FIN6392.0W1
Financial Technology and Blockchain

Group 5

Jessica Chen, Tamo Natarajan, Jane Song,

Kort Suter, and Kristine Wilson

Project 1:

Big Data Alpha Model

Introduction

The big data alpha model is a financial tool that can be used to strategically gain insight into market behavior. Meta data has become increasingly important over recent decades, with collection happening at almost every point in our lives. This process of data creation, storage, and retrieval is growing rapidly, as it has been estimated that at least 1.7Mb of data is created every second for every person on earth (Ahmad 2018). In the finance sector, a large segment of data is leveraged by investment institutions and fintech organizations to predict price movements in securities, bonds, exchanges, and more. These predictions can be formed and fine-tuned with the assistance of a big data alpha model, which can help identify important factors that help explain variation in stock prices. Other, less complex (and data intensive) models exist as well like the capital asset pricing model and the Fama-French model. In this paper our group will explore all three models using R programming to fetch data and construct data frames for factor selection and implementation.

Data Selection and Retrieval

The first step of creating a big data alpha model is to identify and construct factors that can be used for analysis. While done at early stages of the model, it is crucial to pick relevant and meaningful factors that are robust over time and where data can be easily retrieved in appropriate increments like daily, weekly, or monthly to provide the most insight into price movements as possible. MSCI research states that there is strong academic proof that equity portfolio performance and its variations can largely be explained through the use of factors (Bender et al. 2013).

Using R programming to retrieve data from different repositories, our group collected, scrubbed, and organized data from the Federal Reserve, U.S. government, corporate balance sheets, technical indicators, and social media. In total, we analyzed 26 factors that we deemed legitimate and substantial enough to contribute to equity price volatility. The factors tested include: U.S. GDP, CPI, ICS, CFNAI, retail sales, INDPRO, bond yield, PPIACO, OIL/BRENT, D/E ratio, ROA, EPS, PE ratio, Fama-French factors, SMA(3 day), SMA(13 day), SMA(20 day), EMA(14 day), RSI(14 day), MACD, and social media sentiment. While we analyzed and used linear regression on all variables to determine factor relation and importance, only a handful would be used for the alpha model implementation later on and all would be weighted equally. Some factors were even excluded in the process during testing in Excel, as the program returned a #NUM error for MACD, indicating this variable (and related variables like MACDUP and MACDDN) were collinear and would not provide useful results.

Once all potential factors were identified, each group member picked a stock to add to the group's portfolio. The companies chosen were Costco (COST), Home Depot (HD), Southwest Airlines (LUV), Royal Caribbean (RCL), and Tesla (TSLA).

Data Analysis

For our portfolio of stocks, we gathered ten years of daily stock price data from Yahoo Finance for each company, with a start date of September 1st, 2012. This data was converted into average monthly prices to better conform to the timeframe of other datasets used for analysis. Each stock within our portfolio was compared to the SP500, which was used as a benchmark throughout the model.

statistically significant at alpha = 5%
statistically significant at alpha = 10%
not statistically significant

Alpha (Excess Return)				
	Daily	Monthly	Quarterly	Annually
HD	0.03%	0.66%	2.00%	3.30%
COST	0.05%	0.99%	3.12%	10.23%
TSLA	0.14%	2.70%	9.96%	21.63%
RCL	-0.04%	-1.58%	-6.81%	-12.42%
LUV	0.02%	0.33%	0.82%	1.82%

Beta (Systematic Risk)				
	Daily	Monthly	Quarterly	Annually
HD	0.98	1.00	0.97	1.40
COST	0.70	0.74	0.66	0.82
TSLA	1.43	1.73	0.96	2.59
RCL	1.62	2.31	3.15	1.55
LUV	1.06	1.09	1.16	1.14

As seen in the table above, we converted each company's stock price into daily, monthly, quarterly, and annual price points over the ten-year period to determine alpha and beta values at each time. These data points were then cross-referenced with its statistical p-value to identify the significance of the value, with green cells representing that the values were statistically significant at $\alpha = 5\%$, yellow at $\alpha = 10\%$, and red values were not statistically significant at all. From the table, it could be inferred that Costco's excess return was the most stable of the group while maintaining a consistently lower beta value than the rest of the portfolio. This means that the company had less systematic risk and less price volatility (while maintaining significant excess returns) as the market moved compared to its counterparts. Tesla and Royal Caribbean seemed to experience similar fluctuations as the market moved, although alpha values tended in opposite directions.

To further study alpha and beta movements within the portfolio stocks, we used a rolling window approach to calculate these values over the 10-year period. Graphs for each company's cumulative growth and historical alpha/beta values can be seen in the appendix.

Annualized Company Performance Analytics

	Return	VaR (5%)	Max Drawdown	Sharpe Ratio	Kelly Ratio
HD	20.34%	-2.33%	37.99%	0.87	1.96
COST	21.43%	-2.03%	31.40%	1.05	2.57
TSLA	64.90%	-5.59%	60.63%	1.15	1.03
RCL	5.93%	-5.07%	83.30%	0.12	0.37
LUV	16.13%	-3.39%	62.96%	0.48	0.92

Aside from studying each company's alpha and beta values, we also conducted a performance analysis to determine the annualized variables in the table above. These results complement the previous table, highlighting the risk and volatility built into Tesla and Royal Caribbean shares, with the highest annual maximum drawdowns off 60.6% and 83.3% respectively. While Home Depot and Costco's Sharpe ratio hovers around 1, their Kelly ratio values better depict the success seen by these companies over the recent years, with annualized returns of 20.3% and 21.4% respectively.

Portfolio Inputs				
Source from daily pricing over 10-year period				
Stock	**Weight**	**Standard Dev**	**Alpha**	**Beta**
SP500		13.00%	0.00%	1.00
COST	20%	20.46%	0.05%	0.70
HD	20%	23.27%	0.03%	0.98
LUV	20%	33.49%	0.02%	1.06
RCL	20%	49.67%	-0.04%	1.62
TSLA	20%	56.45%	0.14%	1.43
Portfolio	100%	36.67%	0.04%	1.16

As mentioned previously, our group decided to equally weight each stock within the portfolio, so that each stock has a weight of 20%. Using these weights, we were able to calculate the volatility, alpha, and beta values associated with our portfolio. From this breakdown, it can be seen that over the past ten years, our selected portfolio has outperformed the market by earning an additional 0.04% daily return on investment, but at a cost. It has a much higher standard deviation and is more reactive than the market in general, so caution and due diligence should still be taken before investing. Even one of the safest indices on the stock exchange has a standard deviation of 13%, meaning that consistent returns are not guaranteed, with our portfolio at 36% standard deviation being even more uncertain.

