DOW Jones Industrial Average Index Analysis

A Comprehensive Exploration of DOW Jones Performance through Portfolio and Time Series Analysis

By:

Ali Irtaza - G47541925

Rahul Modi - G29861398

Sri Sankeerth Koduru - G42612138

EMSE 6574

December 18,2023

Table of Contents

Introduction:

1.1 Abstract:

This project looks to examine the relationships between individual stocks and the DOW Jones index, aiming to unravel the intricate dynamics that influence overall market fluctuations. Through various analysis methodologies, we seek to identify underlying patterns and correlations, providing valuable insights into the factors shaping market performance. Taking our investigation, a step further, we utilize the LSTM-based time series analysis, our focus extends to forecasting DOW Jones index movements, offering a sophisticated perspective for investors and analysts. This inquiry aspires to contribute to a nuanced understanding of market behavior, fostering informed decision-making in the dynamic realm of financial markets.

1.2 Data Extraction:

To construct the dataset, we initiate the process by extracting the most recent DOW Jones tickers from the source "https://stockmarketmba.com/stocksinthedjia.php" using the 'BeautifulSoup' library. This information is then organized into a comprehensive list. Subsequently, leveraging the 'yfinance' library, we proceed to retrieve stock values dating back to September 1, 2020. This specific start date is chosen for its proximity to the latest index modifications, ensuring we capture relevant changes in the composition of the index. Implementing a loop on the compiled list of tickers and utilizing the pdr.get_data_yahoo() function, we obtain the values of each stock, storing them temporarily in a dataframe. We focus on extracting the adjusted close prices for each stock, creating a structured dataframe named 'df' with columns labeled [‘ticker’ + _close]. This organized dataframe proves versatile for subsequent analysis and can be effortlessly exported to a CSV file. Furthermore, we extend our data gathering efforts to encompass historical data for the DOW Jones Industrial Average (DJI), spanning from January 3, 1984. This historical dataset serves as a crucial foundation for the application of Long Short-Term Memory (LSTM) in time series forecasting, enhancing the depth and context of our analytical endeavors.

1.3 Data Dictionary:

1.3.1 Stocks_in_DJI.csv:

Column Name	Description
Date	Date of stock price
UNH_close	This is the adjusted closing price for UnitedHealthGroup Inc.
MSFT_close	This is the adjusted closing price for Microsoft Corp.
GS_close	This is the adjusted closing price for GoldmanSachs Group Inc.
HD_close	This is the adjusted closing price for Home DepotInc.
MCD_close	This is the adjusted closing price for McDonald’s Corp.
AMGN_close	This is the adjusted closing price for Amgen Inc.
V_close	This is the adjusted closing price for Visa Inc.
CAT_close	This is the adjusted closing price for Caterpillar Inc.
CRM_close	This is the adjusted closing price for Salesforce Inc.
BA_close	This is the adjusted closing price for Boeing Co.
HON_close	This is the adjusted closing price for Honeywell International Inc.

AAPL_close	This is the adjusted closing price for Apple Inc.
TRV_close	This is the adjusted closing price for Travelers Co. Inc.
AXP_close	This is the adjusted closing price for AmericanExpress Co.
WMT_close	This is the adjusted closing price for Walmart Inc.
IBM_close	This is the adjusted closing price for Intl BusinessMachines Corp.
PG_close	This is the adjusted closing price for Procter &Gamble Co.
JPM_close	This is the adjusted closing price for JPMorganChase & Co.
JNJ_close	This is the adjusted closing price for Johnson &Johnson
CVX_close	This is the adjusted closing price for Chevron Corp.
NKE_close	This is the adjusted closing price for Nike Inc.
MRK_close	This is the adjusted closing price for Merck & Co. Inc.
MMM_close	This is the adjusted closing price for 3m Co.
DIS_close	This is the adjusted closing price for Walt Disney Co.
KO_close	This is the adjusted closing price for Coca-Cola Co.
DOW_close	This is the adjusted closing price for Dow Inc.
CSCO_close	This is the adjusted closing price for Cisco Systems Inc.
INTC_close	This is the adjusted closing price for Intel Corp.
VZ_close	This is the adjusted closing price for VerizonCommunications Inc.
WBA_close	This is the adjusted closing price for WalgreensBoots Alliance Inc.
^DJI_close	This is the adjusted closing price for DOW JonesIndustrial Average Index (Market Value)

1.3.2 DJI_Stock_Since_Start.csv:

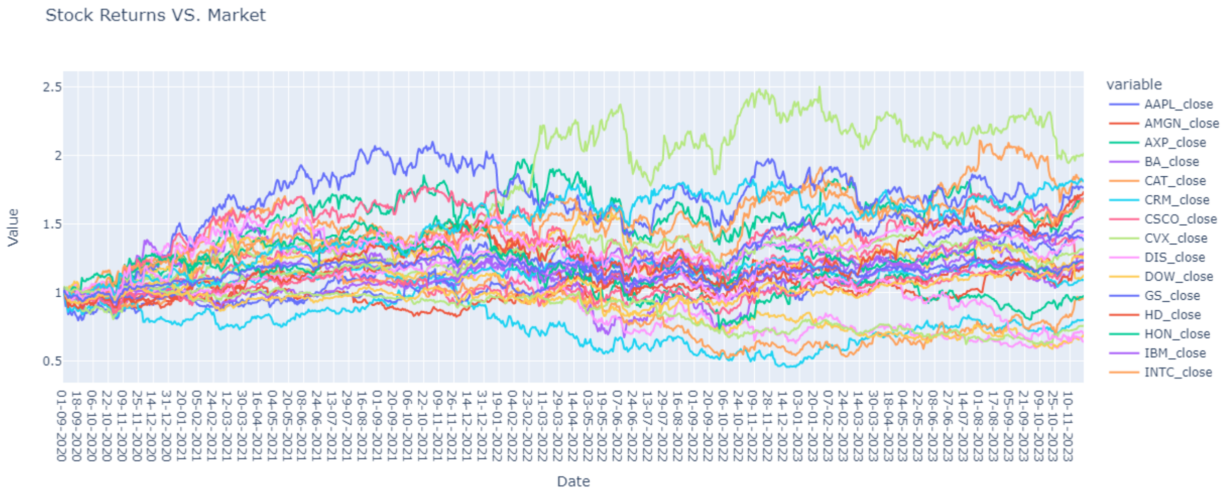
Column Name	**Description**
Date	Date of stock price
Open	This is the opening price for DJI on that day.
High	This is the highest price for DJI on that day.
Low	This is the lowest price for DJI on that day.
Close	This is the closing price for DJI on that day.
Adj_Close	This is the adjusted closing price for DJI on that day.
Volume	Volume of index (stock)

Portfolio Analysis:

2.1 Shares vs Dow Jones Industrial Average:

The Dow Jones Industrial Average has exhibited a cumulative growth rate of 1.2363820652824533, reflecting the overall percentage increase in its value over a specified period. This metric serves as a comprehensive indicator of market performance. Analysis of individual stock performance in comparison to the index provides insights into relative strengths and weaknesses. Notable performers contributing positively to the index include Honeywell International Inc. (HON: 1.2329), Visa Inc. (V: 1.2132), Procter & Gamble Co. (PG: 1.1939), and The Home Depot Inc. (HD: 1.1835). Conversely, some stocks such as Walt Disney Company (DIS: 0.6926), 3M Company (MMM: 0.6876), and Walgreens Boots Alliance Incorporated (WBA: 0.6404) have exhibited growth rates below the index, prompting further investigation into their impact on the overall index performance. Image 1 visually depicts the performance of these shares in relation to the Dow Jones Industrial Average.

