DVE_Assignment2_Report

Ru'aan Maharaj
School of Computer Science
University of the Witwatersrand
Johannesburg, South Africa
<2446659>

<James Stanton>
School of Computer Science
University of the Witwatersrand
Johannesburg, South Africa
<2541773>

<Jarren Downward>
School of Computer Science
University of the Witwatersrand
Johannesburg, South Africa
<2601486>

Preprint. Under review.

1 Question 1: Data Cleaning

An initial inspection of the NBA dataset revealed several data quality issues that required systematic cleaning and validation. The first inconsistency involved the 3P (three-pointers made) column, which was incorrectly typed as an object rather than a numeric variable due to a single malformed entry ("1.s6") for player Bradley Beal. This value was corrected using the basketball identity $3P = 3PA \times 3P\%$, after which the column was cast to a numeric type for consistency across the dataset.

Several derived efficiency metrics, including $FG\%$, $3P\%$, $2P\%$, $eFG\%$, $TS\%$, $3PAr$, and FTr, all contained missing or inconsistent values. Instead of applying naive zero-imputation, each metric was reconstructed from its fundamental components according to standard basketball formulas:

$$\begin{aligned}
FG\% &= \frac{FG}{FGA}, \\
3P\% &= \frac{3P}{3PA}, \\
2P\% &= \frac{2P}{2PA}, \\
eFG\% &= \frac{FG + 0.5 \times 3P}{FGA}, \\
TS\% &= \frac{PTS}{2 \times (FGA + 0.44 \times FTA)}, \\
3PAr &= \frac{3PA}{FGA}, \\
FTr &= \frac{FTA}{FGA}.
\end{aligned}$$

When both the numerator and denominator were zero, which indicates a player made no attempts, the result was set to 0.0 to denote no recorded activity. Cases where made shots were reported but attempts were zero were left as missing values to highlight logical inconsistencies.

To maintain a consistent numeric scale, all rate-based fields reported in percentages (e.g., $ORB\%$, $DRB\%$, $TRB\%$, $AST\%$, $STL\%$, $BLK\%$, $TOV\%$, $USG\%$) were normalized by dividing by 100, converting them to a [0, 1] range. Metrics such as FTr, which can exceed 1.0 when free-throw attempts surpass field-goal attempts, were preserved in their natural ratio form. After normalization, all bounded columns were validated to ensure they fell within expected limits. A few $TS\%$ values slightly above 1.0 were identified as small-sample anomalies and retained to preserve data fidelity.

Finally, arithmetic consistency checks revealed 147 violations of the identity $FG = 2P + 3P$ and 157 violations of $FGA = 2PA + 3PA$. These discrepancies were traced to rounding errors in per-game statistics. To resolve them, season totals were reconstructed using:

$$\text{Total} = \text{round}(\text{PerGame} \times GP),$$

enforcing the identities on these totals, and reconverting them to per-game form. This adjustment eliminated all arithmetic violations, confirming that inconsistencies were artifacts of rounding rather than genuine data errors.

In summary, the resulting dataset was fully numeric, internally consistent, and normalized across all features. The final version satisfied arithmetic constraints, preserved statistical meaning, and provided a robust foundation for downstream analysis, including dimensionality reduction and clustering.

2 Question 2

Following the data cleaning phase, the processed NBA dataset was used to explore various dimensionality reduction and clustering techniques aimed at uncovering latent player groupings and performance structures. High-dimensional sports statistics often contain correlated features, making it difficult to visualize or interpret underlying patterns directly. To address this, multiple approaches were applied, each representing a distinct philosophy of compression and feature extraction.

Section 2.1 implements a sequence of progressively more sophisticated dimensionality reduction methods which include Autoencoder (AE), Autoencoder + Self-Organizing Map (SOM), Autoencoder + t-SNE, Autoencoder + UMAP, and finally a Variational Autoencoder (VAE). Each technique projects the standardized data into a lower-dimensional latent space while preserving as much structural information as possible. The reduced embeddings are then used for subsequent k-Means clustering (Section 2.2), allowing comparison of how each method captures player similarity and group differentiation.

2.1 2.1(a) Autoencoder

A standard autoencoder (AE) was implemented to reduce the 49-dimensional NBA dataset directly to a 2-dimensional latent representation for visualization and clustering analysis.

The architecture consisted of an encoder with progressively narrowing layers using ReLU activations, mirrored by a symmetric decoder with a linear output layer for reconstruction. L2 regularization was applied to mitigate overfitting, and the model was trained using the Adam optimizer with mean squared error (MSE) loss for 50 epochs and a batch size of 32.

The training process (Figure 1) shows a smooth convergence, with both training and validation losses decreasing rapidly within the first 10 epochs before stabilizing around 0.33 MSE. The close alignment between the two curves suggests effective generalization and minimal overfitting despite the significant dimensional compression.

The 2D latent space visualization (Figure 2) displays a dense, roughly elliptical cluster of points centered near the origin, with most players distributed between –10 and 10 on both encoded dimensions. A small number of outliers are visible at the edges, including one extreme case at approximately (–35, –27), indicating players with statistically unique performance profiles.

Reconstruction error analysis (Figure 3) reveals a right-skewed distribution, where most players exhibit low reconstruction errors (mean = 0.326, SD = 0.317). This indicates that the autoencoder effectively captured the dominant structure of the dataset, though extreme players incurred higher reconstruction loss due to their deviation from the population norm.

Overall, the results show that the autoencoder successfully compressed the high-dimensional player statistics into a coherent two-dimensional space. However, the aggressive reduction introduces some information loss, particularly for atypical players. The latent structure suggests that NBA performance metrics exist along continuous gradients rather than forming sharply distinct player categories.

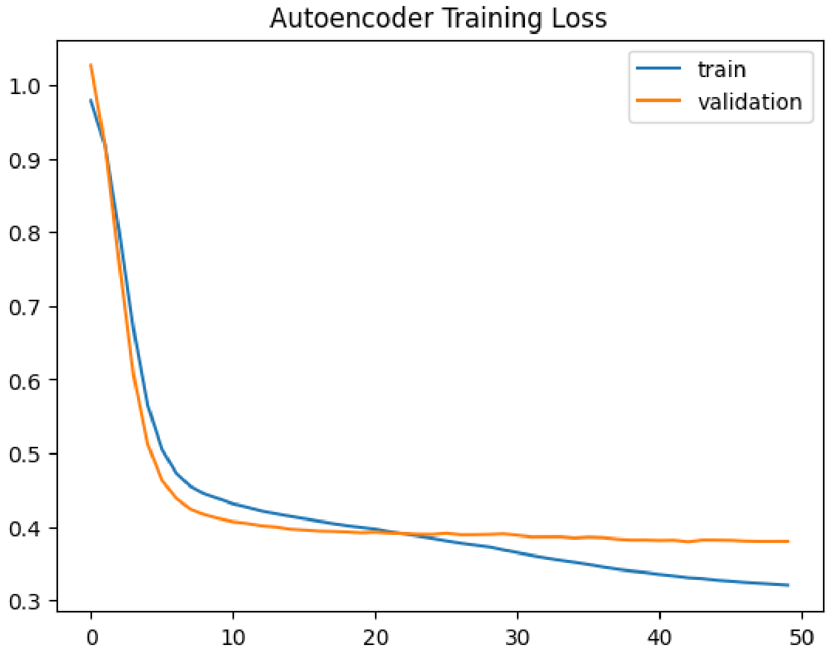

Figure 1: Autoencoder training and validation loss curves.

