A Comparative Tutorial on PCA, t-SNE, and UMAP for Dimensionality Reduction and Clustering on the Mall Customers Dataset

Student Name: Nallamaru Vishnu Vardhan Reddy

Student ID: 24057022

Repository Link: https://github.com/Vishnu181-max/Vishnu_ML_Assignment.git

1. Introduction

Many modern datasets contain features spread across multiple dimensions, making them difficult to visualise or cluster effectively. Human intuition—and most clustering algorithms struggles in high-dimensional spaces, where distances become less meaningful and data points appear uniformly spaced. This phenomenon, known as the *curse of dimensionality*, creates a need for dimensionality reduction techniques that project high-dimensional data into lower-dimensional representations while preserving meaningful structure.

This tutorial explores three of the most widely used dimensionality reduction methods:

- Principal Component Analysis (PCA) – a linear projection method
- t-distributed Stochastic Neighbour Embedding (t-SNE) – a nonlinear, local structure–preserving algorithm
- Uniform Manifold Approximation and Projection (UMAP) – a modern manifold learning approach grounded in topology and graph theory

Our goal is to teach how these methods differ conceptually and practically, and how these differences influence downstream clustering performance. To ground this comparison, we apply each technique to the Mall Customers dataset from Kaggle, which contains demographic and behavioural features used for customer segmentation.

Using K-Means clustering on the 2D embeddings produced by PCA, t-SNE, and UMAP, and evaluating results via silhouette scores, we observe pronounced differences in cluster separation quality. This analysis illustrates the importance of choosing an appropriate dimensionality reduction method—not only for visualisation but also for improving clusterability.

2. Dataset Overview

The Mall Customers dataset consists of 200 customer records, each described by the following features:

- **Gender** (categorical, later encoded numerically)
- **Age**
- **Annual Income (k$)**
- **Spending Score (1–100)**

This dataset is frequently used for teaching customer segmentation, as the combination of income and spending behaviour naturally produces clusters. However, in the raw feature space, these clusters are not always clearly separable. Dimensionality reduction provides a useful way to visualise underlying patterns and enhance clustering algorithms such as K-Means.

Before applying dimensionality reduction, continuous features were standardised using StandardScaler, and the categorical feature *Gender* encoded using LabelEncoder. This ensures all features contribute proportionally to distance-based methods.

---

3. Theoretical Background

A key educational focus of this tutorial is understanding *how and why* PCA, t-SNE, and UMAP produce different embeddings, and when each is appropriate.

3.1 The Curse of Dimensionality

As dimensionality increases:

- Distances between points homogenise
- Local neighbourhoods become less meaningful
- Visualisation becomes impossible without projection
- Clustering algorithms degrade in performance

Dimensionality reduction enables us to compress high-dimensional geometric relationships into 2D or 3D while preserving structure. The effectiveness of this projection directly influences clustering results.

3.2 Principal Component Analysis (PCA)

PCA is a **linear** transformation that identifies orthogonal directions (principal components) capturing maximal variance. Mathematically, PCA solves an eigenvalue decomposition of the covariance matrix and projects data onto the top components.

Strengths:

- Fast and computationally efficient
- Produces interpretable linear components
- Preserves global structure
- Useful for denoising

Limitations:

- Cannot capture nonlinear structure
- Often fails to reveal natural clusters when relationships are curved or manifold-like

In our experiment (see Figure 2), the first two components of PCA explained approximately **60% of the dataset's variance**, indicating moderate information retention.

3.3 t-Distributed Stochastic Neighbour Embedding (t-SNE)

t-SNE is a **nonlinear** technique designed primarily for visualisation. It preserves *local neighbourhoods* by converting pairwise distances into probability distributions and minimising the Kullback Leibler divergence between high-dimensional and low-dimensional similarities.

Strengths:

- Excellent at revealing tight clusters
- Preserves local structure well
- Widely used for visualising embeddings in NLP and computer vision

Limitations:

- Does not preserve global structure
- Sensitive to hyperparameters (e.g., perplexity)
- Time-consuming on large datasets
- Distances between clusters are not meaningful

Despite limitations, t-SNE often produces visually compelling cluster maps.

3.4 Uniform Manifold Approximation and Projection (UMAP)

UMAP is a **state-of-the-art manifold learning algorithm** based on algebraic topology. It constructs a graph representation of the high-dimensional data and optimises a low-dimensional embedding that preserves both local and some global structure.

