Table of Contents:

Executive Summary

Objective

The objective of this analysis was to decipher mutual fund performance, focusing on factors such as portfolio composition, risk-return profiles, and expense ratios. Through comprehensive data exploration and modeling, the aim was to provide valuable insights to guide intelligent decision-making in investment portfolios.

Key Findings

- Diverse Initial Investments: Initial investments spanned from $10 to $5 Billion, with common initial investments clustered around $1,000, $2,500, and $1 Million.

- *Commencement and Maturity Trends*: Notable trends in fund commencement and maturity, with 2015 being a significant year, witnessing 505 funds maturing. Average yield trends fluctuated, reaching the lowest at 0.03% in 1968 and the highest at 22.35% in 2015.

- *ROI and Volatility Trends*: Return on investment (ROI) exhibited volatility in earlier years, with recent times showcasing reduced fluctuations. Notable ROI variations, with 1974 experiencing the highest at 25.27% and the lowest at -0.44%.
- *Morningstar Ratings and Investment Trends*: Morningstar ratings ranged from historical highs in 1975 to a significant drop in 2020. Medium-sized funds held the highest rating within and outside their fund type. Investment amounts varied, with 1994 having the lowest total investment and 2020 holding the largest sum.

- *Avg. Fund Yield by Investment Type and Size*: Yield variations were observed within different fund types and sizes. Small-sized Growth funds had the least average yield, while medium-sized Growth funds peaked overall yield.

- *Sectoral Asset Allocation by Fund Size*: Diverse sectoral allocations were witnessed based on fund size. Financial Services consistently held a significant share of allocated funds across all sizes.

- *Fund Classification - Top and Bottom Performers*: Top performers included funds from the American Beacon and Cohen & Steers families, with AGEIX leading. Bottom performers included FFRLX (Virtus) and FRICX (William Blair), both small-sized funds.

- *Safe Zone* - Historical Returns: ETAHX (Small-sized) emerged as the top performer with the highest overall returns in 3, 5, and 10 years. GHYMX (Small-sized) and ETCTX (Medium-sized) followed with strong overall returns.

Potential Implications of The Key Findings

- *Risk Management*: Assess the implications of investment constraints on fund performance, especially in the context of small-sized funds in the bottom performers.

- *Diversification Strategy*: Utilize insights for informed sectoral allocation to optimize fund performance and consider the performance of these funds for potential inclusion in investment portfolios.

- *Portfolio Optimization*: Use information for informed investment decisions and potential portfolio optimization by considering the performance of these funds for inclusion in investment portfolios.

Introduction

Project Background

The project, titled "FundWise: Deciphering Mutual Fund Performance and Behavior," was initiated with the objective of enhancing portfolio performance when investing in mutual funds. Recognizing the need for analytical insights that guide intelligent decision-making, the project focuses on providing a comprehensive understanding of mutual fund performance through in-depth analysis.

Project Proposal Overview

The project proposal was presented on 20th October 2023 by the following group members:

- Pranav Harish Sharma (002851959)
- Harvineet Singh (002814713)
- Oluwatobi Alao (002645627)

Business Problem

The central business problem addressed in the project is the need to enhance portfolio performance when investing in mutual funds. Stakeholders, when investing money, seek analytical insights to make wiser and more effective financial choices. The project aims to fulfill this need by conducting a thorough analysis, providing visual representations of funds across various criteria, and helping identify the most suitable options for shortlisting.

Motivation

The motivation for the project is twofold:

- *Portfolio Composition*: Evaluating portfolio structure, focusing on asset allocation and sector exposure to gain insights into diversification and risk management.

- *Risk-Return Profile*: Determining the trade-off between costs and benefits in funds and their respective categories, providing a clearer understanding of their financial performance and risk management.

Motivation Sources:

- Investopedia: “How to Achieve Optimal Asset Allocation.”

  Link: https://www.investopedia.com/managing-wealth/achieve-optimal-asset-allocation/

- Vanguard: "The Global Case For Strategic Asset Allocation."

  Link:https://static.vgcontent.info/crp/intl/auw/docs/literature/research/The-global-case-for-strategic-asset-allocation.pdf

Data Source and Nature

The dataset, named "MutualFunds.csv," consists of 23,784 data points and 298 columns representing various features. Each row corresponds to an individual mutual fund, and the columns contain information such as category, total net assets, ratings, annual holding turnover, expense ratio, growth ratio, and more. The data is accessible as an open-source dataset on Kaggle. Link: https://www.kaggle.com/datasets/stefanoleone992/mutual-funds-and-etfs

Data Significance

This dataset includes financial information collected from Yahoo Finance, encompassing all U.S. Mutual Funds and their historical prices. With a diverse set of features, the dataset provides a rich source for analyzing sectoral allocations, predictive modeling for fund returns, expense ratio evaluation, diversification, risk management, and feature engineering.

Data Overview

Overview of the Dataset

The dataset, named "MutualFunds.csv," comprises 23,784 data points and 298 columns representing various features. Each row corresponds to an individual mutual fund, while the columns encompass a wide array of information, including category, total net assets, ratings, annual holding turnover, expense ratio, growth ratio, and more.

Data Description

The dataset provides a comprehensive view of mutual funds, with key features such as category, total net assets, ratings, annual holding turnover, expense ratio, growth ratio, and more. With 23,784 data points, it offers a diverse set of information for in-depth analysis.

```
In [31]: df = pd.read_csv("SelectedDF.csv")
         df.head(2)
```

Out[31]:

	Unnamed: 0	fund_symbol	fund_short_name	investment_type	size_type	initial_investment	subsequent_investment	fund_category	fund_family	management_
0	0	AAAAX	DWS RREEF Real Assets Fund - Cl	Value	Large	1000.0	50.0	World Allocation	DWS	John Vo
1	1	AAAEX	AllianzGl Health Sciences Fund	Blend	Large	1000000.0	NaN	Health	Virtus	Christopher

2 rows × 52 columns

```
In [36]: df.describe()
```

Out[36]:

	Unnamed: 0	initial_investment	subsequent_investment	total_net_assets	year_to_date_return	week52_high	week52_low	fund_yield	morningstar_c
count	23783.000000	2.378300e+04	23783.000000	2.378300e+04	23783.000000	23783.000000	23783.000000	23783.000000	2
mean	11891.000000	3.589909e+06	382.013371	4.917153e+09	0.093485	24.995830	20.352340	0.017481	
std	6865.705062	7.707717e+07	6777.647178	2.108338e+10	0.081580	34.995656	26.473466	0.021854	
min	0.000000	0.000000e+00	0.000000	0.000000e+00	-0.522800	1.700000	1.150000	0.000100	
25%	5945.500000	0.000000e+00	0.000000	1.425418e+08	0.027100	11.400000	10.470000	0.008800	
50%	11891.000000	1.000000e+03	0.000000	6.428231e+08	0.093485	15.810000	13.350000	0.017481	
75%	17836.500000	1.500000e+04	50.000000	2.445104e+09	0.148700	26.450000	21.450000	0.019200	
max	23782.000000	5.000000e+09	500000.000000	7.534100e+11	0.578900	2092.820000	1637.070000	1.258500	

8 rows × 41 columns

Stakeholders

Inferences derived from the analysis of this dataset can benefit a broad range of stakeholders. The following entities are highlighted for the purpose of this proposal:

- *Investment Committee*: Government agencies and industry regulators overseeing the mutual fund sector can utilize the data to ensure compliance with regulations and maintain market integrity.
- *Fund Managers*: Professionals responsible for mutual fund management seek valuable insights into fund performance, asset allocation, and risk management.
- *Individual and Institutional Investors*: Both individual and institutional investors can make well-informed investment decisions to maximize returns and manage risks effectively.
- *Data Analysts and Data Scientists*: Experts in data analysis play a crucial role in extracting insights from the data, processing, and interpreting it to derive actionable insights.

Inferences from Exploratory Data Analysis (EDA)

During the initial phase of the project, exploratory data analysis (EDA) was conducted to establish a thorough grasp of the dataset. Diverse data visualizations, including scatterplots, histograms, and boxplots, were generated to examine the distribution and interconnections among various features. The primary areas of emphasis included fund performance, expenses, and investment strategies.

Data Cleaning and Preprocessing

Data Cleaning Steps

The initial phase of data processing involved thorough data cleaning to ensure data quality and completeness. The following steps were implemented:

- *Handling Null Values*: Null values were meticulously addressed using mean and zero-fill strategies. Columns for which null values were filled include Fund Yield, Morningstar Overall Rating, Morningstar Risk Rating, Annual Holdings Turnover, and others.

- *Outlier Treatment*: Rigorous scrutiny of outliers and irrelevant columns demanded iterative refinement. This involved dropping specified columns that were considered irrelevant for the analysis.

- *Data Imputation*: Mean-fill strategy was employed for specific columns, including Year to Date Return, Week52 High, Week52 Low, and Others. Columns such as Asset Cash, Initial Investment, Subsequent Investment, Fund Bond Maturity, and Total Net Assets were filled with zeros.

