MFIN 7035 2025 – HomeFun 2

Objectives/Comments

1. SQL practice – stop asking me to send you data! You are the junior person go collect yourself! (inching toward real life, sorry)
    a. Practice linking tables
    b. Practice accessing data
    c. You are restricted from doing anything but the graphing or regressions in SQL, the point is that you are being forced to run things efficiently.
2. Data analysis practice
    a. Practice defining metrics, making coherent statistics and explaining them
    b. I am intentionally vague, as often times in the real world (and in academia, mind you I work a lot with companies now too), people just want a reasonable answer, not a perfect answer or there may be no absolutely clear one.
    c. Some large language model practice/NLP practice, if you so please. Because why not, this is the capstone course – unleash all that you know.

3. Due date: April 13

What to do:

- A PDF writeup (Kenny's preference). Submit code notebook as well, but writeup will be expected to be the main thing graded. One submission per group. Ideally, the writeup would describe the thought process you took in the code so Kenny can quickly grade. This does not mean screenshotting the code, but rather outlining (bullet points) or just in plain English.
- Group-size-dependent workload:
    - easy = 1, medium = 2, hard = 3
    - 1-2 people: 2 points. You may do up to 3. Best 2 = 50% weight, worst = 10%.
    - 3-4 people: 3 points, may do up to 4 points. Best 3 = 33.33% weight, worst = 10%.
    - 5-6 people: 4 points, need to hit 6 points for extra credit.
- For groups of size 3 or more, participation form required. Not submitting a participation form is itself a form of non-participation and will result in a grade deduction.

Some notes about where data is:

- Clickhouse (maybe StarRocks, definitely no MySQL)
- Lots of the datare in common_goods.job_postings..*, or the appropriate database (CRSP data is in crsp.*, tiingo data is in tiingo_iex.*)

Capstone oriented SQL/data questions for ToscaFund

I would classify these as hard. It is also probably necessary to use SQL.

- For these questions, follow the steps below to construct a factor that could potentially be used to predict ETF returns.
- Report statistics that show you spotchecked the data's summary statistics to make sure that the number of observations is what you expected, and that temporal and ticker-level coverage matches the thematic fund universe from ToscaFund, and show some way of assessing whether the summary statistics make sense relative to papers in the past or just common sense.

- Then, regress the factor on ETF returns via Fama-MacBeth. Or do a portfolio sort as demonstated in the class slides.

Each one of the points below is a separate task.

- Using OptionMetrics, calculate the changes in call implied volatility per An, Ang, Bali and Cakicki (2013) to see if $\Delta CallIVOL$ is predictive of ETF returns.

- Collect ETF flows data from Bloomberg per Ben David et al (2018) or another data source. CRSP might even be OK, not sure.[1] Compare your summary statistics to theirs. Then, let's see if per Brown, Davies and Riggenberg (2020), there is any predictive relationship within this basket of universe and future fund returns.
    - Bonus point (making it 4 points of 3): could you please compare Bloomberg with Tiingo, CRSP and Eikon to tell me how crazy using these other data sources? Please email me directly your answer as I am actually quite interested.

- Write SQL queries to investigate the following. How many of Tosca's ETF universe can be found in the crsp.holdings database? In turn, how many of the stock-months have coverage in the JKP factor library? What other broad types of factors other than momentum could you use?
    - For a factor or bucket of factors that you choose (for example value), make sure your query cross-sectionally ranks the stock on a percentile basis relative to other stocks at the same time (a P/E ratio has no inherent meaning).
    - Connect these signals and then do a weighted average at the ETF level.
    - Report a regression table of whether or not this factor predicts ETF returns.

- Using tiingo.news, for each ETF in Tosca's universe, come up with a set of keywords that relate to theme.

[1] *Ben-David, I., Franzoni, F., & Moussawi, R. (2018). Do ETFs Increase Volatility? The Journal of Finance, 73 (6), 2471{2535.*

    - Write a SQL query that queries the entire tiingo.news for that theme, then aggregates by some day or week the intensity of news mentions. You may want to consider either absolute volumes or percentages of .
    - One thing you may want to consider is

Capstone oriented SQL/data questions for WorldQuant

- **Assess the feasibility of patent comomentum across borders (medium)**
    - Use the GCPD data to first find all US firms with patents using GCPD data. Do your statistics match those of Lee et al (2020), Technological links and Predictable Returns?
    - Second, collect all Japanese firms with patents in the US using GCPD data.
    - Report summary statistics. If the goal is to sort Japanese stocks into portfolios based on US peers, are there a large enough cross-section to exploit?

