Stock Price Data Analysis Report

Prepared by:

Ruwan Pathiranage Sanduni Nisansala

Internship ID:

CV/A1/45373

Tools Used:

Python, pandas, scikit-learn, matplotlib, statsmodels, seaborn

Project:

Data Analysis using Python

TABLE OF CONTENTS

1. INTRODUCTION

The Stock Price dataset contains daily trading information for multiple companies, including opening, high, low, and closing prices along with trade volume.

Dataset Columns

- Symbol
- Date
- Open
- High
- Low
- Close
- Volume

Project Overview

This project demonstrates essential data analysis and machine learning techniques in Python, focusing on three main analytical tasks:

- **Regression Analysis:** Performing a simple linear regression analysis to predict one variable based on another.
- **Time Series Analysis:** Analyze a time-series dataset (e.g., stock prices, temperature data) to detect trends and seasonality.
- **Clustering Analysis:** Implement K-Means clustering to group similar data points together based on feature similarities.

2. OBJECTIVES

Level 2 – Task 1: Regression Analysis

- Split the dataset into training and testing sets.
- Fit a linear regression model using scikit-learn.
- Interpret the coefficients and evaluate the model using metrics such as R-squared and mean squared error.

Level 2 – Task 2: Time Series Analysis

- Plot time-series data and identify patterns.
- Decompose the series into trend, seasonality, and residuals using statsmodels.
- Perform moving average smoothing and plot the results.

Level 2 – Task 3: Clustering Analysis (K-Means)

- Standardize the dataset (e.g., using StandardScaler).
- Apply K-Means clustering and determine the optimal number of clusters using the elbow method.
- Visualize clusters using 2D scatter plots.

3. TOOLS AND TECHNOLOGIES

Tool	Purpose
Python	Programming language used for all analysis
pandas	Data cleaning, manipulation, and preprocessing
scikit-learn	Machine learning modeling (Regression, K-Means)
statsmodels	Time series decomposition
matplotlib and seaborn	Visualization and plotting
VS Code	Development environment

4. TASK 1: DATA CLEANING AND REGRESSION ANALYSIS

Description

The raw dataset *2) Stock Prices Data Set.csv* was cleaned before performing regression analysis.

Data Cleaning Steps

1. Loaded dataset using pandas.
2. Checked for missing values - found in *open*, *high*, and *low* columns and they were removed missing rows.
3. Checked for duplicates – not found.
4. Converted *date* column to datetime format.
5. Saved cleaned dataset.

Dataset Summary after Cleaning

Detail	Value
Total Rows	497,461
Total Columns	7
Missing Values	None
Duplicate Rows	None

Model Building

- Predict *closing price* based on *opening price* using Linear Regression.

Steps:

1. Defined variables:

   X = df[['open']]

   y = df['close']

2. Split dataset (80% training, 20% testing).

3. Trained Linear Regression model using scikit-learn.
4. Evaluated model with R^2, MSE, and RMSE metrics.

Model Evaluation Metrics

Metric	Value
R-squared (R^2)	0.9997
Mean Squared Error (MSE)	2.7261
Root Mean Squared Error (RMSE)	1.6511
Intercept (b_0)	0.0267
Coefficient (b_1)	0.9999

Results and Interpretation

- **R^2 = 0.9997**

  Excellent model fit, explaining 99.97% of variance in closing prices.

- **Coefficient (b_1 = 0.9999)**

  Strong one-to-one relationship between open and close.

