Project Report

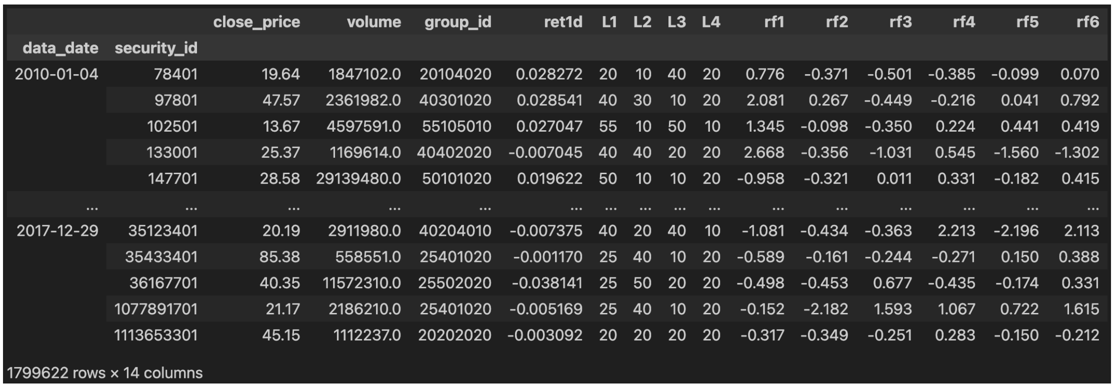
Siqing Guan

Part a: Description of Intuitions and Reasonings

Data Processing and Preliminary Analysis

The initial phase of our approach involved meticulous data cleaning to ensure a robust foundation for analysis. We began by importing security data into a Python environment, converting it to a DataFrame, and indexing it by date. To enhance clarity and facilitate further operations, we dropped unusable data and reindexed the table using ‘security_id’ as a secondary layer.

Our next step was the treatment of the 'group_id' column where we partitioned the data across the levels L1 to L4, segregating them accordingly. Similarly, for risk factors, we applied an equivalent structuring method, reordering the DataFrame by date and ‘security_id’. The culmination of this phase was the merging of these processed datasets into a consolidated frame for comprehensive analysis.

