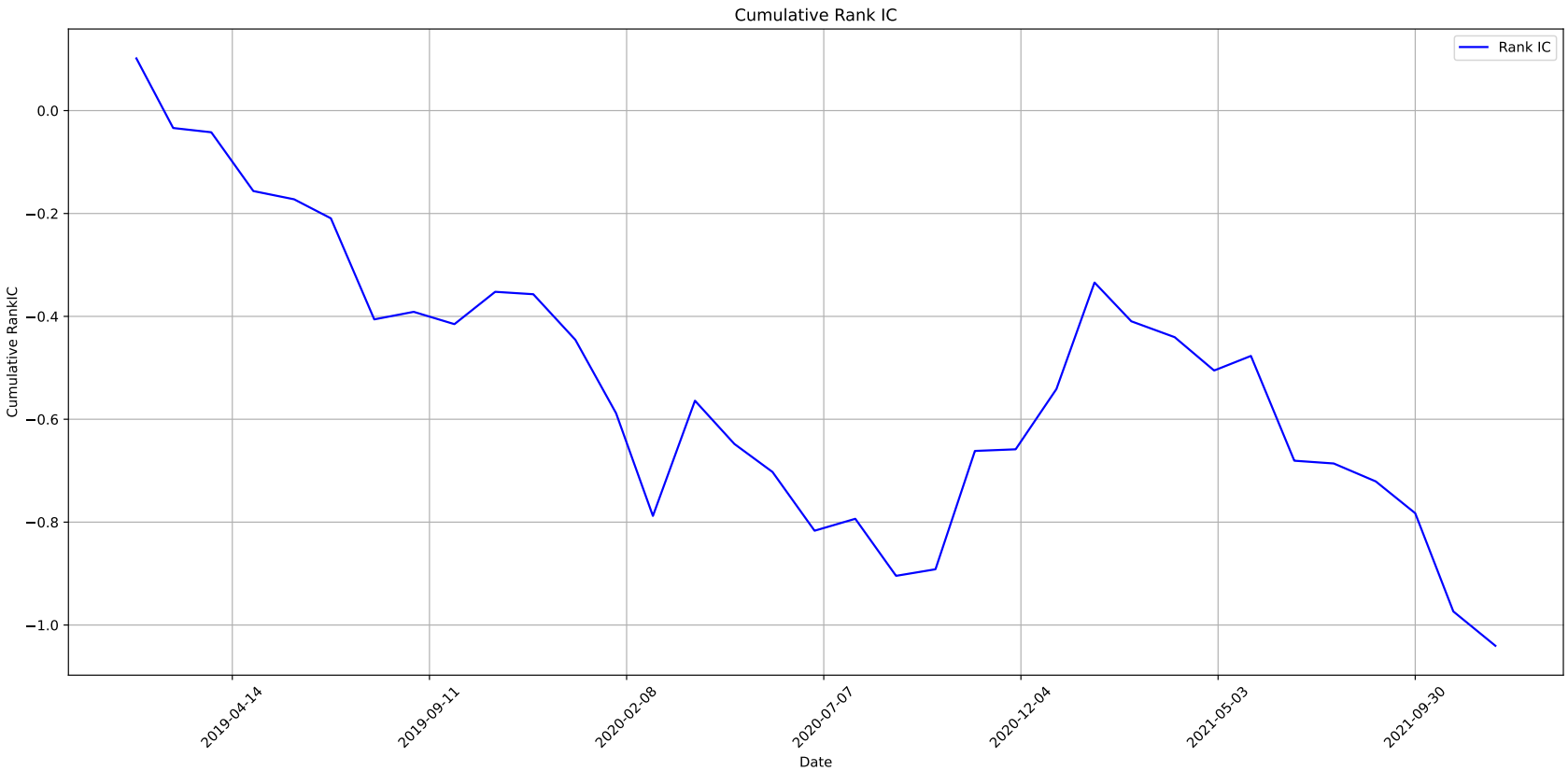
Long-Universe Monthly Statistics

Year	Jan%	Feb%	Mar%	Apr%	May%	Jun%	Jul%	Aug%	Sep%	Oct%	Nov%	Dec%	Total%	YMDD%	Start_date	End_date	Month_odds%
2019	0.74	3.46	0.89	-2.85	-2.94	0.21	-4.48	-2.5	-4.63	-1.39	4.78	4.59	-4.63	-19.7	20190321	20191010	50.0
2020	0.24	1.04	0.75	10.14	6.2	4.21	-0.04	0.8	0.24	-1.88	10.3	2.03	38.71	-9.86	20200116	20200316	83.33
2021	10.06	3.57	-4.58	-3.74	-1.91	4.05	-8.29	2.5	-1.27	-4.05	-7.13	-6.33	-17.23	-32.82	20210216	20211229	33.33

