Are Sector Labels Still Relevant? A Clustering Approach to South African Equity Grouping

Executive Summary

Purpose

This report investigates whether traditional sector classifications (e.g., Financials, Basic Materials) remain effective in grouping South African equities, compared to a **data-driven hierarchical clustering approach** based on **weekly return behavior**.
By evaluating **within-group cohesion** and **between-group separation**—key criteria for effective grouping—the study assesses the statistical robustness and practical relevance of each method. Special attention is given to results before and after removing the **market-wide effect** from returns.

Methodology

- **Data:** Weekly return data from Bloomberg covering the top 100 JSE stocks by market cap (final sample: 90 stocks), from **01-Aug-2010 to 01-Jun-2025**.
- **Preprocessing:**
  - Outliers treated and missing data backfilled using a correlation-based imputation.
  - **Market effect removed** using a linear regression on the first principal component (PC1), isolating idiosyncratic return dynamics.
  - Returns were **standardized (Z-score)** and clipped at ±10 standard deviations.
- **Clustering Models:**
  - **Hierarchical Clustering** (Ward's method), optimized to 10 clusters.
  - Benchmarked against **sector-based groupings**.

Key Findings

1. Raw Returns (Unadjusted)

- **Sector groupings show weak cohesion** (avg. within-sector correlation ≈ 0.09) and poor separation (between ≈ 0.00).
- **Hierarchical clustering performs slightly better**, but the market effect masks deeper structure.

2. Market-Adjusted Returns (Excluding PC1)

- **Hierarchical clustering significantly outperforms sectors**:
  - **Within-cluster cohesion** improves to **0.34 average**, with certain clusters exceeding **0.9**.
  - **Between-cluster separation** strengthens, averaging **–0.08**.
  - Clusters align with **economic themes** (e.g., Gold, Platinum, Rand Hedges), regardless of sector labels.
- **Sector groupings remain weak**, suggesting they do not reflect true return relationships after removing systematic factors.

Implications

- Traditional sectors **fail to capture co-movement** in South African equity returns.
- **Hierarchical clustering enables more accurate groupings** for:
  - Portfolio diversification
  - Risk management
  - Alpha signal construction
  - Thematic investing
- Removing the market effect is **crucial** for revealing true relationships between stocks.

Recommendations

1. **Integrate hierarchical clustering** into portfolio construction and risk models.
2. Use **idiosyncratic return structures**—not just sectors—for building diversified baskets.
3. Apply cluster-based analysis for **thematic investing and long/short strategies**.
4. Rethink performance attribution based solely on sector benchmarks.
5. Enhance internal screening and backtesting processes using clustering logic.

Conclusion

Sector labels, while convenient, are increasingly **insufficient** for understanding equity behavior in dynamic markets like South Africa. **Hierarchical clustering, especially on market-adjusted returns, reveals deeper structure**, enabling smarter grouping strategies that are statistically sound and economically intuitive.

Introduction

Subject

This report looks at whether grouping stocks by their official sectors (like financials, industrials, or technology) is still the best way to understand how they behave. Using return data from the largest companies on the JSE, we tested two alternative methods — KMeans and Hierarchical Clustering — which group stocks based on how similarly their prices move over time. We then compared these groupings to the standard sector groupings by measuring how closely the stocks within each group moved together. The goal is to see if these data-driven methods can reveal better insights for diversification and portfolio construction than traditional sector labels.

Background

In investment management, stocks are typically grouped by sectors such as financials, industrials, or consumer goods. These groupings are widely used for portfolio construction, performance attribution, and risk management. However, sector classifications may not always reflect how stocks actually behave in the market — particularly in terms of how their returns move together.

With advances in data analysis, we now have tools that can group stocks based on actual price behavior rather than predefined labels. Clustering techniques like KMeans and Hierarchical Clustering offer a way to identify stocks that move similarly over time, regardless of their sector.

This report investigates whether these data-driven groupings can provide a more meaningful way to understand stock relationships and construct diversified portfolios, especially in the context of the South African equity market.

Objective

The main objective of this report is to evaluate whether grouping stocks based on how their returns behave over time can provide better insights than grouping them by traditional sector labels. To do this, we:

- Collected weekly return data for the top JSE-listed stocks.
- Cleaned and prepared the data by removing outliers, filling in missing values, and adjusting for the overall market effect.
- Applied two different clustering methods — KMeans and Hierarchical Clustering — to group stocks based on similarities in their return patterns.
- Compared these data-driven clusters to sector-based groupings by measuring how closely stocks in each group moved together (within-group correlation) and how differently they moved from other groups (between-group correlation).
  This analysis aims to help portfolio managers and analysts determine whether return-based groupings offer more useful insights for risk management, diversification, and stock selection.

Methodology

- Weekly return data was sourced from Bloomberg for the top 100 JSE-listed stocks by market capitalisation, covering the period from **01 August 2010 to 01 June 2025**. After removing stocks with significant missing data and extreme outliers, the working dataset consisted of **90 stocks**.
- To address missing values for stocks with partial histories, we implemented a **correlation-based backfilling** approach: for each stock with gaps, we identified its top three most correlated peers (based on available returns) and imputed missing values using the **average return of those peers**.
- To isolate stock-specific return behaviour, we removed the **broad market effect** using **Principal Component Analysis (PCA)**. The **first principal component (PC1)**, representing the primary market factor, was extracted from the return matrix. We then regressed each stock's return series on PC1 using **linear regression** and took the **residuals** as the market-adjusted returns to isolate idiosyncratic movement.
- These residuals were then **standardised into Z-scores**, ensuring each stock had a mean of 0 and a standard deviation of 1. To mitigate the impact of extreme outliers, any values exceeding **±10 standard deviations** were clipped at ±10.
- Two clustering algorithms were applied to the standardised, market-adjusted returns:
  - **KMeans Clustering**, with the optimal number of clusters evaluated using the **Xie-Beni index** and the **elbow method**.
  - **Agglomerative Hierarchical Clustering**, implemented using **Ward's linkage criterion**, which minimizes within-cluster variance during the merging process.
- We determined that the optimal configuration was **10 clusters** using **Hierarchical Clustering with Ward's method**.
- To assess the effectiveness of the return-based clustering, we compared the results to a baseline using **sector-based groupings**. Evaluation was based on:
  - **Within-cluster average pairwise correlations** (i.e., how similar returns were among

stocks within the same cluster).
- **Between-cluster average pairwise correlations** (i.e., how dissimilar returns were across different clusters).
These metrics were computed both **with and without** the market effect, providing a robust comparison between traditional sector labels and data-driven return groupings. **PCA plots** and a **dendrogram** are included to visualise the clustering outcomes.

Technical Notes (for Non-Statisticians)

- **PC1 (First Principal Component):**
PC1 captures the dominant pattern of movement shared across all stocks, similar to a market index. By removing PC1, we strip out market-wide effects and focus only on individual or group-specific return behavior—known as **idiosyncratic risk**.

• Ward's Method (in Hierarchical Clustering):
Ward's method forms clusters by successively merging stocks in a way that **minimizes the increase in within-cluster variance**. This ensures stocks in the same group behave similarly, while maintaining clear differences across clusters.

• Xie-Beni Index:
The Xie-Beni Index is used to determine the **optimal number of clusters**. It favors clusters that are both **internally cohesive** and **well-separated** from one another. A **lower score** indicates better clustering structure.

Scope and Limitations

Scope:

This report focuses on comparing traditional sector-based stock classifications with return-based clustering techniques for the **top 100 market cap stocks listed on the JSE**. The analysis is limited to **weekly total return data** from **01 August 2010 to 01 June 2025**, providing a long-term view across multiple market cycles. Clustering methods applied include **KMeans** and **Hierarchical Clustering (Ward's method)** on standardised, market-adjusted returns. The goal is to assess the effectiveness of each clustering approach using statistical measures of within- and between-group correlation.

Limitations:

- **Universe Constraint**: The analysis is limited to the top 100 JSE-listed stocks by market cap. Smaller-cap stocks or those with insufficient data were excluded, which may bias results toward more stable, liquid companies.

- **Static Clustering**: Clusters were formed using the full historical period, assuming relationships between stocks are stable. In practice, stock behaviours may evolve over time; a dynamic or rolling clustering approach might yield different results.
- **Market Effect Removal Assumption**: The use of **PC1** to remove market-wide effects assumes that a single linear factor captures the dominant driver of co-movement. While common, this method may oversimplify broader macroeconomic influences.
- **Evaluation Metrics**: Correlation-based metrics (within and between clusters) are useful for measuring co-movement but do not capture all dimensions of financial relationships such as tail risk or non-linear dependencies.
- **No Out-of-Sample Testing**: This study does not explicitly test whether the clusters improve portfolio performance, risk-adjusted returns, or asset allocation decisions in a real-world setting.

