Project 3
Seth Beckett, James Sanford-Luevano
CS6830

Project 3: Gathering Data

Introduction

The objective of the following project report is to use statistical analysis techniques to assess the financial data gathered from Yahoo Finance. We decided to use yahoo finance because it has historically been a reputable site for past financial data. Our team decided to approach the data from a financial perspective; focusing primarily on energy. We wanted to analyze the dichotomy between the clean energy and traditional energy markets, as well as isolate the two from greater market trends in the economy overall. Our analysis compares stocks, observes trends during significant world events, and analyzes volume of trades. The results of this project will provide insights into energy markets. The slides for the project can be found here and the code for this project can be found here.

Dataset

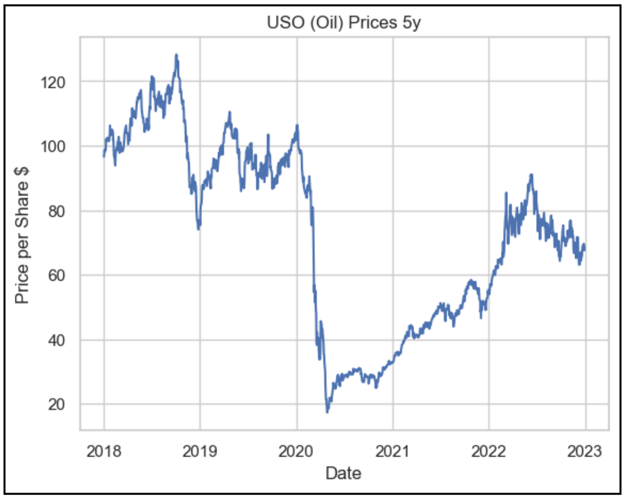
The data we needed was focused on US stock market data. While there are countless energy stocks among the many stock markets across the world, it was most realistic for us to only use the largest US energy ETFs (Exchange-Traded Fund). This is because ETFs are representative of a market as a whole due to their nature of being a pooled investment across a specific sector or commodity. Yahoo Finance has a great API that makes pulling historical stock market and commodity data straightforward, and it’s even simpler when using the yfinance Python library. To obtain each data set, we would simply provide the ticker/symbol, start date, and end date for the data. This returned a table containing the price for open, close, adjusted close, daily high, daily low, as well as containing the volume of trades for each day. We did this for the United States Oil ETF (USO) and the iShares Global Clean Energy ETF (ICLN).

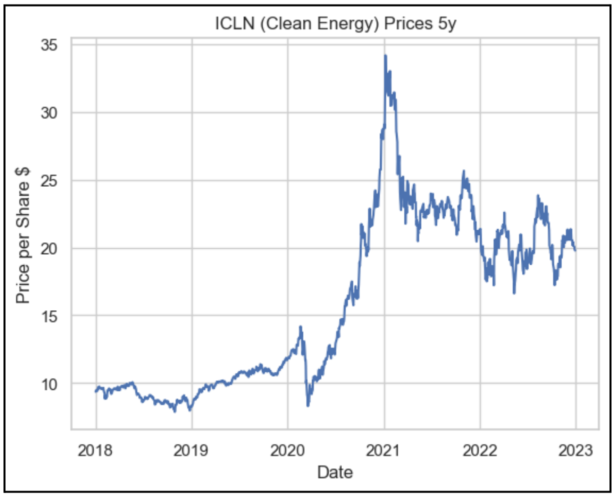
Analysis technique

For basic analysis of our data, we started off by comparing line charts of the ETFs’ prices over the past 5 years, and over the past year. This was suitable as we were showing how a certain variable, price, changes with respect to time. We then found Pearson’s Correlation Coefficient and the associated p-value between each ETF, for each respective time period. This was to see if there was any significant relationship between the two stocks, and potentially predict how one will do based off of the other. We also used histograms to look at the distribution of trading volumes over the two time periods, and used t-tests to see if the trading volumes were significantly different between the two stocks. Finally, we used the standard deviation of stock price to compare volatility between the two stocks.

