Correlating Exchange Traded Funds with Underlying Commodity Futures Contracts

Varun Varanasi and Iris Yang

May 2022

Contents

1 Abstract

The goal of this project is to study the relationship between commodity futures contracts and sector-specific ETFs (and to specifically examine whether commodity futures contracts can provide predictive power into the ETF prices of related sectors). The procedure for doing so involved running a LASSO regression of historical commodity data sets against sector-specific ETFs (Energy, Home-Construction, Industrials, and Technology). By selecting LASSO regression thresholds that kept relatively high null-deviance while minimizing the number of predictors, we identified the significant commodity predictors for each of our ETFs. Specifically, we noticed that Lumber, Palladium, and Platinum were consistently present in each of our models. Due to the distinctiveness of our ETF sectors, we hypothesize that their presence does not support a causality relationship between these three commodities and our ETFs in question. Instead, we believe that their presence is a statistical artifact of our small sample size. Further work would involve rerunning this project with additional commodity and ETF data sets in order to see if the observed trends are specific to our data sets or indicative of larger phenomena.

2 Introduction

The objective of this project is to understand the relationship between exchange-traded funds (ETFs) and the prices of futures contracts for related commodities. Exchange-traded funds are investment securities that function similarly to stocks. Typically, these entities will track particular sectors, commodities, or other assets; however, unlike mutual funds, these securities can be traded on the stock exchange and throughout standard trading hours. Since ETFs are bundles of investable assets, they are often used as portfolio diversification tools. Some of the more common ETFs include the S&P500 and ITOT (total stock market ETFs); however, there exist sector-specific ETFs such as MSOS (US Cannabis industry) and the ITB (US Home Construction) as well.

On the other hand, a futures contract is a contract to buy/sell a commodity at a future date for an agreed-upon price. These often fluctuate with the supply and demand of each commodity. As such, futures contracts are often used by investors speculating on the direction of the underlying asset. Futures contract prices are determined based on current price, risk-free rate of return, time to maturity, storage costs, and dividends, among other financial metrics and values.

The goal of this project is to understand how sector-based ETFs are correlated with commodity futures contracts. We hypothesize that the price of sector-specific ETFs is related to the futures contracts of related commodities. For example, we hypothesize that the price of a coffee shop ETF will be correlated to that of the futures of coffee beans. While this relationship is intuitive, it is likely complicated by multiple layers of financial complexity and causality. The goal of this project is to unpack these layers and see if we can discover a relationship between sector ETFs and associated commodity futures.

3 Methodology

This project will focus on 24 commodity contracts and 4 ETFs.

ETF Data Summary	
Commodities	Brent Crude Oil, Cocoa, Coffee, Copper, Corn, Cotton, Crude Oil, Feeder Cattle, Gold, Heating Oil, Lean Hogs, Live Cattle, Lumber, Natural Gas, Oat, Palladium, Platinum, Gasoline, Silver, Soybean Meal, Soybean Oil, Soybean, Sugar, and Wheat
ETF	Energy, Home Construction, Industrials, Technology

The project can be broken down into three phases: 1) Exploratory Analysis of Commodity Futures Contract Data, 2) Exploratory Analysis of ETF data, and 3) Regression Analysis. The first two phases are concerned with an individual analysis of each data set, while the third and final phase is focused on discovering relationships between our financial securities.

3.1 Commodity Exploratory Analysis

The commodity data is downloaded from a Kaggle collection with a separate .csv for each of the 24 commodities; this data is scraped from Yahoo Finance (for example, historical prices of Brent Crude Oil can be found here. An important detail to note is that commodity pricing is determined in the global market exchange and so they reflect global trends. Since each data set was scraped from Yahoo Finance, they are all structured with the following variables:

Variable	Description
Date	Date the price was measured (YYYY-MM-DD)
Open	The opening price of a single share on the given date
High	The high price of a single share on the given date
Low	The low price of a single share on the given date
Close	The closing price of a single share on the given date
Volume	The number of shares traded on the given date

**all prices in USD*

The original data only contained weekdays, and contained the word "null" for each of the pricing and volume columns for dates where the stock market was closed (i.e. holidays such as Thanksgiving and Christmas). Since the data simply omitted the weekends, we decided to omit these rows as well. In the end, our data set included commodity price information from our latest initial observation, August 30th, 2007, to June 6th, 2021. The original data also included an `Adjusted Close` column in addition to the `Close` column. According to the Yahoo finance website, `Close` price is adjusted for splits, and `Adjusted Close` price is adjusted for splits and dividend and/or capital gain distributions. Upon examination of the data, we discovered that there was no difference in these two columns for any of the dates for any commodity, so we decided to remove the `Adjusted Close` column for simplicity.

Commodity Data Summary		
Commodity	Max Value	Min Value
Brent Crude Oil	$144.49	$19.33
Cocoa	$3374.00	$1780.00
Coffee	$304.90	$86.65
Copper	$4.78	$1.25
Corn	$831.25	$293.50
Cotton	$215.15	$39.14
Crude Oil	$145.18	-$37.63
Feeder Cattle	$242.33	$86.55
Gold	$2015.50	$665.00
Heating Oil	$4.08	$0.61
Lean Hogs	$133.38	$37.33
Live Cattle	$171.00	$79.10
Lumber	$1686.00	$138.10
Natural Gas	$135.51	$1.48
Oat	$557.75	$158.50
Palladium	$2985.40	$162.1
Platinum	$2251.10	$595.90
RBOB Gasoline	$3.56	$0.41
Silver	$48.58	$8.79
Soybean Meal	$548.10	$239.80
Soybean Oil	$72.08	$24.99
Soybean	$1771.00	$783.50
Sugar	$35.31	$9.21
Wheat	$1337.00	$362.00

Once we collected and cleaned our data, our first step was to plot our time series (See page 7). From our initial observations, it is clear that the commodities have a wide range of historical behaviors. Certain commodities like Palladium and Lumber demonstrate a strong upwards trend while other commodities like Wheat and Natural Gas tend downwards. One detail to note is that in April 2020, the Crude Oil futures appear to drop to the negatives. While this may initially seem like an anomaly, upon further research this event actually did occur due to the crash in demand for oil at the beginning of the COVID-19 pandemic. Another interesting detail to note is that the commodities that are associated with crops (Coffee, Oat, Corn, Cotton, and the Soybean commodities) all seem to have reached their peak in the 2011-2013 time range.

