COMMODITY PRICE PREDICTION AND RESEARCH

BRANDON MORETZ

INTRODUCTION

The purpose of this artifact is to memorialize the findings from the exploratory data analysis performed on the commodities historical price data. There will be several revisions to this document as the analysis unfolds and relationships are extracted, distributions are parameterized, and trading strategies are formalized.

DATA PROCESSING AND MARKET STYLIZED FACTS

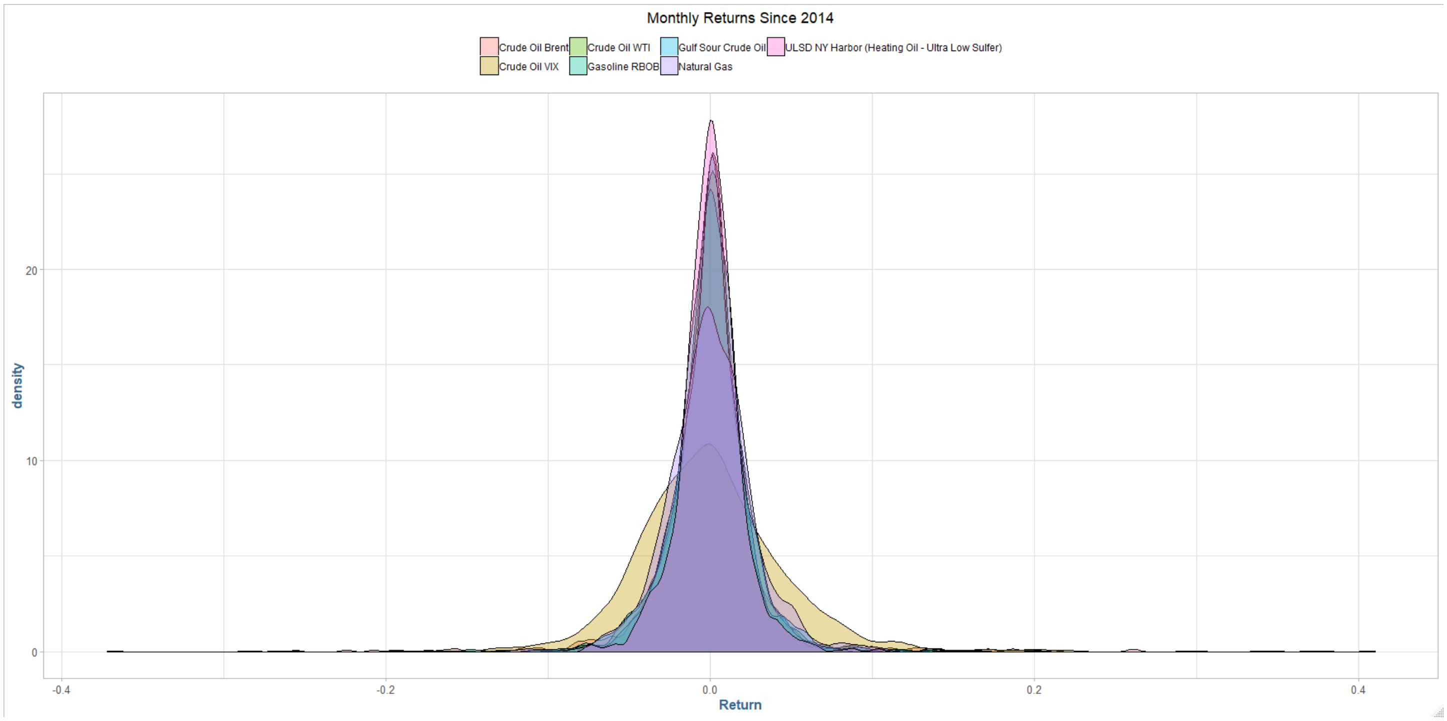
The dataset we are working with in this project is composed of price points on various energy commodities, specifically we have intraday open, high and low and the daily closing price of which we will be exploring in-depth for several commodities. Our initial commodities of interest include Crude (WTI and Brent), Heating Oil (Ultra low Sulfur), Gasoline and Natural Gas. The first step in financial data analysis is typically to take the pricing data and convert these to returns. This transformation serves multiple purposes; however, we are going to specifically callout the two main benefits of this approach.

The first benefit we get from using returns over prices is that this puts all the symbols on a level playing field by process of normalization: measuring all the variables in a comparable metric, thus enabling evaluation of analytical relationships amongst the variables despite originating from prices of unequal values. This is not only a nice mathematical property and convenience; it is a requirement for many of the statistical and mathematical models and procedures we are going to implore in later research.

The second benefit of this normalization process is a corollary from the first stylized fact of market prices, in that they follow an approximately log-normal distribution due to the prices having the mathematical property of being restricted to the domain of greater than or equal to zero. By converting our price series first into simple returns, and then taking the natural log of this series, we transform the data distribution from log-normal, to one that is approximately normal. We can prove this mathematically, however, for brevity we will instead rely here on the mathematical intuition derived from the Central Limit Theorem (CLT), which states that "the sum of a number of independent and identically distributed random variables (R_i) with finite variances will tend to a normal distribution as the number of variables grows." Here, our price series are assumed to be independent and identically distributed, our returns are logtransformed which have the property of time-dependent additivity ($R_{t+1} = R_t + R_{t-1}$), and thereby will be approximately normal. Later, our assumptions of normality will be revised and expanded to include maximum likelihood estimates for parametric versions of the normal distributions to further account for both tail risk and tail dependence, both of which are market stylized facts that will be explored in-depth.

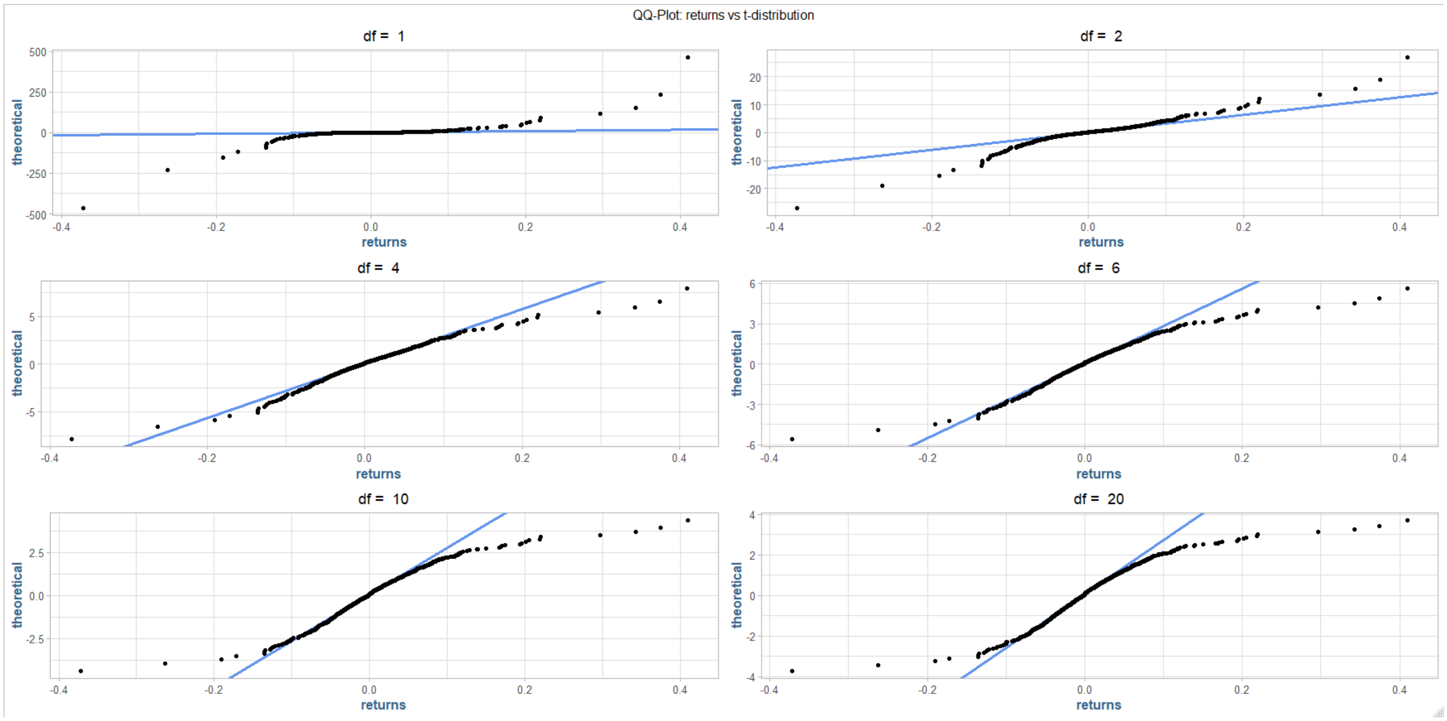
DISTRIBUTIONS & VARIANCE

Looking at the distribution of the returns for the commodities since 2014, we will notice some distinguishing characteristics that may help us in further pattern analysis:

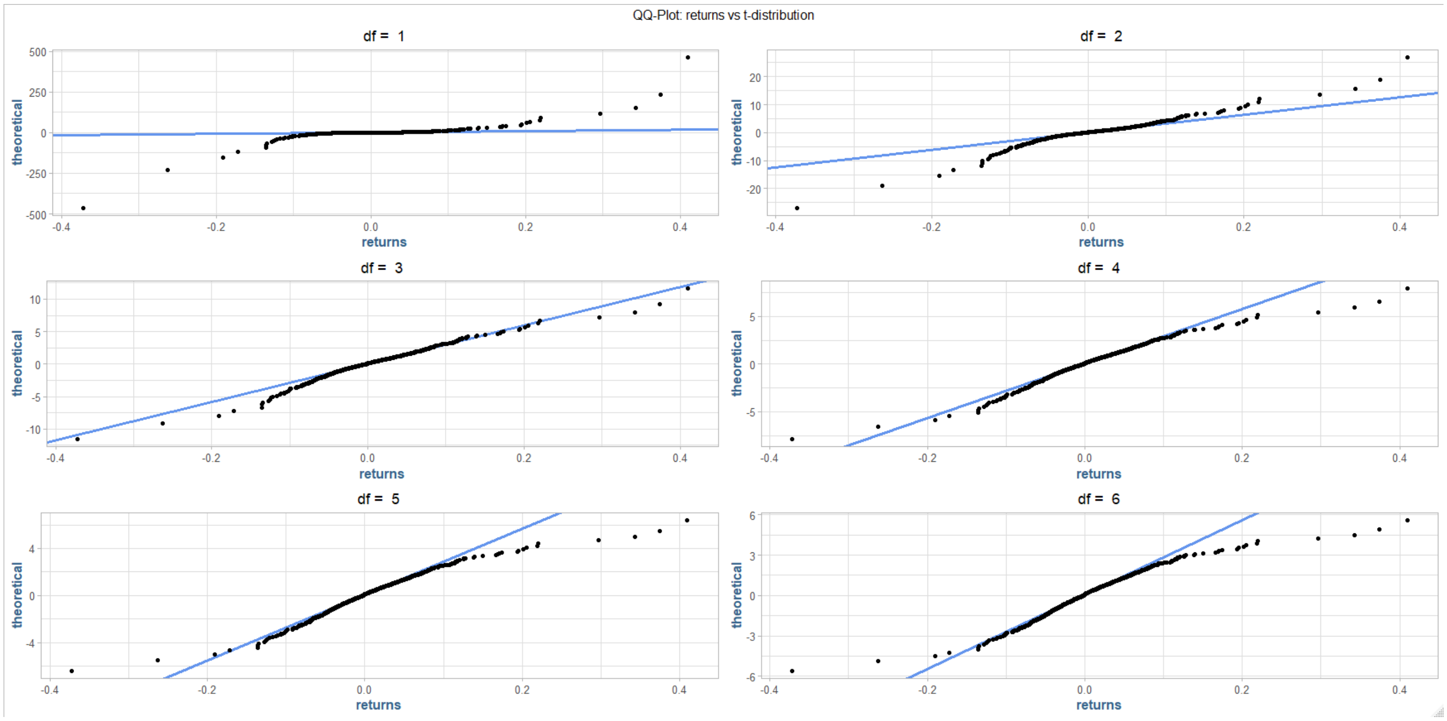


The most obvious fact that jumps out from the above plot is the spread/variance amongst the commodities. The volatility for these assets is obvious, however, the crude VIX is an outlier amongst outliers. It is clear we will need to use a Student's t for each of the assets, as use independent maximum likelihood estimates for the degrees of freedom.

