IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 22, Issue 5, Ser. I (Sep. – Oct. 2026), PP 23-34
www.iosrjournals.org

Behavioural Profiling for portfolio optimization by segmenting indian retail investors using K-means Clustering

Dipali Oza[1], Ravi Gor[2]
[1]*Department of Applied Mathematical Science, Actuarial Science and Analytics, Gujarat University*
[2]*Department of Mathematics, Gujarat University*

Abstract: *Behaviour of investor plays an important role in shaping portfolio management strategies and investment selections. This study identifies distinct investor clusters based on their behavioural preferences and demographic characteristics. Comprehensive online survey was used to gather empirical data of 325 individual investors across the diverse demographic background. The collected responses were passed through K-means clustering for the segmentation of investors into distinct behavioural personas. Investors were classified according to their risk tolerance, investment time horizon, return expectation, and investment priorities. The optimal number of clusters was selected through multi-metric convergence results (Elbow Method, Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index). The structural integrity of these clusters were evaluated through cluster-size analysis and validated with Kruskal-Wallis tests, Chi-Square Test of Independence. Cross-tabulation, grouped bar charts, and cluster–profile analysis provided basis for the development of investor typology. The resulting clusters displayed different type of investors such as – income earners, conservative, optimistic and growth oriented. This study presents that clustering can be implemented in order to shift from one-size-fits-all generalized strategy to the investor-specific investment recommendations. The classification of investor profiles will help researchers and professionals to understand behaviour of investors in today's marketplaces. This behavioural profiling of investors contributes to behavioural finance and acts as a stepping stone for the personalized portfolio optimization.*

Keywords: *K-means clustering, Investors segmentation, Investor preferences, Behavioural profiling*

Date of Submission: 02-09-2026 Date of Acceptance: 13-09-2026

I. INTRODUCTION

Investors have long relied on the traditional investment framework for the purpose of asset allocation and wealth management. The structured approach for the traditional investment framework has been formulated through the principles of Modern Portfolio Theory, introduced by Harry Markowitz [10] in 1952. The Markowitz theory operates on a several assumptions: investors are rational and risk-averse, they only care about mean and variance of returns, they can perfectly estimate expected returns and maximize utility based on mean-variance optimization. While these assumptions give mathematical foundations for portfolio optimization, they diverge from the real-world investors. The empirical evidences from behavioural finance show that the real world investors are not perfectly rational and the emotional responses, financial literacy, and behavioural patterns influence investment decisions. Investors do not always care about mean and variance of returns – factors like diversification, safety, liquidity, advice of financial advisors and recommendations from family and friends also play roles. Therefore, there is a gap between theoretical assumptions and practical realities which can be observed by analyzing different types of investors operating in the financial market.

Behaviour of investor plays major role in investment decisions and portfolio management strategies. Investors exhibit diverse profiles shaped by their demographics, investment experience, investment time horizons, risk taking capacity, return expectations, asset allocation preferences, and reaction to portfolio loss situations. Understanding these different investor types and their preferences is important for the development of personalized investment frameworks. In this research, a survey was conducted on indian investors to know about their behavioural characteristics. Empirical data of 325 investors across the diverse demographic background were collected through a comprehensive online survey. Most of the respondents were from the western regions of india. The survey captures many behavioural dimensions including risk tolerance, return expectations, priority factors for investment (diversification, income generation, higher returns, low risk), investment time horizon (time period for holding investments), reaction for 20% portfolio drop in a month, and asset preferences (gold, equity, both).

This study employed K- means clustering algorithm to classify investors into distinct four clusters. Elbow Method, Silhouette Analysis, Calinski- Harabasz Index, and Davies-Bouldin Index were used to obtain optimal k value. Kruskal- Wallis test and Chi-square tests of independence were conducted to validate significant associations of investor cluster membership with demographic and behavioural variables. Each cluster was characterized using comprehensive demographic (gender, age, annual income, investment experience) and behavioural (priority for investment, asset preferences, portfolio drop reaction) attributes. Through the systematic analysis, four distinct investor clusters were identified: Balanced Income Earners, Conservative Beginners, Aggressive Growth Chasers, and Unrealistic Maximizers. This classification of investor profiles will help researchers and professionals to understand investor behaviour in today's marketplaces.

This behavioural clustering approach specific to the indian market context replaces the traditional 'one size fits all' paradigm with the investor-centric investment frameworks. The real - world decision making patterns along with investors' preferences and financial behaviours move beyond the traditional mean-variance optimization. This study will help investment advisors in providing personalized portfolio recommendations based on the individual's return expectations and risk taking capacity.

II. LITERATURE REVIEW

Wood et al. (2004) [13] surveyed ninety investors and segmented them into four types based on their investing attitudes and behaviour. Investment horizon, risk attitude, confidence, control and personalization of loss were the main behavioral constructs of investors. They performed a hierarchical cluster procedure for the segmentation of investors. They also analyzed each investor segment using their different information sources, and types of stocks held by them.

Clark-Murphy et al. (2005) [1] recognized important characteristics of individuals investing in share market using conjoint analysis, cluster analysis and discriminant analysis. Their study considered attributes such as company quality, strength of firm, behaviour towards stocks in emerging industry sectors, low P/E ratio and income generation for investor preferences. The four subgroups of distinct investors were identified based on their different shares attributes and preferences. They provided significance of investor segmentation for financial consultants and investor- education programs.