One of the more intriguing facets of data inflow was from the social media platform of Reddit. Using R programming, we automated a process to comb through the internet and search Reddit for any threads involving stocks within our portfolio. From there, we used positive and negative

dictionaries to perform sentiment analysis on the gathered information to gain an understanding of consumer's emotions towards each stock. It should be noted that these emotions do not necessarily translate to intention or action and should not be interpreted to do so. Other websites like Hedgechatter and Marketprofit were also viewed to aid in sentiment understanding.

Factor Selection

After using R programming to establish data frames for factors and pricing for each company, we ran a linear regression using two different methods. In our first trial, we converted all values to monthly data for uniformity and exported a .csv file into Excel to run linear regression to determine which variables contributed most to variance in our dependent variable. To create a meaningful dependent variable that would give us insight into how time t affects time t_{+1}, we translated all returns up one time period so that returns for time t_{+1} coincided with factors of time t. Linear regression in Excel had to be run twice for each company because Excel only allows 16 independent variables at a time for analysis, and we had a total of 26 independent variables at our disposal. In our second trial of linear regression, we combined the entirety of our portfolio data into a single data frame to run panel data analysis in R. This method allowed us to run a single regression and had similar results to the first trial but proved to be more informative and accurate.

Overall, we found some interesting results. From our analysis, the factors for personal consumption and housing index (CANDH), FF-R_m-R_f, PE ratio, and crude price of Brent oil had the highest factor exposures with significance. With high exposure values, changes in these selected factors help best explain variability in the equity returns related to our portfolio. Other important and significant factors included Fama-French's factor for HML, consumer sentiment, retail sales, and RSI14. These variables, along with their exposure values, can be found in our appendix. While the factors may appear seemingly unrelated and scattered at first, they all point to similar root causes. It should be noticed that almost all companies included in our portfolio offer goods and services for consumers who have excess discretionary income and are spending in excess, like towards a vacation (using Southwest Airlines or Royal Caribbean), towards higher-end groceries (Costco), or towards more elegant transportation (Tesla). As seen from several of our chosen significant factors, these stocks perform better when the market is experiencing growth, inflation is low, and consumers have more money to spend on their excess wants and desires. On the contrary, when inflation is high and the market is performing poorly consumers will not allocate as much money to the "luxury" products and services that these companies provide. We believe that several stocks in our portfolio could be comparable to a

growth stock, like Tesla and Costco. This would also aid in explaining why stock prices have gone down recently as inflation in the U.S. has soared to nearly 9%.

Big Data Alpha Model

After selecting factors that carry significance to our portfolio, we needed to construct a trading strategy and backtest it with historical data to examine its success and fine-tune the strategy before final implementation. First looking at PE ratio, we saw a strong correlation between a company's reported quarterly PE ratio and its next quarter return. To test this, we developed a strategy to purchase shares from our portfolio stocks when a company would report a higher PE ratio compared to the previous quarter. After the entry condition was met, we held the stock until the end of the quarter, where it would be sold. To determine the viability of this strategy, we tested when $PE_t > PE_{t-1}$ and evaluated the following quarter's return (ret_{t+1}). After doing so we found that a total of 77 trades would be made over the 10-year period of all portfolio data, and 73% of them would result in a profitable return on investment, at an average return of 4.7% and a total return of 363% over the period.

We took advantage of another opportunity to backtest a strategy using the RSI14 factor by using a software called TrendSpider. This program has a built-in strategy testing component that allows the user to create unique entry and exit conditions to backtest against historical stock data. From our factor analysis, we found that RSI14 was a significant and impactful variable for some of the socks in our portfolio. To test the results, we created a strategy that will purchase the stock if the RSI14 crosses up RSI20, or SMA3 is greater than SMA13. After entering into the stock purchase, our specified exit conditions state that we will sell the stock once SMA3 crosses down SMA13. When we applied this strategy, we saw 64% success in a positive return. These results can be seen in more detail from the pictures in our appendix.

Summary

The big data alpha model can be used to identify independent factors that cause some of the most significant variations in our dependent variable of interest, namely returns on equity investments. Our portfolio has shown to be more volatile than the SP500, while yielding a small amount of alpha. The portfolio Sharpe ratio of 0.73 would lend to the idea that our group is carrying additional volatility relative to the excess returns offered by the group of stocks.

We expected social media sentiment to play a larger role in our linear regression models. Using R programming to access twitter API proved to be difficult, and we instead resorted to Reddit

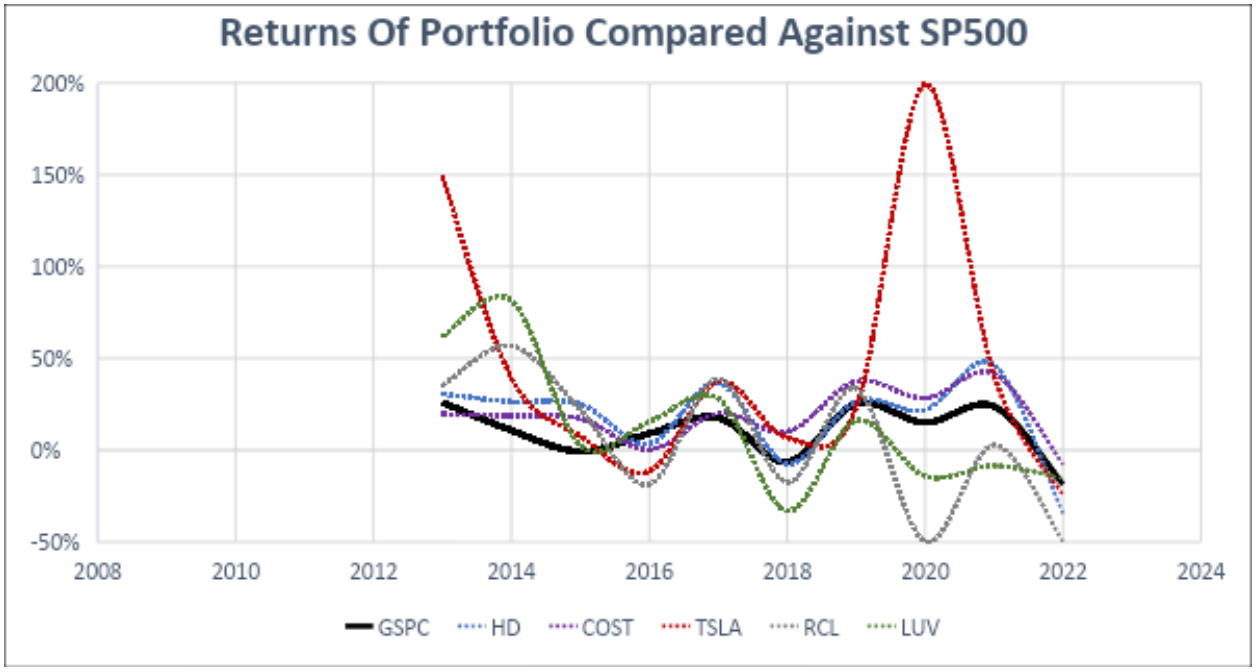
data. While information was available, there may not have been enough volume retrieved to make a significant impact during the sentiment mining process. We would like to continue exploring other methods of sentiment analysis in the future to focus on more accurate depictions of attitudes and feelings on other social media platforms.