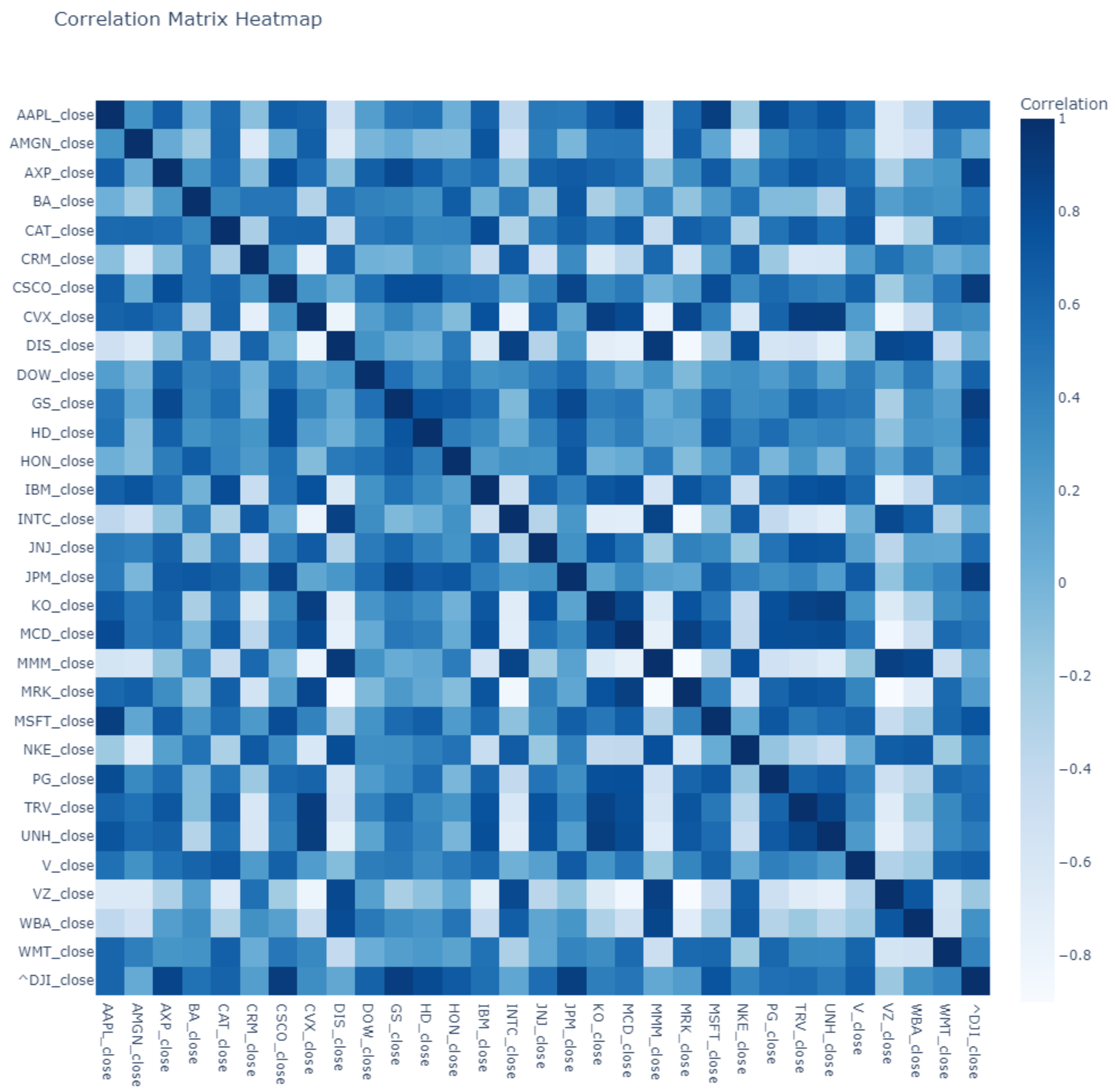
Figure 1: Shares vs Dow Jones Industrial Average



2.2 Correlation:

Certain companies demonstrate strong positive correlations with the market, as evidenced by their correlation coefficients with the Dow Jones Industrial Average Index. Noteworthy among them are American Express (AXP: 0.8544), Cisco Systems (CSCO: 0.9135), Goldman Sachs (GS: 0.9013), The Home Depot (HD: 0.8089), and JPMorgan Chase (JPM: 0.8935). These companies tend to align closely with overall market trends. Conversely, a group of companies exhibits low or negative correlations with the market, including Amgen (AMGN: 0.0852), Salesforce.com (CRM: 0.1872), The Walt Disney Company (DIS: 0.1071), Intel (INTC: 0.1145), 3M Company (MMM: 0.1001), Merck (MRK: 0.2031), and Verizon (VZ: -0.1752). These companies display more independent movements, potentially offering diversification benefits to an investment portfolio. Notably, Verizon (VZ) exhibits a negative correlation, suggesting an inverse behavior in comparison to the market. Further scrutiny of Verizon's performance is warranted, as it may exert an influence on the index, potentially posing challenges or acting as a hindrance to its overall performance. Image 2 below provides a visual representation of the correlation among the shares, as well as the correlation between the shares and the index.

Figure 2: Correlation Matrix



2.3 Market Beta:

The first group of stocks, with beta values greater than 1, indicates higher volatility, presenting opportunities for increased returns during market upswings but heightened risk during downturns. This consists of shares such as Boeing (BA, 1.609), American Express (AXP, 1.489), Salesforce (CRM, 1.342), Disney (DIS, 1.277), Goldman Sachs (GS, 1.271), Intel (INTC, 1.267), Apple (AAPL, 1.266), Nike (NKE, 1.251), Caterpillar (CAT, 1.206), JPMorgan Chase (JPM, 1.202), Microsoft (MSFT, 1.201), Dow Inc. (DOW, 1.169), Visa (V, 1.145), Honeywell (HON, 1.112), Home Depot (HD, 1.048), and Walgreens Boots Alliance (WBA, 1.037).

