# The Financial News impact on Commodities

Hunter Douglas

8/08/2022

## Introduction:

The domain of these data sets is related to finance. The financial news data is from Kaggle https://www.kaggle.com/datasets/heeraldedhia/us-economic-news-articles and uses the Wall Street Journal as its source. The asset price returns data is from Investing.com. There is no data dictionary for either data set. The intended demographic for this data is anyone interested in investing in commodities such as gold, silver and oil. Anyone who can set up an investment account can benefit from this data as gold has been widely known as a safe haven asset when the market is going down or bearish. Both data sets needed some simple cleaning that I used excel for. This involved removing unnecessary data from the financial news data such as ID numbers. For the Commodities I removed the open, close, high, and low prices while just keeping the returns and added the returns of each asset together to make one data set. I also added in the S&P 500 index as a market benchmark and the US Dollar Index(US Dollar value vs six foreign currencies) to show the value of the Dollar during this time.

## Data Goals:

1. Understanding the overall risk of each asset. For this we can use standard deviation as the risk measure. I also want to understand the risk and reward by looking at overall returns as well.
2. See what words are most commonly used together throughout the entire financial news data set. I am using K-means clustering with 5 clusters and choosing the top 5 words from each cluster to begin with to compare against the Alpha(Asset return over the market(S&P500)) of each asset.
3. Take a more in depth look at the 2008 financial crash by creating clusters for this time period.
4. Look at future years by using K-means clustering for each year to see if we can learn anything from this and use it in the future.
5. See what was the best commodity to invest in during the financial crisis.
6. See what keywords are associated with the time period when the market recovered after the financial crash.
7. Looking at the frequency of asset returns as a second way of understanding their risk.

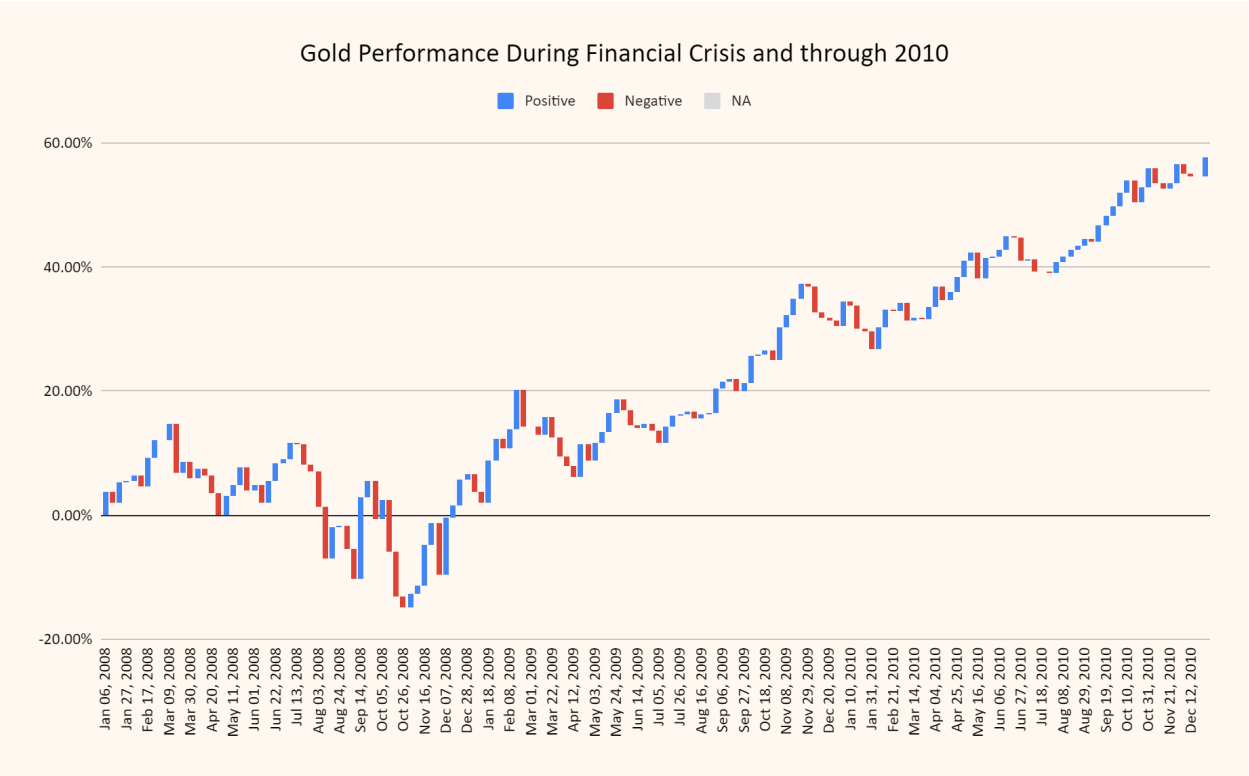
## Data Analysis:

After an extensive data analysis I managed to complete all my goals with this data. I took the sum of all assets for 20 years, the 10 year period for my financial news data, the financial crisis of 2008, and 2009-2010 when the market started to recover. This tells us the overall return for these time periods and we can see from 2000 until the end of 2019 that Gold is the top performer at 224.83% gain. I used standard deviation as my risk measure during each of these time periods as well. This can tell us the risk associated with each asset and we can see that

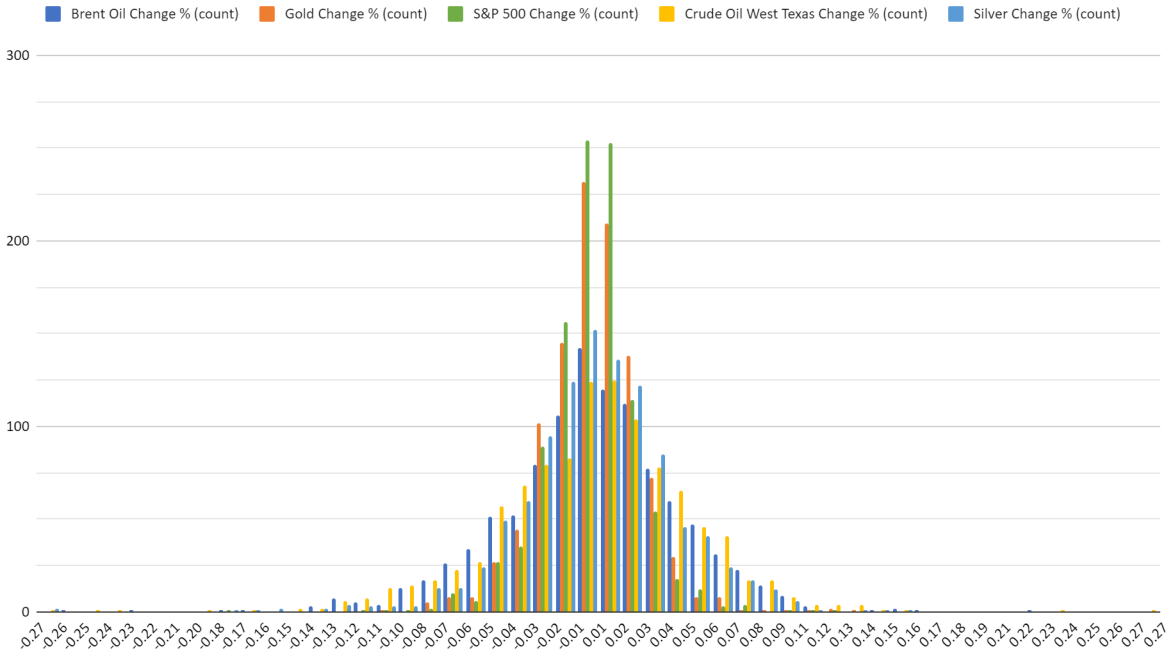
Gold also had a relatively low standard deviation over 20 years at just 2.46%, only 0.08% over the S&P 500.

I also calculated the alpha for each of these time periods. Alpha is the asset returns over a benchmark, in which I will be using the S&P 500 as the benchmark and overall view of the market. Gold was the best performing commodity during 2008 as well being the only commodity with a positive gain at 6.60% and even lower standard deviation than the S&P 500 at 4.43%. Asset returns during the financial crisis had a large spike in both directions right as the banks started failing later in the year around September of 2008 and immediately after in the beginning of 2009. What sticks out here is the over 20% spike from Crude Oil West Texas indicating high demand for oil and possibly a significant amount of investors who went to oil after the market crash. The Frequency of returns throughout the entire data set can show us the S&P 500 having the largest amount of low returns followed by Gold which can be associated with the low risk of each asset. Brent Oil and Crude Oil West Texas however, have a more significant amount of returns in both directions which shows more risk but also larger gains for these 2 assets.

