BM20A6100 Advanced Data Analysis and Machine Learning

Exercise 1 – Nonlinear Dimensionality Reduction

Deadline: 03 November 2025

Arman Golbidi

2025

Repository & Submission Links

GitHub repository: `https://github.com/tembooo/ADML_Part2`
This exercise folder: `https://github.com/tembooo/ADML_Part2/tree/main/1.Exercises/1.HomeworkAdvancedDimensionalityReduction(DL3.11.)`

Submission guidance (course page): Closes: **Monday, 3 November 2025, 8:00 AM**. Create a new GitHub repository (or use an existing one if you have already made one for this course) and return the solutions by adding a link to the GitHub repository in Moodle. Make sure to name the files clearly. You can either create a report PDF or utilize Markdown comment sections in a Jupyter notebook. Comments within the code are insufficient to act as answers to the questions. Prepare solutions to the given tasks and submit a link to your GitHub repository containing the solutions by the given deadline. The material is provided only for the course participants.

Contents

1 Comparing linear and non-linear DR (4 points)

1.1 Data and Preprocessing (Bike Sharing)

The Bike Sharing dataset from OpenML (`data_id = 42712`) contains 17 379 records describing hourly bike rental demand with categorical and numerical features. Categorical attributes (`season, holiday, weekday, workingday, weather`) were label-encoded, while continuous ones (`year, month, hour, temp, feel_temp, humidity, windspeed`) were standardized using:

$$z_i = \frac{x_i - \mu}{\sigma}, \tag{1.1}$$

where μ and σ denote the feature mean and standard deviation.

No missing values were found. Figure 1 shows the distribution of the target variable (`count`), which is right-skewed, implying that most hours have moderate rental demand.

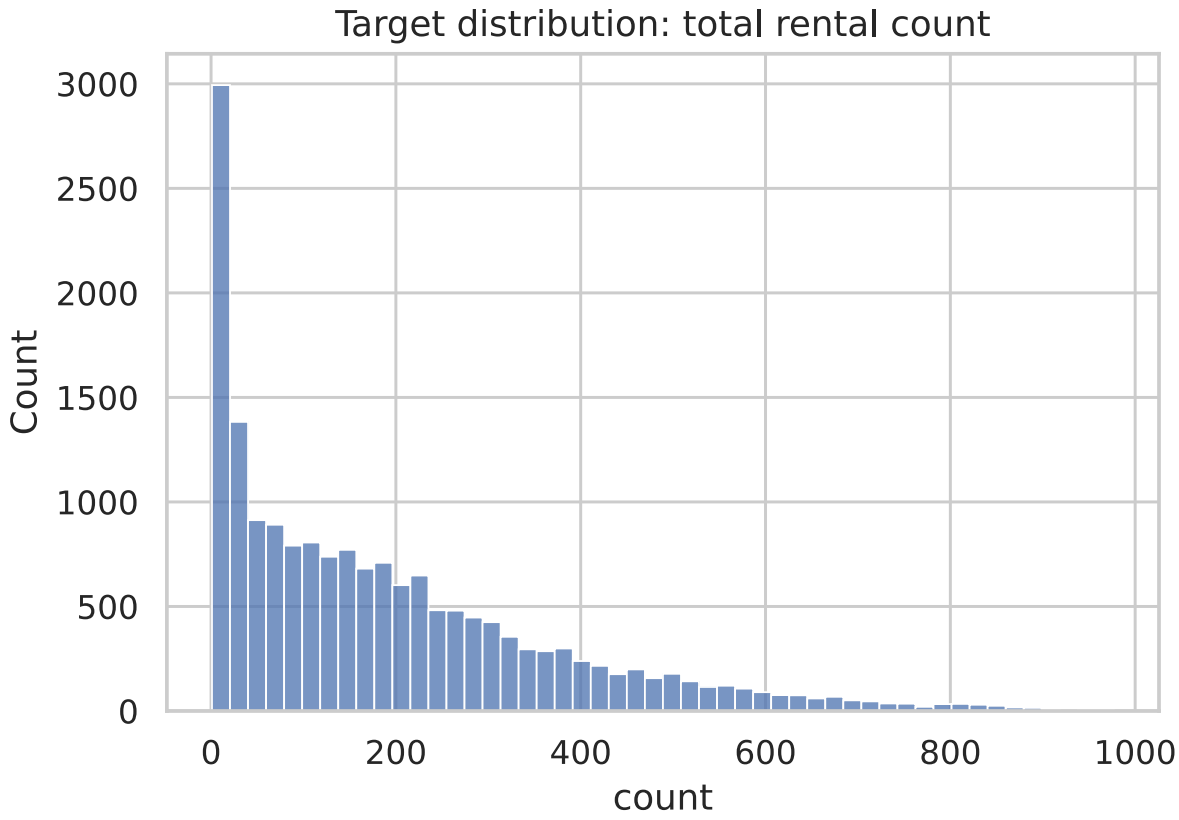

Figure 1: Distribution of total bike rentals (`count`).

To further explore usage patterns, Figures 2–6 display category frequencies.

