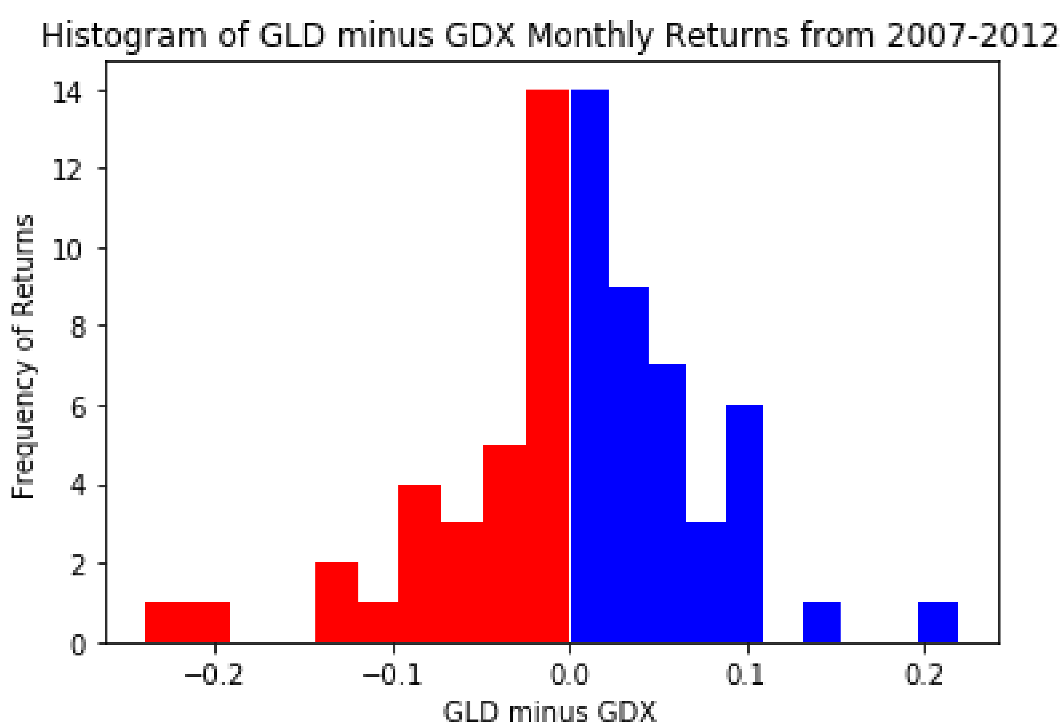
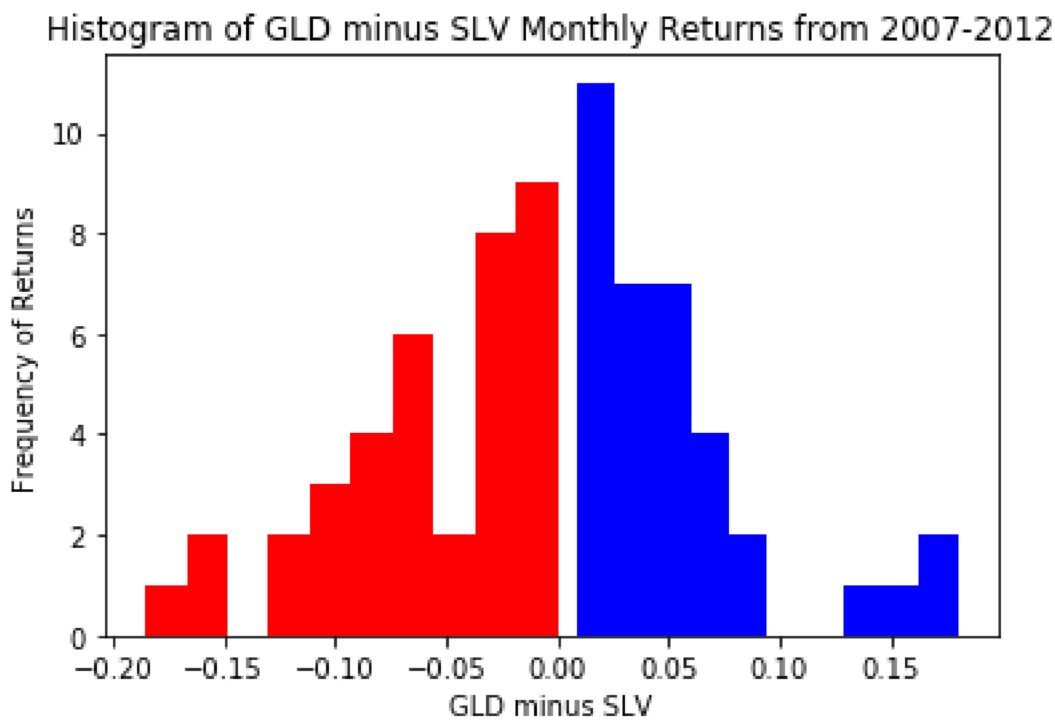
Pairs-Trading Strategy with ETFs

