# Mutual Fund Sector-wise Analysis

Ashwini Dubbewar
Computer Science
University of Colorado Boulder
USA
asdu5808@colorado.edu
Student Id: 111351795

Dhruv Pathak
Computer Science
University of Colorado Boulder
USA dhpa3185@colorado.edu
Student Id: 111364893

Kshitij Jadhav
Computer Science
University of Colorado Boulder
USA
ksja2308@colorado.edu
Student Id: 111369432

## ABSTRACT

This report conducts a comprehensive sector-wise analysis of mutual funds using advanced data mining techniques. By leveraging clustering, classification, and regression, the analysis identifies patterns, trends, and outliers across sectors such as Technology, Healthcare, Energy, and Financials. Our findings offer actionable insights into mutual fund diversification, risk management, and sector rotation strategies, with a focus on metrics like risk-adjusted returns, volatility, and sector-specific performance.

## INTRODUCTION

Mutual funds serve as a versatile investment tool, offering exposure to various economic sectors. Sector-wise analysis reveals how industry-specific trends affect fund performance, particularly during market cycles. Defensive sectors like Healthcare and Utilities outperform during downturns, while cyclical sectors like Technology and Energy excel during growth periods.

Our goal is to utilize data mining techniques to uncover trends, anomalies, and actionable insights, thereby empowering investors with sector-based decision-making tools.

## RELATED WORK

### 1. Previous Studies

Sector Exposure and Performance: Mutual funds exposed to defensive sectors exhibit lower volatility and better performance during downturns. High-risk sectors like Technology yield higher returns but with increased volatility.

Sector-Specific Risk Factors: Regulatory changes, technological advancements, and commodity prices significantly impact sector performance.

Sector Allocation vs. Stock Selection: Superior sector allocation enhances overall performance, underscoring the importance of expertise in mutual fund management.

### 2. Methodologies Used in Prior Studies

Performance Attribution Analysis: Differentiates sector exposure effects from stock selection.

Multi-Factor Models: Evaluates sector contributions to fund risk and returns.

Regression Models: Examines the relationship between sector allocation and fund performance.

## PROJECT MOTIVATION

A significant gap exists in risk-based mutual fund analysis at the sector level. This project aims to fill that void by building a robust framework for sector-wise performance evaluation using data mining techniques. Our objectives include:

- Analyzing large- and mid-cap datasets.
- Identifying sectoral trends, risks, and opportunities.
- Providing actionable insights for investors.

## UNDERSTANDING DATASET

Market Capitalization: Market Capitalization, often referred to as market cap, is the total value of a company's outstanding shares of stock. It is a measure used to determine a company's size and overall market value.

Market Capitalization=Total Number of Outstanding Shares × Current Share Price

Categories of Market Capitalization:

- Large cap (greater than $10 billion): Established, stable companies like Apple or Microsoft.
- Mid cap (between $2 billion and $10 billion): Growing companies with potential for expansion.
- Small cap (between $300 million and $2 billion): Smaller, often younger companies with higher growth potential but also higher risk.

## EXPLORATORY DATA ANALYSIS (EDA)

**Data Cleaning**

1.Handling Missing Values:

•Missing values in Beta and P/E Ratio were imputed with median values.
•Removed rows with excessive missing data.
2.Outlier Detection:
•Outliers in YTD returns and Total Assets were identified using the Interquartile Range (IQR) method.
•Extreme values were capped to prevent skewing results.

**Data Transformation**
• Normalized metrics like Total Assets to ensure comparability.
• Converted RSI and Beta values to standard scores for better clustering.
• Integrated different datasets to ensure consistency and addressing missing and null values to enhance the overall quality of the data

**Visualization Highlights**
1. Boxplots for Key Metrics:
•Beta for Large Cap Growth ranged from -1.82 to 1.55, with most values concentrated near the mean (0.92).
•P/E Ratio in Large Cap Growth showed significant variability, with a max of 51.5.
2.Sector-Wise Trends:
•Defensive sectors like healthcare had consistently lower Beta and higher RSI scores, indicating stability.
•Technology funds exhibited higher volatility but superior YTD growth.

**Challenges and Mitigations:**

- High Variance in Metrics: Managed through scaling and imputation.
- Sector Data Gaps: Addressed using sector averages to fill missing values

## SECTOR-WISE ANALYSIS

**Large Cap Funds**
•Performance: Low Beta, stable returns.
•Key Metrics:
•Beta: 0.92 (mean), P/E Ratio: 8.02, RSI: 58.6.
•Insight: Preferred for low-risk investors seeking stability.