Strengths:

- Faster and more scalable than t-SNE
- Preserves both local and global structure better
- Produces stable, well-separated clusters
- Works well as a preprocessing step for clustering

Limitations:

- More complex mathematically
- Results can vary with hyperparameters

UMAP is increasingly preferred for exploratory data analysis and embedding generation.

4. Experimental Methodology

To compare PCA, t-SNE, and UMAP fairly, we applied each to the standardised dataset using 2-dimensional embeddings. For each embedding, **K-Means clustering with 5 clusters** (a conventional choice for this dataset) was performed.

We then evaluated clustering quality using the **silhouette score**, which measures how similar a point is to its own cluster compared to others. Scores range from:

- **+1** → Perfectly separated clusters
- **0** → Overlapping clusters
- **-1** → Incorrect cluster assignment

All experiments used identical random seeds to ensure reproducibility.

5. Results

5.1 PCA Results

PCA captured a cumulative explained variance of approximately 60% with its first two components (Figure 2). The resulting 2D scatter plot showed only partial

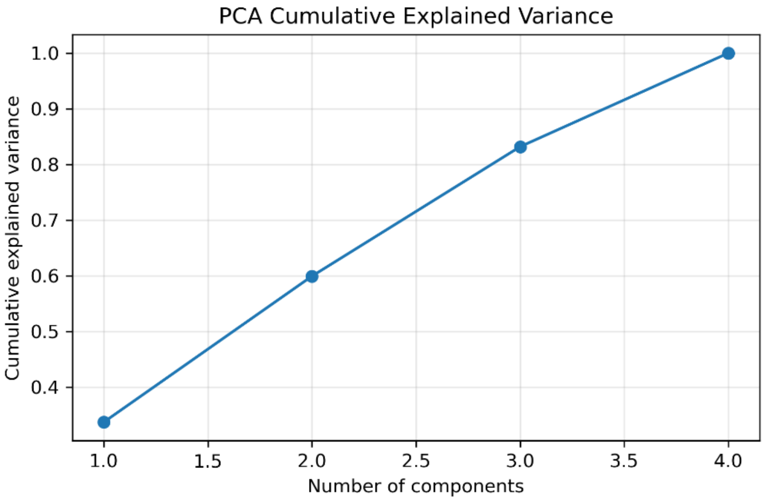

Figure-1: PCA Explained Variance

separation of customer groups, with significant overlap. K-Means on PCA produced a silhouette score of:

- **Silhouette (PCA): 0.404**

This indicates **weak–moderate cluster separation**, consistent with PCA's linear nature.

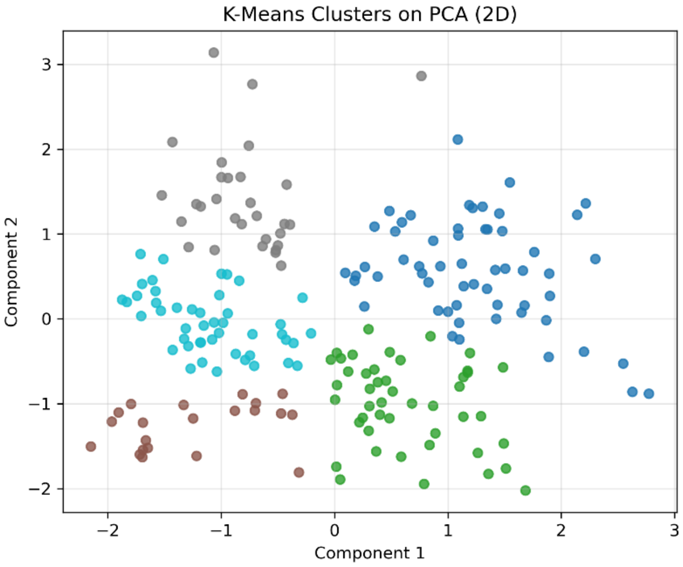

Figure-2: PCA K-means Clustering

5.2 t-SNE Results

t-SNE produced a much clearer visual separation of clusters. Dense, compact neighbourhoods emerged, and distinct clusters were evident. However, distances between clusters could not be meaningfully interpreted due to t-SNE's probability-based embedding mechanism.

Silhouette score:

- **Silhouette (t-SNE): 0.528**

This represents a substantial improvement over PCA and highlights t-SNE's strength in discovering nonlinear structure.

5.3 UMAP Results

UMAP produced the most well-organised clusters with both compact intra-cluster grouping and good overall layout. Visually, it revealed the clearest structure, and K-Means produced:

- **Silhouette (UMAP): 0.644**

This strong score suggests **high-quality cluster separation**, validating UMAP's reputation as an effective preprocessing step for clustering.

