Team 3 - Stock Price Forecasting

Data Sourcing

Dataset 1: Stock prices dataset from Quotemedia

This dataset contains the prices of and volume of stocks sold for each day since 1967. Within this dataset, we are interested in the adjusted closing prices of the companies: Twitter, Google, and maybe another one for the year range 2018-2023. The source of the dataset is the *quandl* library in Python.

Dataset 2: S&P 500 tickers

This dataset has three columns: company, symbol, and weight of the companies making up the S&P 500. The symbol is the ticker (for example, AAPL for Apple).

Dataset 3: Fama-French 5 Factor dataset

This dataset contains five factors for every day since 1963. The five factors are 1) market factor (MKT) which is excess market return; 2) size factor (SMB), the excess return of small versus large companies; and 3) value factor (HML), the excess return of high book/market versus low book/market companies (value stocks versus growth stocks); 4) profitability factor (RMW), the difference between returns of firms with high and low profitability; and 5) investment factor (CMA), the difference in returns between high and low investment firms. This dataset captures can be used to assess the exposure of each stock to the Fama-French risk premium.

Dataset 4: Headlines and sentiment scores

We use GNews, a library in Python, to conduct web scraping of Google News headlines. There is a dataset of headlines for each company. We scrape 100 headlines for each 2 week period across 2018 to 2023. There is some small preprocessing with this to address duplicates, headlines that don't contain the company name, and headlines outside of the data range.

Then, we calculate the sentiment of each headline using the Sentiment Intensity Analyzer from the NLTK library. This function returns a score from 0-1 for the metrics compound, negative, neutral, and positive. For each day, we average the score of each of these 5 metrics and also count the total number of headlines that day.

✓ Data Wrangling 2018 - 2023

Dataset 1: Stock prices dataset from Quotemedia

- Rows: filter for 2018-2023 and companies Google, Amazon, Apple, Nvidia, and Microsoft
- Columns: select adjusted closing price

Dataset 3: Fama-French 5 Factor dataset

- Rows: filter for 2018-2023
- Columns: all

Dataset 4: Headlines and sentiment scores

- Rows:
    - Filter for 2018-2023 and companies Google, Amazon, Apple, Nvidia, and Microsoft
    - Take the average sentiment scores for each day
    - Count the total volume of news headlines each day
    - For days with missing headlines, impute with pos = 0, neu = 1, neg = 0, compound = 0, and volume = 0.
- Columns:
    - pos, neu, neg, compound - sentiment scores
    - volume - number of headlines

Exploratory Data Analysis

1. Selecting columns of interest and target feature(s)

    a. Which columns in your data sets will help you answer the questions posed by your problem statement?

    Since we are using LSTM, the covariates will be themselves time series. To predict the stock price of today, we use the stock prices of previous days. The number of previous days' stock prices will depend on the window length we use. Presumably, some sort of cross validation will be used to determine the window length.

    Also, we will use sentiment scores from news headlines. Similarly, the stock price of today will be predicted with the sentiment scores of previous days.

    b. Which columns represent the key pieces of information you want to examine (i.e. your target variables)?