Sahi et al. (2012) [8] studied indian individual investors and segmented them into distinct groups based on their behavioural tendencies. They conducted cluster analysis on the list of biases like- self-control bias, reliance on expert, overconfidence, budgeting tendency, and socially responsible investing behaviour. They identified four main segments of investors such as- novice learner, competent confirmer, cautious anticipator, and the efficient planner. Their study highlighted that behavioural biases of investors affect significantly on their financial decision making and investment satisfaction.

Fang et al. (2021) [2] applied K- means clustering algorithm in stock prediction and recommendation. They optimized K-means algorithm with the help of artificial fish swarm algorithm (AFSA) and obtained KAFSA. They applied KAFSA on 100 stocks and divided them into high performance stocks and poor performance stocks. They verified the effectiveness of algorithm for stock prediction by comparing stocks in aspects of closing price, return on assets, earning per share and price earning ratio.

Kovács et al. (2021) [9] used a two- stage cluster analysis to explore the investment patterns in retail banking customers. They applied descriptive statistics, machine learning methods Kohonen self- organizing maps (SOM), multiple correspondence analysis, and hierarchical clustering for the analysis of customers' investment preferences, and patterns related to their investment portfolio. In the first stage of procedure, a neural network based Kohonen SOM was used to find out the initial patterns in the data. At the stage second, hierarchical clustering was implemented for the consolidation and interpretation of initial groups. The analysis identified three customer clusters based on their savings profile, risk taking capacity, the way they select financial instruments and seek guidance from professionals.

Gayathri et al. (2025) [3] examined investment behaviour of Generation Y investors in Coimbatore city, India. They conducted a structured questionnaire to find out how financial literacy, investment horizon, risk tolerance, and demographic characteristics influence investment decision and portfolio diversification among investors. They employed descriptive statistics, correlation analysis, and K-means cluster analysis for interpretation and assessment of data. They presented three dominant investor clusters (Conservative, Aggressive and Balanced) based on their behavioural patterns and portfolio characteristics. Their findings offered practical investment solutions for the needs of Gen Y investors.

Gupta et al. (2025) [4] investigated psychological and financial factors influencing retail investors for investment selection. They used K-means cluster analysis on survey responses collected from central india. They identified three investor typologies- Aspirational growth seekers, Conservative high earners, and Independent risk takers. They also performed Cross-tabulation and Chi-square tests to confirm associations between investor types and investment objectives. They emphasized that investor behavior is determined by demographic factors as well as psychological drivers, perception of money, and personal goals.

Oza et al. (2025) [11] provided a comprehensive review of portfolio optimization techniques. They reviewed different mathematical programming approaches used in optimal portfolio design and classified them into classical, advanced and emerging categories. They identified absence of dynamic market conditions and investor psychology in traditional models and suggested hybrid approaches that utilize machine learning to translate qualitative investor behaviour into deterministic constraints for optimization models.

Tan et al. (2025) [12] represented one of the first AI applications in the segmentation of derivative markets. They explored the trading behaviors of traders in Bursa Malaysia's derivative markets with K-means clustering. Analysis over 11 million trade records of derivatives from January to December 2022 was done in their study. They addressed extreme outliers with inverse hyperbolic sine transformation in feature scaling. They identified behavioral patterns among traders and revealed five distinct clusters and validated them using decision tree classifiers.

III. METHODOLOGY

Data collection:

A survey was conducted among indian individual investors to assess their behaviour and preferences for investment. Primary Data was collected through online questionnaire using Google Forms platform. The questionnaire was designed to capture comprehensive profile of investors with their asset preferences, demographic variables, and behavioural metrics.

Table 1: Data Summary Table

Sr. No.	Parameters	Description
1	Research approach	Investors' preferences and behaviour survey – based study
2	Method of data collection	Structured online questionnaire using Google Forms survey
3	Type of data	Primary data
4	Data collection time period	March 2026 and May 2026
5	Sample size	325
6	Respondents of survey	Individual Investors having knowledge of investment and investing in equity markets or commodities
7	Sampling strategy	Convenience and snowball sampling
8	Survey variables	Demographic characteristics (Gender, Age, Annual income), Investment experience, Investment objective, Risk tolerance, Expected target return, Investment horizon, Influencing factors, Behavioural responses and asset preferences
9	Software/tools used	Python, Google Forms

Sample Demographics and Behavioural Characteristics:

After data cleaning, the final sample had 325 respondents. Table 2 represented the demographic distribution of sample data. The demographic context of respondents was described in terms of gender, age, annual income and investment experience using frequencies and percentages.

Table 2: Demographic profile of respondents

Demographic variables	Category	Frequency (n)	Percentage (%)
Gender	Male	214	65.8 %
	Female	111	34.2%
Age	20 - 30 years	190	58.5 %
	30 – 50 years	100	30.8 %
	Above 50 years	35	10.7 %
Annual Income	Below ₹ 5 lakh	158	48.6 %
	₹ 5-10 lakh	86	26.5%
	₹ 10 – 20 lakh	57	17.5%
	Above ₹ 20 lakh	24	7.4%
Investment Experience	1-2 years	182	56%
	2-5 years	80	24.6%
	More than 5 years	63	19.4%

The survey also included behavioural characteristics of respondents. The survey asked for a factor which influences investment decisions, and mix responses were obtained from it. Out of total respondents, 106 (32.6%) selected market news and finance media, 94 (28.9%) were with professional advice, 81 (24.9%) with past performance of assets as influencing factor and 44 (13.5%) followed recommendations from family or friends. 200 (61.5%) respondents had chosen equity market (stocks) and gold both as an preferred investment categories. 84 (25.8%) selected equity market as their preferred asset class while 41 (12.6%) with the gold option only. Capital growth as a primary reason for investment had 157 (48.3%) responses whereas 147 (45.2%) for income generation and 21(6.5%) for capital preservation.

