Clustering and Classification of Russell 2000 Stocks: A Dual-Group K-Means and KNN Analysis

Pushkar Patil

27th October 2025

1 Introduction and Problem Statement

The process of categorizing equities by *size* and *style* (growth vs value) is fundamental to portfolio construction, benchmarking, and style-aware risk management. The **Russell 2000** index is composed primarily of small-cap U.S. equities and presents a practical but challenging domain for such classification because of its breadth (approximately 2,000 constituents) and heterogeneity of company fundamentals.

1.1 Assignment Objectives

This project addresses the following requirements:

1. **Clustering Analysis:** Using market capitalization and price-per-earnings growth (PEG ratio) as metrics, run K-Means clustering with 9 clusters on Russell 2000 companies. Compare cluster positioning with Morningstar ratings and evaluate boundary alignment with expectations.
2. **Label Assignment & Classification:** Assign growth and capitalization labels to clusters, then classify new stocks of interest using the K-Nearest Neighbors (KNN) technique.

This project constructs an automated classification pipeline that:

- groups Russell 2000 stocks into meaningful clusters using unsupervised learning (K-Means),
- maps cluster assignments to Morningstar-style 3×3 style boxes (Size × Style),
- builds a supervised classifier (K-Nearest Neighbors) to label new/unseen stocks, and
- evaluates differences vs. Morningstar categorizations and outlines limitations and improvements.

Primary research questions:

1. Can a dual-feature representation (Market Capitalization with PEG for growth, Market Capitalization with P/B for value) recover industry-recognized style categories?
2. How robust are K-Means clusters in a small-cap dominated universe?
3. Why and where do KNN predictions diverge from Morningstar classifications?

2 Data Description and Preparation

[1]

2.1 Source and Variables

The dataset is provided in an Excel file named `ML_Project.xlsx` and contains 1,962 rows (one per ticker) with 17 columns. Relevant columns used in analysis are:

- **Ticker** (Column A)
- **P/B Ratio** (Column H)
- **1 year EPS Growth (as of latest filing)** (Column K)
- **Market Cap** (Column L) — in millions
- **P/E Ratio** (Column M)

2.2 Cleaning and Type Conversion

Data cleaning follows standard practices for financial datasets:

1. Convert numeric-like columns with `pd.to_numeric(..., errors='coerce')` to force non-numeric tokens to `NaN`.
2. Replace placeholder strings (e.g., "–") with `NaN` to standardize missing values.
3. Focus subsequent analysis only on rows with the required metrics available.

Rationale: financial feeds often include missing, delayed or specially-coded values. Coercing non-numeric entries to `NaN` allows downstream numeric processing and robust filtering.

2.3 Coverage Summary

From the original 1,962 tickers:

- Group 1 companies: 594
- Group 2 companies: 1,177
- Total analyzed: 1,771 (90.3%)

[1] The dataset was gathered collaboratively with Shreyansh Sharma, Soham Joshi, and Partha Sanampudi. The project, code, and report were completed individually by the author.

3 Dual-Group Segmentation Strategy

To respect different investment paradigms (growth vs. value) and accommodate stocks with missing earnings metrics, the dataset is split into two mutually exclusive groups.

3.1 Group 1: Growth View (Market Cap + PEG)

Group 1 contains firms for which market capitalization, trailing P/E, and EPS growth (positive) are available. The derived metric is computed as:

$$\text{PEG} = \frac{\text{P/E Ratio}}{\text{EPS Growth (in \%)}/100}$$

Note: EPS Growth is expressed in percentage terms in the dataset (e.g., 25 for 25%). Before computing PEG, the percentage is converted to a decimal form. **In Excel:** PEG = (Column M `[P/E Ratio]` $\times$ 100) $\div$ Column K `[EPS Growth (%)]`. This ensures consistent scaling between the ratio and growth inputs.

Filtering criteria:

- Market Cap > 0
- P/E > 0
- EPS Growth > 0
- PEG in $(0, 10)$ to exclude extreme outliers

Result: 594 stocks satisfy these constraints.

3.2 Group 2: Value View (Market Cap + P/B)

Group 2 covers the remaining universe (excluding Group 1) that has valid market capitalization and P/B ratios. Filtering:

- Market Cap > 0
- P/B in $(0, 20)$ to avoid outlier influence

Result: 1,177 stocks.

3.3 Rationale for Two-Group Approach

- **Coverage:** Given that Russell 2000 pertains to small-cap companies, many stocks have missing or negative EPS growth; PEG is inapplicable there.
- **Investment semantics:** PEG captures growth-investor preference while P/B is widely used for value analysis, especially when earnings are negative.
- **Model robustness:** The division prevents mixing fundamentally different metrics in a single clustering run, which would dilute interpretability.

- **Assignment alignment:** The assignment specifies using PEG ratio, which we implement for Group 1 (594 stocks). Group 2 uses P/B ratio to extend coverage to the remaining 1,177 stocks that lack reliable earnings data, ensuring comprehensive analysis of the Russell 2000 universe.