Visualization Tools: Rank IC Time series



Visualization Tools: Additive Rank IC



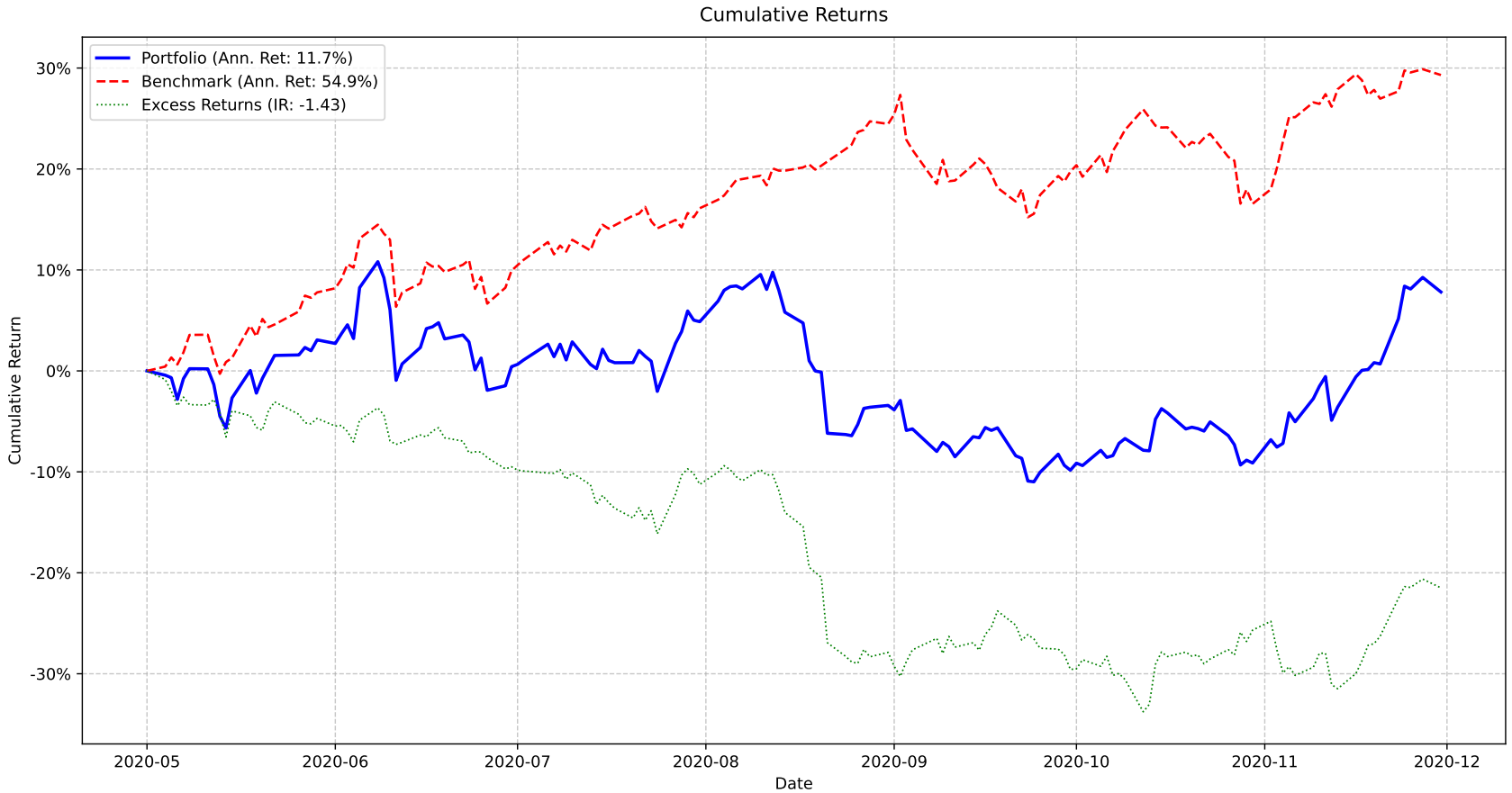
Backtest

- Support fixed investment targets (e.g. buy and hold AAPL and MSFT) and constructing