# Intra-daily 51 Financial Factors Database

Guilherme Masuko
Advisor: Marcelo C. Medeiros

April 2023

**Abstract**

This paper had as main goal to report how we will create an intra daily basis of some of the key financial factors for the US market. In Section 2 we theoretically show how we form 51 factors. The final database will have 51 financial factors in addition to the returns of stocks listed on the US stock exchange NYSE. The main novelty in this database is the intra-daily frequency of these variables.

During the scope of the article, we discussed the use of dimensionality reduction tools that will be needed in the context of model selection. We motivated this discussion by bringing the next steps that include verifying whether these factors help in the predictive power of high-frequency returns.

# 1 Introduction

Within the predictive world of finance, financial economists try to find methods and models that can increasingly more accurately predict stock returns. There are several approaches and tools used in financial analysis. These techniques are used to analyze historical data and try to predict future stock market movements.

On the one hand, the literature walked for a long time trying to discover financial factors that could better explain the returns or that would increase the predictive power of models that already have significant factors when making forecasts of returns. The search for factors that explain the cross section of expected stock returns began a long time ago with its first model, the Capital Asset Pricing Model (CAPM) introduced by Sharpe (1964), Lintner (1965) and Mossin (1966) independently , based on previous work by Markowitz (1952), using only the market factor to explain stocks' excess returns.

The factors are formed through portfolios where there is a long and a short position. For example, to create the market factor, we use a long position on the market index of interest and short on the risk-free interest rate, that is, market return minus risk-free asset return.

Looking for more factors that increase the predictive power, the well-known Fama and French models emerged, respectively, the three-factor model in Fama e French (1992) and the five-factor model in Fama e French (2015).

The first model aggregates the factors related to size, SMB (Small minus Big), and value, HML (High minus Low), based respectively on the variables Market Capitalization and Book-to-Market ratio, keeping the market factor at your model. In the second model, the authors add two more factors related to profitability and investment. Defined analogously to the SMB and HML factors, the RMW (Robust minus Weak) factors are constructed from the difference between the returns of firms with robust (high) and weak (low) operating profitability and the CMA investment factor (Conservative minus Aggressive) constructed from the difference between the returns of companies that invest conservatively and companies that invest aggressively. Another factor related to momentum was also added to the Fama and French three-factor model by Carhart (1997). MOM (Monthly Momentum Factor) can be calculated by subtracting the equal weighted average of the lowest performing companies from the equal weighted average of the highest performing companies.

Now, with hundreds of financial factors released in the literature, the difficulty has become to tame this high dimensionality of factors.

On the other hand, there is still a great use of traditional methods, such as the use of autoregressive regression (AR). This is what is done in Chinco, Clark-Joseph e Ye (2019), which is the article most closely linked to this work. In the article, the authors use the AR regression with three lags as a reference model in their main specifications, although they demonstrate that the number of lags does not matter with the problem addressed. The objective of the article is to show that the identification of predictors in finance using only the intuition of researchers would only work for long-term stable predictors.

In high-frequency finance with modern, large, fast, and complex financial markets, the process of selecting predictors requires efforts that go beyond researchers' intuition. And perhaps these long-term stable predictors are not suitable to explain intraday returns because they are not able to capture the effects that affect stock prices in the middle of a day. The paper's results show that statistical model selection tools can increase the predictive power of one-minute-ahead forecasts of benchmark models using out-of-sample fit and forecast-implied Sharpe ratios as measures of quality.

For model selection, the authors make use of LASSO (Least Absolute Shrinkage and Selection Operator) using three lags of all returns as candidate predictors. They ague the use of a dimension reduction tool because there are many predictor candidates. For example, in the month of January 2003 our sample contains $4500+$ stocks. By using three lags, we then have $3 \cdot 4500+ = 13500+$ candidate predictors. It would take more than 34 trading days to do a simple one-minute OLS (Ordinary Least Squares) estimation (each day has 390 minutes/observations) and test the predictors.

With this high dimensionality of predictor candidates, a reduction tool is necessary and a researcher's intuition does not seem to be the right path for this scale. LASSO fits as a solution because it is able to identify unexpected and short-term predictors, suitable for an intraday financial predictive model. The initial hypothesis for using LASSO is to bet on sparsity, that is, if among the $13500+$ candidate predictors, only a few predictors, say $S$, actually help to predict the returns of an asset, then we can leverage to use LASSO, because it would be needed only a little more than $S$ observations to make predictions, this means that, since we don't have to worry about the weak estimators, LASSO can estimate the remaining parameters with far fewer observations. Thus, if there are only $S$ important predictors at each point in time, LASSO is adequate for estimating unexpected short-term parameters. We explain more about how LASSO works in Appendix A.1.

Although the paper obtains results that refer to gains in predictive power by combining LASSO with Benchmarks models, the selection of predictors is made based only on return lags. The factors literature is neglected and therefore, this work was motivated to verify what happens when we add to this predictive model a base that contains relevant financial factors.

This summer paper therefore aims to explain how the database will be created and report the next steps after that. The work is divided into two more sections. Section 2 addresses how the databases have been built so far and how theoretically the factors will be built. Section 3 will motivate what the next steps will be after having the base ready.

## 2 Data

The first database of this work is TAQ Returns. The TAQ Returns Database is a financial database that contains information about returns for all stocks listed on the New York Stock Exchange (NYSE). We have around ten thousand assets entering and leaving the database from January 2003 to December 2020.

The TAQ Returns database provides return data at different frequencies, ranging from daily frequency to minute and second frequencies. For example, you can get return data for intervals such as 30, 15, 5, and 1 minute, as well as 30, 15, and 5 seconds. This allows for a more detailed and accurate analysis of stock performance over different time frames.

The second database was built using two different sets of data. The first contains data on PERMNO, TAQ_TICKER and some firm characteristics. The second dataset also has the PERMNO column, which is the variable we connect the two datasets with, and several columns that give us the percentiles of companies on where they are located based on financial factors. Companies are grouped into ten percentiles. Thus, each firm receives values between 1 and 10 for each factor, although some factors are binary variables and therefore firms only receive the values 1 and 10. Next, we show how each financial factor was constructed.

1. **Size** ($size$)

   Based on Fama e French (1993).

   $$size = \mathrm{ME}_{\text{Jun}}$$

   The CRSP end of June price times shares outstanding. Rebalanced annually.

2. **Value (annual)** ($value$)

Follows Fama e French (1993).

$$value = \frac{\mathrm{BE}}{\mathrm{ME}}$$

At the end of June of each year, we use book equity from the previous fiscal year and market equity from December of the previous year. Rebalanced annually.

3. **Gross Profitability** ($prof$)

   Follows Novy-Marx (2013).

$$prof = \frac{\mathrm{GP}}{\mathrm{AT}}$$

   where $\mathrm{GP}$ is gross profits and $\mathrm{AT}$ is total assets. Rebalanced annually.

4. **Cash flow duration** ($dur$).

   Follows ?.

$$dur = \frac{\sum_t \mathrm{PV}_0\,(t \times \mathrm{CF}_t)}{P_0}$$

   Present value of expected cashflows. Cashflows' components (from clean surplus identity, ROE and book equity growth) are forecasted using AR(1). Sums are discounted using a constant discount rate. Rebalanced monthly.

5. **Value-Profitability** ($valprof$).

   Follows Novy-Marx (2013).

$$valprof = \mathrm{rank}(value) + \mathrm{rank}(prof)$$

   Sum of ranks in univariate sorts on book-to-market and profitability. Annual book-to-market and profitability values are used for the entire year. Rebalanced monthly

6. **Piotroski's F-score** ($F\text{-}score$)

   Follows Piotroski (2000).

$$
\begin{aligned}
F\text{-}score = & 1_{\text{IB}>0} + 1_{\Delta\text{ROA}>0} + 1_{\text{CFO}>0} + 1_{\text{CFO}>\text{IB}} + 1_{\Delta\text{DTA}<0|\text{DLTT}=0|\text{DLTT}_{-12}=0} \\
& + 1_{\Delta\text{ATL}>0} + 1_{\text{EqIss}\leq 0} + 1_{\Delta\text{GM}>0} + 1_{\Delta\text{ATO}>0}
\end{aligned}
$$

   where $\text{IB}$ is income before extraordinary items, $\text{ROA}$ is income before extraordinary items scaled by lagged total assets, $\text{CFO}$ is cash flow from operations, $\text{DTA}$ is total long-term debt scaled by total assets, $\text{DLTT}$ is total long-term debt, $\text{ATL}$ is total current assets scaled by total current liabilities, $\text{EqIss}$ is the difference between sales of common stock and purchases of common stock recorded on the cash flow statement, $\text{GM}$ equals one minus the ratio of cost of goods sold and total revenues, and $\text{ATO}$ equals total revenues, scaled by total assets. Rebalanced annualy. Binary variable.

7. **Debit Issuance** ($debtiss$)

   Follows Spiess e Affleck-Graves (1999).

$$debtiss = 1_{\text{DLTISS}\leq 0}$$

   Binary variable equal to one if long-term debt issuance indicated in statement of cash flow. Updated annually. Binary variable.

8. **Share Repurchases** ($repurch$)

   Follows Ikenberry, Lakonishok e Vermaelen (1995).

$$repurch = 1_{\text{PRSTKC}>0}$$

   Binary variable equal to one if repurchase of common or preferred shares indicated in statement of cash flow. Updated annually. Binary variable.

9. **Share Issuance (annual)** ($nissa$)

   Follows Pontiff e Woodgate (2008).

$$nissa = \frac{\text{shrout}_{\text{Jun}}}{\text{shrout}_{\text{Jun}-12}}$$

where $\text{shrout}$ is the number of shares outstanding. Change in real number of shares outstanding from past June to June of the previous year. Excludes changes in shares due to stock dividends and splits, and companies with no changes in $\text{shrout}$.

10. **Accruals** ($accruals$)

    Follows Sloan (1996).

$$accruals = \frac{\Delta\text{ACT} - \Delta\text{CIIE} - \Delta\text{LCT} + \Delta\text{DLC} + \Delta\text{TXP} - \Delta\text{DP}}{(\text{AT} + \text{AT}_{-12})/2}$$

where $\Delta\text{ACT}$ is the annual change in total current assets, $\Delta\text{CHE}$ is the annual change in total cash and short-term investments, $\Delta\text{LCT}$ is the annual change in current liabilities, $\Delta\text{DLC}$ is the annual change in debt in current liabilities, $\Delta\text{TXP}$ is the annual change in income taxes payable, $\Delta\text{DP}$ is the annual change in depreciation and amortization, and $(\text{AT} + \text{AT}_{-12})/2$ is average total assets over the last two years. Rebalanced annually.

11. **Asset Growth** ($growth$)

    Follows Cooper, Gulen e Schill (2008).

$$growth = \frac{\text{AT}}{\text{AT}_{-12}}$$

Rebalanced annually.

12. **Asset Turnover** ($aturnover$)

    Follows Novy-Marx (2013).

$$aturnover = \frac{\text{SALE}}{\text{AT}}$$

Sales to total assets. Rebalanced annually.

13. **Gross Margins** ($gmargins$).

    Follows Novy-Marx (2013).

    $$gmargins = \frac{\text{GP}}{\text{SALE}}$$

    where $\text{GP}$ is gross profits and $\text{SALE}$ is total revenues. Rebalanced annually.

14. **Dividend Yield** ($divp$)

    Follows Naranjo, Nimalendran e Ryngaert (1998).

    $$divp = \frac{\text{Div}}{\text{ME}_{\text{Dec}}}$$

    Dividend scaled by price. Both are measured in December of the year $t-1$ or $t-2$ (for returns in months prior to July). Rebalanced annually.

15. **Earnings/Price** ($ep$)

    Follows Basu (1977).

    $$ep = \frac{\text{IB}}{\text{ME}_{\text{Dec}}}$$

    Net income scaled by market value of equity. Updated annually.

16. **Cash Flow / Market Value of Equity** ($cfp$)

    Follows Lakonishok, Shleifer e Vishny (1994).

    $$cfp = \frac{\text{IB} + \text{DP}}{\text{ME}_{\text{Dec}}}$$

    Net income plus depreciation and amortization, all scaled by market value of equity measured at the same date. Updated annually.

17. **Net Operating Assets** ($noa$)

    Follows Hirshleifer et al. (2004).

$$noa = \frac{(\mathrm{AT} - \mathrm{CHE}) - (\mathrm{AT} - \mathrm{DLC} - \mathrm{DLTT} - \mathrm{MIB} - \mathrm{PSTK} - \mathrm{CEQ})}{\mathrm{AT}_{-12}}$$

where $\mathrm{AT}$ is total assets, $\mathrm{CHE}$ is cash and short-term investments, $\mathrm{DLC}$ is debt in current liabilities, $\mathrm{DLTT}$ is long term debt, $\mathrm{MIB}$ is non-controlling interest, $\mathrm{PSTK}$ is preferred capital stock, and $\mathrm{CEQ}$ is common equity. Updated annually.

18. **Investment** ($inv$).

    Follows Chen, Novy-Marx e Zhang (2011).

$$inv = \frac{\Delta\mathrm{PPEGT} + \Delta\mathrm{INVT}}{\mathrm{AT}_{-12}}$$

where $\Delta\mathrm{PPEGT}$ is the annual change in gross total property, plant, and equipment, $\Delta\mathrm{INVT}$ is the annual change in total inventories, and $\mathrm{AT}_{-12}$ is lagged total assets. Rebalanced annually, uses the full period.

19. **Investment-to-Capital** ($invcap$).

    Follows Xing (2008).

$$invap = \frac{\mathrm{CAPX}}{\mathrm{PPENT}}$$

Investment to capital is the ratio of capital expenditure (Compustat item $\mathrm{CAPX}$) over property, plant, and equipment (Compustat item $\mathrm{PPENT}$).

20. **Invetment Growth** ($igrowth$).

    Follows Xing (2008).

$$igrowth = \frac{\mathrm{CAPX}}{\mathrm{CAPX}_{-12}}$$

Investment growth is the percentage change in capital expenditure (Compustat item $(\mathrm{CAPX})$

21. **Sales Growth** ($sgrowth$).

    Follows Lakonishok, Shleifer e Vishny (1994).

$$sgrowth = \frac{\text{SALE}}{\text{SALE}_{-12}}$$

Sales growth is the percent change in net sales over turnover (Compustat item SALE).

22. **Leverage** ($lev$).

    Follows Bhandari (1988).

$$lev = \frac{\text{AT}}{\text{ME}_{\text{Dec}}}$$

    Market leverage is the ratio of total assets (Compustat item AT) over the market value of equity. Both are measured in December of the same year.

23. **Return on Assets (annual)** ($roaa$).

    Follows Chen, Novy-Marx e Zhang (2011).

$$roaa = \frac{\text{IB}}{\text{AT}}$$

    Net income scaled by total assets. Updated annually.

24. **Return on Equity (annual)** ($roea$).

    Follows Haugen e Baker (1996).

$$roea = \frac{\text{IB}}{\text{BE}}$$

    Net income scaled by book value of equity. Updated annually.

25. **Sales-to-Price** ($sp$).

    Follows Jr, Mukherji e Raines (1996).

$$sp = \frac{\text{SALE}}{\text{ME}_{\text{Dec}}}$$

    Total revenues divided by stock price. Updated annually.

26. **Growth in LTNOA** ($gltnoa$).

    Follows Fairfield, Whisenant e Yohn (2003).

$$gltnoa = \text{GRNOA} - \text{ACC}$$

Growth in Net Operating Assets minus Accruals, where

$$\begin{aligned}
\text{GRNOA} &= \text{NOA} - \text{NOA}_{\text{-12}} \\
\text{NOA} &= \frac{\text{RECT} + \text{INVT} + \text{ACO} + \text{PPENT} + \text{INTAN} + \text{AO} - \text{AP} - \text{LCO} - \text{LO}}{\text{AT}} \\
\text{ACC} &= \frac{\Delta\text{RECT} + \Delta\text{INVT} + \Delta\text{ACO} - \Delta\text{AP} - \Delta\text{LCO} - \text{DP}}{(\Delta\text{AT}/2)} \\
\Delta\text{RECT} &= \text{RECT} - \text{RECT}_{-12} \\
\Delta\text{INVT} &= \text{INVT} - \text{INVT}_{-12} \\
\Delta\text{ACO} &= \text{ACO} - \text{ACO}_{-12} \\
\Delta\text{AP} &= \text{AP} - \text{AP}_{-12} \\
\Delta\text{LCO} &= \text{LCO} - \text{LCO}_{-12} \\
\Delta\text{AT} &= \text{AT} - \text{AT}_{-12}
\end{aligned}$$

where RECH = Receivables, INVT = Total Inventory, ACO = Current Assets, AP = Accounts Payable, LCO = Current Liabilities (Other), DP = Depreciation and Amortization, AT = Assets, PPENT = Property, Plant, and Equipment (net), INTAN = Intangible Assets, AO = Assets (Other), LO = Liabilities (Other). Updated annually.

27. **Momentum (6m)** ($mom$).

    Follows Jegadeesh e Titman (1993).

$$mom = \sum_{l=2}^{7} r_{t-l}$$

Cumulated past performance in the previous 6 months by skipping the most recent month. Rebalanced monthly.

28. **Industry Momentum** ($indmom$).

    Follows Moskowitz e Grinblatt (1999).

$$indmom = \text{rank}\left(\sum_{l=1}^{6} r_{t-l}^{\text{ind}}\right)$$

    In each month, the Fama and French 49 industries are ranked on their value-weighted past 6-months performance. Rebalanced monthly.

29. **Value-Momentum** ($valmom$).

    Follows Novy-Marx (2013).

$$valmom = \text{rank}(value) + \text{rank}(mom)$$

    Sum of ranks in univariate sorts on book-to-market and momentum. Annual book-to-market values are used for the entire year. Rebalanced monthly.

30. **Value-Momentum-Profitability** ($valmomprof$).

    Follows Novy-Marx (2013).

$$valmomprof = \text{rank}(value) + \text{rank}(prof) + \text{rank}(mom)$$

    Sum of ranks in umivariate sorts on book-to-market, profitability, and momentum. Annual book-to-market and profitability values are used for the entire year. Rebalanced monthly.

31. **Short Interest** ($shortint$).

    Follows Dechow, Kothari e Watts (1998).

$$shortint = \frac{\text{Shares Shorted}}{\text{Shares Outstanding}}$$

    Updated monthly.

32. **Momentum (1 year)** ($mom12$).

Follows Jegadeesh e Titman (1993).

$$mom12 = \sum_{l=2}^{12} r_{t-l}$$

Cumulated past performance in the previous year by skipping the most recent month. Rebalanced monthly.

33. **Momentum-Reversal** ($momrev$).

    Follows Jegadeesh e Titman (1993).

$$momrev = \sum_{d=14}^{19} r_{t-l}$$

Buy and hold returns from $t - 19$ to $t - 14$. Updated monthly.

34. **Long-term Reversals** ($lrrev$).

    Follows Bondt e Thaler (1985).

$$lrrev = \sum_{l=13}^{60} r_{t-l}$$

Cumulative returns from $t - 60$ to $t - 13$. Updated monthly.

35. **Value (monthly)** ($valuem$).

    Follows Asness e Frazzini (2013).

$$valuem = \frac{\mathrm{BEQ}_{-3}}{\mathrm{ME}_{-1}}$$

Book-to-market ratio using the most up-to-date prices and book equity (appropriately lagged). Rebalanced monthly.

36. **Share Issuance (monthly)** ($nissm$).

    Follows Pontiff e Woodgate (2008).

$$nissm = \frac{\text{shrout}_{t-13}}{\text{shrout}_{t-1}}$$

where $\text{shrout}$ is the number of shares outstanding. Change in real number of shares outstanding from $t-13$ to $t-1$. Excludes changes in shares due to stock dividends and splits, and companies with no changes in $\text{shrout}$.

37. **PEAD (SUE)** ($sue$).

    Follows Foster, Olsen e Shevlin (1984).

$$sue = \frac{\text{IBQ} - \text{IBQ}_{-12}}{\sigma_{\text{IBQ}_{-24}:\text{IBQ}_{-3}}}$$

where $\text{IBQ}$ is income before extraordinary items (updated quarterly), and $\sigma_{\text{IBQ}_{-24}:\text{IBQ}_{-3}}$ is the standard deviation of $\text{IBQ}$ in the past two years skipping the most recent quarter. Earnings surprises are measured by Standardized Unexpected Earnings (SUE), which is the change in the most recently announced quarterly earnings per share from its value announced four quarters ago divided by the standard deviation of this change in quarterly earnings over the prior eight quarters. Rebalanced monthly.

38. **Return on Book Equity** ($roe$).