4 Feature Engineering and Scaling

4.1 Log-Transform of Market Capitalization

Market capitalization spans several orders of magnitude. Use:

$$\text{Market\_Cap\_Log} = \log_{10}(\text{Market\_Cap\_Millions})$$

This transformation equalizes proportional differences: a ten-fold change in market cap maps to a unit increment on the log scale. It prevents large absolute caps from dominating distance computations.

4.2 Standardization

Features are standardized using z-score normalization:

$$z = \frac{x - \mu}{\sigma}$$

where μ and σ are the sample mean and standard deviation for each feature. Standardization is required because K-Means and KNN are distance-based; unstandardized features with larger numerical ranges overwhelm those with smaller ranges.

5 Clustering Methodology (K-Means)

5.1 Algorithm and Hyperparameters

K-Means partitions data into k clusters by minimizing the within-cluster sum of squares (WCSS):

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2$$

Implementation choices:

- `n_init` = 100 (many random starts to escape bad local minima)
- `max_iter` = 500
- `random_state` = 42 (reproducibility)

5.2 Determining k (Elbow and Silhouette)

We examine:

- **Elbow curve (WCSS vs. k)** to identify diminishing returns in cluster reduction of inertia.
- **Silhouette Score**:
$$s = \frac{b - a}{\max(a, b)}$$
where a is mean intra-cluster distance and b is mean nearest-cluster distance. Values near 1 indicate well-separated clusters.

For both groups, automatic diagnostics suggest $k = 3$ or $k = 4$ as natural partitions (largest inertia drop / best silhouette). However, to align with business practice (Morningstar 3×3 style box), $k = 9$ clusters were chosen for interpretability and direct mapping to Size × Style categories.

6 Cluster Labeling and Style Mapping

Three distinct labeling systems are used throughout this study, each serving a different purpose:

1. **K-Means Cluster Labels (data-driven):** Determined purely from numerical similarity in feature space (Market Cap with PEG or P/B). These clusters are unsupervised and represent *natural groupings* discovered by the algorithm.
2. **User-Defined Labels (rule-based):** Applied post-clustering using fixed thresholds on cluster centroids to assign interpretable `Size-Style` names (e.g., `Small-Value`, `Mid-Growth`). These thresholds are explicitly defined by market-cap and ratio cutoffs and serve as a consistent labeling framework.
3. **Morningstar Classifications (external benchmark):** Independent, professional categorizations used only for *validation and comparison*. Morningstar's system integrates multiple valuation and growth factors beyond PEG or P/B.

Each cluster's centroid is then inverse-transformed from standardized features back to original units (log market cap and ratio). The centroid market-cap log is converted back to millions via:

$$\text{Market\_}\widehat{\text{Cap}}\text{\_Millions} = 10^{\text{centroid}_{\text{market\_cap\_log}}}$$

Size bins:

- **Small:** Market cap $<$ \$2,000M
- **Mid:** \$2,000M $\leq$ Market cap $<$ \$10,000M
- **Large:** Market cap $\geq$ \$10,000M

Style bins:

- **Value (Group 1):** PEG < 1
- **Blend (Group 1):** $1 \leq$ PEG < 1.5
- **Growth (Group 1):** PEG ≥ 1.5
- **Value (Group 2):** P/B < 2
- **Blend (Group 2):** $2 \leq$ P/B < 4
- **Growth (Group 2):** P/B ≥ 4

Combine to obtain labels like `Small-Value`, `Mid-Blend`, etc.

6.1 Group 1 Cluster Assignments

The nine clusters produced for Group 1 (PEG-based segmentation) exhibit the following characteristics:

Cluster	Label	Count	Avg Cap ($M)	Avg PEG
0	Small-cap Value	124	629.58	0.31
1	Mid-cap Value	155	2013.05	0.42
2	Mid-cap Growth	14	3016.96	5.10
3	Mid-cap Growth	53	4018.89	2.24
4	Mid-cap Value	85	5483.05	0.51
5	Small-cap Value	97	219.51	0.40
6	Small-cap Growth	16	420.93	3.15
7	Mid-cap Growth	8	2142.70	8.40
8	Small-cap Blend	42	1001.67	1.40

Table 1: Group 1 cluster centroids and classifications.

Key observations: The clusters show a distribution across both size and style categories. Clusters 0 and 5 represent small-cap value stocks with lower market caps and PEG ratios. Clusters 1 and 4 are mid-cap value with higher market capitalizations. Clusters 2, 3, and 7 capture growth stocks with high PEG ratios. Cluster 8 represents small-cap blend stocks with moderate PEG values near 1.4.