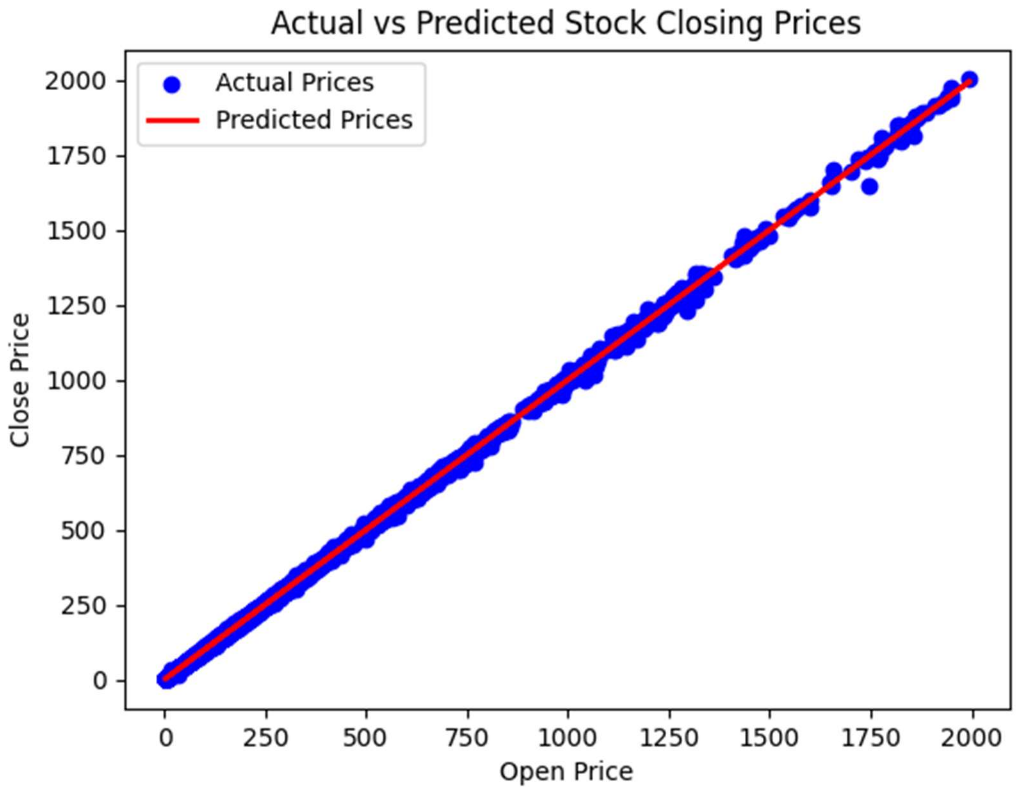
- **Intercept (0.0267)**

  Very small, indicating minimal difference between open and close prices.

- **Low MSE and RMSE**

  Model predicts accurately with minimal error.

Visualization



Above plot clearly shows predicted prices aligning closely with actual prices, proving the strong linear relationship.

5. TASK 2: TIME SERIES ANALYSIS

Objective

To analyze the *close* price trends over time, identify seasonal patterns, and visualize smoothed trends.

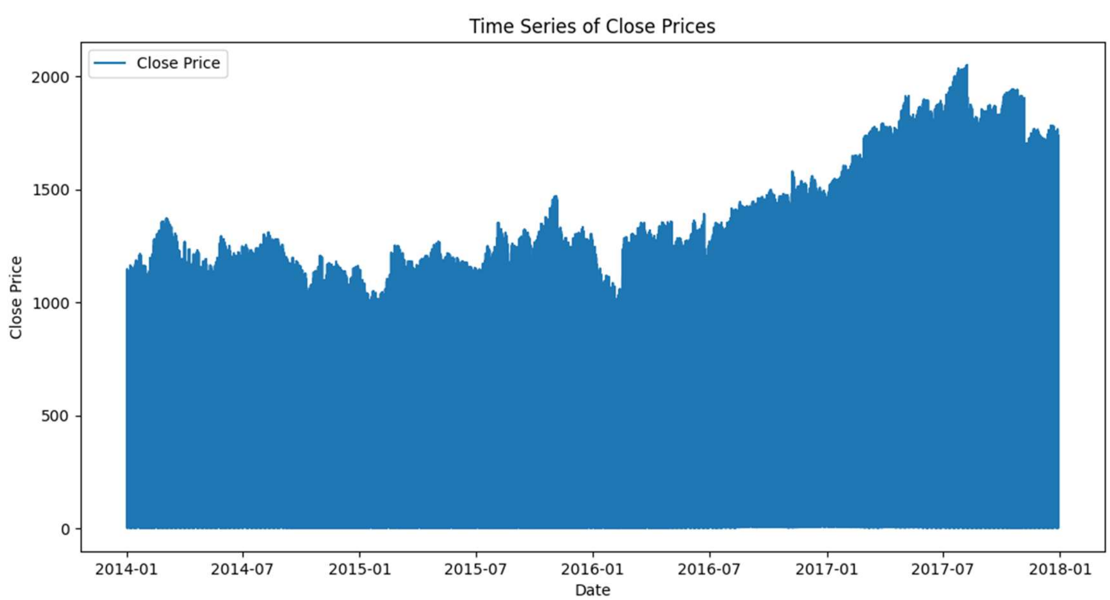
Methodology

1. **Data Preparation**

    - Loaded *Stock_Price_Cleaned.csv*.
    - Converted *date* column to datetime.
    - Sorted by date and set as index.