    Follows Chen, Novy-Marx e Zhang (2011).

$$roe = \frac{\text{IBQ}}{\text{BEQ}_{-3}}$$

where $\text{IBQ}$ is income before extraordinary items (updated quarterly), and $\text{BEQ}$ is book value of equity. Rebalanced monthly.

39. **Return on Market Equity** ($rome$).

    Follows Chen, Novy-Marx e Zhang (2011).

$$rome = \frac{\text{IBQ}}{\text{ME}_{-4}}$$

where $\mathrm{IBQ}$ is income before extraordinary items (updated quarterly), and $\mathrm{ME}$ is market value of equity. Rebalanced monthly.

40. **Return on Assets** ($roa$).

    Follows Chen, Novy-Marx e Zhang (2011).

$$roa = \frac{\mathrm{IBQ}}{\mathrm{ATC}_{-3}}$$

    Net income scaled by total assets. Updated quarterly.

41. **Short-term Reversal** ($strev$).

    Follows Jegadeesh (1990).

$$strev = r_{t-1}$$

    Return in the previous month. Updated monthly.

42. **Idiosyncratic Volatility** ($ivol$).

    Follows Ang et al. (2006).

$$ivol = \mathrm{std}\left(R_{i,t} - \beta_i R_{M,t} - s_i SMB_t - h_i HML_t\right)$$

    The standard deviation of the residual from firm-level regression of daily stock returns on the daily innovations of the Fama and French three-factor model using the estimation window of three months. Lagged one month.

43. **Beta Arbitrage** ($betaarb$).

    Follows Cooper, Gulen e Schill (2008).

$$betaarb = \beta_{t-60:t-1}$$

    Beta with respect to the CRSP equal-weighted return index. Estimated over the past 60 months (minimum 36 months) using daily data and lagged one month. Updated monthly.

44. **Seasonality** ($season$).

Follows Heston e Sadka (2008).

$$season = \sum_{l=1}^{5} r_{t-l\,\mathrm{x}\,12}$$

Average monthly return in the same calendar month over the last 5 years. As an example, the average return from prior Octobers is used to predict returns this October. The firm needs at least one year of data to be included in the sample. Updated monthly.

45. **Industry Relative Reversals** ($indrrev$).

Follows Da, Liu e Schaumburg (2014).

$$indrrev = r_{-1} - r_{-1}^{ind}$$

where $r$ is the return on a stock and $r^ind$ is return on its industry. Difference between a stocks' prior month's return and the prior month's return of its industry (based on the Fama and French 49 industries). Updated monthly.

46. **Industry Relative Reversals (Low Volatility)** ($indrrevlv$).

Follows Da, Liu e Schaumburg (2014).

$$indrrevlv = r_{-1} - r_{-1}^{ind}$$

if vol ¡ NYSE median, where $r$ is the return on a stock and rind is return on its industry. Difference between a stocks' prior month's return and the prior month's return of its industry (based on the Fama and French 49 industries). Only stocks with idiosyncratic volatility lower than the NYSE median for month are included in the sorts. Updated monthly.

47. **Industry Momentum-Reversal** ($indmomrev$).

Follows Moskowitz e Grinblatt (1999).

$$indmomrev = \text{rank}(indmom) + \text{rank}(indrrevlv)$$

Sum of Fama and French 49 industries ranks on industry momentum and industry relative reversals (low vol). Rebalanced monthly.

48. **Composite Issuance** ($ciss$).

    Follows Daniel e Titman (2006).

$$ciss = \log\left(\frac{\text{ME}_{t-13}}{\text{ME}_{t-60}}\right) - \sum^{6} 0_{l=13} r_{t-l}$$

    where $r$ is the log return on the stock and $\text{ME}$ is total market equity. Updated monthly.

49. **Price** ($price$).

    Follows Blume e Husic (1973).

$$price = \log\left(\frac{\text{ME}}{\text{shrout}}\right)$$

    where $\text{ME}$ is market equity and $\text{shrout}$ is the number of shares outstanding. Log of stock price. Updated monthly.

50. **Firm Age** ($age$).

    Follows Barry e Brown (1984).

$$age = \log(1 + \text{number of months since listing})$$

    The number of months that a firm has been listed in the CRSP database.

51. **Share Volume** ($shvol$).

    Follows Datar, Naik e Radcliffe (1998).

$$shvol = \frac{1}{3}\sum_{i=1}^{3}\frac{\text{volume}_{t-i}}{\text{shrout}_t}$$

Average number of shares traded over the previous three months scaled by shares out-standing. Updated monthly.

This database, which we will call Factors, contains daily data from January 2005 to December 2019.

Figure 1 presents a graph that illustrates the intersection of firms between the Factors Database and the TAQ Returns Database, during the period from January 2005 to December 2019. The graph depicts the evolution of the number of stocks present in each of the bases, as well as the quantity of stocks in which there is a match in both bases and that will be used for the formation of the portfolios. There was a structural break for three days, the 29th of October 2009 and the 30th and 31st of March 2010. On these days, the number of tickers drastically decreased compared to the previous and subsequent days. Therefore, these data were discarded.

Using the TAQ Returns and Factors databases, it is possible to create portfolios of various financial factors that are widely used in the literature. However, the construction of these portfolios can vary greatly between studies. To ensure standardization in the creation of these portfolios, we adopted four different strategies. The first two strategies consist of forming portfolios by longing (shorting) stocks that are above (below) the median. The difference between these strategies will be the way of weighting the assets, which can be equal weighted or value weighted. The other two methodologies will use the 30th and 70th percentiles as breakpoints for stock selection, that is, by longing (shorting) stocks in the upper three deciles (lower three deciles).

Let $h_f^b$ (high) be the vector of tickers that are above the breakpoint rule $b$ and $l_f^b$ (low) be the vector of tickers that are below the breakpoint rule $b$ for some specific factor $f$, where the superscript denotes the breakpoint rule $b = m, t$ ($m$ represents the median as the breakpoint rule and $t$ represents the three deciles as the breakpoint rule) and the subscript denotes the factor $f$. Also denote by $r_i$ the return on the share of the firm $i$ and $\omega_i \in \omega$ the weight based on the Market Cap of the firm $i$.

$$\omega_i = \frac{ME_i}{\sum_{i\in v} ME_i}$$

for $v = h_f^b, l_f^b$.

Therefore, portfolios are formed as follows.

- **Equal Weighted Portfolio with Median Breakpoint**

$$f \text{ factor} = \frac{1}{\text{len}(h_f^m)} \sum_{i \in h_f^m} r_i - \frac{1}{\text{len}(l_f^m)} \sum_{i \in l_f^m} r_i$$

- **Value Weighted Portfolio with Median Breakpoint**

$$f \text{ factor} = \sum_{i \in h_f^m} \omega_i \cdot r_i - \sum_{i \in l_f^m} \omega_i \cdot r_i$$

- **Equal Weighted Portfolio with Three Decile Breakpoint**

$$f \text{ factor} = \frac{1}{\text{len}(h_f^t)} \sum_{i \in h_f^t} r_i - \frac{1}{\text{len}(l_f^t)} \sum_{i \in l_f^t} r_i$$

- **Value Weighted Portfolio with Three Decile Breakpoint**

$$f \text{ factor} = \sum_{i \in h_f^t} \omega_i \cdot r_i - \sum_{i \in l_f^t} \omega_i \cdot r_i$$

for $f =$ Size, Value, ..., Share Volume, where $\text{len}(\cdot)$ is a function that returns the amount of elements of some vector.

The figures 2, 3,..., 52 show the return distributions of the portfolios formed for all the financial factors mentioned above. For each financial factor, we will have four portfolios, one for each empirical strategy. Except for the factors whose variable is binary, that is, they only receive the values 1 and 10. These factors have portfolios that are independent of the breakpoint rule.

# 3 Conclusion

The main objective of this summer paper was to document the data manipulation process. We describe how we create portfolios related to some of the key financial factors documented in the literature. The final database contains daily intraday stock returns and financial factor portfolios.

In Section 1, we briefly mentioned the goals after finalizing the database. The intention is to model through LASSO which predictor candidates among stock returns and portfolios of financial factors help to improve the predictive power of high-frequency asset pricing. Along the same line, we will be able to verify whether, unlike Chinco, Clark-Joseph e Ye (2019), when including financial factors there is an increase in predictive power.

Next steps include adding more robust methods to evaluate which predictor candidates increase predictive power, such as the Double Selection LASSO featured in Feng, Giglio e Xiu (2020). This procedure considers model selection errors by selecting predictors that not only help explain the cross-section of expected returns, but are also useful in mitigating the problem of omitted variable bias.

Furthermore, as LASSO is a tool whose predictor selection is based purely on a statistical rule, we would like to check whether there is economic meaning for the selected predictors through twitter data.

# References


ANG, A. et al. The cross-section of volatility and expected returns. *The journal of finance*, Wiley Online Library, v. 61, n. 1, p. 259–299, 2006.

ASNESS, C.; FRAZZINI, A. The devil in hml's details. *The Journal of Portfolio Management*, Institutional Investor Journals Umbrella, v. 39, n. 4, p. 49–68, 2013.

BARRY, C. B.; BROWN, S. J. Differential information and the small firm effect. *Journal of financial economics*, Elsevier, v. 13, n. 2, p. 283–294, 1984.

BASU, S. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The journal of Finance*, Wiley Online Library, v. 32, n. 3, p. 663–682, 1977.

BHANDARI, L. C. Debt/equity ratio and expected common stock returns: Empirical evidence. *The journal of finance*, Wiley Online Library, v. 43, n. 2, p. 507–528, 1988.

BLUME, M. E.; HUSIC, F. Price, beta, and exchange listing. *The Journal of Finance*, JSTOR, v. 28, n. 2, p. 283–299, 1973.

BONDT, W. F. D.; THALER, R. Does the stock market overreact? *The Journal of finance*, Wiley Online Library, v. 40, n. 3, p. 793–805, 1985.

CARHART, M. M. On persistence in mutual fund performance. *The Journal of finance*, Wiley Online Library, v. 52, n. 1, p. 57–82, 1997.

CHEN, L.; NOVY-MARX, R.; ZHANG, L. An alternative three-factor model. *Available at SSRN 1418117*, 2011.

CHINCO, A.; CLARK-JOSEPH, A. D.; YE, M. Sparse signals in the cross-section of returns. *The Journal of Finance*, Wiley Online Library, v. 74, n. 1, p. 449–492, 2019.

COOPER, M. J.; GULEN, H.; SCHILL, M. J. Asset growth and the cross-section of stock returns. *the Journal of Finance*, Wiley Online Library, v. 63, n. 4, p. 1609–1651, 2008.

DA, Z.; LIU, Q.; SCHAUMBURG, E. A closer look at the short-term return reversal. *Management science*, INFORMS, v. 60, n. 3, p. 658–674, 2014.

DANIEL, K.; TITMAN, S. Market reactions to tangible and intangible information. *The Journal of Finance*, Wiley Online Library, v. 61, n. 4, p. 1605–1643, 2006.

DATAR, V. T.; NAIK, N. Y.; RADCLIFFE, R. Liquidity and stock returns: An alternative test. *Journal of financial markets*, Elsevier, v. 1, n. 2, p. 203–219, 1998.

DECHOW, P. M.; KOTHARI, S. P.; WATTS, R. L. The relation between earnings and cash flows. *Journal of accounting and Economics*, Elsevier, v. 25, n. 2, p. 133–168, 1998.

FAIRFIELD, P. M.; WHISENANT, J. S.; YOHN, T. L. Accrued earnings and growth: Implications for future profitability and market mispricing. *The accounting review*, v. 78, n. 1, p. 353–371, 2003.

FAMA, E. F.; FRENCH, K. R. The cross-section of expected stock returns. *the Journal of Finance*, Wiley Online Library, v. 47, n. 2, p. 427–465, 1992.

FAMA, E. F.; FRENCH, K. R. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, Elsevier, v. 33, n. 1, p. 3–56, 1993.

FAMA, E. F.; FRENCH, K. R. A five-factor asset pricing model. *Journal of financial economics*, Elsevier, v. 116, n. 1, p. 1–22, 2015.

FENG, G.; GIGLIO, S.; XIU, D. Taming the factor zoo: A test of new factors. *The Journal of Finance*, Wiley Online Library, v. 75, n. 3, p. 1327–1370, 2020.

FOSTER, G.; OLSEN, C.; SHEVLIN, T. Earnings releases, anomalies, and the behavior of security returns. *Accounting Review*, JSTOR, p. 574–603, 1984.

HAUGEN, R. A.; BAKER, N. L. Commonality in the determinants of expected stock returns. *Journal of financial economics*, Elsevier, v. 41, n. 3, p. 401–439, 1996.

HESTON, S. L.; SADKA, R. Seasonality in the cross-section of stock returns. *Journal of Financial Economics*, Elsevier, v. 87, n. 2, p. 418–445, 2008.

HIRSHLEIFER, D. et al. Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics*, Elsevier, v. 38, p. 297–331, 2004.

IKENBERRY, D.; LAKONISHOK, J.; VERMAELEN, T. Market underreaction to open market share repurchases. *Journal of financial economics*, Elsevier, v. 39, n. 2-3, p. 181–208, 1995.

JEGADEESH, N. Evidence of predictable behavior of security returns. *The Journal of finance*, Wiley Online Library, v. 45, n. 3, p. 881–898, 1990.

JEGADEESH, N.; TITMAN, S. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, Wiley Online Library, v. 48, n. 1, p. 65–91, 1993.

JR, W. C. B.; MUKHERJI, S.; RAINES, G. A. Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal*, Taylor & Francis, v. 52, n. 2, p. 56–60, 1996.

LAKONISHOK, J.; SHLEIFER, A.; VISHNY, R. W. Contrarian investment, extrapolation, and risk. *The journal of finance*, Wiley Online Library, v. 49, n. 5, p. 1541–1578, 1994.

LINTNER, J. Security prices, risk, and maximal gains from diversification. *The journal of finance*, JSTOR, v. 20, n. 4, p. 587–615, 1965.

MARKOWITZ, H. Portfolio selection. *Journal of Finance*, v. 7, n. 1, p. 77–91, 1952.

MOSKOWITZ, T. J.; GRINBLATT, M. Do industries explain momentum? *The Journal of finance*, Wiley Online Library, v. 54, n. 4, p. 1249–1290, 1999.

MOSSIN, J. Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, JSTOR, p. 768–783, 1966.

NARANJO, A.; NIMALENDRAN, M.; RYNGAERT, M. Stock returns, dividend yields, and taxes. *The Journal of Finance*, Wiley Online Library, v. 53, n. 6, p. 2029–2057, 1998.

NOVY-MARX, R. The other side of value: The gross profitability premium. *Journal of financial economics*, Elsevier, v. 108, n. 1, p. 1–28, 2013.

PIOTROSKI, J. D. Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, JSTOR, p. 1–41, 2000.

PONTIFF, J.; WOODGATE, A. Share issuance and cross-sectional returns. *The Journal of Finance*, Wiley Online Library, v. 63, n. 2, p. 921–945, 2008.

SHARPE, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, Wiley Online Library, v. 19, n. 3, p. 425–442, 1964.

SLOAN, R. G. Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting review*, JSTOR, p. 289–315, 1996.

SOLIMAN, M. T. The use of dupont analysis by market participants. *The accounting review*, v. 83, n. 3, p. 823–853, 2008.

SPIESS, D. K.; AFFLECK-GRAVES, J. The long-run performance of stock returns following debt offerings. *Journal of Financial Economics*, Elsevier, v. 54, n. 1, p. 45–73, 1999.

XING, Y. Interpreting the value effect through the q-theory: An empirical investigation. *The Review of Financial Studies*, Society for Financial Studies, v. 21, n. 4, p. 1767–1795, 2008.

# A Appendix

## A.1 LASSO

LASSO is a categorized procedure within the set of penalized regressions. Penalized Least Squares is a punishment procedure that adds to the OLS regression a component that penalizes weak coefficients. The Penalty Least Squares estimator is obtained through

$$\widehat{\boldsymbol{\beta}}(\lambda)=\underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min}\left[\sum_{i=1}^{n}\left(Y_i-\boldsymbol{\beta}'\boldsymbol{X}_i\right)^2+\sum_{j=1}^{p}p_\lambda\left(\left|\beta_j\right|;\boldsymbol{\alpha},\ \text{data}\right)\right]$$

where $p_\lambda\left(\left|\beta_j\right|;\boldsymbol{\alpha},\text{data}\right)$ is a non-negative penalty function indexed by the regularization parameter $\lambda$ and could rely on both the data and additional hyperparameters. For example,

- Ridge

  $p_\lambda\left(\left|\beta_j\right|;\boldsymbol{\alpha},\text{data}\right)=\lambda\left|\beta_j\right|^2$

- LASSO

  $p_\lambda\left(\left|\beta_j\right|;\boldsymbol{\alpha},\text{data}\right)=\lambda\left|\beta_j\right|$

- LASSO and Ridge combined

  $p_\lambda\left(\left|\beta_j\right|;\boldsymbol{\alpha},\text{data}\right)=\alpha\lambda\left|\beta_j\right|+(1-\alpha)\lambda\left|\beta_j\right|^2$

The regularization parameter $\lambda$ controls the number of parameters in the model. If $\lambda=\infty$, then no parameters enter the model, and if $\lambda=0$, then the parameters are simply OLS estimators.

