Table of Contents

[Part A 3](#_Toc167456158)

[A1) 3](#_Toc167456159)

[Introduction 3](#_Toc167456160)

[Methodology 3](#_Toc167456161)

[Traditional Machine Learning Models 3](#_Toc167456162)

[Support Vector Machine (SVM) 3](#_Toc167456163)

[Naive Bayes 4](#_Toc167456164)

[Random Forest 4](#_Toc167456165)

[Deep Learning Models 5](#_Toc167456166)

[Long Short-Term Memory (LSTM) 5](#_Toc167456167)

[Convolutional Neural Network (CNN) 6](#_Toc167456168)

[A2) 7](#_Toc167456169)

[Additional Features for Enhanced Recommendation System 7](#_Toc167456170)

[Feature 1: Sentiment Analysis from User Feedback 8](#_Toc167456171)

[Feature 2: Temporal Dynamics for Content Recommendation 8](#_Toc167456172)

[Part B 9](#_Toc167456173)

[Simple Moving Average (SMA) – Momentum Strategy 12](#_Toc167456174)

[Backtesting 13](#_Toc167456175)

[Maximum Drawdown 14](#_Toc167456176)

[Optimize Strategy 14](#_Toc167456177)

[Recovery from Covid 16](#_Toc167456178)

[Portfolio Optimization 17](#_Toc167456179)

[References 21](#_Toc167456180)

# Part A

## A1)

## Introduction

A crucial component of Natural Language Processing (NLP) is sentiment analysis, which classifies textual opinions to ascertain attitudes towards particular subjects or sentiment in general (Liu, 2012). It is extensively used across many industries to extract insights from unstructured data, which aids businesses in comprehending consumer experiences, public opinion, and brand perception. Sentiment analysis uses a variety of computational techniques, from straightforward rule-based algorithms to sophisticated machine learning methods like CNN and LSTM (Pang and Lee, 2008; Medhat et al., 2014). In order to determine the optimal method for sentiment classification based on metrics such as accuracy, precision, recall, and F1-score, this project investigates and assesses the efficacy of various models on a dataset. This will lay the groundwork for practical applications.

## Methodology

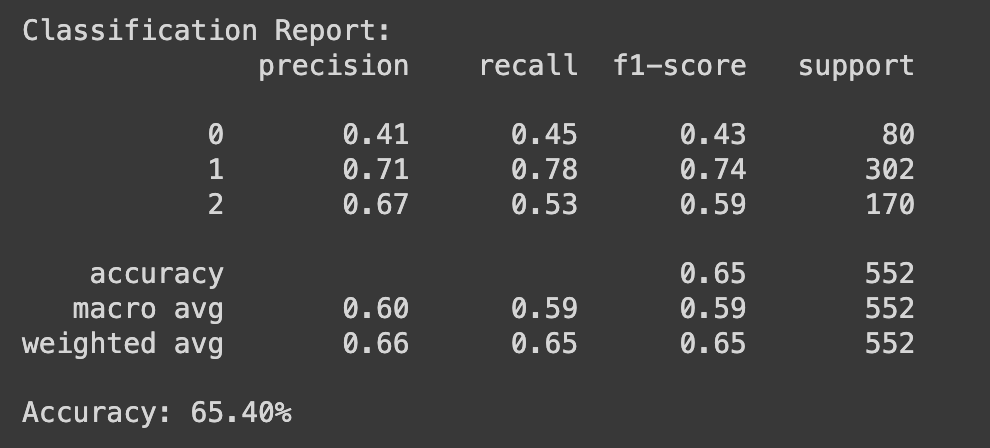
In sentiment analysis, text data is sourced and preprocessed (cleaning, tokenization, stopword removal, and TF-IDF vectorization) to improve machine readability. Sentiment is classified using a variety of machine learning models, including LSTM and CNN, as well as more sophisticated neural networks like Random Forest, SVM, and Naive Bayes. To ensure robustness and prevent overfitting, the dataset is divided into training and testing sets for these models. Metrics such as accuracy, precision, recall, and F1-score are used to assess performance and identify the best performing model after a careful examination of all models' advantages and disadvantages.

## Traditional Machine Learning Models

In sentiment analysis, text data is sourced and preprocessed (cleaning, tokenization, stop word removal, and TF-IDF vectorization) to improve machine readability. Sentiment is classified using a variety of machine learning models, including LSTM and CNN, as well as more sophisticated neural networks like Random Forest, SVM, and Naive Bayes. Because it assumes predictor independence, Naive Bayes is a dependable benchmark because it is quick, makes use of probability theory, and performs well with large datasets (McCallum & Nigam, 1998). By highlighting feature significance and reducing overfitting, Random Forest's ensemble of decision trees improves model interpretation (Breiman, 2001).

### Support Vector Machine (SVM)

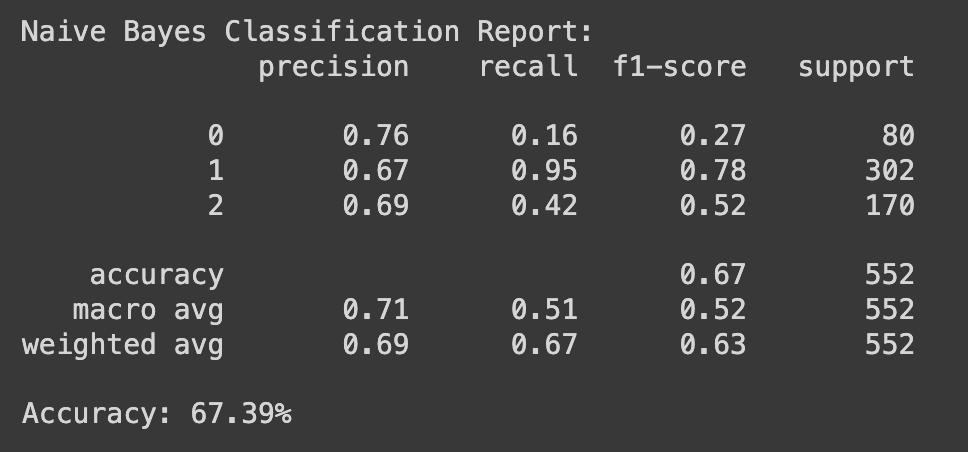
Support Vector Machines (SVM) are perfect for sentiment analysis because they perform well in high-dimensional spaces, like the ones created by TF-IDF vectorized text data. SVM builds hyperplanes to efficiently divide sentiment categories, guaranteeing accurate classification. This ability is essential for effectively classifying texts into positive, negative, and neutral sentiments and for detecting minute variations in sentiment (Cristianini & Shawe-Taylor, 2000).



#### Fig 1: SVM Statistics

### Naive Bayes

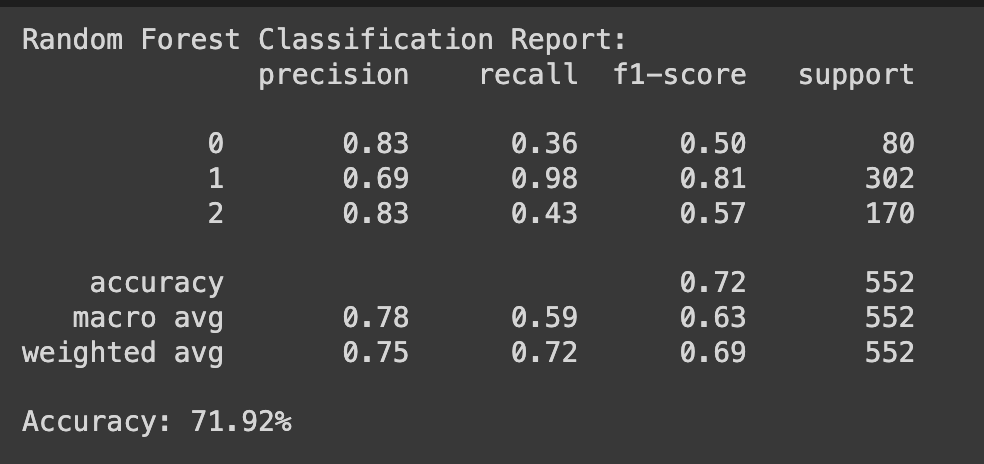
Naive Bayes classifiers are popular due to their ease of use and efficiency when dealing with large datasets, making them especially useful for text classification tasks such as sentiment analysis. This model uses probability theory and assumes predictor independence, resulting in quick and effective sentiment categorization. Naive Bayes' straightforward probabilistic approach efficiently distinguishes between positive, negative, and neutral sentiments in text data, making it a practical solution for large-scale sentiment analysis (McCallum & Nigam, 1998).



#### Fig 2: Naïve Bayes Statistics

### Random Forest

Random Forest classifiers are used for their ability to handle complex datasets with intricate decision boundaries, making them ideal for sentiment analysis tasks. This ensemble method, which consists of multiple decision trees, uses averaging to improve predictive accuracy while minimising overfitting. Random Forest's ability to manage large feature sets while providing insights into feature importance makes it capable of distinguishing between different sentiment categories in text data (Breiman, 2001).



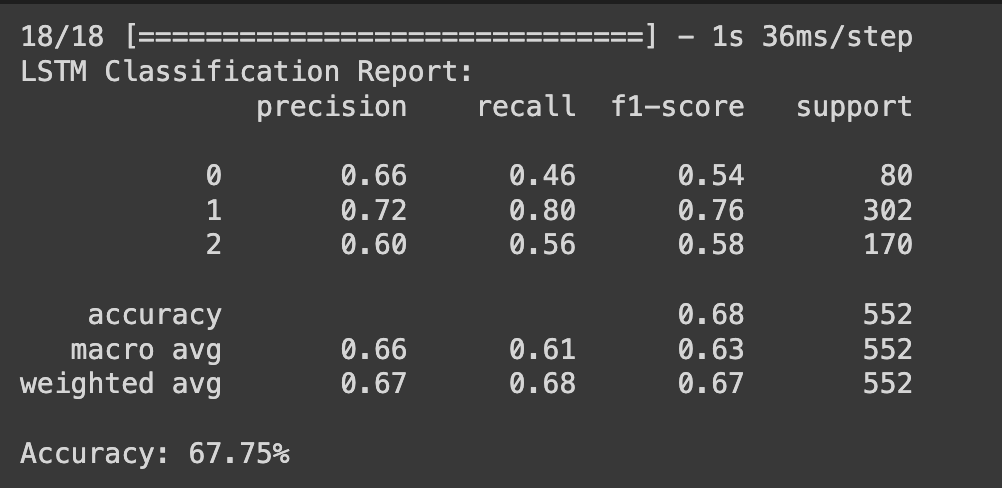
#### Fig 3: Random Forest Statistics

## Deep Learning Models

Deep learning models, such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), provide advanced approaches for sentiment analysis in Part A1, capturing context and semantic relationships in text data. LSTM, a type of recurrent neural network, excels at processing data sequences while maintaining context over long texts, which is critical for accurate sentiment prediction (Hochreiter & Schmidhuber, 1997). CNNs, which have been adapted from image processing to NLP, use convolutional filters to detect hierarchical patterns and local dependencies, allowing them to identify sentiment-indicative phrases in larger texts. Kim (2014) demonstrated the efficacy of CNNs in sentence-level classification by extracting meaningful patterns. These deep learning models significantly improve sentiment analysis by handling complex and large-scale textual data more effectively than traditional models.

