MidSem Evaluation

Data Science Project

GitHub - https://github.com/himangg/Data-Science-Project

Group Name - HVN

Himang Chandra Garg, 2022214
Nishil Agarwal, 2022334
Vansh Aggarwal, 2022558

DATASET DESCRIPTION

Our dataset consists of 235,192 entries detailing daily trading and price information for NIFTY50-listed companies stocks. Key columns include Symbol (company ticker), Prev Close, Open, High, Low, Last, and Close prices, providing insights into daily price movements. VWAP (Volume Weighted Average Price) reflects the average trading price weighted by volume, while Volume and Turnover capture the total shares and value traded daily. Trades and Deliverable Volume give further trading insights, with %Deliverable showing the percentage of shares intended for delivery. We also had the data for the companies metadata corresponding to symbol and we joined both the tables. Also some symbols that were old were replaced with the new ones for the companies. For one company we did not have any data.

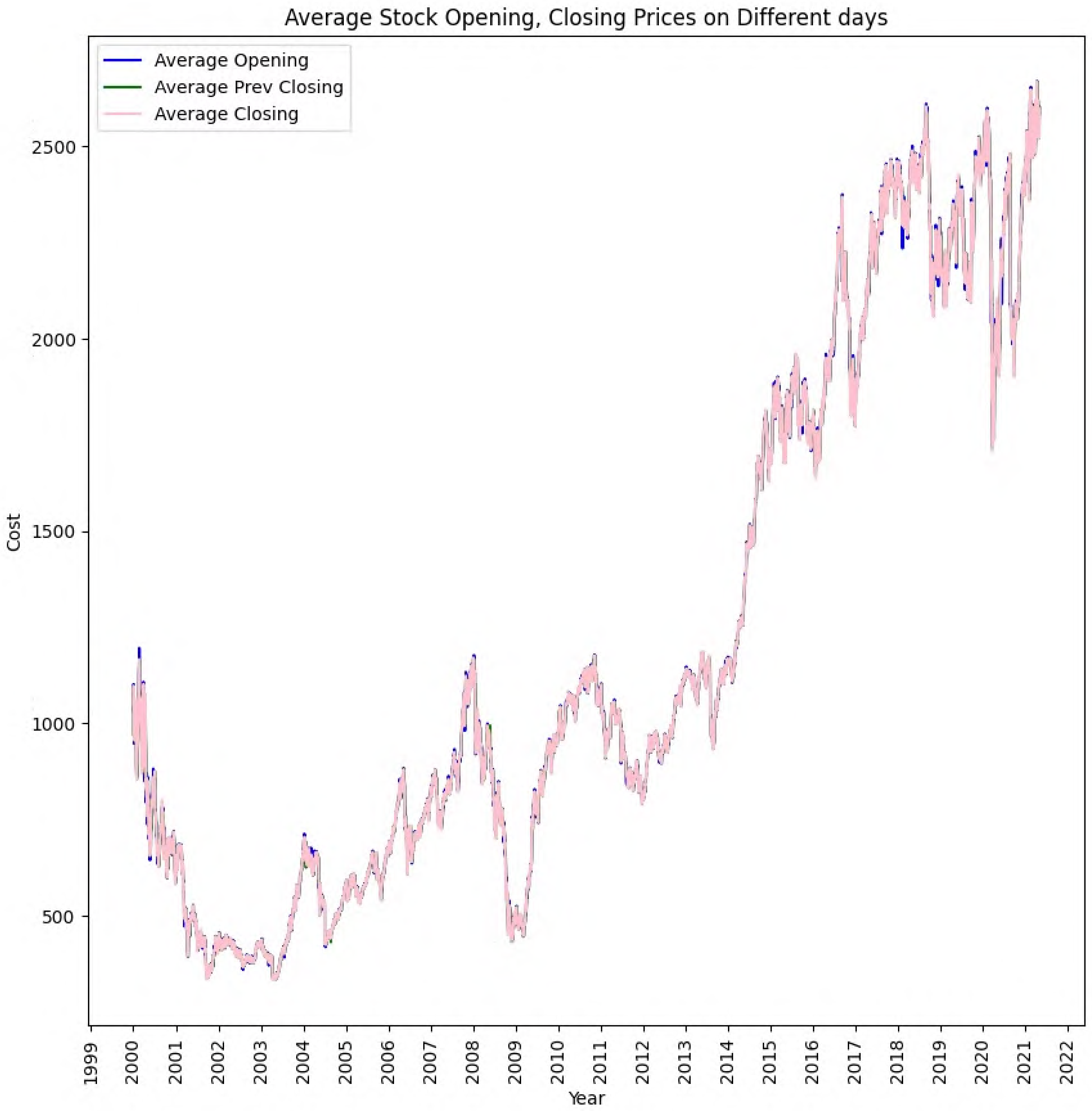
PROBLEM STATEMENT

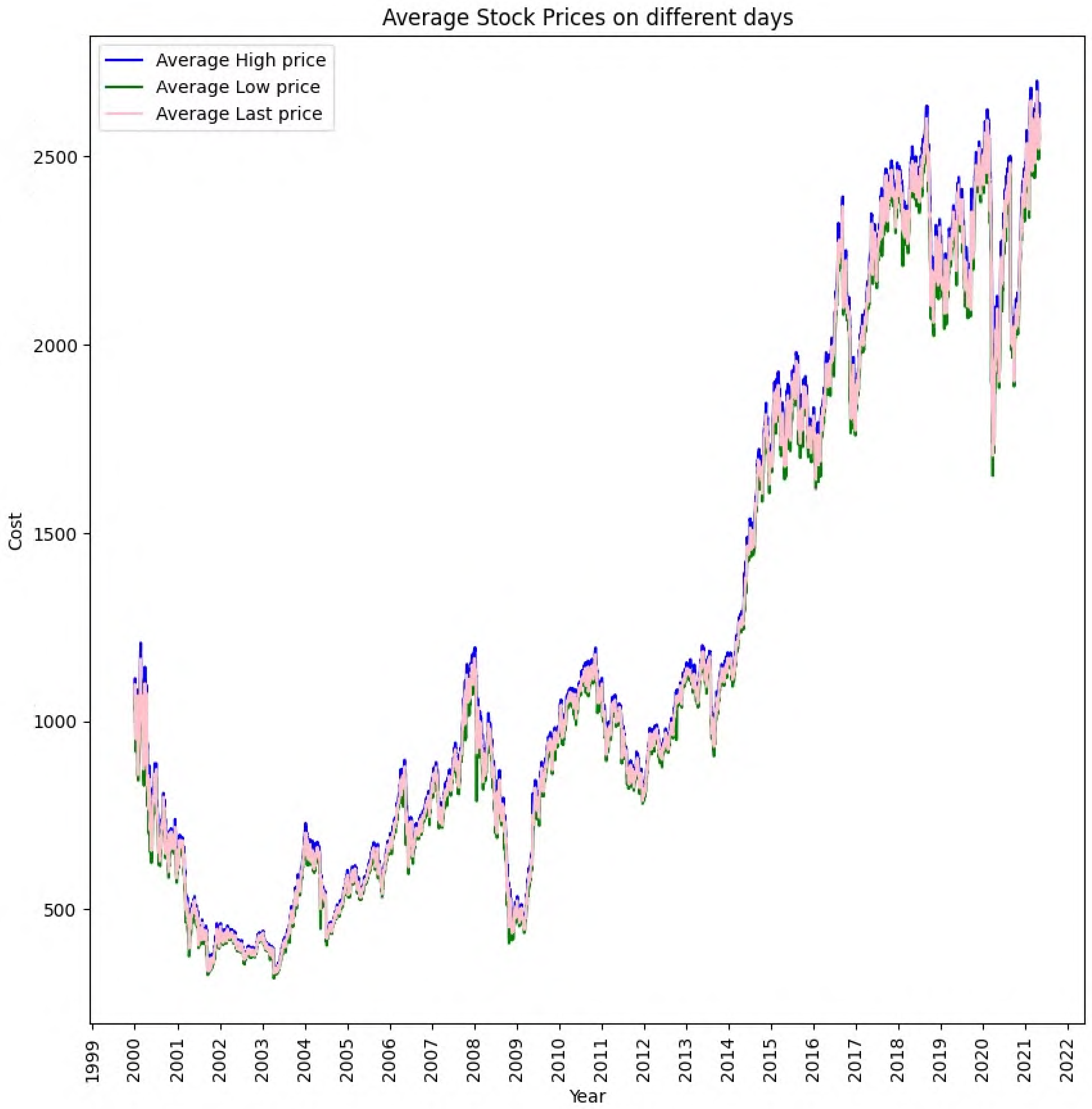
The goal of this project is to develop a predictive model for daily stock price movements of NIFTY50-listed companies. Specifically, the objective is to forecast closing prices or price trends based on historical market data, including past prices, trading volume, volatility, and moving averages. Understanding and accurately predicting these trends is crucial for investors and financial analysts to make informed decisions, optimize investment strategies, and manage risk effectively. Reliable price forecasts can help investors capitalize on favorable market trends and avoid losses from adverse movements.

Exploratory Data Analysis

PLOT OF STOCK PRICES

We plotted the stock opening, closing prices and low, high, last prices from different days respectively . This helped us identify if there exists some patterns in the trend of stock prices. Also it helped notice relationships between the 3 attributes.