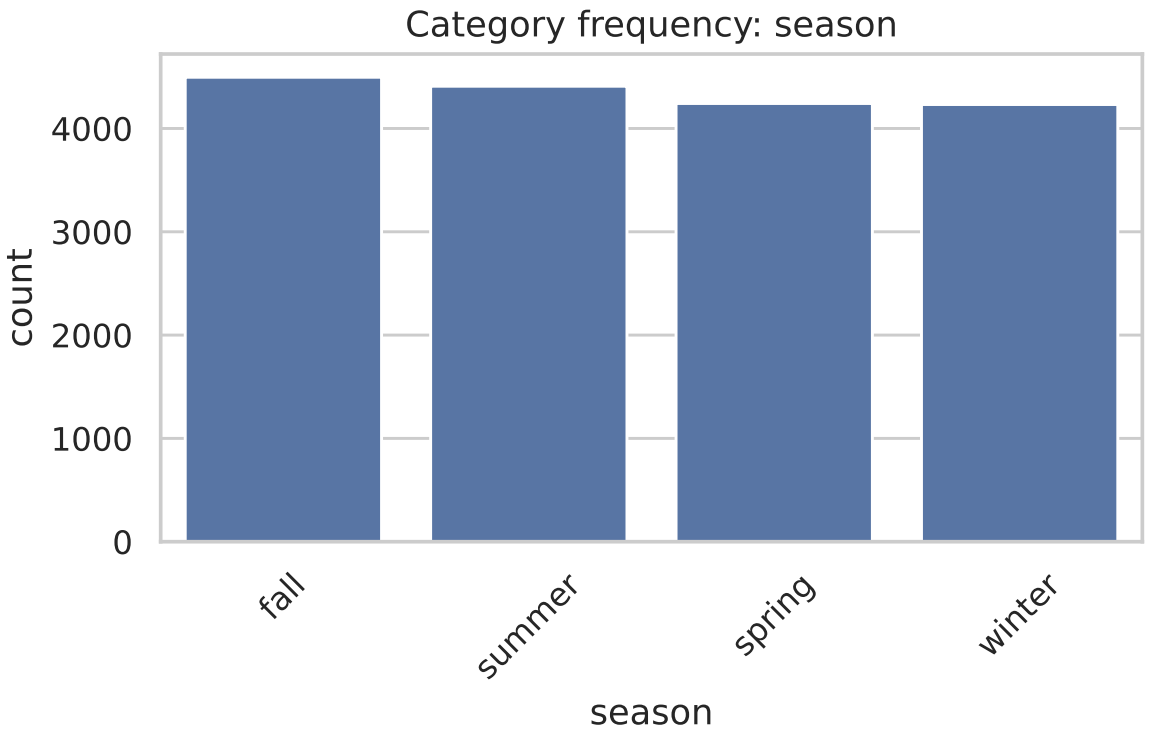

Figure 2: Rental frequency by season.

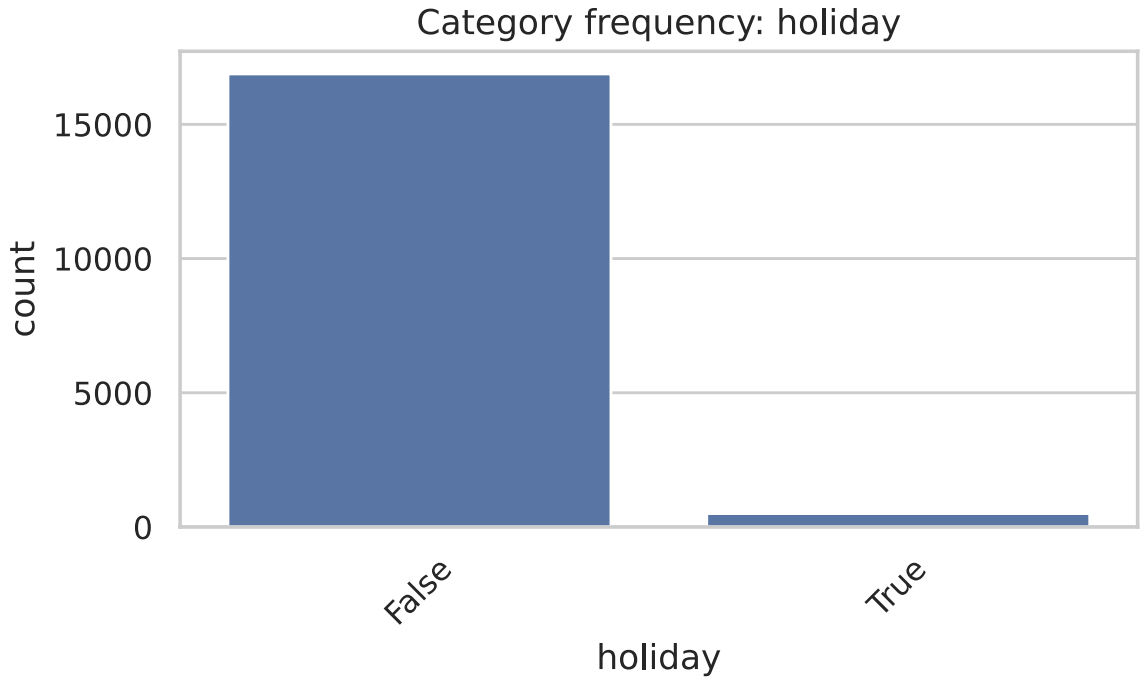

Figure 3: Rental frequency by holiday status.