### Long Short-Term Memory (LSTM)

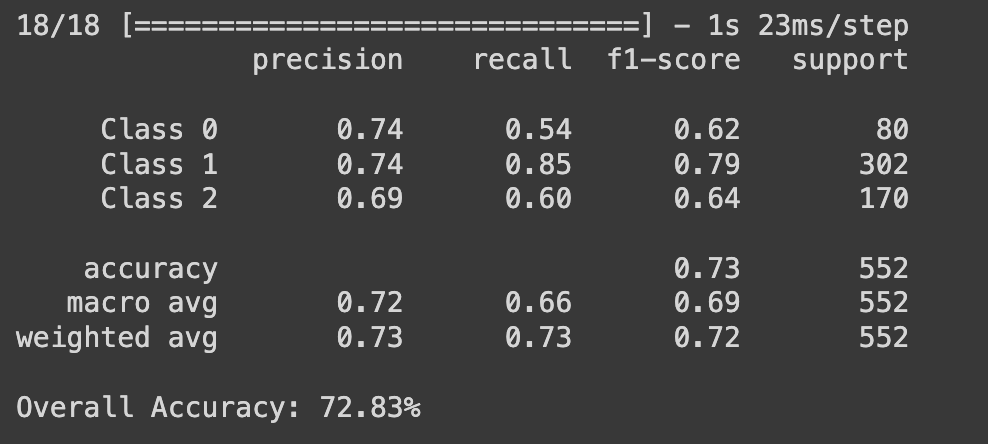
Long Short-Term Memory networks (LSTM) are especially useful for sentiment analysis because of their ability to process sequences and retain information over long periods of time, capturing the context within text data that is critical for understanding sentiments. This type of recurrent neural network is particularly effective in situations where the relationship between words and phrases has a significant impact on the sentiment expressed. LSTMs can handle the sequential nature of text, making them good at detecting underlying sentiment trends and nuances across long text passages (Hochreiter & Schmidhuber, 1997).



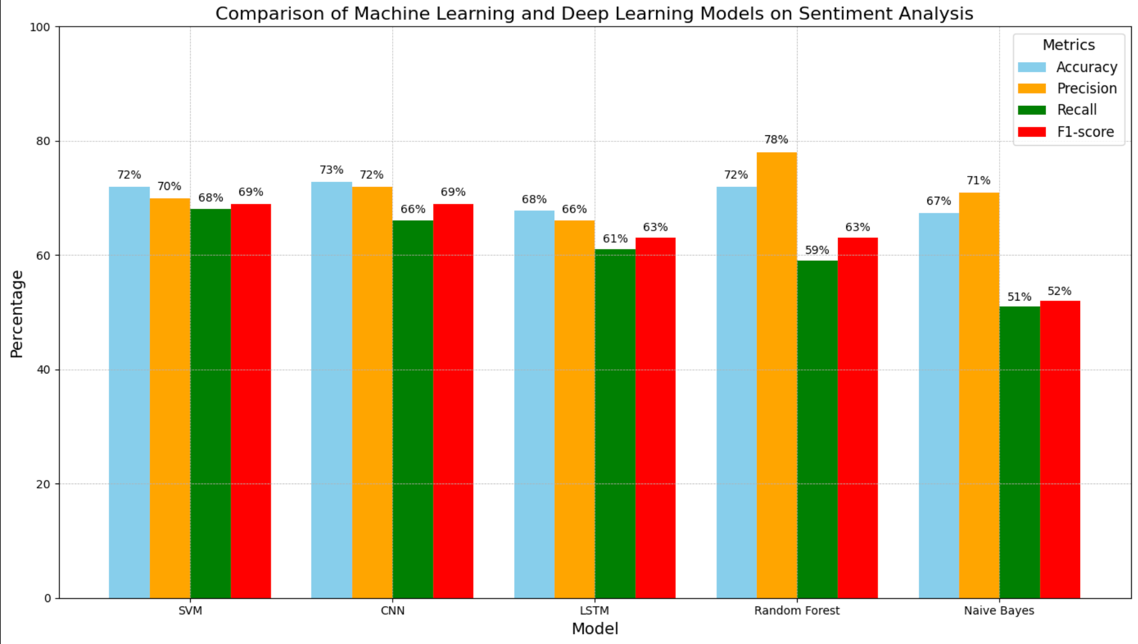
#### Fig 4: LSTM Statistics

### Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are widely used in sentiment analysis because of their ability to extract local contextual features from text, making them ideal for detecting sentiment-related patterns. CNNs use convolutional layers to detect specific n-gram features, which are critical in understanding the sentiment expressed in phrases and sentences. This approach is especially useful for capturing subtle nuances in text that influence sentiment, thereby improving classification accuracy. Kim (2014) emphasises the success of CNNs in sentence-level classification tasks, demonstrating their ability to extract meaningful patterns from textual data.

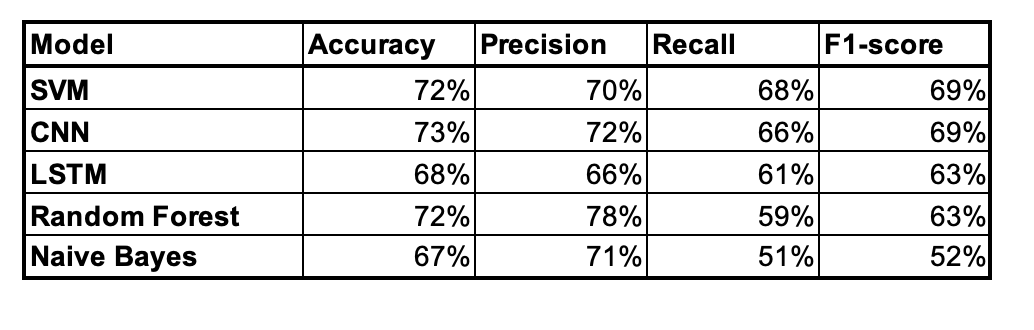


#### Fig 5: CNN Statistics



#### Fig 6: Model accuracy comparison

The results of the ANOVA test and performance metrics indicate that the Convolutional Neural Network (CNN) is the most effective model for sentiment analysis with this dataset. With a 73% overall accuracy, strong precision of 72%, and an F1-score of 69%, CNN performed best overall, demonstrating a balanced approach to handling false positives and false negatives. Its recall (66%) is not the best, but it's still competitive in terms of finding pertinent cases. The statistical significance of the differences in model performances is confirmed by the ANOVA test (F-statistic = 4.26, p-value = 0.005). Although Random Forest exhibits high precision (78%), its overall balance is impacted by its lower recall (59%). With an accuracy of 72%, precision of 70%, recall of 68%, and F1-score of 69%, Support Vector Machine (SVM) is a strong competitor as well. LSTM networks and Naive Bayes, on the other hand, perform poorly. LSTM networks exhibit moderate performance (68% accuracy, 66% precision, 61% recall, and 63% F1-score), while Naive Bayes displays the lowest metrics (67% accuracy, 71% precision, 51% recall, and 52% F1-score). As a result, CNN is the most efficient model, followed by SVM and Random Forest. Because of their relatively poorer performance, LSTM and Naive Bayes are less desirable.



#### Table 1: Model Accuracy Comparison Table

## A2)

## Additional Features for Enhanced Recommendation System

### Feature 1: Sentiment Analysis from User Feedback

**Feature**: Sentiment analysis on user feedback (e.g., reviews, comments) in order to improve content recommendations assuring the sentiment and preferences of a user as follows:

**Data Source**:

* **User Reviews and Comments**: Derived from feedback sections of the platform.
* **Social Media**: It can extract the sentiment data from social media interaction with respect to content of that platform.

**Techniques**:

* **Natural Language Processing (NLP)**: One can use NLP to analyse the sentiment of user feedback and differentiate content from positive, neutral negative sentiment. or based on Medhat et al. 2014
* **Sentiment-Based Filtering**: Leveraging sentiment scores within recommendation algorithms enables accurate recommending content to users that matches their preferences as well as sentiments (Pang and Lee, 2008).

**Implications**:

* **Improved User Experience**: Customized offerings by mining the user sentiments will lead to more satisfied and active users.
* **Increased Retention**: Users are more likely to stay engaged with content that resonates with their sentiments.

### Feature 2: Temporal Dynamics for Content Recommendation

**Feature**: Take into account temporal dynamics such as seasonal trends and time-of-day patterns for providing users with opportune contextually relevant content

**Data Source**:

* **Platform Usage Logs**: Internal logs tracking user activity patterns over time.
* **Calendar Events**: Integrate with public and personal calendars to understand relevant events and holidays.

**Techniques**:

* **Time-Series Analysis**: Use time-series forecasting methods to predict content popularity trends based on historical data (Zhou et al., 2019).
* **Seasonal and Temporal Models**: With planning, implement models such as Seasonal ARIMA (AutoRegressive Integrated Moving Average) that can model and forecast seasonal patterns in the consumption of your product content (Hyndman & Athanasopoulos, 2018).

