FACTSET 〉 SEE THE ADVANTAGE

TRUST THE PROCESS, PART I: AN IN-DEPTH GUIDE TO EXPLORING 2iQ INSIDER TRANSACTIONS DATA

By Joe Importico, CFA, FRM

Contents

Introduction

"Trust the Process" is a phrase commonly associated with the Philadelphia 76ers' former general manager, Sam Hinkie, who helped popularize the importance of data analytics and advanced statistics in the game of basketball. Hinkie's reliance on analytics not only reformed the 76ers' approach to the National Basketball Association (NBA), it has also taken them from a non-factor to a perennial championship contender.

Hinkie's insistence on having a framework for decision making backed by data transformed how the entire organization was run, from how they evaluated talent, all the way to how offensive and defensive schemes were executed. Perhaps most importantly, it taught the entire organization to focus on the process and not get bogged down with individual outcomes that don't go according to plan. Having a sound process doesn't guarantee successful outcomes, but it certainly increases the odds of achieving one.

While trusting the process may be inextricably linked to Sam Hinkie and the Sixers, the concept is ubiquitous in analytical circles. Many, if not all, evidence-based practices rely on a sound process to guide future decisions. Given the importance of having a sound process in successfully conducting empirical research, we'll take a page out of Hinkie's book by creating a rubric that is objective and repeatable and of course, we'll trust the process along the way.

This is the first of several papers where we will undertake an end-to-end exploration of the 2iQ Global Insider Transaction dataset.[1] We will follow these steps:

1. Frame the Problem
2. Identify Relevant Variables
3. Gather and Explore Data
4. Feature Engineering
5. Predictive Modeling
6. Conclusion and Next Steps

Starting by framing the problem is akin to the scientific method where an observation is made and a question posed to test the validity of an observation. Our observation is that individuals deemed to be "insiders" have an informational advantage relative to the field. We will evaluate this statement by testing the influence that insider transactions have on the cross-section of future asset returns. Said more plainly, can insider transactions help investors predict asset price movements?

Before jumping into the analytical rigor that accompanies data exploration, we need to frame our problem, identify relevant variables, develop a shared vocabulary, and then detail the tools that will be used to guide this workflow by gathering, preparing, and modeling the data. With more than 15 years of history, over 10 million unique transactions, and a universe spanning more than 60,000 companies, we clearly have our work cut out for us; so, let's get started.

[1] Open:FactSet, *2iQ Global Insider Transaction Data*, https://open.factset.com/products/2iq-global-insider-transaction-data/en-us (accessed Apr. 25, 2019).

Frame the Problem

The title of this article implies that we are going to analyze the transactions executed by insiders; but who exactly are these insiders? According to the U.S. Securities and Exchange Commission (SEC):[2]

> *The federal securities laws require certain individuals (such as officers, directors, and those that hold more than 10% of any class of a company's securities, together we'll call, "insiders") to report purchases, sales, and holdings of their company's securities by filing Forms 3, 4, and 5.*[3]

This SEC document goes on to describe why investors may care to investigate insider activity:

> *Many investors believe that reports of insiders' purchases and sales of company securities can provide useful information as to insiders' views of the performance or prospects of the company.*

Relative to investors that aren't categorized as "insiders," insiders are deemed to have an asymmetric informational advantage. After all, who knows more about the inner workings of a company than the executives? That informational asymmetry is what we're after, not just here, but whenever investors are trying to understand which factors influence an asset's future price movement. In short, for purposes of this analysis, we want to know whether insider transactions influence the cross-section of asset returns.

Gather and Explore Data

Data Acquisition

Before we begin querying and exploring the data, we need to be thoughtful about the types of information we need. With so much data available, we need to distill our requirements down to the lowest level, then identify and track the relevant variables for the exercise at hand.

Data Requirements

Data	Investable Universe	Time Horizon

• Insider Transactions (2iQ)
• Identifier Mappings
• ETF Constituents
• FactSet Prices
• Sector Taxonomy (RBICS)
• Traditional Factors
  ○ Fundamentals
  ○ Pricing
  ○ Estimates
	• iShares Russell 3000 Index	• Date Range:
		  ○ 12/2003 – 12/2018
		• Frequency: Daily

Data Exploration

With our list of requirements handy, we can begin our investigation. Our initial query retrieves insider transactions from the 2iQ dataset for all securities in the iShares Russell 3000 Index (Ticker: IWV) between December 2003 and December 2018.

---

[2] U.S. Securities and Exchange Commission, https://www.sec.gov/, (accessed May 2020).
[3] U.S. SEC, *Insider Transactions and Forms 3, 4, and 5*, https://www.sec.gov/files/forms-3-4-5.pdf, (accessed May 2020).

As shown below, we queried transaction-level data for each company over the 15-year horizon, amassing over three million rows. Seeing the properties of this dataset highlights the depth and completeness of the 2iQ data.