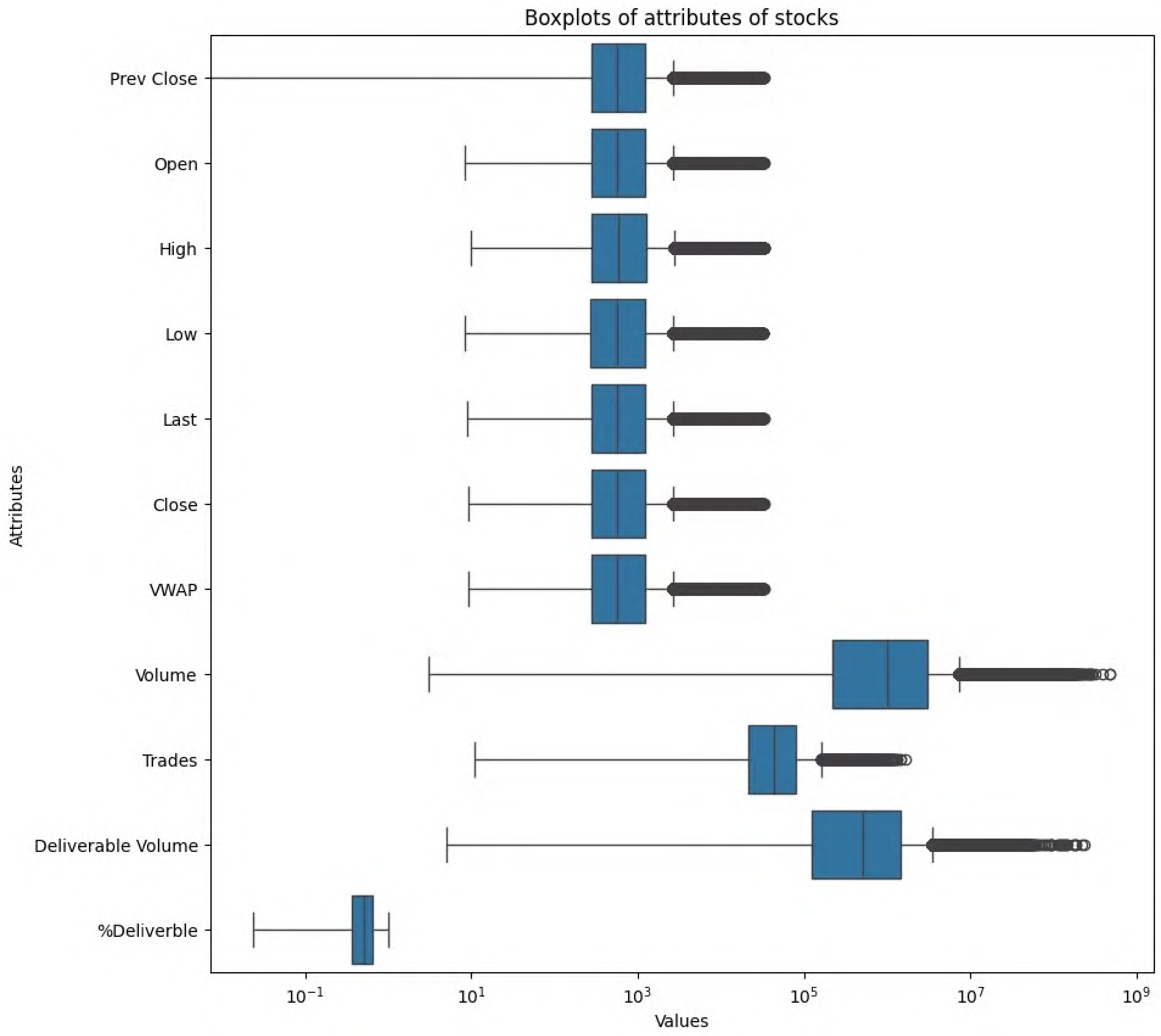


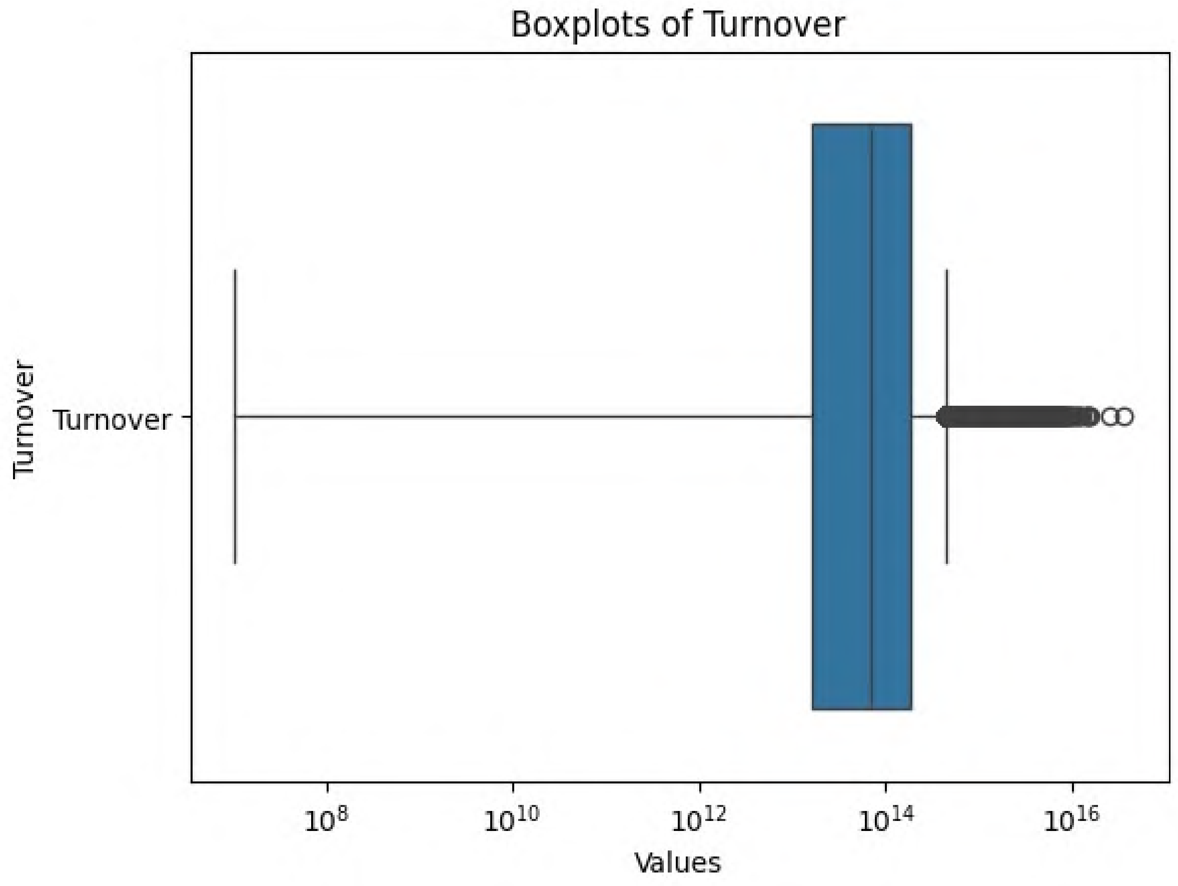


Try Pitch

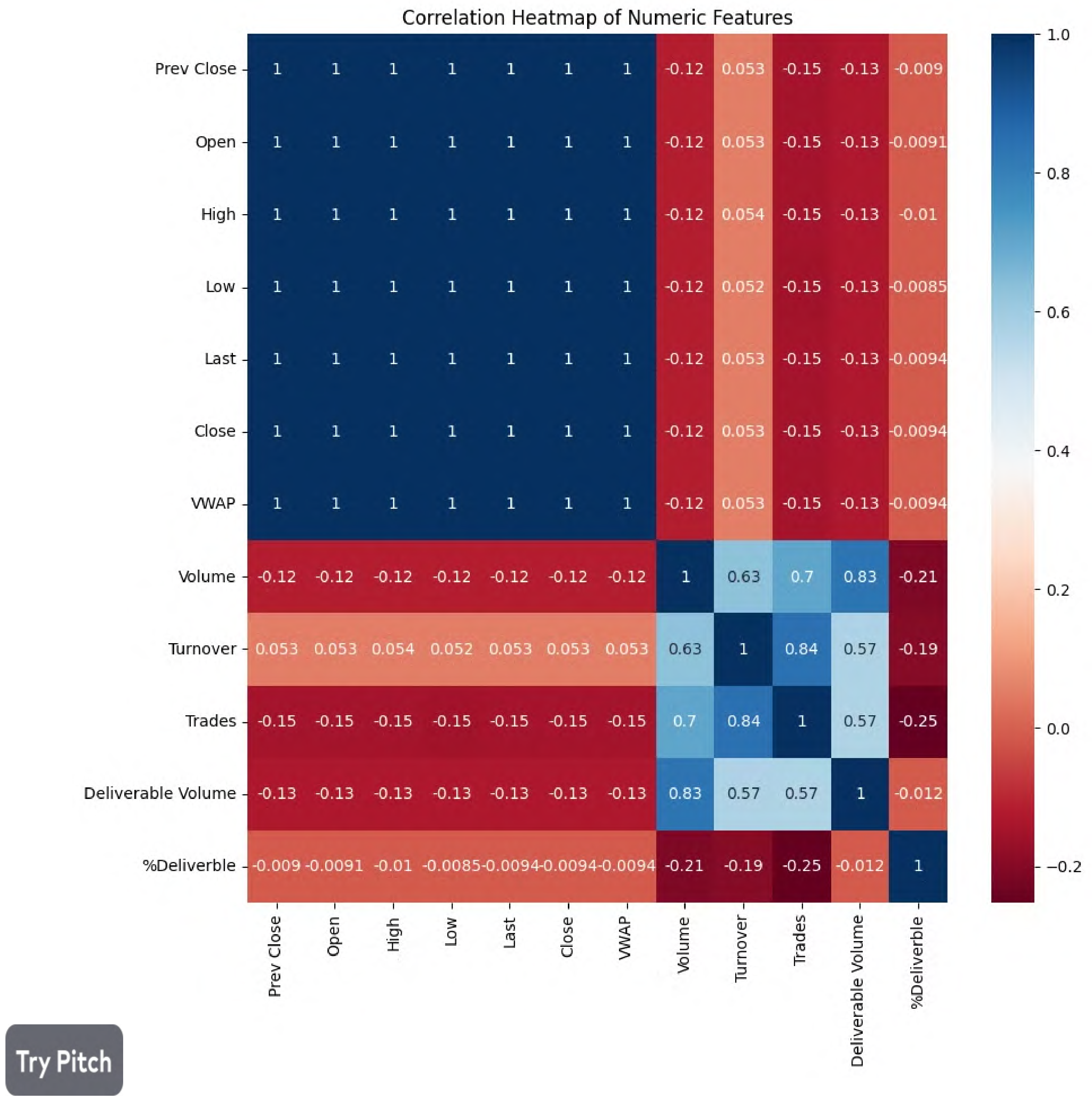
BOX PLOTS OF ALL ATTRIBUTES

We plotted box plots of all attributes so as to get to know the different range of values all of them cover. We also can see from the graph that we have some extreme outliers differing in magnitudes of powers of 10 from IQR!





