Exploratory Data Analysis on

Conservative Hybrid Mutual Fund

by

Group 26

Chaitri Vadaviya
ID: 202301243
Course: BTech(ICT)

Utsav Tala
ID: 202303018
Course:
BTech(MnC)

Parthiv Bhesaniya
ID: 202303037
Course:
BTech(MnC)

Course Code: IT 462
Semester: Autumn 2025

Under the guidance of

Dr. Gopinath Panda

November 30, 2025

Acknowledgment

I am writing this letter to express my heartfelt gratitude for your guidance and support throughout the duration of our project titled "Exploratory Data Analysis on Conservative Hybrid Mutual Funds." Your invaluable assistance, insightful feedback, and constant encouragement have played a crucial role in the successful completion of this work.

I am extremely grateful for your expertise and the time you dedicated to mentoring us. Your constructive suggestions and clear direction greatly enhanced our understanding of the subject and helped us refine the overall quality of our project.

Furthermore, I would like to extend my appreciation to Dhirubhai Ambani University (DAU) for providing the academic environment, resources, and support essential for completing this project. The opportunities and facilities offered by the institution contributed significantly to our research and analysis.

I would also like to express my gratitude to my peers and colleagues who have been supportive throughout this journey. Their valuable input and camaraderie have been a constant source of motivation.

Completing this project has been a tremendous learning experience, and we are confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for our future academic and professional pursuits.

Once again, thank you for your continuous guidance, support, and encouragement throughout this project.

Sincerely,
Chaitri Vadaviya (202301243)
Parthiv Bhesania (202303037)
Utsav Tala (202303018).

DECLARATION

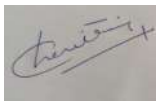
PLEASE CHANGE THIS FOLLOWING STATEMENTS AS PER YOUR REQUIREMENT.

We, [...] hereby declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

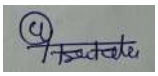
We acknowledge that the data used in this project is obtained from the [...] site. We also declare that we have adhered to the terms and conditions mentioned in the website for using the dataset. We confirm that the dataset used in this project is true and accurate to the best of our knowledge.

We acknowledge that we have received no external help or assistance in conducting this project, except for the guidance provided by our mentor Prof. Gopinath Panda. We declare that there is no conflict of interest in conducting this EDA project.

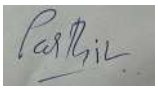
We hereby sign the declaration statement and confirm the submission of this report on 2nd July, 2023.



Chaitri Vadaviya
ID: 202301243
Course: BTech(ICT)



Utsav Tala
ID: 202303018
Course:
BTech(MNC)



Parthiv Bhesaniya
ID: 202303037
Course:
BTech(MNC)

CERTIFICATE

This is to certify that **Group 26**, comprising **Chaitri Vadaviya (202301243)**, **Parthiv Bhesania (202303037)**, and **Utsav Tala (202303018)**, has successfully completed an Exploratory Data Analysis (EDA) project on **Conservative Hybrid Mutual Funds**, using data obtained from `amfiindia.com`.

The EDA project presented by Group 26 is their original work and has been completed under the guidance of the course instructor, **Dr. Gopinath Panda**, who has provided continuous support and direction throughout the duration of the project. The project involves a detailed analysis of NAV values and portfolio datasets, and all results presented in the report are based solely on the data processed and interpreted during the study.

This certificate is issued to recognize the successful completion of the EDA project on Conservative Hybrid Mutual Funds, which demonstrates the analytical skills, understanding, and competence of Group 26 in the field of data analysis.

Dr. Gopinath Panda
Course Instructor

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

November 30, 2025

Contents

List of Figures

Abstract

This project presents an extensive Exploratory Data Analysis (EDA) of Conservative Hybrid Mutual Funds, focusing on evaluating their performance, risk characteristics, portfolio composition, and long-term stability. Data was collected from multiple sources and combined into three major datasets: fund analytics, daily NAV values, and portfolio asset allocation. These datasets were thoroughly inspected, cleaned, standardized, and preprocessed to handle missing values, inconsistent labels, date-format issues, and numerical conversion errors. The analysis provides a comprehensive understanding of fund performance across short-term, medium-term, and long-term return windows, including annual, quarterly, monthly, and weekly returns. Additionally, NAV time-series analysis was performed after identifying and removing non-trading dates and holiday gaps. The project further explores risk metrics such as Sharpe Ratio, Sortino Ratio, Beta, Alpha, Standard Deviation, and Fund Risk Grade, enabling a deeper evaluation of the risk-return profile of Conservative Hybrid Funds. Portfolio-level attributes such as equity allocation, debt exposure, cash positions, number of holdings, credit quality, maturity profile, and turnover ratio were analyzed to understand the underlying investment strategy and stability of the funds. The project also attempts predictive insights by experimenting with machine-learning models to explore potential relationships between NAV trends, fund returns, and portfolio characteristics. Overall, this EDA provides a clear, data-driven view of how Conservative Hybrid Funds behave in terms of performance, risk, and portfolio construction. The insights obtained can support investors and analysts in evaluating fund stability, understanding market behavior, and making more informed investment decisions.

Chapter 1. Introduction

1.1 Your Project idea

The primary objective of this project is to perform a comprehensive Exploratory Data Analysis (EDA) on Conservative Hybrid Mutual Funds to understand their performance behaviour, risk characteristics, and portfolio allocation patterns. The specific goals include: 1. Data Understanding Preparation To collect, inspect, clean, and preprocess multiple datasets related to mutual fund analytics, NAV history, and portfolio composition. To standardize inconsistent column names, convert data types, handle missing values, and organize time-series NAV data. 2. Performance Analysis To evaluate the funds' short-term, medium-term, and long-term returns, including annual, quarterly, monthly, and weekly performance. To compare return patterns across different Conservative Hybrid Funds. 3. NAV Trend Analysis To analyze daily NAV trends after removing holidays and invalid date entries. To identify the growth patterns and volatility present in the NAV time series. 4. Risk Assessment To study key risk indicators such as Sharpe Ratio, Sortino Ratio, Beta, Alpha, Standard Deviation, and overall fund risk grades. To determine the risk–return relationship within the Conservative Hybrid Funds category. 5. Portfolio Structure Analysis To examine the equity, debt, cash, and other asset allocations of each fund. To evaluate credit quality, maturity profile, turnover ratios, and the number of holdings. 6. Insight Generation To derive meaningful insights regarding fund stability, investment strategy, and long-term suitability. To identify which factors contribute most to consistent returns and lower risk. 7. Exploratory Modeling (Optional) To explore predictive relationships using machine-learning models for understanding NAV or return behaviour.

1.2 Data Collection

1. **NAV Data – Collected from MFAPI**
   We used MFAPI (a free API for Indian mutual fund data). From MFAPI, we downloaded daily NAV values for every Conservative Hybrid Fund scheme. This data became our `nav_values.csv` file. It included thousands of date columns with NAV values for each scheme.

2. **Analytics Data – Collected from Value Research**
   We used the Value Research Online website for performance and analytics data. Value Research provides return percentages, SIP returns, annual/quarterly/monthly returns, risk ratios, AUM, NAV details, and fund manager information. Because direct download was not available, we created this CSV manually:

   - Applied different filters on Value Research
   - Downloaded multiple smaller CSV files

- Then manually merged all these files

This combined sheet became our `analytics.csv` file.

3. **Portfolio Holdings Data – Collected from Moneycontrol**
We used Moneycontrol to get portfolio-related information for every fund. This included equity allocation, debt allocation, cash holding, number of stocks, credit rating, turnover ratio, etc. The data was downloaded from each scheme page. Then all the scheme-level portfolio sheets were merged together. This created our `portfolio_assets.csv` file.

4. **Cross-Verification with AMFI**
We used AMFI (Association of Mutual Funds in India) to verify:

   - Scheme names
   - NAV update dates
   - Fund categories

   AMFI helped ensure the data we collected from other websites was correct and consistent.

5. **Cleaning and Preprocessing the Collected Data**
After collecting all three CSV files, we performed data cleaning:

   - Removed symbols like "–" and converted them to proper missing values (NaN)
   - Fixed inconsistent column names
   - Corrected date formats in the NAV dataset
   - Removed blank and holiday NAV columns
   - Converted text-based numbers into numeric format
   - Standardized return columns (monthly, weekly, quarterly, annual)
   - Cleaned portfolio percentage fields (removed "%" and converted to numbers)
   - Merged scheme names properly

   This preprocessing made our datasets clean, structured, and ready for Exploratory Data Analysis (EDA).

1.3 Dataset Description

In this project, we worked with three different datasets, each providing a unique type of information related to Conservative Hybrid Mutual Funds. These datasets were collected from four major financial platforms: **MFAPI**, **Value Research**, **AMFI**, and **Moneycontrol**. Each dataset played an important role in understanding fund performance, NAV movement, and portfolio structure.

1.3.1 analytics.csv

This dataset contains fund-level performance and analytics information for all selected Conservative Hybrid Funds. It was created by downloading multiple filtered CSV files from Value Research and manually merging them into one combined file.

Important Columns

- **Fund Information:** Scheme Name, Plan, Category
- **Return Data:**
  - Short-term: 1 Week, 1 Month, 3 Months, 6 Months
  - Long-term: 1 Year, 3 Years, 5 Years, 10 Years, 20 Years
  - SIP Returns: 3-Year SIP, 5-Year SIP, 10-Year SIP
  - Annual Returns: Individual financial year returns
  - Quarterly & Monthly Returns: Q1, Q2, Q3, Q4 and major monthly returns
- **Risk Metrics:** Sharpe Ratio, Sortino Ratio, Alpha, Beta, Standard Deviation
- **Fund Attributes:** Expense Ratio, AUM, Market Cap Allocation, Risk Grade, Return Grade
- **Fund Manager Details**
- **NAV Summary:** Latest NAV, 52-Week High/Low

Purpose of the Dataset

This dataset helps in analysing performance, risk, and return behaviour across all funds in the Conservative Hybrid category.

1.3.2 nav_values.csv

This dataset includes daily NAV (Net Asset Value) history for each fund, collected from MFAPI, which provides date-wise NAV values for Indian mutual funds.

Important Characteristics

- Each row represents a fund scheme.
- Each column (after the scheme name) represents a specific trading day.
- Contains NAV values across hundreds of dates.
- Includes missing values for non-trading days or holidays.
- Required major preprocessing such as:
  - Identifying valid date columns
  - Cleaning invalid or blank NAV entries
  - Sorting dates
  - Removing holiday-only columns

Purpose of the Dataset

This dataset was used to study NAV trends, price movement patterns, and time-based fund performance.

1.3.3 portfolio_assets.csv

This dataset contains information about the portfolio composition of each fund. It was downloaded from Moneycontrol, where each scheme's portfolio page lists its asset allocation and holding details.

Important Columns

- **Asset Allocation:** Equity, Debt, Cash, MF Holdings, Other Holdings
- **Holding Structure:** Number of Equity Stocks, Number of Debt Instruments
- **Additional Attributes:** Turnover Ratio, Credit Rating, CRISIL Rating, Portfolio Maturity

Purpose of the Dataset

This dataset helps in understanding each fund's investment strategy, exposure to equity vs. debt, and the overall diversification and stability of the portfolio.

1.4 Packages required

To perform data cleaning, analysis, and visualization for this project, several Python libraries were used. These packages helped in reading the datasets, preprocessing them, handling missing values, performing exploratory data analysis, and creating graphical visualizations. The main packages used are listed below:

1. **pandas**
   Used for loading CSV files and performing all data manipulation tasks. Helpful for cleaning, merging, filtering, and transforming the datasets. Provides functions such as `read_csv()`, `dropna()`, `fillna()`, `groupby()`, and many more.
2. **numpy**
   Used for numerical operations and array-based calculations. Helps in handling missing values, converting columns to numeric types, and performing mathematical operations during analysis.
3. **matplotlib**
   Used for creating basic visualizations such as line charts, bar charts, scatter plots, and histograms. Very useful for plotting NAV trends and distribution graphs.
4. **seaborn**
   Built on top of matplotlib and used for advanced and visually appealing charts. Used for correlation heatmaps, boxplots, KDE plots, and category-wise comparisons.
5. **datetime**
   Used for handling date formats, converting strings to datetime objects, and sorting NAV values in chronological order.

6. **sklearn (Scikit-learn)**
   Used only for optional exploratory modeling. Provides functions for:

   - Train-test splitting
   - Linear regression
   - Feature scaling
   - Model performance evaluation

   (If the final report excludes the modeling section, this may be optional.)

7. **warnings**
   Used to suppress unnecessary warning messages during preprocessing so that the output looks clean.

8. **re (Regular Expressions)**
   Used during preprocessing to detect which column names in the NAV dataset represent valid dates. Helps identify and clean inconsistent formats.

Chapter 2. Data Cleaning

Data cleaning is one of the most important steps in any analytics project. Since our datasets were collected from different financial websites (MFAPI, Value Research, Moneycontrol), each file had different formats, missing values, inconsistent labels, and noisy data. To ensure accurate analysis, all three datasets were cleaned, standardised, and converted into usable formats before performing Exploratory Data Analysis (EDA). This chapter explains the complete cleaning procedure applied to the datasets `analytics.csv`, `nav_values.csv`, and `portfolio_assets.csv`.

2.1 Missing data analysis

Missing data is one of the most common challenges when working with real-world financial datasets. Since our information was collected from multiple external platforms such as MFAPI, Value Research, and Moneycontrol, each dataset contained missing values in different forms. Identifying and understanding these missing values was an important step before performing any further analysis.

This section describes the patterns of missing data present in the three datasets and the steps taken to handle them.

2.1.1 Identifying Missing Values

To detect missing values, the following methods were used:

- `isnull().sum()`
- `info()`
- Manual inspection for unusual characters such as "–", "N.A", or empty strings

Types of missing values found:

- Blank cells
- Cells containing "–" instead of numbers
- Missing NAV values for holidays or non-trading days
- Missing portfolio values not provided by Moneycontrol
- Missing risk metrics for some funds on Value Research

2.1.2 Missing Data in `analytics.csv`

The `analytics.csv` dataset contained missing values mainly in the following areas:

- **Risk Metrics:** Sharpe Ratio, Sortino Ratio, Beta, Alpha
- **SIP Returns:** Long-term SIP values missing for newer schemes
- **Annual Returns:** Gaps for recently launched funds
- **Fund Manager Information:** Some entries incomplete or not available

How we handled it:

- These missing values were not filled, because they reflect genuine unavailability.
- Leaving them as NaN keeps the dataset realistic and avoids wrong assumptions.

2.1.3 Missing Data in `nav_values.csv`

This dataset had the highest amount of missing values because it contains NAV data across hundreds of dates.

Reasons for missing values:

- Market holidays
- NAV not published on certain dates
- Newly launched funds with shorter NAV history
- Invalid columns generated during extraction

Treatment:

- Removed entire date columns where all values were missing
- Retained partially missing NAV values because they are a natural part of time-series data
- No interpolation was applied (NAV should not be artificially created)

2.1.4 Missing Data in `portfolio_assets.csv`

This dataset had missing values in the following areas:

- **Allocation Percentages:** Some funds did not disclose debt or cash allocation
- **Credit Quality and Ratings**
- **Number of Equity / Debt Instruments**
- **Turnover Ratio**

Treatment:

- Removed % symbols and converted valid numbers to numeric type
- Missing values were kept as NaN to avoid misrepresenting the fund's actual portfolio
- Text-based missing values were cleaned and standardized

2.2 Imputation

Imputation refers to the process of replacing missing or invalid values in a dataset with meaningful substitutes. However, in financial datasets—especially those involving mutual fund NAVs, risk metrics, and portfolio compositions—imputation must be approached very carefully to avoid distorting the results.

In this project, a selective and minimal imputation strategy was followed. Only values that could be safely estimated without affecting financial accuracy were imputed. For most fields, missing values were intentionally left as NaN to preserve the authenticity of the data.

2.2.1 Why Imputation Was Needed

A few fields contained missing or incorrect entries due to:

- Data extracted from multiple websites (Value Research, MFAPI, Moneycontrol)
- Missing NAV values on certain dates
- Portfolio fields not reported by the source
- Symbols like "–" used instead of numbers
- Newly launched schemes lacking return history

To ensure consistency, some controlled imputation was required.

2.2.2 Columns Where Imputation Was Applied

Only those fields were imputed where doing so would not affect the financial logic.

1. **Formatting Errors**
   Values like "–" or "—" were replaced with NaN. Empty strings were converted into proper null values.

2. **Numeric Conversion**
   Columns containing numbers with symbols (e.g., "15%") were cleaned and converted to numeric. Missing percentages or numeric fields were kept as NaN, but invalid characters were removed.

3. **Non-critical Portfolio Fields**
   Some portfolio fields such as:

   - Number of Stocks
   - Number of Debt Instruments

   contained text-like entries or blanks. These were converted to numeric, and blanks were set to NaN. No artificial values were inserted.

2.2.3 Columns Where Imputation Was NOT Performed

In most financial fields, imputation was intentionally avoided because filling missing values could create misleading results.

1. **NAV Values (`nav_values.csv`)**
   Missing NAV values occur on holidays, non-trading days, or launch days. Imputing NAVs would distort trend analysis, so values were kept untouched.

2. **Risk Metrics (`analytics.csv`)**
   Sharpe Ratio, Sortino Ratio, Alpha, Beta These metrics depend on historical performance. If missing, it means the value does not exist — hence no imputation.

3. **SIP Returns & Annual Returns**
   These values are missing due to insufficient historical NAV data. They were left as NaN to maintain accuracy.

4. **Portfolio Allocation (`portfolio_assets.csv`)**
   Missing debt or cash allocation does not imply "0". Missing values were therefore left unchanged.

2.2.4 Reason for Minimal Imputation

Imputation was used only where necessary because:

- Financial values should not be guessed or assumed
- NAV trends would become inaccurate if filled
- Risk metrics cannot be computed without full data
- Portfolio values missing from the source represent genuine missing data
- Incorrect imputation could lead to misleading EDA results

Thus, a **minimal and safe imputation approach** was adopted for the project.