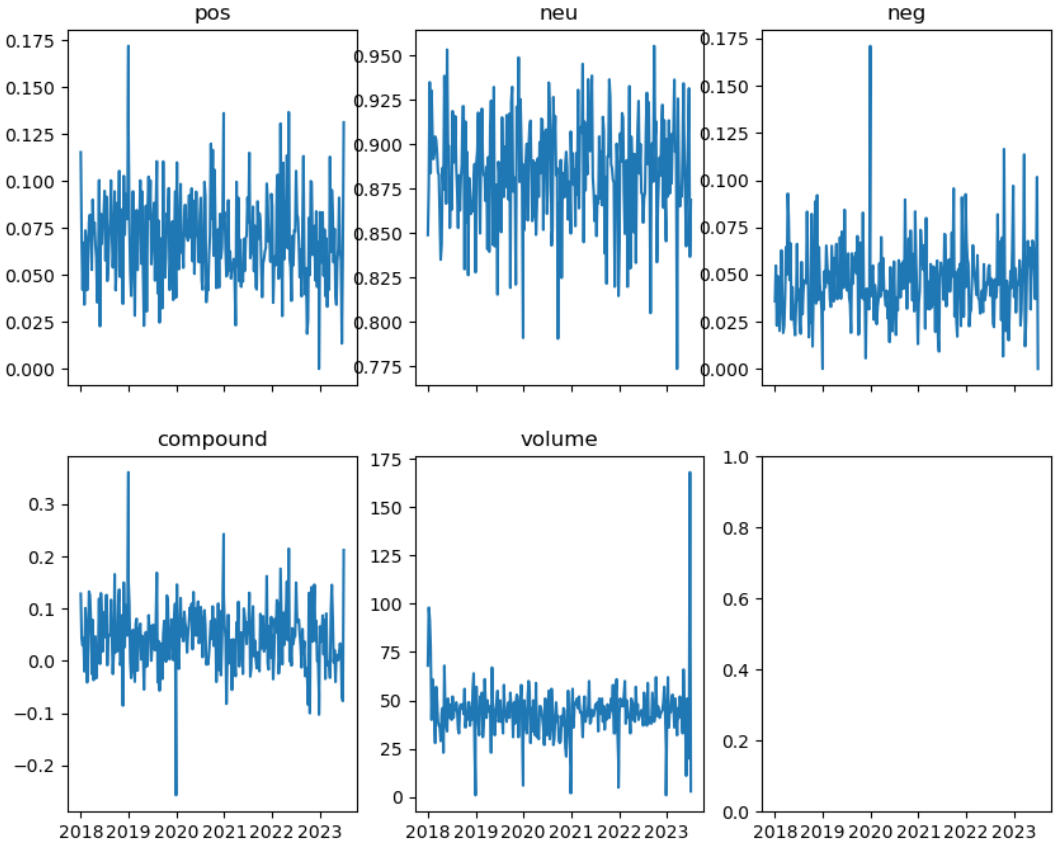
    Target variable is stock price.

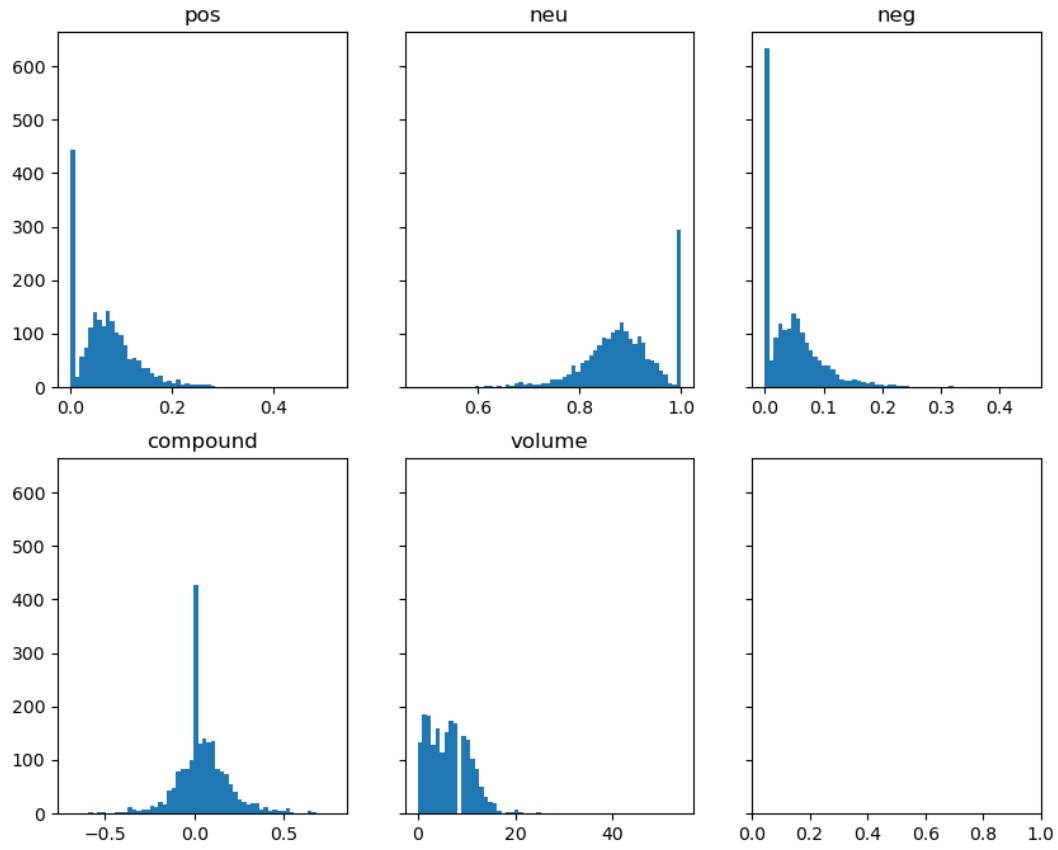
    c. How many numerical, textual, datetime etc. columns are in your dataset?