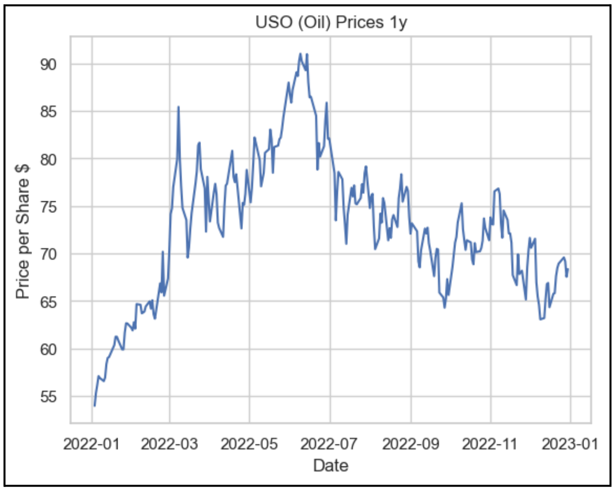
Results

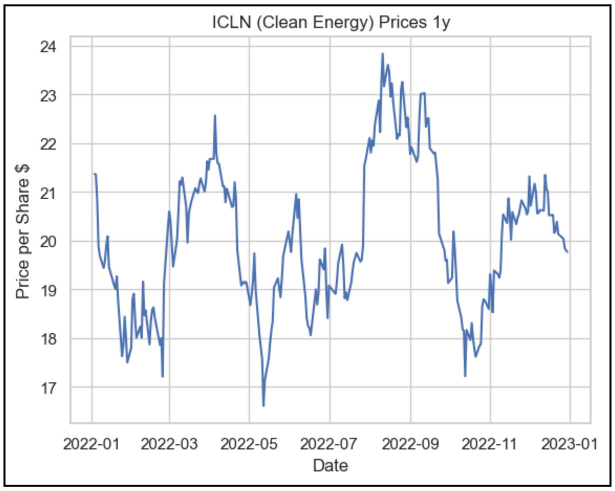
The first result we will discuss is the effect of a worldwide pandemic on energy stocks. In the left chart below, we can see that the COVID caused a severe decline in the price of oil, that still has not quite recovered. Based on this data, we would recommend to an audience of investors to sell oil stocks as soon as any new global pandemics are announced. On the bottom right, we can see that clean energy was also originally negatively affected after COVID, but had a quick and steep recovery, far surpassing its previous value. This could imply to investors that lockdowns might increase the value of clean energy stocks, so if restrictions resurface then invest in clean energy. We found that over this 5 year time frame, USO and ICLN had a statistically significant correlation of -0.65 (p-value = 0).



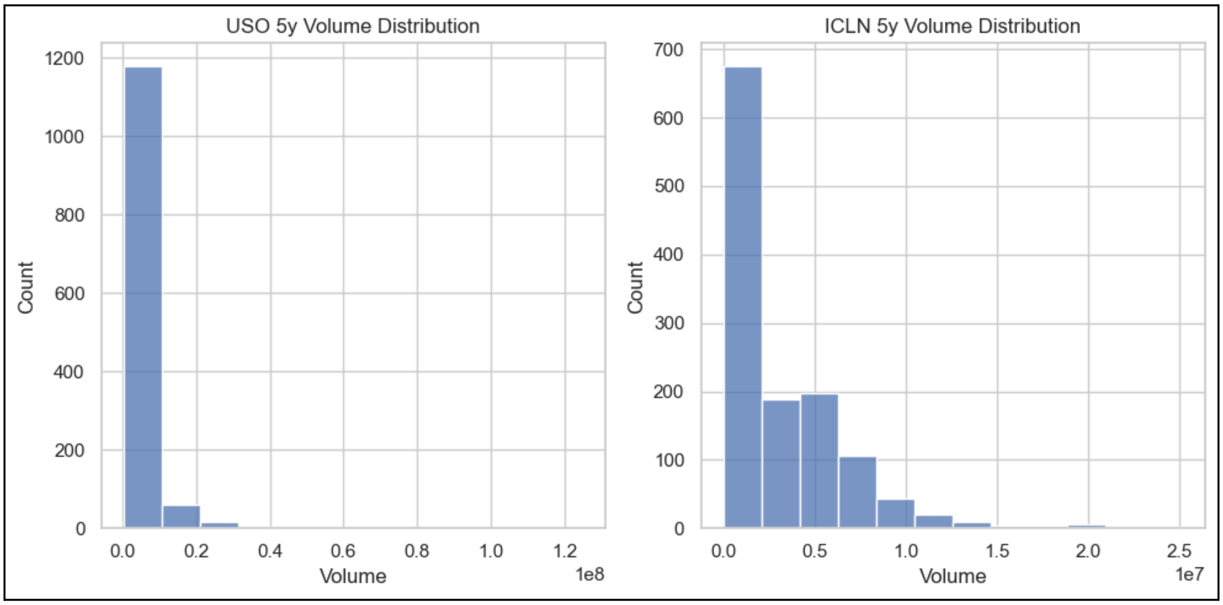


We then look at the same ETFs on a timeframe of only 2022. The clean energy ETF is seemingly random in trend, while the oil ETF steadily increases til about June, then steadily decreases. With this time period, we got an insignificant correlation between the two ETFs. A large geopolitical event took place in February of 2022— the Russia-Ukraine war. This would likely affect almost all stocks, and could explain why they seemingly became uncorrelated over this year. We would advise investors to be cautious when trading during large wars, as using techniques such as basing trades off of previously-proven correlations with other stocks can fail to continue working.

