**STOCK PRICE PREDICTION (Phase 2)**

**Madras Institute of Technology**

2021506088 – Santhi Priya D N

2021506096 - Shiek Sajnathul Faizana

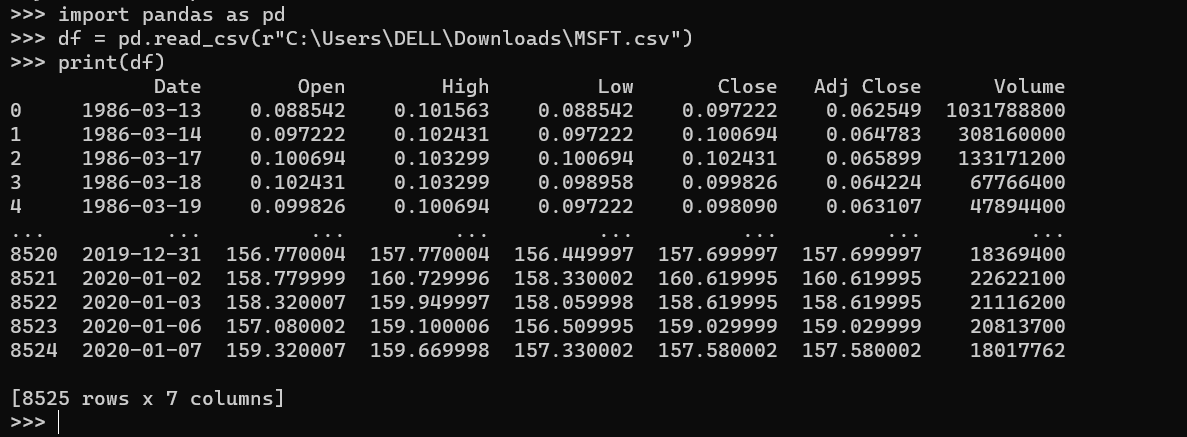
2021506103 – Sowmiya J

2021506323 - Revanth P

**1.INTRODUCTION:**

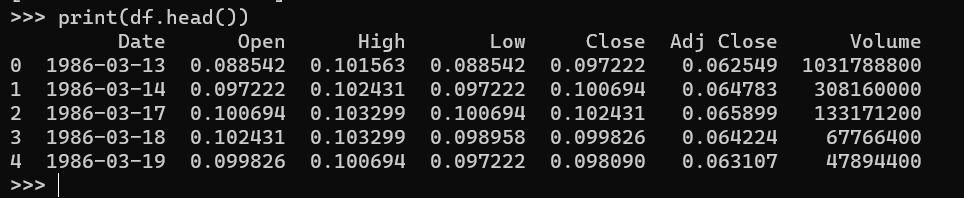
This phase aims to clean, transform, and engineer features in a way that maximizes the model's ability to capture patterns and make accurate predictions. Through careful data preparation and feature engineering, we enhance the quality of input fed into the models. Effective preprocessing lays the foundation for improved predictive models. **The given dataset has been pre-processed and the outputs are attached with snap shots.**

**2.IMPORTING LIBRARIES AND LOADING DATA:**

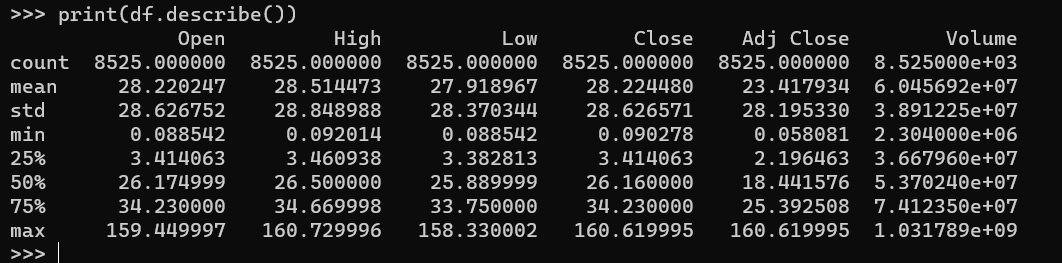
For Pre-Processing the given dataset, the pandas library is used. The given csv file is uploaded to pandas as follows:  