Chapter 3. Visualization

Visualization is an essential part of Exploratory Data Analysis (EDA) because it helps convert raw numerical data into intuitive graphs and patterns. In this project, visualizations were used to understand the behaviour of Conservative Hybrid Mutual Funds in terms of their performance, NAV movement, risk patterns, and portfolio allocation. Different types of charts such as line plots, bar charts, histograms, heatmaps, and boxplots were used to uncover underlying insights. This chapter explains all major visualizations created during the analysis.

3.1 Univariate analysis

Univariate analysis focuses on examining one variable at a time to understand its individual behavior, distribution, and range. This step helps in identifying basic patterns, spotting irregularities, and understanding how each feature behaves independently before comparing it with other variables.

In the context of Conservative Hybrid Funds, univariate analysis was applied to performance metrics, NAV values, risk indicators, and portfolio allocations.

3.1.1 Analysis of Fund Returns

Return-related attributes from the `analytics.csv` dataset were analysed individually.

Variables Included

- 1 Month, 3 Month, 6 Month Returns
- 1 Year, 3 Year, 5 Year, 10 Year Returns
- SIP Returns (3Y, 5Y, 10Y)
- Annual Returns

Key Observations

- Short-term returns (1M–6M) showed high variability, indicating sensitivity to recent market conditions.
- Long-term returns (3Y–10Y) had smoother and stable distributions, which is expected for Conservative Hybrid Funds.
- SIP returns displayed consistent upward patterns, reflecting lower volatility in systematic investments.

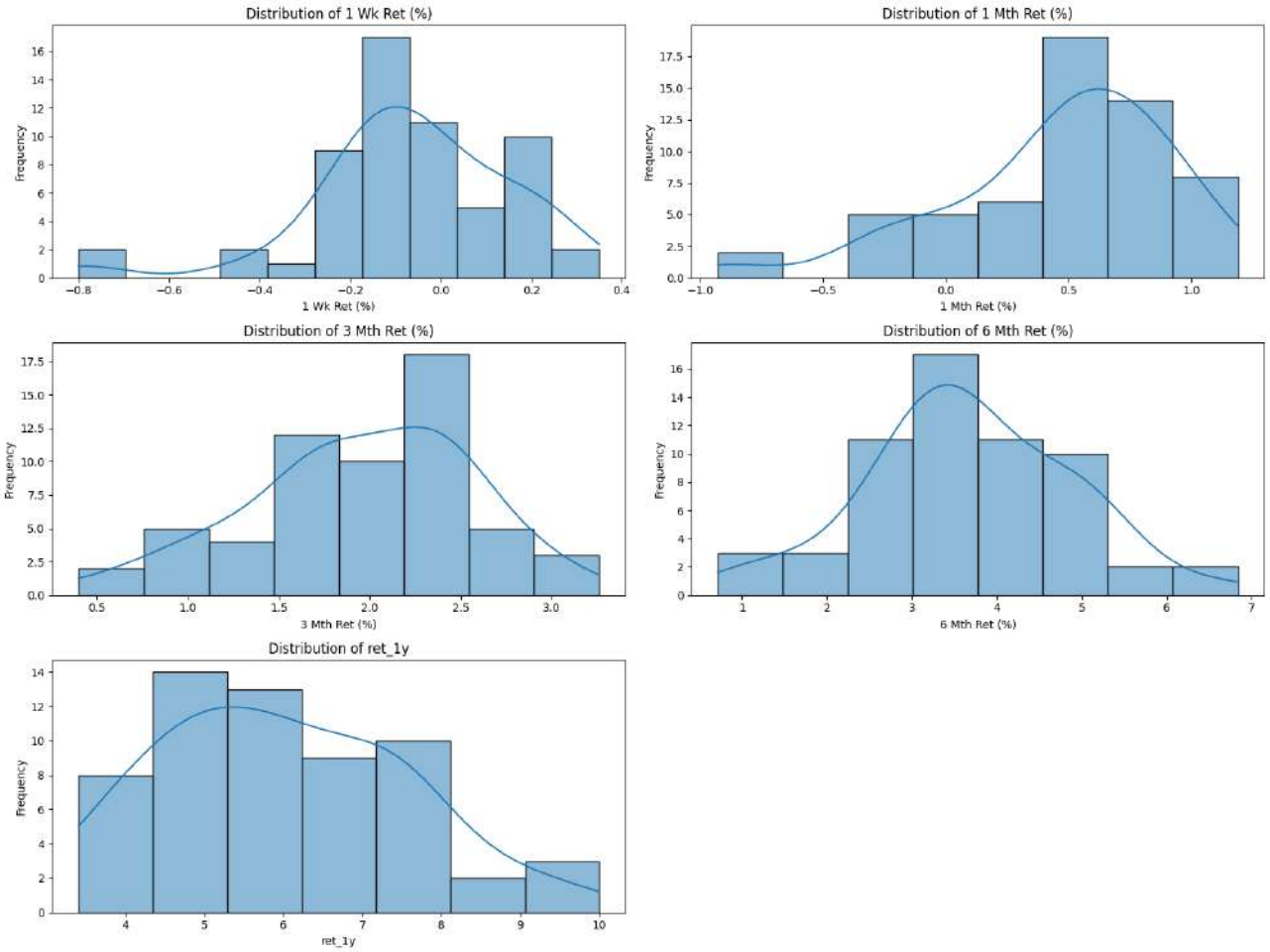


Figure 3.1: Distribution of Return for different time

3.1.2 Short-Term Return Boxplot

Boxplot of Short-Term Return Variables

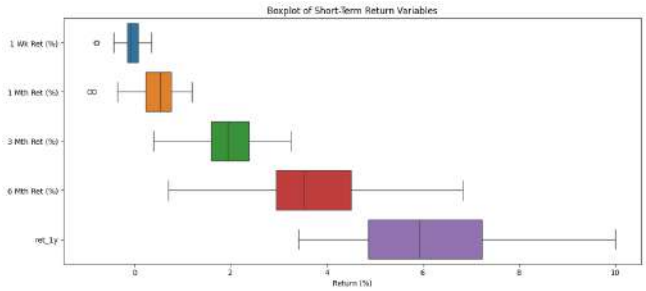


Figure 3.2: Box Plot of Short Term Return Variables

The boxplot clearly shows the spread and outliers for all short-term returns.

- 1-week and 1-month returns contain some negative outliers.
- The median increases as we move from 1 week to 1 year.

- This confirms that returns become more stable as the duration increases.

3.1.3 Long-Term Return Distributions

Histogram Plots of:

- 3-Year Return
- 5-Year Return
- 10-Year Return

Explanation:
Long-term returns show smooth and consistent distributions:

- 3-year and 5-year returns are tightly grouped around 8–11%.
- 10-year returns also show stability, with very few high outliers.
- Long-term return patterns prove that Conservative Hybrid Funds are steady and reliable for longer durations.

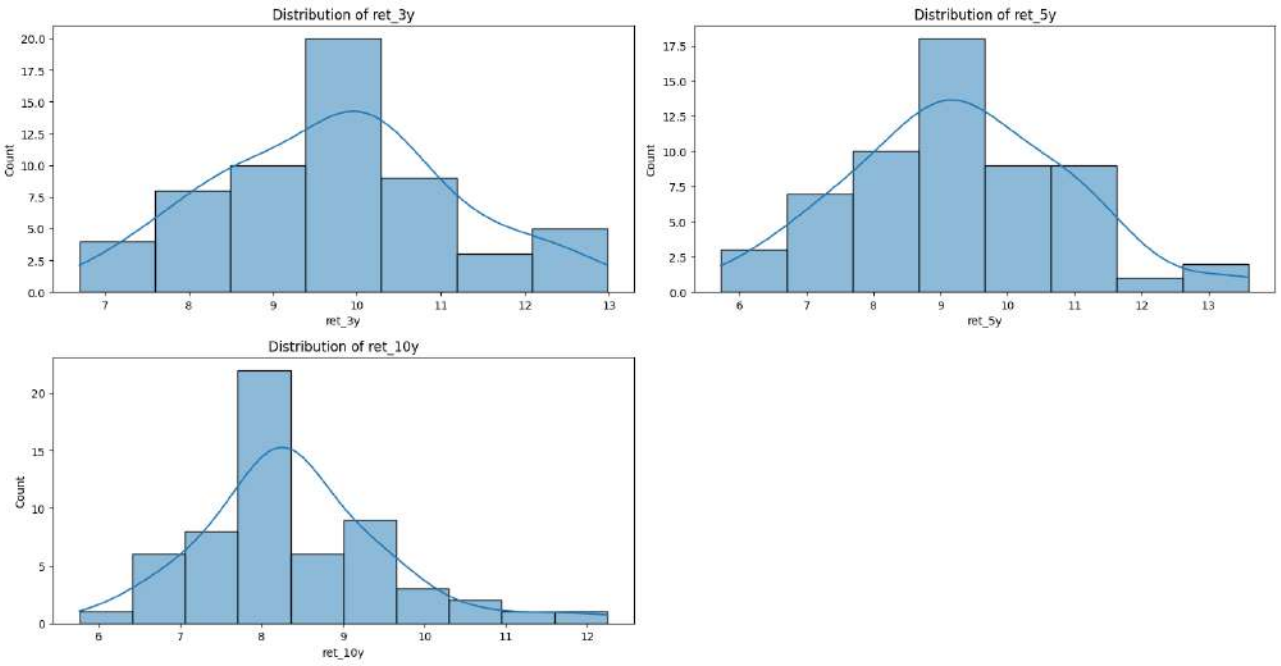


Figure 3.3: Distribution of return over long term

3.1.4 Long-Term Return Boxplot

Boxplot — Long-Term Returns

Explanation:

The boxplot shows how returns change over 3, 5, and 10 years:

- There are a few high outliers in the 10-year category.
- Overall, long-term returns have low spread, meaning lower risk and more consistency.

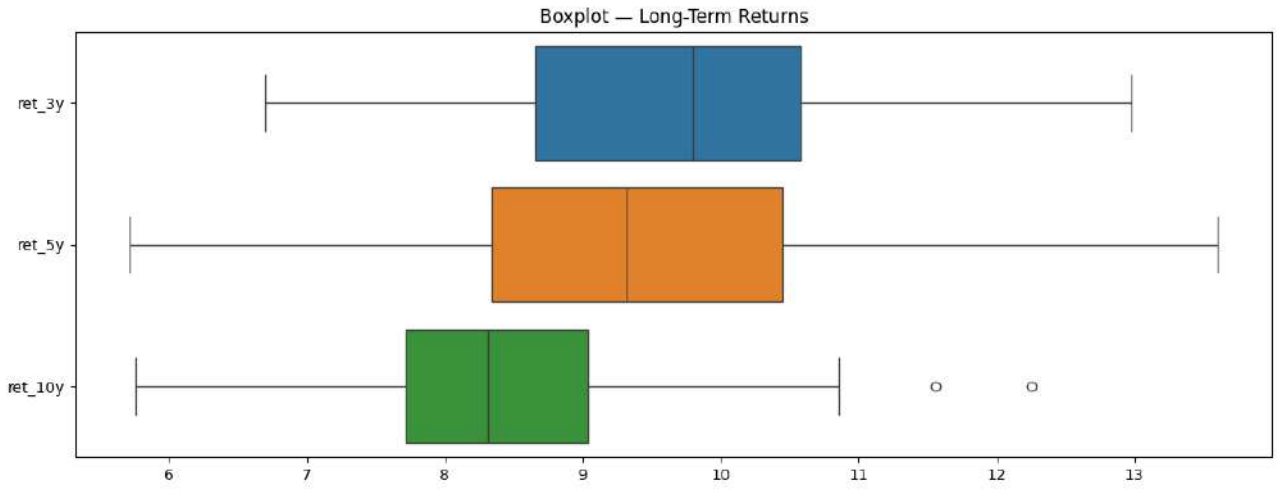


Figure 3.4: Boxplot of Long-Term Return Variables (3, 5, 10 Years)

3.1.5 Risk Metric Distributions

Histograms for:

- Standard Deviation
- Beta
- Sharpe Ratio
- Sortino Ratio
- Alpha
- Information Ratio

Explanation:

- Standard deviation is mostly between 3–4, meaning low volatility.
- Beta is close to 1 for most funds, showing low sensitivity to market movements.
- Sharpe and Sortino ratios are healthy for most schemes, showing good risk-adjusted returns.
- Alpha and Information Ratio show both positive and negative values, meaning performance varies across funds.

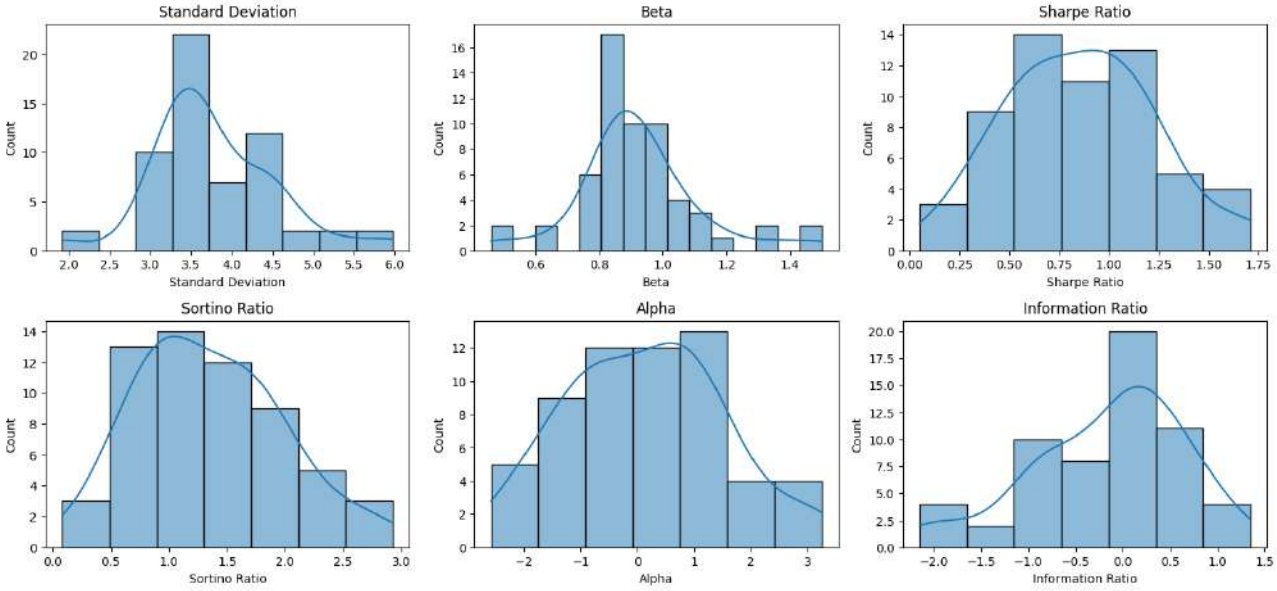


Figure 3.5: Risk Metric Distributions

3.1.6 Risk Metrics Boxplot

Risk Metrics — Boxplot

Explanation:

The boxplot shows outliers clearly in Alpha and Standard Deviation.

- Most risk metrics lie within a small range.
- This confirms that Conservative Hybrid Funds operate with low overall risk.

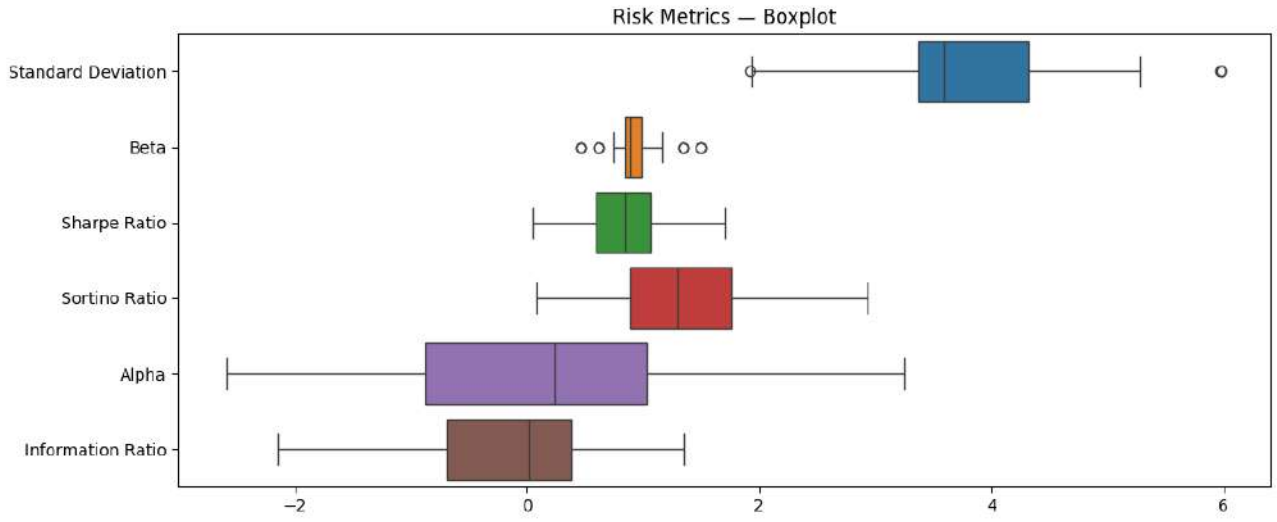


Figure 3.6: Boxplot of Risk Metrics (Alpha, Standard Deviation, etc.)

3.1.7 Summary of Univariate Analysis

- Short-term returns show small fluctuations with a few outliers.
- Long-term returns are stable and tightly grouped.
- Risk metrics indicate low volatility and controlled downside risk.
- Overall, Conservative Hybrid Funds behave consistently and predictably across all variables.

3.2 Bivariate Analysis

Bivariate analysis helps examine how two variables are related. Scatter plots and boxplots are used to understand relationships between returns, risk metrics, expenses, and fund characteristics.