In [32]:
```
# List of columns for which you want to fill null values with the mean
columns_to_fill = ['fund_yield', 'morningstar_overall_rating', 'morningstar_risk_rating', 'annual_holdings_turnover'
                   'fund_annual_report_net_expense_ratio', 'category_annual_report_net_expense_ratio',
                   'fund_year3_expense_projection', 'fund_year5_expense_projection', 'fund_year10_expense_projectio
                   'asset_cash', 'fund_sector_basic_materials', 'fund_sector_utilities', 'fund_sector_technology',
                   'fund_sector_real_estate', 'fund_sector_industrials', 'fund_sector_healthcare',
                   'fund_sector_financial_services', 'fund_sector_energy', 'fund_sector_consumer_defensive',
                   'fund_sector_consumer_cyclical', 'fund_sector_communication_services', 'fund_price_earning_ratio
                   'fund_price_sales_ratio', 'fund_year3_earnings_growth', 'fund_bond_maturity',
                   'morningstar_return_rating', 'fund_return_ytd', 'fund_return_3years', 'fund_return_5years',
                   'fund_return_10years', 'environment_score', 'esg_score', 'sustainability_rank', 'governance_scor

# Fill null values in the specified columns with the mean
df[columns_to_fill] = df[columns_to_fill].fillna(df[columns_to_fill].mean())
```

In [33]:
```
# Columns to fill with mean
mean_fill_columns = ['year_to_date_return','week52_high','week52_low','morningstar_overall_rating','morningstar_risk

# Columns to fill with 0
zero_fill_columns = ['asset_cash','initial_investment', 'subsequent_investment', 'fund_bond_maturity', 'total_net_as

# Fill null values with mean
df[mean_fill_columns] = df[mean_fill_columns].fillna(df[mean_fill_columns].mean())

# Fill null values with 0
df[zero_fill_columns] = df[zero_fill_columns].fillna(0)

df.head(2)


# Drop specified columns
df2 = df.drop(['fund_symbol', 'fund_short_name', 'fund_category', 'fund_family', 'investment_type','size_type','inve
```

Feature Engineering

Feature engineering was a crucial aspect of the data preprocessing phase. Notable steps included:

- *Creation of Total Investment Column*: The total investment column was engineered by combining initial and subsequent investments. This new feature provides a holistic view of the overall investment in each mutual fund.

Descriptive Statistics:

- Employed the describe() function to acquire fundamental statistical metrics for the dataset.
- *Rationale*: Descriptive statistics offer a rapid summary of the dataset's central tendency, spread, and shape, facilitating comprehension of variable distributions.

Visualizations for Univariate, Bivariate, and Multivariate Analysis:

- Utilized a diverse set of visualizations, including count plots, histogram plots, bar plots, scatter plots, tree maps, heatmaps, and more, to comprehensively explore the distributions of various features within the dataset.
- Rationale: Employing a variety of visualizations facilitates a nuanced examination of the distribution patterns of individual features, contributing to a more insightful and comprehensive understanding of the dataset.

Data Cleaning and Preprocessing Summary

The data cleaning and preprocessing phase aimed to ensure the quality and completeness of the dataset. By addressing null values, outliers, and irrelevant columns, the dataset was refined for subsequent analysis. Feature engineering, particularly the creation of the total investment column, added a valuable dimension to the dataset for more nuanced insights.

Analysis Steps and Shifts in Direction

Purpose and Choices

The analysis embarked on a comprehensive exploration of the risk-return profile and portfolio composition within the context of mutual fund investments. The primary choices and objectives guiding the analysis were:

- *Risk-Return Profile Exploration*: The analysis delved into understanding the trade-off between risk and return in mutual funds, unraveling insights into the risk and return behavior of various funds.

- *Portfolio Composition Analysis*: Emphasis was placed on deciphering how portfolio composition impacts asset allocation and risk within the mutual fund landscape. This involved examining the sectoral allocation of funds and its implications.

Predictive Modelling for Fund Returns

Utilizing advanced statistical and machine learning techniques, the analysis aimed to develop predictive models for fund returns. Specific steps included:

- *Random Forest Regressor*: A Random Forest regressor was employed to predict fund returns, leveraging an ensemble of decision trees for improved predictive accuracy.

- *Decision Tree Regressor*: A Decision Tree regressor was utilized to model the non-linear relationships between various features and fund returns.

- *Extra Tree Regressor*: The Extra Tree regressor, an ensemble learning method, was employed to further enhance the accuracy of predicting fund returns.

Expense Ratio Evaluation

The assessment of expense ratios in mutual funds was a focal point of the analysis. Linear regression analysis was conducted to explore the relationship between expenses and fund returns, determining whether lower expense ratios correlate with better performance.

Diversification and Risk Management

The analysis included an examination of diversification and risk management strategies within mutual fund portfolios. This encompassed sector and category analysis, correlation analysis, and exploration of the risk and return trade-offs in funds.

Challenges Faced

The journey of analysing mutual fund performance and behavior was not without its set of challenges. The following obstacles were encountered during the course of the project:

- *Data Quality and Completeness*: The dataset exhibited instances of missing or incomplete data, requiring meticulous handling to ensure the reliability of the analysis.

- *Dynamic Market Conditions*: The financial landscape is dynamic, influenced by ever-changing market conditions that can impact the behavior of mutual funds.

- *Complexity of Financial Data*: Financial data, inherently intricate, presented challenges in terms of understanding and interpreting complex patterns.

- *Risk Assessment and Management*: Assessing and managing risks associated with mutual fund investments, a crucial aspect of the analysis, presented inherent difficulties.

- *Model Interpretability*: Ensuring the interpretability of the models used in predictive analytics posed a challenge, especially when dealing with complex machine learning models.

- *Overfitting in Prediction Models*: Overfitting, a common issue in predictive modeling, could potentially compromise the generalization ability of the models.

How Challenges Were Overcome

Addressing these challenges required a combination of methodical approaches and strategic decision-making:

- *Data Quality and Completeness*: The dataset underwent a thorough data cleaning process, where missing values were handled using mean and zero-fill strategies. Iterative refinement was conducted to ensure the dataset's completeness.

- *Dynamic Market Conditions*: The analysis considered the dynamic nature of the market by incorporating macroeconomic indicators and external factors that might influence fund behavior.

- *Complexity of Financial Data*: Collaboration with domain experts and financial analysts played a crucial role in interpreting complex financial metrics. Visualization techniques were employed to simplify data patterns.

- *Risk Assessment and Management*: Risks associated with small-sized funds, especially those identified as bottom performers, were rigorously assessed to provide insights for potential investors.

- *Model Interpretability*: The use of models with better interpretability, such as linear regression, was emphasized. Feature importance analysis techniques were employed to enhance the interpretability of complex models.

- *Overfitting in Prediction Models*: Rigorous model validation techniques, including cross-validation, were implemented. Regularization methods were considered to prevent overfitting and improve model generalization.

The successful navigation through these challenges contributed to the robustness and reliability of the analysis, ensuring that the insights derived are meaningful and actionable.

Tools and Technology Used

The analysis of mutual fund performance and behavior involved the utilization of a diverse set of tools and technologies to ensure a comprehensive and insightful exploration. The primary tools and technologies employed in this project include:

Environment

- *Jupyter Notebooks*: Provided an interactive and collaborative environment for coding in Python, allowing for a seamless integration of code, visualizations, and explanatory text.

Programming Languages

- *Python*: A versatile and widely-used programming language, served as the core language for data analysis, machine learning, and statistical modeling.

Libraries

- *Scikit Learn*: Scikit Learn, a powerful machine learning library for Python, played a pivotal role in implementing various predictive models and statistical analyses.

- *Matplotlib*: A comprehensive 2D plotting library, was utilized for creating a variety of static, interactive, and animated visualizations.

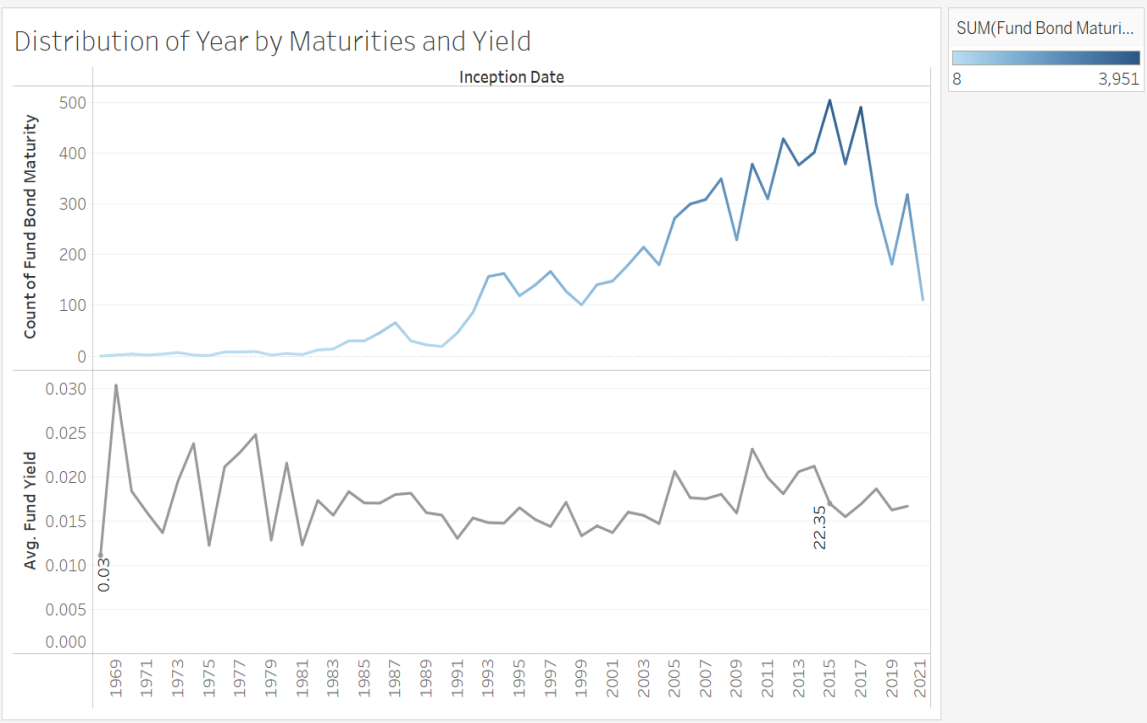
- *Seaborn*: Built on top of Matplotlib, enhanced the visual appeal of statistical graphics and provided additional functionality for data visualization.