2iQ Dataset Properties

```
RangeIndex: 3532046 entries, 0 to 3532045
Data columns (total 38 columns):
startdate                        object
enddate                          object
fsym_id                          object
tiq_start_datetime               datetime64[ns]
tiq_filing_date                  datetime64[ns]
lag_availability                 int64
entity_proper_name               object
tiq_transaction_id               int64
tiq_rev_id                       int64
tiq_person_id                    object
tiq_insiderrelation              object
tiq_insiderlevel_code            object
tiq_connectiontype               object
tiq_connectiontypespecified      object
tiq_transactiontype_code         object
tiq_shares                       float64
tiq_trade_date                   datetime64[ns]
txn_value                        float64
tiq_valueusd                     float64
tiq_source_code                  object
tiq_tradesignificance_code       int64
tiq_holdings                     float64
tiq_optionrelated_flag           int64
tiq_automated_flag               int64
tiq_taxrelated_flag              int64
tiq_otc_flag                     int64
tiq_capitalincrease_flag         int64
tiq_initial_flag                 int64
tiq_mergerrelated_flag           int64
tiq_offering_flag                int64
tiq_intention_flag               int64
tiq_privateplacement_flag        int64
tiq_dividendreinvestment_flag    int64
tiq_remuneration_flag            int64
tiq_shareplan_flag               int64
tiq_private_flag                 int64
tiq_forced_flag                  int64
l1_name                          object
dtypes: datetime64[ns](3), float64(4), int64(19), object(12)
memory usage: 1.0+ GB
```

Source: FactSet Research Systems Inc.

The subset of fields we queried can be broadly characterized as dates, transaction metadata, and flags. Let's summarize what we know about this data:

1. Available data types include objects (many appear to be strings), dates, and integers/floating point numbers.
2. This is a time-series dataset that requires us to understand which date(s) should be used to form our perspective. We'll use this date to approximate when we knew this data became available historically.
3. There are several dimensions available when considering the available fields, such as time, company name, individual, transaction, trade reason, etc.

Of the three observations above, we should explore the date attribute before moving on.

Dates

The date used to form our perspective on the dataset is integral to the rest of the analysis, so we should understand the implications of choosing one date versus another.

Below is a preview of the date fields included in the 2iQ dataset:

- **tiq_start_datetime:** 2iQ captured timestamp of the transaction
- **tiq_filing_date:** Date of the transaction filing
- **tiq_trade_date:** Date of the transaction

The **tiq_start_datetime** field captures the date on which 2iQ collected the transaction and at first glance appears to be the field we're looking for. 2iQ provides a revision flag denoting each revision that has been applied to a transaction. In this study, we limited revisions to the first time a transaction was stored in the dataset and did not overwrite subsequent entries.

Think of **tiq_start_datetime** as the point-in-time (PIT) date or the date users of the 2iQ dataset became aware of the transaction. Let's explore the available dates a bit further before settling on that field.

Comparison of 2iQ Date Fields

	tiq_start_datetime	tiq_filing_date	tiq_trade_date
count	3532046	3532046	3532046
unique	4182	4082	6152
top	2013-09-02 08:00:00	2007-08-06 00:00:00	2007-04-30 00:00:00
freq	1197104	5879	5640
first	2013-09-02 08:00:00	2004-03-31 00:00:00	1982-03-01 00:00:00
last	2020-02-24 20:00:00	2018-12-31 00:00:00	2018-12-31 00:00:00

Source: FactSet Research Systems Inc.

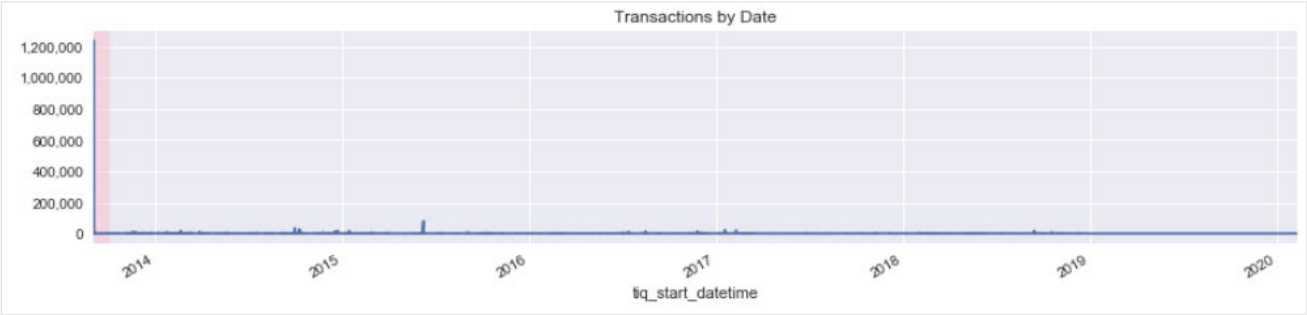
The descriptive statistics above highlight a few important features surrounding the three date fields:

1. The **tiq_start_datetime** field begins on September 2, 2013. Since our analysis begins prior to 2013, we may need to pursue a new date that can be used to form our perspective date.
2. The **tiq_filing_date** field aligns more closely with the start and end dates of our analysis. This makes sense since that field was specifically called out in the initial query for the 2iQ data, where filing dates were retrieved within the date range 2004–2018.
3. The **tiq_trade_date** field has a first date that is well before that of our start date (2004).

Let's examine these date fields further by plotting the number of transactions over time when grouped by each of these date fields.

Looking at the **tiq_start_datetime** field, we see that the number of transactions taking place on September 2, 2013, is orders of magnitude higher than those on subsequent dates, indicating that 2iQ probably began collecting the PIT date field on September 2, 2013.