Appendix A

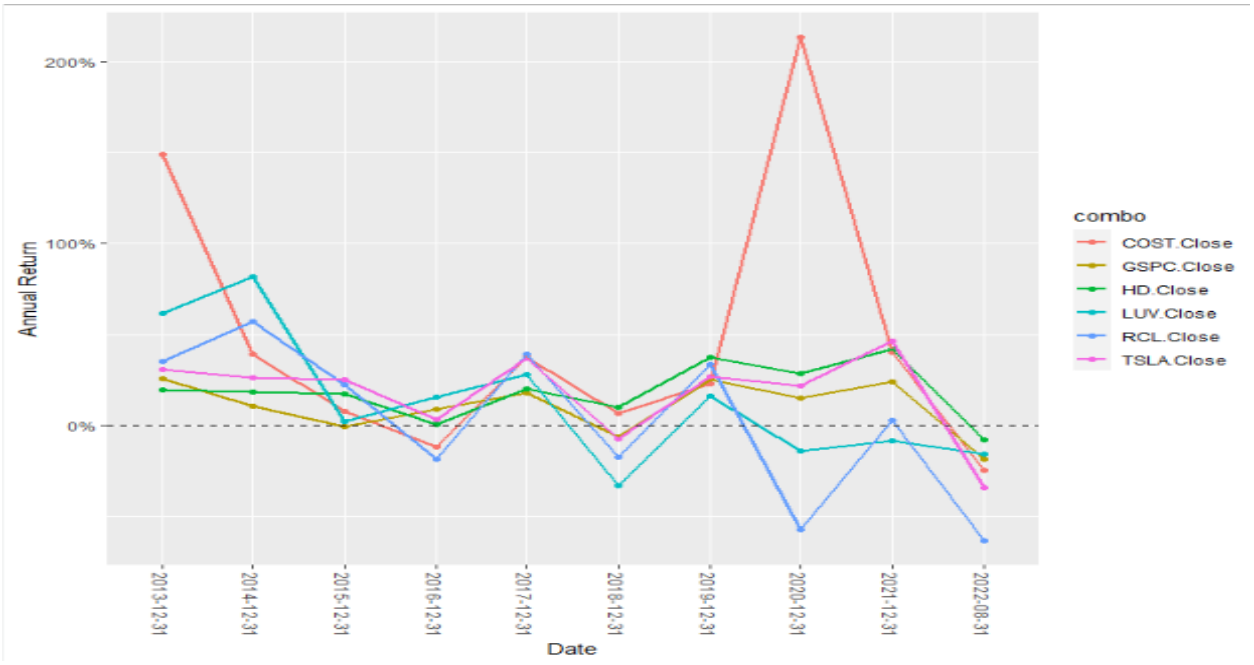
Ahmad, Irfan. “How Much Data Is Generated Every Minute?” *SocialMediaToday*, 15 June 2018, https://www.socialmediatoday.com/news/how-much-data-is-generated-every-minute-infographic-1/525692/. Accessed 30 Sept. 2022.

Bender, Jennifer, et al. MSCI, 2013, *Foundations of Factor Investing*, https://www.msci.com/documents/1296102/1336482/Foundations_of_Factor_Investing.pdf/004e02ad-6f98-4730-90e0-ea14515ff3dc. Accessed 30 Sept. 2022.