**3.UNDERSTANDING THE DATASET:**

**df.head:**

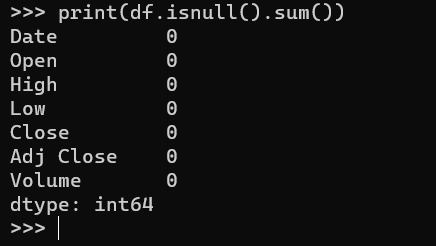


**df.describe:**

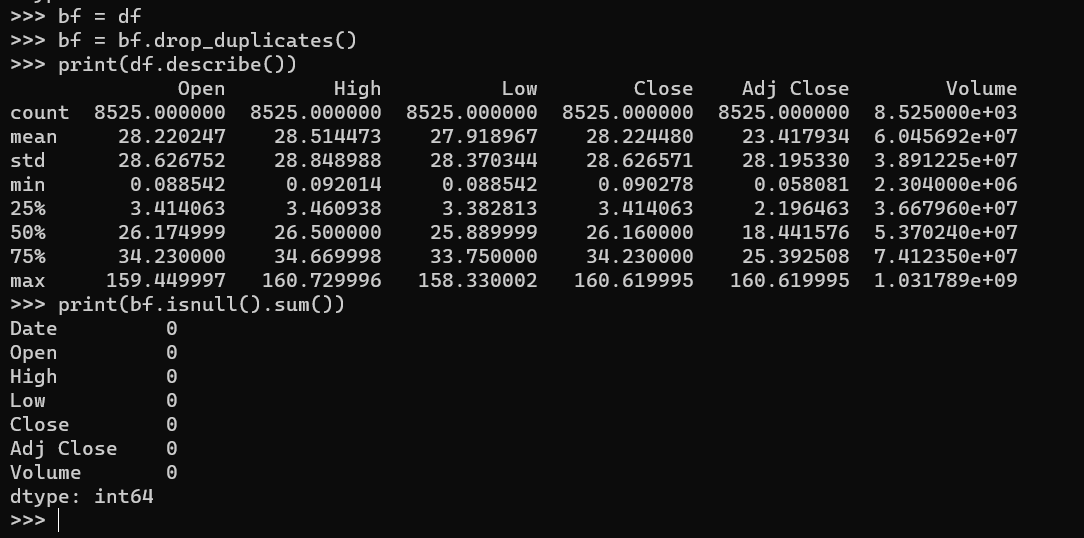


**Isnull:**

This function is used to identify missing values in the dataset. Since there in no null value, there is no need for handling the missing data.