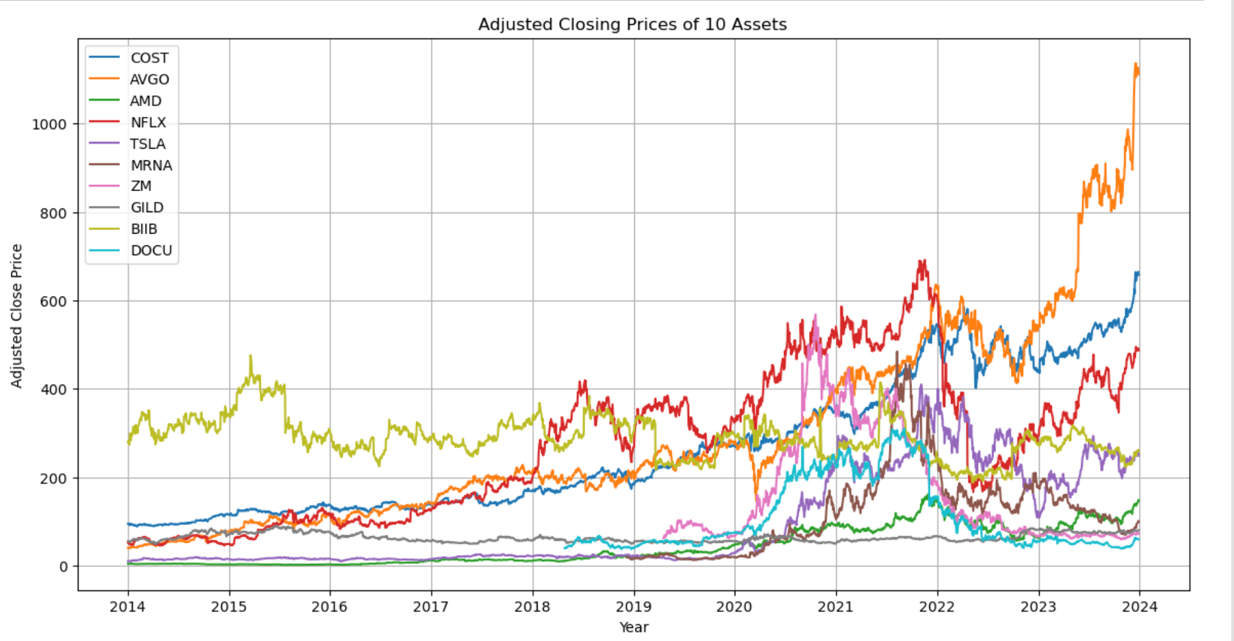
**Implications**:

* **Enhanced Relevance**: The recommendation of new content that corresponds with the consumer’s latest interests and behavior will help make it more relevant.
* **Dynamic Adaptability**: Add in the mix that a recommendation system is built into this for it to also fine-tune and adapt based on current patterns of how content is consumed. Ultimately, improving user satisfaction resulting in higher retention

We believe all of these features will help improve the recommendation system using newer and more sophisticated data sources in new analytical methods that can make recommendations more relevant, timely to users and eventually result in improved engagement/ operational metrics.

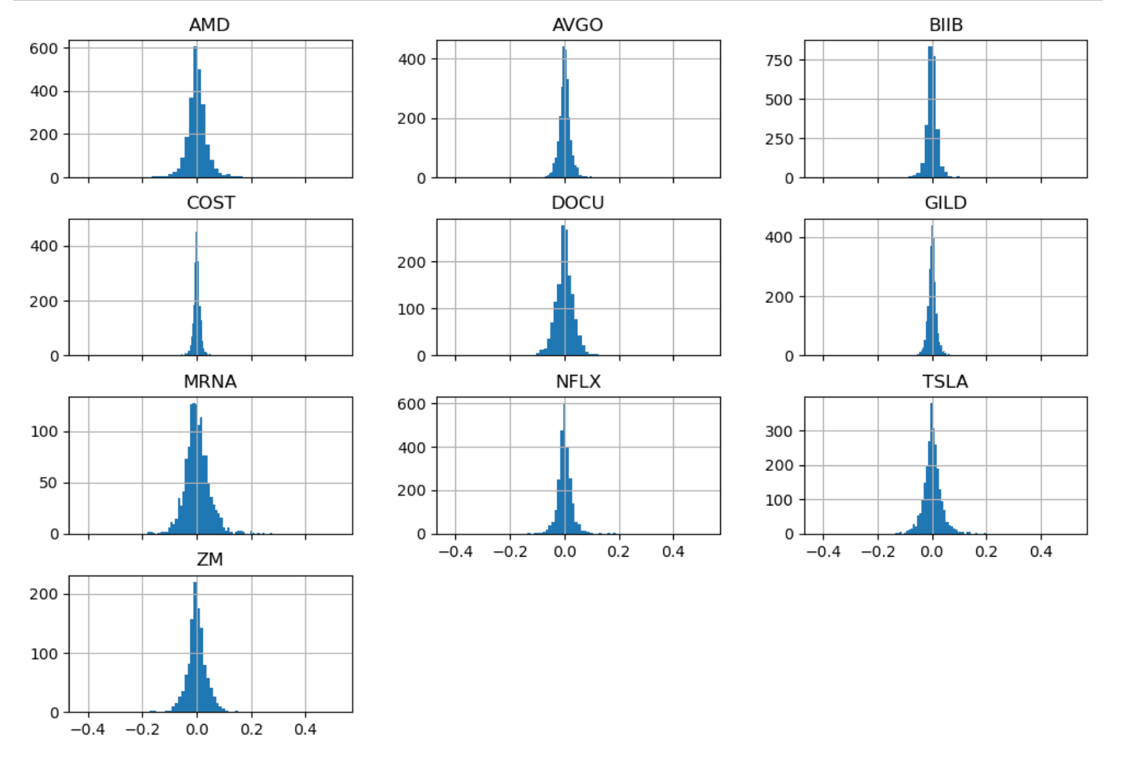
# Part B

The list comprising COST, AVGO, AMD, NFLX, TSLA, MRNA, ZM, GILD, BIIB, and DOCU represents some of the top-performing assets on the NASDAQ stock exchange, carefully selected based on their performance from January 1, 2014, to January 1, 2024. This selection was made after a comprehensive review of data sourced from Yahoo Finance, focusing on their growth, market influence, and investment returns. To perform the data analysis, all required Python packages have been installed, ensuring the ability to handle extensive data manipulation and visualization tasks effectively. These stocks were analysed to understand market trends and investment strategies within this ten-year period, highlighting their significant impact on portfolios and the broader financial market.

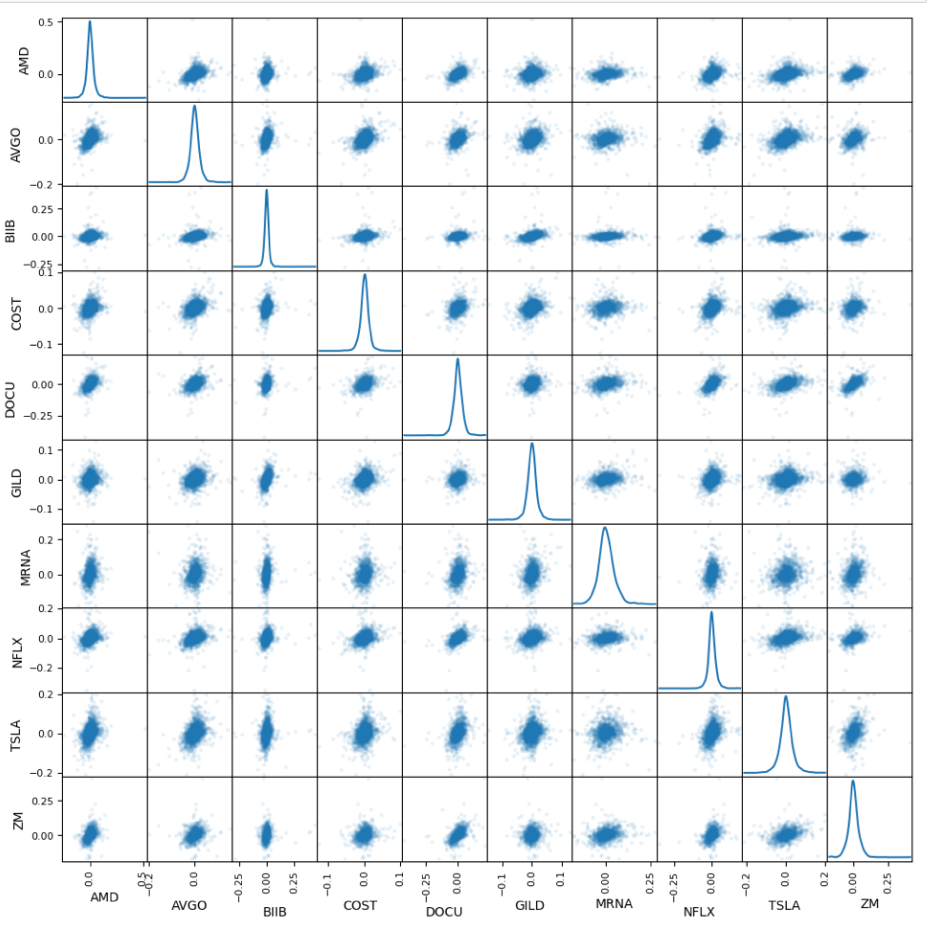


#### Fig 7: Adjusted Closing Price of all assets

The adjusted close price reflects a stock's value after accounting for dividends, stock splits, and other corporate actions, providing an accurate measure of its historical performance. In this graph, TSLA (Tesla) appears to be the most volatile asset, with dramatic fluctuations and a steep rise towards the end of the observed period.