A quick visual inspection of the Crude VIX vs some standard t-distributions will be useful to gather a quick estimate on the degrees of freedom we'll need to model these further:



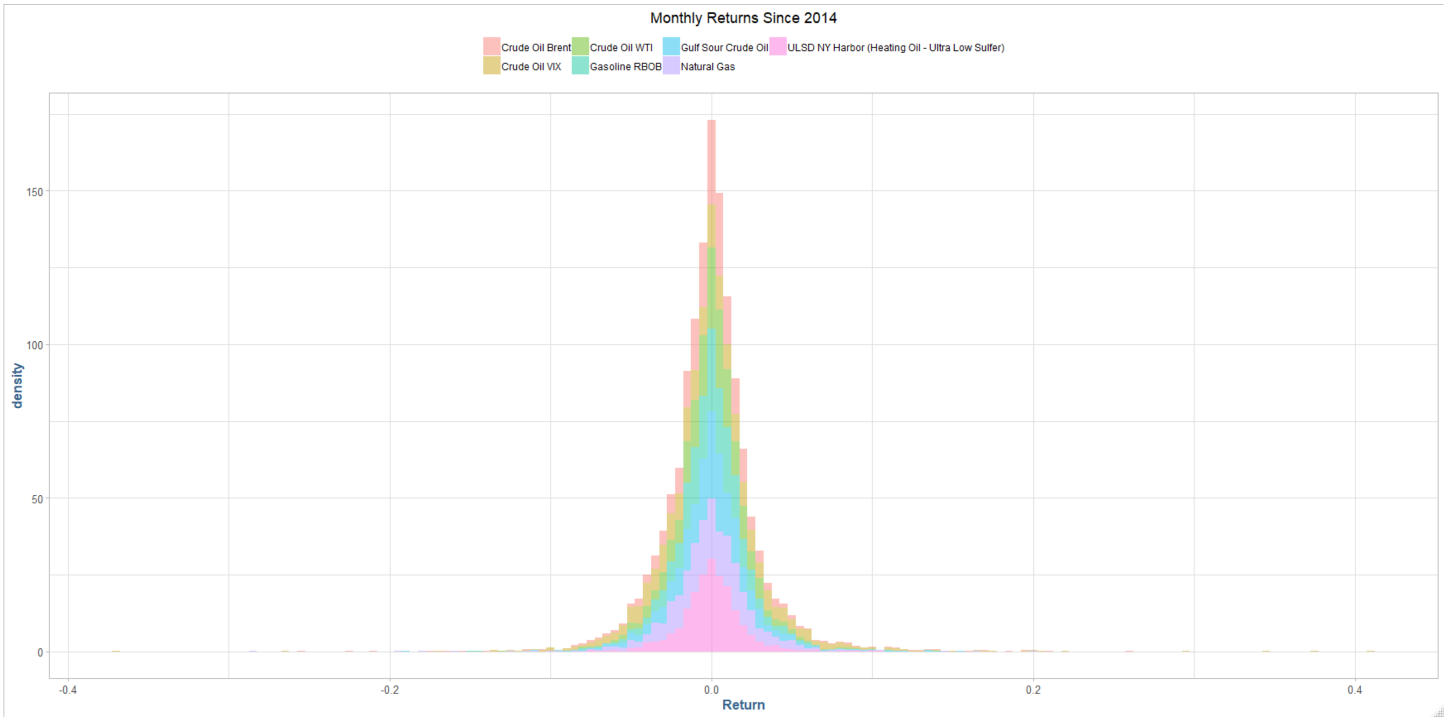
It appears that the Crude VIX will need a an extremely low degree of freedom parameter, as the variance is almost undefined (v <= 2 is “infinite” variance).



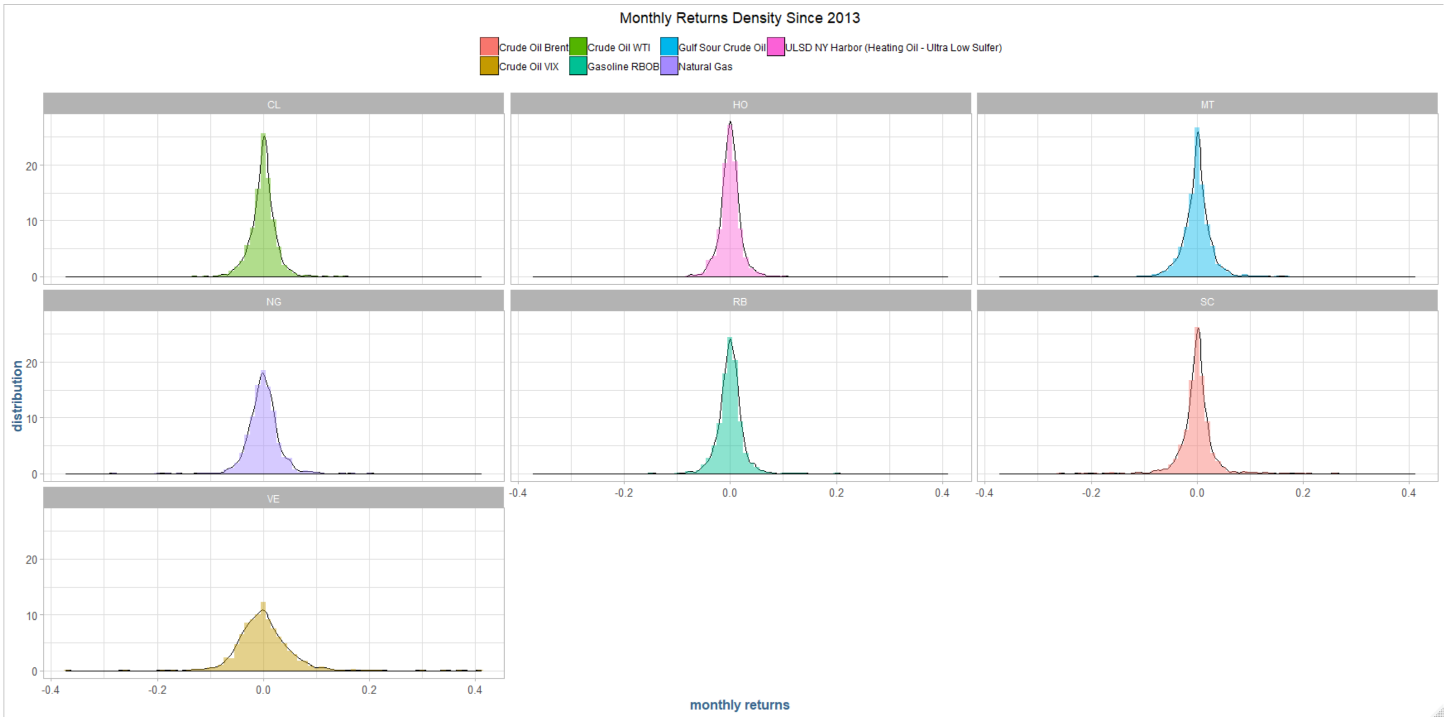
It looks like we can get close with around ~3 degrees of freedom. We will have to revisit these later when we perform the maximum likelihood estimation.

For reference, a QQ-plot fit to a random normal distribution can be found in the appendix.

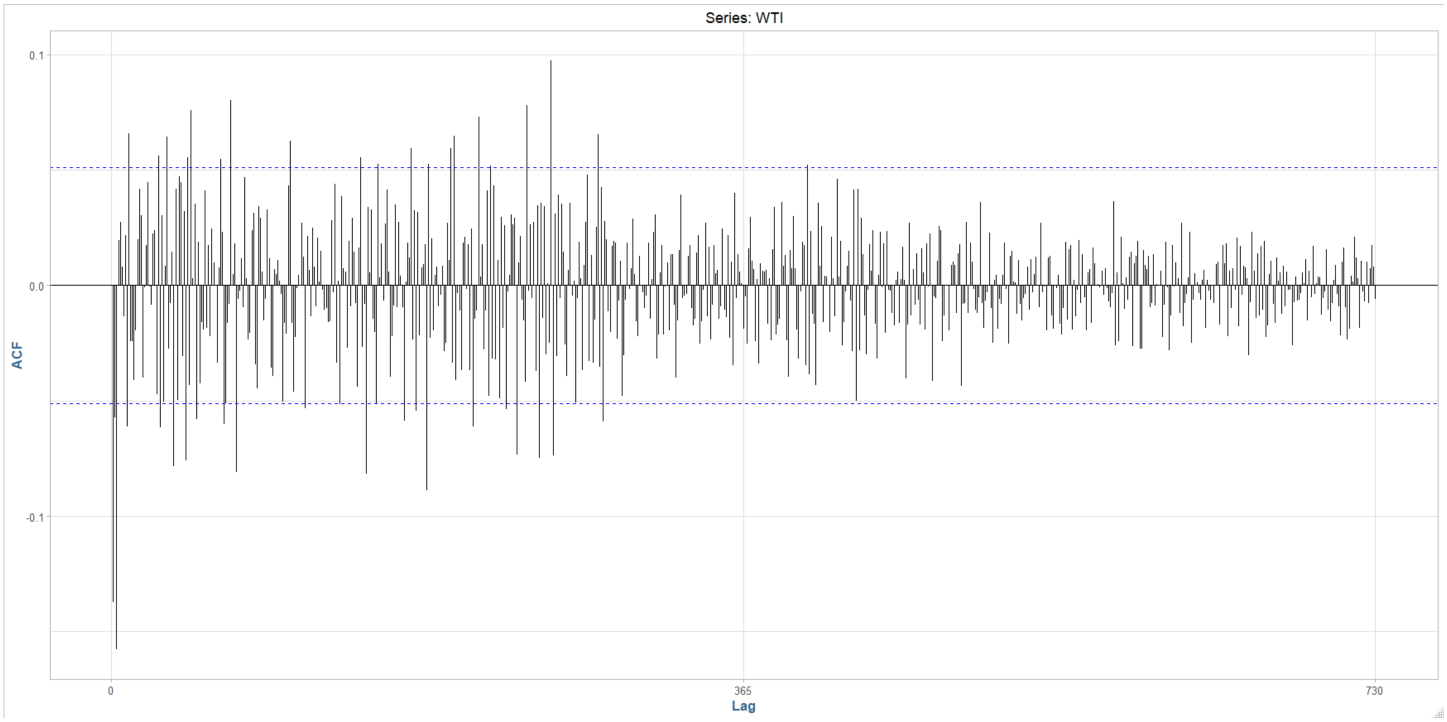
A similar look at the clustering behavior of the returns yield some additional interesting observations:

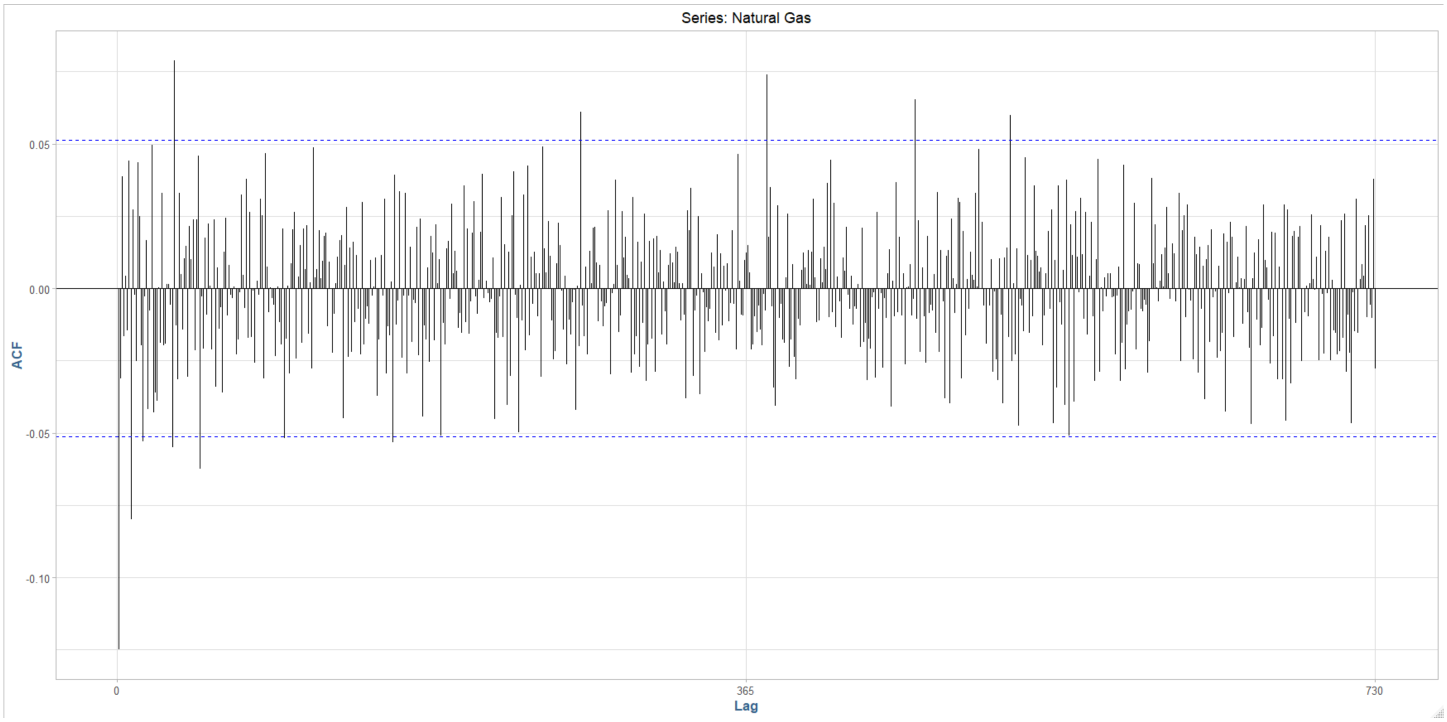


We can see the Brent and WTI appear to be relatively stable amongst the assets, having the most returns cluster towards the mean. We can also break it out by commodity, to get an individual sense of the spreads.

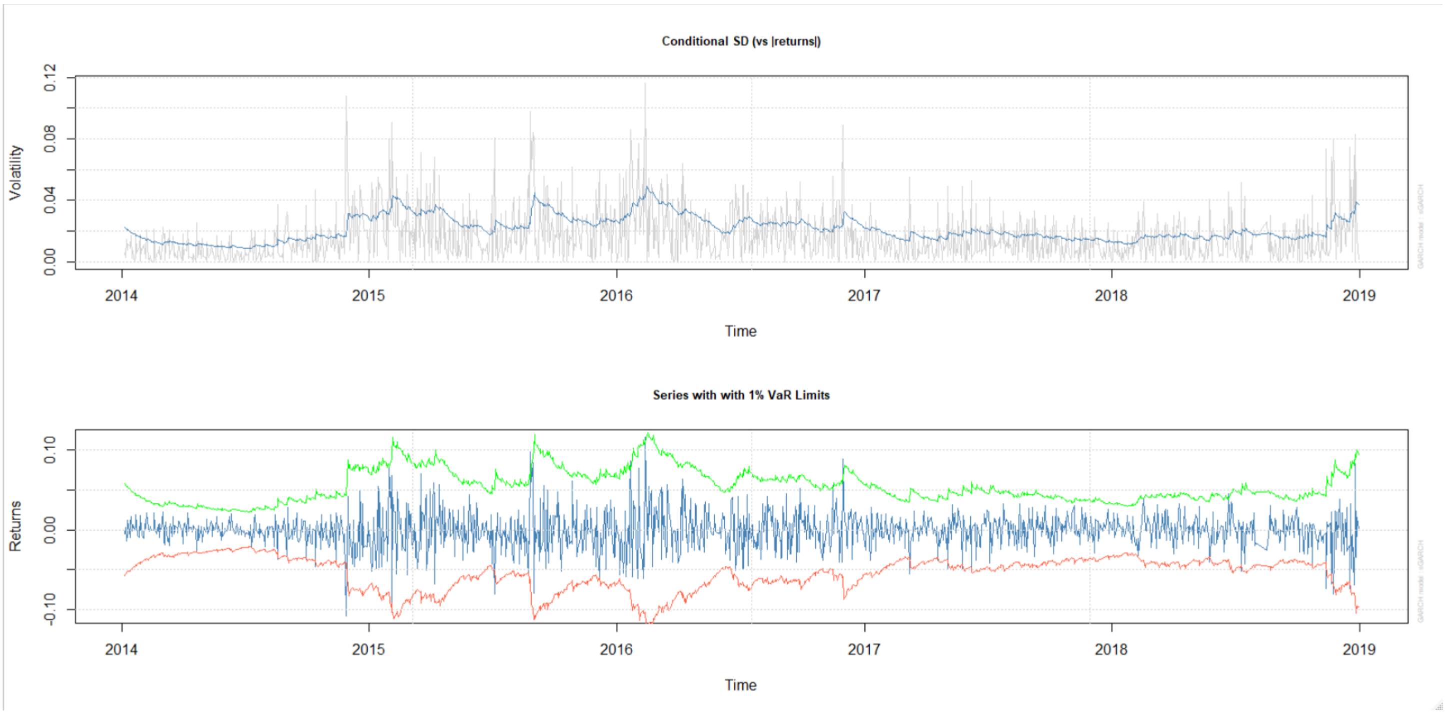


Individually, we will examine the time-series information in these instruments to look for any potential stationary process.

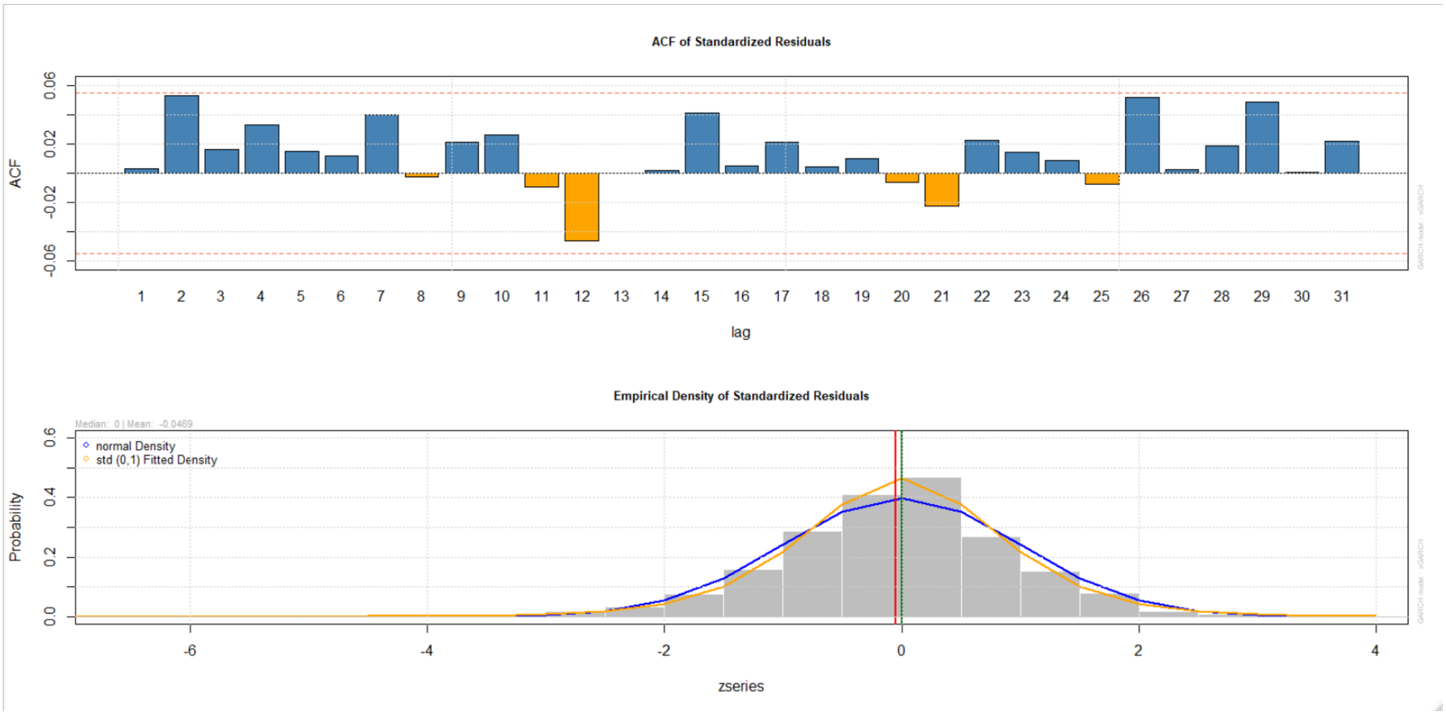




The volatility in these assets is obvious. We will attempt to capture these extreme movements in a GARCH process. Below is the historical volatility of WTI, with 2 SD conditionals & 1% value-at-risk lines superimposed on the training data, along with the empirical density of the model residuals:

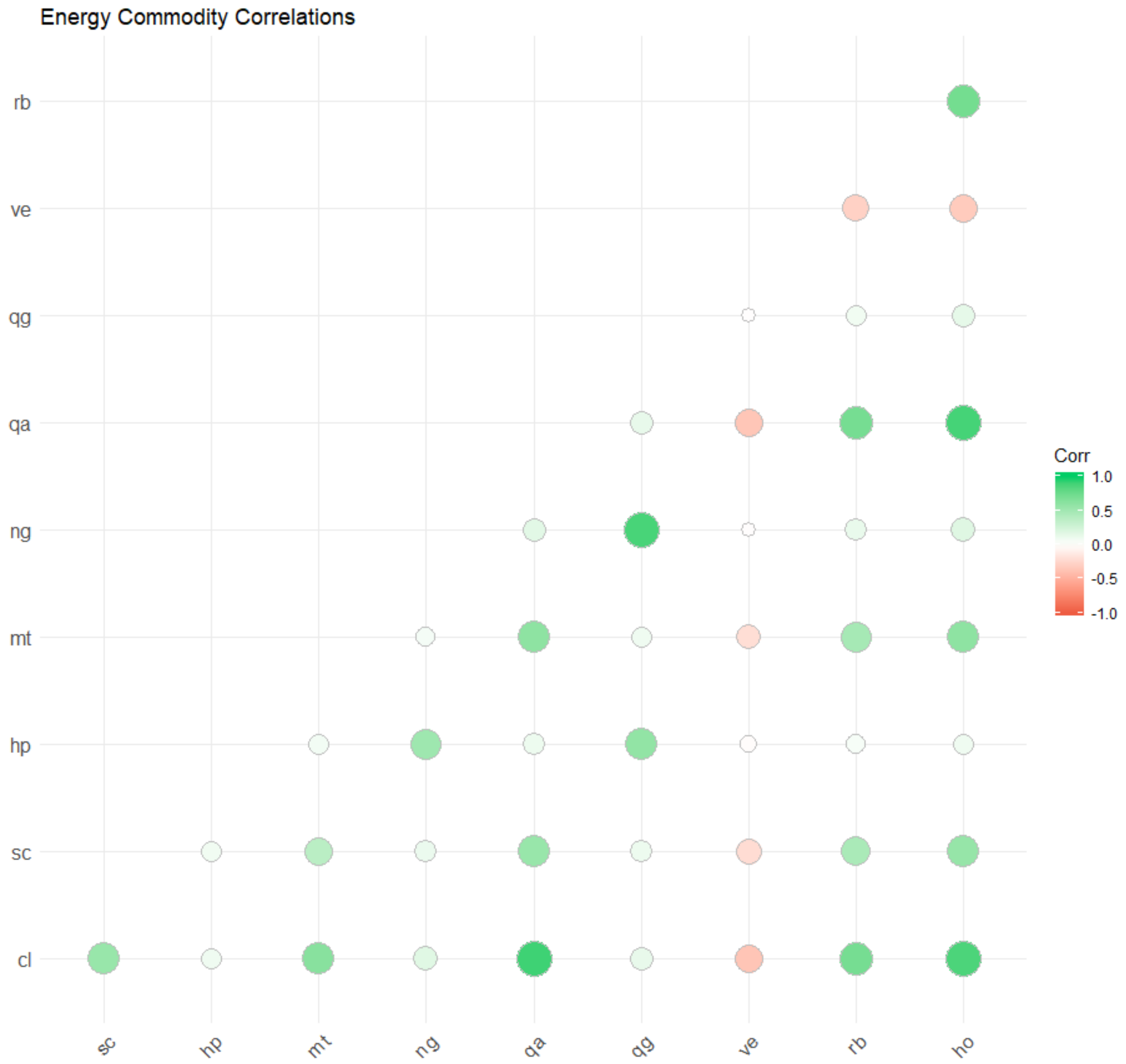


And a look at the fitted standardized residuals for WTI:



RELATIONSHIPS

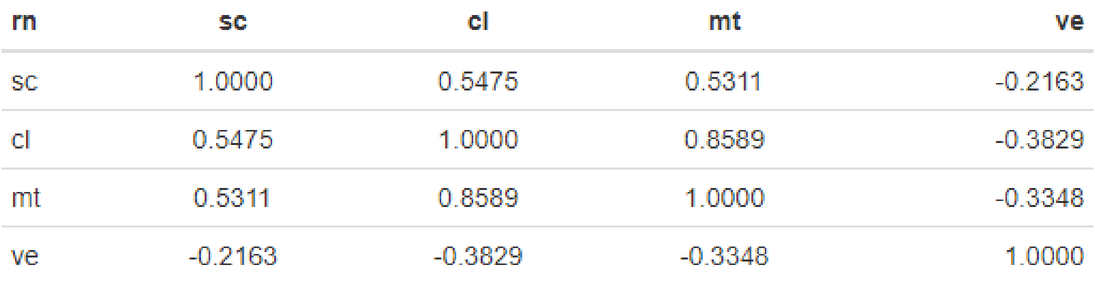
An important aspect of any quantitative trading strategy is to establish the strength of relationships between assets, instruments and markets. Relationships are important for many reasons including, but not limited to, creating potential arbitrage opportunities, hedging risks, managing drawdowns and expectation of returns and volatility. For example, if our strategy holds instruments with highly positive correlations, especially in the tails, then we would expect a high degree of volatility in our strategy. In the below chart, we can see the correlations of the returns of each of the commodities under analysis. These high-level relationships will drive further in-depth analysis.



In the above correlation matrix, we see the simple "linear" (Pearson's) correlation only produce negative correlations between VE, which is the volatility index for crude. Incidentally, the VE is a lagging measure of the rolling volatility in crude oil products, and the negative correlation to the VE is only prominent in the four crude symbols, CL, SC, MT and QA.

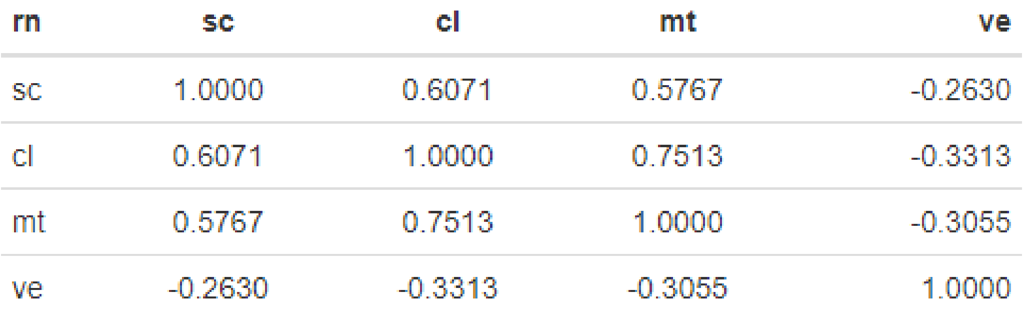
CRUDE CORRELATIONS

Person's Correlation:



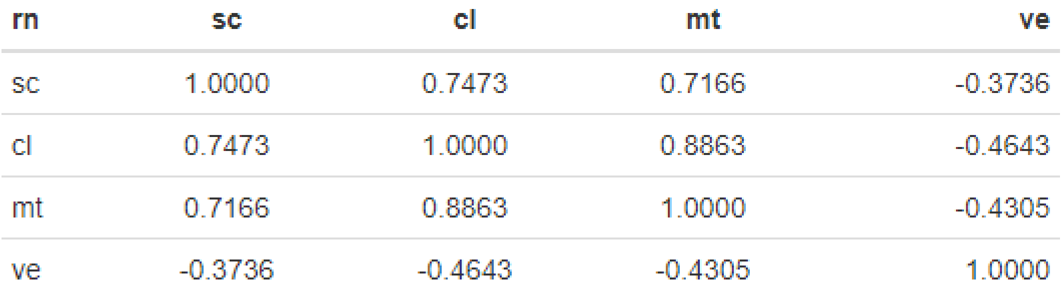
rn	sc	cl	mt	ve
sc	1.0000	0.5475	0.5311	-0.2163
cl	0.5475	1.0000	0.8589	-0.3829
mt	0.5311	0.8589	1.0000	-0.3348
ve	-0.2163	-0.3829	-0.3348	1.0000

Kendall's Tau:



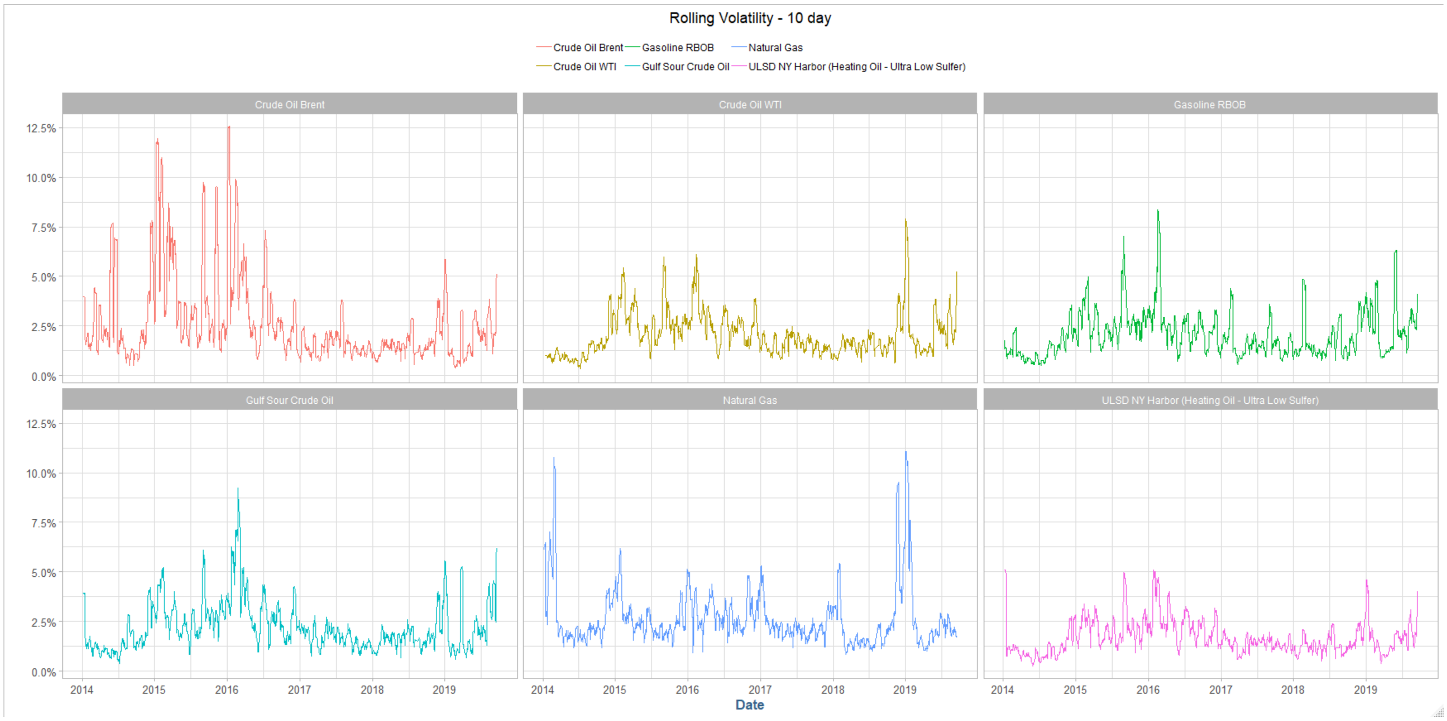
rn	sc	cl	mt	ve
sc	1.0000	0.6071	0.5767	-0.2630
cl	0.6071	1.0000	0.7513	-0.3313
mt	0.5767	0.7513	1.0000	-0.3055
ve	-0.2630	-0.3313	-0.3055	1.0000

Spearman's Rho:

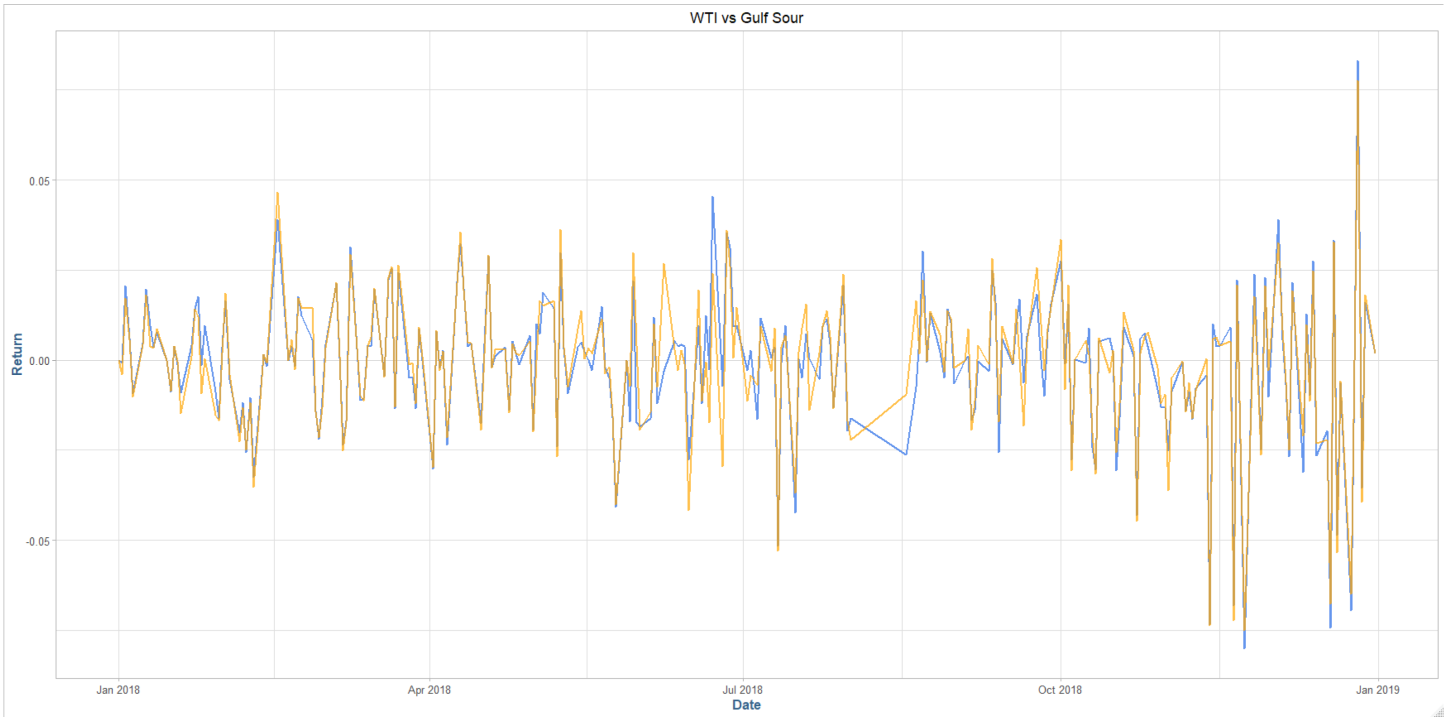


rn	sc	cl	mt	ve
sc	1.0000	0.7473	0.7166	-0.3736
cl	0.7473	1.0000	0.8863	-0.4643
mt	0.7166	0.8863	1.0000	-0.4305
ve	-0.3736	-0.4643	-0.4305	1.0000

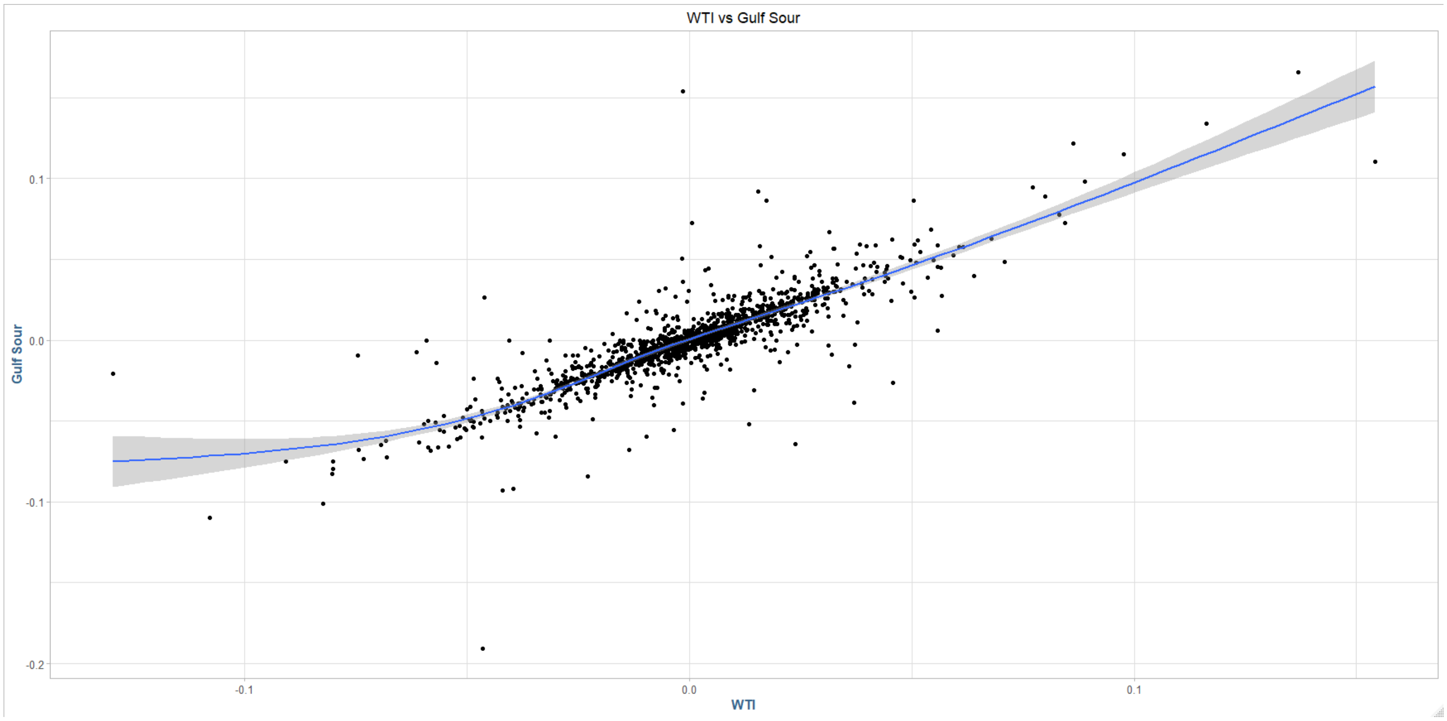
In addition to the correlations which measure the strength the commodities move in the same direction overall; we can compute a rolling standard deviation for a specified period. This process helps smooth out the variance over time, which cuts down on the noise and thereby increasing the transparency of the relationships in volatility:



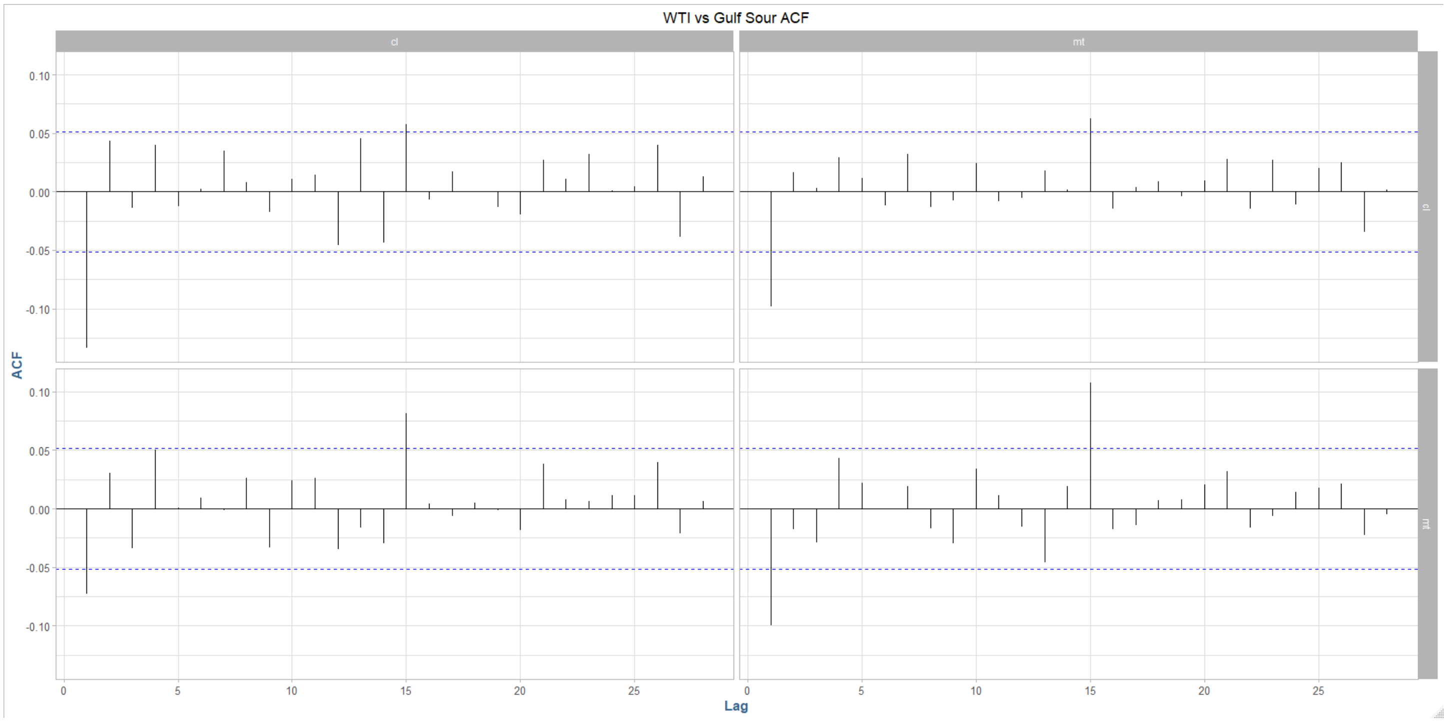
In the correlations we notice the Spearman's Rho for WTI and Gulf Sour seem to be directionally correlated about 89% of the time. Additionally, there is a strong linear relationship indicated by Pearson's correlation coefficient: Let's zoom in a highly correlated period, 2018:



A scatterplot of the returns reveals additional clustering, and confirms the near linear relationship:



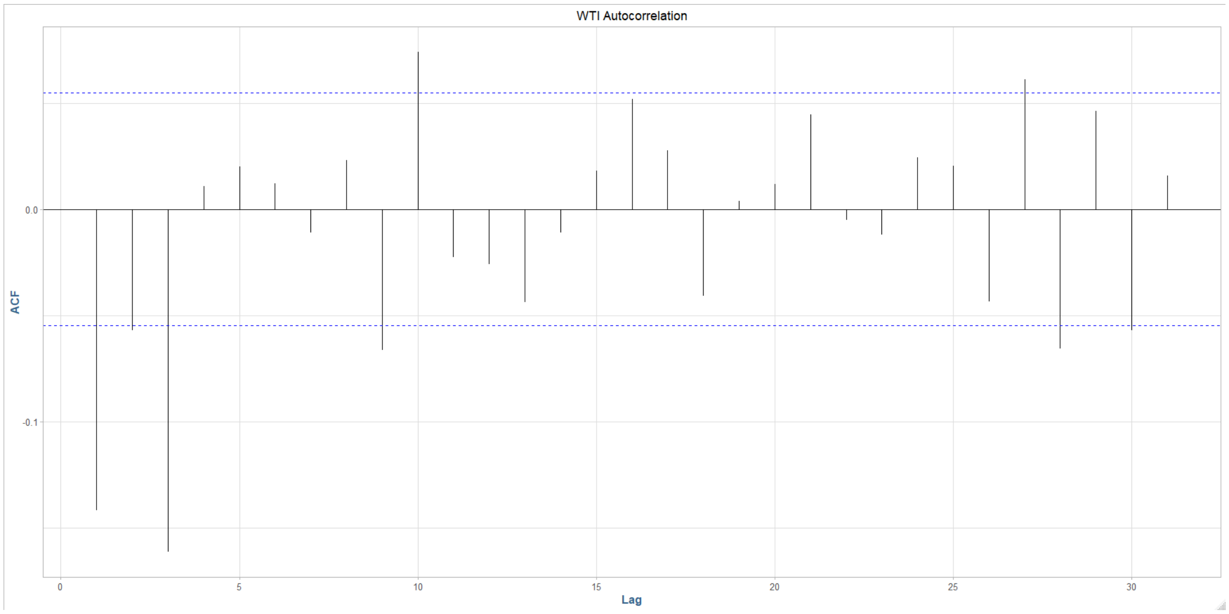
On a time-series basis, we see strong negative autocorrelations at 1, 4 and 15 days out, indication there could be a strong leading indicator here, therefore a transactional arbitrage opportunity:



There are additional strong relationships further out, however, the chance of those relationships being generated purely by chance are quite high.