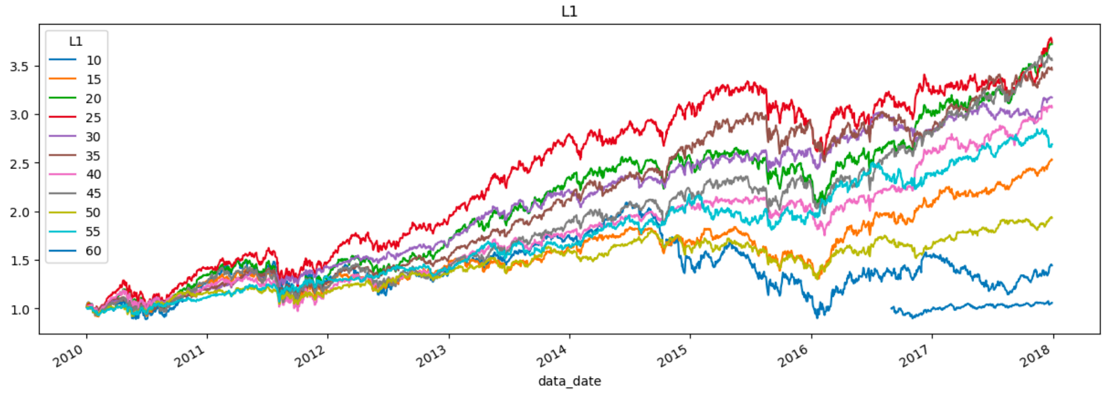


		close_price	volume	group_id	ret1d	L1	L2	L3	L4	rf1	rf2	rf3	rf4	rf5	rf6
data_date	security_id														
2010-01-04	78401	19.64	1847102.0	20104020	0.028272	20	10	40	20	0.776	-0.371	-0.501	-0.385	-0.099	0.070
	97801	47.57	2361982.0	40301020	0.028541	40	30	10	20	2.081	0.267	-0.449	-0.216	0.041	0.792
	102501	13.67	4597591.0	55105010	0.027047	55	10	50	10	1.345	-0.098	-0.350	0.224	0.441	0.419
	133001	25.37	1169614.0	40402020	-0.007045	40	40	20	20	2.668	-0.356	-1.031	0.545	-1.560	-1.302
	147701	28.58	29139480.0	50101020	0.019622	50	10	10	20	-0.958	-0.321	0.011	0.331	-0.182	0.415
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2017-12-29	35123401	20.19	2911980.0	40204010	-0.007375	40	20	40	10	-1.081	-0.434	-0.363	2.213	-2.196	2.113
	35433401	85.38	558551.0	25401020	-0.001170	25	40	10	20	-0.589	-0.161	-0.244	-0.271	0.150	0.388
	36167701	40.35	11572310.0	25502020	-0.038141	25	50	20	20	-0.498	-0.453	0.677	-0.435	-0.174	0.331
	1077891701	21.17	2186210.0	25401020	-0.005169	25	40	10	20	-0.152	-2.182	1.593	1.067	0.722	1.615
	1113653301	45.15	1112237.0	20202020	-0.003092	20	20	20	20	-0.317	-0.349	-0.251	0.283	-0.150	-0.212

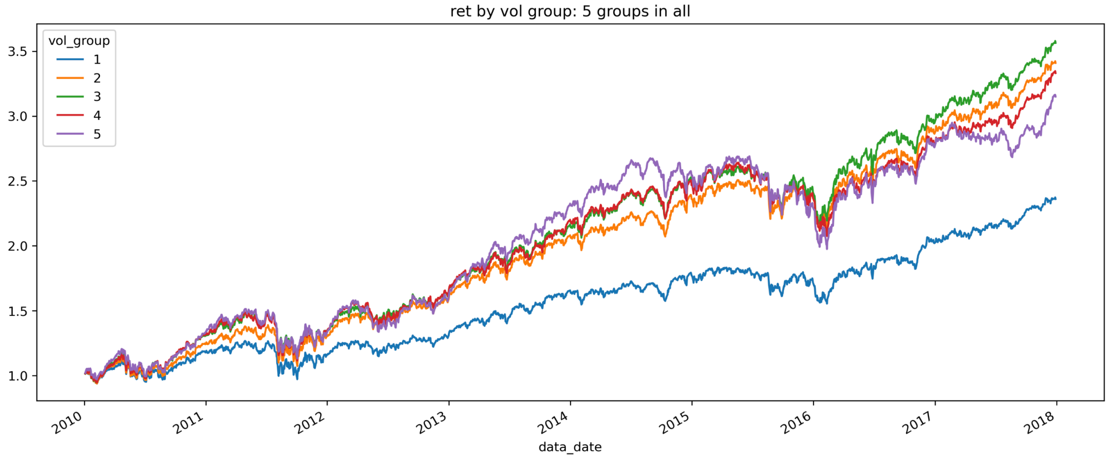
1799622 rows × 14 columns

Data Analysis and Strategic Decision-Making

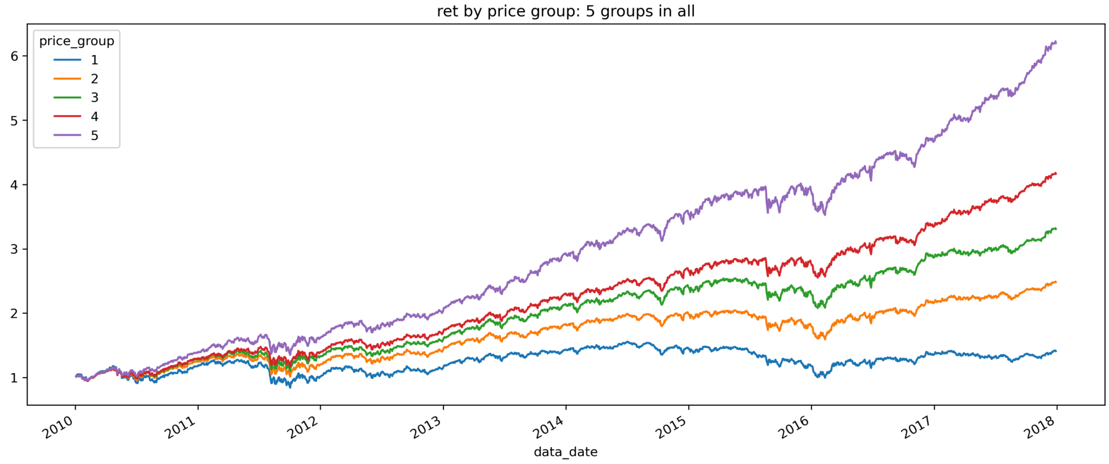
Our analytical strategy leaned towards a selection-based arbitrage approach. We conducted an in-depth evaluation of 'group_id', discerning the largest performance disparities at level L1. Within this level, group 'L1=25' exhibited more stable performance, earmarking it as our long position of choice.



