Final Assignment (linear regression, classification, and NLP)

Data file: stockdailyhlnews.csv

date	weekday	president	sp500	ibm	news
8/8/2008	Friday	republican	1296.32	87.77	b"Georgia
8/11/2008	Monday	republican	1305.32	86.26	b'Why wo
8/12/2008	Tuesday	republican	1289.59	85.32	b'Rememb
8/13/2008	Wednesda	republican	1285.83	85.72	b' U.S. refu
8/14/2008	Thursday	republican	1292.93	86.50	b'All the e

Columns:

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

Weekday: day of the week (Monday, Tuesday, Wednesday, Thursday, Friday)

President: Which political party is in the White House? Republican or democrat

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 6 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) Converts weekday and the president columns to dummy variables
    a. Add the dummy variables (columns) to the original data frame df
3) 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
4) Can we predict whether Republican or Democrat will be in the White House based on s&p 500 index, ibm stock price, and sscore?
    a. Use sklearn for this classification problem:
        i. from sklearn.model_selection import train_test_split
        ii. from sklearn.linear_model import LogisticRegression

iii. split the data into training and test datasets using 80% training & 20% test and a random seed value of 10
iv. store the logistic model score in a variable named logmodel_score