## A.2 Figures

Figure 1: Number of stocks in each Database

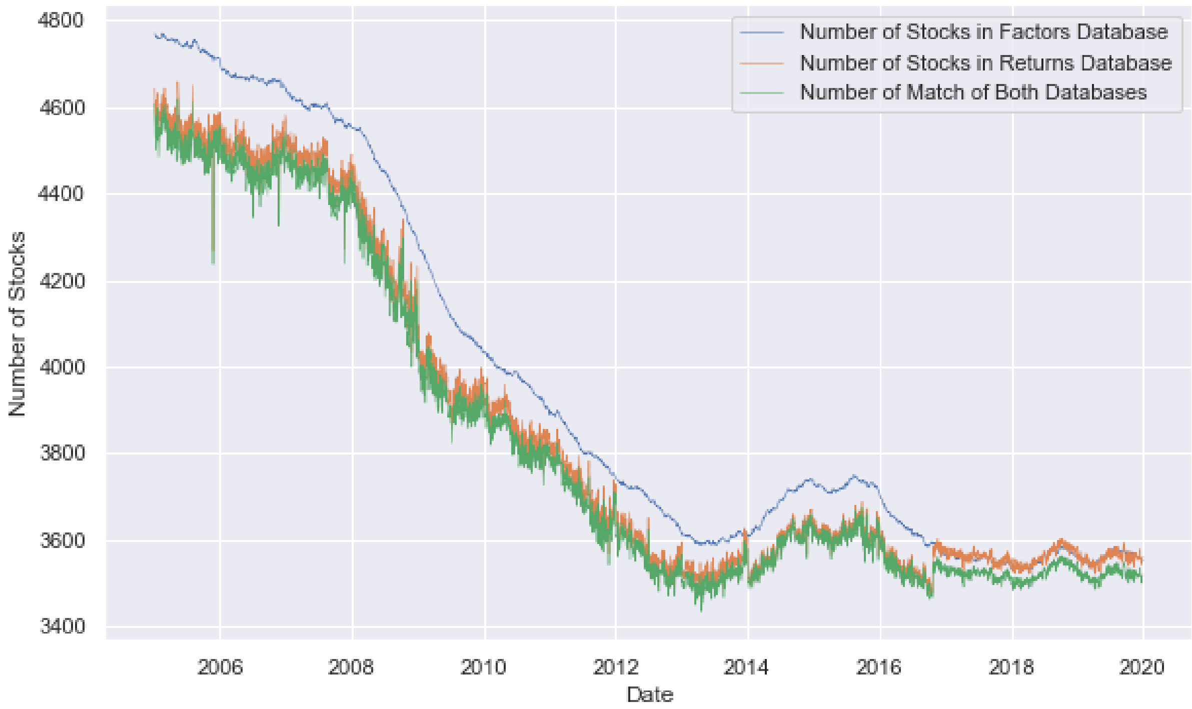


Figure 2: Size Factor Returns Distribution

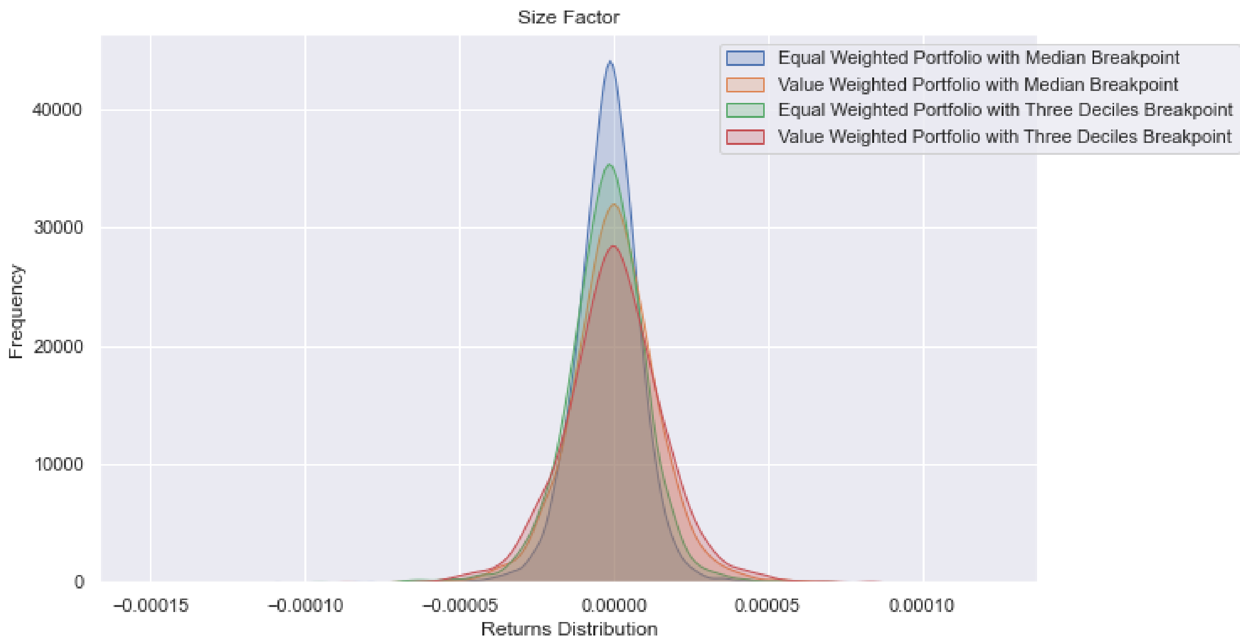

Figure 3: Value (annual) Factor Returns Distribution

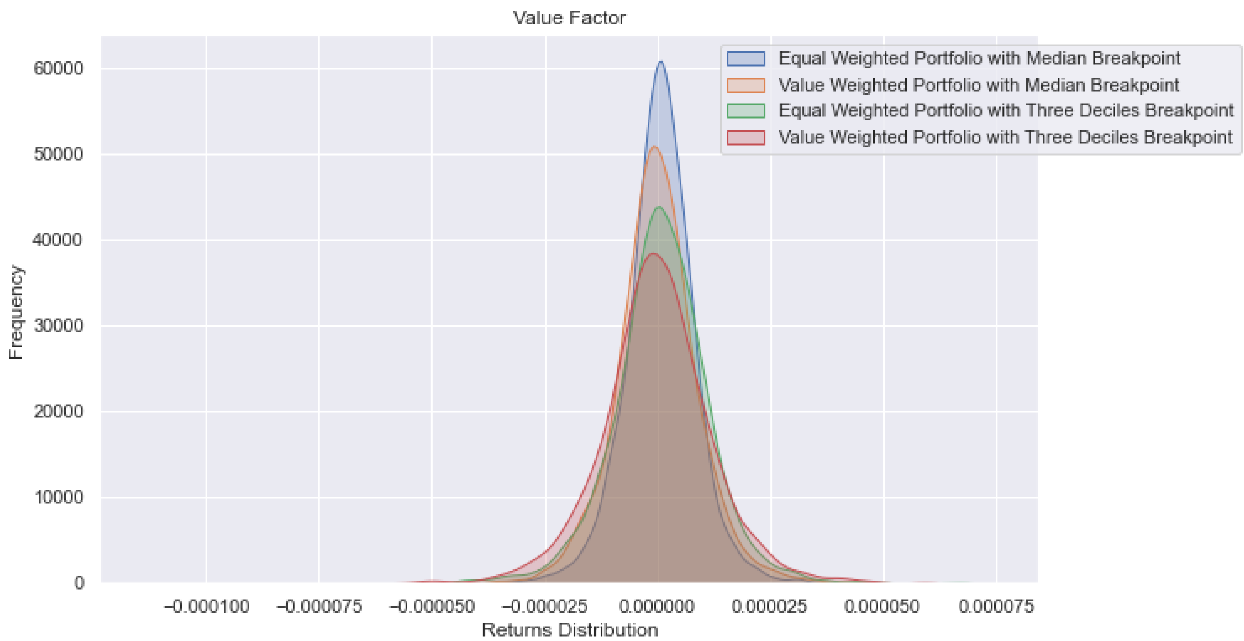


Figure 4: Gross Profitability Factor Returns Distribution

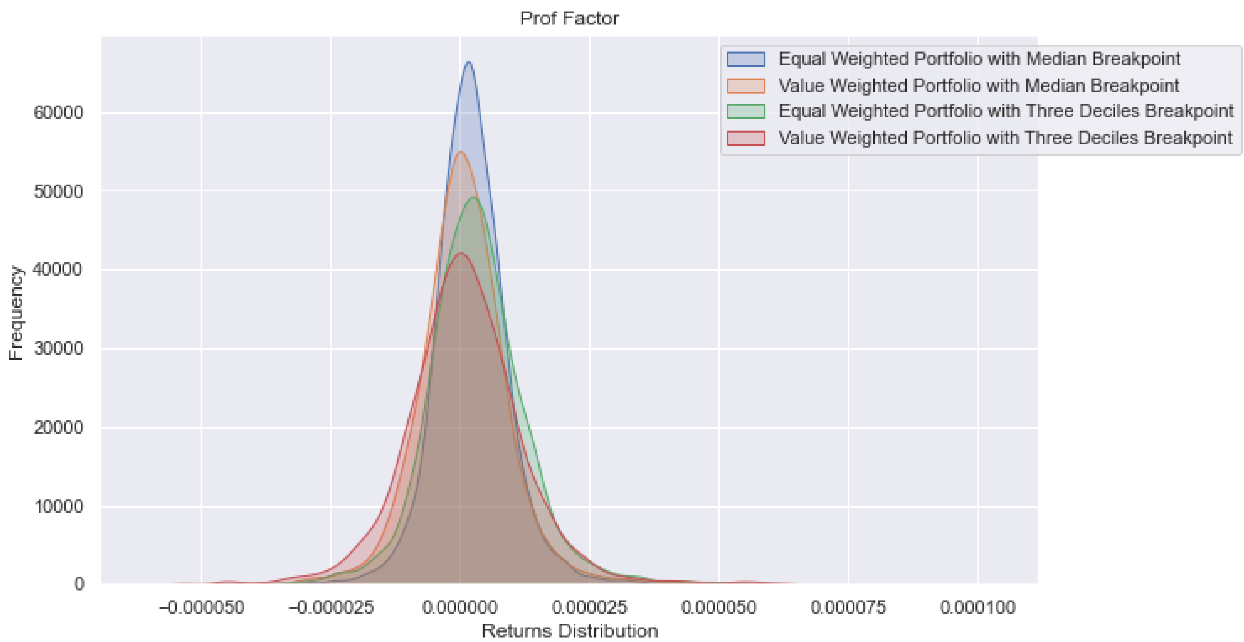

Figure 5: Cash Flow Duration Factor Returns Distribution

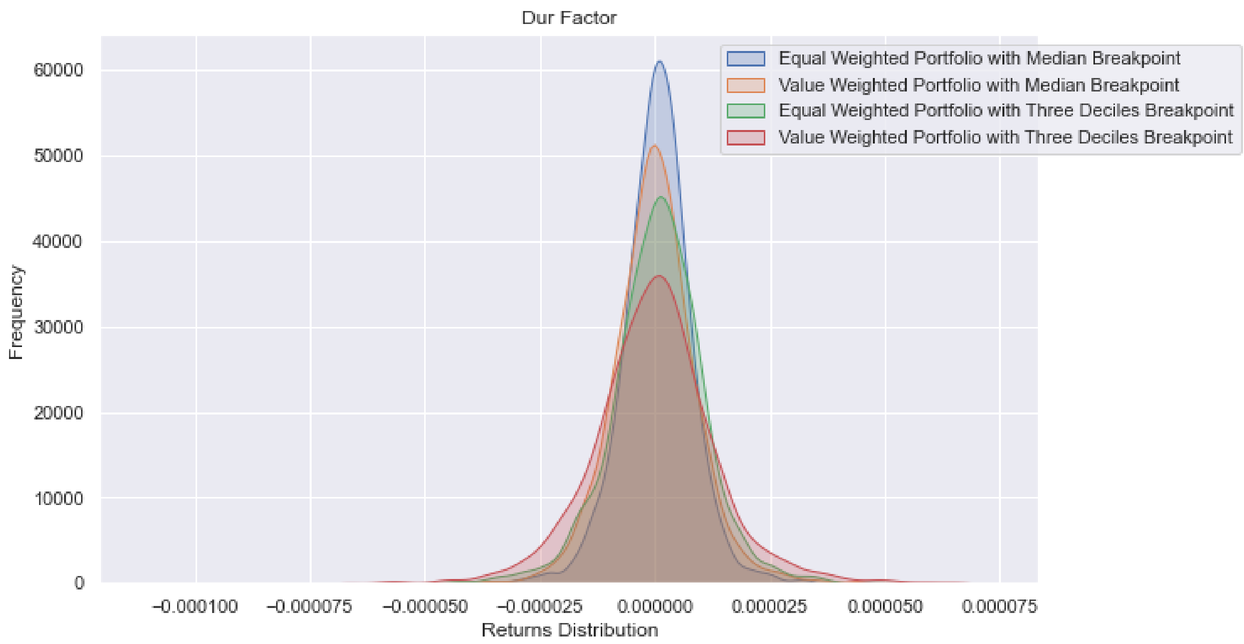


Figure 6: Value-Profitability Factor Returns Distribution

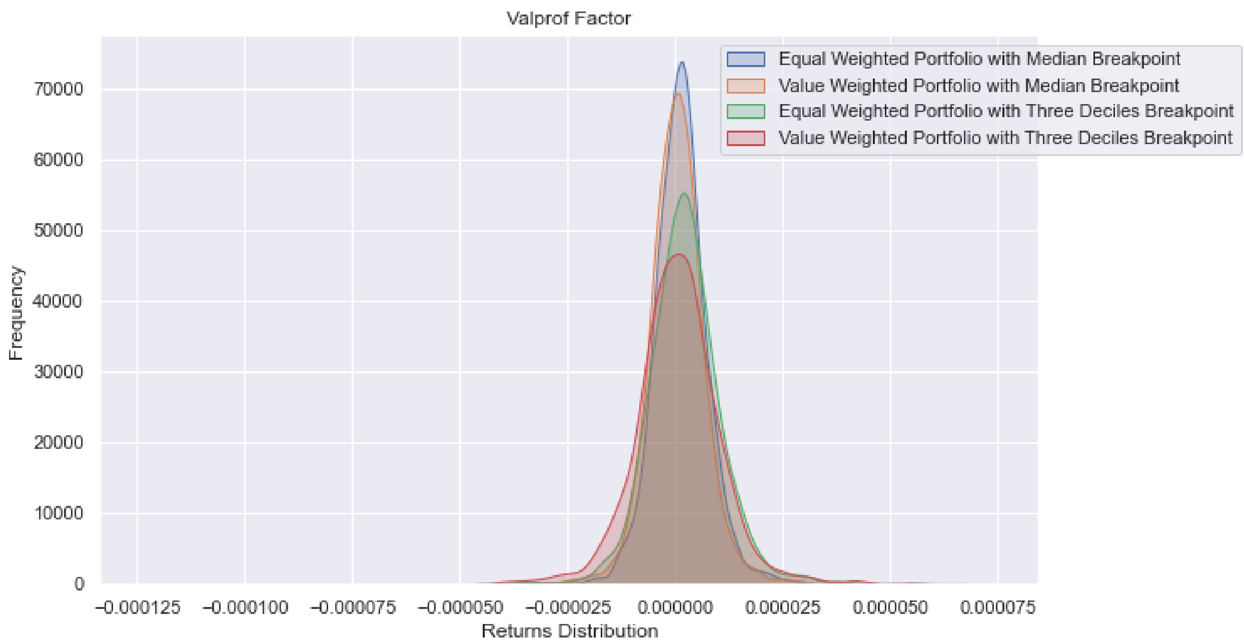

Figure 7: Piotroski's F-score Factor Returns Distribution

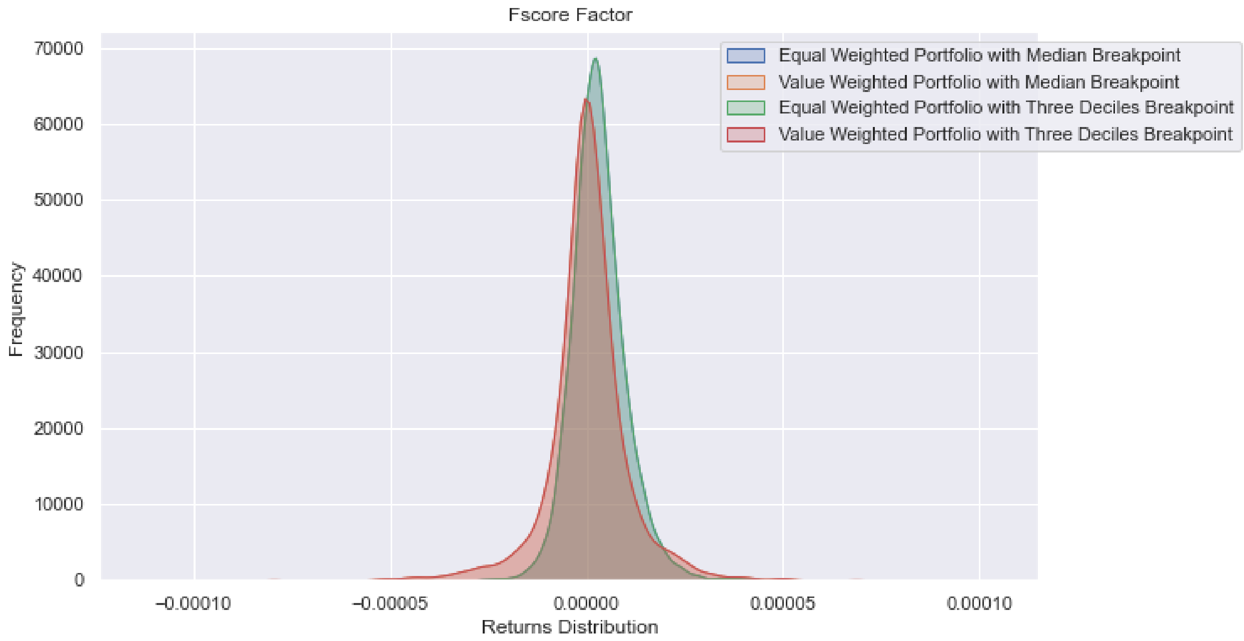


Figure 8: Debit Issuance Factor Returns Distribution

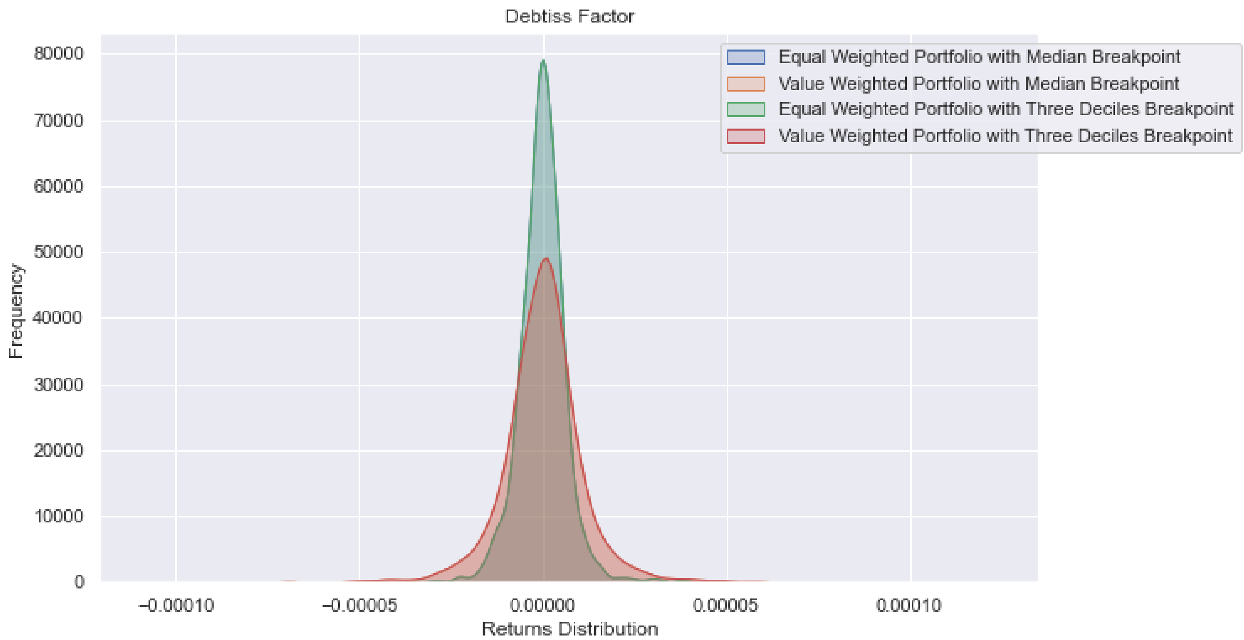

Figure 9: Share Repurchases Factor Returns Distribution