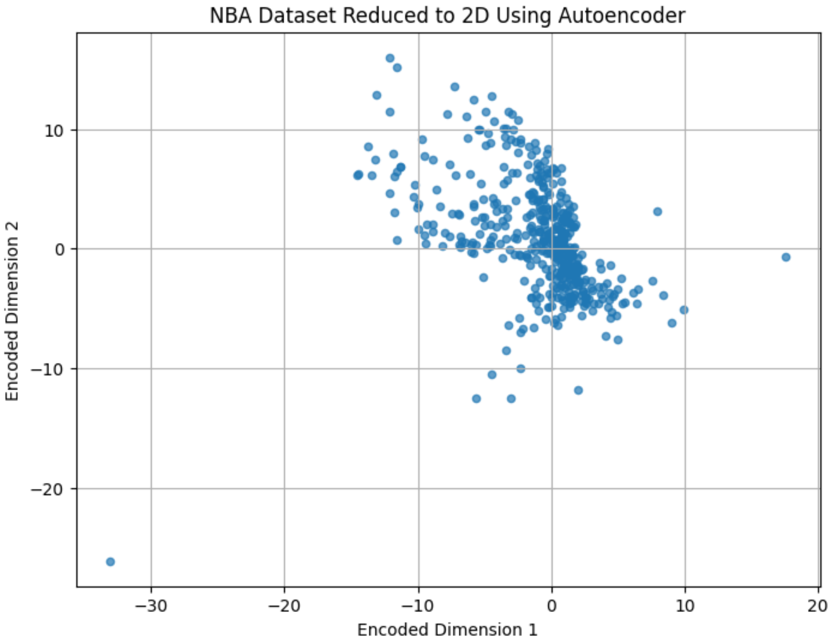

Figure 2: 2D latent space projection of the NBA dataset.

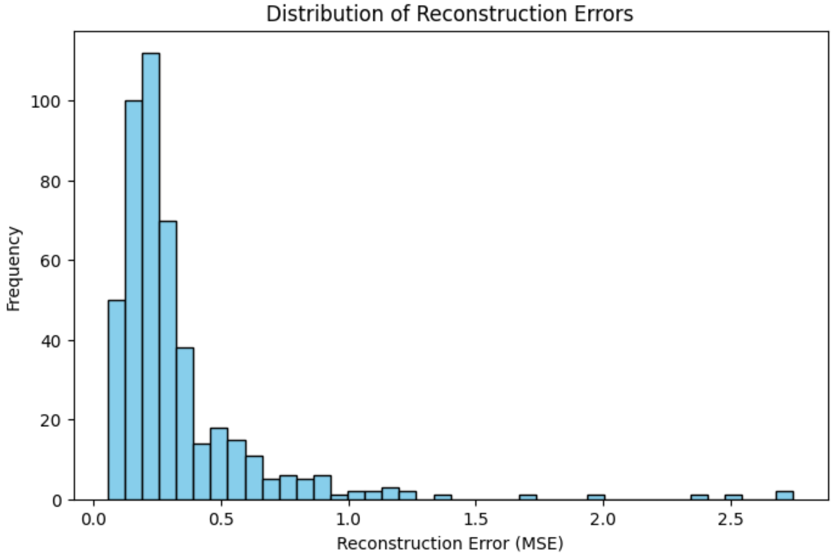

Figure 3: Distribution of autoencoder reconstruction errors across players.

2.2 2.1(b) Autoencoder + Self-Organizing Map (SOM)

To capture topological structure in the autoencoder's latent space, a Self-Organizing Map (SOM) was trained on the 2-dimensional encoded features. The SOM clusters players according to nonlinear similarity, preserving neighborhood relationships between latent representations and exposing distinct regions within the manifold.

The U-Matrix visualization (Figure 4) shows average neighbor distances across the SOM grid. Darker regions represent areas of high player density, while lighter boundaries indicate separations between statistically distinct player groups. The cluster projection (Figure 5) overlays these groups onto the autoencoder's latent space, illustrating how the SOM discretized the continuous embedding into interpretable categories.

2.2.1 Cluster Statistics and Reconstruction Performance

The SOM identified six player clusters of varying size and reconstruction accuracy:

Table 1: SOM cluster statistics and reconstruction performance.

Cluster	Count	Mean MSE	Max MSE	Interpretation
0	162	0.25	1.98	Large, moderate-error group-typical balanced players
1	54	0.31	1.71	Slightly higher error-low-impact, inconsistent players
2	54	0.33	0.85	Stable, high-usage scorers
3	161	0.37	2.75	Efficient specialists with broader variability
4	10	0.74	2.38	Small outlier cluster-statistical anomalies
5	26	0.35	0.91	Balanced contributors near the dataset mean

Clusters 0, 2, and 5 dominate the latent space, representing the majority of the league, while Cluster 4 isolates only 10 players with unusually high reconstruction errors, and shows a clear evidence of statistical outliers that the autoencoder could not model well.

2.2.2 Feature-Level Differentiation (Z-Scores)

Z-score analysis of numeric features clarifies how each cluster diverges from the overall population mean:

- **Cluster 0 – Balanced starters:** Slightly positive scoring ($FG \approx +0.10$, $FGA \approx +0.15$) and moderate defensive contributions ($DWS \approx +0.26$). These players fit the archetype of reliable rotation regulars with well-rounded metrics.
- **Cluster 1 – Low-efficiency performers:** Negative scoring ($FG \approx -0.23$, $FGA \approx -0.43$) but high $FG\%$ ($+1.00$) caused by small sample size; under performing in minutes ($MP \approx -0.29$) and win shares ($WS \approx +0.13$). Represents low-usage bench players or limited-minutes specialists.
- **Cluster 2 – High-volume offensive stars:** Strong positives across $FG(+1.87)$, $FGA(+1.99)$, $3P(+1.57)$, and usage rate ($+1.50$). Above-average offensive metrics ($OBPM \approx +1.09$, $WS \approx +1.11$). These correspond to primary scorers and focal offensive options.
- **Cluster 3 – Defensive contributors with lower usage:** Negative offensive metrics ($FG \approx -0.81$, $FGA \approx -0.75$) but below-average minutes ($MP \approx -0.89$) and efficiency ($FG\% \approx -0.52$). They maintain modest positive defense ($DWS \approx -0.84 \rightarrow$ relative depth role) but generally represent lower-impact contributors.
- **Cluster 4 – Anomalous outliers:** High z-scores in $OBPM(+2.15)$, $DBPM(+2.23)$, and $BPM(+2.58)$, despite very low playing time ($MP \approx -1.52$). Likely statistical noise from small-sample or exceptional situational players.
- **Cluster 5 – Efficient high-minute performers:** Strong positives in $FG(+1.36)$, $FGA(+0.89)$, $FG\%(+1.23)$, and $MP(+1.10)$. Solid overall production with $WS(+1.92)$ and moderate usage ($+0.52$). This group defines the dataset's "core" player profile—consistent, high-impact contributors.

2.2.3 Interpretation

The SOM successfully mapped the autoencoder's continuous latent space into six discrete, interpretable groups, distinguishing outliers (Cluster 4) and elite scorers (Cluster 2) from the broader population of balanced players (Clusters 0 and 5).

Overall, the AE + SOM approach demonstrates that non-linear manifold learning preserves player similarity relationships while the SOM imposes a meaningful topological segmentation, revealing the statistical archetypes embedded within the NBA dataset.