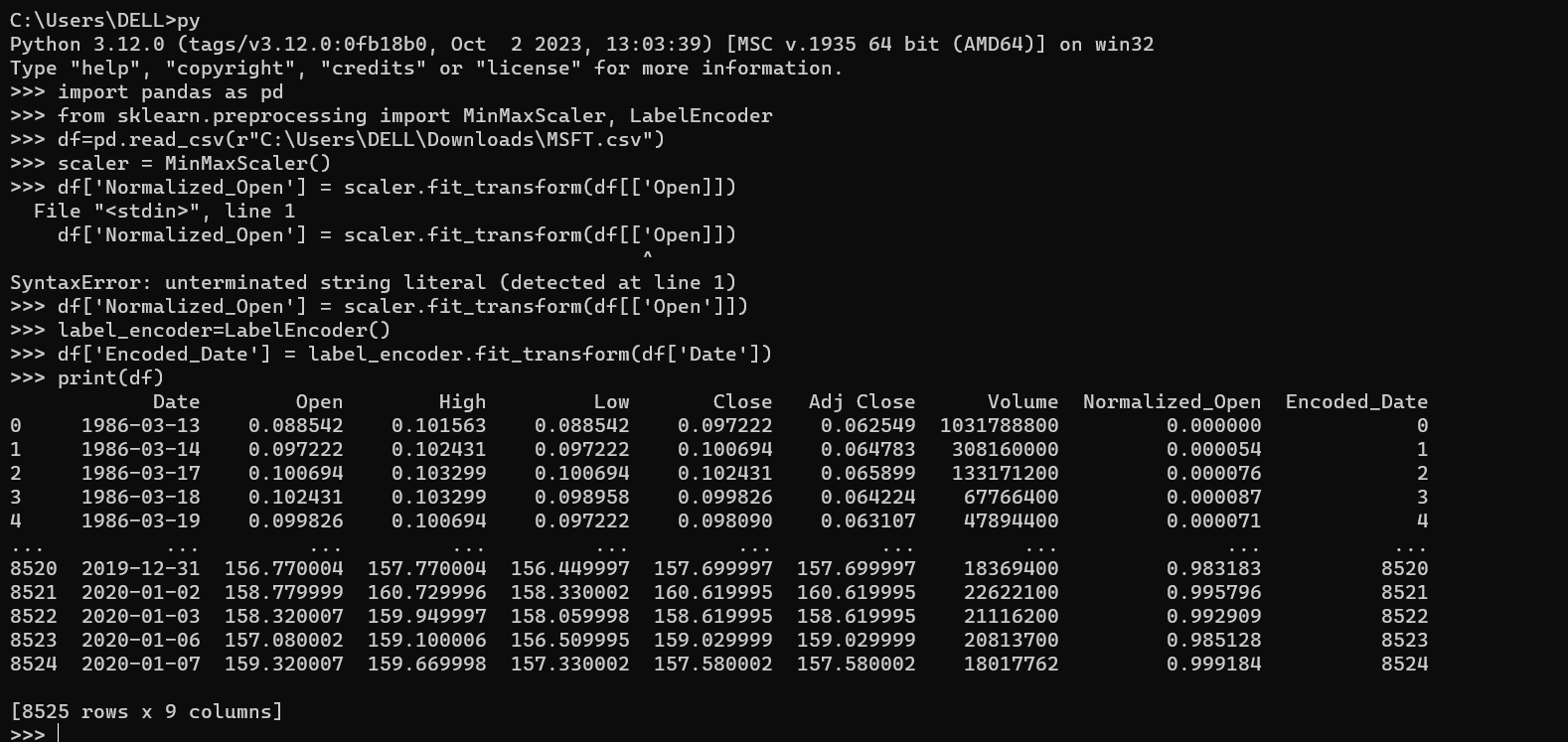


**4.REMOVING DUPLICATES:**

If any row is duplicated in the given dataset, the following code will identify it and remove it. The given dataset does not contain any duplicates and hence the dataset is the same as before.  


**5.DATA TRANSFROMATION:**

**Normalizing Data:**



**Z-Score Standardization (for column – high):**

Z-score standardization, also known as "z-score normalization" or "z-score scaling," is a statistical method used to standardize or normalize features in a dataset. It's a process that transforms the features by scaling them to have a mean of 0 and a standard deviation of 1. This makes it easier to compare and analyze variables with different units or scales.

The formula to calculate the z-score for a given data point

X in a feature is: **z=X- μ/ σ**

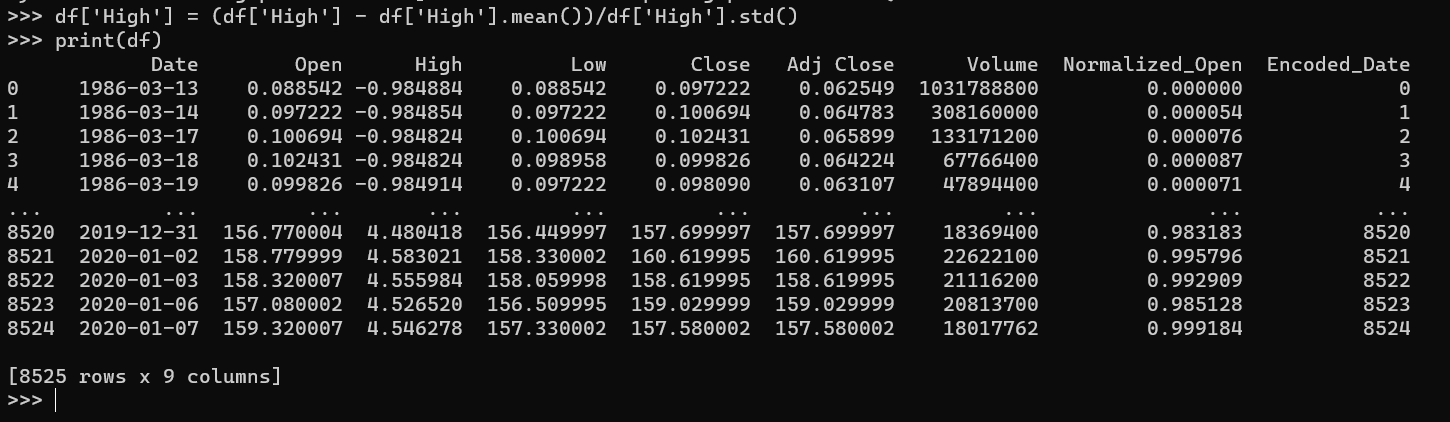
where:

X is an individual data point.

μ is the mean of the feature.

σ is the standard deviation of the feature.

The z-score measures how many standard deviations a data point is from the mean. A positive z-score indicates that the data point is above the mean, while a negative z-score indicates it's below the mean.

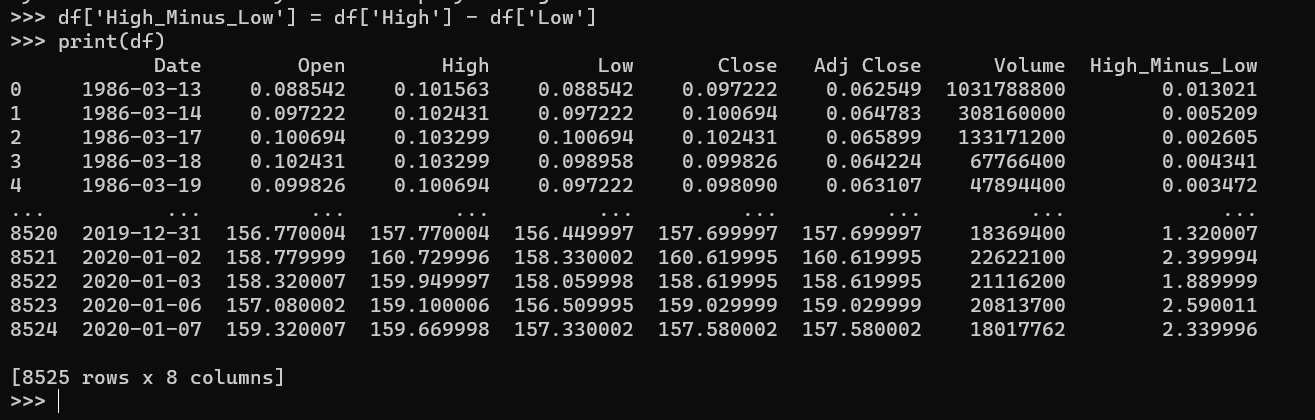


**6.FEATURE ENGINEERING:**

Feature engineering is the process of creating new features from the existing raw data or modifying the existing features in a way that enhances the performance of machine learning models. The goal is to provide more informative, representative, and discriminative features to improve the model's ability to learn and generalize.

In the given dataset, a new column(feature) High minus Low is obtained by subtracting low from high, which can be used for further analysis.