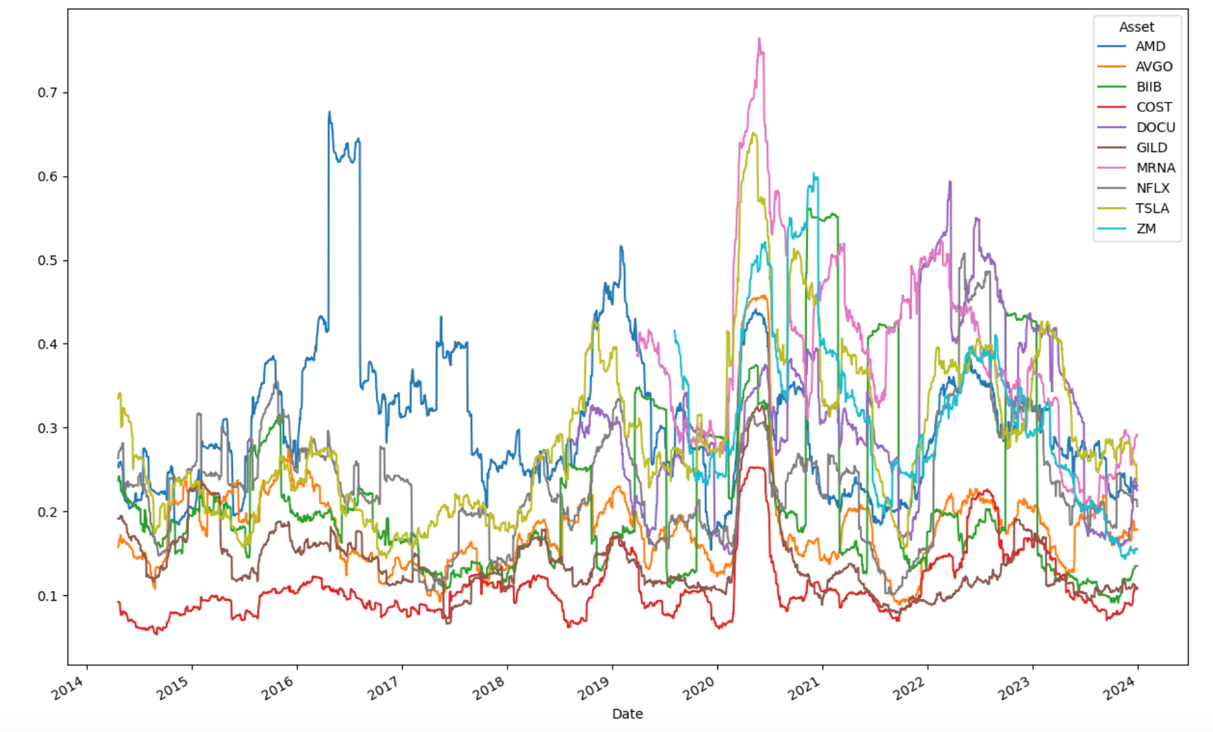


#### Fig 8: Returns



#### Fig 9: Returns

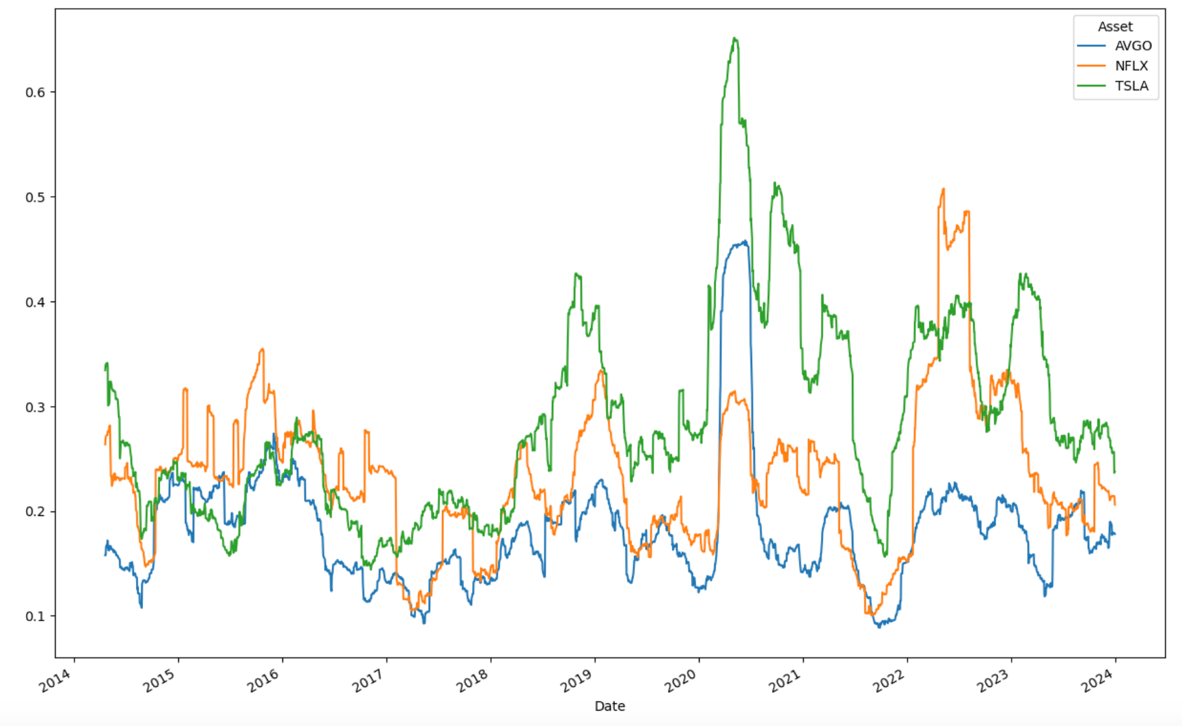
Fig. 8 displays histograms of daily returns for various NASDAQ-listed stocks, illustrating the frequency and distribution of returns, with some showing wider spread indicating higher volatility. Fig. 9 is a matrix of scatter plots comparing the daily returns between each pair of these stocks, along with their individual distributions, highlighting correlations and the relative volatility among them.



#### Fig 10: Volatility



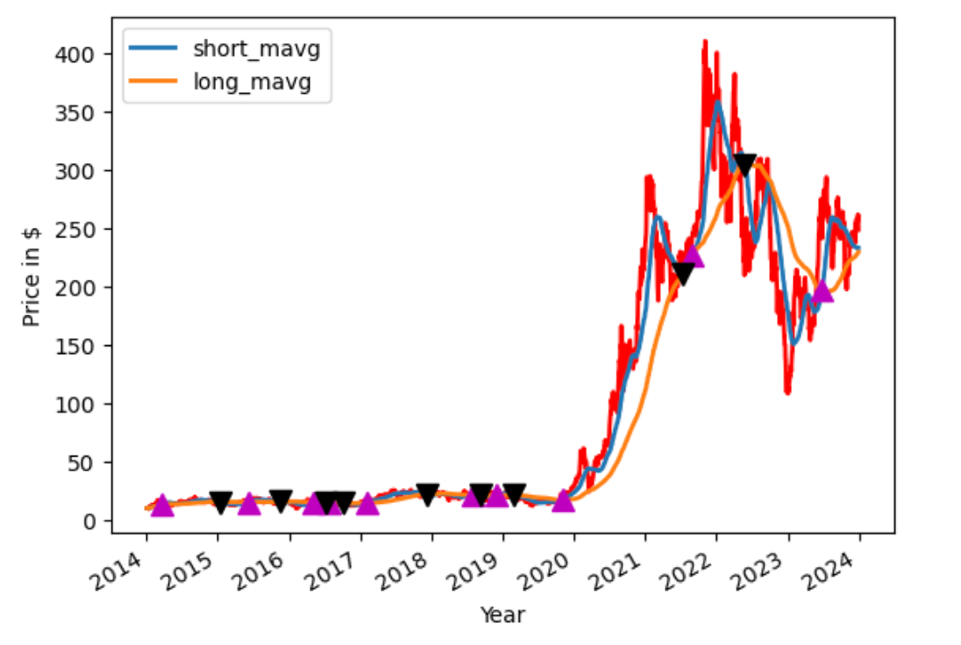
#### Fig 11: Returns of Top 3



#### Fig 12: Volatility of Top 3

### Simple Moving Average (SMA) – Momentum Strategy

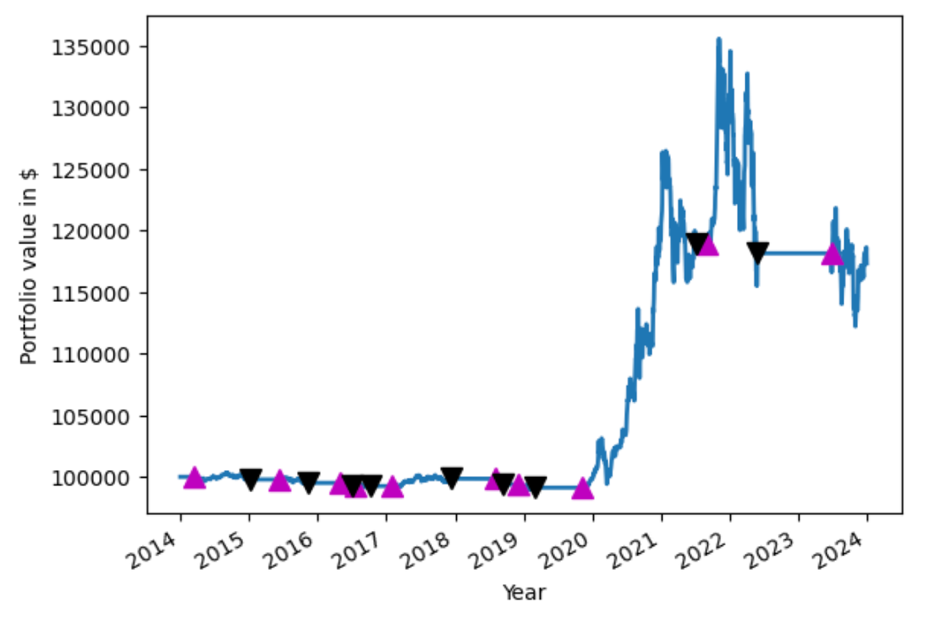
The image depicts a practical application of a momentum strategy to a stock's historical price data over a decade, utilising short-term (10 days) and long-term (45 days) simple moving averages to determine optimal buying and selling points. This method is depicted with the short-term average in blue and the long-term average in orange, with the former crossing above the latter indicating a buying opportunity and an upward trend. A crossover below suggests a selling point, indicating a downward trend. The various coloured triangles on the graph represent the buy and sell signals generated by rolling window calculations of the moving averages. This strategic approach emphasises the significance of trend analysis in investment decision-making, allowing traders to potentially capitalise on the continuation of price movements (Jegadeesh & Titman, 1993).



#### Fig 13: Momentum Strategy

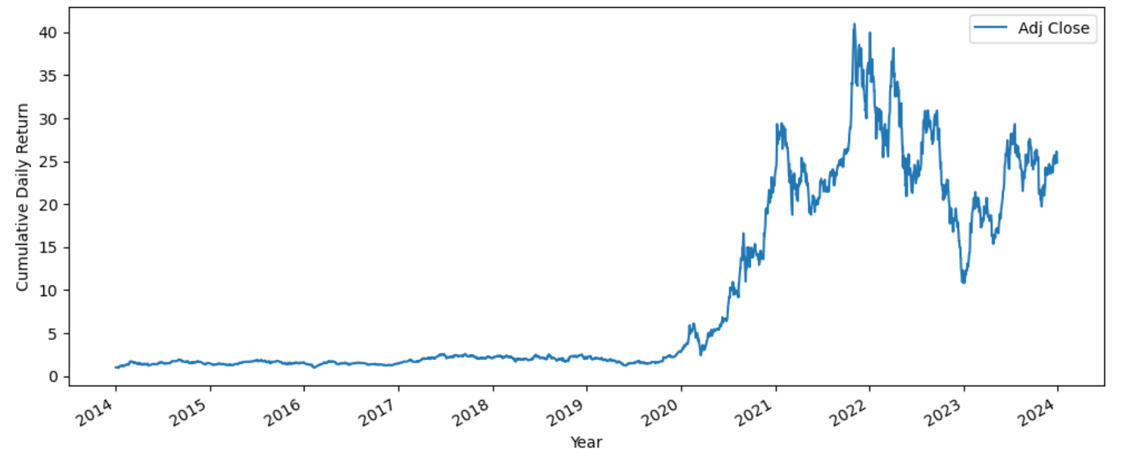
### Backtesting

Backtesting is a technique used to assess the viability of a trading strategy by simulating its application on historical data. In this case, the strategy entails managing an initial capital of $150,000 and carrying out trades involving 150 shares each based on predefined buy and sell signals over a 10-year period. This process not only assists in understanding the strategy's potential under different market conditions, but it also identifies periods of underperformance or exceptional returns. The Sharpe ratio is used to evaluate the strategy's performance. It calculates risk-adjusted returns by comparing the investment's returns to a risk-free rate. A higher Sharpe ratio indicates that the returns compensate adequately for the risks taken, making it an important metric for determining the effectiveness of a trading strategy. Through backtesting, traders can refine their strategies, adjusting parameters to enhance profitability and reduce risk, offering a comprehensive tool for strategic decision-making in financial markets.