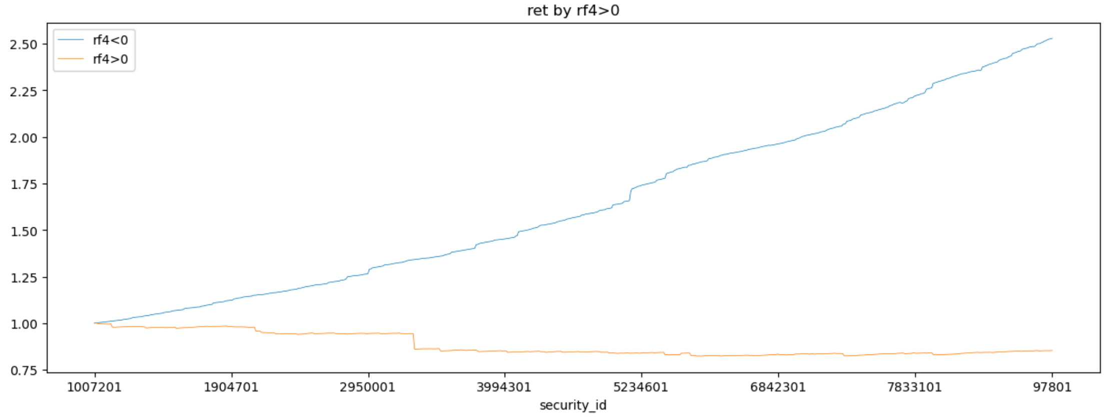
The volume analysis followed, where securities were categorized into groups based on transaction volume. Through iterative comparison across 2 to 6 group segments, we observed a stark underperformance in one segment when partitioned into quintiles. Consequently, we resolved to trade within the top 80% volume segment.

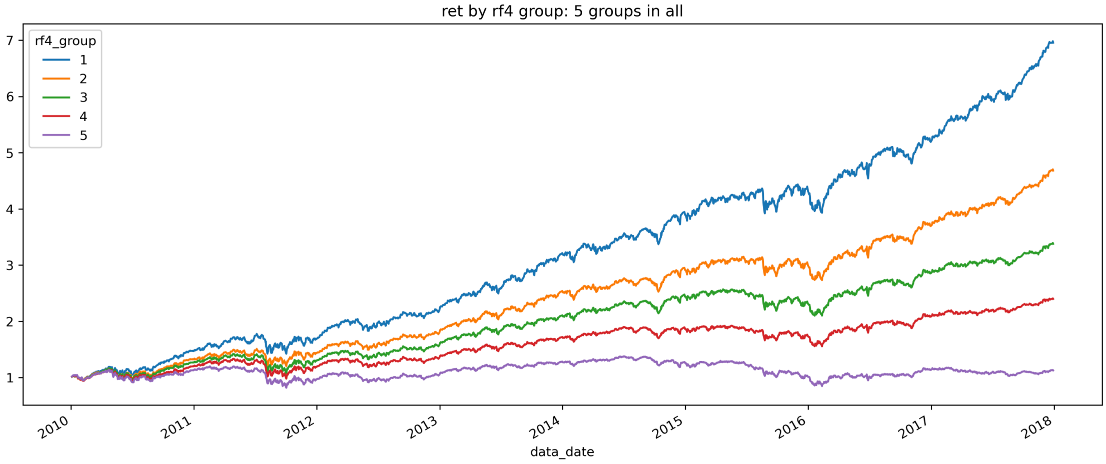


The 'close_price' analysis mirrored the previous steps, leading us to a strategy that involves long only securities in the upper 20% of price while considering short positions for those in the lower 20th percentile.



