Evaluating PCA and t-SNE for Dimensionality Reduction on Bike Sharing Data

Bahadir Bagci

November 2025

1 Introduction

Dimensionality reduction (DR) methods aim to keep most of the information from data while working in fewer dimensions. A main difference between these methods is that some use linear projections, while others follow nonlinear relations in the data. As a linear method, Principal Component Analysis (PCA) performs this reduction by using the variation in the data to form new axes that summarize the main structure. In contrast, t-distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear method that focuses on preserving the neighbourhood relationships between data points in the lower-dimensional space. In this work, we compare both the two-dimensional visualization and the predictive performance of Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) using the Bike Sharing Rental dataset. The following sections of this report are organized as follows: the Methodology section describes the applied steps including preprocessing, PCA, and t-SNE implementation; the Results section presents both visual and predictive comparisons; and finally, the Conclusion summarizes the main findings.

2 Methodology

2.1 Data

The dataset includes 17,379 observations collected from the years 2011 and 2012. It contains information about the number of bike rentals, weather conditions, and time-related features. The total number of rented bikes (count) is used as the target variable, while all other fields serve as input features for the analysis.

2.2 Feature Selection and Data Preparation

Month was removed because it carries the same seasonal information as season. Casual and registered were also excluded since they directly relate to the target variable. Hour was grouped into three parts (0–8, 8–16, 16–24) to show daily

patterns. All features were converted into numeric form and normalized before applying PCA and t-SNE.

2.3 Implementation Steps

Two dimensionality reduction methods were applied: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). In the first stage, these methods reduced the dataset into two dimensions for visualization, and the results were interpreted based on how well they represented the data structure. In the second stage, their predictive performances were compared with each other and with a baseline model built on the original features.

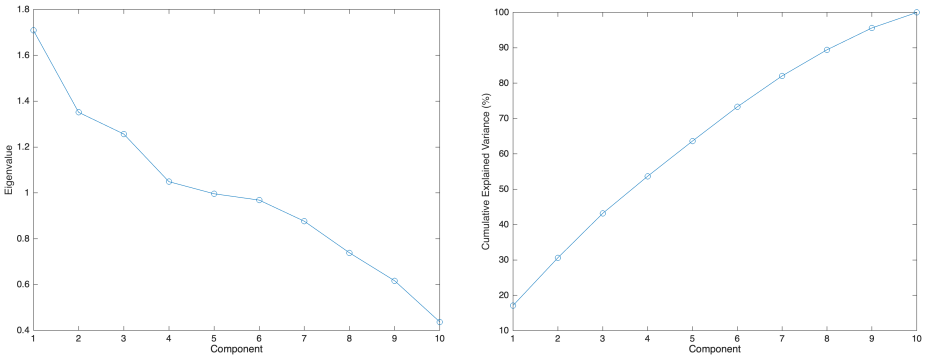
2.4 Predictive Modeling

A regression model using the bagged ensemble method was applied. The model was trained on the PCA features, t-SNE features, and the original data to compare their predictive performance.

3 Results

3.1 Visualizations in 2 Dimension

When looking at Figure 1, the curve appears almost as a straight line rather than having a clear elbow. No single component dominates the variance. This problem suggests that the data cannot be expressed well in a linear way, and that PCA does not capture a clear underlying structure.



(a) Scree Plot (b) Cumulative Explained Variance

Figure 1: Scree Plot and Cumulative Explained Variance

Figure 2 supports this: the 2D PCA projection is a single dense cloud and high–low rental values are mixed, so linear PCA cannot separate the observations.

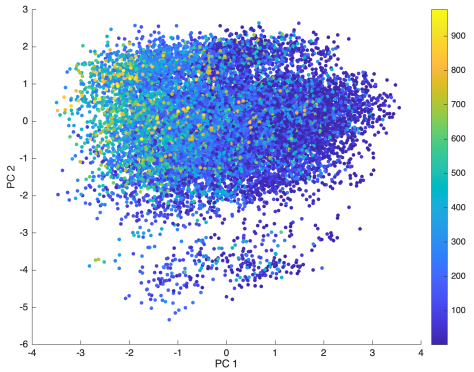


Figure 2: 2D PCA Projection

Figure 3 shows the t-SNE projection using Euclidean distance. Compared to PCA, some local clusters appear more clearly, and the points are spread more evenly in the space. Still, the clusters are not fully separated, but the result suggests that t-SNE captures certain local patterns that the linear PCA could not.

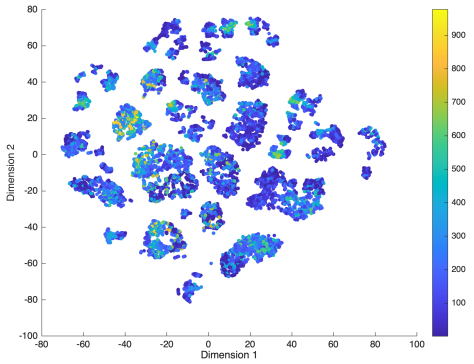


Figure 3: 2D t-SNE Projection

3.2 Predictive Performance

Based on Table 1, the baseline model reached an R^2 of 0.79, while PCA gave a slightly lower score of 0.76. The t-SNE–based model performed best with an R^2 of 0.80 and the lowest RMSE (84.0), showing a small improvement over both PCA and the baseline model.

Method	R^2	RMSE
Base	0.787	85.35
PCA	0.760	95.99
t-SNE	0.795	84.00

Table 1: Predictive performance comparison between models.

4 Conclusion

Overall, the results of this study show that PCA and t-SNE provide similar outcomes for this dataset. However, both the two-dimensional visualization and the predictive performance indicate that t-SNE performs slightly better than PCA.