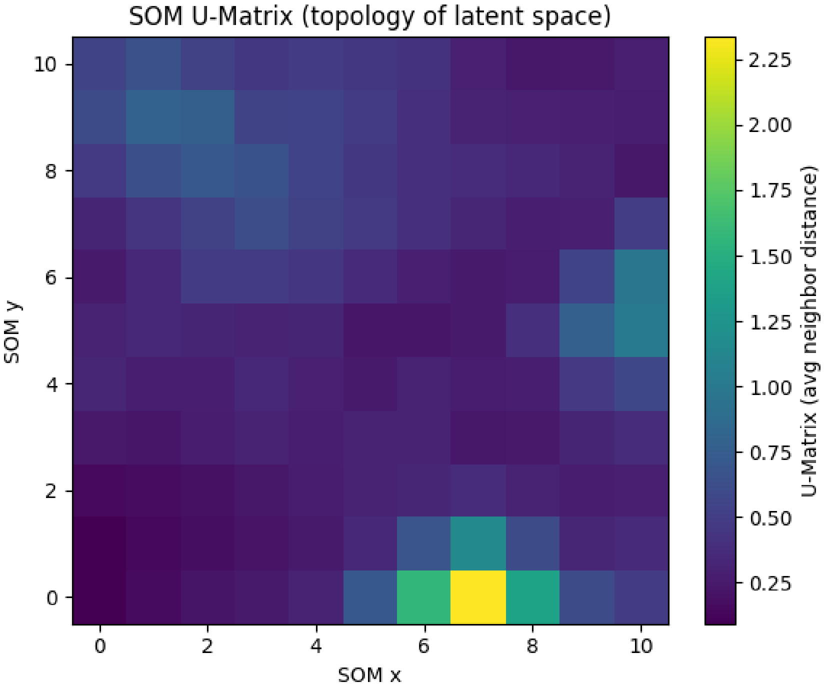

Figure 4: SOM U-Matrix (topology of latent space).

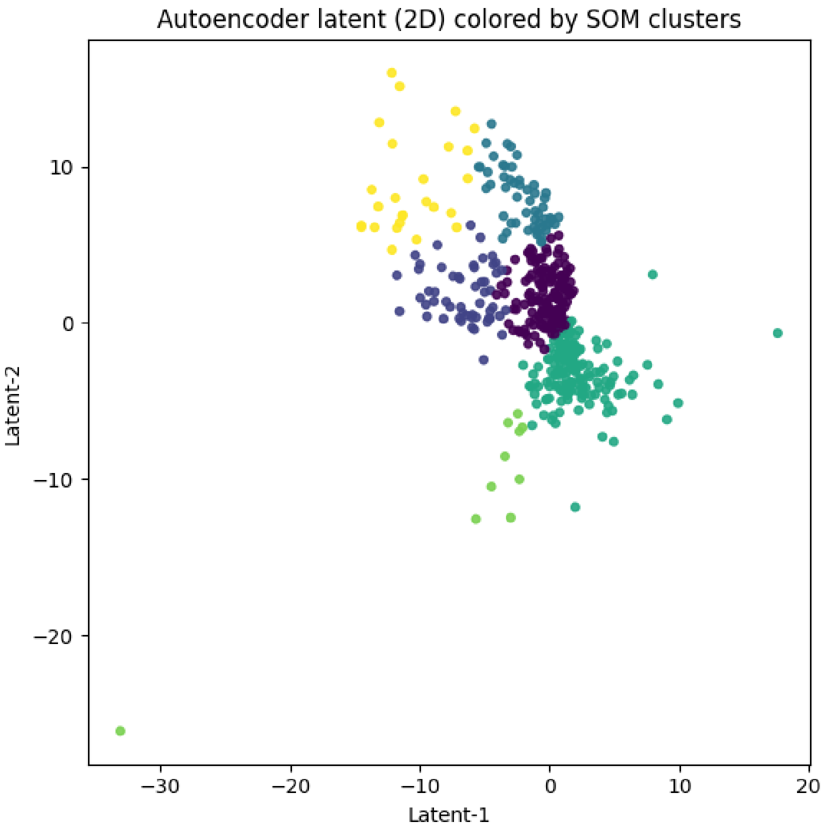

Figure 5: Autoencoder latent (2D) colored by SOM clusters.

2.3 2.1(c) Autoencoder + t-SNE

To visualize the nonlinear structure of player statistics in a lower-dimensional space, the 49-dimensional NBA dataset was first compressed into a 16-dimensional latent representation using the trained autoencoder. The encoded features were then projected to 2D using t-Distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear manifold learning algorithm optimized for local neighborhood preservation.

The autoencoder was trained for 100 epochs with the Adam optimizer and mean-squared-error (MSE) loss (Figure 7). Training converged smoothly, with validation loss stabilizing around ≈ 0.15 MSE and no signs of overfitting. The average reconstruction error across the dataset was 0.068 ± 0.069, indicating that most player profiles were accurately reconstructed from their compressed latent codes. The reconstruction-error histogram (Figure 7, right) confirms a highly right-skewed distribution, with the vast majority of players exhibiting errors below 0.2 MSE and only a few isolated outliers with higher deviation.

The resulting t-SNE embedding (Figure 6) produced a well-distributed two-dimensional manifold with visible density gradients rather than discrete clusters. This reflects the algorithm's emphasis on preserving local rather than global structure: similar players are embedded near one another, while distant relationships are not necessarily preserved linearly. The scattered, organically shaped cloud suggests that NBA player statistics exist on continuous performance spectra, for example, smooth transitions between low-usage defenders, balanced all-rounders, and high-volume scorers, rather than in sharply separated categories.

The t-SNE map therefore complements the SOM analysis by providing a fine-grained local view of the autoencoder's latent space. Whereas the SOM imposed discrete topological segmentation, t-SNE emphasizes micro-structure, revealing nuanced statistical neighborhoods that correspond to subtle stylistic differences among players.

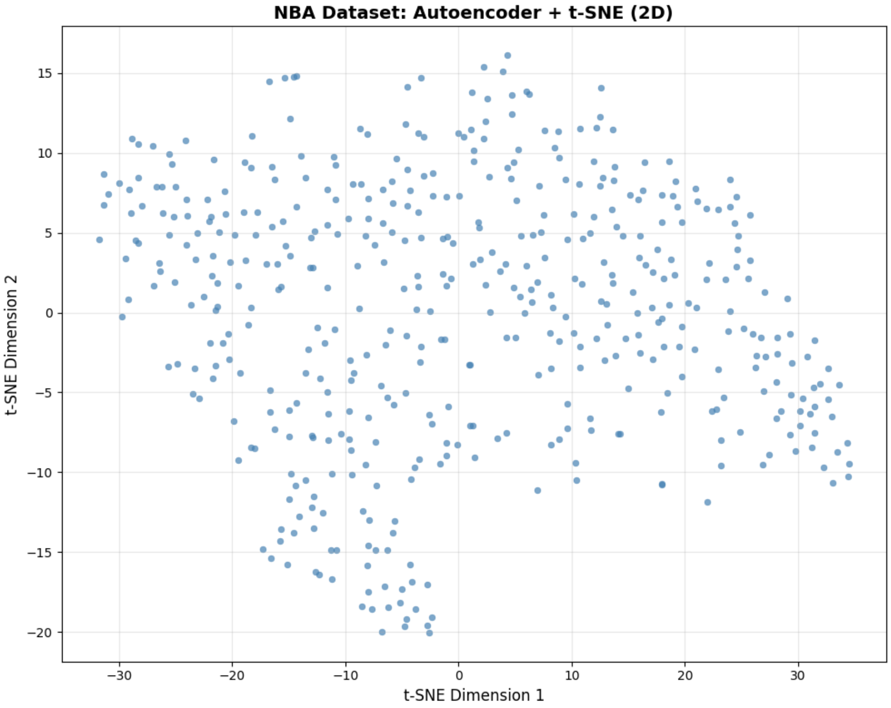

Figure 6: NBA dataset: Autoencoder (16D) + t-SNE (2D) embedding.