portfolio by buying top ranking stops according to factors
- Each day a detailed trading log and portfolio snapshot is provided.

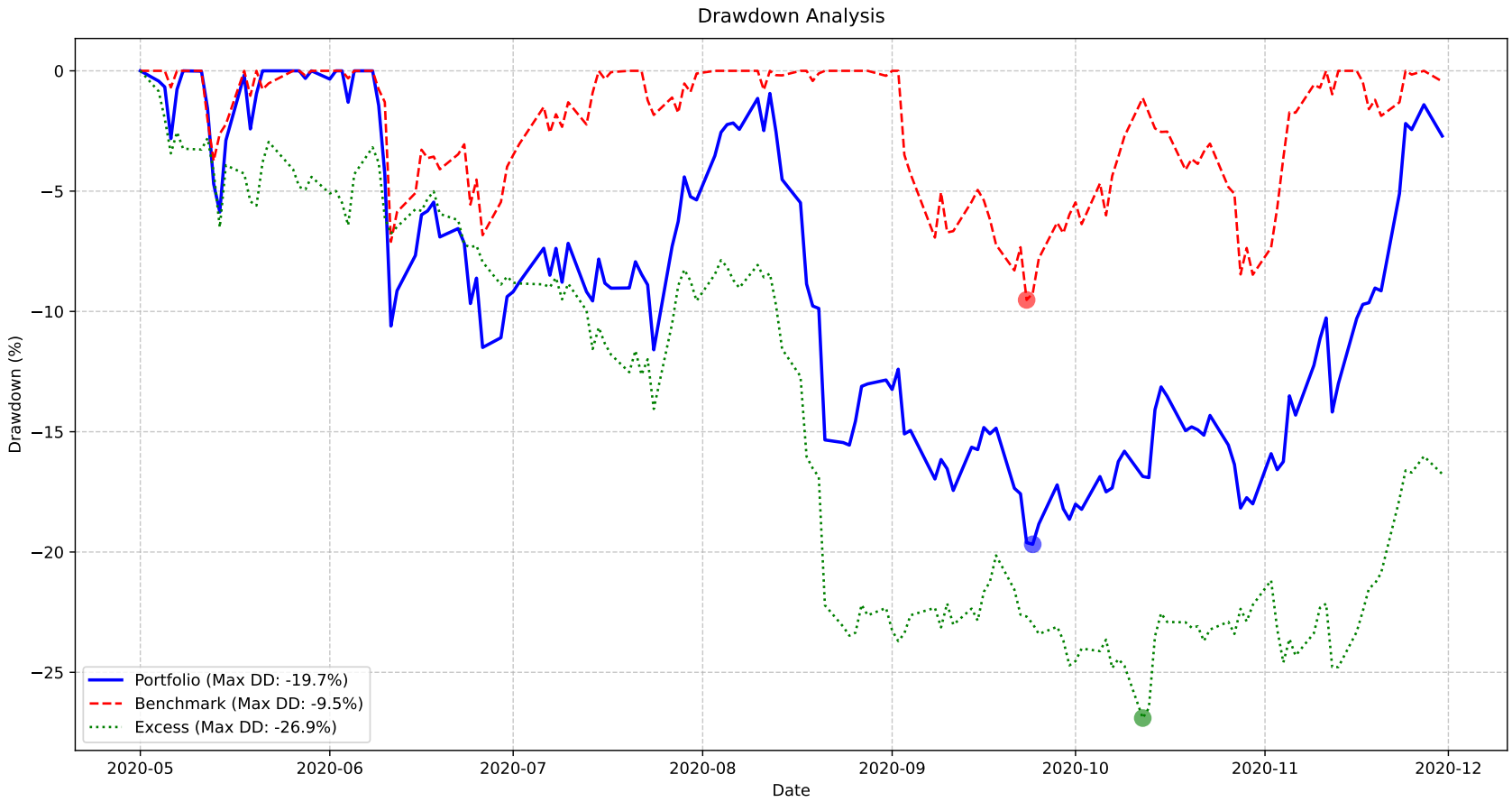
Visualization Tools: Stats of trading strategy