6.2 Group 2 Cluster Assignments

The nine clusters for Group 2 (P/B-based segmentation) are distributed as follows:

Cluster	Label	Count	Avg Cap ($M)	Avg P/B
0	Mid-cap Blend	169	3896.11	2.40
1	Small-cap Value	202	187.98	1.49
2	Small-cap Growth	50	521.46	7.61
3	Small-cap Value	279	514.85	1.43
4	Small-cap Value	232	1436.93	1.42
5	Mid-cap Growth	26	4247.45	15.82
6	Mid-cap Growth	74	3108.15	7.70
7	Small-cap Growth	32	607.49	14.17
8	Small-cap Growth	113	1254.86	4.20

Table 2: Group 2 cluster centroids and classifications.

Key observations: Group 2 shows clear segmentation across size and style categories. Clusters 1, 3, and 4 are the largest, representing small-cap value stocks with low P/B ratios (around 1.4-1.5). Cluster 0 is the sole mid-cap blend cluster with 169 stocks. Growth stocks are identified in clusters 2, 5, 6, 7, and 8, with cluster 5 showing the highest average P/B ratio of 15.82. The distribution suggests that while small-cap value stocks dominate numerically, the clustering successfully identifies distinct growth segments.

7 Results

7.1 Group 1 (Growth) — Summary

- Stocks analyzed: 594
- Automatically-detected optimal k (elbow): 3
- Best silhouette (diagnostic): $k = 3$
- Selected k for business mapping: 9
- Silhouette at $k = 9$: 0.4347
- Inertia: 111.09

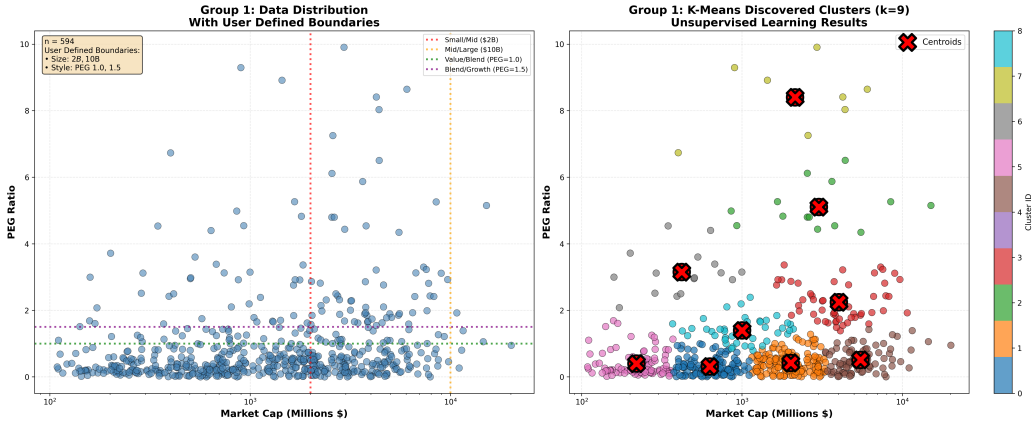

Figure 1: Group 1: Distribution and K-Means clustering (Market Cap vs. PEG). Centroids (red X) and cluster IDs are annotated.

Interpretation: The majority of Group 1 observations span across small-cap to mid-cap ranges with varying PEG ratios. The clustering identifies distinct growth and value segments, with silhouette score of 0.4347 indicating moderate to good clustering quality.

Understanding the Dotted Boundaries: The left panel shows user-defined boundaries (dotted lines) that represent rule-based thresholds: vertical lines at \$2B and \$10B separate small/mid/large-cap categories, while horizontal lines at PEG = 1.0 and 1.5 separate value/blend/growth styles. These are *prescriptive* boundaries imposed on the data. In contrast, the right panel shows K-Means discovered clusters without knowledge of these boundaries—the algorithm groups stocks based on their natural similarity in feature space, not predetermined rules.

Why No Large-Cap Clusters? K-Means allocated zero clusters to the large-cap region (market cap > \$10B) because the Russell 2000 contains very few stocks in this range—most large-caps graduate to the Russell 1000 index. The algorithm intelligently concentrated all 9 clusters in regions with substantial data density (small and mid-cap zones), rather than wasting clusters on sparse regions. This demonstrates K-Means' data-adaptive behavior: it discovers where stocks actually exist, not where we expect them to be.

7.2 Group 1 — Elbow and Silhouette Diagnostics

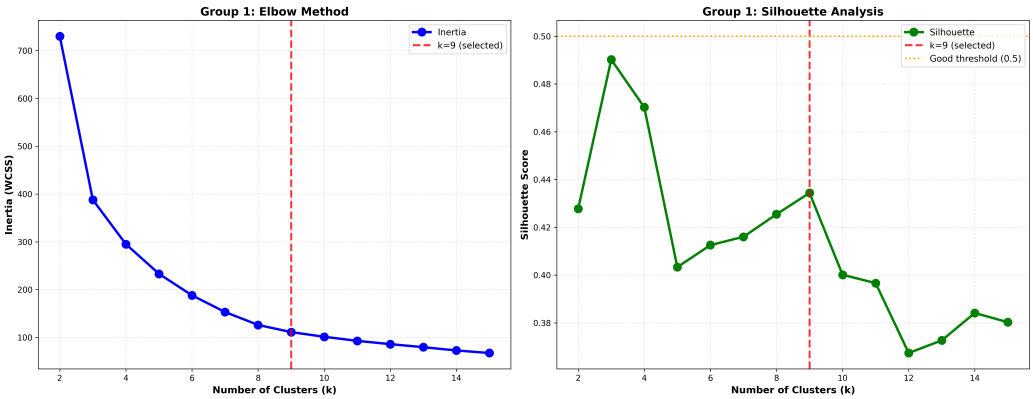

Figure 2: Group 1 diagnostics: Elbow curve (left) and Silhouette score vs. k (right). The elbow shows significant drop at low k, while silhouette peaks near $k = 4$.