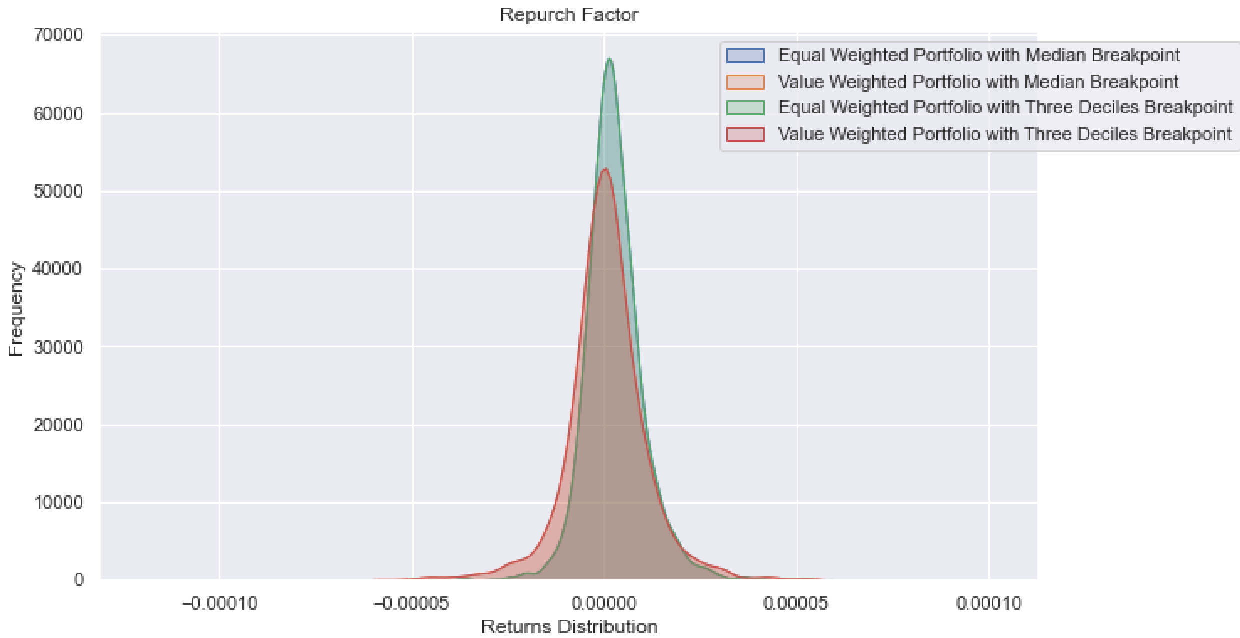


Figure 10: Share Issuance (annual) Factor Returns Distribution

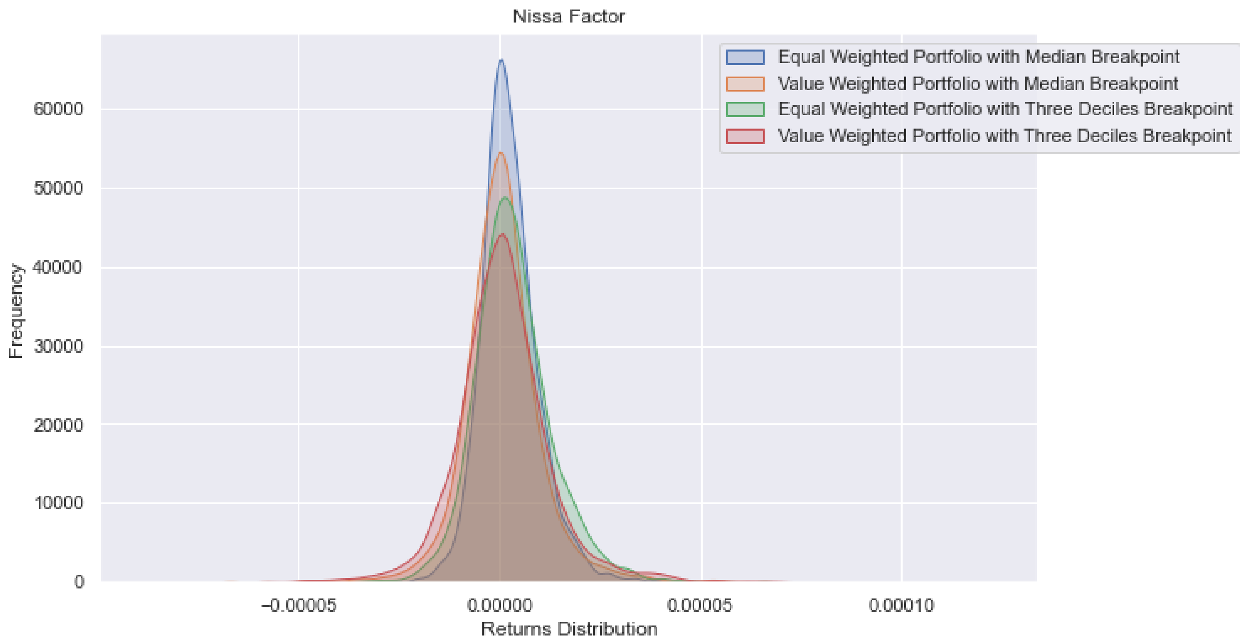

Figure 11: Accruals Factor Returns Distribution

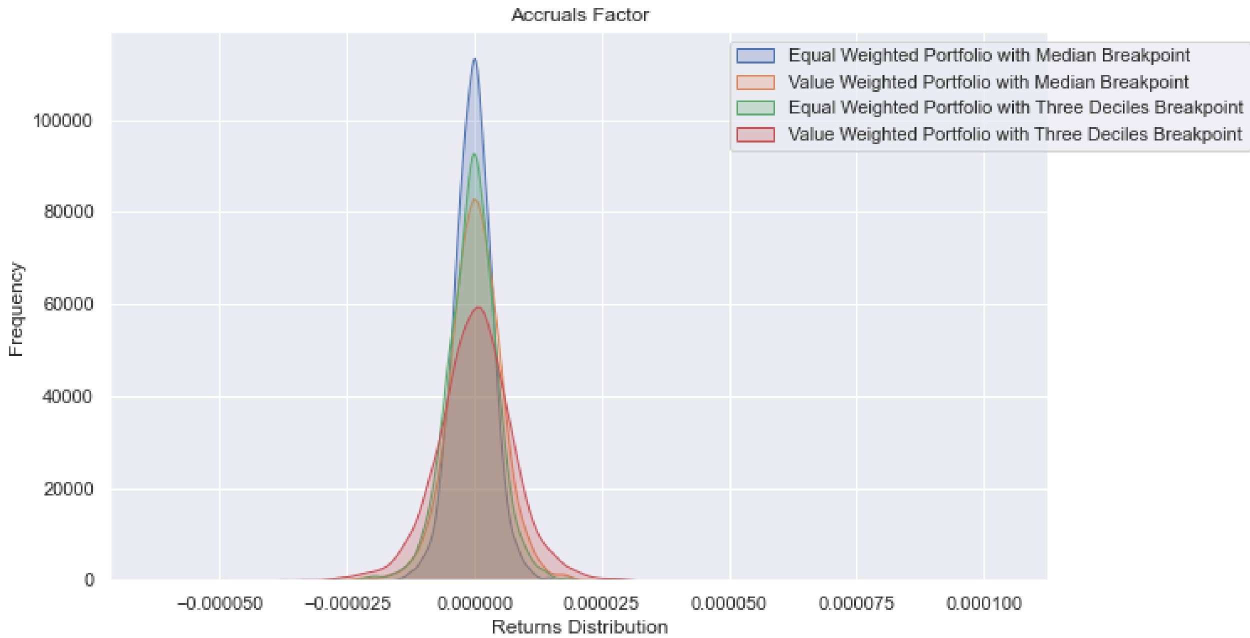


Figure 12: Asset Growth Factor Returns Distribution

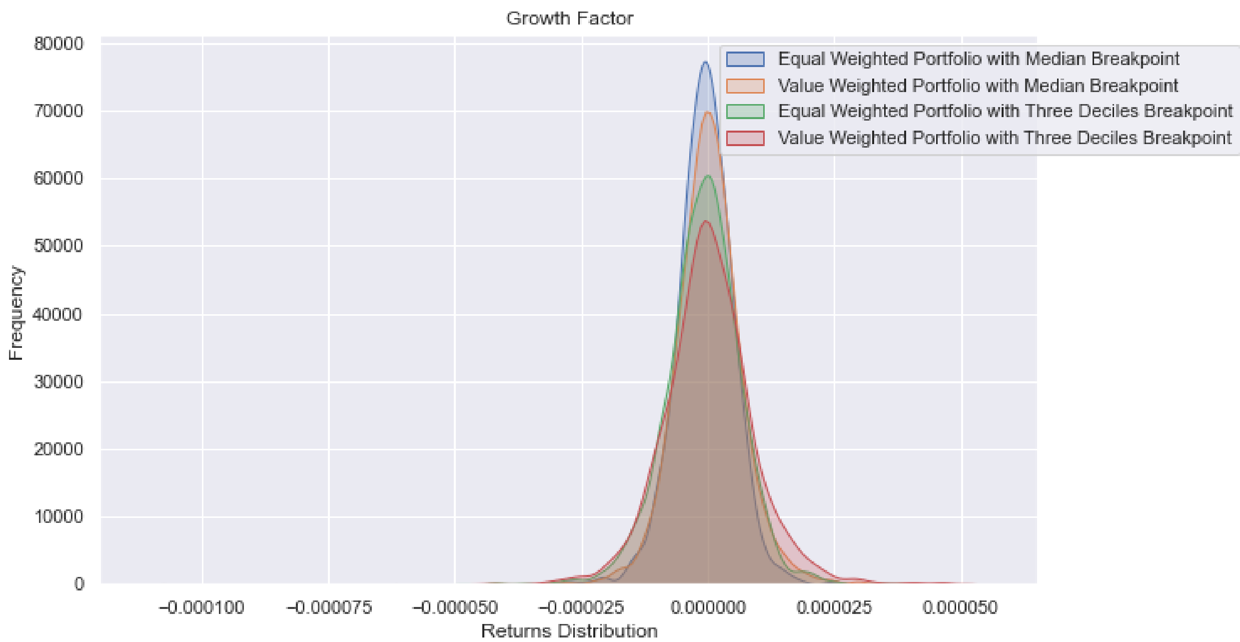

Figure 13: Asset Turnover Factor Returns Distribution

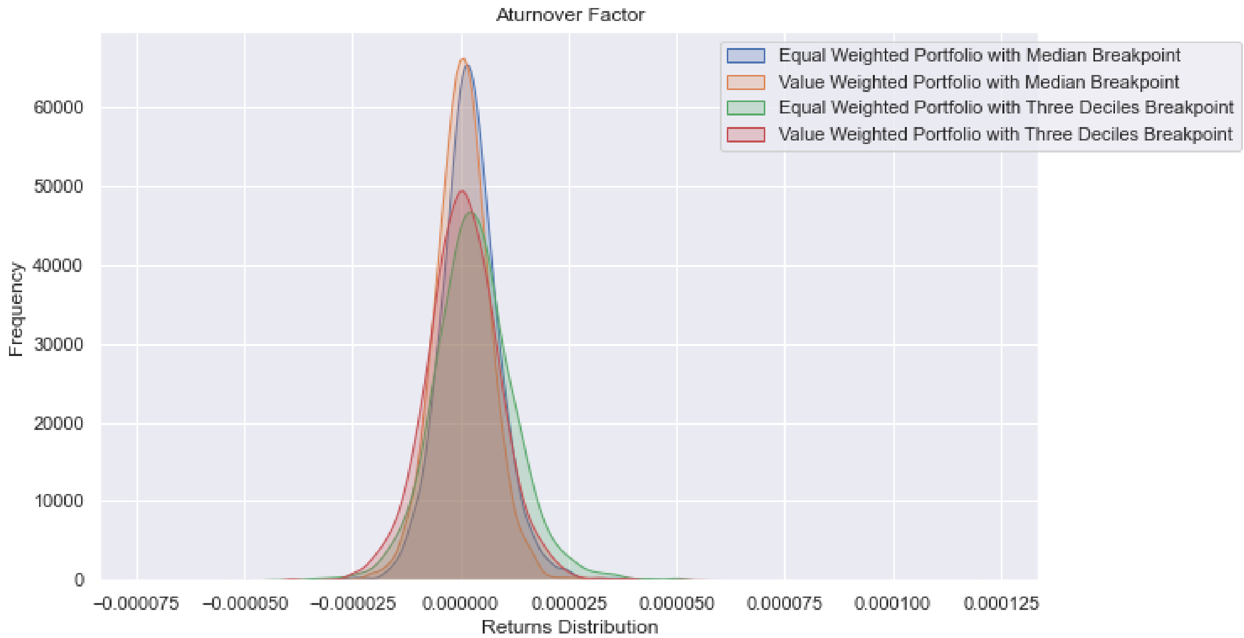


Figure 14: Gross Margins Factor Returns Distribution

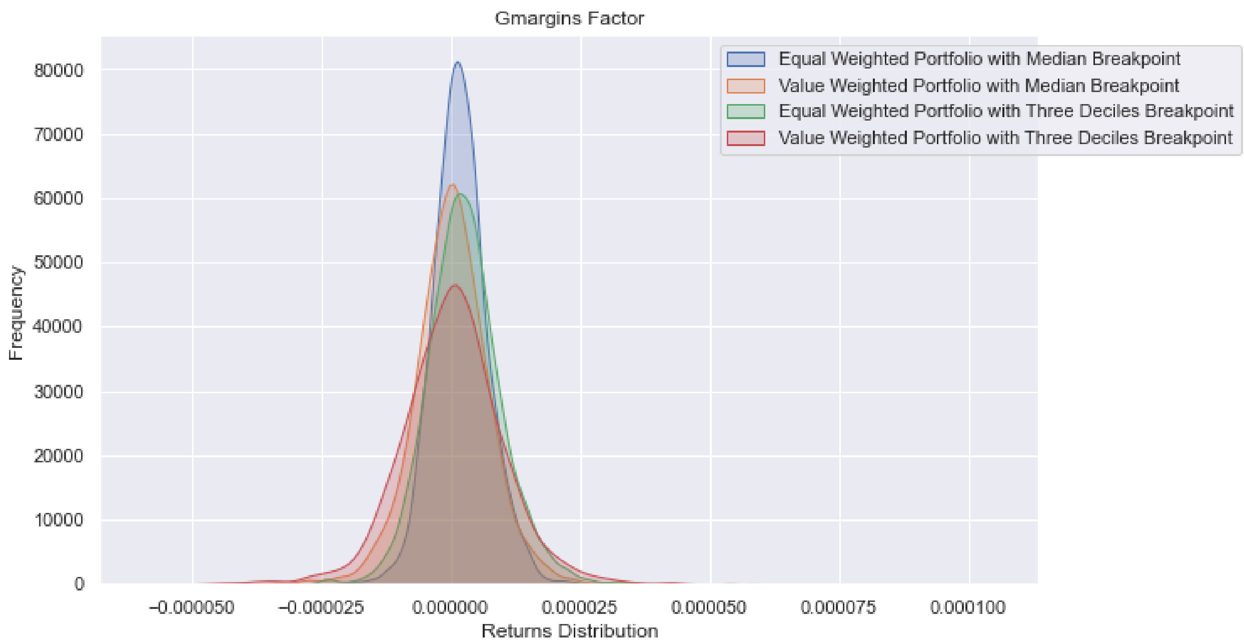

Figure 15: Dividend Yield Factor Returns Distribution

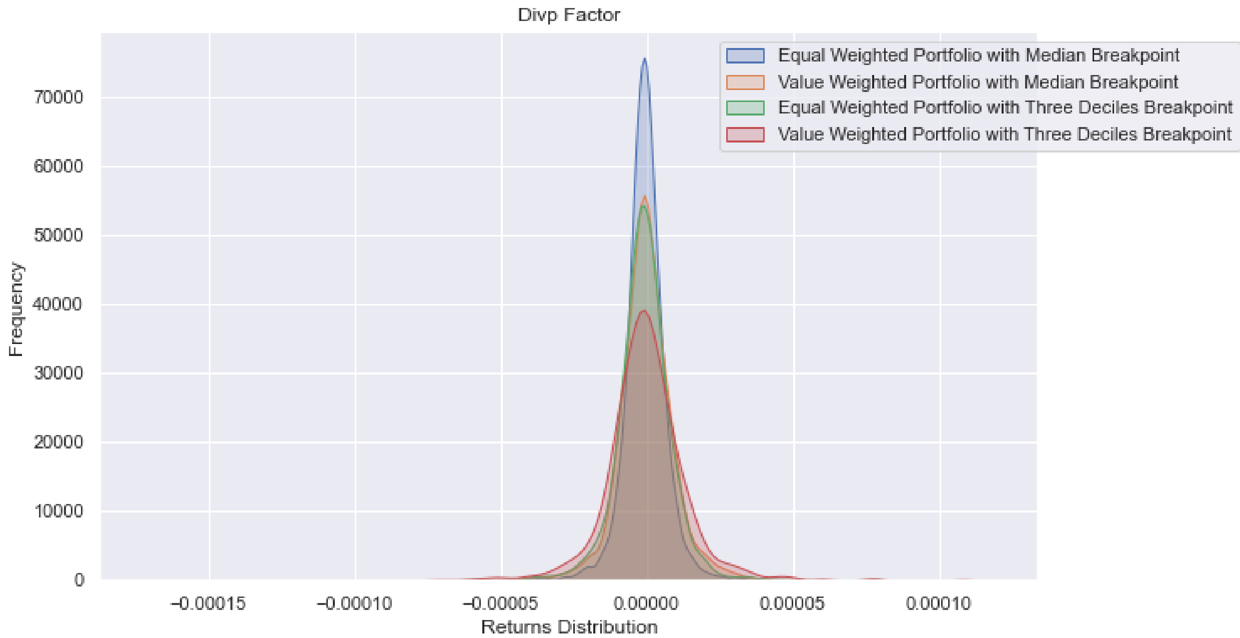


Figure 16: Earnings/Price Factor Returns Distribution

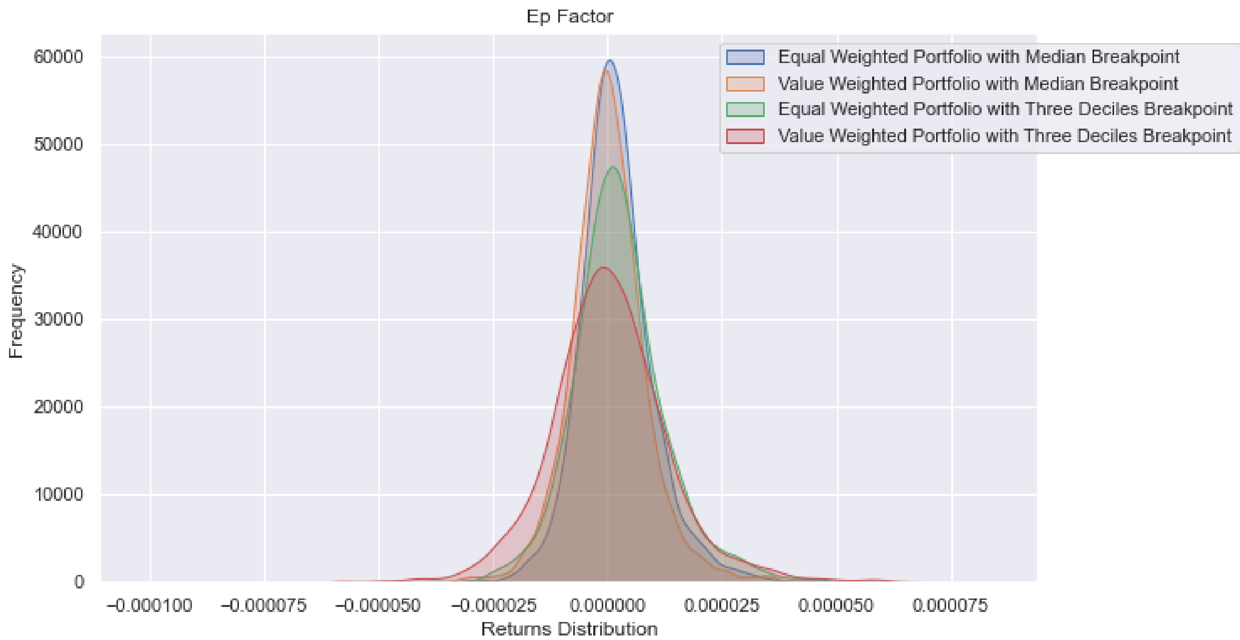

Figure 17: Cash Flow / Market Value of Equity Factor Returns Distribution

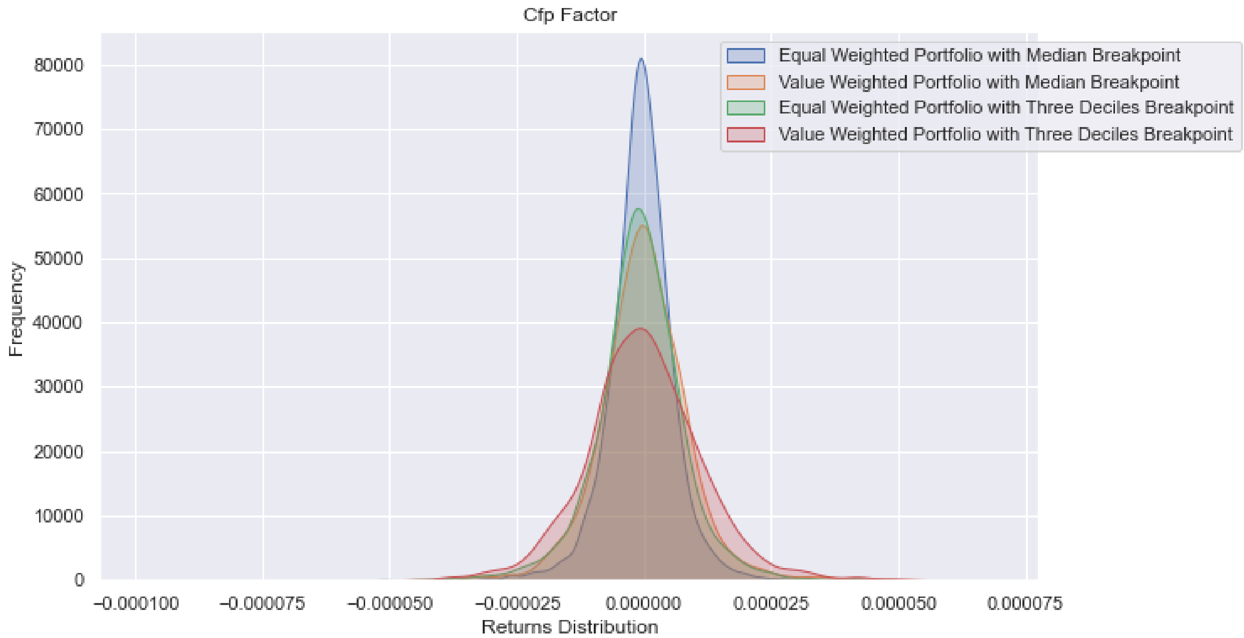