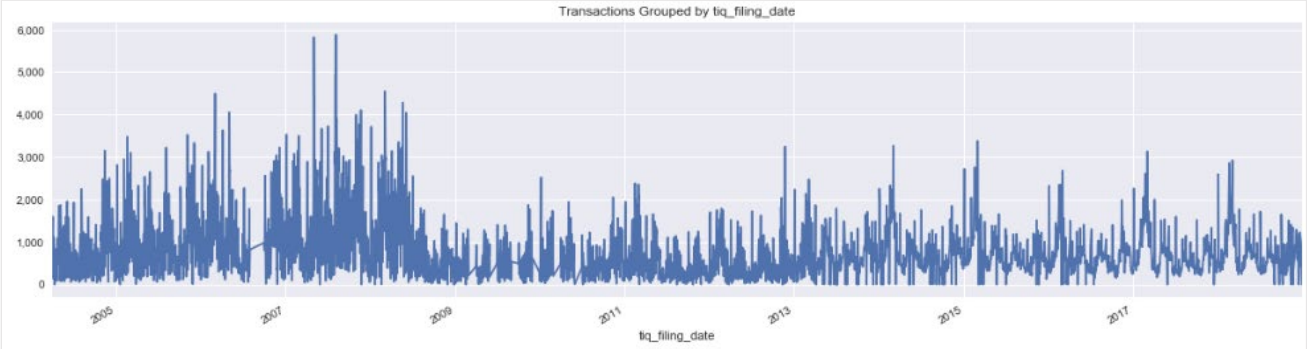
Transactions for tiq_start_datetime



Source: FactSet Research Systems Inc.; 2iQ Research

As expected, the **tiq_filing_date** is the most complete date field from a longevity perspective as it is required and regularly maintained as part of the regulatory process around insider transactions. The rationale for relying on this date as a filter in the initial query was two-fold: it ensured that we captured all transactions occurring during the sampling period (2004–2018) and it provided an additional data point to consider as we work to construct an "as-was" view of the world for our analysis.

Transactions for tiq_filing_date



Source: FactSet Research Systems Inc.; 2iQ Research

Lastly, the **tiq_trade_date** is relatively aligned with the **tiq_filing_date** except for a few transactions that occur well before the start date of the analysis.

Transactions for tiq_trade_date

Source: FactSet Research Systems Inc.; 2iQ Research

Let's dig into the trade date a bit more. Of the transactions queried (more than three million), 75% of the filing dates occur within four days of the trade date. The gap between the trade and filing date begs a question: How much of a gap are we comfortable keeping in the analysis?

We need to set a threshold with which we're comfortable. We set ours at 45 days, removing all trades that took place more than 45 days before the filing date. The good news is that this doesn't happen very often—less than 2% of the transaction volume falls into that category.

Before moving forward, let's take stock of where we are and decide how the perspective date will be determined.

- The **tiq_filing_date** encompasses the entire sampling period and is the most complete date field available prior to the collection of the **tiq_start_datetime**.
- After removing a few records, we have a series of trade dates (**tiq_trade_date**) with which we are comfortable (i.e., those that occur within a reasonable amount of time to the filing date).
- The start date (**tiq_start_datetime**), representing the date 2iQ loaded the transactions, begins on September 2, 2013.

We can use the **tiq_start_datetime** as the perspective date beyond September 2, 2013, but what should we use prior to that date? One solution is to apply a heuristic, typically based on the collection methodology employed by the vendor, to avoid look-ahead bias.[4] Forming that heuristic or baseline adjustment requires understanding the origin story behind the data. Ideally, that would entail getting input directly from the vendor as to how the data was collected prior to a true PIT date being made available. That's ideal but often impractical for a variety of reasons. For purposes of this exercise, we'll look to the data to select a conservative adjustment.

To measure the amount of time that passes between when a transaction is filed and when it is entered into the 2iQ dataset, we created a new field titled "lag_availability." The table below shows that 2iQ gets to nearly all transactions very quickly, accounting for over 75% of the observations within one day of the filing date and 99% of the observations within three days.

Measuring Lag Time in Data Availability

	lag_availability
count	1,007,225.00
mean	4.66
std	64.03
min	-1.00
25%	1.00
50%	1.00
75%	1.00
95%	3.00
99%	3.00
max	2,151.00

Source: FactSet Research Systems Inc.

Our methodology is relatively conservative, where we will use **tiq_filing_date** + three days to define the perspective date prior to the collection of the PIT date (September 2, 2013). From September 2, 2013 forward, we will rely on the PIT date. For more information on this process, please refer to the corresponding Jupyter Notebook that goes through how these dates get applied in more detail.

[4] Kenton, Will. "Look-Ahead Bias," *Investopedia*. February 16, 2020. https://www.investopedia.com/terms/l/lookaheadbias.asp, (accessed May 2020).

Flags

2iQ collects a variety of flags containing rich contextual information about the transactions placed by insiders. Before exploring the transactions, we should better understand why they are occurring in the first place. 2iQ's documentation has more information on the importance of these flags, but the crux of it is that not all transactions are created equal. Insiders place trades for a variety of reasons (as indicated by a large number of flags) and not all imply an insider's sentiment on their company's prospects. Take for example an insider that places a trade to offset capital gains on other investments; this trade doesn't necessarily suggest that they're bearish on their company's stock. There are many reasons that insiders trade stock; each of these reasons needs to be considered before combing through the data for that informational asymmetry or "edge" we spoke about earlier. 2iQ has a great methodology guide[5] that details how those with access to their content can identify relevant trades.

With this information in mind, let's look at the most common reasons insiders trade their company's stock. The most frequently occurring trades are related to options, automated trades, and taxes.

