I. TASK 0

First, we read the data from the csv file into a pandas dataframe. Then we check for missing values. We check for the number of null values in each column. There are no missing values in the dataset. If there were missing values, appropriate missing value handling techniques would have been applied. The rows that consist of more than or equal to 30% missing values will be dropped. Then forward fill would be applied to fill in the missing values with the previous available value. Forward fill is typically used to fill missing stock prices as the missing stock price is expected to be similar to the most recent observation.

II. TASK 1

Appropriate analysis of the data is started after handling of missing values. First, visual inspection of data is done with plotting the stock prices of 'KO' with time. This demonstrates us that data is non-stationary and has an increasing trend with time. The visual plot of our data is given below:

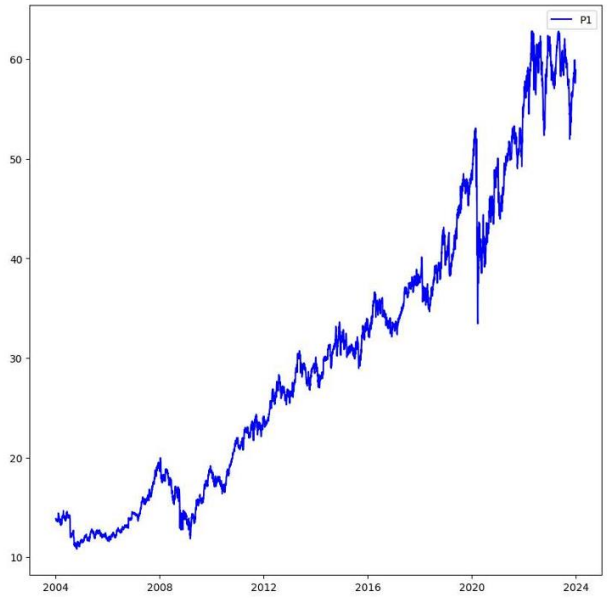


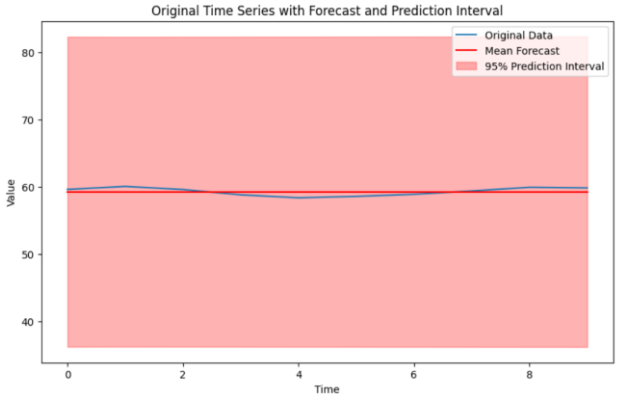
Figure 1: Our time series data

Then to further confirm, an Augmented Dickey-Fuller Test is performed on the data. As expected, it confirms time series has a unit root, indicating it is non-stationary. So, differencing needs to be carried out on the dataset. We choose Auto-regressive Integrated Moving Average(ARIMA) model for our particular dataset as it takes into account the differencing needed for non-stationary data as well as the AR and MA parts. It is good model as it handles trends and does not assume stationarity. Furthermore, it is suitable for our case as we do automated model selection using Bayesian Information Criteria(BIC) technique to select appropriate parameters for the model. Different models with varying p, d and q are compared and the model with the lowest BIC value is chosen as lowest BIC value means best trade-off between the fit and complexity. BIC is a widely used statistical measure for model selection. We also plot ACF and PACF for our selected model to verify the accuracy of the BIC technique as an extra layer of verification. As aforementioned there is a q parameter. ACF is used to select that parameter and PACF is used to select the p parameter. GARCH model is applied to predict the volatility of the stock returns. GARCH stands for Generalized Autoregressive Conditional Heteroskedasticity. It is used to estimate the volatility of time series data. The same BIC technique is used for choosing appropriate parameters for the model. The table of the final predictions are given below:

Table 1:final preictions

ASSET	Mean forecast	Average volatility
KO	59.27	11.77
PG	148.69	21.86
MSFT	338.00	98.04
WMT	52.01	6.22
JPM	142.42	20.82

The plots are more informative and visually appealing and comparing with the original data to evaluate our model. Then we plot the mean and variance forecasts within 95% prediction volatile interval and compare with the original data. The 95% prediction intervals indicate a range within which future observations are expected to fall with 95% probability. The plots indicate our models are predicting relatively accurately and closely to the original data. Figure 1 shows us the mean and volatility forecasts for the 'KO' stock price.