**Technology Funds**
•Performance: High volatility, significant YTD growth.
•Key Metrics:
•YTD: 19.4 (mean), P/E Ratio: 15.1, RSI: 60.2.
•Insight: Suitable for risk-tolerant investors during growth phases.

**Healthcare Funds**
•Performance: Defensive, consistent during downturns.
•Key Metrics:
•Beta: < 1, RSI: High (> 60).
•Insight: Ideal for capital preservation during volatile markets.

## FINANCIAL TERMS

- **Total Assets:** This represents the size of the fund, indicating the total amount of money invested. Larger funds may have more stability and resources, which can influence fund performance.
- **YTD (Year-to-Date):** Shows the fund's performance so far this year, helping investors assess its recent growth or decline.
- **Avg Volume:** Indicates the average number of shares traded daily, which gives an idea of liquidity. A higher average volume suggests easier buying and selling without impacting the price.
- **Annual Dividends:** The yearly payout per share to investors, representing the income generated by the fund, which is particularly relevant for income-focused investors.
- P/E Ratio (Price-to-Earnings Ratio): Helps evaluate the fund’s valuation. A lower P/E ratio can indicate that the fund is more "affordable" compared to its earnings.
- **Beta:** Measures how volatile the fund is relative to the market. A beta above 1 means the fund is more volatile, while below 1 suggests it's less volatile than the market.
- **RSI (Relative Strength Index):** A momentum indicator that shows if the fund is overbought (high RSI) or oversold (low RSI), which can be useful for timing buys and sales.
- **Liquidity Rating:** This rating shows how easily the ETF can be traded. A higher liquidity rating generally means more efficient trading with minimal price changes.

**Volatility Rating:** Indicates how much the price of the fund fluctuates. Higher ratings mean greater price swings, which could signify higher risk.
**ESG Score:** Represents the fund’s commitment to Environmental, Social, and Governance principles. Higher scores are more attractive to socially responsible investors who prioritize sustainability and ethical practices.

## PROPOSED WORK

### *1. Data Collection and Preprocessing*

Dataset: Historical data (2010–2023) from Kaggle, including sector indices like S&P 500 and NASDAQ.

Preprocessing:

- Handling missing values with mean imputation.
- Normalization for comparability.
- Dimensionality reduction using Principal Component Analysis (PCA).

## *2. Data Warehousing*

Preprocessed datasets are stored in a centralized, structured data warehouse for efficient querying and retrieval. This ensures quick access to large datasets during analysis phases and enables scalable computations.

## *3. Data Mining Techniques*

**Clustering**: K-Means to group mutual funds by sector exposure and performance.

**Classification**: Decision Trees and Random Forests to predict fund categories (high, medium, low performance).

**Regression**: Regression models are employed to quantify the relationships between independent variables and dependent variables. This approach predicts fund performance and highlights key drivers influencing outcomes.

# MILESTONES

- Week 1-2: Data Collection and Cleaning – Data sources are identified, and missing data are inputted. Outliers are detected and removed.

- Week 3: Sector Performance Analysis – Sector-wise performance metrics are calculated, and clustering is performed to group similar mutual funds.

- Week 4: Index vs. Mutual Fund Return Comparison – A comparison between mutual funds and sector-specific index returns is performed to evaluate active management.

- Week 5: Risk and Diversification Assessment – An assessment of mutual fund risk, including beta and sector diversification strategies, is conducted.

**-** Week 6: Outlier Detection and Final Report Preparation – Outliers are identified using statistical and machine learning techniques, and the final report is compiled.

# EVALUATION

1. Performance Metrics

   We will evaluate mutual fund performance using a variety of financial metrics:

   - Sharpe Ratio: This measures risk-adjusted return, allowing us to evaluate whether the fund's returns are compensating for the level of risk.

   - Alpha: This metric measures the excess return of a fund relative to a benchmark index, accounting for risk.

   - Maximum Drawdown: This represents the largest peak-to-trough decline in a fund's value during the time period.

   - Volatility: The measure of how much the mutual fund's returns fluctuate over time.

2. Experiments and Results

   We will perform sector-wise analysis on mutual funds across sectors like Healthcare, Technology, Financials, Energy, and Consumer Discretionary. The analysis will focus on comparing the performance of mutual funds within these sectors over the period 2010–2023.