Flags Indicating Reasons for Stock Trades

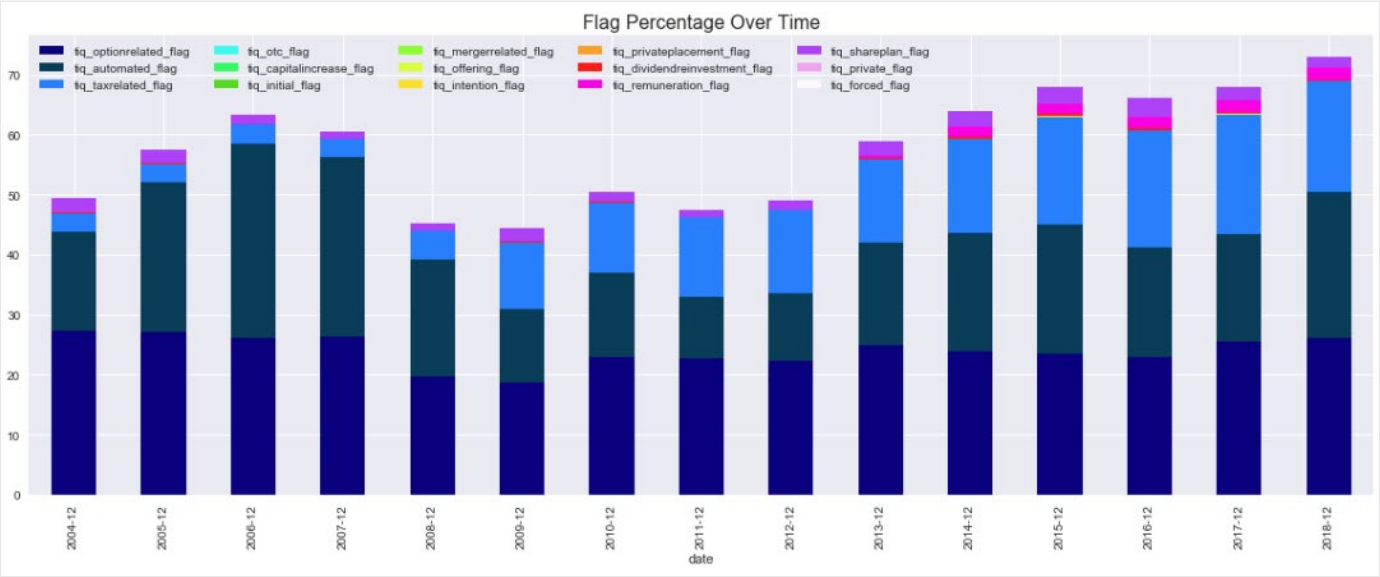
	Percent Available
tiq_optionrelated_flag	25.48%
tiq_automated_flag	22.47%
tiq_taxrelated_flag	11.77%
tiq_shareplan_flag	2.51%
tiq_remuneration_flag	0.86%
tiq_dividendreinvestment_flag	0.62%
tiq_offering_flag	0.13%
tiq_mergerrelated_flag	0.07%
tiq_private_flag	0.04%
tiq_forced_flag	0.02%
tiq_initial_flag	0.01%
tiq_privateplacement_flag	0.01%
tiq_otc_flag	0.00%
tiq_capitalincrease_flag	0.00%
tiq_intention_flag	0.00%

Source: FactSet Research Systems Inc.

The story remains consistent over time. Across our time frame, anywhere from 40–60% of transactions are tagged as related to options, automated trades, and taxes.

[5] Open:FactSet, "2iQ Global Data."

Reasons for Stock Trades



Source: FactSet Research Systems Inc.; 2iQ Research

Knowing why transactions are taking place is helpful and it provides more food for thought down the road when we can consider filtering out certain categories of transactions (i.e., Should we exclude trades flagged as being automated or tax related?). We can consider applying filters later, once we get past data exploration and begin feature engineering exercises.

Transaction Metadata

Before delving into the transactions, let's review what we already know about insiders with sizeable amounts of company stock at their disposal:

- According to the SEC, insiders are defined as officers, directors, or those holding more than 10% of a company's stock. Insiders meeting those guidelines have to file Forms 3, 4, and 5 with the SEC.
- It is a relatively common practice for companies to compensate their employees by issuing stock, stock options, or restricted stock options (RSUs).
- As indicated by the flags we learned about earlier, insiders trade for a wide range of reasons, each of which influences how relevant a trade may be to outsiders looking to glean an informational advantage.

Knowing that most companies use stock to compensate their employees means that we should expect to see far more sales than purchases. It's common for individuals to use proceeds from their stock issuance to meet living expenses and other personal obligations. Beyond that, investors with sizeable allocations to their employer's stock may be looking to diversify their positions.