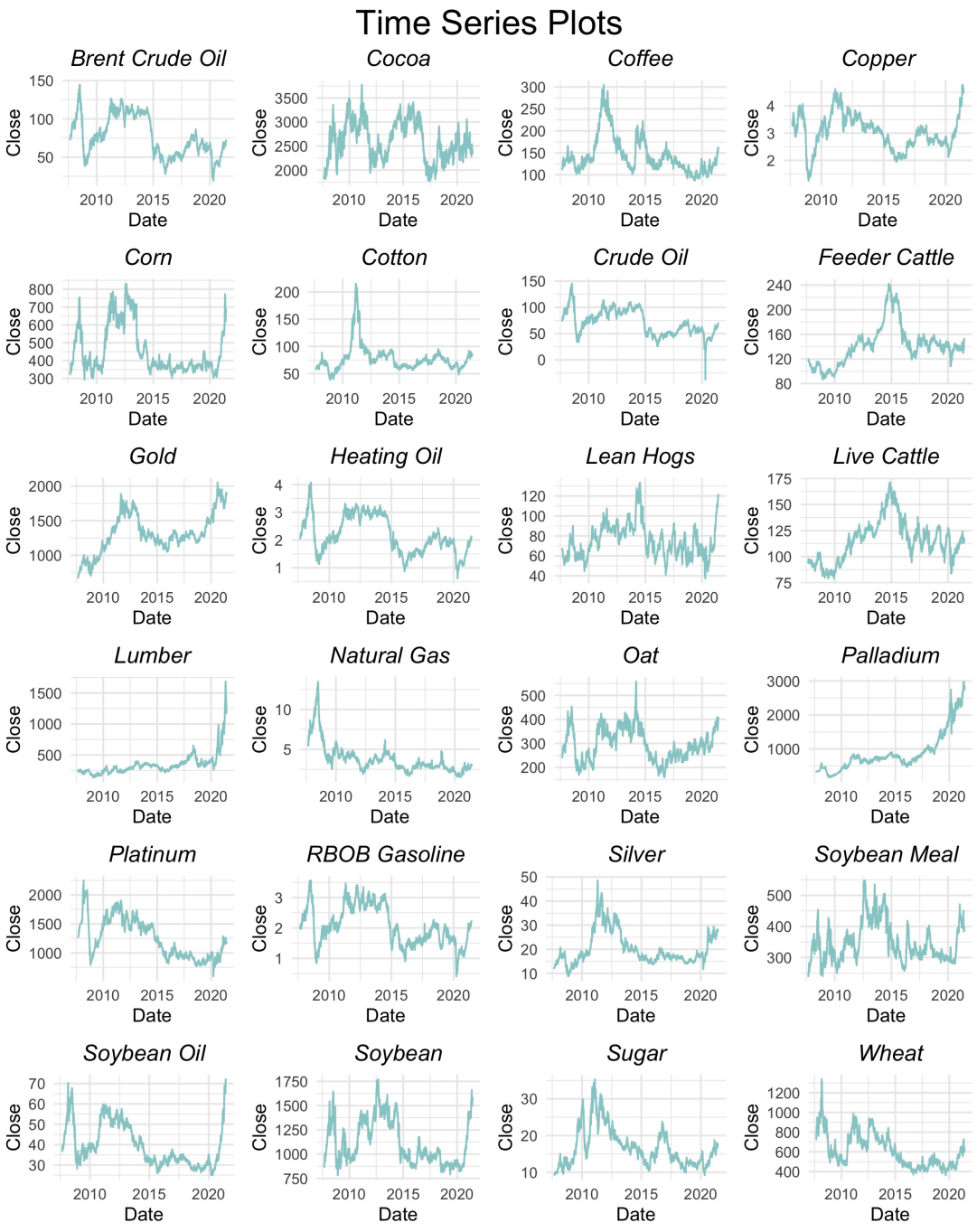

Figure 1: Commodities Time Series

Another metric worth evaluating is the distribution of our commodity prices. Consider a violin plot of the commodity prices:

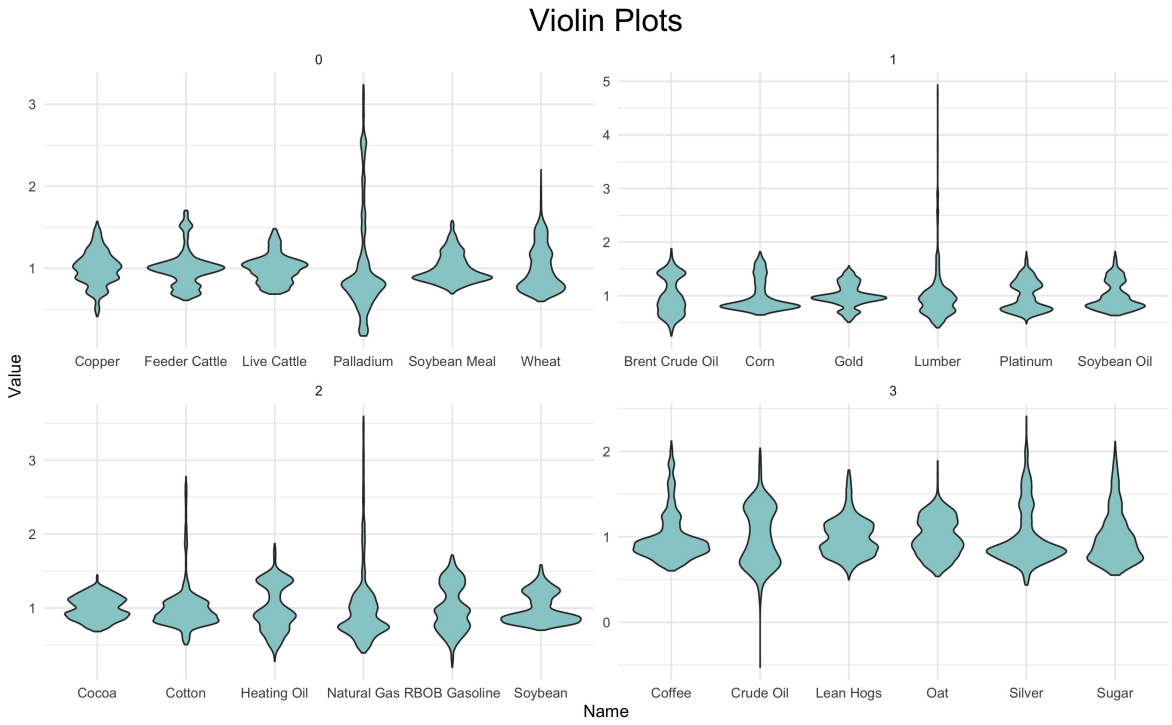

Figure 2: Commodities Violin Plot

Notice that for the most part our commodity prices seem to be tightly distributed; however, Palladium, Lumber, Cotton, Natural Gas, and Crude Oil stand against this pattern. These commodities seem to exhibit extreme values in relation to the remainder of the commodities. If we recall the time series plots on the previous page, both Palladium and Lumber exhibited strong growth in the last few years; this likely accounts for the long upward tails in their violin plots. Similarly, the negative values discussed in the Crude Oil time series likely explain the downwards tail present in the Crude Oil violin plot. While these are interesting observations, as of now, they do not indicate that we need to take any special considerations in our analysis.

3.1.1 Volatility

Volatility is a common financial marker used to indicate the change in an underlying asset over a given period of time. From a mathematical standpoint, volatility is a measure of the standard deviation of the data set. In layman's terms, volatility is simply a measure of the expected deviation from the average value. As such, volatility is often used in risk analysis to understand the stability of a given security or index.

With its long history in financial modeling, we decided to add the volatility of our commodities as an additional predictor in our analysis. Specifically, we created an additional column containing the volatility of the given metric over the past 30 days (representing approximate monthly volatility) using the formula below:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{30}}$$

Volatility encodes similar information to the underlying price, but is directionally invariant. For that reason, while volatility may help improve our model, it may also introduce multicollinearity. Therefore, it is prudent to consider whether the calculated volatility data encodes similar information to the original price. Our first pass attempt at characterizing this phenomena was to plot the volatility data alongside the underlying commodity price (See page 10).

Closer inspection of these graphs reveals to us that the variance plot seems to exaggerate changes in the underlying plot and encode sudden increases and decreases as spikes within the data. Since the data seems more or less comparable, we decided to perform our regression with volatility and price in separate models to avoid any potential multicollinearity issues.