Total Return (%)	Annualized Return (%)	Sharpe Ratio	Information Ratio	Max Drawdown (%)	Benchmark Ann. Return (%)	Std of Excess Return (%)	Avg Daily Turnover (%)	Number of Trades
6.70	11.67	0.53	-1.43	-19.68	54.90	21.47	4.43	66

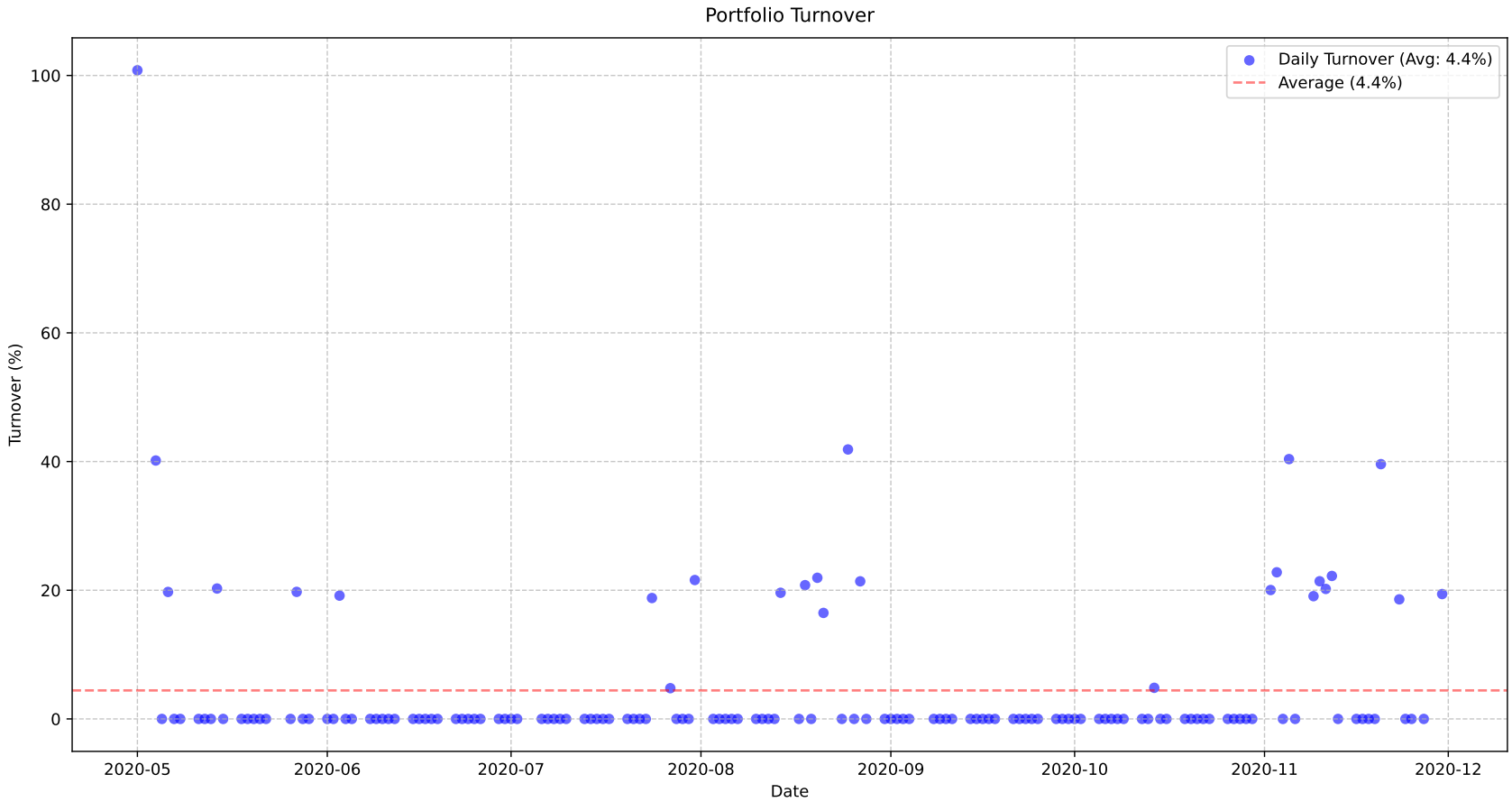
Visualization Tools: Return of trading strategy