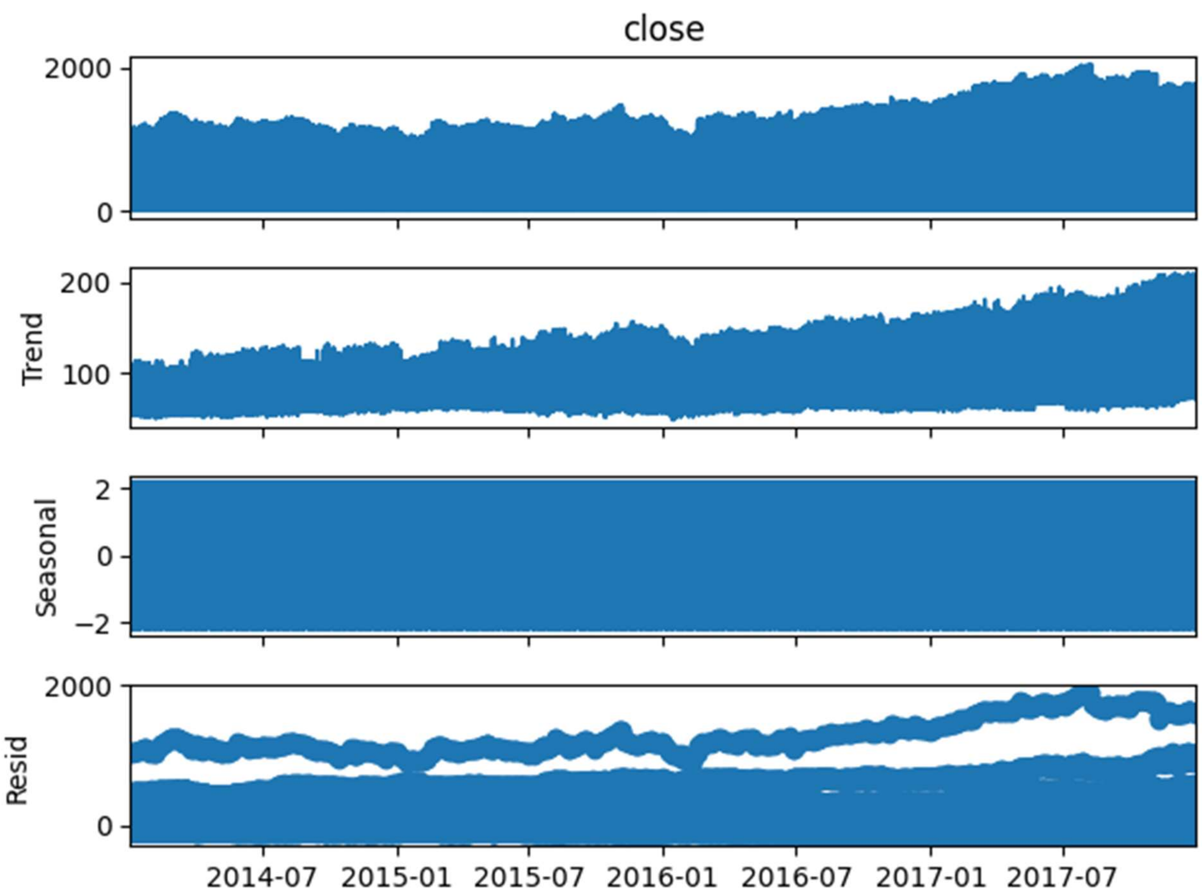
2. **Visualization**

    - Plotted *close* prices against date to show daily price fluctuations.