   Key Experiments:

   - Sector-wise Performance Comparison: We will evaluate the performance of mutual funds within each sector, focusing on risk-adjusted returns, volatility, and outlier detection.

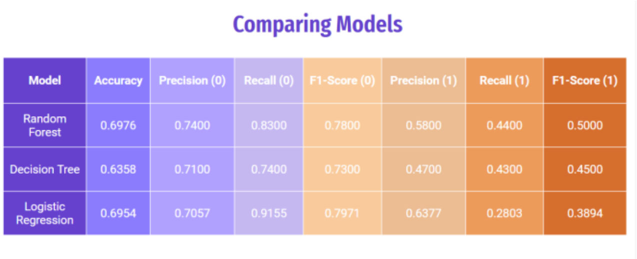
   - Comparison with Index Funds: We will compare mutual fund performance to sector-specific index funds to assess the effectiveness of active management.

   - Sector Rotation Strategies: An analysis of how mutual funds adjust sector allocations during different market cycles will be conducted to evaluate the impact on returns.

3. Expected Results

   The expected results of the analysis will include sector-wise performance trends and insights into how sector allocation affects mutual fund performance. High/Low Performers: Identification of high-performing and underperforming mutual funds within each sector. Regression models uncover the impact of key independent variables.

# Model Training:



Comparing Models

| Model | Accuracy | Precision (0) | Recall (0) | F1-Score (0) | Precision (1) | Recall (1) | F1-Score (1) |
|---|---|---|---|---|---|---|---|
| Random Forest | 0.6976 | 0.7400 | 0.8300 | 0.7800 | 0.5800 | 0.4400 | 0.5000 |
| Decision Tree | 0.6358 | 0.7100 | 0.7400 | 0.7300 | 0.4700 | 0.4300 | 0.4500 |
| Logistic Regression | 0.6954 | 0.7057 | 0.9155 | 0.7971 | 0.6377 | 0.2803 | 0.3894 |

To predict high or low returns for ETFs based on historical performance and sector exposures, we trained and evaluated three machine learning models: **Logistic Regression**, **Decision Tree**, and **Random Forest**.

- **Logistic Regression**:
- Achieved an accuracy of **69.54%** and excelled in recall for low-return funds (91.55%), making it effective for identifying poor-performing ETFs.
- However, the model struggled with precision and recall for high-return predictions due to its linear nature.
- **Decision Tree**:
- While highly interpretable, it had the lowest accuracy **(63.58%**) and recall values, indicating overfitting and limited generalizability.
- Precision and recall for predicting high returns were notably lower compared to other models.

- **Random Forest**:
- Outperformed both models with an accuracy of **69.76%** and the highest F1-Score for high returns (0.50), indicating its ability to handle non-linear relationships and feature interactions.
- It balanced predictions for both high and low returns effectively, making it the most robust model.

## *Observations:*

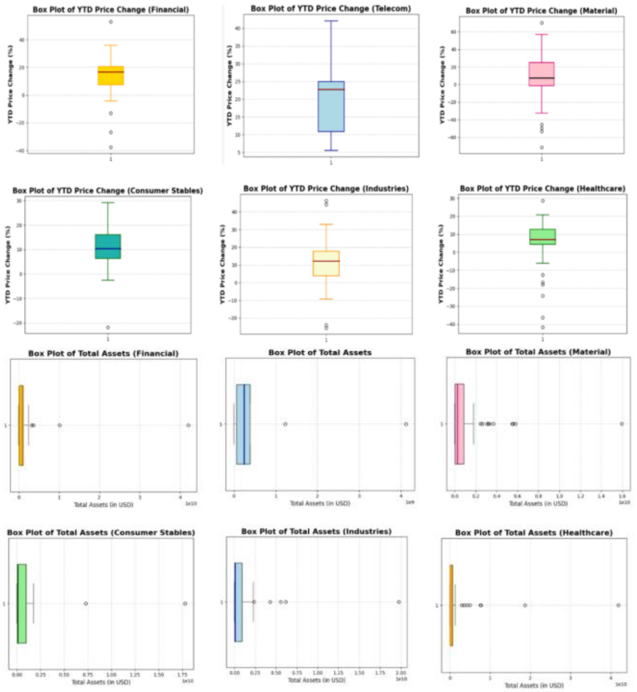
- **Feature Impact**: Key features like fund size, dividend yield, and sector exposure were significant contributors to model predictions.
- **Performance Trade-Off**: Logistic Regression excelled at low-return recall, while Random Forest provided balanced predictions across both classes.
- **Overfitting in Decision Tree**: Its lower accuracy and F1-Score suggest the need for further hyperparameter tuning to improve generalizability.