#### Fig 14: Backtesting

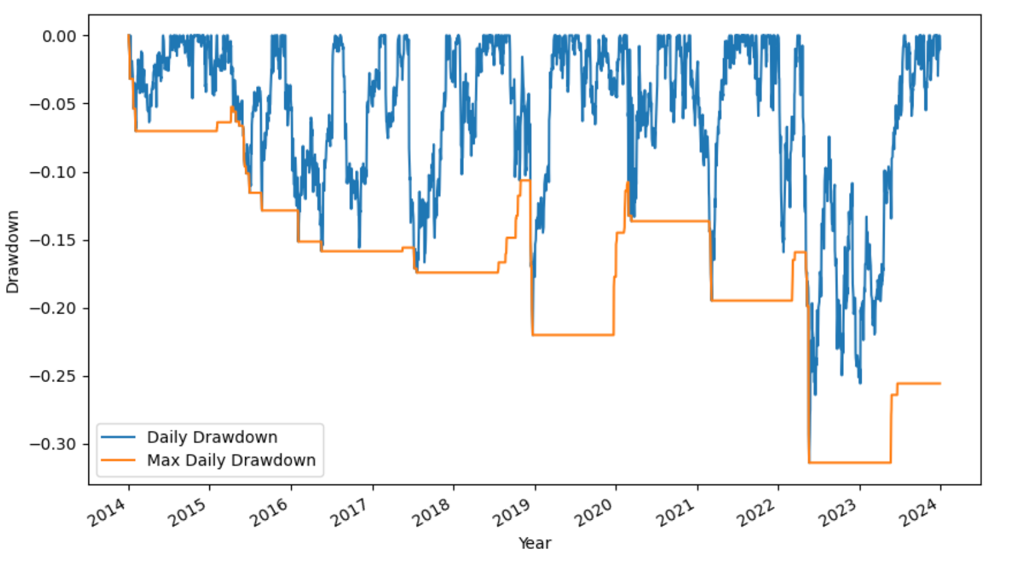
The Sharpe ratio is calculated to support the evaluation.



#### Fig 15: Cumulative Returns

### Maximum Drawdown

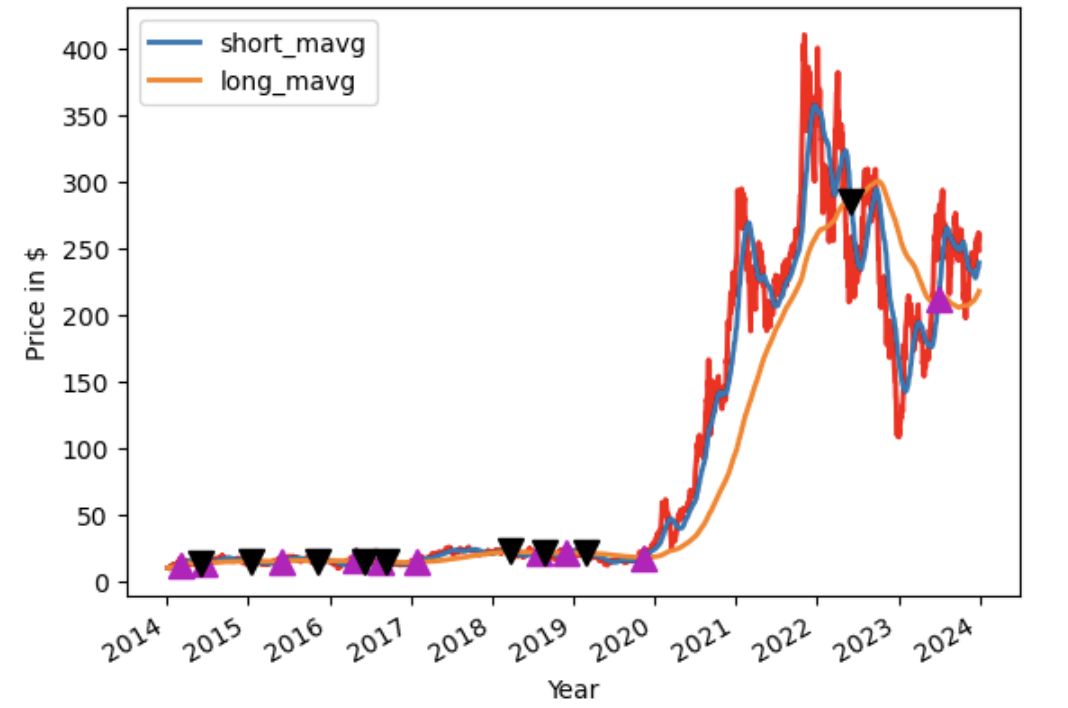
Fig. 16 specifically shows the Maximum Drawdown experienced by a portfolio consisting of 10 selected assets over a decade, with the most severe drawdown reaching -30%. This significant figure highlights the high-risk nature of the strategy employed, particularly evident in assets such as TSLA and NFLX, which are known for their volatility. The pronounced drawdown reflects substantial potential declines from peak portfolio values, illustrating periods of high volatility and the associated risk of significant losses. While the strategy has managed to secure a high Compound Annual Return of 41.2%, the steep drawdowns indicate a considerable risk exposure, requiring investors to be cautious and prepared for potential substantial capital reductions, especially during turbulent market phases. This graph effectively underscores the need for careful risk management when engaging in high-return strategies in volatile asset classes.



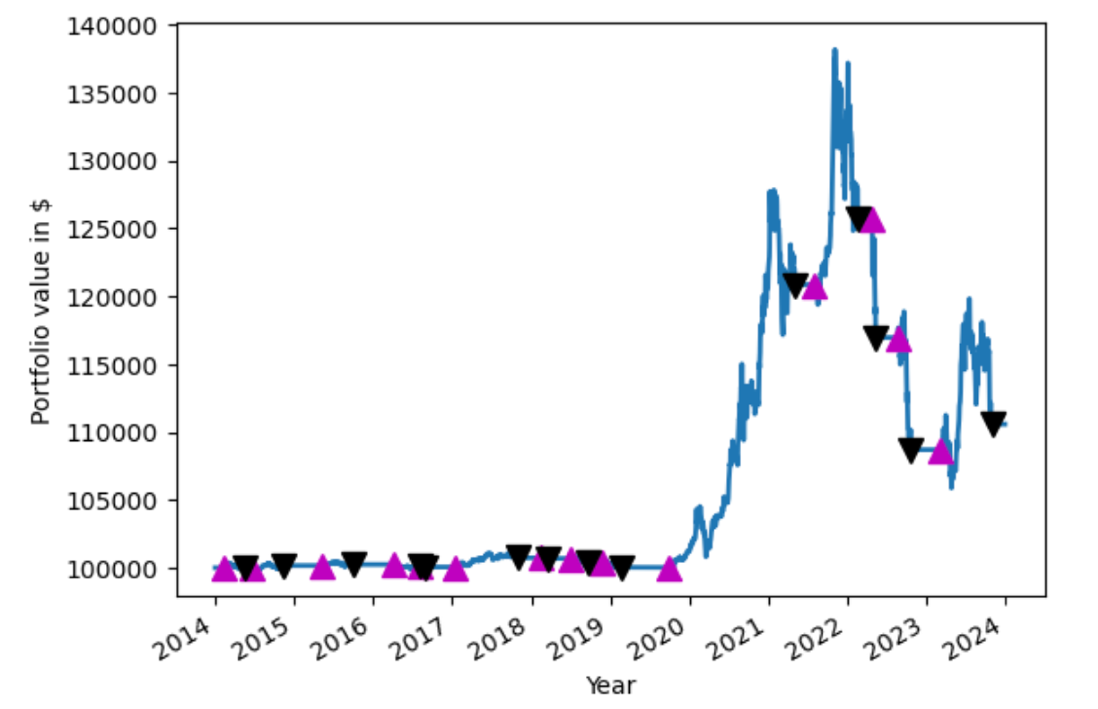
#### Fig 16: Maximum Drawdown

### Optimize Strategy

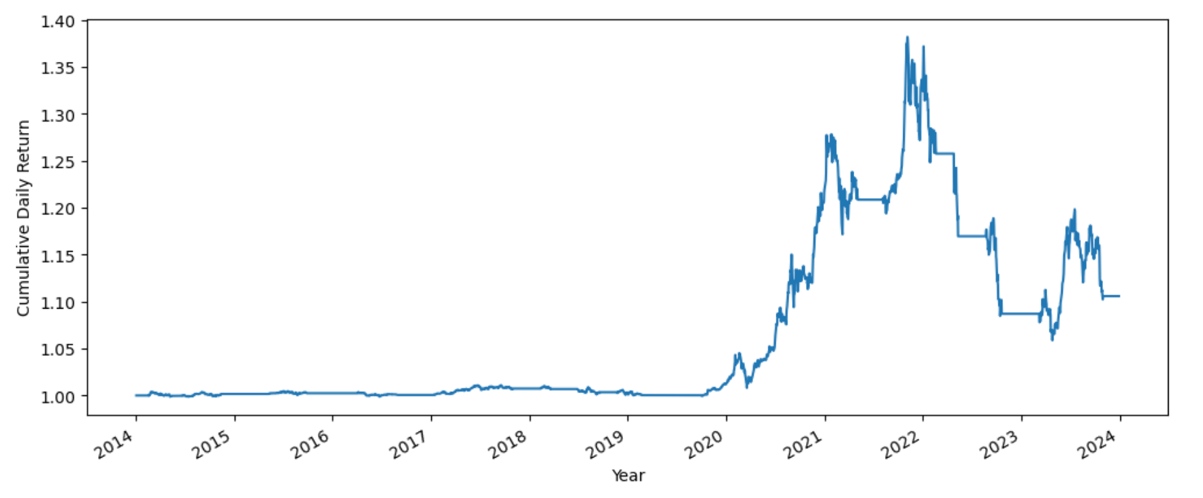
To optimize the trading strategy and enhance its performance, we adjusted the moving average windows to more responsive settings—shifting the short moving average to 30 days and the long moving average to 120 days. This modification allowed the strategy to react more swiftly to market fluctuations, enabling it to capitalize on profitable opportunities more effectively. As a result of these changes, the Sharpe Ratio improved from 0.2916 to 0.3548, indicating a more efficient balance between return and risk. The tighter moving average windows significantly enhance the strategy’s sensitivity to trend changes, offering potential for quicker entries and exits which is crucial in volatile markets. This adaptive approach not only optimizes the performance but also increases the strategy's robustness, making it more resilient against market uncertainties and fluctuations, thereby safeguarding the investment against potential downturns.



