## Names: Michael Lees and Olivia Venuti

## NetID: mlees and ovenuti

## Dataset Choice 1: Price Volume Data for All US Stocks and ETFs

The chosen dataset encompasses a comprehensive collection of stock price and trading volume data, specific to United States-based companies and Exchange-Traded Funds (ETFs), dating up to the year 2017. This dataset is particularly valuable for financial analysis and predictive modeling projects. For our purposes, we will delve into the stock market information pertaining to three key entities, Broadwind Inc (BWEN), CSI Compressco LP (CCLP), and IAC Inc (IAC).

This dataset includes daily transactional data points such as opening price (Open), highest price point of the day (High), lowest price point (Low), closing price (Close), the volume of stocks traded (Volume), and open interest in the security (OpenInt). These columns serve as vital indicators for predicting stock market trends and are therefore of primary interest in our regression analysis. Each row in the data represents a trading days’ worth of information for a given stock or ETF.

In terms of technical specifics, the dataset contains thousands of rows — each corresponding to a trading day — and the seven columns mentioned earlier. The dependent variable in the context of a regression model would typically be the stock's closing price, as it reflects the final market valuation at the end of a trading day. Conversely, the remaining columns could be treated as independent variables that influence the dependent variable.

The dataset's integrity is ensured by its generation process, which involves collating officially released stock trade information. This process guarantees that the dataset reflects accurate market behaviors and trends.

[Find the dataset here.](https://www.kaggle.com/datasets/borismarjanovic/price-volume-data-for-all-us-stocks-etfs/)

## Project Goal:

The principal objective of our project is to construct a robust regression model tailored for forecasting the closing prices of U.S. stocks and ETFs, with an emphasis on accuracy and reliability. By utilizing the comprehensive dataset from Kaggle, which includes price volume data for all U.S. stocks and ETFs up to the end of 2017, we aim to dissect and understand the nuances of market trends and forces. Our model will endeavor to predict future closing prices based on historical open, high, low, close, volume, and open interest data, thereby providing a valuable tool for investors and market analysts to anticipate market movements and enhance their decision-making process.

The motivation behind this project stems from the complex nature of financial markets and the perpetual quest for better predictive tools. With this model, we hope to capture the multifaceted factors that drive stock prices and provide a quantitative foundation for investment strategies. Our plan is to meticulously preprocess the data, perform exploratory data analysis to identify patterns and correlations, and then apply advanced regression techniques to develop the prediction model. We anticipate that the outcome will not only serve as an academic exercise but also pave the way for practical applications in financial analysis and trading.