Visualization Tool

- *Tableau*: A leading data visualization tool, was employed to create interactive and dynamic visualizations that facilitate a more intuitive understanding of complex patterns in the data.

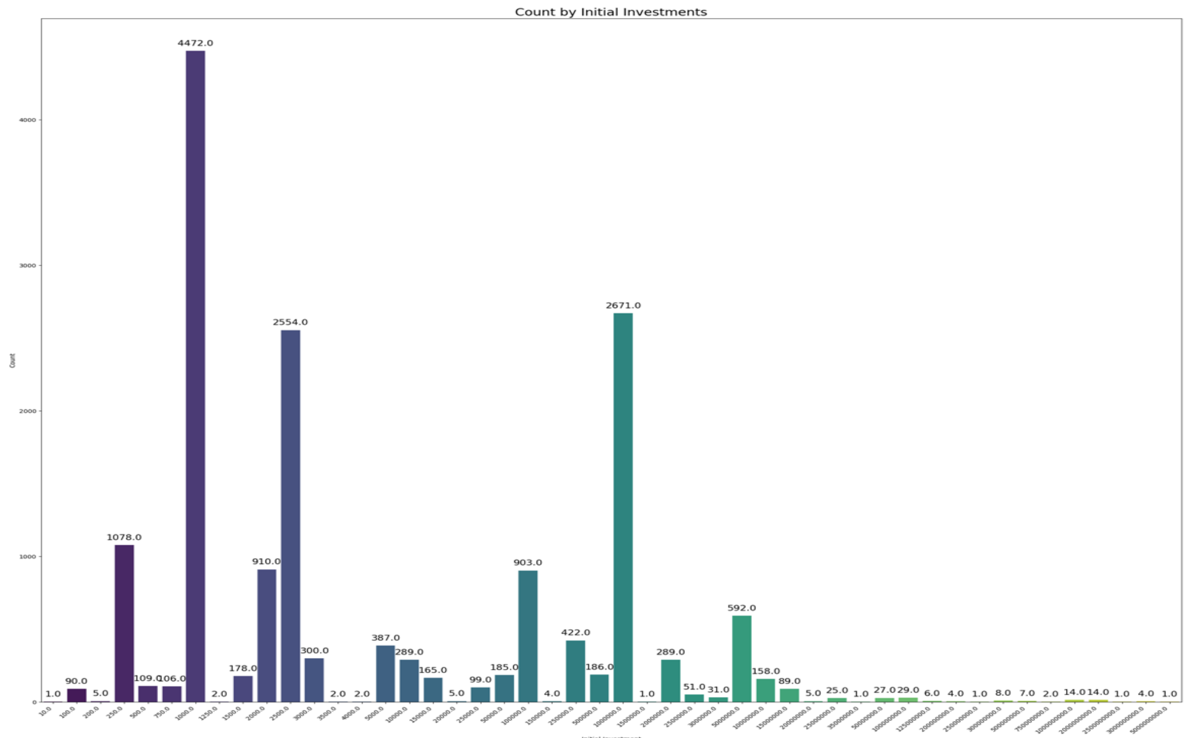
Visualizations

1. **Commencement and Maturity Trends:**



The analysis of commencement and maturity trends revealed a significant year in 2015, where 505 funds matured. Additionally, the examination of average yield trends highlighted notable fluctuations over the years. Specifically, in 1968, the average yield was at its lowest, measuring 0.03%. In contrast, the year 2015 experienced the highest average yield, reaching 22.35%. These trends indicate dynamic shifts in the mutual fund landscape, with potential implications for investor decision-making and portfolio management.

2. Distribution of Initial Investments



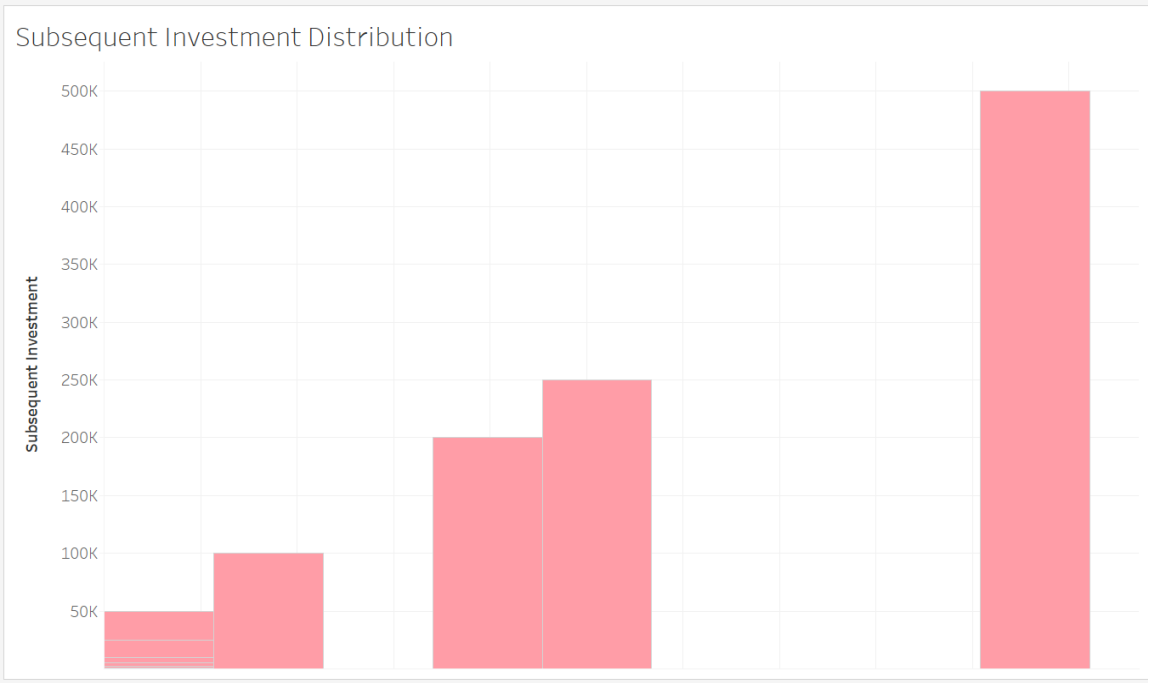
The analysis of initial investments revealed a diverse landscape, ranging from as low as $10 to a substantial $5 billion. Common initial investments were clustered around $1,000, $2,500, and $1 million. Uncommon investments, such as $5 billion, were minimal.

Financial Implications

- *Diverse Initial Investments*: The presence of a wide range of initial investments demands strategic planning for investors to optimize their portfolio based on their financial capacity and risk tolerance.

- *Recognizing Common Initial Investments*: Understanding the spectrum of common initial investments aids potential investors in comprehending the available investment options and making informed decisions aligned with their financial goals.

3. **Distribution of Subsequent Investments**

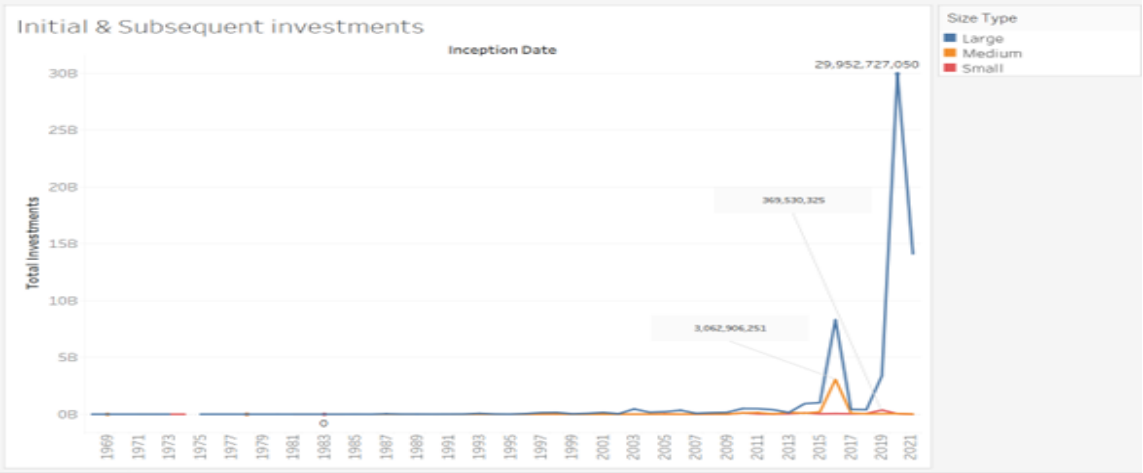


The examination of subsequent investments revealed additional insights into the financial behaviour of mutual funds. The minimum and maximum subsequent investments were not explicitly mentioned, but it is understood that they play a role in the overall financial dynamics of the funds.