The second group of stocks, with beta values greater than 0.5 but less than 1, indicates moderate volatility, offering a more stable investment environment. However, they may not participate as strongly in market upswings, potentially limiting overall returns. Notable companies in this category include Cisco (CSCO, 0.979), 3M (MMM, 0.965), Chevron (CVX, 0.959), Travelers (TRV, 0.844), UnitedHealth Group (UNH, 0.813), IBM (IBM, 0.785), Coca-Cola (KO, 0.681), McDonald's (MCD, 0.659), Amgen (AMGN, 0.658), Procter & Gamble (PG, 0.569), Walmart (WMT, 0.556), Verizon (VZ, 0.519), and Johnson & Johnson (JNJ, 0.517).

In the final category of beta values less than 0.5, exemplified by Merck (MRK, 0.465) this stock exhibits lower volatility, providing a stable investment environment but potentially missing out on opportunities for higher returns during bullish market conditions.

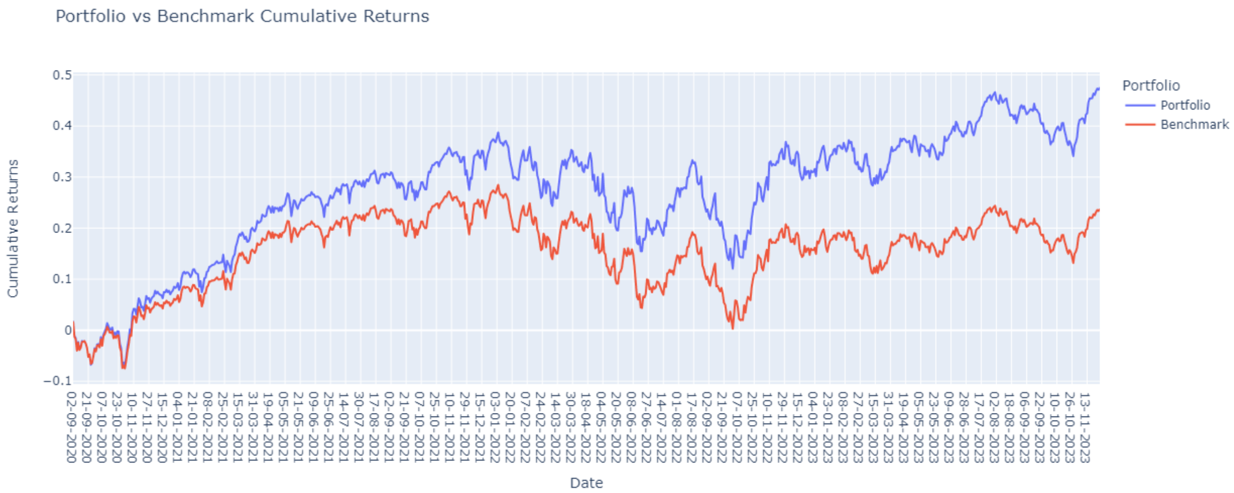
In a nutshell, when picking stocks with different beta values, it's crucial for investors to think about how much risk they're comfortable with and what they want to achieve. By doing this, they can build a mix of stocks that matches their preferences and financial goals, creating a portfolio that works for them.

2.4 Portfolio Analysis:

Verizon (VZ) appears to exert a downward influence on the broader market, exhibiting a negative correlation with overall market trends. Compounding this, VZ is characterized by a relatively low growth rate. Consequently, a comprehensive portfolio analysis is warranted, involving a comparative assessment of cumulative growth between the entire market and a portfolio that excludes Verizon. We're looking at how Verizon's performance affects the overall market and exploring ways to improve portfolios with different investments.

When doing the portfolio analysis, the VZ-excluded DOW Jones Industrial Average portfolio has markedly outperformed the benchmark, delivering a total return of 47.46% and an annualized return of 13.28%, surpassing the benchmark's 23.64% and 7.76%, respectively. With a higher Sharpe Ratio of 0.8338, signifying superior risk-adjusted returns, the portfolio reflects effective investment strategies. Despite comparable volatility levels (15.93% for the portfolio and 15.46% for the benchmark), the portfolio's overall performance indicates a more favorable investment outcome. . Image 3 below visually depicts the comparison between the VZ-excluded DOW Jones Industrial Average and the benchmark.