Variable Selection and Transformation for K- means Clustering:

The primary objective of this empirical survey is to classify individual investors based on their financial preferences and investment behaviour. K- means clustering is selected as the method for the classification of investors into distinct typologies. Risk capacity, target annual return, and investment time horizon were selected as input variables for this algorithmic segmentation method. These variables represent the fundamental dimensions of investors' financial preferences. The other demographic and behavioural questions from the survey were used for the post-clustering analysis and profiling of clusters.

Responses for risk and return were collected as ordinal ranges in the survey. 5% - 10%, 10% - 15%, 15% - 20%, and 20% - 25% these four options were provided for the risk capacity. For the target annual return 6% - 8%, 8% - 10%, 10% - 12%, and 12% - 15% were the options. Since K-means clustering requires numerical input data to calculate Euclidean distances between observations, these categorical ranges were transformed into numerical midpoint values. 7.5%, 12.5%, 17.5% and 22.5% were numerical values for risk capacity and 7%, 9%, 11%, 13.5% were values of target annual return after the transformation. For the investment time horizon 1 year, 3 years, 5 years and 7 years options were taken from the data.

Spearman rank-order correlation:

To analyze the relationships between clustering input variables, Spearman rank correlation analysis was conducted. It is non- parametric test which verifies the multidimensional independence of the three continuous variables.

Table 3: Spearman Correlation Matrix of Variables

Variables	Risk	Target Annual Return	Time Horizon
Risk	1.000	0.457	0.233
Target Annual Return	0.457	1.000	0.342
Time Horizon	0.233	0.342	1.000

Spearman's rank correlation values are given in above Table 3. Along with this, $P < 0.001$ for all correlations shows statistically significant results. Table 3 shows very weak positive correlation between Risk and Target Annual Return, Target Annual Return and Time Horizon, and Risk and Time Horizon. This justifies that all three variables for K-means clustering capture distinct but related dimensions of investor behaviour.

K-means Clustering and Optimal K selection:

In this study, K-means clustering algorithm was employed for the segmentation of investors. It groups similar data set into clusters without labelling the clusters. 'K' represents the number of clusters in classification of dataset. It is partitioning method that groups data points based on similarity, minimizing the sum of squared distances between data-points and their respective cluster centroids and maximizing inter-cluster separation. This unsupervised machine learning algorithm operates through iterative process and uses squared Euclidean distance as a measurement. All the input variables were standardized using StandardScaler and scikit learn's KMeans function was used for the clustering process. Following are the steps for K-means clustering algorithm:

(1) X = $[x_1, x_2, \dots, x_m]$ is a set of m input data points of individual investors. The algorithm seeks to divide a dataset into K distinct clusters denoted as $C = \{C_1, C_2, \dots, C_K\}$. The process begins with the randomly selected K centroids $\mu_1, \mu_2, \dots, \mu_k$ from the dataset X.

(2) Each data point x_i in X is assigned to the cluster C_j whose centroid μ_j yields the minimum Euclidean distance. The distance formula is defined as:

$$d(x_i,\ \mu_j)^2 = \parallel x_i - \mu_j \parallel^2 = \ \sum_{n=1}^{N} (x_{in} - \ \mu_{jn})^2$$

Where,

$d(x_i, \mu_j)^2$ represents squared Euclidean distance between data point x_i (individual investor) to the nearest centroid μ_j (mathematical center) of cluster C_j.

N represents the total number of features used in clustering as input variables.

x_{in} and μ_{jn} represent the value of n^{th} feature for an individual investor and the centroid of cluster, respectively.

(3) The centroids are recalculated as the mean of all data points in each cluster.

(4) Repeat steps 2 and 3 until the centroids are no longer changing further.

Optimal K selection:

Table 4 provided four methods used in determining the optimal number of clusters K. The range of K values is selected from 2 to 8.

Table 4: Optimal K selection methods

Optimal K selection method	Description
The Elbow Method	It evaluates the within-cluster sum of squares (WCSS), also known as the inertia. Inertia shows closeness of data points to their cluster centroids. It plots inertia against K values. The Inertia value decreases as K increases, but at certain point the rate of decrease slows down and form "elbow" point of the curve. This point gives optimal K.
Silhouette Score Analysis	This score measures similarity between each data point to its own cluster compared to other clusters. Values range from -1 to +1 and higher score indicates well separated clusters.
Calinski- Harabasz (C-H) Index	This index measures the ratio of the sum of inter- cluster dispersion to the sum of intra-cluster dispersion. A higher C-H index value indicates better clustering as data points in each cluster are close together and provides distinct clusters.
Davies-Bouldin (D-B) Index	This index assesses the clustering structure by computing the ratio of within- cluster dispersion to between- cluster separation. A lower D-B index gives compact and well-separated clusters.

Statistical validation hypotheses:

Kruskal-Wallis H-test was used to validate whether the clusters are distinct and statistically different groups. Chi-square tests of independence were conducted to check the relationship between clusters and demographic and behavioural variables.

- **Kruskal- Wallis H-Test Hypothesis:**

Null Hypothesis: There is no significant difference in the distribution of variables across the clusters.
Alternative Hypothesis: There is a significant difference in the distribution of variables across the clusters and clusters represent distinct groups.