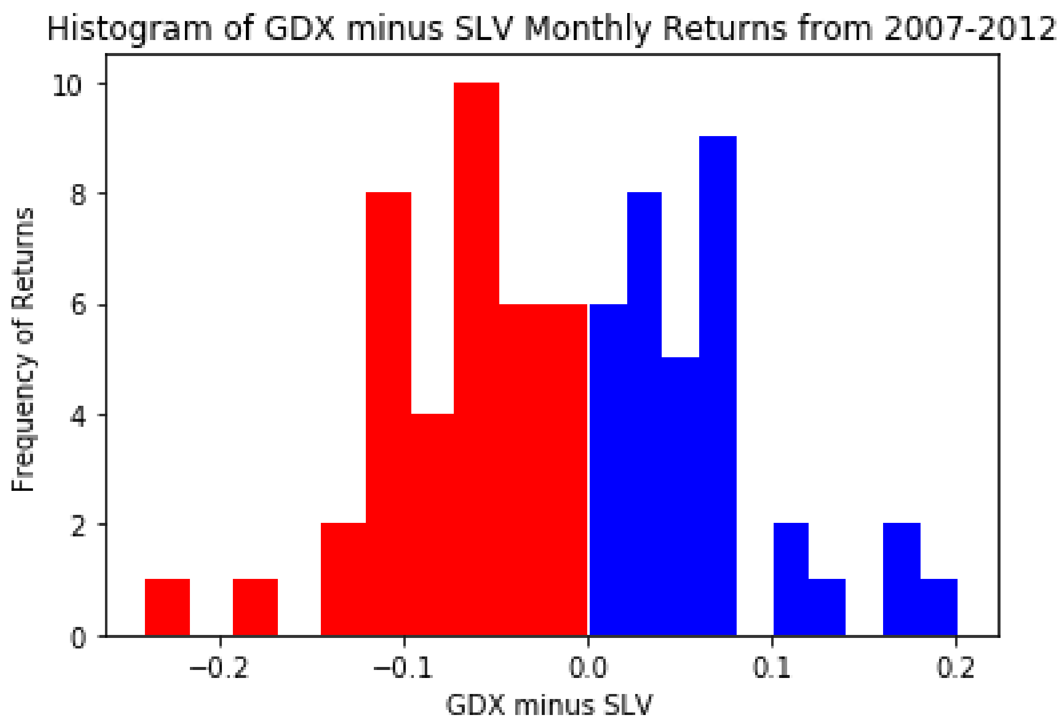
I developed a bunch of trading strategies going long-short all combinations of the following ETFs: GLD, GDX, and SLV. I followed through in most of what was asked, but I did a bit extra in forming additional strategies, all in the same vein of looking at the distribution of returns at different frequencies of monthly, quarterly, and yearly. The backtests and out-of-sample tests performed could be classified into 3 groups: 1) predictive average – as outlined by you; 2) looking at different statistical metrics, such as correlation and volatility, skewness, and kurtosis of the long-short pair; and 3) same as 2), but overlaying that with how relatively speaking the ETF was doing based on a certain statistic compared to its pair. I will analyze each of these strategies in detail.

Backtests for all frequencies – monthly, quarterly, and yearly – were done from 2007 – 2012, while the out-of-sample tests were from 2013 – 2016. In all strategies, I look at every possible long-short combination (not permutation). There were certain relevant criteria for a pair to be traded, which will be different for each strategy. Since we are looking at the distribution of returns and seeing if it has predictive value, this can be seen as a form of timing, based on relative levels in the backtest period of each long-short pair, as opposed to relative levels cross-sectionally amongst all ETF pairs. That is, we enter a trade, when a certain level is reached, as that is a signal, which we hypothesize to be predictive in the backtest.

For the first strategy, which I label as Predictive Averages, I look at every possible combination of long-short ETFs for each frequency: monthly, quarterly, and yearly and place them in return buckets for both positive and negative returns, 10 for each. In the backtest, I look for two things: 1) how many returns were there in a decile and 2) the average return in a decile if there is one. Here are histograms that include information about the amount of returns in each bucket for each ETF pair for each frequency:







Monthly	Negative Returns - Count	Positive Return - Count
GLD – GDX	31	41
GLD – SLV	37	35
GDX – SLV	38	34

Monthly		Min	Mean	Max
GLD – GDX	Negative	-25.45%	-5.96%	-0.1%
	Positive	0.10%	5.59%	21.86%
GLD – SLV	Negative	-18.55%	-4.92%	-0.00%
	Positive	0.00%	4.73%	18.00%
GDX – SLV	Negative	-24.00%	-6.43%	-0.00%
	Positive	0.00%	6.46%	31.89%

The average returns and counts in each decile, for both positive and negative returns, in the backtest is given in the excel spreadsheet for the next month. The count and distribution of negative to positive returns is roughly similar. It is difficult to make any discerning statements about the distribution of returns considering there being a total of 20 buckets for the backtest. Part of the problem is that there might be a large magnitude return in a particular decile, but it might not have very many returns, which is expected for the outer deciles, such as for Decile 1 for Negative Returns and Decile 10 for Positive Returns. For monthly returns, I wanted to see about 10% of the data in a decile, before I thought of investing in it the next month, so about 7. Additionally, the average return in that decile should be greater than 1% in magnitude, as we can always short the pair.

The distribution of returns in each bucket at best seems quite random. Successful backtest deciles included:

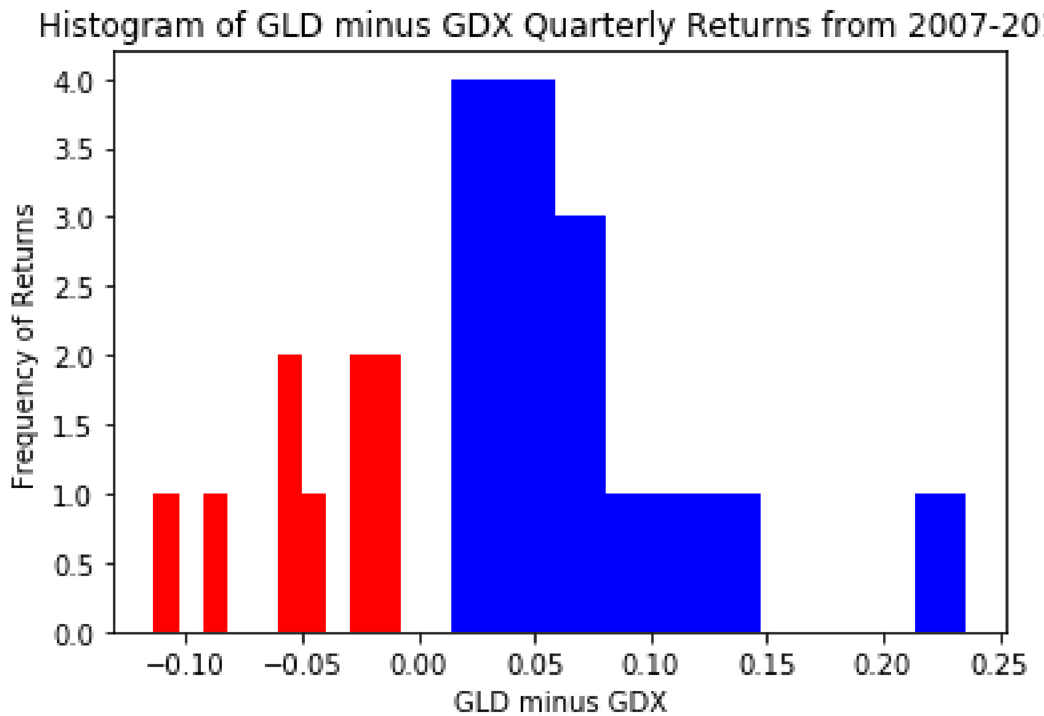
ETF Pair	Positive or Negative	Decile	Return Direction
GLD – GDX	Negative	10	+
GLD – SLV	Positive	1	-
GLD – SLV	Negative	10	-
GDX – SLV	Positive	2	-
GDX – SLV	Negative	6	-