3.2.1 Standard Deviation vs 1-Year Return

Std Dev vs 1-Year Return

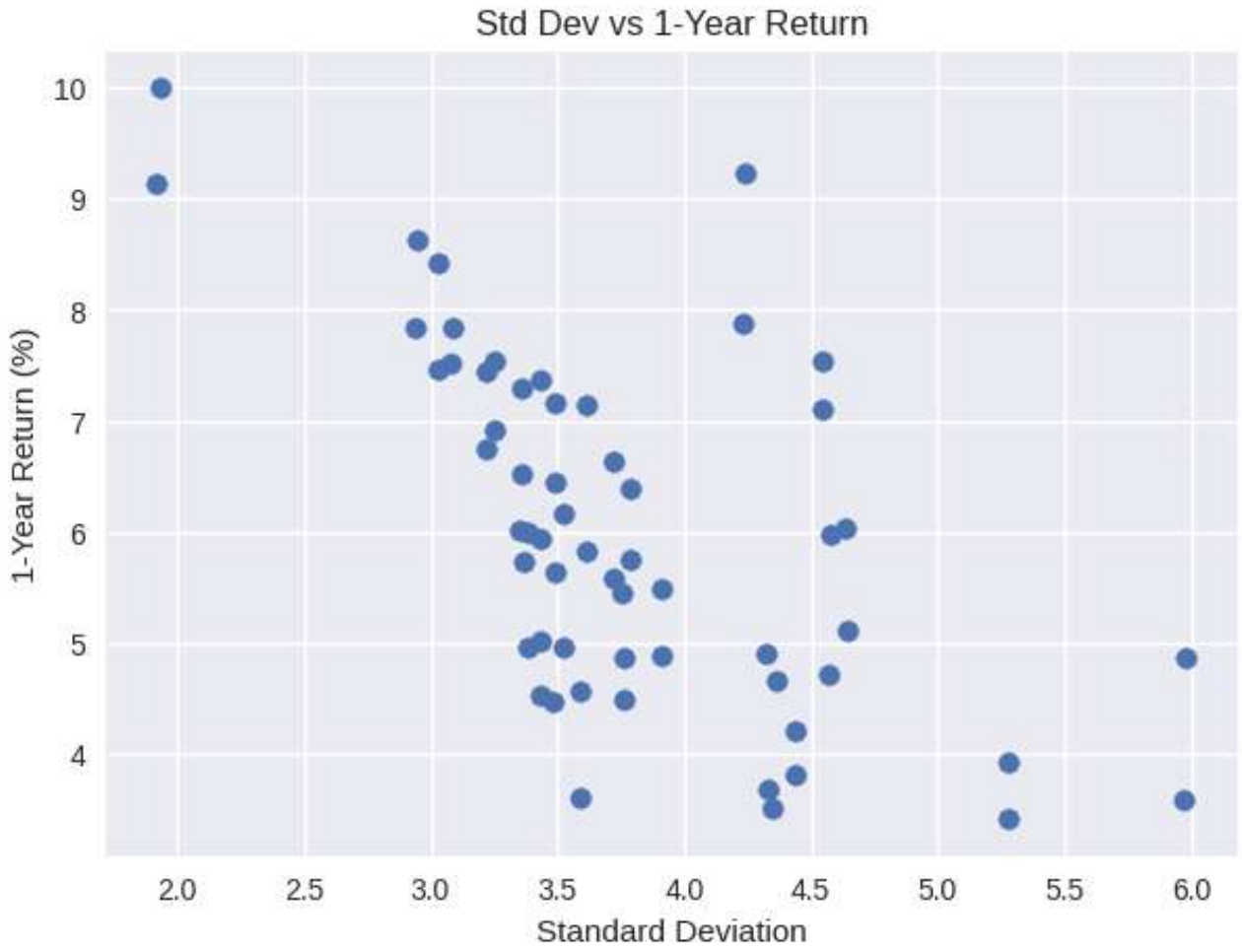


Figure 3.7: Standard Deviation vs 1-Year Return

Analysis:

- There is a clear negative relationship between Standard Deviation and 1-Year Return.
- Funds with lower volatility (Standard Deviation 2.5 – 3.5) tend to show higher returns.
- Higher volatility funds usually deliver lower 1-year returns.
- This confirms that in Conservative Hybrid Funds, lower risk often leads to better short-term outcomes.

3.2.2 Sharpe Ratio vs 3-Year Return

Sharpe Ratio vs 3-Year Return

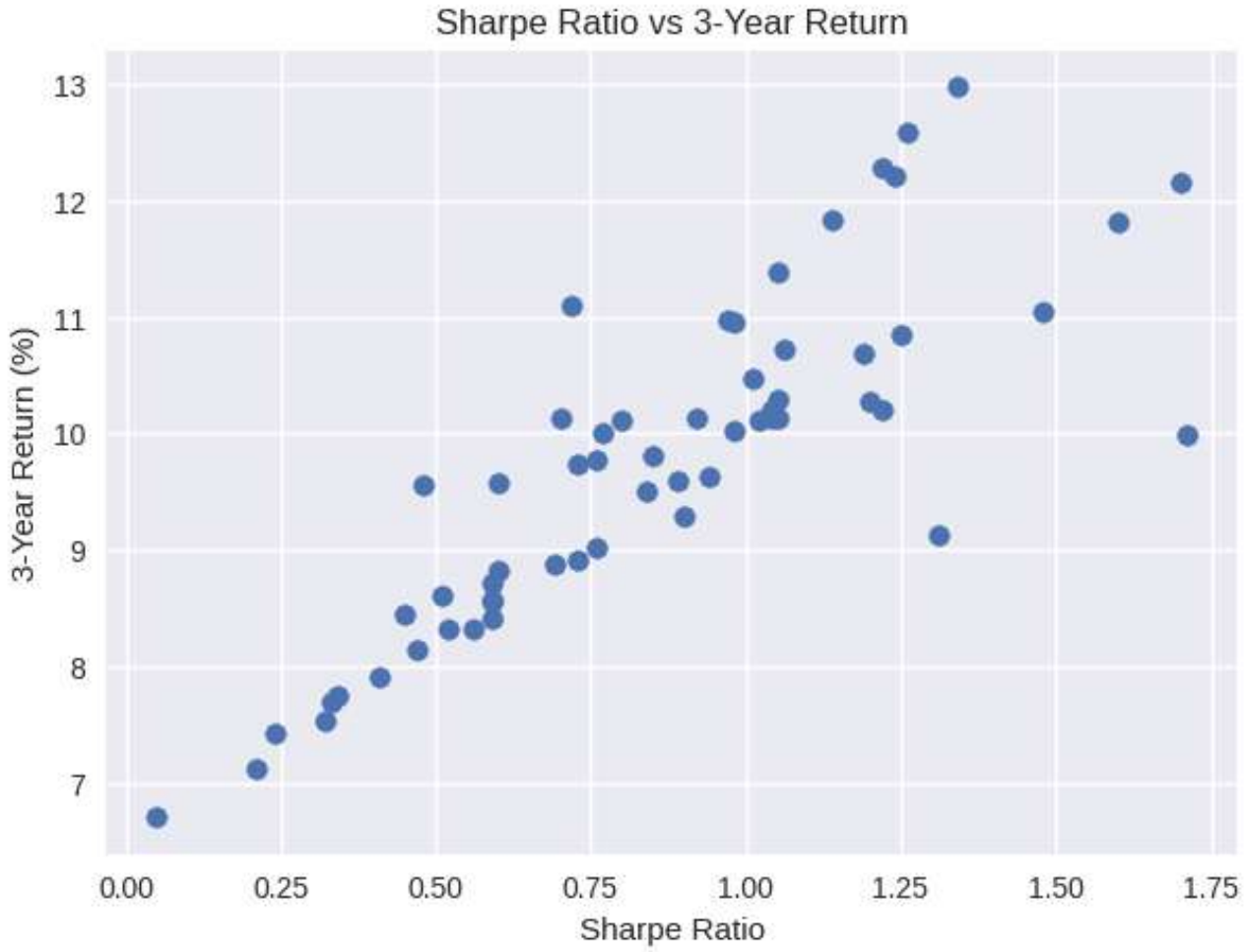


Figure 3.8: Sharpe Ratio vs 3-Year Return

Analysis:

- A strong positive correlation exists between Sharpe Ratio and 3-year returns.
- Higher Sharpe values (1.0 – 1.7) correspond to better long-term performance (10–13%).

- This indicates that funds with better risk-adjusted returns consistently perform better over the long term.

3.2.3 Beta vs 5-Year Return

Beta vs 5-Year Return

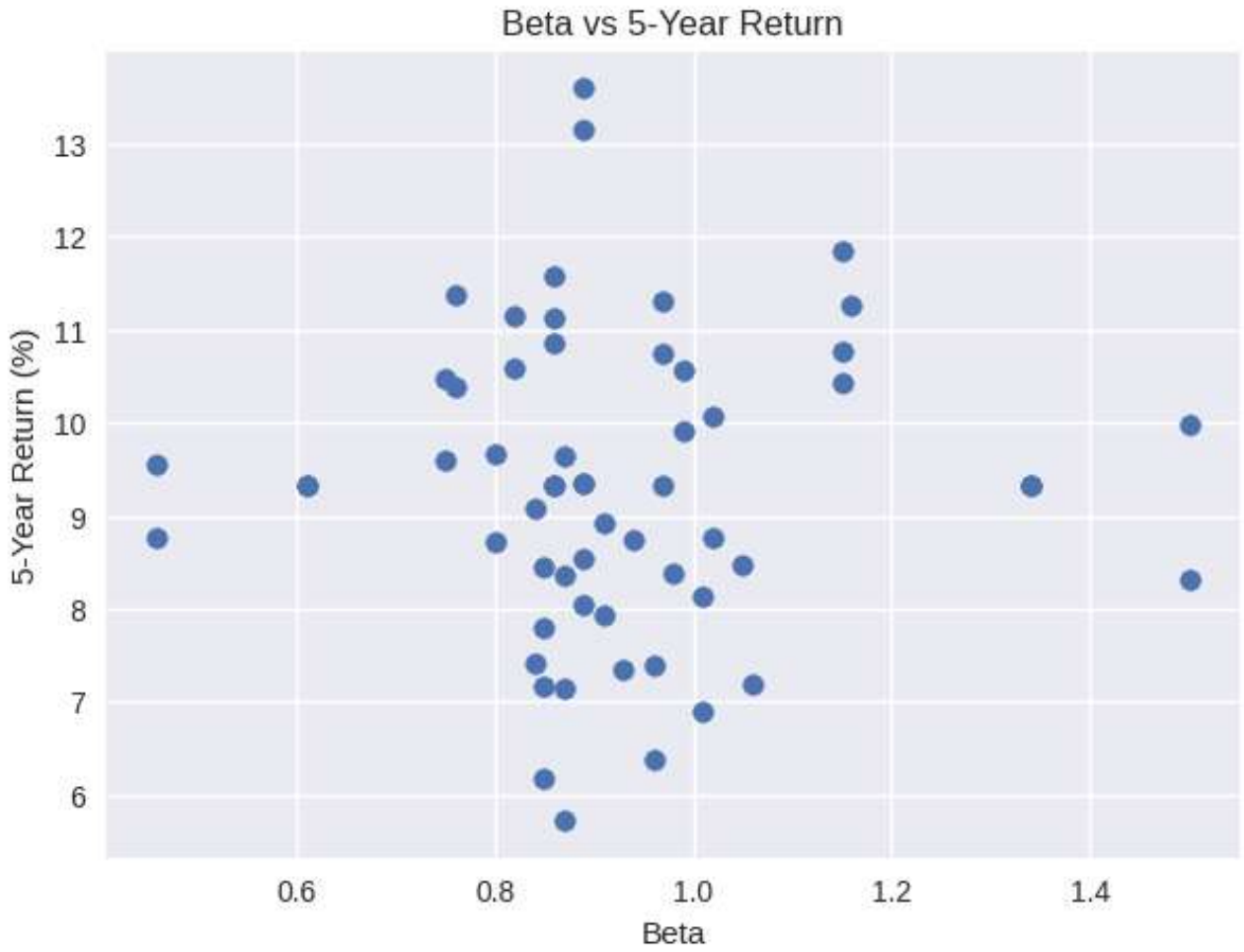


Figure 3.9: Beta vs 5-Year Return

Analysis:

- There is no strong relationship between Beta and 5-year returns.
- Funds with both low Beta (0.6) and high Beta (1.4) show similar return ranges.
- This indicates that Conservative Hybrid Funds are less dependent on market movement, so Beta does not heavily affect returns.

3.2.4 Category vs 1-Year Return

Category vs 1-Year Return (Boxplot)

Analysis:

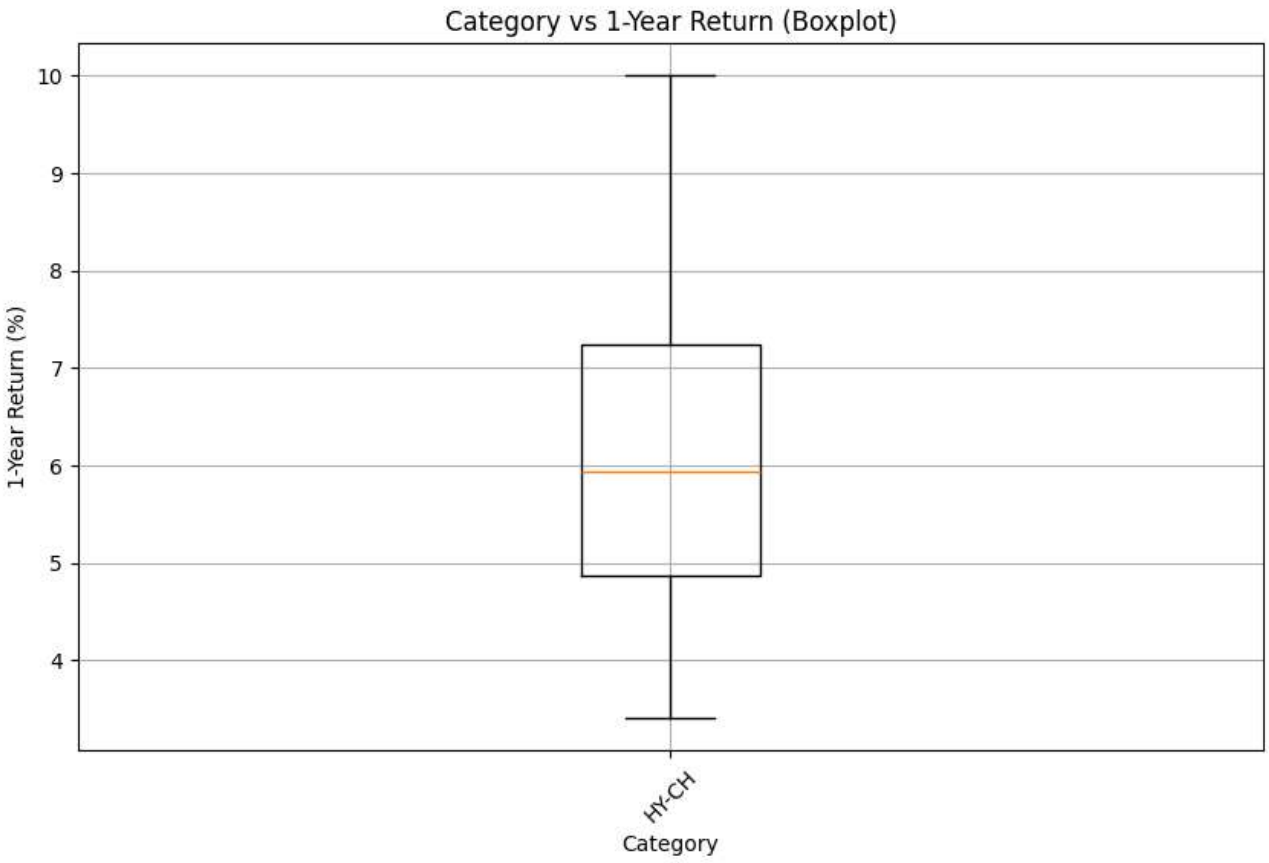


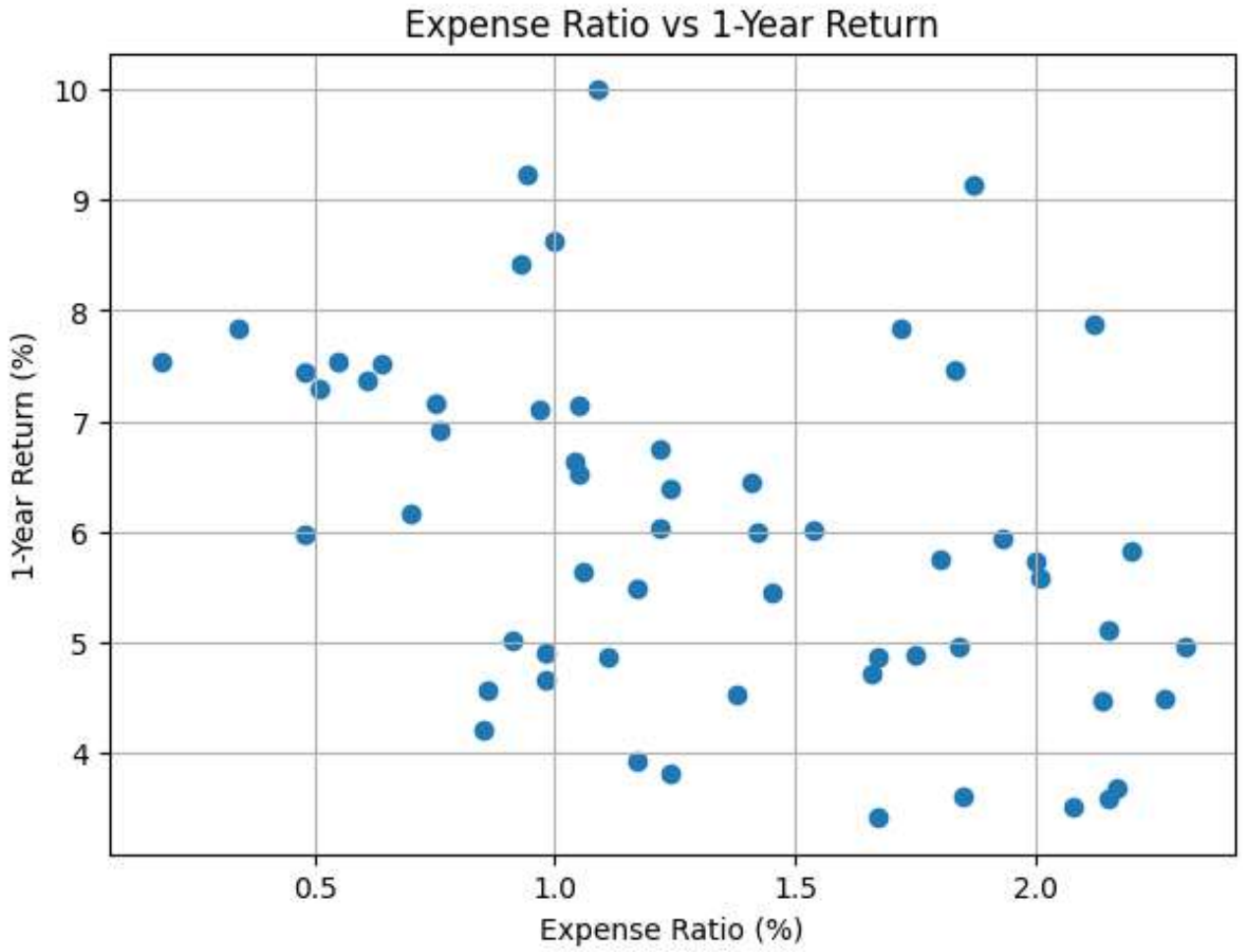
Figure 3.10: Category vs 1-Year Return (Boxplot)

- All schemes belong to the Conservative Hybrid category.
- Most 1-year returns lie between 5% and 7%, with a few outliers.
- This shows that the category provides stable returns with low variability.

