# Stock Price Prediction-Comprehensive Exploratory Data Analysis Report

## 1. Data Loading and Initial Exploration

The analysis began with loading the dataset, time-series financial data containing stock price information spanning multiple years. Upon initial inspection, the dataset contained historical records with varying degrees of quality and completeness.

## 2. Data Cleaning and Preprocessing

### 2.1 Column Selection

After examining the dataset structure, unnecessary columns were removed to focus on the most relevant features for analysis. This step helped streamline the dataset and improve computational efficiency.

### 2.2 Temporal Filtering

The data was aggregated and counted by year and month to understand the temporal distribution. Analysis revealed significant quality issues in older historical records, with more errors and inconsistencies present. Therefore, a decision was made to focus on the most recent 20 years of data to ensure higher quality analysis.

### 2.3 Handling Missing Values

The dataset contained null values across various columns. After careful consideration, all rows with null values were dropped, as the dataset still contained more than 4,000 records after this operation, providing sufficient data for meaningful analysis and model training.

### 2.4 Zero Volume Handling

Records with zero trading volume were identified and examined, as these could represent either market closures or data recording errors. The presence of zero volume entries was checked both in the original dataset and in the 20-year filtered dataset.

### 2.5 Duplicate Detection

The dataset was checked for duplicate rows to ensure data integrity. No duplicate records were found, confirming the uniqueness of each observation.

## 3. Exploratory Data Visualization

### 3.1 Time Series Visualization

Interactive line charts were created using Plotly to visualize all key features across the time dimension. This provided insights into the overall patterns, trends, and potential relationships between different

variables.

### 3.2 Trend Analysis with Rolling Means

Rolling means were calculated to smooth out noise and identify underlying trends in the data. This technique helped distinguish between short-term fluctuations and long-term patterns.

### 3.3 Close Price Analysis

Special attention was given to the 'Close' price, as it is a key indicator of market performance. Visualizations focusing solely on this metric helped identify critical trends and patterns that could inform trading strategies or predictive models.

## 4. Feature Correlation and Selection

Correlation analysis was performed to understand relationships between different features. This analysis informed feature selection decisions for subsequent modeling.

## 5. Data Preparation for Modeling

### 5.1 Train-Test Split

The dataset was split into training and testing sets, maintaining the chronological order of observations by avoiding random shuffling. This approach is crucial for time series data to prevent data leakage and ensure that models are trained on past data and tested on future data.

## 6. Insights

### 6.1 Key Findings

- **The dataset showed clear cyclical patterns and trends in price movements**
- **Volume data revealed periods of high trading activity that often corresponded with significant price changes**
- **Rolling means effectively smoothed out noise while preserving important trend information**
- **The 20-year filtering approach significantly improved data quality for analysis**

**Data_Dominators Task 4 part1**