Financial Implications

- *Optimizing Portfolio Based on Subsequent Investments*: Investors can strategically plan and optimize their portfolio by considering subsequent investments in addition to initial investments. This holistic view contributes to more informed decision-making.

4. Size-wise Distribution of Funds

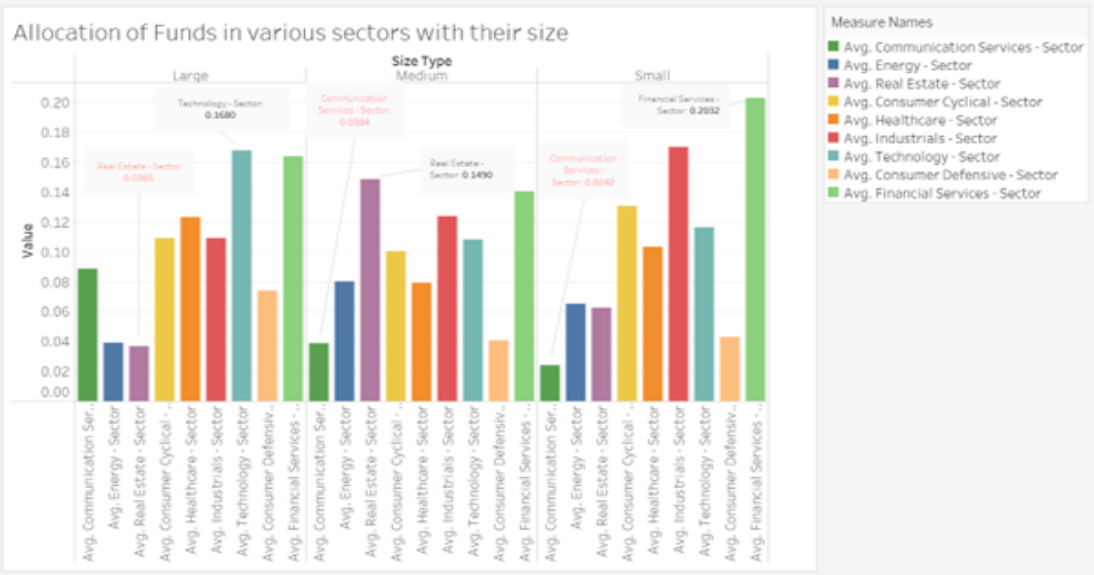


The analysis of size-wise distribution and allocations of total investments provided valuable insights into the mutual fund landscape. Overall, large-sized funds dominated through the years.

Financial Implications

- *Strategic Sectoral Allocation*: Understanding the dominant sectors within each size category allows investors and fund managers to strategically allocate assets, potentially enhancing fund performance and managing risk effectively.

5. Sectoral Allocation of Funds



The analysis further delved into sectoral allocations based on fund size. Financial Services consistently held a significant share of allocated funds across all sizes. This insight aids in understanding the sectoral preferences and allocations within the mutual fund landscape.

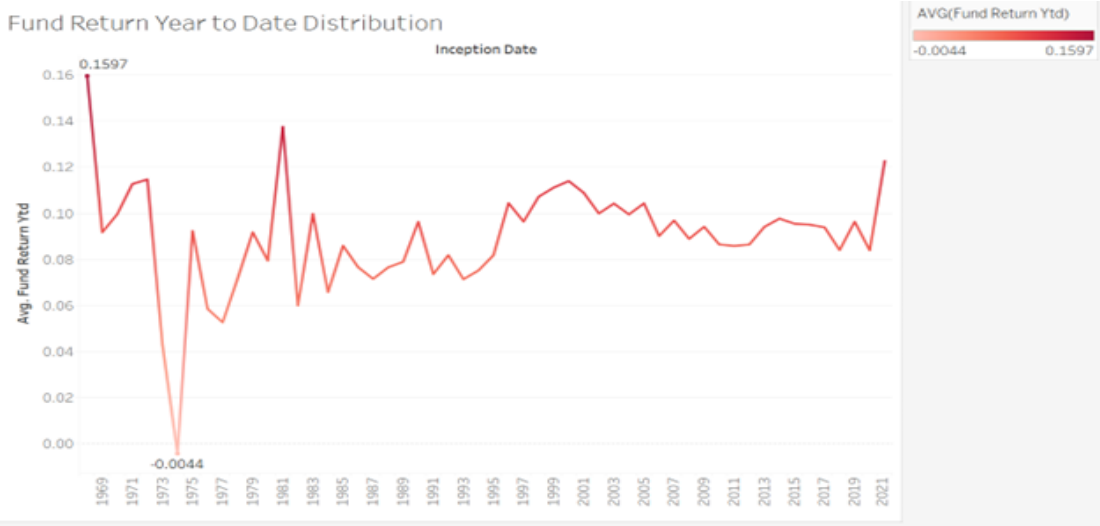
Within each size category, sectoral allocations exhibited distinct trends:

- Small Size:
  - Top Sector: Financial Sector.
  - Least Sector: Communication Services.
- Mid Size:
  - Top Sector: Real Estate.
  - Least Sector: Communication Services.
- Large Size:
  - Top Sector: Technology.
  - Least Sector: Real Estate.

Financial Implications

- *Sectoral Preferences*: Investors and fund managers can utilize information on sectoral preferences within different fund sizes to make informed decisions regarding portfolio diversification and risk management.

6. ROI and Volatility Trends



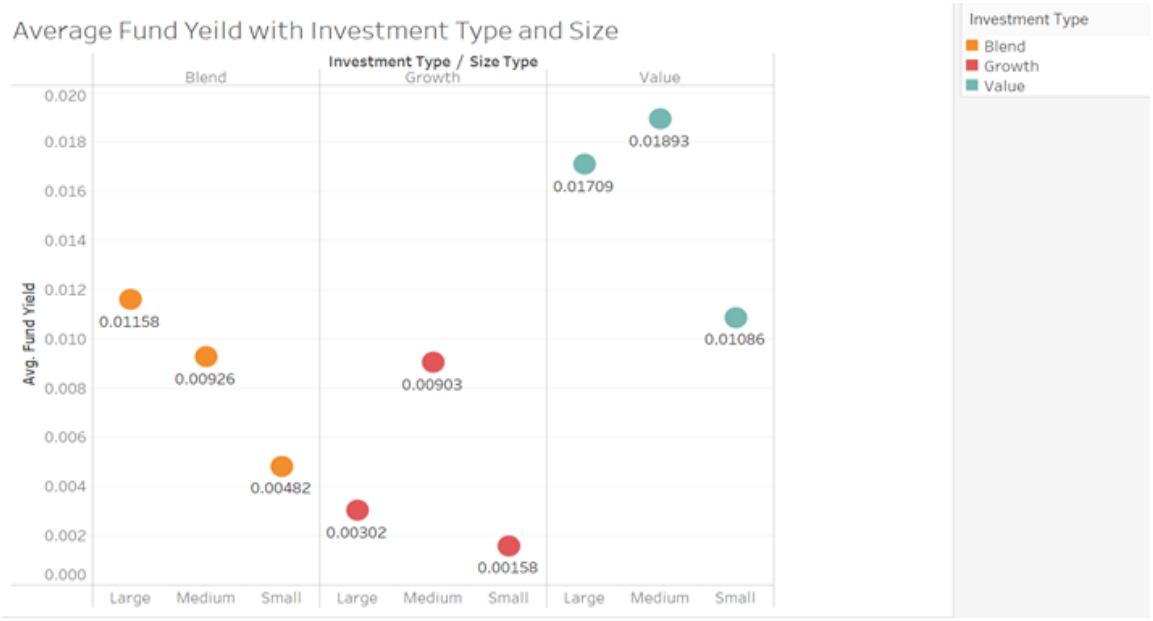
The analysis of Return on Investment (ROI) and Volatility Trends revealed noteworthy patterns in the performance of mutual funds over time. Return on Investment exhibited volatility in earlier years, with recent times showcasing reduced fluctuations. Specifically, the year 1974 experienced the highest ROI at 25.27%, while the lowest was observed in the same year at -0.44%.

Financial Implications

- *ROI Fluctuations*: Fluctuations in Return on Investment (ROI) suggest a dynamic and evolving market environment. Rapid changes in ROI indicate potential market shifts and uncertainties.

- *Reduced Volatility*: Decreased volatility implies a potential trend towards market stabilization. Investor Confidence: Reduced fluctuations may enhance investor confidence in the market's stability and predictability.

7. Average Fund Yield Across Investment Types and Size Categories



The analysis of Average Fund Yield Across Investment Types and Size Categories provided insights into yield variations based on investment types (Value, Growth, Blend) and size categories (Large, Medium, Small).

Key observations include:

- Growth Funds:
  - Small Size: Least Average Yield at 0.158%.
  - Medium Size: Peaked overall yield within Growth at 0.903%.