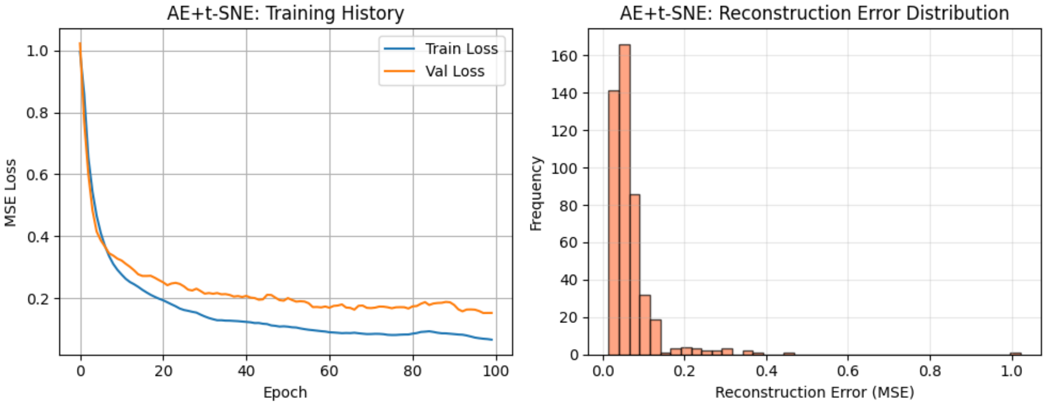

Figure 7: AE + t-SNE training history (left) and reconstruction-error histogram (right).

2.4 2.1(d) Autoencoder + UMAP

To complement the t-SNE analysis, Uniform Manifold Approximation and Projection (UMAP) was applied to the autoencoder's 16-dimensional latent representations. UMAP is a graph-based dimensionality reduction technique that preserves both local and global structure, often producing more interpretable and stable embeddings than t-SNE. It models the latent space as a weighted graph of neighborhood relationships and optimizes a low-dimensional layout that retains these topological connections.

The algorithm was executed with 500 training epochs and a fixed random seed for reproducibility, resulting in a 2D embedding of 467 player representations (Figure 8). Unlike t-SNE's scattered, cloud-like map, the UMAP projection revealed a distinct elongated manifold with multiple curved bands, suggesting smooth statistical transitions across player archetypes. This pattern implies that the underlying player space is topologically continuous, which means that players form performance gradients rather than isolated clusters.

UMAP's layout captures both the broad global structure (e.g., offense-oriented vs. defense-oriented spectra) and finer local neighborhoods (e.g., players with similar shooting efficiency or usage rates). Dense regions correspond to typical player archetypes, while sparse stretches at the periphery likely represent rare or extreme profiles (such as high-volume scorers or low-minute specialists).

Overall, the AE + UMAP combination produced a balanced representation: it maintained the local consistency of t-SNE while recovering a more interpretable large-scale geometry of player similarity. The resulting embedding highlights that NBA performance data occupy a continuous, curved manifold where small variations in playstyle and efficiency gradually shift one archetype into another.

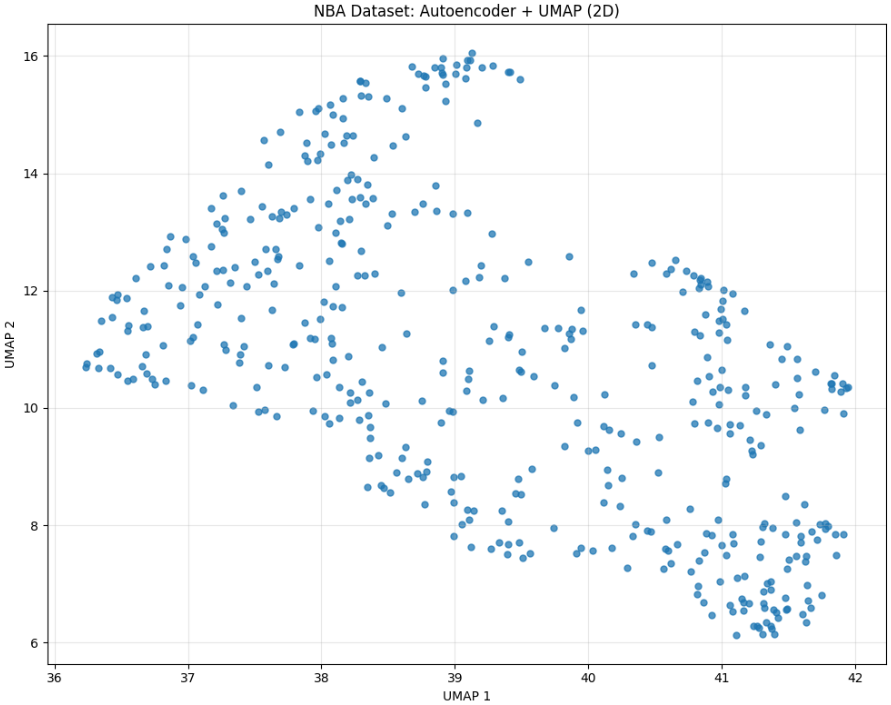

Figure 8: NBA dataset: Autoencoder (16D) + UMAP (2D) embedding.

2.5 2.1(e) Variational Autoencoder (VAE)

The Variational Autoencoder (VAE) extends the standard autoencoder architecture by introducing a probabilistic latent space, enabling both dimensionality reduction and generative modeling. Instead of mapping each input to a fixed latent vector, the encoder estimates two parameters for each latent dimension—the mean (μ) and log-variance ($\log \sigma^2$)—which define a Gaussian distribution from which latent samples are drawn via the reparameterization trick:

$$z = \mu + \sigma \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1)$$

This stochastic formulation regularizes the latent space to follow a standard normal prior, ensuring continuity and smoothness that facilitate meaningful interpolation and sampling.

The encoder consisted of fully connected layers with progressively decreasing sizes (128 $\rightarrow$ 64 $\rightarrow$ 32) followed by separate 2D latent outputs for z_{mean} and z_{logvar}. The decoder mirrored this structure in reverse, reconstructing the original 49 features through ReLU activations and a linear output layer. The VAE was trained for 150 epochs using the Adam optimizer with a learning rate of 0.001, minimizing a total loss composed of:

1. **Reconstruction Loss (MSE):** measures fidelity between input and reconstruction.
2. **KL Divergence:** regularizes the learned latent distributions toward the unit Gaussian prior.

As shown in Figure 9, the total and reconstruction losses decreased sharply during the first 20 epochs before plateauing around epoch 100, indicating convergence. The KL divergence term stabilized around 3.6, confirming a well-regularized latent space. The final model achieved a mean reconstruction error of 0.372 ± 1.08, suggesting reasonable fidelity despite the enforced probabilistic constraints.

The resulting 2D latent representation (Figure 10) formed a dense, roughly Gaussian cloud centered near the origin ($\mu \approx [0.10, 0.05]$, $\sigma \approx [1.10, 1.01]$), consistent with the theoretical $\mathcal{N}(0, I)$ prior. Unlike the deterministic autoencoder, the VAE's points occupy a smooth and continuous latent surface with no gaps or sharp boundaries, and each player representation blends seamlessly into others. This reflects the model's ability to capture the underlying manifold of player statistics rather than discrete partitions, enabling the generation of synthetic yet plausible player profiles through sampling from the latent distribution.

Overall, the VAE effectively regularized the latent space while retaining the major variance in player performance metrics. Though its reconstruction accuracy was slightly lower than that of the standard autoencoder, the resulting probabilistic embedding offers a richer, interpretable foundation for subsequent clustering analysis.

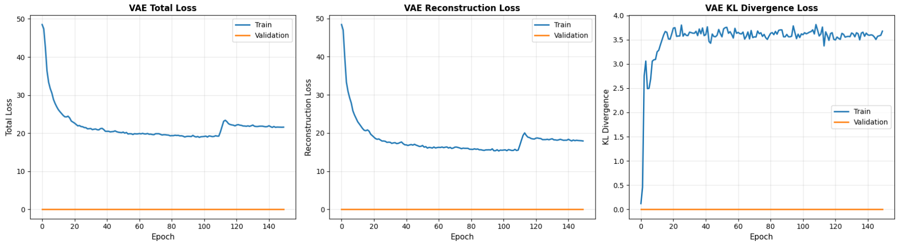

Figure 9: VAE training history: total, reconstruction, and KL divergence losses.