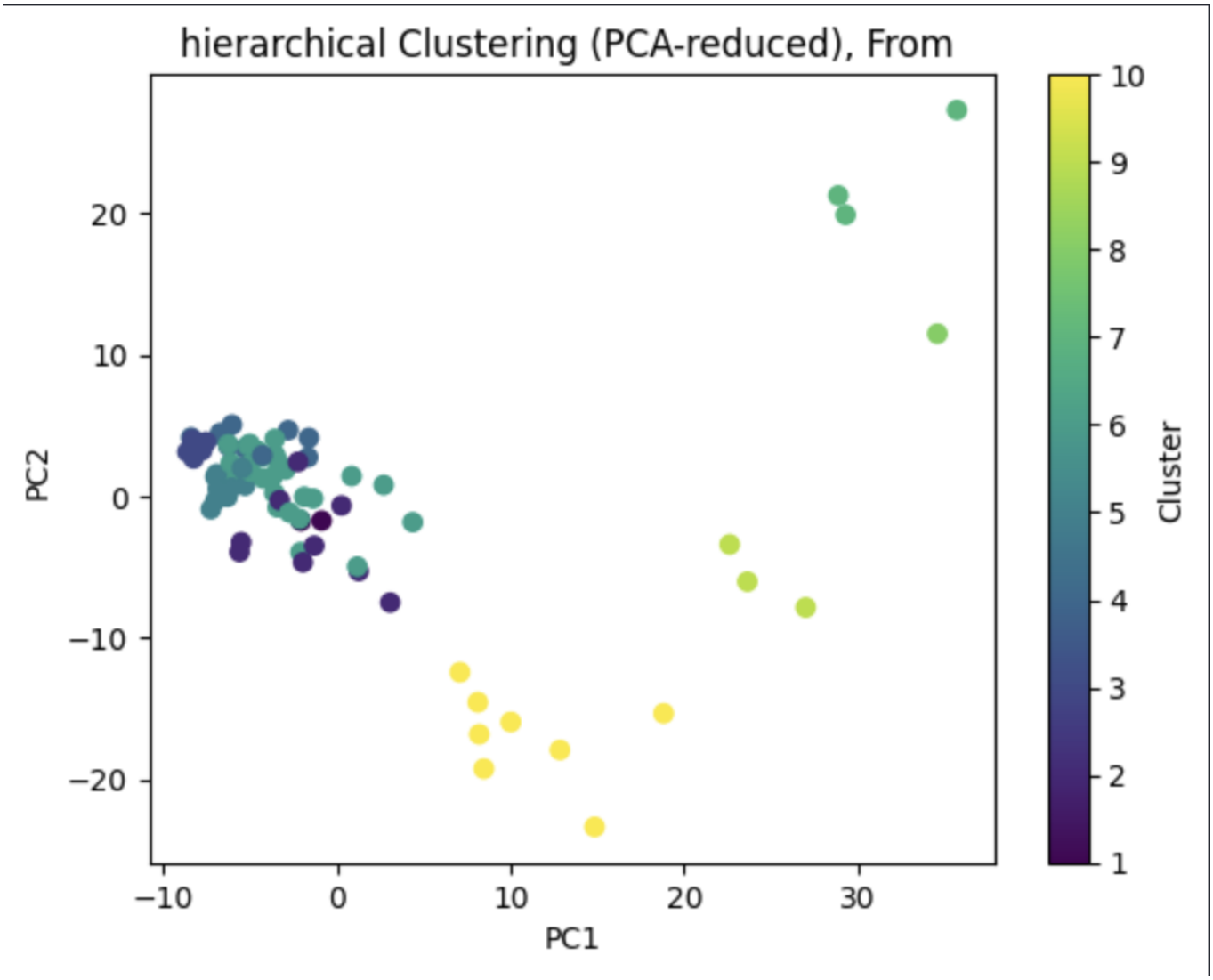
Despite these limitations, the analysis offers a robust comparison of traditional and data-driven equity classification methods and provides a foundation for further research or integration into portfolio construction frameworks.

PCA Comparison of Return-Based vs Sector-Based Clustering

To compare the structure and cohesiveness of return-based clustering against traditional sector classifications, we applied **Principal Component Analysis (PCA)** to the standardised, market-adjusted weekly returns of JSE-listed equities. This dimensionality reduction technique projects each stock into a two-dimensional space (PC1 vs PC2), capturing the most significant variance in return behavior.

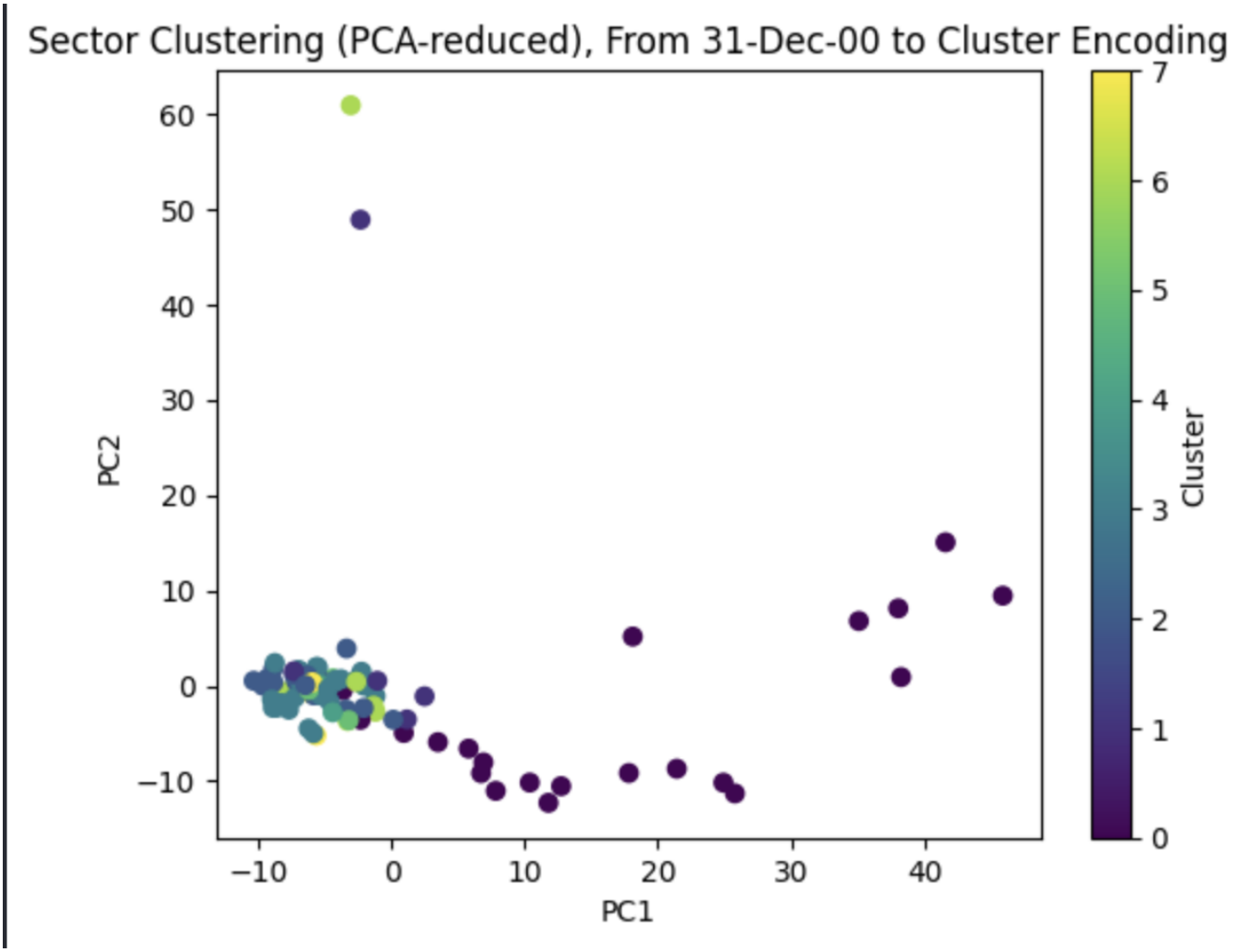
Each point represents a stock, and the color indicates its assigned cluster under either the hierarchical or sector-based approach.

Return-Based Clustering (Hierarchical, Ward's Method)



- The first PCA plot shows cluster assignments based on **weekly return behavior**, grouped using **hierarchical clustering** with Ward's method (10 clusters).
- **Observations**:
  - **Clear separation** exists between several clusters, especially along PC1. For example, **Cluster 10** (likely diversified miners) and **Cluster 9** (platinum miners) form visibly distinct groups.
  - Most clusters are **tightly packed**, indicating a **high degree of within-cluster similarity** and relatively **low overlap** with other clusters.
  - The spread across both PC1 and PC2 suggests the model captured **multi-dimensional differences** in return dynamics — including volatility, sensitivity to macro factors, and sector-independent movements.

Sector-Based Clustering



- This plot displays PCA-reduced positions of the same stocks, colored by their **GICS sector** classification.
- **Observations**:
  - The vast majority of stocks are concentrated in a **single dominant cluster** (dark purple), despite representing multiple sectors.
  - While some outliers are visible — potentially corresponding to idiosyncratic or globally exposed stocks — the lack of clear separation indicates that sector-based clustering **fails to capture significant return-based differentiation**.
  - Overlapping sector behavior may result from shared macroeconomic influences or outdated assumptions about risk similarity within sectors.

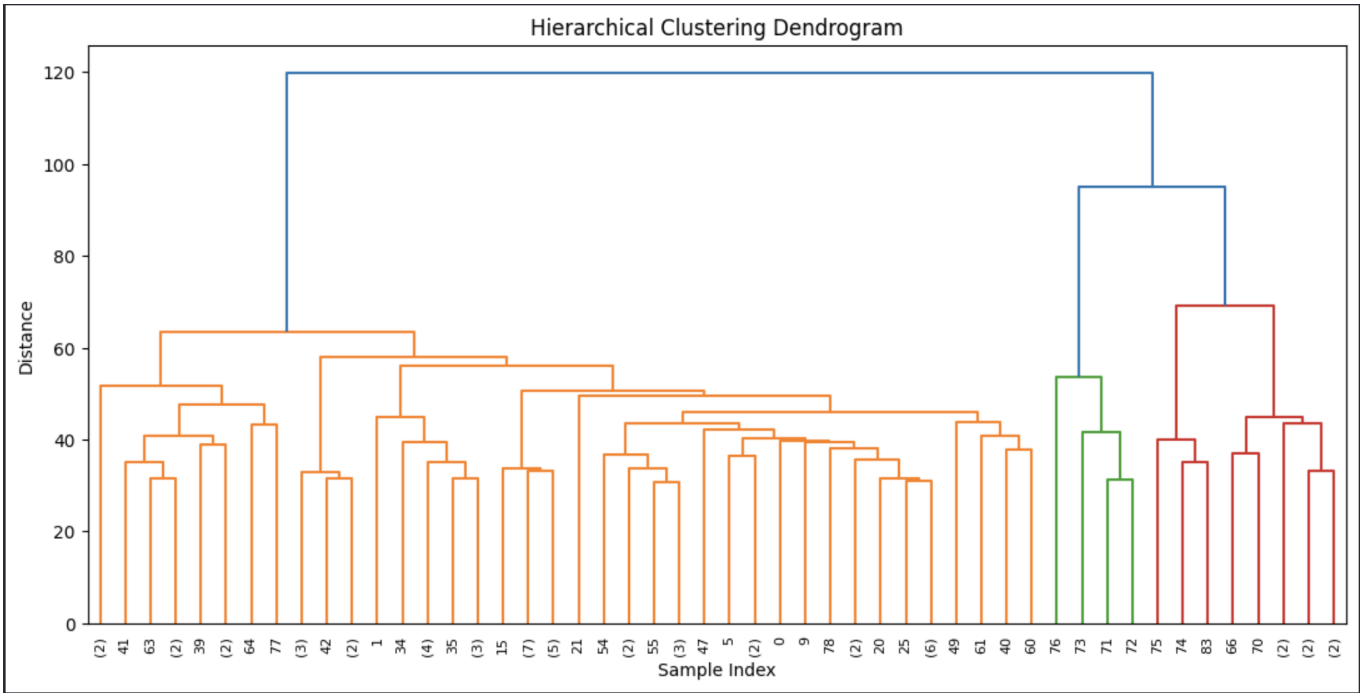
Key Takeaways:

Aspect	Return-Based Clustering (Hierarchical)	Sector-Based Clustering
Cluster separation	Strong — distinct visual groups	Weak — high overlap

Aspect	Return-Based Clustering (Hierarchical)	Sector-Based Clustering
Intra-cluster cohesion	High — stocks cluster tightly	Low — broad dispersion
Economic meaning	Emergent from data (e.g., Rand hedge, miners, REITs)	Predefined, may not reflect return behavior
Outlier identification	Clear — e.g., Sibanye forms a unique cluster	Ambiguous

The visual evidence strongly supports that **return-based clustering provides a more coherent and statistically meaningful grouping of stocks** than traditional sector labels. It identifies **latent structures** in market behavior that sector classifications fail to capture — particularly relevant in the South African context, where sector concentration and dual listings can obscure actual stock behavior.