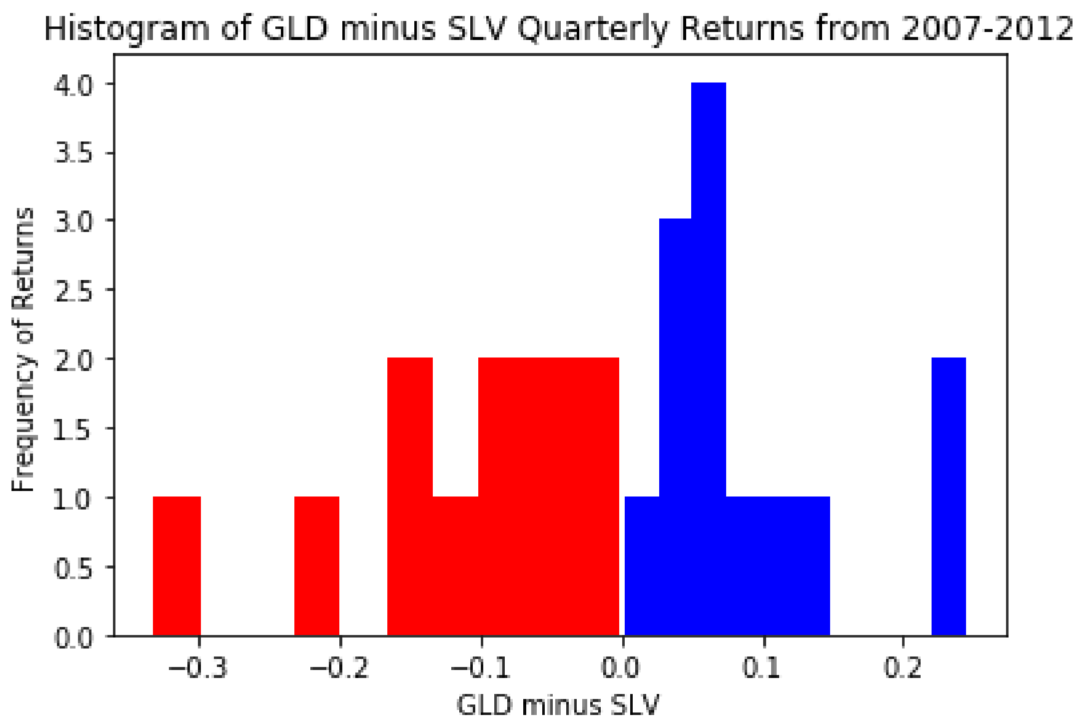
Looking to see how they did on the test set:

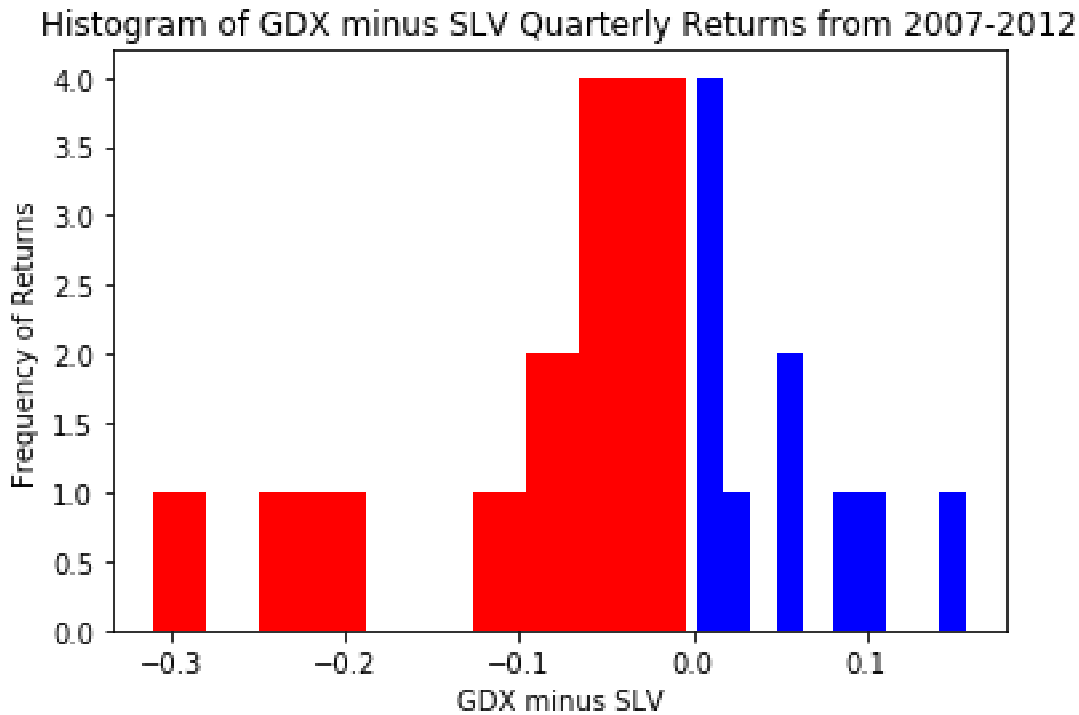
ETF Pair	Positive or Negative	Decile	Returns
GLD – GDX	Negative	10	36.72%
GLD – SLV	Positive	1	-23.31%
GLD – SLV	Negative	10	-2.58%
GDX – SLV	Positive	2	0.00%