Figure 3: Portfolio Analysis

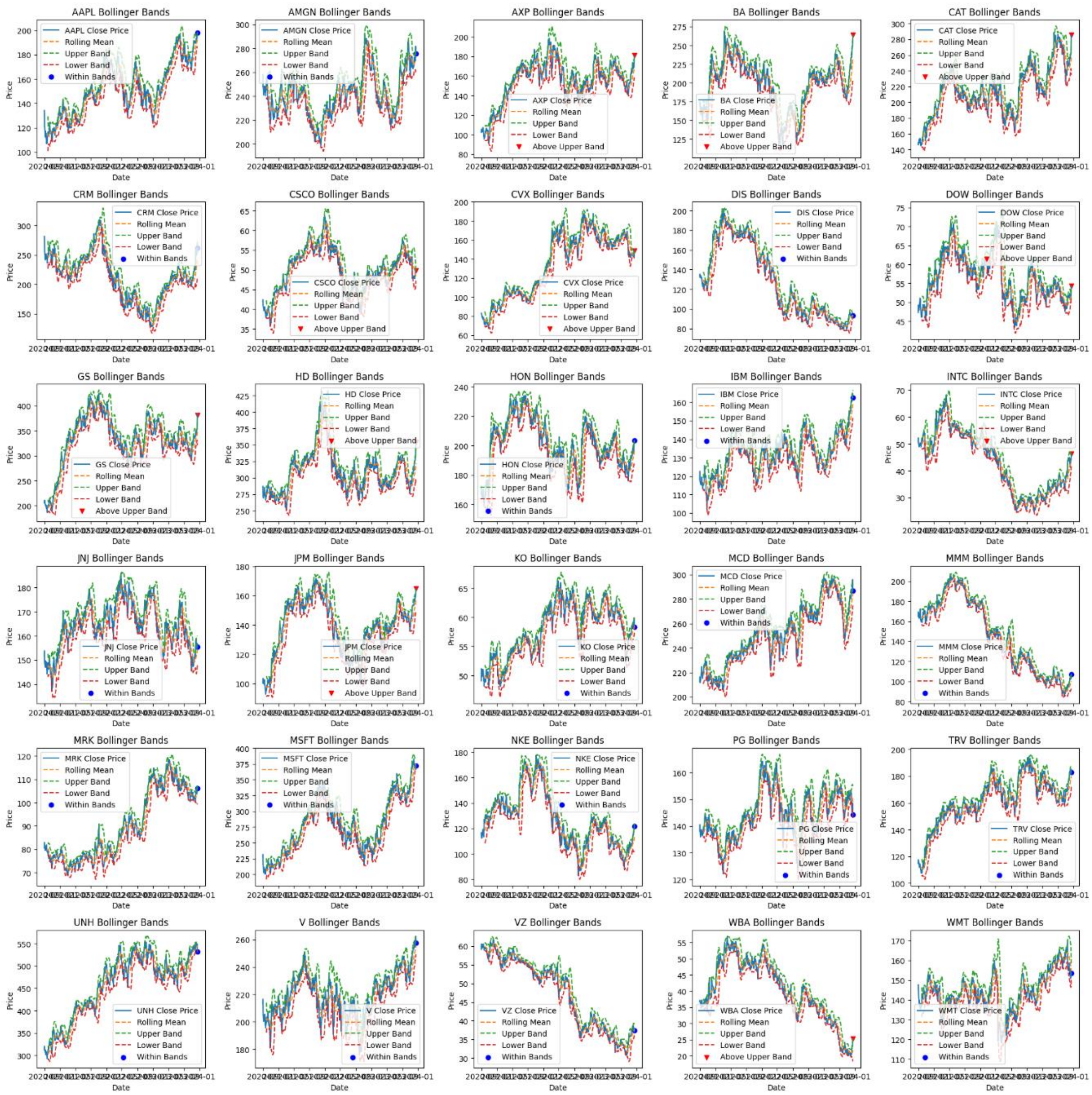


2.5 Bollinger Bands:

Bollinger Bands is a statistical tool used to identify the optimal entry and exit points in stock trading. By analyzing the closing price's movement in relation to the upper and lower bands on the graph, a more informed decision can be made. A breach of the upper band indicates overbought conditions, signaling a potential price decrease and prompting an exit. Conversely, breaking the lower band suggests underbought conditions, signaling a potential price increase and presenting an entry opportunity. To devise a comprehensive strategy for enhancing the DOW Jones Industrial Average Index, Bollinger Bands are employed for both short-term (20-day moving average with a standard deviation of 2) and long-term (50-day moving average with a standard deviation of 2.5) analyses.

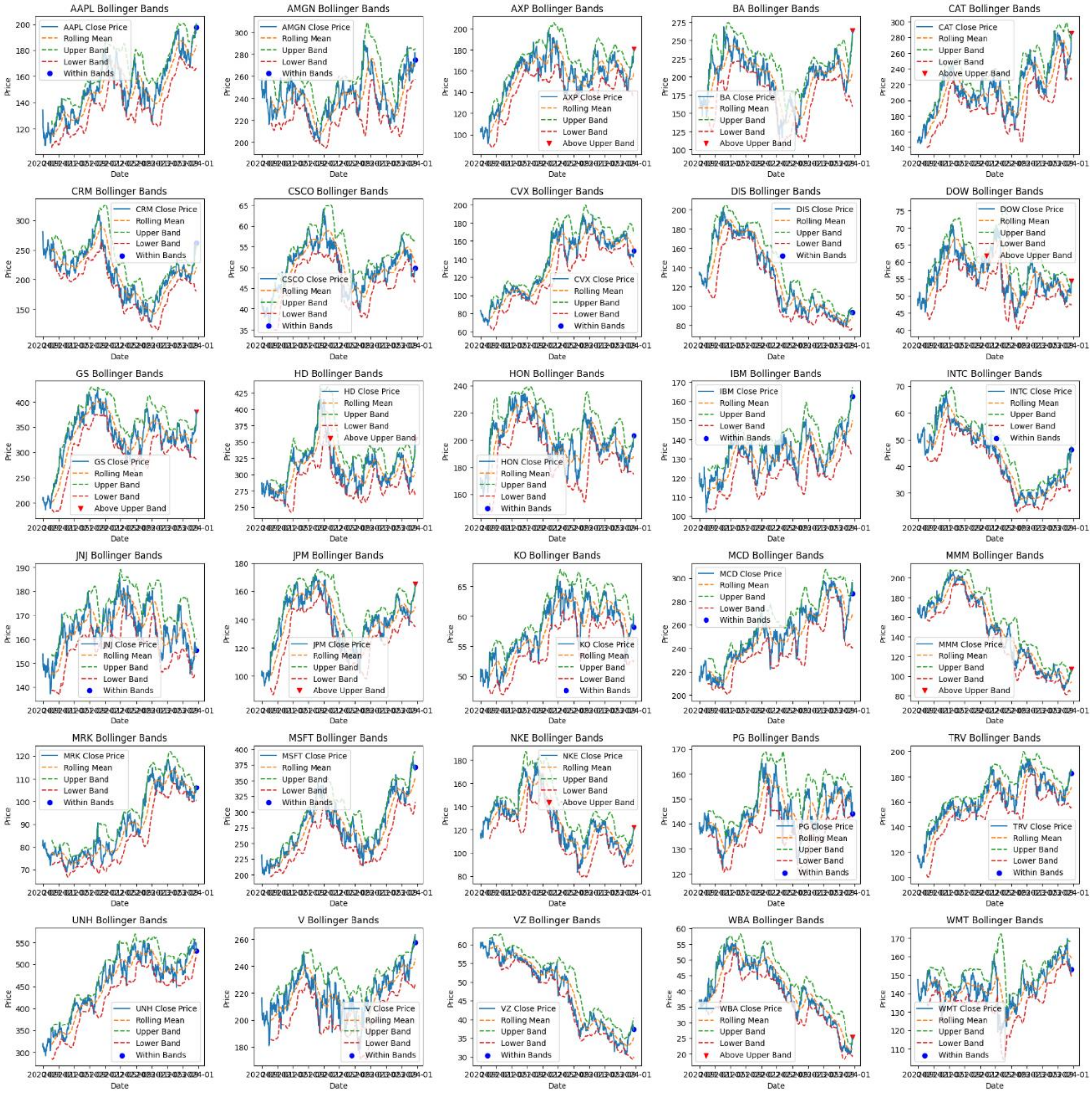
By analyzing the short-term Bollinger Bands, we see that the majority of the shares in the DOW Jones Industrial Average Index fall within the bounds. Furthermore, none of the shares are breaking the lower band suggesting that there may not be any shares which are underbought. On the other hand, several stocks, including American Express, Boeing, Caterpillar, Cisco Systems, Chevron, Dow Inc., Goldman Sachs, Home Depot, Intel, JPMorgan Chase, and Walgreens Boots Alliance, are breaking the upper bounds, signaling potential overbought conditions, and warranting consideration for divestment, exposure limitation, or hedging. The Bollinger Bands for these stocks are depicted in Table X below.

