

Literature Review for Portfolio Recommender:

1. Nabipour, P. Nayyeri, H. Jabani, S. Shahab, A. Mosavi, Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis, IEEE Access 8 (2020) 150199–150212.
 - **Model:** Decision Tree, Bagging, Random Forest, Adaboost, Gradient Boosting, XGBoost (Tree-Based models). ANN, RNN, LSTM (Neural Networks).
 - **Dataset:** Diversified financials, petroleum, non-metallic minerals, and basic metals from the Tehran Stock Exchange.
 - **Methodology:** Ten technical indicators are utilized as inputs to the models. The study includes two different approaches for inputs, continuous data and binary data, to investigate the effect of preprocessing; the former uses stock trading data (open, close, high and low values) while the latter employs preprocessing step to convert continuous data to binary one.
 - **Conclusion/Results:** LSTM was the top performer with the lowest amount of error and best fitting ability. But specifically for Tree-Based models, Adaboost Regressor stood out the most with accuracy, fit, and runtime.
 - **Link:** [http://refhub.elsevier.com/S2772-6622\(21\)00010-2/sb38](http://refhub.elsevier.com/S2772-6622(21)00010-2/sb38)
2. X. Zhong, D. Enke, Predicting the daily return direction of the stock market using hybrid machine learning algorithms, Financ. Innov. 5 (1) (2019) 1–20
 - **Model:** Deep neural networks (DNNs) and traditional artificial neural networks.
 - **Dataset:** S&P 500 2003–2013
 - **Methodology:** This paper presents a comprehensive big data analytics process to predict the daily return direction of the SPDR S&P 500 ETF (ticker symbol: SPY) based on 60 financial and economic features. DNNs and traditional artificial neural networks (ANNs) are then deployed over the entire pre-processed but untransformed dataset, along with two datasets transformed via principal component analysis (PCA), to predict the daily direction of future stock market index returns. While controlling for overfitting, a pattern for the classification accuracy of the DNNs is detected and demonstrated as the number of the hidden layers increases gradually from 12 to 1000.
 - **Conclusion/Results:** The trading strategies guided by the DNN classification process based on PCA-represented data perform slightly better than the baseline model.
 - **Link:** <http://jfin-swufe.springeropen.com/articles/10.1186/s40854-019-0138-0>
3. Vivek Palaniappan, Neural networks to predict the market, 2018
 - **Model:** MLP and LSTM (RNN)
 - **Dataset:** Past ten days of stock price data were obtained from Yahoo Finance. Keras (a software library, open-sourced). It also helped with the Python Interfaces for the networks.
 - **Methodology:** The entire Data is divided by 200, which helps by making the weights in the neural network not too large. AdaGrad and RMSProp are used to maintain the performance.
 - **Conclusion/Results:** The LSTM model proved quite satisfactory but could do better with further working on the algorithm. The stock prices of Apple were forecasted and almost matched.

- **Link:** <https://towardsdatascience.com/neural-networks-to-predict-the-market-c4861b649371>
4. M. Gurjar, P. Naik, G. Mujumdar, T. Vaidya, Stock market prediction using ann, Int. Res. J. Eng. Technol. 5 (3) (2018) 2758–2761
 - **Model:** Moving Averages, Stochastic Oscillator, Standard, Deviation, On-Balance-Volume
 - **Dataset:** The stock prices of NSE like NIFTY 50, CNX, and S&P
 - **Methodology:** Project tries to predict future stock prices using machine learning techniques on the NSE. It uses linear regression and SVM regression. Linear regression will be used for predicting open price of the stock for the next day using close price of the stock for the previous day. SVM regression will be used for predicting the difference between close and open prices of the stock for the next day. External factors like foreign exchange rate, NSE index, moving averages, relative Strength index etc are used to get maximum accuracy. Exchange rates, moving averages, etc., are taken into consideration for comparison.
 - **Conclusion/Results:** The model can predict prices maximum for the coming five days. The model proved able to handle more than 50 stocks.
 - **Link:** [IRJET-V5I3634.pdf](https://www.researchgate.net/publication/328136344)
 5. M. Qiu, Y. Song, Predicting the direction of stock market index movement using an optimized artificial neural network model, PLoS One 11 (5) (2016) e0155133
 - **Model:** ANN algorithm. Two types of input variables were chosen so that the results could be comparable
 - **Dataset:** Predicting Nikkei 225 Index of the Japanese stock market with the help of ANN
 - **Methodology:** The model was tweaked with the help of G.A. to make it less convergent and more precise.
 - **Conclusion/Results:** The model proved to have an 81.27% accuracy rate. This model was further compared with other models and proved to be the best for prediction.
 - **Link:** <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0155133>
 6. Chih-Fong Tsai Ming-Lun Chen, Credit rating by hybrid machine learning techniques
 - **Model:** ANN + Decision trees (D.T.) to forecast stock price movements.
 - **Dataset:** The stock prices of the electron industry in Taiwan. All the indexes (fundamental, technical, and macroeconomic) were collected from TEJ.
 - **Methodology:** One hidden layer was used in the ANN. 1 and 0 could be the answers to the decision model (1 being prices rise, 0 being prices fall). Then ANN and D.T. are combined. Also, a DT+DT model is created for comparison.
 - **Conclusion/Results:** The ANN + Decision Tree model had a total prediction accuracy of 77%. Only DT gave 65% accuracy, and only ANN gave 59% accuracy. DT+DT gave an average of 67%.
 - **Link:** <https://www.sciencedirect.com/science/article/abs/pii/S1568494609001215>
 7. Stock Price Direction Prediction Using Artificial Neural Network Approach: The Case of Turkey
 - **Model:** Eight ANN models were prepared for seven different Prediction Systems