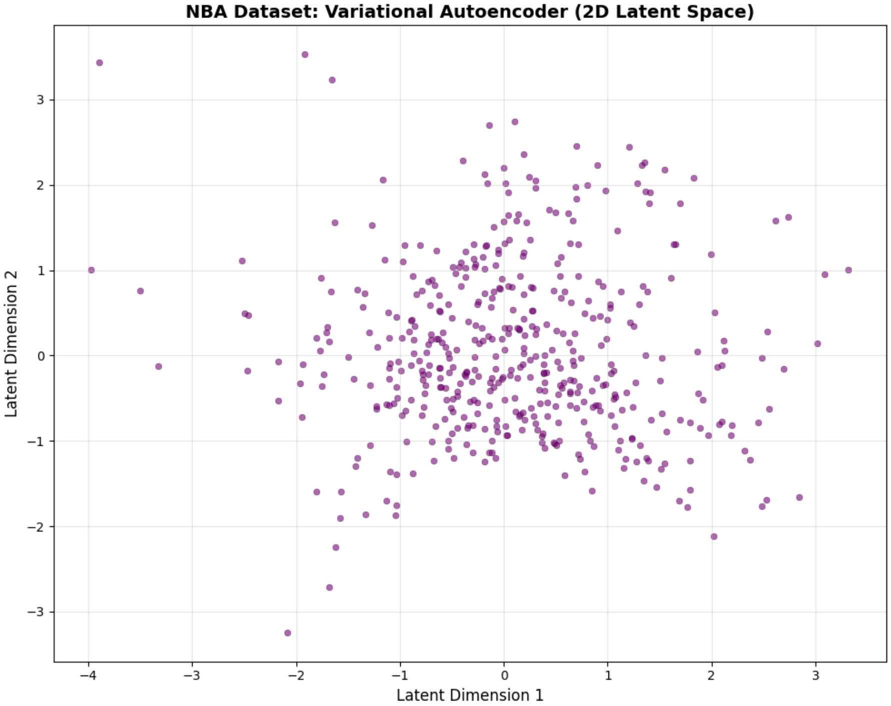

Figure 10: 2D latent space representation learned by the Variational Autoencoder.

2.6 2.2(a) AE + K-Means

Elbow and Silhouette Analysis

Figure 11 presents the elbow and silhouette plots for K-Means clustering applied to the autoencoder's 2D latent space. The elbow plot flattens notably after $k = 2$ or $k = 3$, indicating diminishing returns which is a typical pattern for moderately separated data. The silhouette score peaks at 0.456 for $k = 2$, suggesting that two clusters provide the most coherent partitioning of the latent representation.

Cluster Interpretation

As visualized in Figure 12, two distinct clusters emerge from the AE-encoded player space:

- **Cluster 0 - Big Scorers:** Includes elite players such as Embiid, Lillard, Giannis, Tatum, and SGA. These athletes exhibit strong two-point metrics ($2P$, $2PA$, TRB, DRB, FG), reflecting high-impact, efficient inside scorers with extended playing minutes and overall dominance.
- **Cluster 1 - Mid-Tier or Role Players:** Characterized by elevated three-point activity ($3PAr$, $3P\%$) and comparatively lower field-goal and rebounding stats. These are support shooters and rotational contributors rather than core offensive anchors.

The AE's latent representation effectively distinguishes high-impact stars from average or low-impact players, confirming that the model captured the non-linear structure of NBA performance metrics.

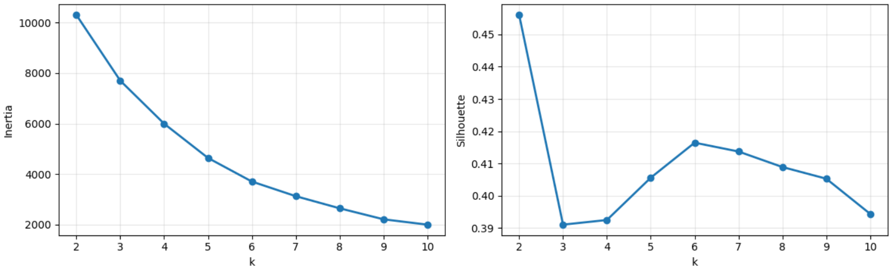

Figure 11: Elbow and silhouette analysis for AE + K-Means clustering.

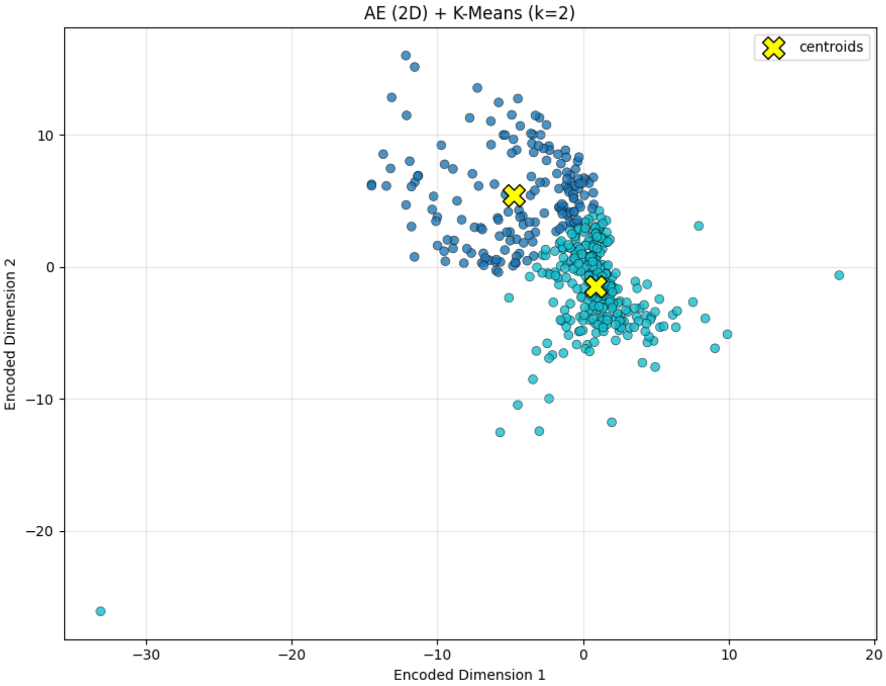

Figure 12: 2D cluster visualization of AE latent space using K-Means.

2.6.1 2.2(b) Autoencoder + Self-Organizing Map (SOM) + K-Means

After obtaining the 2-dimensional latent representation from the autoencoder, a Self-Organizing Map (SOM) was trained to project the latent vectors onto a topology-preserving grid. The SOM compresses non-linear relationships and maintains local neighborhood structure, allowing similar player profiles to map to nearby cells on the 10×10 grid. K-Means clustering was then applied to the SOM Best Matching Unit (BMU) coordinates to group players according to these learned topological relationships.

Cluster Evaluation

Figure 13 shows the elbow (inertia) and silhouette analyses for K-Means over the SOM BMU coordinates. Inertia decreased sharply until $k \approx$ 5–6 and then flattened, indicating diminishing returns beyond that point. The silhouette score peaked at ≈ 0.664 for $k = 10$, suggesting well-separated yet coherent clusters; scores for $k = 9$ and $k = 2$ were also strong (≈ 0.66 and 0.65), confirming consistent structure across resolutions.

Table 2: Cluster evaluation metrics for SOM BMU coordinates.

Metric	**Observation**
Elbow (Inertia)	Sharp decrease until $k \approx$ 5–6, then flattening
Silhouette Score	Peak ≈ 0.664 at $k = 10$; $k = 9$ and $k = 2$ also strong (≈ 0.66, 0.65)

Overall, the SOM representation supports both coarse and fine-grained separations of the data, producing more clearly defined clusters than the AE-only embedding in Section 2.2(a).