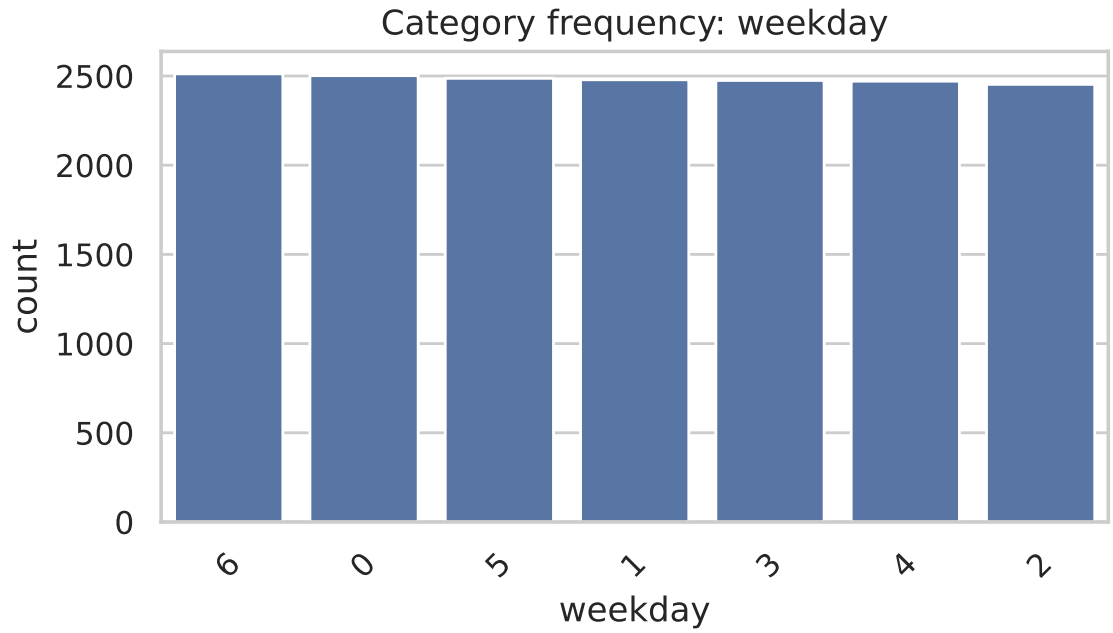

Figure 4: Rental frequency by weekday.

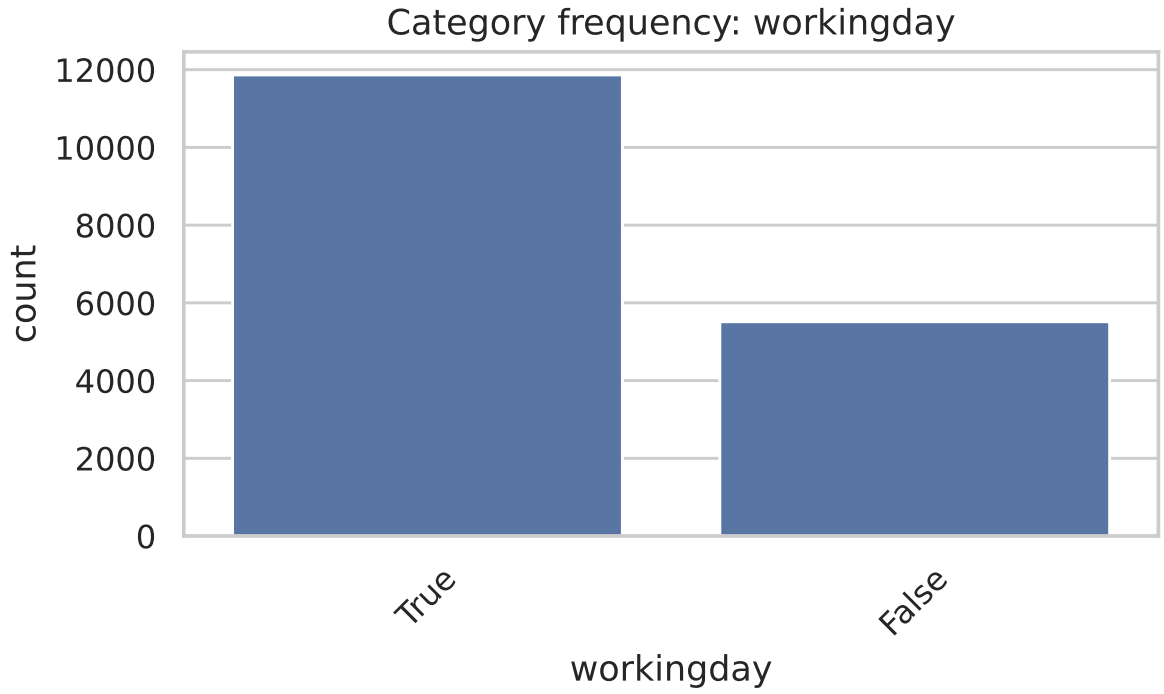

Figure 5: Rental frequency by working day.