The test set incorporates information about how the decile did for each ETF pair in the backtest and uses that to go long or short. This means that if the backtest for a decile was negative, then I went short it in the test set, while if it had a positive return in the backtest, I went long the differential in the test set. What this means is that decile 1 for GLD – SLV in positive returns, was long SLV and short GLD in the test set, as the return in the backtest was negative. In short, if the backtest results were suitable to the test set period, then all the returns should have the same sign, positive. Given that information, GDX – SLV never went into decile 6 when the differential pair was negative, which isn't unusual. A lot of deciles were empty in the test set. When the differential between GLD – GDX was slightly negative, corresponding to be in decile 10, there was significant outperformance, but when GLD – SLV was slightly positive, corresponding to decile 1, there was significant negative performance. Most of the magnitude of returns in the test set were larger than the backtest, which is probably a function of lesser data in the test set.

Turning to quarterly frequency, here are plots of histograms for each ETF pair:







Quarterly	Negative Returns – Count	Positive Return - Count
GLD – GDX	9	15
GLD – SLV	11	13
GDX – SLV	14	10

Quarterly		Min	Mean	Max
GLD – GDX	Negative	-31.25%	-8.79%	-0.10%
	Positive	0.10%	7.60%	23.53%
GLD – SLV	Negative	-33.17%	-9.10%	-0.00%
	Positive	0.00%	7.02%	24.43%
GDX – SLV	Negative	-30.98%	-8.55%	-0.00%
	Positive	0.00%	7.57%	34.26%

The results of the backtest are in the spreadsheet. Once again, the counts of negative and positive returns are about the same. For a decile to be chosen, it needed to have about 4 counts and an average return greater than 2.5% in magnitude. This roughly accounted for about 16.67% of the backtest data, as there wasn't as much.

The distribution of returns in each bucket at best seems quite random. Successful backtest deciles included:

ETF Pair	Positive or Negative	Decile	Return Direction
GLD – GDX	Positive	1	+
GLD – GDX	Positive	2	-
GLD – SLV	Positive	3	-

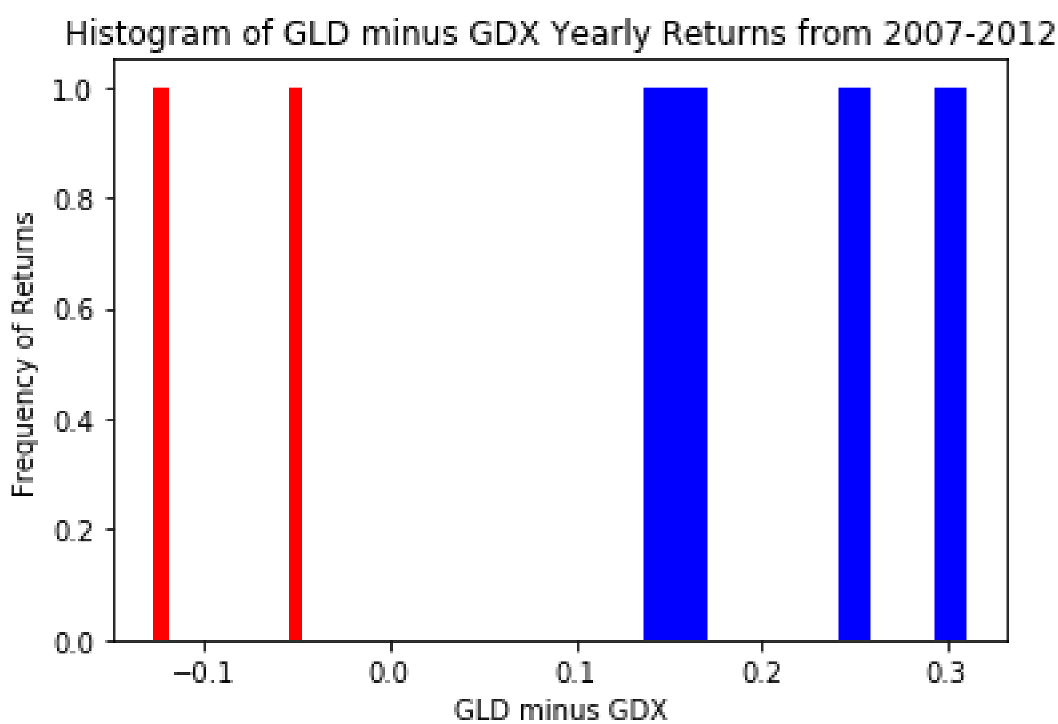
GDX – SLV	Positive	1	-
GDX – SLV	Negative	9	-