**High Minus Low:**

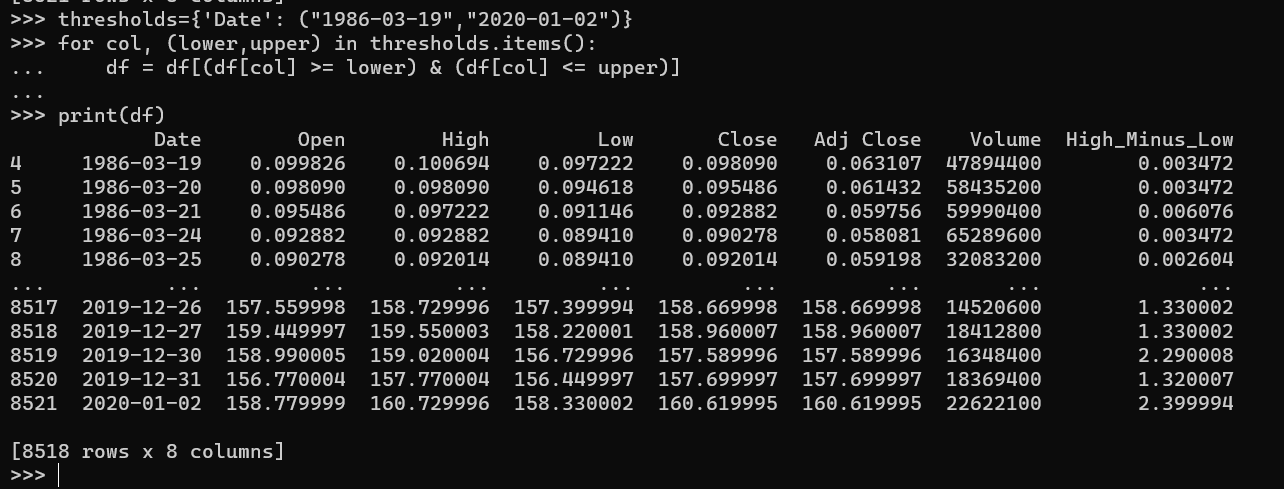


**7.HANDLING OUTLIERS:**

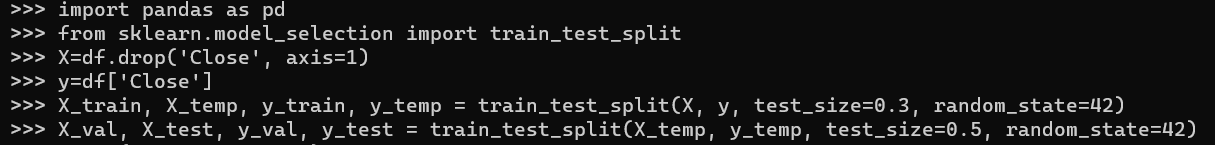
Outliers are data points that significantly differ from other observations in a dataset, deviating markedly from the overall pattern or distribution. They can be unusually high or low values that don't conform to the typical behaviour of the dataset.

The threshold fixed are the end points or outliers, all the values above and below are range are excluded and this process is called handling outliers.

**Date fixed as threshold:**



**8.DATA SPLITTING:**

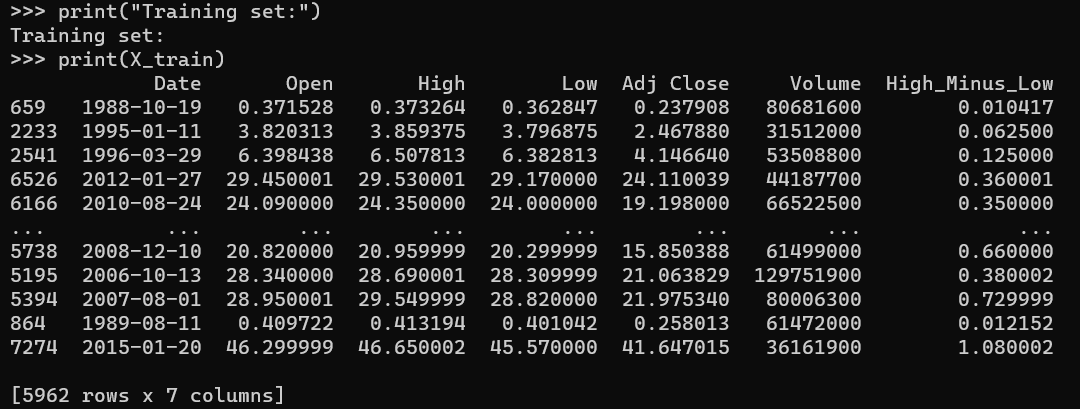
Outliers are data points that significantly differ from other observations in a dataset, deviating markedly from the overall pattern or distribution. They can be unusually high or low values that don't conform to the typical behavior of the dataset.  