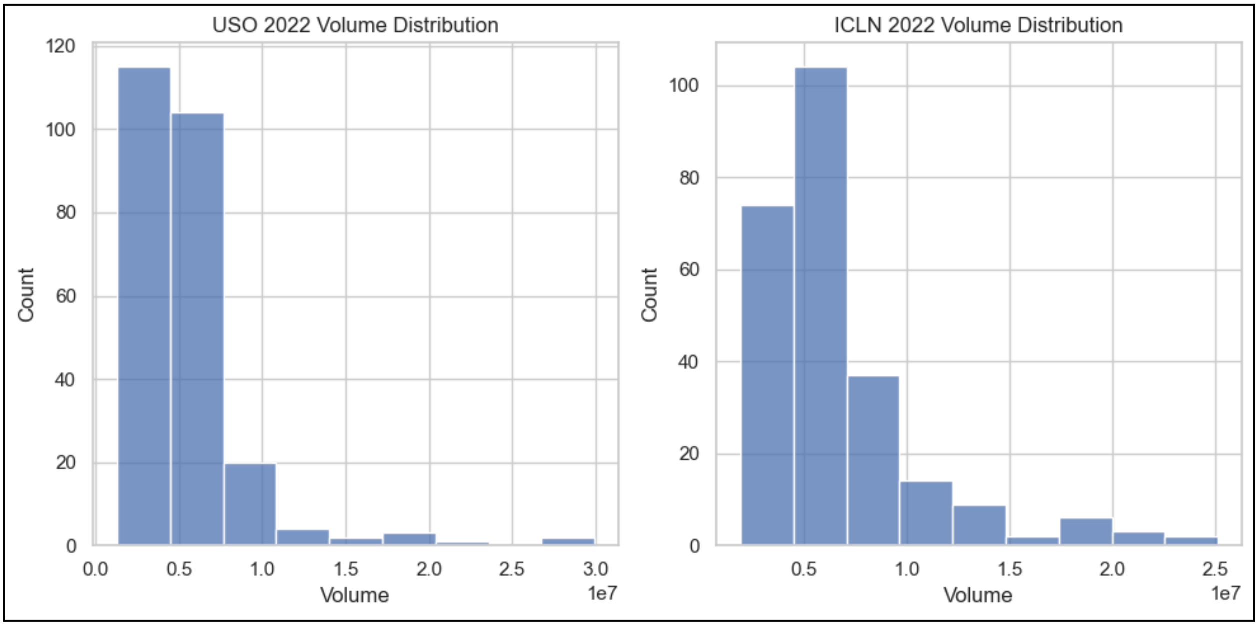




We also analyzed the trading volumes of the USO and ICLN ETFs, again over a five year and one year span. The distributions of them can be seen below. In the five year time span, the mean trading volume for USO was 5,266,591, while the ICLN mean trading volume was 3,010,962. These means were proven to be significantly different using a t-test.



Interestingly, the story flips when we look at their trading volumes in 2022 exclusively. USO had a mean volume of 5,418,460 while ICLN had a mean volume of 6,667,941. These means were again statistically significant. The large increase in mean volume of ICLN trades could be due to increasing popularity of clean energy, demand for energy independence from other countries, and other factors including the war in Ukraine. We would encourage investors to pay close attention to clean energy as it seems to be growing, while oil has remained relatively stagnant.



Do standard deviation by pct and add sections? Or nah ignore it

Technical

The yahoo finance api provides relatively clean data, that for the most part the data did not require extensive preprocessing. Data extraction was straightforward, in that the api provided clean ready to use data, all we needed to do was specify the stock and the range of data we were interested in using.

Yet again the analysis for this project was fairly simple. Basic data aggregation and averaging techniques were used. These were a good way to get a sense of the shape of the data, and to find interesting areas to investigate. Following this, the project expanded on the graphing techniques learned in class. To find the correlation between two variables, the Pearson Correlation Coefficients were calculated, along with the p values, to ensure that the Coefficients were statistically significant. This was particularly helpful for the various scatter plots generated by this group. They made trends more apparent, and clarified if the trends were more than a coincidence.

We did not face many challenges when analyzing the data, although initially we looked at single stock and had a hard time finding correlations we shifted to observing ETFs, and were then able to make some fairly intuitive estimates that turned out to be supported by our findings.