MUTUAL FUND ANALYSIS

Data science case study

Shyamlal S

MTECH AI

CB.EN.P2AIE22017

ABSTRACT

In this project, we conducted a data analysis on a mutual funds dataset to gain insights into the performance and characteristics of various mutual funds. The dataset included columns such as scheme_name, min_sip, min_lumpsum, expense_ratio, fund_size_cr, fund_age_yr, fund_manager, alpha, sd, beta, sharpe, risk_level, amc_name, rating, category, sub_category, returns_1yr, returns_3yr, and returns_5yr.

We started by performing exploratory data analysis (EDA) to understand the distribution and relationships of the variables in the dataset. We used histograms, bar plots, box plots, scatter plots, correlation matrices, and other visualizations to analyse the data. Through EDA, we gained insights into the distribution of numerical variables, relationships between variables, trends over time, and comparisons across categories.

Additionally, we utilized linear regression models to analyse the relationship between fund managers and returns. We calculated coefficients for each fund manager and visualized them through bar plots to identify the fund managers with positive and negative impacts on returns.

Throughout the project, we presented our findings through various plots and visualizations, allowing for clear interpretation and communication of the results. We highlighted top-performing funds based on different time periods and risk levels, and we examined the relationship between risk level and returns.

INTRODUCTION

The aim of this project is to perform a comprehensive data analysis on a mutual funds dataset. Mutual funds are investment vehicles that pool money from multiple investors to invest in various securities such as stocks, bonds, and money market instruments. Analysing the performance and characteristics of mutual funds is crucial for investors and financial professionals to make informed investment decisions.

By conducting a thorough analysis of this dataset, we aim to uncover valuable insights and trends related to mutual funds. This analysis will involve exploring the relationships between various factors and the performance of mutual funds, identifying top-performing funds based on returns, assessing the impact of fund managers on fund performance, examining the relationship between risk and returns, and performing hypothesis testing to validate our findings.

DATA ANALYSIS AND FINDINGS

In the initial phase of the project, data pre-processing was performed to ensure the quality and integrity of the dataset. The following steps were undertaken:

Checking Data Types: The data types of the attributes in the dataset were examined to understand the nature of the data. This step helps in identifying whether the attributes are categorical or numerical, which is essential for further analysis.

Handling Null Values: The presence of null values in the dataset was assessed to ensure that the data is complete. Specifically, the attributes "returns_3yr" and "returns_5yr" were found to have null values. To address this issue, the missing values were filled using appropriate techniques such as mean and median imputation. Mean imputation was applied to fill the null values in "returns_3yr", while median imputation was used for "returns_5yr".

Descriptive Statistics: To gain a deeper understanding of the dataset, a descriptive analysis was performed. This involved computing the five-number summary (minimum, 1st quartile, median, 3rd quartile, and maximum) for each attribute in the dataset. This summary provides an overview of the distribution and range of values for each attribute, allowing for better insights into the dataset.

After examining the category distribution using the value_counts() function, it was observed that the dataset is imbalanced, with a higher number of funds in the "Debt" category compared to the other categories. The category distribution is as follows:

- Debt: 409 funds
- Equity: 323 funds
- Hybrid: 128 funds
- Other: 88 funds
- Solution Oriented: 30 funds