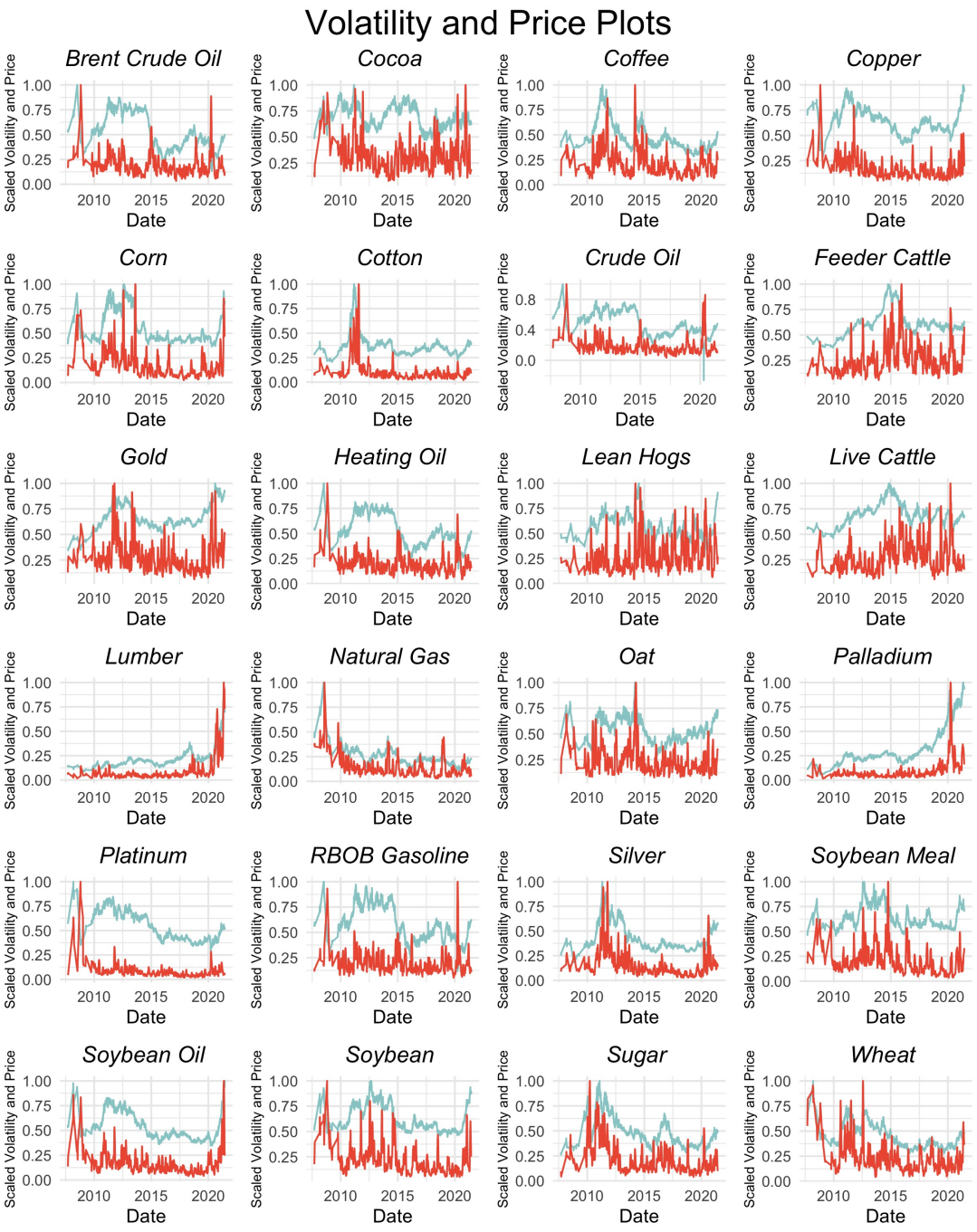

Figure 3: Commodities (Blue) and Volatility (Red) Time Series

3.2 ETF Exploratory Analysis

In our brief exploration of exchange-traded funds we decided to focus on four sectors: Energy, Home Construction, Industrials, and Technology. In the interest of consistency in our data, we decided to source all of our ETFs from iShares by Blackrock. Specifically, we used the iShares-US-Energy-ETF (IYE), iShares-US-Home-Construction-ETF (ITB), iShares-US-Industrials (IYJ), and the iShares-US-Technology (IYW) funds. The data sets contained historical data until April 29th, 2022 and the associated Net Asset Value per share for each date. While not exactly the price of a share, the NAV per Share works in an analogous fashion by the calculating the ratio between the total asset value and the number of outstanding shares (henceforth referred to as price). A summary of our data can be found below:

ETF Data Summary			
FUND	Start Date	Max Value	Min Value
Energy	June 12th, 2000	$57.72	$11.68
Home Construction	May 1st, 2006	$83.05	$11.68
Industrials	June 12th, 2000	$115.06	$6.49
Technology	May 15th, 2000	$117.17	$6.27

Once we collected and cleaned our data, our first step was to plot our time series.

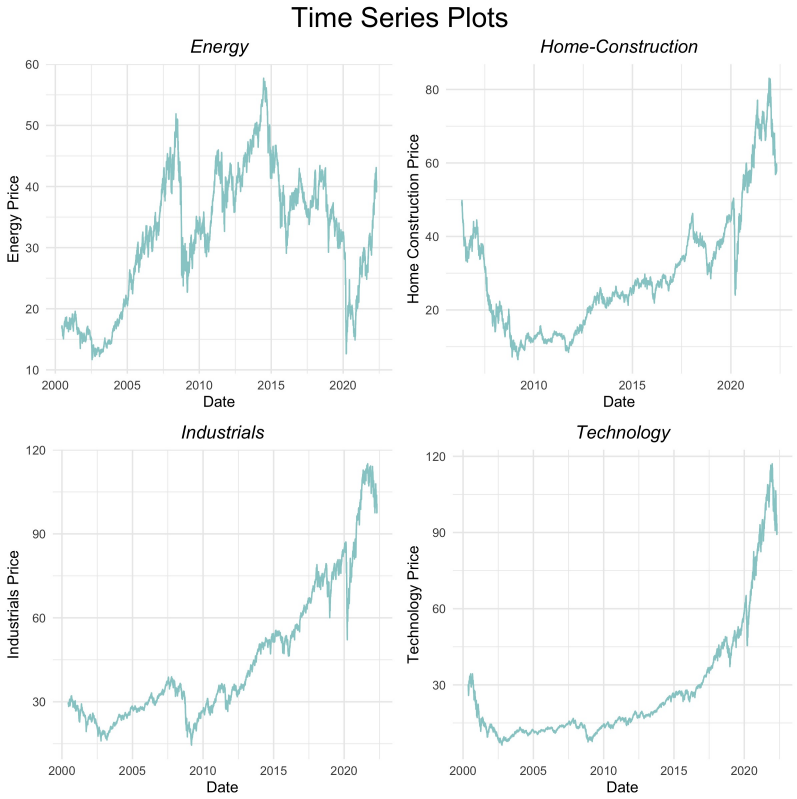

Figure 4: ETF Time Series

A quick look at the time series plots shows us that most of the four ETFs have been on a steady upward trend. The only exception is the Energy ETF, which has had a more inconsistent trajectory over the last twenty years. In each data set we can see dips associated with 2008 financial crisis and the 2020 COVID-19 pandemic. We can also see that the Home Construction ETF took the longest to recover from the 2008 housing crisis. Furthermore, all four ETFs seem to have recovered from the most recent drop. The Technology ETF appears to be the most resilient to downward trends, with only small drops associated with each event. Other historical events also appear to be encoded within our data as seen by the peak in the

Technology ETF in the early 2000s which likely corresponds to the dot.com bubble. This quick analysis demonstrates to us the correlation between ETF prices and external/historical factors.

An exploration into the distribution of our ETF values (via violin plots) can provide us a better understanding of our data and help us contextualize our models in the next phase of our analysis.