Appendix B



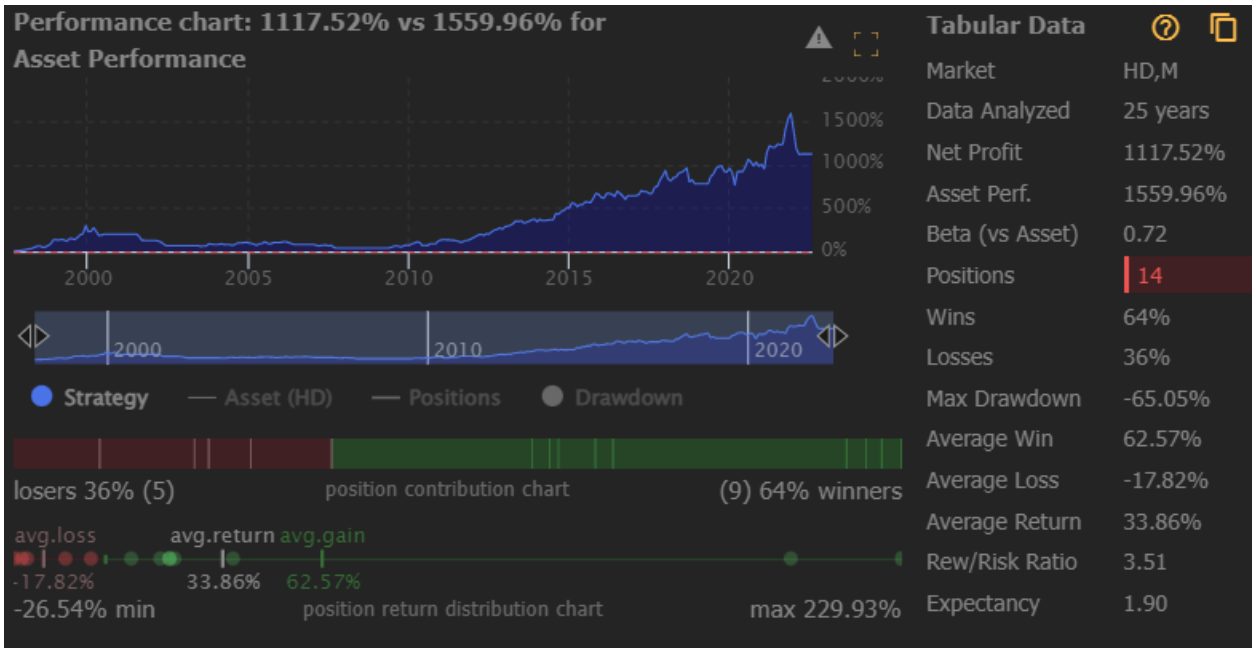
The graph above was calculated in Excel

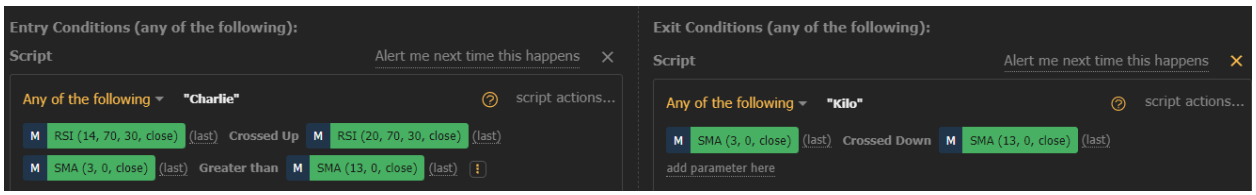


The graph above was calculated in R

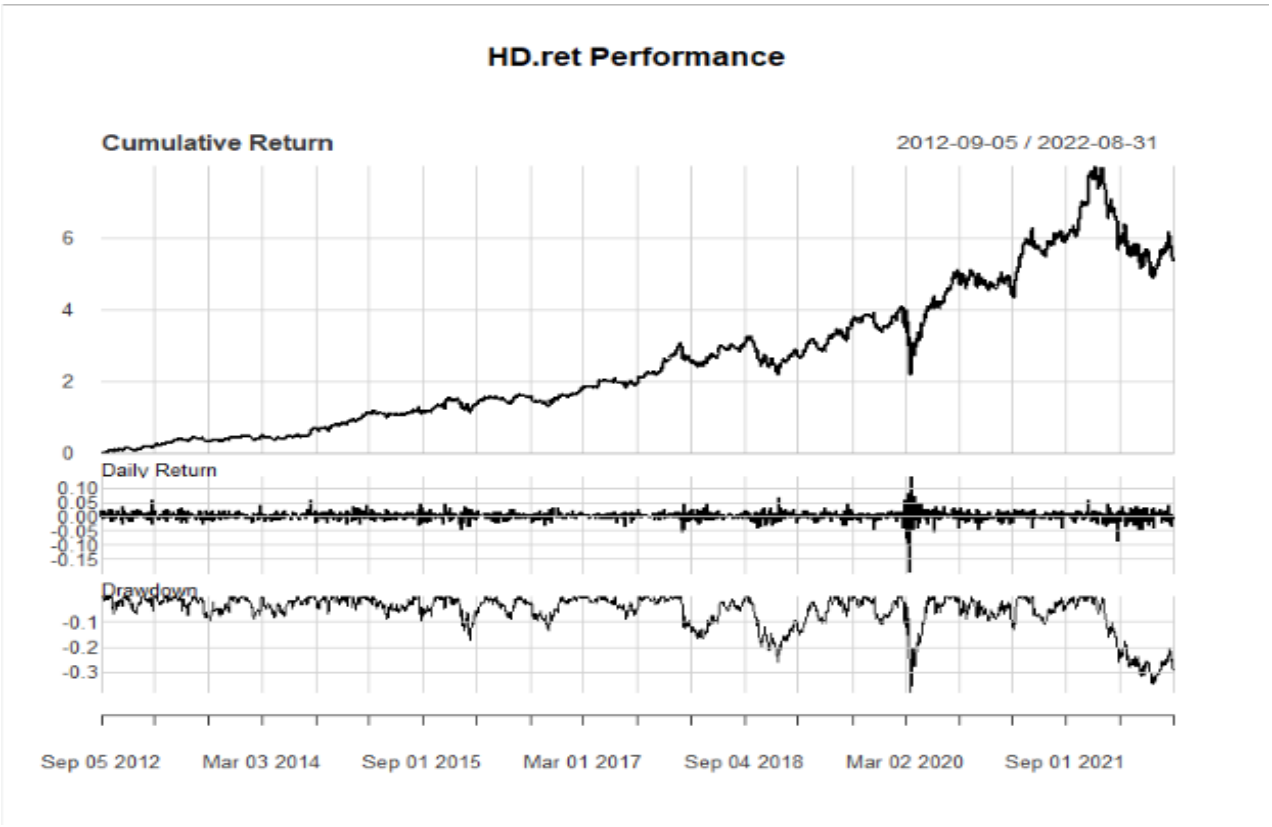
Appendix C

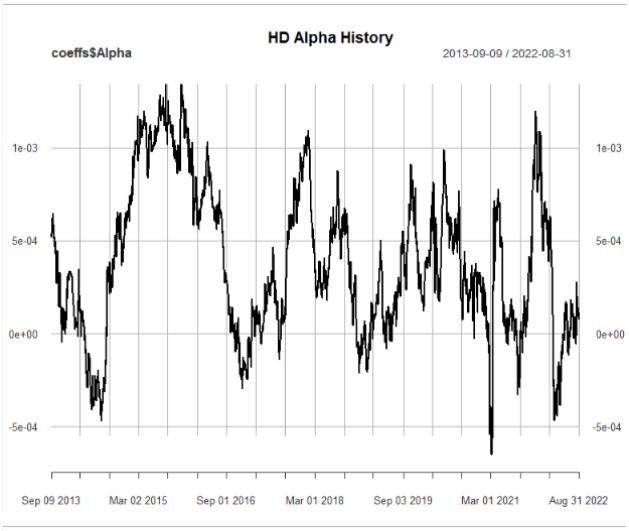

TrendSpider Sep 30, 2022 5:34PM EDT
HD, Monthly, Candles chart (Logarithmic scale)
Current strategy, Monthly
Drawings, Monthly
Backtest started here
Entry, long 17.38
Exit +197.84% ("Kilo")
Entry, long 50.37
Entry, long 41.08
Entry, long 38.90
Entry, long 40.74
Entry, long 23.16
Exit +56.79% ("Kilo")
Entry, long 30.21
Exit +18.50% ("Kilo")
Entry, long 39.22
Exit +229.93% ("Kilo")
Entry, long 144.91
Exit +18.57% ("Kilo")
Entry, long 189.85
Exit +15.79% ("Kilo")
Entry, long 250.51
Exit +19.92% ("Kilo")
275.94
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023
©2022 TrendSpider