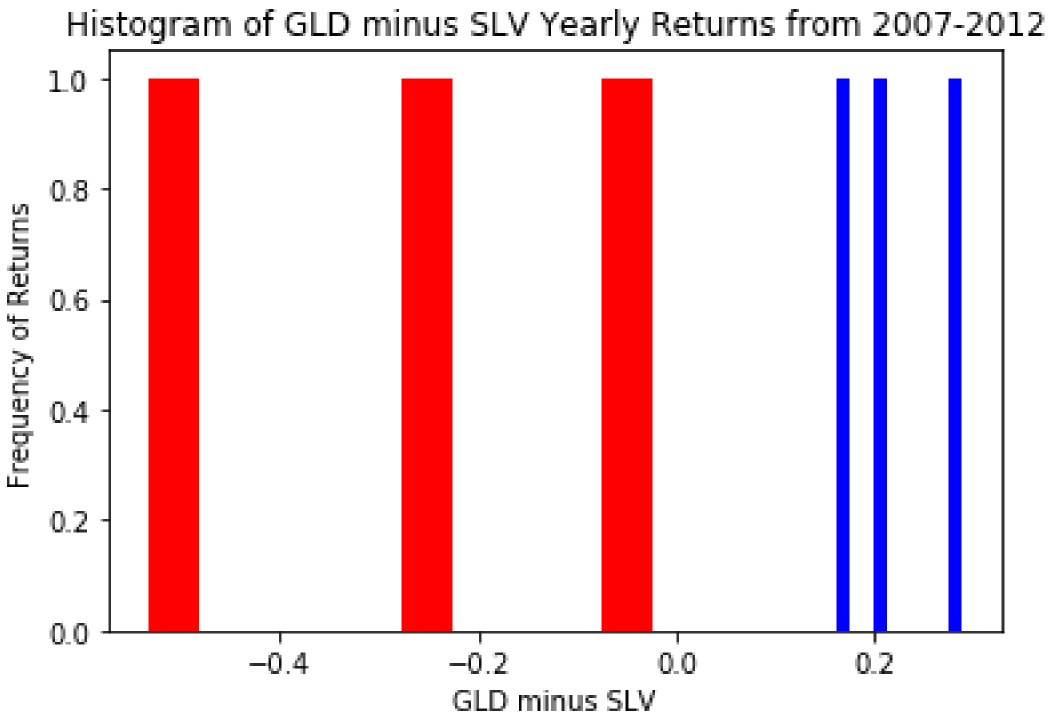
Looking to see how they did on the test set:

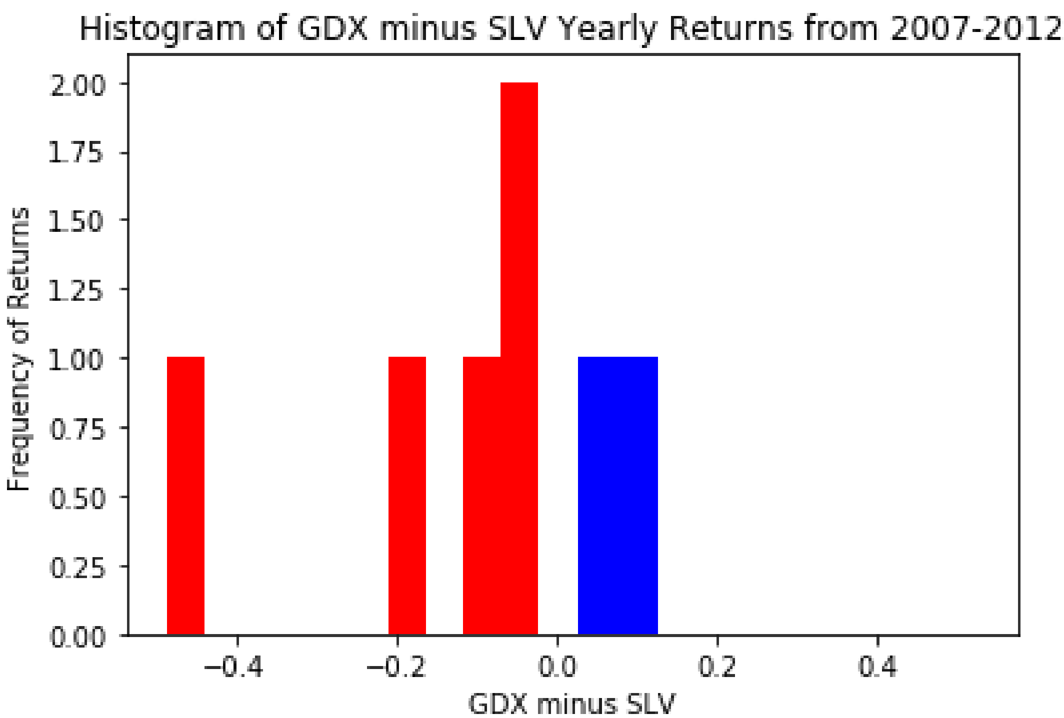
ETF Pair	Positive or Negative	Decile	Returns
GLD – GDX	Positive	1	17.73%
GLD – GDX	Positive	2	-7.61%
GLD – SLV	Positive	3	5.71%
GDX – SLV	Negative	9	8.89%

This time around, the test set returns seem to be much better. Only GLD – GDX decile 2 for positive returns did not come into the backtest, which isn't surprising. Most of the returns are positive, so the backtest would have gotten the direction of the trade correct. All the returns are rather large in magnitude, compared to the criteria of 2.5% in magnitude for inclusion in the backtest. The magnitudes of the returns are in fact larger than the backtest, which is good for all except the case for GLD – GDX decile 2 for positive returns. However, this is probably more a result of luck than skill, as the test set is smaller and the few returns that could have fallen into a decile were more pronounced. Had the test set been larger, it would not be unusual to see these magnitudes fall.

Turning to yearly frequency, here are plots of histograms for each ETF pair:







Yearly	Negative Returns - Count	Positive Return - Count
GLD – GDX	2	4
GLD – SLV	3	3
GDX – SLV	4	2

Yearly		Min	Mean	Max
GLD – GDX	Negative	-44.89%	-20.74%	-4.65%
	Positive	10.26%	19.40%	31.00%
GLD – SLV	Negative	-53.20%	-21.45%	-2.42%
	Positive	1.75%	15.38%	28.72%
GDX – SLV	Negative	-48.55%	-16.45%	-2.28%
	Positive	2.52%	15.98%	38.36%

The results of the backtest are in the spreadsheet, along with counts. Dealing with yearly data is a bit more complicated, as there were only 6 points in the backtest, and so even 1 observation in a decile is enough for a signal, provided that the average return, or just return for the most part, is greater than 8% in magnitude.