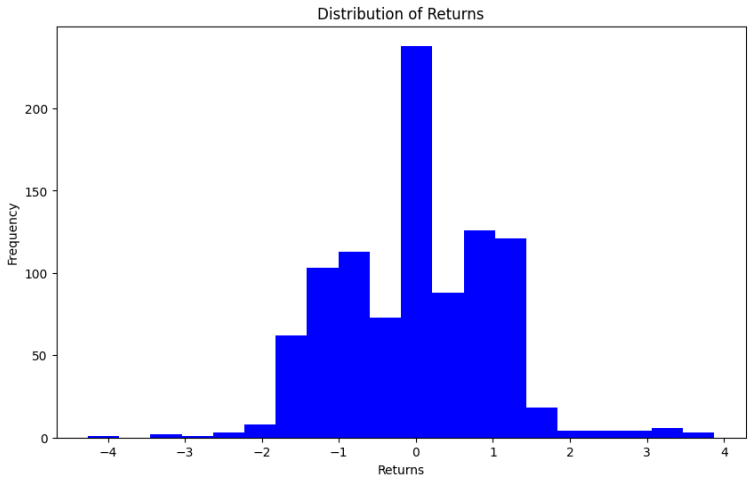
To address the issue of imbalance, under-sampling was performed. Under-sampling involves randomly selecting a subset of the majority class (in this case, "Debt") to match the number of samples in the minority classes (i.e., "Equity" and "Hybrid"). The under-sampling process resulted in an equal number of funds in each category, with 128 funds in each category .we can reduce the bias towards the majority class ("Debt") and ensure that the analysis is not skewed towards any particular category. This balanced dataset can now be used for further analysis and modelling, providing a more fair and unbiased representation of the different fund categories.

Normalization using the StandardScaler was performed on the following columns: 'fund_size_cr', 'returns_1yr', 'returns_3yr', 'returns_5yr'. The StandardScaler scales the data such that each feature has a mean of 0 and a standard deviation of 1.

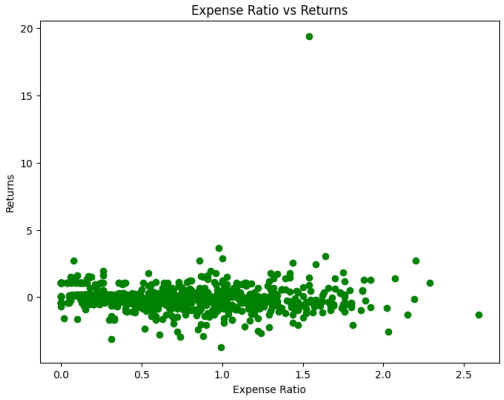
The purpose of normalization is to bring all the features to a similar scale, which is particularly important when performing certain machine learning algorithms or statistical analyses that are sensitive to the scale of the variables. By scaling the features, we ensure that they have a similar influence on the analysis and prevent any particular feature from dominating the results due to its larger magnitude.

fund_size_cr	returns_1yr	returns_3yr	returns_5yr
0.008413	0.148970	4.464590	3.875319
-0.069607	-0.541953	3.057435	3.577274
-0.362107	0.179678	3.620297	3.249426
0.237092	-0.772261	2.793060	0.000000
0.666948	-0.373061	2.034049	2.623532

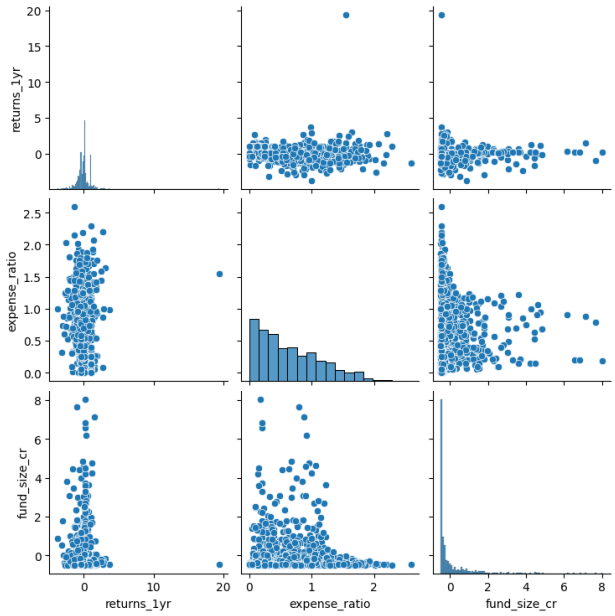
The histogram plot shows the distribution of returns over a 5-year period. The majority of mutual funds seem to have returns within a certain range, with a peak around the center of the distribution. This suggests that there is a cluster of funds with similar returns.



