IT-462 : Exploratory Data Analyis

Conservative Hybrid Funds

Group 26

Chaitri Vadaviya - 202301243
Parthiv Bhesaniya - 202303037
Utsav Tala -202303018

Overview

- Analyze performance, risk, NAV trends, and portfolio allocation of Conservative Hybrid Mutual Funds.
- Perform full EDA: summary statistics, missing value analysis, outlier detection, correlations, and time-series NAV trends.
- Engineer features: return stability, equity–debt ratio, LS vs SIP stability gape, diversification score, total holdings, etc.
- Evaluate short-term returns, long-term returns, SIP, and risk metrics across all schemes.
- Build predictive models for NAV & return forecasting.
- Solve the final problem: Identify the top 3 schemes offering the highest returns per unit of risk.

Understanding Conservative Hybrid Funds

- **Conservative Hybrid Funds invest predominantly in debt instruments (75–90%) and a smaller portion in equity (10–25%), to provide stable income with moderate growth.**
- **Their strategy focuses on capital preservation, low volatility, and steady long-term returns, making them suitable for risk-averse investors.**
- **They offer a balanced risk–return profile, delivering better returns than pure debt funds while maintaining lower risk than equity-heavy hybrids.**

What are we solving?

Investors often face difficulty in evaluating the performance and risk level of Conservative Hybrid Funds due to fluctuating NAV movements, changing portfolio allocations, and market volatility. Without proper analysis, it becomes challenging to understand whether these funds truly provide stable and low-risk returns.

This project applies Exploratory Data Analysis (EDA) on multiple fund-related datasets to study their NAV trends, portfolio structure, return behaviour, and volatility, enabling data-driven assessment of fund stability and performance.

Data Import & Preprocessing

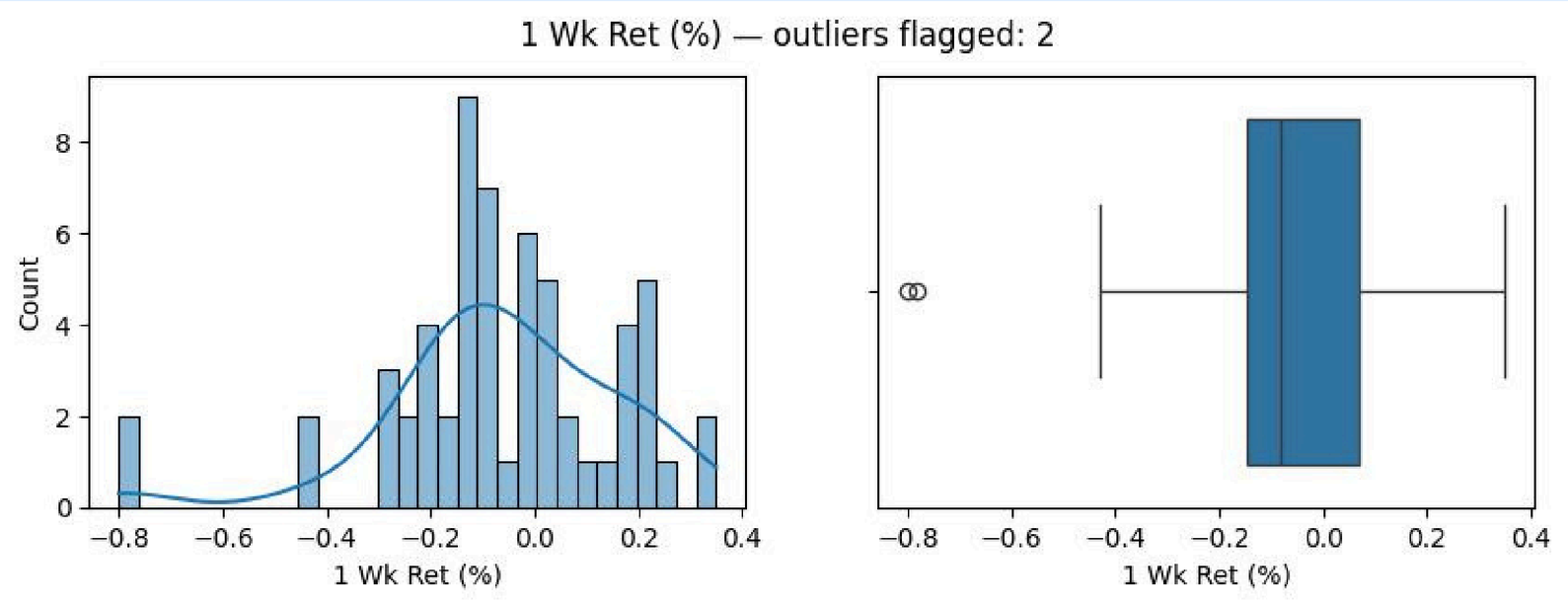
- Loaded all raw files: analytics, NAV values, and portfolio assets.
- Inspected structure using head(), info(), describe() to understand datatypes and missingness.
- Fixed messy/unlabeled columns (“Unnamed”) by extracting labels from row 0 and assigning meaningful names.
- Standardized column names and cleaned numeric fields (removed “%”, commas, “--”).
- Dropped redundant columns and aligned datasets using Scheme Name.
- Saved cleaned, structured versions for further EDA and modeling.

Missing Value Detection & Handling

- Detected missing values using isnull().sum(), descriptive stats, and dataset-wide inspection.
- Found heavy missingness in several return blocks (annual, quarterly, monthly) and NAV columns.
- Removed columns with >50% missing data using the cleaning function.
- Replaced placeholders like "--" and empty strings with NaN for consistency.
- Applied mean/median imputation for numeric fields and constant-value imputation for categorical columns.
- For NAV data, dropped rows with all-zero numeric values and filled remaining missing points using automated imputation.
- Saved cleaned datasets: analytics_cleaned.csv, nav_values_cleaned.csv, p_assets_cleaned.csv for further analysis.

Outlier Detection

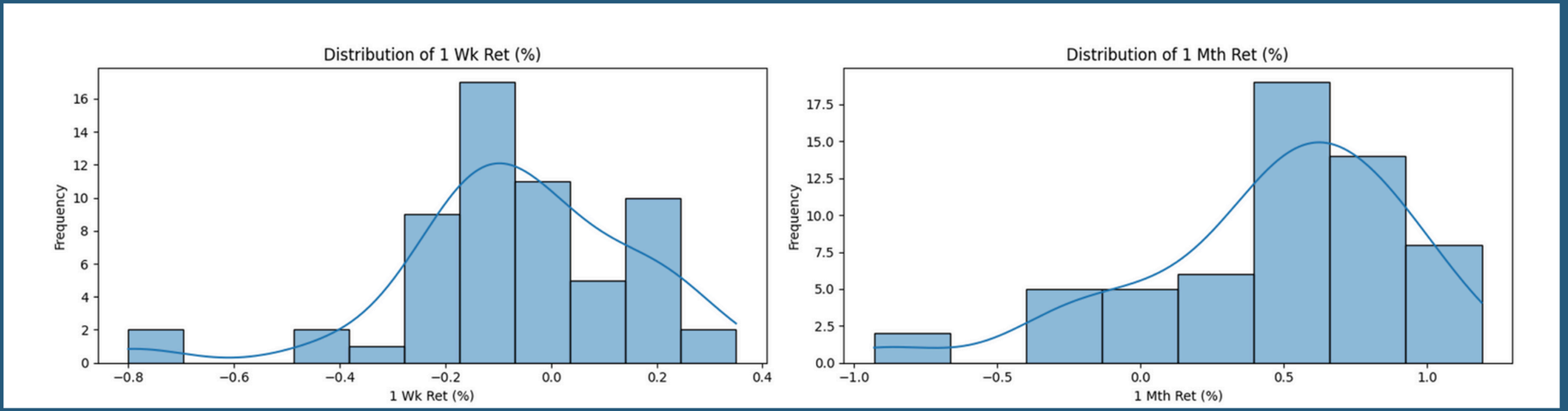
- Identified numeric columns and checked outliers using IQR and visual plots (histograms, boxplots).
- Analyzed outliers across return metrics, risk metrics, NAV values and Expanse Ratio.
- Flagged extreme values for review but did not remove them to retain real financial behavior.