ETF Pair	Positive or Negative	Decile	Return Direction
GLD – GDX	Positive	1	+
GLD – GDX	Positive	2	+
GLD – GDX	Positive	7	+
GLD – GDX	Positive	10	-
GLD – GDX	Negative	10	+
GLD – SLV	Positive	1	+
GLD – SLV	Positive	10	-
GLD – SLV	Negative	1	+
GDX – SLV	Negative	6	-
GDX – SLV	Negative	7	-
GDX – SLV	Negative	9	-
GDX – SLV	Negative	10	-

This is how the test-set performed:

ETF Pair	Positive or Negative	Decile	Returns
GLD – GDX	Positive	1	-44.89%
GLD – GDX	Positive	7	10.26%
GLD – SLV	Positive	1	1.75%
GDX – SLV	Negative	7	-7.07%

It is easy to understand why there were so many tests, as the barrier to enter the test set was low. However, only 4 deciles had results from the test set, which once again isn't very surprising. It seems that information from the backtest was not as informative for the test set, and only two of the returns from the test set had beaten the threshold return of 8% for inclusion. The direction of returns does not seem very informative either in this case. Additionally, none of the deciles consistently appear across all frequencies.

In the next strategy, I look at correlation between the two ETF pairs, and volatility, skewness, and kurtosis of the long-short ETF pair and see if that is predictive in nature. I do this first by calculating the relevant statistics from daily data up to the noted frequency. I then proceed to create buckets for those statistics based on the backtest period data. I see if any of the deciles have predictive value in the backtest and then test them in the test set.

For performance measurement, I calculate the average return, standard deviation and risk-adjusted return for each decile in the backtest period in assessing predictive value of the upcoming period. Risk-adjusted return is simply the average return divided by the standard deviation. The statistics for a decile to be included in a test set are 1) the amount of returns in a decile; and 2) the magnitude of risk-adjusted return above a certain threshold. This table indicates the minimum thresholds for inclusion of a decile for each frequency:

	Minimum Counts	Risk-Adjusted Return Magnitude
Monthly	7	0.2887
Quarterly	4	0.5
Yearly	1	1

It is important to remember that the test set is not very large and with a lot of buckets for a return to be placed in, some buckets will remain empty in the test set. What additionally doesn't help is the fact that the backtest period was much more volatile than the test set period. Once again, there is also the issue of having a large risk-adjusted return in magnitude accompanied by a small amount of returns that excludes some deciles from being tested.

I went out to examine how the level of correlation between two ETFs can affect the future realized returns for a couple of reasons. For one, correlation is a largely used statistic in the investment industry. Since we are going long-short, the correlation statistic should have an intimate relation with the differential between going long one ETF and short the other. More precisely, when the correlation is relatively high, I wouldn't expect the differential to be very large in magnitude concurrently. If one expects the correlation to decrease, then there are more chances for the differential to increase and profit, but if one expects the correlation to increase, then the chances for the differential to increase is less. What is important here would be to see if future correlation can be predicted.

Next I went to examine how the volatility of the differential in long-short ETF pairs can be used to trade profitably. It is typically known that higher volatility is indicative of negative future returns and lower volatility of more stable future returns. The reasoning being that with higher volatility there is greater chances for outsized movements in either direction. Typically, investors are much more sensitive about bad news and so there is a negative bias there. The formula for volatility for the ETF pair is:

$$var(r_{long} - r_{short}) = var(r_{long}) + var(r_{short}) - 2Cov(r_{long}, r_{short})$$

Correlations between related ETF pairs tends to be positive, so there should be a meaningful reduction in total volatility when trading both long short. In the event of high or rising correlations, we would expect the total volatility to be rather muted, especially since rising correlations can be associated with negative sentiment. In the event of low or decreasing correlations, one would expect the total volatility to be elevated, giving rise to greater chances of outsized returns. This would be concurrent during more neutral/positive market environments. In some respects, this signal is the opposite of both the correlation signal and the association of elevated volatility with positive outcomes. However, the volatility of the long and short sides is both dynamic and so that would offset some of the opposing effects mentioned with regards to correlation and market sentiment. Taking these dynamics into account, it can be insightful to know what exactly is driving the combined volatility of the positions, which could be key to analyzing the predictability of future returns.

I also look at the skewness of the combined position. A positive skewness would be indicative of the bulk of return distribution being negative or in this case less than the mean, and a large tail that extends in the positive direction. The opposite is true for negative skewness. The importance of this statistic is that if on average we know where spreads are going to be centered around and having some indicator that might be able to tell when there might be large divergence, can help create a profitable trading strategy. I do not explore what that signal could be that thoroughly, as it is most likely dependent on other exogenous factors. The most important feature in this type of situation is to get the direction of the trade right, otherwise outsized losses can be experienced.

Finally, I look at kurtosis of the positions. In this report, I used Fisher's definition of kurtosis, which corresponds to a value of 0 belonging to a normal distribution, as opposed to 3. High kurtosis values are said to have fatter tails and could thus generate more extreme spreads, whereas low kurtosis values have thinner tails and the spreads generated will tend to be more balanced. Unlike skewness, there isn't much indication as to the direction of the outsized returns. Much more important for the backtest is to see when there might be chances of such returns. Later, these statistics are overlaid to see if that can add value.

Backtest results of different strategies based on correlation and volatility, skewness, and kurtosis of spreads are provided in the spreadsheet. Also available are the number of months attributable to a certain statistic. First, monthly returns are evaluated based on different signals and here is a backtest of the results of all the strategies:

ETF Pair	Strategy	Decile	Return Direction
GLD – GDX	Correlation	6	+
GDX – SLV	Correlation	8	-
GDX – SLV	Correlation	9	-
GLD – GDX	Differential Volatility	2	+
GLD – SLV	Differential Volatility	3	-
GLD – GDX	Differential Skewness	5	+
GLD – GDX	Differential Skewness	6	-
GLD – GDX	Differential Skewness	7	-
GLD – SLV	Differential Skewness	2	+
GDX – SLV	Differential Skewness	2	-
GDX – SLV	Differential Skewness	3	-
GDX – SLV	Differential Skewness	7	+
GLD – GDX	Differential Kurtosis	1	+
GLD – GDX	Differential Kurtosis	5	+