Dendrogram Interpretation and Cluster Themes



Dendrogram Interpretation

The dendrogram shown above illustrates the hierarchical clustering structure generated from the market-adjusted and standardised weekly return data for the top 90 JSE stocks. Each leaf on the horizontal axis represents a stock (indexed), and the vertical lines represent the

distances at which clusters merge, using **Ward's method**, which minimises the total within-cluster variance.

Key Observations:

- **Clear Hierarchical Structure**: Stocks are grouped progressively based on similarity in return behavior. Clusters at the bottom merge into larger groups as the distance threshold increases.
- **Dissimilar Outliers**: The dendrogram shows **a few branches with longer vertical distances before merging** (notably near the right-hand side), suggesting that these stocks behave very differently from the rest. These may include highly volatile or sector-specific stocks such as commodity producers or tech-heavy firms.
- **Natural Cutoff for 10 Clusters**: A cut around a vertical distance of ~60–70 results in approximately 10 distinct clusters. This supports the choice of 10 clusters in the final model, as it balances granularity with interpretability.

Cluster Themes

The dendrogram resulting from hierarchical clustering using Ward's method was used to identify 10 distinct clusters based on the similarity in **market-adjusted weekly return profiles**. These clusters reveal groupings that often cut across traditional sector lines and instead reflect **common economic exposures**, **market factors**, or **idiosyncratic risk drivers**. Below is a thematic summary of the identified clusters:

Cluster 1 — Naspers & Prosus (Tencent Exposure)

This cluster contains **only two companies**: **Naspers** and **Prosus**. Both are heavily exposed to **Tencent** and global tech through a similar investment structure. Their return profiles are nearly indistinguishable and highly idiosyncratic relative to the rest of the market.

Cluster 2 — Rand Hedge Stocks

This group includes multinationals such as **AB InBev, British American Tobacco, Richemont, Investec plc**, and **Mondi** — companies with **large offshore earnings** and strong exposure to **currency fluctuations** and **global macroeconomic trends**. These stocks tend to behave similarly due to their **rand-hedge characteristics**.

Cluster 3 — Domestic Retailers

This cluster consists of **South African-focused retailers** such as **Pepkor, Mr Price, Woolworths, Foschini**, and **Truworths**. Their returns likely reflect **consumer sentiment**,

interest rate cycles, and **local spending trends**, making them sensitive to domestic economic conditions.

Cluster 4 — REITs and Property Stocks

This group is dominated by **real estate investment trusts (REITs)** such as **Growthpoint, Redefine, Vukile**, and **Equites**, whose returns are **interest-rate sensitive** and affected by **bond yields**, **property cycles**, and **macroeconomic stability**.

Cluster 5 — Traditional Financials

This cluster includes large **South African banks, insurers**, and **investment holding firms**: **FirstRand, Capitec, Standard Bank, Nedbank, ABSA, Sanlam**, and **Old Mutual**, among others. Their return patterns are **correlated through systemic financial exposure** and domestic credit conditions.

Cluster 6 — Mixed Domestic & Defensive Stocks

This is the most diverse cluster, containing firms from **telecoms, retail, healthcare, asset management**, and **industrials**. Notable constituents include **Vodacom, MTN, Shoprite, Aspen, Clicks, Santam**, and **Coronation**. While sectorally diverse, these firms may share **stable earnings, defensive characteristics**, and **mid-cap risk profiles**, leading to similar return patterns despite different sectors.

Cluster 7 — Gold Miners

Includes **Anglogold, Gold Fields**, and **Harmony**. These companies are tightly linked through **gold price exposure**, **currency sensitivity**, and **commodity cycles**.

Cluster 8 — Sibanye Stillwater

Sibanye forms its own cluster, likely due to its **unique exposure to both gold and platinum**, **labour unrest**, and **idiosyncratic risk events** that differentiate it from pure-play miners.

Cluster 9 — Platinum Miners

Includes **Impala, Northam**, and **Valterra Platinum** — all focused on **PGMs (Platinum Group Metals)** with similar cost structures, exposure to industrial demand, and sensitivity to the **auto sector** (catalytic converters).

Cluster 10 — Diversified Miners

Large diversified mining firms such as **BHP, Anglo American, Glencore**, and **Kumba Iron Ore** form this cluster. These firms' return dynamics are shaped by **bulk commodity prices**, **global**

demand, and **supply chain shifts**.

Within-Cluster vs. Between-Cluster Correlations

This section evaluates how effective the clustering methods are by examining the **average correlation between stocks within the same cluster** (cohesion) and **across different clusters** (separation).

We compare both clustering methods — **return-based (hierarchical)** and **sector-based** — across two conditions:

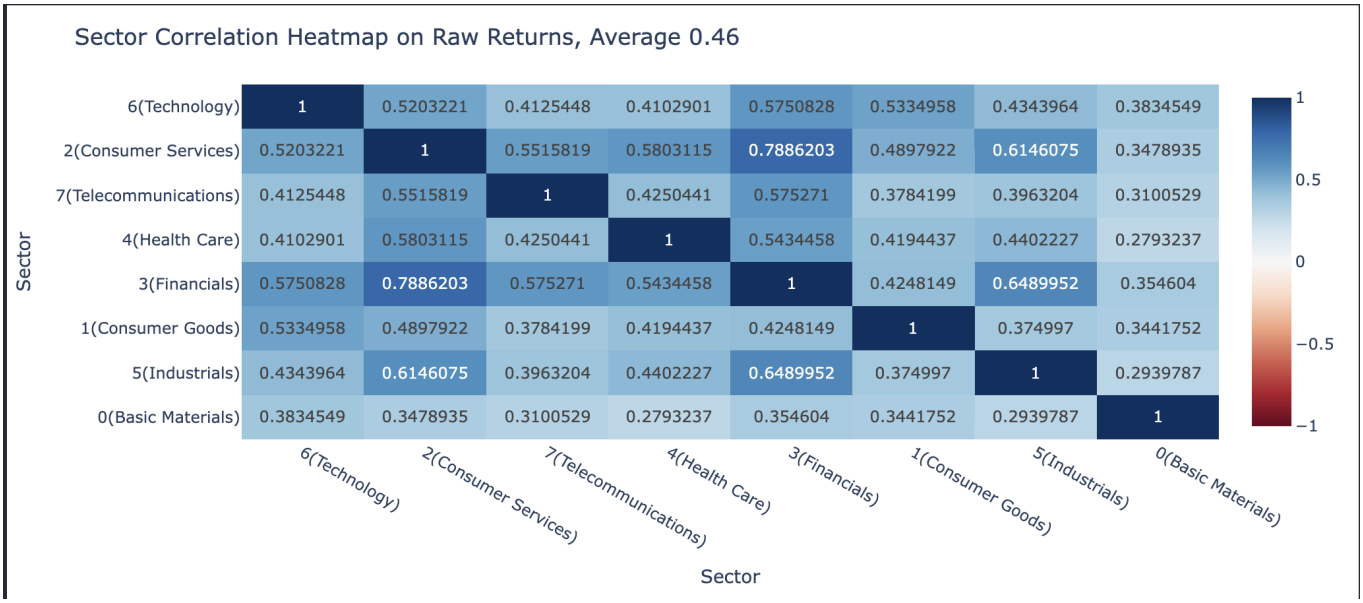
1. Using raw weekly returns
2. Using **market-adjusted returns**, where the market effect was removed by regressing each stock's return on the first principal component (PC1) and taking the residuals.

Raw Returns

Between Cluster/Section Correlations (separation)

To evaluate the separation between groups, we compare the average pairwise correlations across clusters and sectors. A lower average between-group correlation indicates greater distinctiveness between clusters/sectors, which is desirable for a clustering methodology aiming to create heterogeneity across groups.