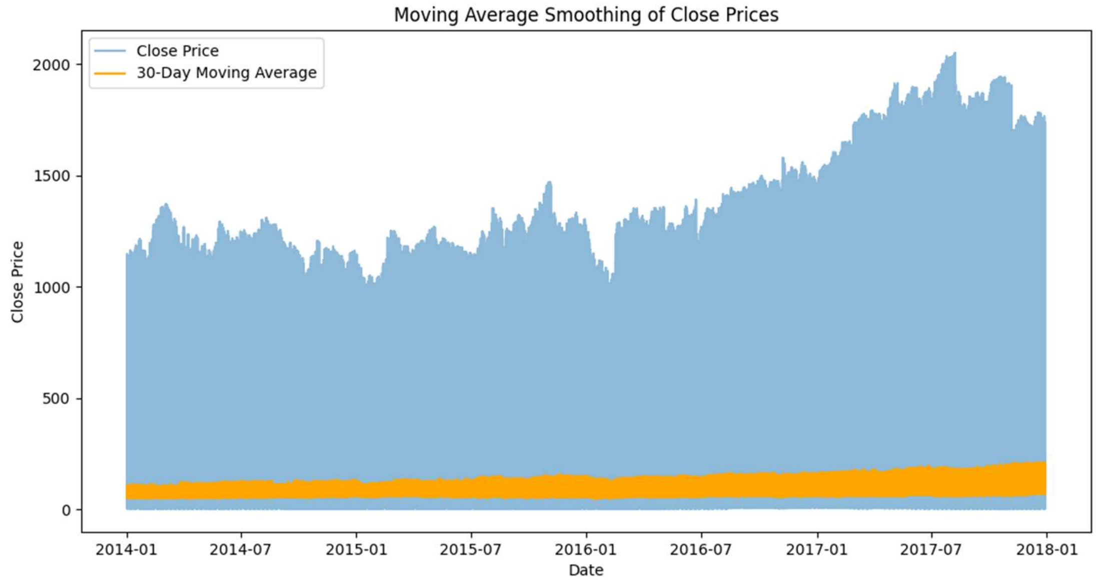
3. **Decomposition**

- Applied additive seasonal decomposition using seasonal_decompose().
- Extracted trend, seasonal, and residual components.



4. **Moving Average**

- Calculated 30-day moving average.
- Overlaid with actual prices to smooth short-term fluctuations.



5. **Saving Outputs**

Results and Interpretation

- Time series plot shows high frequency fluctuations over time.
- Decomposition highlights no seasonal component and reveals the underlying long-term movement.
- Moving average is barely visible at the bottom and suggesting it’s much flatter than the raw close price.

6. TASK 3: CLUSTERING ANALYSIS (K-MEANS)

Objective

To group stocks with similar price and volume characteristics into meaningful clusters.

Methodology

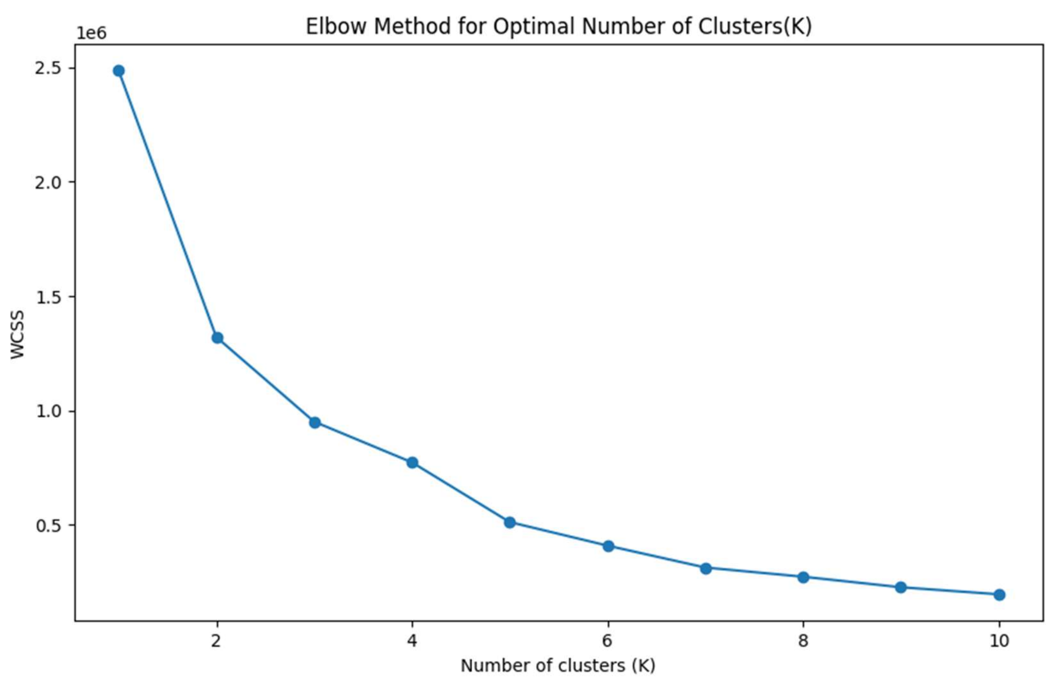
1. **Feature Selection**

    - Selected numerical features: *open, high, low, close, volume.*

2. **Feature Standardization**

    - Applied StandardScaler() to standardize the features.

