Stock Prediction Analysis

Group 8

Data Science Capstone Project
Data Acquisition and Preprocessing Report

Date:

02/10/2025

Team Members:

Name: Robert Lignowski - rml345@drexel.edu

Name: Uditi Shah - us54@drexel.edu

Name: Ahmad Javed -aj3235@drexel.edu

Name: Steven Sullivan - sas683@drexel.edu

Identifying Data

Data Sources:

Our project utilizes historical stock data obtained from Investing.com. The data was downloaded from the "General - Historical Data" section of the website, allowing us to access comprehensive stock market information. We have selected stock data for four major automobile manufacturers: Ford, Volkswagen, Toyota, and Tesla. These companies were chosen due to their significant presence in the global automotive industry, representing a mix of traditional and electric vehicle manufacturers. By analyzing their stock trends, we aim to gain insights into market behavior, investor sentiment, and the financial performance of these companies over time. Investing.com was selected as our primary data source because it provides reliable, up-to-date, and structured historical stock data, making it a suitable choice for our analysis.

Acquisition Process:

The dataset for our project is readily available for download from Investing.com. To acquire the data, we navigated to the "General - Historical Data" section for each stock: Ford, Volkswagen, Toyota, and Tesla and manually downloaded the historical stock prices in CSV format. Since the data is available for direct download, we did not need to write custom code for web scraping or API access.

As we are using multiple data sources (one for each stock), integration is necessary to ensure consistency in our analysis. After downloading the CSV files, we preprocessed the data by standardizing date formats, handling missing values, and aligning time frames to ensure comparability across different stocks. Additionally, we performed initial data cleaning, such as removing unnecessary columns and renaming headers for uniformity. This structured approach ensures that the dataset is well-prepared for further analysis and modeling.

Issues:

Tesla's data acquisition issues include missing values in key columns, inconsistent volume formatting, and potential outliers affecting analysis accuracy. Ensuring data completeness and consistency remains a challenge for reliable trend analysis.

Volkswagen's dataset had volume values formatted with 'K' (thousands) and 'M' (millions), requiring conversion to a consistent numerical format. Additionally, some missing values were detected in the dataset, which were filled using the median of the respective column. The stock also exhibited significant volatility between 2021-2022, likely due to market events related to Volkswagen's EV expansion.

Toyota had similar issues to Volkswagen, requiring conversion to a consistent numerical format. Missing values were filled using the median of the respective column. We specifically chose to use the Toyota data from the New York Stock Exchange rather than that of Toyota's native Japan to ensure consistency.

Data-Processing

Missing Data:

- Some columns may have missing values.
- **Solution**: Fill missing values with the median of each respective column.

Noise:

- The dataset may contain outliers or extreme fluctuations.
- **Solution**: Identify and handle outliers using statistical methods or visualization (e.g., box plots).

Data Format:

- Columns like 'Vol.' and 'Change %' may have non-numeric characters (e.g., 'M', '%').
- **Solution**: Remove non-numeric characters and convert the columns to numeric types.

Date Column:

- The 'Date' column may not be in the correct datetime format.
- **Solution**: Convert the 'Date' column to datetime format.

Data Consistency:

- Ensure consistency in units, date formats, and naming conventions if using data from multiple sources.
- **Solution**: Standardize data before merging or combining datasets.

Appendix

For TSLA:

Sample Data

Date	Price	Open	High	Low	Vol.	Change %
01/17/2025	426.50	421.50	439.74	419.75	94.99M	3.06%
01/16/2025	413.82	423.49	424.00	409.13	68.34M	-3.36%
01/15/2025	428.22	409.90	429.80	405.66	81.38M	8.04%

Data Definition

- **Date**: The date of the stock data entry in the format (MM/DD/YYYY).
- **Price**: The closing price of Tesla stock on the given date.
- **Open**: The opening price of Tesla stock on the given date.
- **High**: The highest price reached by Tesla stock during the trading day.
- **Low**: The lowest price reached by Tesla stock during the trading day.
- **Vol.**: The volume of Tesla shares traded on the given date (in millions).
- **Change %**: The percentage change in Tesla stock price compared to the previous trading day.