HEATMAP OF CORRELATION BETWEEN ATTRIBUTES



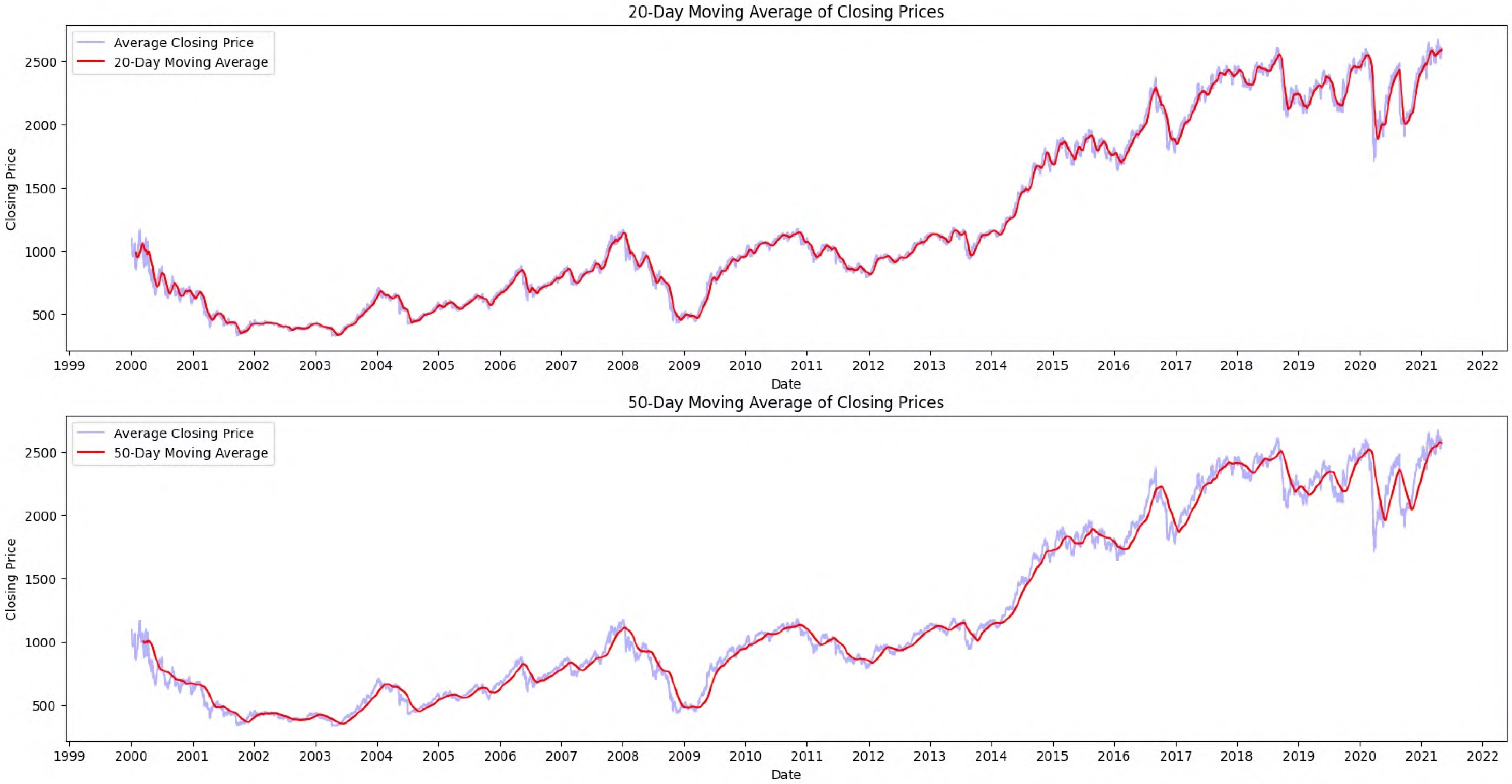
We plotted correlation heatmap between different attributes of the data. Thus, we managed to gain insight into the relationship between various attributes of different stocks.

Here we can see that a lot of features are extremely closely related to each other and would probably be redundant when training the model if used together.

We also can notice that Turnover and Volume are indeed closely related with correlation of 0.63.

PLOT OF ROLLING AVERAGES ACROSS 20/50 DAYS

We plotted graphs of rolling averages across 20 and 50 days respectively to smoothen the trend between stock price changes. This helped us generalize our insights and check for any patterns within the trend of data.



Data Preprocessing

HANDLING MISSING VALUES

Task

Impute missing values for features like ‘Trades’ and ‘Deliverable Volume’using linear interpolation.which estimates values based on adjacent points.

Why

Linear interpolation is particularly useful in time series because it maintains trends and continuity, making it more effective than simply filling with zeros, especially when data points change gradually over time.

Effect

Ensures a consistent dataset, improves model training by eliminating gaps, and avoids errors during analysis

DATE PROCESSING

Task

Convert date column to a datetime format and extract day, month, and year, as well as any day-of-week information.

Why

Enables seasonal and temporal analysis, helping identify patterns related to specific months, days, or quarters.

Effect

Supports seasonality analysis and may reveal recurring trends that can improve prediction accuracy.

OUTLIER DETECTION AND NORMALISATION

Task

Normalise continuous variables, especially trading volume and price

Why

Extreme values can distort the model's understanding of normal patterns and introduce noise. Normalisation brings all features to a similar scale, ensuring no single feature dominates due to magnitude.

Effect

Improves model stability and prediction reliability by focusing on meaningful patterns in the data.

FEATURE ENGINEERING

Task

Create new features such as moving averages (MA-20), returns, and volatility and add lagged features for past days' prices and volumes.

Why

These engineered features provide insight into past behavior and trends, which is crucial for time series prediction.

Effect

Enhances model's ability to detect patterns over time, improves forecast accuracy, and provides a richer feature set.

Hypothesis Testing

DAY-OF-THE-WEEK EFFECT ON PRICE CHANGES

Objective

Test if average returns vary across different days to identify any “day-of-the-week effect” in stock returns.

Hypothesis

Null Hypothesis (H0): Mean returns are the same across all days of the week.
Alternative Hypothesis (H1): Mean returns differ significantly across days.

Experiment Conducted

Returns were grouped by weekday and analyzed with ANOVA (Analysis of Variance)

Result

Since we reject the null hypothesis, this suggests that returns vary by day, indicating a potential “day-of-the-week” effect. This insight could help in time-aware prediction models by accounting for weekday-specific return patterns.

SEASONALITY IN MONTHLY RETURNS

Objective

Determine if there is a significant seasonality effect in monthly returns, which could impact stock price trends.

Hypothesis

Null Hypothesis (H0): Monthly returns are uniformly distributed (no seasonality).
Alternative Hypothesis (H1): Monthly returns are not uniformly distributed (seasonality exists).

Experiment Conducted

Returns were grouped by month and analyzed with Chi-Square Test for Goodness of Fit.

Result

We fail to reject the null hypothesis, indicating no statistically significant seasonality effect in monthly returns for the dataset analyzed.

PROPORTION OF DAYS WITH POSITIVE RETURNS

Objective

Determine if the proportion of days with positive returns differs significantly from a 50% benchmark.

Hypothesis

Null Hypothesis (H0): Proportion of days with positive returns is 0.5.
Alternative Hypothesis (H1): Proportion of days with positive returns is not 0.5.

Experiment Conducted

Z-Test for Proportions

Result

The null hypothesis is rejected, showing that the proportion of positive-return days significantly differs from 50%. This could imply a bias towards positive or negative returns, which is useful for modeling trends.

CORRELATION BETWEEN LAGGED VOLUME AND PRICE CHANGE

Objective

Test if the previous day’s trading volume correlates with the current day’s price change.

Hypothesis

Null Hypothesis (H0): No correlation between lagged volume and price change.
Alternative Hypothesis (H1): There is a correlation between lagged volume and price change.

Experiment Conducted

Pearson Correlation Test

Result

The null hypothesis is not rejected, suggesting no significant correlation. This indicates that previous day volume may not be a strong predictor of the next day’s price change in this dataset.

IMPACT OF VOLUME ON RETURNS

Objective

Determine if high-volume days yield significantly larger returns than low-volume days.

Hypothesis

Null Hypothesis (H0): No difference in returns between high and low-volume days.
Alternative Hypothesis (H1): High-volume days have greater average returns than low-volume days

Experiment Conducted

Independent T-Test

Result

We reject the null hypothesis, suggesting that high-volume days are associated with larger returns.
This insight supports the inclusion of volume as a feature in prediction models.