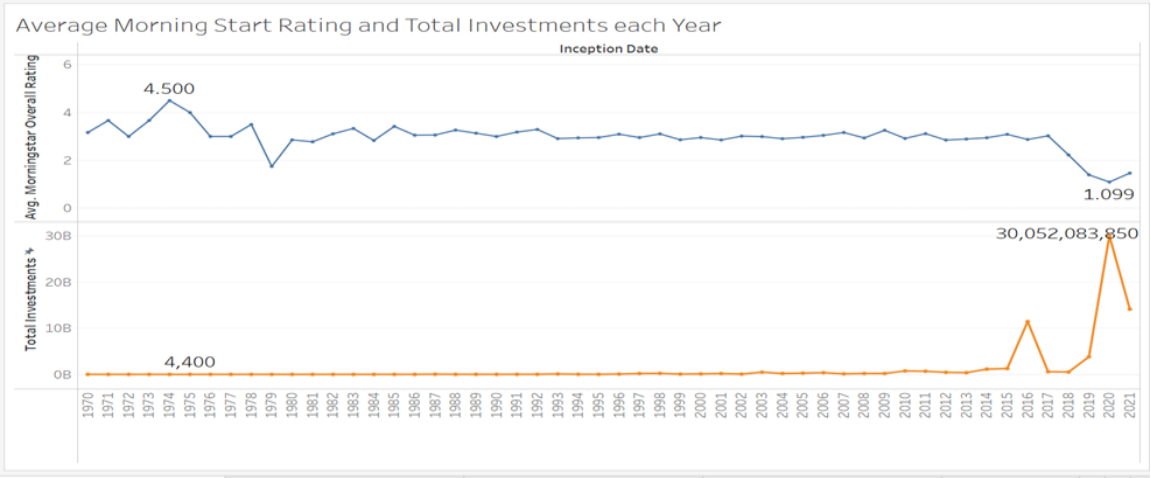
- Blend Funds:
  - Direct Relationship with Size.

- Value Funds:
  - Medium Size: Highest overall yield at 1.893%.
  - Small Size: Least Average Yield.

Financial Implications

- *Investment Type Impact*: Understanding the yield variations among different investment types and sizes allows investors to align their investment strategies with their risk tolerance and return expectations.

8. Morningstar Ratings



The Morningstar Ratings analysis highlighted trends and patterns in overall ratings over time. Key findings include:

- Morningstar Overall Ratings:

  - Highest Rating: 1975, with a Morningstar overall average rating of 4.50 out of 5.
  - Lowest Rating: 2020, at 1.099 out of 5.

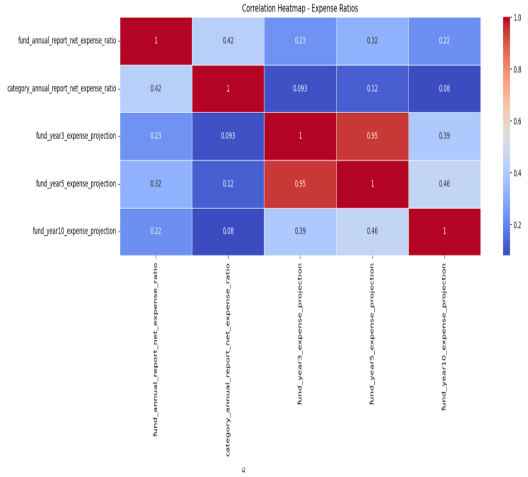
- Investment Amount Trends:

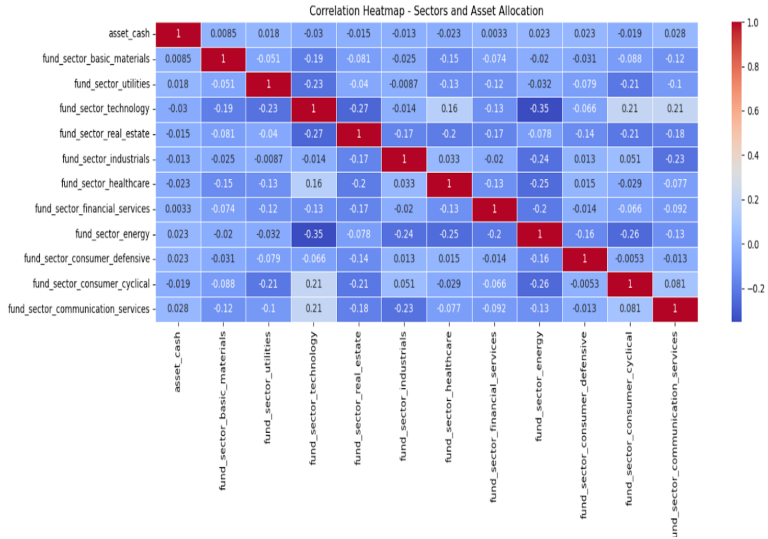
  - Lowest Total Investment: 1994, at $4,400.
  - Lowest Overall Rating Year (2020): Highest total investment, holding $30,052,083,850.

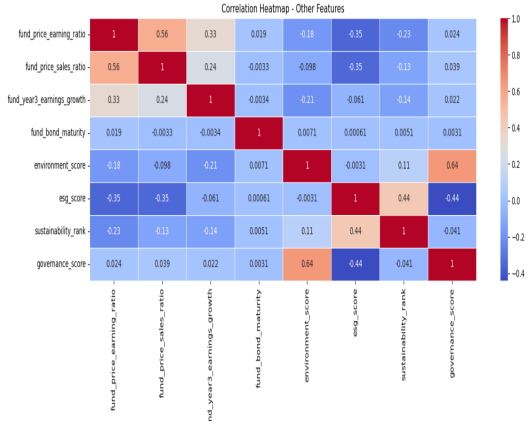
Financial Implications

- *Historical Rating Trends*: Understanding historical highs and lows provides context for fund performance. Discrepancy in 1994 and 2020 highlights the non-linear relationship between ratings and investment amounts.

9. Correlation Heatmaps







The Correlation heatmap analysis focused on exploring relationships between various performance metrics, expense ratios, sectors, and other features. Key insights include:

- *Performance Metrics*

  - A weak positive correlation exists between fund_yield and morningstar_risk_rating (0.104).

  - Morningstar Overall Rating has a strong positive correlation with morningstar_return_rating (0.907).

  - Morningstar Return Rating and Fund Return 3 Years exhibit a strong positive correlation (0.301).

- Expense *Ratios*:

  - Fund Annual Report Net Expense Ratio has a strong positive correlation with Category Annual Report Net Expense Ratio (0.423).

  - Strong correlations exist among various expense projections (fund_year3_expense_projection, fund_year5_expense_projection, fund_year10_expense_projection).

- *Sectors and Asset Allocation*:

  - Financial Services consistently holds a significant share across all fund sizes.

  - Negative correlations are observed between Asset Cash and Fund Sector Technology (-0.03) and Fund Sector Energy (-0.03).

- *Other Features*:

  - Fund Price Earning Ratio and Fund Price Sales Ratio have a moderate positive correlation (0.565).

  - Negative correlations are observed between ESG-related scores (Environment Score, ESG Score) and Governance Score.

Financial Implications

- *Performance Metrics*: The positive correlation between Morningstar Overall Rating and Morningstar Return Rating suggests that higher overall ratings are associated with better return ratings. This can guide investors in selecting funds with stronger Morningstar Overall Ratings for potentially better returns.

- *Expense Ratios*: The strong positive correlation between Fund Annual Report Net Expense Ratio and Category Annual Report Net Expense Ratio (0.423) indicates that changes in one are likely to influence the other. Investors should consider expense ratios carefully as they play a crucial role in overall fund performance.

- *Sectors and Asset Allocation*: The consistent significant share of funds in the Financial Services sector across all fund sizes indicates the sector's importance in fund allocations. Investors may assess their portfolios for exposure to this sector based on the fund size.

- *Other Features*: The moderate positive correlation between Fund Price Earning Ratio and Fund Price Sales Ratio (0.565) suggests that these metrics move together, influencing each other. Investors may consider these ratios in conjunction for a more comprehensive analysis of fund valuation.

Overall, these financial implications underscore the importance of considering interrelationships between various metrics when making investment decisions.

10. Model Development and Testing

Fund Price Earning Ratio Analysis.

1) **ML Model:** Random Forest

   Target Variable: fund_price_earning_ratio

In [6]:

```
#Random Forest -> fund_price_earning_ratio


X = df2.drop(['fund_price_earning_ratio'],axis=1)
y = df2['fund_price_earning_ratio']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("For Random Forest : ")

print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

```
Mean Squared Error: 2.024959654063318
R-squared: 0.9509406215240864
```

2) **ML Model**: Extra Trees Regressor

   Target Variable: fund_price_earning_ratio

In [7]:

```
#Extra Trees Regressor -> fund_price_earning_ratio

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = ExtraTreesRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)


print("For Extra Tree Regressor : ")
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

```
Mean Squared Error: 1.587166214981995
R-squared: 0.9615471903903178
```

3) **ML Model:** Decision Tree Regressor

   Target Variable: fund_price_earning_ratio

```
In [8]: #DecisionTreeRegressor -> fund_price_earning_ratio

        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

        # Initialize and train the decision tree regressor model
        model = DecisionTreeRegressor(random_state=42)
        model.fit(X_train, y_train)

        # Make predictions on the test set
        y_pred = model.predict(X_test)

        # Evaluate the model
        mse = mean_squared_error(y_test, y_pred)
        r2 = r2_score(y_test, y_pred)

        print("For Decision Tree Regressor : ")
        print(f'Mean Squared Error: {mse}')
        print(f'R-squared: {r2}')

        Mean Squared Error: 4.9759866078179895
        R-squared: 0.8794450991681899
```

Amongst the above the Extra Tree Regressor has the highest R-squared value - **96.17%**

```
For Decision Tree Regressor :
Mean Squared Error: 4.918874307192528
R-squared: 0.8808287780807017
```

```
For Random Forest :
Mean Squared Error: 2.019810806149673
R-squared: 0.951065364393903
```