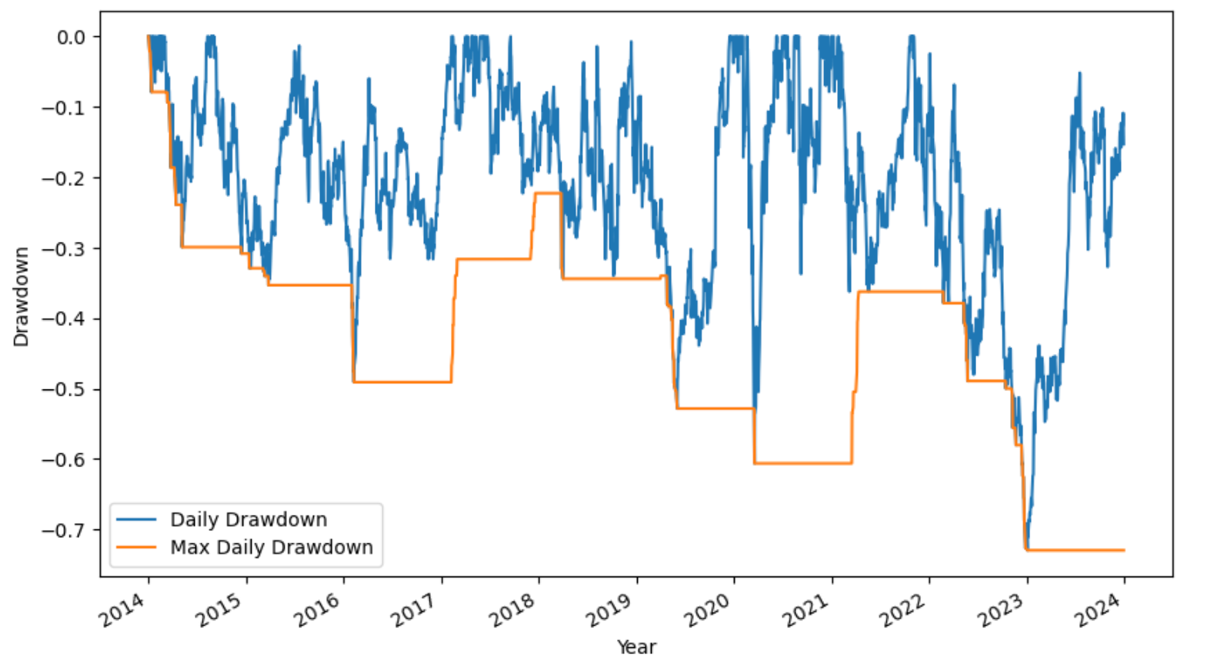
#### Fig 17: Optimized Momentum Strategy



#### Fig 18: Optimized Backtesting



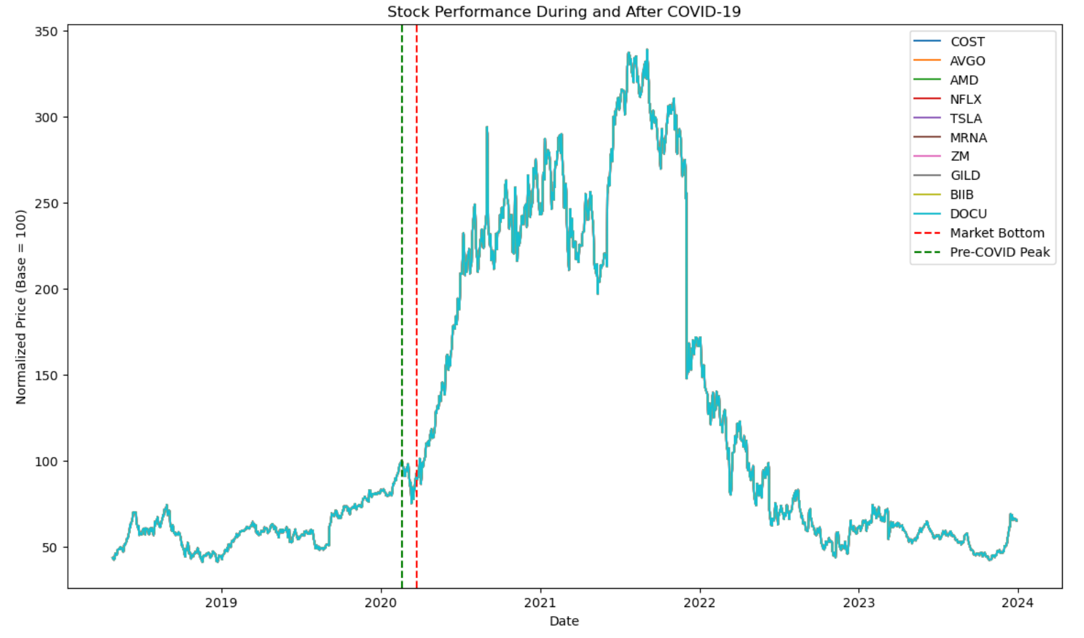
#### Fig 19: Optimized Cumulative Returns



#### Fig 20: Optimized Maximum Drawdown

### Recovery from Covid

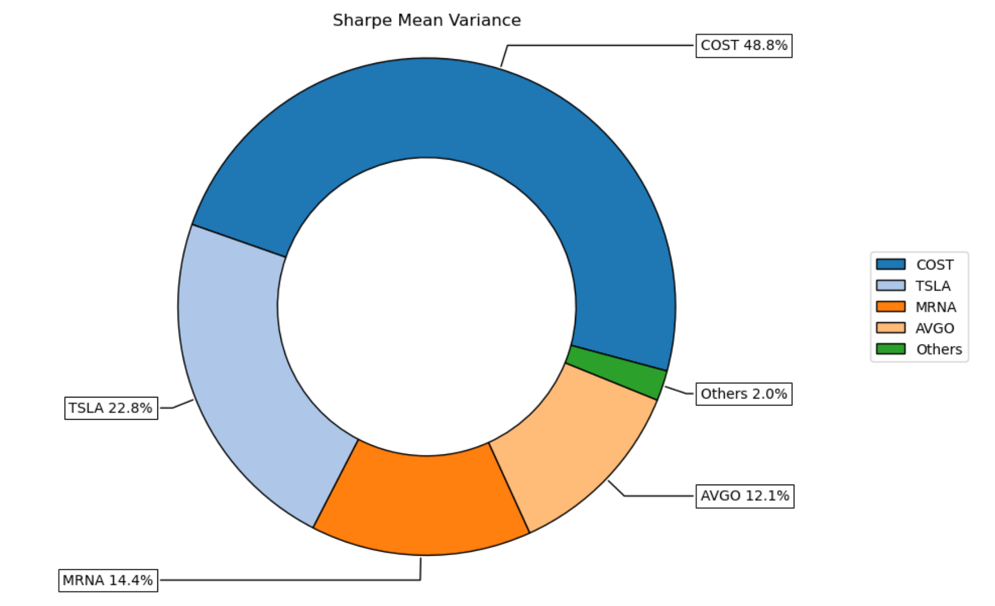
The graph depicting stock performance during and after COVID-19 shows distinct recovery trajectories across sectors, with noticeable differences in how quickly different industries recovered. Moderna (MRNA) has made an exceptional recovery, with its stock price nearly tripling from the market bottom in March 2020 to peak values in late 2020, highlighting its critical role in vaccine development. Tesla (TSLA) also demonstrated remarkable resilience, with its stock price doubling from around $100 in early 2020 to more than $200 by the end of the year, capitalising on the growing interest in electric vehicles and sustainable technologies. In contrast, Costco (COST) and Gilead Sciences (GILD) demonstrated more gradual recoveries. Costco's stock increased by about 30% from 140 to 180 over several months after the market bottom, owing to consistent demand for consumer staples, whereas Gilead's stock increased by about 20% in the first few months after the market bottom due to its role in developing COVID-19 therapeutic solutions, but then fluctuated in response to the evolving pandemic landscape. This data shows that companies directly involved in pandemic response or in high-growth tech sectors not only recovered quickly, but also capitalised on new market opportunities and consumer trends triggered by the pandemic.



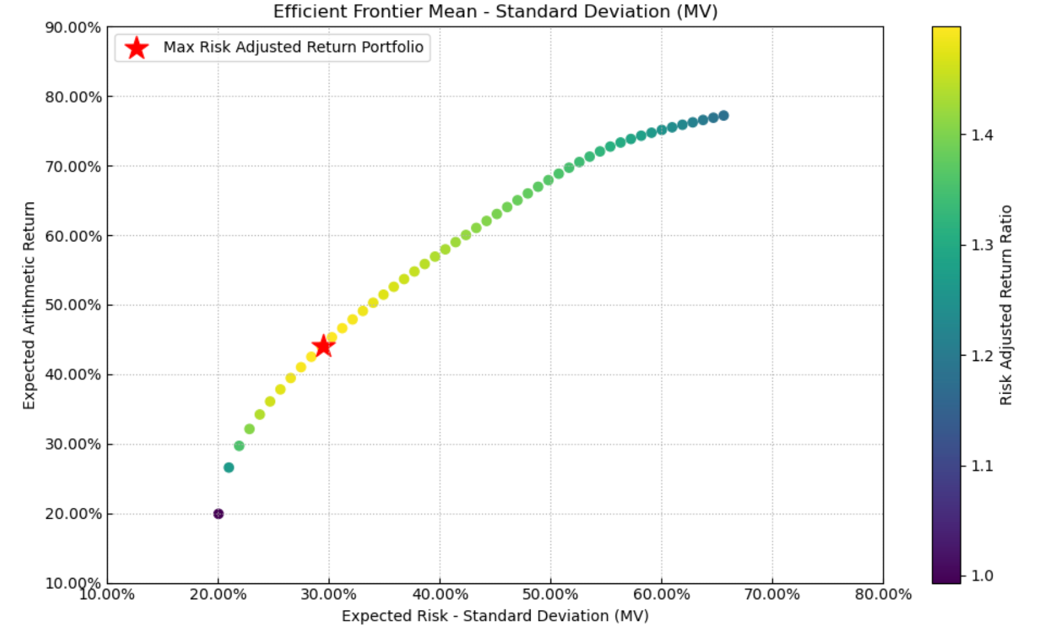
#### Fig 21: Stock Performance During and After Covid