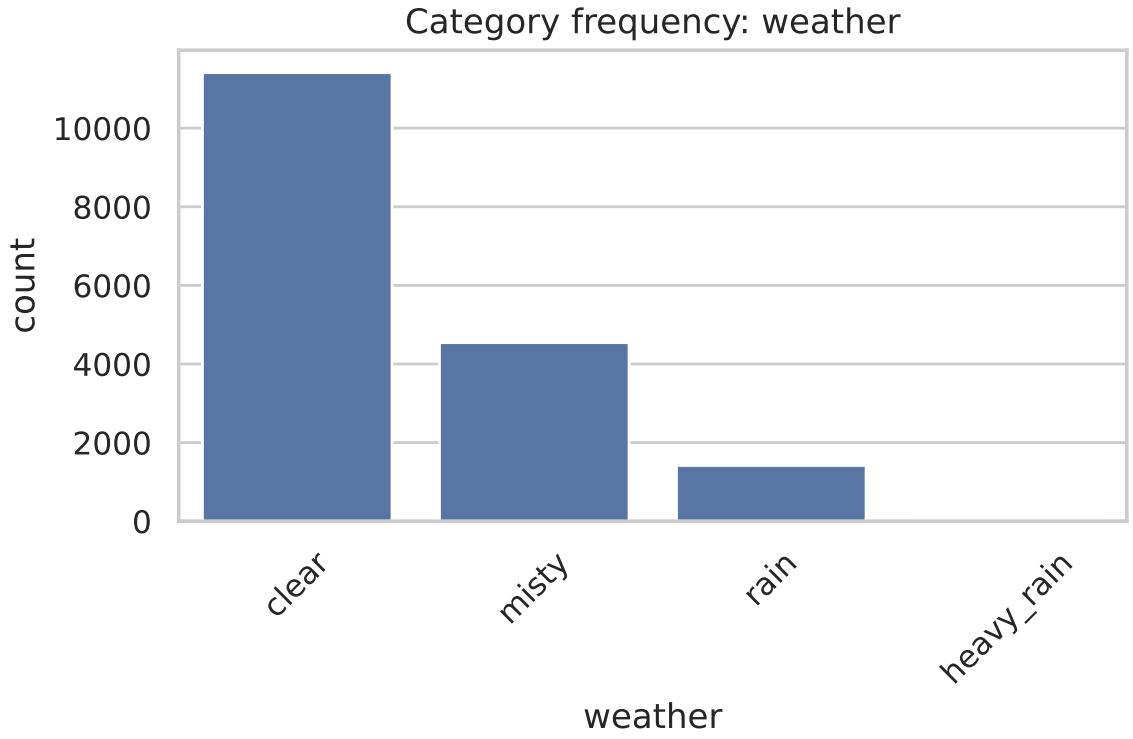

Figure 6: Rental frequency by weather condition.

1.2 Dimensionality Reduction: PCA vs t-SNE

Principal Component Analysis (PCA) was first applied to the standardized features. It projects the data onto orthogonal directions that maximize variance:

$$\mathbf{Z} = \mathbf{X}\mathbf{W}, \quad \text{where } \mathbf{W} = \arg\max_{\mathbf{W}^\top \mathbf{W} = \mathbf{I}} \mathrm{Var}(\mathbf{X}\mathbf{W}). \tag{1.2}$$

Only the first two components were retained for visualization.