- **Dataset:** Istanbul Stock Exchange (ISE-30). Daily closing prices of each stock were collected and employed to calculate indicators of P.S. algorithms.
 - **Methodology:** Stock prices- - Decrease when output is greater than or equal to 0 and less than 0.5 - Stay same when output is equal to 0.5 - Increase when output is greater than 0.5 and equal to or less than 0 -Stopping criteria = 10 000 - Activation Function = Linear Sigmoid -Learning Rate = 0.2.
 - **Conclusion/Results:** 78.47% was the success rate averaged out, 50% was the minimum success rate for each stock, showing high predicting capability. The best ANN topology was ANNM3.11.1 (with three inputs, 11 hidden neurons, and one output).
 - **Link:** [Stock Price Direction Prediction Using Artificial Neural Network Approach: The Case of Turkey \(scialert.net\)](#)
8. Reza Aghababaeyan, TamannaSiddiqui, NajeebAhmadKhan, Forecasting the Tehran Stock Market by Artificial Neural Network
- **Model:**
 - **Dataset:** Mobarakeh-Steel Co. tries from the Tehran Stock exchange. Data was used from Mar 15, 2007, to Feb 14, 2011.
 - **Methodology:** The model had three layers, trained to use fast backpropagation algorithm. Ten neurons are comprised of the hidden layer.
 - **Conclusion/Results:** The results were that the algorithm formed was 97% accurate. But overall, the algorithm was marked to be 83% correct during any new news released regarding the particular company.
 - **Link:** [Forecasting the Tehran Stock Market by Artificial Neural Network \(thesai.org\)](#)
9. DA. Puspitasari, Z. Rustam, Application of SVM-KNN using SVR As Feature Selection on Stock Analysis for Indonesia Stock Exchange, Vol. 2023, 2018, 20205.
- **Model:** SVM and K Nearest Neighbor (KNN)
 - **Dataset:** Indonesian Stock Prices from the company P.T. Waskita Karya (Persero) Tbk. Data used was from January, 2013 to December, 2016
 - **Methodology:** First, the class labels were predicted. Then, with the help of SVM, indicators were chosen to help anticipate stock price movement and predict the future trend of the stock.
 - **Conclusion/Results:** The research ends with three indicators representing results that show the good capacity to predict prices.
 - **Link:** <http://dx.doi.org/10.1063/1.5064203>
10. IK. Nti, AF. Adekoya, BA. Weyori, Efficient stock-market prediction using ensemble support vector machine, Open Comput. Sci. 10 (1) (2020) 153–163,
- **Model:** A homogeneous ensemble classifier known as GASVM, which is based on the G.A. This helps select and optimize SVM and its factors.
 - **Dataset:** In the study, 10-day price movements were predicted from the Ghana Stock Exchange (GSE). Data was used from Jun 25, 2007, to Aug 27, 2019.
 - **Methodology:** Other models like RMSE, MAE, AUC, Accuracy, Recall were used to compare their efficiency
 - **Conclusion/Results:** In the end, GASVM showed the most promising results and predicted the prices of stocks from the GSE more accurately. The model developed provided an accuracy of 93.7%.

- **Link:** <http://dx.doi.org/10.1515/comp-2020-0199>
11. J. John, A. Kumar, A. Abhishek, TA. Dhule4 , A. Roy, A. Jha, Stock market prediction using machine learning
- **Model:** SVM model is implemented in Python Programming Language.
 - **Dataset:** Dataset is obtained from Web Scrapping to train the algorithm Data has been scrapped from Yahoo Finance. Apple. Inc's data has been used from Jan 1, 2013, to Dec 30, 2019.
 - **Methodology:** Pandas Datereader is imported to the header of Python Code to train the dataset. Pandas for Data manipulation and analysis, NumPy for core scientific computations, sklearn to import SVM and Matplotlib for 2-D plots f array.
 - **Conclusion/Results:** SVM proved to be the best choice to carry out the experiment as it can handle a large pool of data and trains faster than many other algorithms. SVM also improves results with more training.
 - **Link:** [IJRAR2001262.pdf](#)
12. AF. Sheta, SEM. Ahmed, H. Faris, A comparison between regression, artificial neural networks and support vector machines for predicting stock market index, Int. J. Adv. Res. Artif. Intell. 4 (7) (2015)
- **Model:** Multiple Linear regression (MLR), SVM with RBF Kernel.
 - **Dataset:** SVM model predicts the S and P 500 stock Index. The Data contains 27 features and 1192 days from Dec 7, 2009, to Sept 2, 2014.
 - **Methodology:** The Data was sampled weekly. The samples were divided into 100 Samples as training sets and 43 as testing sets.
 - **Conclusion/Results:** SVM outperforms MLP and MLR models in training as well as testing. SVM also has the advantage of using multiple kernels, which makes the model flexible.
 - **Link:** [F:/Unix/Papers 2001-2015/papers2015/IJARAI2015/Final version/IJARAI2015\(143\).dvi \(thesai.org\)](#)
13. Yancong Xie Hongxun Jiang, Stock Market Forecasting Based on Text Mining Technology: A Support Vector Machine Method
- **Model:** SVM and Support Vector Classifier (SVC)
 - **Dataset:** Twenty stocks were selected from the Chinese Stocks as testing data. Stocks were chosen based on trade volumes, stock sectors, publishing date, and price.
 - **Methodology:** Text mining Technology was used on SVM. The same was done with SVC so that both models can be compared on a common basis. News heavily affected the results as the articles were also taken into consideration.
 - **Conclusion/Results:** SVM's forecasting proved to be superior to B.P. due to reduced NMSE and MAE and increased D.S., CP, and CD than B.P. SVM even trains faster than B.P.
 - **Link:** [http://refhub.elsevier.com/S2772-6622\(21\)00010-2/sb60](http://refhub.elsevier.com/S2772-6622(21)00010-2/sb60)
14. Y. Hao, Q. Gao, Predicting the trend of stock market index using the hybrid neural network based on multiple time scale feature learning, Appl. Sci. 10 (11) (2020)

