Big Data in Finance - Part IV

CRSP and Compustat: an Application to Quant Finance

Kirsten Burr*

March 28, 2025

During this course you are going to learn how to work on two of the most widely used datasets in finance: CRSP and Compustat. The course will be based on the hands-on application of quantitative investing and is designed to give students exposure to up-to-date academic empirical work in asset pricing and its application to the practice of quantitative finance. The course is especially appropriate for students contemplating analytical finance and quantitative money management, and provides many tools and concepts that are essential for a career in quantitative investments. The course is also appropriate for students contemplating writing empirical academic papers in asset pricing or corporate finance using data on US publicly traded firms.

Class Material

All class material is available in a GIT repository in GitHub. You can clone or download here.

Schedule

- Lecture 7: Introduction, CRSP and Compustat basics
- Lecture 8: Application to asset pricing: Factor Replication and Quantitative Investing

Prerequisites

1. Install Anaconda
    - I use use **Anaconda**, one of the most popular Python distributions.
    - Anaconda is easy to install and comes with a great package management system called *conda*.
    - In class, I am going to use **Python 3.9.13**.

*Columbia Business School. I am greatly indebted to Lira Mota and Ritt Keerati for sharing their teaching materials for this course with me. If you find typos, or have any comments or suggestions, then please let me know via kburr26@gsb.columbia.edu.

- Follow this link to install Anaconda. Make sure you choose the correct operating system.

2. Working knowledge with Python.
   - Choose an IDE in which you can run python code (.py) and notebooks. I recommend using Jupyter Lab or PyCharm. You can install it using Anaconda or, for PyCharm, it can be downloaded for free at download PyCharm.
   - Student and faculty members license is for free, you only need to apply at PyCharm license.
3. WRDS direct connection with Python.
   - WRDS has built a Python module that allows direct download of data sets from WRDS services in Python. This is very convenient and we are going to use this tool in class.
   - In order to use the direct download you need to setup your connection beforehand by following the instructions here.
   - We will use some SQL synax when querying data from WRDS Server.
     - SQL is powerful tool for managing relational databases, and many intitutions utilize to maintain and process information.
     - If you aren't familiar with SQl queries in general you may want to take a look at this cheatsheet.
     - Although we won't cover it further, if you are interested in better understanding SQL, please feel free to take a look at this Harvard CS50 lecture on SQL as it is also quite informative. (I highly recommend this series for those interested in programming in general! – though unrelated to this course)
   - The WRDS support is very responsive, so make sure to email them if you need help to setup your connection.
4. Working knowledge with GIT.
   - All course material will be available in GitHub repository. Access here.
   - Make sure to setup a GitHub account here and study the GIT basics here and here.
   - Using GIT will change the way you collaborate in research projects, making it much easier to organize and keep track of changes made by you or your colleagues.
5. Power-up your Jupyter Notebook and verify packages installed. (Pre-Class Assignment)
   - Notebooks are great to produce documents you intend to present, and we are going to use notebooks during class.
   - Please run the todo_before_class.ipynb notebook before we meet on 3/28 and submit the html on Canvas.
   - This assignment is not for a grade. It is simply a check to ensure everyone has the necessary packages.
   - Here you can find a description of very useful plugins for Jupyter Notebooks. I highly recommend that you install the suggested plugins.

Homeworks

There will be two homeworks.

1. Due Thursday, 4/3: Exploring CRSP and Compustat.
2. Due Thursday, 4/10: Asset pricing factors replication.

Lectures

Lecture I: CRSP and Compustat Basics

1. Introduction
   (a) WRDS basics
   (b) How to download data into Python
2. CRSP
   (a) Securities File Monthly
   (b) Securities File Daily
   (c) Events Table
   (d) Stock Header Info
3. Compustat
   (a) Fundamentals Annual
   (b) Fundamentals Quarterly
   (c) Names Table
4. CRSP and Compustat merge

Lecture II: Asset Pricing Factor Replication

1. An overview of Fama and French factor construction technology
2. Characteristics Construction: Fama and French (2015) + Momentum
   (a) Size (CRSP)
   (b) Book to Market (Compustat)
   (c) Profitability (Compustat)
   (d) Investment (Compustat)
   (e) Momentum (CRSP)
3. Replicate Fama and French (2015) five factors and momentum factor.
4. Performance evaluation: Alpha evaluation and Fama-MacBeth regressions
   (a) Reversal
   (b) Momentum
   (c) Characteristic efficient portfolios

References:

- **Daniel, Kent and Tobias Moskowitz, "Momentum Crashes,"** 2016, *Journal of Financial Economics*, 122(2), 221-247.
- **Daniel, Kent, Lira Mota, Simon Rottke and Tano Santos, "The Cross-Section of Risk and Returns,"** 2020, *The Review of Financial Studies,* 33(5), 927–1979.
- **Fama, Eugene and Kenneth French, "Common Risk Factors in the Returns on Stocks and Bonds,"** 1993, *Journal of Financial Economics*, 33, 3-56.

- **Fama, Eugene and Kenneth French, "A Five-Factor Asset Pricing Model,"** 2015, *Journal of Financial Economics*, 116, 1-22.
- **Harvey, Campbell, Yan Liu and Heqing Zhu, " . . . and the Cross-Section of Expected Returns",** *Review of Financial Studies*, 2016, 29, 5-68.
- **Jegadeesh, Narasimhan and Sheridan Titman, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency,"** 1993, *Journal of Finance*, 48, 65-91.
- **Korajczyk, Robert and Ronnie Sadka, "Are Momentum Profits Robust to Trading Costs?,"** 2004, *Journal of Finance*, 59(3), 1030-1082.
- **McLean, David and Jeff Pontiff, "Does Academic Publication Destroy Stock Return Predictability?"** 2016, *Journal of Finance*, 71, 5-32.
- **Novy-Marx, Robert and Mihail Velikov, "A Taxonomy of Anomalies and Their Trading Costs,"** *Review of Financial Studies*, 2017.