Finally, the risk factor analysis, albeit less straightforward due to unknown variable specifics, was approached by plotting performances based on their polarity. We found 'rf4' to exhibit the greatest differences, indicating its potential as a primary focus. Further achieved by replicating the volume and price analyses for 'rf4' indicated an optimal strategy to long the bottom 20% and short the top 20% in this risk factor.





Part b: Plot and Analysis

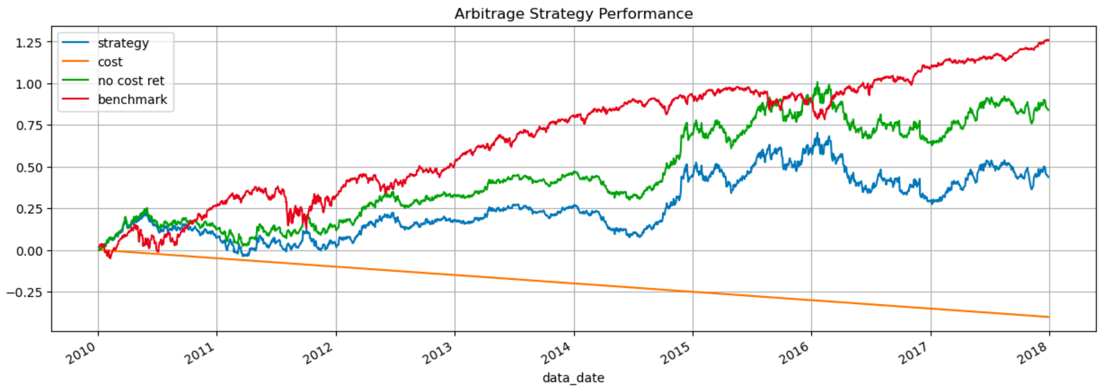
Portfolio Strategy and Performance Analysis

The essence of our investment approach was encapsulated by a multi-faceted strategy, which was as follows:

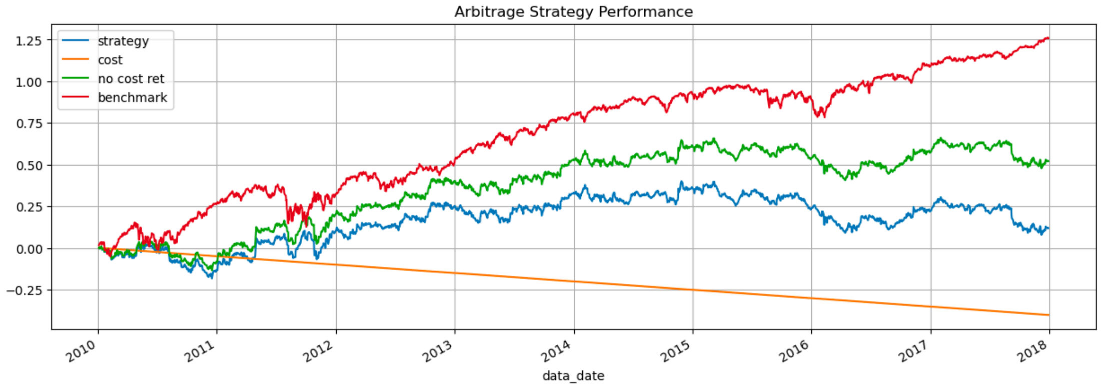
1. Risk Factor (rf4): long positions in the bottom 20% and short positions in the top 20% based on rf4 values.
2. Price Strategy: long positions in the top 20% and short positions in the bottom 20% based on 'close_price'.
3. Volume Strategy: My focus was on securities with a trading volume in the top 80% based on the volume values.
4. Group Strategy: long positions exclusively in group 25 (L1).