- **Model:** LSTM, other models like SVM, CNN, NFNN, and Multiple Pipeline models were used to compare results
 - **Dataset:** S and P 500 stock prices were taken as prediction data. Data was taken for 20 years, from Jan 30, 1999, to Jan 30, 2019.
 - **Methodology:** Multiple LSTMs were taken to learn the time dependencies of features of different time scales. All the information was combined to predict the closing price in the future.
 - **Conclusion/Results:** The LSTM model proved to be better than other models due to its impressive capability of efficiently processing the data. The model had an accuracy of 74.55% over one month.
 - **Link:** [http://refhub.elsevier.com/S2772-6622\(21\)00010-2/sb78](http://refhub.elsevier.com/S2772-6622(21)00010-2/sb78)
15. P. Gao, R. Zhang, X. Yang, The application of stock index price prediction with neural network, *Math. Comput. Appl.* 25 (3) (2020).
- **Model:** MLP, LSTM, CNN and Uncertainty-Aware Attention (U.A.).
 - **Dataset:** CSI300 from China, S and P 500, Nikkei225 of Tokyo. All the stock data is selected from July, 2008 to September 2016. 90% of the Data was used as training data, while 10% was used as test data.
 - **Methodology:** The LSTM model had 140 hidden layers. Each model was predicted using MAPE, RMSE, and R.
 - **Conclusion/Results:** LSTM model performed well in predicting CSI300 but overall had a higher MAPE, RMSE, R. U.A. performed better than all the others, MLP performed the worst.
 - **Link:** [http://refhub.elsevier.com/S2772-6622\(21\)00010-2/sb78](http://refhub.elsevier.com/S2772-6622(21)00010-2/sb78)
16. G. Ding, L. Qin, Study on the prediction of stock price based on the associated network model of LSTM, *Int. J. Mach. Learn. Cybern.* 11 (6) (2019) 1307–1317
- **Model:** LSTM, LSTM-based deep recurrent neural network (DRNN), Associated Neural Network (Associated Net).
 - **Dataset:** Shanghai Composite Index, PetroChina, ZTE with 6112, 2688, and 4930 historical Data, respectively.
 - **Methodology:** MSE and MAE are used for comparison.
 - **Conclusion/Results:** LSTM model did not perform well, and it had too many deviations from the actual data. DRNN and Associated Net provided almost identical results with very little to no deviation.
 - **Link:** <http://link.springer.com/article/10.1007/s13042-019-01041-1>
17. JMT. Wu, Z. Li, N. Herencsar, B. Vo, JCW. Lin, A graph-based CNN-LSTM stock price prediction algorithm with leading indicators, *Multimedia Syst.* (2021) 3
- **Model:** LSTM, CNN, Stock Sequence Array Convolution LSTM (SACLSTM)
 - **Dataset:** AAPL, IBM, MSFT, F.B., AMZN from American stock exchange and CDA, CFO, DJO, DVO, IJO from Taiwan stock exchange. Data is from October 2018 to October 2019
 - **Methodology:** 60% of data was employed to train the model, 20% was used to test the model, and the other 20% was for verification purposes. TensorFlow was the model which gave effect to CNN and LSTM
 - **Conclusion/Results:** SACLSTM performed the best with minimum error and definitive price movement graphs. CNN and LSTM did not show good results.
 - **Link:** <http://dx.doi.org/10.1007/s00530-021-00758-w>

18. W. Budiharto, Data science approach to stock prices forecasting in Indonesia during Covid-19 using long short-term memory (LSTM), J. Big Data 8 (1) (2021)

- **Model:** R language-based LSTM model
- **Dataset:** P.T. Bank Central Asia Tbk and P.T. Bank Mandiri. 80% of data was used for training and the remaining 20% for testing
- **Methodology:** Results of the model were calculated based on different periods and varying numbers of epochs.
- **Conclusion/Results:** The most promising results were obtained using a period of 1 year and a total of 100 epochs. The accuracy summed up to 94.59%.
- **Link:** <http://dx.doi.org/10.1186/s40537-021-00430-0>

19. R. Nandakumar, R. UK, Y.V. Lokeswari, Stock price prediction using long short term memory, Int. Res. J. Eng. Technol. (2018)

- **Model:** Comparison between ANN and LSTM. To compare results, RMSE was used.
- **Dataset:** Dixon Hughes, Cooper Tire and Rubber, PNC financial, CitiGroup, Alcoa Corp.
- **Methodology:** The LSTM model had 1 Input layer having five neurons, 'n' hidden layers, and one output layer.
- **Conclusion/Results:** LSTM had a much better prediction accuracy. For every company, RMSE was lower when predicted through LSTM and slightly higher through ANN.
- **Link:** [IRJET-V5I3788.pdf](https://www.researchgate.net/publication/343212879_Financial_Literacy_and_Financial_Risk_Tolerance_of_Individual_Investors_Multinomial_Logistic_Regression_Approach)

20. Financial Literacy and Financial Risk Tolerance of Individual Investors: Multinomial Logistic Regression Approach by Yılmaz Bayar, H. Funda Sezgin, Ömer Faruk Öztürk and Mahmut Ünsal Şaşmaz.

- **Model:** Multinomial logistic regression to analyse the impact of financial literacy level.
- **Dataset:** Income Data for USAK University's Staff
- **Methodology:** Here, they have used multinomial logistic regression to analyse the impact of financial literacy level and demographic characteristics on individual risk tolerance, using data from individuals. Hypotheses suggest relationships between financial literacy level, demographic characteristics, and risk tolerance, with financial literacy expected to influence risk tolerance, and demographic factors such as age, education level, and income expected to have varying effects on risk tolerance.
- **Conclusion/Results:** The study highlights the positive influence of financial literacy and education on financial risk tolerance, alongside demographic factors such as gender, age, and income. Findings underscore the importance of targeted financial education programs, especially for young individuals and women, to enhance risk tolerance and contribute to the development of financial sector.
- **Link:** https://www.researchgate.net/publication/343212879_Financial_Literacy_and_Financial_Risk_Tolerance_of_Individual_Investors_Multinomial_Logistic_Regression_Approach

21. Stock prediction and mutual fund portfolio management using curve fitting techniques by Giridhar Maji, Debomita Mondal, Nilanjan Dey, Narayan C. Debnath and Soumya Sen.

- **Model:** Curve fitting/regression
- **Dataset:** Mutual funds Indian markets
- **Methodology:** The research utilizes regression analysis to predict individual company share prices, considering recent data with higher weightage. The experiment utilized historical closing stock prices of 20 companies from August 2004 to December 2015 for training and data from January 2016 to December 2016 for validation. Regression analysis was performed on the historical prices to select the best-fitting curve for each company. Funds were then allocated across industrial sectors based on stock performance, followed by allocation to individual companies within each sector.
- **Conclusion/Results:** A total capital of Rs. 1,000,000 was invested to evaluate returns over 30 months. Results showed that the proposed methodology outperformed top-performing mutual funds, demonstrating good capital appreciation and return on investment despite market fluctuations, following a buy-and-hold strategy.
- **Link:** <https://link.springer.com/article/10.1007/s12652-020-02693-6>

22. Fusion in stock market prediction: A decade survey on the necessity, recent developments, and potential future directions by Ankit Thakkar, Kinjal Chaudhari.

