INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

Web Scraping and Exploratory Data Analysis of Mutual Funds - Groww

Web Scraping, Data Cleaning, and Exploratory Data Analysis

Kesarapu Teja - 81021004325

Introduction

INNOMATICS RESEARCH LABS

What is a Mutual Fund?

- A mutual fund is a pool of money collected from multiple investors.
- This money is managed by professional fund managers.
- The fund is invested in:
  - Stocks
  - Bonds
  - Gold
  - Money market instruments
- Investors earn returns based on fund performance.

Key Characteristics

- Diversification reduces risk.
- Professionally managed portfolios.
- Suitable for both small and large investors.
- Returns depend on market performance.

Problem Statement

- Thousands of mutual funds are available in the market.
- Investors often rely only on ratings without deeper analysis
- Risk and returns must be evaluated together for proper decisions
- Manual comparison of hundreds of funds is difficult
- Financial data is scattered and not easy to interpret visually
- Need for a structured and analytical approach to evaluate funds
- Data analytics helps transform raw financial data into meaningful insights.

Objectives

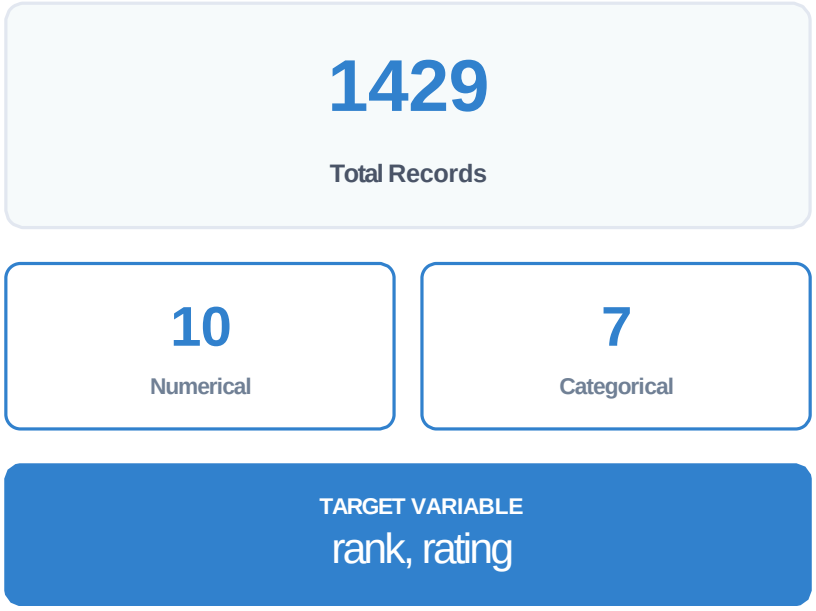
- Collect real-world mutual fund data using web scraping.
- Clean and pre-process raw financial datasets.
- Perform exploratory data analysis (EDA).
- Analyze relationships between:
  - Risk
  - Returns
  - Ratings
- Identify patterns and trends in fund performance.
- Present insights using data visualization.

Why Data Analytics in Finance?

- Large volume of financial data generated daily.
- Investors need data-driven decision-making tools.
- Analytics helps:
  - Detect trends
  - Compare investments
  - Reduce decision bias

Dataset Overview

- **Data source:** Groww
- **Initial Records:** 1643; **Variables:** 7 (numeric: 1, categorical: 6)
- **Final Records:** 1429; **Variables:** 17 (numeric: 10, categorical: 7)
- **Engineered features:** average return, risk score, risk-adjusted return, rank
- **Target variable:** rank, rating
- **Dataset type:** Structured financial performance data



Technology Stacks

Programming Language:

- Python

Development Environment

- Jupyter Notebook

Libraries Used:

- Requests – sending HTTP requests to webpages
- BeautifulSoup – extracting structured data
- HTML
- Pandas – data cleaning and transformation
- NumPy – numerical operations
- Matplotlib, Seaborn – visualization
- Regex – Match pattern in the text

Data Cleaning & Preparation

- **Parsed text fields:** Converted 1Y, 3Y, 5Y returns from percentage strings to numeric values; standardized risk and category fields.
- **Missing values:** Missing long-term returns handled using median imputation, where required. Median used instead of mean because financial metrics are often skewed.
- **Duplicates:** Checked and removed duplicate fund records.
- **Outliers:** Extreme return values were examined through distribution analysis, and the outliers were retained since, in mutual fund studies, they represent meaningful data rather than noise.
- **Result:** Clean, structured dataset ready for exploratory data analysis.