Univariate Analysis

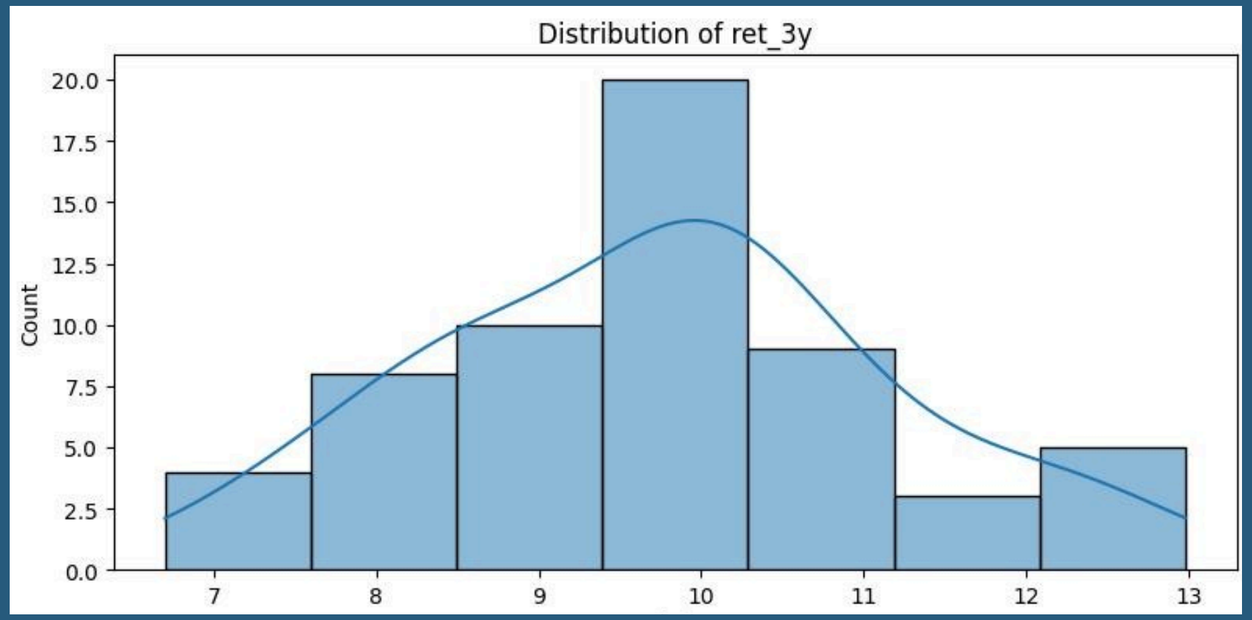
For Short Term Returns

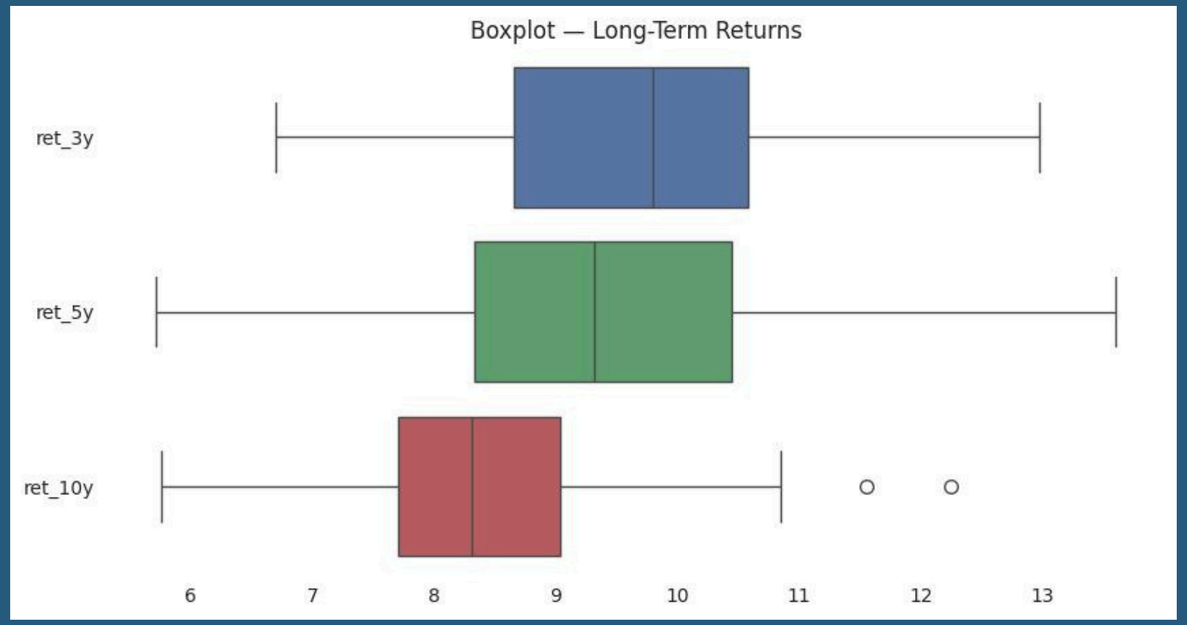
- Analyzed short-term return metrics: 1-week, 1-month, 3-month, 6-month, and 1-year returns.
- Used histograms + KDE curves to understand distribution shapes.
- Created boxplots to check spread and identify extreme values.
- Observed patterns in volatility, skewness, and fund performance consistency in the short term.



Univariate Analysis

For Long Term Returns





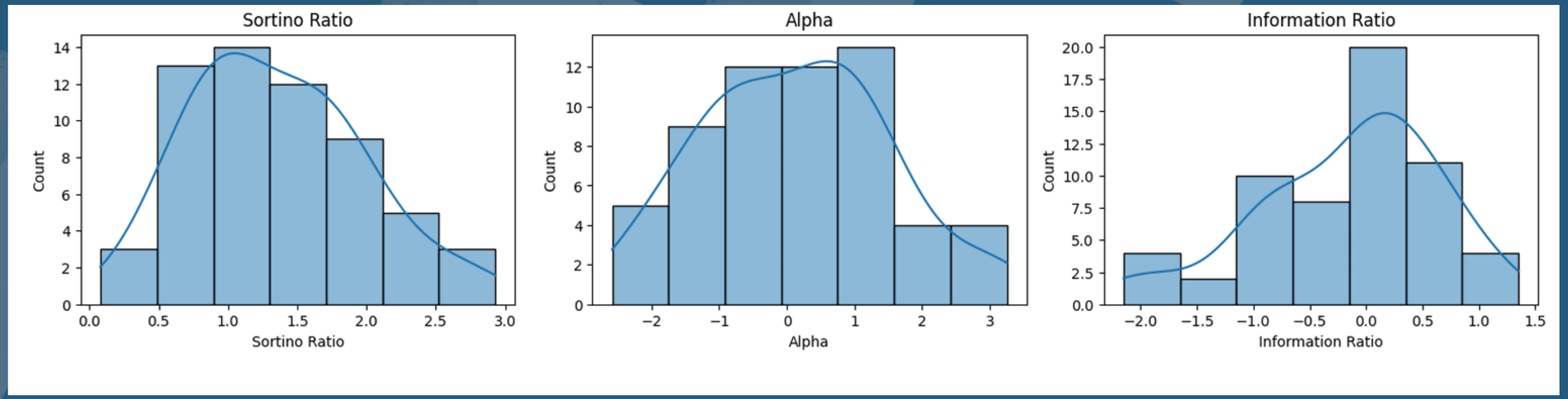
- Studied 3-year, 5-year, and 10-year return distributions.
- Calculated summary statistics: mean, median, standard deviation, quartiles, and skewness.

- Plotted histograms and boxplots to visualize long-term performance patterns
- Insights showed stability trends and long-term return behavior across funds.