### Portfolio Optimization

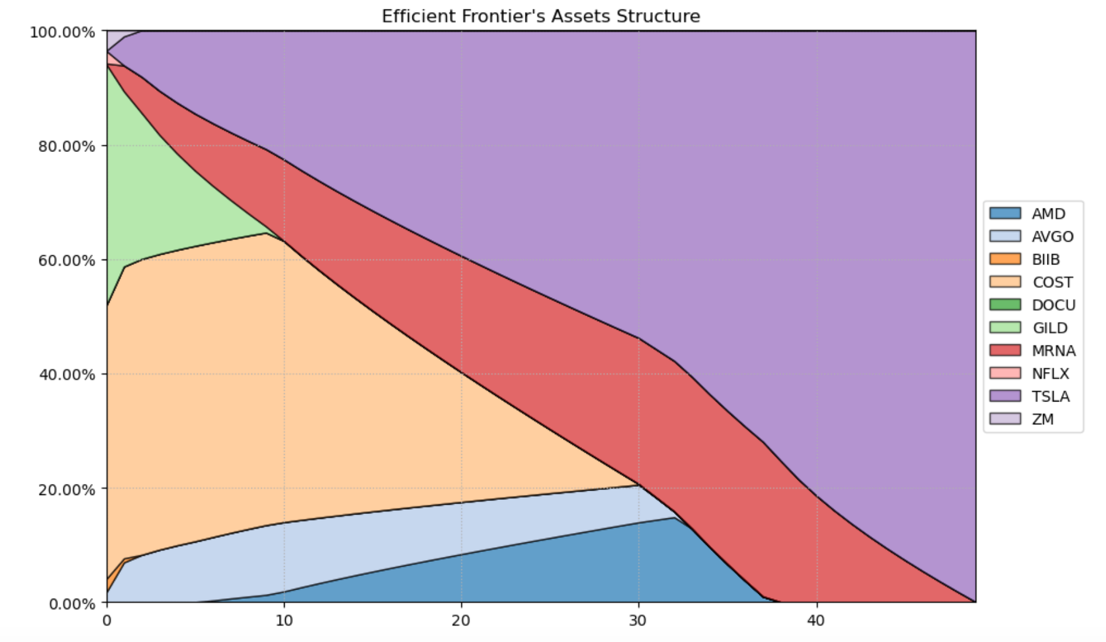
To build and optimise a portfolio for several risk measures, we examined the performance of various assets such as COST, TSLA, MRNA, and AVGO. Initially, the portfolio's Sharpe Ratio was 0.2916, but it was increased to 0.3548 by adjusting the moving average windows to 30 and 120 days. The Sharpe Mean Variance graph shows that COST (48.8%) and TSLA (22.8%) make significant contributions to the portfolio's risk-adjusted return, followed by MRNA (14.4%) and AVGO (12.1%). The Efficient Frontier graph with standard deviation shows that the optimal portfolio has a 60% expected arithmetic return and a risk level of about 35% standard deviation. The asset structure graph shows the allocation changes, with TSLA and MRNA becoming more prominent as risk-adjusted returns improve. Furthermore, the heatmap reveals insights into the best asset allocation across various risk measures. For example, COST dominates mean-variance (MV) with 48.8%, maximum drawdown (MDD) with 37.16%, and average downside deviation (ADD) with 58.27%. TSLA outperforms in the conditional value at risk (CVaR) with 23.26% and the worst return (WR) with 23.49%. This diversified allocation ensures a balanced approach by leveraging high-performing assets to improve overall portfolio performance in changing market conditions.



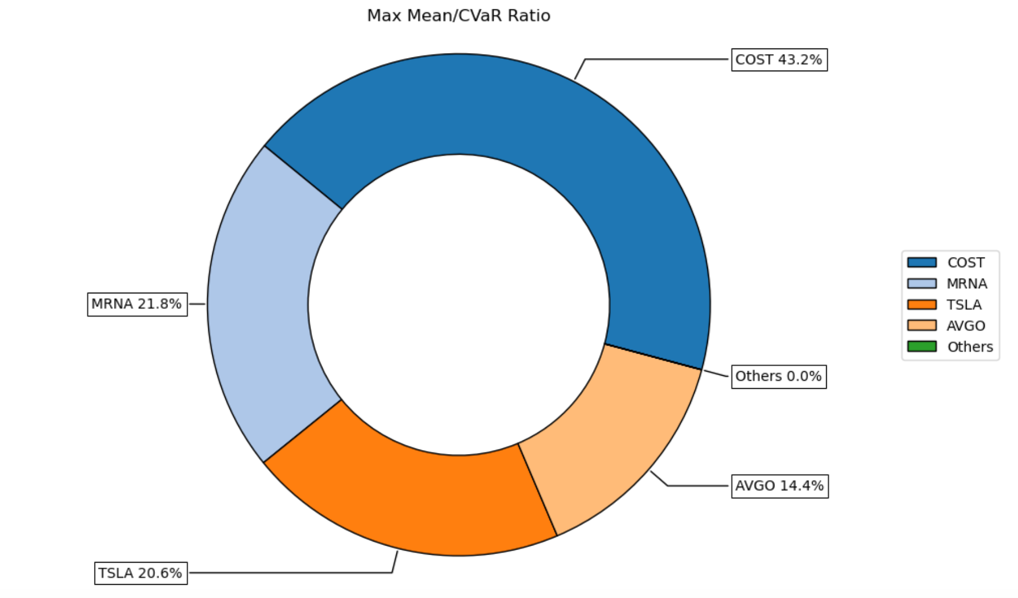
#### Fig 21: Composition of Portfolio Sharpe Mean Variance



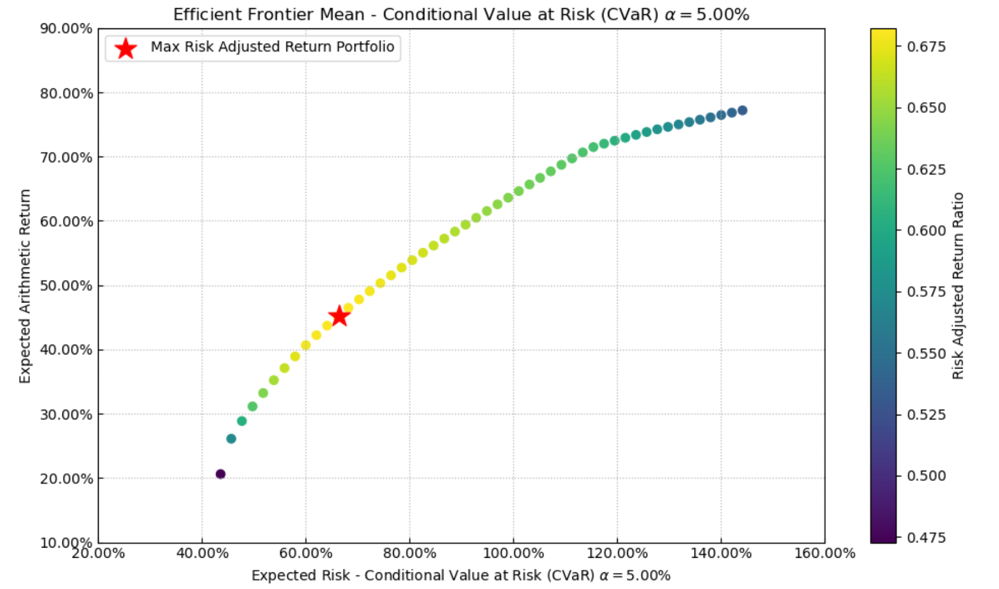
#### Fig 21: Efficient Frontier Mean - Standard Deviation (MV)



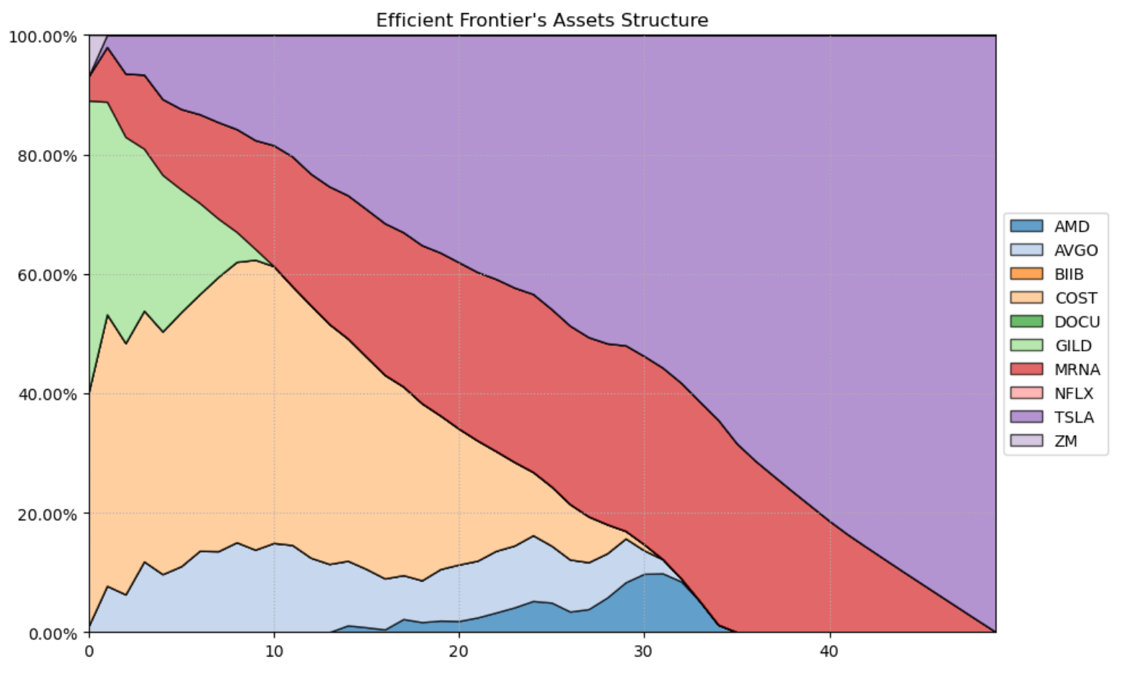
#### Fig 21: Composition of Efficient Frontier