- **Chi-Square Tests of Independence Hypotheses:**

Table 5: Chi-Square Tests of Independence Hypotheses

Sr. No.	Variables	Hypotheses
1	Gender	Null Hypothesis: Gender and cluster membership are independent of each other. Alternative Hypothesis: Gender and cluster membership are not independent of each other.
2	Age	Null Hypothesis: Age and cluster membership are independent of each other. Alternative Hypothesis: Age and cluster membership are not independent of each other.
3	Annual Income	Null Hypothesis: Annual Income and cluster membership are independent of each other. Alternative Hypothesis: Annual Income and cluster membership are not independent of each other.
4	Investment Experience	Null Hypothesis: Investment Experience and cluster membership are independent of each other. Alternative Hypothesis: Investment Experience and cluster membership are not independent of each other.
5	Priority Factor of Investors	Null Hypothesis: Priority Factor of Investors and cluster membership are independent of each other. Alternative Hypothesis: Priority Factor of Investors and cluster membership are not independent of each other.
6	Reaction to 20% portfolio drop in a month	Null Hypothesis: Reaction to 20% portfolio drop in a month and cluster membership are independent of each other. Alternative Hypothesis: Reaction to 20% portfolio drop in a month and cluster membership are not independent of each other.
7	Asset Preference	Null Hypothesis: Asset Preference and cluster membership are independent of each other. Alternative Hypothesis: Asset Preference and cluster membership are not independent of each other.

Behavioural profiling methodology:

Cross-Tabulation Method:

Cross- tabulation method is employed to study the intersection between the mathematical output of the K-means algorithm and behavioural as well as demographic variables of investors. It displays joint frequency distributions of behavioural characteristics across the K clusters. For the systematic comparison of clusters, cell frequencies and column percentages are computed for each variable (Gender, Age, Annual income, Investment experience, Reaction to 20% portfolio drop in a month, Priority of investor across clusters, and Asset preferences for investment).

Interpretation framework:

At last, a qualitative framework was formed to transform the abstract clusters into distinct investor personas. This interpretive step provides clustering results with dominant behavioural characteristics that distinguishes it from other clusters. The four clusters were given names based on their variable percentage distributions within each cluster and overall investment attitude that reflects empirical behavioural aspects rather than arbitrary labels. Investor profiles constructed using this framework can serve as a connection between survey-driven behavioural segmentation and personalized investment strategies.

IV. RESULTS AND DISCUSSION

Optimal cluster selection:

The optimal number of clusters was determined by collective results of four evaluation metrics: The Elbow method, Silhouette score, Calinski- Harabasz (C-H) Index and Davies-Bouldin (D-B) Index. The convergence of these four metrics independently supported for K = 4 as optimal cluster count. The elbow method calculates the inertia (WCSS) across different K values. As shown in below figure 1, the elbow is identified at K = 4 and adding clusters beyond K = 4 would not provide any improvement in variance reduction. Higher silhouette score indicates better separated clusters. In figure 2, Silhouette score reaches its maximum value of 0.40 at K = 7, but this cluster number creates excessive fragmentation for investor behaviour structure. Therefore, to maintain practical interpretability K = 4 is selected with 0.38 value. C-H index recorded a maximum value 218 at K = 4 in figure 3. This high variance ratio confirms that the four distinct clusters are well-separated and individual data points remained close to their respective centers. In figure 4 the minimum D-B index value 0.96 at K = 7 demonstrates the best balance between minimizing intra-cluster variance and maximizing inter-cluster distance. However, the D-B index score 0.97 at K = 4 is nearly optimal number of clusters. Therefore, based on multi-metric convergence K = 4 is selected as optimal number for clusters. Quality metrics for optimal K selection is given below in Table 6.

