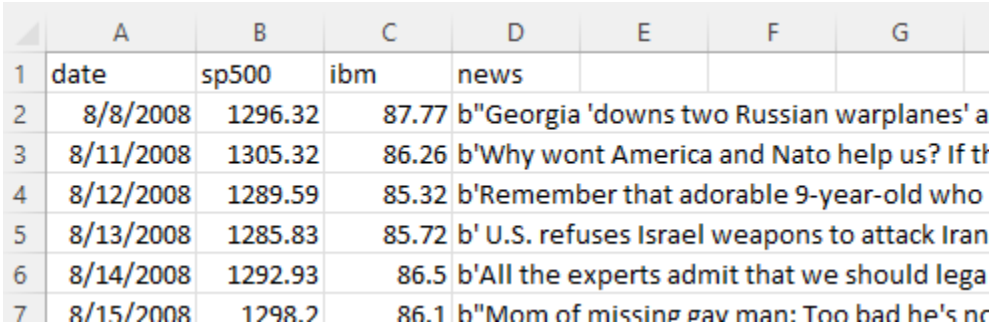
E213 Final Assignment (linear regression and NLP)

Data file: stockdailyhlnews.csv



	A	B	C	D	E	F	G
1	date	sp500	ibm	news			
2	8/8/2008	1296.32	87.77	b"Georgia 'downs two Russian warplanes' a			
3	8/11/2008	1305.32	86.26	b'Why wont America and Nato help us? If th			
4	8/12/2008	1289.59	85.32	b'Remember that adorable 9-year-old who			
5	8/13/2008	1285.83	85.72	b' U.S. refuses Israel weapons to attack Iran			
6	8/14/2008	1292.93	86.5	b'All the experts admit that we should lega			
7	8/15/2008	1298.2	86.1	b"Mom of missing gay man: Too bad he's no			

Columns:

Date: date from 8/8/2008 to 7/1/2016

Sp500: S&P 500 daily index

Ibm: daily close price of IBM stock

News: daily headline news from multiple news sources

Your Tasks:

1) Load the data from the csv file into a Python Pandas data frame named df
    a. Initially, df should have 1989 rows and 4 columns
    b. Add a column called sscore to df
        i. Fill the sscore column with the 'compound' sentiment analysis score based on the daily headline news for each day.
        ii. Calculate the average (mean) 'compound' score for the column sscore and store this average number in a variable named avgsscore.
2) Is the IBM stock price influenced by the sentiment compound score and/or s&p 500 index?
    a. Use *from statsmodels.formula.api import ols* for this linear regression task
    b. Store adjusted rsquaures in a variable named adj_rsquared
    c. Store pvalue of f-statistics in a variable named f_pvalue
    d. Store pvalue of sscore in a variable named sscore_pvalue
    e. Store pvalue of sp500 in a variable named sp500_pvalue
    f. If a relationship exists between sscore and ibm stock price, then store a boolean value of True in a variable named sscore_rel; otherwise, sscore_rel should be set to False
    g. If a relationship exists between s&p 500 index and ibm stock price, then store a boolean value of True in a variable named sp500_rel; otherwise, sp500_rel should be set to False