- **Assess the feasibility of supply-chain across borders (medium)**
    - Examine the Factset Revere database, and report the number of unique companies with *gvkey* (you have to figure out how to link these different throughout the fset_revere_debt, factset_common_2024, and various Compust globial tables).
    - Plot the number of stocks per year in Japan that have publicly listed US customers. Plot also the number of unique US firms that comprise those customers. Is there enough variation to form a quintile portfolio?

- **Assess the feasibility of past comovement implied networks (very hard, 4 points)**
    - Based on now-portfolio manager and former HKUST finance prof Binying Liu
    - This problem is classified is very hard because it is computationally quite difficult. This is a good problem, however, because if you are seriously interested in big data, I don't think there's a way to do this without being smart about computation.
    - First, you must residualize Japanese and US returns, I would presume best to residualize with respect to local factor models or perhaps a US + Japan factor model. In other words, perhaps best to residualize to both $mkt_{US} - r_f$ and $mkt_{JPN} - r_f$. For ease of calculation, you can perhaps use the past 24 or 36 months instead of weeks per the original paper.[2]
    - Second, post residualization, you need to calculate pairwise t-statistics to ask whether two stocks are significantly positively correlated, or Binying Liu equivalently assumes that a correlation above 26% or below 26% is statistically significant based on 100 weeks – you may have to adjust this threshold based on monthly returns.
    - How sparse is this network? Does it allow for lead-lag returns?

- **Assess the feasibility of analyst networks across borders (medium)**
    - Using IBES, are there any stocks where the ANALYS (ID for an analyst) is shared across Japanese and US firms?

[2] If instead you need daily US/Japan returns, I can put them on Clickhouse.

- **Information discreteness for comomentum (hard)**
    - Compute the information discreteness measure of “Frog in the Pan: Continuous Information and Momentum” and show that continuous momentum outperforms discrete momentum over the original sample period of the paper. As discussed in class, this will help you because if you create a comomentum measure, the discreteness should interact adversely with momentum strategies while continuity of returns the opposite (read Shiyang’s paper “A Frog in Every Pan..”).

- **Patent comomentum (medium)**
    - The original patent comomentum paper uses patent classes assigned by USPTO. Lots of academics have created their own measures, can you search the web and find comprehensive patent classifications that can alternatively classify patents? Load them into the SQL database (also send me a note).
- **Patent comomentum (hard)**
    - Hard: re-do the original paper with citations, defining peer firms as those who a firm cited. In other words, if I am Microsoft, my peer is Google if I cite their patents but not Samsung if I don’t cite their patents. Think about a logical way I would weight the firms – would I citation-weight or value-weight?
    - Re-do the original technological links and predictable returns paper using patent citations instead of tech classes.
    - How many Japanese firms cite American patents, and vice versa? You can find the gvkey-patent number mapping in patents.gpcd_data

- **Hard: the accounting differences project**
    - In order to target accounting differences where accounting data treatment might be suspect in standard databases, it may be helpful, as noted in class, to monitor the decline in performance of a factor in the US in Japan that uses accounting data.
    - Use SQL to rank stocks cross-setionally (per lecture note 2) and construct long-short portfolios for Japanese stocks for all factors using the JKP library, and likewise for US stocks. Calculate the average long-short return for each factor in the US/Japan.
    - Plot them against each other and draw a regression line. In general, is there a correlation between what works well in the US and what works well in Japan?
    - Rank the factors by the largest decline in performance in Japan relative to the US. If the top 10 is primarily all factors such as momentum, maybe only do this analysis for Japanese

Capstone oriented data questions for BIlby

Since they give you commodity-level factors, you need commodity level returns. Unfortunately I don’t have any handy.

- **For Bilby project (medium)**
    - Find the returns to commodity ETFs in CRSP. Find one commodity ETF per commodity of interest, if possible. Extract the history. I would think it’s best to either

find an ETF that is meant to track the spot price or which rolls the one month futures.

  - You can also find ETFs for stock sectors that may have been impacted

- **Hard: construct a time series of commodity 1 month futures using Eikon or Bloomberg.** Double check the returns to make sure they are reasonable based on some sort of online data source. Send me the data so I have it, ideally you would write some Python code to systematically pull from Python. If you need an Eikon account I can send mine over.