CORRELATION BETWEEN VOLUME AND DELIVERABLE VOLUME

Objective

Test if day’s trading volume correlates with the Deliverable Volume.

Hypothesis

Null Hypothesis (H0): No correlation between the day's trading Volume and the Deliverable Volume.

Alternative Hypothesis (H1): There is significant correlation between Volume and the Deliverable Volume.

Experiment Conducted

Pearson Correlation Test

Result

We reject the null hypothesis, suggesting that there is significant correlation between Volume and the Deliverable Volume.

CORRELATION BETWEEN TRADES AND TURNOVER

Objective

Test if day’s Trades correlates with the Turnover.

Hypothesis

Null Hypothesis (H0): No correlation between the day's Trades and Turnover.
Alternative Hypothesis (H1): There is significant correlation between Trades and Turnover.

Experiment Conducted

Pearson Correlation Test

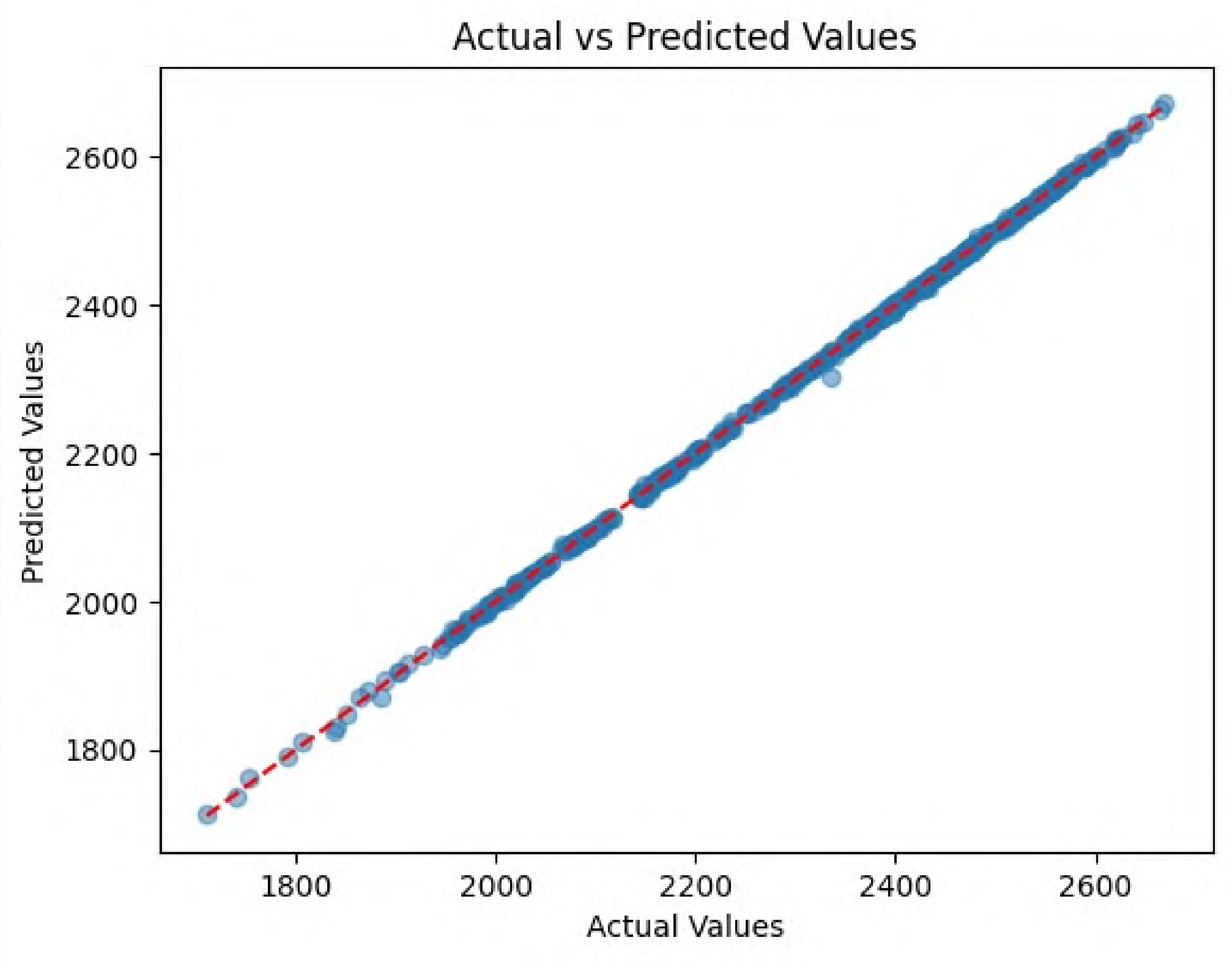
Result

We reject the null hypothesis, suggesting that there is significant correlation between Trades and Turnover.

Model Description

MODEL USED : LINEAR REGRESSION

We have used Linear Regression to predict stock Closing values in our project.



Linear Regression is a simple model which fits a linear equation to the data in order to predict target values.

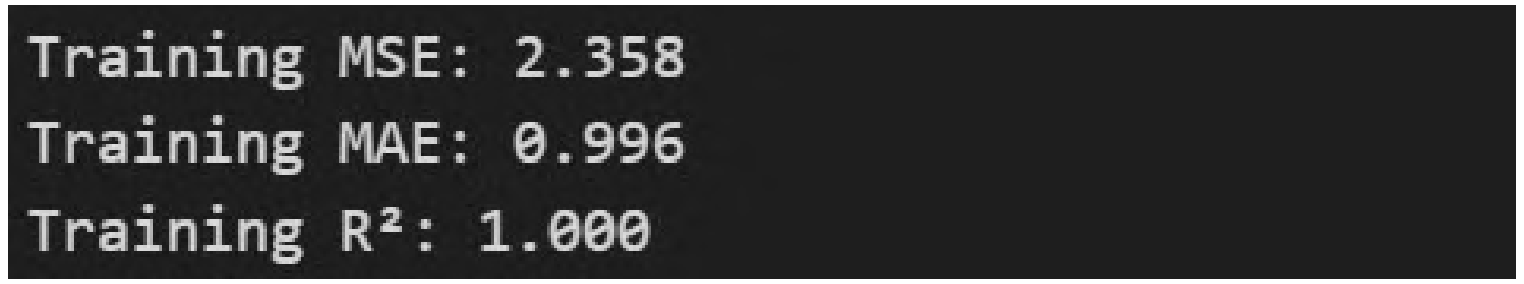
Stock prices often exhibit linear trends over short time periods. Linear regression can capture trends effectively and help predict short-term price changes.

It can help make an informed decision on when to buy stock to have a profit at the end of the day.

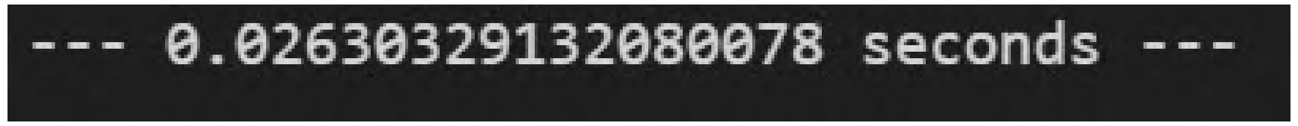
MODEL FORMATION : LINEAR REGRESSION

When using the model without any scaling techniques, we are feeding the model attributes like Open, Trades, Volume etc. to predict the Closing price of the stock.

We are checking the accuracy of our model through metrics like MAE, MSE and R2 score which were received as follows :



```
Training MSE: 2.358
Training MAE: 0.996
Training R²: 1.000
```



```
--- 0.02630329132080078 seconds ---
```

The model got trained in only 0.027 seconds due to linear regression's lightweight nature.

The time complexity of linear regression is O(np^2 + p^3). The formula for it is →

where n=number of data points and p = number of features

$$\beta = (X^T X)^{-1} X^T y$$

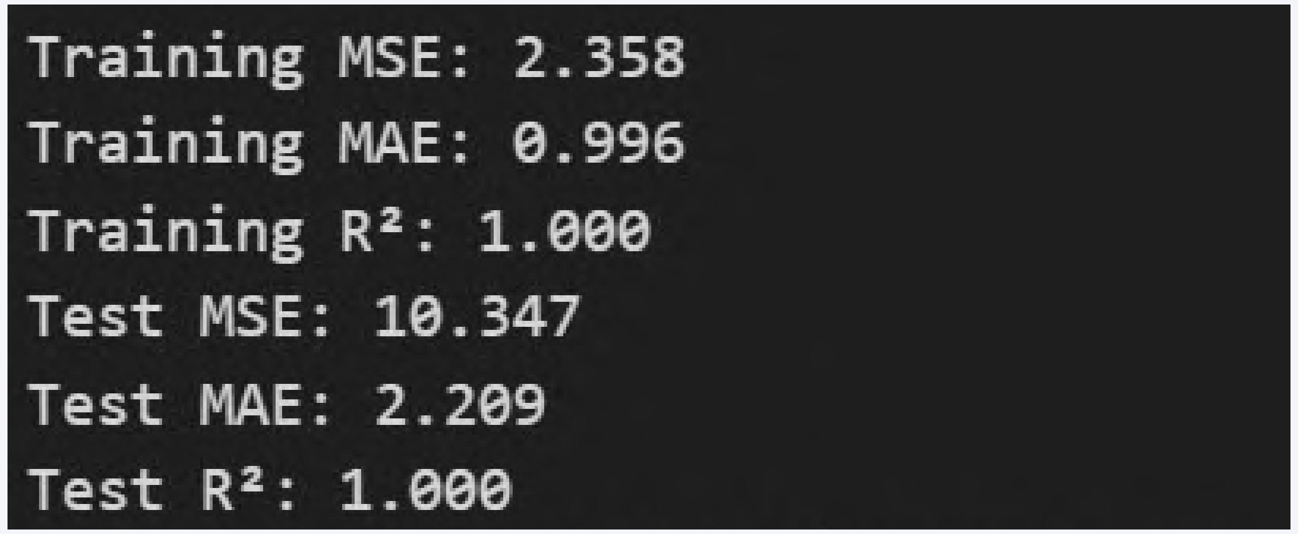
Matrix multiplication takes O(np^2) and matrix inversion takes O(p^3)

MODEL REFINEMENT : LINEAR REGRESSION

To ensure our model is generalized we have used TimeSeriesSplit instead of cross validation to generate Train and Test splits.