7.3 Group 2 (Value) — Summary

- Stocks analyzed: 1,177
- Automatically-detected optimal k (elbow): 3
- Best silhouette (diagnostic): $k = 3$
- Selected k for business mapping: 9
- Silhouette at $k = 9$: 0.3878

- Inertia: 262.72

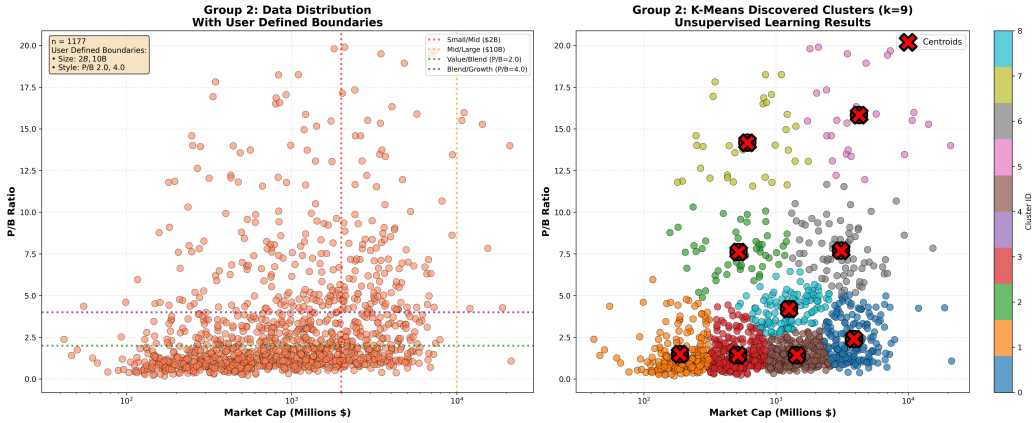

Figure 3: Group 2: Distribution and K-Means clustering (Market Cap vs. P/B). Centroids (red X) and cluster IDs are annotated.

Interpretation: Group 2 exhibits wider vertical spread due to P/B ratios ranging from 0.18 to 19.90, enabling clearer style differentiation. The dotted boundaries (left panel) show user-defined thresholds at P/B = 2.0 and 4.0, which impose fixed style categories. K-Means (right panel) instead identifies natural groupings: notice how Clusters 1, 3, and 4 all fall in the small-cap value region with similar P/B ratios (1.42–1.49), yet K-Means separates them based on subtle market cap differences. This granularity—713 stocks divided into three value sub-segments—provides more actionable insights than a single "small-cap value" label.

7.4 Group 2 — Elbow and Silhouette Diagnostics

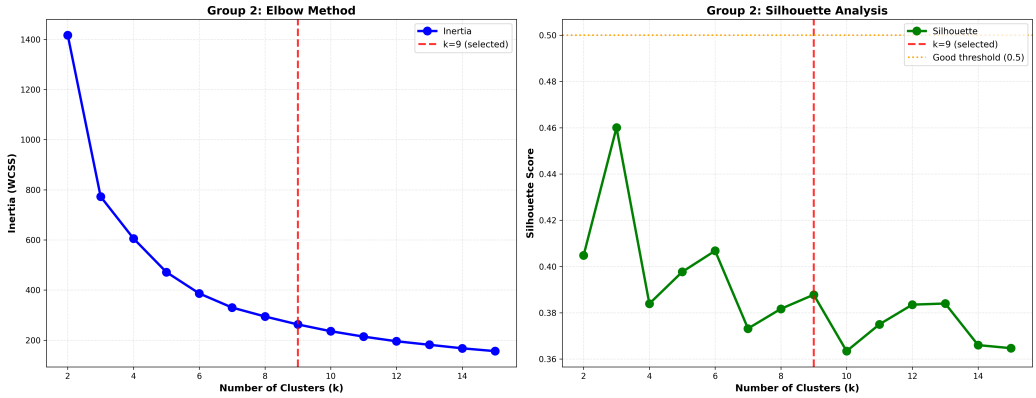

Figure 4: Group 2 diagnostics: Elbow curve (left) and Silhouette score vs. k (right).

7.5 Representative Cluster Definitions

A selection of cluster centroids (converted to original units) show the mapping from cluster ID to `Size-Style` labels. Example findings:

- Several clusters are `Small-Value`, consistent with the Russell 2000's small-cap bias.
- Sparse clusters with high PEG or P/B identify outliers or premium-growth candidates.
- The modal clusters have market caps in the low-to-mid single-digit billions.