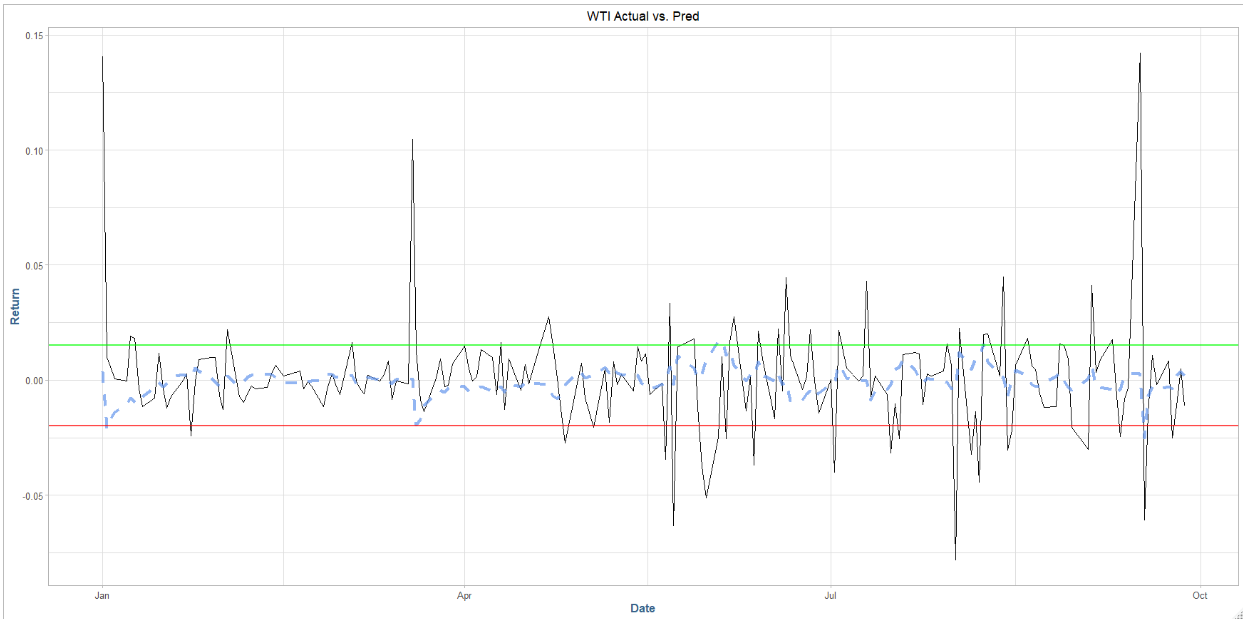
MODELING

The first model we are going to look at is one that is focused solely on predicting the returns for **WTI** as a stand-alone asset as the returns seem to exhibit the characteristics of a stationary process at lag = 4.



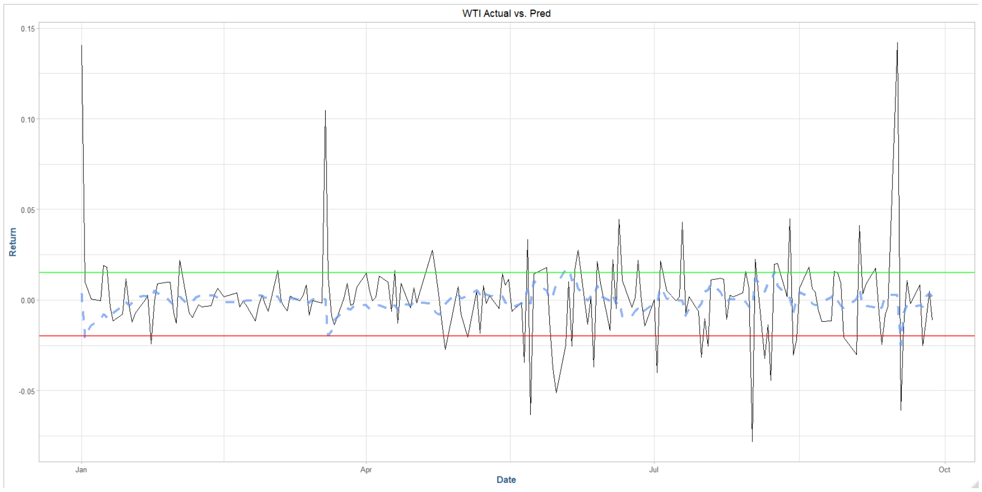
A Ljung-Box test at lag = 4 yield a p-value beyond the level of strongly significant at the .01 level, and therefore we reject the null hypothesis of a random-walk process and conclude there is indeed self-correlation at the lag interval = 4.

We split the data into test and training sets, where we train the model on the data prior to 2019 and the test on data in 2019. Our fitted model is an ARIMA(1, 0, 1) with mean zero with auto-regressive coefficient 0.5595 and moving average coefficient 0.78506. The Bayesian information criterion is -4,925.8 which is a good indication of fit. The performance of the model predicted (**blue**) compared to the actual WTI (**black**): Our trading bounds from the predictions are in green and red, indicating the direction we want to go with the model.



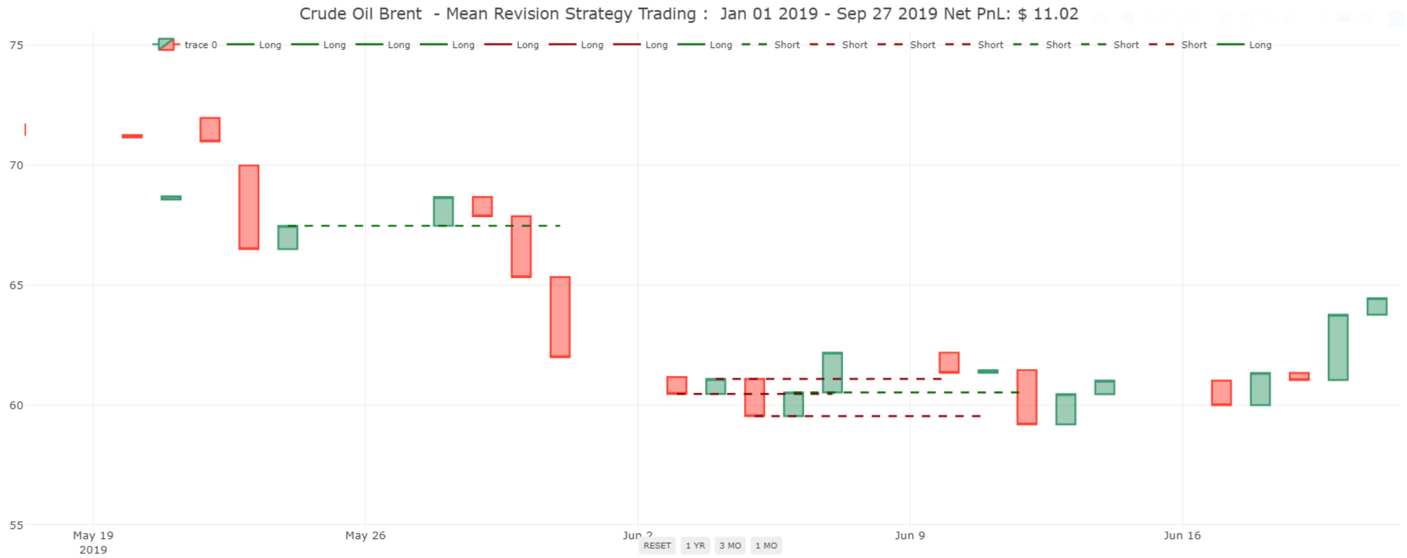
WTI

For the WTI mean revision strategy, we looked at the returns predicted from the model:



From this predicted data, our trading strategy is simple: we look for peaks and troughs in the returns, areas that are relatively large outliers (for this example, we're using a threshold of 0.0015, where the EV = 0). The basic idea is that because we have a stationary time-series, when the returns take a sharp up/down-turn, we enter the position (the direction is the **opposite** from the sign of the predicted return: positive, we **sell**, negative we **buy**). Since we found a statistically significant autocorrelation at lag = 4, we will take the inverse side and close out our position in **4 trading days from entry**, *always*.

For an example of the strategy in action, see the below plot for multiple positions entered over the period from May 19th to June 16th, 2019:

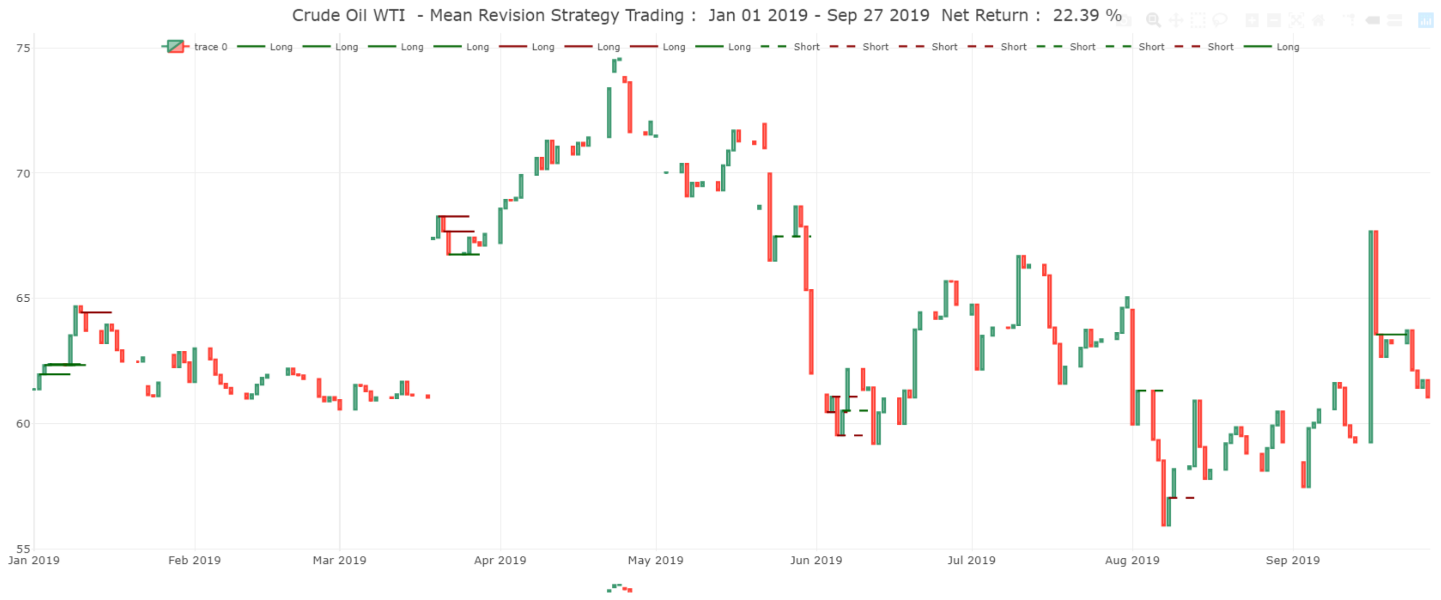


Some notation about the above chart:

Solid lines: Long position (buy low -> sell high)

Dashed lines: Short position (sell high -> buy low)

The line representing the position is through its entire holding period, and its colored green if it made money, or red if it lost money. Looking at one trade in detail:

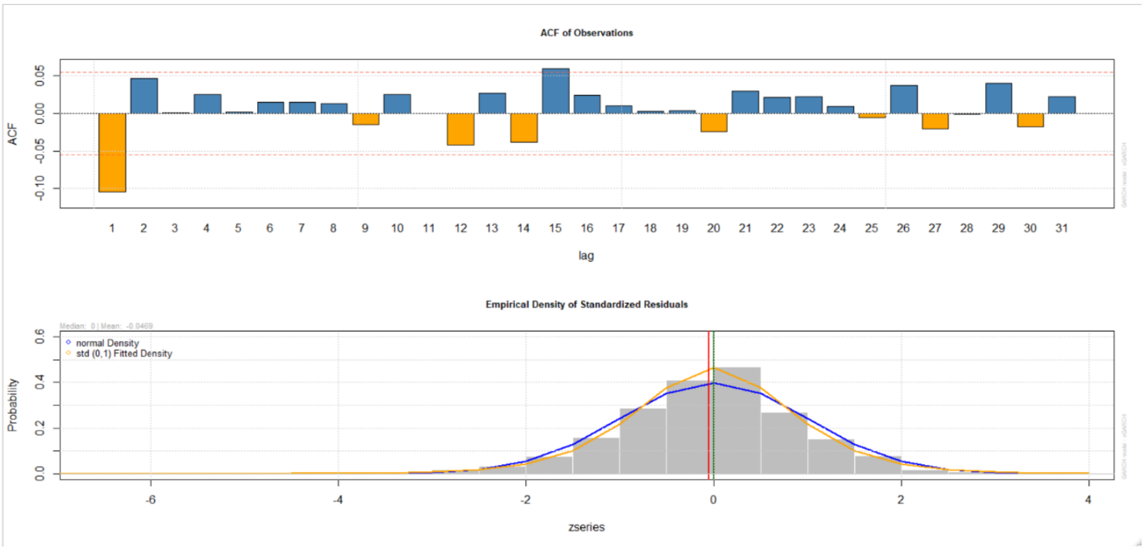


Here, we entered in the position on 5/25 with a short sell of WTI Crude at $67.47, and then issued a cover short to close out our position at $61.99, yielding a net profit of $5.48 per barrel. Assuming 44,000 barrels per contract, this transaction netted our portfolio a gross profit of $241,120.