MINIMAL CODE EXAMPLE

```
# Convert percentage columns
df['1y_return'] = df['1y_return']
    .str.replace('%','')
    .astype(float)

df['3y_return'] = df['3y_return']
    .str.replace('%','')
    .astype(float)

# Handle missing values
df.fillna(df.median(numeric_only=True), inplace=True)

# Remove duplicates
df.drop_duplicates(inplace=True)
```

Feature Engineering

INNOMATICS RESEARCH LABS

- Calculated **average return**
- Created **risk score mapping**
- Calculated **risk-adjusted return**
- Created **return bands** for performance segmentation
- Computed **return variability (standard deviation)**

Explanation

Feature engineering helps transform raw financial data into more meaningful variables. These derived features improve comparison between funds and enable deeper analytical insights.

Exploratory Data Analysis Approach

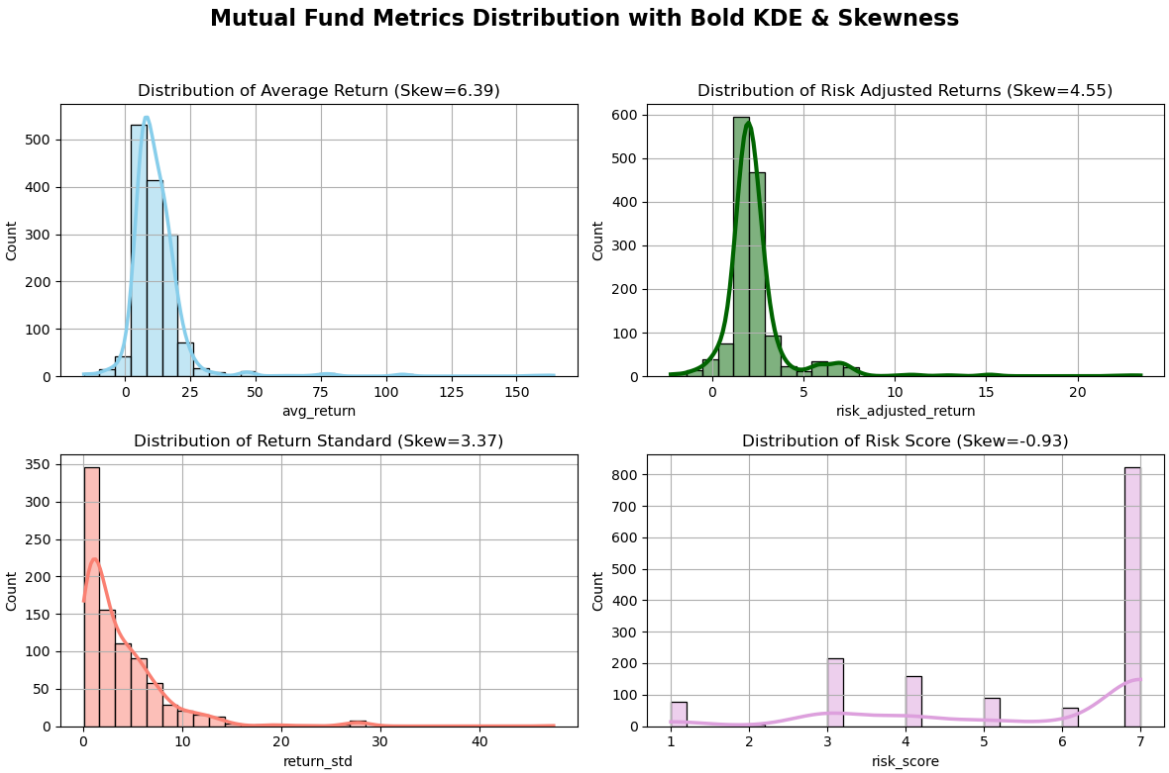
Analysis Types

- Univariate analysis
- Bivariate analysis
- Category-wise comparisons

Focus Areas

- Risk distribution
- Return patterns
- Category performance

Mutual Fund Metrics Distribution with Bold KDE & Skewness (Univariate Analysis)