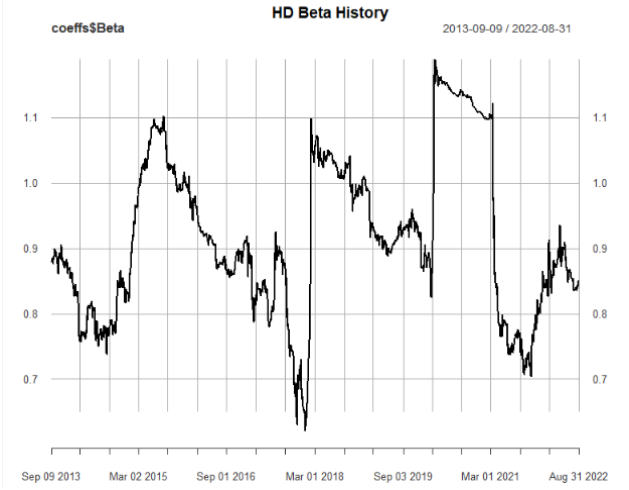

Performance chart: 1117.52% vs 1559.96% for Asset Performance
1500%
1000%
500%
0%
2000
2005
2010
2015
2020
Strategy
Asset (HD)
Positions
Drawdown
losers 36% (5)
position contribution chart
(9) 64% winners
avg.loss
avg.return
avg.gain
-17.82%
33.86%
62.57%
-26.54% min
position return distribution chart
max 229.93%
Tabular Data
Market HD,M
Data Analyzed 25 years
Net Profit 1117.52%
Asset Perf. 1559.96%
Beta (vs Asset) 0.72
Positions 14
Wins 64%
Losses 36%
Max Drawdown -65.05%
Average Win 62.57%
Average Loss -17.82%
Average Return 33.86%
Rew/Risk Ratio 3.51
Expectancy 1.90

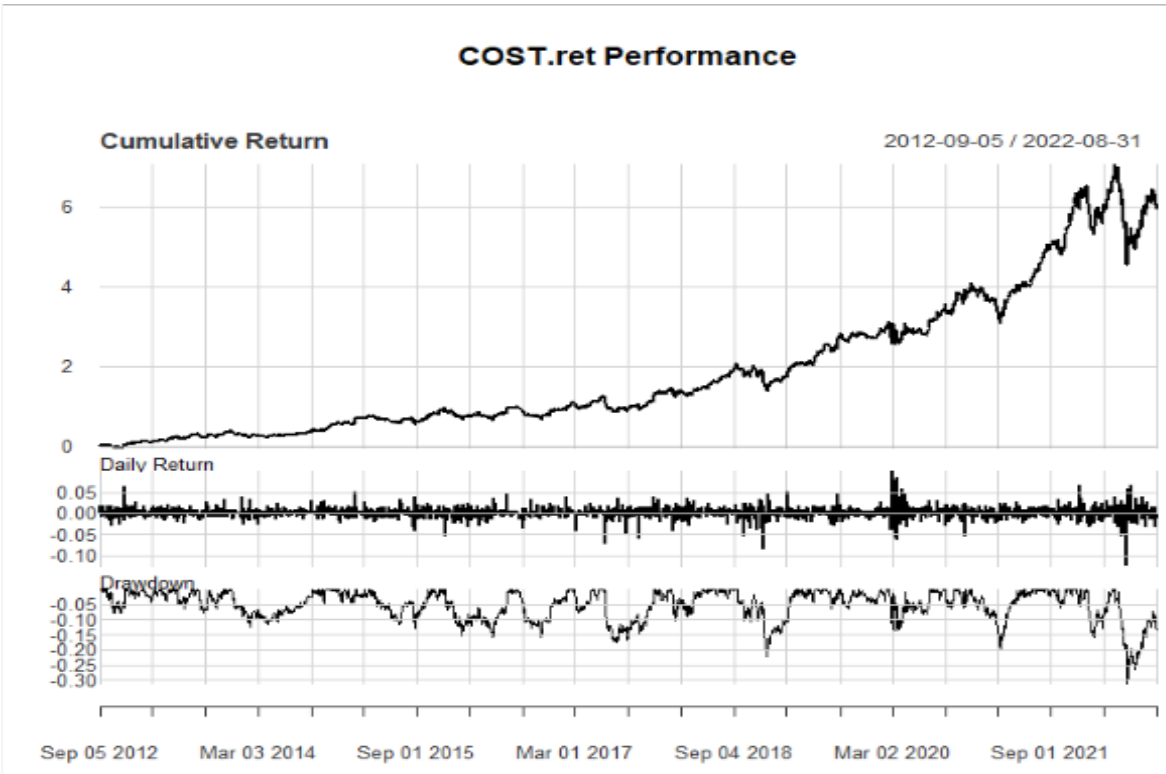

Entry Conditions (any of the following):
Script
Alert me next time this happens
Any of the following
"Charlie"
script actions...
RSI (14, 70, 30, close) (last) Crossed Up RSI (20, 70, 30, close) (last)
SMA (3, 0, close) (last) Greater than SMA (13, 0, close) (last)
Exit Conditions (any of the following):
Script
Alert me next time this happens
Any of the following
"Kilo"
script actions...
SMA (3, 0, close) (last) Crossed Down SMA (13, 0, close) (last)
add parameter here