3.2.5 Expense Ratio vs 1-Year Return

Expense Ratio vs 1-Year Return Analysis:

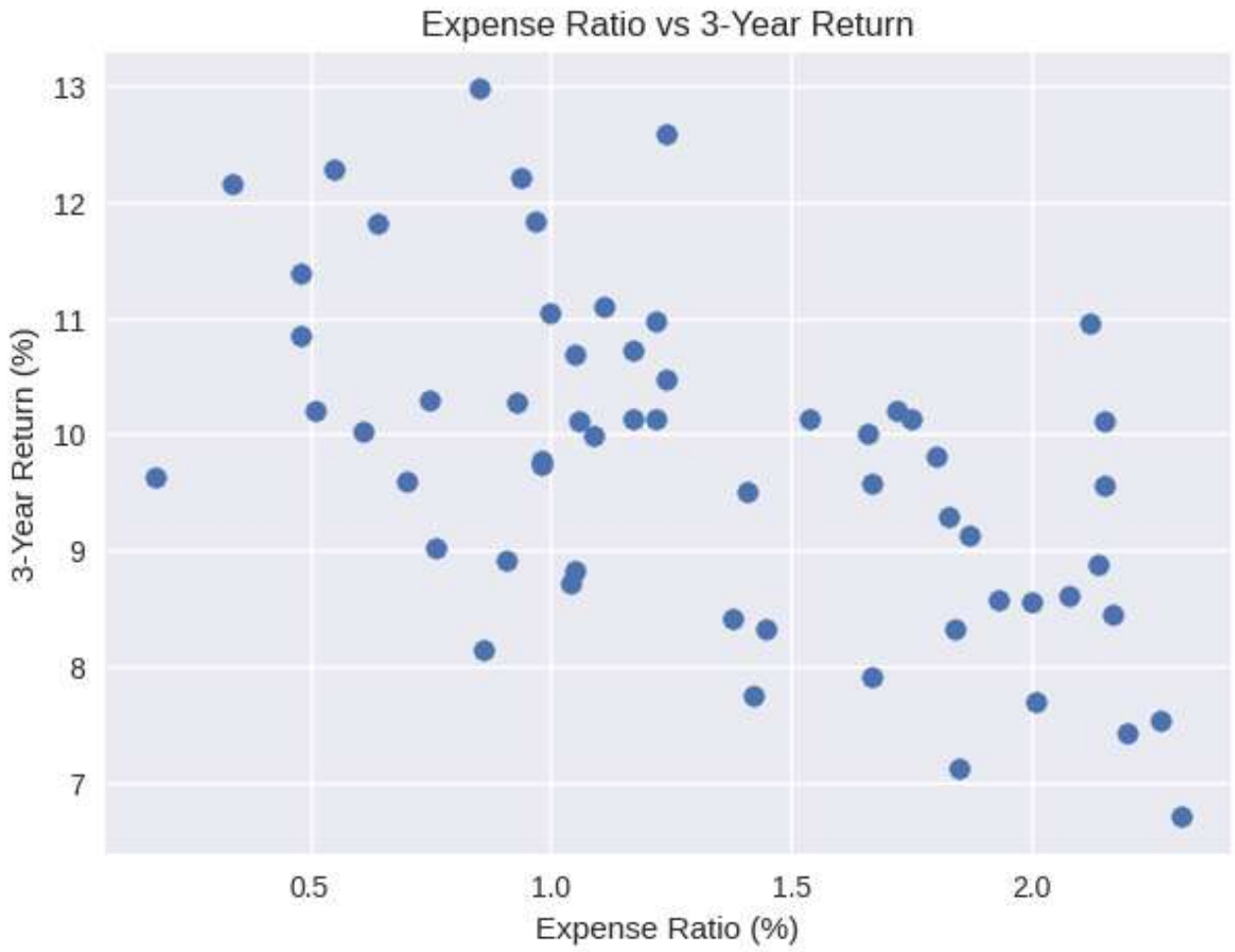
- A slightly negative trend is visible.
- Funds with lower expense ratios (0.4%–1.0%) achieve higher 1-year returns.
- Higher costs reduce investor profits, confirming that lower expense funds perform better.



Expense Ratio vs 1-Year Return

3.2.6 Expense Ratio vs 3-Year Return

Expense Ratio vs 3-Year Return



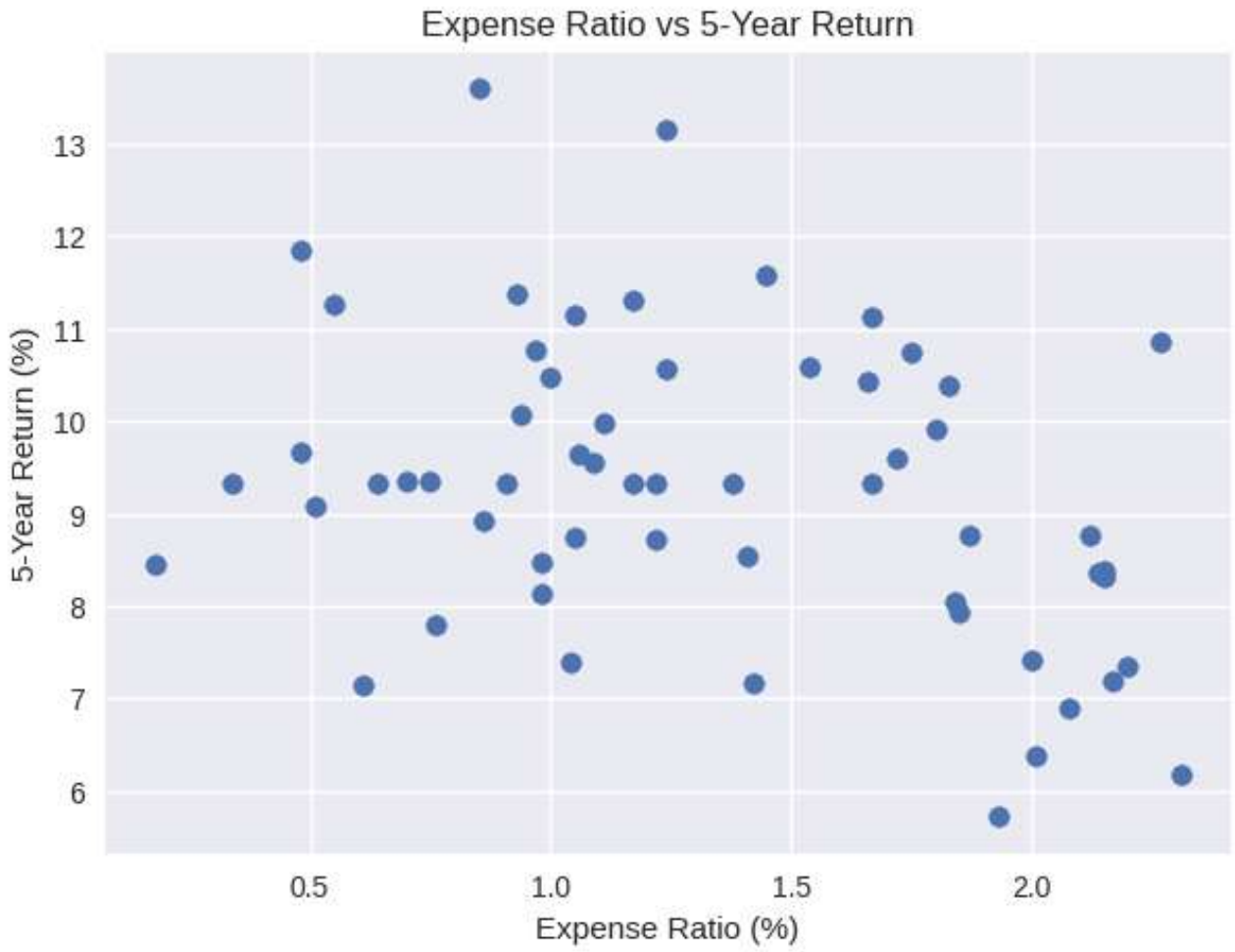
Expense Ratio vs 3-Year Return

Analysis:

- The negative impact continues for longer periods.
- Funds with lower expense ratios show superior 3-year performance.
- Expense ratio is an important factor affecting long-term returns.

3.2.7 Expense Ratio vs 5-Year Return

Expense Ratio vs 5-Year Return



Expense Ratio vs 5-Year Return

Analysis:

- Same pattern as 1-year and 3-year returns.
- Higher expense ratio funds underperform in the long run.
- This proves that cost efficiency is very important for hybrid fund performance.

3.2.8 NAV Volatility vs Riskometer Level

NAV Volatility vs Riskometer Level (Boxplot)

Analysis:

- Higher risk levels correspond to higher NAV volatility.
- ”High” rated funds show a wider spread and more outliers.
- ”Moderate” rated funds show the least volatility.
- This matches expectations and SEBI risk ratings.

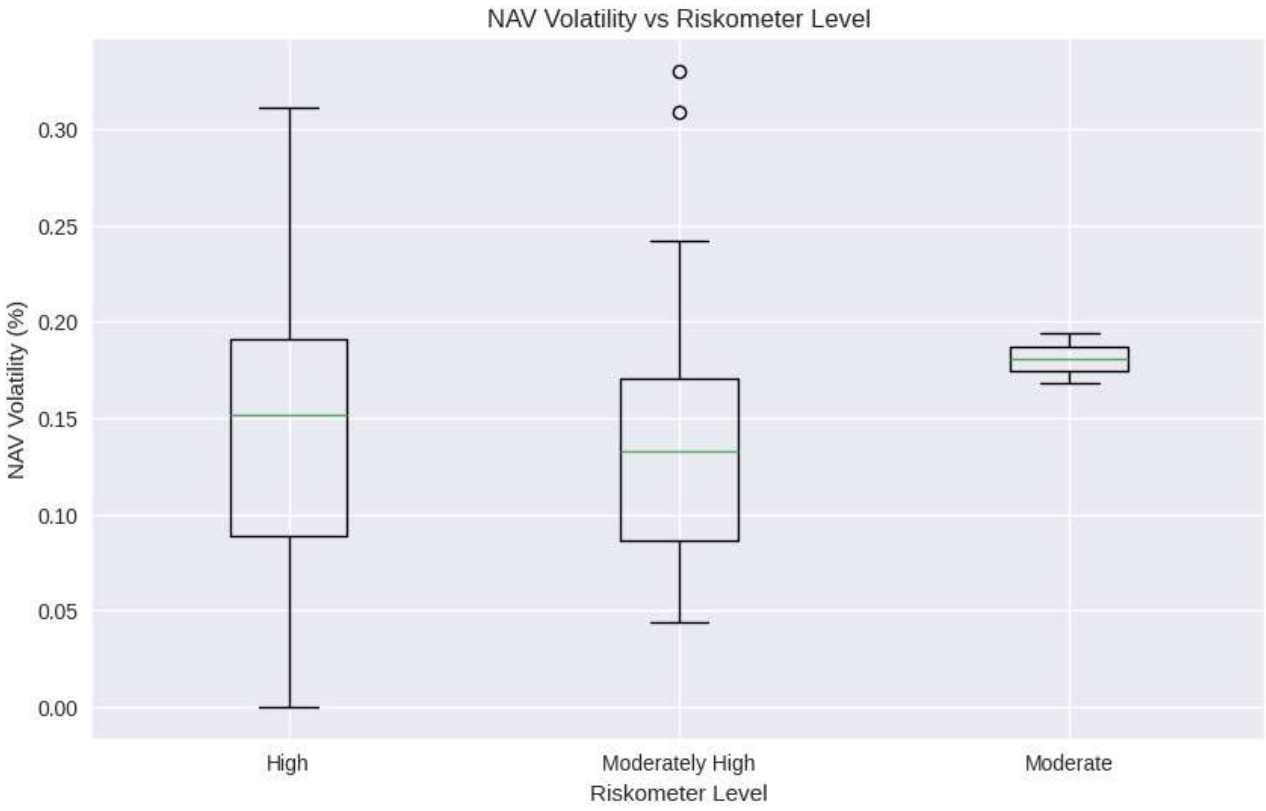


Figure 3.11: NAV Volatility vs Riskometer Level (Boxplot)

3.2.9 Market Cap vs 1-Year Return

Market Cap vs 1-Year Return

Market Cap vs 1-Year Return

Analysis:

- No strong relationship exists between fund size (Market Cap) and returns.
- Both small and large funds perform within a similar range.
- This means AUM/size does not strongly influence performance in Conservative Hybrid Funds.

3.3 Multivariate Analysis

Multivariate analysis is a statistical technique used to study the relationship between more than two variables at the same time. While univariate analysis looks at one variable and bivariate analysis examines two variables, multivariate analysis helps identify combined effects, patterns, and interactions across multiple factors.

It is especially useful in mutual fund analysis because fund performance depends on many variables together, such as returns, volatility, Sharpe ratio, expense ratio, asset allocation, and market conditions.

3.3.1 Common Multivariate Techniques

- **Correlation Matrix:** Shows how multiple variables are related to each other.
- **Multiple Regression:** Examines how several independent variables influence one dependent variable.
- **Principal Component Analysis (PCA):** Reduces many variables into a smaller set of important components.
- **Cluster Analysis:** Groups similar funds based on multiple characteristics.
- **Heatmaps:** Visual representation of correlations between variables.

In summary, multivariate analysis helps in understanding the overall behavior of mutual funds by analyzing several attributes simultaneously, providing deeper insights compared to univariate or bivariate approaches.

3.3.2 Correlation Matrix

The correlation matrix is a multivariate analysis technique used to understand how multiple variables are related to each other at the same time. In this project, correlation heatmaps were created for:

- Return metrics
- Risk metrics
- Cross-correlation between risk and return variables

These heatmaps help identify patterns, strong relationships, weak relationships, and negative correlations across the fund dataset.

Correlation Heatmap — Return Metrics

Description:

- Short-term returns (1 week, 1 month, 3 months) show moderate to strong positive correlation among each other (0.63–0.72).
- 3-month, 6-month, and 1-year returns have strong correlation (0.57–0.73), indicating consistent performance patterns.
- Long-term returns (3-year, 5-year, 10-year) have very strong correlation (0.83–0.93), meaning long-term performance is stable across funds.
- SIP returns also strongly correlate (0.75–0.89) with long-term lump-sum returns.
- Weekly and monthly returns have weak or negative relationship with long-term returns, showing that short-term fluctuations do not impact long-term performance significantly.

Conclusion: Short-term metrics behave differently, while long-term returns move together strongly. This indicates stable long-term performance in Conservative Hybrid Funds.