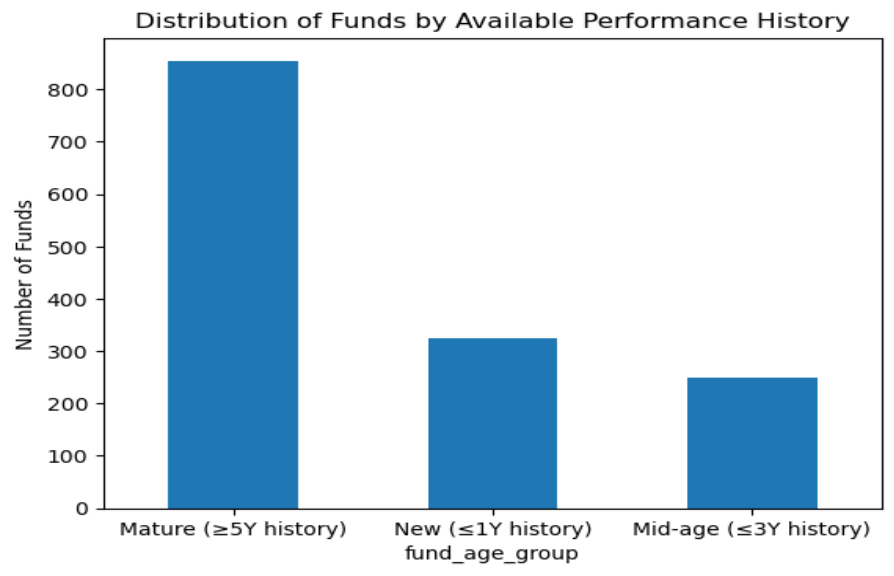


Key Insight

- **Outlier-driven performance:** Both average and risk-adjusted returns are skewed by a small number of standout funds, while most funds deliver modest outcomes
- **Volatility concentration:** Most funds are relatively stable, but a minority carry extreme volatility, which investors should watch closely.
- **Risk profile imbalance:** The negative skew in risk scores shows that the fund universe leans toward higher risk categories, limiting safe options.

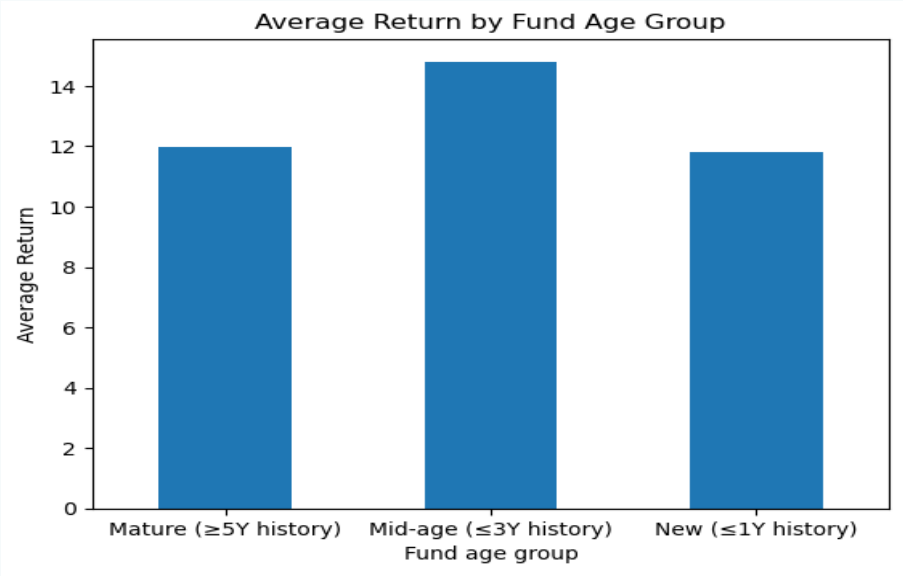
- The distributions reveal that while most funds deliver modest returns and moderate volatility, a small set of outliers drive extreme performance, and the overall fund universe skews toward higher risk, making careful selection essential.

Distribution of Funds by Available Performance History



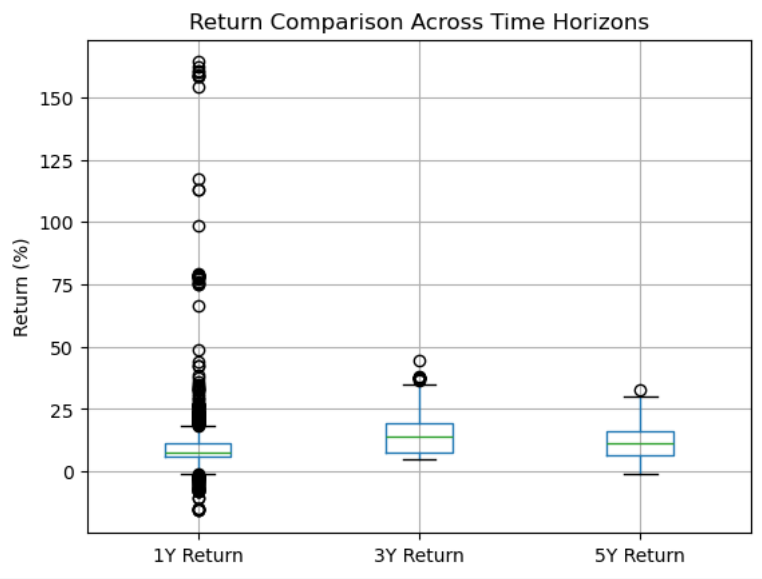
- Mature funds (≥5Y history) dominate the dataset, with over 850 funds.
- New funds (≤1Y history) are the second largest group, around 330 funds.
- Mid-age funds (≤3Y history) are the smallest group, about 260 funds.

Average Return by Fund Age Group



- **Performance sweet spot:** Mid-age funds appear to be in a “sweet spot” — they’ve survived the initial volatility of new funds and are still agile enough to capture growth opportunities.
- **Stability vs. growth trade-off:** Mature funds provide stability and reliability due to their long track record, but their average returns suggest they may be less aggressive in chasing growth compared to mid-age funds.
- **New fund risk:** New funds underperform on average, likely due to lack of established strategies, higher start-up costs, or market testing phases. This reinforces the need for caution when investing in funds with very limited history.