- **Medium: Write SQL queries to generate a handful of features using the Bilby data. Calculate the following:**

  - Calculate a rolling disagreement measure between the last 20 pieces of legislation, using a window function, where disagreement is defined as (positive-negative)
  - Using SQL, calculate the number of regulations in the last 20 days
  - Using SQL write down five other features you think might actually predict returns

  and then run a pooled OLS regression with a time fixed effect – or a Fama-Macbeth regression which basically

Capstone oriented data questions for Benjamin.ai

Instead of Tesla, you can use whatever arbitrary stock you want but humour me and pull the data from the course server using SQL.

- **Easy:** Here’s a simple unit test. Extract data for Tesla using SQL and calculate the market beta in Python. Remember that you must subtract rf. Extract this data, write it to a .csv file or some format that Benjamin.ai can read, and you can use this as one example test for Benjamin.ai to verify (i.e. can you properly calculate the beta)
- **Medium:** write a query using SQL that extracts the **rolling beta** for Tesla. Then have Benjamin.ai verify it.
- **easy:** use SQL to calculate the cumulative return to the SPY ETF since 2000 or some starting period in the crsp.dsf. Then, based on that same SQL query, calculate the arithmetic sharpe ratio annualized. Have the AI verify it.
- **Medium:** calculate the historical in-sample returns to SPY, TLT and GLD and MCHI over say a 10 year period. supply said data to Benjamin.ai, and calculate the optimal portfolio assuming in-sample returns and covariances, and verify using your own calculation.
- **Easy (feel free to do this 3 times):** propose your own unit test, but write a SQL query (a dash of Python is OK), justify why this is a useful thing to do, and verify using Benjamin.ai

Capstone oriented questions for Premialab

- **Hard:** For whatever calculation you were going to do in Pandas, try to use instead Polars, chDB (clickhouse package) or DuckDB. Compare the timings of DuckDB or Clickhouse or Polars. You can also feel free to add one more performant package I may be unaware of. Report ten different data operations and the time it takes to use pandas versus two of these solutions. Create a table similar to clickbench where, for a given set of ten queries, you have three columns comparing the timings of these different methods. Verify also that the calculation is the exact same in all languages.
- **Medium:** Given some reporting data, think of a way that generative AI LLMs can be used to summarize information in an appealing way for the user of a dashboard. Demonstrate that using an API call of your choice.

Practice with LLMs

Matching using LLMs?

- **Medium:** Use my Perplexity API key in the Dropbox (or your own) and classify CRSP asking for the correct ticker/CUSIP for the company name for 100 stocks in the CRSP universe in 2024. How accurate is it? (class example on Confluence using older model)
    - What is the accuracy rate if you assume CRSP is absolutely correct?
    - Conceptual question: do you expect higher accuracy for 2014, using Perplexity or any LLM for that matter?
- **Hard:** write a sensible set of prompts that checks the results ejecting false positives, perhaps using a large language model, or do something else that is meant to guarantee accuracy of the match.
    - How does the accuracy rate improve?

Can Deepseek or some other LLM predict bitcoin? (hard)

- I have some news data for the last 3 months. Write a prompt that interprets the news and produces a prediction for the next day. How accurate is this prediction and plot the profit/loss curve? I suppose this would mean combining the multiple news articles into your prompt. Note: probably best not to use a search-enabled LLM for obvious reasons.
- There's also price data in SQL, feel free to use the EOD price or some price pulled off of the intraday.
- Conceptual question: Would we better off extending this backtest back to 2020? In the case of Deepseek's training data, why or why not?

Evaluate some KOL on the web? (hard) if anyone wants to do this more expansively as a final project, go for it, expanding along the dimension of a few KOLs.. I would be super curious!

- There is skepticism among industry quants as well as academics about technical trading (although maybe I'm wrong, I dunno lol). Go to https://www.youtube.com/c/TradeBrigade or another KOL of similar stature, collect one video per week over the last year (can stop at 20-30 videos), and download a transcript. Then, use an LLM to classify the prediction for perhaps the S&P 500.
    - What is the hit rate? Note that sometimes their predictions are conditional (if this weird shape crosses another weird shape etc then my crystal ball tells me to buy, but if not my crystal ball tells me to sell, etc.)
    - What are the list of justifications used (RSI, moving average, etc.)?