## *Evaluation:*

The **Random Forest** model emerged as the most effective, achieving the best overall balance between precision, recall, and F1-Score for high-return predictions. Logistic Regression remains valuable for identifying low-performing funds due to its high recall.

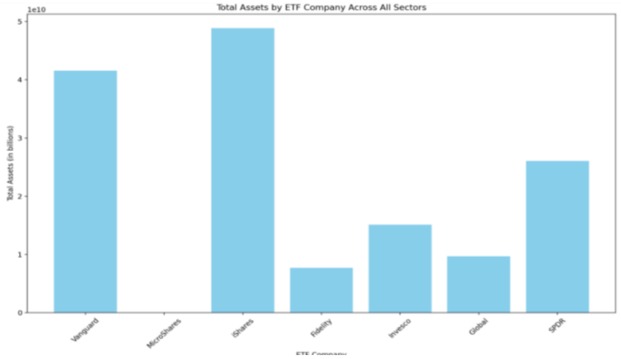
By leveraging these models, we successfully demonstrated the ability to predict ETF performance based on historical and sector-specific data. Random Forest proved to be the most reliable tool for robust and accurate predictions, aligning with our goal of empowering investors with actionable insights for sector-based decision-making.

## Sector-Wise Analysis of Total Assets and Performance Trends:



## Evaluation:

- The initial analysis focused on understanding the distribution of total assets and Year-to-Date (YTD) price changes across various sectors using box plots and bar charts.
- Key techniques like outlier detection and interquartile range (IQR) analysis were employed to identify trends, outliers, and concentration of data within sectors.
- Major ETF companies were also evaluated for their dominance in terms of total assets.



## *Insights:*

- **Outliers in Total Assets:** Significant outliers in sectors like Financials, Healthcare, and Materials suggest a few funds dominate the sector with disproportionately large assets.
- **Sector Performance Trends:**

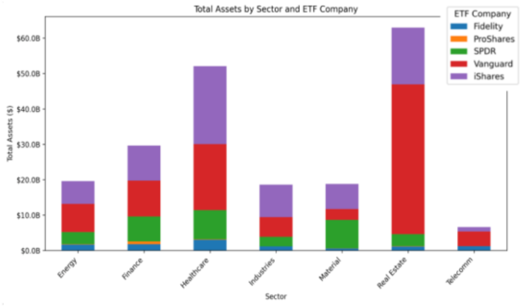
- Defensive sectors like Healthcare and Consumer Stables show stability with fewer extreme performers and narrower spreads.
- Cyclical sectors like Materials and Telecom exhibit high variability, indicating greater risk but also the potential for higher returns.
- **ETF Company Dominance:** iShares, Vanguard, and SPDR collectively hold a significant share of total assets, highlighting their market leadership. Smaller players like Fidelity and Global may focus on niche opportunities.

## *Observations:*

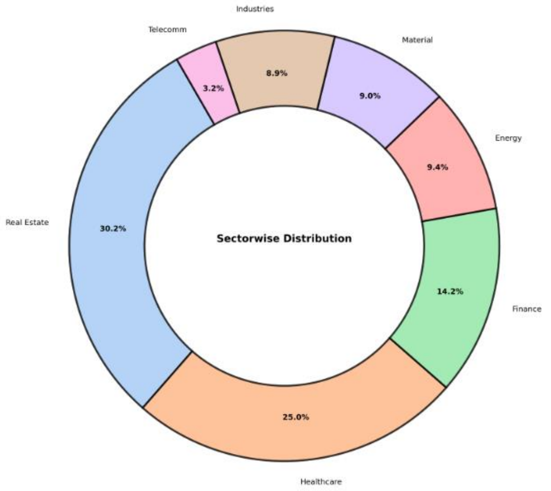
- **Concentration of Assets:** Most sectors have a majority of funds clustered at the lower end of the asset range, reflecting a concentration of smaller funds. Outliers with very high assets impact sector averages.
- **YTD Performance Spread:**
- Materials and Telecom sectors show a broader spread of YTD returns, with extreme positive and negative outliers.
- Healthcare and Consumer Stables sectors display a narrow spread, reflecting stability.
- **Market Concentration:** The significant disparity between the top ETF providers (iShares, Vanguard, SPDR) and smaller firms indicates a highly concentrated market.