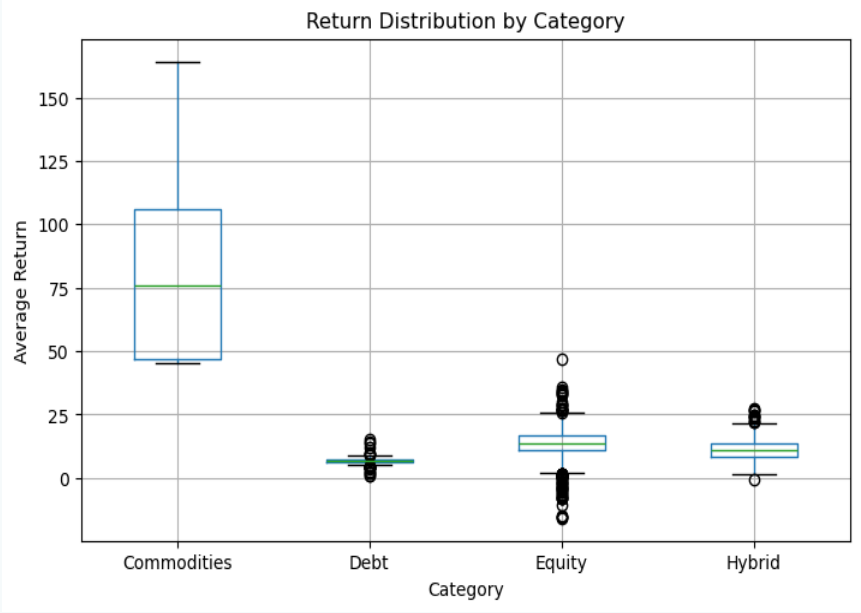
Return Comparison Across Time Horizons



Insights:

- **Volatility decreases with time:** The box plots clearly show that short-term returns are erratic, while medium- and long-term horizons smooth out extremes.
- **Risk–reward trade-off:** 1Y investing offers the chance of outsized gains but carries high risk. Longer horizons reduce risk but also temper the possibility of extreme upside.
- **Consistency improves with maturity:** By 5Y, returns cluster tightly around the median, making long-term investing more reliable.