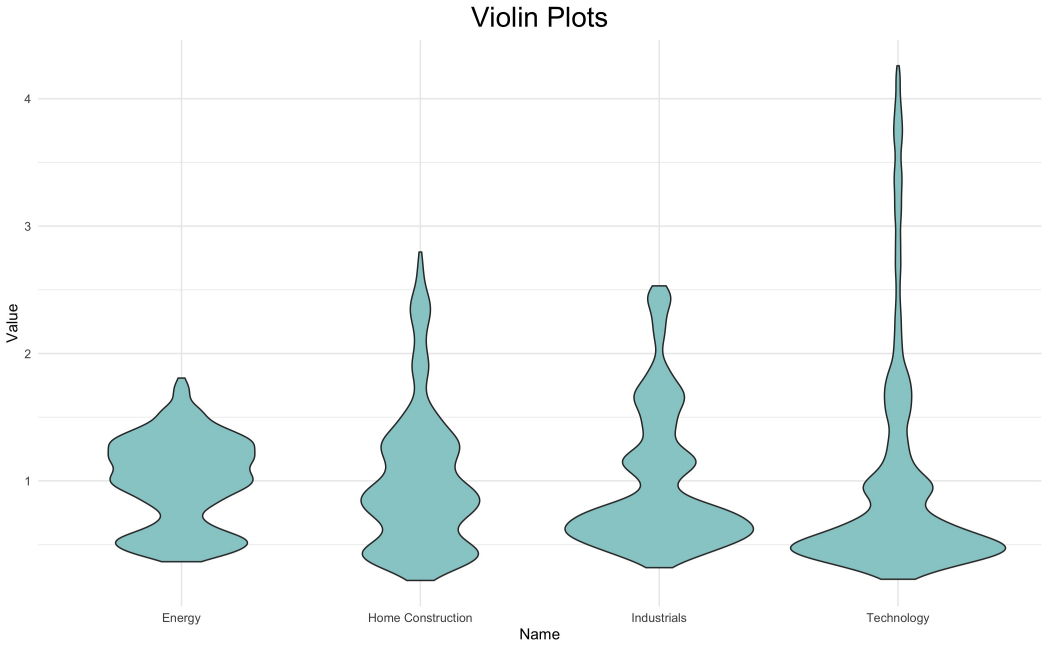

Figure 5: ETF Violin Plots

All four ETFs seem to have mostly similar distributions, with the Energy ETF having the smallest normalized range and the Technology ETF differentiated by its long upwards tail. This tail is likely a result of the Technology ETFs strong growth in recent years. Another point of interest is the almost ubiquitous large bulge at the bottom of each ETF violin plot, with the Technology violin plot having the widest bulge at the bottom (corresponding to the Technology time series spending the most time at lower values of its range). As with the commodities data sets, this analysis only provides us better context on our data and does not impact our approach to regression.

3.3 Regression

With our cleaned and explored data sets, our next step was to test the initial hypothesis. Specifically, we wanted to develop a model that tells us which, if any, commodities are useful in predicting a given ETF. Since our goal was to remove all unnecessary predictors, a LASSO regression was the natural choice. Therefore, our experimental procedure was as follows:

1. Isolate ETF time series data
2. Select predictor data set (Volatility vs. Commodity price)
3. Conduct a LASSO regression of ETF vs. predictors
4. Plot LASSO coefficients vs. Mean-Squared-Error
5. Evaluate various models based on number of coefficients remaining and their corresponding deviance ratios (R-squared equivalent)

4 Results

We repeated the procedure above a total of eight times using volatility and raw price as predictors for each of the four commodities. The results of each experiment are as follows.

4.1 Energy ETF

4.1.1 Experiment 1: Energy ETF vs. Commodity Price

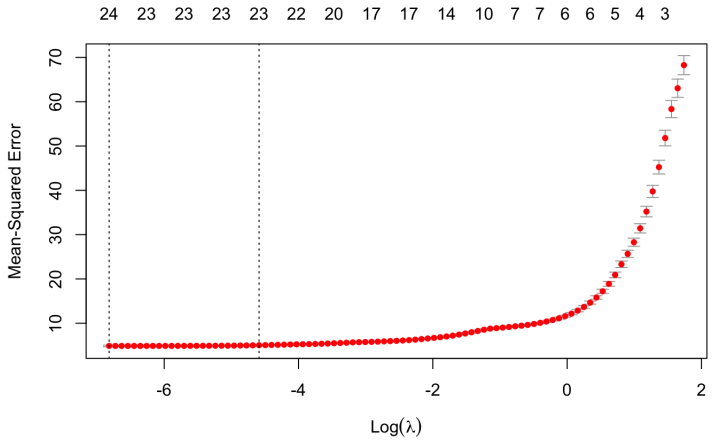

Figure 6: MSE vs. LASSO Coefficient for Energy ETF vs. Commodity Prices

Our model maintains consistent MSE values for LASSO coefficients in the range of log(λ) between -7 and -2.

Energy vs. Commodity Price Summary			
log(λ)	Null Deviance	Number of Predictors	5 Most Significant Predictors
Minimum MSE	0.930	24	1. Heating Oil 2. Gasoline 3. Copper 4. Soybean Oil 5. Sugar
1 SD	0.928	23	1. Heating Oil 2. Gasoline 3. Copper 4. Sugar 5. Soybean Oil
-2	0.904	15	1. Heating Oil 2. Gasoline 3. Copper 4. Sugar 5. Live Cattle

4.1.2 Experiment 2: Energy ETF vs. Commodity Volatility

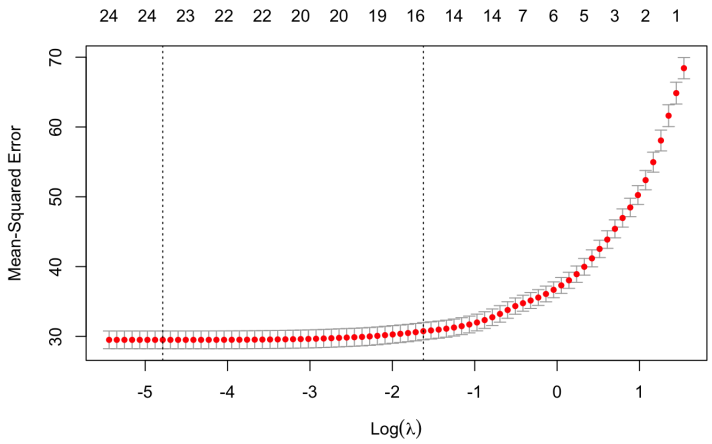

Figure 7: MSE vs. LASSO Coefficient for Energy ETF vs. Commodity Volatilities

Our model maintains consistent MSE values for LASSO coefficients in the range of $\log(\lambda)$ between -6 and -1.

