Factor Investing with Python #1 Data

Mr. Q Follow
Jan 8 · 6 min read ★

Data comes first when we are talking about Finance and Python. However, I am not only talking about capturing the raw data but also the way to structure it and ideally creating a financial analytics API around the data.

> Financial Analytics Data API (not raw data) should help us do research quicker and manage the process more consistantly.

This is such an important and fundamental piece of the modern asset management stack. This is the reason why big asset managers are spending millions on such Data APIs.

What exactly is needed?

Point-in-time Data

First of all, we need raw data and it should be point-in-time data, or short for PIT data (at least we call it this way in the industry). So what exactly is it?

> It is the data points always associated with an as of date.

You may ask, hold on a second, isn't it all the financial data has an as of date? For example, you use Yahoo Finance to download price series and obviously it has got a date for open, close, high and low.

> Yes, the daily pricing data is Point-in-time already and it is the simplest data set you may find in the financial industry.

But think about fundamental data, for example, net income or EBITDA. They are coming from financial reports. You may still use Yahoo Finance downloading the data from the financial reports, but what you have is a Fiscal Period Date, e.g. 2019 Annual (2019–12–31). This is **NOT** an as of date, because the company is not going to release the financial reports on the Fiscal Period Date. Most companies report some time after the Fiscal Period Date and usually about several weeks later.

> Fundamental Data, which is the most used for Factor Research, is not widely available as Point-in-time. Remember the Fiscal Period Date is not as of date.

But why is that important?

The key to Factor Investing is Factor Backtesting and Strategy Backtesting. No matter what backtesting you are going to do, you DO NOT want to pretend to know the future, otherwise, your return is good simply because you got the future number to pick up the stocks. As we know, the actual release date of the financial reports are all after the Fiscal Period Date, so we CAN NOT it as if it's as of date.

Except for the Fundamental Data, there are other data types we need with as of date. Broker Estimate is one of them and the most used one for the Factor Research. For Broker Estimate, we need to know when is the estimate is made and what is the target fiscal period (quarterly or annually). There are other things we need consider such as does the estimate include any special items and is it comparable to the reported figures. That is why usually Broker Estimate is one of the most expensive financial data you may find in the market.

> Broker Estimate data is expensive and not widely available for reasons!

Finally, one more typical data people may think it's point-in-time by mistake is Economic Data. Like financial reporting data, when you see the GDP number for 2019, it is never released on December end of 2019 and we need to know the reporting date to be able to backtest with it. Also, similarly, the Estimated Economic number needs a consensus date too.

Easy to use API

As I said in the beginning, the right raw data is not enough, we need excellent financial analytics API to access the data. If you only worked with pricing data, you may not feel

so but for Factor Investing, we need to deal with fundamental data, which makes the API essential.

Let's see an example. Say we want to use the PE Ratio as one of our factors. Imagine we only have raw data files or a database giving us the following tables:

> ***income_statement_ttm****: [ticker, release_date, preiod_date, period, currency, sales, cost, net_income, ebitda…]*
>
> ***balance_sheet_ttm****: [ticker, release_date, preiod_date, period, currency, total asset, total liability, total equity, share_outstanding…]*
>
> ***table_price****: [ticker, date, currency, close…]*
>
> ***table_fx****: [fx, date, rate]*

To use those tables working out a PE Ratio as of a certain date, you will need some works. Taking ebitda from the income statement, share_outstanding from the balance sheet and price from price table, given the as of date. Then you need to check the currency and do the conversion with fx table accordingly before doing the division for the PE Ratio.

This may not seem to be that complicated but what if we want to use an average PE, or current PE against the historical average, or look at the fiscal period trend of PE, etc. We probably don't want to work with SQL and code that many lines all the time, especially this is only a single factor on a specific date and what need is multiple factors backtesting (going back to the history).

A typical analytics API way maybe like the following:

> *get_fundamental(ibm, pe_ratio, ftype=ttm, poffset=0, dates=2019–12–31)*

I am just making it up, but the point is that if we have such an analytics API, it would be much easier to construct our Factor Backtesting and Strategy Backtesting more consistently. Also, it is easier to check out if there's any issue.

How do we get the Data and API?

Option 1: Data Providers

Bloomberg, Factset and Refinitives, those are big names in the financial data industry. They sell not only the data but also the technology i.e. ready-to-use analytics API to users as a one-stop solution. Products are amazing but expensive.

There are other smaller data providers, such as IEX Cloud. They may offer data for limited markets and usually only provide the raw data. So to be ready to use, we still need to do some works. However, they are much cheaper and affordable by individual investors or small teams.

> Too expensive for an individual or a small team. But if you have enough budget, this is the best solution, because the analytics API is ready for you directly.