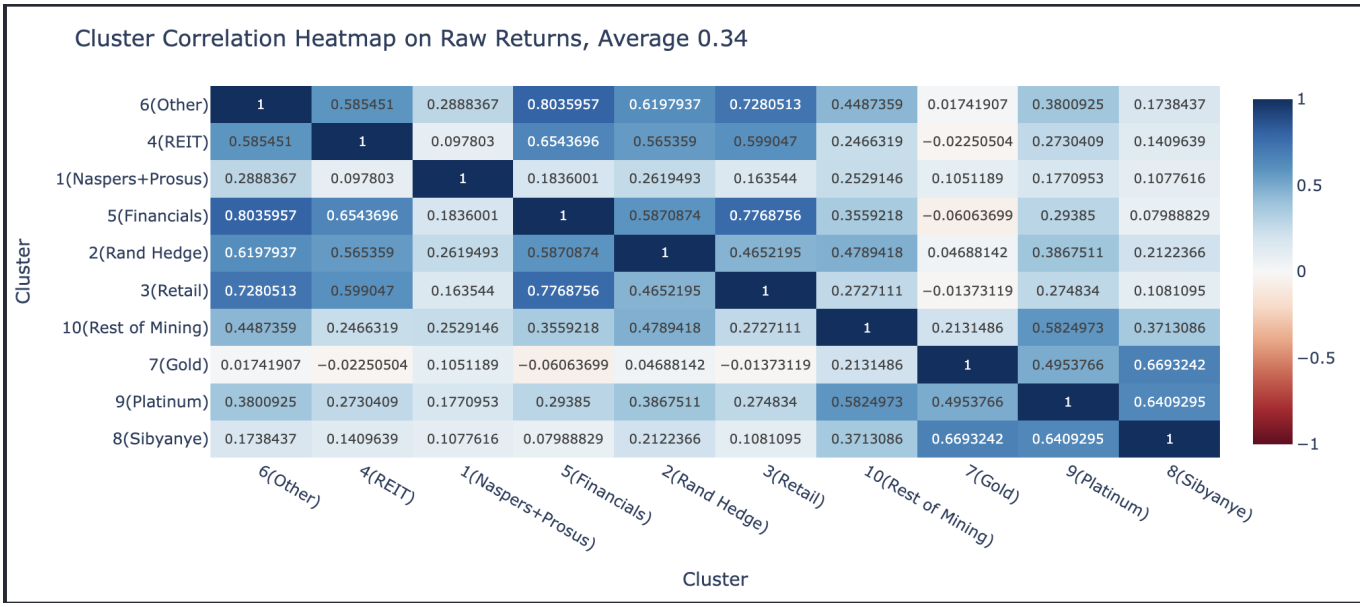
Sector-Based Groupings



From the **Sector Correlation Heatmap**, we observe that the average pairwise correlation across sectors is **0.46**. Certain sector pairs such as *Consumer Services–Financials (0.79)* and

Industrials–Financials (0.65) show elevated cross-sector correlation, indicating that traditional GICS sector definitions may group together stocks that co-move significantly. However, many other sector pairs maintain moderate to low correlation, demonstrating partial separation across sectors.

Hierarchical Clustering-Based Groupings



In contrast, the **Hierarchical Clustering Correlation Heatmap** reveals an overall **lower average between-cluster correlation of 0.34**. Notably, clusters such as *(1) Naspers+Prosus*, *(7) Gold*, and *(8) Sibanye* exhibit minimal correlation with most other clusters, indicating clear separation and improved distinctiveness relative to sector-based groupings. This is particularly apparent for mining-related clusters which were automatically segmented out by the unsupervised model.

Key Takeaway

These findings suggest that the hierarchical clustering methodology identifies more internally homogeneous and externally heterogeneous groupings than the static sector definitions. This supports the notion that return-based clustering methods may offer a more nuanced segmentation of stocks for applications like portfolio construction or thematic investing.

Within Cluster/Section Correlations (cohesion)

To evaluate how cohesive each grouping was in terms of price behaviour, we computed the average pairwise correlation of weekly returns within each cluster or sector. This helps assess how similarly the stocks in each group moved over time — the higher the average correlation, the more internally consistent or "cohesive" the group.

Sector-Based Groupings

Using traditional GICS-based sector labels, the average within-sector pairwise correlation across all sectors was **0.28**. The most cohesive sectors were:

- **Financials**: 0.386
- **Health Care**: 0.372
- **Basic Materials**: 0.321
- **Consumer Services**: 0.312
  Some sectors showed notably low internal coherence:
- **Consumer Goods**: 0.155
- **Technology**: 0.164
  This suggests that traditional sector labels do not always reflect actual behavioural similarity in stock returns, particularly for more heterogeneous or diversified sectors.

Sector Raw Weekly Returns Table, With Average: 0.28

Cluster	Within Cluster Average Pairwise Correlation
6(Technology)	0.164
2(Consumer Services)	0.312
7(Telecommunications)	0.273
4(Health Care)	0.372
3(Financials)	0.386
1(Consumer Goods)	0.155
5(Industrials)	0.29
0(Basic Materials)	0.321

Hierarchical Clustering-Based Groupings

In contrast, the data-driven hierarchical clustering approach yielded a significantly higher average within-cluster correlation of **0.57**, indicating much stronger group cohesion. Several clusters demonstrated extremely high internal correlations, such as:

- **Gold Cluster**: 0.771
- **Platinum Cluster**: 0.737
- **Retail Cluster**: 0.607
  This result highlights that clustering based on historical return patterns leads to groupings that better reflect actual co-movement, especially among commodity-linked and uniquely-behaving stocks.
  Some smaller clusters, such as the “Other” cluster (0.222), were less cohesive — likely due

to their role as residual groupings for stocks with less obvious similarity.

Cluster Raw Weekly Returns Table, With Average: 0.57

Cluster	Within Cluster Average Pairwise Correlation
6(Other)	0.222
4(REIT)	0.409
1(Naspers+Prosus)	0.853
5(Financials)	0.585
2(Rand Hedge)	0.328
3(Retail)	0.607
10(Rest of Mining)	0.584
7(Gold)	0.771
9(Platinum)	0.737
8(Sibyanye)	null

Key Takeaway

The hierarchical clustering method captured more meaningful intra-group relationships than traditional sector classification. This suggests that **return-based clustering is a superior method for forming groups of stocks with similar return dynamics**, which may be more useful in portfolio construction, risk budgeting, or thematic investing frameworks.

Excluding Market Effect

To better assess the true co-movement structure among stocks, it was important to isolate idiosyncratic return relationships from systematic market-wide influences. This was achieved by **removing the first principal component (PC1)** of the return matrix — a well-established proxy for the broad market factor — effectively filtering out the common variance shared across all assets.

By excluding the market effect, we obtain a clearer picture of how stocks or sectors co-move independently of general market trends. This allows us to more robustly evaluate the **cohesion within clusters** (whether the grouped assets still move together when the market is taken out) and the **separation between clusters** (whether the clusters remain distinct from one another in a purely idiosyncratic return space).

This adjustment is particularly important in the South African market context, where a small number of large-cap stocks and global macro linkages often dominate correlation structures. After removing PC1, we reassess both the within-cluster pairwise correlations (as a measure of internal cohesion) and the between-cluster correlations (as a test of discriminative power).

The following subsections present these results and offer a comparison between **sector-based groupings** and **K-means-derived clusters**, revealing how much of their structure is driven by

true economic similarity versus shared market exposure.

Within Cluster/Section Correlations (cohesion)

Sector-Based Cohesion:

The average within-sector pairwise correlation fell significantly to **0.09**, indicating limited co-movement once the market factor was removed. This suggests that sectors—defined by traditional GICS classifications—exhibit relatively low fundamental synchrony on a residual basis. Notably, **Health Care** (0.254) and **Consumer Goods** (0.195) retained some moderate cohesion, while others like **Basic Materials** (0.023) and **Industrials** (0.024) displayed almost no intra-sector alignment.

Sector Weekly Returns Excluding Market Effect Table, with Average: 0.09

Cluster	Within Cluster Average Pairwise Correlation
6(Technology)	0.058
2(Consumer Services)	0.061
7(Telecommunications)	0.103
4(Health Care)	0.254
3(Financials)	0.033
1(Consumer Goods)	0.195
5(Industrials)	0.024
0(Basic Materials)	0.023

Hierarchical Cluster Cohesion:

In contrast, the clustering methodology produced significantly stronger internal cohesion after the market effect was excluded. The average within-cluster correlation was **0.34**, nearly four times higher than that of sectors. Particularly cohesive clusters include **Naspers + Prosus** (0.945), **Platinum** (0.601), and **Gold** (0.535), all of which likely reflect structural or commodity-linked return patterns not captured by sector labels.

Cluster Weekly Returns Excluding Market Effect Table, with Average: 0.34

Cluster	Within Cluster Average Pairwise Correlation
6(Other)	0.071
4(REIT)	0.137
1(Naspers+Prosus)	0.945
5(Financials)	0.162
2(Rand Hedge)	0.06
3(Retail)	0.278
10(Rest of Mining)	0.269
7(Gold)	0.535
9(Platinum)	0.601
8(Sibyanye)	null

Takeaway

These findings demonstrate that **data-driven clustering outperforms traditional sector classifications** in grouping stocks with shared residual dynamics. The persistence of high intra-cluster correlations post-market adjustment affirms the validity of the clustering method in capturing latent commonalities beyond market exposure.

This enhanced cohesion suggests that clustering offers a **more statistically meaningful and economically intuitive segmentation** of the market, which may benefit both portfolio construction and risk monitoring processes.

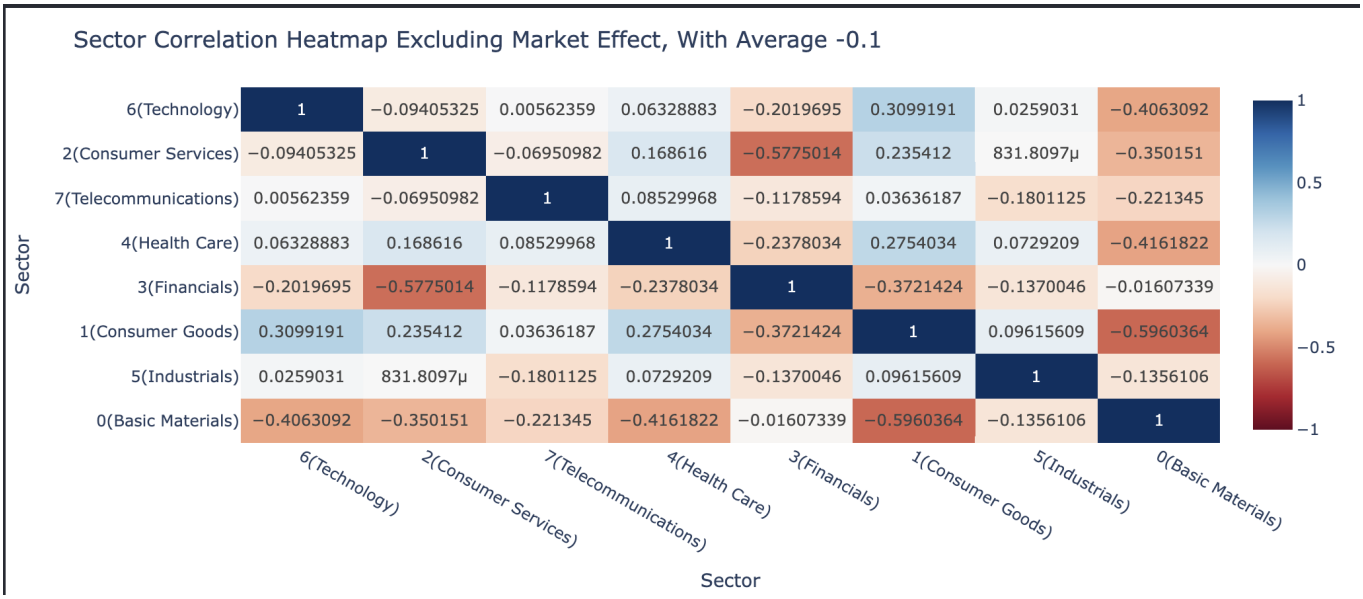
Between Cluster/Section Correlations (separation)

To assess the degree of **inter-cluster differentiation**, we examined the **pairwise correlations between clusters or sectors** after removing the common market component, typically attributed to broad macroeconomic or systemic movements. This adjustment isolates **idiosyncratic or sector-specific return co-movements**, allowing a clearer view of genuine diversification.