Year to date in 2019, our testing data range, we issue exactly 16 positions (32 transactions), leaving us with a net zero position at the end of the holding period, for a total profit of $408,000 assuming 50,000 contracts per transaction.

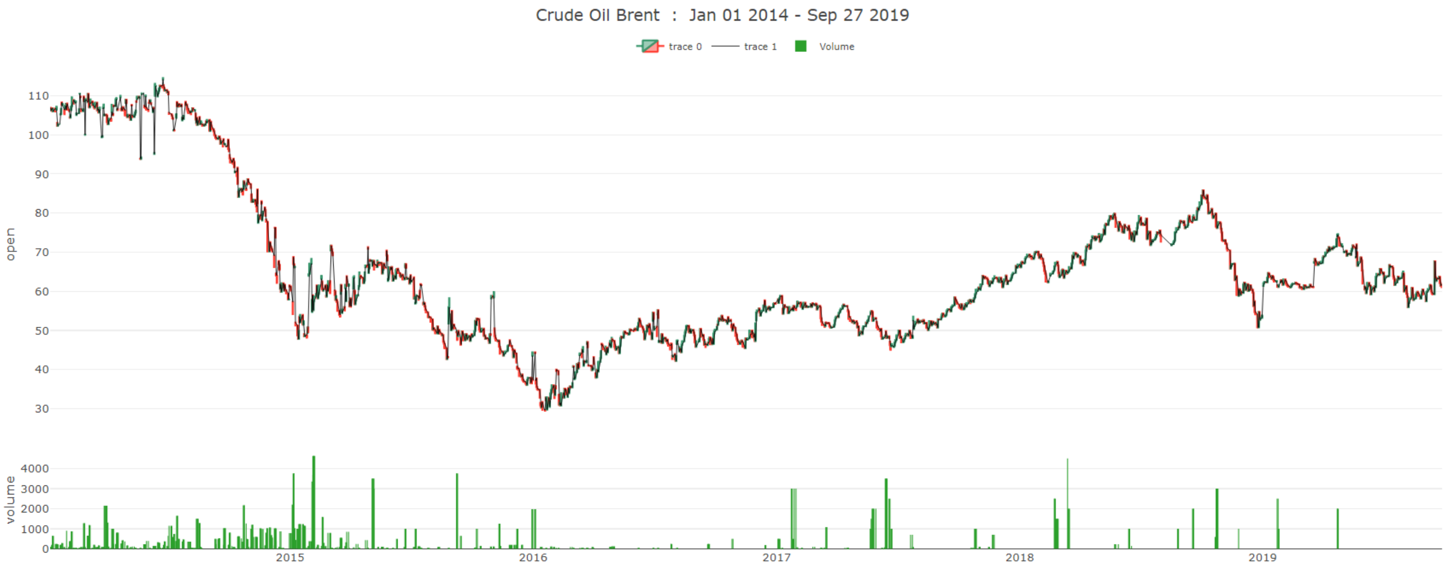
BRENT CRUDE

The next commodity we are going to look at is Brent crude. First thing we are going to do is look at the autocorrelations of the time series:



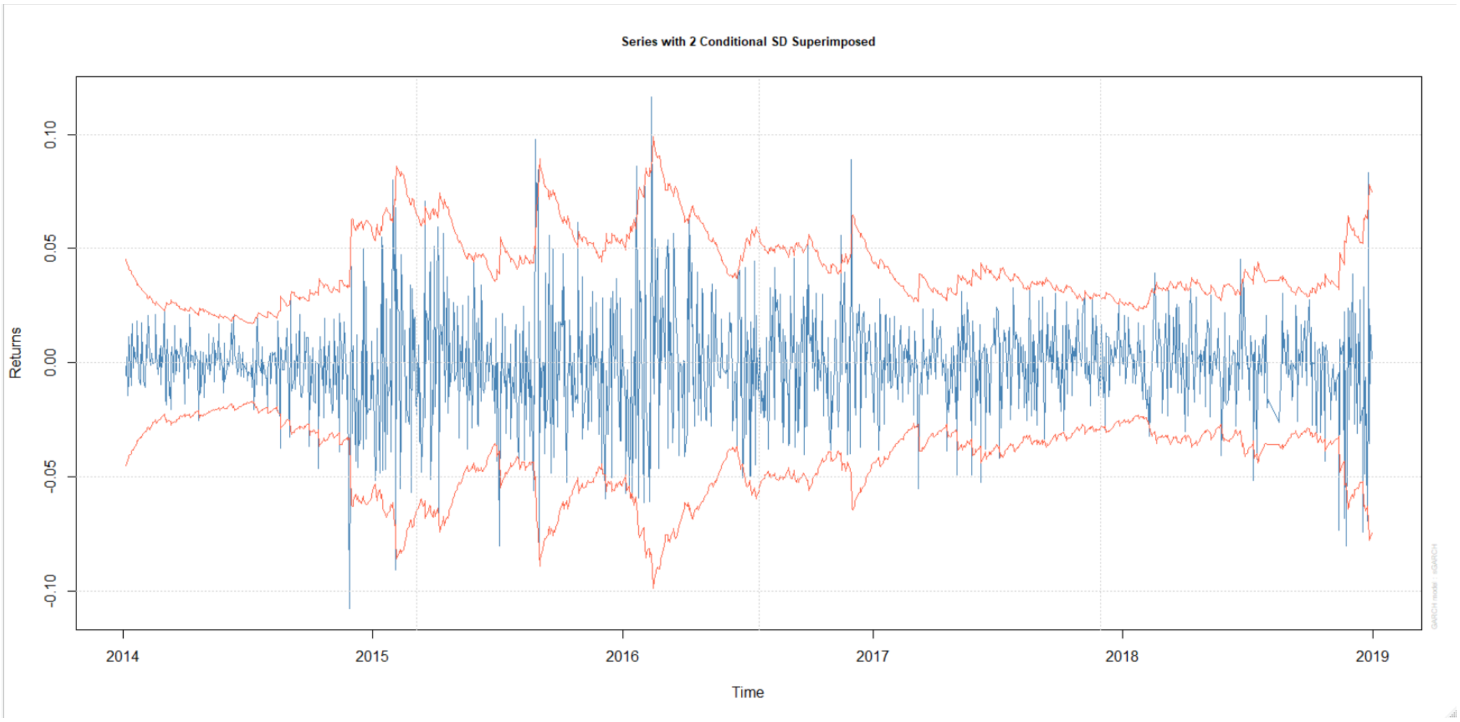
The first thing we notice is that at lags 2 and 15 there are significant positive autocorrelations. At lag 15 we have statistical significance with the Ljung-Box tests at a strong level, however, due to the extreme volatility of the series we are going to further explore the 1 period lag.

Notice the extreme volatility of the Brent series, and the momentum like behavior of the price series:

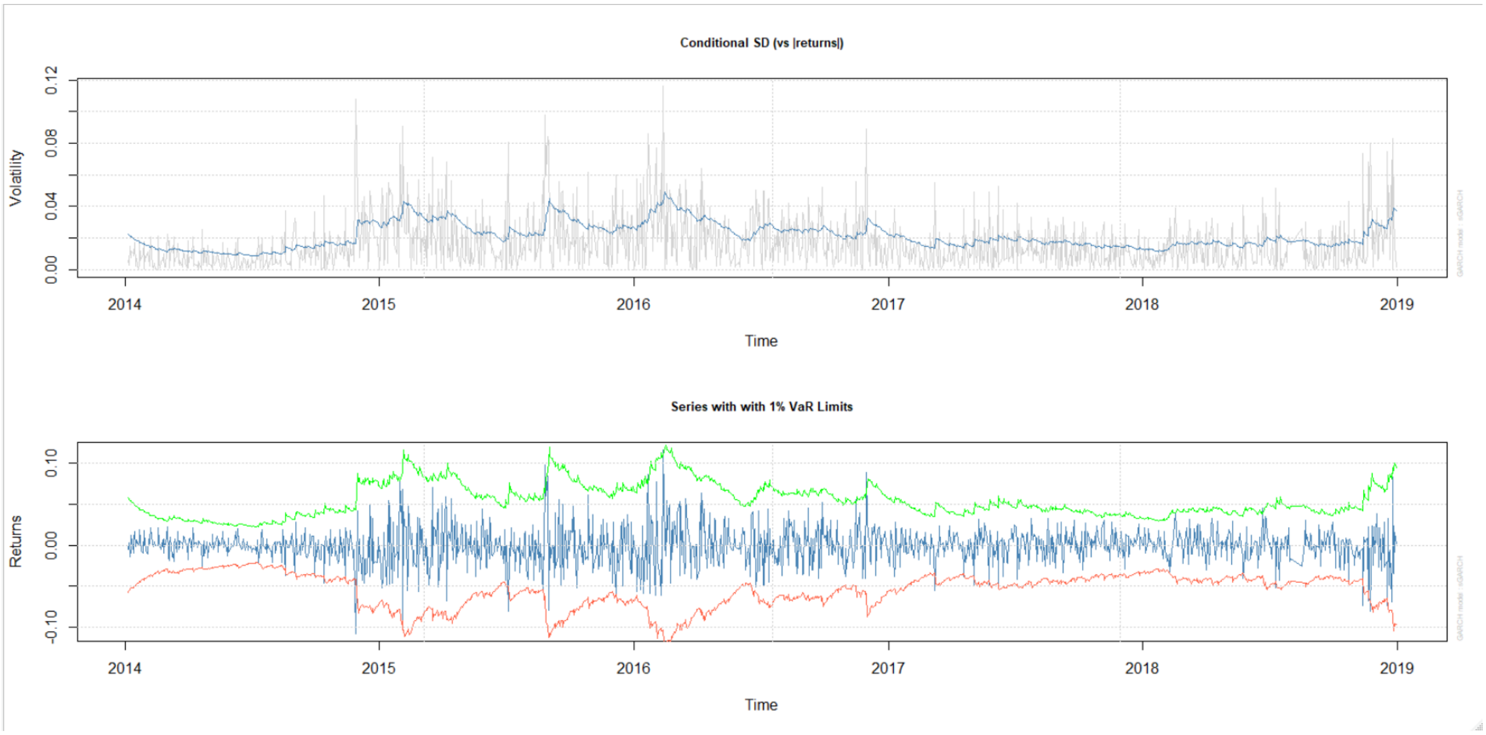


We are going to use an Autoregressive/Moving Average model with drift (non-stationary), to model the random walk behavior we see from the prices. We are going to attempt to model the volatility of our series as a separate GARACH process, that we will use in conjunction with our ARIMA(1, 1, 0) model for the forecasting. Our fitted model is an ARIMA(1, 1, 0) with drift (mean != 0), with an autoregressive coefficient of -0.586.