The initial analysis highlights sector-specific fund behavior, with Healthcare and Consumer Stables offering stability, while Materials and Telecom show higher variability and risk-reward potential. Outliers significantly impact asset distribution and performance, making median-based evaluations more reliable. Dominance by leading ETF providers emphasizes the need for smaller players to focus on niche strategies.

# ETF and Sector Wise Insights:



- **Sector Strengths:** Healthcare emerged as a high-potential sector, especially during market downturns, offering stability with lower Beta values.
- **High-Risk Sectors:** The Materials sector was deemed high-risk due to its volatility, characterized by significant Beta and P/E variations.



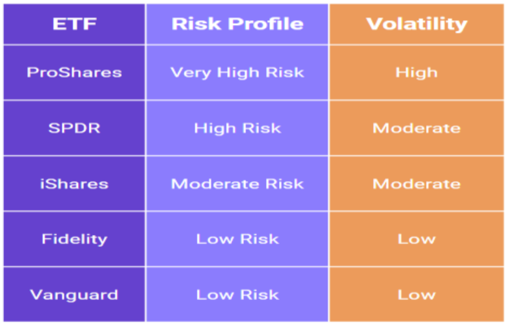
## *Observations:*

- **ETF Dominance:** iShares and ProShares were identified as leading ETF providers, with significant market presence and assets under management.
- **Asset Distribution:** A skewed distribution was observed in total assets, with larger ETFs dominating market capitalization, while smaller funds focused on niche strategies.

## *Evaluation:*

- **Risk Detection:** Bollinger Band analysis highlighted ETFs at risk due to overbought conditions. Higher Beta and P/E ratios correlated with increased sector risk, demanding caution.
- **Sector Variability:** Material sectors exhibited higher variability and unpredictability compared to defensive sectors like Healthcare, indicating distinct investment profiles.

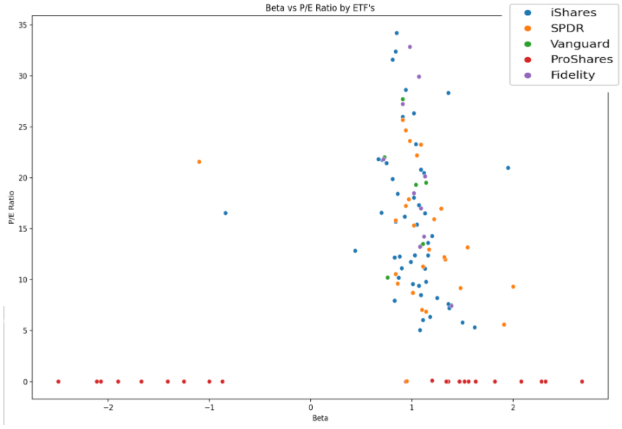
## *P/E and Beta Based Risk Detection:*



| ETF | Risk Profile | Volatility |
|---|---|---|
| ProShares | Very High Risk | High |
| SPDR | High Risk | Moderate |
| iShares | Moderate Risk | Moderate |
| Fidelity | Low Risk | Low |
| Vanguard | Low Risk | Low |

## *Insights:*

- **Risk-Return Tradeoff:** Higher Beta values correlate with greater market volatility, often accompanied by higher P/E ratios, indicating the potential for both risk and reward.
- **Sector Outliers:** Certain ETFs showed exceptionally high P/E ratios but moderate Beta values, suggesting overvaluation relative to their market risk.



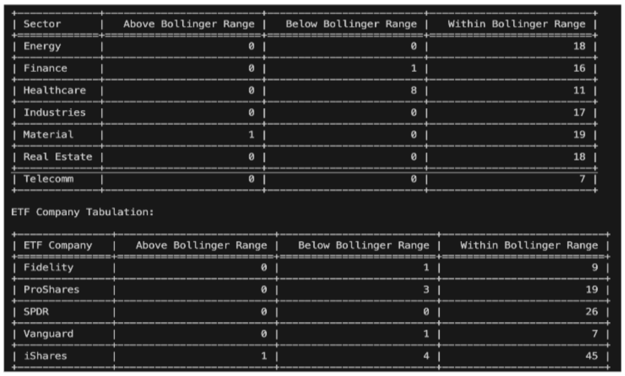
## *Observations:*

- **Healthcare Stability:** ETFs in the Healthcare sector generally had lower Beta values, indicating stable performance, with P/E ratios reflecting moderate valuation.
- **Materials Sector Risk:** ETFs in the Materials sector demonstrated high Beta values paired with relatively higher P/E ratios, highlighting their susceptibility to market swings and potential overvaluation.