The forecasts for the other four assets are:

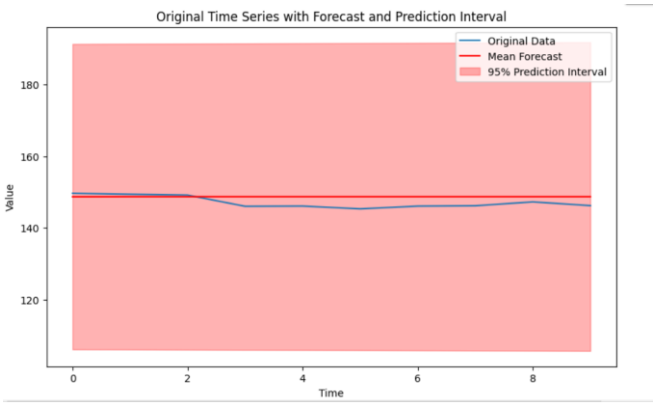


Figure 2: forecast for PG

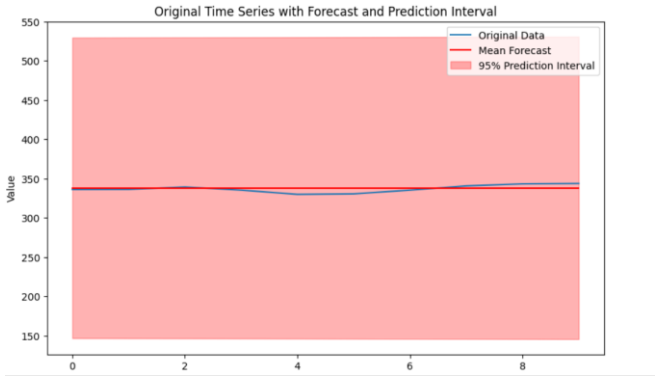
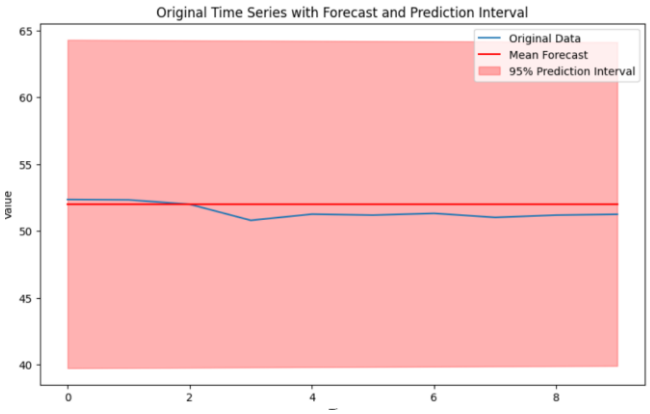
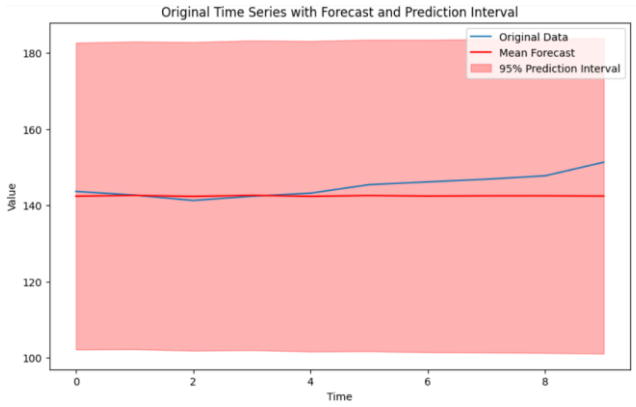