Univariate Analysis

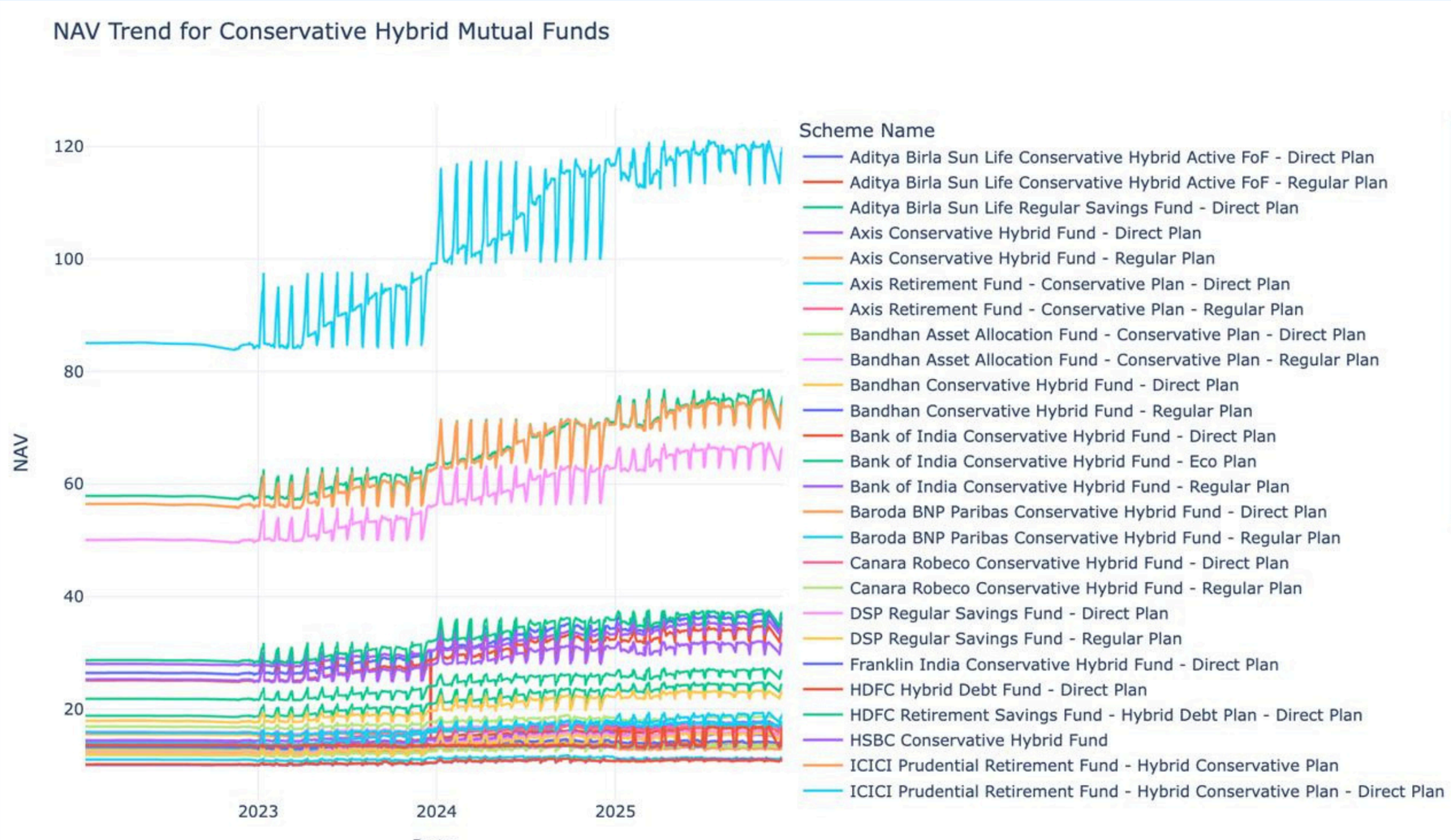
For Risk Matrics

- Analyzed key risk indicators: Standard Deviation, Beta, Sortino Ratio, Alpha, Information Ratio.
- Converted all risk columns to numeric and summarized using describe().
- Used visualizations to understand variability, tail behavior, and distribution differences.
- Provided a clear picture of how conservative hybrid funds differ in their risk levels.



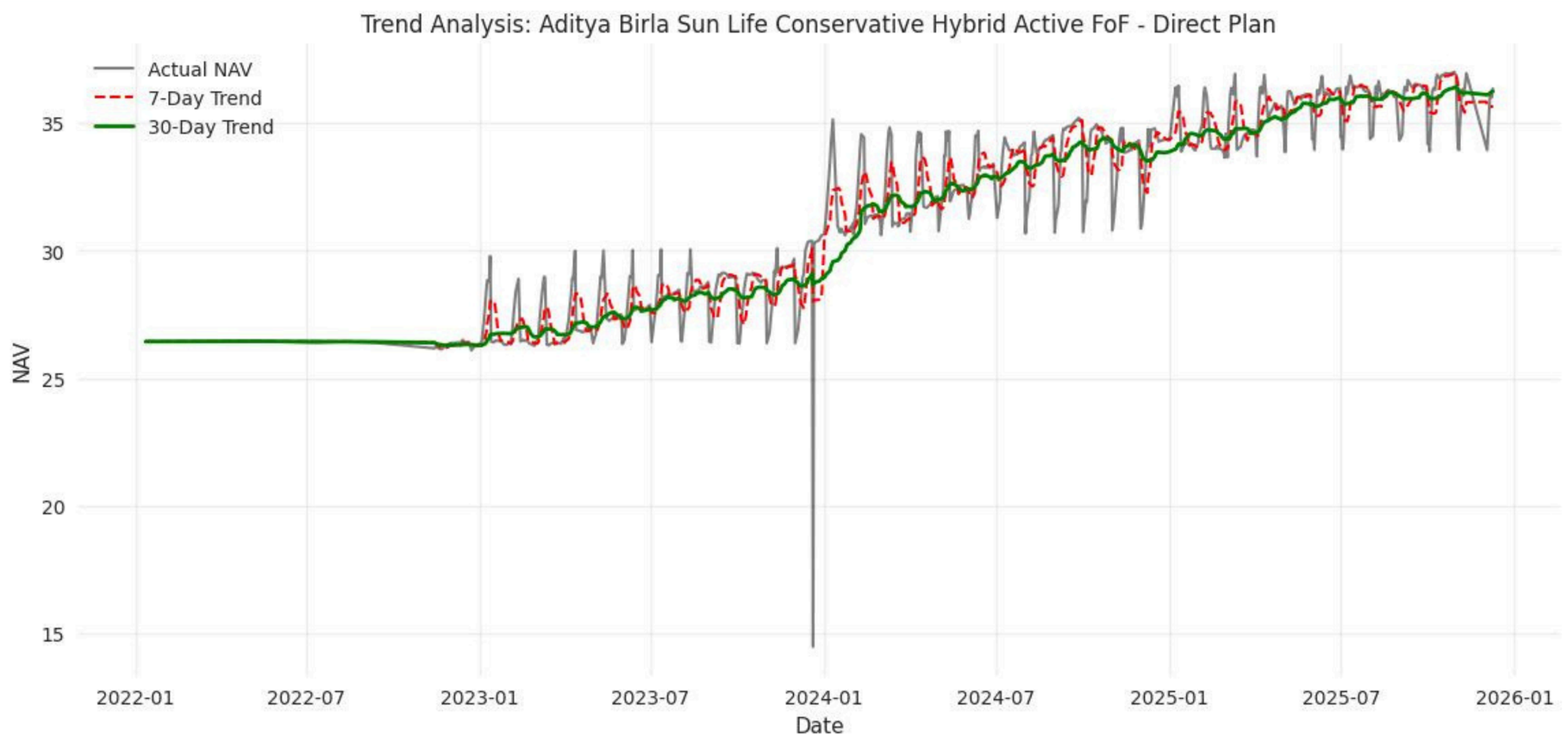
NAV Trend Time Series Analysis

- Plotted NAV trend lines across the entire period (2022–2025) for all schemes.
- Identified patterns such as gradual growth, short-term dips, and overall stability of conservative hybrid funds.



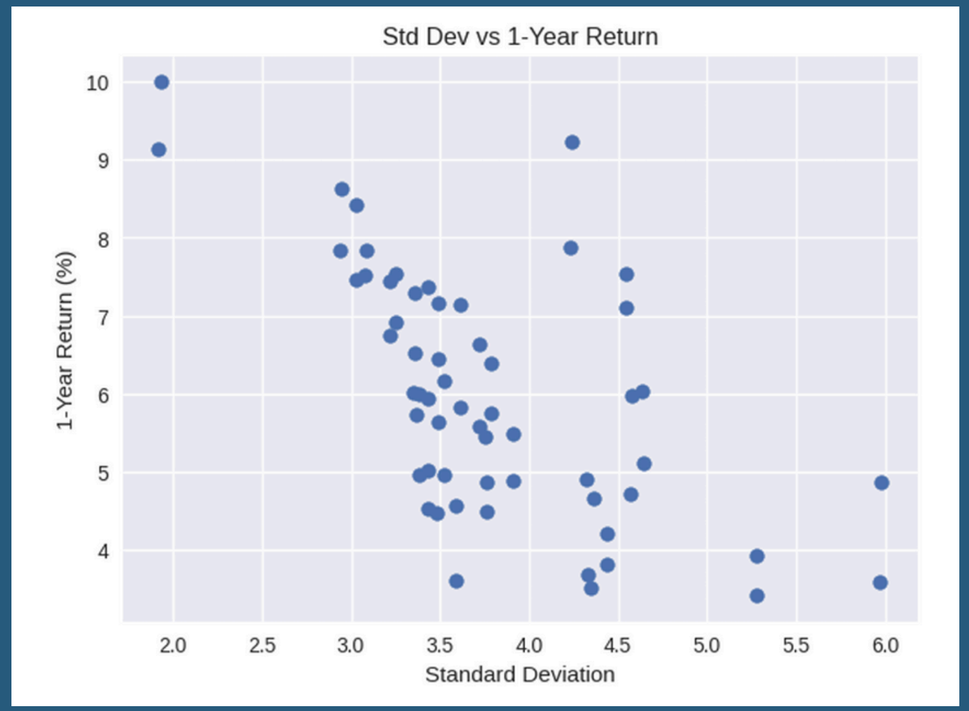
NAV Trend Time Series Analysis

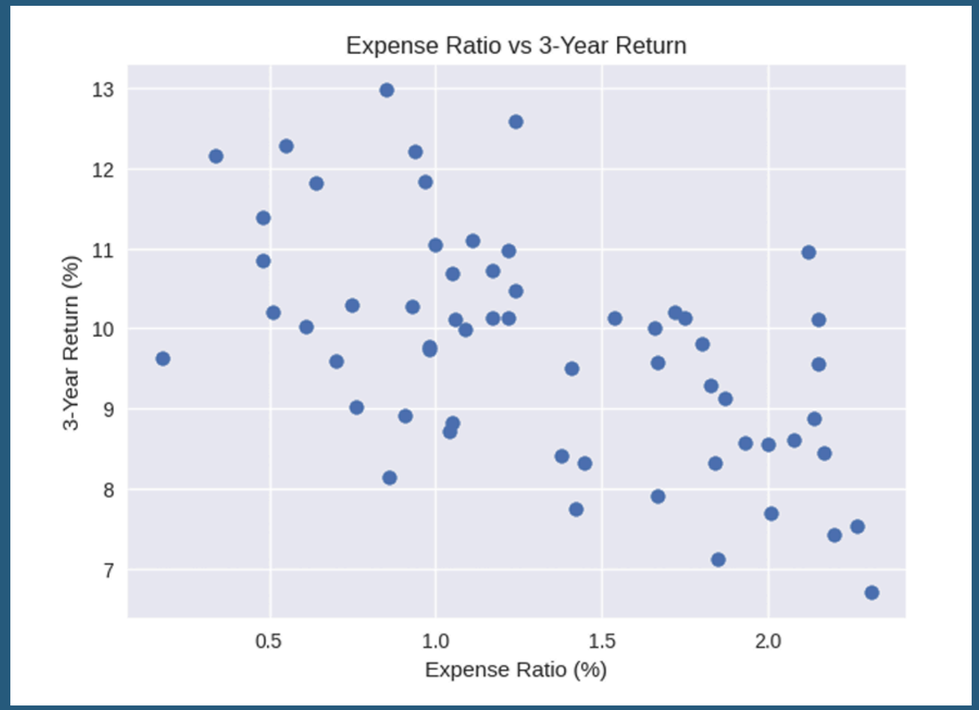
- Computed daily returns, 30-day rolling returns, and 90-day rolling volatility to capture short-term market behavior.
- Performed seasonality analysis by resampling NAV returns into monthly averages.
- Conducted drawdown analysis to measure worst declines from peak NAV levels.