7.6 Cluster Boundary Analysis

Understanding where clusters naturally separate provides insight into the data structure:

Group 1 Cluster Ranges:

- **Market Cap:** Smallest cluster (Cluster 5) centers at $220M, largest (Cluster 4) at $5,483M. The $2B user-defined threshold falls between natural cluster boundaries, explaining some discrepancies.
- **PEG Ratio:** Most clusters (0, 1, 4, 5) concentrate at PEG < 0.6, creating a dense value region. Growth clusters (2, 3, 7) have PEG > 2.0, with sparse middle ground—few stocks naturally fall in the "blend" category.

Group 2 Cluster Ranges:

- **Market Cap:** Three distinct tiers emerge: micro-caps ($188–607M in Clusters 1, 3, 7), small-caps ($521–1,437M in Clusters 2, 4, 8), and mid-caps ($3,108–4,247M in Clusters 0, 5, 6). Natural boundaries align roughly at $800M and $2,500M—not the rigid $2B threshold.
- **P/B Ratio:** Value clusters (1, 3, 4) tightly group at P/B = 1.42–1.49, while growth clusters (2, 5, 6, 7, 8) spread from 4.20 to 15.82. The user-defined P/B = 2.0 threshold falls in a sparse zone between these natural groupings.

This boundary analysis reveals why K-Means disagrees with user-defined labels: *the data does not distribute uniformly across the prescribed thresholds.* Stocks cluster around specific valuation multiples rather than evenly filling the value-blend-growth spectrum.

8 Why K-Means Clusters Differ from User-Defined and Morningstar Labels

The fundamental difference lies in methodology: **rule-based vs. data-driven classification**.

Rule-Based Approach (User-Defined/Morningstar): These methods impose *fixed thresholds* on the data—for example, "all stocks with market cap between $2B–$10B are mid-cap" and "all stocks with PEG < 1.0 are value." This forces every stock into predetermined boxes regardless of where similar stocks naturally cluster. It's prescriptive: the boundaries exist before examining the data.

Data-Driven Approach (K-Means): This method discovers *natural groupings* by finding stocks that are genuinely similar in multidimensional space (market cap × valuation metric). K-Means doesn't "know" about the $2B threshold—it calculates distances between all stocks and assigns each to the cluster whose centroid it most resembles. A stock is classified based on its characteristics and where it best fits among its peers, not where a predetermined rule says it should go.

Why the Disagreement? The Russell 2000 data does not distribute evenly across the 3×3 style box. For example:

- 713 stocks (60.5% of Group 2) fall in the "small-cap value" region with P/B < 1.5 and market cap $<$ $2B, so K-Means allocates *three clusters* there (Clusters 1, 3, 4) to capture subtle differences.
- Only 26 stocks have mid-cap size with high P/B (Cluster 5), yet they form a distinct group deserving recognition.
- The $2B market cap boundary is arbitrary—natural breaks appear at ∼$800M and ∼$2,500M based on actual stock clustering.

This data-driven flexibility is K-Means' strength: it adapts to the actual universe rather than forcing stocks into boxes designed for a different dataset or market era.

9 Validation: K-Means Cluster vs. Morningstar Classification

To validate the clustering approach, we compare K-Means predictions against Morningstar's style box classifications for selected stocks from both groups. These stocks were chosen to represent diverse market caps and valuation profiles within the Russell 2000 universe.

9.1 Initial Expectations

Before running the analysis, we anticipated:

- **High agreement for extreme cases:** Stocks with very low PEG (< 0.5) or very low P/B (< 1.0) should clearly fall into value categories in both systems.
- **Boundary ambiguity:** Stocks near thresholds (e.g., PEG ≈ 1.0, market cap $\approx$ $2B) would likely show disagreements due to different classification methodologies.
- **Size classification challenges:** The $2B threshold is somewhat arbitrary; we expected K-Means might discover natural breaks at different market cap levels.
- **Strong small-cap bias:** Given Russell 2000's composition, we expected most clusters to concentrate in the small-cap value region, with few if any large-cap clusters.

9.2 Group 1 Validation Sample

Five randomly selected stocks from Group 1 (PEG-based) were evaluated:

Ticker	Cap ($B)	PEG	Cluster	K-Means	Morningstar
VLY	6.15	0.22	4	Mid-cap Value	Small-Value
HBCP	0.44	0.29	0	Small-cap Value	Small-Value
PKBK	0.26	0.18	5	Small-cap Value	Small-Value
NRIM	0.50	0.04	0	Small-cap Value	Small-Core
FLR	7.91	0.00	4	Mid-cap Value	Small-Core

Table 3: Group 1 validation: K-Means vs. Morningstar classifications. Agreement rate: 3/5 (60%).