3. **Finding Optimal K (Elbow Method)**
    - Computed Within-Cluster Sum of Squares (WCSS) for K = 1–10.



- Observed elbow point at **K = 3** (Optimal number of clusters)

- Saved plot.

4. **K-Means Clustering**

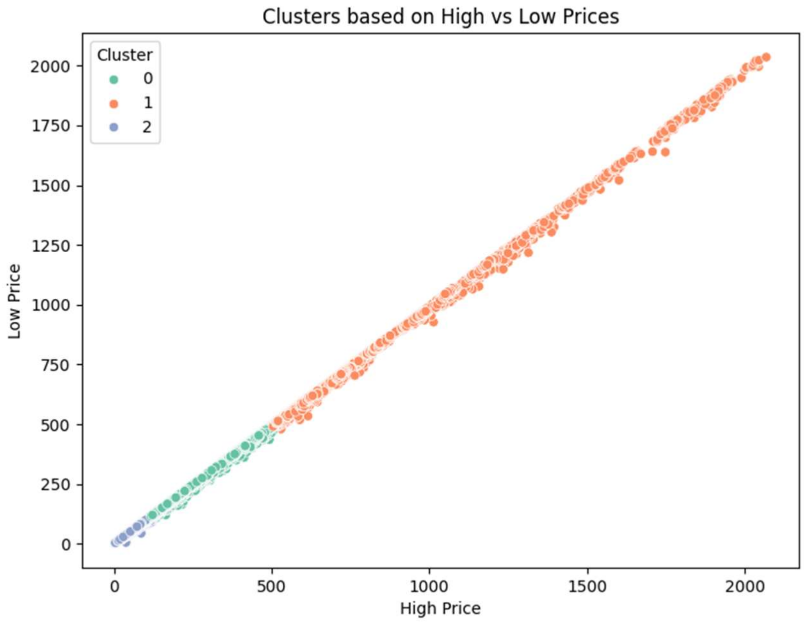
- Applied KMeans(n_clusters=3) and assigned cluster labels.
- Added a new column Cluster to the dataset.

5. **Cluster Visualization**

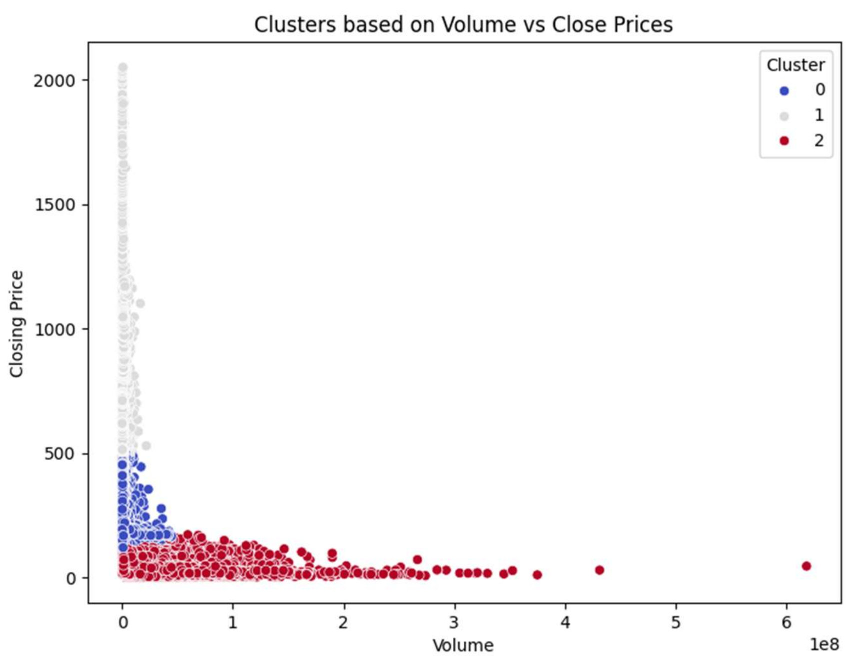
- Visualized clusters using multiple feature pairs:

    ✓ Open vs Close

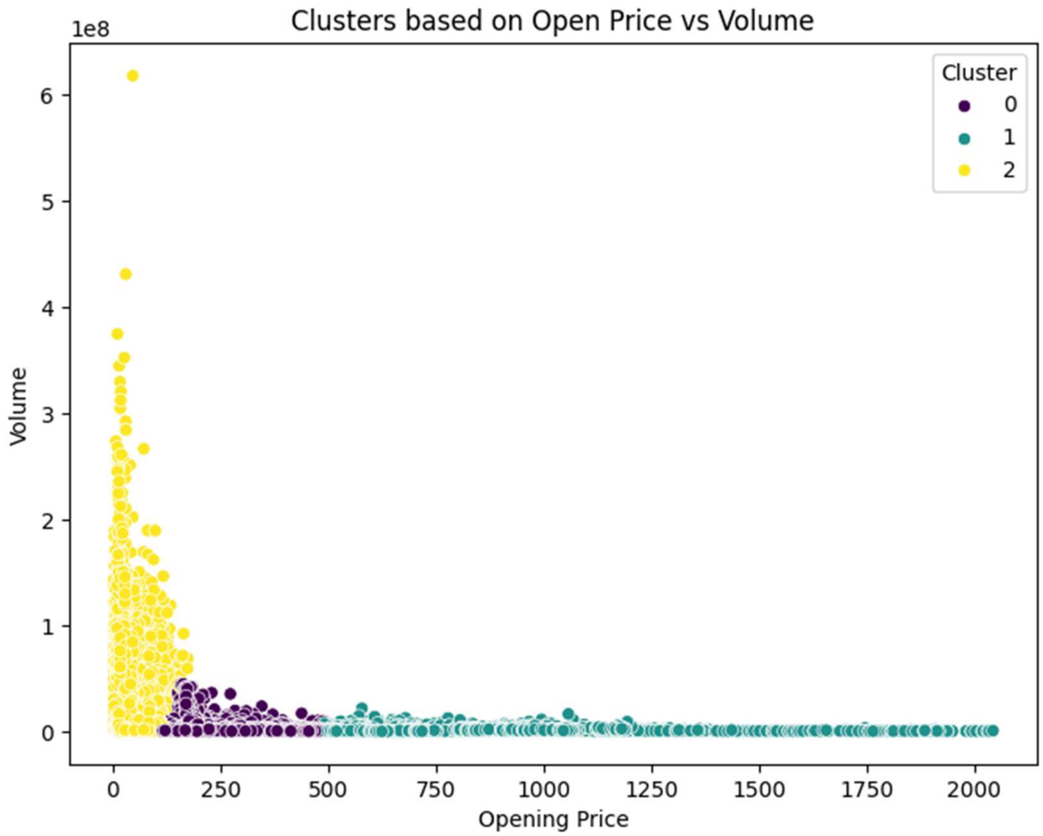
✓ High vs Low



✓ Volume vs Close



✓ Open vs Volume



- Saved clustered dataset
-

Results and Interpretation

- The Elbow Method confirmed 3 distinct stock clusters.
- Cluster visualizations revealed:
  - Distinct separations based on stock price ranges and trade volumes.
  - Clear grouping of similar-performing stocks.
- Clusters indicate behavioral patterns of the data set.

7. CONCLUSION

This project successfully explored stock price data through three analytical approaches:

- **Regression Analysis:**
  Accurately predicted closing prices with an R^2 of 0.9997, confirming a strong linear relationship.
- **Time Series Analysis:**
  Revealed seasonality, trends, and smoothed fluctuations through decomposition and moving average.
- **Clustering Analysis:**
  Grouped similar stocks into three clusters, highlighting it's patterns in trading behavior and price similarity.

8. LIST OF FIGURES

Figure	File Name
Regression Plot	regression_plot.png
Time Series Decomposition	time_series_decomposition.png
Moving Average Plot	moving_average_smoothing.png
Elbow Method Plot	elbow_method.png
Open vs Close Clusters	kmeans_clusters.png
High vs Low Clusters	kmeans_high_low_clusters.png
Volume vs Close Clusters	kmeans_volume_close_clusters.png
Open vs Volume Clusters	kmeans_open_volume_clusters.png

End of the Report