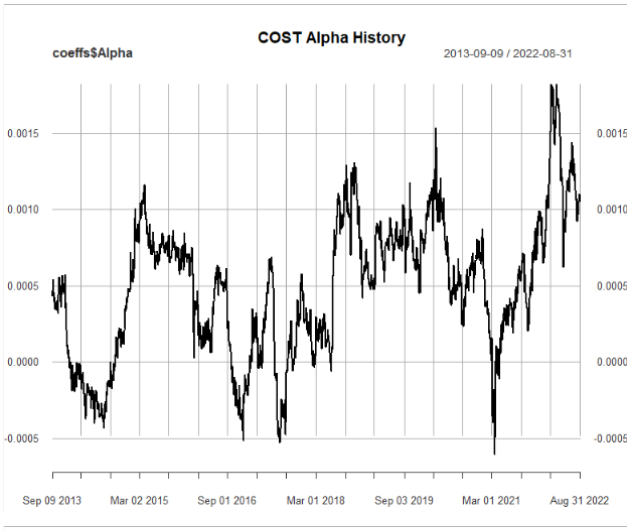
Appendix D

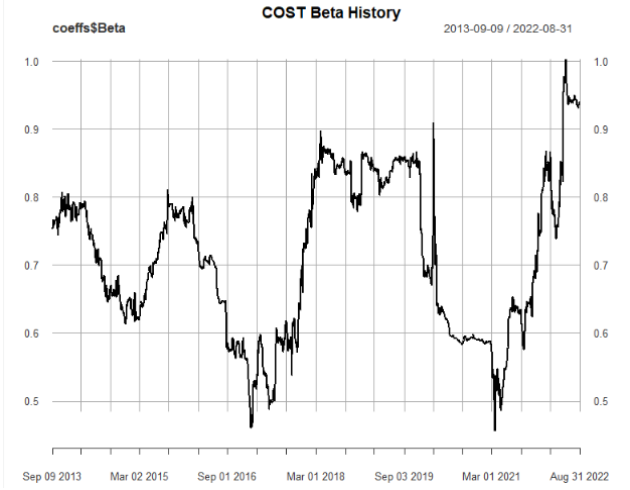

HD.ret Performance
Cumulative Return
2012-09-05 / 2022-08-31
6
4
2
0
Daily Return
0.10
0.05
0.00
-0.05
-0.10
-0.15
Drawdown
-0.1
-0.2
-0.3
Sep 05 2012
Mar 03 2014
Sep 01 2015
Mar 01 2017
Sep 04 2018
Mar 02 2020
Sep 01 2021

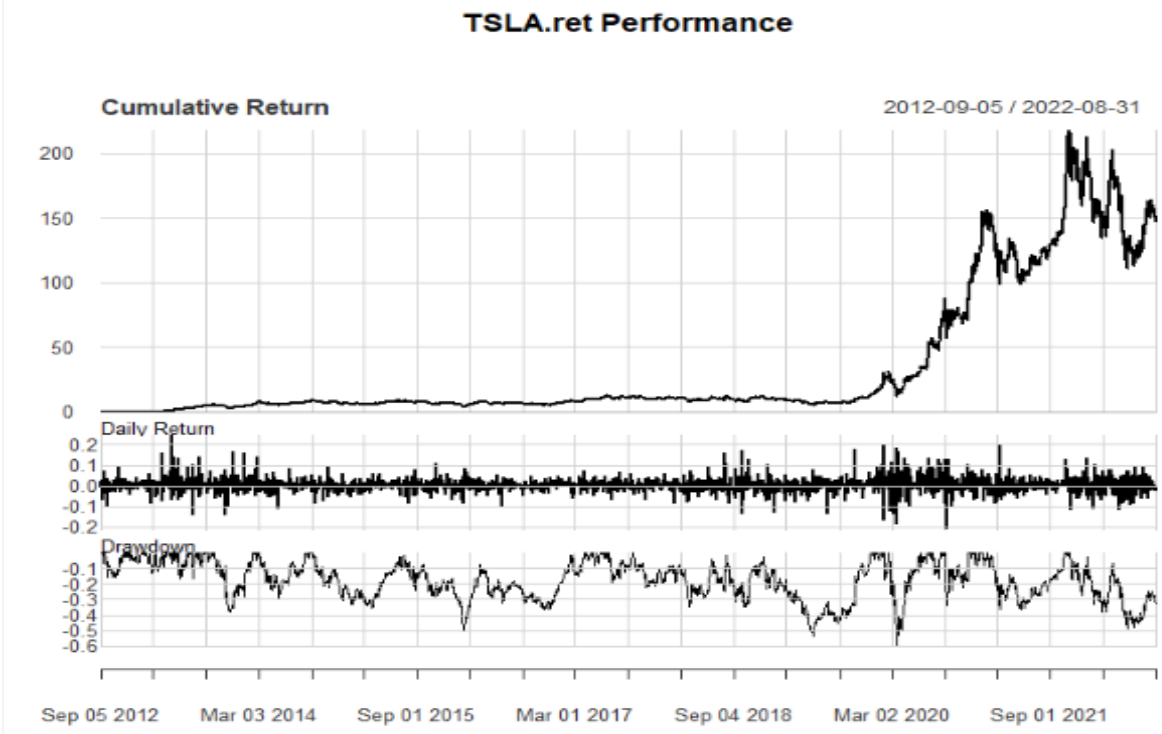

HD Alpha History
coeffs$Alpha
2013-09-09 / 2022-08-31
1e-03
5e-04
0e+00
-5e-04
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022

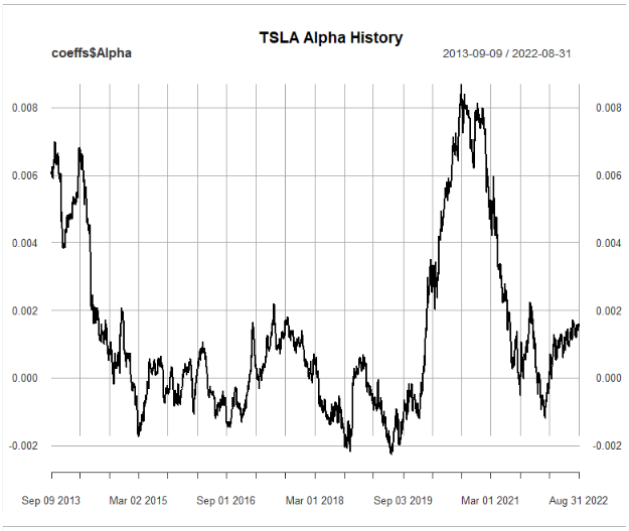

HD Beta History
coeffs$Beta
2013-09-09 / 2022-08-31
1.1
1.0
0.9
0.8
0.7
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022

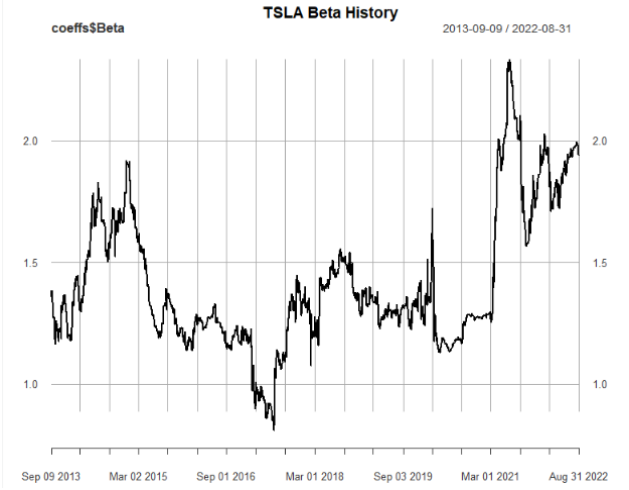

COST.ret Performance
Cumulative Return
2012-09-05 / 2022-08-31
6
4
2
0
Daily Return
0.05
0.00
-0.05
-0.10
Drawdown
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
Sep 05 2012
Mar 03 2014
Sep 01 2015
Mar 01 2017
Sep 04 2018
Mar 02 2020
Sep 01 2021