Here is the test set of the strategies that made it into the backtest:

ETF Pair	Strategy	Decile	Returns
GLD – GDX	Correlation	6	10.18%
GDX – SLV	Correlation	8	-17.39%
GDX – SLV	Correlation	9	-0.37%
GLD - GDX	Differentials Volatility	2	23.67%
GLD - SLV	Differentials Volatility	3	1.85%
GLD - GDX	Differentials Skewness	5	24.58%
GLD - GDX	Differentials Skewness	6	19.98%
GLD - GDX	Differentials Skewness	7	32.37%
GLD - SLV	Differentials Skewness	2	2.19%
GDX - SLV	Differentials Skewness	2	8.08%
GDX - SLV	Differentials Skewness	3	14.10%
GDX - SLV	Differentials Skewness	7	19.43%
GLD - GDX	Differentials Kurtosis	1	15.43%
GLD - GDX	Differentials Kurtosis	5	-23.67%

The returns for correlation in the test set seem to be inconsistent with the backtest. One of the deciles is positive, while another is negative, and the last one is close to 0. This could entail that the backtest did not provide much if any information as to the direction of trade, which could imply it is difficult to predict correlations. It is necessary to consider the fact that backtest period was quite different from the test set period, due to the financial crisis.

The results for differential volatility were a bit more promising. Both deciles had positive spreads, although one of them did not yield much. At least the backtest got the direction of the trade right. It might be worthwhile to note that the deciles and ETF pairs selected were different. This is important, because of the intimate relation of volatility of positions in ETF pairs with correlation of those pairs. This maybe speaking to the fact of the volatility of the individual ETF or sums of volatilities of the individual ETFs having a more important role than correlation, which might be surprising considering the environment of the backtest. Of course, this was tested on just 3 ETF pairs, so it would be rash to generalize this result to other less correlated pairs.

Differential skewness spreads were far more exciting, as there were many deciles, all of which were traded in the test set, and their magnitudes were significant for the most part. The backtest had important information regarding the direction of the trade. It appears that the mid-deciles were the sweet spot, averaging around 24.09% for deciles 4-7. For the ETF pairs, GLD and GDX and GDX and SLV, this corresponded to a skewness that was slightly to moderately positive. It would seem that spreads were generated by being positive on average and taking chances on tail risk on the downside. This would have been significant given the environment of the backtest period, although nothing nearly as damaging happened in the test set period.

Differential kurtosis spreads were both positive and negative. It is difficult to say much about them, given the limited number of deciles in both the backtest and the test set, but they were large in magnitude. With the lower kurtosis deciles, extreme spreads were much more unlikely. It might be noteworthy to consider that for the ETF pair, GLD and GDX, low deciles for differential volatility and differential kurtosis were traded and yielded significant positive spreads in both the backtest and test set. Both measures are related, but not exactly, as volatility looked at the overall distribution of returns, whereas kurtosis more heavily weighed large differentials.

Now quarterly returns are examined. These are the results of the backtest:

ETF Pair	Strategy	Decile	Return Direction
GLD – SLV	Correlation	7	-
GLD – SLV	Correlation	9	+
GDX – SLV	Correlation	9	-
GDX – SLV	Correlation	10	-
GLD – GDX	Differential Volatility	1	+
GLD – SLV	Differential Volatility	2	+
GDX – SLV	Differential Volatility	2	-
GLD – GDX	Differential Skewness	5	+
GLD – SLV	Differential Skewness	7	+
GDX – SLV	Differential Skewness	3	-
GLD – GDX	Differential Kurtosis	3	+
GDX – SLV	Differential Kurtosis	1	-
GDX – SLV	Differential Kurtosis	2	-

These are the results of the test set:

ETF Pair	Strategy	Decile	Returns
GLD – SLV	Correlation	7	2.77%
GLD – SLV	Correlation	9	17.22%
GLD – GDX	Differential Volatility	1	61.78%
GLD – SLV	Differential Volatility	2	-3.09%
GDX – SLV	Differential Volatility	2	-11.98%
GLD – GDX	Differential Skewness	5	0.83%
GLD – SLV	Differential Skewness	7	-5.71%
GDX – SLV	Differential Skewness	3	-38.31%
GLD – GDX	Differential Kurtosis	3	29.5%
GDX – SLV	Differential Kurtosis	1	1.61%
GDX – SLV	Differential Kurtosis	2	12.12%

Correlation strategy yielded positive results, in that the backtest was informative about the direction of the trade on average. It was the higher deciles that entered the backtest and successfully to some extent traded on the test set. The deciles for correlation were smaller in range than monthly, which can be expected as more data was used to generate those deciles.

For differential volatility, lower deciles entered the backtest and traded on the test set, just as in the monthly case. Only the ETF pairs GLD and GDX traded successfully in the test set with an outsized return of 61.78%. The thinking of how correlation would be inversely related to volatility of trading the ETF pairs long short, might be roughly present here. For ETF ticker pairs, GLD and SLV, the seventh and ninth decile traded in the test set, while for the same ETF ticker pairs, differential volatility of the second decile traded. These are not completely inversely related, but there is a negative coefficient in front of the covariance/correlation in the volatility of the long short position. The not so perfect relation is captured by the negative spread in the test set for differential volatility as opposed to positive for correlation. The relation is a rough approximation for a couple of reasons: 1) the time periods corresponding to a decile for correlation probably weren't the same as that for a decile for volatility; and 2) individual volatilities of the ETF pairs can change also. Once again, the range for the buckets is smaller as compared to monthly.

The results for differential skewness were negative, although like the monthly case, mid-deciles were selected. The backtest was in general not informative of the direction of the trades. Interestingly, the deciles that traded both in monthly and quarterly test sets yielded vastly different results. That is not to say they traded the exact same periods, but it could mean that this signal can be very sensitive to timing. Skewness was negative for the ETF pairs GLD and GDX, but positive for the other two. The large negative return shown with ETF pairs GDX and SLV is an example of the tail risk of having positive skewness, whereas for GLD and GDX, the return was barely positive. Once again the range of the buckets for skewness was smaller for quarterly than monthly.