Practice with industrial grade ML (hard!)

- Using older Tiingo data, use an API to perform textual embedding on the news on BTC (perhaps Mistral API, perhaps BERT embeddings), run a ridge regression splitting the sample into training/validation set.
- Run that same model out of sample using the last 3 months. Would such a model have done well in predicting performance?

What are the most annoying complaints about the HSBC app? (hard)

- Take a sample of 100-200 reviews, read a few, and extract the topics of the complaints people make. To randomly sample, you may want to generate a uniform random number and sample the top 100.
- To validate your topics, run a regression of the presence of the topic against the score. If it's a complaint, then the regression should show a negative relationship. If it's a positive, then mentions of this feature should show a positive relationship with
- What are the most frequent topics in the random sample?

Alternative data

Go into Eikon or Bloomberg

- **medium:** Report back every alternative data metric available for Netflix, the world's most alt-data'ed stock. Screenshot the terminal and extract the metrics from the terminal to prove you did it. Also okay: Luckin Coffee, BYD, BABA, Tencent, or some other generally retail-oriented company.
    - What metrics is each alternative data indicator meant to approximate?
    - What metrics are available that we did not discuss in class? What metrics intersected with what we discussed in class?
    - Collect the time series, plot them, and make one or two graphs and put them in pptx. Selfishly, if they are good, I can use this for teaching ☺
- **Hard:** take one or two of the numbers, and run a regression of changes of Netflix visits quarter over quarter over contemporaneous changes in Netflix revenues. Or, if the indicator

is not meant to approximate revenues but rather something else, such as CAPX, you can do that instead. How accurate is this?

Go into Dewey, pull the data for Similarweb

- **Medium:** Pull the data for Netflix from Similarweb in the dewey database. How similar are the numbers for user growth to Netflix reported numbers? Feel free to swap out for a Chinese company like Baidu or Ctrip, but obviously don't do something that won't work like Luckin Coffee since the digital datasets won't match.
  - What is incomplete about the web traffic data?
  - Is the data an under or overestimate of users?
- **Hard:**
  - Run a regression of changes of Netflix visits quarter over quarter over contemporaneous changes in Netflix revenues. How accurate is this?
  - Do a single variable regression of web traffic first
  - Add in features from the app table, and re-run this regression. How much does it improve?
  - Is there any data from Safegraph, the mobile store visit dataset? Why or why not?
  - Conceptual question: would you be better off pooling multiple firms together to estimate this model or estimating a company-specific model?
- **Very hard (4 of 3 points):**
  - Go into IBES and compare your model forecast implied growth to the expected growth rates for SALES from the ibes.statsum_us file. Would you have guessed the direction of the surprise correctly using your forecast model for the last few forecasts? Ideally, you'd separate your forecast model from your predictions, so that if for example you're predicting for Q1 2024 you aren't training said model using data from Q1 2024 or Q2 2024 (i.e. lookahead).

Machine learning

I gave code for Gu, Kelly and Xiu (2020). Take the US model my PhD student exogenously gave. I will probably add a few questions very soon.

Hard: Allan Timmerman the wrongly spelled Alan shows that tree models way overfit, yet industry practitioners often say XGBoost outperforms linear models by far.

- Why does XGBoost overfit?
- Although not academically elegant, how could one robustify the objective function of the model to prevent overfitting?

Hard: The traditional GKX (2020) setup minimizes validation set MSE, modify the validation set to optimize on a different statistic designed to pick the most profitable hyperparameters. Train a

model that gives the best MSE in the validation set versus the new metric. Does your OOS performance improve?

More SQL / data questions

- **Medium:** This question helps you match. How many US publicly listed firms are represented in the dewey.similarweb_daily_traffic database on a daily basis? Think about how this linking could occur, given the tables you have. For example, what about Compustat or using the company_ref file?

Hard: department of governmental efficiency (hard)

- **Medium:** I have data from a company called that specializes in government contract data, which maps contracts up until about 2020 or so. This data has been mapped to firms, you might assume that, if the linktables are stable, you can map a firms overall government contract exposure until that point. Note that government contracts are long term contracts and repeating, so its best to not calculate exposures using just 2024 data.
- The govt_contract data up until about 2024 can be loaded onto the server. Or otherwise collected by usa_spending.gov
- **Hard:** Collect data from the course server or Tiingo API to check whether the returns so far since the DOGE campaign can be explained by government contract exposure.