Cluster Structure

The optimal $k = 10$ solution yielded the following distribution:

The optimal $k = 10$ solution produced a diverse yet interpretable cluster distribution. Cluster 0, comprising 44 players, exhibited strong shooting efficiency ($TS\%, eFG\%, 3P\%, FG\%$) but lower playing time and games played. These are efficient role players, limited-minute sharpshooters and specialists.

Cluster 1, the largest group with 110 players, showed elevated two-point and rebounding statistics ($2P$, $2PA$, FG, PTS, DRB), identifying interior scorers and dominant forwards or centers.

Cluster 2 (81 players) displayed high three-point attempt ratios ($3PAr$) and turnover rates ($TOV\%$) coupled with below-average accuracy ($FG\%, TS\%$), representing high-volume but inefficient perimeter shooters.

Cluster 9, a smaller group of 39 players, stood out for strong perimeter metrics ($3P$, $3PA$, AST, STL) but weak rebounding, playmaking guards and floor-stretching wings.

Several other smaller clusters captured subtle variations in usage efficiency and defensive contribution, reflecting fine-grained player archetypes across the latent space.

Discussion

The AE+SOM combination captures smooth transitions between player archetypes. Unlike direct AE clustering, the SOM topology helps maintain local consistency in which players with statistically similar styles are grouped into neighboring regions rather than forced into arbitrary boundaries. The resulting K-Means clusters achieve higher silhouette scores, indicating improved separation and interpretability. This suggests that SOM provides a more structured latent space, effectively distinguishing between efficient scorers, volume shooters, interior stars, and secondary contributors.

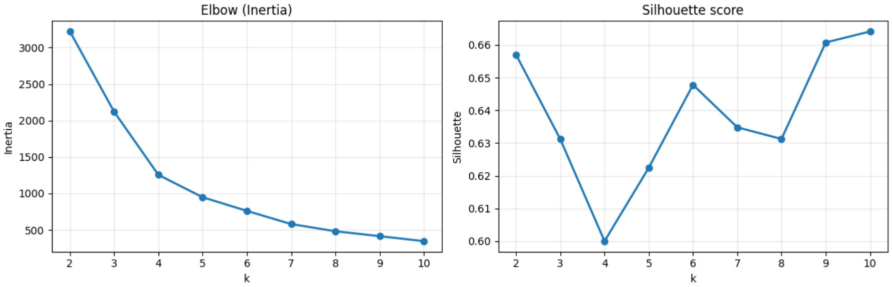

Figure 13: Elbow and silhouette analyses for SOM BMU coordinates clustered with K-Means.

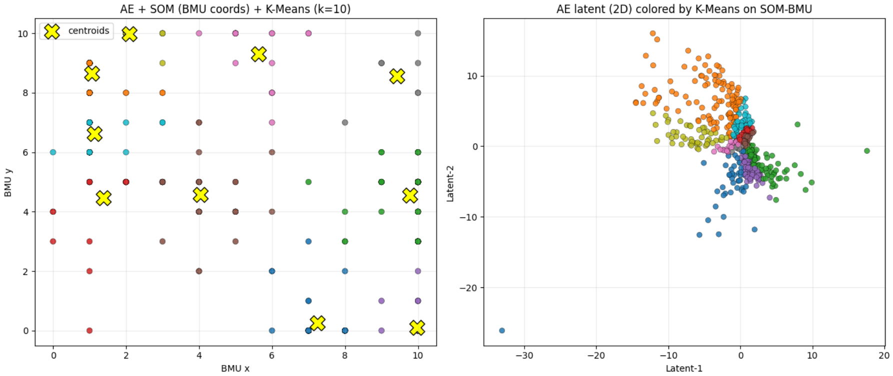

Figure 14: 2D cluster visualization for SOM BMU coordinates clustered via K-Means ($k = 10$).

2.7 2.2(c) Autoencoder + t-SNE + K-Means

To further evaluate the latent space representation, t-Distributed Stochastic Neighbor Embedding (t-SNE) was applied to the 16-dimensional encoded vectors produced by the autoencoder, reducing them to two dimensions for clustering and visualization. Unlike the SOM, which preserves global topology, t-SNE focuses on maintaining local similarity, ensuring that nearby points in high-dimensional space remain close in the 2D projection.

K-Means clustering was then performed on the t-SNE embeddings to identify distinct player archetypes.

Cluster Evaluation

Figure 15 shows the elbow and silhouette analyses for K-Means applied to the t-SNE representation. The elbow curve flattens around $k = 5$–6, indicating diminishing improvement in within-cluster variance reduction beyond this point. The silhouette score peaked at 0.493 for $k = 2$, followed by 0.448 for $k = 4$, suggesting moderate separation between clusters.

These silhouette scores are notably lower than those observed for the SOM approach (≈ 0.66), implying that while t-SNE effectively preserves local neighborhoods, it introduces overlapping regions that blur cluster boundaries. Consequently, clusters in this embedding are less sharply defined but remain interpretable at a finer level of local structure.

Cluster Structure ($k = 6$)

The optimal clustering solution with $k = 6$ produced interpretable groupings of players based on their statistical profiles. Cluster 0, containing about 60 players, included those with high points, salaries, field-goal attempts, and free throws, all clearly representing high-volume scorers and offensive leaders. Cluster 1, comprising 94 players, showed elevated rebounding and efficiency metrics ($ORB\%$, $TRB\%$, $FG\%$), identifying them as efficient interior finishers and rebounders. Clusters 2 through 5 captured mixtures of moderate scorers, defenders, and low-usage contributors, supporting or rotational players with specialized skill sets. The resulting t-SNE map (Figure 16) showed well-separated regional groups with some transitional overlaps, reflecting blended player roles and smooth stylistic gradients.

Discussion

Compared to the SOM-based approach, t-SNE produced smoother local groupings but weaker global structure. The lower silhouette scores suggest that t-SNE emphasizes micro-clusters over broad archetypal separations. Nevertheless, the visualization offers intuitive insight into stylistic relationships among players, highlighting how similar statistical profiles form overlapping yet coherent groups. In summary, AE + t-SNE + K-Means captures fine-grained local differences in player behavior but shows reduced separation between broader archetypes, complementing the SOM approach by revealing nuanced internal structure within the player manifold.

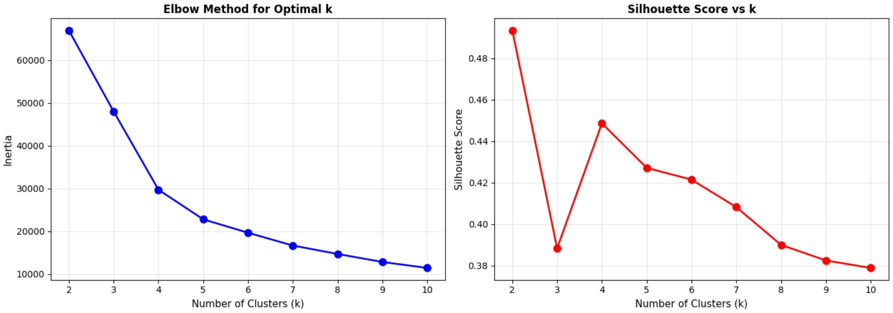

Figure 15: Elbow and silhouette analyses for AE + t-SNE + K-Means clustering.