Pseudocode

1. Load the Tesla stock dataset from a CSV file.

2. Clean the 'Vol.' column:

   - Remove the 'M' character and convert the values to float.

   - Multiply by 1,000,000 to convert to actual volume.

3. Clean the 'Change %' column:

   - Remove the '%' sign and convert the values to float.

4. Convert the 'Date' column to datetime format.

5. Handle missing values:

   - Fill missing values with the median of the respective column.

6. Perform Exploratory Data Analysis (EDA):

   a. Display basic statistics of the dataset.

   b. Check for missing values.

   c. Plot the time series of 'Price', 'Vol.', and 'Change %'.

For VWAGY:

Sample Data

Date	Price	Open	High	Low	Vol.	Change %
01/02/2025	9.08	9.22	9.23	9.03	307210	-2.47
12/31/2024	9.31	9.36	9.41	9.27	451220	-0.53
12/30/2024	9.36	9.42	9.48	9.36	759570	-0.74
12/27/2024	9.43	9.44	9.49	9.40	427290	1.51
12/26/2024	9.29	9.30	9.37	9.20	436210	-0.21

1. Load the Volkswagen (VWAGY) stock dataset from CSV file.

2. Convert the 'Vol.' column:

   - If it contains 'M', multiply by 1,000,000.

   - If it contains 'K', multiply by 1,000.

   - Ensure all values are numeric.

3. Convert the 'Change %' column:

   - Remove the '%' symbol.

- Convert it to a floating-point number.

4. Convert the 'Date' column to datetime format.

5. Handle missing values:

- Fill missing values with the median of the respective column.

6. Perform Exploratory Data Analysis (EDA):

a. Display basic statistics of the dataset.

b. Check for missing values.

c. Plot the time series of 'Price', 'Vol.', and 'Change %'.

For TM:

Sample Data

	Date	Price	Open	High	Low	Vol.	Change %
1279	1/2/2020	142.24	142	142.35	141.53	118090	1.21
1278	1/3/2020	140.75	141.23	141.42	140.3	174890	-1.05
1277	1/6/2020	140.77	139.46	140.89	139.46	150390	0.01
1276	1/7/2020	141.51	142.13	142.43	141.36	123670	0.53
1275	1/8/2020	141.16	140.91	141.58	140.6	117710	-0.25
1274	1/9/2020	140.51	141.16	141.16	140.11	116170	-0.46

Psuedocode:

1. Load data from CSV file
2. Convert “Vol.” column
    a. If it contains “K” multiply by 1000
    b. If it contains “M” multiply by 1000000
    c. Convert all values to float
3. Remove “%” from “Change%” column
4. Convert “Date” column to Datetime format
5. Fill missing values with the median of the respective column.
6. Perform Exploratory Data Analysis
    a. Display basic statistics of the dataset.
    b. Check for missing values.

c. Plot the time series of 'Price', 'Vol.', and 'Change %'

For F (Ford):

Sample Data:

	Date	Close	Open	High	Low	Volume
0	1/2/2020	$9.42	$9.29	$9.42	$9.19	43,432,239
1	1/3/2020	$9.21	$9.31	$9.37	$9.15	45,059,915
2	1/6/2020	$9.16	$9.10	$9.17	$9.06	43,380,677
3	1/7/2020	$9.25	$9.20	$9.25	$9.12	45,334,552
4	1/8/2020	$9.25	$9.23	$9.30	$9.17	46,003,049
...	...	...	...	...	...	...
1255	12/27/2024	$10.03	$10.03	$10.20	$9.98	52,899,918
1256	12/30/2024	$9.88	$9.95	$9.98	$9.82	47,116,735
1257	12/31/2024	$9.90	$9.91	$10.01	$9.84	54,104,154
1258	1/2/2025	$9.65	$9.91	$9.96	$9.64	67,156,220
1259	1/3/2025	$9.88	$9.69	$9.95	$9.53	77,245,879

Pseudocode:

1. Import libraries
    a. pandas
    b. numpy
    c. matplotlib
2. Import CSV files as dataframes
3. Identify all column types
    a. All columns were original ‘object’ type
4. Clear all symbols and commas from values
    a. Replace ‘$’ with ‘’ for ‘Close’ Column
    b. Replace ‘$’ with ‘’ for ‘Open Column
    c. Replace ‘$’ with ‘’ for ‘High’ Column
    d. Replace ‘$’ with ‘’ for ‘Low’ Column
    e. Replace ‘,’ with ‘’ for ‘Volume’ Column
5. Convert ‘Date’ column from object type to datetime type
6. Identify any missing or null values in dataframe
    a. No column contained any missing or null values
7. Plot Line graph for Close Price vs Date
    a. X-axis = ‘Date’
    b. Y-axis = ‘Close’

8. Complete additional analysis on values in dataframe
    a. Most common stock price for High, Low, Close, and Open

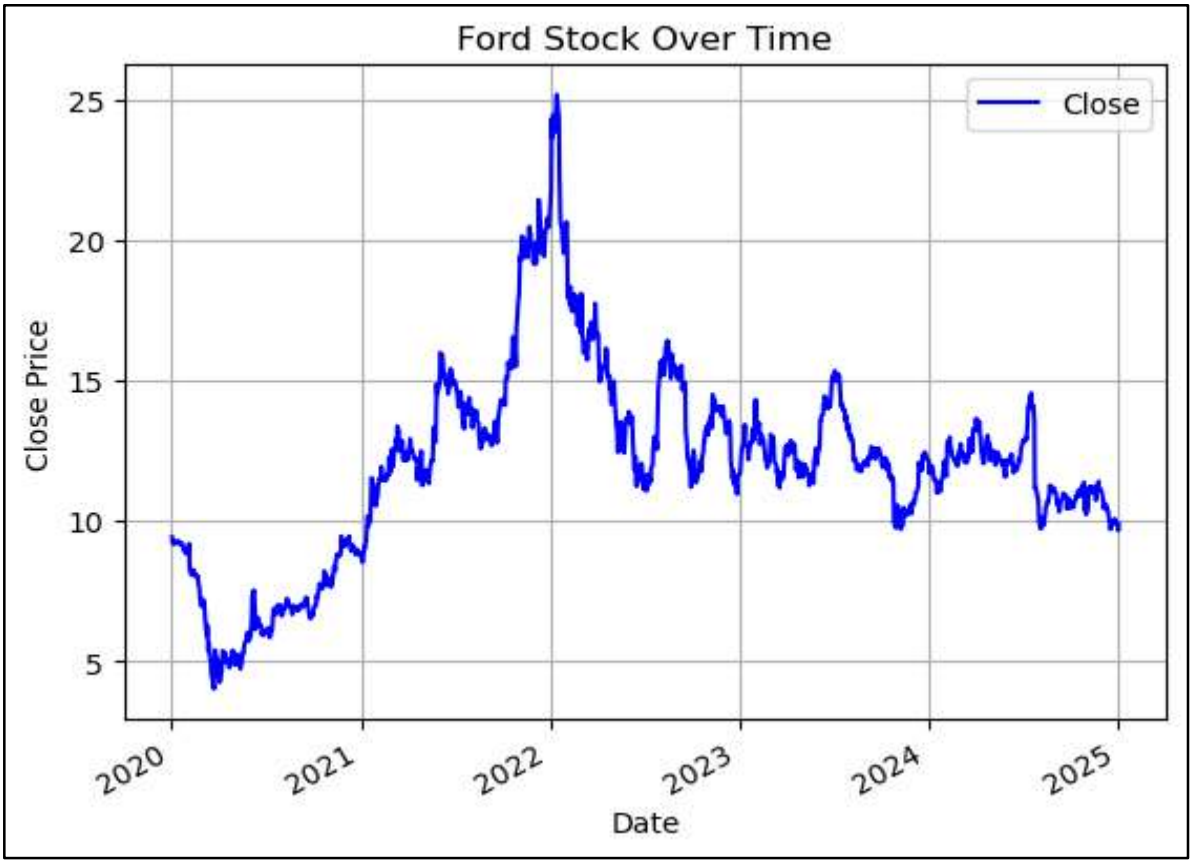

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	Writing	Editing
1	**Data Sources**	**All Members**	**Uditi, Steven**
2	**Data Pre-Processing**	**All Members**	**Ahmad, Robert**
3	**Appendix**	**All Members**	**All Members**

Grading

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.