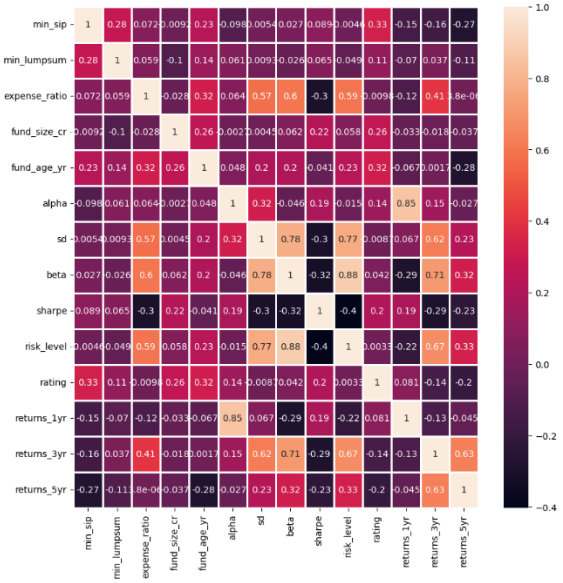
The scatter plot shows the relationship between the expense ratio of mutual funds and their corresponding returns over a 1-year period. There are a few outliers where mutual funds have high returns despite having high expense ratios. These funds may have some unique characteristics or strategies that contribute to their performance.



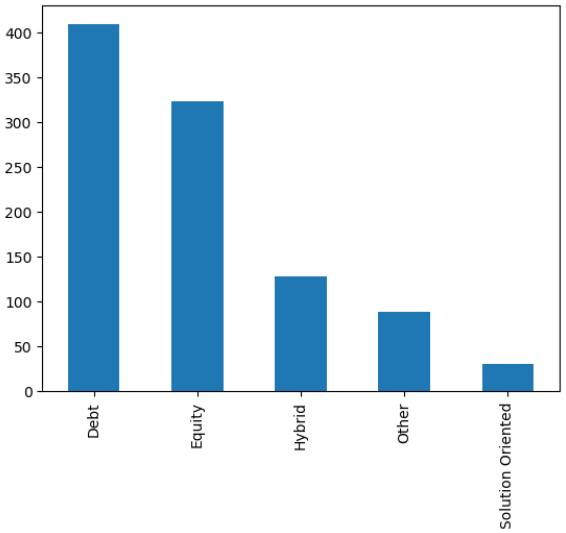
The pair plot provides a visual representation of the relationships between the variables 'returns_1yr', 'expense_ratio', and 'fund_size_cr' through scatter plots and histograms.



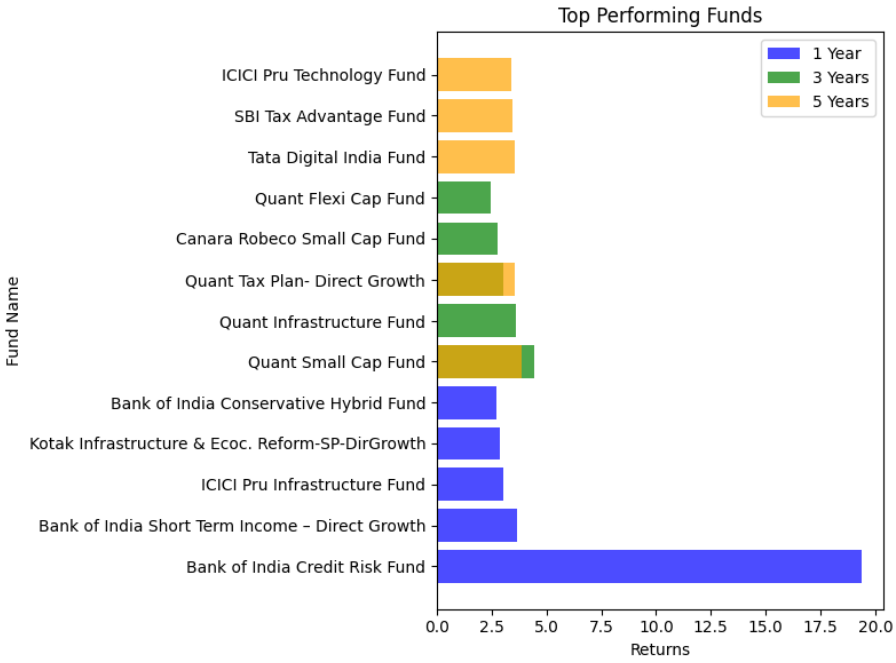
The correlation matrix and heatmap provide insights into the pairwise correlations between different numerical variables in the dataset.