- **Model:** Information, feature, and model fusion
- **Dataset:** Indian stock exchange
- **Methodology:** Paper reviews fusion techniques in stock market prediction, spanning information, feature, and model fusion, showcasing their applications across stock price prediction, risk analysis, index forecasting, and more. It emphasizes the significance of integrating diverse data sources for enhancing prediction accuracy while highlighting challenges in feature selection and model suitability.
- **Conclusion/Results:** The paper explores the significance of fusion techniques in stock market prediction from 2011 to 2020, covering information fusion, feature fusion, and model fusion. It assesses fusion's impact on diverse aspects such as stock price prediction, index forecasting, and risk/return analysis.
- **Link:** <https://www.sciencedirect.com/science/article/pii/S1566253520303481>

Literature Review Comparison Table

Paper	Model	Dataset	Method	Results
1	Decision Tree, Naive Bayes, ANN, RNN, LSTM	Tehran Stock Exchange	Decision Trees. ANN and RNN (RNN and LSTM) algorithms were created and trained.	LSTM and RNN: ~90% Naive Bayes and Decision Tree: ~85%
2	Deep neural networks (DNNs)	S&P 500	By controlling the overfitting and observing DNNs as the number of the hidden layers increased gradually from 12 to 1000 for different no. of PC's.	DNN with PCs = 60 hidden layers = 16: 58.9% DNN with PCs = 31 hidden layers = 16: 60.1%
3	MLP and LSTM	Apple stock (10 days)	Data is divided by 200, maintaining the weights. AdaGrad and RMSProp are used to maintain the performance.	The stock prices of Apple were forecasted and almost matched using LSTM compared to MLP.
4	Linear Regression and SVM	NSE Nifty-50, CNX, S&P 500	Linear regression was used for predicting open price of the stock and SVM regression for predicting the difference between close and open prices.	Model can predict prices maximum 5 days. It can handle more than 50 stocks.

5	ANN	Nikkei 225 (Japan)	Adjusted the weights and biases of the ANN model using the GA algorithm and evaluated on the input data.	Model accuracy: ~81.2%
6	ANN + Decision trees	Taiwan Stock Exchange	Hidden layer was used in the ANN. 1 and 0 could be the answers to the decision model along with combinations.	ANN + DT: ~77% DT: ~65% ANN: ~59% DT+DT: ~67%
7	Different ANN models	Istanbul Stock Exchange (ISE-30)	Eight ANN models were prepared for seven different Prediction Systems (P.S.)	Best Accuracy: ~78.47% Avg. Accuracy: ~50%
8	Linear Regression and FFNN with back-propagation.	Tehran Stock exchange	The model had three layers, trained to use fast backpropagation algorithm	Accuracy: ~83%
9	SVM and KNN	Indonesian Stock Prices	With the help of SVM, indicators were chosen to help anticipate stock price movement and price	Accuracy: ~91%
10	GASVM	Ghana Stock Exchange	GASVM based on SVM enhanced with GA for feature-selection and kernel optimisation for prediction.	Accuracy: ~93.7%
11	SVM	Apple stock (6 years)	Paper proposed an application of ML and used SVM to predict upcoming Stock Prices	SVM proved to be the best choice to carry out the experiment
12	MLR, SVM with RBF Kernel	S & P 500	The samples were divided into 100 Samples(weely) as training sets and 43 as testing sets.	SVM outperforms MLP and MLR models in training as well as testing
13	SVM & SVC	Chinese Stock Exchange	Text mining was used on both SVM and SVC so that both models can be compared on a common basis.	SVM's forecasting proved to be superior
14	LSTM, SVM, CNN, NFNN, Pipeline models	S&P 500	Multiple LSTMs were taken to learn the time dependencies of features of different time scales to predict the closing price in the future.	LSTM: ~74.55%
15	MLP, LSTM, CNN and Uncertainty-Aware Attention (U.A.)	CSI300 from China, S and P 500, Nikkei225 of Tokyo	The LSTM model had 140 hidden layers. Each model was predicted using MAPE, RMSE, and R.	U.A. performed better than all the others
16	LSTM, LSTM-based deep recurrent neural network (DRNN),	Shanghai Composite Index	MSE and MAE are used for comparison	DRNN and Associated Net provided almost identical results with very little to no deviation
17	LSTM, CNN, Stock Sequence Array Convolution LSTM (SACLSTM)	American and Taiwan stock exchange	TensorFlow was the model which gave effect to CNN and LSTM	SACLSTM performed the best with minimum error and definitive price movement graphs
18	LSTM	P.T. Bank Central Asia & Mandiri	Results of the model were calculated based on different periods and varying numbers of epochs.	Accuracy: ~94.59%
19	ANN and LSTM	Indian Stock Exchange	The LSTM model had 1 Input layer having five neurons, 'n' hidden layers, and one output layer	LSTM had a much better prediction accuracy
20	Multinomial logistic regression (MLR)	USAK University's Staff's income data	They have used MLR to analyse the impact of gender, age, and income on individual risk tolerance.	The study highlights the positive influence of gender, age, and income on financial risk tolerance level.
21	Curve fitting/regression	Mutual funds Indian markets	The research utilizes regression analysis to predict individual company share prices, considering recent data with higher weightage.	Results showed that the proposed methodology outperformed top-performing mutual funds.
22	Information, feature, and model fusion	Indian stock exchange	Paper reviews fusion techniques in stock market prediction, spanning information, feature, and model fusion.	Assesses fusion's impact on diverse aspects such as stock price prediction, index forecasting, and risk/return analysis.

Data Collection:

The data is mainly divided into three parts, the income data, equity data and mutual funds data. The income and mutual funds were directly downloaded from Kaggle the link for the same is mentioned below:

Income Data: <https://www.kaggle.com/datasets/vishwas199728/credit-card>

Mutual Funds Data: <https://www.kaggle.com/datasets/ravibarnawal/mutual-funds-india-detailed>

The equity data was downloaded from the official website of Bombay Stock Exchange (BSE). All the data on BSE is open source and is available to the public on demand through its website. The data obtained for this research is of daily data for the top 100 large cap stocks, top 100 mid cap stocks and top 100 small cap stocks from 1st January 2014 to 30th April 2024. By searching equity names under the Security Name tab on the website and providing the timeframe the data was downloaded under csv format. Simultaneously, I have also maintained a metadata for all the equity data consisting of the equity names, sector/industry and its ticker symbol in BSE and NSE.

Equity Data: <https://www.bseindia.com/markets/equity/EQReports/StockPrcHistori.html?flag=0>

Data Description:

1. **Income Data:** The income data consists of individuals Income and dependency data. It covers a wide variety of Occupations and Industries. It overall gives an outlook of the investor population of India.