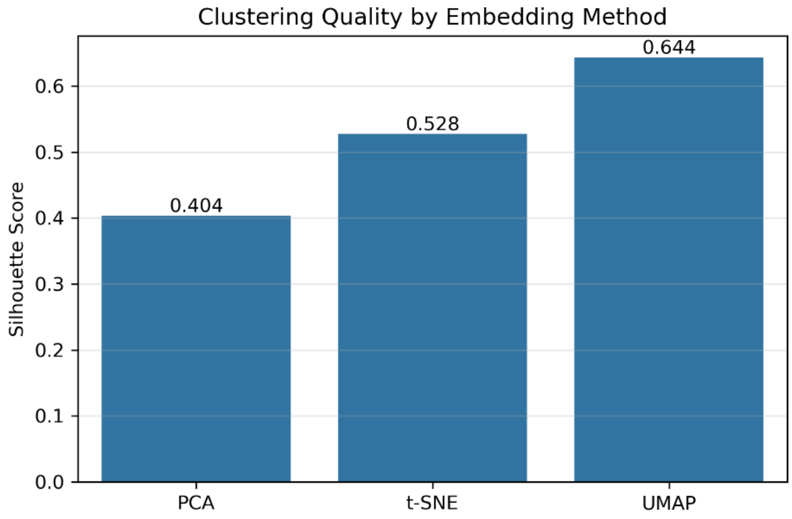

Figure-3: Silhouette Score comparison

5.4 Silhouette Score Comparison

Insert Figure: Silhouette Bar Chart Here

The silhouette bar chart clearly demonstrates:

- UMAP significantly outperforms both PCA and t-SNE
- t-SNE performs notably better than PCA
- PCA provides limited separation due to its linear nature

This empirical hierarchy aligns with theoretical expectations: nonlinear manifold preservation enables t-SNE and UMAP to uncover structure PCA misses.

6. Discussion

The differences in silhouette scores and embeddings highlight a fundamental lesson:

The choice of dimensionality reduction method directly impacts clustering quality.

Why PCA underperforms

- PCA is restricted to linear relationships
- Real-world customer behaviour often lies on nonlinear manifolds
- Linear projection compresses structure and overlaps clusters

Thus, PCA is excellent for interpretability but poor for discovering complex patterns.

Why t-SNE improves clustering

t-SNE effectively preserves *local neighbourhood relationships*, drawing similar points together while exaggerating cluster boundaries. This makes clusters visibly distinct and easier for algorithms like K-Means to identify.

However:

- t-SNE distorts global structure
- Clusters may appear artificially separated
- Distances between clusters lack interpretability

Therefore, t-SNE is powerful for visualisation but not always ideal for downstream tasks requiring global structure.

Why UMAP performs best

UMAP combines advantages of PCA and t-SNE:

- Preserves local neighbourhoods
- Retains some global structure
- Captures manifold geometry
- Produces stable, meaningful embeddings

This leads to:

- Strong cluster compactness
- Better K-Means performance, Highest silhouette score in our experiment.

UMAP is therefore an excellent default choice for exploratory analysis and clustering workflows.

7. Conclusion

This tutorial has demonstrated how PCA, t-SNE, and UMAP differ in theory and practice when applied to customer segmentation. Using the Mall Customers dataset, we showed that:

- **PCA** provides limited cluster separation due to its linear nature
- **t-SNE** preserves local structure well and provides clearer clusters
- **UMAP** offers the best overall embedding for clustering, achieving the highest silhouette score (0.644).

These findings reinforce a crucial insight for practitioners:

Choosing the right dimensionality reduction technique is essential, as it determines whether clustering algorithms reveal meaningful patterns or misleading structures.

For small and medium-sized datasets with nonlinear relationships, **UMAP is generally the most effective**. For visualisation focused on local structure, **t-SNE is excellent**, while **PCA remains valuable** for speed, interpretability, and noise reduction.

This study illustrates how dimensionality reduction serves not merely as a visualisation tool but as a foundational component of effective exploratory data analysis.

References

1. van der Maaten, L. & Hinton, G. (2008). *Visualizing Data using t-SNE.* Journal of Machine Learning Research.
2. McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* arXiv:1802.03426.
3. Jolliffe, I. (2002). *Principal Component Analysis.* Springer Series in Statistics.
4. Aggarwal, C. (2018). *Machine Learning for High-Dimensional Data.* Springer.