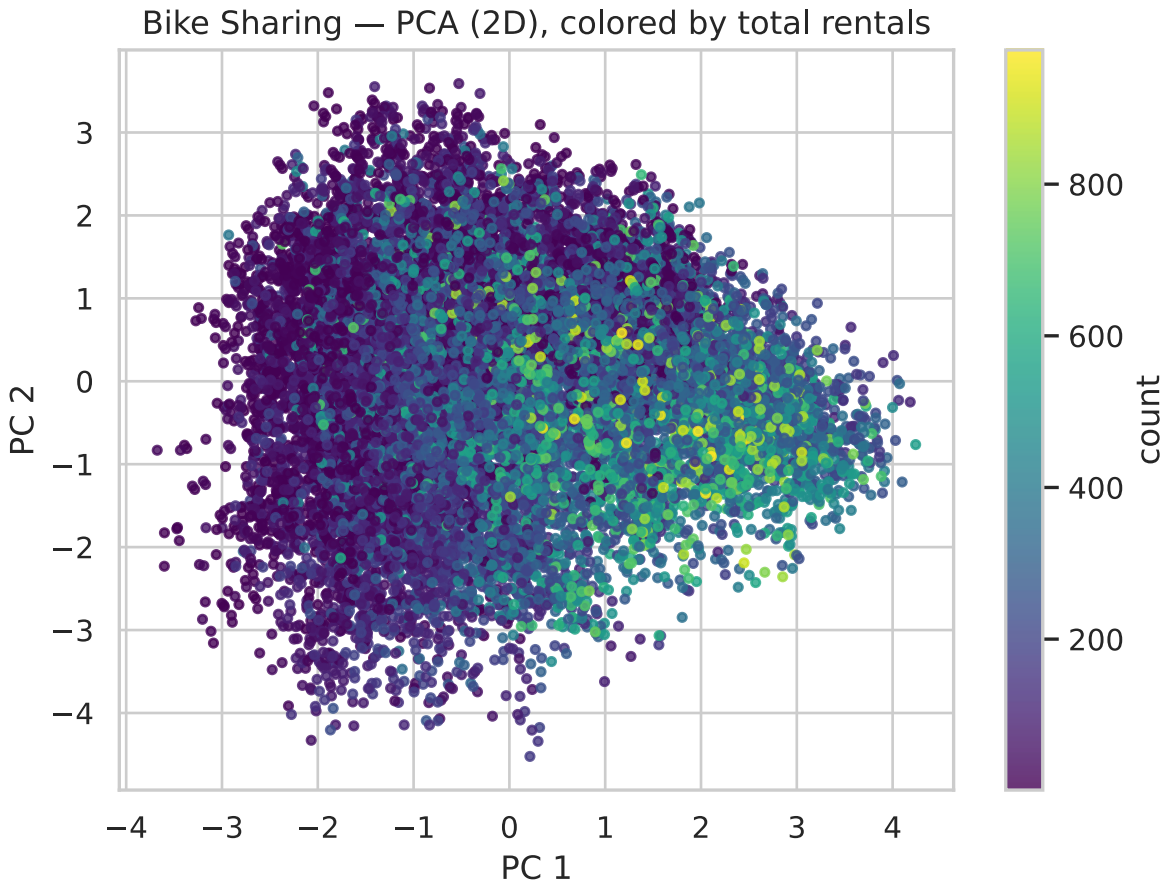

Figure 7: PCA (2D) projection of the Bike Sharing dataset.

Interpretation: Figure 7 displays a dense, overlapping cloud of samples with a mild color gradient along PC1, indicating only weak linear correlation between total rentals and the main direction of variance. PCA efficiently summarizes overall variability but cannot clearly separate different demand regimes, implying that nonlinear or feature-interaction effects dominate.

Next, t-Distributed Stochastic Neighbor Embedding (t-SNE) was applied with perplexity $= 30$ and $1\,000$ iterations. t-SNE minimizes the Kullback–Leibler divergence between pairwise similarity distributions in high and low dimensions:

$$\mathrm{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{1.3}$$

where p_{ij} and q_{ij} denote neighborhood affinities in the original and embedded spaces, respectively.

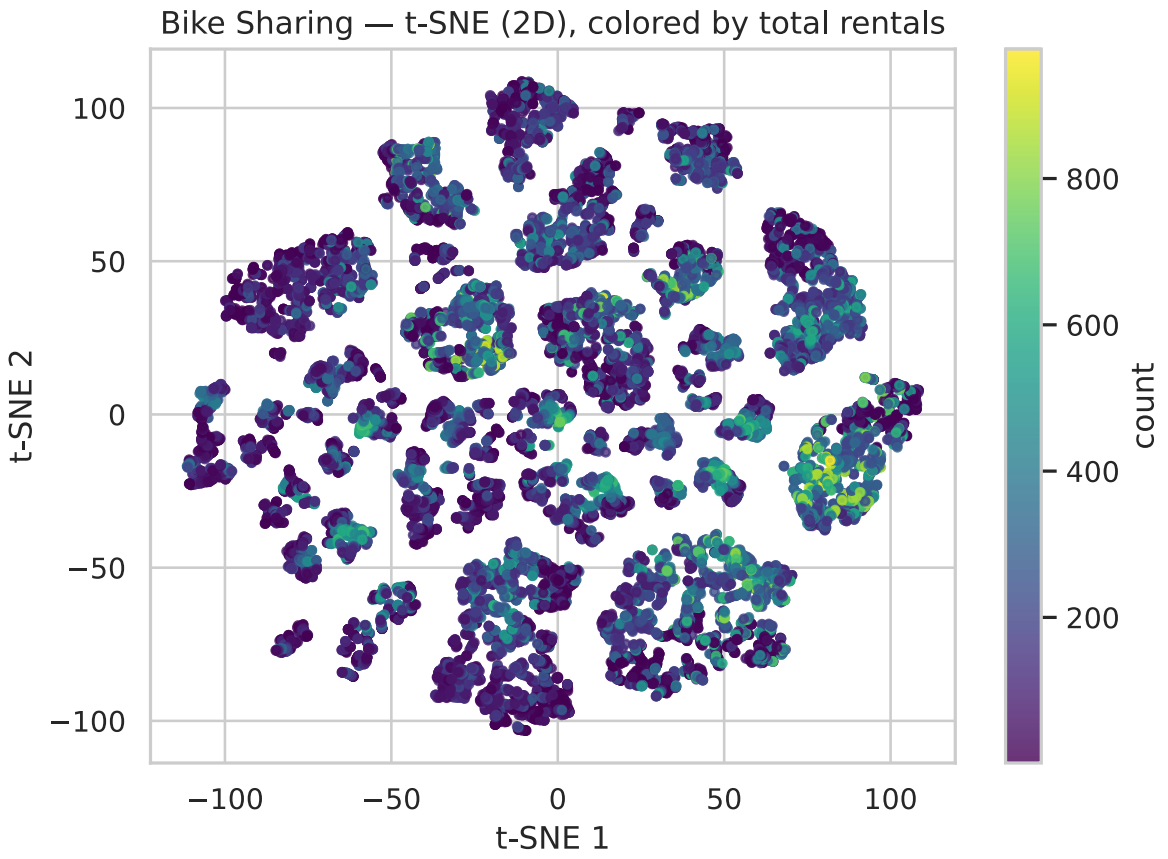

Figure 8: t-SNE (2D) projection of the Bike Sharing dataset.

Interpretation: The t-SNE plot reveals multiple compact local clusters, where nearby samples exhibit similar rental counts. Smooth color transitions within clusters and sharper boundaries between them indicate the presence of distinct behavioral regimes (e.g., weekday commuting, weekend leisure, or weather-driven variations). Compared with PCA, t-SNE captures a richer nonlinear structure that aligns more closely with demand intensity.