Columns:

- Ind_ID: Client ID
- Gender: Gender information
- Car_owner: Having car or not
- Propert_owner: Having property or not
- Children: Count of children
- Annual_income: Annual income
- Type_Income: Income type
- Education: Education level
- Marital_status: Marital_status
- Housing_type: Living style
- Birthday_count: Use backward count from current day (0), -1 means yesterday.
- Employed_days: Start date of employment. Use backward count from current day (0). Positive value means, individual is currently unemployed.
- Mobile_phone: Any mobile phone
- Work_phone: Any work phone
- Phone: Any phone number
- EMAIL_ID: Any email ID
- Type_Occupation: Occupation
- Family_Members: Family size

2. **Mutual Funds Data:** Mutual funds data consists of around 800 different types of mutual funds data with their detailed information about their performance over the years, type and industry in which they operate. It gives an overall aspect of how the fund is performing in the market.

Columns:

- Scheme Name: Name of the mutual fund scheme
- Min sip: Min sip amount required to start.
- Min lumpsum: Min lumpsum amount required to start.
- Expense ratio: calculated as a percentage of the Scheme's average Net Asset Value (NAV).

- Fund size: the total amount of money that a mutual fund manager must oversee and invest.
 - Fund age: years since inception of scheme
 - Fund manager: A fund manager is responsible for implementing a fund's investment strategy and managing its trading activities.
 - Sortino : Sortino ratio measures the risk-adjusted return of an investment asset, portfolio, or strategy
 - Alpha: Alpha is the excess returns relative to market benchmark for a given amount of risk taken by the scheme
 - Standard deviation: A standard deviation is a number that can be used to show how much the returns of a mutual fund scheme are likely to deviate from its average annual returns.
 - Beta: Beta in a mutual fund is often used to convey the fund's volatility (gains or losses) in relation to its respective benchmark index
 - Sharpe: Sharpe Ratio of a mutual fund reveals its potential risk-adjusted returns
 - Risk level:
 - 1- Low risk
 - 2- Low to moderate
 - 3- Moderate
 - 4- Moderately High
 - 5- High
 - 6- Very High
 - AMC name: Mutual fund house managing the assets.
 - Rating: 0-5 rating assigned to scheme
 - Category: The category to which the mutual fund belongs (e.g. equity, debt, hybrid)
 - Sub-category : It includes category like Small cap, Large cap, ELSS, etc.
 - Return_1yr (%): The return percentage of the mutual fund scheme over 1 year.
 - Return_3yr (%): The return percentage of the mutual fund scheme over 3 year.
 - Return_5yr (%): The return percentage of the mutual fund scheme over 5year.
- 3. Equity Data:** This data consists of details of individual equity's day wise open and close price from 2014-2024. It also includes the volume and weighted average volume (WAP) for that day.

Columns:

- Date: Date
- Open: Open price on that particular date.
- High: High price on that particular date.
- Low: Low price on that particular date.
- Close: Close price on that particular date.
- WAP: Weighted Average Price on that particular date.
- Volume: Volume of shares traded on that particular date.
- No. of Trades: Total trades made on that day.
- Total Turnover (Rs.): Total market cap of that equity on that day.
- Deliverable Quantity: Total deliverable quantity on that day.
- % Deli. Qty to Traded Qty: Ratio of deliverable quantity to traded quantity.
- Spread High-Low: Difference of High and Low price for that day.

- Spread Close-Open: Difference of Close and Open price for that day.
4. **Equity Metadata:** This file consists of the metadata of all the equities available in our data corpus with its sector/industry in which the company operates.

Columns:

- Company name: Equity company name.
- Industry: Equity company sector of operations.
- BSE: BSE ticker symbol
- Symbol: NSE ticker symbol
- Market Cap: Market Cap of the equity(Large, Mid, Small)

Data Cleaning:

1. Income Data:

- Dropped all the rows having “nan” as annual income as “nan” values in income column would result in difficulty for getting a proper financial risk level for that individual.
- All the “nan” values in birthday_count column was handled by interpolate linear method to get age for individual.
- Created a new “AGE” column and stored the age of all the individuals by dividing the birthday_count by 365.25.
- All “nan” values in “Gender” column were replaced by random gender between “M” and “F”.
- Dropped columns: 'Mobile_phone', 'Work_Phone', 'Phone' and 'EMAIL_ID' which won't be required for calculating the Financial Risk Tolerance of an individual.
- Created a new column “Employed” which takes binary values and stores the employment status of the individual based on the “Employed_days” column.
- Multiplied the Annual income of all the individuals by 3.75 (the rate of Price Parity Ratio) between US and India.
- Updated all the column names to standard column names: 'Ind_ID', 'Gender', 'Car_Owner', 'Property_Owner', 'Children', 'Annual_income', 'Income_source', 'Education', 'Marital_status', '_type', 'Birthday_count', 'Employed_days', 'Occupation_type', 'Family_Members', 'Age', 'Employed'.
- Saved the updated file to another folder.

jupyter Exploratory_Data_Analysis_IncomeData Last Checkpoint: 05/15/2024 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

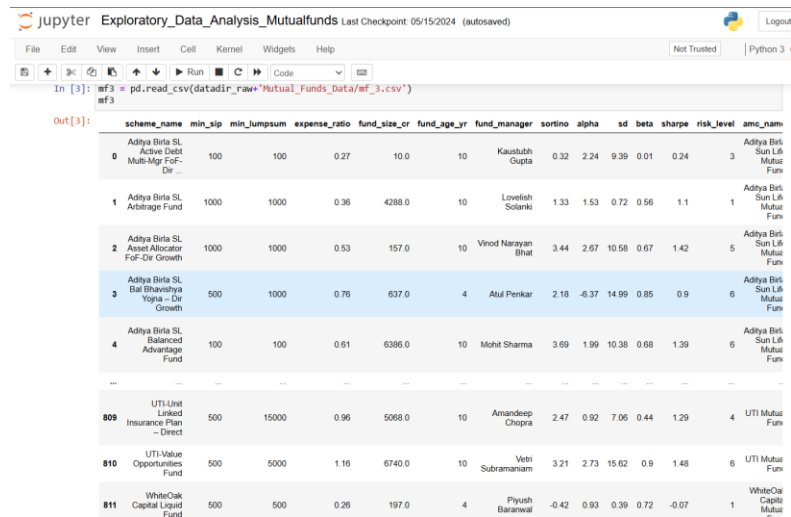
```
In [3]: income = pd.read_csv(datadir_raw+'Income_Data/Income_data.csv')
income
```

Out[3]:

	Ind_ID	GENDER	Car_Owner	Property_Owner	CHILDREN	Annual_income	Type_Income	EDUCATION	Marital_status	Housing_type	Birthday_count	Em
0	5008827	M	Y	Y	0	180000.0	Pensioner	Higher education	Married	House / apartment	-18772.0	
1	5009744	F	Y	N	0	315000.0	Commercial associate	Higher education	Married	House / apartment	-13557.0	
2	5009746	F	Y	N	0	315000.0	Commercial associate	Higher education	Married	House / apartment	NaN	
3	5009749	F	Y	N	0	NaN	Commercial associate	Higher education	Married	House / apartment	-13557.0	
4	5009752	F	Y	N	0	315000.0	Commercial associate	Higher education	Married	House / apartment	-13557.0	
...
1543	5028645	F	N	Y	0	NaN	Commercial associate	Higher education	Married	House / apartment	-11957.0	
1544	5028655	F	N	N	0	225000.0	Commercial associate	Incomplete higher	Single / not married	House / apartment	-10229.0	
1545	5115992	M	Y	Y	2	180000.0	Working	Higher education	Married	House / apartment	-13174.0	
1546	5118219	M	Y	N	0	270000.0	Working	Secondary / secondary special	Civil marriage	House / apartment	-15262.0	
1547	5053790	F	Y	Y	0	225000.0	Working	Higher education	Married	House / apartment	-19601.0	

1548 rows x 18 columns

2. Mutual Funds Data: The mutual funds data does not require any additional data preprocessing or cleaning as the data from Kaggle is already cleaned and pre-processed and in the expected format for the research.

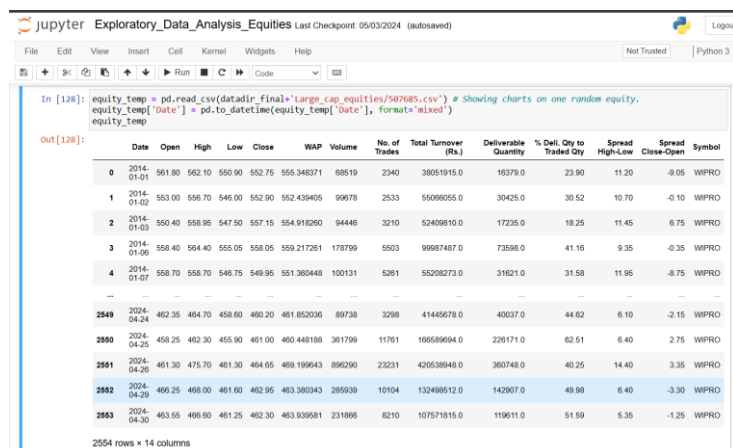


The screenshot shows a Jupyter Notebook titled 'Exploratory_Data_Analysis_Mutualfunds'. The code cell contains the following Python code:

```
In [3]: mf3 = pd.read_csv(datadir_raw+'Mutual_Funds_Data/mf_3.csv')
mf3
```

The output is a DataFrame with 112 rows and 14 columns. The columns are: scheme_name, min_sip, min_lumpsum, expense_ratio, fund_size_cr, fund_age_yr, fund_manager, sortino, alpha, sd, beta, sharpe, risk_level, and amc_nam. The data includes various mutual fund schemes such as Aditya Birla SL Active Debt Multi-Sip FoF-Direct, Aditya Birla SL Arbitrage Fund, Aditya Birla SL Asset Allocator FoF-Direct Growth, Aditya Birla SL Bai Bhavishya Yojna - Dir Growth, Aditya Birla SL Balanced Advantage Fund, UTI-Unit Linked Insurance Plan - Direct, UTI Value Opportunities Fund, and WhiteOak Capital Liquid Fund.

3. Equity Data: All the equity files are stored by their BSE ticker symbol name in csv format. So, for cleaning the data I looped through all the BSE ticker symbols in the metadata file based on the equities market cap and in each iteration performed the following steps:
 - Converted the Date column to Datetime column for easy processing.
 - Got the NSE symbol name from the metadata file and created a new column "Symbol" with its NSE symbol name.
 - Sorted the whole file by Date in Ascending Order i.e. 01/01/2014 – 30/04/2024.
 - Changed few column names for easy recognition for different packages. Updated column names: 'Date', 'Open', 'High', 'Low', 'Close', 'WAP', 'Volume', 'No. of Trades', 'Total Turnover (Rs.)', 'Deliverable Quantity', '% Del. Qty to Traded Qty', 'Spread High-Low', 'Spread Close-Open', 'Symbol'.
 - Making the date column as index for the file for handling missing values.
 - Handling all missing values using interpolates time method, adjusting the nan values based on the previous and forthcoming data.
 - Resetting the index and saving the file in new folder.



The screenshot shows a Jupyter Notebook titled 'Exploratory_Data_Analysis_Equities'. The code cell contains the following Python code:

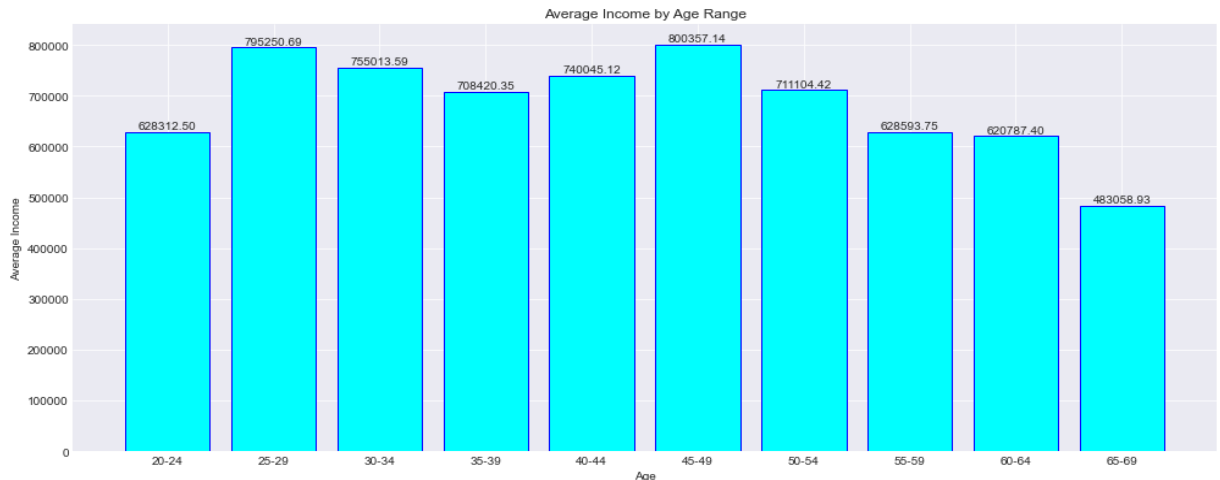
```
In [128]: equity_temp = pd.read_csv(datadir_Final+'Large_cap_equities/507685.csv') # Showing charts on one random equity.
equity_temp['Date'] = pd.to_datetime(equity_temp['date'], format='%m/%d/%Y')
equity_temp
```

The output is a DataFrame with 2554 rows and 14 columns. The columns are: Date, Open, High, Low, Close, WAP, Volume, No. of Trades, Total Turnover (Rs.), Deliverable Quantity, % Del. Qty to Traded Qty, Spread High-Low, Spread Close-Open, and Symbol. The data includes equity information for various companies, with the Symbol column showing 'WIPRO' for all entries.