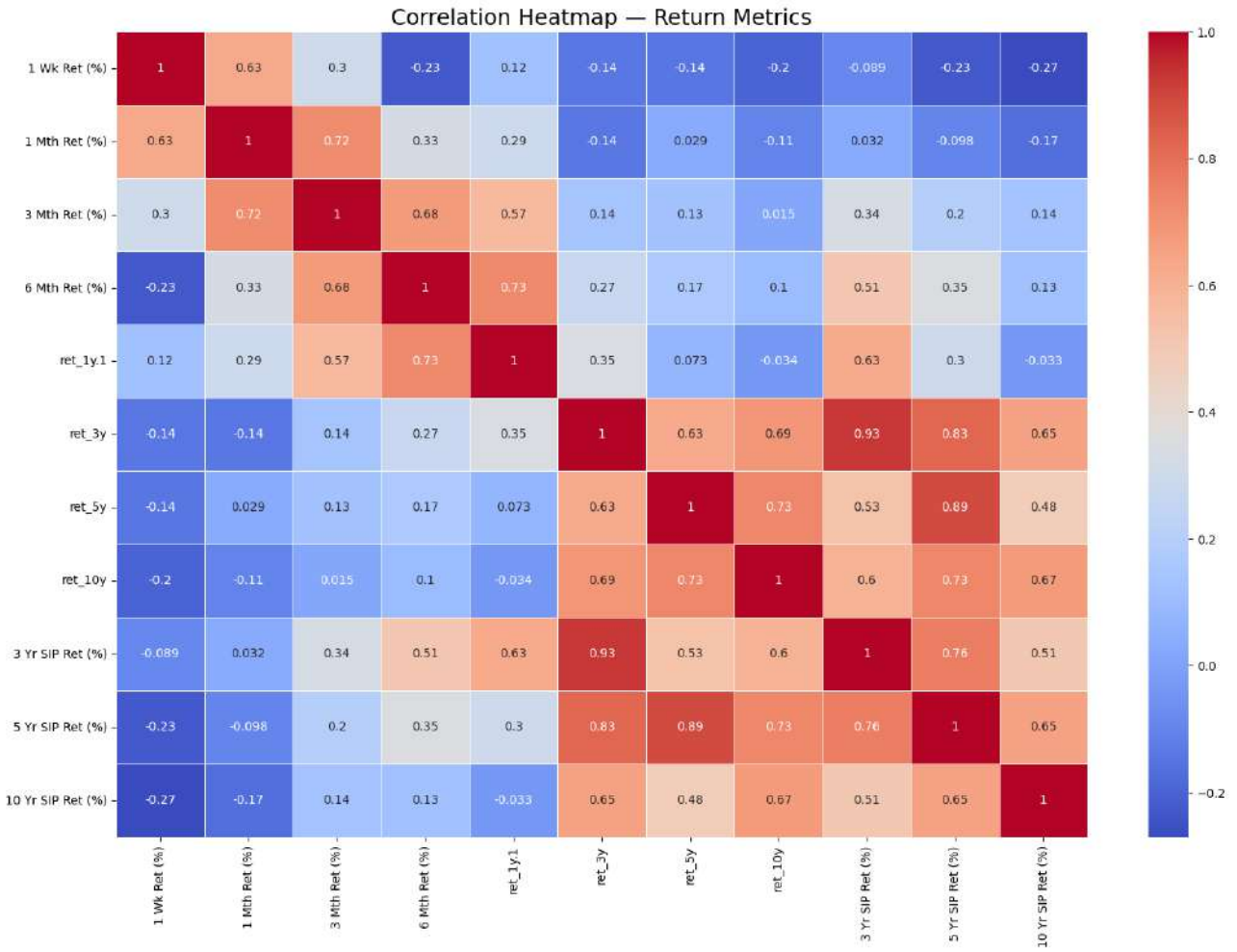


Figure 3.12: Correlation Heatmap — Return Metrics

Correlation Heatmap — Risk Metrics

Description:

- Sharpe Ratio, Sortino Ratio, and Alpha show very strong positive correlation (¿0.90). Funds with good risk-adjusted returns also generate higher Alpha.
- Standard Deviation and Beta also have strong positive correlation (0.94). Higher Beta funds naturally show higher volatility.
- Sharpe and Sortino ratios show negative correlation with Standard Deviation (-0.29 to -0.37), indicating that lower volatility leads to better risk-adjusted performance.
- R-Squared has moderate connections with other risk metrics, meaning it behaves independently.

Conclusion: Risk metrics form two clusters:

- **Volatility cluster** → Standard Deviation & Beta
- **Performance cluster** → Sharpe, Sortino, Alpha, Information Ratio

This clean separation is typical in hybrid mutual funds.

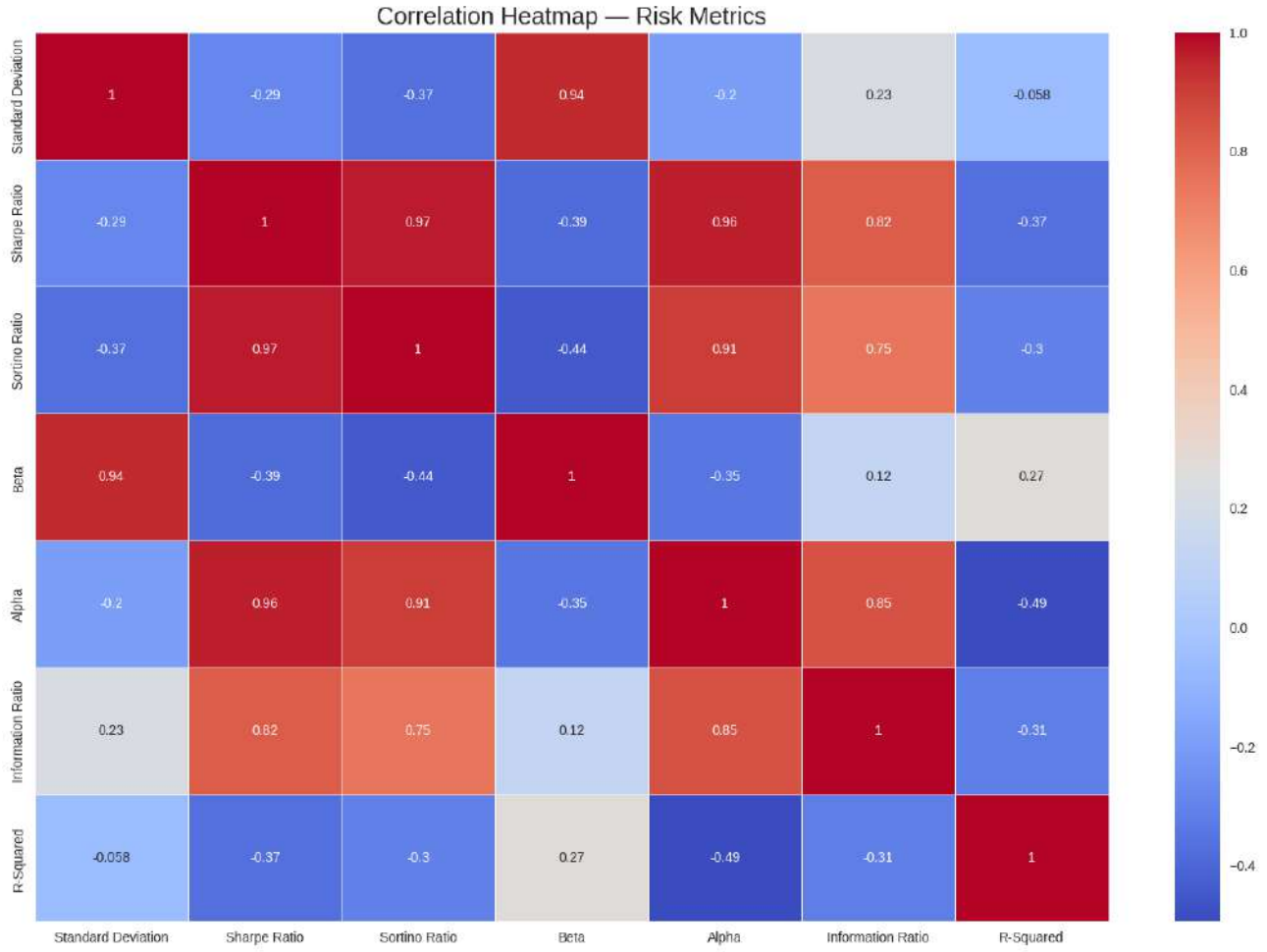


Figure 3.13: Correlation Heatmap — Risk Metrics

Cross-Correlation Heatmap — Risk Metrics vs Return Metrics

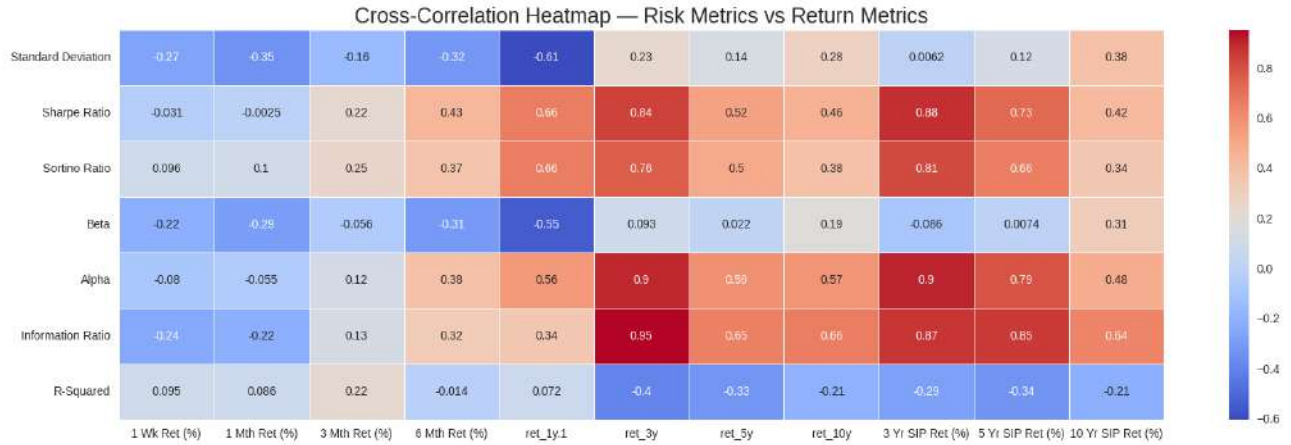


Figure 3.14: Cross-Correlation Heatmap — Risk Metrics vs Return Metrics

Description:

- Sharpe Ratio, Sortino Ratio, Alpha, and Information Ratio have strong positive correlation (0.66–0.95) with long-term returns (3Y, 5Y, 10Y). Better risk-adjusted funds consistently perform better.
- Standard Deviation and Beta show negative correlation with long-term returns (-0.31 to -0.61). More volatile funds tend to perform worse.
- Short-term returns (1W, 1M, 3M) show very weak correlation with risk metrics, meaning day-to-day fluctuations do not reflect risk behavior.

3.3.3 Multiple Regression

Multiple Regression is a multivariate statistical technique used to study how two or more independent variables jointly influence a single dependent variable. It extends simple linear regression by allowing multiple predictors, making it useful for analyzing complex financial datasets.

The general form is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{3.1}$$

Where:

- Y = dependent variable (example: 5-Year Return)
- $X_1, X_2, \ldots, X_n$ = independent variables (example: Standard Deviation, AUM, Plan Type, Tenure)
- β = coefficients that show how strongly each variable affects Y
- ϵ = error term

Why it is used in Mutual Fund Analysis

- To understand what factors influence returns
- To compare risk-return trade-offs
- To check whether AUM, plan type, or risk levels impact performance
- To identify significant predictors of long-term returns

Multiple regression helps create a clearer picture of how risk, size, and fund characteristics influence performance beyond simple pairwise relationships.

Bubble Regression Visualizations

The following bubble charts visually represent the relationship between risk (Standard Deviation) and returns (5-Year Return) while including additional variables such as AUM, Riskometer Level, Max Tenure, and Plan Type. These act as multivariate regression-style visual tools.

Description:

- Volatility (Std Dev) influences 5-year returns, bubble size represents AUM.
- Returns generally decline as Standard Deviation increases.

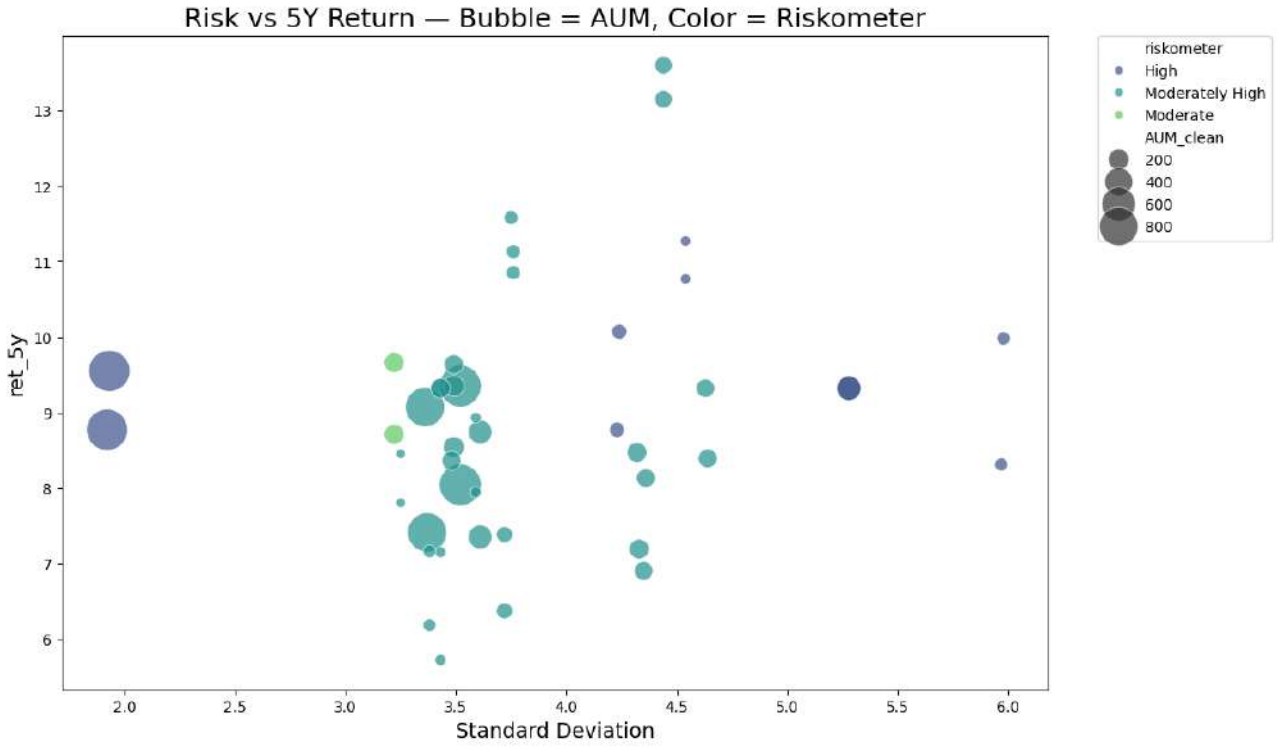
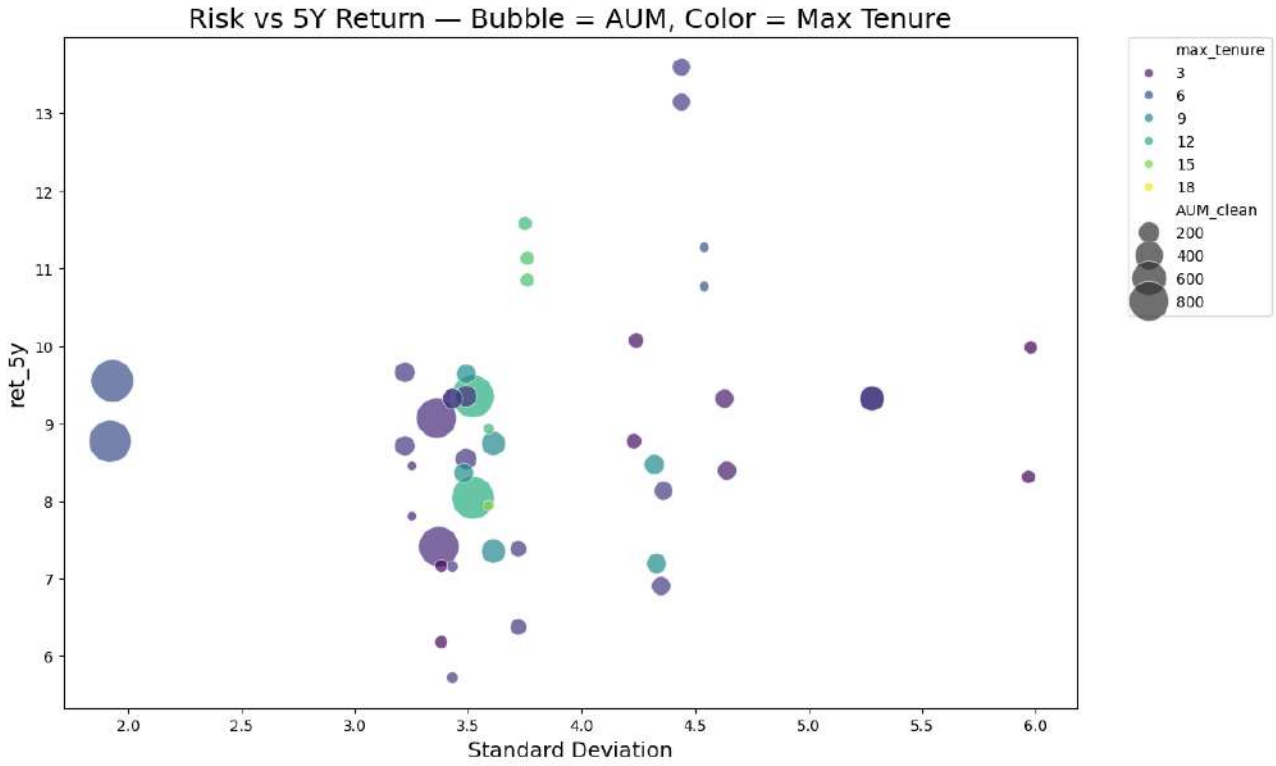


Figure 3.15: Risk vs 5-Year Return — Bubble = AUM, Color = Riskometer

- Larger funds (big bubbles) are mostly found in the Moderately High risk level.
- High-risk funds show more scattered performance and inconsistent returns.
- Moderate-risk funds cluster in the stable return zone (8%–10%).

Interpretation: Riskometer category and AUM both add explanatory power and behave like additional predictors in a multiple-regression setting.