Energy vs. Commodity Volatility Summary			
$\log(\lambda)$	Null Deviance	Number of Predictors	5 Most Significant Predictors
Minimum MSE	0.576	23	1. Heating Oil 2. Gasoline 3. Sugar 4. Natural Gas 5. Soybean Oil
1 SD	0.556	16	1. Sugar 2. Natural Gas 3. Gasoline 4. Crude Oil 5. Coffee
-1	0.540	14	1. Sugar 2. Natural Gas 3. Coffee 4. Soybean Meal 5. Crude Oil

4.2 Home Construction ETF

4.2.1 Experiment 3: Home Construction ETF vs. Commodity Price

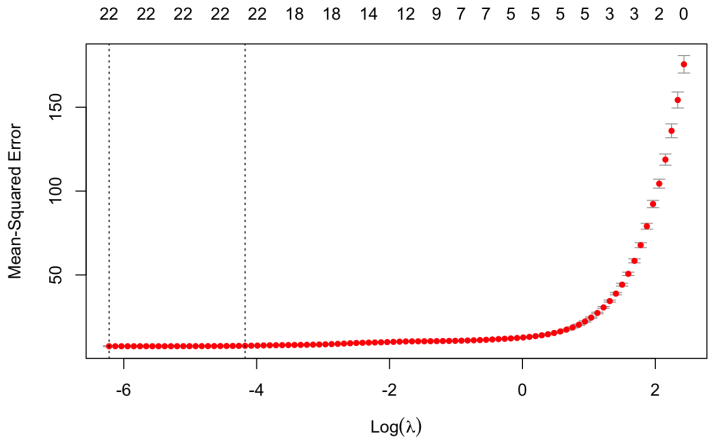

Figure 8: MSE vs. LASSO Coefficient for Home Construction ETF vs. Commodity Prices

Our model maintains consistent MSE values for LASSO coefficients in the range of $\log(\lambda)$ between -6.5 and 0.

Home Construction vs. Commodity Price Summary			
$\log(\lambda)$	Null Deviance	Number of Predictors	5 Most Significant Predictors
Minimum MSE	0.958	22	1. Gasoline 2. Heating Oil 3. Copper 4. Natural Gas 5. Brent Crude Oil
1 SD	0.957	22	1. Gasoline 2. Heating Oil 3. Copper 4. Natural Gas 5. Brent Crude Oil
0	0.928	5	1. Sugar 2. Lumber 3. Coffee 4. Platinum 5. Palladium

4.2.2 Experiment 4: Home Construction ETF vs. Commodity Volatility

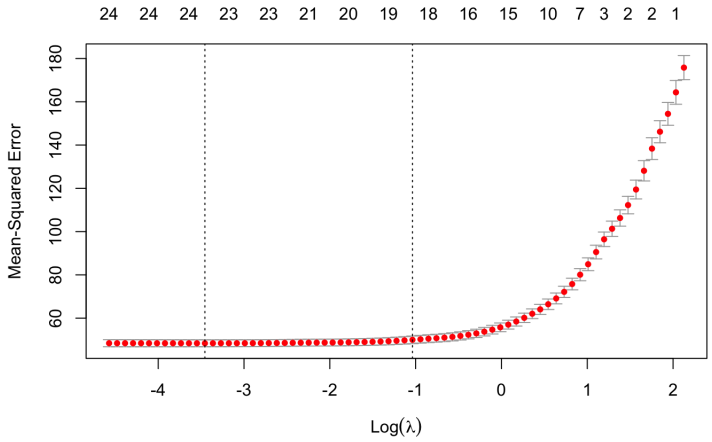

Figure 9: MSE vs. LASSO Coefficient for Home Construction ETF vs. Commodity Volatilities

Our model maintains consistent MSE values for LASSO coefficients in the range of $\log(\lambda)$ between -5 and 0.

Home Construction vs. Commodity Volatility Summary			
$\log(\lambda)$	Null Deviance	Number of Predictors	5 Most Significant Predictors
Minimum MSE	0.734	24	1. Heating Oil 2. Copper 3. Sugar 4. Soybean Oil 5. Gasoline
1 SD	0.723	19	1. Sugar 2. Silver 3. Natural Gas 4. Copper 5. Brent Crude Oil
0	0.686	16	1. Sugar 2. Silver 3. Natural Gas 4. Brent Crude Oil 5. Lumber

4.3 Industrials ETF

4.3.1 Experiment 5: Industrials ETF vs. Commodity Price

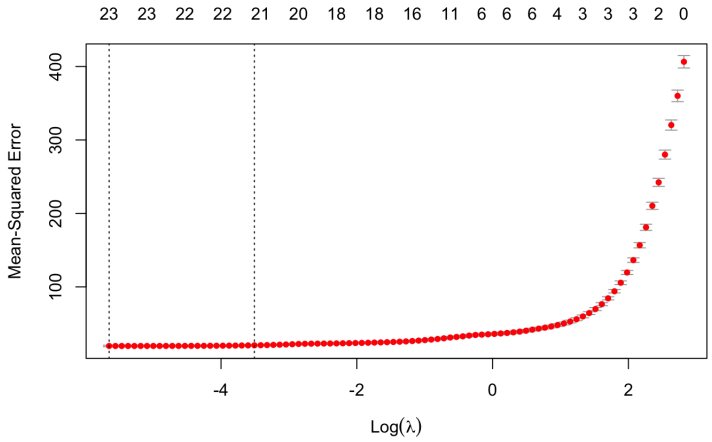

Figure 10: MSE vs. LASSO Coefficient for Industrial ETF vs. Commodity Prices

Our model maintains consistent MSE values for LASSO coefficients in the range of $\log(\lambda)$ between -6 and 0.

Industrials vs. Commodity Price Summary			
$\log(\lambda)$	Null Deviance	Number of Predictors	5 Most Significant Predictors
Minimum MSE	0.953	24	1. Heating Oil 2. Gasoline 3. Copper 4. Natural Gas 5. Brent Crude Oil
1 SD	0.951	21	1. Heating Oil 2. Gasoline 3. Copper 4. Natural Gas 5. Brent Crude Oil
0	0.913	6	1. Sugar 2. Live Cattle 3. Lumber 4. Platinum 5. Palladium

4.3.2 Experiment 6: Industrials ETF vs. Commodity Volatility

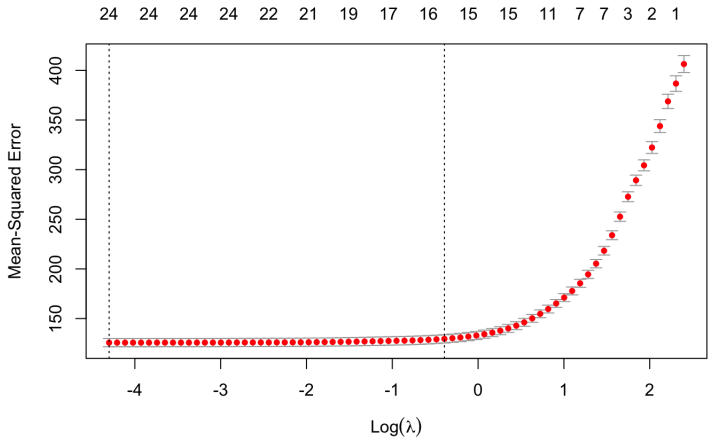

Figure 11: MSE vs. LASSO Coefficient for Industrials ETF vs. Commodity Volatilities

Our model maintains consistent MSE values for LASSO coefficients in the range of $\log(\lambda)$ between -5 and 0.