Figure 3:forecast for MSFT

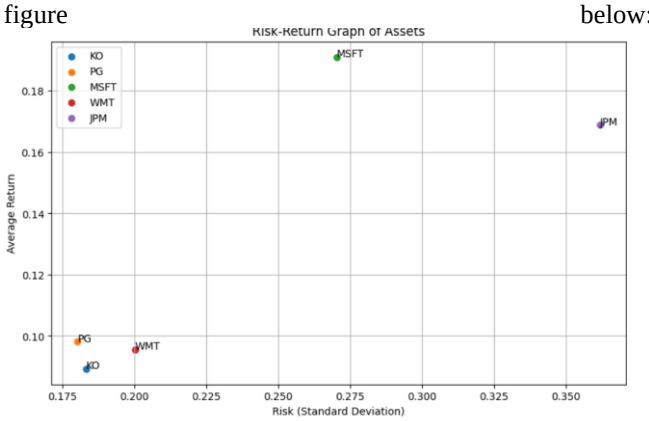




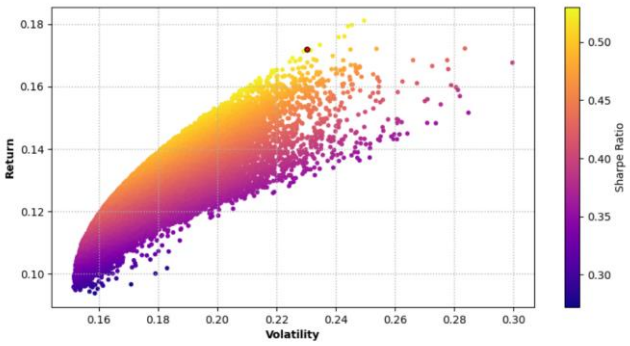
The above figures suggest that the predicted means lie very close to the original data and this is true atleast for the short term. All future observations are expected to fall within the 95% prediction interval. This is in line with our original data, which also falls within the 95% prediction interval. This suggests that the model takes a very good account of the volatility of stock prices, i.e. the risks involved in each asset. Investors using our model will always be safe from risks.

III. TASK 2

Historical value has been chosen for the dataset as risk and return predicted from task 1 depends on the accuracy of our model. So there is a significant uncertainty involved. Historical data for the last twenty years is readily available and provides a comprehensive view of the performance of the selected stocks. Historical data is more robust as it reflects actual market behaviour. Furthermore, ARIMA model is not good for long term forecasting or prediction. We choose five years of daily data from the dataset as it is recent enough to reflect current market conditions and long enough to smooth out short term volatility. It provides a large enough sample for robust statistical analysis. Then we visually inspect the time series of each stock price over time for all the five stocks. We calculate the daily returns and average return and risk for each asset, and then plot them on a risk and return graph, which provides with a good comparison about different risks and returns of different assets. This is demonstrated in the figure below:



Before doing the Monte Carlo simulation, we set the random seed to 42 so that the work can be replicated. Then the monte-carlo simulation is carried out. Monte Carlo simulation is a popular method used in portfolio optimization [3]. In this simulation a large number of portfolios are generated and in each portfolio, random weights are assigned to the five stocks. The sum of the weights equals 1. Then the return and risk of each portfolio is calculated using the weighted sums of individual returns and covariance matrix. The expected return for a portfolio is calculated by finding the average daily return for each asset and then by annualizing it(multiplying by 252) and then with the portfolio weight. Similarly, the expected volatility of a portfolio is found by applying the weights to the annualized covariance matrix. Volatility is the standard deviation of a company's stock [2]. Covariance measures the directional relationship between the returns on two assets [2]. Then portfolio weights are applied to the annualized covariance matrix to get a vector of weighted contributions to the variance. These weighted contributions are aggregated with another dot product with the weights. The variance is converted to volatility by taking its square root, i.e. standard deviation. Volatility is the standard deviation which represents the portfolio's annualized risk. Then the risk is plotted against the return for all simulated portfolios. This is illustrated in the figure below:

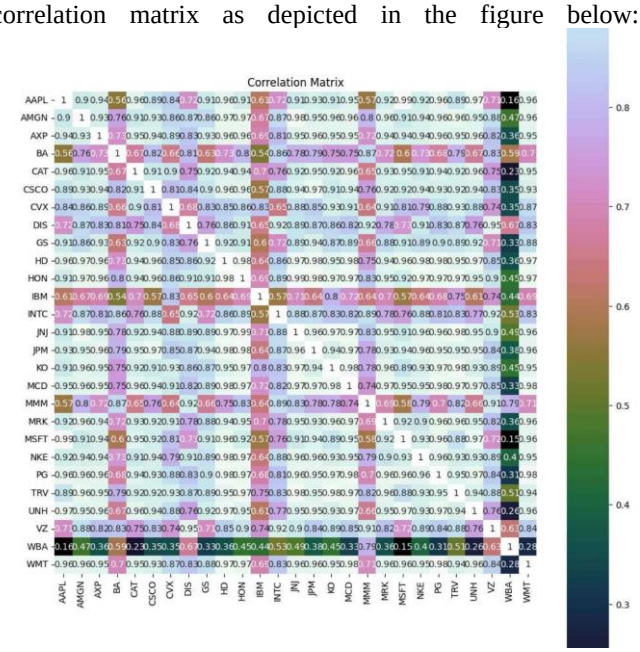


The efficient frontier is given by the portfolios with the highest return for a given level of risk. The sharpe ratio measures the performance of an investment compared to a risk-free asset In portfolio management, the utility function

represents an investor's preference for risk and return. The tangency portfolio is the one with the maximum sharpe ratio the weights for Alice of 'KO', 'PG', 'MSFT', 'WMT', 'JPM' are 0.00314261, 0.02477545, 0.73873334, 0.17885711, 0.05449148 respectively. Bob's portfolio weights are calculated using the same monte carlo simulation as before, but this this using the utility function instead of the sharpe ratio, and the risk aversion factor is set to 1. The portfolio weights for Bob on 'KO', 'PG', 'MSFT', 'WMT', 'JPM' are 0.03717646, 0.09253257, 0.81431634, 0.04332192, 0.01265271 respectively. This is the utility with the maximum portfolio.

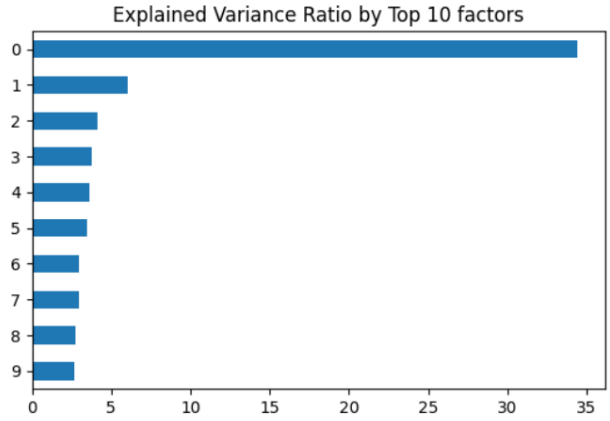
IV. TASK 3

We load the stock price data from the excel file into a pandas dataframe. The correlation matrix is calculated based on these returns. Visualizing these correlation matrix helps us in understanding the relationship between the stocks. Higher correlation suggests that they move together and lower correlations suggest more independent movement. Principal Component analysis (PCA) is then applied to it. PCA is a dimensionality reduction technique used to transform a large set of correlated variables into a smaller set of uncorrelated variables called Principal components. It is used because it reduces the number of variables while preserving as much variance as possible. It identifies the most important features that explain the most variance in the data by eliminating less important components that maybe considered as noise. Better diversification is achieved when allocating investments among uncorrelated components. We calculate the correlation matrix as depicted in the figure below:

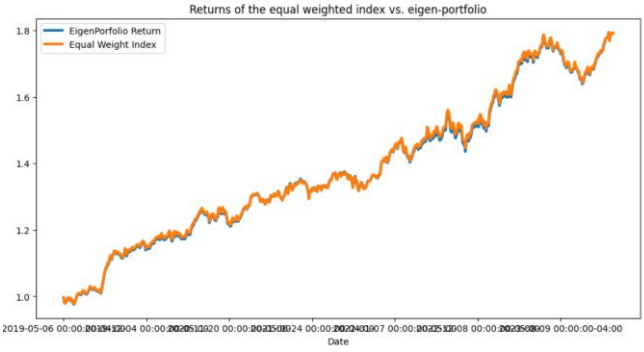


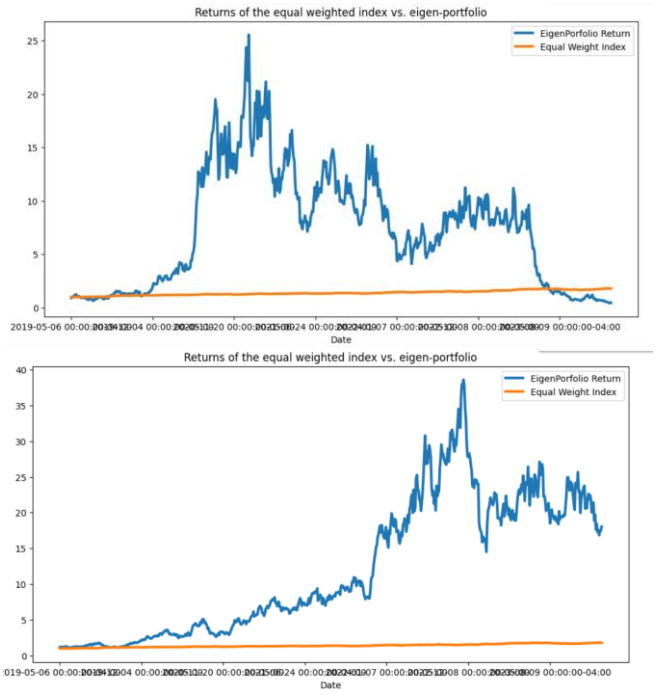
As it can be seen in chart above, there is a significant positive correlation between the stocks. Then we calculate the daily log returns of the data. Outliers are handled by removing points beyond 3 standard deviation and use Standardization to transform attributes to a standard normal distribution with a mean of 0 and standard deviation of 1. This is needed as all the variables should be on the same scale before applying visualization, otherwise a feature with large values will dominate the result. After standardization, the data is split into 80% training set and 20% testing set. The training set is used in the portfolio construction and the testing set is used for thorough evaluation of the model. The model is fitted on the training set. The PCA model computes the mean vector and Covariance matrix of original data points. After calculating the covariance matrix the eigenvectors and eigenvalues of the covariance matrix are computed. Then it selects the top k eigenvectors and projects original data points onto the subspace spanned by them. After calculating the covariance matrix we calculate the eigenvalues and eigenvectors, which are then sorted in descending order of eigenvalues. Then the top principal components are selected based on the explained variance. In this way it finds a k-dim projection that best preserves variance.

The variance ratio plot is given below:



This explained variance plot helps to understand how much of the total variance in the data is captured by each principal component. The first few principal components explain a large portion of the variance. Component 0 captures a large portion of the variance. We selected the value of k as 10 and it is well justified by the diagram, as the values We extract the eigenvectors corresponding to the principal components. Traditional Normalization approach is used where each component is divided by the sum of components, ensuring that the weights sum to 1. We construct the returns for each of three eigen portfolio and the equal-weighted portfolio. To evaluate our model thorough backtesting is performed where three eigen portfolios are selected and their performance is compared with the equal weighted index by plotting. The plots are given below:

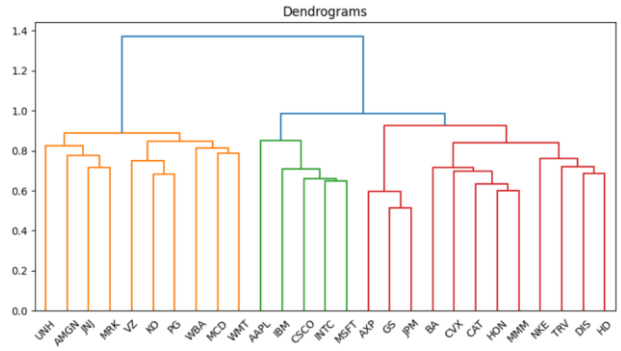




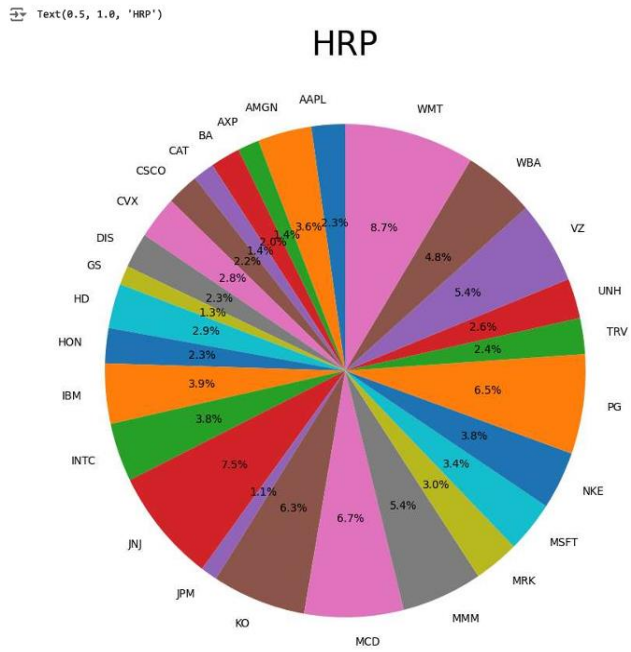
These plots suggests that of our eigen portfolio is either equal to and often greater than the equal weighted index. This indicate the model is performing well.