Bivariate Analysis

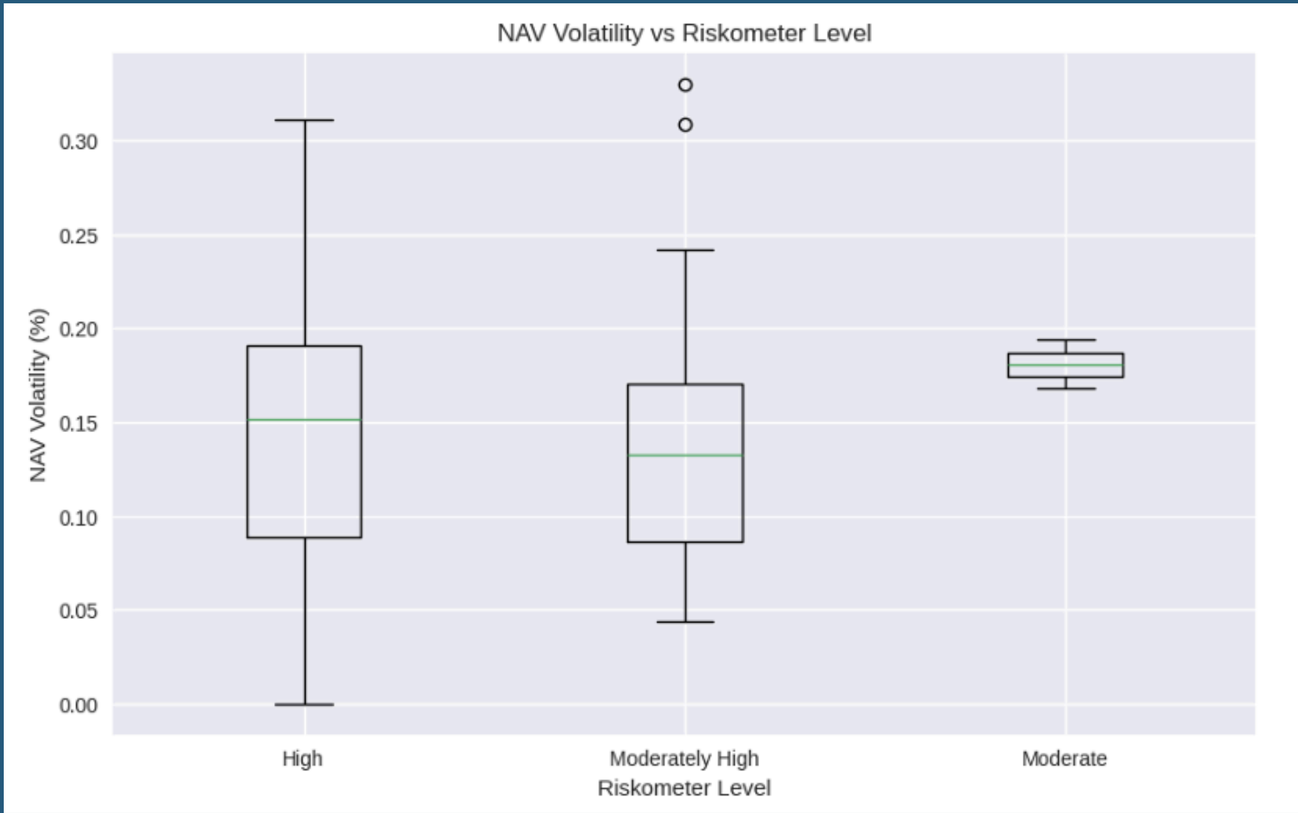
- Studied relationships between risk metrics and returns using scatter plots.
- Analyzed patterns such as:
  - Standard Deviation vs 1-Year Return (risk vs short-term performance)
  - Sharpe Ratio vs 3-Year Return (risk-adjusted efficiency)
  - Beta vs 5-Year Return (market sensitivity vs long-term returns)

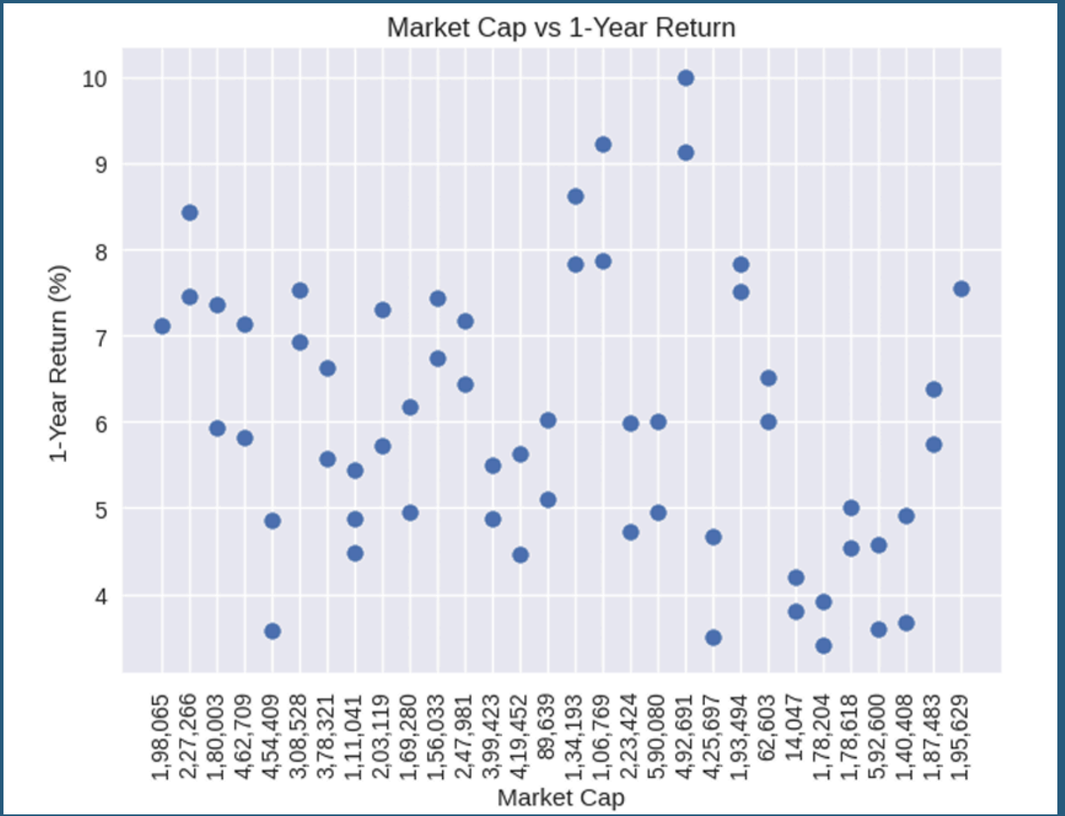




Bivariate Analysis

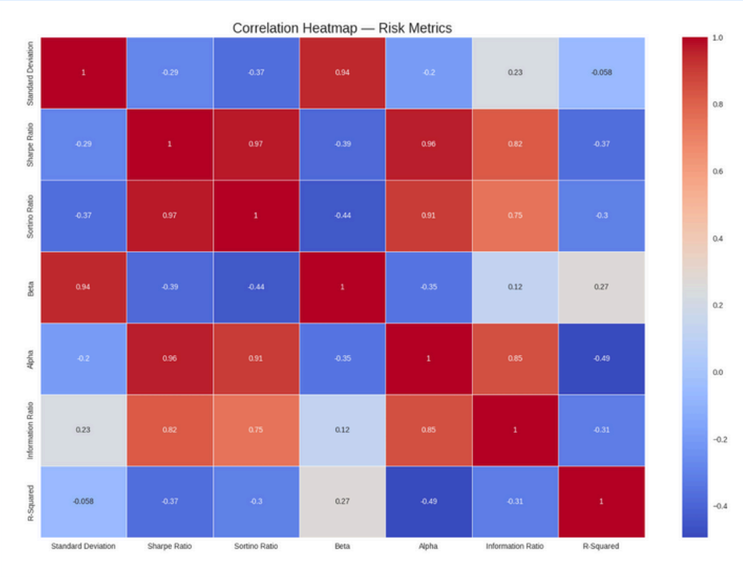
- Plotted box-plots for Riskometer Level vs NAV volatility and scatter plots for Market Cap vs Return values.
- Examined Expense Ratio vs Returns to understand fee–performance relationships (observed no positive correlation).