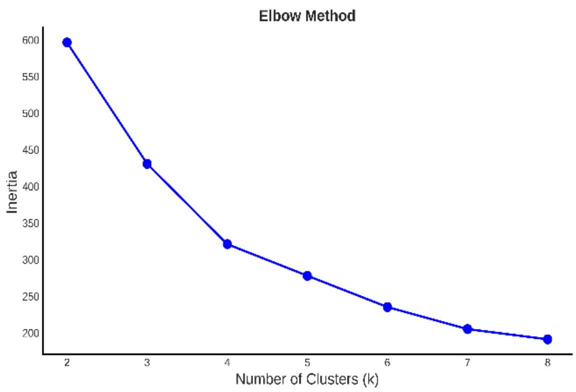

Figure 1:The Elbow Method

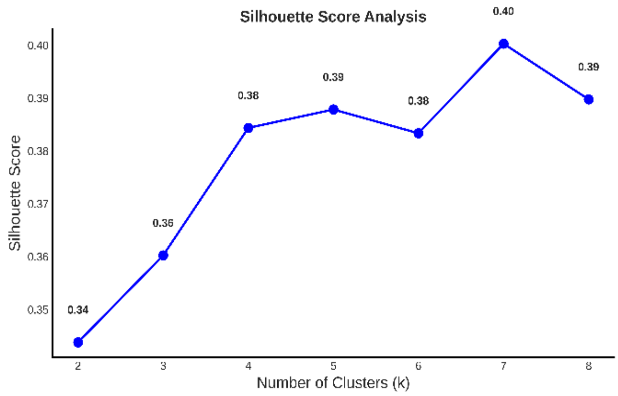

Figure 2: Silhouette Score Analysis

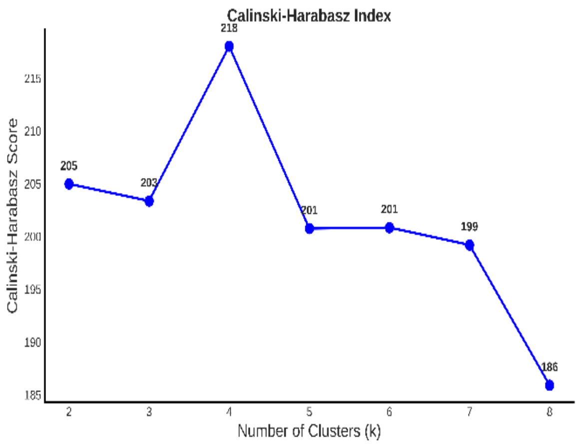

Figure 3: Calinski-Harabasz Index

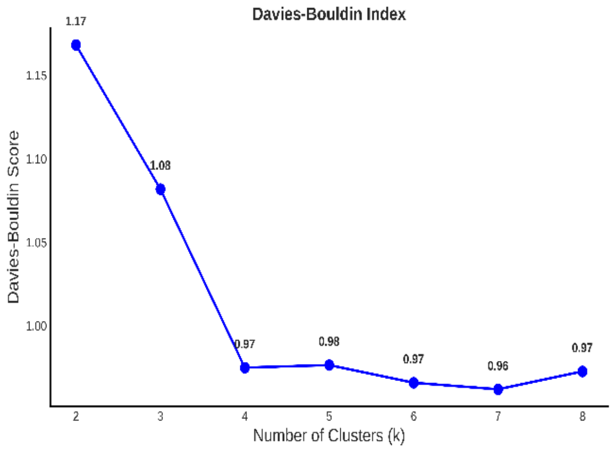

Figure 4: Davies-Bouldin Index

Table 6: Clustering Quality Metrics

Quality Metrics of Clusters for K = 4	Score
Inertia	320.97
Silhouette score	0.38
Calinski – Harabasz Index	218.0
Davies – Bouldin Index	0.97

Cluster Centroids and Cluster Size Summary:

The following Table 7 presents the size of the four identified clusters. The cluster centroids (mean values) are computed for the three behavioural variables.

Table 7: Cluster Size and Centroids Summary

Clusters	Size	Population (%)	Centroid: Risk (%)	Centroid: Return (%)	Centroid: Time Horizon (Years)
1	104	32.0 %	10.05	11.01	6.40
2	98	30.2 %	9.34	7.98	2.02
3	64	19.7 %	19.53	12.40	5.66
4	59	18.1 %	10.72	12.19	1.98
Total	325	100 %			

Kruskal – Wallis test:

Kruskal-Wallis test is employed to validate these clusters as statistically distinct groups by accessing differences in behavioural variables across clusters. The H statistic indicates the dispersion of ranked data across the clusters. Value of $P < 0.001$ for all three variables provides strong reason to reject the null hypothesis. As shown in Table 8, Time horizon with the largest H statistic 232.770 is the main variable that strongly divides investors into clusters.

Table 8: Kruskal-Wallis test results

Input Variables	H- statistic	P value	Significance
Risk Tolerance	174.424	P < 0.001	Very Significant (Reject Null)
Target Annual Return	199.066	P < 0.001	Very Significant (Reject Null)
Time Horizon	232.770	P < 0.001	Very Significant (Reject Null)

Chi-Square Test of Independence:

Table 9 gives results of Chi-square test of independence conducted for demographic variables (gender, age,annual income, investment experience) and three categorical behavioural variables (priority of investors, reaction to 20% portfolio drop in a month, asset preference). A high Chi- square value shows that clusters from survey data are not random and there is relationship between clusters and variables. $P < 0.001$ indicates strong associations between variables and cluster membership. For the demographic variable Age, value of $P > 0.001$ presents zero relationship between investors' clusters and their Age.

Table 9: Results of Chi-square test of independence

Variables	Chi-Square value	Degree of freedom (df)	P value	Significance
Gender	42.51	3	P < 0.001	Significant relation (Reject Null) Gender influences which cluster an investor belongs to.
Age	10.41	6	**P > 0.001**	No strong relation (Accept Null) Age is distributed evenly across all clusters.
Annual Income	36.35	9	P < 0.001	Significant relation (Reject Null) Annual Income influences which cluster an investor belongs to.
Investment Experience	48.40	6	P < 0.001	Significant relation (Reject Null) Investment Experience influences which cluster an investor belongs to.
Priority of Investors	41.42	9	P < 0.001	Significant relation (Reject Null) Priority of Investors influences 6cluster membership.
Reaction to 20% portfolio drop in a month	61.28	6	P < 0.001	Significant relation (Reject Null) Reaction to 20% portfolio drop in a month influences cluster membership.
Asset Preference	33.08	6	P < 0.001	Significant relation (Reject Null) Asset Preference of investors influences cluster membership.

Cross tabulation:

Results of Chi-square tests of independence validated the relationship between variables and cluster membership. Cross-tabulation shows how this relationship looks with grouped bar charts and percentage distribution analysis. In graphical format, x-axis represents the clusters and y-axis represents the percentage value for each category of variables. This representation highlights the dominant behavioural preferences, demographic characteristics and contrasts between the cluster profiles.

The demographic charts in below figures show that the clusters composition vary in gender, annual income and investment experience variables. Age distribution appears relatively balanced across the clusters. Cluster 2 marks higher female representation with 58.2%, 77.6% of respondents with 1 to 2 years investment experience, and a larger proportion of respondents with below ₹ 5 lakh annual income. Cluster 3 indicates comparatively lower female representation with 10.9%, along with large share of individuals with above 5 years investment experience, and investors with above ₹ 20 lakh annual income. Both cluster 1 and cluster 4 are identical in gender representation and have similar pattern in investment experience categories. This suggests that clusters 1 and 4 differ more on behavioural attributes than by demographic characteristics.