Figure 4: Short-term Bollinger Bands



By analyzing the long-term Bollinger Bands, we see that the majority of the shares in the DOW Jones Industrial Average Index fall within the bounds. Furthermore, none of the shares are breaking the lower band suggesting that there may not be any shares which are underbought. On the other hand, several stocks, including American Express, Boeing, Caterpillar, Dow Inc., Goldman Sachs, Home Depot, JPMorgan Chase, 3M, Nike, and Walgreens Boots Alliance, are breaking the upper bounds, signaling potential overbought conditions, and warranting consideration for divestment, exposure limitation, or hedging. The Bollinger Bands for these stocks are depicted in Table Y below.

Figure 5: Long-term Bollinger Bands



Considering evaluations using both long-term and short-term Bollinger Bands, it is evident that American Express, Boeing, Caterpillar, Dow Inc., Goldman Sachs, Home Depot, JPMorgan Chase, and Walgreens Boots Alliance shares are currently exhibiting overbought conditions. A comprehensive examination is recommended to mitigate potential risks to the index, with actions such as strategically divesting or implementing hedging measures for identified stocks, ultimately contributing to a well-informed risk management strategy for the portfolio.

2.6 Portfolio Analysis Conclusion:

A better understanding of investors is warranted to develop a more comprehensive strategy for improving the DOW Jones Industrial Average. However, certain conclusions can already be drawn from the analysis.

Excluding Verizon (VZ) has proven beneficial for the portfolio, surpassing the Dow Jones Industrial Average with a total return of 47.46% and an annualized return of 13.28%. The notably higher Sharpe Ratio (0.8338) underscores superior risk-adjusted returns, positioning the portfolio favourably. It is crucial to acknowledge that Verizon is the sole telecommunications company in the portfolio, and its removal may compromise diversification. Given that Verizon represents only 0.67% of the index, a strategic reallocation to another telecommunications company exhibiting a stronger correlation with the index is advisable. This adjustment aims to sustain a well-balanced and diversified portfolio.

Furthermore, Caution is advised regarding overbought conditions in stocks such as American Express, Boeing, Caterpillar, Dow Inc., Goldman Sachs, Home Depot, JPMorgan Chase, and Walgreens Boots Alliance. Heightened vigilance is essential, prompting a consideration of potential divestment or the implementation of risk mitigation strategies to ensure sustained portfolio optimization.

Time Series Analysis and Predicting DJIA using LSTM:

3.1 Time Series Analysis on DJIA:

Time series analysis is a crucial methodology for investigating and extracting meaningful patterns from temporal data, making it an invaluable tool for understanding the dynamics of financial markets such as the DJIA. In the context of the DJIA, time series analysis involves examining historical price movements, identifying trends, and uncovering potential patterns that can assist in predicting future market behaviour.

3.1.1 Visualization and plotting of DJIA over time:

The below diagram shows DJIA Stock Price over Time. For DJIA, we have data from 1992 until 2023. The DJIA is a stock market index that measures the stock performance of 30 large companies listed on stock exchanges in the United States. From the graph, we can see a noticeable upward trend in stock price over this period, indicating an increase in the value of the DJIA. However, there are also fluctuations visible on the line indicating periods of both increase and decrease in stock value. These fluctuations could be due to a variety of factors, including changes in the economy, political events, and changes in the companies included in the index.

Figure 6: Plotting DJIA over time.

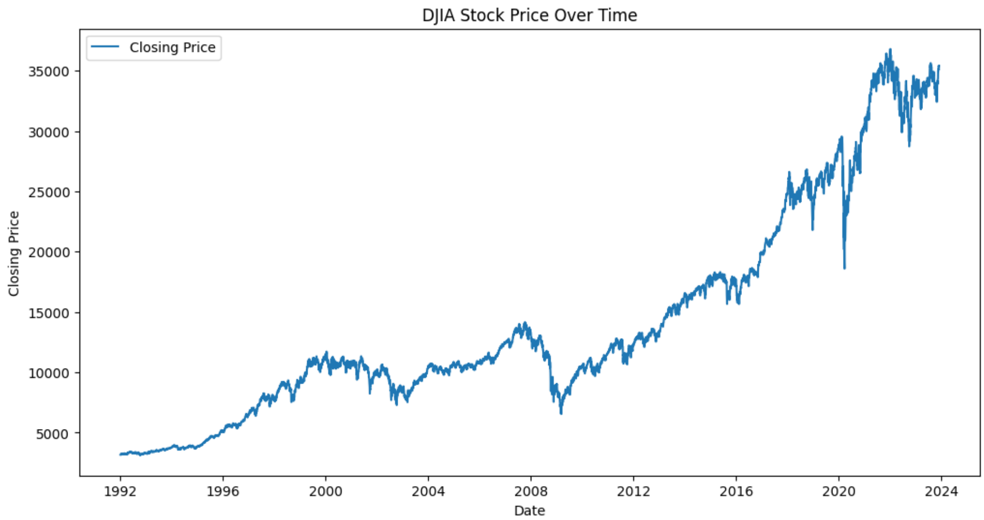
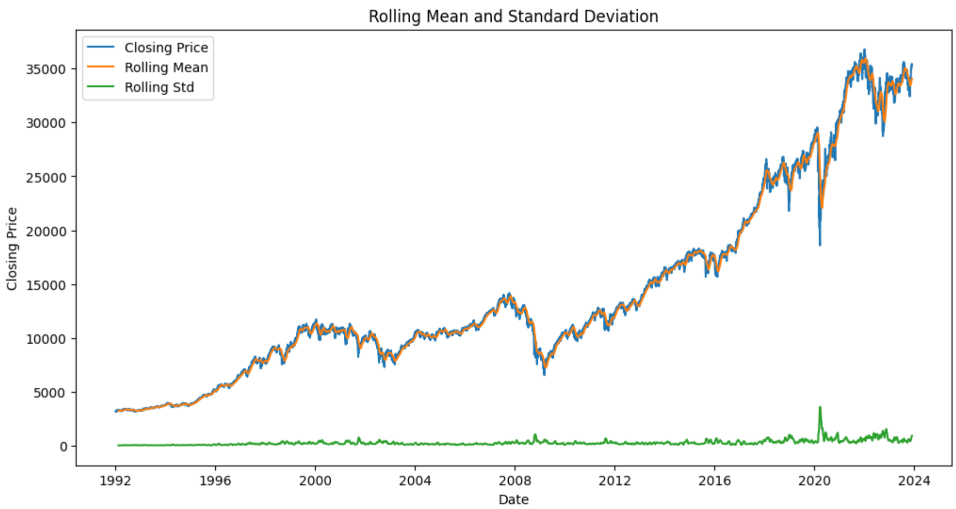


Figure 7: DJIA Close Price, rolling mean and rolling std.