For differential kurtosis, the results on average were good. The backtest was indicative of the direction of the trades and the spreads were significant. Lower deciles were chosen, which makes intuitive sense, to minimize the risk of spreads vastly increasing in magnitude. It is interesting to see how low deciles for the ETF pairs GLD and GDX had large outperformance for both differential volatility and differential kurtosis, but this wasn't the case with GDX and SLV. The reason for this isn't exactly obvious. The range of the buckets for kurtosis was smaller for quarterly than monthly once again.

Next would be yearly frequency. There were some deciles in certain strategies that had met both the risk-adjusted return and return counts requirements for inclusion in the test set, but out of the 12 strategies, only 4 deciles did. Additionally, since there were only 4 data points in the test set, none of the strategies had their respective metrics fall within the corresponding decile and so no returns were generated. Consequently, there are no returns to report.

The final set of strategies overlays correlation and volatility, skewness, and kurtosis of the spread with a comparison of the decile of the volatility, skewness, and kurtosis of the individual ETF returns. The only condition for trading the pair is that there should be sufficient observations in the backtest, which is the same as stated before. I hypothesize the following and see what the backtest says to determine the direction of trade in the test set: go long the ETF pair that falls in a relatively lower bucket and short the other pair. For example, if we are considering the volatility of the spread between two ETFs, say GLD and GDX, I would go long GLD if the decile of the kurtosis of its returns was less than the decile of the kurtosis of GDX's returns in the backtest. If they both had the same decile, I didn't trade them. If this generates a positive return in the backtest, then in the test set I maintain the direction of the trade, otherwise I reverse it. I report the results of the backtest and test set in the spreadsheet.

As seen earlier, the deciles included in the backtest tended to be higher for correlation, lower for volatility of the spread, mid to lower for skewness of the spread, and lower for kurtosis of the spread. With this backdrop, the rationale for going long the pair with the lower bucket should be examined. For volatility, going long the ETF that has relatively lower volatility relative to its distribution can mean that there are less chances for outsized movements, compared to its counterpart's relative distribution. This does not by any means say that the overall volatility of the ETF that you are long is less than the ETF you are short. Going long the ETF that has relatively lower skewness relative to its distribution can mean that the bulk of the returns would be above its mean, which is presumed to be positive on average, and take chances on large losses, relative to its distribution. Once again, this does not mean that the overall skewness of the ETF that you are long is less than the ETF you are short. Having a long position on lower kurtosis relative to the ETF's distribution can mean that the bulk of the returns are concentrated around its mean, which again is presumed to be positive, giving less chances for large losses or gains. Of course, this does not mean that the overall kurtosis of the ETF you are long is less than the ETF you are short.

The number of strategies is considerable. First, there are 4 statistics that look at ETF pairs together: correlation, volatility, skewness, and kurtosis. For each of those statistics, I compute volatility, skewness, and kurtosis of each ETF in the pair and I compare what bucket they are in, for a total of 3 strategies for each of the 4 statistics. This yields 12 strategies, but there are 3 frequencies, which yields a total of 36 strategies. On top of this, the number of deciles per each strategy is variable with some not having any, while others having multiple, up to 10, which would be unlikely. The test set tests those deciles, but only if the relevant statistic, which could be correlation, or volatility, skewness, and kurtosis of the spread, falls into the relevant decile. Due to the variable nature of the number of tests performed, I have only mentioned the average of all deciles within a particular ETF long-short pair for each strategy in the spreadsheet.

In looking at the spreadsheet, 57 out of the 108 strategies had positive returns in the test set. However, not all the strategies were traded, specifically those that had matched on the deciles or for certain yearly frequencies where there wasn't a trigger to make a move. Eliminating those yielded 95 strategies. To see if the backtest has informational value, I tested the probability of it choosing the right direction with 50% probability, although it was observed to be right 60% of the time. The question arises of statistical significance, and it is seen from the analysis that the p-value is 6.42%, making it significant at the 10% level. However, it would be rash to jump to conclusions so quickly, before having a more granular view.

After looking at how the strategies performed at each frequency and observing statistical significance, one may have reservations. Here is a table showing the amount of strategies that had positive and negative returns for each frequency:

	Positive Returns	Negative Returns
Month	21	15
Quarter	21	15
Year	15	8

For both monthly and quarterly returns, positive returns occurred 58.33% of the time, whereas for yearly returns, positive returns occurred 65.22% of the time. I evaluate the statistical significance of both the direction of the returns and the difference between the magnitude of positive and negative returns. It is important to be right more times than not, but also you want to be more right than wrong on average. This table provides the statistical significance of the different frequencies:

	Directional P-Value	Magnitude P-Value
Month	0.405	0.832
Quarter	0.405	0.710
Year	0.210	0.057

From this table, it appears that the backtest does not hold significant information regarding the direction of trades for any frequency. This does change the initial perspective that the backtest does hold significant information, most likely due to the decrease in the sample size, as the success ratio does not appear to be too far off for directional bets. The other way to profit is to be more right when you are right than wrong when you are wrong, which essentially is testing the difference in magnitude of returns of positive and negative returns. As can be seen, only the yearly strategies have a significant p-value at the 10% level for differences in magnitude between positive and negative returns. Given the limitations of that frequency (mostly limited data), the level of confidence placed on trading yearly should not be very high. It can be concluded that the signals generated on average were not significant, although individual signals may be.

I have researched whether statistical factors have predictive value over related ETF pairs. The results are mixed on the surface, but a more thorough analysis that might question the economic and fundamental rationale of these strategies would not place great confidence in them. An important thing to note is that the backtest period was quite different from the test set, as it included the financial crisis. It might be more insightful to include variables that might be indicative of how similar the environment you are testing in is compared to periods in your backtest. I am not of the view that statistical factors do not have value, but possibly further research can be done to see where that value can be unlocked. It might be worthwhile seeing how these results fair with less related ETFs and also to see if these statistics can be modeled.