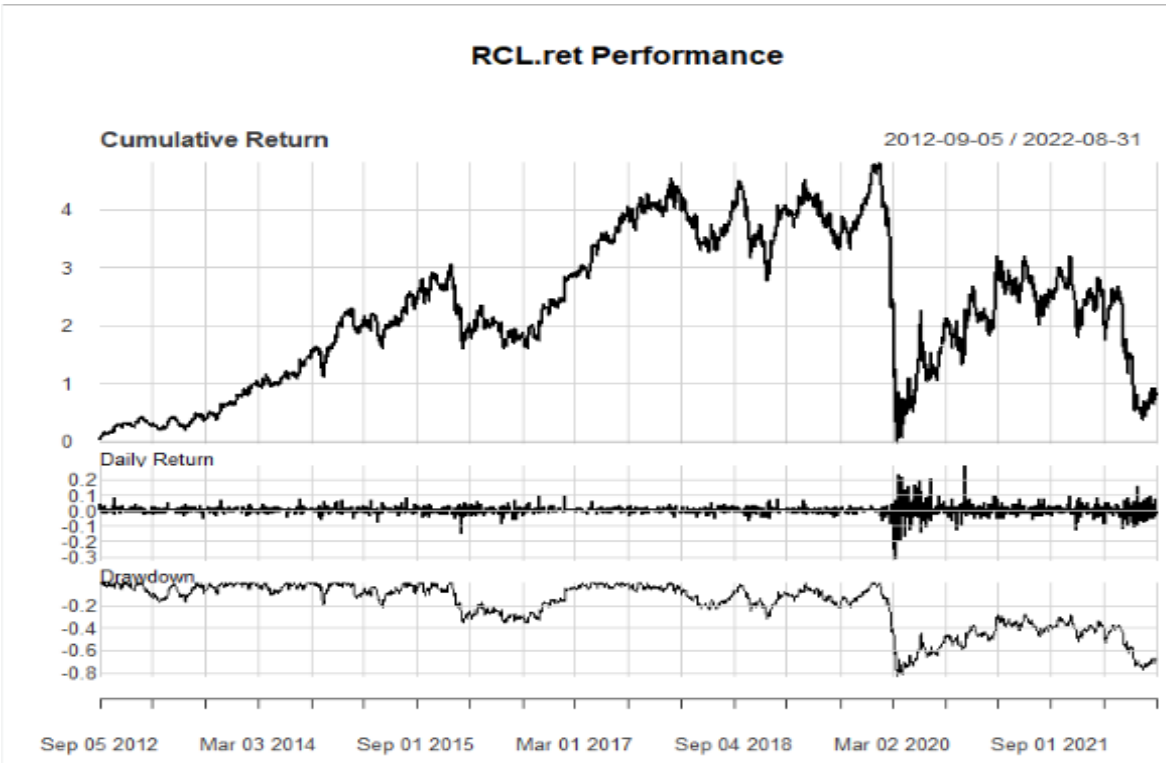

COST Alpha History
coeffs$Alpha
2013-09-09 / 2022-08-31
0.0015
0.0010
0.0005
0.0000
-0.0005
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022

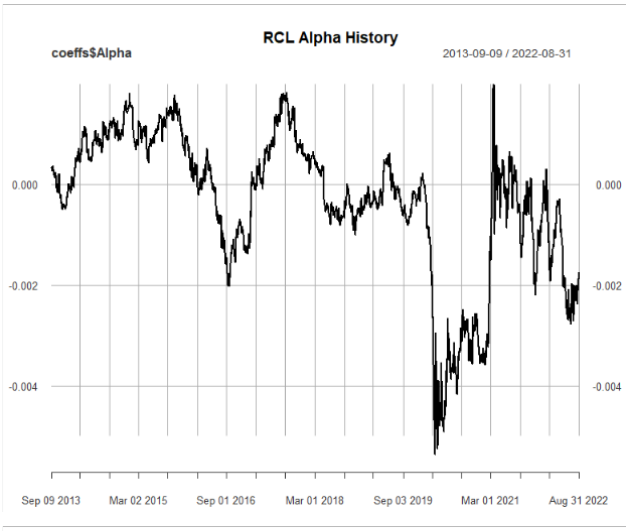

COST Beta History
coeffs$Beta
2013-09-09 / 2022-08-31
1.0
0.9
0.8
0.7
0.6
0.5
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022

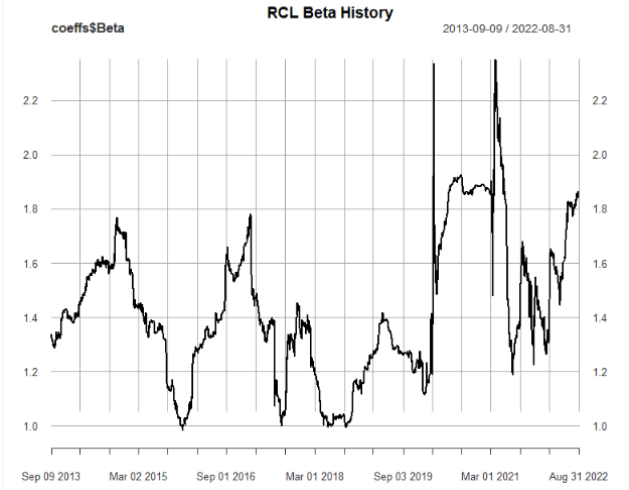

TSLA.ret Performance
Cumulative Return
2012-09-05 / 2022-08-31
200
150
100
50
0
Daily Return
0.2
0.1
0.0
-0.1
-0.2
Drawdown
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
Sep 05 2012
Mar 03 2014
Sep 01 2015
Mar 01 2017
Sep 04 2018
Mar 02 2020
Sep 01 2021