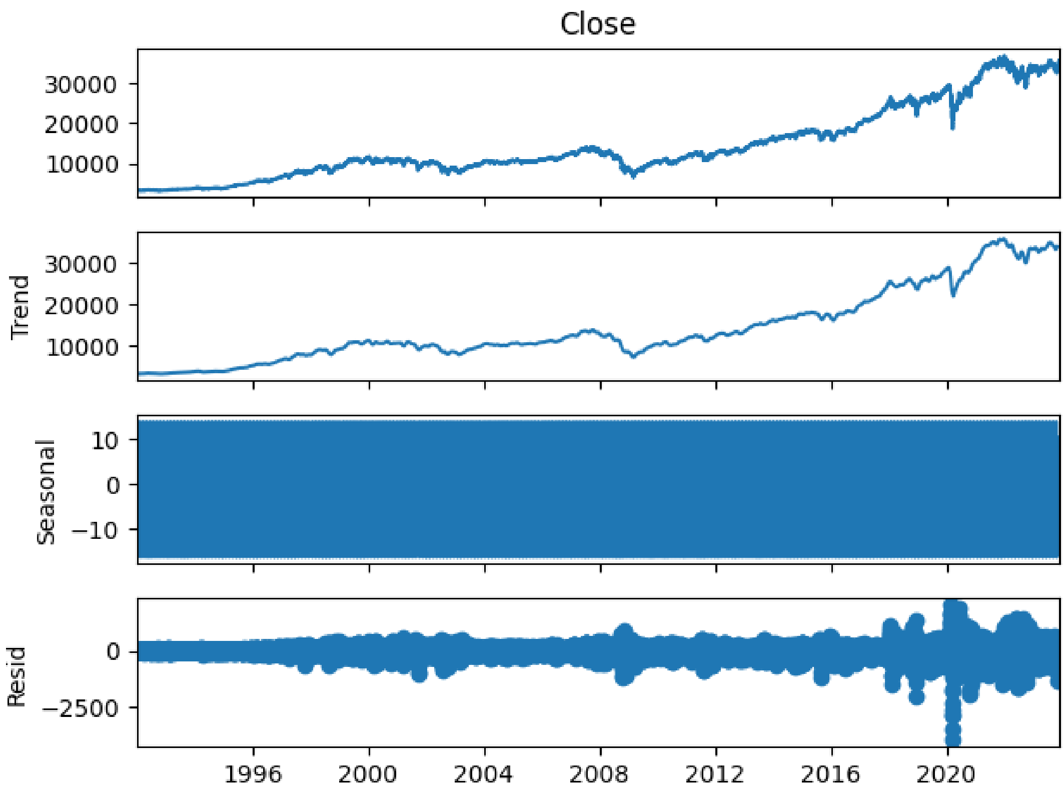
3.1.2 Decomposing DJIA Time Series:

Decomposition of a time series data into three components: trend, seasonal, and residual.

Here's what each component represents:

- Trend: This graph shows the underlying trend in the data when seasonality is removed. It's a smoothed version of the Close graph showing only long-term movements.
- Seasonal: This graph illustrates regular fluctuations that occur at specific intervals. In this case, it appears as a horizontal line indicating no significant seasonal pattern detected in this dataset.
- Resid: This graph displays the residual component, which is what remains after extracting the trend and seasonal components from the original data. It indicates irregularities and anomalies not accounted for by trend or seasonality.

Figure 8: Decomposition of DJIA



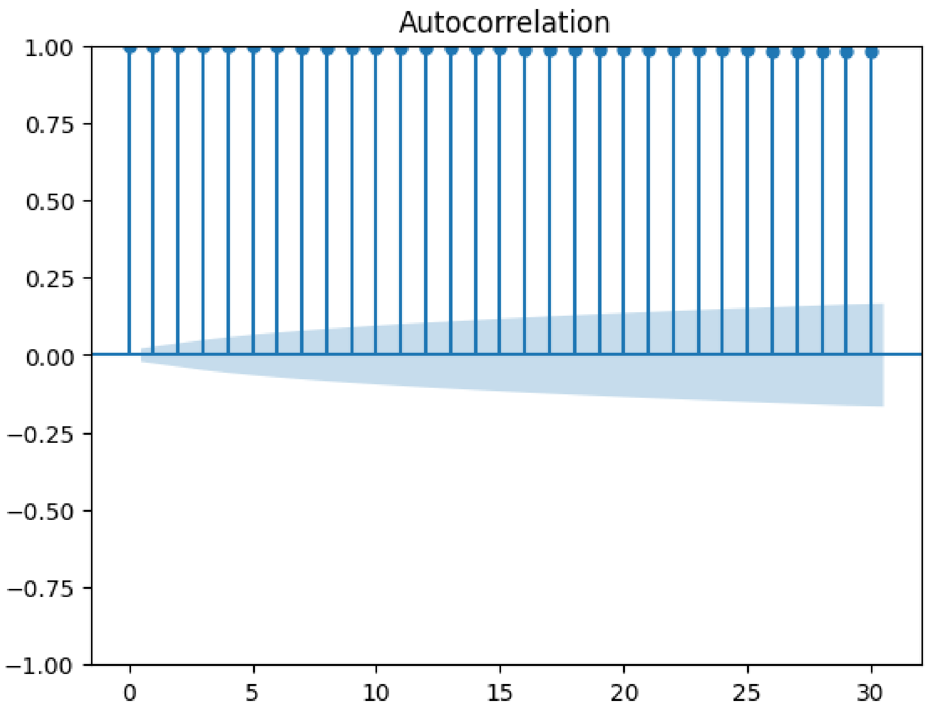
3.1.3 Autocorrelation and Partial Autocorrelation:

3.1.3.1 Autocorrelation and Analysis:

Autocorrelation, also known as serial correlation, is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations of a random variable as a function of the time lag between them. It is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time interval.

There are vertical blue lines extending from the x-axis up to various heights, representing positive autocorrelations at different lags. A horizontal line at y=0 serves as a baseline for distinguishing positive and negative correlations. The area below y=0 is shaded in light blue but there are no lines extending into this area, indicating no significant negative correlations were observed. The autocorrelation starts at 1 when the delay is zero and decreases as the delay increases, indicating that values further apart in time are less correlated. This is a common pattern in autocorrelation plots and suggests that the data may be appropriately modelled using an autoregressive or moving average model.