To evaluate and compare the efficacy of various strategies, we created a function, stg_plot, which allowed us to visualize the strategy's performance and compute associated metrics more effective.

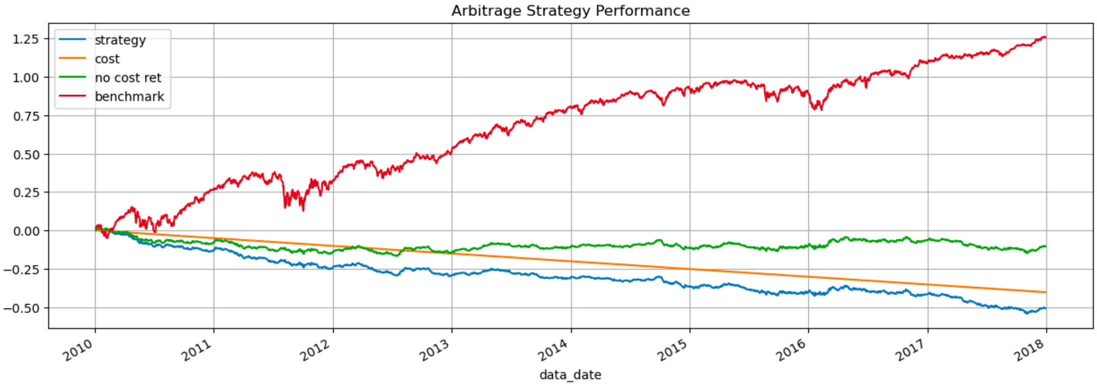
strategy 1: group arbitrage, long group 25, short group 10



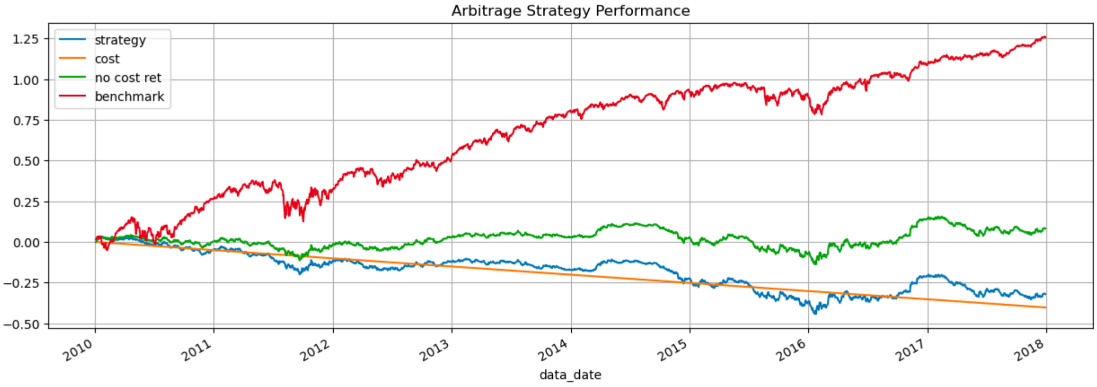
strategy 2: group arbitrage, long group 25, short group 10 (L4)