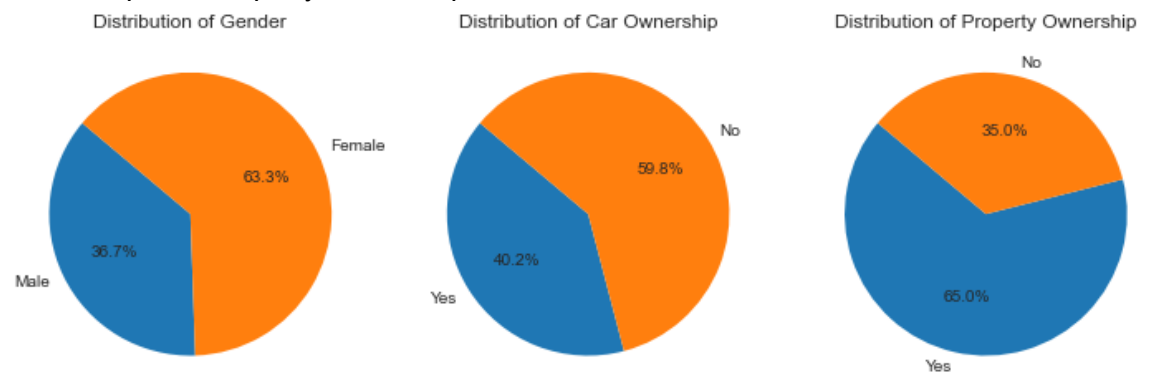
Exploratory Data Analysis:

1. Income data:

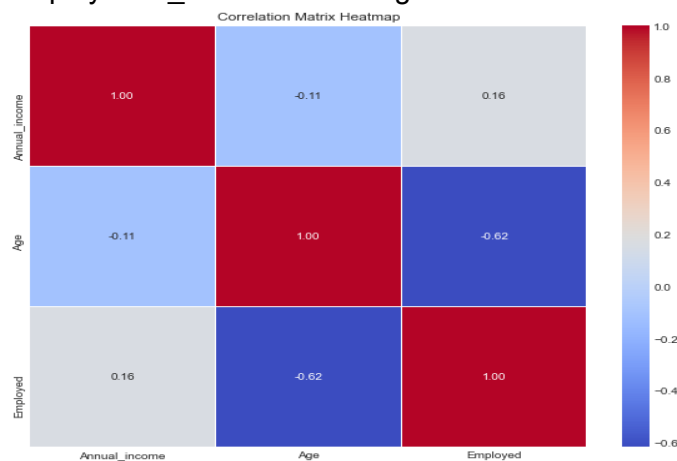
- The chart below shows the Average Income of individuals based on their Age range.



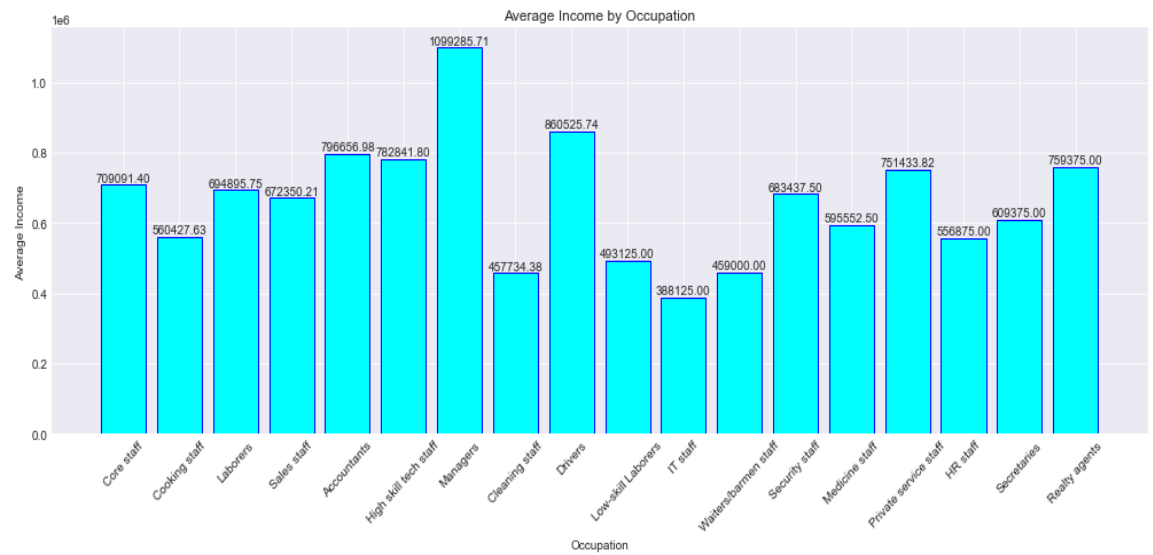
- The Pie chart below shows the proportion of multiple factors like Gender, Car Ownership and Property Ownership in the dataset:



- The heatmap below shows the correlation between Annual Income, Employment_status and the Age of individuals