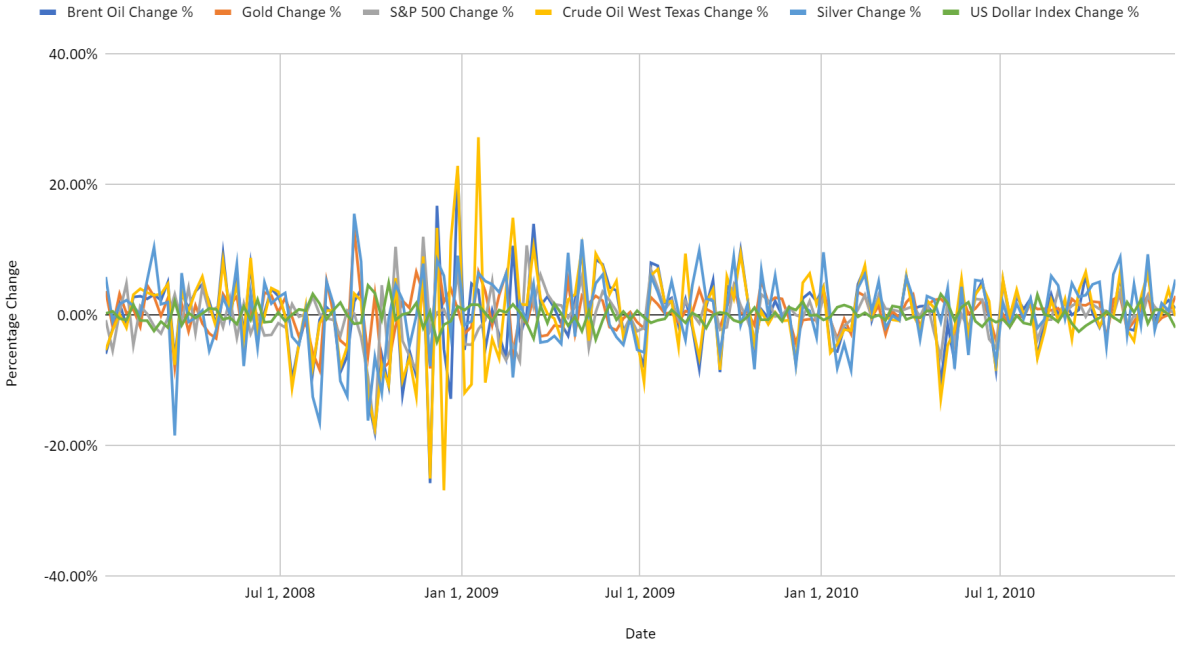
## **Excel Data Mining:**



## Frequency of Returns throughout entire data set




## Asset Return Behavior during Financial Crisis and through 2010



## K-means Clustering Assessment:

For K-Means clustering, since the financial news data is only a 10 year period, I used this algorithm during 2004-2014. 2008, 2009-2010, 2011, and 2012. The only new time periods here are 2011 and 2012. This is to simply do a deeper data analysis and give us a better understanding of what keywords are associated with asset returns so maybe we can pay more attention to these words in the future. The operators I used were Select Attributes, Filter Example, Process Documents from Data, Tokenize, Transform Cases, Filter Stopwords, Filter Tokens, K-Means Clustering, and Performance. Looking at the entire data set from 2004-2014 we can already see some keywords that stand out based on previous data analysis. In the First cluster, we see gold and crude. Gold had the second best overall performance at 123.30% during this time period right behind Silver at 159.16%. Crude Oil had the lowest performance here but the second largest overall return during 2009-2010 at 84.74%. The other clusters seem to be more geared to the overall economy.

During the financial crisis we immediately see banks and debt together in cluster one which can be associated with the significance of the banks failing in late 2008. In the second cluster we see some keywords, dollar, inflation, rate, growth, euro, unemployment, and rates. With this information we can also see the US Dollar index still managed to rise meaning the value of the US Dollar remained higher than other nation currencies. As the market starts recovering in 2009 and 2010, we see Gold in cluster 2 followed by Oil within the same cluster several rows down. Then looking at overall returns and alpha, Gold and Oil both performed well and significantly outperformed the market. Looking at cluster 4 we see percent, nasdaq, profit, consumers, and gain which could be a good sign for the market and the S&P 500. Using this past Cluster information, we might assume that Oil and Gold will continue to be good investments and the market may soon start seeing large gains. Looking first at the 2011 alpha we see Brent Oil had the largest Alpha at 14.65% even as the S&P 500 saw a 2.13% gain. Gold also had a 10.92% gain beating the market as well. Now looking at 2011 clusters we see in cluster one, quarter, sales, earnings, company profit, growth and revenue. We also see Gold and Oil in cluster two but below dollar and inflation. Cluster one seems to show us some significance in potential positive market performance. The following year in 2012, we noticed that not a single commodity outperformed the market, meaning cluster one might have helped us allocate more of our investment portfolio to the S&P 500 and out of Gold and Oil.

## Conclusion:

It seems that only a few words within 2 out of 4 clusters have any significance in relation to the commodity returns. There is also a possibility of confirmation bias and look ahead bias at play here. A more extensive walk forward analysis with more financial news data could give us better insight to what assets might be the better investment for the following year. Overall I believe this data analysis gives us a better understanding of what we could possibly pay attention to when reading the wall street journal or any other financial news source.