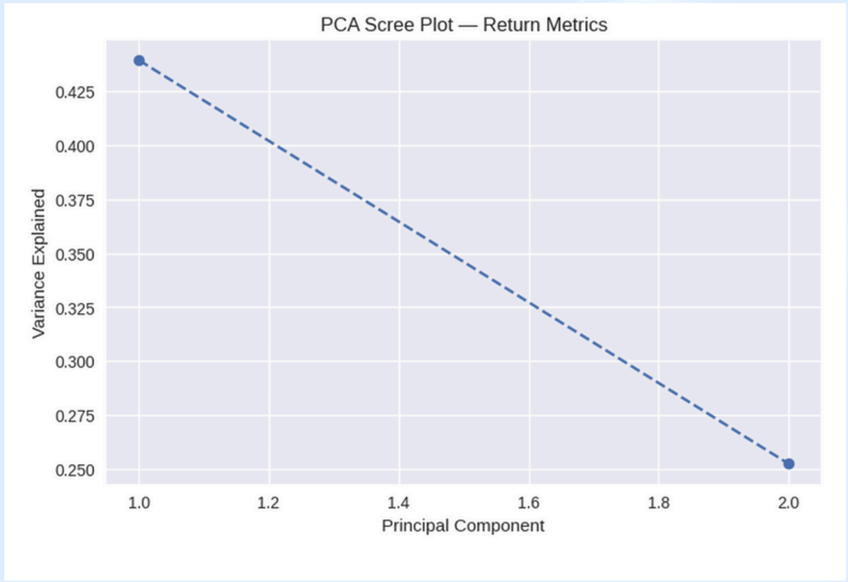


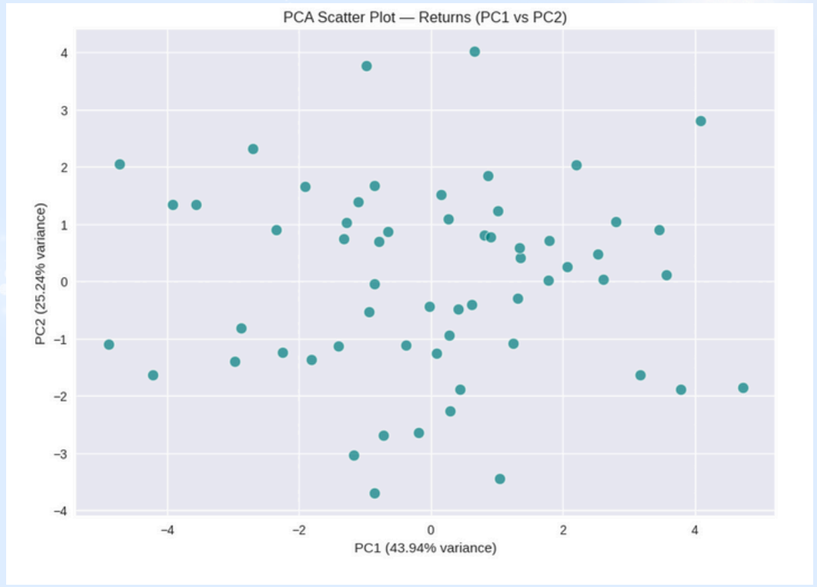


Multivariate Analysis

- Combined return metrics, risk metrics, and other numerical features to study overall fund behavior.
- Generated correlation heatmaps for:
  - Return metrics
  - Risk metrics
  - Cross-correlation between risk & return
- Performed PCA on portfolio composition to understand variance drivers such as equity–debt balance and diversification depth.