Sector-Based Grouping:

In the traditional sector classification (first heatmap), the **average inter-sector correlation falls to -0.10** once the market effect is excluded. This **negative average implies mild decoupling** between sectors under idiosyncratic conditions. Some sectors, such as **Financials and Consumer Services (-0.577)** or **Basic Materials and Consumer Goods (-0.596)**, display notable negative relationships, suggesting potential for diversification. However, many sectors remain **weakly related or ambiguous**, with values near zero, undermining the assumption that

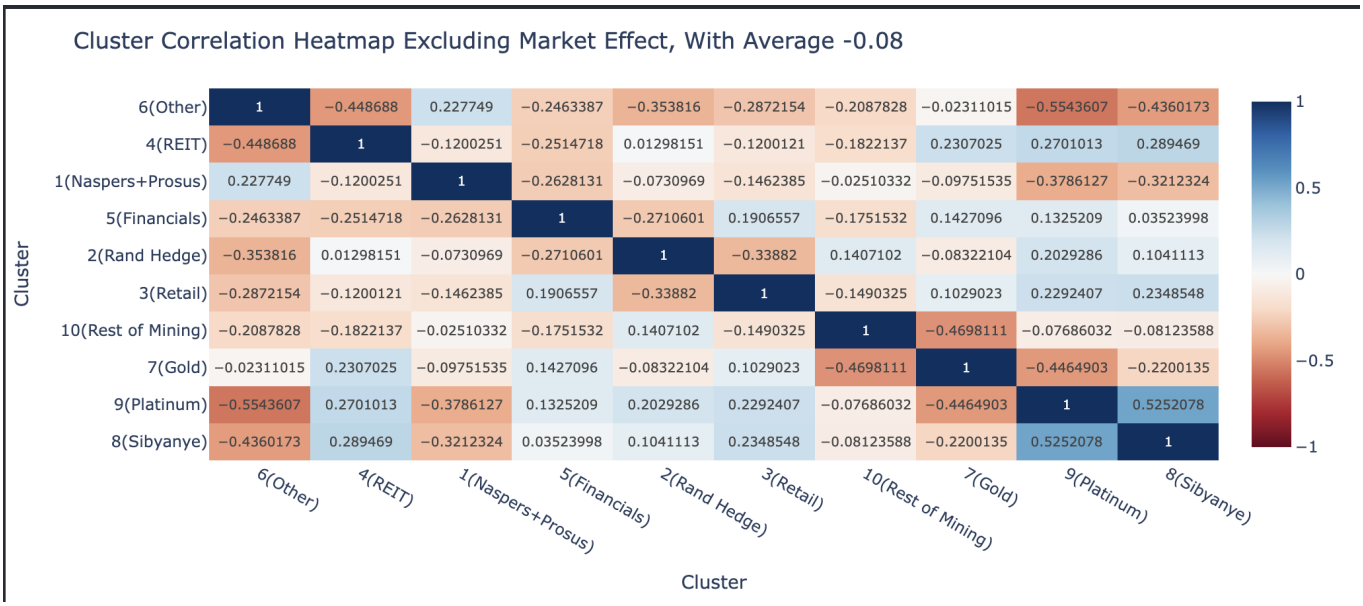
sector labels inherently imply meaningful diversification.



Hierarchical Cluster-Derived Clusters:

By contrast, the Hierarchical Cluster-based clusters (second heatmap) display an **average inter-cluster correlation of -0.08**, suggesting similarly low alignment between clusters. However, **some cluster pairs show notably low or even inverse co-movement**, such as **Cluster 6 (Other) vs Cluster 9 (Platinum) at -0.554** and **Cluster 10 (Rest of Mining) vs Cluster 7 (Gold) at -0.470**. These results reflect **better dispersion and potential for portfolio diversification** than sectors in certain cases.

Importantly, some clusters such as **Platinum and Sibanye (0.525)** or **REIT and Sibanye (0.289)** show moderate alignment, which is expected given operational or commodity linkages. Nevertheless, the **overall lower average correlation and the broader dispersion of values suggest that KMeans clusters separate return streams more effectively**, especially after controlling for market-wide movements.



Takeaway

This implies that **Hierarchical Cluster clustering may yield superior diversification insight compared to traditional sectors**, especially when isolating for firm-specific or sub-sector effects.

Conclusions

Raw Returns Analysis

Sector-Based Groupings

- **Within-Sector Cohesion**: Weak, with average correlations clustering around **0.09**
- **Between-Sector Separation**: Also weak, with average between-sector correlation close to **0.00**, indicating **minimal differentiation**
- Sectors failed to isolate unique return behavior; co-movement is largely driven by overall market moves.
  Implication:
  Sector groupings based on GICS labels are insufficient in capturing coherent return structures in raw returns. Many sectors displayed **internal heterogeneity**, suggesting they are too broad for return-based risk modeling or alpha grouping.

Hierarchical Clustering

- **Within-Cluster Cohesion**: Slightly stronger than sectors, with select clusters showing values above **0.15**, particularly those grouping mining-related stocks.

- **Between-Cluster Separation**: Still limited, with average correlations near zero, though slightly lower than sectors.
- The clustering begins to group together companies based on **latent return co-movements**, not explicitly captured by sector labels.
  Implication:
  Hierarchical clustering outperforms sector groupings marginally in the raw return space, indicating that even with market-driven noise, data-driven methods extract meaningful structures. However, market-wide effects still obscure deeper relationships.

Market-Adjusted Return Analysis (Excluding Market Effect)

Sector-Based Groupings

- **Within-Sector Cohesion**: Still weak. The average within-sector correlation was **0.09**, marginally improved, but most sectors remained noisy or internally diverse.
- **Between-Sector Separation**: Average correlation of **-0.10**, showing some dispersion between sectors, but not significantly better than chance.
  Implication:
  Sectors still fail to offer tightly knit groups post-adjustment. This suggests that many stocks within the same sector respond to **different idiosyncratic drivers**, and market effect alone does not explain the internal sector noise.

Hierarchical Clustering

- **Within-Cluster Cohesion**: Significantly improved. Average correlation rose to **0.34**, with clusters like:
  - **Platinum stocks** at **0.601**
  - **Gold stocks** at **0.535**
- **Between-Cluster Separation**: Strengthened, with average correlation dropping to **-0.08**. Clusters became more distinct and isolated in behavior.
  Implication:
  Hierarchical clustering performs **substantially better** than sector classification once the market effect is removed. The method can isolate **pure, economically intuitive themes** (e.g., commodity cycles, retail, rand hedge dynamics), which are lost in broad GICS sectors. This is particularly important for:
- **Long/short equity strategies**
- **Risk clustering in portfolio construction**
- **Building orthogonal investment themes**

Key Insights

Return Space	Method	Cohesion (↑ good)	Separation (↓ good)	Result
Raw Returns	Sector Grouping	Low (~0.09)	Poor (~0.00)	Weak structure overall
	Hierarchical Clustering	Moderate	Moderate	Slight improvement over sector grouping
Excl. Market Effect	Sector Grouping	Low (~0.09)	Slightly Better	Sectors still noisy
	Hierarchical Clustering	High (~0.34 avg)	Good (-0.08 avg)	Strongest result – tight, distinct clusters

Hierarchical clustering is a **superior tool** for uncovering latent structure in stock return behavior —particularly in the **idiosyncratic return space**. It delivers **greater within-group consistency (cohesion)** and **clearer between-group divergence (separation)** than sector-based classification.
This has **broad applications** in asset management, including:

- **Alpha signal generation** from statistically coherent groups
- **Risk modeling** using truly co-moving groups
- **Portfolio diversification** using economically distinct, uncorrelated clusters
- **Factor investing** beyond standard GICS classifications
  In the South African market, where sector classifications can span heterogeneous businesses (e.g., diversified miners, conglomerates), a **data-driven approach like hierarchical clustering is essential** to accurately reflect the underlying return dynamics of listed equities.

Recommendations

1. Adopt Hierarchical Clustering for Risk Grouping and Theme Formation

- **Replace or complement GICS sector classifications** with data-driven hierarchical clusters when constructing factor exposures, sector tilts, or thematic sleeves.
- Hierarchical clustering yielded significantly **higher within-cluster cohesion** and **lower between-cluster correlations** (particularly after controlling for market-wide effects), making

it a more robust framework for building diversified portfolios with **clear risk silos**.

> *Implication:* Risk clustering using hierarchical methods will likely lead to more reliable diversification and better alpha attribution in multi-asset or equity-only mandates.

---

2. Build Portfolios on Idiosyncratic Return Structure, Not Just Market Beta

- The clustering performance improved substantially **after adjusting for the market effect**, indicating that the market beta component masks important relationships between stocks.
- Decompose returns into **systematic vs. idiosyncratic components**, and use clustering on the residuals to design baskets of stocks that share similar alpha drivers.

> *Implication:* Enhancing portfolio construction using clusters based on **pure alpha signals** can help reduce noise, improve Sharpe ratios, and sharpen factor exposures.

---

3. Develop Targeted Long/Short or Market-Neutral Strategies

- Clusters identified from the residual space (e.g., Platinum, Gold, Retail) were not only cohesive but also economically intuitive, suggesting an opportunity to exploit **structural dislocations or thematic divergence**.
- Strategies that go **long high-cohesion clusters** and **short uncorrelated or anti-correlated groups** could enhance both return predictability and downside protection.

> *Implication:* Hierarchical clusters can serve as a foundational input for **market-neutral or factor-tilted products** that aim to isolate structural alpha sources.

---

4. Enhance Risk Monitoring and Stress Testing

- Traditional sector labels fail to reflect **true latent co-movements** between stocks, which can lead to underestimation of tail-risk exposure.
- Use hierarchical clustering outputs as a **risk lens** to understand exposure concentration, correlated drawdowns, or systemic contagion potential within the portfolio.