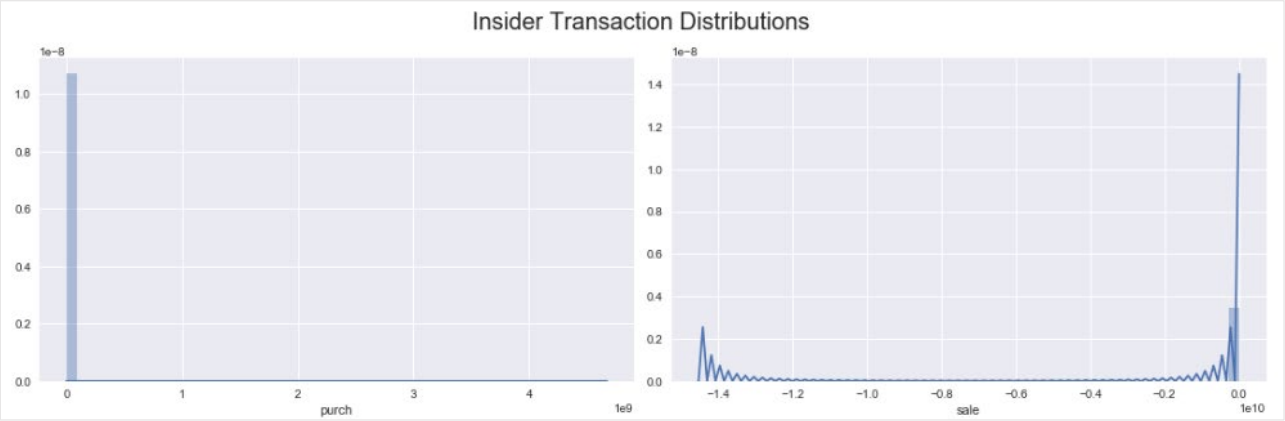
Does the data support this assumption? Insiders sell around five times as often as they buy their company's stock and in much larger quantities.

Insiders' Stock Purchases vs. Sales

	txn_value	purch	sale
count	3,464,781.00	502,141.00	2,729,156.00
mean	-466,817.09	341,285.28	-655,437.91
std	18,163,823.64	12,866,189.03	19,703,333.28
min	-14,535,156,874.00	0.01	-14,535,156,874.00
25%	-123,300.00	2,625.00	-187,550.00
50%	-21,580.00	12,312.00	-45,018.57
75%	-1,942.00	48,600.00	-10,844.00
max	4,673,244,381.44	4,673,244,381.44	-0.02

Source: FactSet Research Systems Inc.

This is confirmed in the charts below, which plot the distribution of purchases and sales.

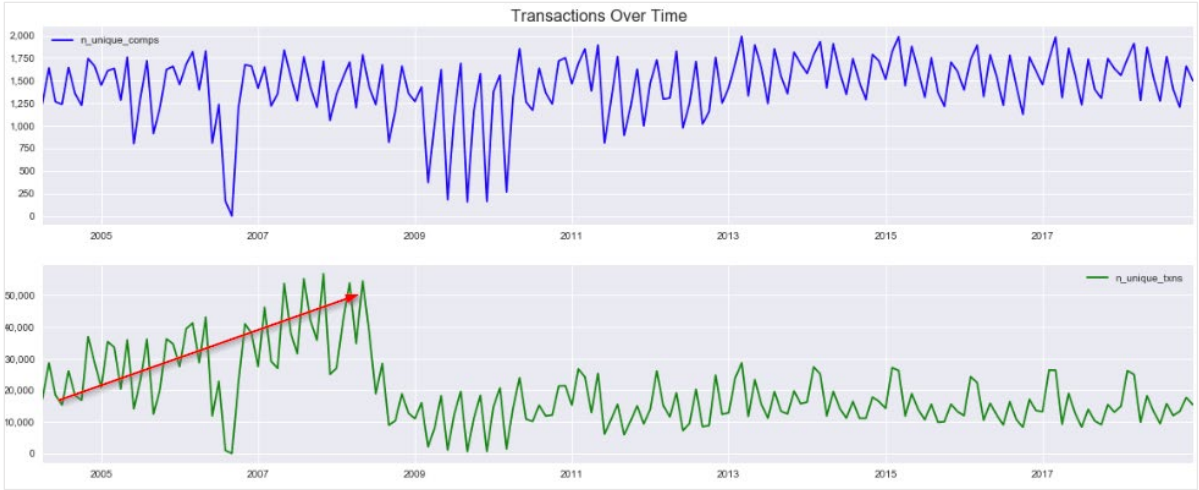


Source: FactSet Research Systems Inc.; 2iQ Research

The trade flags tell us why trades are placed but we still want to identify who the insiders are and when their trades occur. Revealing more information on the "who" and "when" around insider transactions will provide clarity on important questions such as:

- How tightly coupled or evenly distributed have trades been throughout the sampling period? Are there seasonal trends in the data?
- How concentrated/unconcentrated has the trading volume been with respect to the number of firms in the universe over time?

To begin answering these questions, we down sampled (i.e., converted) the data to a monthly frequency and analyzed the transaction volume over time in two ways. The first chart shows the number of companies where at least one trade was placed throughout the month and the second shows the number of unique transactions placed.



Source: FactSet Research Systems Inc.; 2iQ Research

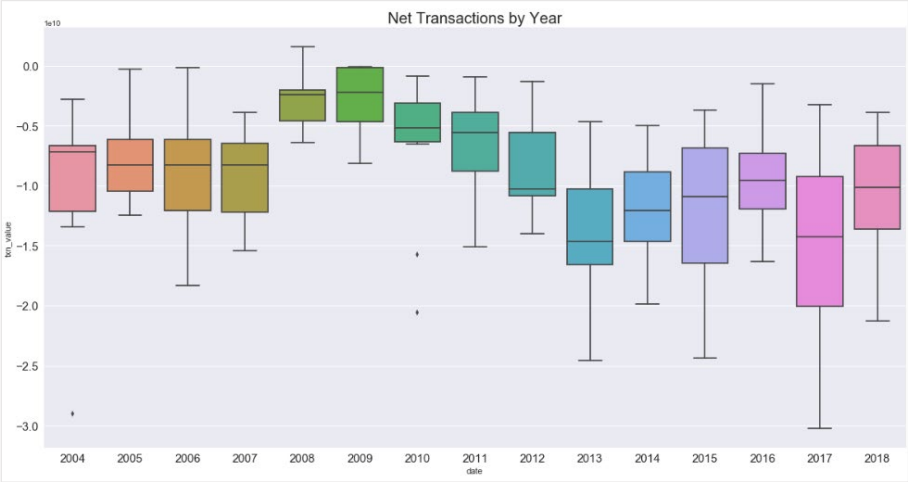
Observations from the charts above:

1. Transaction volume looks relatively consistent across companies in the index with approximately 1,500 companies having at least one transaction per month. Additionally, the oscillating peaks and troughs may indicate a seasonal pattern.
2. Transaction volume increases from 2004–2008 before reaching an all-time high near 60,000 and dropping precipitously, settling in a range of 10,000 to 20,000 per month.

As shown in the chart above, the lightest transaction volume over the 15-year sampling period came in 2008 and 2009. The authors of Identifying Profitable Insider Transactions[6] postulate that the steep decline in insider transaction volume was catalyzed by the global financial crisis (GFC), causing many insiders to hold onto their firm's stock with hopes that the passage of time would allow their remuneration packages to regain their value. With approximately one-third of insider activity tied to stock options (as shown using the insider flags discussed earlier), you can get a sense for just how many insiders were left with underwater options (i.e., those with no intrinsic value).

[6] Pageant Media Ltd. "Identifying Profitable Insider Transactions." *The Journal of Investing* 21 (2012): 61-75. https://joi.pm-research.com/content/21/2/61, (accessed May 2020).

The boxplots below display the distribution of net transaction values over time, highlighting the tightening of dispositions throughout 2008 and 2009.



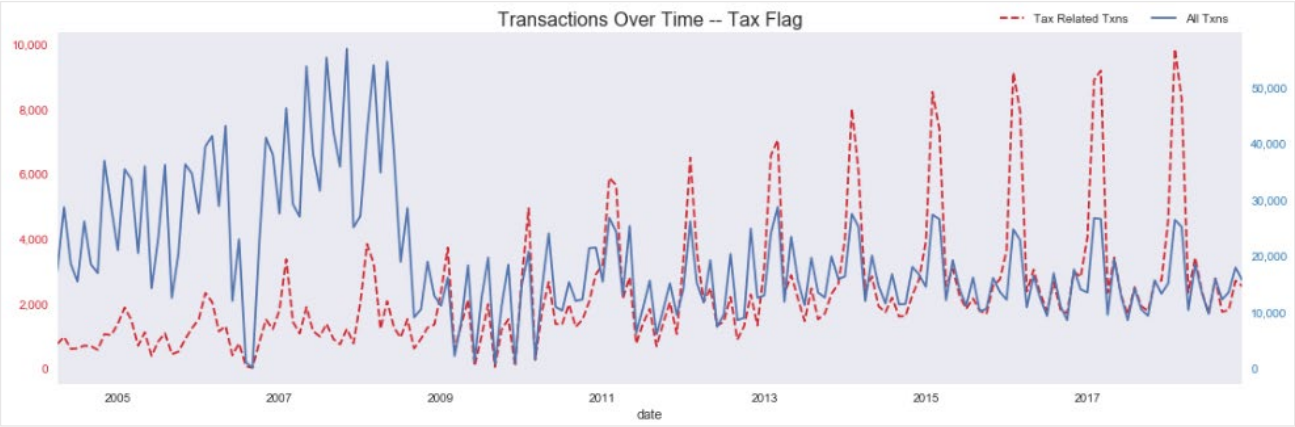
Source: FactSet Research Systems Inc.; 2iQ Research

We can explain the seasonal pattern in the data by thinking through when insiders are awarded their company's stock and when dispositions of existing allocations can be made. Issuances are commonly tied to year-end remuneration packages where the year-end is tied to the company's fiscal calendar (many, but not all, align with the calendar year-end).

Both stock acquisitions and dispositions are subject to what is commonly referred to as a "blackout period"[7] that prevents equity recipients from trading around the company's report dates. That is a slight, albeit important, distinction as it offsets the seasonal pattern from quarter end and helps explain why on average the largest volume takes place right before or after the calendar quarter end.