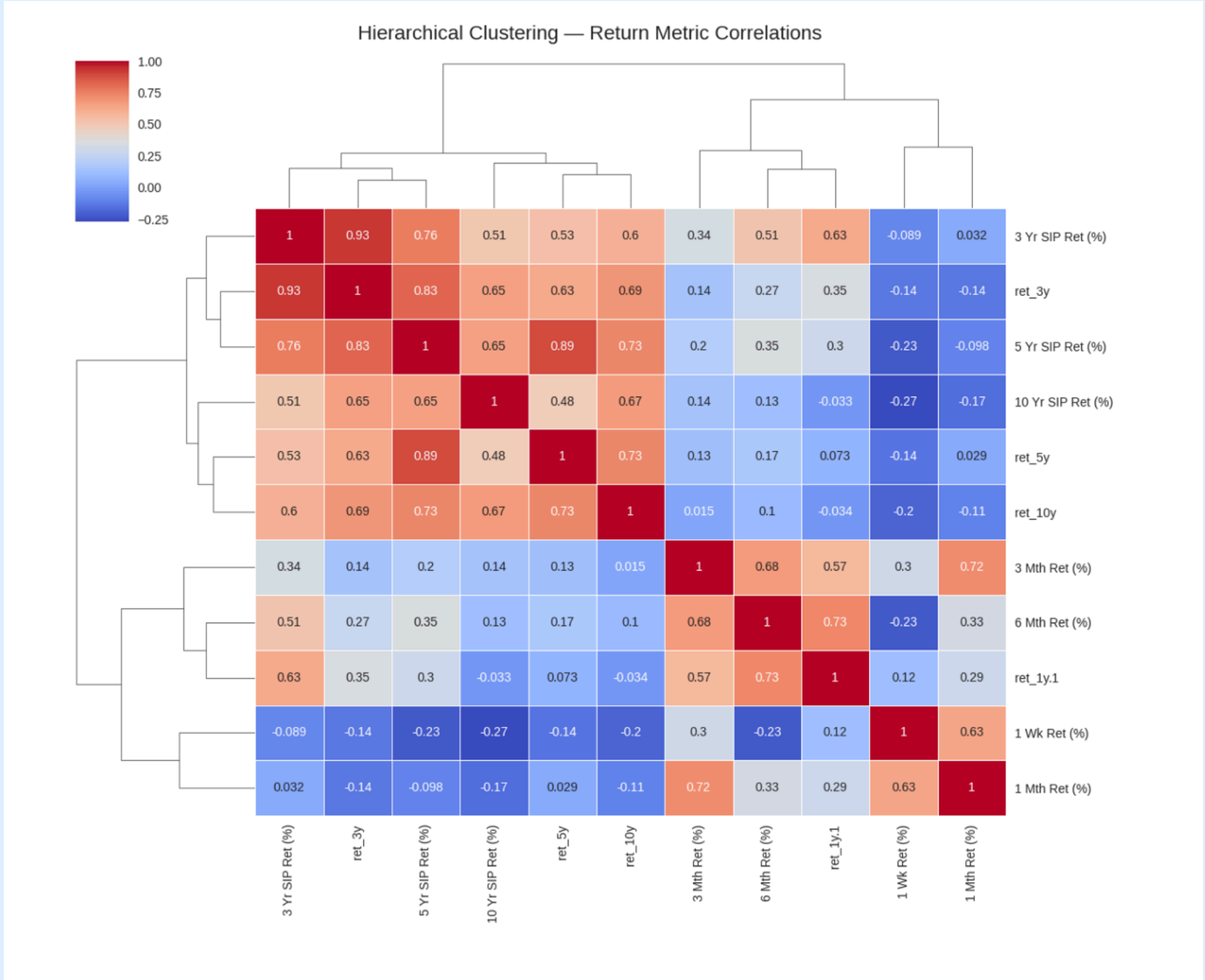






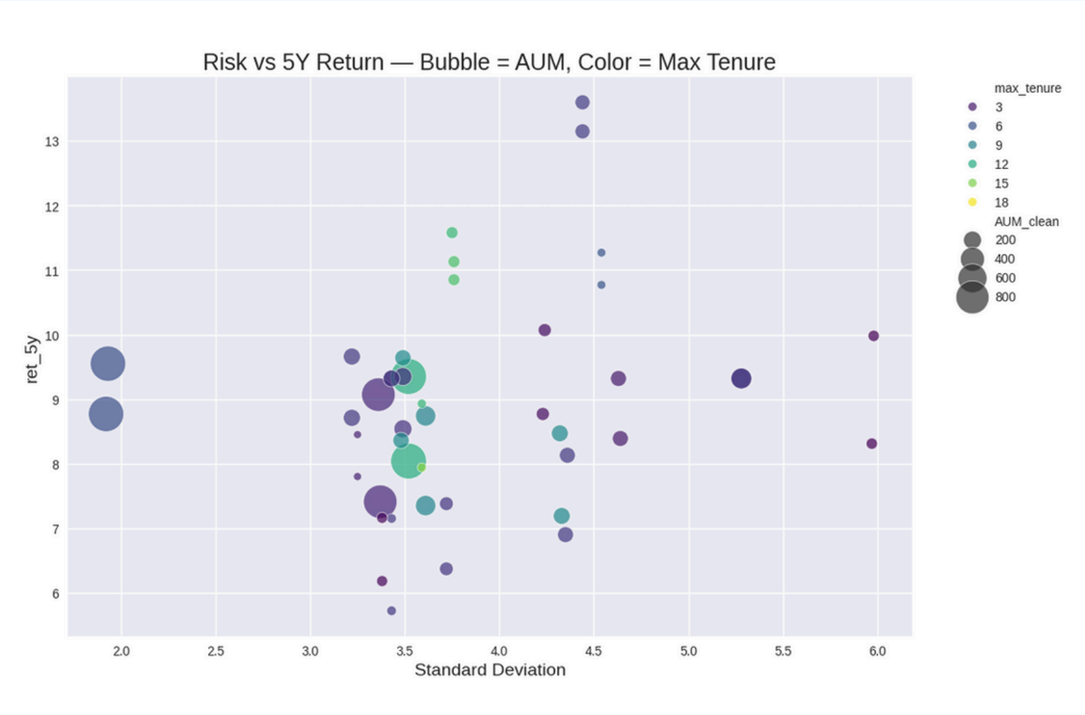
Multivariate Analysis

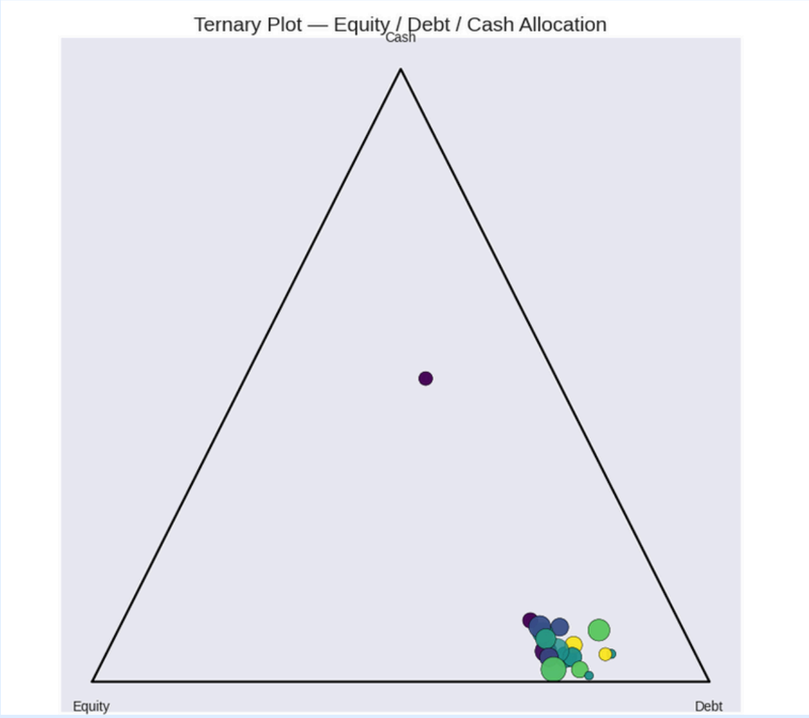
- Applied hierarchical clustering on key portfolio features such as equity %, debt %, cash %, number of stocks, number of debt holdings.
- Generated a dendrogram to visualize how schemes group together based on portfolio similarity.
- Identified three clear fund clusters:
  - Debt-Heavy / Low-Risk Funds
  - Balanced / Growth-Oriented Funds
  - High-Cash / Tactical Allocation Funds



Multivariate Analysis

- Visualized the relationship between risk and return using bubble charts (e.g., 5-Year Return vs Standard Deviation).
- Bubble size represented AUM, showing how large funds differ in performance.
- Bubble color captured categories such as Riskometer level, Fund Manager Tenure, or Plan Type (Direct/Regular).
- Also plotted Ternary Graph for Equity/Debt/Cash allocation





Feature Engineering

Feature Extraction

- Derived new metrics such as Avg_LongTerm_Return, Return_Stability, Risk_Adjusted_1Y, Sharpe_Adjusted_3Y, and Performance_Score.
- Created portfolio-level features: Equity–Debt Ratio, Diversification Score, Portfolio Concentration.
- Standardized numeric fields and cleaned percentage/ratio columns.