Figure 18: Net Operating Assets Factor Returns Distribution

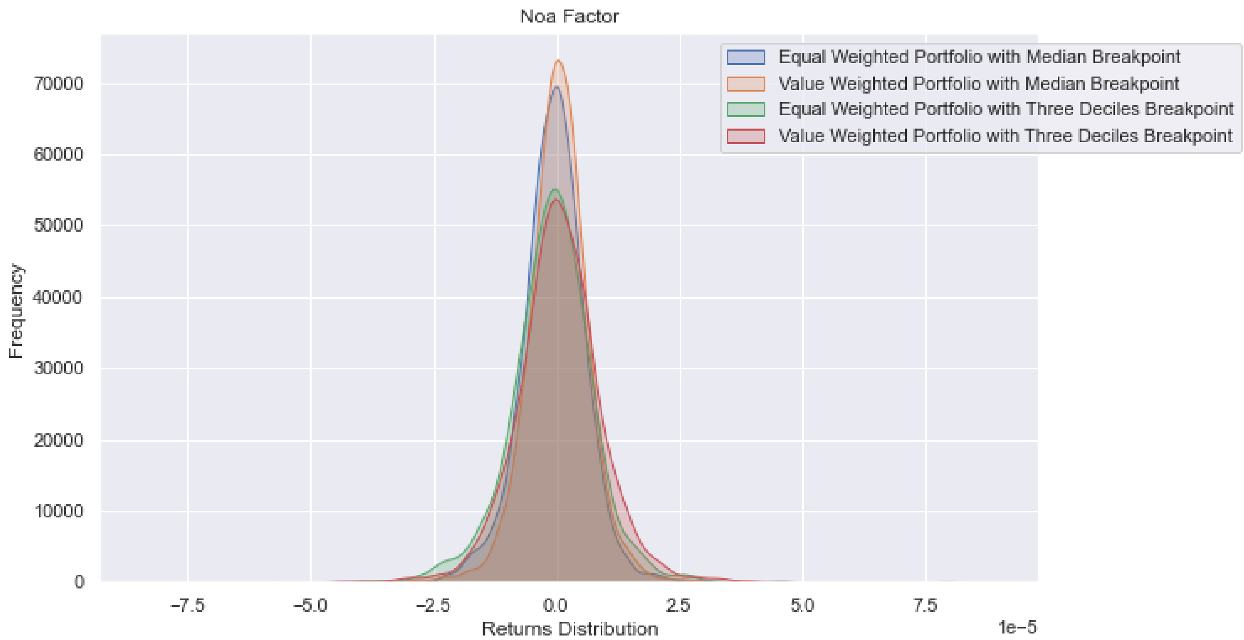

Figure 19: Investment Factor Returns Distribution

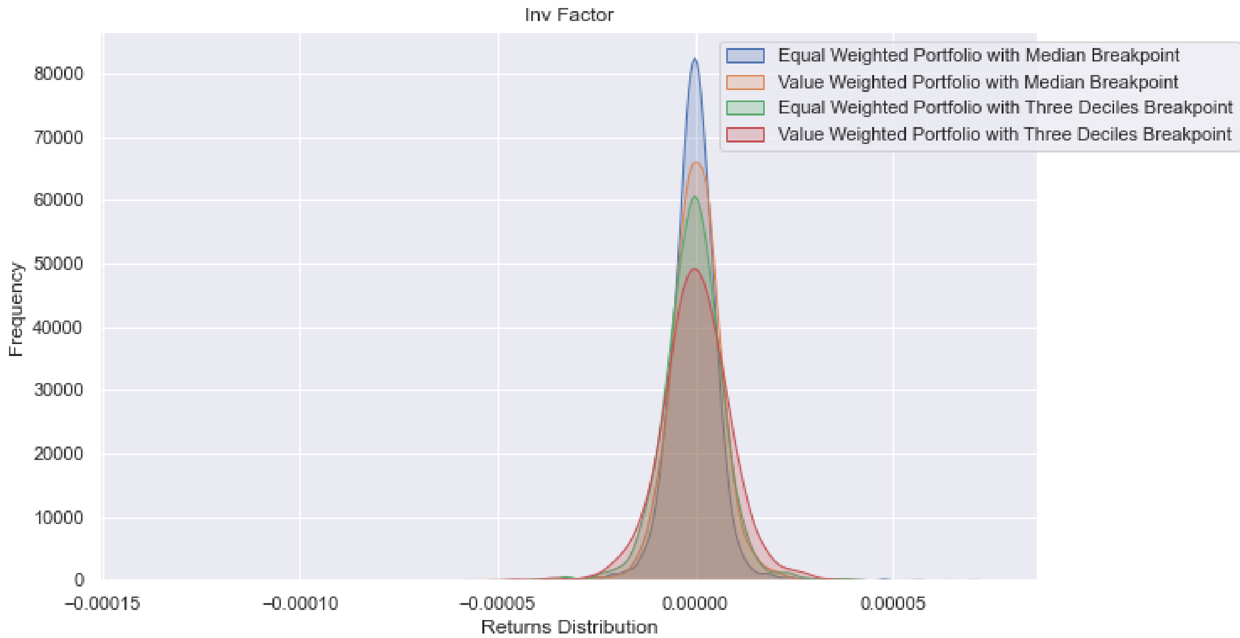


Figure 20: Investment-to-Capital Factor Returns Distribution

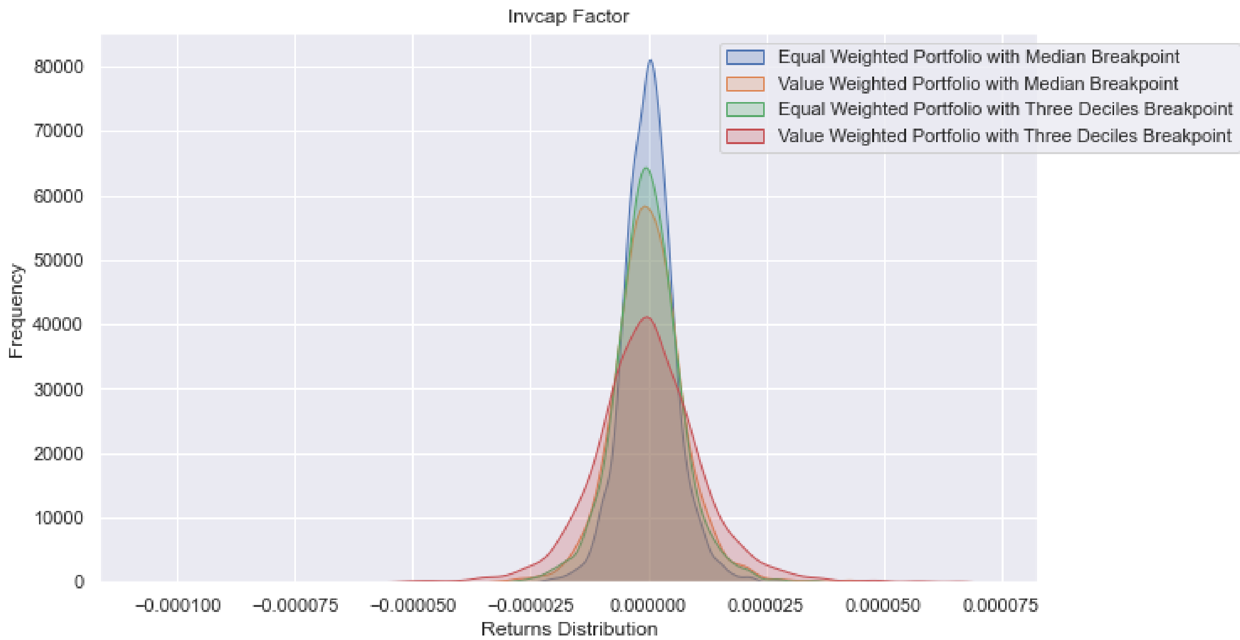

Figure 21: Investment Growth Factor Returns Distribution

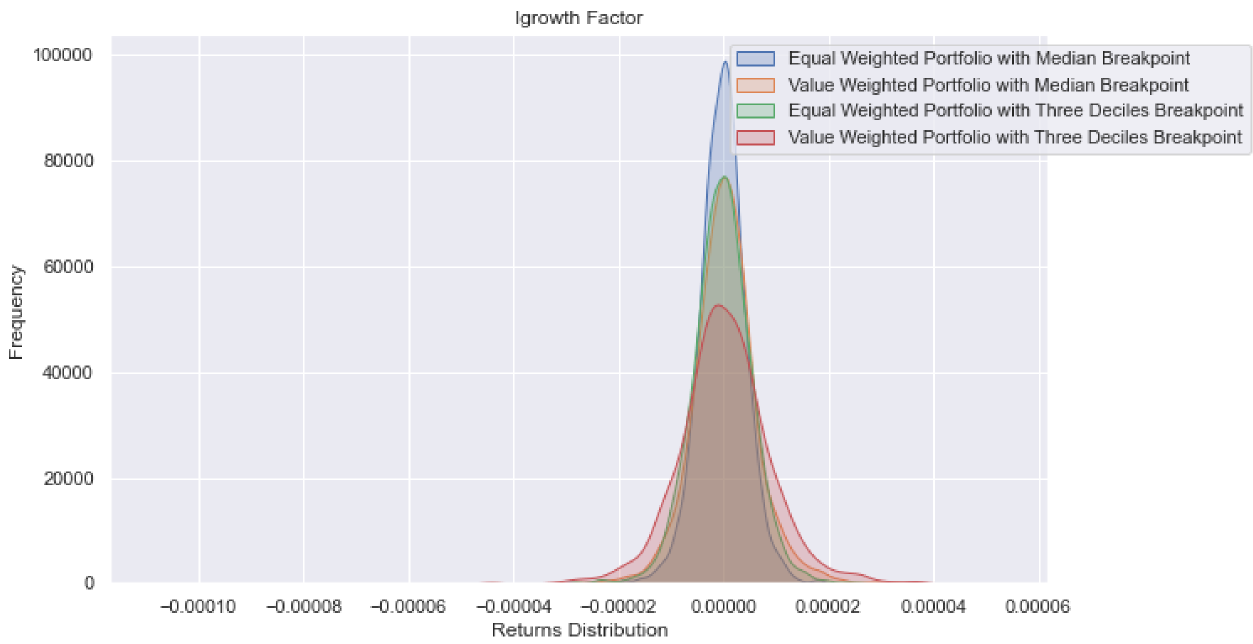


Figure 22: Sales Growth Factor Returns Distribution

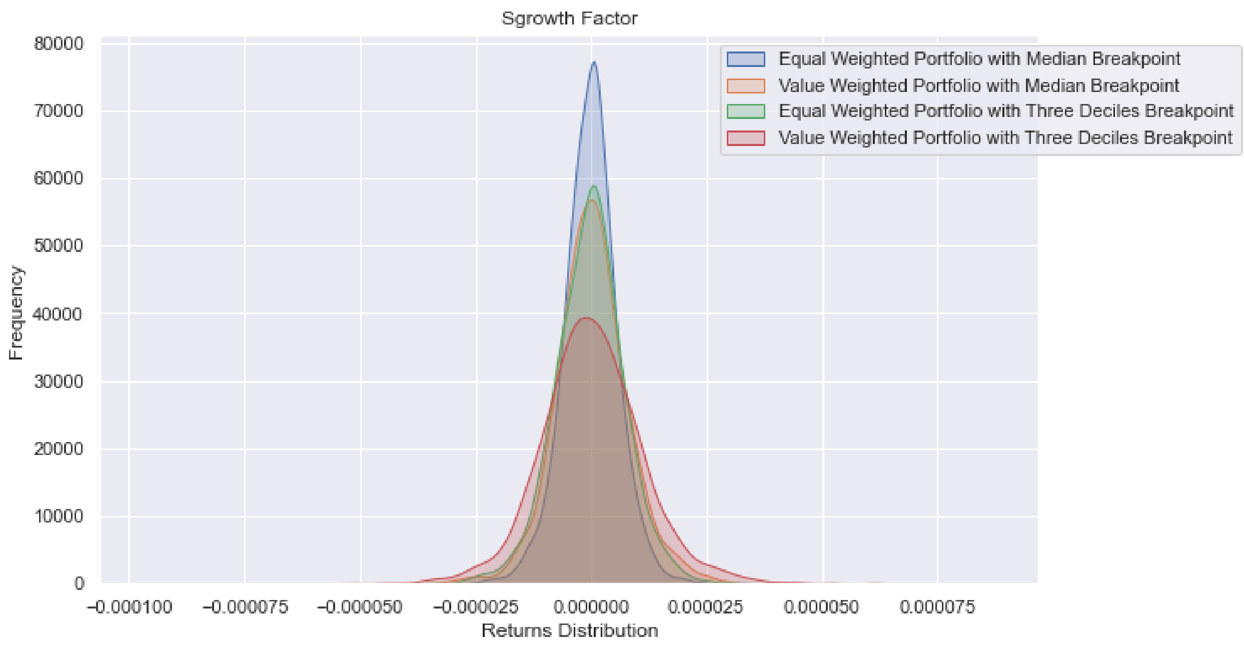

Figure 23: Leverage Factor Returns Distribution

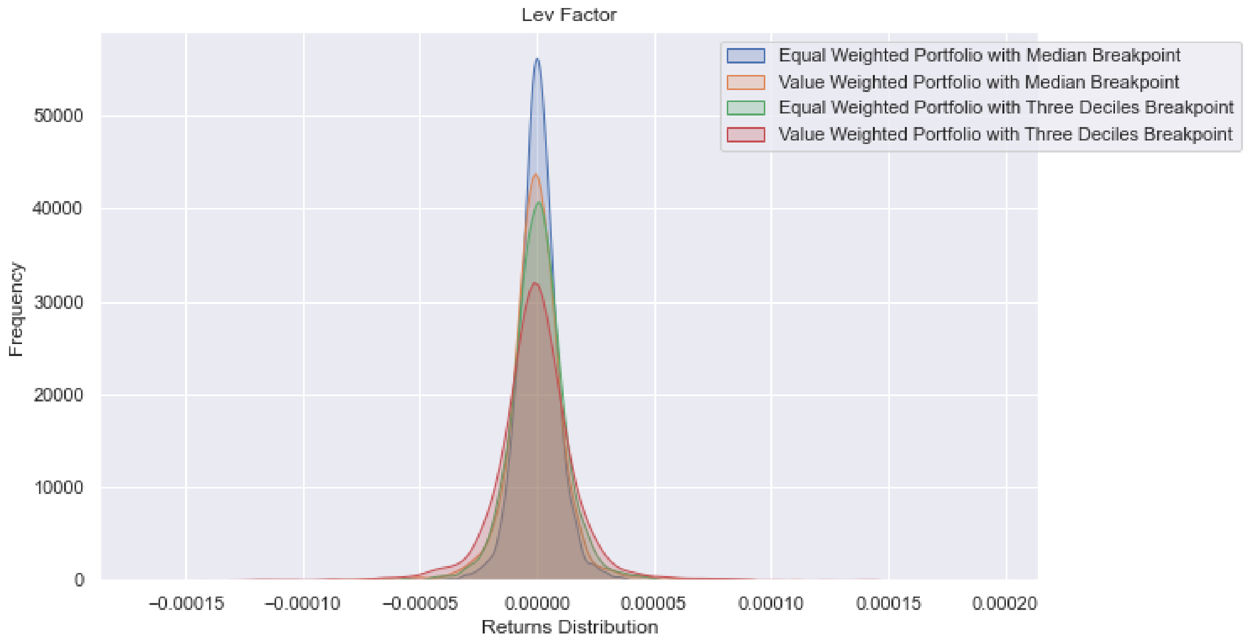


Figure 24: Return on Assets (annual) Factor Returns Distribution

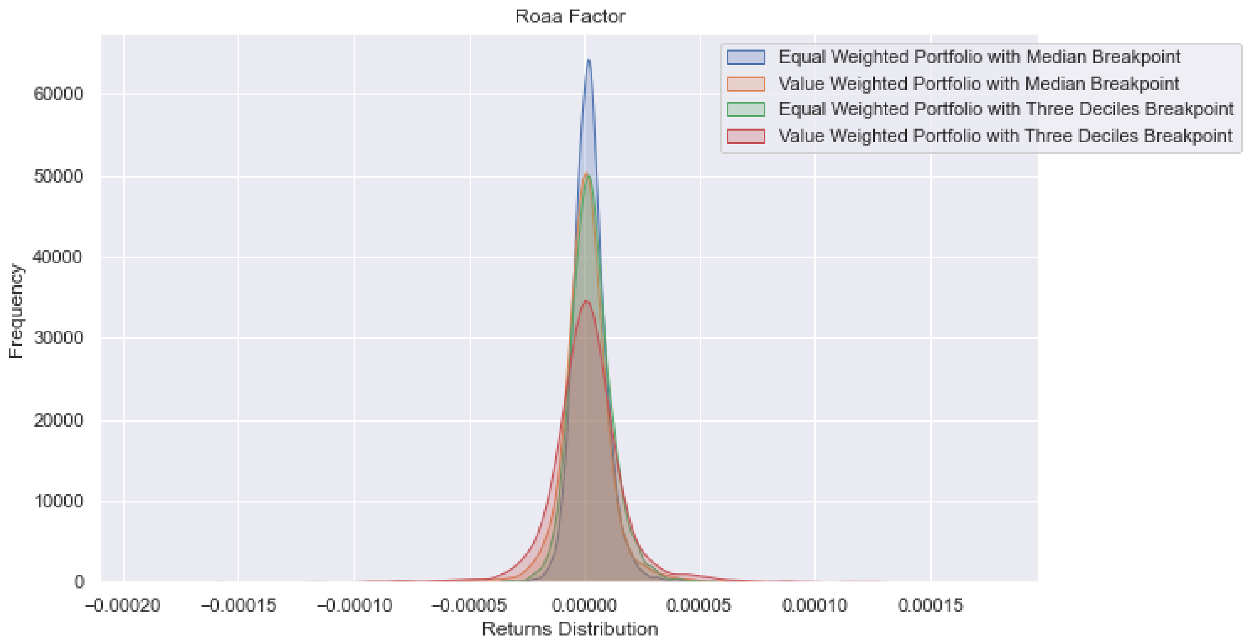

Figure 25: Return on Equity (annual) Factor Returns Distribution

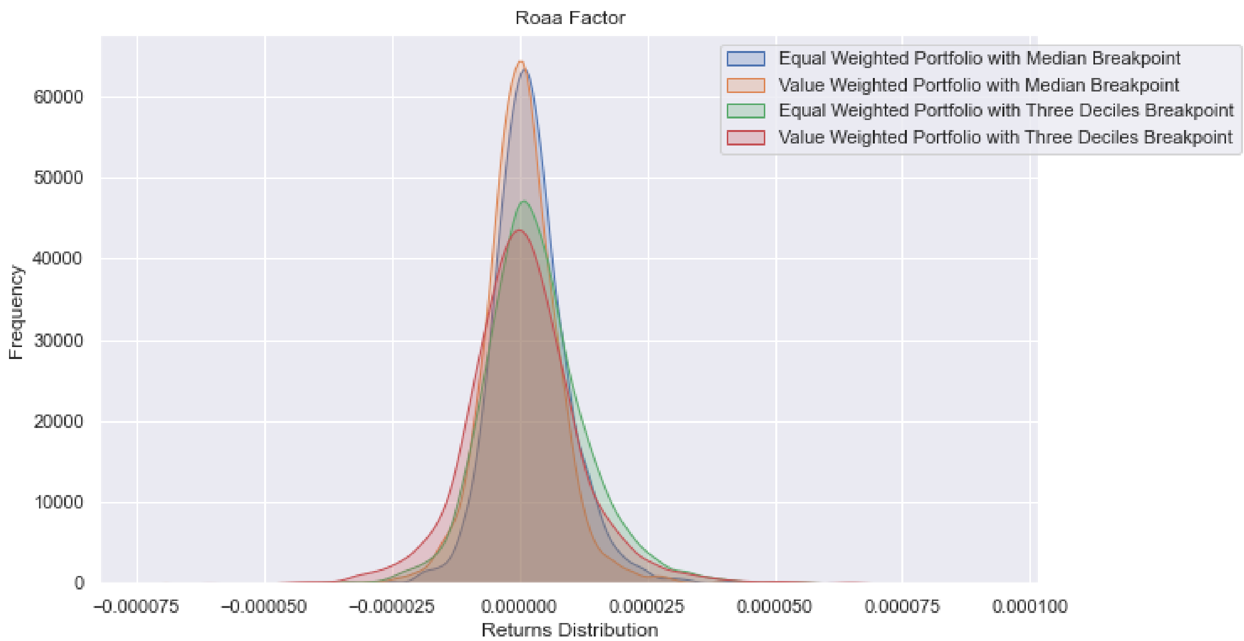


Figure 26: Sales-to-Price Factor Returns Distribution