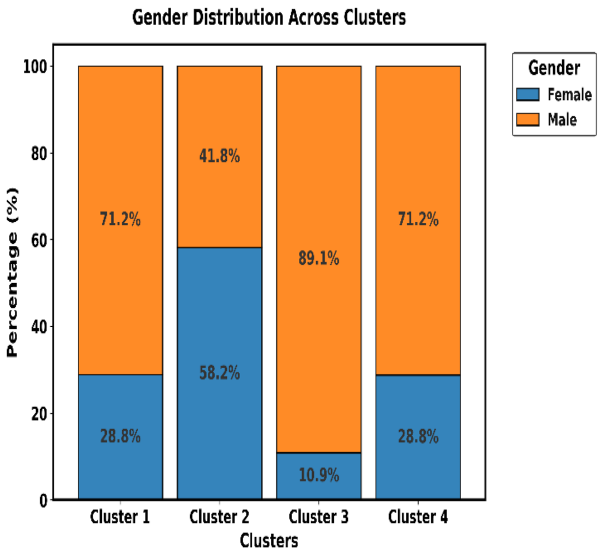

Figure 5: Gender distribution across clusters

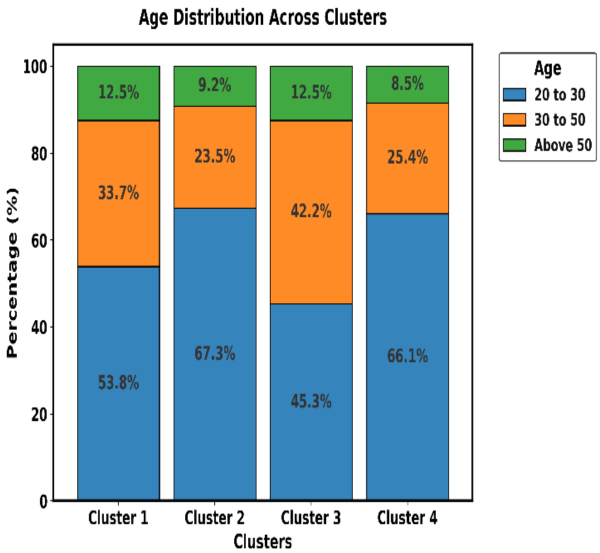

Figure 6: Age distribution across clusters

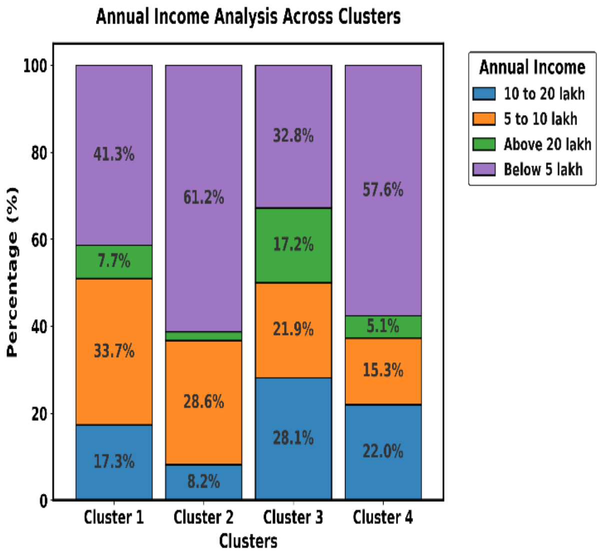

Figure 7: Annual income (in ₹) analysis across clusters

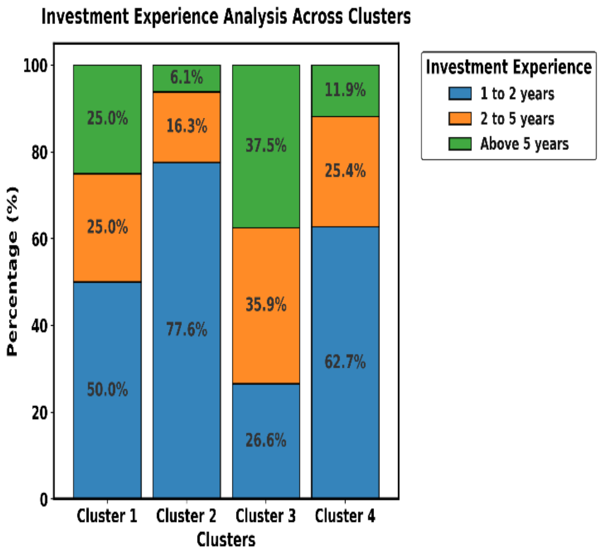

Figure 8: Investment experience analysis across clusters

The following behavioural charts reveal the cluster-wise separation in investment priority, reaction to 20% portfolio drop in a month, and asset preferences. Majority of investors in cluster 2 have priority of keeping risk low, 50% with tendency to hold and wait for recovery and 27.6% shows reaction of immediate selling in portfolio drop situation, and gold is not most preferred investment in this but this cluster has the highest percentage of gold among all four clusters. Cluster 3 invests more as a reaction to portfolio drop of 20% in a month. Majority of respondents in cluster 3 have priority of achieving diversification in investment. 32.7% of respondents select getting stable income as their priority in cluster 1. Low risk and higher returns are priorities for cluster 4. Stocks and gold both are preferred as investment choice by all four clusters with different percentage distributions.
These visual patterns add support to statistical results and help to convert the clusters into more interpretable distinct investor profiles.

