2.11.2025 Ina Laurila

Advanced Data Analysis and Machine Learning
Period 2, Homework 1: Nonlinear Dimensionality Reduction

Task 1 - Comparing linear and non-linear DR

The objective was to explore how the different features are shown in the dimensionality-reduction components and to compare the performance of a prediction model with the different DR techniques: Principal Component Analysis (PCA) representing linear relationships, and t-distributed Stochastic Neighbor Embedding (t-SNE) capturing non-linear similarities. The Bike Sharing Rental dataset was used in the task. Dataset contains daily rental counts and weather-related variables such as temperature, humidity, wind speed and season.

The PCA biplot (Figure 1) shows how the different weather and calendar variables are related to each other. The first principal component (PC1) is dominated by temperature and feeling temperature, both strongly positively correlated with bike rentals. Humidity and windspeed are pointing quite opposite direction meaning that different weather attributes drive usage in different directions.

The second component relates more to humidity and month, indicating variation between different times of the year and weather conditions. Together, the first two components explained about 33% of the total variance.

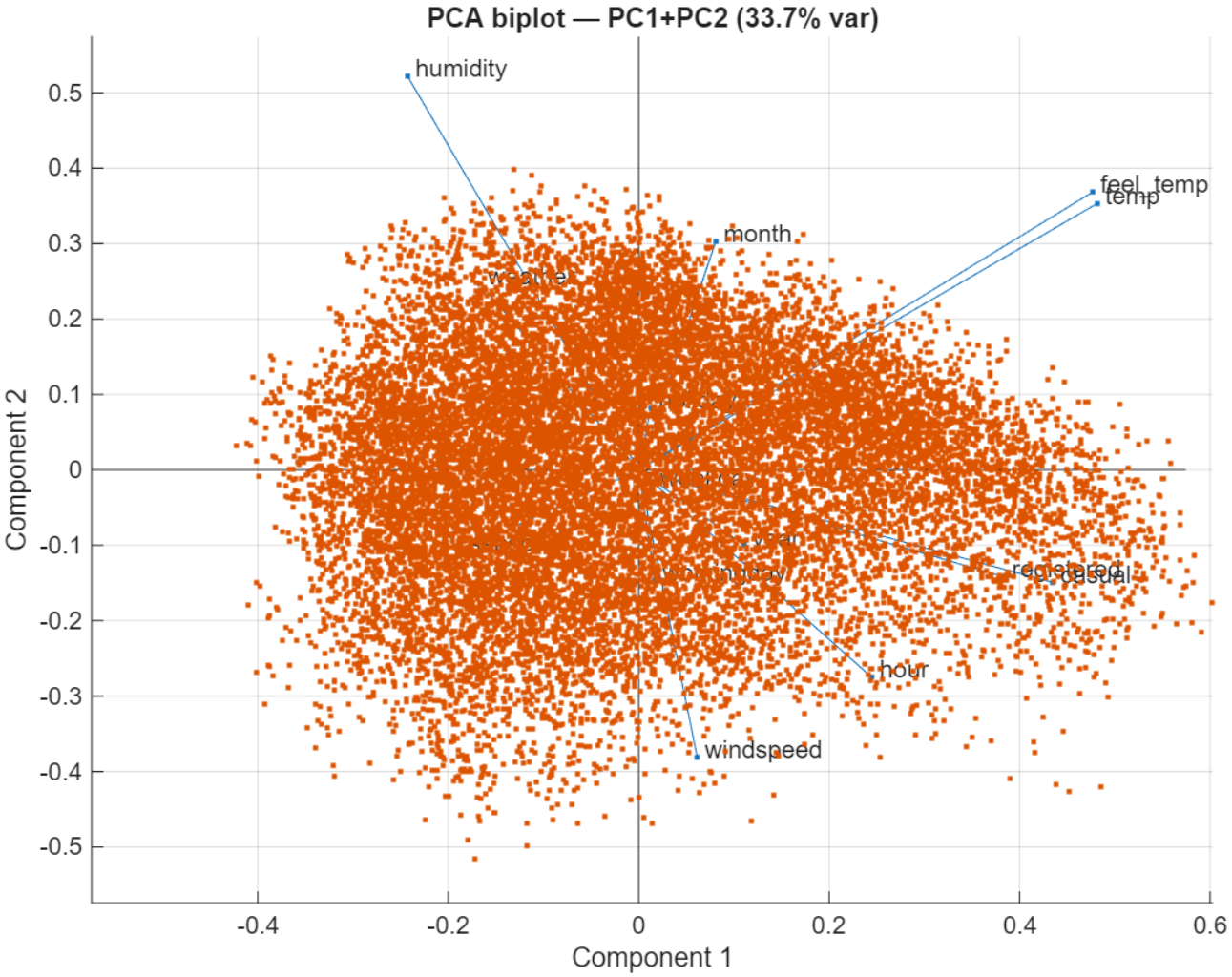

Figure 1. PCA biplot

The loading plots (Figure 2) confirm the same pattern. For PC1, the strongest positive loadings come from temperature, feeling temperature, and the rental variables (casual and registered), while humidity and windspeed have clear negative loadings. For PC2, humidity and windspeed dominate, while month and feel_temp also contribute positively. This means that PC1 mainly reflects overall weather favourability, and PC2 distinguishes between humid and seasonal effects.

