**Report: Stock Market Analysis on Twitter Dataset**

**Tweets dataset**

1. **Overview**

* This dataset contains 3,653 rows and 4 columns. 4 columns are Date, closing\_price, adj\_close\_price, Tweets

1. **Columns Description**

| **Column Name** | **Description** |
| --- | --- |
| Date | Date of record. |
| closing\_price | Closing stock price of that day. |
| adj\_close\_price | Adjusted closing stock price. |
| Tweets | Text data that containing tweets. |

**Make tweets dataset Unstructured to Structured**

**1. Problem Identification**

* The original dataset contained a column that name Tweets.
* Some tweets are written as multi-line text within a single cell, separated by line breaks.
* This made the data inconsistent and difficult to analyse, as each tweet was not cleanly represented in a single row.

**2. Objective**

* Ensure each row in the dataset contains one full tweet in a single row, single line.
* Preserve the associated metadata: Date, closing\_price, and adj\_close\_price.
* Remove unwanted line breaks and whitespace from tweets.

**3. Tools Used**

* Python with the pandas library for data manipulation.
* CSV files used for both input (unstructured) and output (structured).

**4. Steps of making tweets\_dataset Unstructure to Structure**

* Load original tweets\_dataset.
* Prepare for clean data.
* Itearate every row & remove row’s empty space & append break tweet line in a single row

to make dataset structure.

* Create a New DataFrame that containing the clean data.
* Save this DataFrame to a New CSV file.

**5. Result**

* Every tweet data is now in a single row, single line, free of line breaks.
* Date, closing\_price, and adj\_close\_price data also maintained their line following

Tweets data.

**Report on Exploratory Data Analysis**

**1. Closing Price Over Time**

* A line plot was created to visualize the trend of the closing price over time using the sns.lineplot() function from Seaborn. This visualization helps identify fluctuations in the stock price.

**2.** **Adjusted vs Closing Price**

* To compare the 'Adjusted Price' and 'Closing Price' over time line plot was generated. This visualization helps understand the difference and relationship between those two price metrics.

**3. Tweet Length Distribution**

* A new column 'Tweet\_Length' was created to store the length of each tweet.
* The distribution of tweet lengths was plotted using a histogram sns.histplot(), providing insights into the length distribution of tweets.

**4. Correlation Analysis**

* A correlation matrix was calculated for 'closing\_price', 'adj\_close\_price', and 'Tweet\_Length' using df.corr().
* A heatmap was generated using seaborn's sns.heatmap() to visualize the correlations between these variables. This helps understand potential relationships or dependencies between variables.

**Conclusion:**

* Exploratory Data Analysis provides valuable insights from the dataset, particularly through the visualizations, plotting and correlation heatmap.