News Article Data

january_news.json

Data is retrieved from the Alpaca News API for each day between January 6 and January 31, 2025, specifically for Nvidia. For every day in that range, an API request is made to fetch the news data. For each news article in the daily results, only the relevant fields—the publication timestamp, the headline, and the summary—are kept. These cleaned news items are aggregated into a single JSON object under the key “news” and saved as january_news.json.

Unit of Observation

Each record in the "news" list represents a single news article related to Nvidia stock. Every article includes a timestamp, headline, and summary. This dataset contains only articles published between January 6 and January 31.

Variables

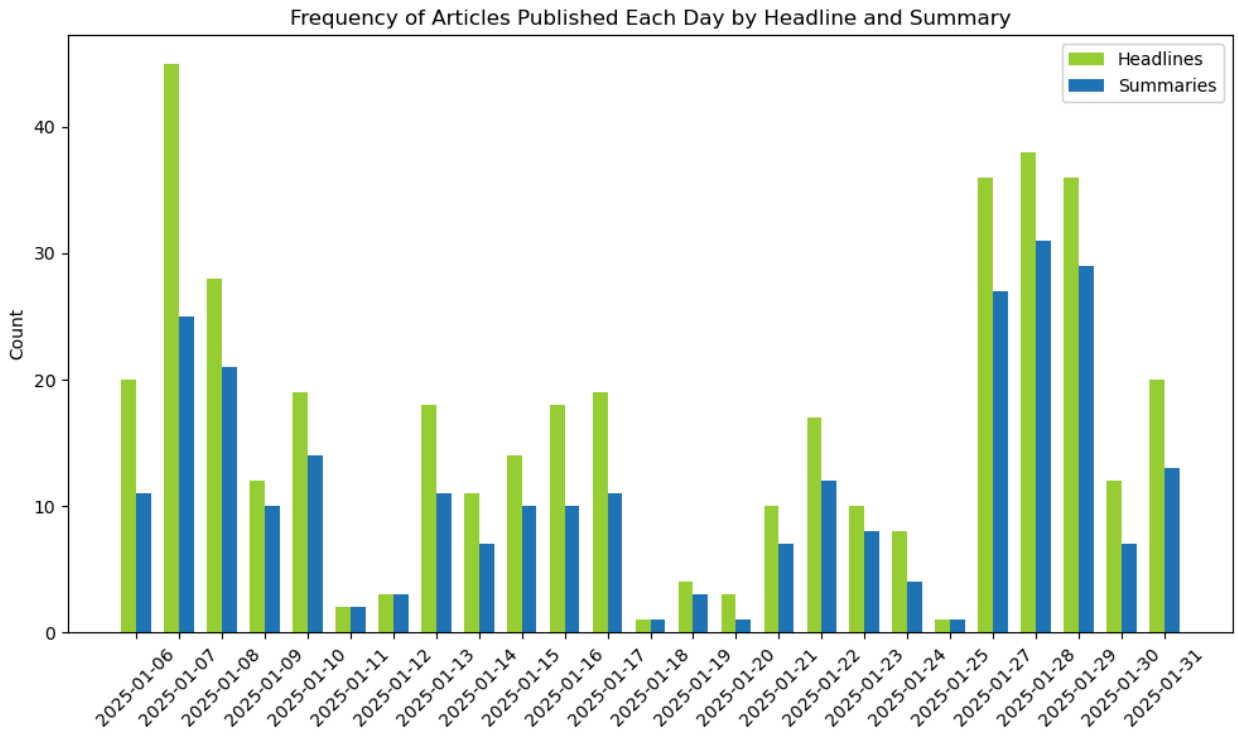
1. created_at
    a. The date and time when the news article was published (UTC timezone, ISO 8601 format)
    b. Data type: String
    c. Missing values: 0 (405(0))
2. headline
    a. The title of the news article
    b. Data type: String
    c. Missing values: 0 (405(0))
3. summary
    a. A brief summary of the news article content
    b. Data Type: String
    c. Missing values: 126 (405(126))

Frequency Table

Date Published	Headline Count	Summary Count
2025-01-06	20	11
2025-01-07	45	25
2025-01-08	28	21
2025-01-09	12	10