Option 2: Web Scraping

Just put this option as a theoretical solution, but you really don't want to do so, it's too time-consuming. Unless you like organising the data and inventing is not your main goal, you would not like to go for this solution.

> Wouldn't recommend it and too time-consuming, still may not be giving you the right data.

Option 3: Hybrid + Compromise

There are many free data to use, such as Yahoo Finance. But for some fundamental data, you may still want to buy from small data providers. Then in some cases, we had to make some assumption to avoid big data bill but still doing sensible research. For example, I may have to assume all the company will report the earnings three weeks later. Not perfect solution, but makes sense given the constraints of the cost for research.

> Most people with less budget for research would do.

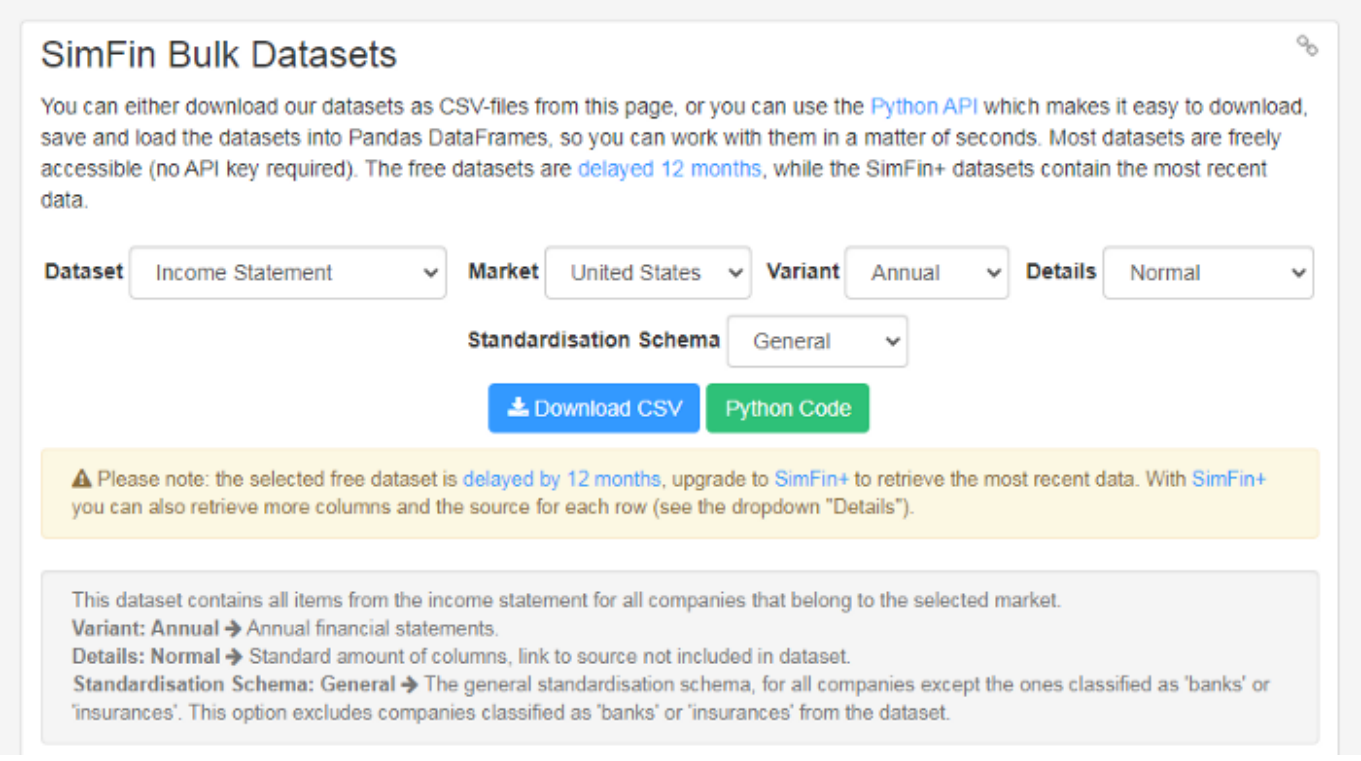
What we will use for this series?

We will use SimFin US market sample data for learning purpose. The latest data will be a premium product and they only have US and Germany data, but for learning purpose, it should be more than enough to use their FREE samples.

> The most important thing: SimFin's fundamental data has the as of date, i.e. point-in-time and backtesting ready.

The bulk data needed

To prepare for the next session, where we will illustrate how to create simple analytics API around the raw data, we will need those bulks and you may download from SimFin.



> *US Companies*
>
> *US Share Prices*

> *Industries*
>
> *US Balance Sheet, Income Statement, Cashflow (Annual, Quarter, TTM — Trailing 12 Months)*

Once you download the data, please unzip them and save the csv files into a folder, so that we are ready to go further from our next session.