Figure 9: Autocorrelation Plot

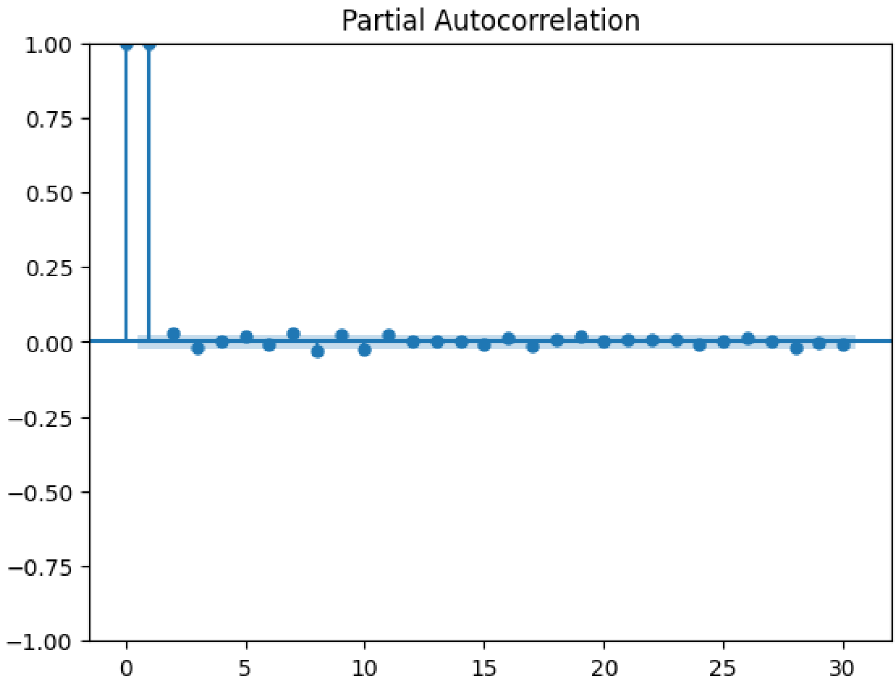


3.1.3.2 Partial Autocorrelation and Analysis:

In time series analysis, the partial autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags. Partial autocorrelation is a measure of the correlation between a time series and a lagged version of itself after controlling for the effects of the other lags.

There are blue dots indicating the partial autocorrelation at each lag. A significant spike is visible at lag 0, reaching up to a correlation value of 1.00. This is expected as a variable is perfectly correlated with itself. Two blue horizontal lines serve as boundaries for significance levels; one is located just above 0.75 and another just below -0.25 on the y-axis. Beyond lag 0, all other correlations are close to zero and fall within the significance boundaries. This suggests that there is no significant correlation between the variable and its lags beyond lag 0.

Figure 10: Partial Autocorrelation



3.1.4 Testing for Stationarity – ADF Test:

Augmented Dickey-Fuller (ADF) test is a common statistical test used to determine whether a given Time series is stationary or not. Here's what the output of ADF test means:

- ADF statistic is 0.711919.
- P-value: This is the probability that you would see these results if the null hypothesis were true. In the context of the ADF test, the null hypothesis is that the time series is not stationary (has some time-dependent structure). The p-value in your case is 0.990073. If the p-value is less than a threshold (commonly 0.05), then you reject the null hypothesis and infer that the time series is indeed stationary. However, since p-value is much larger than 0.05, we fail to reject the null hypothesis and infer that time series is not stationary.
- Critical Values: These are the test statistic values at which you can reject the null hypothesis if your test statistic is less than the critical value. If the ADF Statistic is less than the critical value, we can reject the null hypothesis (aka the series is stationary). When the ADF statistic is greater than the critical value, we fail to reject the null hypothesis (which means the series is not stationary).

In our case, because the ADF statistic is greater than all of the critical values, the test suggests that the series is non-stationary.

Figure 11: ADF Test on DJIA

```python
from statsmodels.tsa.stattools import adfuller

X = df['Close'].values

result = adfuller(X)
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
print('Critical Values:')
for key, value in result[4].items():
    print('\t%s: %.3f' % (key, value))
```

```
ADF Statistic: 0.711919
p-value: 0.990073
Critical Values:
	1%: -3.431
	5%: -2.862
	10%: -2.567
```

3.2 Predicting DJIA using LSTM:

3.2.1 LSTM:

LSTM, or Long Short-Term Memory, is a specialized tool in the realm of artificial intelligence used for predicting future stock prices, particularly in the context of indices like DJIA. It excels in capturing and understanding complex patterns and trends inherent in stock price data over time. What makes LSTM unique is its ability to retain and utilize crucial information for extended periods, enabling more informed predictions. In essence, LSTM acts as an intelligent assistant in the field of finance, enhancing our ability to make more accurate forecasts about potential movements in the stock market.

3.2.2 Why LSTM:

LSTM models outshine traditional methods like ARIMA in time series analysis due to their knack for grasping intricate patterns and long-term dependencies in sequential data, which is crucial for real-world datasets like stock prices. Unlike ARIMA, which assumes linear relationships, LSTMs excel in handling dynamic and non-linear connections. Their automatic learning capability eliminates the need for manual feature engineering, making them adept at managing large datasets with high complexity, such as financial time series. LSTMs adapt to changing relationships between variables over time, addressing challenges that conventional models may face. In essence, LSTM models offer versatility, proficiency in capturing intricate patterns, and the ability to navigate non-linear relationships, making them a powerful tool for time series analysis, particularly in domains like stock market forecasting.

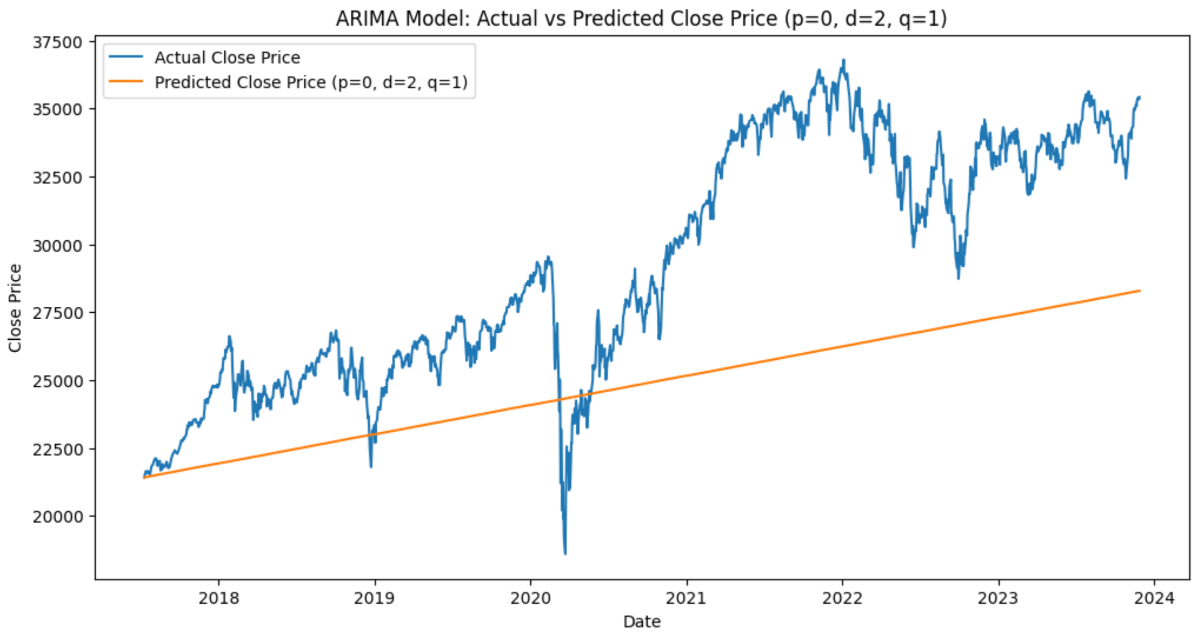
3.2.3 How does LSTM works:

LSTM is a type of recurrent neural network (RNN) designed to tackle the challenges of learning and predicting sequences, making it well-suited for time series data like stock prices. Its architecture features memory cells, gates, and connections that enable it to capture and store important information over extended periods. The memory cells act as containers, holding and releasing information when needed. The gates regulate the flow of data into and out of the cells, allowing the network to decide what information is crucial for making predictions. This unique design empowers LSTM to effectively capture long-term dependencies and intricate patterns in sequential data, enabling more accurate and nuanced predictions in complex scenarios like stock market forecasting.

3.2.4 Comparison - Prediction using ARIMA model:

First we used traditional ARIMA model to predict DJIA in order to compare it with LSTM. The ARIMA model with parameters (p=0, d=2, q=1) was used to predict the DJIA's closing price. However, with a high Mean Squared Error (MSE) of 27353804.65 and RMSE of 5230.08, the model's predictions show significant deviation from the actual prices. As you can see, model is unable to capture trends and just predicting the straight line.

Figure 12: Prediction of DJIA using ARIMA Model

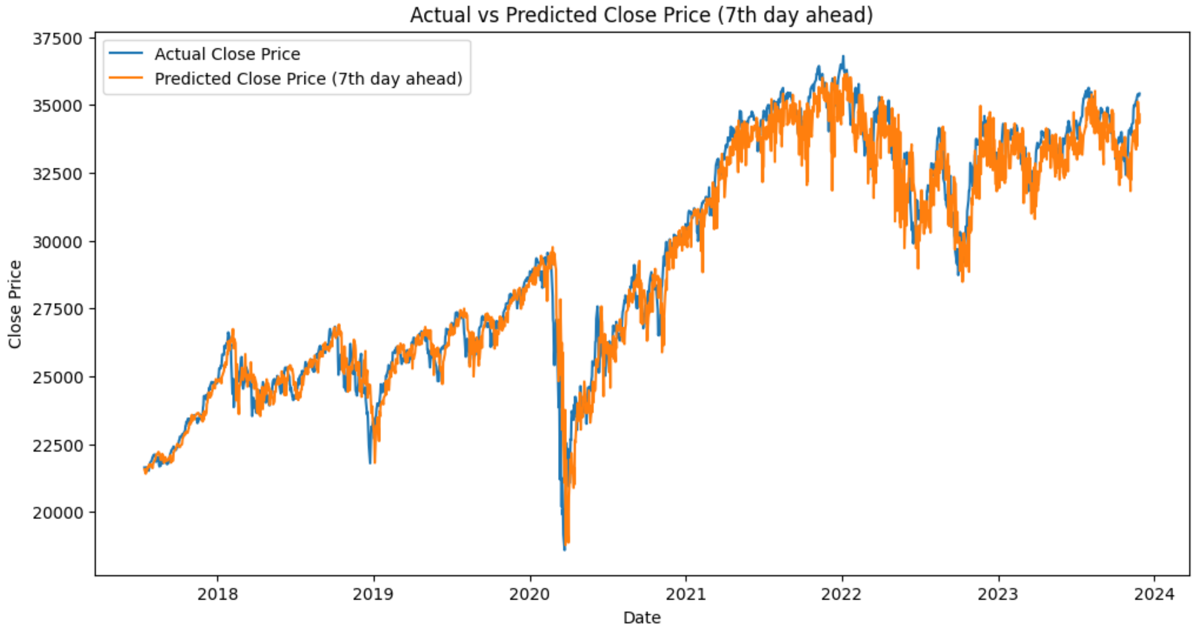


3.2.5 Predicting DJIA 7th Day Ahead (One Week Ahead) using LSTM:

We started by preprocessing the data, including sorting by date and normalizing the closing prices. We created sequences of historical data with a sequence length of 10 and aimed to predict the stock price for the 7th day ahead. The LSTM model, consisting of multiple layers, was trained on 80% of the data, and its performance was evaluated on the remaining 20%. The mean squared error (MSE) and root mean squared error (RMSE) were calculated to assess the model's accuracy. The visualization of the predicted and actual closing prices showcased the model's forecasting capabilities. The choice of LSTM was justified by its ability to capture complex patterns and dependencies in time series data, making it a powerful tool for stock price prediction.

The LSTM model achieved a Mean Squared Error (MSE) of approximately 848,586.390 and a Root Mean Squared Error (RMSE) of about 921.187 on the test set. As we can see these values are lower than ARIMA model hence showing better prediction.

Figure 13: Predicting DJIA One week ahead using LSTM



3.3 Conclusion

In conclusion, the analysis of DJIA stock price over time revealed a noticeable upward trend with fluctuations indicating periods of both increase and decrease. The decomposition of the time series data into trend, seasonal, and residual components provided insights into the underlying patterns. Autocorrelation and partial autocorrelation functions were employed to explore the temporal dependencies in the data, with autocorrelation suggesting potential for autoregressive or moving average modeling.

Comparing traditional ARIMA and advanced LSTM models for DJIA closing price prediction, it was evident that the ARIMA model with parameters (p=0, d=2, q=1) struggled to capture trends, resulting in a high Mean Squared Error (MSE) of 27,353,804.65 and RMSE of 5,230.08. In contrast, the LSTM model, trained on sequences of historical data with a sequence length of 10 and predicted the close price for one week ahead, demonstrated superior forecasting capabilities. With a lower MSE of approximately 848,586.39 and RMSE of about 921.19 on the test set, the LSTM model outperformed the ARIMA model, showcasing its ability to capture complex patterns and dependencies in time series data. The LSTM model is thus deemed a more effective tool for stock price prediction in this scenario.