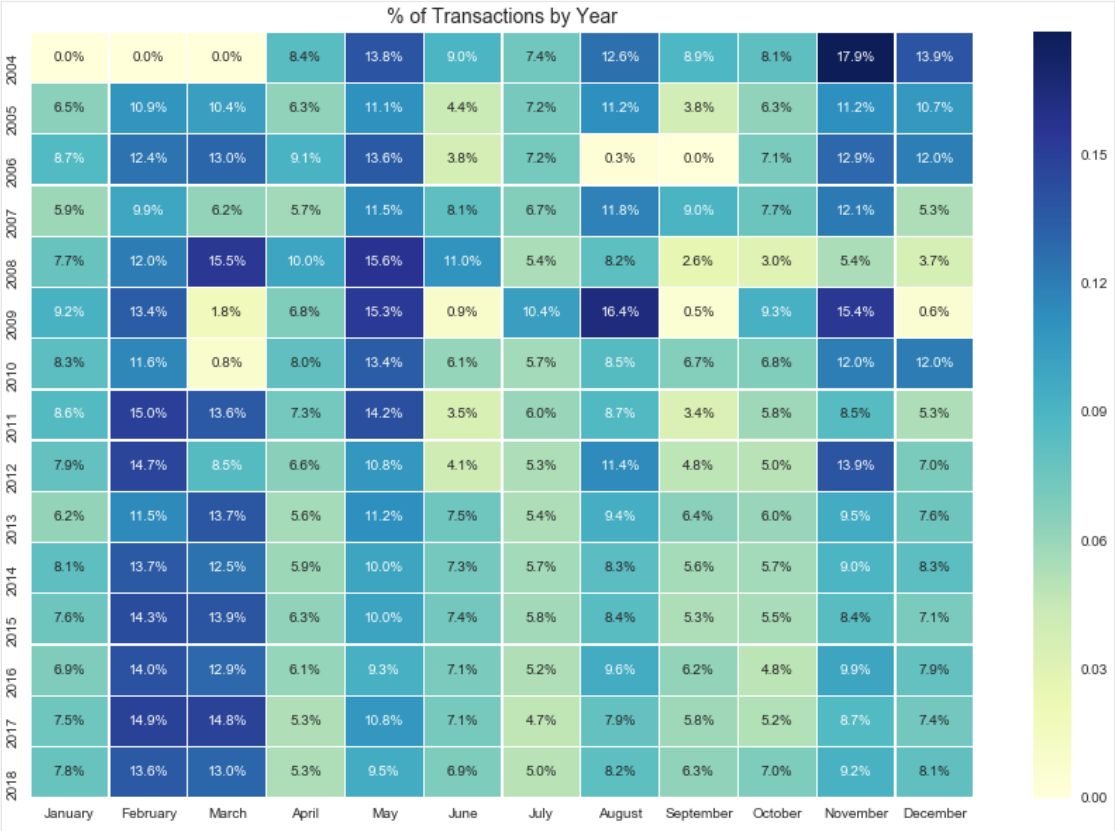
This seasonal pattern is perhaps most clearly demonstrated when viewing the number of transactions tagged as being related to tax settlements. In the chart below, we see all transaction activity (blue) vs. transaction activity flagged as tax related (red). The annual spike occurs each February/March (depending on the year) and reflects the fact that many transactions are accompanied by dispositions to cover to cover tax-related settlements.

[7] Schneider, Bob. "What Is A Black-Out Period?" *Investopedia*. May 27, 2018. https://www.investopedia.com/ask/answers/08/blackout-period.asp, (accessed May 2020).



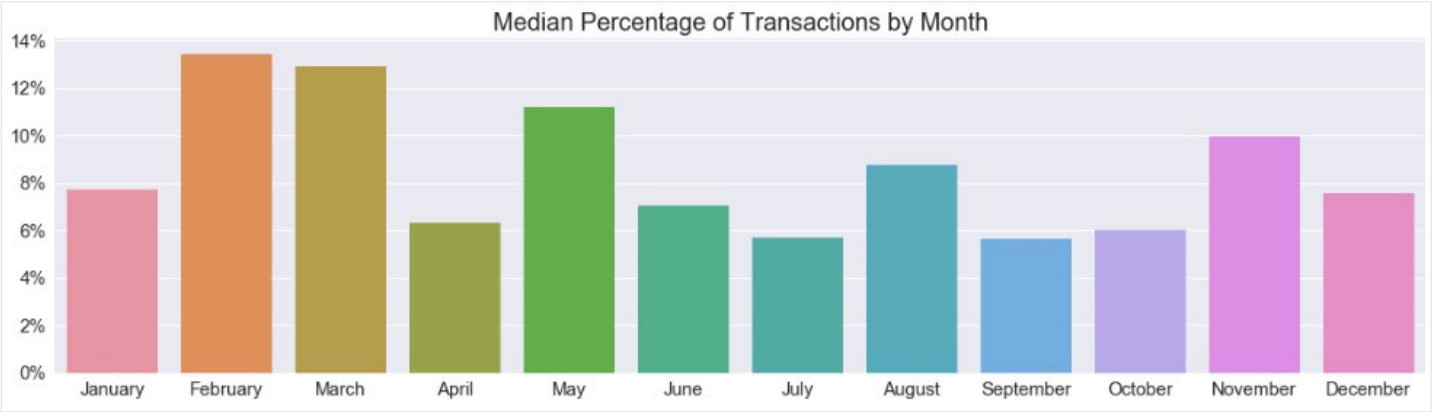
Source: FactSet Research Systems Inc.; 2iQ Research

The year-by-month matrix highlights this seasonal effect as well. We see the percentage of transactions occurring each month as a proportion of the entire year with the highest transaction volume consistently taking place in February and March, slightly after year-end. In addition, we see a decline in volume beginning in August 2008 and continuing through the end of the year.



Source: FactSet Research Systems Inc.; 2iQ Research

Summarizing the matrix by taking the median value for each month allows us to identify the months where insiders consistently place their trades. As described earlier, this elevated activity takes place slightly off calendar quarter ends (i.e., February, March, May, August, and November).