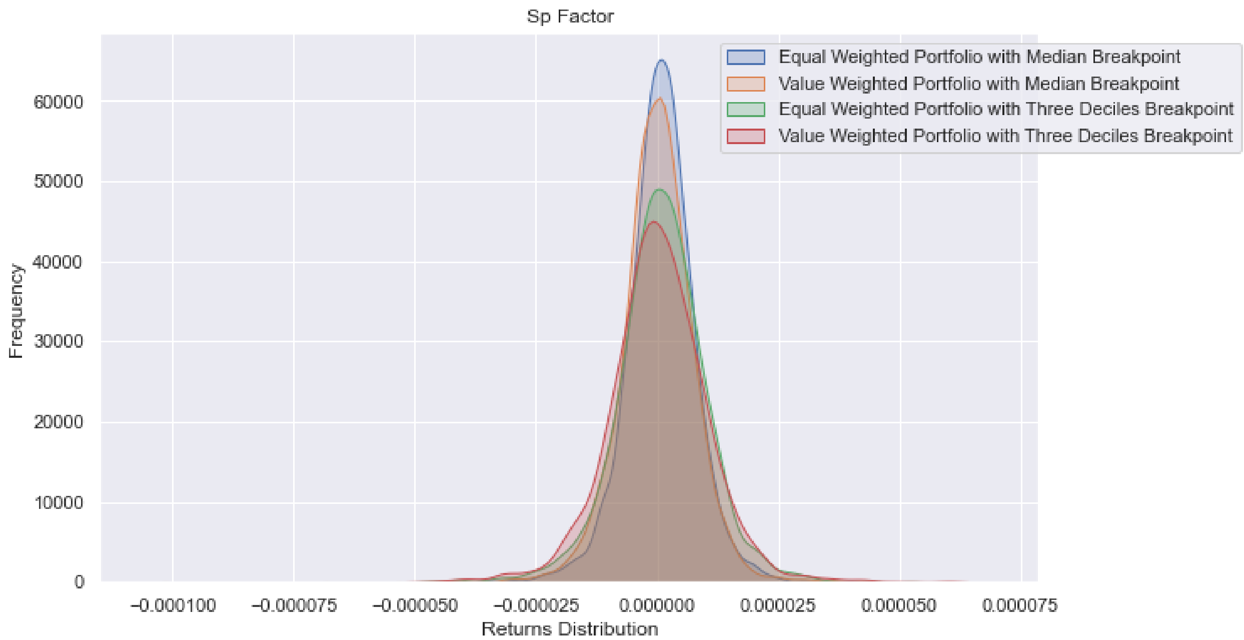

Figure 27: Growth in LTNOA Factor Returns Distribution

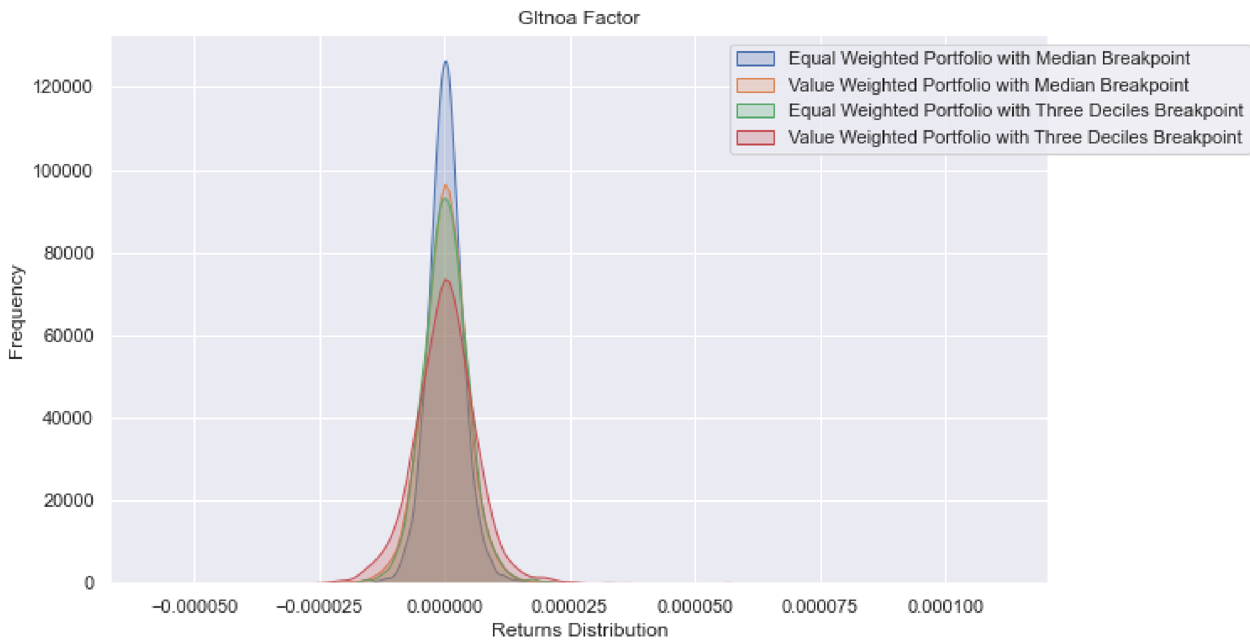


Figure 28: Momentum (6m) Factor Returns Distribution

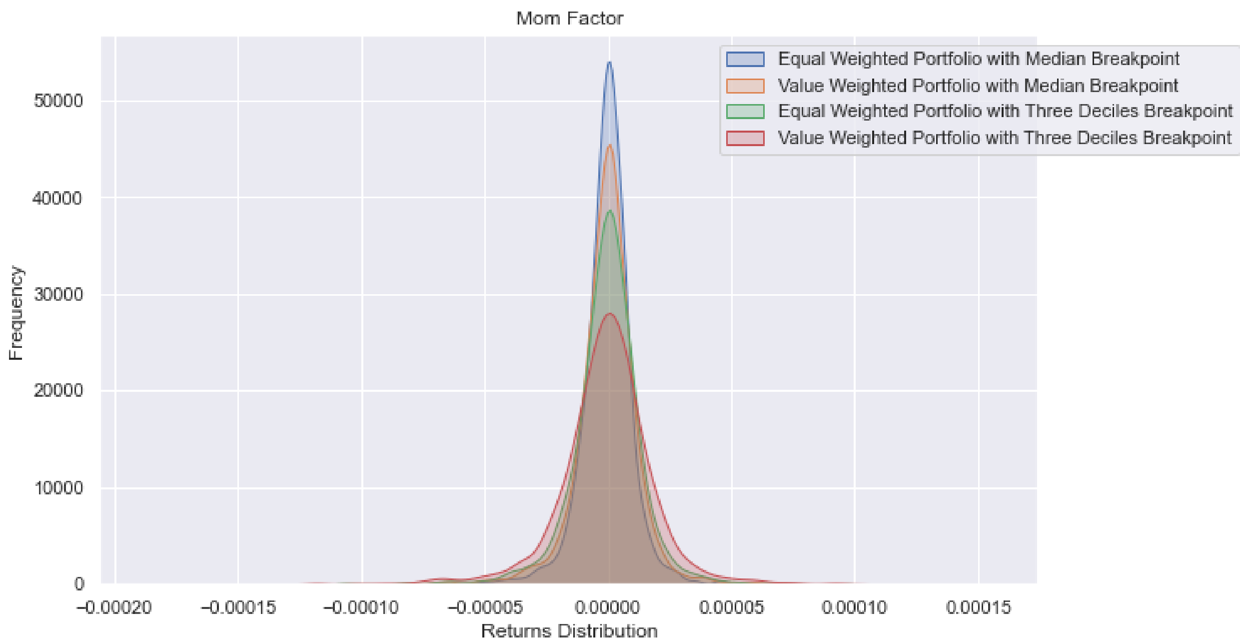

Figure 29: Industry Momentum Factor Returns Distribution


Figure 30: Value-Momentum Factor Returns Distribution

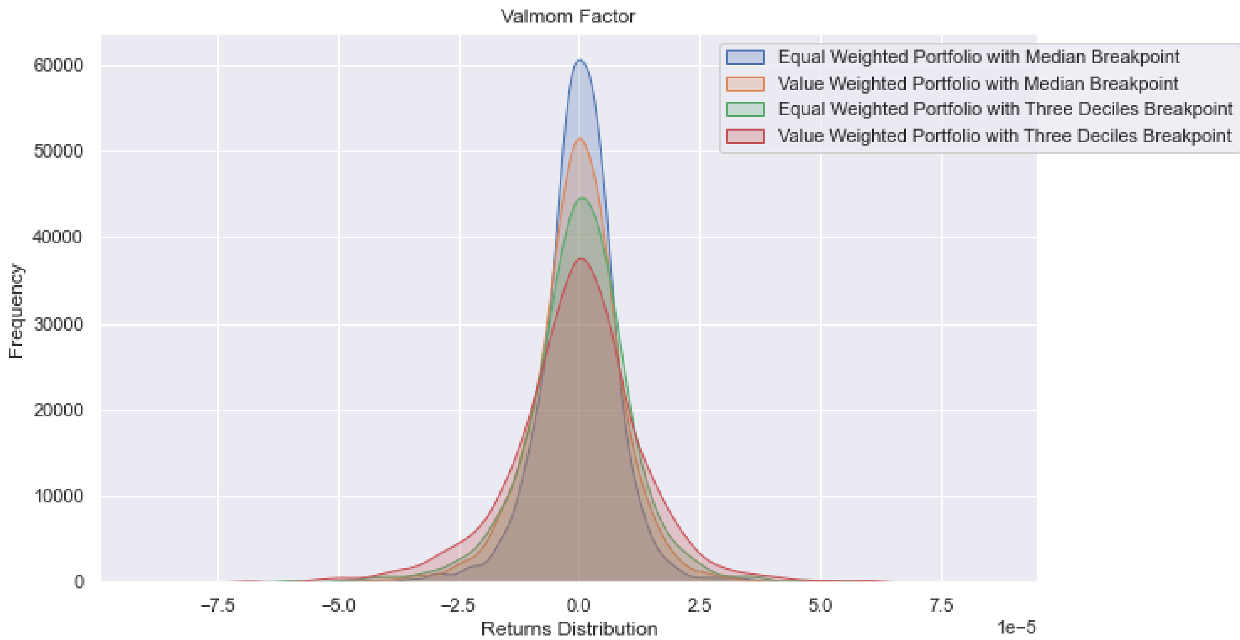

Figure 31: Value-Momentum-Profitability Factor Returns Distribution

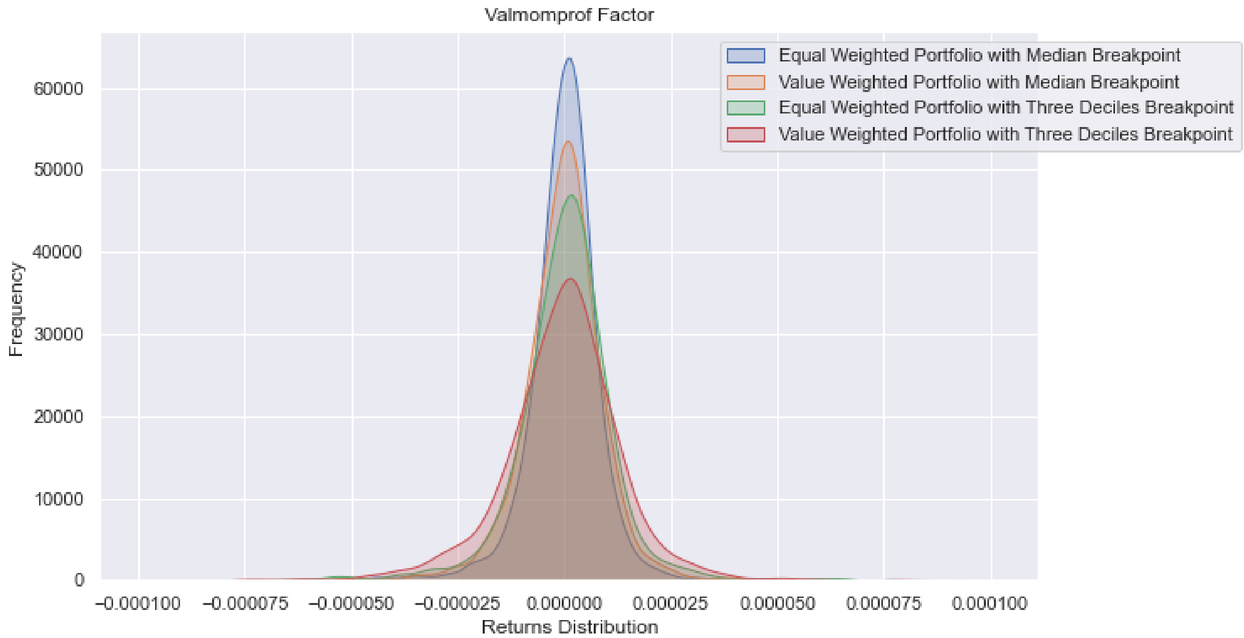


Figure 32: Short Interest Factor Returns Distribution

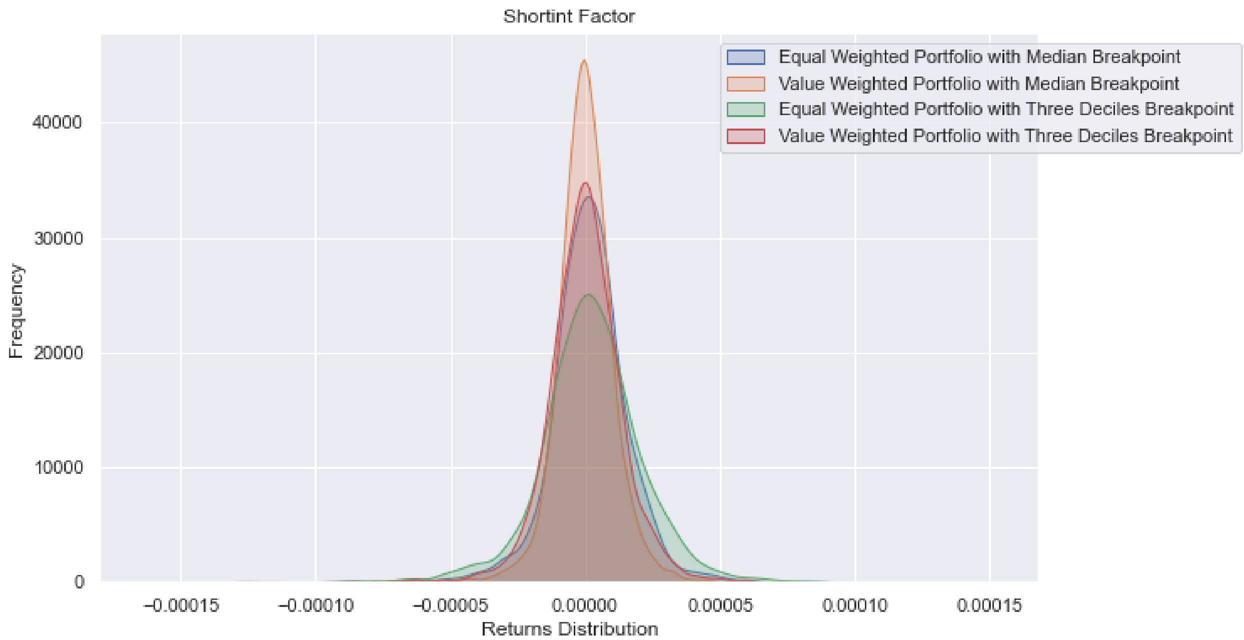

Figure 33: Momentum (1 year) Factor Returns Distribution

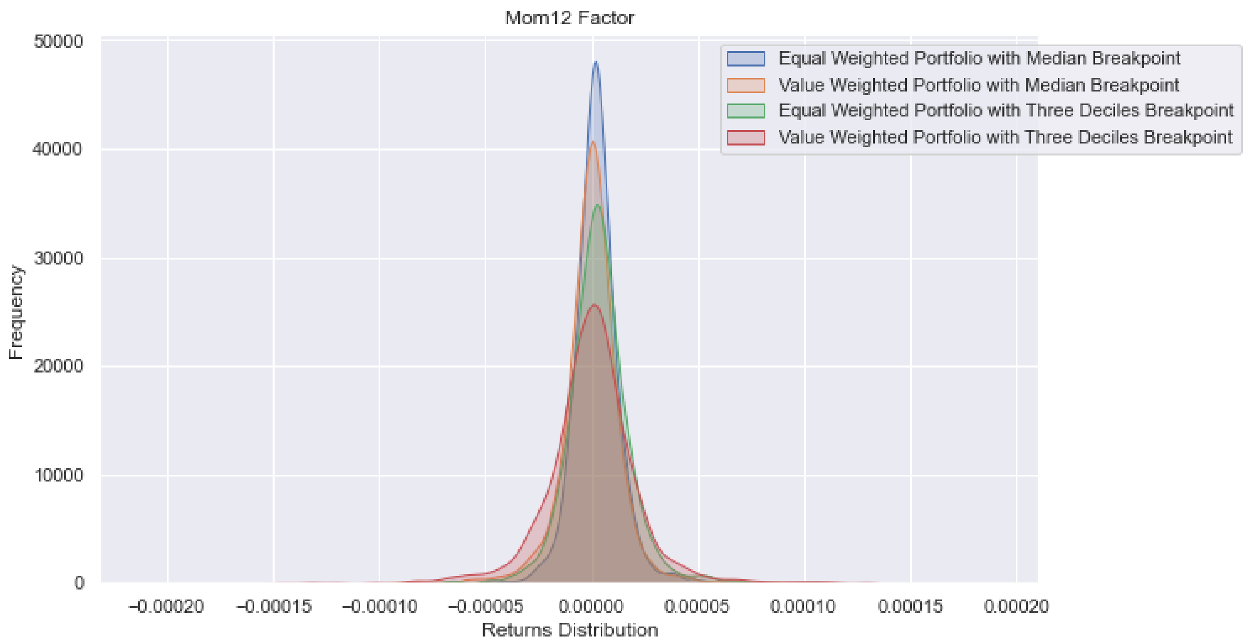


Figure 34: Momentum-Reversal Factor Returns Distribution

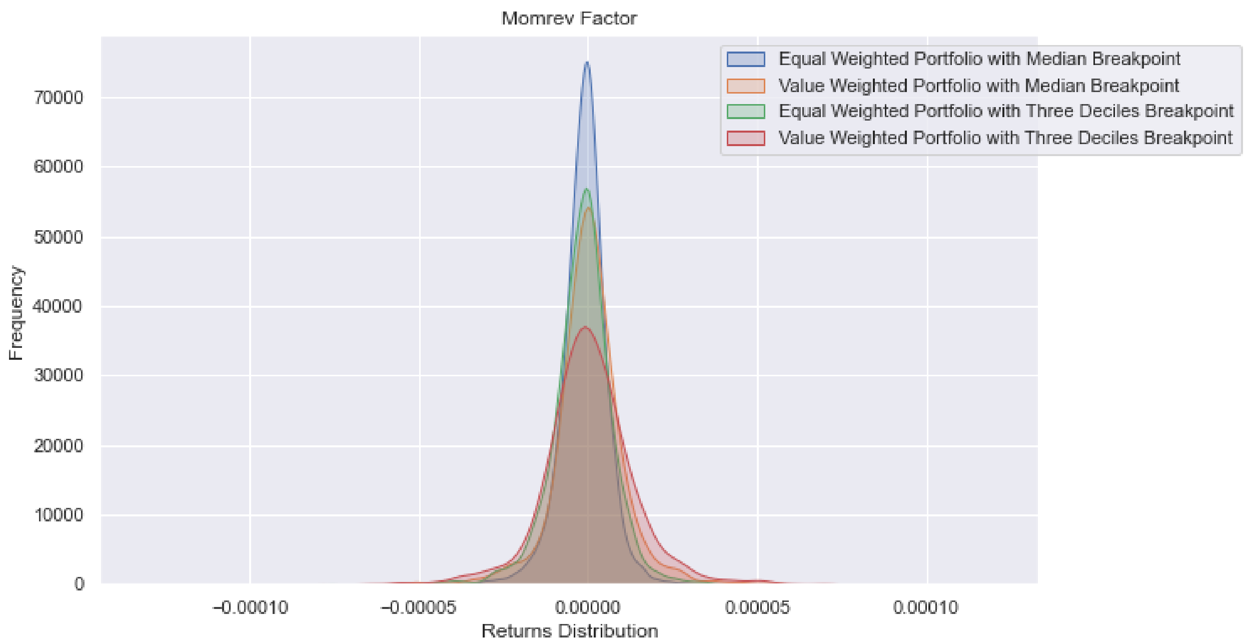

Figure 35: Long-term Reversals Factor Returns Distribution