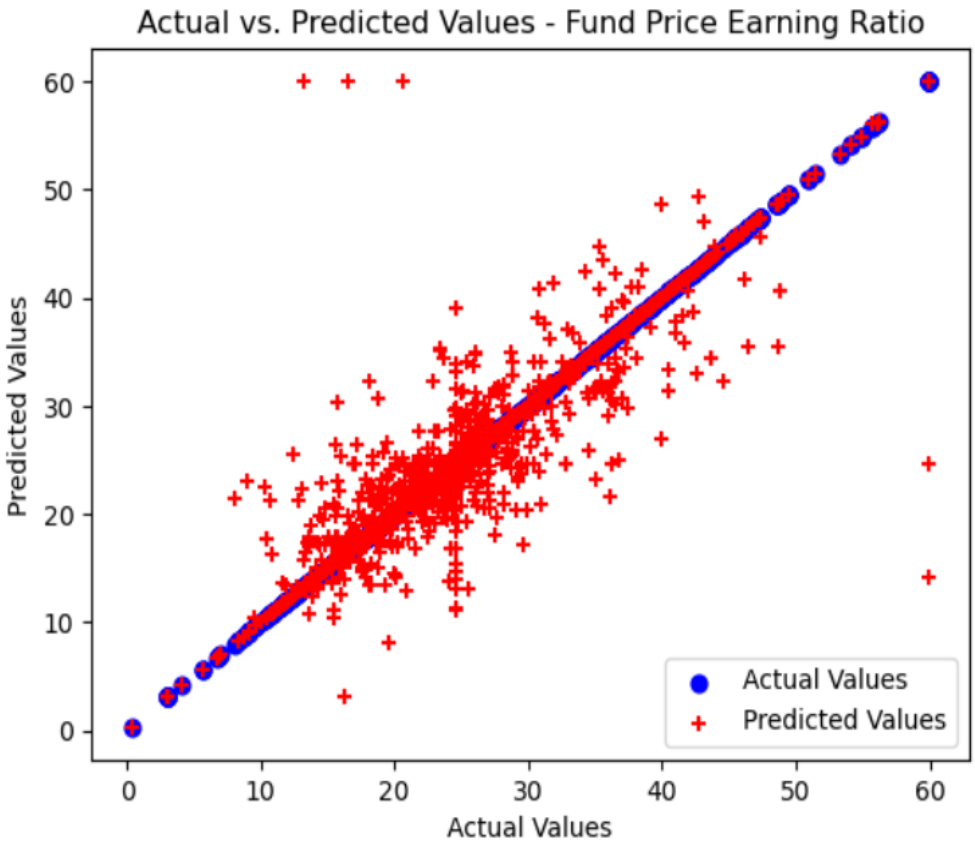
```
For Extra Tree Regressor :
Mean Squared Error: 1.578700981323357
R-squared: 0.9617522804527853
```

- The Extra Tree Regressor outperforms both Random Forest and Decision Tree Regressors in predicting the fund return over 5 years. It has the lowest MSE and the highest R-squared, indicating superior predictive accuracy.

- While Random Forest performs well with an R-squared of 0.9072, Extra Tree Regressor slightly surpasses it with an R-squared of 0.9149. Decision Tree Regressor, although decent, lags behind with an R-squared of 0.8334.

The choice of the regression model is crucial for accurate predictions, and in this context, Extra Tree Regressor demonstrates its effectiveness in modeling fund return.

Actual vs. Predicted Values - Fund Price earning Ratio



The scatter plot illustrates a generally positive correlation between actual and predicted values for fund return over 5 years.

Implication

- The consistent positive correlation suggests that the models, particularly the Extra Tree Regressor, effectively capture the underlying patterns in the data.

- The model predictions align with the actual fund returns, indicating reliability in forecasting future returns. Investors can use these models to make informed decisions based on the predicted fund returns.

Year to Date return Analysis

The machine learning models were used by the target variable year to date return

A) **ML Model**: Random Forest
Target Variable: year_to_date_return

```
#Random Forest -> Year to Date Returns



X = df2.drop(['year_to_date_return'],axis=1)
y = df2['year_to_date_return']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("For Random Forest : ")
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

```
Mean Squared Error: 4.371642870809334e-06
R-squared: 0.9993428814710024
```

B) ML Model: Extra Tree Regressor
Target Variable: year_to_date_return

```
#Extra Tree Regressor -> Year to Date Returns


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = ExtraTreesRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("For Extra Tree Regressor : ")
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

```
Mean Squared Error: 4.408069702307986e-06
R-squared: 0.999337406013231
```

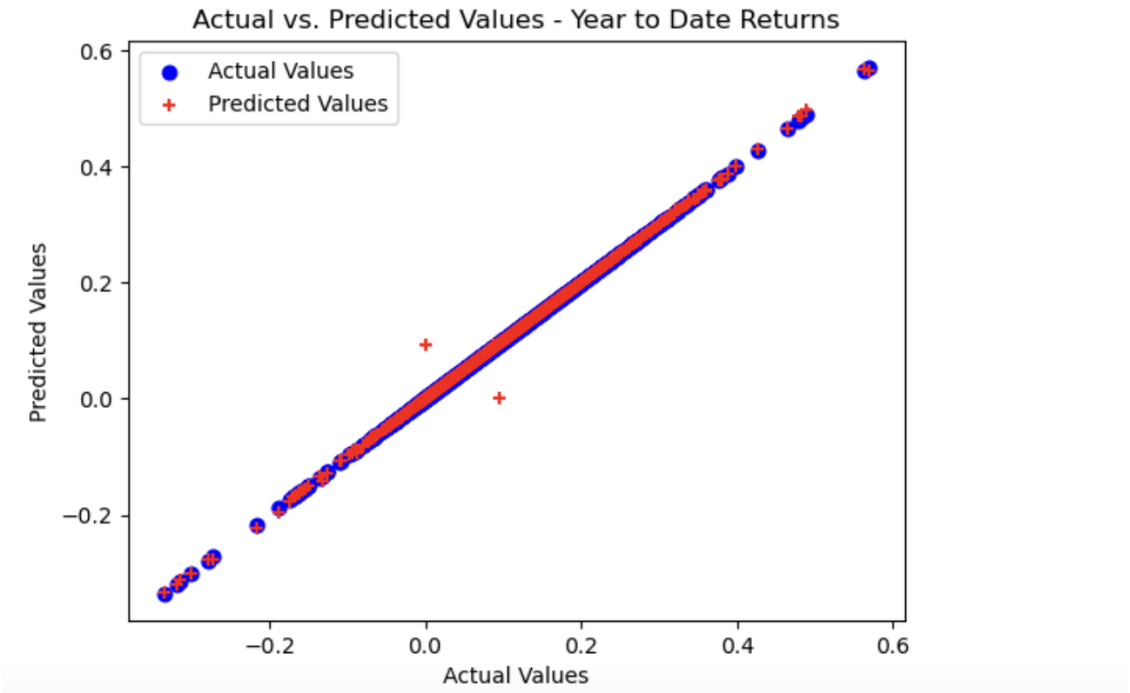
C) **ML Model:** Decision Tree Regressor
Target Variable: year_to_date_return

```
#DecisionTreeRegressor -> Year to Date Returns

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train the decision tree regressor model
model = DecisionTreeRegressor(random_state=42)
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("For Decision Tree Regressor : ")
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

```
Mean Squared Error: 5.651310504569858e-06
R-squared: 0.9991505296851971
```

Actual vs Predicted : Fund Year to Date Returns

```
# Scatter plot for actual values
plt.scatter(y_test, y_test, color='blue', label='Actual Values', marker='o', s=40)

# Scatter plot for predicted values
plt.scatter(y_test, y_pred, color='red', label='Predicted Values', marker='+', s=30)

plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.title("Actual vs. Predicted Values - Year to Date Returns")
plt.legend()
plt.show()
```



Fund Return 5 Years Analysis:

1) **ML Model** : Random Forest
   Target Variable: fund_return_5Years

```
#Random Forest -> fund_return_5years


X = df2.drop(['fund_return_5years'],axis=1)
y = df2['fund_return_5years']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("For Random Forest : ")
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

```
Mean Squared Error: 0.0002055200746461377
R-squared: 0.9071822781306205
```

2) **ML Model:** Extra Tree Regressor
   Target Variable: fund_return_5years

```
#Extra Tree Regressor -> fund_return_5years


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = ExtraTreesRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print("For Extra Tree Regressor : ")
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

```
For Extra Tree Regressor :
Mean Squared Error: 0.00018836948813910574
R-squared: 0.9149278882421743
```

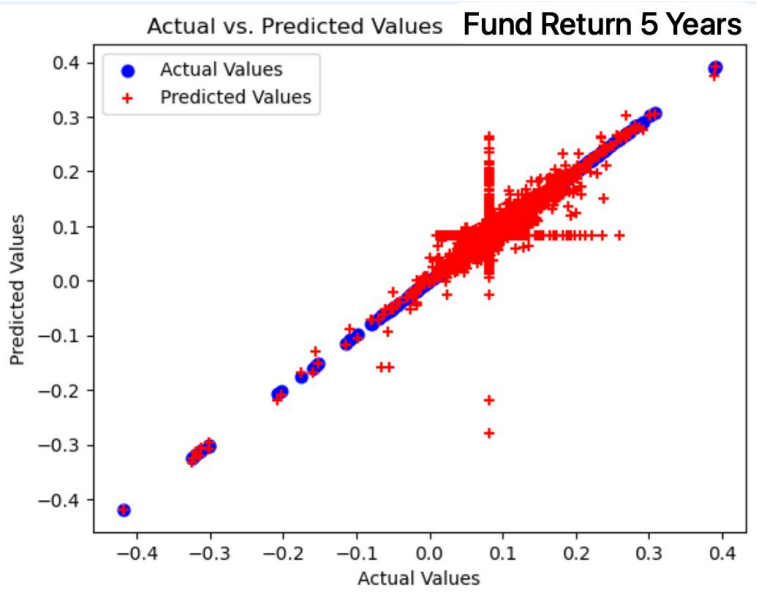
3) **ML Model**: Decision Tree Regressor

Target Variable:

```
#DecisionTreeRegressor -> fund_return_5years

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train the decision tree regressor model
model = DecisionTreeRegressor(random_state=42)
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("For Decision Tree Regressor : ")
print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

```
For Decision Tree Regressor :
Mean Squared Error: 0.0003689135462141956
R-squared: 0.8333899256055024
```

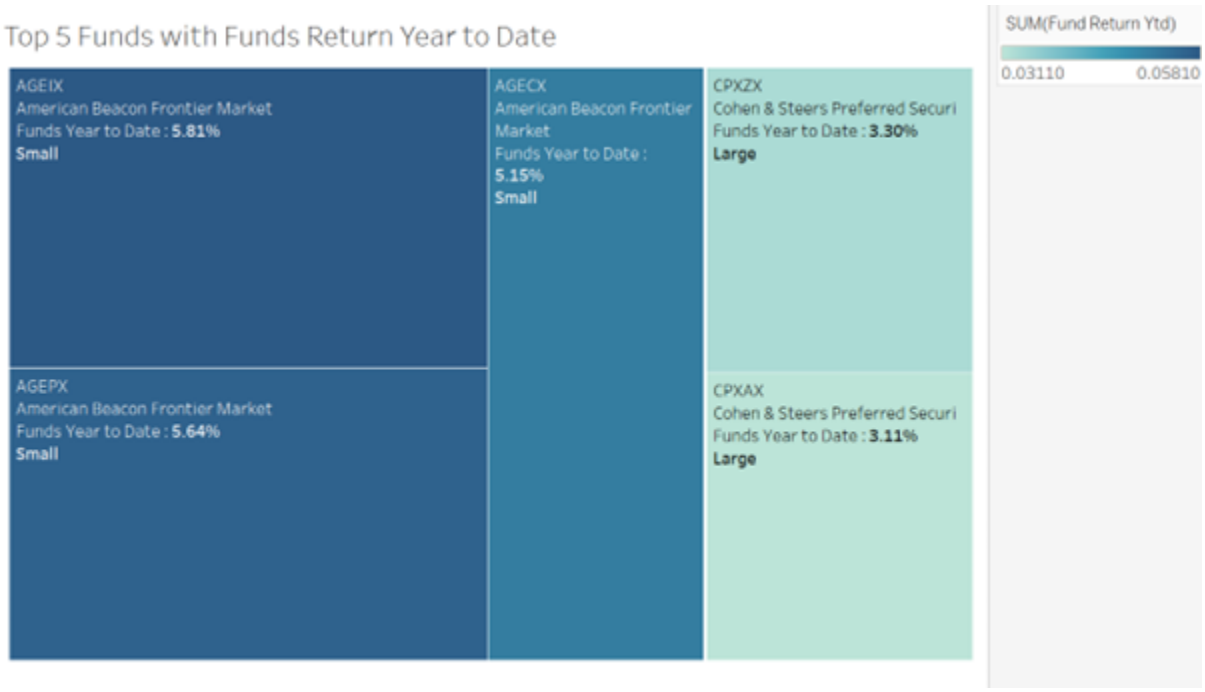
Actual vs Predicted : Fund Year 5 Years



Overall R-Squared for the models for target variables

Sr. No.	Models	Earning Ratio	YTD Returns	Fund Return 5 Years
1	Random Forest	95.09%	99.93%	90.71%
2	**Extra Tree Regressor**	**96.15%**	**99.93%**	**91.49%**
3	Decision Tree Regressor	87.94%	99.91%	83.33%

11. Top 5 Funds by YTD return



- AGEIX emerges as the top performer with a Return Ratio of 0.05810%.
- CPXAX is noted as the least performer.

Fund Family Performance

- American Beacon Fund Family dominates the top positions in small-sized funds.

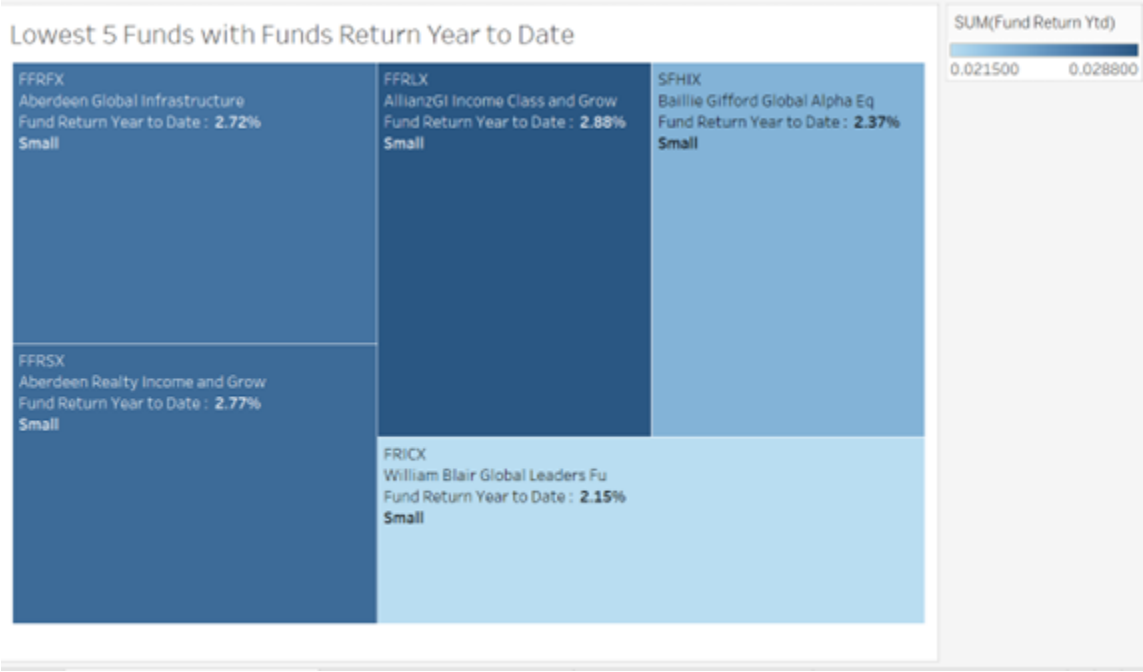
- Cohen & Steers Fund Family showcases strong performance within its family and fund size.

Implications

- American Beacon Fund Family is a notable performer in the small-sized fund category, holding the top three positions.

- Cohen & Steers Fund Family demonstrates consistent strong performance, securing two positions in the top five, indicating stability and reliability in their funds.

12. Bottom 5 Funds by YTD return



- FFRLX from the Virtus Family is identified as the top lagger.
- FRICX from William Blair holds the last position among the bottom performers.

Fund Family Performance

- Aberdeen Family Performance:
    - Two funds from the Aberdeen Family secure the 2nd and 3rd positions in the bottom performers.

Size Type Performance

- Size Category Analysis:
    - All funds in the bottom 5 are small-sized.

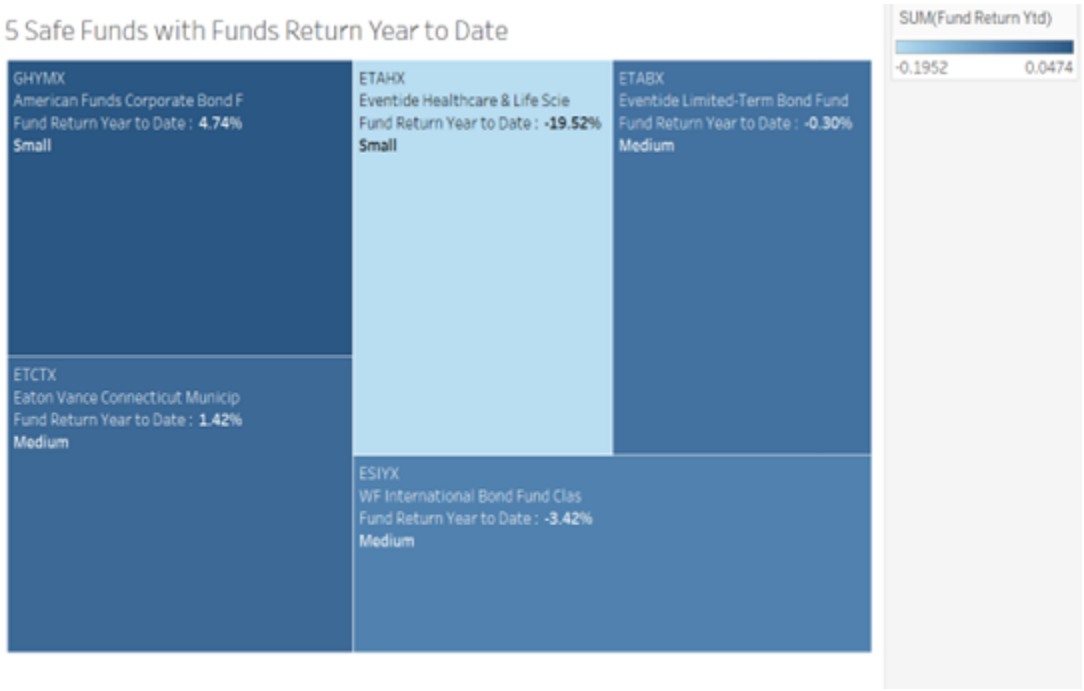
Implications

- Virtus Family (FFRLX) stands out as the lead performer among the bottom funds, suggesting potential challenges or issues with fund management.

