## Introduction:

This project is developed for the needs of investors in Indian Mutual funds market. It considers using association rule mining algorithm and collaborative filtering to provide personalized investment recommendations to users based on their risk profile, investment objectives, and financial preferences. This system assists users in making informed investment decisions in mutual funds. By analyzing transactional data of mutual funds and client attributes, the system aims to match users with the most suitable mutual fund options, thereby optimizing their investment portfolio and maximizing returns.

## Scope:

- Collection and preprocessing of mutual fund data from the Value Research website.
- Generation of synthetic client data.
- Correlation analysis between mutual fund risk factors and client risk profiles.
- Creation of transactional dataset.
- Implementation of association rule mining and item-item collaborative filtering techniques for recommendation generation.

## Dataset description:

The MF data is taken from Value research website, about Indian mutual funds' performance and attributes. The dataset includes funds from various categories, each represented by separate CSV files containing approximately 50 funds of a specific category.

Attributes of the mutual fund data include returns (1 wk, 3 m, 6 m, 1 yr, 3 yr, 5 yr, and 10 yr), market capitalization, turnover, net assets, riskometer, standard deviation, expense ratio, and other relevant metrics.

Client data is randomly generated and labeled based on current requirements. It contains attributes employment status, income range, age, investment horizon, investment objective, and risk appetite rating derived from psychological and market-based questions.

The recommendation process involves matching the risk profiles of MF with those of clients and allocating suitable funds to client baskets. The resulting transactional data is then binary encoded for further processing.

Data cleaning: Filling missing value in MF data.
Data encoding: Categoric to numeric transformation in case of MF categories, riskometer etc.
Data transformation: Binary encoded transactional data for applying asso. rule mining.

## Novelty:

- **Multiple Data Sources**: We used data from MF and client both, with multiple attributes.
- **Recommendation Techniques and Analysis**: Apriori, FP growth and collaborative filtering, to generate personalized investment recommendation.
- **Risk Profile Matching**: A key novelty of our approach is the incorporation of risk profile matching between MF and clients.
- **Dynamic Recommendation Generation:** If new user or MF joins the data, using regression and classifier, values of derived attributes can be predicted.

## Algorithm:

**Algorithm 1** Mutual Fund Recommendation Algorithm

**Require:** Mutual fund data $D_{\text{funds}}$, Client data $D_{\text{clients}}$
**Ensure:** Recommended funds for each client
1: Calculate risk $\rho_f$ for all mutual funds in $D_{\text{funds}}$
2: Calculate risk $\rho_c$ for all clients in $D_{\text{clients}}$
3: **for** each client $C_i$ in $D_{\text{clients}}$ **do**
4:     **for** each mutual fund $MF_j$ in $D_{\text{funds}}$ **do**
5:         Calculate compatibility between $\rho_f$ and $\rho_c$
6:     **end for**
7:     Select top 5 $MF$ for $C_i$
8: **end for**
9: Transaction data $\xrightarrow{\text{convert}}$ Binary encoded data
10: Apply Apriori
11: Apply FP-Growth
12: Time analysis of Apriori and FP-Growth
13: Correlation analysis and Chi-square analysis on the association rules

**Algorithm 2** Item-Item Collaborative Filtering

**Require:** Mutual fund data $D_{\text{funds}}$, Client data $D_{\text{clients}}$
**Ensure:** Recommended funds for each client
Calculate risk $\rho_f$ for all mutual funds in $D_{\text{funds}}$
Calculate risk $\rho_c$ for all clients in $D_{\text{clients}}$
Generate user-item matrix $M$ based on risk matching as in step 3,4,5 in Algo.1
**for** each client $C_i$ in $D_{\text{clients}}$ **do**
    Extract client investments from $M$
    Predicted ratings using item-item collaborative filtering
    Select top 5 MF for client $C_i$
**end for**
Analysis and Inference

**Algorithm 3** PrefixSpan Algorithm

**Require:** Sequences $S$, Minimum support threshold $min\_support$
**Ensure:** Frequent sequential patterns $freq\_patterns$
$freq\_patterns \leftarrow \{\}$
PrefixSpan$projected_db$, $prefix$, $min\_support$
$item\_count \leftarrow \{\}$
**for** each sequence $seq$ in $projected\_db$ **do**
    **for** each item $item$ in $seq$ **do**
        **if** $item$ exists in $item\_count$ **then**
            Increment count of $item$ in $item\_count$
        **else**
            Add $item$ to $item\_count$ with count 1
        **end if**
    **end for**
**end for**
**for** each item $item$, count $count$ in $item\_count$ **do**
    **if** $count \geq min\_support$ **then**
        $new\_prefix \leftarrow prefix \cup \{item\}$
        Add $new\_prefix$ to $freq\_patterns$
        $projected\_db\_new \leftarrow \{seq[seq.index(item) + 1 :]$ for $seq$ in $projected\_db$ if $item$ in $seq\}$
        **if** $projected\_db\_new$ is not empty **then**
            PrefixSpan($projected\_db\_new$, $new\_prefix$, $min\_support$)
        **end if**
    **end if**
**end for**
PrefixSpan($S$, $\{\}$, $|S| \times min\_support$)