## *Evaluation:*

- **Portfolio Diversification:** The spread of ETFs across Beta and P/E axes reinforces the importance of diversification, allowing investors to balance high-risk, high-reward options with more stable, undervalued funds.
- **High-Risk Indicators:** ETFs with both high Beta and P/E values demand cautious consideration, as they might underperform during market downturns despite their growth potential.

## Bollinger Range Risk Detection:



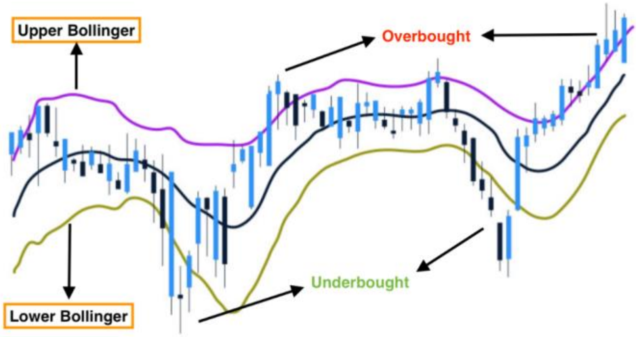
```
+-------------+------------------------+------------------------+-------------------------+
| Sector      |   Above Bollinger Range |   Below Bollinger Range |   Within Bollinger Range |
+=============+========================+========================+=========================+
| Energy      |                      0 |                      0 |                      18 |
+-------------+------------------------+------------------------+-------------------------+
| Finance     |                      0 |                      1 |                      16 |
+-------------+------------------------+------------------------+-------------------------+
| Healthcare  |                      0 |                      8 |                      11 |
+-------------+------------------------+------------------------+-------------------------+
| Industries  |                      0 |                      0 |                      17 |
+-------------+------------------------+------------------------+-------------------------+
| Material    |                      1 |                      0 |                      19 |
+-------------+------------------------+------------------------+-------------------------+
| Real Estate |                      0 |                      0 |                      18 |
+-------------+------------------------+------------------------+-------------------------+
| Telecomm    |                      0 |                      0 |                       7 |
+-------------+------------------------+------------------------+-------------------------+

ETF Company Tabulation:

+---------------+------------------------+------------------------+-------------------------+
| ETF Company   |   Above Bollinger Range |   Below Bollinger Range |   Within Bollinger Range |
+===============+========================+========================+=========================+
| Fidelity      |                      0 |                      1 |                       9 |
+---------------+------------------------+------------------------+-------------------------+
| ProShares     |                      0 |                      3 |                      19 |
+---------------+------------------------+------------------------+-------------------------+
| SPDR          |                      0 |                      0 |                      26 |
+---------------+------------------------+------------------------+-------------------------+
| Vanguard      |                      0 |                      1 |                       7 |
+---------------+------------------------+------------------------+-------------------------+
| iShares       |                      1 |                      4 |                      45 |
+---------------+------------------------+------------------------+-------------------------+
```

## *Insights:*

- **Overbought ETFs:** ETFs approaching or exceeding the **Upper Bollinger Band** indicated high price levels, suggesting a potential overbought condition and increased risk of price correction.
- **Undervalued ETFs:** ETFs near the **Lower Bollinger Band** highlighted undervalued conditions, signaling potential investment opportunities for long-term growth.