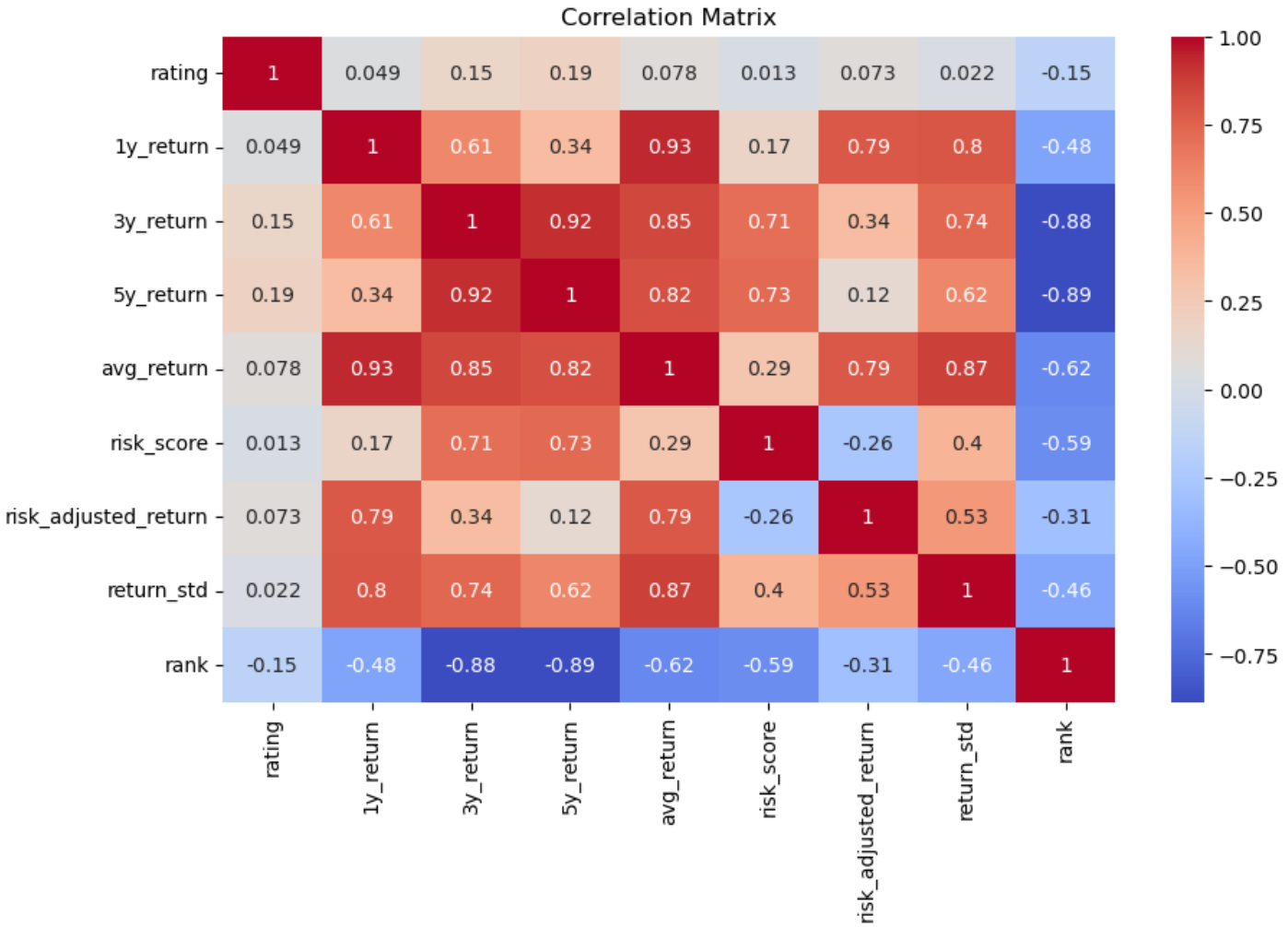
Return Distribution by Category



Insights:

- **Risk–return trade-off across categories:** Commodities deliver the highest returns but also the widest variability, while Debt offers stability at the cost of growth.
- **Equity vs Hybrid:** Equity provides stronger median returns but with more volatility; Hybrid smooths some of that risk while sacrificing a bit of upside.

What factors influence fund ranking?

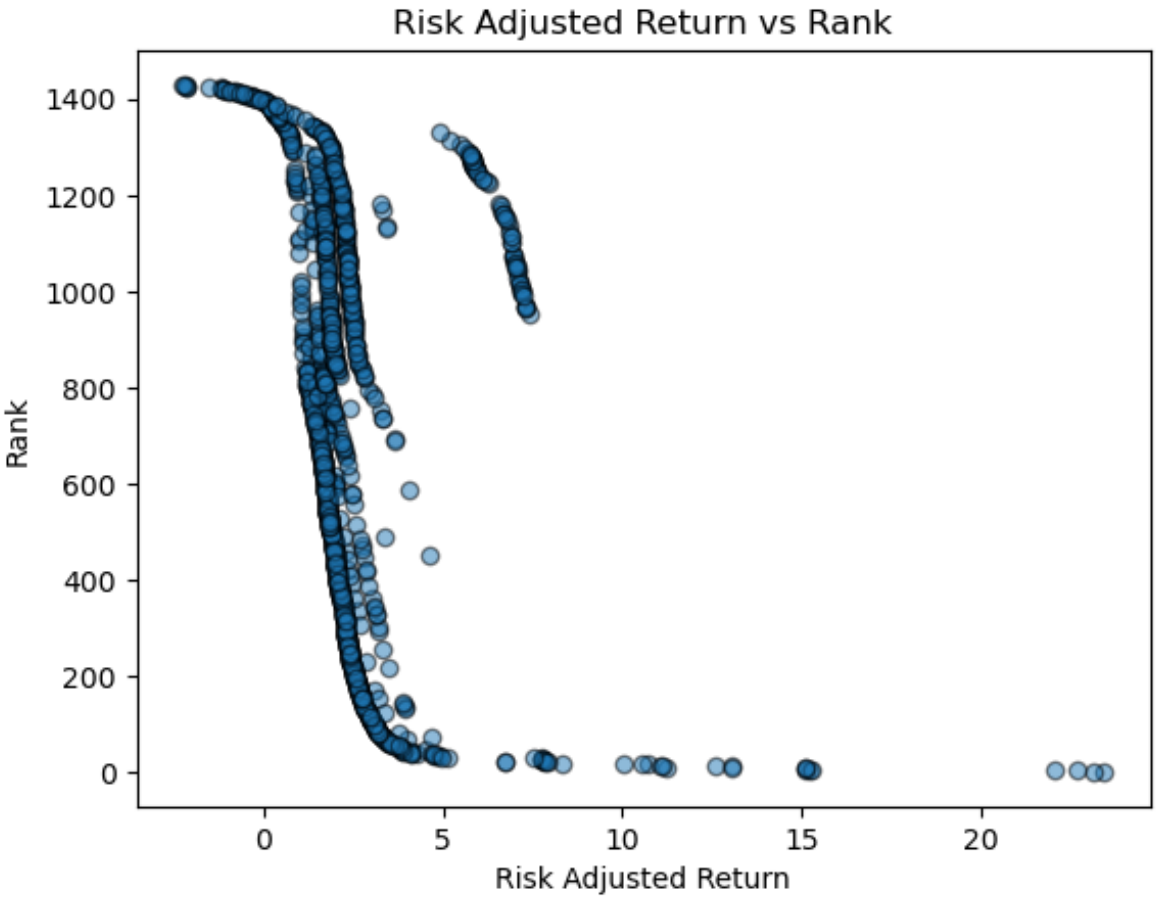


Insights:

- **Risk–reward trade-off:** Higher risk scores align with both higher average returns and higher volatility, reinforcing the classic risk–return principle.
- **Return consistency:** The high correlations among 1Y, 3Y, and 5Y returns suggest that funds performing well in one horizon often perform well across others.
- **Risk-adjusted efficiency:** Some funds manage to deliver strong returns while maintaining favourable risk-adjusted scores, making them particularly attractive.

“The correlation matrix highlights that medium-term returns are the most reliable predictor of long-term success, risk scores align closely with volatility, and fund rankings effectively capture performance, while short-term returns remain noisy and less dependable.”

Which mutual fund characteristics drive higher rank and better risk-adjusted performance?



Insights:

- **Ranking credibility:** The strong inverse relationship shows that the ranking methodology is consistent better risk-adjusted performers are rewarded with superior ranks.
- **Market skew:** The dense clustering at low returns highlights that most funds fail to deliver strong risk-adjusted performance, making top performers relatively rare.

❑ **Dense cluster at low risk-adjusted returns with high ranks:**

- Many funds sit in the lower risk-adjusted return zone and correspondingly have weaker ranks.
- This suggests that a large portion of funds are average or underperforming when risk is considered.

❑ **Clear inverse relationship:**

- As risk-adjusted return increases, rank decreases (improves).
- This validates the ranking system as being aligned with efficiency in balancing risk and reward.

INNOMATICS RESEARCH LABS

1–5Y Return Band Comparison

INNOMATICS RESEARCH LABS

Key Observations:

1-Year Returns:

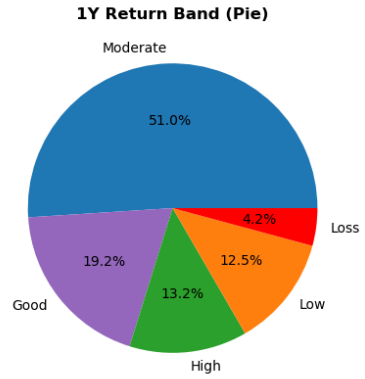
- Majority of funds fall into the Moderate band (51%).
- Smaller but notable shares in Good (19%) and High (13%) bands.
- Losses are present (≈4%), showing short-term volatility.

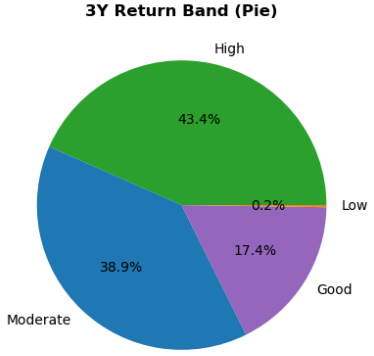
3-Year Returns:

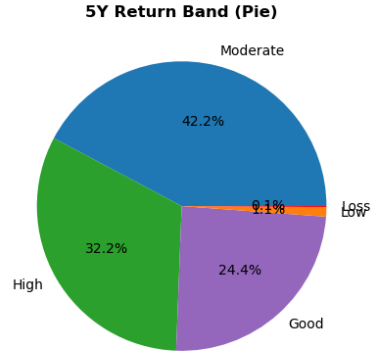
- Strong shift toward High band (43%), with Moderate (39%) close behind.
- Good band (17%) remains meaningful.
- Losses almost disappear, showing improved stability over time.