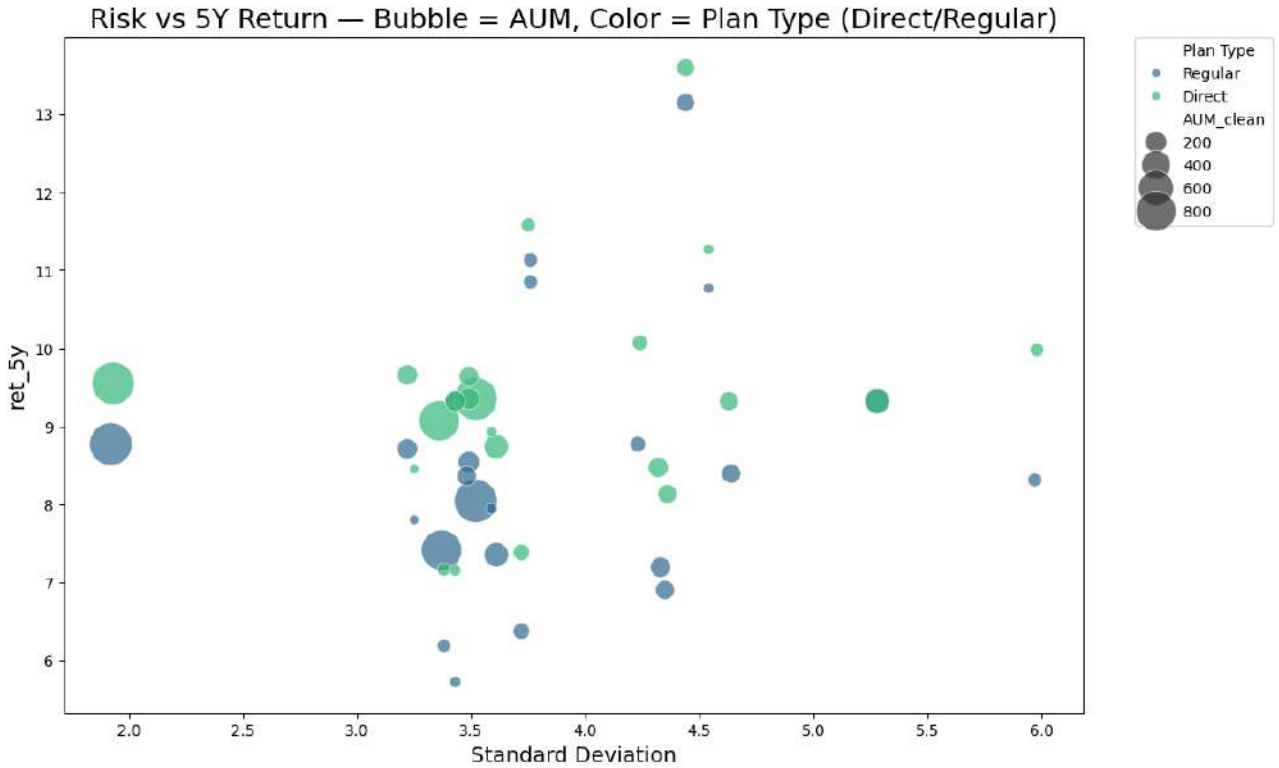


Risk vs 5-Year Return — Bubble = AUM, Color = Max Tenure

Description:

- Color gradient reflects maximum scheme age (tenure).
- Older funds appear slightly higher in return range.
- Younger funds show greater spread, suggesting less historical stability.
- AUM size does not dominate performance, but older funds with moderate risk tend to be more stable.

Interpretation: Tenure acts as a positive predictor of long-term return when combined with risk measures.



Risk vs 5-Year Return — Bubble = AUM, Color = Plan Type

Description:

- Direct plans (green) generally lie slightly above Regular plans for the same risk level.
- Regular plans (blue) cluster slightly lower for similar Standard Deviation values.
- AUM for Direct plans is usually smaller, but returns tend to be higher.

Interpretation: Plan type is a significant predictor of returns — Direct plans outperform, even after adjusting for risk and size.

3.3.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality-reduction technique that helps simplify large multivariate datasets by converting the original variables into a smaller set of new variables called Principal Components (PCs).

Why PCA is Useful

- Removes noise and multicollinearity.
- Helps visualize high-dimensional data in 2D.
- Identifies hidden patterns in data.
- Shows which combinations of variables explain most variation.

How PCA Works (Simple Explanation)

- PCA looks at all variables together.
- It finds combinations of variables that have maximum variance.
- PC1 explains the highest variance.
- PC2 explains the next highest, and so on.
- These components help summarize the entire dataset using fewer dimensions.

PCA on Return Metrics

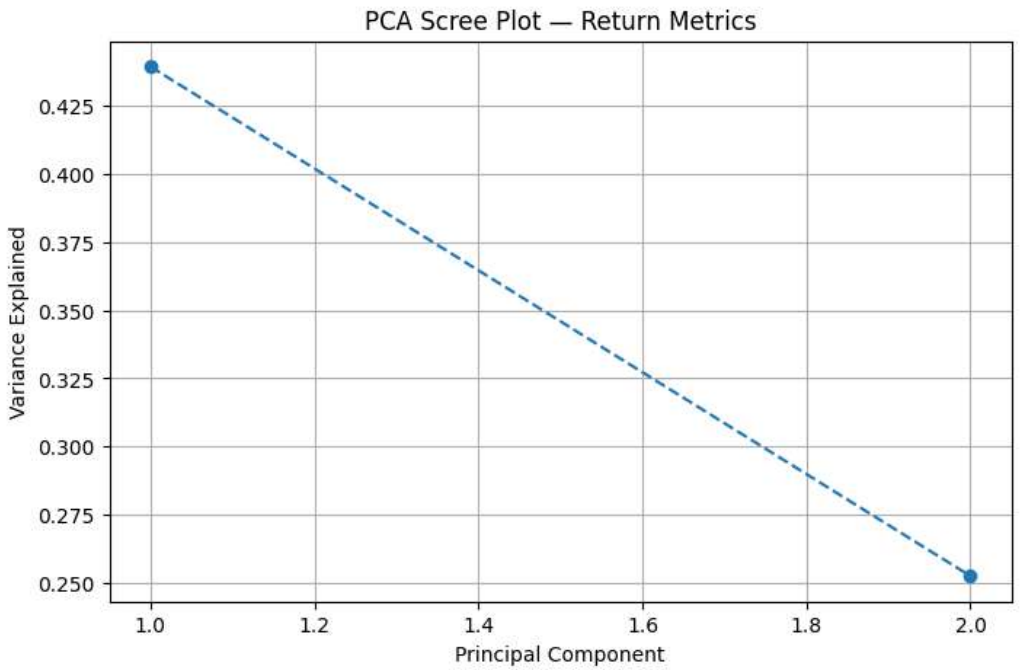


Figure 3.16: PCA Scree Plot — Return Metrics

1. PCA Scree Plot — Return Metrics Description:

- PC1 explains around 44% of total variance.
- PC2 explains around 25% variance.
- Together, PC1 + PC2 capture almost 70% of the total information from return variables.
- After PC2, the drop is sharp, meaning additional components are less useful.

Interpretation: Two components are enough to summarize the majority of the variation in short-term and long-term returns.

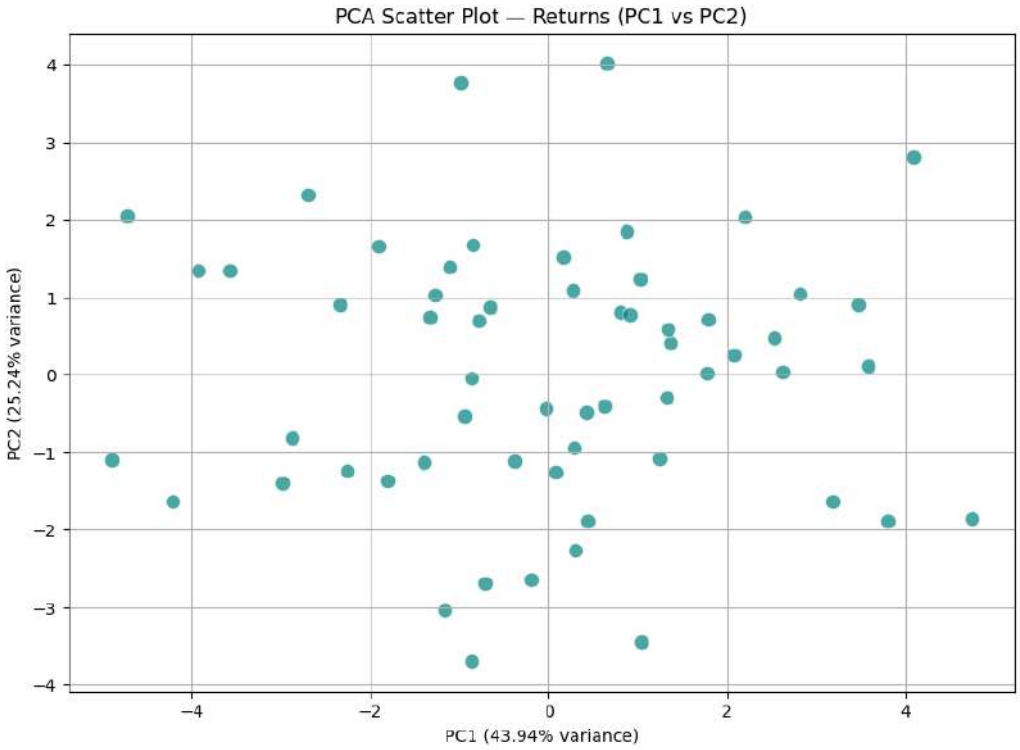


Figure 3.17: PCA Scatter Plot — Returns

2. PCA Scatter Plot — Returns (PC1 vs PC2) Description:

- Each point represents a fund plotted using PC1 and PC2 scores.
- Funds are spread out widely, indicating significant variation across schemes.
- Some grouping appears based on long-term/short-term return patterns.
- Funds on the right side have higher loadings on return-related variables.

Interpretation: Return profiles differ across funds — PCA helps visualize this variation clearly in 2D space.

PCA on Portfolio Composition

3. Ternary Plot — Equity / Debt / Cash Allocation Description:

- Most points are concentrated near the Debt corner, showing majority debt allocation.
- Equity and Cash proportions remain small as expected.
- A few schemes differ slightly, showing moderate deviation in allocation.

Interpretation: Portfolio compositions are quite similar, with debt dominating the mix, aligning with SEBI's definition of Conservative Hybrid Funds.

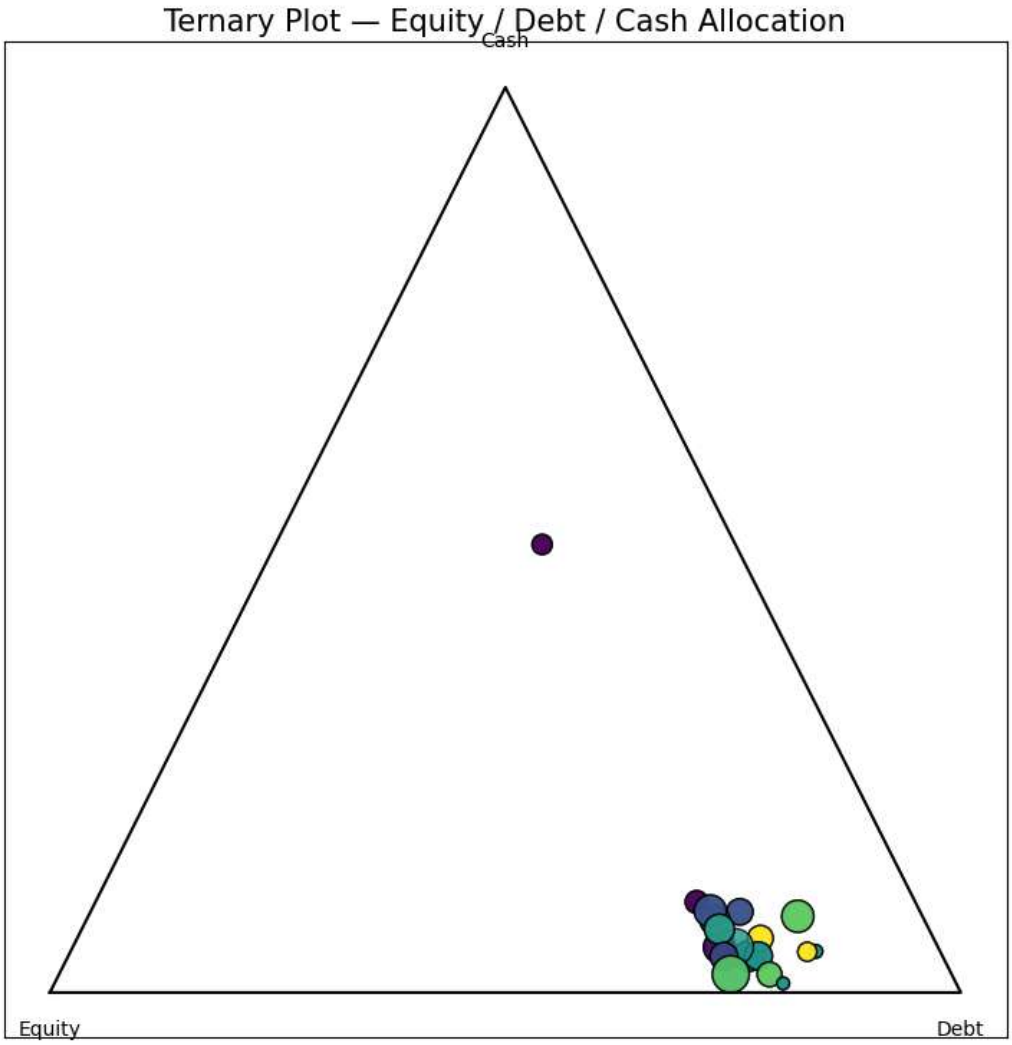


Figure 3.18: Ternary Plot — Equity/Debt/Cash Allocation

4. PCA Scree Plot — Portfolio Composition Description:

- PC1 explains about 44% of variance.
- PC2 explains about 31%.
- Combined, the first two components capture 75% of portfolio variation.
- Clear steep drop after PC2 → only two components are essential.

Interpretation: Portfolio allocations can be summarized very well using just two principal components.

5. PCA Scatter Plot — Portfolio Composition (Colored by CRISIL Rating) Description:

- Points (funds) are plotted based on their PC1 and PC2 values.
- Colour represents CRISIL ratings (1 to 5).

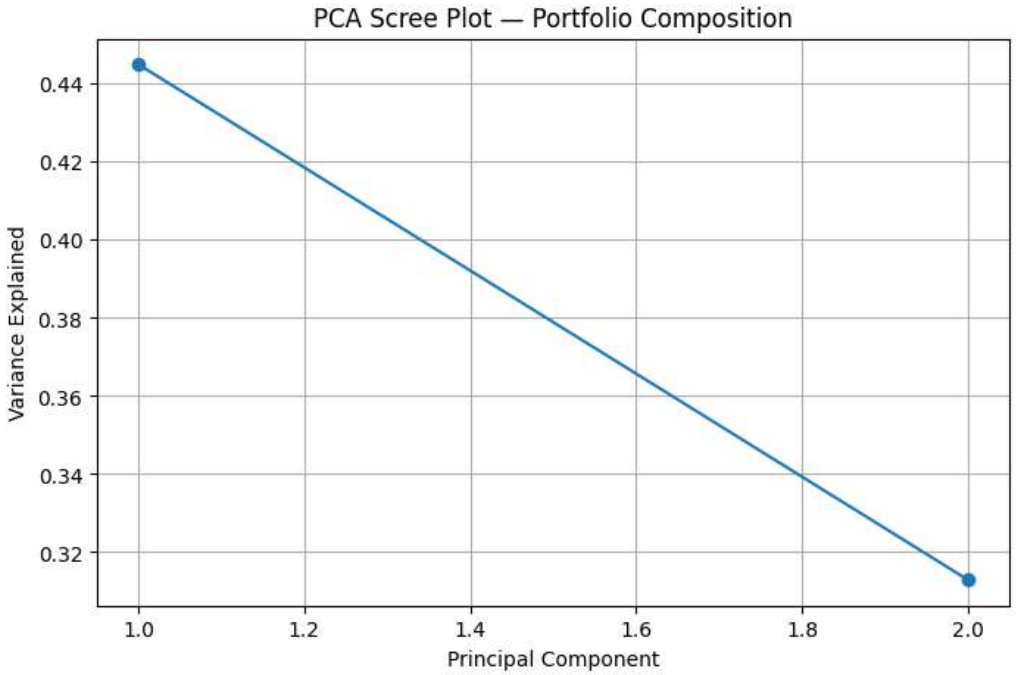


Figure 3.19: PCA Scree Plot — Portfolio Composition

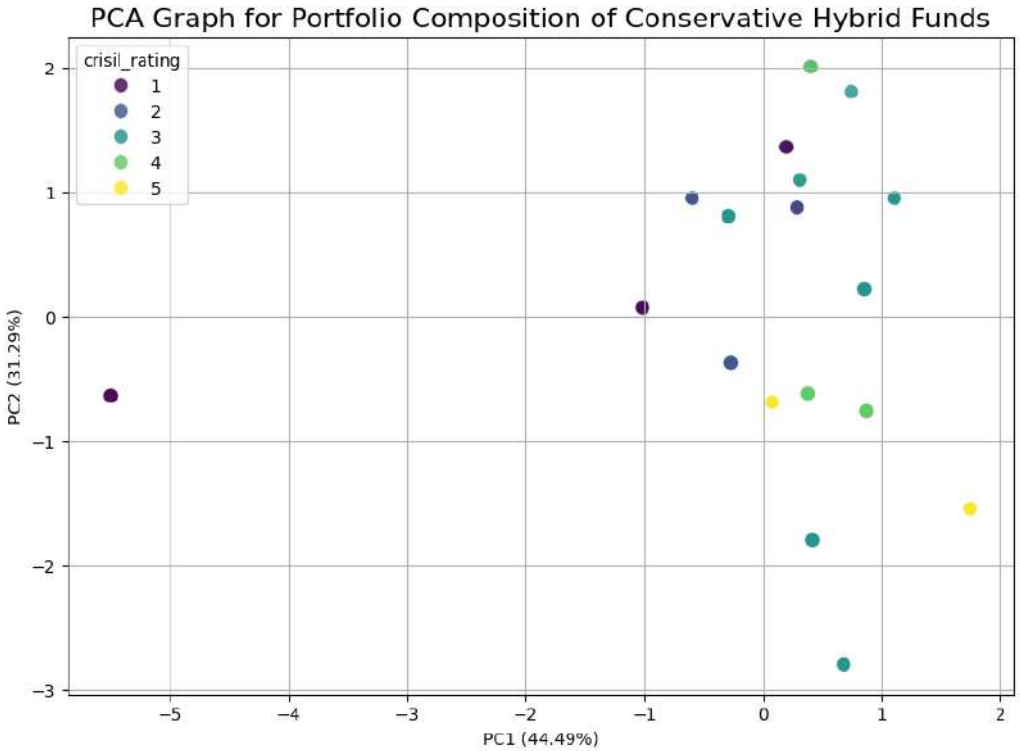


Figure 3.20: PCA Scatter Plot — Portfolio Composition (CRISIL Rating)

- Higher-rated funds cluster slightly on one side.
- Lower-rated funds move toward extremes.

Interpretation: Higher-rated funds tend to have more stable and balanced compositions, which PCA highlights through clustering.