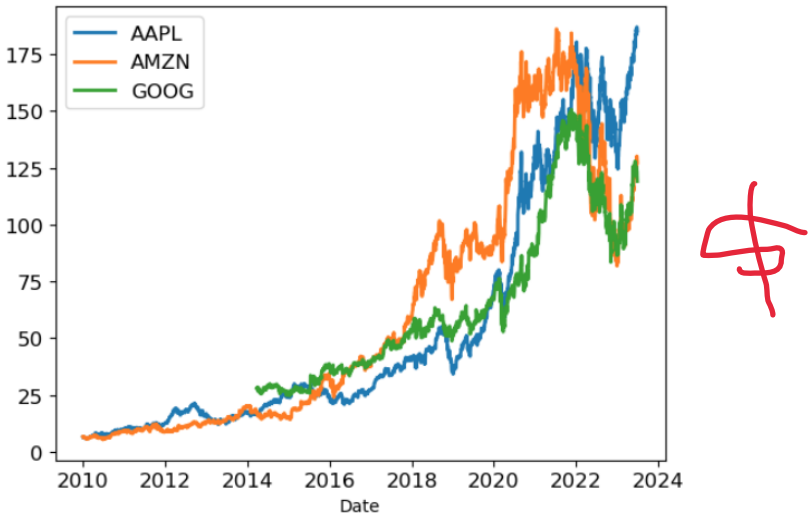
    Numerical columns: stock prices, sentiment scores, volume of headlines
    Datetime column
    Text columns: name of company (Apple, Amazon, Google, Nvidia, Microsoft)

2. Explore Individual columns for preliminary insights

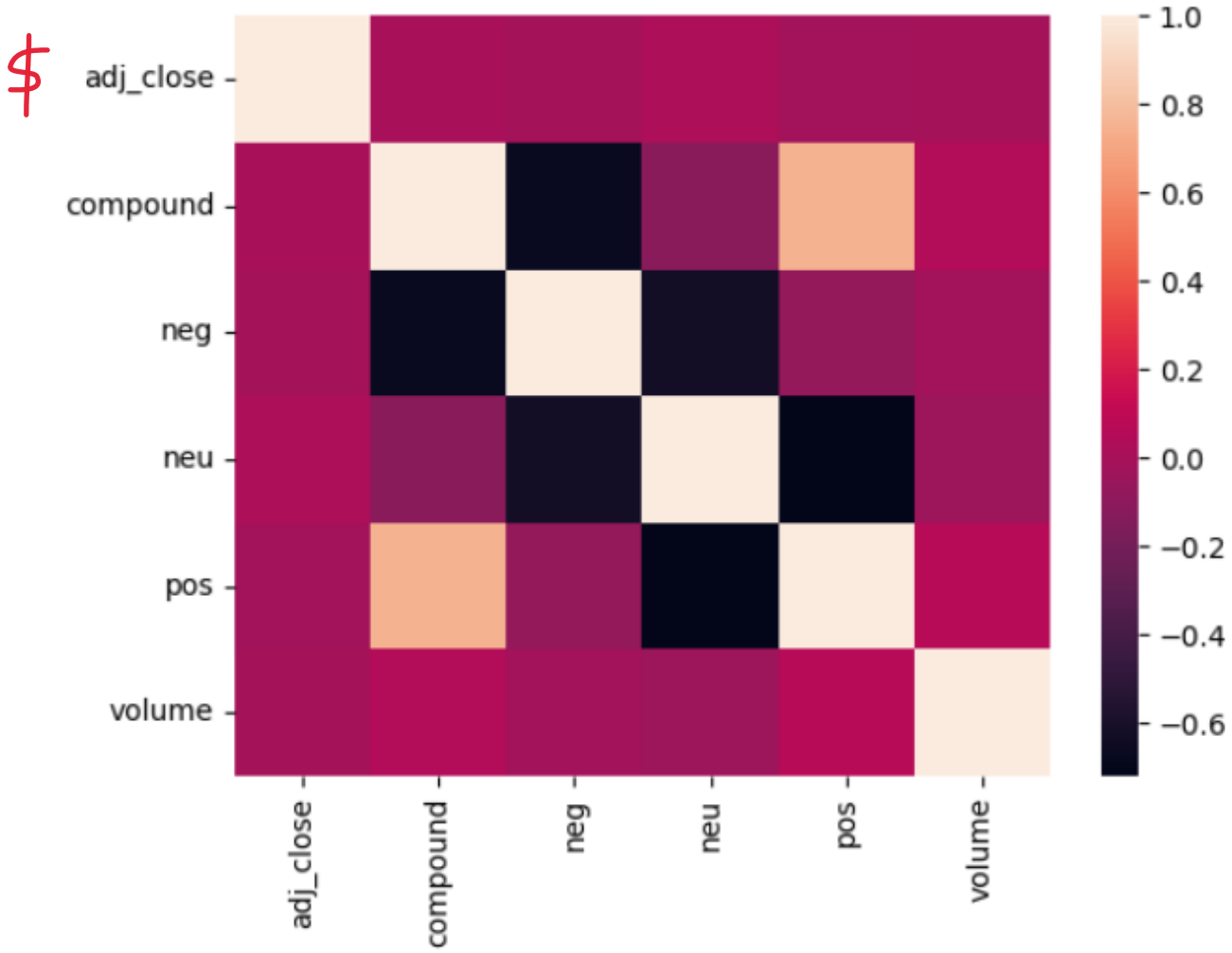
    We only have 5 years of Google headline data because we haven't retrieved the headlines for the other companies.


