Series of mini-projects to underscore the different and important phases in financial data science –from acquiring a dataset, cleaning and analyzing it as well as iteratively building impactful insights out of it.

**The deadline for submissions is June 18th (Tuesday) at 6pm. Please submit 2 items:**

1. **Jupyter notebook of code (and comments), and**
2. **a powerpoint slide.**

# Course Project 1: Data Cleaning and Exploration

Data Scientists and Analysts spend almost 80% of their time cleaning and analyzing datasets. Currently, we are working with an external research firm who specializes in the application of machine learning to forecasting prices of financial instruments. This firm has developed a proprietary system, that we would like to investigate.

To demonstrate the effectiveness of their forecasting system, the vendor has sent us the attached sample dataset (QF632\_Project\_1.csv). The dataset includes signal values generated by the proposed system as well as historical prices for a well-known broad market ETF.

Before using the data in our production systems, we need to run through a few things:

1. Review the quality of the data, list any potential errors, and propose corrected values. Please list each quality check error and correction applied.
   1. Check object types
   2. Check for nan values
      1. Interpolate or forward fill. If interpolate, be careful not to include future rows
   3. Difference all first
   4. Scatter plot/boxplot for outliers
      1. See if need outlier handling (If no, maybe justify why)
2. Please analyze the signal’s effectiveness or lack thereof in forecasting ETF price, using whatever metrics you think are most relevant.
   1. Cross correlation
   2. Kurtosis & skew
   3. R2 score from Lasso, Ridge & RandomForest
   4. ACF, PACF of signal, ETF
3. Run any exploratory data analysis you think is important and highlight any interesting insights you come across.
   1. Uhhh same as parts 1 &2 unless yall can think of smth else
4. Write a summary for the team addressing your observations about the efficacy and believability of the product, and recommendation for next steps.
   1. Write all of the above in slides
5. Please include all the intermediate steps, and lay out your thinking as well.

# Course project 2: Data acquisition and investigative analysis and exploration

I will provide a few csv files of data science jobs and salaries (QF632\_Project\_2\_a/b/c/d).csv, and I would want you to **harvest similar type data from the website h1bdata.info - focus on data science roles**. This is very broad - data scientists, data engineers, data storytellers, data insights etc. are a variety of possible roles that encompasses similar scopes of work. Please extract them from the website (h1bdata.info). Part of the exploratory work will be in understanding the dataset first (recall broad elements of what you did for Project 1).

*This will be open ended, and you can creatively think of what type of metrics and consequently insights you would like to draw from the combination of datasets.*

*For example, if I were to be a data scientist, which industry would be best? Is being a data scientist or a data engineer more rewarding? Does this depend on level of experience? Does location matter? Which companies are best?*

*There are many ways to aggregate and stratify the data. But you need to start by thinking about the right (or better) questions to formulate first (very important) so that you can lay out your thought process in finding the right sources of data. The quality of insights would depend on a few things - how much useful information can you extract from the data, how much data was the analysis based on, how transparent/details/clarity is there in the narrative. How would you improve the insights if you had more time/money/resources - what other types of data do you think that's out there that you can tap on?*

Proposed Parts:

1. Webscraping & Data Cleaning
   1. Check object types
   2. Check for nan values
      1. Interpolate
   3. Scatter plot/boxplot for outliers
      1. See if need outlier handling (If no, maybe justify why)
2. EDA\*\*
   1. Cross correlation
   2. Kurtosis & skew
   3. ACF, PACF
   4. PCA
   5. May need to analyse data by categories (i.e., Histogram of salaries by year vs loc)
      1. Use the kmeans elbow method?? (not sure)
   6. See wat else prof present
3. Prediction/Modellings
   1. Gridsearch on polynomial degree (1-3), and look at features selected
   2. R2 score from Lasso, Ridge & RandomForest
   3. RandomForest & Feature importance