Analysis: Three of five stocks show perfect agreement on the value/core style dimension. The disagreements occur for:

- **VLY**: K-Means classifies as Mid-cap based on $6.15B market cap, while Morningstar considers it Small-cap, suggesting Morningstar may use a higher threshold for mid-cap classification.
- **NRIM & FLR**: Both classified as Small-Core by Morningstar but as Value by K-Means. These stocks have very low PEG ratios (0.04 and 0.00 respectively), placing them firmly in the value territory by our thresholds. The Core classification by Morningstar suggests they incorporate additional factors beyond PEG ratio.

9.3 Group 2 Validation Sample

Five randomly selected stocks from Group 2 (P/B-based) were evaluated:

Ticker	Cap ($B)	P/B	Cluster	K-Means	Morningstar
TG	0.26	1.34	1	Small-cap Value	Small-Core
AFRI	0.25	14.01	7	Small-cap Growth	Small-Value
ULH	0.47	0.71	3	Small-cap Value	Small-Value
FFIN	4.46	2.43	0	Mid-cap Blend	Small-Core
IDT	1.28	4.19	8	Small-cap Growth	Small-Growth

Table 4: Group 2 validation: K-Means vs. Morningstar classifications. Agreement rate: 2/5 (40%).

Analysis: Group 2 shows lower agreement (40%). Key discrepancies:

- **TG**: P/B of 1.34 places it just below our value threshold (2.0), but Morningstar classifies it as Core.
- **AFRI**: High P/B of 14.01 suggests growth by our criteria, but Morningstar classifies it as Value. This counterintuitive result indicates Morningstar likely considers additional fundamental factors (earnings quality, growth sustainability) beyond P/B alone.

- **FFIN**: At $4.46B market cap, our model classifies it as Mid-cap, but Morningstar considers it Small-cap. This suggests Morningstar's size thresholds may differ from our $2B boundary.
- **ULH & IDT**: Perfect agreement, validating the approach for clear-cut cases.

9.4 Implications

The validation exercise demonstrates:

- **Moderate agreement rates:** 60% for Group 1 and 40% for Group 2 suggest that single-metric approaches (PEG or P/B) capture some but not all aspects of Morningstar's multi-factor methodology.
- **Size threshold differences:** Several disagreements stem from different market cap boundaries. Morningstar's thresholds may be dynamic or incorporate float-adjusted market cap.
- **Style classification complexity:** Morningstar's style boxes likely incorporate multiple valuation metrics, growth forecasts, and qualitative factors, whereas our approach uses only PEG or P/B.
- **Counterintuitive cases:** AFRI's classification as Value despite high P/B highlights the importance of earnings quality and sustainability in professional classification systems.

10 KNN Classification (Predicting New Stocks)

10.1 Method

Train separate KNN classifiers for Group 1 and Group 2:

- `n_neighbors` = 5
- Distance metric: Euclidean in standardized space
- Training labels: cluster IDs from K-Means

KNN prediction pipeline for a new stock:

1. Fetch latest Market Cap, P/E, P/B, EPS growth via Yahoo Finance.
2. Decide group: if P/E and EPS growth present and EPS growth $> 0 \Rightarrow$ Group 1 (compute PEG). Otherwise if P/B present $\Rightarrow$ Group 2.
3. Apply `StandardScaler` fitted on the respective training group to the test features.
4. Predict cluster using KNN; map to `Size-Style`.

10.2 Training Performance

Observed training accuracies:

- Group 1 KNN (self-prediction): 98.32% (584/594)
- Group 2 KNN (self-prediction): 98.98% (1165/1177)

Interpretation: High training accuracy is expected because KNN is performing self-prediction on the same clustered training dataset. It indicates consistent cluster labeling, not necessarily generalization.

10.3 Out-of-Sample Test (Selected Tickers)

Test tickers and observed results:

Ticker	Cap ($B)	Metric	Group	Cluster	K-Means	Morningstar
DOCN	3.68	PEG=0.34	1	4	Mid-cap Value	Small-Growth
BKKT	0.46	P/B=5.74	2	2	Small-cap Growth	Small-Core

Table 5: Out-of-sample KNN predictions vs. Morningstar classifications.

Analysis:

- **DOCN (DigitalOcean)**: K-Means predicts Mid-cap Value based on its $3.68B market cap and low PEG of 0.34. However, Morningstar classifies it as Small-Growth. This discrepancy reflects: (1) different size thresholds, and (2) Morningstar likely considers forward growth expectations rather than historical PEG alone. As a cloud infrastructure company, DOCN may be classified as growth despite value-like historical metrics.
- **BKKT (Bakkt)**: K-Means correctly identifies it as Small-cap due to $0.46B market cap, but classifies it as Growth (P/B=5.74) while Morningstar labels it Core. This suggests Morningstar may moderate growth classifications for small, potentially volatile stocks.