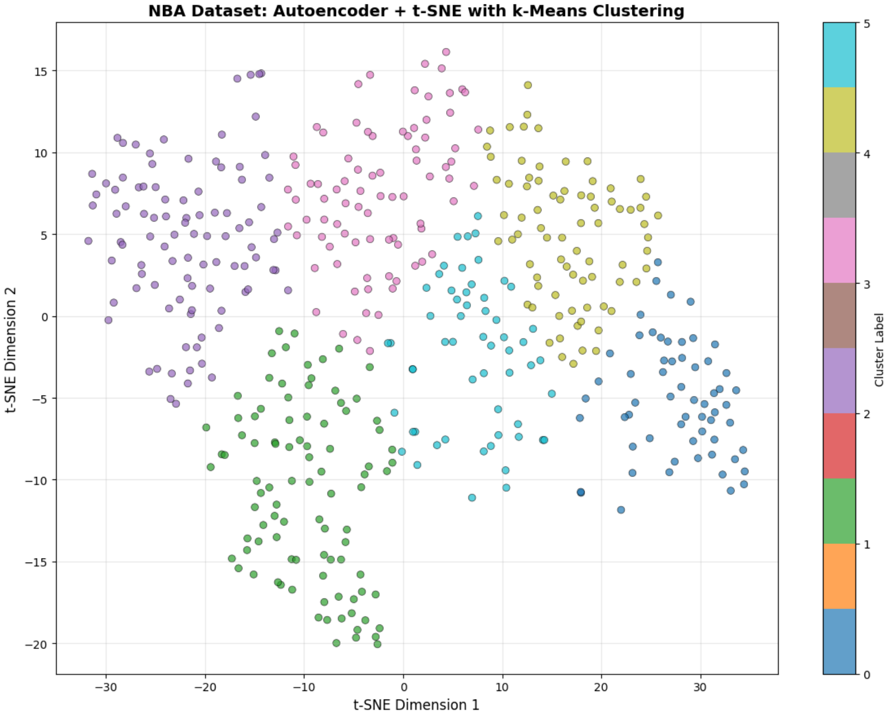

Figure 16: 2D t-SNE embedding clustered using K-Means ($k = 6$).

2.8 2.2(d) Autoencoder + UMAP + K-Means

UMAP (Uniform Manifold Approximation and Projection) was applied to the autoencoder's 16-dimensional latent representation to perform non-linear dimensionality reduction. UMAP balances local and global structure more effectively than t-SNE, preserving neighborhood relationships while maintaining the overall geometry of the data. The resulting two-dimensional embedding was then clustered using K-Means.

Cluster Evaluation

As shown in Figure 17, the elbow curve exhibited diminishing returns after $k \approx 4$–5, suggesting that the intrinsic structure of the data stabilizes beyond this point. The silhouette score reached a maximum of 0.483 at $k = 2$, with values remaining stable around 0.47 for $k = 4$–5, indicating consistent and moderately strong cluster cohesion.

Although UMAP did not achieve the higher silhouette values observed in the SOM-based clustering (≈ 0.66), it outperformed t-SNE in balancing local smoothness and global structural integrity. These results suggest that the UMAP embedding produces more interpretable and spatially coherent clusters overall.

Cluster Structure ($k = 2$)

The optimal two-cluster solution identified two clear player groups. Cluster 0 included high-minute, high-activity players with elevated values in total minutes, games played, field-goal attempts, and defensive win shares, all are essentially consistent starters with strong on-court presence and defensive contribution. Cluster 1 comprised players with higher offensive and total rebounding rates ($ORB\%$, $TRB\%$) and turnover rates ($TOV\%$) but lower playing time and shot volume, representing lower-minute or bench players with moderate productivity in limited roles.

The UMAP projection (Figure 18) showed a clear visual separation between these groups, emphasizing the contrast between primary contributors with heavy workloads and secondary or rotational players with lighter statistical footprints.

Discussion

The AE + UMAP + K-Means approach produced stable and interpretable clusters that effectively capture the distinction between high-usage and supporting players. Compared to t-SNE, UMAP maintained a more coherent global structure and smoother cluster boundaries, avoiding the crowding issue often associated with t-SNE visualizations. While the silhouette scores remained moderate, the visual clarity and interpretability of the embedding make UMAP a strong candidate for understanding the overall landscape of player roles in the NBA dataset.

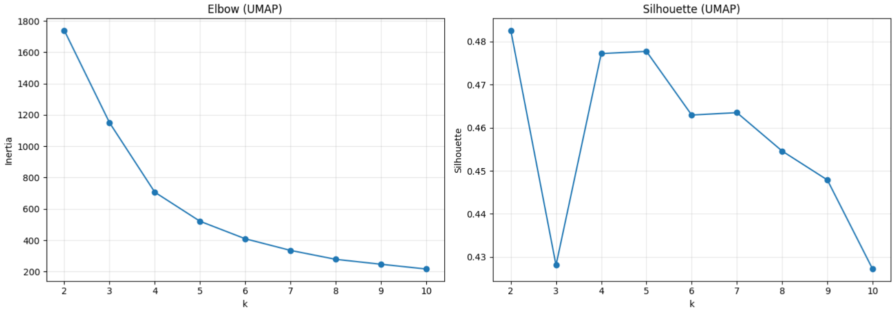

Figure 17: Elbow and silhouette analyses for AE + UMAP + K-Means clustering.

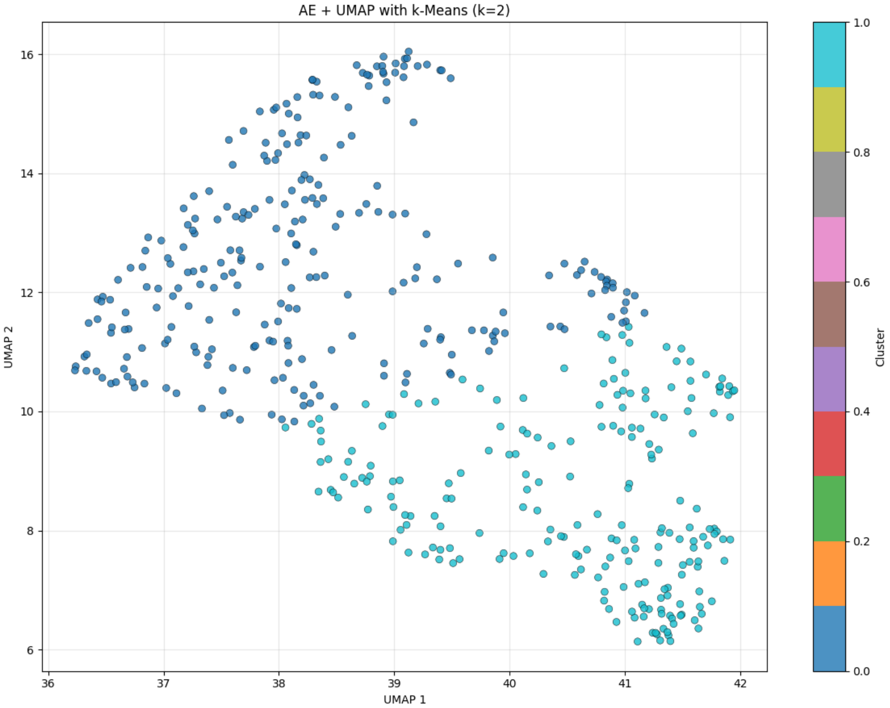

Figure 18: 2D UMAP embedding clustered with K-Means ($k = 2$).

2.9 2.2(e) Variational Autoencoder (VAE) + K-Means

In this final experiment, a Variational Autoencoder (VAE) was trained to learn a continuous latent distribution of the NBA dataset. Unlike the standard autoencoder, which maps each input to a fixed latent point, the VAE learns a probabilistic latent space that captures the underlying data variance more smoothly. This probabilistic modeling enables better generalization and interpretability of the latent dimensions.

After training, the 2-dimensional latent representations were clustered using K-Means to identify natural groupings in the VAE manifold.