2025-01-10	19	14
2025-01-11	2	2
2025-01-12	3	3
2025-01-13	18	11
2025-01-14	11	7
2025-01-15	14	10
2025-01-16	18	10
2025-01-17	19	11
2025-01-18	1	1
2025-01-19	4	3
2025-01-20	3	1
2025-01-21	10	7
2025-01-22	17	12
2025-01-23	10	8
2025-01-24	8	4
2025-01-25	1	1
2025-01-27	36	27
2025-01-28	38	31
2025-01-29	36	29
2025-01-30	12	7
2025-01-31	20	13

Frequency Distribution



news_data.cvs

Data is retrieved from the Alpaca News API for each day between January 6 and January 31, 2025, specifically for Nvidia. For every day in that range, an API request is made to fetch the news data. For each news article in the daily results, only the relevant fields—the publication timestamp, the headline, and the summary—are kept. These values are then aggregated into a list of dictionaries, converted into a Pandas DataFrame, and then saved as news_data.csv.

Unit of Observation

This file contains the same content as the json file described above. Each row represents a single news article related to Nvidia stock and includes its timestamp, headline, and summary. The dataset covers articles published between January 6 and January 31, except that no articles about Nvidia were published on January 26.

Variables

- date
    - The date and time when the news article was published (UTC timezone, ISO 8601 format)
    - Data type: Object
    - Missing values: 0 (405(0))
- headline

    - The title of the news article
    - Data type: Object
    - Missing values: 0 (405(0))
- summary
    - A brief summary of the news article content
    - Data Type: Object
    - Missing values: 126 (405(126))

Frequency Table

Same as the one for january_news.json.

Frequency Distribution

Same as the one for january_news.json.

news_data_processed.csv

Data from news_data.csv: The ‘date’ variable is trimmed so that only the date portion (YYYY-MM-DD) remains. The variable ‘text’, is created by concatenating each article’s headline and summary using a “: ” separator, with any missing summary replaced by an empty string. Additionally, because weekends are not trading days, articles published on Saturday or Sunday are aggregated with those from the preceding Friday. This adjustment makes it easier to analyze how articles posted on Friday and over the weekend may predict Monday’s opening prices.

Unit of Observation

Each row in this file represents an aggregated news summary for a specific date, covering January 6 through January 31. For each date, the headlines (and summaries, when available) of all articles published on that day are combined into a single text block. This aggregation is performed so that sentiment analysis can generate an average sentiment score—ranging from 0 (bearish) to 1 (bullish)—for each day.

Variables

- date
    - The date (formatted as YYYY-MM-DD) for which news items have been aggregated
    - Data Type: Object
    - Missing values: 0 (25(0))
- text
    - A concatenated text string that aggregates multiple headlines (and summaries) for the corresponding date.
    - Datatype: Object

    - Missing values: 0 (25(0))

Nvidia Stock Price Data

prices_data.cvs

Data is retrieved from the Alpaca API for each day between January 6 and January 31, 2025, specifically for the Nvidia stock. For every day in that range, an API request is made to fetch the price bar data. For each price bar, we extracts:

- The timestamp as the date. This timestamp is then converted to a string and truncated to the first 10 characters (yielding the YYYY-MM-DD format).
- The opening price, which is stored as the price.

Unit of observation

Each row represents the recorded opening price of Nvidia stock for a trading day between January 6 and January 31. Note that there is no price data available for weekends.