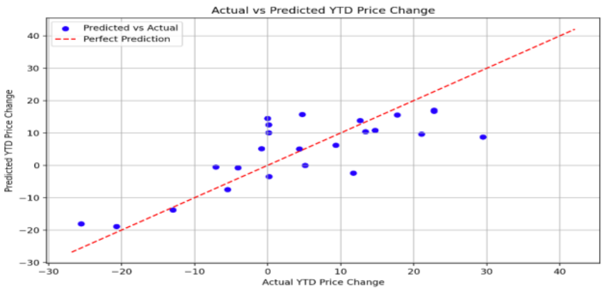
## *Observations:*

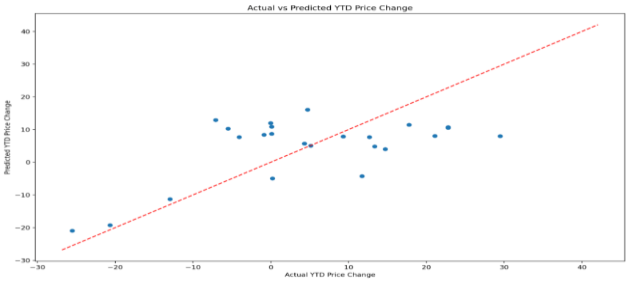
- **Sector Risk Profile:** ETFs in high-volatility sectors like Technology and Materials frequently tested the Upper Bollinger Band, reflecting speculative behavior and heightened market activity.
- **Defensive Stability:** Healthcare sector ETFs consistently hovered near the middle or Lower Bollinger Band, showcasing stability and reduced risk during market fluctuations.

## *Evaluation:*

- **Volatility as a Signal:** ETFs with broader Bollinger Band ranges demonstrated significant price volatility, requiring cautious evaluation of their short-term investment prospects.
- **Market Timing Tool:** Bollinger Band levels effectively indicated market timing opportunities, with funds near the Upper Band signaling exit points and those near the Lower Band indicating entry points.

## Predictive Model Training:





## Insights:

- **Linear Regression:** Highlighted the linear relationships between key financial metrics (e.g., Beta, P/E Ratio) and fund performance, providing a baseline understanding of feature influences.
- **Gradient Boosting Regression:** Outperformed linear regression by capturing complex, non-linear relationships among features, offering improved predictive accuracy for fund returns.

## *Observations:*

- **Feature Importance:** Total Assets and Beta emerged as the most significant predictors across both models, underscoring their role in driving fund performance.
- **Model Accuracy:** Gradient Boosting Regression demonstrated a higher R² score, reflecting superior fit and predictive capability compared to Linear Regression.

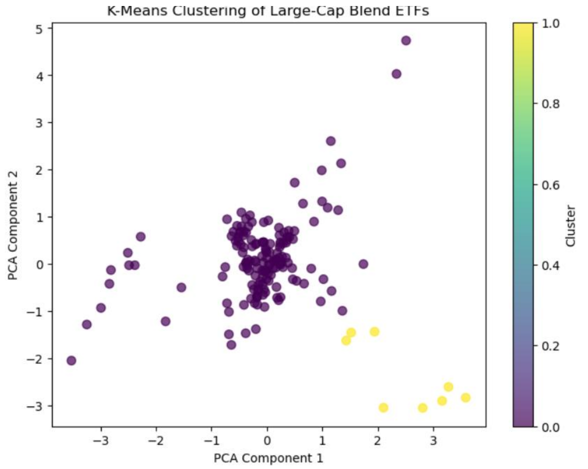
## *Evaluation:*

- **Model Strengths:** Linear Regression provided interpretability, making it easier to identify direct feature impacts. Gradient Boosting Regression delivered robust predictions by effectively handling non-linear dependencies and interactions between variables.
- **Limitations:** Both models' accuracy was sensitive to outliers and data preprocessing quality, highlighting the importance of clean, normalized input data.

## Analyzing Large Cap:

## Insights on Large Cap Blend:

- **Cluster Characteristics:** Two primary clusters were identified:
- **Cluster 0:** Represents conservative funds with lower risk, smaller sizes, and moderate returns, appealing to risk-averse investors.
- **Cluster 1:** Encompasses aggressive growth strategies or large-cap-focused funds with higher volatility and superior returns, catering to risk-tolerant investors.