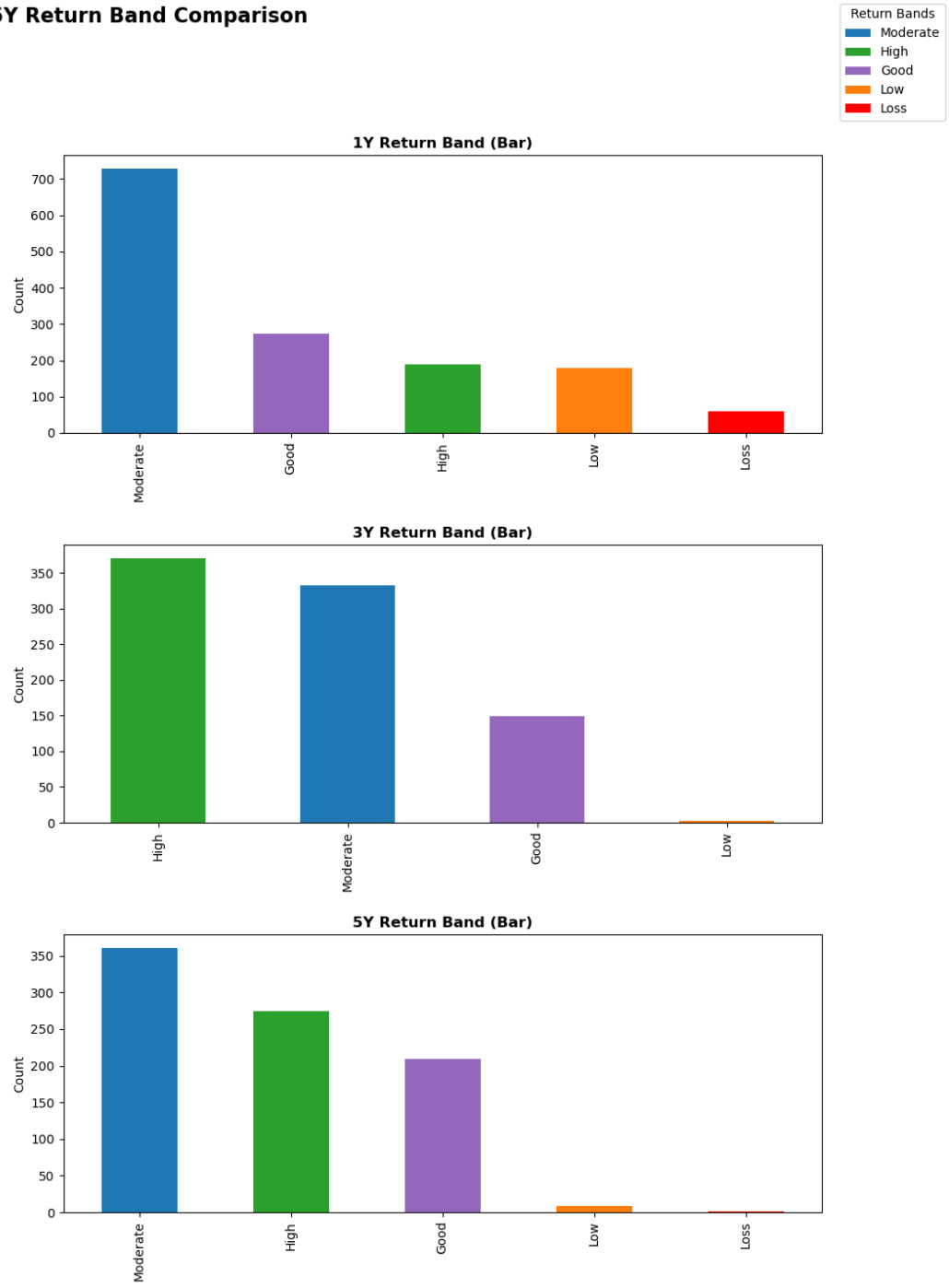
5-Year Returns:

- Distribution balances out: Moderate (42%), High (32%), and Good (24%).
- Losses are negligible (≈0.1%), confirming long-term resilience.
- Very few funds in the Low band, indicating most achieve at least moderate returns.