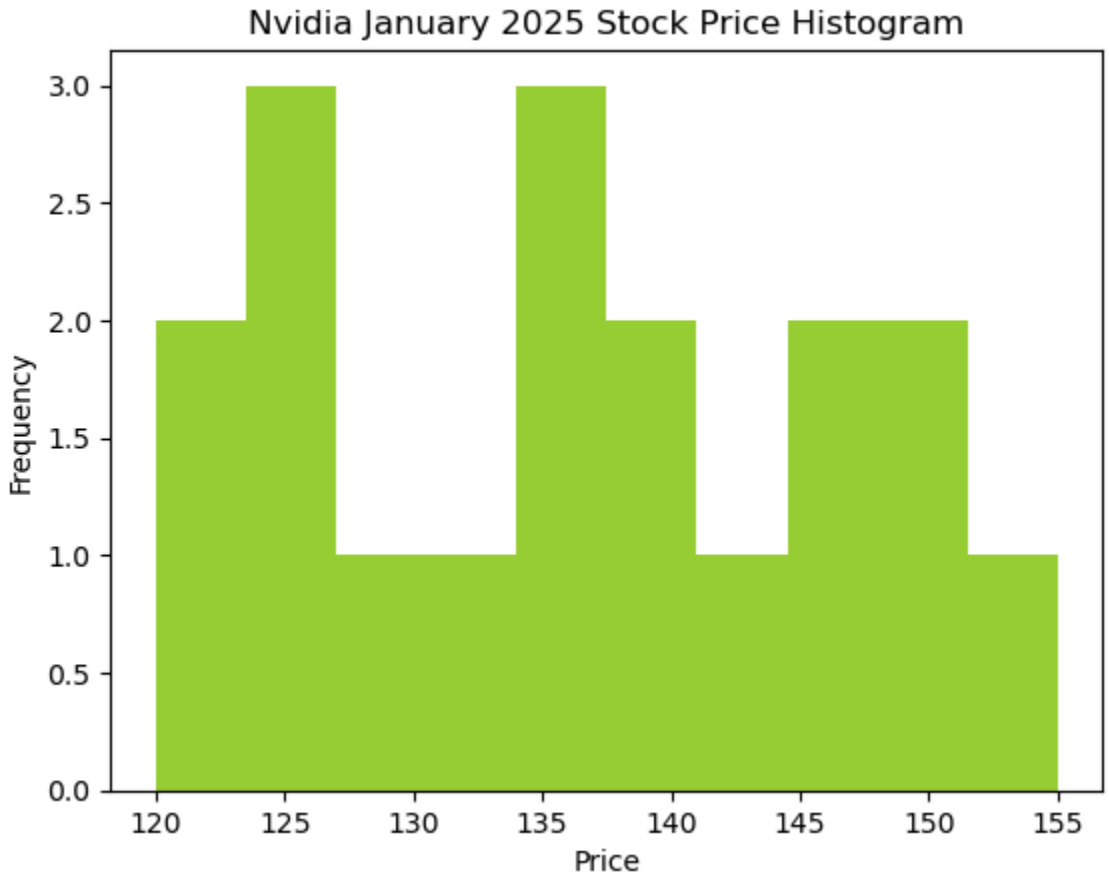
Variables

- date
    - The date for which the Nvidia opening price is recorded
    - Data Type: Object
    - Missing values: 0 (20(0))
- price
    - The opening price Nvidia stock of a given date
    - Data Type: Float
    - Missing values: 0 (20(0))

Summary Statistics

Count	18.00
Mean	136.33
Standard deviation	9.68
Minimum	121.81
25%	127.37
50%	137.07
75%	144.14
Maximum	153.03

Histogram



Sentiment Analysis and Comparative Analysis Data

sentiment_scores.csv

This file is constructed by first reading aggregated daily news text from news_data_processed.csv. For each date, the aggregated text is formatted and sent to the GPT-4-turbo API with a prompt to generate a sentiment score between 0 (bearish) and 1 (bullish). The resulting sentiment score is then paired with its corresponding date, and all records are compiled into sentiment_scores.csv for subsequent analysis.

Unit of observation

Each row represents a specific date for which news articles were processed and a sentiment score was computed.