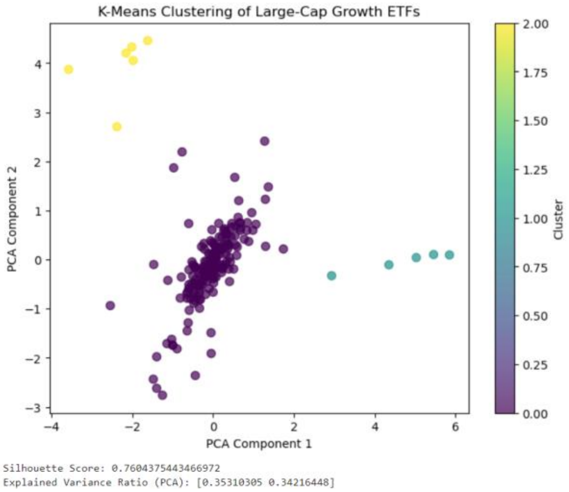
```
Silhouette Score: 0.6242139151185025
Explained Variance Ratio (PCA): [0.36303226 0.34026153]
```

## *Observations:*

- **Clustering Effectiveness:** The Silhouette Score of 0.62 confirmed well-separated and cohesive clusters, validating the segmentation approach.
- **Optimal Cluster Count:** The Elbow Method demonstrated diminishing returns in WCSS reduction beyond two clusters, supporting the selection of two groups.

## *Evaluation:*

- **Risk-Return Balance:** Cluster 0 highlights stability and moderate growth, suitable for long-term capital preservation, while Cluster 1 offers high-risk, high-reward opportunities, aligning with short-term aggressive strategies.
- **Investor Alignment:** The cluster characteristics provide a valuable decision-making framework, enabling investors to align funds with their risk appetite and investment goals.



```
Silhouette Score: 0.7604375443466972
Explained Variance Ratio (PCA): [0.35310305 0.34216448]
```

## *Insights:*

- **Cluster Characteristics:** Three primary clusters were identified:
- **Cluster 0:** Represents ETFs with moderate risk (Beta) and average YTD returns, suited for balanced investment strategies.
- **Cluster 1:** Consists of high-growth ETFs with superior YTD returns, indicating aggressive growth potential but with higher volatility.
- **Cluster 2:** Includes large-scale ETFs with very high Total Assets and relatively stable Beta and YTD values, representing established and less volatile options.
- **Dimensionality Reduction:** PCA was used for visualization, with the first two components capturing 69% of the variance in the data, ensuring meaningful clustering representation.

## *Observations:*

- **Optimal Clustering:** The Silhouette Score of 0.76 indicated cohesive and well-separated clusters, while the Elbow Method confirmed three clusters as optimal.
- **Risk-Return Profiles:** High-growth clusters exhibited strong YTD performance but were accompanied by higher Beta values, suggesting elevated risk.

## *Evaluation:*

- **Investment Alignment:** Cluster 0 provides stability for moderate-risk investors, Cluster 1 appeals to those seeking high returns with higher risk, and Cluster 2 offers a safer, large-cap-focused option.
- **Model Effectiveness:** The clustering approach effectively segmented ETFs, enabling tailored investment strategies based on risk tolerance and market behavior.

### Conclusion:

This project has successfully demonstrated how sector-specific mutual fund analysis can be enhanced using advanced data mining techniques. Through a comprehensive study, we have uncovered actionable insights that align closely with our problem statement and objectives.

- **Sector-Specific Investment Strategies:** Our analysis confirms that different economic sectors exhibit unique performance characteristics during market cycles. Defensive sectors like Healthcare and Utilities provide stability and consistent returns during downturns, while cyclical sectors such as Technology and Materials offer significant growth potential during expansionary periods.
- **Risk and Return Analysis:** By evaluating metrics such as Beta, P/E Ratio, and Bollinger Bands, we provided a robust framework for understanding risk-return profiles. These tools help investors balance their portfolios effectively by combining low-risk stable funds with high-risk, high-reward opportunities.
- **Market Timing and Risk Detection:** The application of Bollinger Band analysis proved to be a valuable technique for identifying overbought or oversold funds and understanding price volatility. These insights allow investors to make informed decisions about market entry and exit points.
- **Predictive Modeling for Performance Insights:** Using Linear Regression and Gradient Boosting Regression models, we achieved actionable predictions for mutual

fund performance. These models offered complementary strengths, with one excelling in interpretability and the other in accuracy, thereby enabling tailored investment strategies for varying risk profiles.

- **Clustering for Diversification:** K-Means clustering analysis on Large Cap Blend and Growth funds revealed diverse fund profiles, providing a deeper understanding of mutual fund characteristics. This segmentation supports strategic portfolio diversification and better alignment with investor preferences.
- **Comprehensive Investment Insights:** The combination of these techniques provided a holistic view of mutual fund performance across sectors. By uncovering trends, identifying anomalies, and deriving actionable insights, this project equips investors with the tools needed to make data-driven decisions in alignment with market conditions and individual goals.