Google sentiment scores and volume aggregated by week
pos
neu
neg
compound
volume
2018 2019 2020 2021 2022 2023

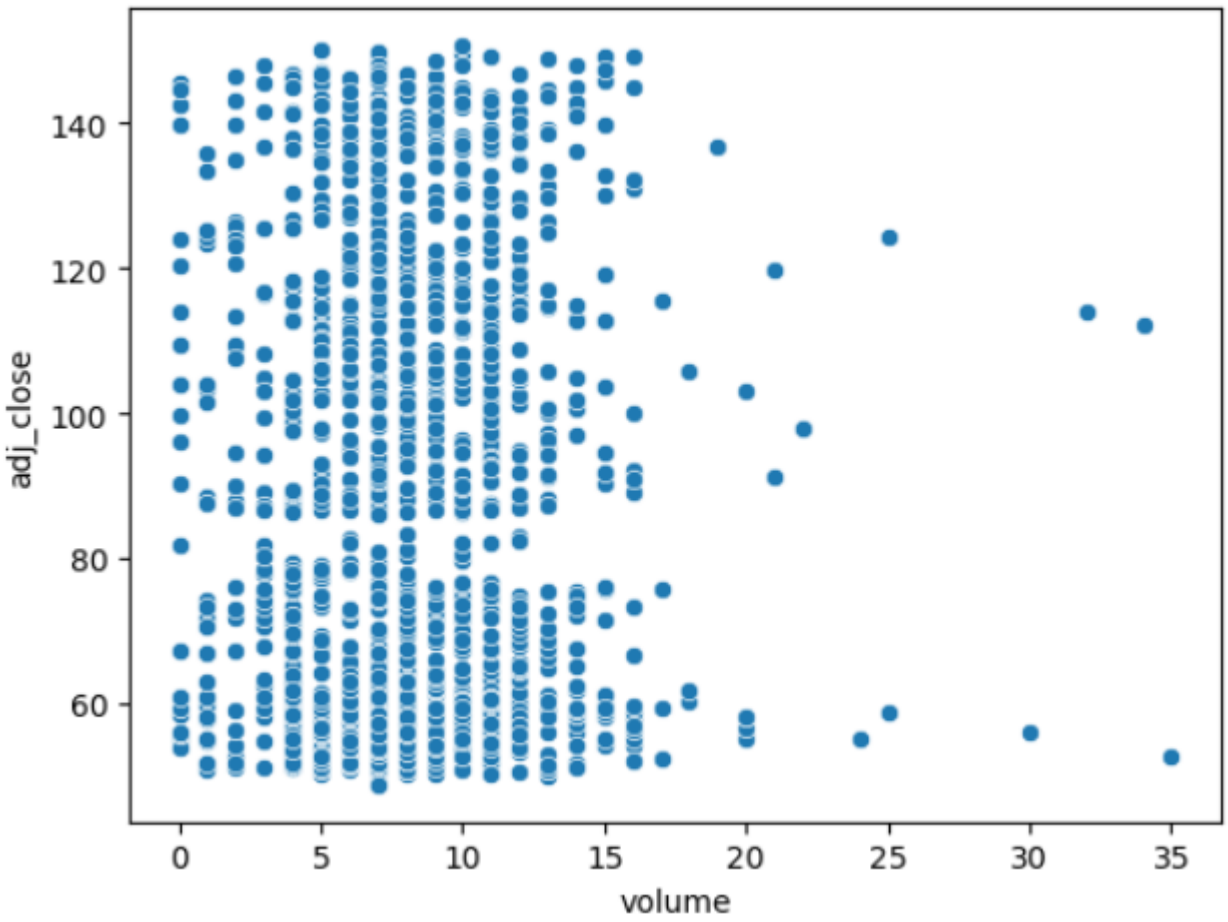


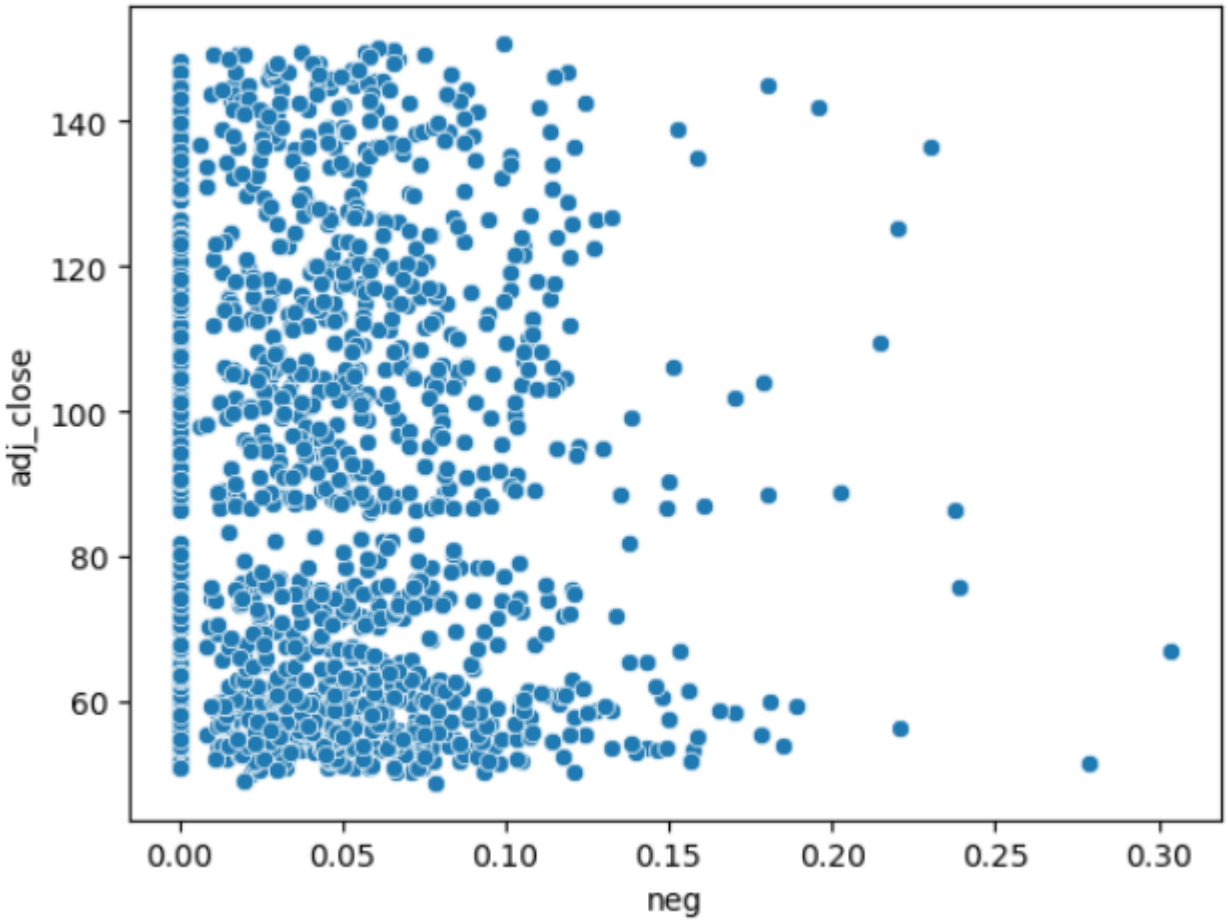


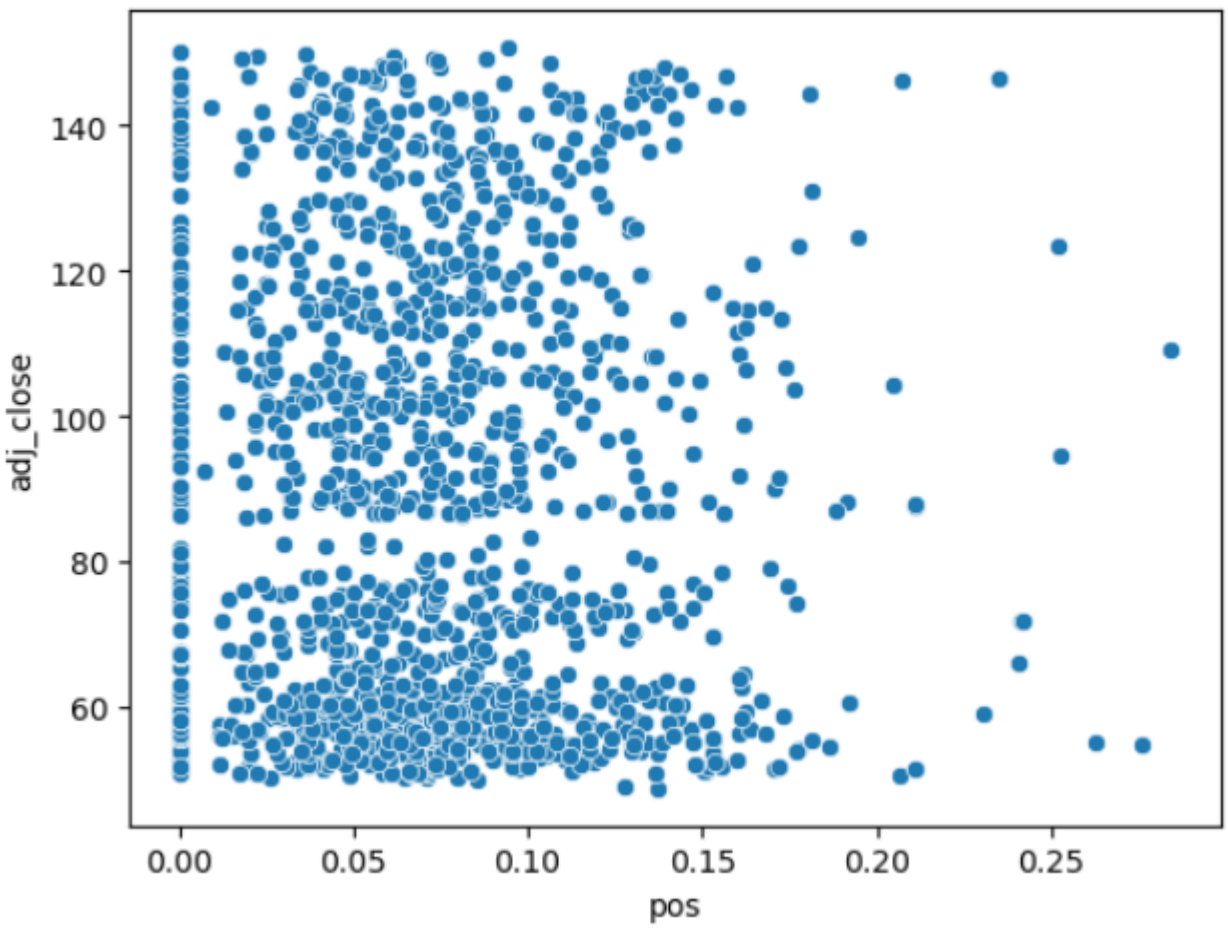
Correlation plots and scatter plots of the adjusted closing price (target variable) and the sentiment analysis variables. The correlation is not too strong.



	adj_close	compound	neg	neu	pos	volume
adj_close	1.00	0.00	-0.01	0.02	-0.02	-0.01
compound	0.00	1.00	-0.68	-0.11	0.75	0.05
neg	-0.01	-0.68	1.00	-0.63	-0.08	-0.02
neu	0.02	-0.11	-0.63	1.00	-0.72	-0.04
pos	-0.02	0.75	-0.08	-0.72	1.00	0.06
volume	-0.01	0.05	-0.02	-0.04	0.06	1.00


adj_close
140
120
100
80
60
0
5
10
15
20
25
30
35
volume


adj_close
140
120
100
80
60
0.00
0.05
0.10
0.15
0.20
0.25
0.30
neg


adj_close
140
120
100
80
60
0.00
0.05
0.10
0.15
0.20
0.25
pos