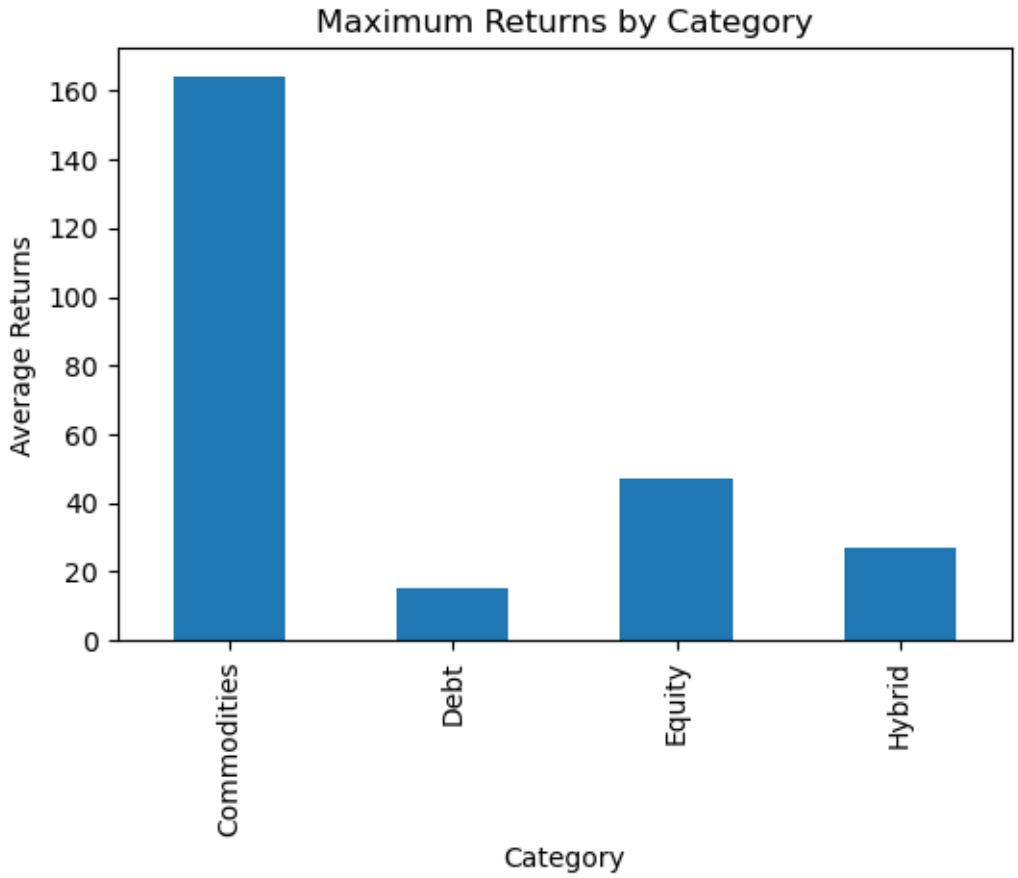




Finally, the most Investor asked question, which fund to invest?

Insights:

- **Risk vs. Reward:** Commodities may look attractive, but their extreme outperformance often comes with significant risk. Equities and hybrids provide more balanced options.
- **Portfolio Diversification:** A mix of equity and hybrid could offer a safer growth path, while commodities can be used selectively for aggressive growth.
- **Investor Profile Fit:**
  - Conservative → Debt/Hybrid
  - Moderate → Equity/Hybrid



Commodities dominate:

- With average returns around 165, commodities far outperform all other categories. This suggests high potential gains but likely comes with higher volatility and risk.

Equity is the next best:

- At ~47, equities provide solid returns, though far below commodities. This aligns with the general expectation that equities outperform debt and hybrid instruments over time.

Hybrid offers moderate returns:

- Averaging ~28, hybrids balance risk and reward, making them suitable for investors seeking stability with some growth.

Debt is the lowest:

- At ~17, debt instruments provide the most stable but least lucrative returns, consistent with their role as safe, low-risk investments.

Key Insights

Performance Sweet Spot

Mid-age funds appear to be in a performance "sweet spot." They have moved beyond the volatility of newly launched funds while still maintaining flexibility to capture growth opportunities.

Fund Lifecycle Insight

Stability vs Growth Trade-off

Mature funds provide stability and reliability due to their long track record. However, their average returns indicate they may be less aggressive in pursuing high growth compared to mid-age funds.

Risk-Return Balance

New Fund Risk

New funds tend to underperform on average. This may be due to limited historical data, evolving strategies, or higher initial operational costs.

Investor Caution

Risk and Return Relationship

Higher-risk funds generally produce higher returns, reinforcing the fundamental financial principle of risk-return trade-off.

Financial Principle

Category Performance Differences

Commodity funds dominate the high-return segment, equity funds provide solid growth potential, while debt and hybrid funds deliver steadier but more moderate performance.

Category Analysis

Challenges faced

Web Scraping Issues

Website structure varied across pages and some elements were dynamically loaded. Pagination loops and HTML inspection were used to reliably extract fund details.

Scraping Logic

Large Dataset Handling

Scraping produced large datasets across multiple pages. Vectorized Pandas operations and CSV storage improved performance and memory usage.

Optimization

Inconsistent Formatting

Category and risk labels varied in formatting. Text normalization and standardization ensured uniform categorical values.

Preprocessing

Missing Values

Several funds had missing long-term returns. Median imputation and selective handling ensured dataset consistency.

Data Cleaning

Percentage Conversion

Return columns were stored as percentage strings. String replacement and numeric casting enabled statistical analysis.

Data Transformation

Applications and future scope

INNOMATICS RESEARCH LABS

Applications

- Investor decision support using performance metrics
- Portfolio comparison and screening tools
- Financial dashboards for visualization and insights

Real-World Use Cases

Future Scope

- Predictive modeling for fund ranking
- Interactive dashboards (Power BI)

Next Steps

Conclusion

This project demonstrated a complete data analytics workflow—from web scraping and data cleaning to exploratory analysis and insight generation. The analysis revealed meaningful patterns in mutual fund performance and highlighted the importance of risk-adjusted evaluation in financial decision-making.

Process

End-to-end pipeline including scraping, cleaning, feature engineering, and visualization.

Insights

Identified relationships between risk, returns, and fund lifecycle patterns.

Impact

Demonstrated how analytics can support investor decision-making and portfolio evaluation.

Thank You

tejakesarapu@gmail.com

github.com/TEJAKESARAPU

linkedin.com/in/tejakesarapu

INNOMATİCS RESEARCH LABS