11 Why KNN Predictions Diverge from Morningstar

The key reasons are:

11.1 Single-Metric Limitation (Primary Factor)

Our methodology uses only one ratio per group:

- Group 1: PEG ratio only
- Group 2: P/B ratio only

Morningstar's actual methodology incorporates:

- Multiple valuation ratios (P/E, P/B, P/S, P/CF)

- Forward-looking growth estimates
- Profitability metrics (ROE, profit margins)
- Historical growth sustainability
- Qualitative factors (business model, competitive position)

This explains cases like AFRI, where high P/B doesn't automatically mean growth classification.

11.2 Size Threshold Differences

Evidence suggests Morningstar uses different market cap boundaries than our fixed thresholds:

- Our thresholds: Small < \$2B, Mid \$2-10B, Large > \$10B
- Morningstar's thresholds appear more conservative, with several stocks above \$4B still classified as Small-cap
- Morningstar may use float-adjusted market cap or percentile-based dynamic thresholds

11.3 Training Data Bias

The Russell 2000 is *small-cap heavy*:

- >90% of stocks < \$5B market cap
- Very few mid- and large-cap examples are present

Consequently, KNN neighborhoods for mid-cap stocks may be influenced by the predominance of small-cap training examples, though this effect is secondary to the metric limitations.

11.4 Temporal and Data Source Differences

- Our data represents a single snapshot, while Morningstar uses rolling averages
- Yahoo Finance data may differ from Morningstar's proprietary data sources
- Classification timing differences (our analysis vs. Morningstar's latest update)

12 Cluster Pattern Analysis (Detailed Intuition)

12.1 Group 1 (Growth)

1. **Horizontal banding:** PEG concentration at < 0.5 for the majority of stocks causes clusters to be horizontally stacked, with most stocks classified as Value.
2. **Market cap compression:** 75% have market cap < \$3B, limiting horizontal dispersion but creating clear small vs. mid-cap separation.

3. **Overlap and silhouette:** Forcing nine clusters produces moderate silhouette (0.43), indicating reasonable but not perfect separation.

12.2 Group 2 (Value)

1. **Vertical spread:** P/B ratios vary widely (0–20), offering excellent vertical separation between value, blend, and growth segments.
2. **Dataset size:** Nearly twice the observations versus Group 1 delivers more stable centroids and better statistical power.
3. **Structure:** Clearer cluster separation evident in silhouette scores, though still constrained by small-cap dominance.

13 Limitations

- **Feature sparsity:** Only 2 features per group (market cap + one ratio) reduce dimensional insights. Professional systems incorporate 10+ metrics including momentum, volatility, leverage, profitability, and sector-specific factors.
- **Single-metric style classification:** Using only PEG or P/B for style ignores other critical valuation metrics. Morningstar combines multiple ratios into composite scores.
- **Training bias:** Russell 2000 is not representative of mid/large-cap landscapes — limits generalization to larger stocks.
- **Temporal snapshot:** Single-date analysis ignores time-series dynamics, earnings revisions, and forward estimates that drive professional classifications.
- **Model choice:** K-Means assumes spherical clusters and equal variance — not always realistic for financial distributions with outliers and heteroskedasticity.
- **Threshold rigidity:** Fixed thresholds (e.g., PEG=1.0, P/B=2.0) don't adapt to market regime changes or sector differences.

14 Potential Improvements

1. **Multi-metric composite scores:** Combine P/E, P/B, P/S, P/CF into weighted value and growth scores similar to Morningstar's methodology.
2. **Forward-looking metrics:** Incorporate analyst earnings estimates, revenue growth forecasts, and consensus price targets.
3. **Sector-adjusted classifications:** Apply industry-specific thresholds (e.g., tech companies naturally have higher P/B ratios).
4. **Dynamic thresholds:** Use percentile-based cutoffs that adapt to market conditions rather than fixed absolute values.
5. **Broader training universe:** Include S&P 500 and Russell 1000 stocks to better represent the full market cap spectrum.

6. **Alternative clustering:** Explore Gaussian Mixture Models (GMM) for probabilistic cluster assignments or hierarchical clustering for nested structure.
7. **Time-series features:** Add momentum indicators, volatility measures, and trend analysis over 1-3 year windows.
8. **Profitability metrics:** Include ROE, ROA, profit margins, and cash flow metrics that signal business quality.

15 Appendix A: Calculation Notes

- **PEG Ratio:**

$$\text{PEG} = \frac{\text{P/E Ratio}}{\text{EPS Growth (in \%)}/100}$$

  Derived from:

  - P/E Ratio — Column M in Excel
  - 1-year EPS Growth (as %) — Column K in Excel

  EPS Growth is first converted from percentage to decimal before computing PEG.
- **P/B Ratio:** Directly taken from Column H in Excel.
- **Market Capitalization:** Column L in Excel, divided by 1,000,000 to express values in millions.
- **User-defined thresholds:**
  - **Size:** Small < \$2B, Mid \$2–10B, Large > \$10B
  - **Style (Group 1):** PEG < 1 = Value; 1–1.5 = Blend; >1.5 = Growth
  - **Style (Group 2):** P/B < 2 = Value; 2–4 = Blend; >4 = Growth