This was done as Stock Prediction data is sensitive to data ordered by date and we needed to maintain the continuity of data while making these splits.

The Splitting of data was necessary to ensure that the model is generalized and not overfitting.



```
Training MSE: 2.358
Training MAE: 0.996
Training R²: 1.000
Test MSE: 10.347
Test MAE: 2.209
Test R²: 1.000
```

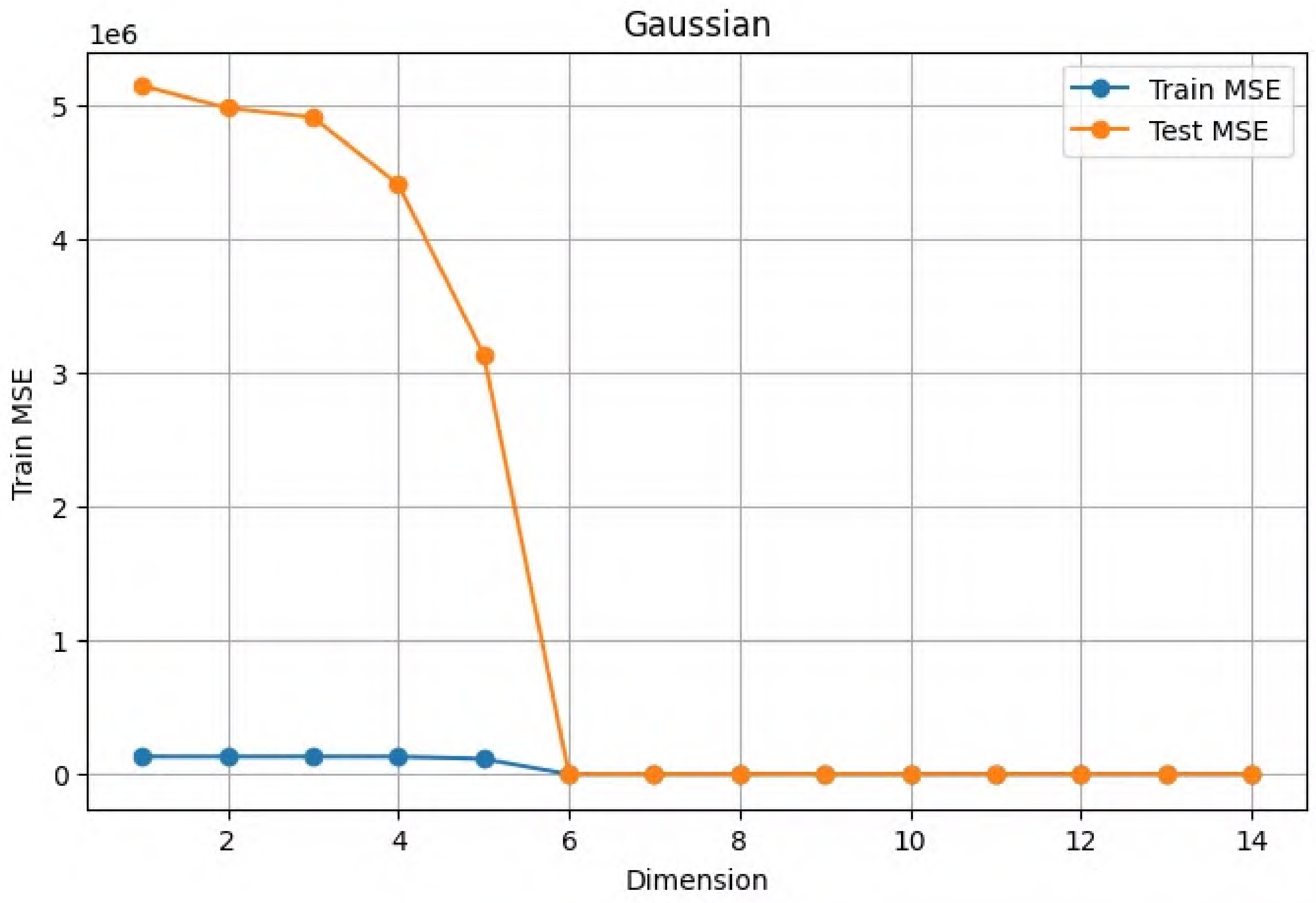
Scaling Techniques

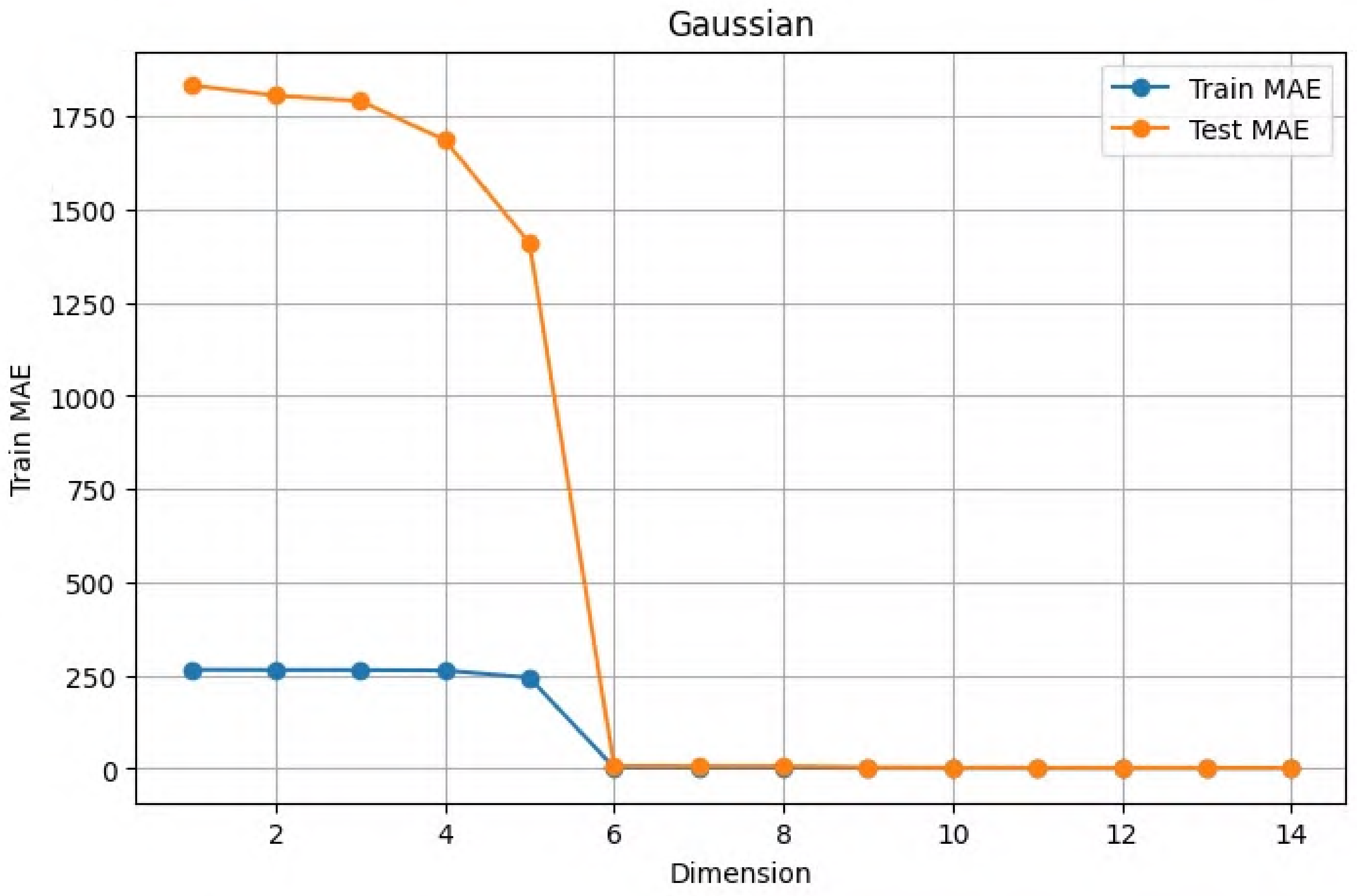
JL LEMMA

The Johnson-Lindenstrauss (JL) lemma performs dimensionality reduction while preserving pairwise distances. We have used both **Gaussian Random Projection** and **Sparse Random Projection**. GRP uses a dense random matrix with Gaussian-distributed entries which gives high accuracy but at a higher computational cost. In contrast, SRP uses a sparse matrix with mostly zero entries which improves the efficiency for large datasets. The formula used for JL lemma no. of reduced dimension calculation was giving a larger dimension than the original dimension of 15. Thus, we have set the number of reduced dimensions(k) manually and have checked it for all the **possible values of k** and plotted the results and found the appropriate vale of k using the elbow method in the graph.