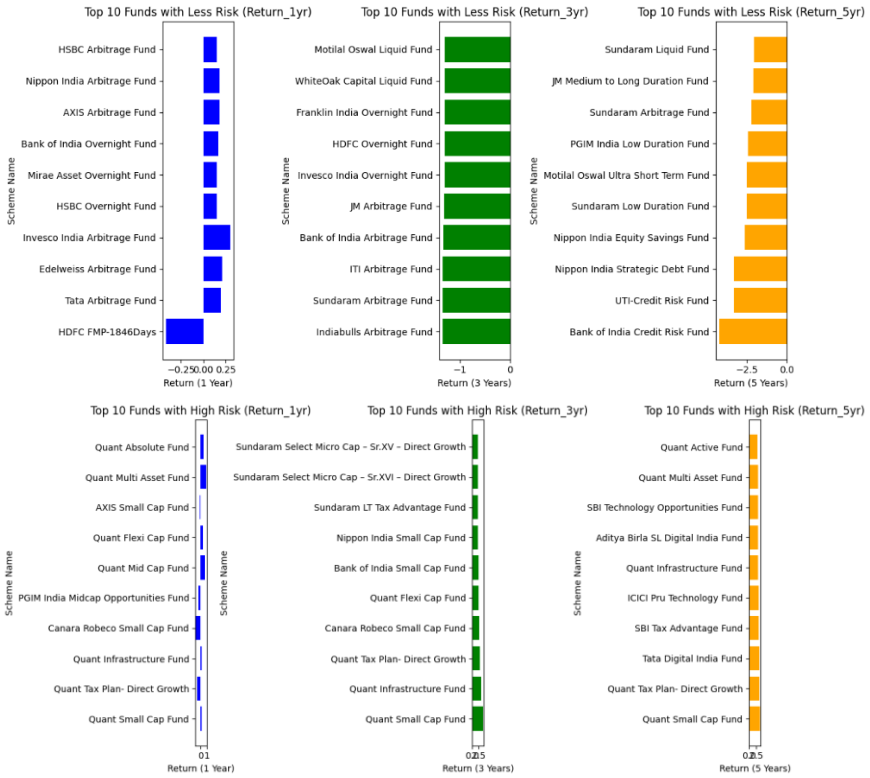
The bar plot of the frequency distribution of the 'category' variable provides insights into the distribution of mutual funds across different categories.



The top performing funds based on returns of all the years will give investors to know which mutual fund is best to invest in for the corresponding years.