> *Implication:* Cluster-based overlays can improve the precision of **Value-at-Risk (VaR)**, **stress testing**, and **scenario analysis frameworks**.

---

5. Re-Evaluate Fund Benchmarking Practices

- Given the lack of coherence within many GICS sectors, fund performance or attribution tied strictly to sector benchmarks may be **misleading or overly simplistic**.
- Consider augmenting traditional benchmarking frameworks with **cluster-based peers** to more accurately measure manager skill and stock selection quality.

> *Implication:* This can help firms develop more **nuanced performance attribution models**, especially for high-conviction or unconstrained equity strategies.

---

6. Institutionalize the Use of Clustering in Research and Screening Tools

- Integrate hierarchical clustering into internal research pipelines to better group stocks when conducting:
  - Cross-sectional signal backtesting
  - Event studies
  - Regime classification
- Clusters can help refine universe screening, especially in **quantamental models** and **machine learning pipelines**.

> *Implication:* Teams can enhance idea generation, backtesting robustness, and improve signal-to-noise ratios by working within structurally coherent clusters.

Appendix

Cluster Constituents

Cluster 6(Other)

	Code	Company	Market Cap	Sector
13	VOD SJ Equity	Vodacom Group Limited (JSE:VOD)	288425000000	Telecommunications
14	MTN SJ Equity	MTN Group Limited (JSE:MTN)	256660000000	Telecommunications
17	SHP SJ Equity	Shoprite Holdings Limited (JSE:SHP)	167337000000	Consumer Services
21	BID SJ Equity	Bid Corporation Limited (JSE:BID)	155313000000	Consumer Services
35	CLS SJ Equity	Clicks Group Limited (JSE:CLS)	87717100000	Consumer Services
42	TBS SJ Equity	Tiger Brands Limited (JSE:TBS)	53817500000	Consumer Goods
43	APN SJ Equity	Aspen Pharmacare Holdings Limited (JSE:APN)	52907700000	Health Care
44	MCG SJ Equity	MultiChoice Group (JSE:MCG)	51924400000	Consumer Services
47	SNT SJ Equity	Santam Limited (JSE:SNT)	49436300000	Financials
57	AVI SJ Equity	AVI Limited (JSE:AVI)	33235700000	Consumer Goods
58	BOX SJ Equity	Boxer Retail Limited (JSE:BOX)	32599400000	Consumer Services
60	N91 SJ Equity	Ninety One plc (JSE:N91)	28178900000	Financials
61	DCP SJ Equity	Dis-Chem Pharmacies Limited (JSE:DCP)	27909700000	Consumer Services
62	KST SJ Equity	PSG Konsult Limited (JSE:KST)	27545000000	Financials
72	BAW SJ Equity	Barloworld Limited (JSE:BAW)	21922600000	Industrials
73	SPP SJ Equity	Spar Group Limited (JSE:SPP)	21546400000	Consumer Services
75	LHC SJ Equity	Life Healthcare Group (JSE:LHC)	21173800000	Health Care

	Code	Company	Market Cap	Sector
77	PIK SJ Equity	Pick n Pay Stores Limited (JSE:PIK)	20483200000	Consumer Services
78	NTC SJ Equity	Netcare Limited (JSE:NTC)	18996000000	Health Care
79	ADH SJ Equity	ADvTECH Limited (JSE:ADH)	18485600000	Consumer Services
87	CML SJ Equity	Coronation Fund Managers (JSE:CML)	15370200000	Financials
88	DTC SJ Equity	Datatec Limited (JSE:DTC)	15037900000	Technology
90	ITE SJ Equity	Italtile Limited (JSE:ITE)	13216500000	Consumer Services
91	OMN SJ Equity	Omnia Holdings Limited (JSE:OMN)	12858800000	Basic Materials
93	WBO SJ Equity	Wilson Bayly Holmes-Ovcon Limited (JSE:WBO)	12712300000	Industrials
96	SUI SJ Equity	Sun International Limited (JSE:SUI)	12165500000	Consumer Services
97	NY1 SJ Equity	Ninety One Limited (JSE:NY1)	11709000000	Financials
98	JSE SJ Equity	JSE Limited (JSE:JSE)	11354900000	Financials
99	AFH SJ Equity	Alexander Forbes Group Holdings Limited (JSE:AFH)	11045800000	Financials

	Code	Company	Market Cap	Sector
49	GRT SJ Equity	Growthpoint Properties Limited (JSE:GRT)	47276200000	Financials
56	RDF SJ Equity	Redefine Properties Limited (JSE:RDF)	33708200000	Financials
67	FFB SJ Equity	Fortress REIT Limited - B (JSE:FFB)	24675900000	Financials
68	VKE SJ Equity	Vukile Property Fund Limited (JSE:VKE)	24631200000	Financials

	Code	Company	Market Cap	Sector
69	SRI SJ Equity	Supermarket Income REIT (JSE:SRI)	23940300000	Financials
71	RES SJ Equity	Resilient REIT Limited (JSE:RES)	22982300000	Financials
76	BYI SJ Equity	Bytes Technology Group (JSE:BYI)	21031200000	Technology
83	HYP SJ Equity	Hyprop Investments Limited (JSE:HYP)	17486600000	Financials
85	LTE SJ Equity	Lighthouse Capital Limited (JSE:LTE)	17067200000	Financials
92	EQU SJ Equity	Equites Property Fund Limited (JSE:EQU)	12720300000	Financials

	Code	Company	Market Cap	Sector
0	PRX SJ Equity	Prosus N.V. (JSE:PRX)	2287790000000	Technology
6	NPN SJ Equity	Naspers Limited - N (JSE:NPN)	881232000000	Technology

	Code	Company	Market Cap	Sector
9	FSR SJ Equity	Firstrand Limited (JSE:FSR)	419646000000	Financials
10	CPI SJ Equity	Capitec Bank Holdings Limited (JSE:CPI)	415417000000	Financials
12	SBK SJ Equity	Standard Bank Group (JSE:SBK)	372867000000	Financials
16	SLM SJ Equity	Sanlam Limited (JSE:SLM)	188024000000	Financials
20	ABG SJ Equity	Absa Group Limited (JSE:ABG)	160630000000	Financials
23	DSY SJ Equity	Discovery Limited (JSE:DSY)	149836000000	Financials
25	OUT SJ Equity	OUTsurance Holdings Limited (JSE:OUT)	121798000000	Financials

	Code	Company	Market Cap	Sector
26	NED SJ Equity	Nedbank Group Limited (JSE:NED)	118317000000	Financials
34	REM SJ Equity	Remgro Limited (JSE:REM)	89072500000	Financials
36	BVT SJ Equity	Bidvest Group (JSE:BVT)	81240500000	Industrials
39	OMU SJ Equity	Old Mutual Limited (JSE:OMU)	57261700000	Financials
50	MTM SJ Equity	Momentum Metropolitan Holdings (JSE:MTM)	46336800000	Financials
66	WBC SJ Equity	We Buy Cars Holdings Limited (JSE:WBC)	24956100000	Technology

	Code	Company	Market Cap	Sector
2	ANH SJ Equity	Anheuser-Busch Inbev (JSE:ANH)	2201800000000	Consumer Goods
3	BTI SJ Equity	British American Tobacco (JSE:BTI)	1987650000000	Consumer Goods
4	CFR SJ Equity	Compagnie Fin Richemont (JSE:CFR)	1783080000000	Consumer Goods
24	MNP SJ Equity	Mondi plc (JSE:MNP)	127083000000	Basic Materials
33	INP SJ Equity	Investec plc (JSE:INP)	90532500000	Financials
40	SOL SJ Equity	Sasol Limited (JSE:SOL)	56935100000	Basic Materials
53	INL SJ Equity	Investec Limited (JSE:INL)	38567000000	Financials
54	HMN SJ Equity	Hammerson plc (JSE:HMN)	34955300000	Financials
82	SAP SJ Equity	Sappi Limited (JSE:SAP)	17625300000	Basic Materials

	Code	Company	Market Cap	Sector
28	PPH SJ Equity	Pepkor Holdings Limited (JSE:PPH)	103966000000	Consumer Services
38	MRP SJ Equity	Mr Price Group (JSE:MRP)	57588200000	Consumer Services
45	WHL SJ Equity	Woolworths Holdings Limited (JSE:WHL)	51768100000	Consumer Services
51	TFG SJ Equity	The Foschini Group (JSE:TFG)	45380500000	Consumer Services
59	TRU SJ Equity	Truworths International Limited (JSE:TRU)	30184000000	Consumer Services
81	MTH SJ Equity	Motus Holdings Limited (JSE:MTH)	17955300000	Consumer Services