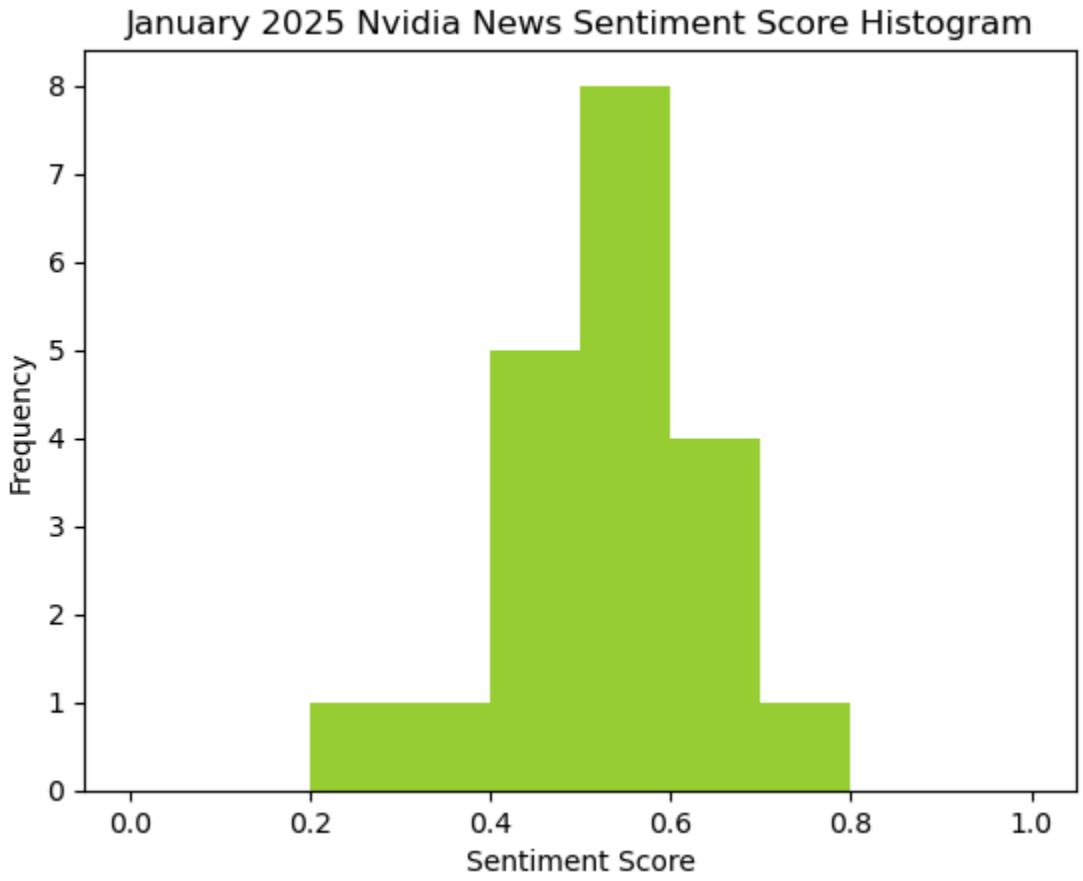
Variables

- date
    - The date on which the news articles were analyzed, formatted as YYYY-MM-DD.
    - Data Type: Object
    - Missing values: 0 (20(0))
- Sentiment Score
    - A numerical value between 0 and 1 representing the overall sentiment of the news articles for a given date.
        - 0 corresponds to extremely negative sentiment.
        - 1 corresponds to extremely positive sentiment.
    - Data Type: Float
    - Missing values: 0 (20(0))

Summary Statistics

Count	20.000
Mean	0.529
Standard deviation	0.107
Minimum	0.280
25%	0.470
50%	0.540
75%	0.590
Maximum	0.760

Histogram



scores_with_prices.csv

Data from sentiment_scores.csv (daily sentiment scores) and prices_data.csv (daily opening prices) are merged by date. The variable ‘Next Day Price’ is created by shifting the price data upward by one day and then ‘Price Delta’ (difference between the next day’s price and the current day’s price) is calculated. This merged and adjusted dataset—with records for date, sentiment score, price, next day price, and price delta—is saved as scores_with_prices.csv.

Unit of observation

Each row represents a trading day for which both a news sentiment score and stock price data are available.

Variables

- date
    - The calendar date of the trading date, formatted as YYYY-MM-DD
    - Data Type: Object

    - Missing values: 0 (17(0))
- Sentiment Score
    - A floating-point number between 0 and 1 representing the average sentiment of news articles on NVIDIA stock for the trading day
    - Data Type: Float
    - Missing values: 0 (17(0))
- price
    - The opening stock price for the trading day, expressed in U.S. dollars (USD).
    - Data Type: Float
    - Missing values: 0 (17(0))
- Next Day Price
    - The opening stock price of the subsequent trading day, expressed in USD.
    - Data Type: Float
    - Missing values: 0 (17(0))
- Price Delta
    - The change in the opening price from the current trading day to the next, calculated as: Price Delta = Next Day Price - price
    - Data Type: Float
    - Missing values: 0 (17(0))