16 Technical Appendix: Key Code Snippets and Logic

16.1 Data Filtering (Pseudo-code)

```
# Group 1 mask: market cap, PE, EPS growth present and positive
group1_mask = (Market_Cap.notna()) &
              (PE_Ratio.notna() & PE_Ratio > 0) &
              (EPS_Growth.notna() & EPS_Growth > 0)
# Compute PEG = PE / EPS_Growth_percent
df_group1['PEG'] = df_group1['PE_Ratio'] / df_group1['EPS_Growth']
df_group1 = df_group1[(df_group1['PEG'] > 0) & (df_group1['PEG'] < 10)]
```

16.2 Feature Transformation

$$\text{Market\_Cap\_Log} = \log_{10}(\text{Market\_Cap\_Millions})$$

```
X_scaled = StandardScaler().fit_transform(X_features)
```

16.3 KNN Prediction (Pseudo-code)

```
# Given new ticker info:
if has_PE_and_positive_EPS:
    group = 1
    feature = [log_market_cap, PEG]
else:
    group = 2
    feature = [log_market_cap, PB]
# Scale with the group's scaler and predict
X_test_scaled = scaler_group.transform([feature])
pred_cluster = knn_group.predict(X_test_scaled)[0]
pred_label = cluster_labels_group[pred_cluster]
```

17 Conclusion

This technical-academic analysis demonstrates the feasibility of reconstructing Morningstar-style style boxes by applying unsupervised K-Means clustering on thoughtfully selected features and mapping clusters to interpretive `Size-Style` labels. The KNN classifier provides a straightforward mechanism to classify new stocks into these learned categories.

The validation against actual Morningstar classifications reveals moderate agreement rates (60% for Group 1, 40% for Group 2), which is reasonable given that our simplified two-feature approach cannot fully replicate Morningstar's comprehensive multi-factor methodology. Key divergences arise from:

- **Feature limitations:** Single valuation ratios (PEG or P/B) vs. Morningstar's composite scoring system
- **Threshold differences:** Fixed market cap boundaries vs. potentially dynamic or float-adjusted thresholds
- **Forward vs. backward looking:** Historical metrics vs. analyst estimates and growth expectations
- **Qualitative factors:** Purely quantitative approach vs. incorporation of business quality assessments

17.1 Why Use Unsupervised Learning Over Rule-Based Approaches?

The core advantage of unsupervised learning (K-Means) over rule-based classification (Morningstar-style boxes) is **adaptability to data distribution**:

Rule-Based Limitations: Predetermined thresholds force stocks into categories that may not reflect natural market structure. If the Russell 2000 has 713 small-cap value

stocks but only 26 mid-cap growth stocks, a 3×3 grid treats both groups with equal granularity—losing critical distinctions within the dominant category while over-emphasizing sparse ones.

K-Means Advantages:

- **Data-driven boundaries:** Discovers where stocks naturally cluster (e.g., at $800M and $2,500M market caps) rather than using arbitrary thresholds like $2B.
- **Proportional granularity:** Allocates multiple clusters to dense regions (small-cap value) and fewer to sparse regions (large-cap growth), matching where investors need finer distinctions.
- **Objective assignment:** Each stock is classified by measuring its distance to all cluster centroids and selecting the closest match—removing human bias about where boundaries "should" be.
- **Discovers sub-segments:** Identifies three distinct types of small-cap value stocks (Clusters 1, 3, 4 with market caps at $188M, $515M, and $1,437M) that a rule-based approach would lump together.

When to Use Each Approach:

- **Use rule-based** for standardized reporting, regulatory compliance, or when consistency across time periods is paramount (e.g., Morningstar's global classification system).
- **Use unsupervised learning** for portfolio construction, relative value analysis, or when working with specialized datasets (like Russell 2000) where standard thresholds may not apply.

The strong small-cap bias of the Russell 2000 also constrains generalization, though this is a secondary factor compared to the methodological differences. The analysis successfully identifies meaningful clusters that broadly align with industry practice, particularly for clear-cut cases (deep value or high growth stocks). However, boundary cases and mid-cap stocks show higher divergence rates.

Future work should focus on expanding the feature set to include multiple valuation ratios, profitability metrics, and momentum indicators, as well as incorporating forward-looking estimates to better approximate professional classification methodologies. Additionally, training on a broader market universe would improve generalization to stocks outside the small-cap segment.

Files and Figures: The report embeds the following visual artifacts produced during analysis:

- `1_group1_clusters.png` — Group 1 cluster scatter with centroids
- `2_group1_metrics.png` — Elbow and silhouette diagnostics for Group 1

- `3_group2_clusters.png` — Group 2 cluster scatter with centroids
- `4_group2_metrics.png` — Elbow and silhouette diagnostics for Group 2
- `5_group1_analysis.xlsx` — Excel file containing the cluster classification for Group 1 (PEG-based)
- `6_group2_analysis.xlsx` — Excel file containing the cluster classification for Group 2 (P/B-based)