GAUSSIAN RANDOM PROJECTION

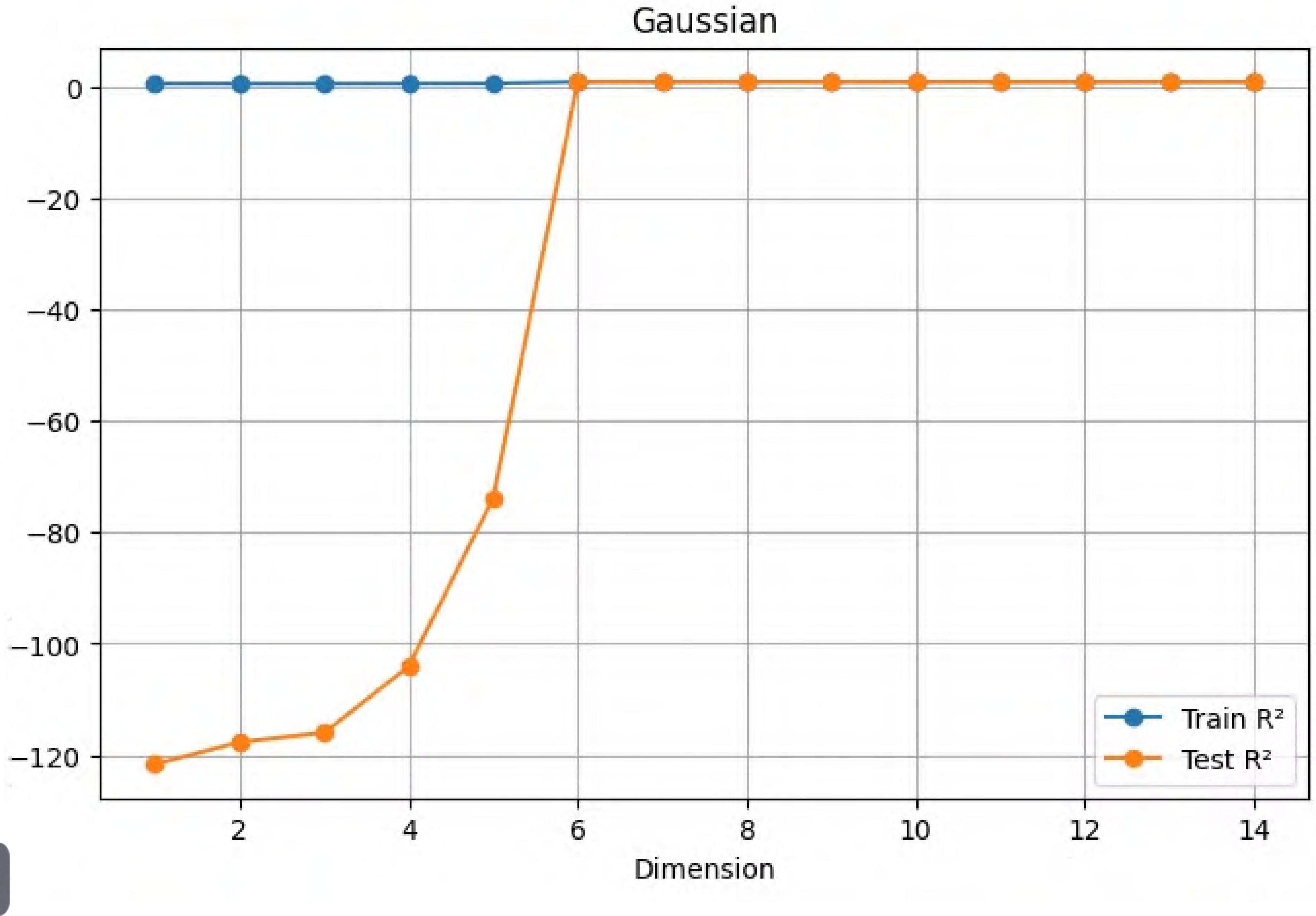
Gaussian Random Projection uses a dense random matrix with Gaussian-distributed entries. We have plotted a graph of train MSE, test MSE, train MAE and test MAE vs the number of dimensions as shown below.



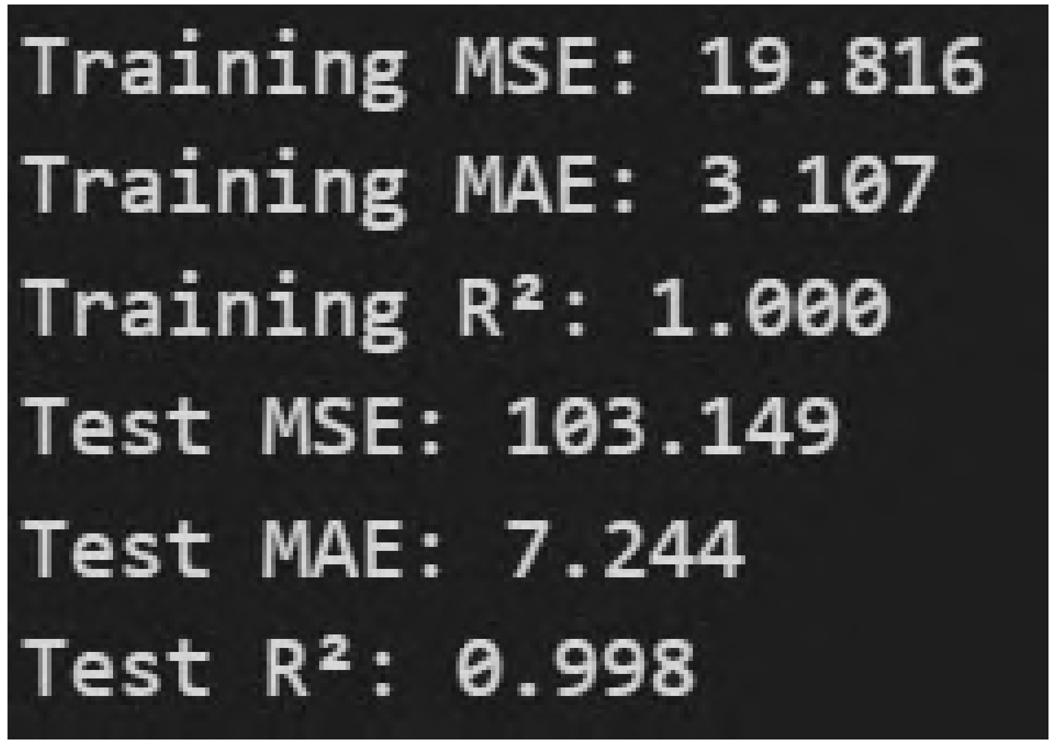


GAUSSIAN RANDOM PROJECTION

We have also plotted a graph of train R2 score and test R2 score vs the number of dimensions. We can clearly observe that the elbow point was found at k=6 as after that the value of test R2 become almost equal to 1.



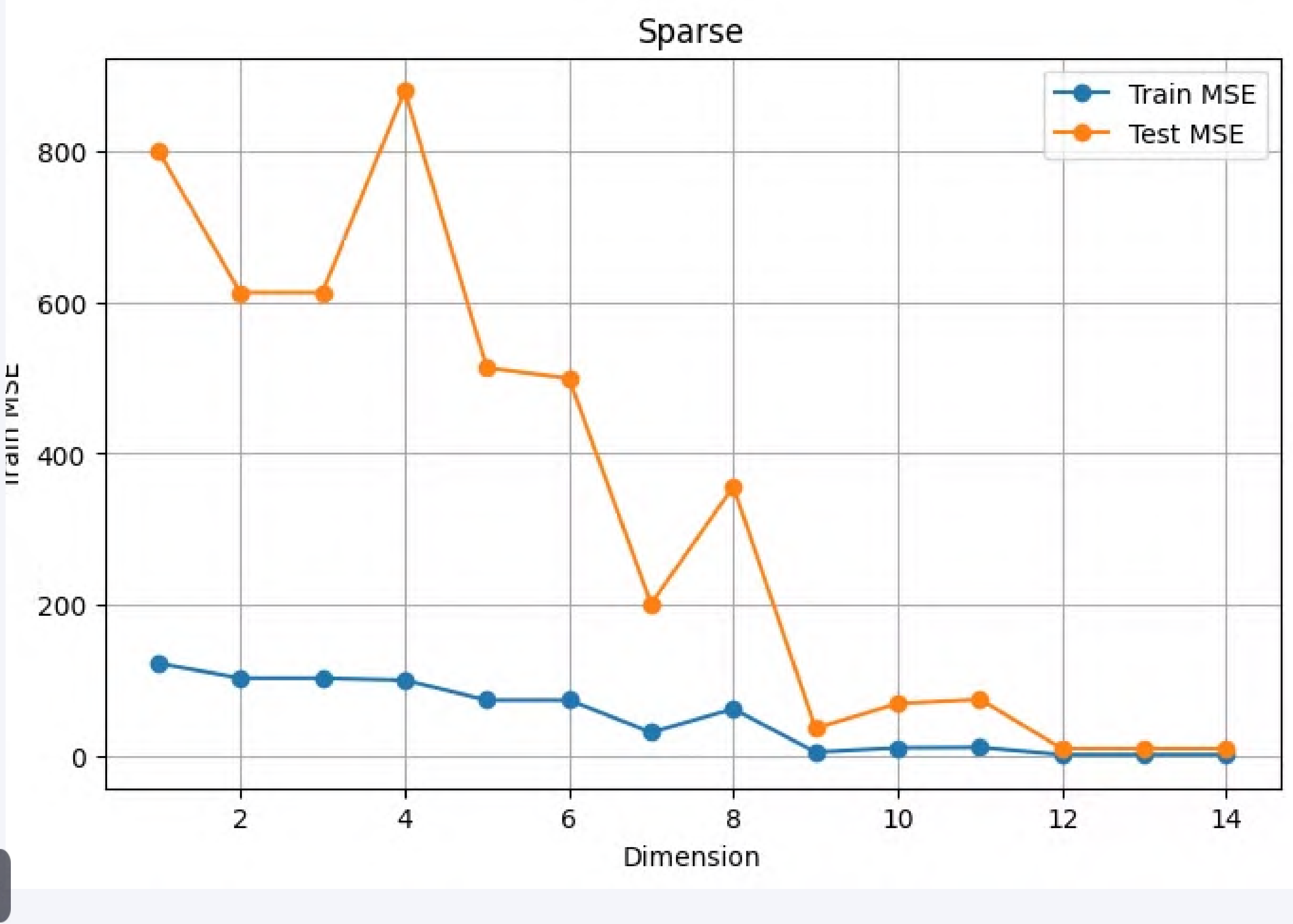
Below are the results that we observed for running our model for k=6:

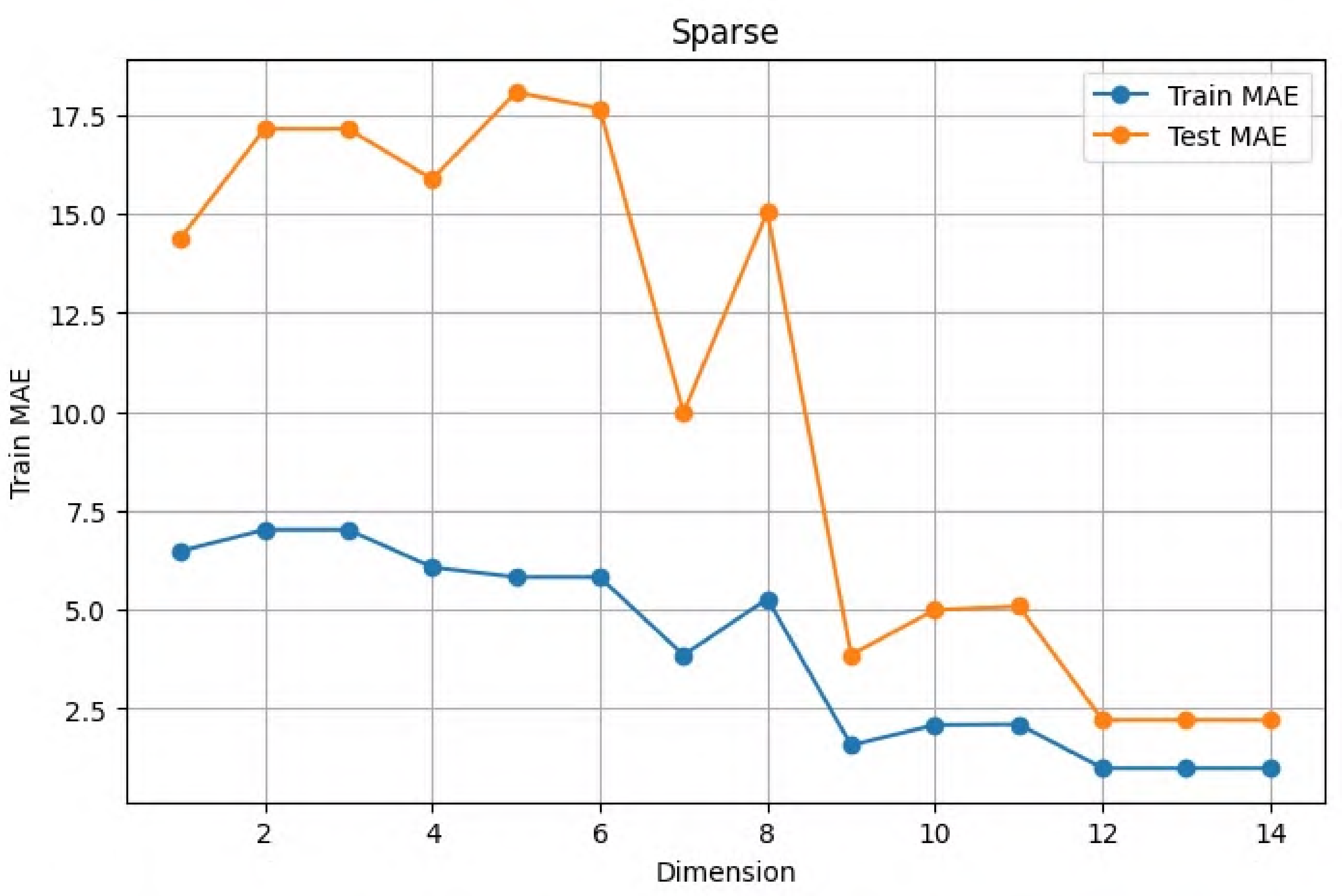


```
Training MSE: 19.816
Training MAE: 3.107
Training R²: 1.000
Test MSE: 103.149
Test MAE: 7.244
Test R²: 0.998
```

SPARSE RANDOM PROJECTION

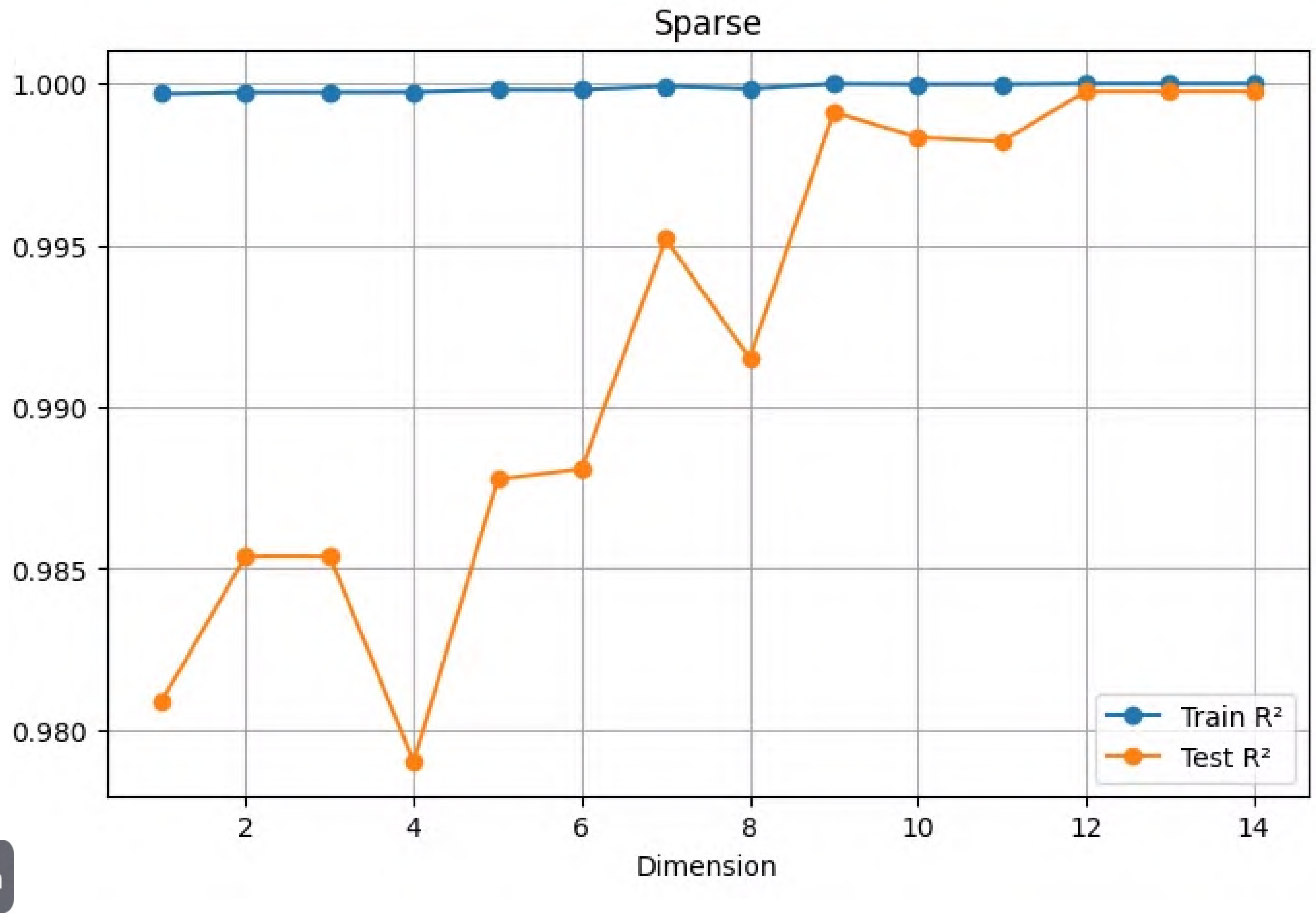
Sparse Random Projection uses a sparse matrix with mostly zero entries. We have plotted a graph of train MSE, test MSE, train MAE and test MAE vs the number of dimensions as shown below.