1.3 Prediction Model and Evaluation

A Random Forest Regressor was trained using 80/20 train-test splits on the two-dimensional PCA and t-SNE representations. Model accuracy was assessed via the coefficient of determination (R^2) and Root Mean Squared Error (RMSE):

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{1.4}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}. \tag{1.5}$$

Table 1: Performance comparison between PCA and t-SNE features.

Model / Features	R^2	**RMSE**
Random Forest (PCA 2D)	0.180	161.14
Random Forest (t-SNE 2D)	0.598	112.76

Result Summary: Table 1 shows that t-SNE yields a substantially higher R^2 and lower RMSE, demonstrating that nonlinear embedding preserves structure more predictive of rental demand. However, t-SNE's axes lack explicit meaning, making it unsuitable for scalable prediction pipelines. For practical modeling, retaining more PCA components or using the original standardized features would better maintain the global variance information.

1.4 Feature structure in DR spaces (per assignment)

Beyond the target-colored views, we examined how original features appear in the two DR spaces. For PCA with standardized inputs, the loading of feature f on component k equals the correlation between feature f and the PC-k scores:

$$\mathrm{corr}(x_f, z_k) \;=\; w_{f,k}\,\sqrt{\lambda_k}, \tag{1.6}$$

where $w_{f,k}$ is the f-th entry of eigenvector $\mathbf{w}_k$ and λ_k its eigenvalue (explained variance of PC-k). In practice, *hour*, *temp*, and *season* showed the largest absolute loadings on PC1/PC2, indicating that diurnal and weather factors drive the dominant linear variance.

For t-SNE, axes are not linear directions; hence we assessed features by *coloring* the embedding with individual feature values. This revealed banded patterns for *hour* (commute-related neighborhoods), smooth arcs for *temp/feel_temp*, and fragmented islands for *weather/humidity*. These feature-wise overlays align with the clusters visible in Figure 8 and explain why the Random Forest trained on t-SNE coordinates captures demand regimes more effectively than the PCA-2D representation.

1.5 Discussion

Overall, PCA provided a broad linear overview of data variance, while t-SNE revealed hidden nonlinear manifolds reflecting daily and contextual variations. The stronger cluster formation in Figure 8 explains the Random Forest improvement in Table 1. Nevertheless, t-SNE should be regarded primarily as an exploratory visualization tool rather than a feature extractor for predictive modeling due to its computational cost and lack of inverse mapping.

2 Visualizing with SOM (3 points)

2.1 SOM Setup (MNIST-784)

A Self-Organizing Map (SOM) was trained on a 20 000-sample subset of the MNIST-784 dataset. The map grid size was 15×15, using Gaussian neighborhood and learning rate $\alpha = 0.5$. Weights were initialized randomly, and training proceeded for $1\,000$ iterations.

Data acquisition and normalization. We fetch MNIST-784 from OpenML and split features/labels. Grayscale pixel intensities are normalized to $[0, 1]$ by

$$x'_p = \frac{x_p}{255}, \qquad p = 1, \ldots, 784. \tag{2.1}$$

This scale stabilizes training and emphasizes shape over absolute brightness.

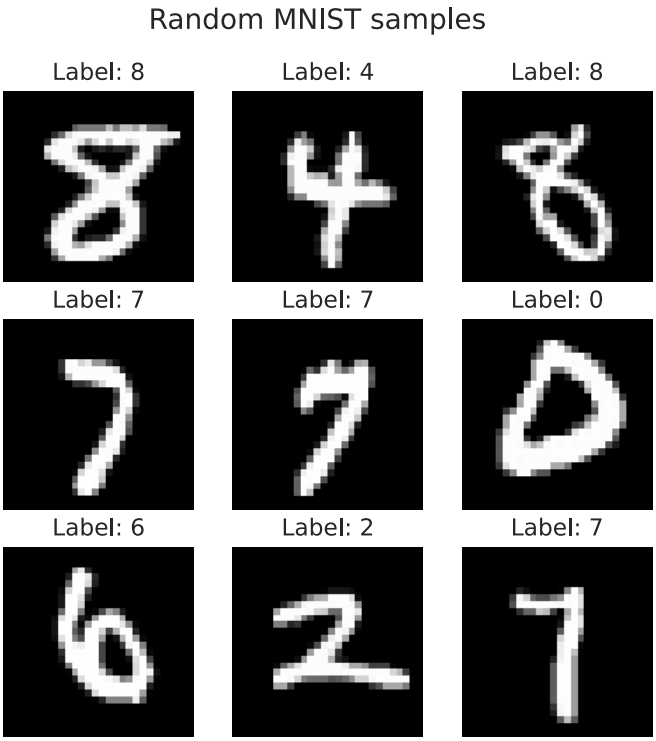

Figure 9: Random MNIST samples (9 images of size 28×28). Variability in stroke, tilt, and curvature is evident even within the same class.

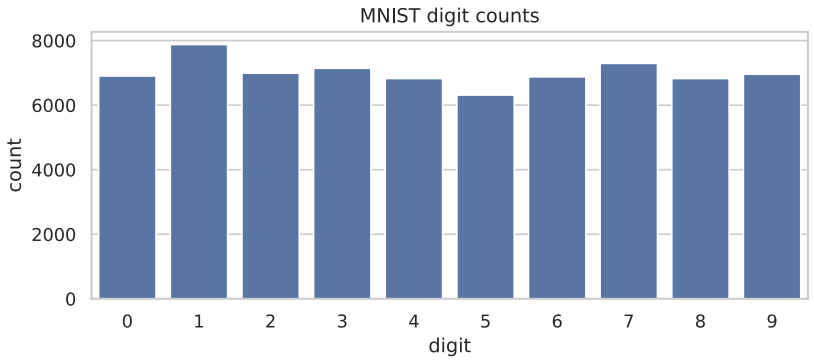

Figure 10: Label counts for digits 0–9. The distribution is nearly uniform, reducing sampling bias across classes.