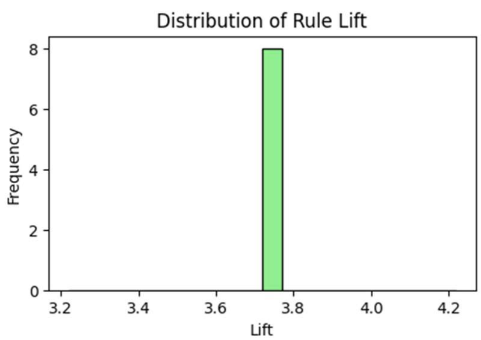
## Results and Discussion:

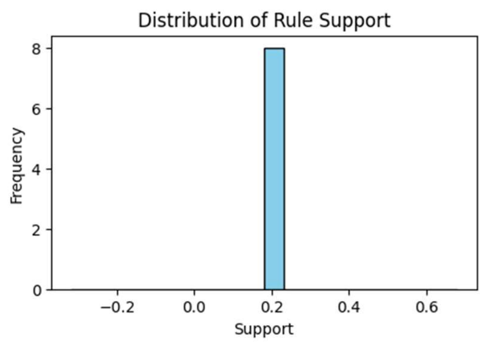
Time Analysis:
FP-Growth generally outperformed Apriori in terms of execution time, especially for larger datasets and lower support thresholds. This can be attributed to the efficient data structure (tree) used by the FP-Growth algorithm, which reduces the computational overhead associated with candidate generation and pruning.

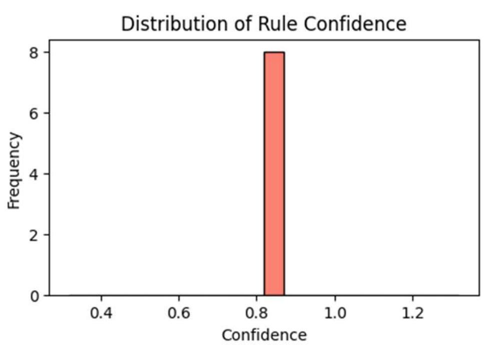
Correlation analysis:
Correlation coefficients were computed for pairs of mutual funds based on their co-occurrence in transactions. The results were visualized using a heatmap, highlighting strong positive and negative correlations between funds.

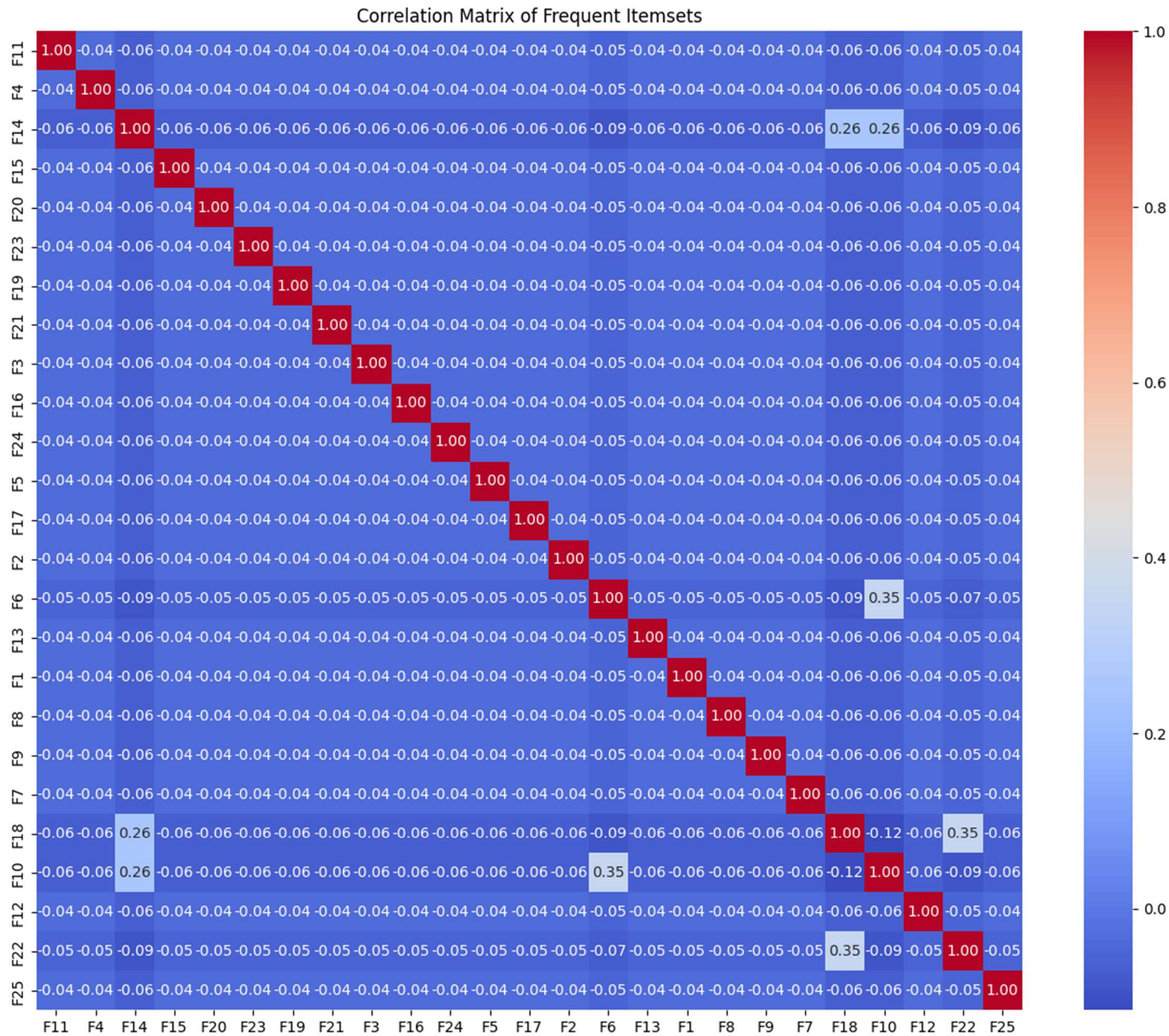
Chi-Square analysis:
It was performed to assess the statistical significance of association rules. Contigency tables formed by the presence and absence of mutual funds in transactions. The resulting p-values were used to determine the strength of association between funds, with lower p-values indicating higher significance.

