**Midterm Project: Fraud Detection**

The goal of this assignment is to make you comfortable with using two frequently used datasets in financial analysis – Computat and Audit Analytics, and comfortable analyzing and assessing the earnings quality of firms. The earnings quality measures will feed into some other work we do later in the semester, so this is a great learning opportunity as to how to estimate those using the universe of firms listed on the US exchanges. For this assignment, you will need 2012-2022 financial data (Compustat) at the annual level and 2012-2022 Audit Analytics (Restatement) data.

1. Use the restatement data and financial data from 2012-2018 and examine whether restating firms have lower earnings quality compared to non-restating firms, where earnings quality (EQ) is estimated using any of the three different models covered in class (Jones, Modified Jones, Dechow Dichev, Teoh). Report whether you find any difference between restating and non-restating firms. For restatements, use two different outcomes – all restatements or only sec investigated restatements (the variable is available in the dataset).

2. In addition to abnormal accruals, there are other firm-level factors (i.e., market-based incentives like prior stock performance, need for financing, etc.) that may predict fraud. Include at least 2 of these additional factors in the model above along with accruals and examine whether these factors explain restatements. Explain your findings.

3. Apply the coefficients you estimated from 2) to 2019, 2020, and 2021 data, and predict 100 firms that are most likely to restate earnings in the future; Verify your prediction using 2019, 2020, and 2021 restatement data - how many of them really restate financial statements in each year? Share the story.

4. Create an annual EM index by first sorting firms into deciles based on the three models of EQ (lowest EQ to highest EQ, with lowest firms being in the last decile), and then adding deciles across the three EQ measures to create a total EM Index score for each firm. Re-estimate #2 but using the EM Index (instead of individual EQ measures) and other variables you originally had in 2 and apply the estimated coefficients to 2021 financial data. Identify 10 firms most likely to restate financials. Using qualitative assessment, narrow down the list to 5 firms you believe have the highest potential to restate 2021 financials. Share the story.

Note 1: The restatement dataset has restatement begin and restatement end date variables. Those capture the period of “incorrect” financial statements that were later restated. You will need to convert the dates to an indicator variable where if a firm had a restatement covering any period of that fiscal year, restatement = 1, otherwise 0.

Note 2: Company\_fkey variable in the Restatement dataset is equivalent to CIK code in Compustat dataset. Note that the variable may have to be converted to numerical type to match the formatting across the two datasets.