

Data columns explanation

Column	Description
Company	The company to which the row belongs to
Date	The date on which the filing or press release was published
Adj Close	Adjusted Close indicates the stock closing price for that given day.
Volume	The number of stock shares that were traded on this day
Press release	The CEO quotes that were scraped for a given company and a given date
Price change 12mo	This column presents the adjusted close price change. For example if the stock price in the row is 100 and the price after 12 months is 150, then the 12 month price change for this row will be 50.
Price change 8mo	Simillary, this price change indicates the same but over a time span of 8 months instead of 12.
Price change 4mo	This price change indicates the change over a time span of 4 months.
Price change pct 12mo	The change in percentage of price change over a 12 month time period
Price change pct 8mo	The change in percentage of price change over a 8 month time period
Price change pct 4mo	The change in percentage of price change over a 4 month time period
Price change sign 12mo	This column contains either a value of -1 for when the price change 12mo is negative or a 1 for when the price change 12mo is positive.
Price change sign 8mo	This column contains either a value of -1 for when the price change 8mo is negative or a 1 for when the price change 8mo is positive.
Price change sign 4mo	This column contains either a value of -1 for when the price change 4mo is negative or a 1 for when the price change 4mo is positive.
Compound	The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive). This score has been calculated with the vader sentiment analysis. Vader is an abbreviation for Valence Aware Dictionary and Sentiment Reasoner. It is a lexicon and rule-based feeling analysis instrument. With a mix of highlighted tokens VADER marks a score to mark each token as either positive

	and negative with a score between -1 and 1. The next three columns are derived from this.
Negative	A negative score is a score that has a compound score lower than -0.05
Neutral	A neutral score is a score that has a compound score between -0.05 and 0.05
Positive	A positive score is a score that has a compound score high than 0.05
Polarity	The polarity score is a float between 1 and -1. This score indicates a negative sentiment for -1 and a positive sentiment for 1. The library textblob provides a simple API to conduct polarity and subjectivity scores.
Subjectivity	The subjectivity score is also a float between -1 and 1. According to the textblob library it refers to the general emotion or opinion.
Text length	Text length indicates the length of the text that was scraped.
Word count	The word count is the amount of words that occur for the given text. More specifically, it is the text length split on whitespace.
Word density	The word density is a value that has been calculated by dividing the text length by the word count.
Punctuation count	This column represents the amount of punctuation that occurs per given text in the press release column.
Upper case count	The amount of upper cases that occur
Stopword count	The amount of stop words that occur. The stop words that were taken into account are the stop words that occur in the spaCy library as stopwords.
Readability dale chall	The readability score is calculated with the Dale Chall readability metric. This metric can be found in the readability library [27]. The Dale Chall score is derived from a formula that is based on the use of familiar words, rather than syllable or letter counts. It indicates how easy it is to read a certain text based upon the familiarity of the words that occur in said text
Readability flesch reading ease	This readability score is based on the Flesch Reading Ease. It is a standard test of readability used for the U.S. Department of Defense and insurance policies.
Noun count	The amount of nouns that occur in the relevant text. The library TextBlob was used to identify each Part of Speech.
Verb count	The amount of verbs that occur in the relevant text.

Adj count	The amount of adjectives that occur in the relevant text.
Adv count	The amount of adverbs that occur in the relevant text.
Pron count	The amount of pronouns that occur in the relevant text.
Top words	Top words is a list of tuples. Each tuple consists of a word and the amount of occurrences for that word in the press release text.
Word calculation	The word calculation is calculated by calculating for each word how likely it is to appear in a text that belongs to a company that has a rising stock price after a year. Then this value per word is multiplied by each other.
'Continued', 'record', 'cash', 'results', 'customers', 'fiscal', 'billion', 'cloud', 'growth', 'strong', 'fourth', 'performance', 'quarter', 'business', 'sales'	Every word is a separate column and counts the frequency of occurrences of that specific word in the press release instance.