SOM objective and learning rule. Given an input $\mathbf{x} \in \mathbb{R}^{784}$ and neuron weights $\{\mathbf{w}_j\}$ arranged on a 2D grid at positions $\{\mathbf{r}_j\}$, the Best Matching Unit (BMU) index b is

$$b \;=\; \arg\min_{j} \left\|\mathbf{x} - \mathbf{w}_j\right\|_2. \tag{2.2}$$

With Gaussian neighborhood

$$h_{bj}(t) \;=\; \exp\!\left(-\frac{\|\mathbf{r}_b - \mathbf{r}_j\|_2^2}{2\,\sigma(t)^2}\right), \tag{2.3}$$

weights update as

$$\mathbf{w}_j(t{+}1) \;=\; \mathbf{w}_j(t) \;+\; \alpha(t)\, h_{bj}(t)\,\big(\mathbf{x} - \mathbf{w}_j(t)\big), \tag{2.4}$$

where $\alpha(t)$ and $\sigma(t)$ decay over iterations to shrink the effective neighborhood.

2.2 Results and Interpretation

U-Matrix (inter-neuron distances). The unified distance matrix highlights boundaries between clusters. For neuron (i, j) with weight $\mathbf{w}_{i,j}$ and neighbor set $\mathcal{N}_{i,j}$, we visualize

$$U_{i,j} \;=\; \frac{1}{|\mathcal{N}_{i,j}|} \sum_{(p,q)\in\mathcal{N}_{i,j}} \|\mathbf{w}_{i,j} - \mathbf{w}_{p,q}\|_2\,. \tag{2.5}$$

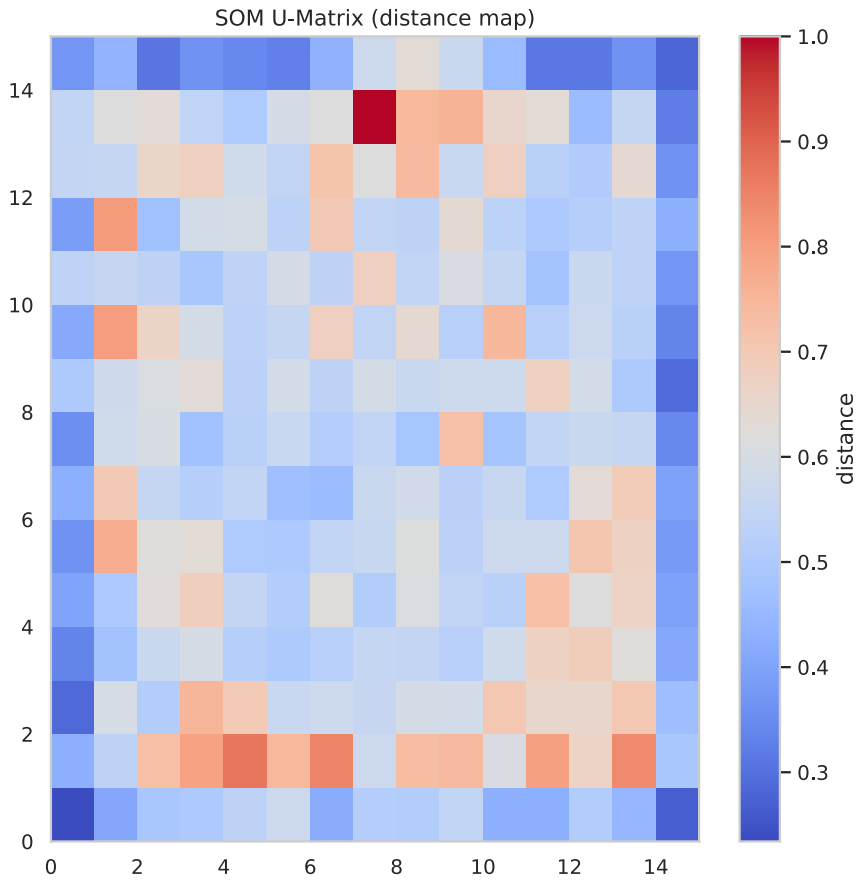

Figure 11: SOM U-Matrix on a 15×15 map trained with 20 000 samples. Warm ridges indicate sharp transitions; cool basins indicate coherent clusters.

Neuron-wise predominant labels. Assigning each training image to its BMU and coloring neurons by the majority label makes clusters explicit and reveals semantic adjacency (e.g., curved digits such as 3 and 8 placed nearby, slender 1 forming a narrow band).