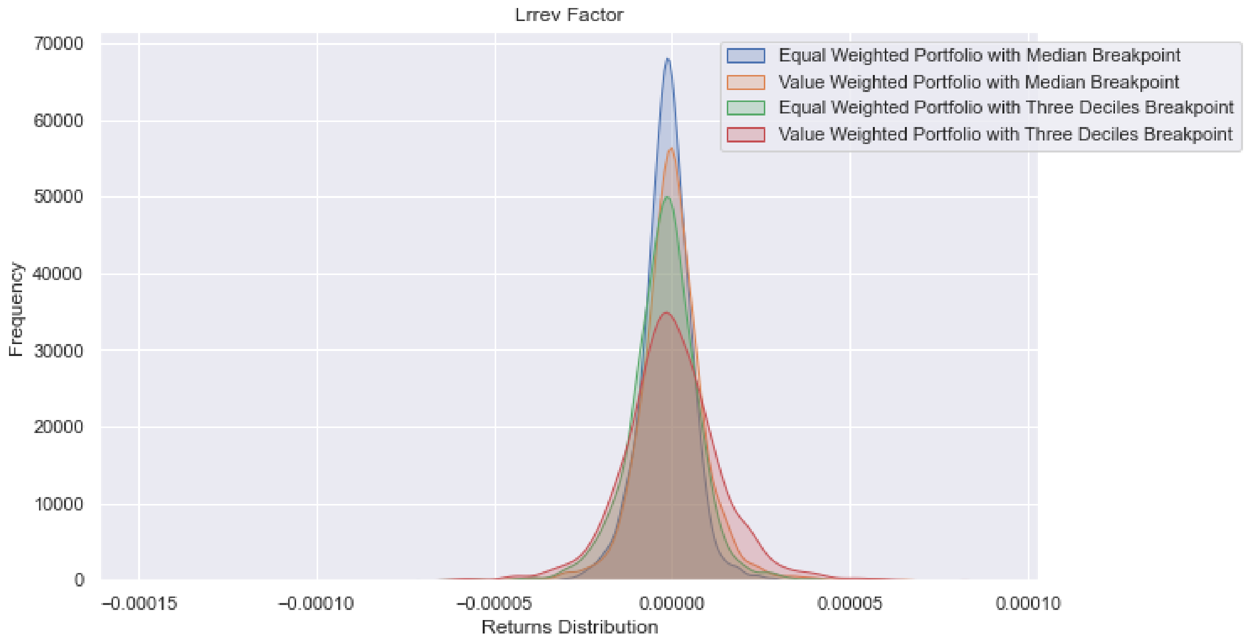


Figure 36: Value (monthly) Factor Returns Distribution

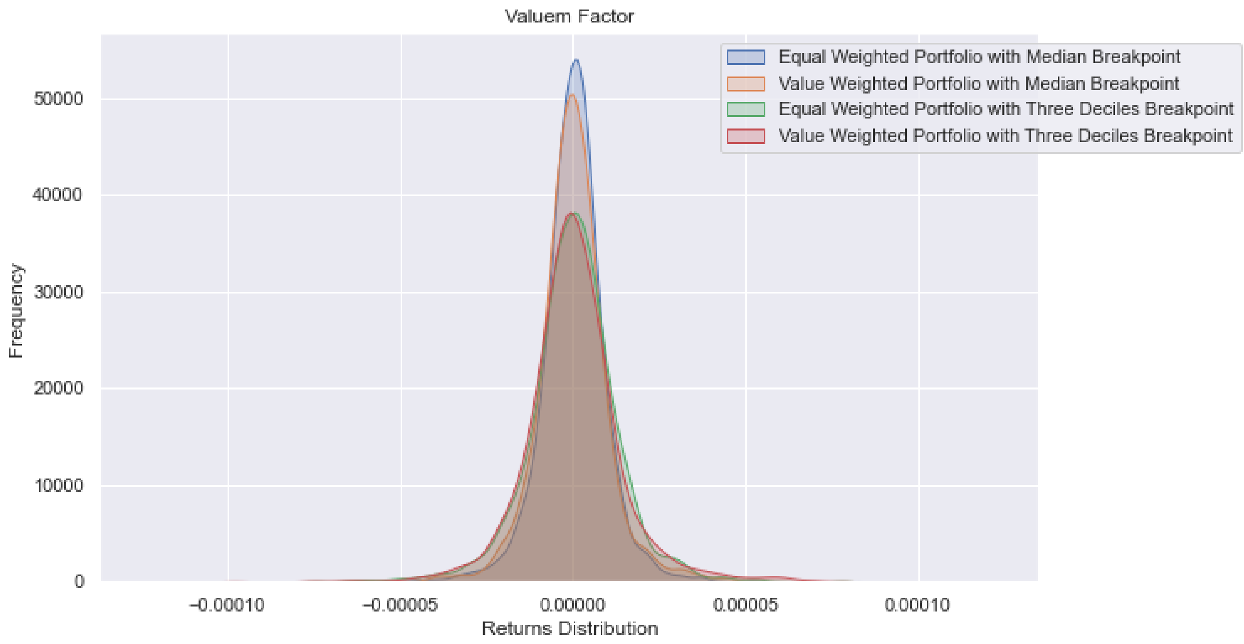

Figure 37: Share Issuance (monthly) Factor Returns Distribution

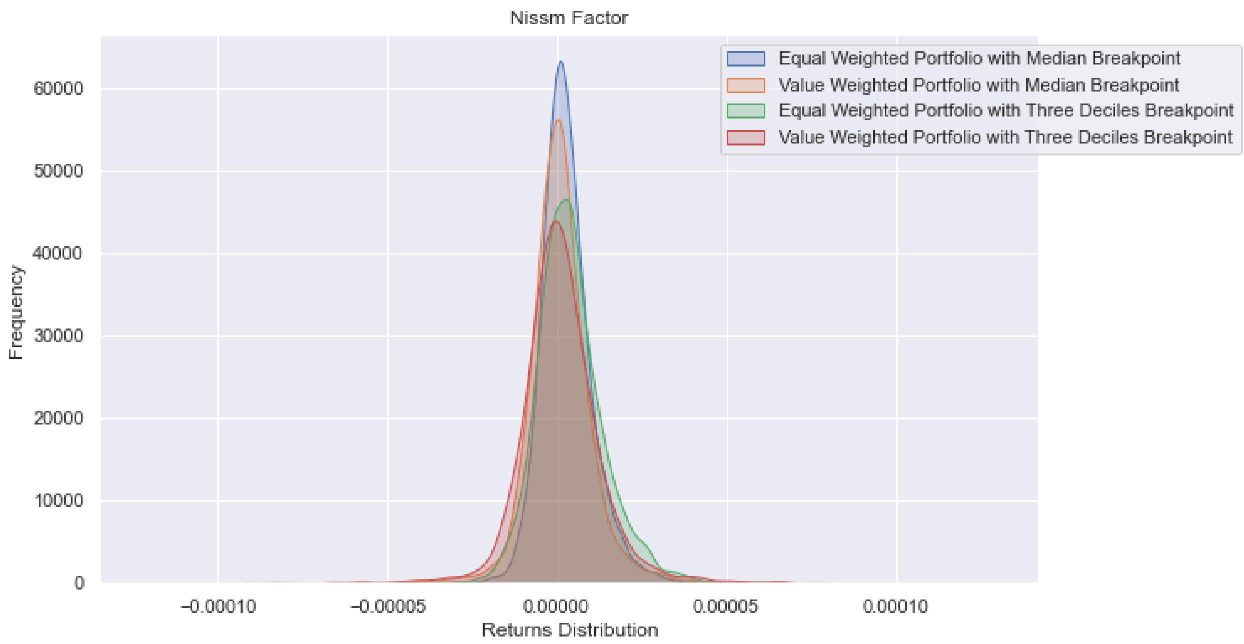


Figure 38: PEAD (SUE) Factor Returns Distribution

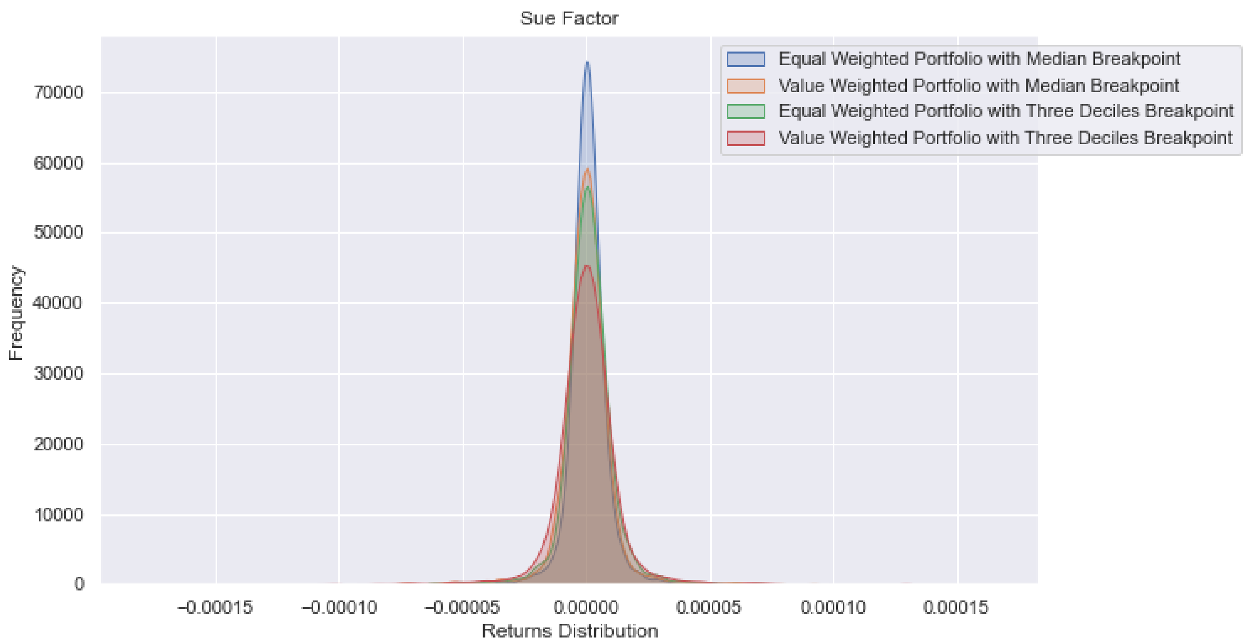

Figure 39: Return on Book Equity Factor Returns Distribution

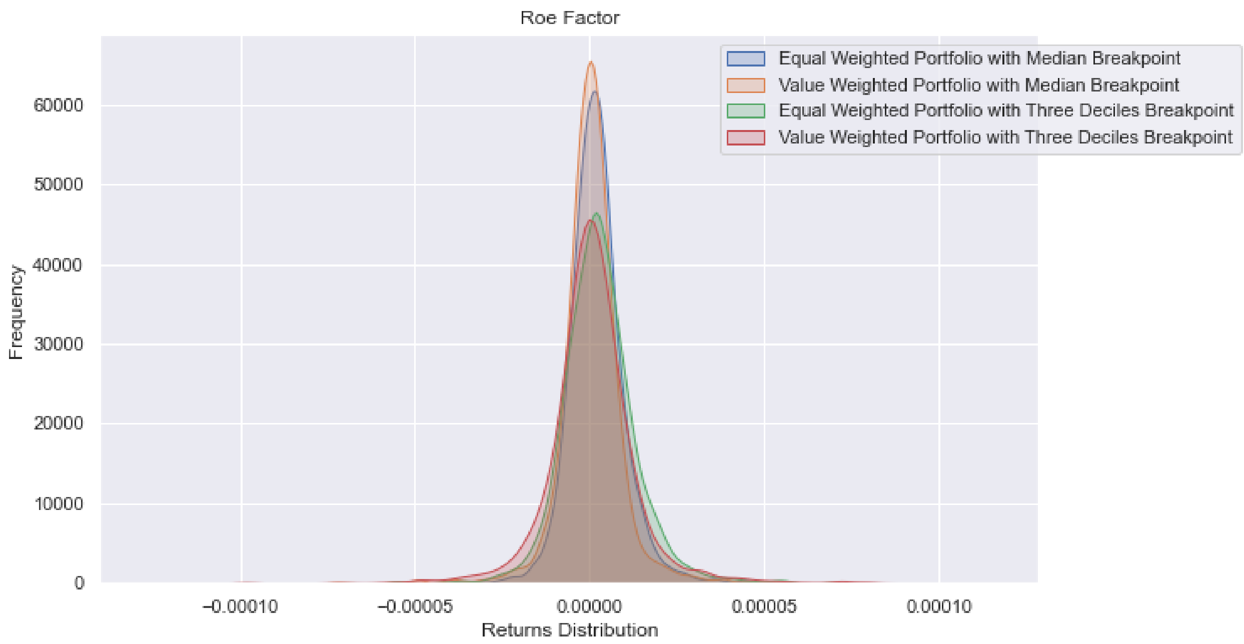


Figure 40: Return on Market Equity Factor Returns Distribution

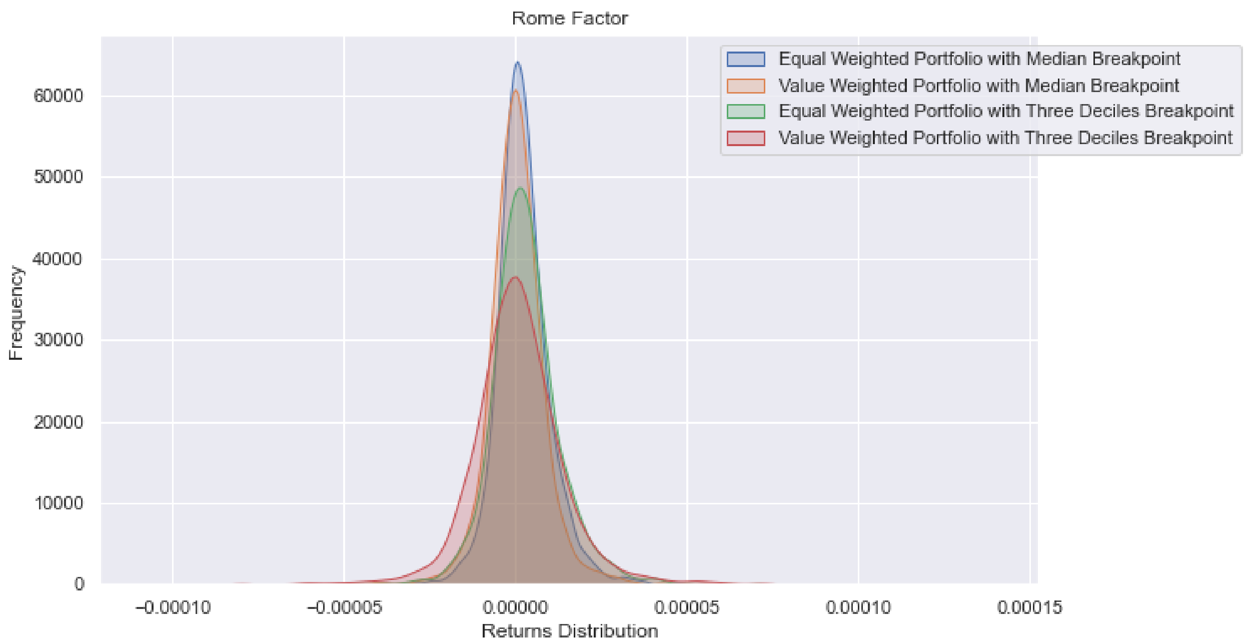

Figure 41: Return on Assets Factor Returns Distribution

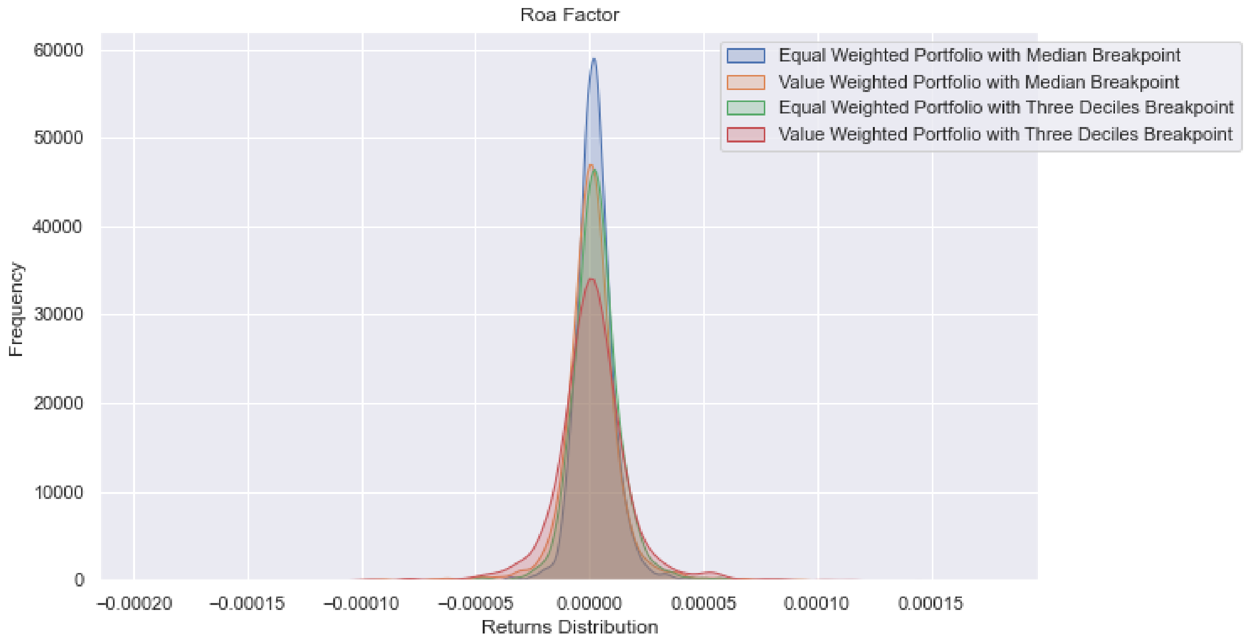


Figure 42: Short-term Reversal Factor Returns Distribution

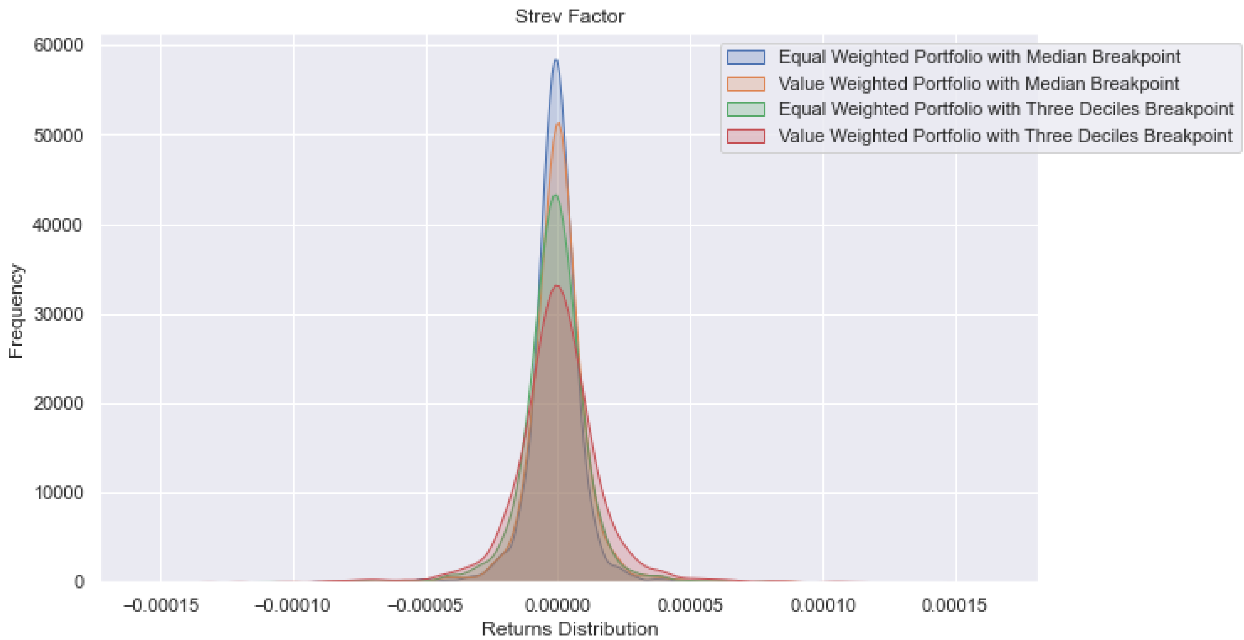

Figure 43: Idiosyncratic Volatility Factor Returns Distribution

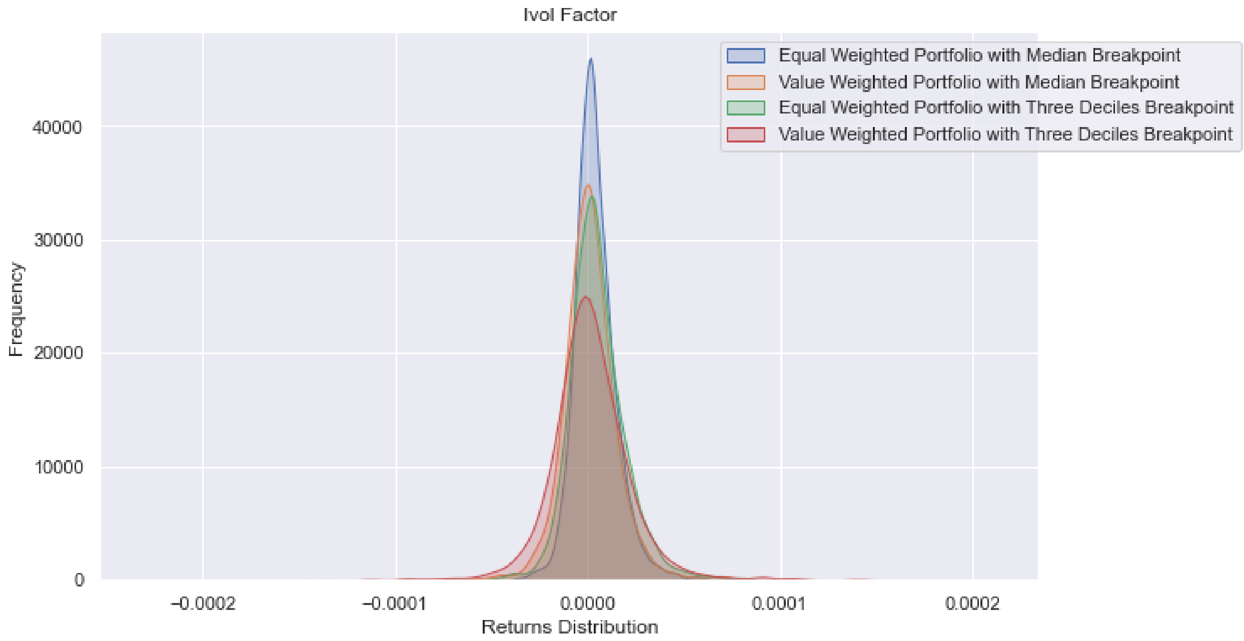