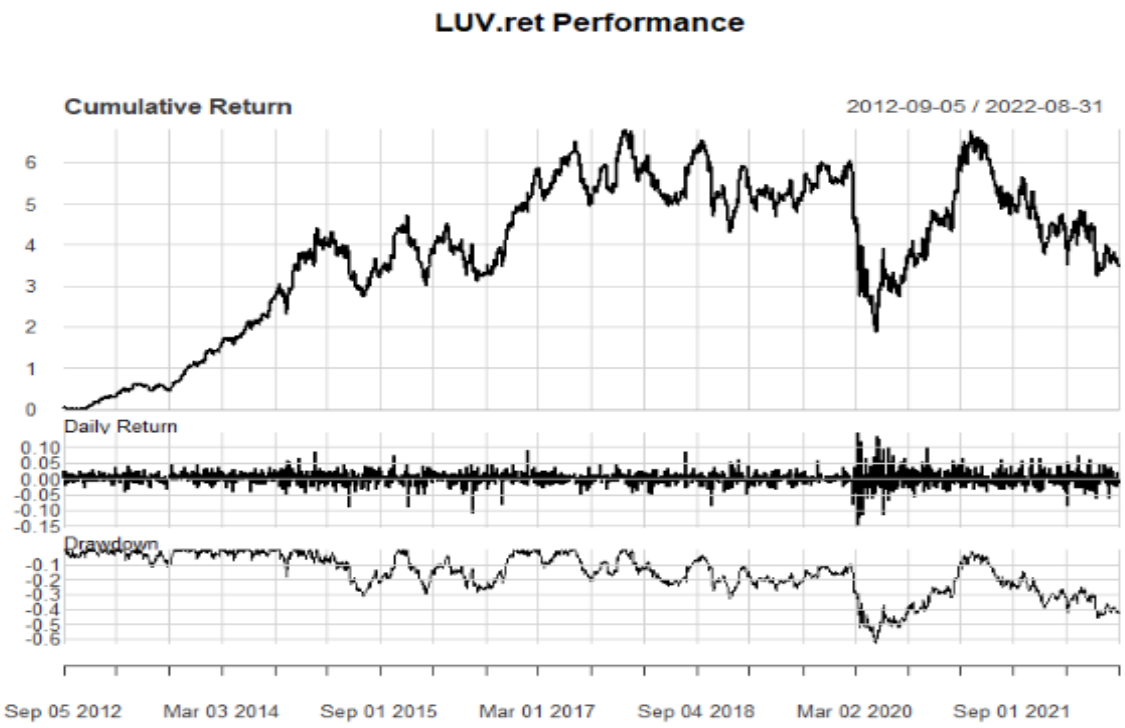

TSLA Alpha History
coeffs$Alpha
2013-09-09 / 2022-08-31
0.008
0.006
0.004
0.002
0.000
-0.002
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022

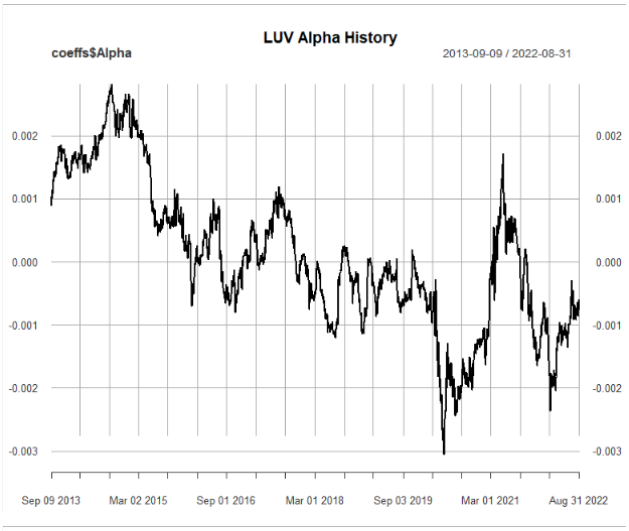

TSLA Beta History
coeffs$Beta
2013-09-09 / 2022-08-31
2.0
1.5
1.0
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022

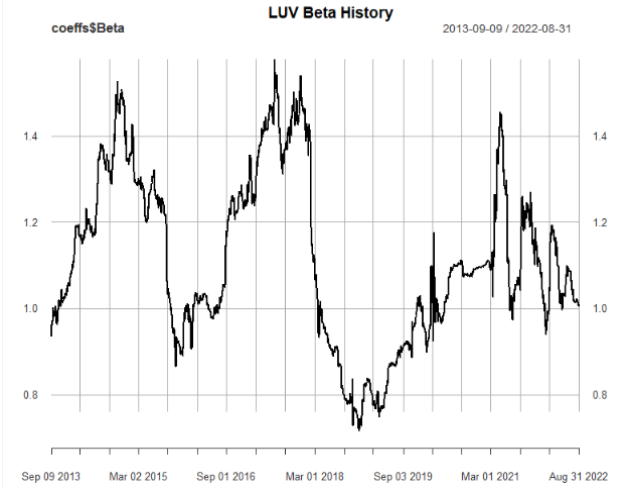

RCL.ret Performance
Cumulative Return
2012-09-05 / 2022-08-31
4
3
2
1
0
Daily Return
0.2
0.1
0.0
-0.1
-0.2
-0.3
Drawdown
-0.2
-0.4
-0.6
-0.8
Sep 05 2012
Mar 03 2014
Sep 01 2015
Mar 01 2017
Sep 04 2018
Mar 02 2020
Sep 01 2021


RCL Alpha History
coeffs$Alpha
2013-09-09 / 2022-08-31
0.000
-0.002
-0.004
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022


RCL Beta History
coeffs$Beta
2013-09-09 / 2022-08-31
2.2
2.0
1.8
1.6
1.4
1.2
1.0
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022


LUV.ret Performance
Cumulative Return
2012-09-05 / 2022-08-31
6
5
4
3
2
1
0
Daily Return
0.10
0.05
0.00
-0.05
-0.10
-0.15
Drawdown
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
Sep 05 2012
Mar 03 2014
Sep 01 2015
Mar 01 2017
Sep 04 2018
Mar 02 2020
Sep 01 2021


LUV Alpha History
coeffs$Alpha
2013-09-09 / 2022-08-31
0.002
0.001
0.000
-0.001
-0.002
-0.003
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022


LUV Beta History
coeffs$Beta
2013-09-09 / 2022-08-31
1.4
1.2
1.0
0.8
Sep 09 2013
Mar 02 2015
Sep 01 2016
Mar 01 2018
Sep 03 2019
Mar 01 2021
Aug 31 2022