3.3.5 Cluster Analysis

Cluster Analysis is a multivariate technique used to group similar observations based on their characteristics. In this project, hierarchical clustering is used to identify:

- Groups of return metrics that behave similarly
- Groups of mutual fund schemes with similar performance patterns

Why Hierarchical Clustering?

- Does not require a pre-defined number of clusters
- Visualizes similarity using dendrograms
- Works well for correlation-based grouping
- Helps detect natural patterns in financial data

How it Works (Simple Explanation)

1. Start with each variable or scheme as its own cluster.
2. Combine the closest pair step by step.
3. Continue merging until all become one big cluster.
4. The dendrogram shows how clusters are formed.

Hierarchical Clustering on Return Metrics

1. Heatmap + Dendrogram — Return Metrics Description:

- The heatmap shows correlations between all return variables (1-week, 1-month, 3-month, SIP returns, 3-year, 5-year, 10-year, etc.).
- The dendrogram on top and left groups variables based on similarity.
- Short-term returns (1W, 1M, 3M) cluster together.
- Long-term rolling returns (5Y SIP, 10Y SIP, 3Y SIP) form another strong cluster.
- Medium-term returns (1Y, 3Y, 5Y) appear in the middle cluster.
- High correlations among long-term SIP returns indicate consistent investment performance.

Interpretation: Return variables behave in three main groups:

1. Short-term performance cluster

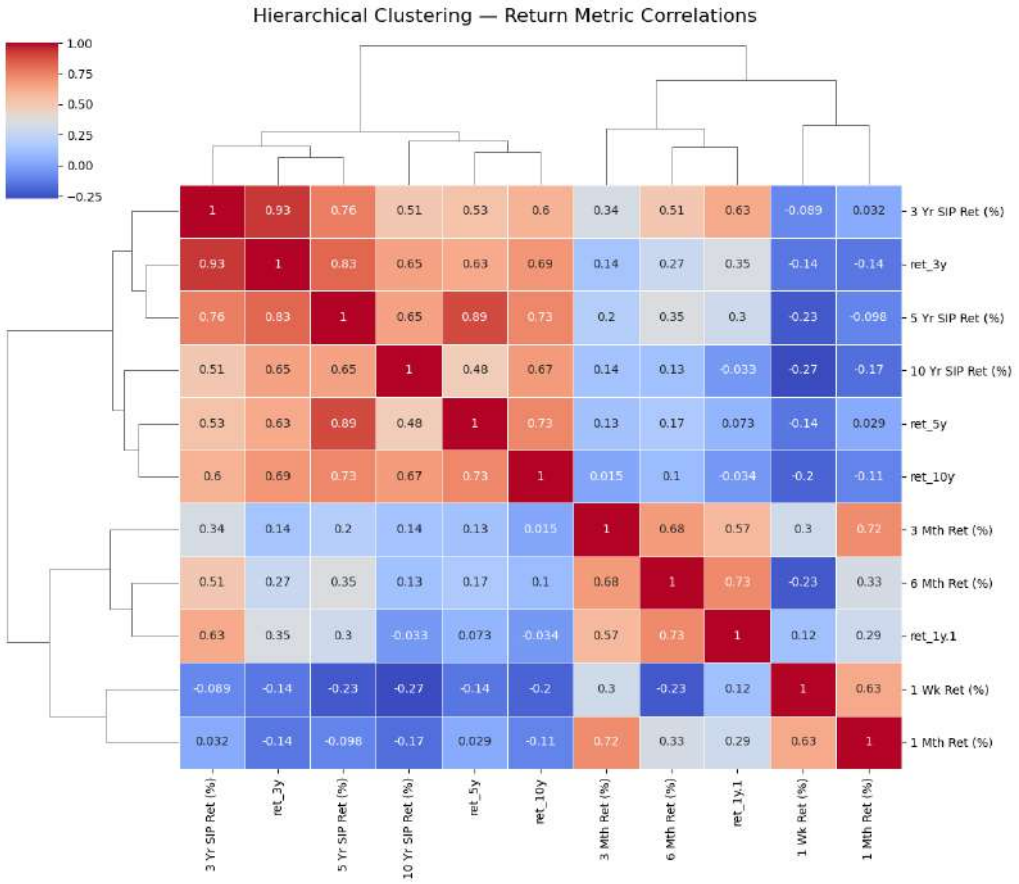


Figure 3.21: Hierarchical Clustering — Return Metric Correlations

2. Medium-term performance cluster
3. Long-term SIP performance cluster

This helps understand return behaviour across different time horizons.

Hierarchical Clustering on Fund Schemes

2. Dendrogram — Scheme-Level Clustering Description:

- Each branch represents a mutual fund scheme.
- The y-axis shows the distance between clusters (Euclidean distance).
- A horizontal cut at the dashed line (7 distance) forms 3 major clusters of funds.
- Schemes in the same cluster share similar risk-return behaviour.
- Some funds merge early (low distance) → very similar performance, others join later → very different performance patterns.

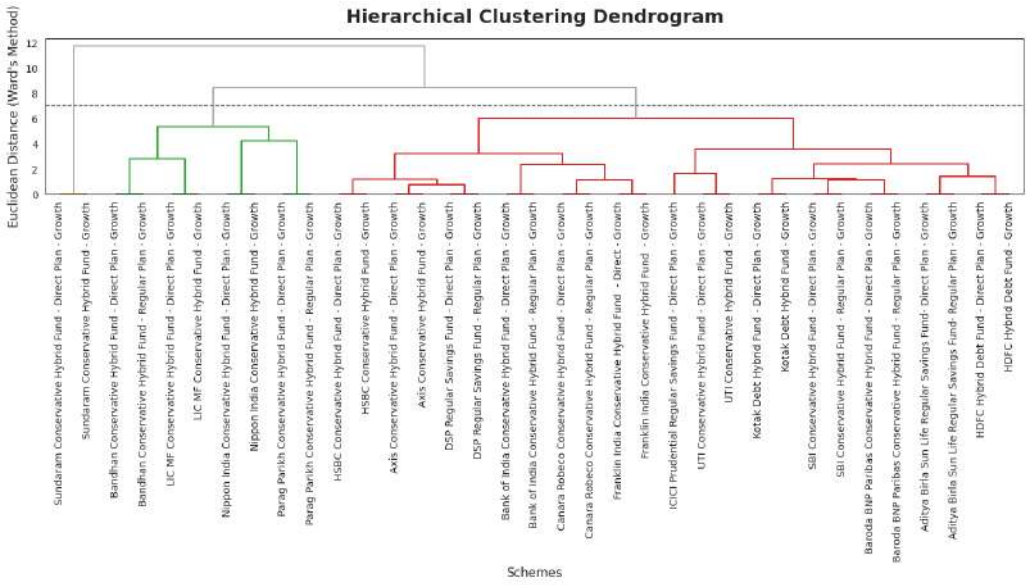


Figure 3.22: Hierarchical Clustering — Fund Schemes

Cluster Interpretation

Cluster 1 — Low Risk / Stable Returns: Funds with lower standard deviation, moderate but stable returns, higher AUM in many cases, often Regular plan variants, mostly from large AMCs.

Cluster 2 — Medium Risk / Medium Returns: Balanced return patterns, slightly higher volatility, many conservative hybrid category funds fall here, mix of Direct and Regular plans.

Cluster 3 — High Risk / Mixed Performance: Higher standard deviation, highly scattered return behaviour, often Direct plans or newer schemes, lower AUM commonly seen.

Conclusion: This segmentation helps investors understand which schemes behave similarly and how risk levels influence clustering.

3.4 Correlation Heatmap — Portfolio Composition

What this heatmap shows

This heatmap illustrates the **strength and direction of relationships** between key portfolio features. Correlation values range from −1 to +1:

- **+1**: perfect positive correlation
- **−1**: perfect negative correlation
- **0**: no correlation

Important Relationships

Debt holding vs Cash holding: –0.98 Very strong negative correlation. When debt allocation is high, cash allocation is usually very low, and vice versa. This is expected because funds with higher fixed-income exposure tend to stay fully invested.

Equity holding vs CRISIL rating: –0.54 Moderate negative correlation. Higher equity exposure is associated with lower CRISIL ratings. Equity-heavy portfolios carry higher market risk, reducing credit-quality ratings.

Equity holding vs Number of equity stocks: +0.49 Positive correlation. Funds with higher equity allocation tend to diversify across more equity stocks. Growth-oriented conservative hybrids often hold more individual stocks.

Debt holding vs CRISIL rating: +0.50 Positive correlation. Higher debt allocation corresponds to better CRISIL ratings, reflecting safer and more stable portfolios.

Cash holding vs CRISIL rating: –0.45 Slight negative correlation. Higher cash exposure is linked with marginally lower CRISIL ratings, possibly because high cash levels imply less active investment.

Number of debt holdings vs CRISIL rating: +0.30 Mild positive correlation. More diversified debt portfolios tend to have better credit quality.

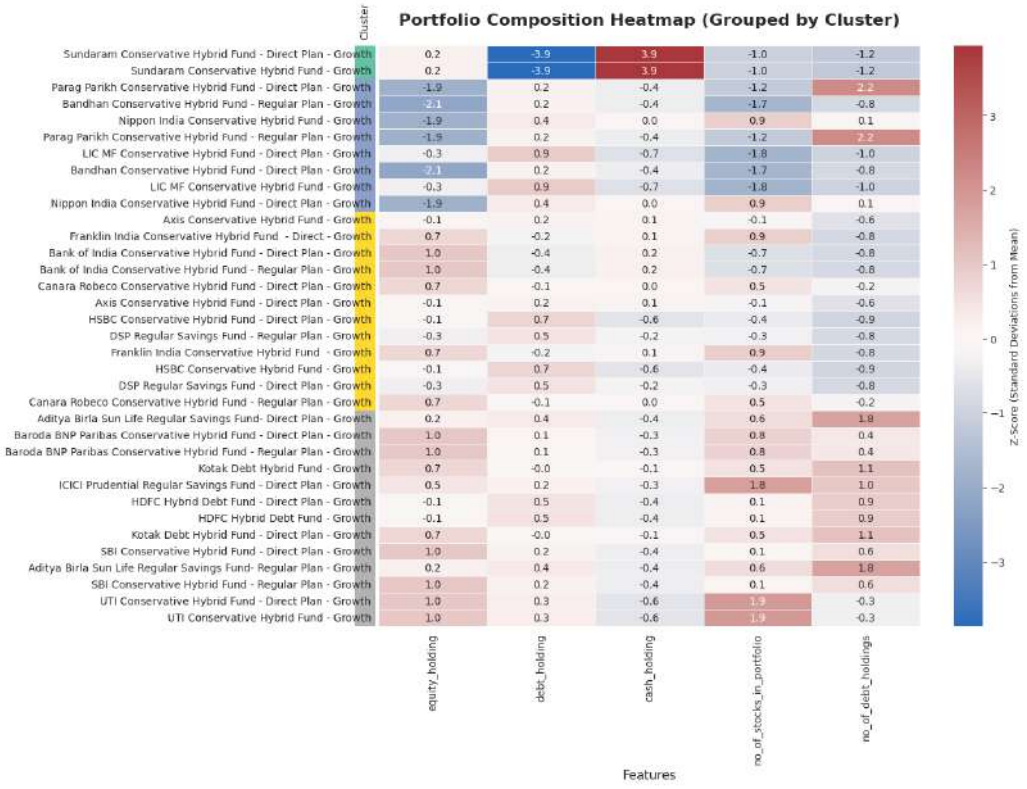


Figure 3.23: Portfolio Composition Heatmap

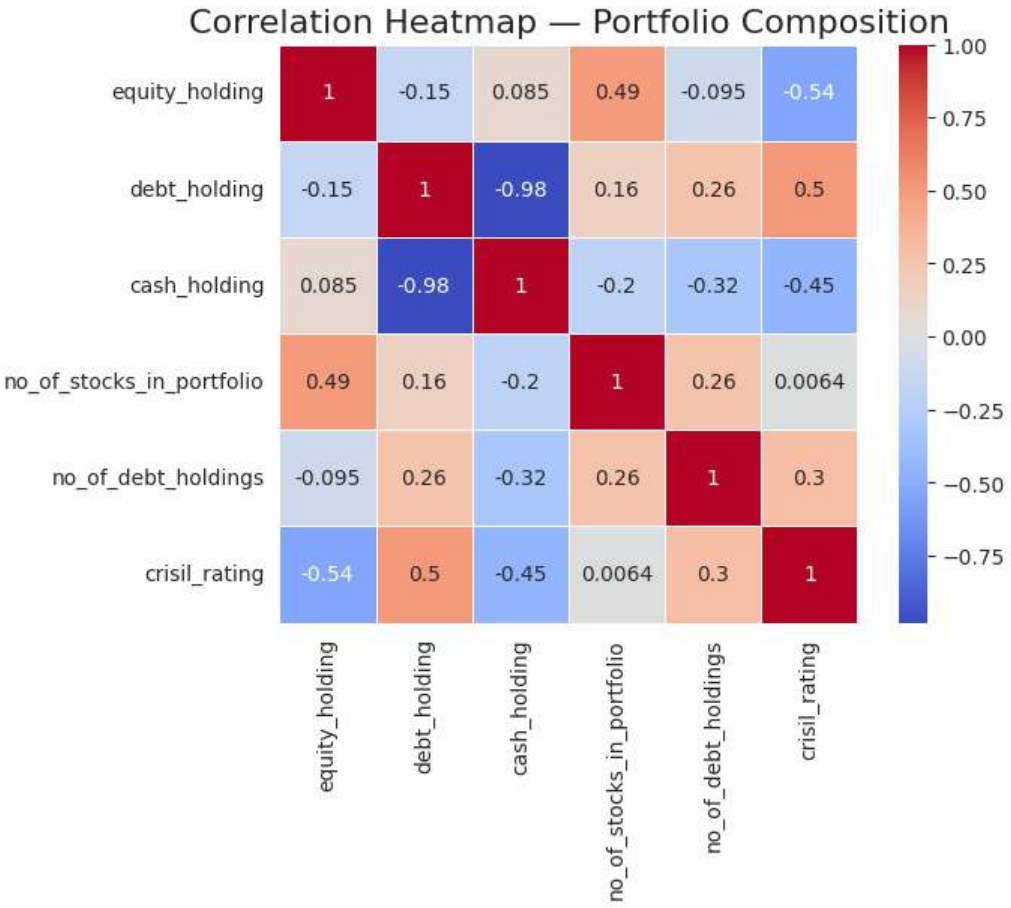


Figure 3.24: Correlation Heatmap - Portfolio Composition

This heatmap shows how different Conservative Hybrid Funds allocate their portfolios across equity, debt, cash, and number of holdings. Funds with similar allocation patterns are grouped together. Red cells indicate above-average values, while blue cells indicate below-average values.

This heatmap shows how portfolio features relate to each other. Key relationships include: Debt vs Cash = –0.98: Higher debt allocation almost always means lower cash. Equity vs CRISIL rating = –0.54: Higher equity exposure tends to reduce credit quality rating. Equity vs No. of stocks = +0.49: More equity → more stock holdings.

Chapter 4. Feature Engineering

Feature Engineering refers to the process of transforming raw data into meaningful and structured features that improve the quality of analysis. In this project on **Conservative Hybrid Fund Analysis**, feature engineering played an important role in converting financial datasets into analytical insights.

4.0.1 1. Date-Based Feature Creation (NAV Dataset)

The NAV dataset contained historical Net Asset Values along with their respective dates. To analyze time-based performance trends, several date-derived features were engineered:

- **Year** – for annual trend analysis
- **Month** – to identify monthly behaviour and seasonality
- **Quarter** – to compare quarterly performance
- **Daily/Monthly Returns** – percentage change in NAV values

These features supported the study of growth cycles and fund stability over time.

4.0.2 2. Return Metrics

From NAV values, additional return-related features were created:

- Daily Return
- Cumulative Return
- Rolling Returns (7-day, 30-day, 90-day)

These features helped observe short-term and long-term performance patterns.

4.0.3 3. Volatility Features

To measure the stability and risk of the fund, the following volatility-oriented features were engineered:

- Rolling Standard Deviation (7-day, 30-day)
- Coefficient of Variation
- Return Volatility

These metrics enabled us to assess whether the fund behaves consistently with conservative investment expectations.

4.0.4 4. Drawdown Features

Drawdown-related features provided insight into worst-case performance scenarios:

- Peak NAV
- Drawdown Value (drop from recent peak)
- Maximum Drawdown

4.0.5 5. Portfolio Allocation Features

Using the portfolio dataset, several allocation-based features were derived:

- Equity Allocation Percentage
- Debt Allocation Percentage
- Cash Allocation Percentage
- Top Holdings Concentration

These features helped analyze how asset distribution impacts NAV performance and returns.

4.0.6 6. Correlation Features

Correlation matrices were generated to study relationships among:

- NAV values
- Returns
- Asset Allocation Percentages
- Volatility Measures

These correlations provided deeper insight into how NAV behaviour is influenced by portfolio structure.

4.0.7 7. Cleaning and Standardization as Feature Engineering

Some preprocessing operations also contributed to feature engineering, such as:

- Converting textual numeric values to numerical format
- Standardizing date formats
- Handling missing values
- Normalizing percentage fields
- Removing inconsistencies in scheme/fund names

These steps ensured that the data was consistent and analysis-ready.

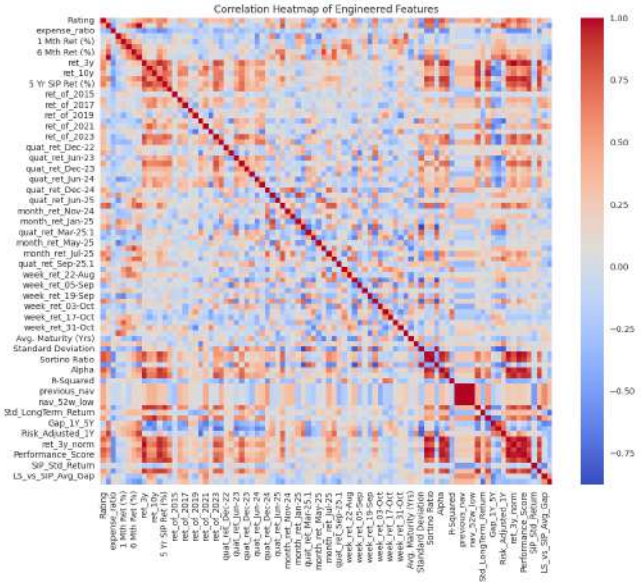
4.0.8 Summary

Feature Engineering transformed raw NAV values, analytics data, and portfolio allocation data into structured analytical features. The engineered features—such as returns, volatility, drawdowns, date-based attributes, and allocation percentages—served as the foundation for extracting meaningful insights during the Exploratory Data Analysis.