**TRAINING SET:**

Purpose: Used to train the model, allowing it to learn patterns and relationships in the data.

Size: Largest portion of the dataset (e.g., 70-80%).

Importance: Fundamental for model training, ensuring the model learns from a variety of examples

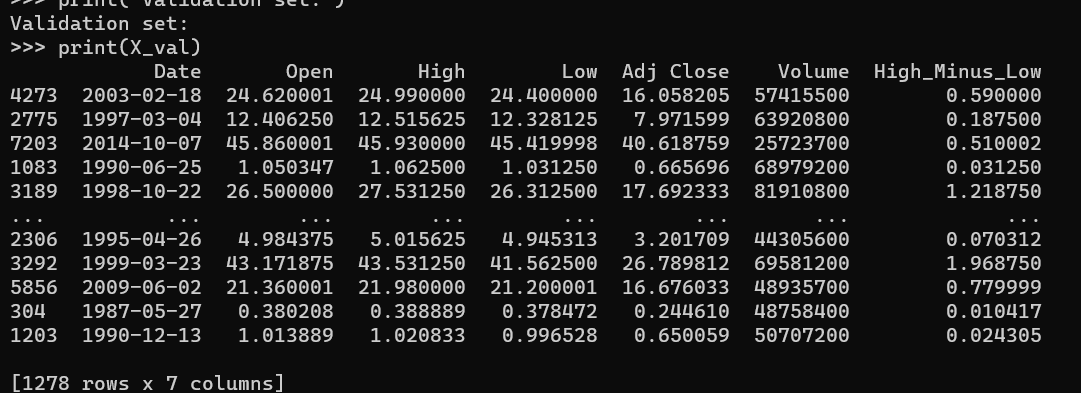


**VALIDATION SET:**

Purpose: Used to fine-tune the model's hyperparameters, aiding in model selection and preventing overfitting.

Size: Smaller portion of the dataset (e.g., 10-15%).

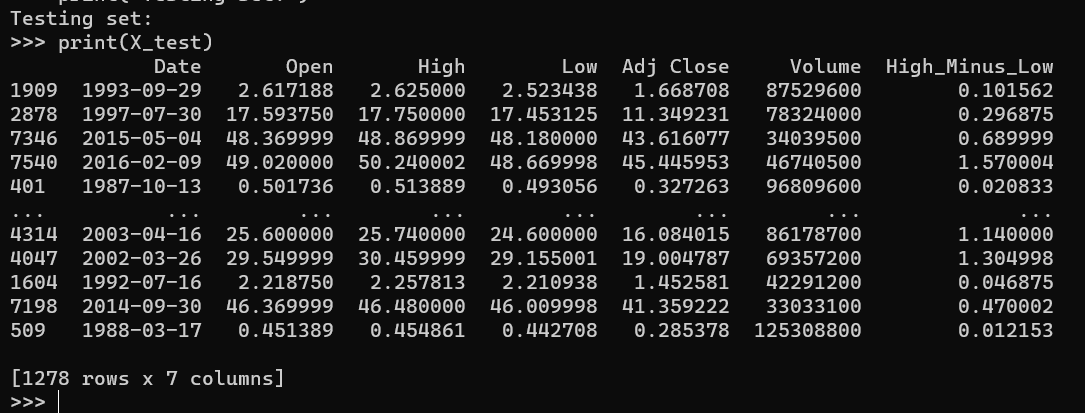
Importance: Helps optimize the model's performance and generalization.



**TESTING SET:**

Purpose: Used to evaluate the model's performance on unseen data after training and validation.

Size: Smaller portion of the dataset (e.g., 10-15%).

Importance: Provides an unbiased evaluation of the model's performance and generalization to new data.  


**9.SAVING:**



**10.CONCLUSION:**

In the third phase, the dataset has been preprocessed, which is fundamental to building accurate and reliable predictive models. This involved handling missing values, scaling, encoding categorical features, and possibly applying other transformations like feature engineering or selection. The preprocessed dataset is now ready for the subsequent phases, where it will be utilized to train and validate models