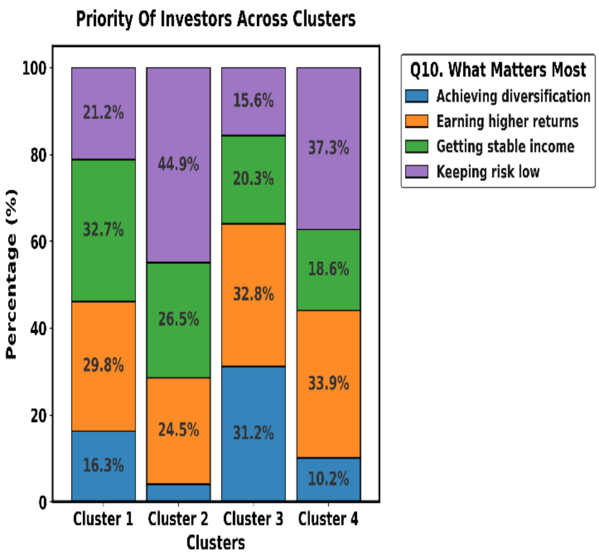

Figure 9: Priority of investors across clusters

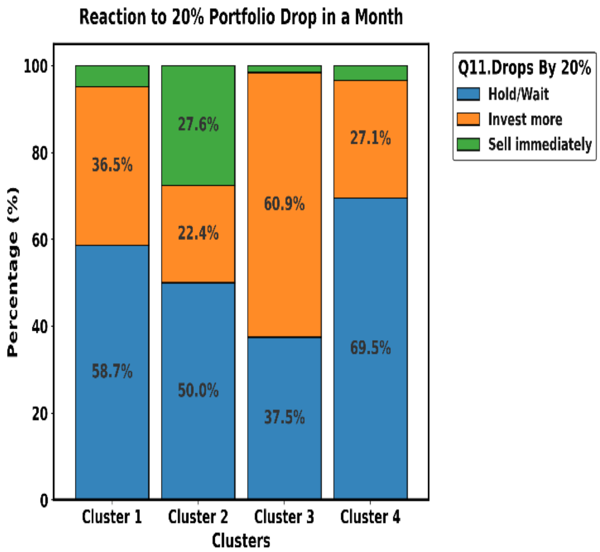

Figure 10: Reaction to 20% portfolio drop in a month

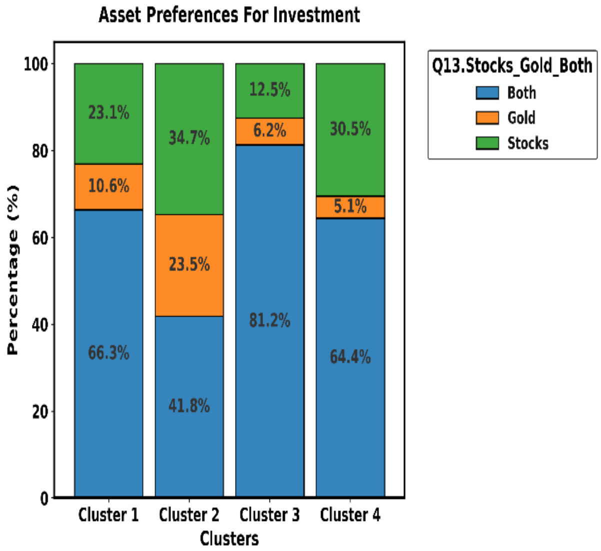

Figure 11: Asset preferences for investment

Cluster Naming and Investor Typology:

Based on the observed cluster size distribution, cluster centroids, statistical test results, cross-tabulation results, and grouped bar charts, each cluster was given a descriptive name that represents its behavioural profile. The primary patterns found in target annual return, risk tolerance, investment time horizon, investment priority, reaction to 20% portfolio drop in a month, annual income, and investment experience served as the basis for the naming procedure. The cluster names were derived by identifying the distinguishing behavioural percentage within each cluster and overall investment attitude profile.

As shown in below Table 10, each cluster name represents the empirical behavioural characteristics and distinct practical profile rather than just an arbitrary labels.

Table 10: Cluster names and key characteristics

Key Variables	Cluster 1 Balanced Income Earners	Cluster 2 Conservative Beginners	Cluster 3 Aggressive Growth Chasers	Cluster 4 Unrealistic Maximizers
Mean Risk	10.05%	9.34%	19.53%	10.72%
Mean Expected Return	11.01%	7.98%	12.4%	12.19%
Mean Time Horizon For Investment	6.4 years	2.02 years	5.66 years	1.98 years
Priority of Investors	Getting Stable Income	Keeping Risk Low	Earning Higher Returns, Achieving Diversification	Keeping Risk Low, Earning Higher Returns
Behavioral Reaction to 20% Portfolio Drop In a Month	Hold the Investment and wait for recovery	Hold the Investment and wait for recovery, Sell immediately to avoid further losses	Invest more to take advantage of lower prices	Hold the Investment and wait for recovery
Investment Experience	1 to 2 years	1 to 2 years	Above 5 years	1 to 2 years, 2 to 5 years
Annual Income	Below ₹ 5 lakh, ₹ 5 to 10 lakh	Below ₹ 5 lakh	Below ₹ 5 lakh, ₹ 5 to 10 lakh, ₹ 10 to 20 lakh and above ₹ 20 lakh	Below ₹ 5 lakh, ₹ 5 to 10 lakh, ₹ 10 to 20 lakh

Cluster 1: Balanced Income Earners

This cluster was named based on respondents' priority of getting stable income, combined with moderate value of mean risk (10.05%) and expected mean return (11.01%) and balanced 6.4 years of mean time horizon for investment. Investors show patience for 20% portfolio drop situation by avoiding immediate action and waiting for recovery. All these factors indicate the steady income seeking balanced cluster of investors.