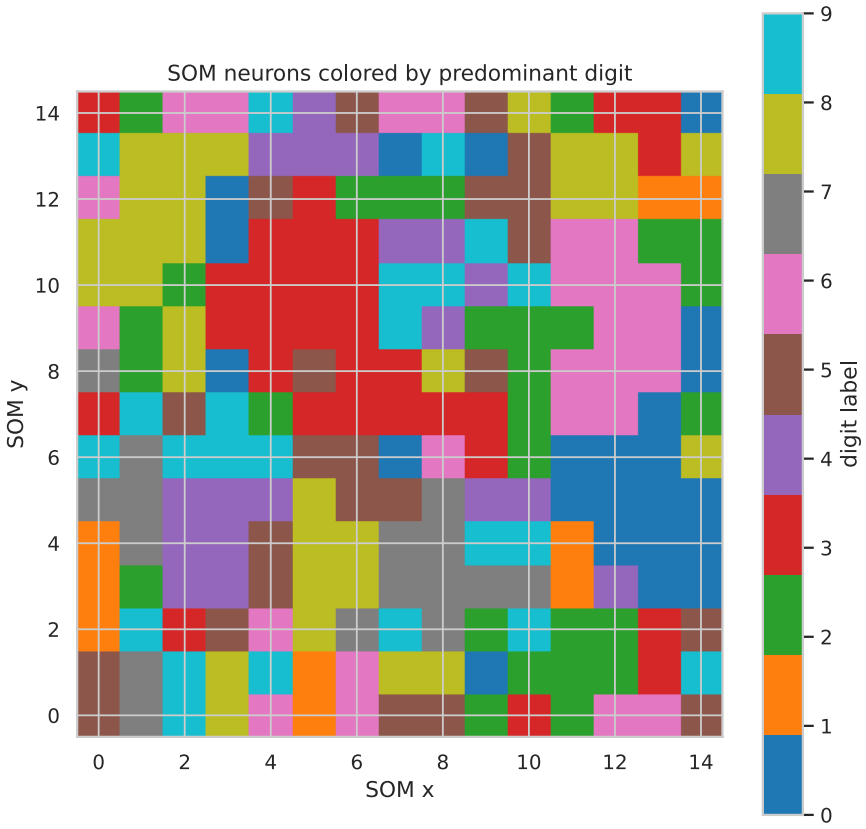

Figure 12: Neuron coloring by predominant digit label. Compact, contiguous regions reflect topology preservation of the SOM.

Prototype inspection. Sampling representatives from selected neurons confirms high intra-neuron consistency (stroke thickness, tilt, curvature) and inter-neuron diversity.

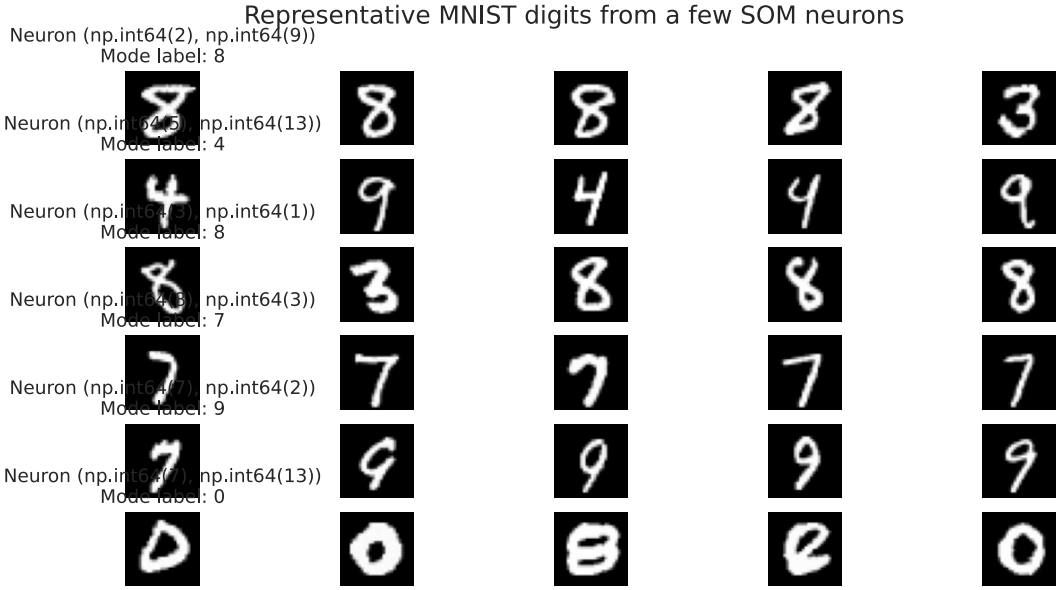

Figure 13: Representative MNIST digits from several SOM neurons. Each row corresponds to one neuron; titles show neuron coordinates and mode label.

2.3 Limitations and Future Work

SOMs require manual choices of grid size, iterations, and decay schedules. Resolution is limited relative to modern deep embeddings (e.g., autoencoders, UMAP). A practical extension is to learn features with a CNN (or an autoencoder) and then train a SOM on the feature space to obtain higher-fidelity maps with preserved interpretability.

Final summary. PCA provided a linear, global overview of Bike Sharing variability, whereas t-SNE uncovered tighter local neighborhoods aligned with rental intensity, leading a simple Random Forest on t-SNE coordinates to outperform the same model on two PCA components. For production prediction, however, more PCA components or original features are preferable. On MNIST, the SOM produced an interpretable 2D atlas: the U-Matrix exposed boundaries; neuron-wise majority labels revealed cluster structure and semantic adjacency; and prototype grids validated coherent local representations—together offering a topology-preserving, human-readable view that complements standard DR tools.