* assetCode(object) - a unique id of an asset. *This is the ticker symbol as traded on the exchange at the given time. It is possible, though not very common, that a company will change its ticker symbol over time.*
* assetName(category) - the name that corresponds to a group of assetCodes. These may be "Unknown" if the corresponding assetCode does not have any rows in the news data. *This is the name of the company. It's interesting that they say "the name that corresponds to a group of assetCodes" which implies that this column is a permanent identifier. That's not true in the real world as companies can and do change their names (though also not often).*
* universe(float64) - a boolean indicating whether or not the instrument on that day will be included in scoring. This value is not provided outside of the training data time period. The trading universe on a given date is the set of instruments that are avilable for trading (the scoring function will not consider instruments that are not in the trading universe). The trading universe changes daily. *It is a float64 and not an int or boolean to, I presume, account for NaNs. This is the selection of companies out of the entire set for a day that you are supposed to predict against. As discussed in a couple other posts, it looks like the included names are the names with generally the highest dollar trading volume.*
* volume(float64) - trading volume in shares for the day. *For the given day and stock, this is the number of total shares which traded on the exchange.*
* close(float64) - the close price for the day (not adjusted for splits or dividends). *This is essentially, the last traded price at 4pm US/Eastern, which is when US exchanges shut for the day. The values are not necessarily comparable for the same stock over time because they are unadjusted for splits (when a company issues new shares) or dividends (when a company makes a cash distribution). You can infer the split and/or dividend impact per day by comparing close/close.shift(1) -1 toreturnsClosePrevRaw1. This may seem trivial, but it can be important. Sometimes exchanges close on the half day (rarely) which may affect you if you are trying to sync up the news timestamps.*
* open(float64) - the open price for the day (not adjusted for splits or dividends). *This is essentially, the first traded price of the day at 9:30am US/Eastern. Also note not adjusted for splits. Also appears there are more bad data items in this field than in others.*
* returnsClosePrevRaw1(float64) - see returns explanation above. *This is the 1-day close-to-close return adjusted for splits and dividends; if there are no splits/divs then it is just close/close.shift(1)-1.*
* returnsOpenPrevRaw1(float64) - see returns explanation above. *This is the 1-day open-to-open return adj for splits and divs; if there are no splits/divs then it is just open/open.shift(1)-1.*
* returnsClosePrevMktres1(float64) - see returns explanation above. *The description of this is a little unclear but we can infer. Essentially, this is the component of the return (close-to-close) after stripping out the return of the overall market (think S&P500) for the day. So if you are looking at AAPL and it went up 2% on the day, and the S&P500 went up 1.5%, this number is 0.50%. This might be more complicated, but the intuition here is correct. The res is short for residualized; so think of a linear regression; you take out the trend and are left with the residuals. You might expect that a news article is specific to a particular stock; to test/measure that you would likely want to look at residualized fields because they have the portion of the return which is stock-specific.*
* returnsOpenPrevMktres1(float64) - see returns explanation above. *Like the previous; residualized open-to-open return.*
* returnsClosePrevRaw10(float64) - see returns explanation above. *Essentially close/close.shift(10) -1 and adjusted for splits/divs.; should be equal to the trailing returnsClosePrevRaw1 compounded for 10 days.*
* returnsOpenPrevRaw10(float64) - see returns explanation above. *Essentially open/open.shift(10) -1 and adj for splits/divs.*
* etc. on the other return variables.

Information on the Training Data

* There are no Unknown assetName in news\_train\_df, but there are 24 479 rows with Unknown as the assetNamein market\_train\_df. Merging by assetCode leaves out Unknown rows, which could be problematic.
* Volume has the highest correlation in terms of returnsOpenNextMktres10.
* Merging by just assetCodes greatly increases the dataframe (with just 100k rows, it has turned into 10 million rows), although merging by assetCodes and time greatly decrease the original dataframe.