4.0.9 Correlation Heatmap of Engineered Features

What It Shows

This heatmap displays the correlation between all engineered features, including:



- Monthly, weekly, and annual return features
- Risk metrics (Standard Deviation, Sharpe Ratio, Sortino Ratio)
- NAV-based metrics
- Stability indicators

Meaning

- **Red:** strong positive correlation (both features increase together)
- **Blue:** strong negative correlation (one increases while the other decreases)
- Clusters of red/blue blocks indicate groups of features that move together

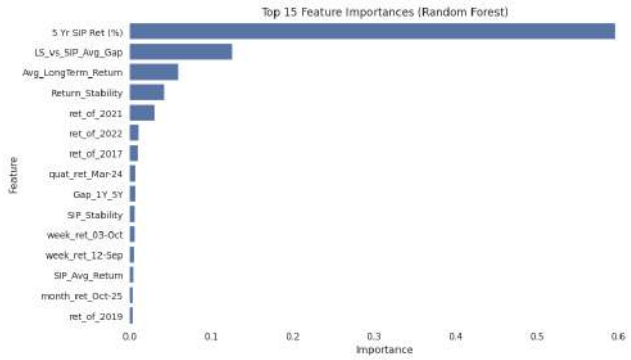
Key Insight

Several return-based features (weekly, monthly, yearly) form strong clusters, indicating similar behavior. Risk metrics (Standard Deviation, Sharpe, Sortino) form a distinct cluster, showing stable internal relationships.

Top 15 Feature Importances (Random Forest)

What It Shows

This bar chart presents the most important features selected by a Random Forest model for predicting fund performance (e.g., long-term return or stability).



Enter Caption

Meaning

- Longer bars represent features with higher predictive power.
- Shorter bars indicate features with low importance.

Key Insight

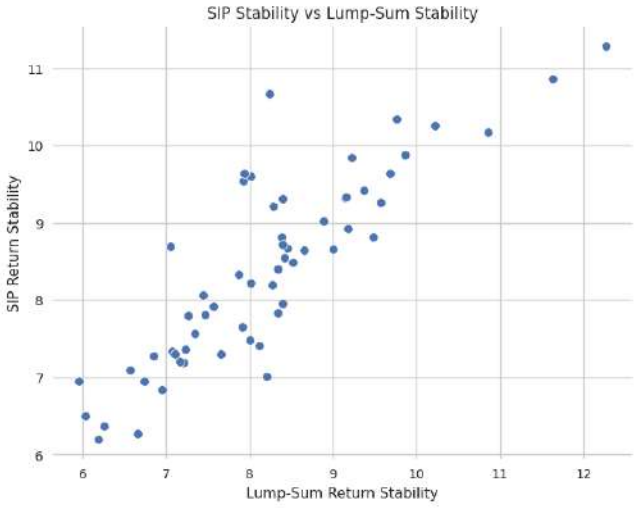
- **5-Year SIP Return (%)** is the single strongest predictor of overall performance.
- Features such as `LS_vs_SIP_Avg_Gap`, `Avg_LongTerm_Return`, and `Return_Stability` also have high importance.
- Weekly or monthly returns contribute minimally to long-term performance prediction.

SIP Stability vs Lump-Sum Stability

What It Shows

A scatter plot comparing:

- SIP Return Stability — smoothness of returns under monthly investments
- Lump-Sum Return Stability — volatility when invested all at once



Enter Caption

Meaning

The points form a clear upward trend: funds with stable lump-sum performance also show high SIP stability.

Key Insight

SIP performance is strongly correlated with overall fund stability. Stable funds remain stable regardless of whether the investment is lump-sum or SIP.

Risk vs Return Plot (Colored by Sharpe Ratio)

What It Shows

A 2D risk–return scatter plot with:

- X-axis: Standard Deviation (Risk)
- Y-axis: 5-Year Return (%)
- Color: Sharpe Ratio (Risk-Adjusted Return)

Enter Caption

Meaning

- Funds in the upper-left region (high return, low risk) are ideal.
- Greenish colors indicate higher Sharpe Ratio, meaning better risk-adjusted performance.

Key Insight

Some funds achieve higher returns even at lower risk, indicated by bright green points. Higher-Sharpe funds tend to lie above the main cluster, demonstrating superior performance per unit of risk.

4.1 Feature extraction

Feature Extraction is a key component of the Feature Engineering process. It focuses on deriving new, meaningful variables from existing raw data. The purpose of feature extraction is to summarize complex financial information into simplified indicators that better represent performance trends, risk, and behavioural characteristics. In the context of our project on Conservative Hybrid Fund Analysis, feature extraction allowed us to convert NAV series, portfolio holdings, and analytical metrics into compact and insightful features suitable for detailed exploratory analysis.

4.1.1 Return-Based Feature Extraction

To understand the performance behaviour of the fund, several return-oriented features were extracted from the NAV dataset:

- **Daily Return** – percentage change in NAV from the previous day.
- **Monthly Return** – aggregated NAV change over a month.
- **Cumulative Return** – total NAV growth since the initial period.
- **Rolling Returns** (7-day, 30-day, 90-day) – short- and medium-term return windows.

These extracted features convert NAV fluctuations into clear and interpretable performance indicators.

4.1.2 Volatility and Risk Feature Extraction

To evaluate the stability and risk level of the fund, the following statistical features were extracted:

- **Rolling Standard Deviation** – measures short-term and medium-term volatility.
- **Return Volatility Index** – overall variability in returns.
- **Coefficient of Variation** – normalized measure of dispersion.

These features help assess how consistently the fund performs over time.

4.1.3 Drawdown Feature Extraction

Drawdown analysis quantifies downside risk by identifying declines from historical peaks. From the NAV series, the following features were extracted:

- **Peak NAV** – highest NAV reached.
- **Drawdown Value** – drop from the most recent peak.
- **Maximum Drawdown** – largest recorded historical decline.

These features capture the worst-case performance scenarios crucial for risk assessment.

4.1.4 Time-Based Feature Extraction

Using the date field from the NAV dataset, several temporal features were extracted:

- **Year**
- **Month**
- **Quarter**
- **Day of Week**
- **Time Index** – numerical index useful for modelling trends.

These features support time-series analysis and enhance period-wise comparison.

4.1.5 Portfolio Structure Feature Extraction

From the portfolio dataset, several allocation-based features were derived:

- **Equity Allocation Percentage**
- **Debt Allocation Percentage**
- **Cash Allocation Percentage**
- **Category-wise Exposure**
- **Top Holdings Concentration**

These features help correlate fund performance with portfolio composition.

4.1.6 Correlation-Based Feature Extraction

To understand the relationships between key variables, correlation features were generated:

- Correlation between NAV and returns.
- Correlation between NAV and asset allocation.
- Correlation between returns and volatility indicators.

These extracted relationships highlight the underlying drivers of fund performance.

4.1.7 Importance of Feature Extraction

Feature Extraction played an essential role in this project because it:

- Reduced complex datasets into concise analytical indicators.
- Simplified interpretation of performance, volatility, and risk trends.
- Enabled understanding of how portfolio allocation affects NAV movement.
- Transformed raw NAV values into statistically meaningful signals.

Overall, feature extraction established a strong foundation for conducting Exploratory Data Analysis of Conservative Hybrid Funds.

4.2 Feature selection

Feature Selection is the process of identifying the most relevant variables from the available dataset that contribute significantly to the analysis. While feature extraction creates new features, feature selection focuses on choosing the most meaningful ones and eliminating irrelevant, redundant, or highly correlated variables. This step helps improve interpretability, reduce noise, and enhance the overall quality of the exploratory analysis.

In the context of our Conservative Hybrid Fund Analysis, feature selection was essential because the datasets (NAV values, portfolio allocation data, and analytics data) contained multiple fields, some of which did not contribute meaningfully to the study. Selecting the right features helped us focus on metrics that describe fund performance, stability, and risk behaviour effectively.

4.2.1 Selection of Performance-Related Features

From the NAV and analytics dataset, the following features were selected as they represent core performance indicators:

- Historical NAV values
- Daily, monthly, and cumulative returns
- Rolling returns (7-day, 30-day, 90-day)

These features directly reflect the historical growth and behaviour of the fund.

4.2.2 Selection of Risk and Volatility Features

To evaluate the stability and risk profile of the Conservative Hybrid Fund, the following features were selected:

- Standard deviation of returns
- Rolling volatility measures
- Coefficient of variation
- Maximum drawdown and drawdown series

These features help understand the risk–return trade-off and assess whether the fund aligns with conservative investment expectations.

4.2.3 Selection of Portfolio Allocation Features

Portfolio composition plays a major role in determining fund performance. Hence, the following allocation-based features were selected:

- Equity allocation percentage
- Debt allocation percentage
- Cash or liquid asset percentage
- Category-wise portfolio weights

These features enable correlation analysis between portfolio structure and NAV behaviour.

4.2.4 Use of Correlation Analysis in Feature Selection

Correlation matrices were used to identify redundant features:

- Highly correlated features were flagged to avoid duplication of information.
- Weakly correlated or non-contributing features were excluded from deeper analysis.

This ensured that the selected feature set provided unique and meaningful information.

4.2.5 Importance of Feature Selection

Feature selection was crucial for the following reasons:

- It reduced dimensionality and improved clarity of the analysis.
- It enabled focus on the most relevant performance and risk indicators.
- It eliminated noise and redundant features that could mislead interpretation.
- It improved efficiency of visualisation and statistical examination.

Overall, feature selection ensured that the final analytical model focused only on the variables that significantly contributed to understanding the performance, structure, and risk characteristics of Conservative Hybrid Funds.

Chapter 5. Model fitting

Pipeline A – Predict 1-Year Return (`ret_1y`)

Goal

Use past returns, risk metrics, AUM, expense ratio, category, and other fund characteristics to predict the **1-year return** of a scheme.

What It Does

- Loads `analytics_cleaned.csv`.
- Cleans columns and selects `ret_1y` as the prediction target.
- Uses features such as:
  - Short-term and long-term return metrics
  - Standard Deviation, Sharpe Ratio, Beta
  - AUM, expense ratio
  - Rating, category, plan type, riskometer
- Handles missing values, scales numeric columns, and one-hot encodes categorical features.
- Trains the following models:
  - Linear Regression
  - Random Forest Regressor
  - Ridge Regression
  - Lasso Regression
- Evaluates each model using:
  - MAE (Mean Absolute Error)
  - RMSE (Root Mean Squared Error)
  - R^2 Score
  - Spearman Rank Correlation
- Plots **Actual vs Predicted** values of `ret_1y` for every model.

Meaning

This pipeline builds a regression system that uses a fund's historical returns, risk metrics, AUM, expenses, ratings, and category information to estimate its upcoming 1-year return. It transforms the cleaned analytics dataset, handles missing values, scales numerical columns, and encodes categorical attributes. Several regression models—Linear, Ridge, Lasso, and Random Forest—are trained and evaluated with MAE, RMSE, R^2, and Spearman correlation. The output includes Actual vs Predicted plots that reveal how accurately various model families capture the relationship between fund characteristics and near-term performance.

Pipeline B – NAV Time-Series Forecasting (Per Scheme)

Goal

Use historical NAV values to predict the **next-day NAV** for a specific scheme.

What It Does

- Loads `nav_values_cleaned.csv` and reshapes it from wide to long format:

$$(\text{date},\ \text{scheme},\ \text{nav})$$

- Selects a single scheme and builds a daily NAV time series.
- Generates engineered features:
  - Lag features (NAV_{t-1}, NAV_{t-3}, NAV_{t-7}, ...)
  - Rolling means and rolling standard deviations (7-day, 30-day)
  - Calendar features: day of week, day of month
- Defines the target as next-day NAV.
- Splits data into:
  - Training set (older data)
  - Test set (most recent data)
- Trains three models:
  - Random Forest Regressor
  - Ridge Regression
  - Lasso Regression
- For each model, plots:
  - Actual vs Predicted NAV over time (last few months)
  - Clean and interpretable y-axis tick labels

Meaning

This pipeline focuses on per-scheme NAV prediction, using historical NAV observations to forecast the NAV of the next trading day. It converts the NAV dataset to long format, selects one scheme, and constructs a time series enriched with lag values (1-day, 3-day, 7-day, etc.), rolling averages and volatilities, and calendar patterns like day of week. The data is split chronologically into training and recent testing periods. Random Forest, Ridge, and Lasso models are trained to forecast the next-day NAV, and each model's output is visualized through Actual vs Predicted NAV curves.

Chapter 6. Conclusion & future scope

Conclusion

This project provided a comprehensive Exploratory Data Analysis (EDA) and machine-learning study of **Conservative Hybrid Mutual Funds**. The analysis revealed clear relationships between returns, risk metrics, and portfolio structure.

Key Findings

- Short-term returns, long-term returns, and risk indicators (Standard Deviation, Sharpe Ratio) show meaningful patterns that differentiate fund performance.
- Higher equity exposure increases volatility and lowers CRISIL credit quality, whereas higher debt exposure improves stability.
- Random Forest–based feature importance revealed that **5-Year SIP Return**, **Long-Term Return**, and **Return Stability** are the strongest predictors of overall fund performance.
- Regression models showed that predicting 1-year returns is feasible to a moderate extent, with non-linear models (such as Random Forest) outperforming linear models.
- NAV time-series forecasting demonstrated that machine-learning models can effectively track short-term NAV movements when lag features and rolling statistics are used.

Overall, the project successfully combined financial understanding with statistical analysis, data engineering, and machine-learning pipelines to better understand the dynamics of Conservative Hybrid Funds.

Future Scope

This work opens several opportunities for further improvement and extension:

1. **Build a complete multi-fund NAV forecasting model** Extend the single-scheme NAV forecasting approach to all schemes using multi-index time series or deep learning frameworks.
2. **Implement advanced sequence models** Models such as LSTMs, GRUs, and Temporal Convolutional Networks could improve NAV prediction accuracy.

3. **Hyperparameter tuning** The models (Random Forest, Ridge, Lasso, Linear Regression) used default settings; applying GridSearchCV or Bayesian Optimization may significantly enhance predictive performance.

4. **Add macroeconomic indicators** Incorporating interest rates, inflation, crude prices, bond yields, and economic indices could improve return prediction accuracy.

5. **Create a recommendation system** Using risk–return analysis, cluster segmentation, and Sharpe Ratio, a system can be developed to recommend the "best-fit" Conservative Hybrid Fund for an investor.

6. **Expand feature engineering** Introduce more advanced features such as volatility clustering, trend indicators, rolling beta, and sentiment analysis to strengthen both regression and time-series models.

7. **Use Explainable AI** Apply SHAP or LIME to interpret model predictions and better understand feature contributions.

Group Contribution

Member 1: Chaitri Vadaviya (202301243)

- Data Extraction
- Multivariate Analysis
- Model Fitting
- Presentation Slides
- Report Writing
- Google Colab Notebook Development

Member 2: Parthiv Bhesaniya (202303037)

- Data Preprocessing
- Bivariate Analysis
- Feature Engineering
- Presentation Slides
- Report Writing
- Google Colab Notebook Development

Member 3: Utsav Tala (202303018)

- Missing Values Detection
- Missing Values Handling
- Univariate Analysis
- NAV Time Series Analysis
- Presentation Slides

- Report Writing
- Google Colab Notebook Development

Short Bio

1. **Chaitri Vadaviya (202301243)** is a dedicated learner with strong interests in data analysis, financial research, and analytical modelling. She has worked on projects involving Exploratory Data Analysis, portfolio evaluation, and visualization, especially within the domain of mutual funds and financial markets.

Throughout her academic journey, Chaitri has developed proficiency in Python, pandas, NumPy, and statistical methods to derive meaningful insights from data. She approaches problems methodically and enjoys simplifying complex ideas through structured thinking.

In addition to her technical skills, she values discipline, conceptual clarity, and continuous improvement. Outside academics, she explores topics in finance, mathematics, and computing, focusing on personal growth and skill development.

2. **Parthiv Bhesaniya (202303037)** is a motivated B.Tech student with strong skills in Python, C++, Java, and analytical problem-solving. He has hands-on experience in data preprocessing, feature engineering, and EDA for financial datasets, including mutual fund analysis.

Parthiv frequently works with tools such as pandas, NumPy, and Scikit-learn and uses visualization libraries to interpret data effectively. His interests include algorithmic thinking, mathematical modelling, and system-level programming.

He follows a practical and detail-oriented approach to learning and consistently strengthens his technical expertise through hands-on experimentation and exploration across multiple domains.

3. **Utsav Tala (202303018)** is a B.Tech student in Mathematics and Computing at Dhirubhai Ambani University (DAU), with strong interests in Data Science, Machine Learning, statistical modelling, and software development.

He works with programming languages such as Python, C++, and Java, and often uses tools like pandas, NumPy, Matplotlib, and Scikit-learn for analysis and visualization. Utsav also enjoys performing time-series analysis, solving analytical problems, and learning through practical experimentation.

Alongside academics, he practices competitive programming on platforms like Codeforces and LeetCode and is building skills in software engineering and data-driven technologies through personal projects and continuous learning.

References

[1] Mutual Fund Knowledge. *Mutual Fund*. *URL:* https://en.wikipedia.org/wiki/Mutual_fund

[2] Association of Mutual Funds in India (AMFI). *Official Website*. *URL:* https://www.amfiindia.com/

[3] Panda, Gopinath. *Class Notes and Course Material*. Dhirubhai Ambani University (DAU), Autumn 2025.

[4] Tukey, John W. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[5] Downey, Allen B. *Think Stats: Exploratory Data Analysis*. O'Reilly Media, 2014.

[6] ChatGPT. *Language refinement assistance and informational support*. *URL:* https://chatgpt.com