Visualization Tools: Drawdown of trading strategy



Visualization Tools: Turnover of trading strategy



Sample Trading Activity log

For illustrative purposes only, data is simulated

Type	Symbol	Shares	Price	Commission	Margin
Sell	123456	300	3.50	2.0	1,200
Sell	234567	250	12.45	1.5	2,300
Sell	345678	450	4.25	2.5	950
Sell	456789	5,000	0.85	20.0	2,500
Sell	567890	150	7.80	1.0	500
Sell	678901	100	18.00	1.0	0
Buy	789012	5,000	2.00	15.0	1,000
Buy	890123	120	30.25	2.0	3,500
Buy	901234	700	4.75	3.0	2,200

Sample Position Snapshot log

For illustrative purposes only, data is simulated

Symbol	Status	Cost	Shares	Current Price	Change (%)	Position Value	Position (%)
112233	ACTIVE	150.50	7,000	145.25	-3.42	1,016,750.00	11.50
223344	ACTIVE	10.15	55,000	10.35	1.97	570,250.00	6.45
334455	ACTIVE	9.25	42,500	12.00	29.73	510,000.00	6.00
445566	ACTIVE	0.75	400,000	0.70	-6.67	280,000.00	3.20
556677	SUSPENDED	7.80	70,500	6.50	-16.67	458,250.00	5.00
667788	ACTIVE	30.25	10,200	27.75	-8.19	282,750.00	3.45
778899	ACTIVE	80.00	3,000	85.00	6.25	255,000.00	3.10
889900	ACTIVE	70.50	3,750	72.00	2.12	270,000.00	3.05

Sample Corporate Action Processing (Simulated Event)

```
Processing corporate action for XYZ111:
Distribution Flag: CS
Found 2 distribution events
Processing events in sequence:
Processing event sequence number: 1
Processing Distribution Event for XYZ111:
Event Type: SD
Detail Type: SDIV
Cash Distribution:
Amount per share: $1.5200
Tax Rate: 38.0%
Gross Amount: $230,000.00
Net Amount: $142,600.00
Processing payment immediately
Processing event sequence number: 2
Processing Distribution Event for XYZ111:
Event Type: FRS
Detail Type: STKSPL
Stock Split:
Split Factor: 0.130000
```

Sample Corporate Action Processing (Simulated Event)

```
Processing Split for XYZ111:
Before Split:
Shares: 150,000
Cost Basis: $2.00
Current Price: $2.10
After Split:
Shares: 19,500
Cost Basis: $15.38
Current Price: $16.15
Updated last valid price from $2.10 to $16.15
```

Disclaimer

- This is not investment advice—just an open-source research tool for educational purposes.
- Wharton Research Data Services (WRDS) was used in preparing this analysis. The data and tools provided constitute valuable intellectual property and trade secrets of WRDS and/or its third-party suppliers.
- To protect proprietary information and comply with data usage agreements, all figures have been aggregated and anonymized. No raw or sensitive data is revealed in these slides, and the displayed results are intended solely for illustrative and educational purposes.