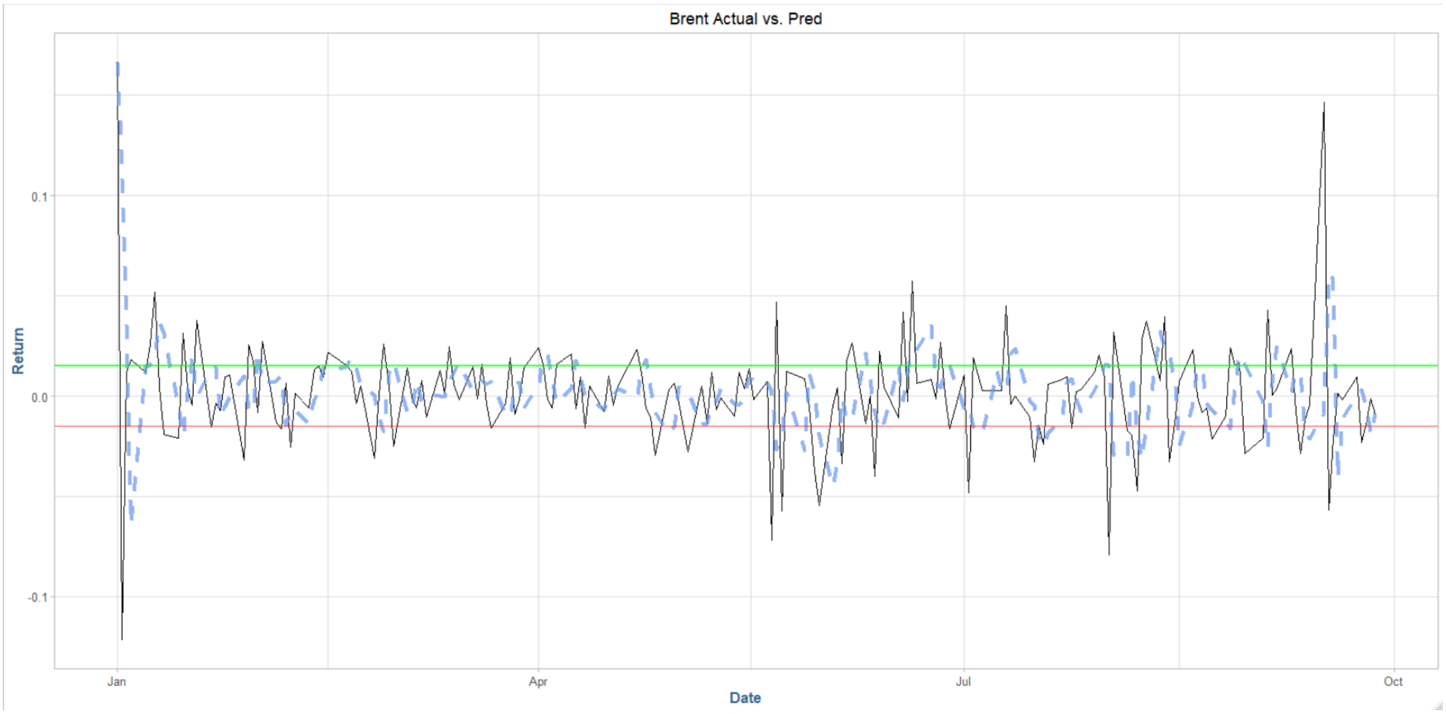
The volatility of the returns can be seen below, where we superimposed a 2 standard deviation fit to the returns to bring to the forefront how extreme volatility is in this series. The volatility can be used to our advantage, if we build a model that considers the explosive momentum seen in large periods of the instrument's prices. This strategy is going to be one that has a short holding period, one day, and the trading signals are going to come from a low threshold from the models predicted values.



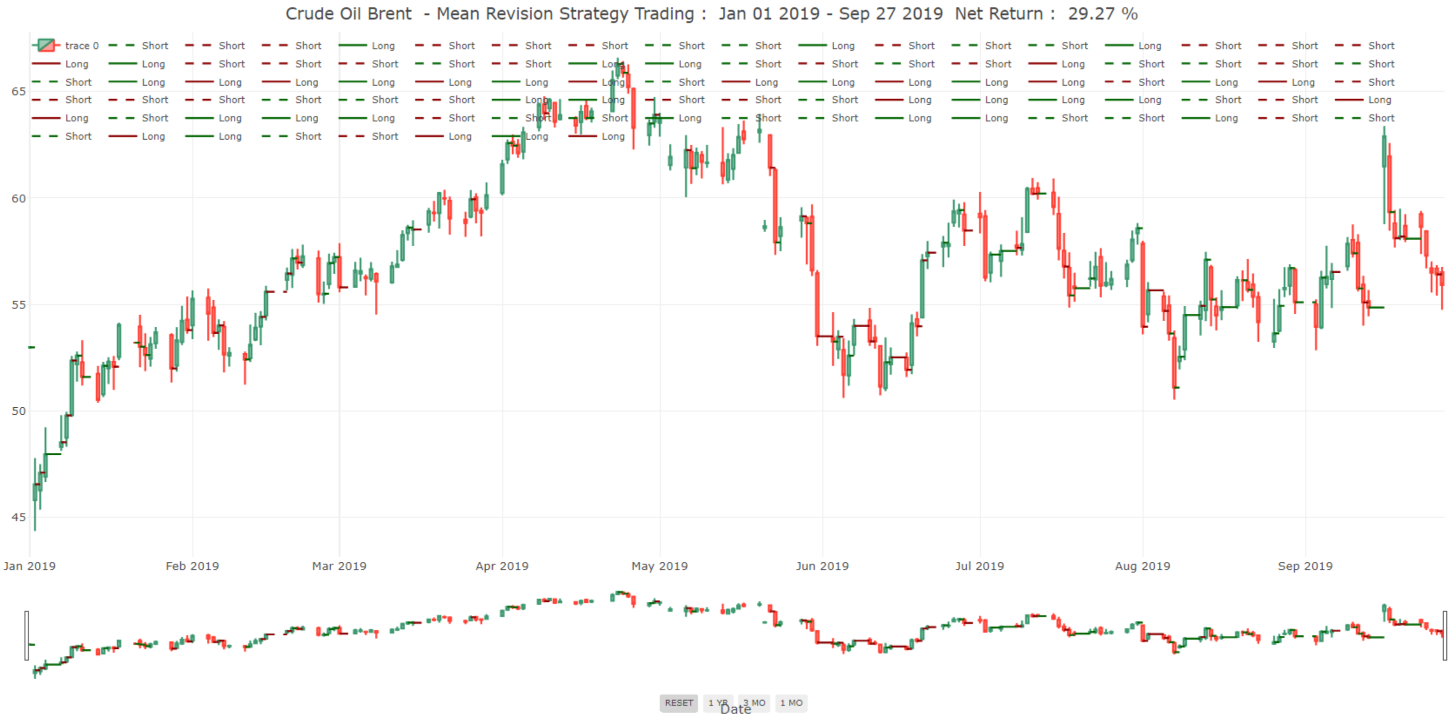
We are going to use an Autoregressive/Moving Average model with drift (non-stationary), to model the random walk behavior we see from the returns. We are going to attempt to model the volatility of our series as a separate GARACH process, that we will use in conjunction with our ARIMA(1, 1, 0) model for the forecasting. Our fitted model is an ARIMA(1, 1, 0) with drift (mean != 0), with an autoregressive coefficient of -0.586.



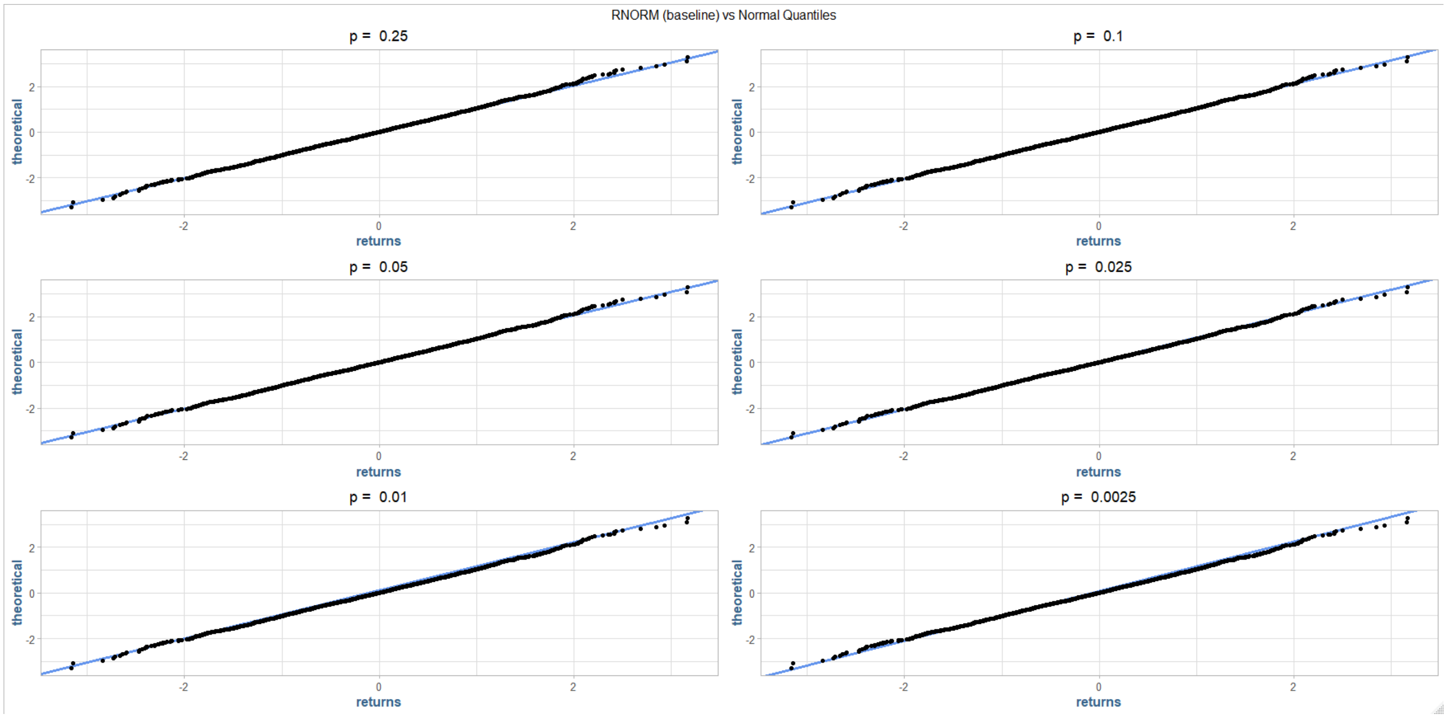
This model is a more aggressive attempt to capture the oscillations of the return series and profit from the extreme levels of volatility. We are going to first use a threshold of .02, as the threshold for our trading rule, and our direction will be inverse to the direction of the movement, due to the strong positive autocorrelation behavior we have observed.



With a low trading threshold and a short-period, explosive volatility time series, the resulting strategy is a one with a 1 day holding period per transaction. The annualized return on the strategy is 29.27%, resulting from 97 holdings over the period.



Normal Quantiles, Benchmark fit


RNORM (baseline) vs Normal Quantiles
p = 0.25
p = 0.1
p = 0.05
p = 0.025
p = 0.01
p = 0.0025
theoretical
returns