Industrials vs. Commodity Volatility Summary			
$\log(\lambda)$	Null Deviance	Number of Predictors	5 Most Significant Predictors
Minimum MSE	0.697	24	1. Heating Oil 2. Sugar 3. Copper 4. Natural Gas 5. Silver
1 SD	0.685	16	1. Sugar 2. Natural Gas 3. Silver 4. Palladium 5. Crude Oil
0	0.676	15	1. Sugar 2. Natural Gas 3. Silver 4. Crude Oil 5. Lean Hogs

4.4 Technology ETF

4.4.1 Experiment 7: Technology ETF vs. Commodity Price

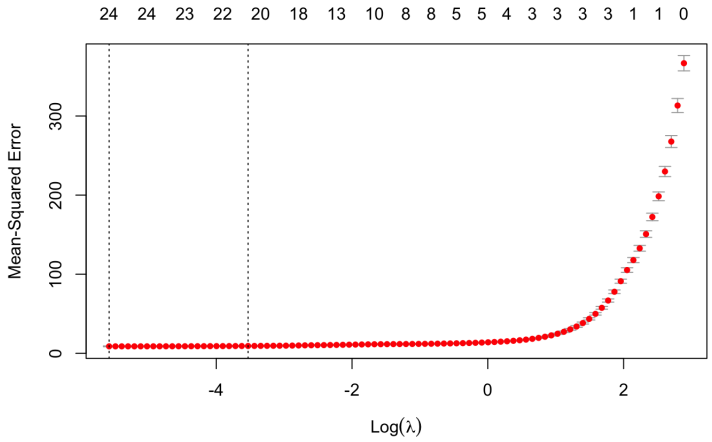

Figure 12: MSE vs. LASSO Coefficient for Technology ETF vs. Commodity Prices

Our model maintains consistent MSE values for LASSO coefficients in the range of $\log(\lambda)$ between -6 and 1.

Technology vs. Commodity Price Summary			
$\log(\lambda)$	Null Deviance	Number of Predictors	5 Most Significant Predictors
Minimum MSE	0.976	24	1. Heating Oil 2. Gasoline 3. Copper 4. Natural Gas 5. Soybean Oil
1 SD	0.975	20	1. Heating Oil 2. Copper 3. Gasoline 4. Natural Gas 5. Sugar
1	0.935	3	1. Palladium 2. Lumber 3. Platinum

4.4.2 Experiment 8: Technology ETF vs. Commodity Volatility

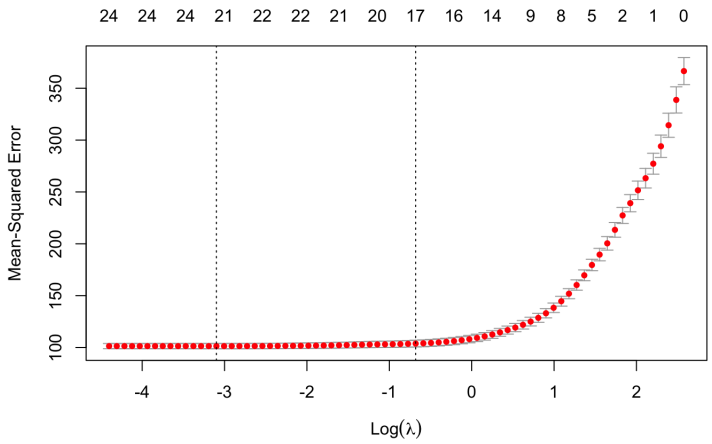

Figure 13: MSE vs. LASSO Coefficient for Technology ETF vs. Commodity Volatilities

Our model maintains consistent MSE values for LASSO coefficients in the range of $\log(\lambda)$ between -5 and 0.

Technology vs. Commodity Volatility Summary			
$\log(\lambda)$	Null Deviance	Number of Predictors	5 Most Significant Predictors
Minimum MSE	0.733	21	1. Gasoline 2. Copper 3. Sugar 4. Natural Gas 5. Soybean Oil
1 SD	0.722	17	1. Gasoline 2. Sugar 3. Heating Oil 4. Silver 5. Natural Gas
0	0.709	15	1. Gasoline 2. Sugar 3. Heating Oil 4. Silver 5. Lumber

5 Analysis & Conclusion

5.1 Price vs. Volatility

A common theme in each of our regression models was that the commodity price model outperformed its volatility based counterpart. This observation is most apparent in the Energy ETF models, where the null deviation in the minimum MSE models for the price based model was 0.930 while the volatility based model was 0.576. This reduction in model efficacy is likely due to the lack of directionality in the volatility metric. As noted in our exploratory analysis of the ETF datasets, the Energy ETF appeared to exhibit the largest fluctuations and inflections of all 4 ETF data sets. As such, since the volatility data set did not encode the directional changes of the underlying commodities, it was unable to translate directional trends from the commodity data set into our predictive model. While all ETFs exhibited upwards and downwards trends, since the Energy ETF was the most volatile, this effect was most pronounced in the large null deviation difference. With this observation in mind, the remainder of our analysis was focused on the price-based models for each ETF.

5.2 Regression Coefficients

Since the main goal of this project was to elucidate the relationship between commodity futures contracts and sector-specific ETFs, an important note of consideration was the value of the regression coefficients in each our models. We displayed our coefficients for the minimum MSE models in a heat map to visualize trends in the data.