Figure 44: Beta Arbitrage Factor Returns Distribution

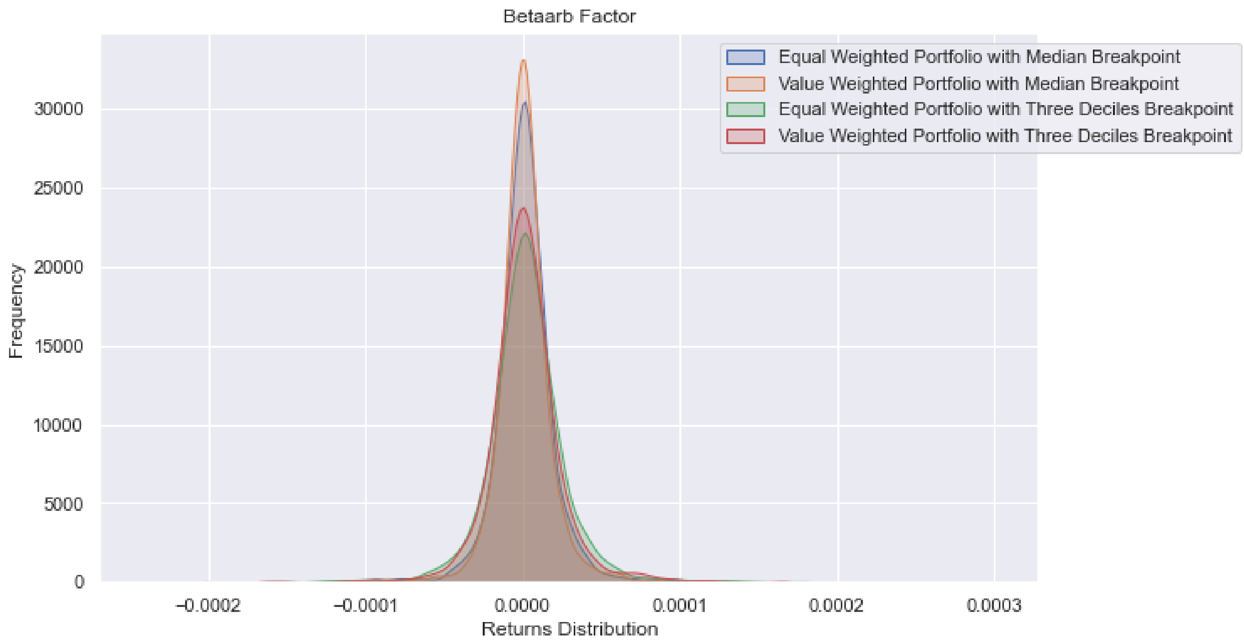

Figure 45: Seasonality Factor Returns Distribution

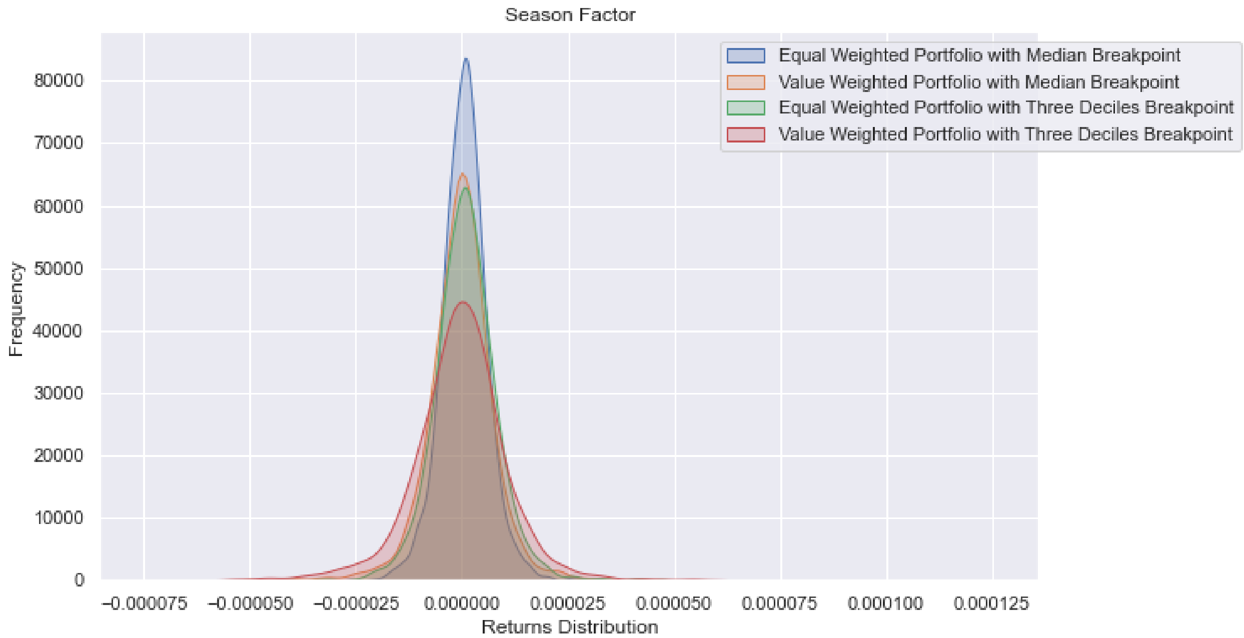


Figure 46: Industry Relative Reversals Factor Returns Distribution

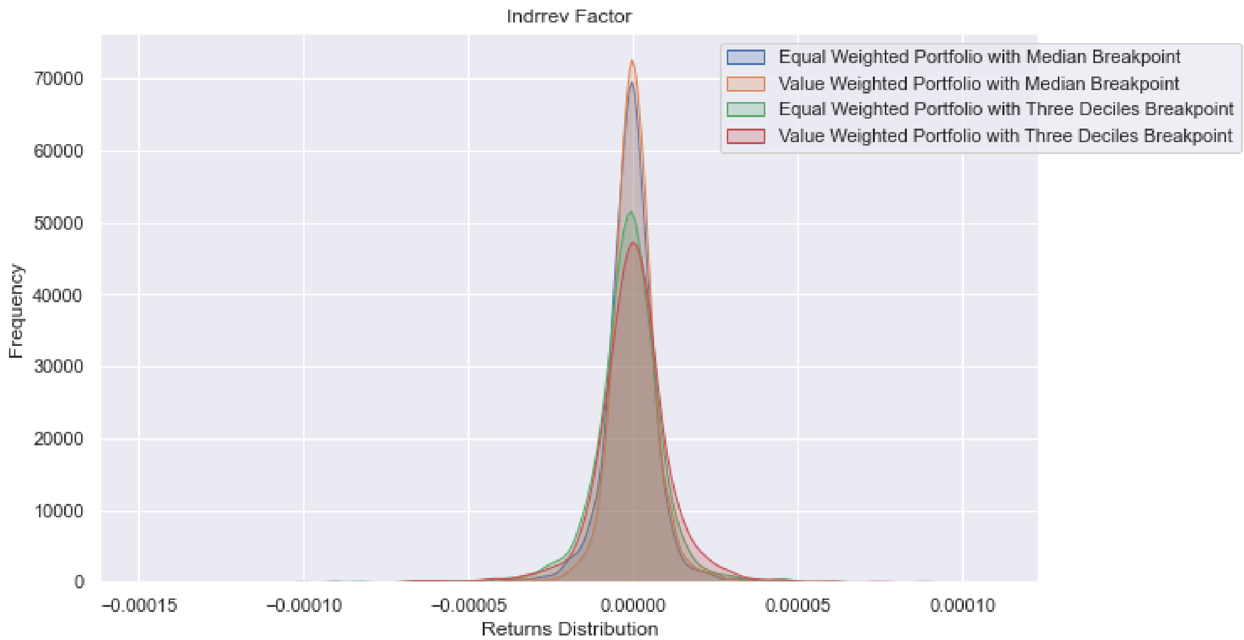

Figure 47: Industry Relative Reversals (Low Volatility) Factor Returns Distribution

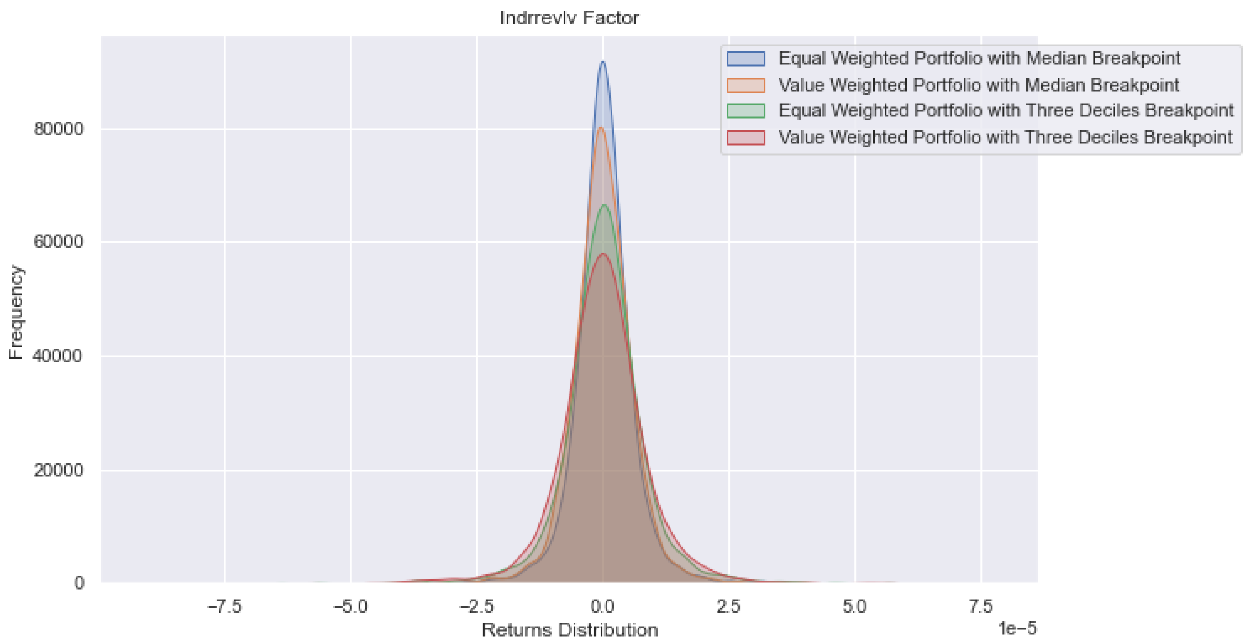


Figure 48: Industry Momentum-Reversal Factor Returns Distribution

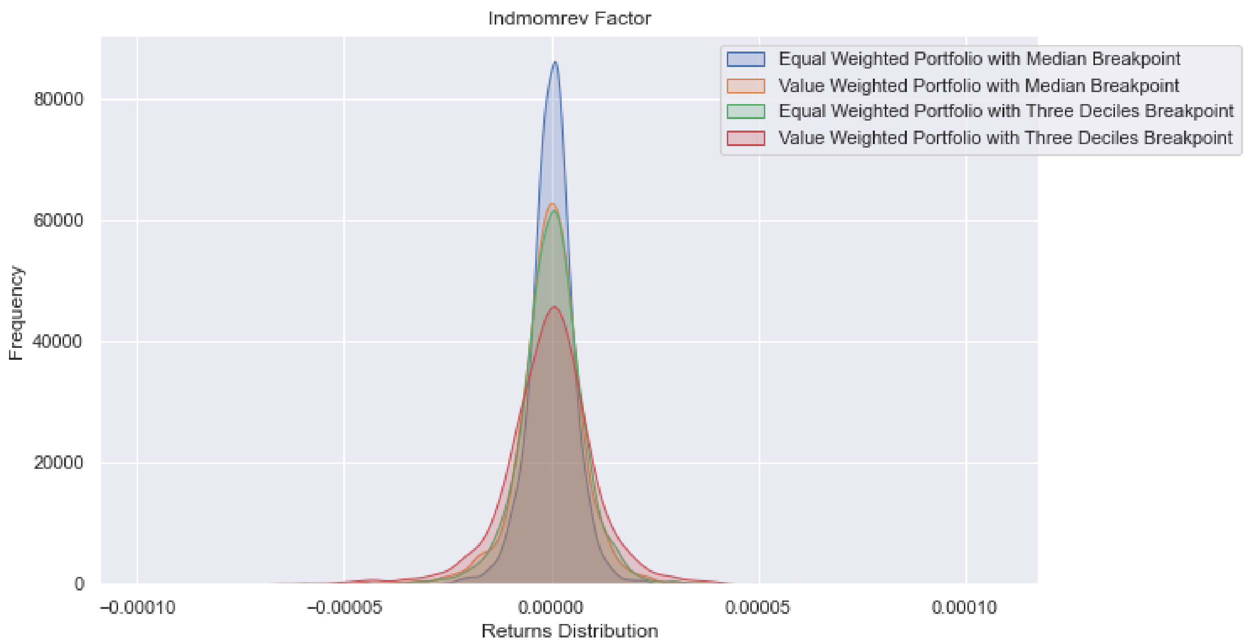

Figure 49: Composite Issuance Factor Returns Distribution

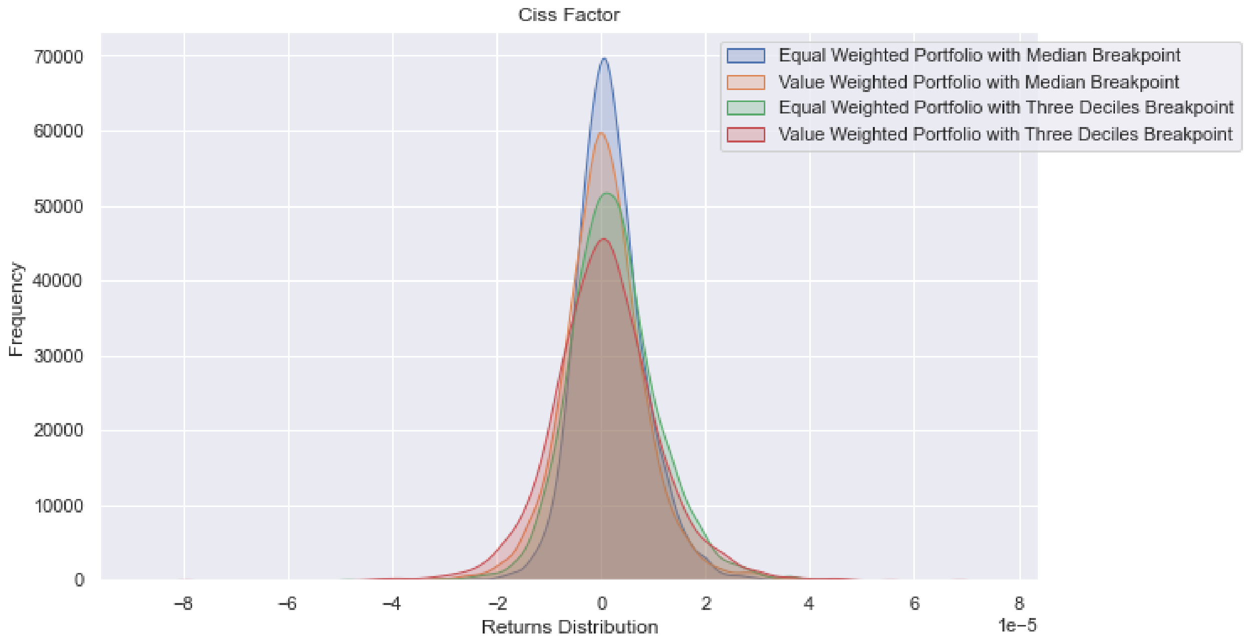


Figure 50: Price Factor Returns Distribution

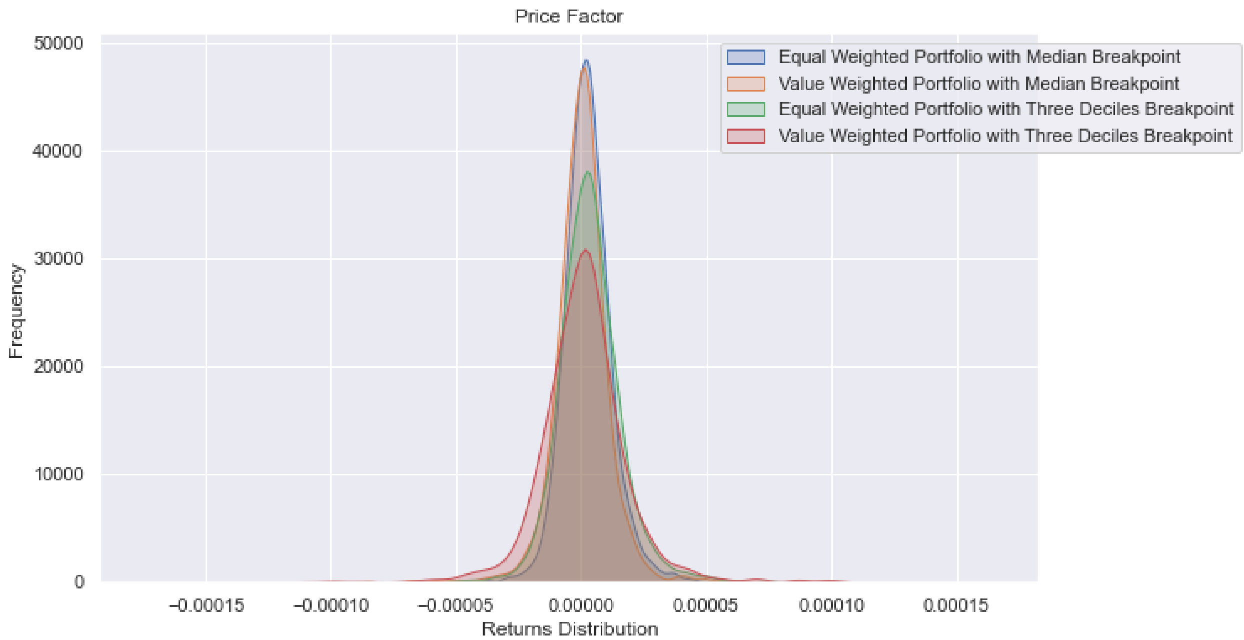

Figure 51: Firm Age Factor Returns Distribution


Figure 52: Share Volume Factor Returns Distribution