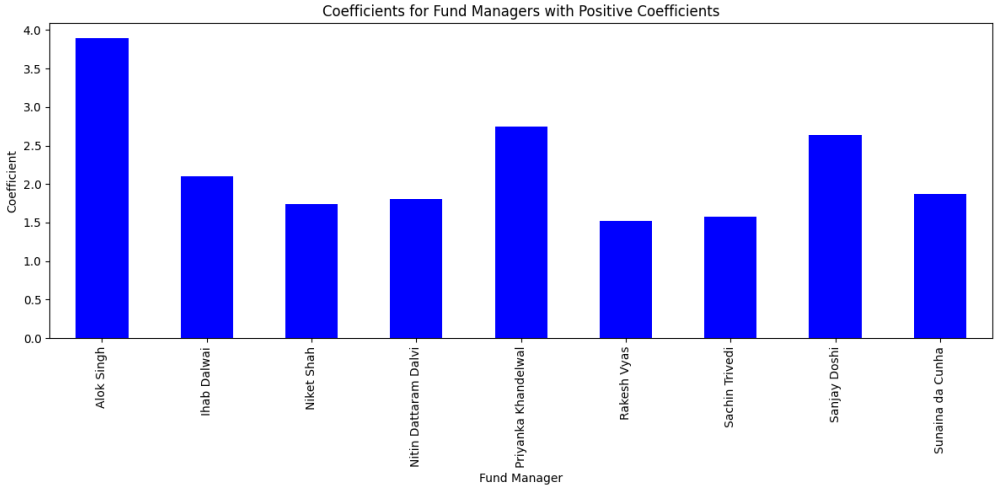


Also the top high performing and top less performing mutual funds based on risk level is analysed. This will help the investors to make right decisions according to their perspective. That is whether they need a less risk mutual fund or even it's a high risk mutual fund they have no problem.



From the analysis, the fund managers with positive and negative coefficients is identified using the linear regression model.

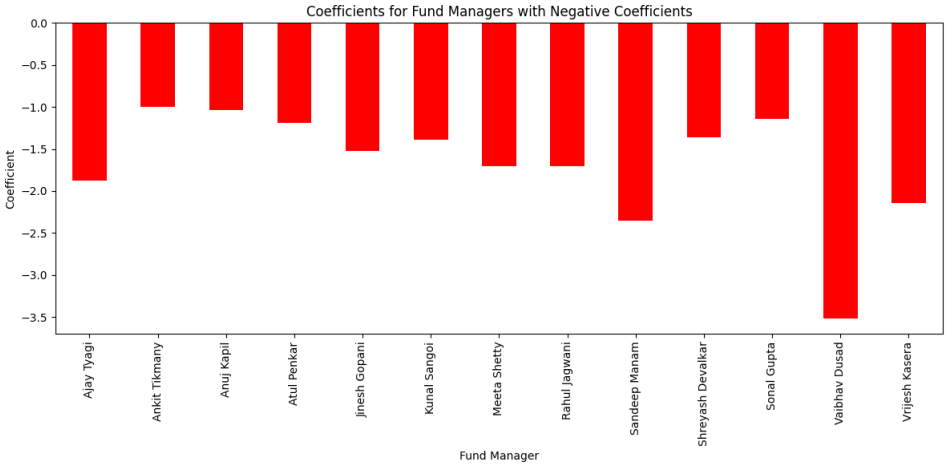
For fund managers with positive coefficients greater than 1.5, it indicates that they have a significant positive impact on the returns of the mutual funds. These fund managers are considered influential in generating higher returns for the funds they manage.



Based on the analysis, the fund managers with positive coefficients (> 1.5) are:

- Alok Singh
- Ihab Dalwai
- Niket Shah
- Nitin Dattaram Dalvi
- Priyanka Khandelwal
- Rakesh Vyas
- Sachin Trivedi
- Sanjay Doshi
- Sunaina da Cunha

On the other hand, fund managers with negative coefficients less than -1 indicate a significant negative impact on the returns of the mutual funds. This suggests that these fund managers may have struggled to generate positive returns for the funds they manage.



The fund managers with negative coefficients (< -1) are:

- Ajay Tyagi
- Ankit Tikmany
- Anuj Kapil
- Atul Penkar
- Jinesh Gopani
- Kunal Sangoi
- Meeta Shetty
- Rahul Jagwani
- Sandeep Manam
- Shreyash Devalkar
- Sonal Gupta
- Vaibhav Dusad
- Vrijesh Kasera

CONCLUSION

In this project, mutual funds dataset is analysed to gain insights for investment decisions. We performed data pre-processing, explored distributions of returns and expense ratios, and identified top-performing funds based on returns and risks. Also, the impact of fund managers on returns and identified positive and negative coefficients is examined. This analysis provides valuable information for investors to make informed choices in selecting mutual funds.