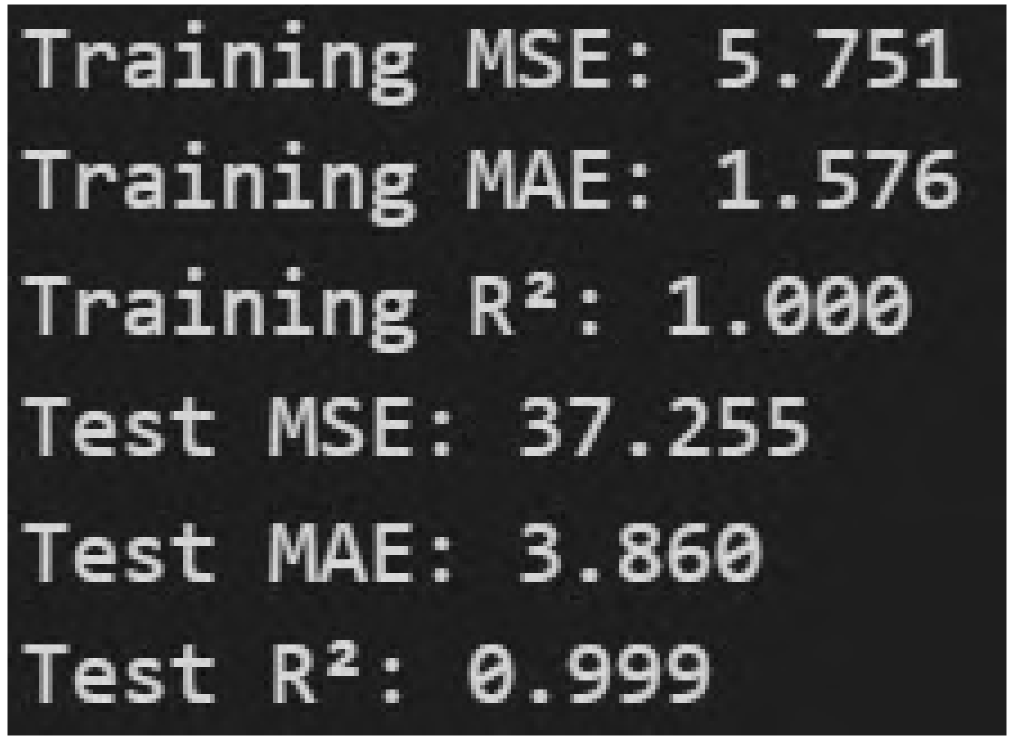


SPARSE RANDOM PROJECTION

We have also plotted a graph of train R2 score and test R2 score vs the number of dimensions. We can clearly observe that the elbow point was found at k=9 as after that the value of test R2 almost becomes stagnant after that.



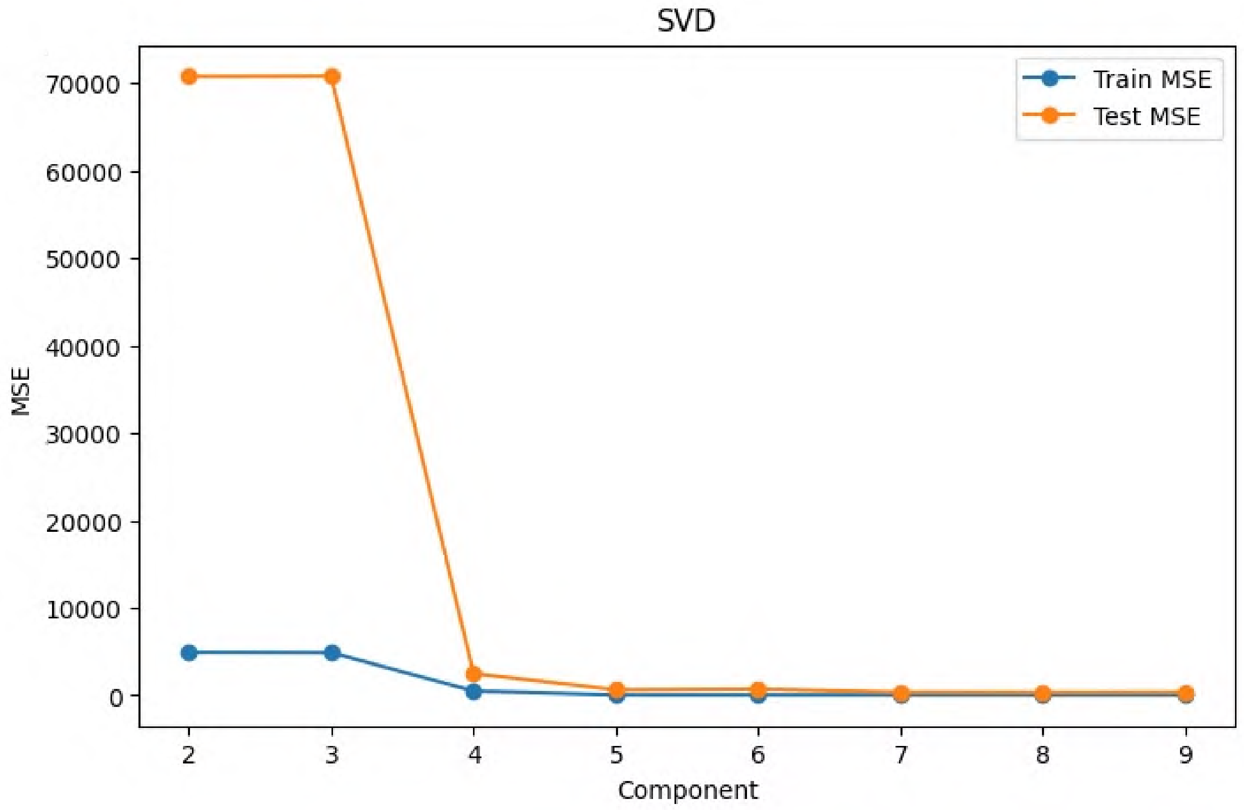
Below are the results that we observed for running our model for k=9:



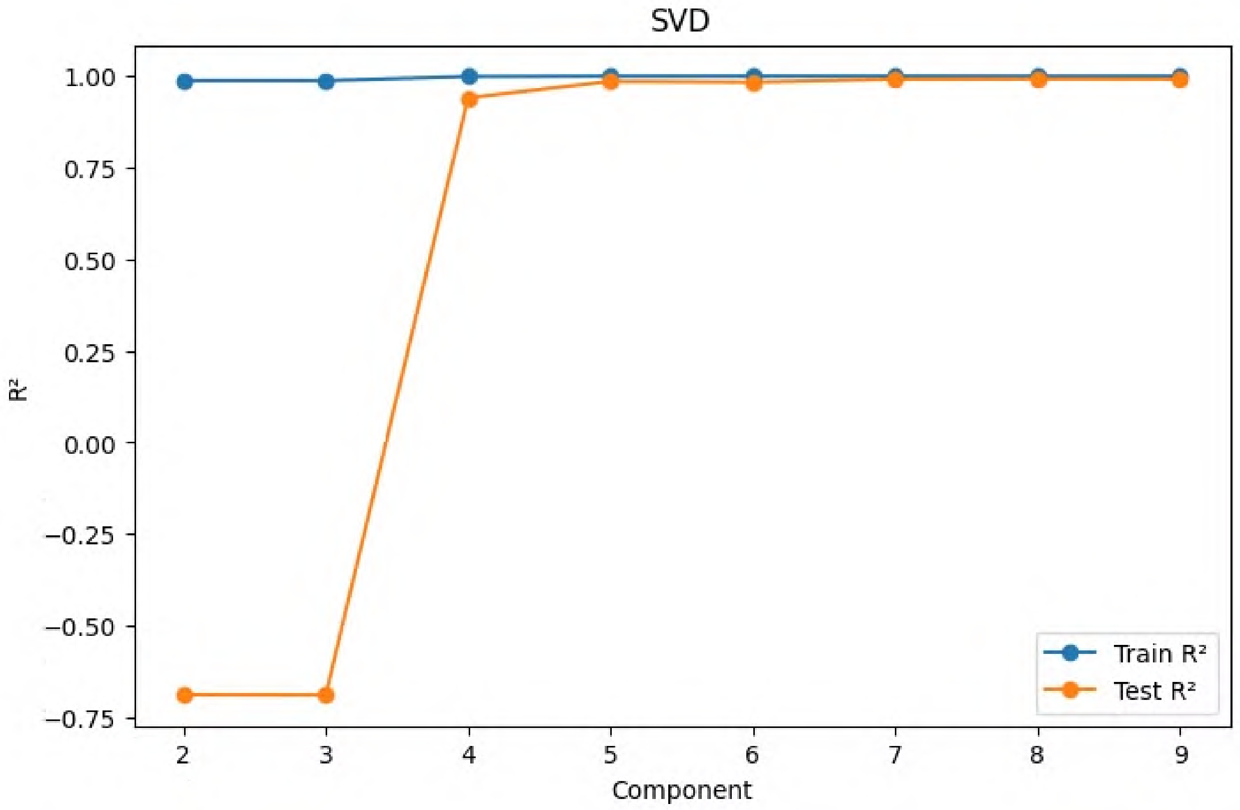
```
Training MSE: 5.751
Training MAE: 1.576
Training R²: 1.000
Test MSE: 37.255
Test MAE: 3.860
Test R²: 0.999
```

SVD

SVD is a mathematical technique used to decompose a matrix into components (singular vectors and singular values). It is effective in reducing the dimensions of a dataset while preserving as much information as possible. By keeping only the top components, we can remove redundant or less significant features.
Features were highly correlated, SVD resolved multicollinearity by transforming the features into an orthogonal space.



As we can see that even 4 components of SVD could easily make stock predictions with a good R2 score and low MSE comparable to our original model.



CONCLUSION

This project focused on stock price prediction using a combination of dimensionality reduction techniques and Linear Regression. By leveraging methods like Gaussian Random Projection for feature scaling, the model achieved efficient processing and reasonable prediction accuracy. The results demonstrated the importance of selecting meaningful features and still achieving a similar amount of accuracy.

FUTURE WORK

Linear regression is limited to short-term trends and is not suitable for long-term predictions or highly volatile markets due to its inability to capture complex patterns. Its a light weight model and can't be used to effectively predict long term price changes which might help in getting better profits.
So in future we may use more sophisticated models like RNN or LSTMs for long term Stock prediction.

Thank You