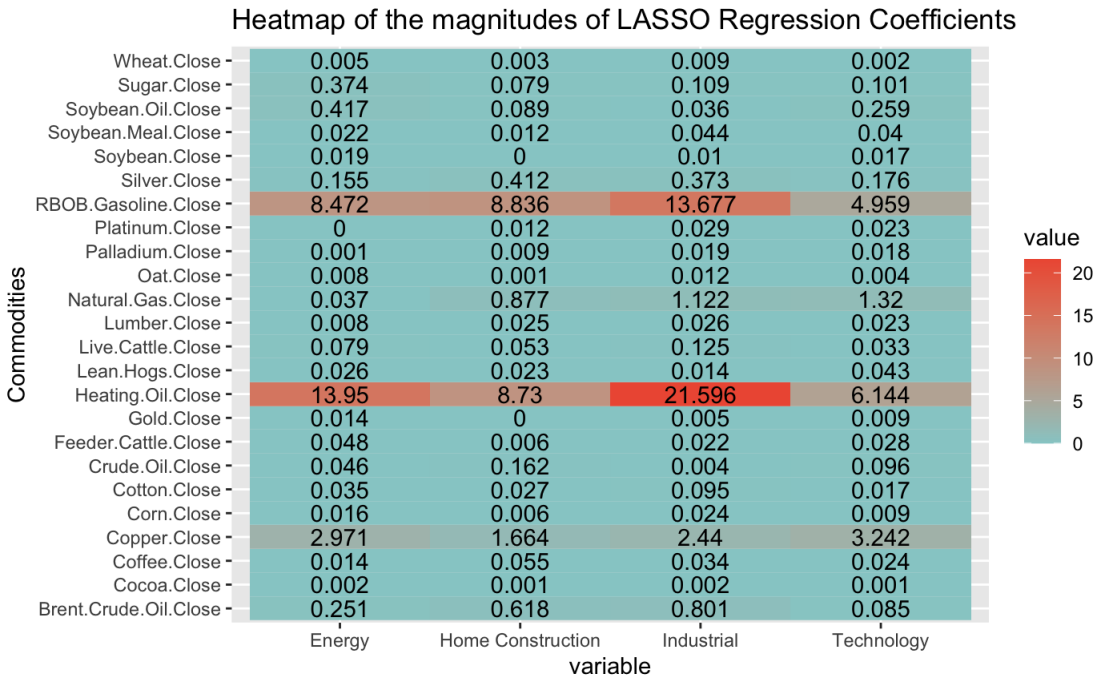

Figure 14: Heatmap of LASSO Regression Coefficients

A quick look at the regression coefficients shows that Gasoline and Heating Oil dominate all other regression coefficients in almost every model. Copper seems to come in third with a Natural Gas in a close 4th. These observations are also noted in the 5 most significant predictors column of our Results section (in which these 4 commodities showed up a total of 14 times across the 4 models).

Referring back to our earlier plots, we can see that these 4 commodities are among the commodities with the smallest prices. Therefore, it is likely that these points require large regression coefficients to contribute

to the model. To verify their importance to the model, we can increase our LASSO regression coefficient and consider models with fewer predictors (as seen in the tables in the Results section).

5.3 Minimal Predictor Models

Using the LASSO coefficient cutoffs explored in the results section we see that we can drastically reduce the number of predictors at minimal expense to our null deviation for each of the ETF models.

Strict LASSO Regression Coefficient Summary			
ETF	log(λ)	Null Deviance Change	Number of Predictors
Energy	-2	-0.26	15
Home Construction	0	-0.30	5
Industrials	0	-0.40	6
Technology	1	-0.41	3

**Null Deviance Change calculated in reference to the minimum MSE model*

Once again, we see that these models have a common set of commodities. Specifically, Palladium, Lumber, and Platinum are present in every model except the Energy regression. We offer two possible explanations for the common presence of Palladium, Lumber, and Platinum in each of minimal predictor models. First, we hypothesize that the presence of these parameters is a reflection of their capacity to model the limited sample size of ETFs we studied in this project. We could easily verify this by rerunning these procedures on a larger set of ETF values and seeing whether the trend continues. If so, we can make an argument that these three commodities reflect general market trends and their presence in sector-specific ETFs is a relic of that phenomena. Alternatively, if we fimd that these three commodities are specific to the ETFs we studied, we can support our claim that they are correlated with the ETFs of our study. Either way, to further evaluate the relationship between these commodities and our ETFs, we must run this analysis on a larger sample size.

The second explanation that we posit is that these three commodities were chosen because of their generally monotonic trends. As noted in our ETF exploratory analysis, the Industrial and Technology ETFs exhibit an almost steady positive growth. Similarly, the Home-Construction ETF, exhibits an initial drop and then a steady increase. Taking a look at the time series plots for the commodities in question, we notice similar trends where Platinum exhibits an early drop while Palladium and Lumber exhibit a steady increase. Due the geometry of the times series, we hypothesize that they can be used in a linear combination that effectively replicates the time series in question. If this is true, we cannot make any fundamental claims of causality or predictive power about these commodities on general ETFs. This idea is further supported by the lack of distinctiveness of our sector ETFs. Since they are from such different areas, it is likely that the presence of common commodities across the ETFs is indicative of statistical artifact rather than an actual relationship. Once again, this can only be verified by further study of a larger data set of ETFs and commodities.

6 Code

The files in the folder are arranged as follows:

File	Description
cleaning_commodity_data.Rmd	Exploring and cleaning commodity data
cleaning_ETF_data.Rmd	Clean ETF data
commodity_vis.Rmd	Creating time series plots and violin plots for commodity data
ETF_vis.Rmd	Creating time series plots and violin plots for ETF data
lasso_reg.Rmd	Running LASSO regression on commodity and ETF data
volatility_calculation.Rmd	Calculating volatility for commodity data

The original commodity data is in the "commodities" folder (with a separate .csv for each of the 24 commodities); the cleaned commodity data is in the "cleaned_commodities" folder; and the cleaned commodity data with dates earlier than the latest overall start date removed are in the "dates_cleaned_commodities" folder. The "ETF" folder contains a "Cleaned_Data" folder and a "Raw_Data" folder (each with a separate .csv for each ETF) in addition to a Total_ETF.csv (with all the ETF data merged into a single file for ease of access). The "plots" folder contains all the plots.