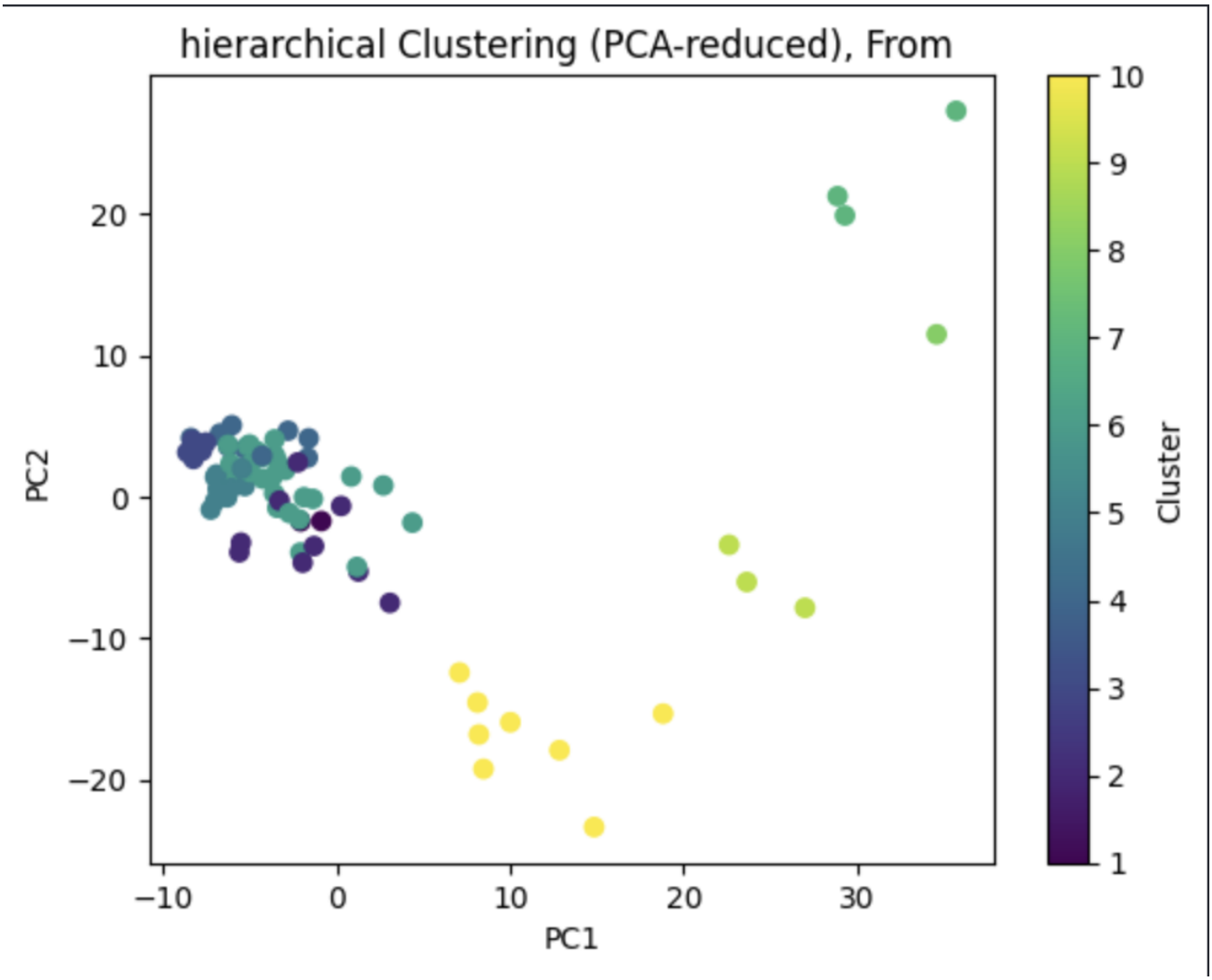
	Code	Company	Market Cap	Sector
1	BHG SJ Equity	BHP Group Limited (JSE:BHG)	2238510000000	Basic Materials
5	GLN SJ Equity	Glencore plc (JSE:GLN)	964367000000	Basic Materials
7	AGL SJ Equity	Anglo American plc (JSE:AGL)	626228000000	Basic Materials
18	S32 SJ Equity	South32 Limited (JSE:S32)	161631000000	Basic Materials
32	KIO SJ Equity	Kumba Iron Ore (JSE:KIO)	96261800000	Basic Materials
41	EXX SJ Equity	Exxaro Resources Limited (JSE:EXX)	54104100000	Basic Materials
52	ARI SJ Equity	African Rainbow Minerals (JSE:ARI)	39660000000	Basic Materials
95	TGA SJ Equity	Thungela Resources Limited (JSE:TGA)	12361900000	Basic Materials

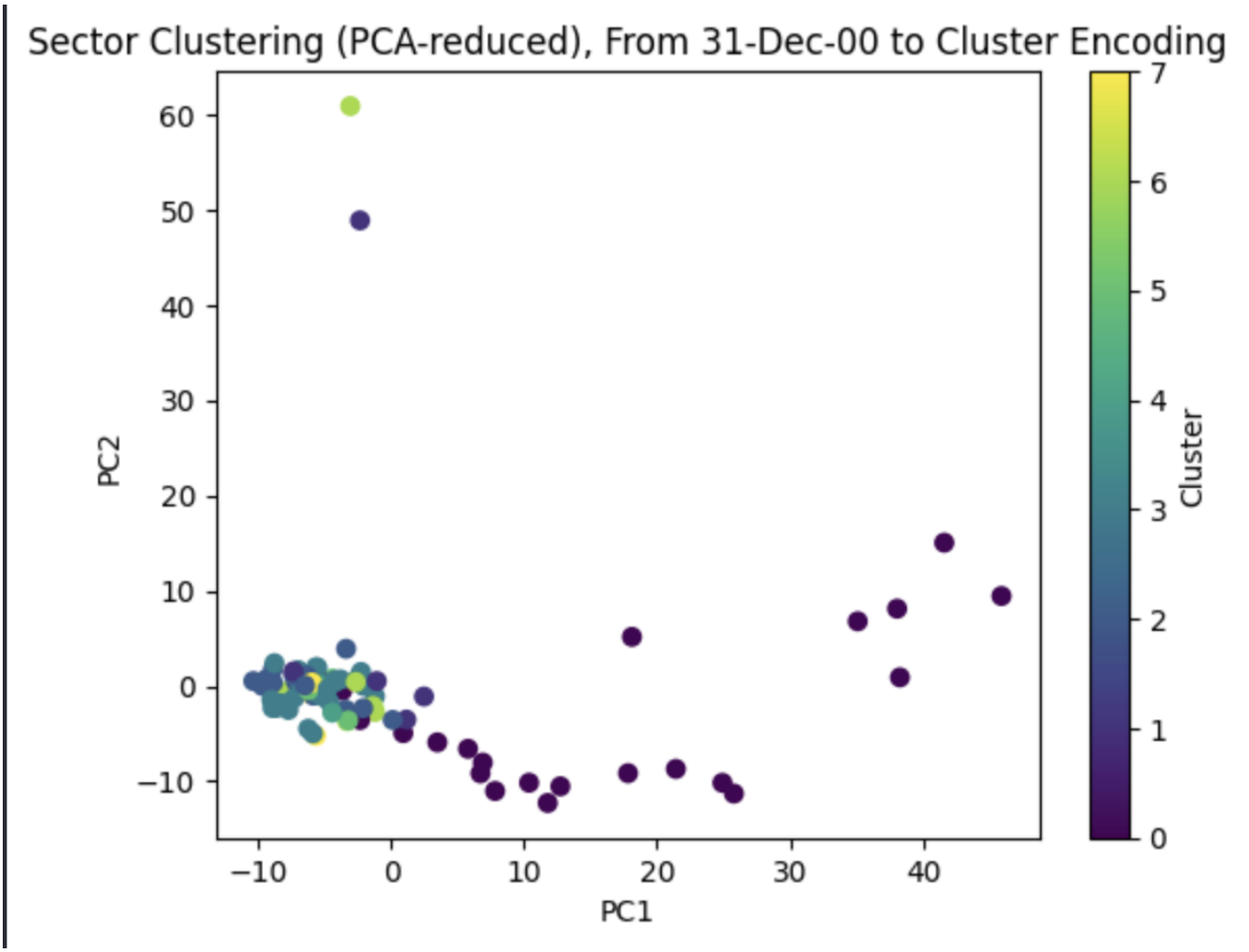
	Code	Company	Market Cap	Sector
8	ANG SJ Equity	Anglogold Ashanti (JSE:ANG)	419908000000	Basic Materials
11	GFI SJ Equity	Gold Fields Limited (JSE:GFI)	387134000000	Basic Materials
19	HAR SJ Equity	Harmony Gold Mining Company (JSE:HAR)	161155000000	Basic Materials

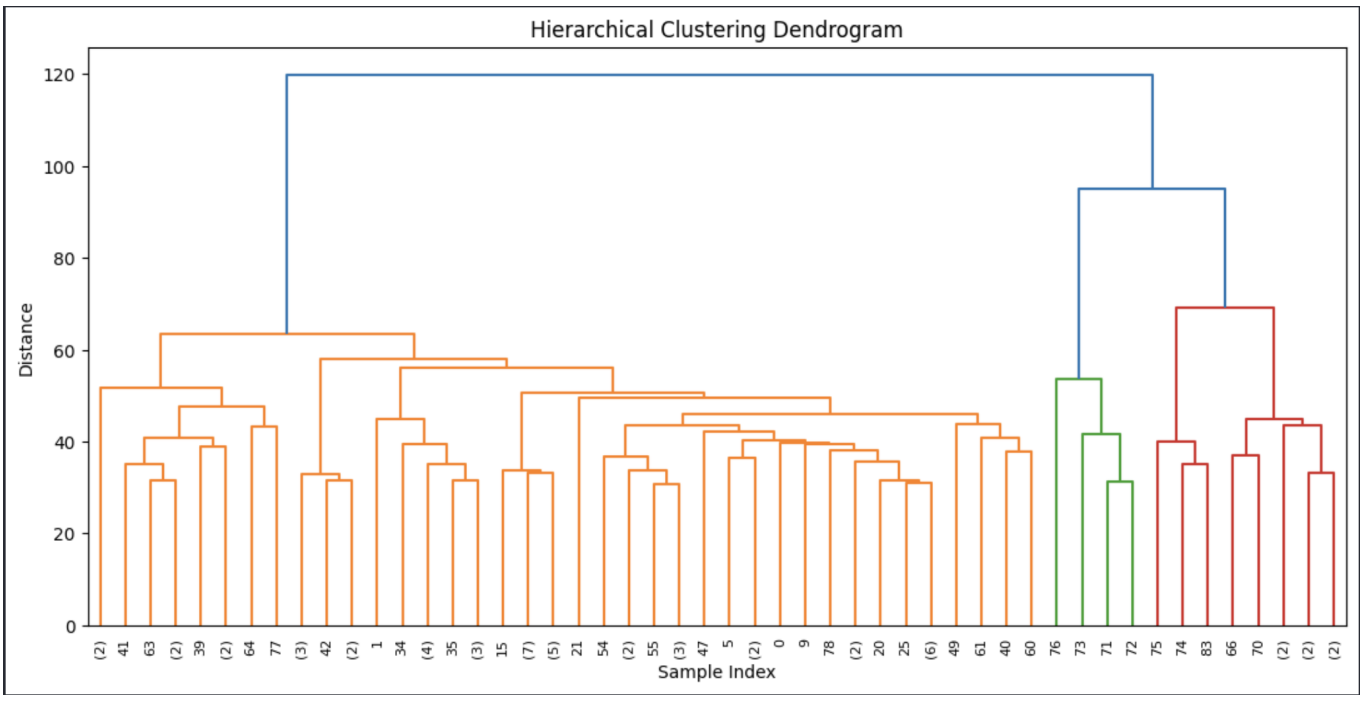
	Code	Company	Market Cap	Sector
15	VAL SJ Equity	Valterra Platinum Limited (JSE:VAL)	225498000000	Basic Materials
22	IMP SJ Equity	Impala Platinum Holdings Limited (JSE:IMP)	154737000000	Basic Materials
37	NPH SJ Equity	Northam Platinum Holdings Limited (JSE:NPH)	79780500000	Basic Materials