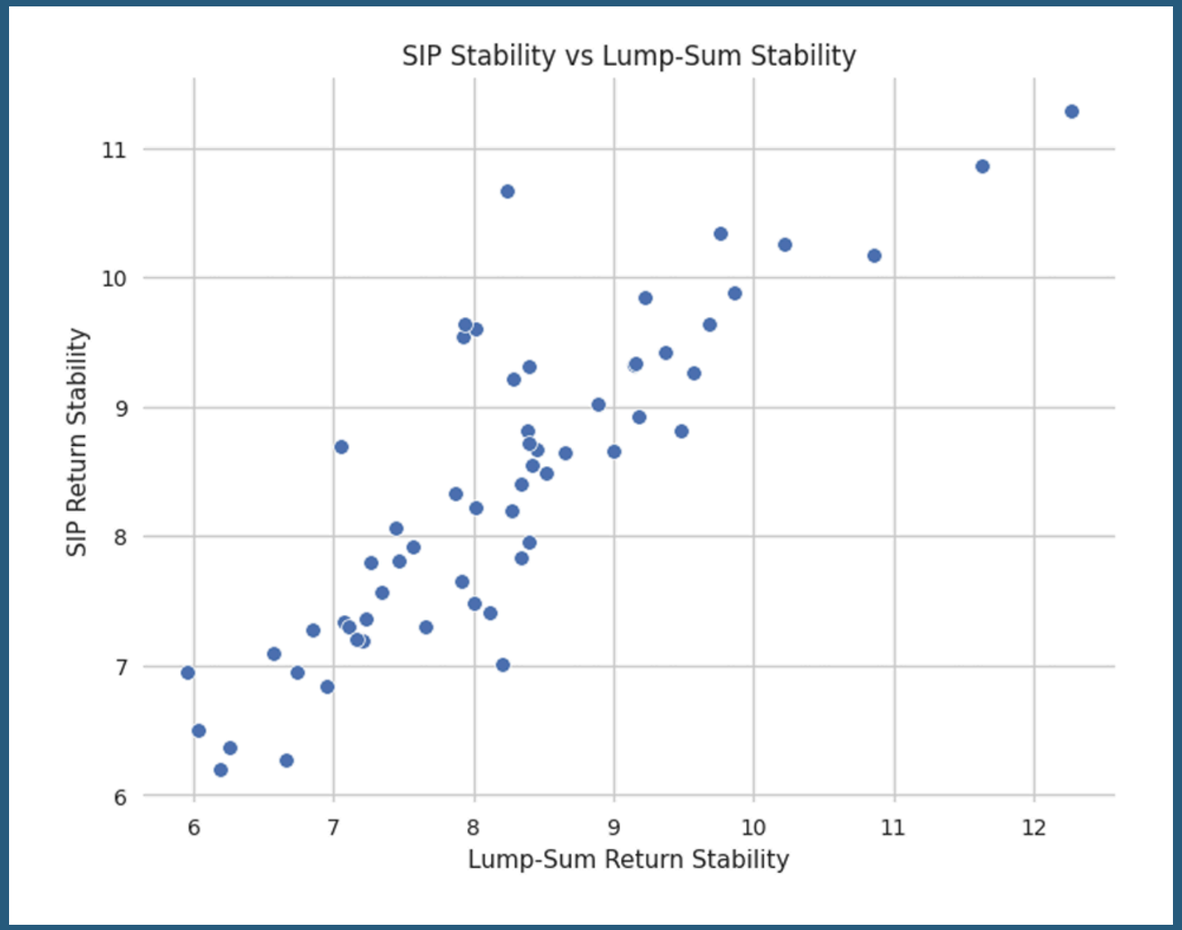
Feature Selection

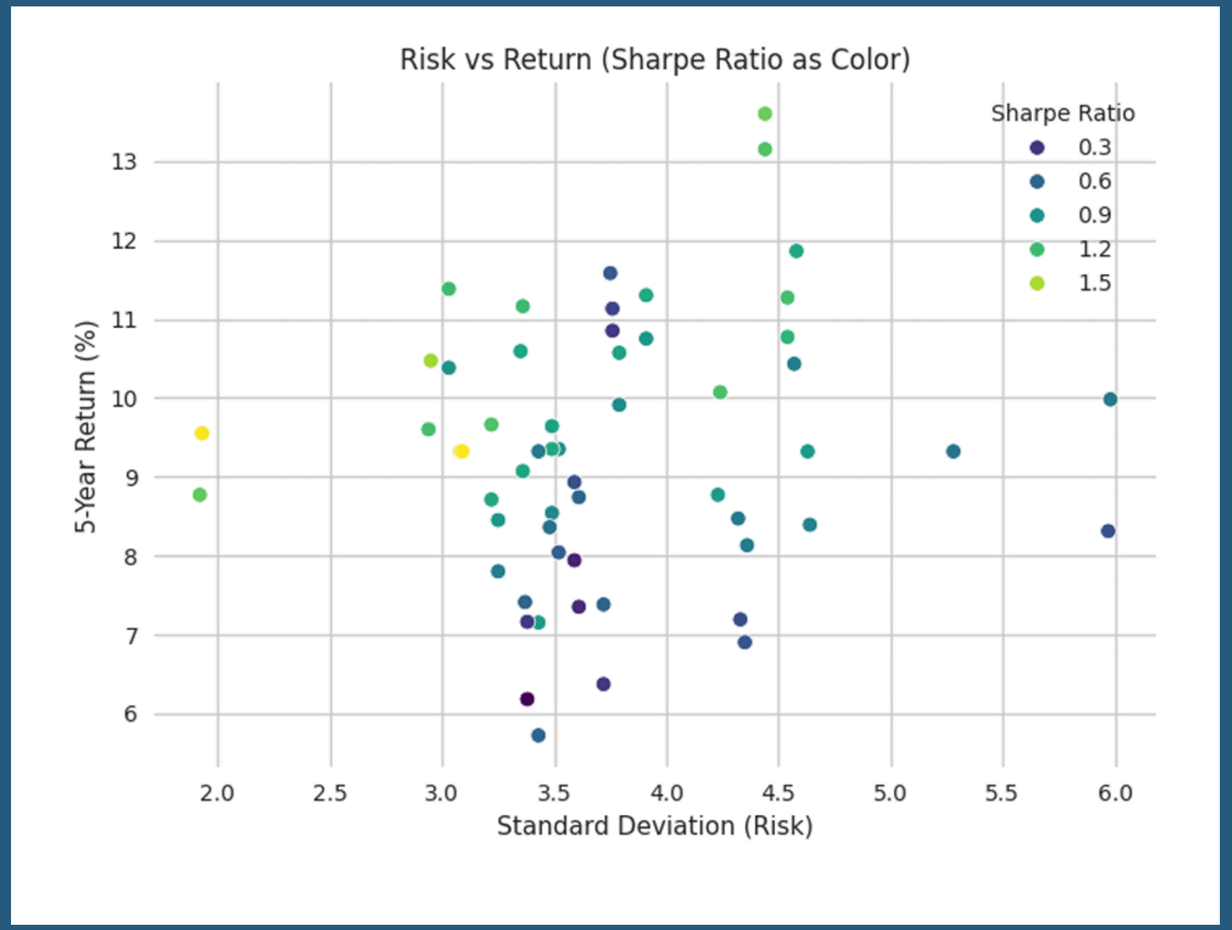
- Performed feature selection using correlation, redundancy (>0.9), and Random Forest importance on the feature-engineered dataset.
- Removed irrelevant, redundant, and low-quality features, keeping only the most predictive numeric variables for modeling.

Feature Visualization

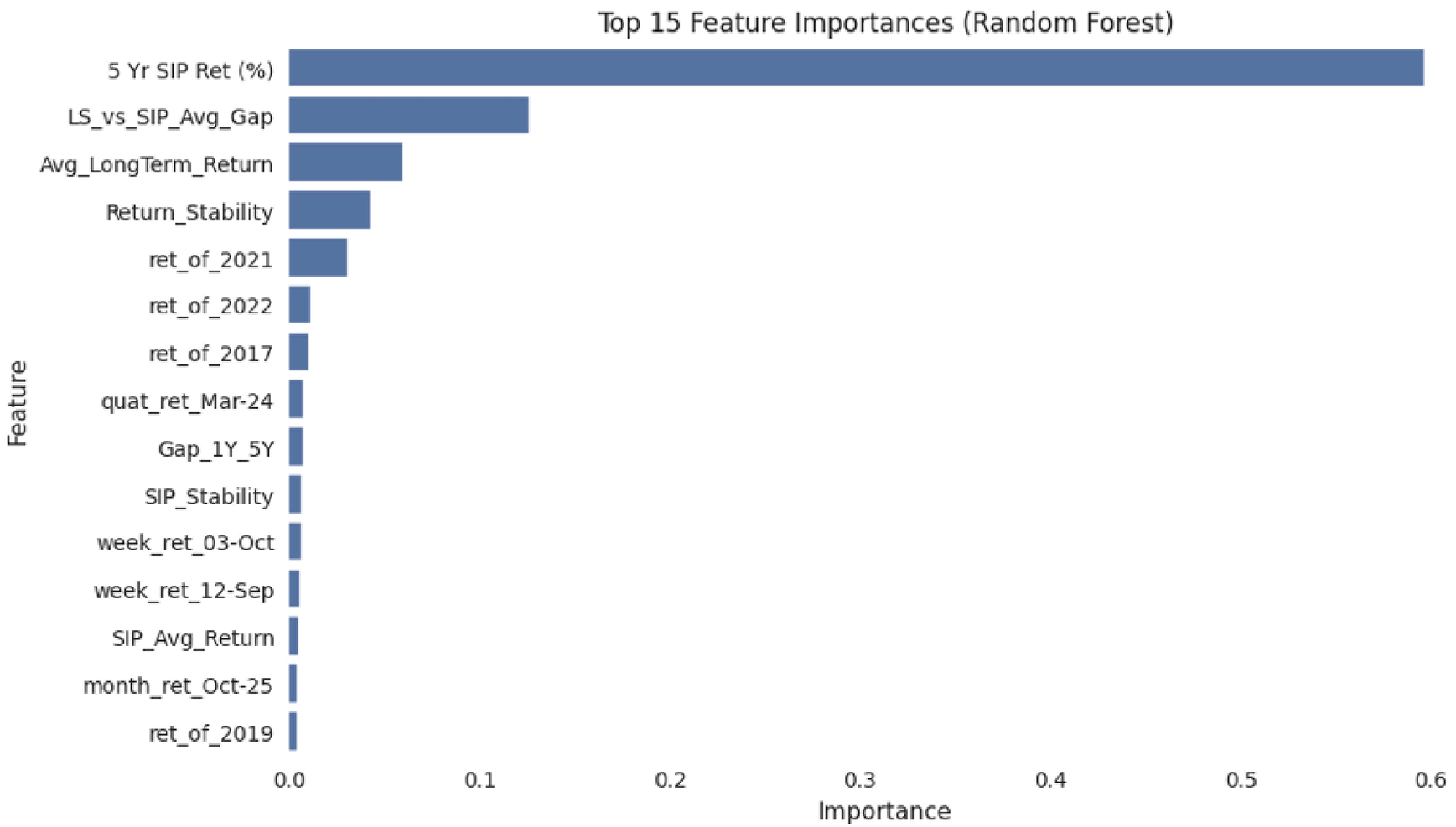
- Plotted heatmaps to view correlations among engineered features.
- Used scatterplots and bubble charts to compare risk-adjusted performance.
- Visualized portfolio features using distributions and composition heatmaps.