V. Task 4

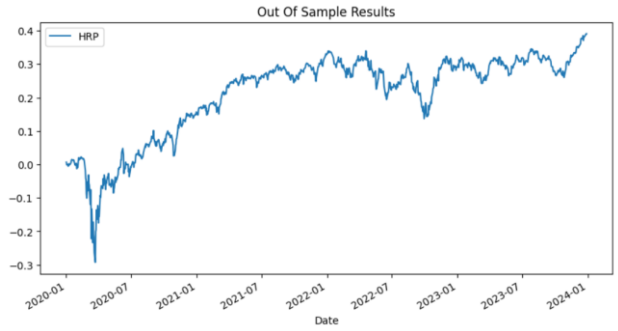
Risk parity is an approach to portfolio management that focuses on the allocation of risk rather than the allocation of capital. The risk parity approach asserts that when asset allcations are adjusted to the same risk level, the portfolio can achieve a higher sharpe ratio and can be more resistant to market downturns. Hierarchical clustering is a popular method of cluster analysis. We calculate the daily returns. First, correlation matrix is calculated then based on it, distance matrix is calculated and then the closest pairs of clusters are merged together iteratively to create form dendograms that records the merges, and the distance matrix is updated. The figure below shows us the plot of the output dendrogram:



This gives us the covariance matrix. Quasi-diagonalization is a process used in Hierarchical risk parity to reorder the covariance matrix so that the largest values lie along the diagonal. This makes sure that similar assets are placed together. The differences between Quasi-diagonalization and clustering is reflected in its purpose, methodology and Outcome. The purpose of clustering in HRP is to identify groups of assets that exhibit similar behaviour, i.e. they are highly correlated with each other, whereas the purpose of Quasi-diagonalization is to reorder the covariance matrix based on the clustering results. This reordering aims to reveal a block-diagonal structure which simplifies the subsequent allocation process. The outcome of the clustering is a hierarchical tree(dendgram) that represents the nested grouping of assets based on their correlations, whereas in quasi-diagonalization the outcome is a reordered covariance matrix that facilitates the applicatioon of recursive bisection for risk allcation. Recursive bisection is used to allocate weights to the portfolio based on hierarchical clustering by splitting allocations through recursive bisection of the reordered matrix. It does this by calculating weighting factor of the sub clusters and assigning the weights to the left and right clusters. The weight of the left and right clusters are updated respectively. This steps are performed recursively until all weights are assigned to the stocks. This gives us ur final portfolio weights. The figure below gives us a visualization of our portfolio weight allcation:



The out of sample prediction is given below:



The out of sample ratio is 0.497421 as shown below:

	stdev_oos	sharp_ratio_oos
HRP	0.196639	0.497421

VI. TASK 5

High-quality data is crucial for the success of machine learning models in portfolio management. Data quality is one of the factors that need to be addressed [6]. The performance of financial decision-making directly concerns both businesses and individuals[8]. So, poor data quality can lead to wrng decisin processes having huge impacts Financial markets generate vast amounts of data, but not all of it is reliable or relevant. The challenge lies in ensuring that the data fed into ML models is accurate, timely, and representative of the underlying market conditions. Poor data quality can lead to overfitting, where the model captures noise instead of underlying patterns, resulting in poor out-of-sample performance. Moreover, financial data often contains missing values, outliers, and non-stationarities (e.g., sudden market regime changes), complicating the training process. Data quality is very important in financial modeling [7]. Cleaner and more precise data significantly improve model performance in predicting stock returns [7]. . Data quality variables are Accuracy, timeliness, completeness, Consistency and data relevance [12]. space shuttle *Challenger* and the USS *Vincennes*/Iranian Airbus disasters were due to poor data quality [12]. A classification of data quality problems is presented in [13]. To effectively manage data quality issues, we can categorize them into a hierarchical framework with four levels: multiple data sources, multiple relations, single relation, and attribute/tuple. At the attribute/tuple level, common problems include missing values, misspellings, and syntax violations. Moving up to the single relation level, issues like approximate and inconsistent duplicate tuples are prevalent. At the multiple relation level, challenges such as referential integrity violations or incorrect references arise. Finally, at the level of multiple data sources, we encounter heterogeneity in syntaxes, measurement units, and data representation, among other issues. This hierarchical organization helps in systematically identifying and addressing data quality problems at each level. Financial time series data has a higher chance of becoming inaccurate. This is because often they include a very high volume of data, for instance, 20 years of daily prices of certain stocks. One method to remove outliers in this kind of data could be apply rules based on financial domain knowledge to identify values that are clearly erroneous like stock prices that deviate drastically without a known cause. We could implement automated scripts to validate incoming data against predefined rules and historical patterns. We should verify such data from multiple financial sources to ensure accuracy and consistency. Discrepancies can be flagged for further investigation. Oultiers impact data and robust statistical techniques should be used to mitigate their effects. Implementing rigorous data cleaning procedures, including the imputation of missing values, removal of outliers, and normalization of data. Utilizing methods like robust regression and regularization techniques (e.g., Lasso, Ridge) to handle anomalies and improve model robustness. Enhancing data quality through the creation of more informative features. This involves transforming raw data into a more suitable form for modeling, often capturing underlying market trends and patterns more effectively. Interpretability remains a significant hurdle in integrating complex ML models into portfolio management. Financial professionals must understand and trust the decision-making process of these models, especially when they are used to manage substantial sums of money. Traditional models like linear regression or logistic regression offer high interpretability but might lack the predictive power of more complex models like deep neural networks or ensemble methods. Complex ML models are often viewed as "black boxes," making it difficult to explain why certain decisions were made. This opacity can lead to a lack of trust and reluctance to adopt these models in practice, especially in regulated industries where explainability is crucial. A paper by Doshi-Velez and Kim (2017) provides a comprehensive overview of the interpretability problem in ML, emphasizing the need for models whose decisions can be easily understood by humans . Additionally, a study by Lipton (2016) discusses the tension between accuracy and interpretability in ML models and explores various methods to enhance the latter without sacrificing performance .Techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) can provide post-hoc interpretations of complex models by approximating their behavior with simpler, more interpretable models.[8] Using inherently interpretable models such as decision trees or linear models with interaction terms. While these models might not capture all complexities, they offer a good balance between interpretability and performance. Developing advanced visualization tools to help stakeholders understand model predictions and their driving factors. Tools such as partial dependence plots and individual conditional expectation plots can be valuable. The integration of machine learning into portfolio management is fraught with challenges, particularly regarding data quality and interpretability. Addressing these challenges requires a combination of rigorous data preprocessing, robust statistical techniques, and advanced interpretability methods. By focusing on these areas, it is possible to harness the power of machine learning while maintaining the trust and confidence of financial professionals and stakeholders.

REFERENCES

[1] https://medium.com/analytics-vidhya/arima-garch-forecasting-with-python-7a3f797de3ff

[2] https://www.machinelearningplus.com/machine-learning/portfolio-optimization-python-example/?utm_content=cmp-trueI. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] Mukherjee, A., Singh, A.K., Mallick, P.K. and Samanta, S.R., 2022. Portfolio Optimization for US-Based Equity Instruments Using Monte-Carlo Simulation. In *Cognitive Informatics and Soft Computing: Proceeding of CISC 2021* (pp. 691-701). Singapore: Springer Nature Singapore.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[4] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[5] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[6] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[7] Frank, Edwin & Ray, Ramon. (2024). Machine learning in portfolio management and asset allocation.

[8] Kogan, Shimon & Levin, Dimitry & Routledge, Bryan & Sagi, Jacob & Smith, Noah. (2009). Predicting Risk from Financial Reports with Regression.. 272-280. 10.3115/1620754.1620794.

[9] Avramov, Doron and Guofu Zhou. “Bayesian Portfolio Analysis.” *Review of Financial Economics* 2 (2010): 25-47.

[10] Du, J. and Zhou, L., 2012. Improving financial data quality using ontologies. *Decision Support Systems, 54*(1), pp.76-86.

[11] https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models

[12] Fisher, C.W. and Kingma, B.R., 2001. Criticality of data quality as exemplified in two disasters. *Information & Management*, *39*(2), pp.109-116.

[13] P. Oliveira, F. Rodrigues and P. R. Henriques, "A formal definition of data quality problems", *IQ*, 2005.