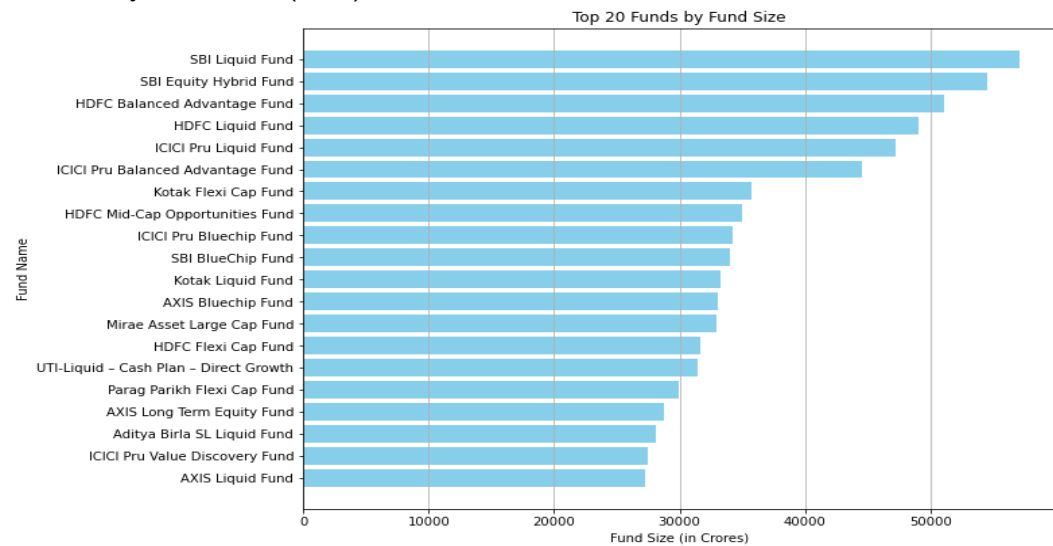


- The bar chart below shows the average income by profession:

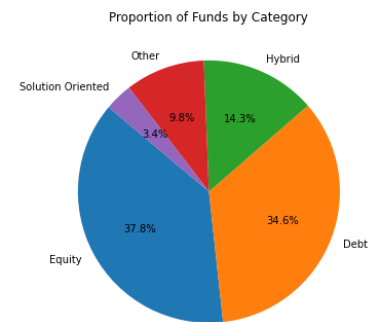
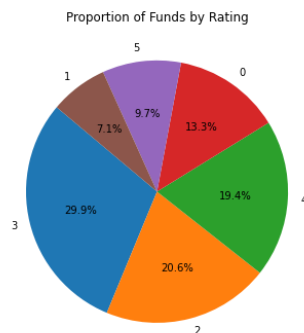
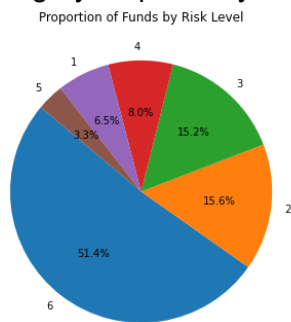


2. Mutual funds data:

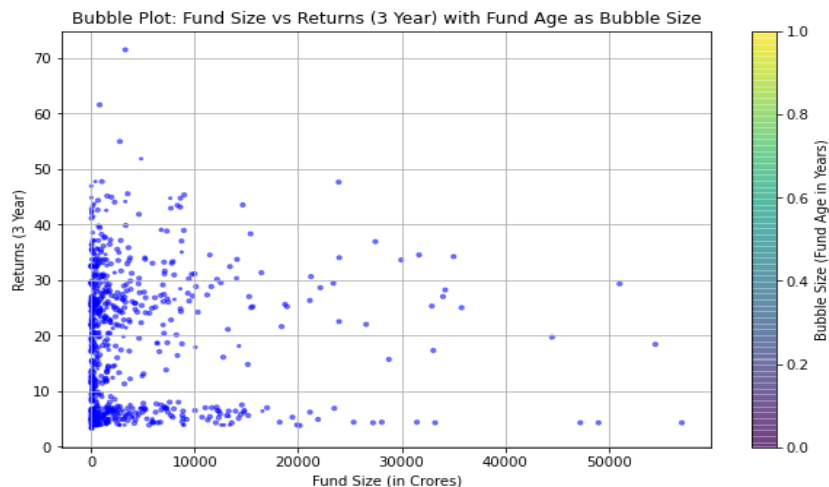
- The horizontal bar chart below shows the top 20 mutual funds in the whole dataset by Fund size (in cr)



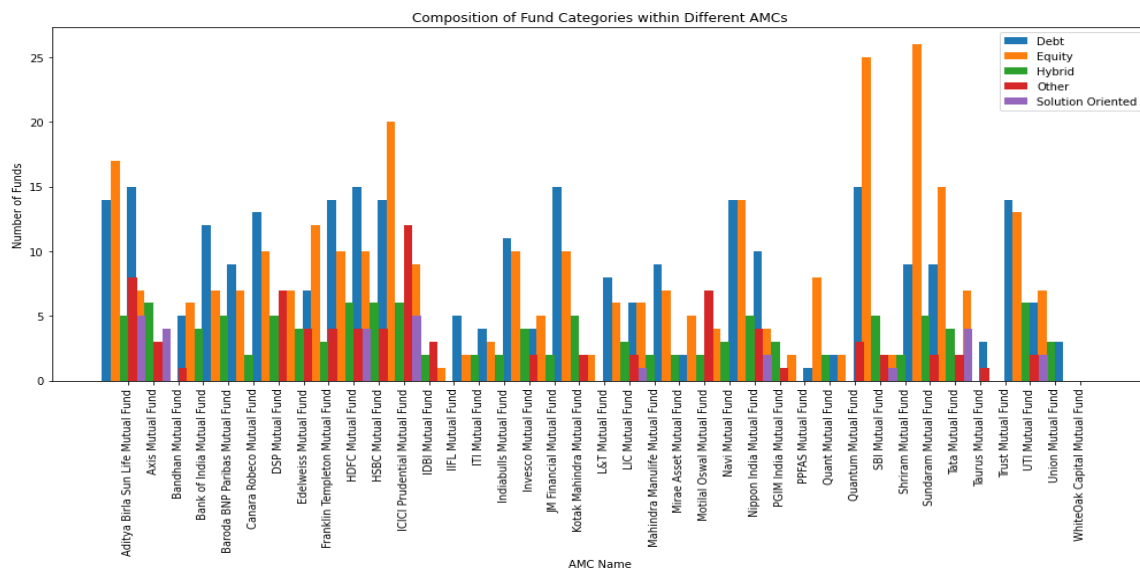
- The pie charts below show the proportion of Risk level, Fund Rating and Fund Category respectively:



- The bubble plot shows the Fund size vs the 3 year returns which the fund has provided with the bubble size as the Fund Age:



- The multi-bar chart shows the count of different types of mutual funds an AMC owns:



3. Equity data:

- The charts below show the candle stick chart with volume bars from 01-01-2024 to 30-04-2024 for different fund size equities namely Wipro Ltd (Large cap), Ashok Leyland (Mid Cap) and VIP Industries (Small Cap) respectively.