- William Blair (FRICX) is positioned as the least performer, requiring attention and evaluation of its investment strategy.

- Aberdeen Family's representation in the bottom 5 indicates a need for a thorough review of their fund management practices.

13. 5 Safe Funds - Year to Date



- Top Performer (Small)
  - GHYMX emerges as the top performer with a strong YTD return.

- Last Position (Small)
  - ETAHX is identified as the least performer among the safe funds.

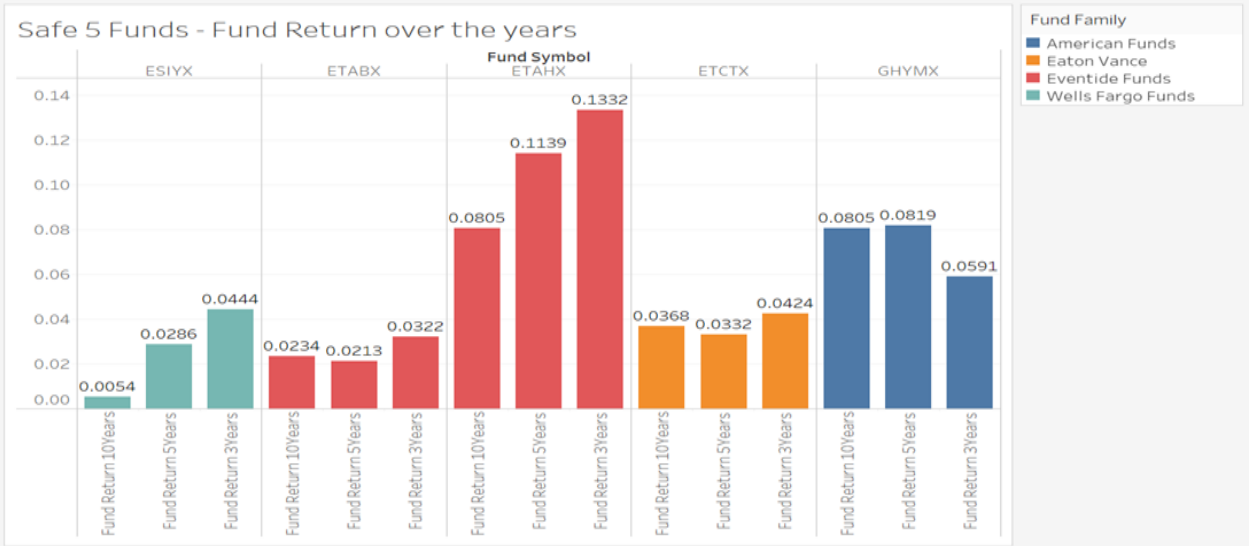
- Eventide Funds Family
  - Two out of the five funds in the 'Safe' zone belong to the Eventide Funds Family.

Implications

- GHYMX and ETAHX represent the extremes within the 'Safe' zone, both being small-sized funds.

- Eventide Funds Family's strong representation in the 'Safe' zone indicates the family's consistent focus on stability and positive returns in their funds.

14. Safe Funds - Historical Returns



Top Safe Funds - Overall Returns (3, 5, 10 Years)

- ETAHX (Small-Sized)
  - Dominates with the highest overall returns across all periods.
- GHYMX (Small-Sized)
  - Follows with strong overall performance.

- ETCTX (Medium-Sized)
  - Third-highest overall returns.

- ESIYX and ETABX (Medium-Sized):
  - Compete for the fourth and fifth spots alternatively.

Implications

- ETAHX stands out as the dominant performer, showcasing consistent strong returns across various timeframes.

- GHYMX maintains its position as a top performer, contributing to the overall stability and positive performance of the safe funds.

- ETCTX, ESIYX, and ETABX, being medium-sized funds, demonstrate competitive performance, offering a balanced approach to investors seeking safe investment options.

Conclusion and Reflection

Summary of Key Findings

- Diverse Initial Investments
  - Initial investments ranged from $10 to $5 billion.
  - Common initial investments clustered around $1,000, $2,500, and $1 million.
  - Uncommon investments, such as $5 billion, were minimal.

- Commencement and Maturity Trends:
  - 2015 was a significant year, with 505 funds maturing.
  - Average yield trends fluctuated, with 1968 witnessing the lowest at 0.03%, and 2015 the highest at 22.35%.

- ROI and Volatility Trends
  - ROI exhibited volatility in earlier years, with recent times showcasing reduced fluctuations.
  - Notable ROI variations, with 1974 experiencing the highest at 25.27% and the lowest at -0.44%.

- Morningstar Ratings and Investment Trends
  - Morningstar ratings ranged from historical highs in 1975 to a significant drop in 2020.
  - Medium-sized funds held the highest rating within and outside their fund type.

- Investment amounts varied, with 1994 having the lowest total investment and 2020 holding the largest sum.

- Avg Fund Yield by Investment Type and Size
  - Yield variations observed within different fund types and sizes.
  - Small-sized Growth funds had the least average yield, while medium-sized Growth funds peaked in overall yield.

- Sectoral Asset Allocation by Fund Size
  - Diverse sectoral allocations were witnessed based on fund size.
  - Financial Services consistently held a significant share of allocated funds across all sizes.

- Fund Classification - Top and Bottom Performers
  - Top performers included funds from the American Beacon and Cohen & Steers families, with AGEIX leading.
  - Bottom performers included FFRLX (Virtus) and FRICX (William Blair), both small-sized funds.

- Safe Zone - Historical Returns:
  - ETAHX (Small-sized) emerged as the top performer with the highest overall returns in 3, 5, and 10 years.
  - GHYMX (Small-sized) and ETCTX (Medium-sized) followed with strong overall returns.

Strategic Considerations

- Risk Management
  - Assess the implications of investment constraints on fund performance.
  - Evaluate the potential risks associated with small-sized funds in the bottom performers.

- Diversification Strategy
  - Utilize insights for informed sectoral allocation to optimize fund performance.

    - Consider the performance of these funds for potential inclusion in investment portfolios.

- Portfolio Optimization
    - Use information for informed investment decisions and potential portfolio optimization.
    - Consider the performance of these funds for potential inclusion in investment portfolios.

Future Work

The comprehensive analysis of the investment fund dataset has provided valuable insights and strategic considerations. However, there are several avenues for future work and exploration:

- *Deep Learning Models*: Consider exploring the application of deep learning models, such as neural networks, to enhance predictive modeling accuracy. Deep learning can capture intricate patterns and relationships in financial data.

- *Dynamic Portfolio Optimization*: Develop dynamic portfolio optimization strategies that adapt to changing market conditions. Incorporate real-time data and advanced algorithms to continuously optimize investment portfolios.

- *Alternative Data Sources*: Explore the integration of alternative data sources, such as social media sentiment analysis, economic indicators, or geopolitical events, to enhance the predictive power of the models.

- *Enhanced Risk Management*: Further refine risk management strategies by incorporating additional risk factors and stress-testing models to assess their robustness in extreme market scenarios.

- *Investor Sentiment Analysis*: Incorporate sentiment analysis of investor news, forums, and opinions to gauge market sentiment and potential impacts on fund performance.

- *Interactive Dashboard*: Develop an interactive and user-friendly dashboard using advanced visualization tools for stakeholders to dynamically explore and analyze fund performance, trends, and predictions.

- *External Validation*: Conduct external validation of models by comparing predictions with industry benchmarks or consulting with financial experts to ensure the models align with market expectations.

- *Ethical and Responsible Investing*: Integrate ethical and responsible investing criteria into the analysis, considering environmental, social, and governance (ESG) factors to align investments with sustainable and responsible practices.

- *Investment Strategy Simulation*: Implement a simulation framework to test different investment strategies based on historical data, allowing for the exploration of optimal strategies under various market conditions.

- *Continuous Monitoring*: Establish a continuous monitoring system to keep track of changes in the financial landscape, regulatory environment, and economic indicators, ensuring the models remain relevant and effective.

By exploring these avenues, the analysis can evolve to meet the dynamic challenges of the financial landscape, providing more accurate predictions and actionable insights for investors and fund managers.

References

- Malkiel, B. (Random Walk Down Wall Street)
- Bogle, J. C. (Common Sense on Mutual Funds)
- Markowitz, H. (Portfolio Selection: Efficient Diversification of Investments)
- Bernstein, W. J. (The Intelligent Asset Allocator: How to Build Your Portfolio to Maximize Returns and Minimize Risk)
- Bodie, Z., Kane, A., & Marcus, A. J. (Investments)"
- https://www.kaggle.com/datasets/stefanoleone992/mutual-funds-and-etfs
- *https://static.vgcontent.info/crp/intl/auw/docs/literature/research/The-global-case-for-strategic-asset-allocation.pdf*
- https://www.investopedia.com/managing-wealth/achieve-optimal-asset-allocation/