Cluster 2: Conservative Beginners

This cluster reflects risk aversion behaviour of investors. It has lowest mean values of risk and return among all four clusters. Keeping risk low is priority of this cluster with predominant beginner characteristics (77.6% of respondents with 1-2 years investment experience). It has defensive approach for portfolio drop situation with mix reactions of investment holding and waiting for recovery and selling immediately to avoid further losses. Hence, the name for this cluster is conservative beginners.

Cluster 3: Aggressive Growth Chasers
This cluster has highest risk taking capacity of 19.53% among all the four clusters. The main focus of cluster is on earning higher returns and achieving diversification. It demonstrates proactive response of investing more during portfolio drop of 20% in a month. This indicates aggressive opportunistic behaviour by taking advantage of falling prices for further investment.

Cluster 4: Unrealistic Maximizers
This cluster has lowest mean time horizon for investment with mean values of expected return 12.19% by taking risk 10.72%. It shows that respondents of this cluster desire for quick returns with moderate risk level. Priority of investors in this cluster are mainly keeping risk low and earning higher returns. Therefore, investors in this cluster gave least 5.1% preference to gold and choose stocks with gold for investment.
The final Table 10 consolidated the statistical, behavioral, and demographic evidences from the investors, and represented the clusters as four investor profiles with distinct investment orientations. By translating the priorities derived from empirical data into quantitative parameters, personalized optimization models can be created.

V. CONCLUSION

This study suggested that investors are not a homogeneous group of people; rather, they vary in their attitudes toward market fluctuations and preferred investment choices. The clustering analysis determined four investor groups with different characteristics across return expectation, risk tolerance, investment time horizon, reaction to 20% portfolio drop in a month, asset preference, and other categorical variables. The statistical tests and cross-tabulation with bar charts validated the distinctiveness of the identified clusters. Taken together, these results exhibited a complete picture of each cluster with the preferences and behaviour of investors.

These investor profiles can help as a basis for constructing customized investment strategies. For instance, conservative beginners preferring safety and low level of risk can be provided to more capital preserving investment strategies, whereas aggressive growth chasers with the higher risk taking capacity can be taken into consideration for a higher return strategies. In a same manner, investing time horizon, priority of investors and portfolio drop reactions can be taken into account while making decisions for asset allocations, liquidity needs and rebalancing practices. Thus, each cluster provides practical foundation for investment firms and financial professionals to shift from generalized strategy to the investor-specific investment recommendations. The characteristics of each persona can be translated into definitive mathematical constraints for the investor-centric investment framework.

VI. FUTURE SCOPE

Future research can expand this study by testing the proposed methodology on massive and more diverse populations covering different regions, market conditions, financial literacy, and types of investors. The replication of results across multiple demographics would be helpful to assess whether the derived clusters remain stable and relevant beyond the present sample data. Further studies can incorporate these behavioural profiles with portfolio optimization models and evaluate cluster- specific portfolios. However, any personalized investment strategies are required to incorporate periodic reassessment and financial fitness checks. The performance of these customized portfolios can be compared with the traditional portfolio strategies. Other clustering algorithms can be used for segmentation and may also be assessed against K-means to evaluate the robustness of the cluster formation.

BIBLIOGRAPHY

[1] Clark-Murphy, M., & Soutar, G. (2005). Individual investor preferences: A segmentation analysis. The Journal of Behavioral Finance, 6(1), 6-14.
[2] Fang, Z., & Chiao, C. (2021). Research on prediction and recommendation of financial stocks based on K-means clustering algorithm optimization. Journal of Computational Methods in Science and Engineering, 21(5), 1081-1089.
[3] Gayathri, A., & Sathish, S. (2025). A Study On Investment Decision And Portfolio Management Among Generation Y Investors In Coimbatore City. Journal of Applied Bioanalysis, 11(9s), 93-102.
[4] Gupta, R. C., Mittal, M., & Gupta, R. (2025). Behavioural Drivers and Investor Typologies in Investment Selection: Evidence from Retail Investors in Central India. Adhyayan: A Journal of Management Sciences, 15(02), 42-53.
[5] https://towardsdatascience.com/calinski-harabasz-index-for-k-means-clustering-evaluation-using-python-4fefeeb2988e/
[6] https://www.geeksforgeeks.org/machine-learning/davies-bouldin-index/
[7] https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/
[8] Kalra Sahi, S., & Pratap Arora, A. (2012). Individual investor biases: a segmentation analysis. Qualitative Research in Financial Markets, 4(1), 6-25.

[9] Kovács, T., Ko, A., & Asemi, A. (2021). Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis. Journal of Big Data, 8(1), 141.
[10] Markowitz, H. M., (1952). Portfolio selection. The Journal of Finance 7, 77-91.
[11] Oza, D., & Gor, R. (2025). Mathematical Programming in Portfolio Optimization: A Three- Part Review of Classical, Advanced, and Emerging Techniques. IOSR Journal Of Economics and Finance (IOSR-JEF), 16(5), 19-32.
[12] Tan, E. H. L., Hamed, Y., Daud, H., Abdul Wahab, M. A. F., Azhar, A. A. A., & Tan, S. Y. (2025). Profiling investor behavior in the Malaysian derivatives market using K-means clustering. Frontiers in Artificial Intelligence, 8, 1640776.
[13] Wood, R., & Zaichkowsky, J. L. (2004). Attitudes and trading behavior of stock market investors: A segmentation approach. The Journal of Behavioral Finance, 5(3), 170-179.