**Stock market Trend Analysis**

**Statistical/Hypothetical Question:** Our primary objective of this analysis is to understand various stock market trends that will help to determine the relationship between the variables of the stock market open prices and close prices for the ticker along with the high, low and trading volume. I have conducted the regression analysis, identifying outliers, computing PMF, CDF, analyze distribution for the dataset , correlation, regression and liner regression analysis, hypothesis testing is conducted to find whether there is a statistically significant difference between opening and closing prices of the stocks. So understanding the relation between various stock variables is important for the investors and stock traders to get the informed decisions and develop predictive models for the stock price movements.

**Outcome of exploratory data analysis:**

Our dataset from the Kaggle has flat file has the historical S&P 500 dataset. This data source from Kaggle has a huge flat file with data such as data, open, high, low, close, volume, name. Here the name indicates the ticker name.

Link: <https://www.kaggle.com/datasets/camnugent/sandp500>

EDA done in this project included with the histograms, descriptive analysis, outlier detection , Cumulative distribution function, probability mass functions, correlation matrices, scatter plots, regression analysis. Our dataset contains 619,029 records and we have selected the five variables Open, Close, High, Volume, Low.

A screenshot of a computer screen

AI-generated content may be incorrect.

Correlation between open and the close variables for the stock trend data analysis is 0.99987 which is nearly a linear relationship.

**A number of numbers and a number of numbers

AI-generated content may be incorrect.**

Our Regression analysis confirmed that the Opening prices, high prices , low prices were strong predictors of the closing prices as the R squared value is almost to 1.

A screenshot of a computer

AI-generated content may be incorrect.

The presence of the outliers in Open, Close, High, Low, volume variables is suggested that the stock market is highly volatile and this can cause the less accuracy of the prediction model.

Correlation in the volume with closing prices is of the negative value suggests that the trading volume alone doesn’t impact the pricing changes. We need the additional variables.

**Missed aspect with analysis:**

Out dataset analysis with the market stock trend covers the major statistical measures, but additional measures such as the external economic indicator, market trend and other company specific news, social trends such as Twitter, FB or other major social media platforms trends are not analyzed in this analysis. Time bases trend analysis also not concluded in our data analysis if we could have used that in our analysis then this could have given us with the deeper insights into the stock price fluctuations over the different time periods.

Also adding the seasonal trends and daily trending patterns could have given us with the cyclic trends with the stock and this would have strengthened our analysis.

If we add the social trends such as FB, Twitter etc. that would have given us with an insights for how the external factors affects the stock market trends.

**Additional Variables :**

Incorporating additional variables such as moving averages, earning reports, relative strength index could have improved our analysis. Also the macroeconomic factors such as inflation rates, S&P 500 trends, interest rates might have improved our analysis if these parameters are considered with additional datasets in our analysis.

Also the market fundamentals such as earnings per share, company revenue, big investors or other new investment projects in the company. Price to earnings ratio per share could have provided the broader view in our analysis.

**Incorrect Assumptions:**

One of the biggest assumptions made here is that stock market would follow the normal distribution.

A blue line with a red line

AI-generated content may be incorrect.

But the Q-Q plot suggest that there are some deviations which suggests that stock market may not be perfectly the normal. So we need additional distributions. Additionally we assumed that the linear relationship between volume and closing price may be incorrect as we noticed that the volume shows weak or negative relationship with the correlation with price movements.

We have also assumed that the stock market behavior remains consistent with multiple tickers. Difference companies stocks follows various trends as it will not be consistent with one another so grouping all these stocks together might be obscured individual differences.

**Challenges faced, Uncertainties:**

As we see in the outliers results this is the hard one to handle as there are extreme fluctuations in the stock market on any particular day. Our selected variable outliers outcome also confirmed that the high market volatility making the prediction of the model harder. Another challenge is with the negative correlation for the volume with the prices changes indicates it is weak. So understanding the best approach for our regression modeling is giving the high collinearity among Open, High, Low, close prices and this is indicating the difficulty level. Also differentiating causality and correlation is complex here. So we need additional variables are required for the causation to more effective.

**Conclusion:**

Overall our analysis provided insights into the stock market price movements but it could be expanded with the additional factors.

Including additional external factors and variables are the key successes for the accuracy in the prediction model. Additionally exploring various regression model for linear, collinear in stock price movements can improve the predictability. My future work on this will be adding more datasets and machine learning techniques which could influence the predictability of our model