Feature Engineering



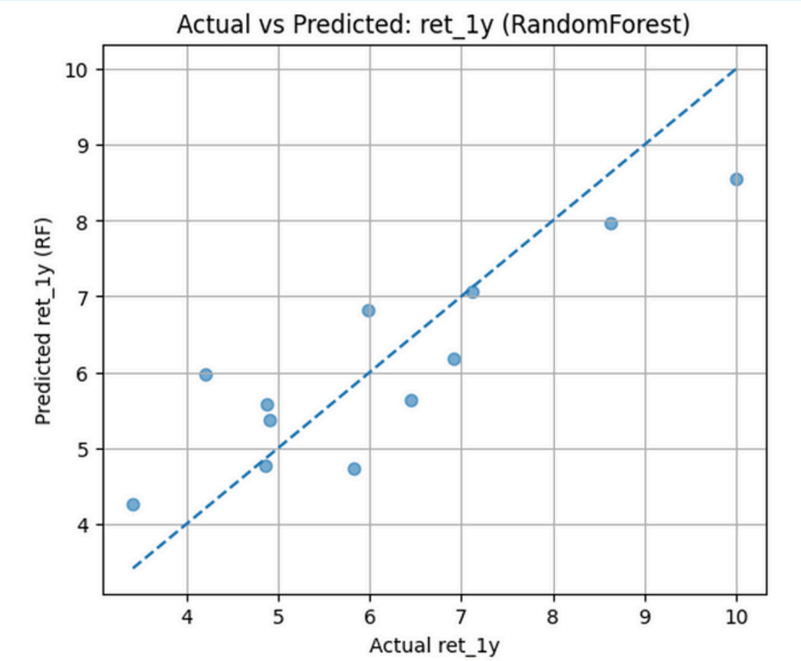


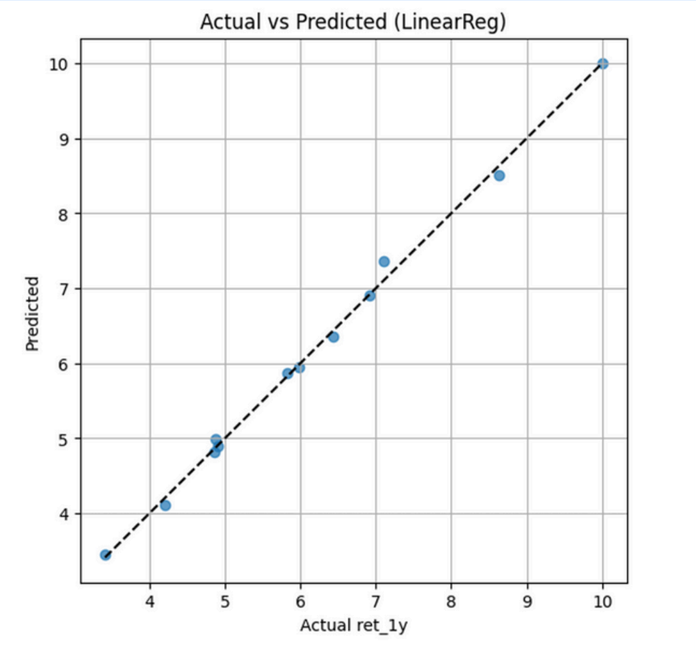
Feature Engineering

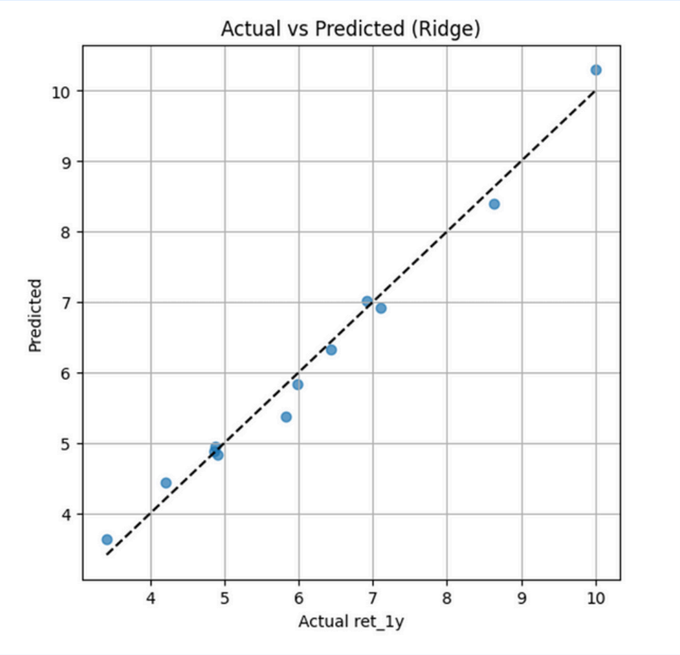


Model Fitting

- Prepared a modeling dataset by selecting key return metrics, risk metrics, NAV features, and engineered features.
- Split data into train–test sets using an 80/20 ratio.
- Built Three regression models to predict 1-Year Return (ret_1y)
  - Random Forest Regressor
  - Ridge Regressor
  - Lasso Regressor

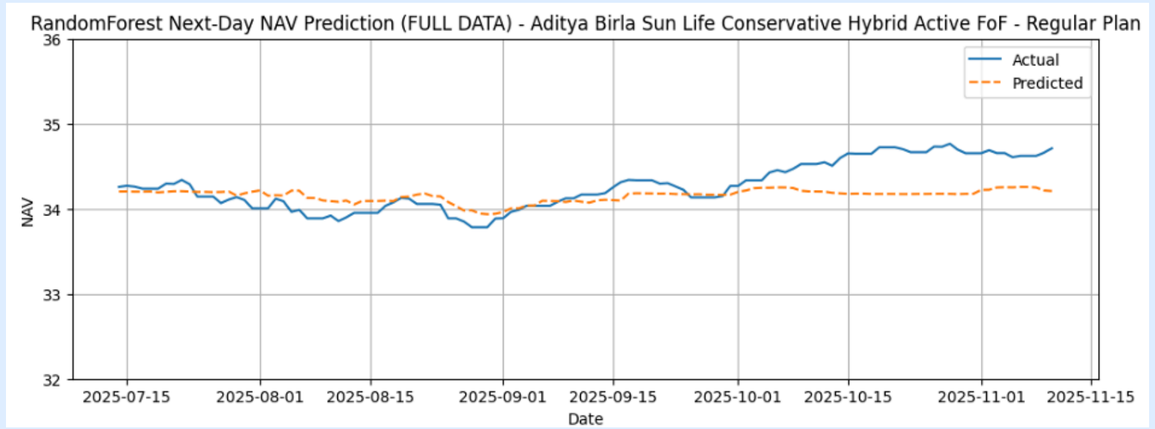


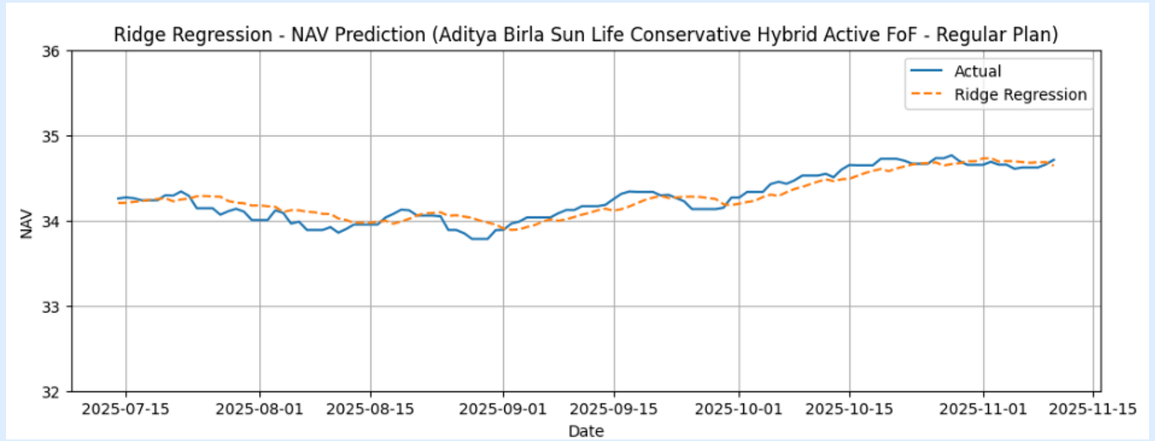


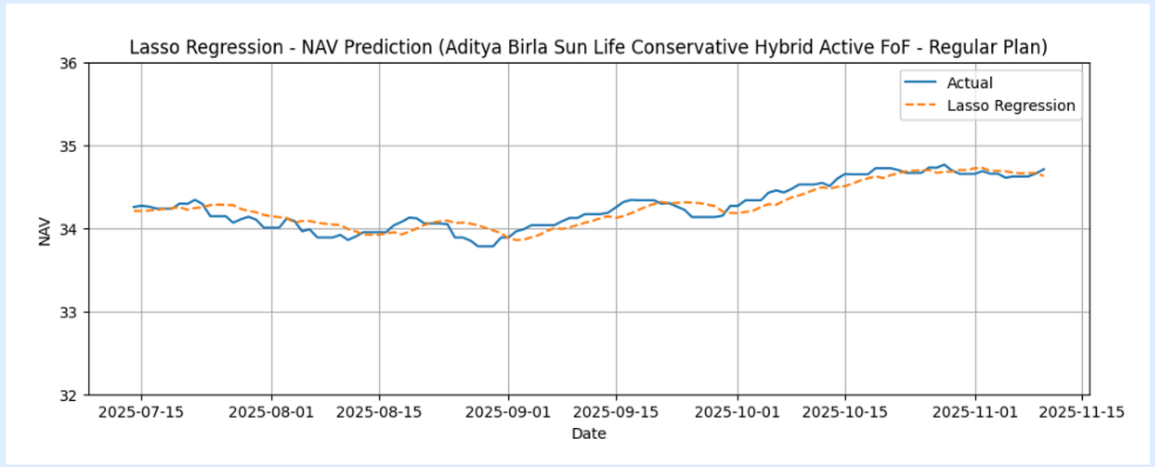


Model Fitting

- Built forecasting models to predict next-day / short-term NAV movement, using:
  - Random Forest Regressor
  - Ridge Regressor
  - Lasso Regressor
- Ridge and Lasso performed better than Random Forest.
- This is because the dataset has very limited rows (~55-65 schemes), where linear models (Ridge/Lasso) handle small sample sizes and low-noise patterns more effectively.
- Random Forest requires large datasets (typically 200+ observations) to learn complex patterns, so it underperformed on our small time-series.







Problem Statement

Our goal is to identify the Top 3 Conservative Hybrid Fund schemes that provide the highest possible returns for the lowest possible level of risk, using risk-adjusted performance metrics.

- To find the Top 3 best schemes, we used the Sharpe Ratio, which measures how much return a fund gives per unit of risk taken.
- We calculated:
  - Fund Return (average return over time)
  - Risk-Free Rate (return with zero risk)
  - Volatility (Standard Deviation)
- Using the formula:
  - Sharpe Ratio = (Fund Return−Risk-Free Rate)/Volatility
- Funds with the highest Sharpe Ratio offer the best risk-adjusted performance.
- So, we ranked all schemes by Sharpe Ratio and selected the Top 3 schemes that provide the highest return for the lowest risk.

Key Insights

- Long-term returns are stable, while short-term returns show higher variation.
- Risk metrics (Std Dev, Sharpe, Beta) strongly influence fund performance.
- Higher expense ratios do not lead to better returns.
- NAV trends show steady growth with limited drawdowns.
- Clustering reveals three portfolio styles: debt-heavy, balanced, and tactical.
- Ridge & Lasso outperform Random Forest for NAV prediction due to small sample size.