Source: FactSet Research Systems Inc.; 2iQ Research

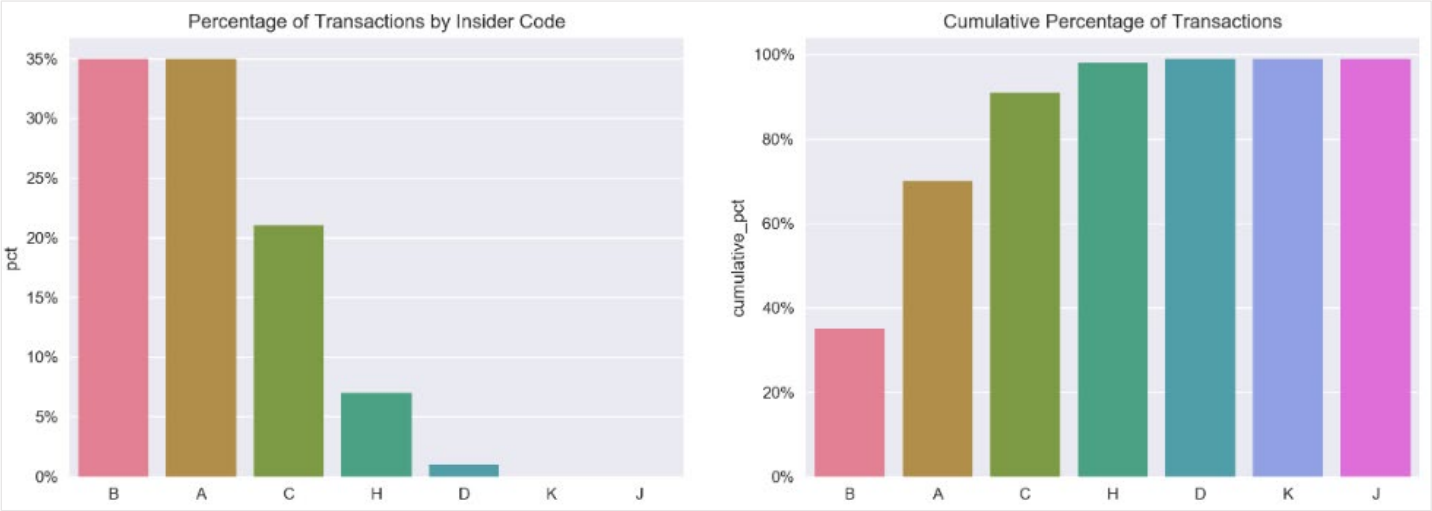
2iQ has a well-defined taxonomy for classifying insiders into several levels, from A-S, each of which represents a different level of the corporate hierarchy present in the dataset. This taxonomy simplifies the process of identifying the "who" behind the transaction data. Insiders belonging to class A represent the most senior levels of the taxonomy, typically consisting of the C-Suite (e.g., CEO, CFO) and subsequent classes continue to partition out the corporate hierarchy monotonically (based on seniority). The table below from 2iQ displays the insider levels along with descriptions for each class.

2iQ Insider Levels

insiderLevel	Description
A	Top insider (executive board , chairman, Top 5)
B	Upper level management (executive committee, Top 20)
C	Non executives, supervisory board, Board of Directors
D	Lower level executives
E	Legal entities, funds and trusts
F	Outsider (Finland only, not included in the feed)
G	Family and other relatives
H	Partner, large shareholder, founder, investor, family holding
J	Custodian
K	Government
S	Issuer (not included in the feed)

Source: 2iQ Research

When grouping transactions by insider levels, we see more than 90% of the volume stemming from levels A, B, and C. This information will come in handy in the next part of this series where we'll limit our analysis to the first three levels of the taxonomy. For a more granular look at the titles attributed to these levels, please refer to the corresponding Jupyter notebook.



Sources: FactSet Research Systems Inc.; 2iQ Research

Conclusion and Next Steps

In Part I of our exploration of the 2iQ Global Insider Transaction dataset, we covered a lot of ground. Staying true to the process we defined at the outset, we began by framing the problem, gathering the data and relevant variables to test the problem statement, and exploring the data that was subsequently collected. That last step required us to understand the temporal properties of our data, making decisions on how to best handle perspective dates, and analyzing the resultant dataset.

In Part II, we'll build upon this analysis and move to feature engineering where we'll work with the transaction-level data and our knowledge of the subject matter to construct informative variables required to create a predictive model.

Visit the Trust the Process: Exploring 2iQ Insider Transactions webpage for more details and resources on Part I, including the corresponding Jupyter Notebook, SQL queries, and utility modules.

Important Notice

The information contained in this report is provided “as is” and all representations, warranties, terms and conditions, oral or written, express or implied (by common law, statute or otherwise), in relation to the information are hereby excluded and disclaimed to the fullest extent permitted by law. In particular, FactSet, its affiliates and its suppliers disclaim implied warranties of merchantability and fitness for a particular purpose and make no warranty of accuracy, completeness or reliability of the information. This report is for informational purposes and does not constitute a solicitation or an offer to buy or sell any securities mentioned within it. The information in this report is not investment advice. FactSet, its affiliates and its suppliers assume no liability for any consequence relating directly or indirectly to any action or inaction taken based on the information contained in this report. No part of this document (in whole or in part) may be reproduced, displayed, modified, distributed or copied (in whole or in part), in any form or format or by any means, without express written permission from FactSet.

About FactSet

FactSet (NYSE:FDS | NASDAQ:FDS) delivers superior content, analytics, and flexible technology to help more than 128,000 users see and seize opportunity sooner. We give investment professionals the edge to outperform with informed insights, workflow solutions across the portfolio lifecycle, and industry-leading support from dedicated specialists. We're proud to have been recognized with multiple awards for our analytical and data-driven solutions and repeatedly ranked as one of Fortune's 100 Best Companies to Work For® and *Best Workplace* in the United Kingdom and France. Subscribe to our thought leadership blog to get fresh insight delivered daily at insight.factset.com. Learn more at http://www.factset.com and follow us on Twitter: www.twitter.com/factset.