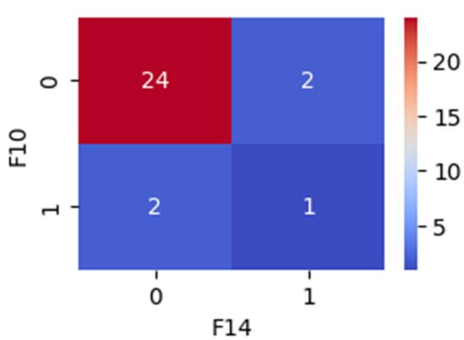
Plots:



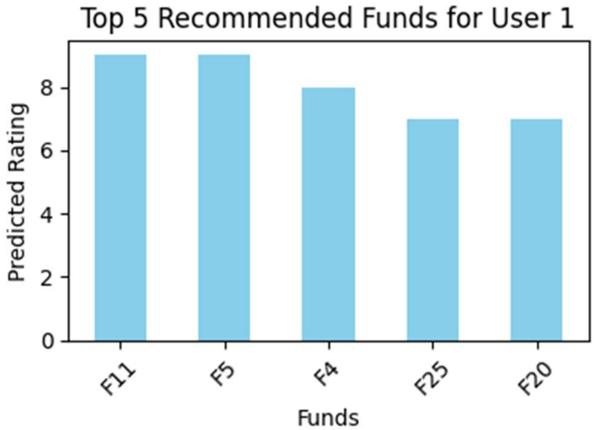



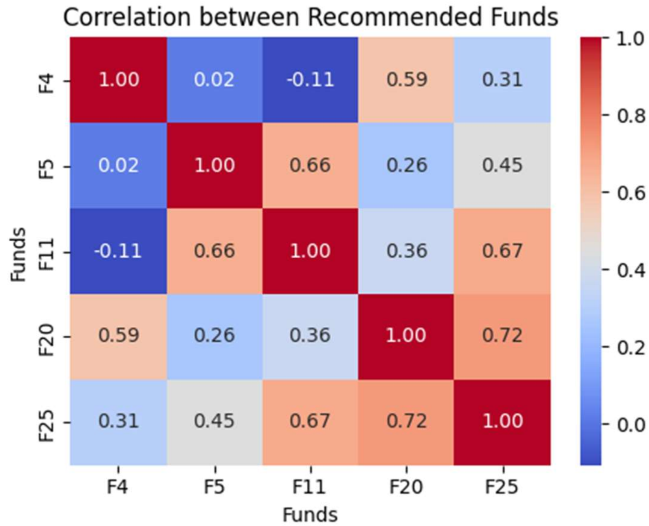






Contigency table for F10 and F14




RMSE for predicted ratings:

```
rmse = calculate_rmse(df_nan, ratings)
print("RMSE:", rmse)
```

```
RMSE: 3.866908285771863
```

Overall, the results demonstrate the effectiveness of the Mutual Fund Recommendation System in generating meaningful investment recommendations based on association rule mining techniques. Further refinement and optimization of the system can be DONE to enhance its accuracy.