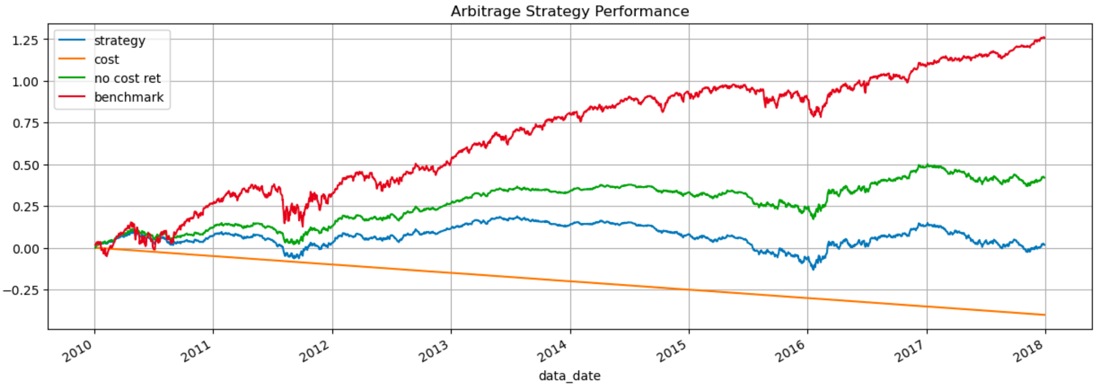
strategy 3: volume strategy



strategy 4: rf4 strategy



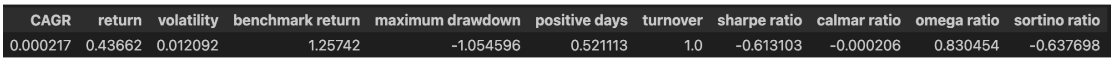
strategy 5: price strategy



Our comparative analysis across five distinct strategies revealed that 'strategy 1,' which involves a group arbitrage with long positions in group 25 and short positions in group 10, yielded consistently positive returns. This arbitrage was unique in its ability to produce stable gains in contrast to other tested strategies.

Risk Management and Statistical Computation

For comprehensive risk assessment and performance quantification, we developed a function, stg_stats, and applied it to 'strategy 1'. This utility function facilitated the computation of a suite of performance statistics including:
CAGR, Return, Volatility, Benchmark Return, Maximum Drawdown, Positive Days(%), Turnover, Sharpe Ratio, Calmar Ratio, Omega Ratio, Sortino Ratio



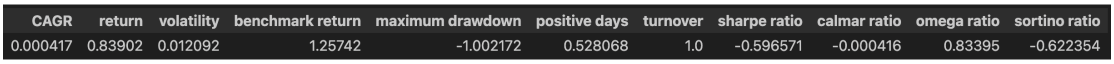
CAGR	return	volatility	benchmark return	maximum drawdown	positive days	turnover	sharpe ratio	calmar ratio	omega ratio	sortino ratio
0.000217	0.43662	0.012092	1.25742	-1.054596	0.521113	1.0	-0.613103	-0.000206	0.830454	-0.637698

A plot of the portfolio's cumulative return was constructed, illustrating the growth of the investment value over time based on daily arithmetic returns. This plot is an essential visual tool in understanding the trajectory and fluctuations in the portfolio's value.

Through this analytical lens, we have been able to illustrate not only the raw performance metrics of our chosen strategy but also how it weathers different market conditions, thereby informing future strategic decisions and potential adjustments.

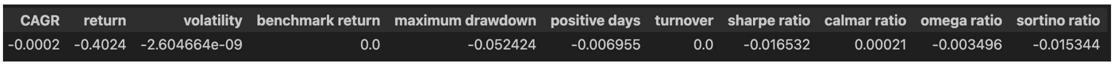
Part c: Impact of Trading Cost and Rebalance Frequency

We compared various statistical measures of the portfolio under two scenarios: one accounting for trading costs and the other disregarding them. The results are as follows: The results without trading cost:



CAGR	return	volatility	benchmark return	maximum drawdown	positive days	turnover	sharpe ratio	calmar ratio	omega ratio	sortino ratio
0.000417	0.83902	0.012092	1.25742	-1.002172	0.528068	1.0	-0.596571	-0.000416	0.83395	-0.622354

The differences between these two conditions:



CAGR	return	volatility	benchmark return	maximum drawdown	positive days	turnover	sharpe ratio	calmar ratio	omega ratio	sortino ratio
-0.0002	-0.4024	-2.604664e-09	0.0	-0.052424	-0.006955	0.0	-0.016532	0.00021	-0.003496	-0.015344

We can see relatively big difference in returns. If we consider the trading cost, the return will drop 40% of the one without taking it into consideration.

Part d: Future Potential Improvement and Concerns

In the U.S. stock market, which is a semi-efficient market, zero-cost arbitrage strategies are virtually ineffective. If we have higher initial capital, our strategy may work better.