Cluster Evaluation

As shown in Figure 19, the elbow curve flattened around $k = 5$–6, suggesting an optimal cluster count near six. The silhouette score peaked at approximately 0.367 for $k = 3$, with stable but moderate values between 0.33 and 0.36 across higher k values. These results indicate that while the VAE captures global structure effectively, its inherent stochasticity produces overlapping clusters, reflecting the smooth nature of its latent space. Overall, the resulting clustering structure was more diffuse than that observed in the deterministic AE-based models, consistent with the probabilistic behavior of the VAE.

Cluster Structure ($k = 6$)

The six-cluster configuration yielded several meaningful player archetypes within the VAE's latent space. Cluster 0 (31 players) exhibited higher turnover and offensive rebound rates ($TOV\%$, $ORB\%$) and free-throw efficiency ($FT\%$) but lower true-shooting percentages ($TS\%$) and win shares per 48 minutes ($WS/48$), identifying inefficient, high-turnover players. Cluster 1 (104 players) displayed elevated defensive and rebounding metrics ($BLK\%$, $ORB\%$, $DRB\%$), representing defensively oriented players with solid inside presence. Cluster 3 (100 players) included high scorers with strong shooting and salary indicators (PTS, FG, FT, $Salary$), clearly marking offensive anchors. Cluster 5 (118 players) featured consistently active contributors with high minutes, games played, and defensive win shares, which are players who serve as dependable workhorses within their teams. Other clusters reflected moderate trade-offs between offensive and defensive contributions, capturing balanced or specialized player roles.

The VAE's clusters exhibited broad thematic distinctions but softer boundaries, a hallmark of its probabilistic encoding scheme.

Discussion

The VAE-based clustering provides a more flexible and generative representation of player archetypes. While its silhouette scores are lower than those achieved with AE + SOM or AE + UMAP (≈ 0.66 and ≈ 0.48 respectively), the VAE excels at modeling gradual variations in player characteristics rather than enforcing rigid boundaries. The continuous nature of its latent space reveals smooth transitions between archetypes, for example, from defensive specialists to balanced contributors, which captures the fluid spectrum of player roles in the NBA.

This generative capability positions the VAE as a powerful foundation for future work, such as simulating hypothetical player profiles or exploring latent interpolations between player types.

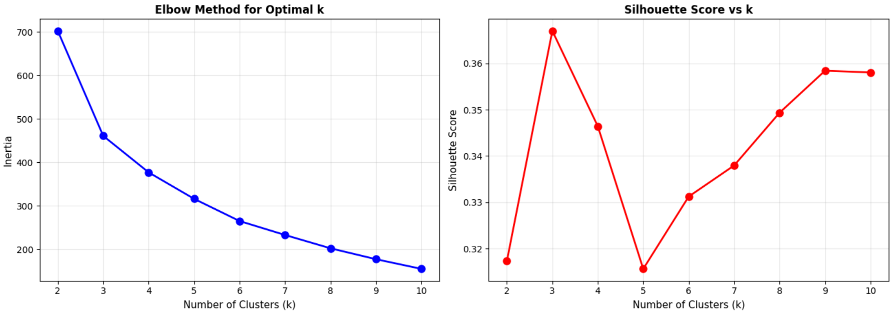

Figure 19: Elbow and silhouette analyses for VAE latent representations clustered with K-Means.

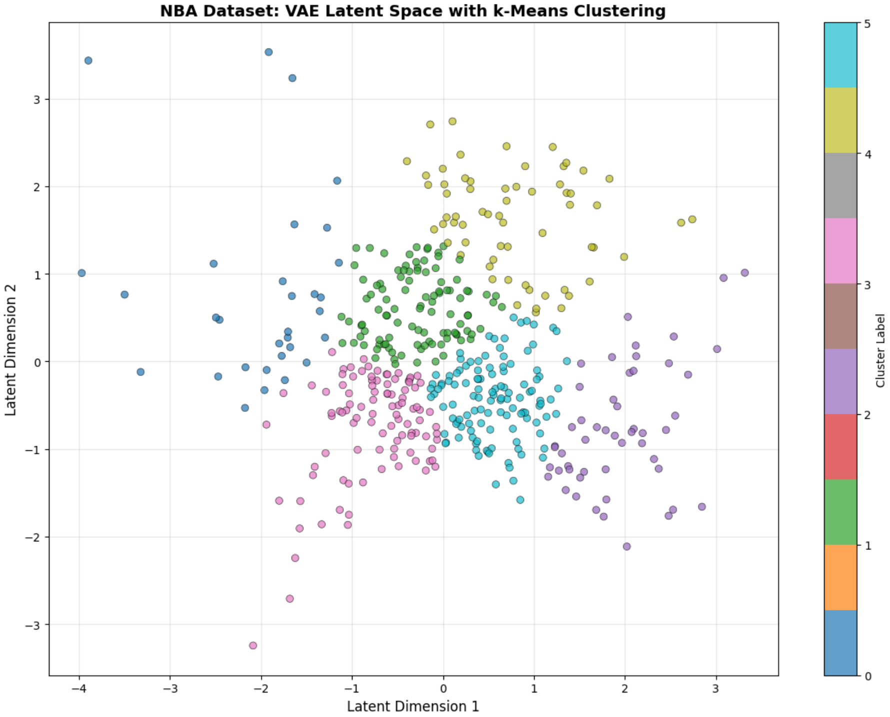

Figure 20: 2D VAE latent space clustered using K-Means ($k = 6$).

2.10 2.2 Overall Comparison and Conclusion

To evaluate the effectiveness of different dimensionality reduction and clustering techniques, all models were compared based on cluster compactness (silhouette score), interpretability, and visual separability. Each approach provides a unique perspective on the underlying player structure within the NBA dataset.

Table 3: Comparison of Dimensionality Reduction + K-Means Results.

Method	Best k	Highest Silhouette	Cluster Characteristics	Interpretation Quality	Comments
(a) AE (2D)	2	0.456	Clear two-group split (high vs. low performance)	4/5	Simple, effective baseline; smooth latent representation.
(b) AE + SOM	10	0.664	Ten well-separated role-based clusters	5/5	Best overall separation; preserves global and local topology.
(c) AE + t-SNE	6	0.493	Moderate separation; overlapping local groups	3/5	Strong local structure, weaker global separation.
(d) AE + UMAP	2	0.483	Distinct high-usage vs. low-usage player split	4/5	Balanced global & local preservation; clear two-cluster division.
(e) VAE (2D)	3	0.367	Soft boundaries; gradual transitions	3/5	Captures continuous skill gradients; less distinct clusters.

Key Insights

- **Best overall performer:** The AE + SOM model achieved the highest silhouette score (≈ 0.66) and produced the most interpretable and well-distributed clusters. Its ability to map local similarities while maintaining a global grid structure made it ideal for identifying distinct player archetypes.
- **Most interpretable latent space:** The UMAP and AE (2D) embeddings provided intuitive visualizations that clearly separated high-usage, star-level players from low-minute or role-specific players.
- **Most detailed micro-clusters:** The t-SNE approach effectively captured subtle local patterns but at the cost of overall structural coherence.
- **Most generative representation:** The VAE produced smooth, continuous transitions between clusters, reflecting player variability rather than strict categorical divisions. While its clusters were less distinct, this flexibility makes it well-suited for generative or exploratory analysis.

Overall Conclusion

Each technique offered complementary perspectives on the player landscape:

- **SOM** excels in structure and interpretability, ideal for categorical insights.
- **UMAP** balances smoothness and global cohesion.
- **t-SNE** reveals fine-grained substructures within local neighborhoods.
- **VAE** models continuity and player similarity transitions.
- **Baseline AE** provides a stable, clear segmentation benchmark.

In summary, **AE + SOM** stands out as the most robust clustering framework, while the **VAE** adds conceptual depth for understanding continuous distributions of player performance. Together, these approaches provide both discrete archetypes and continuous mappings of NBA player roles across statistical dimensions, forming a comprehensive view of performance variability within the league.