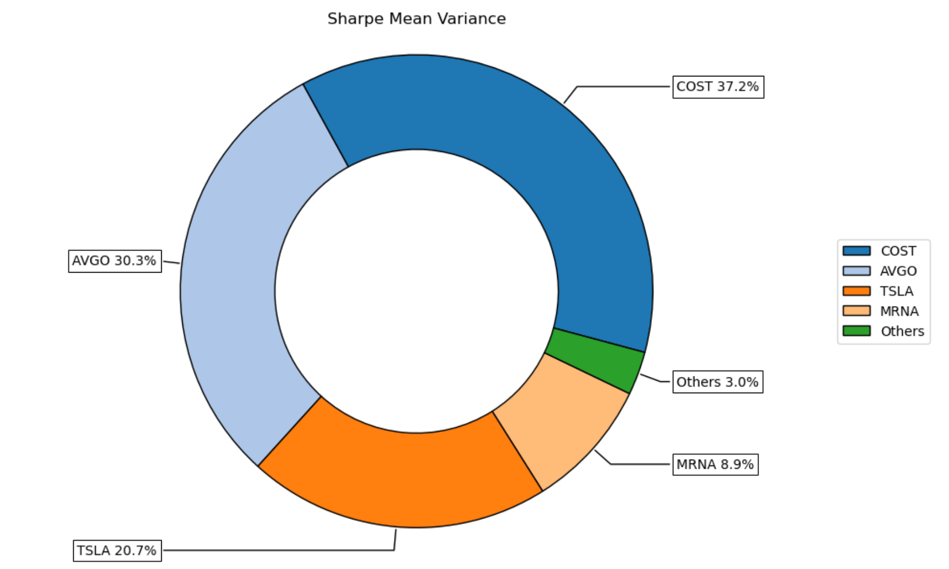
#### Fig 21: Max Mean/CVaR Ratio



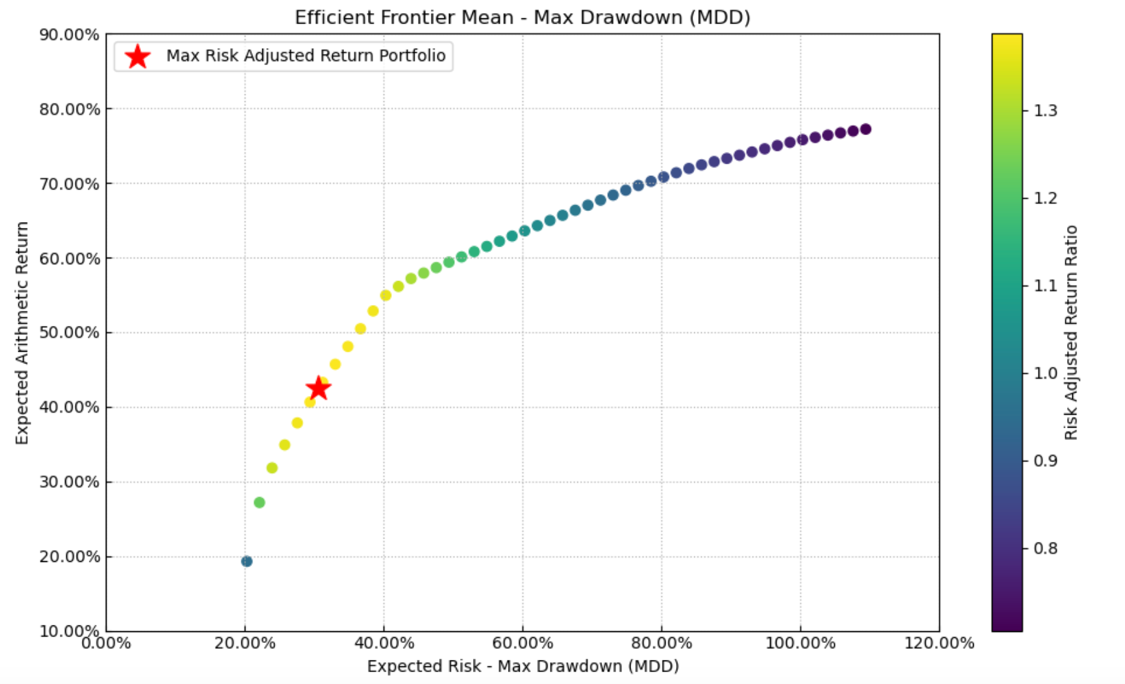
#### Fig 21: Efficient Frontier Mean – Conditional Value at Risk



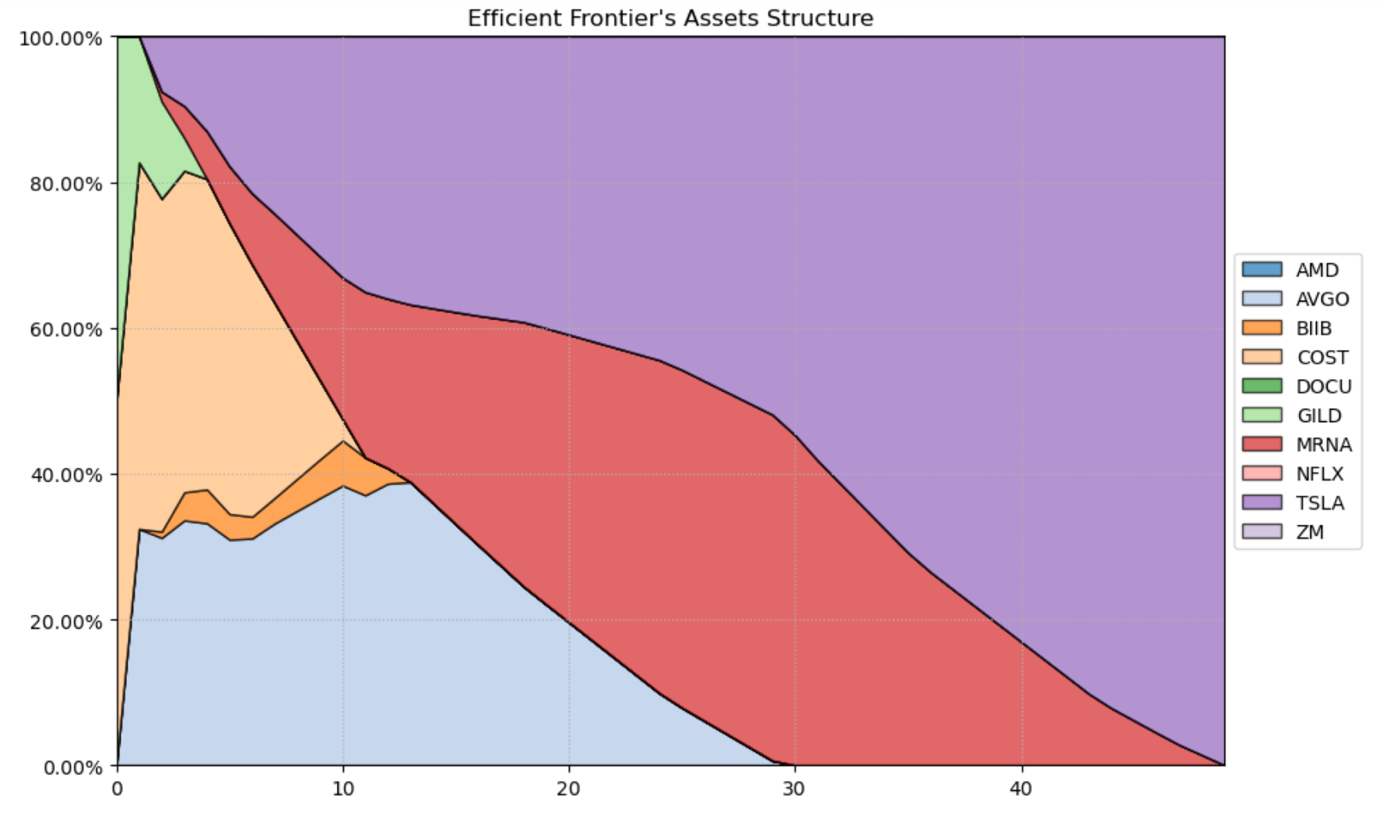
#### Fig 21: Efficient Frontier Asset Structure



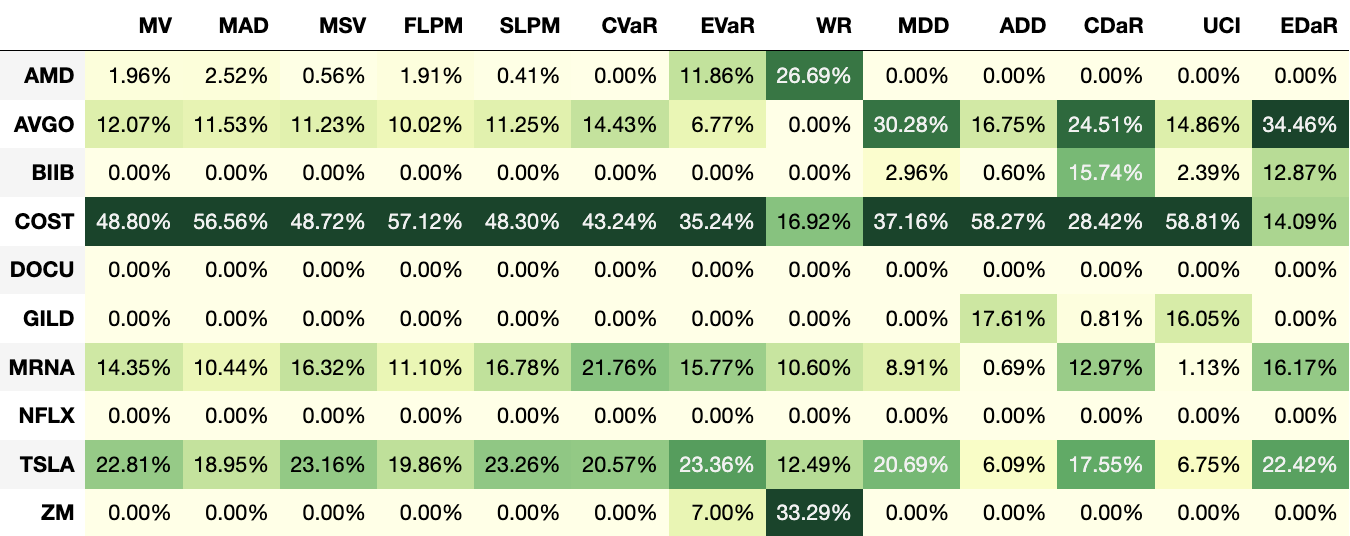
#### Fig 21: Sharpe Mean Variance



#### Fig 21: Efficient Frontier Mean – Max Drawdown



#### Fig 21: Efficient Frontier Asset Structure



#### Fig 21: Optimal Portfolio for Several Risk Measures

# References

Breiman, L., 2001. Random Forests. Machine Learning, 45(1), pp.5-32.

Cristianini, N. and Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press.

Feldman, R., 2013. Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), pp.82-89.

Hochreiter, S. and Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation, 9(8), pp.1735-1780.

Hyndman, R.J. and Athanasopoulos, G., 2018. Forecasting: principles and practice. Melbourne: OTexts.

Jegadeesh, N. and Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance, 48(1), pp.65-91.

Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv:1408.5882.

Liu, B., 2012. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

McCallum, A. and Nigam, K., 1998. A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization. pp.41-48.

Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), pp.1093-1113.

Pang, B. and Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), pp.1-135.

Zhou, Q., Wu, X., Ding, S. and Wang, Z., 2019. A survey on time series data mining. Pattern Recognition Letters, 131, pp.38-54.