	Code	Company	Market Cap	Sector
30	SSW SJ Equity	Sibanye Stillwater Limited (JSE:SSW)	98900000000	Basic Materials

Figures


hierarchical Clustering (PCA-reduced), From
PC2
PC1
Cluster
20
10
0
−10
−20
−10
0
10
20
30
10
9
8
7
6
5
4
3
2
1


Sector Clustering (PCA-reduced), From 31-Dec-00 to Cluster Encoding
PC1
PC2
Cluster
60
50
40
30
20
10
0
−10
−10
0
10
20
30
40
7
6
5
4
3
2
1
0


Hierarchical Clustering Dendrogram
Distance
Sample Index
120
100
80
60
40
20
0
(2) 41 63 (2) 39 (2) 64 77 (3) 42 (2) 1 34 (4) 35 (3) 15 (7) (5) 21 54 (2) 55 (3) 47 5 (2) 0 9 78 (2) 20 25 (6) 49 61 40 60 76 73 71 72 75 74 83 66 70 (2) (2) (2)

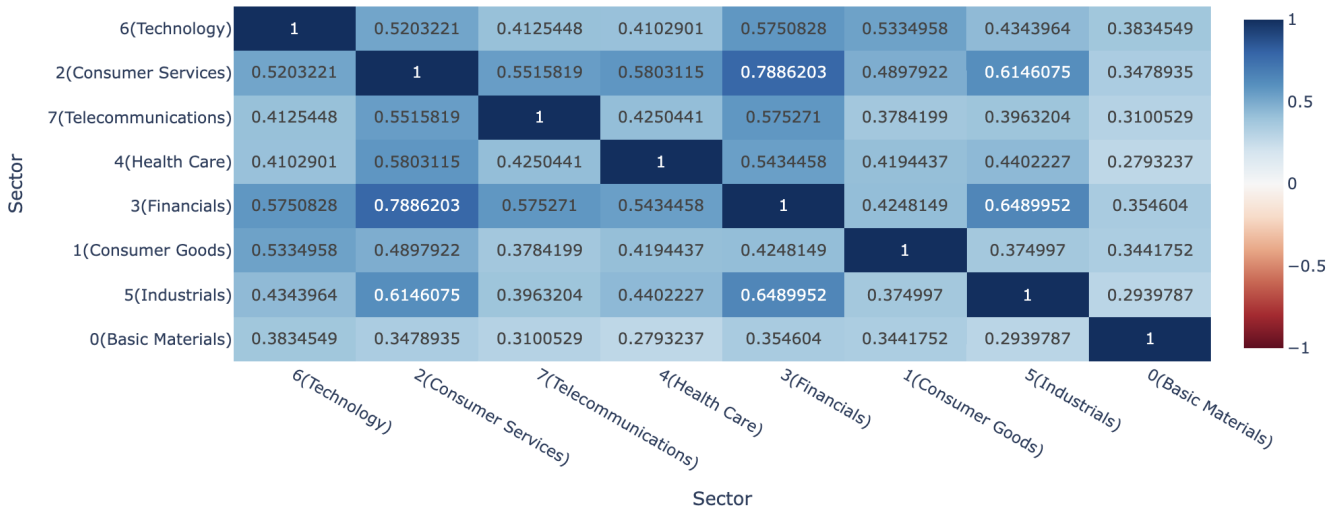
Sector Raw Weekly Returns Table, With Average: 0.28

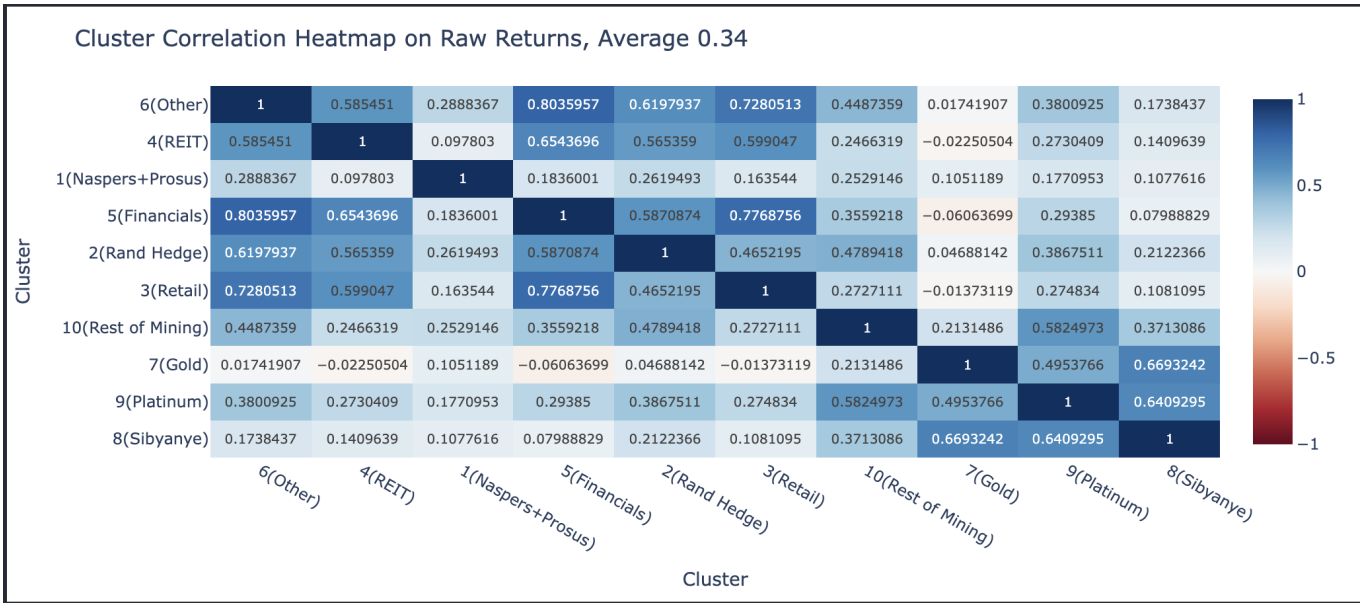
Cluster	Within Cluster Average Pairwise Correlation
6(Technology)	0.164
2(Consumer Services)	0.312
7(Telecommunications)	0.273
4(Health Care)	0.372
3(Financials)	0.386
1(Consumer Goods)	0.155
5(Industrials)	0.29
0(Basic Materials)	0.321

Cluster Raw Weekly Returns Table, With Average: 0.57

Cluster	Within Cluster Average Pairwise Correlation
6(Other)	0.222
4(REIT)	0.409
1(Naspers+Prosus)	0.853
5(Financials)	0.585
2(Rand Hedge)	0.328
3(Retail)	0.607
10(Rest of Mining)	0.584
7(Gold)	0.771
9(Platinum)	0.737
8(Sibyanye)	null

Sector Correlation Heatmap on Raw Returns, Average 0.46





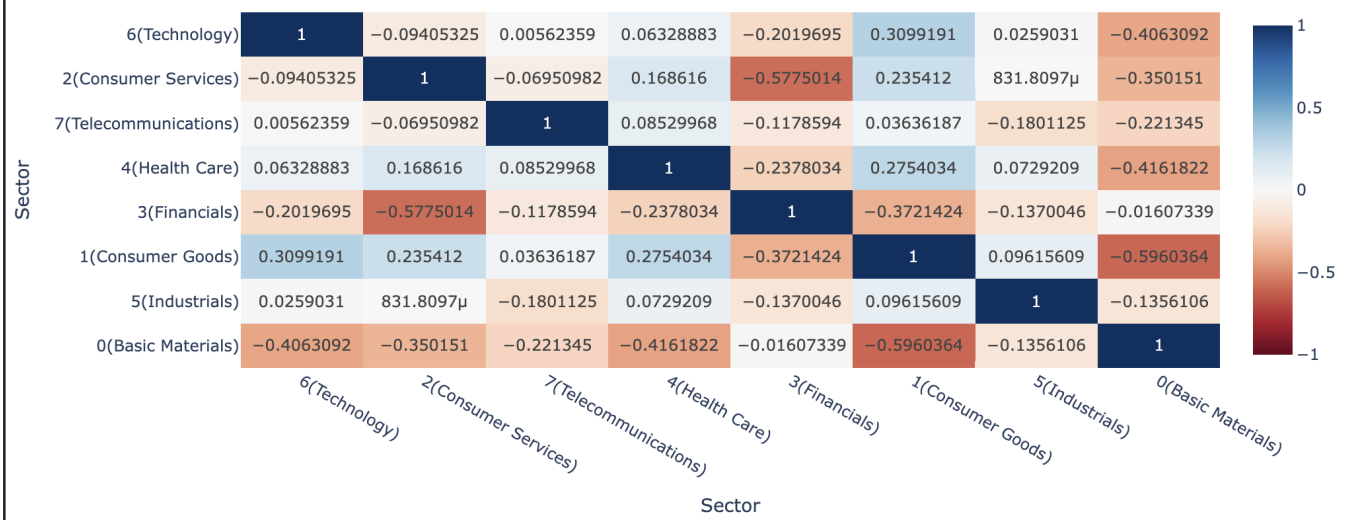
Sector Weekly Returns Excluding Market Effect Table, with Average: 0.09

Cluster	Within Cluster Average Pairwise Correlation
6(Technology)	0.058
2(Consumer Services)	0.061
7(Telecommunications)	0.103
4(Health Care)	0.254
3(Financials)	0.033
1(Consumer Goods)	0.195
5(Industrials)	0.024
0(Basic Materials)	0.023

Cluster Weekly Returns Excluding Market Effect Table, with Average: 0.34

Cluster	Within Cluster Average Pairwise Correlation
6(Other)	0.071
4(REIT)	0.137
1(Naspers+Prosus)	0.945
5(Financials)	0.162
2(Rand Hedge)	0.06
3(Retail)	0.278
10(Rest of Mining)	0.269
7(Gold)	0.535
9(Platinum)	0.601
8(Sibyanye)	null

Sector Correlation Heatmap Excluding Market Effect, With Average -0.1



Cluster Correlation Heatmap Excluding Market Effect, With Average -0.08