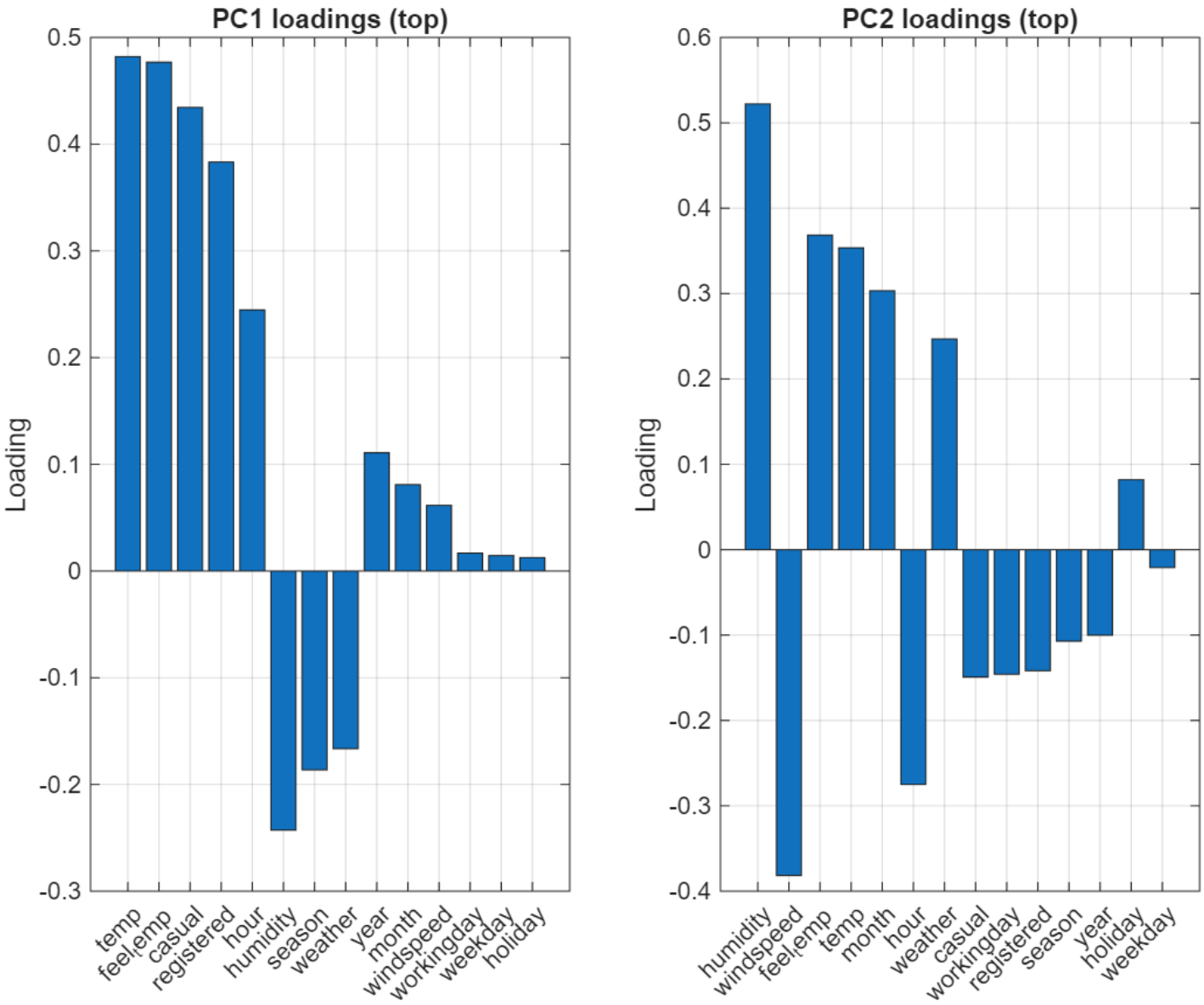

Figure 2. PCA component loadings

The t-SNE visualisation complements PCA by revealing more detailed, non-linear clusters. The most interesting variables based on PCA were chosen for t-SNE visualisations: *feel_temp, humidity, windspeed, season, month* and *registered*.

In the plots in Figure 3, points with similar weather and seasonal conditions appear close to each other, forming visible clusters. The clearest separation is seen with the *season* variable, which forms distinct clusters. The *month* variable follows the same pattern, showing a

gradual transition through the seasons. Other variables show more scattered distribution within those seasonal clusters.

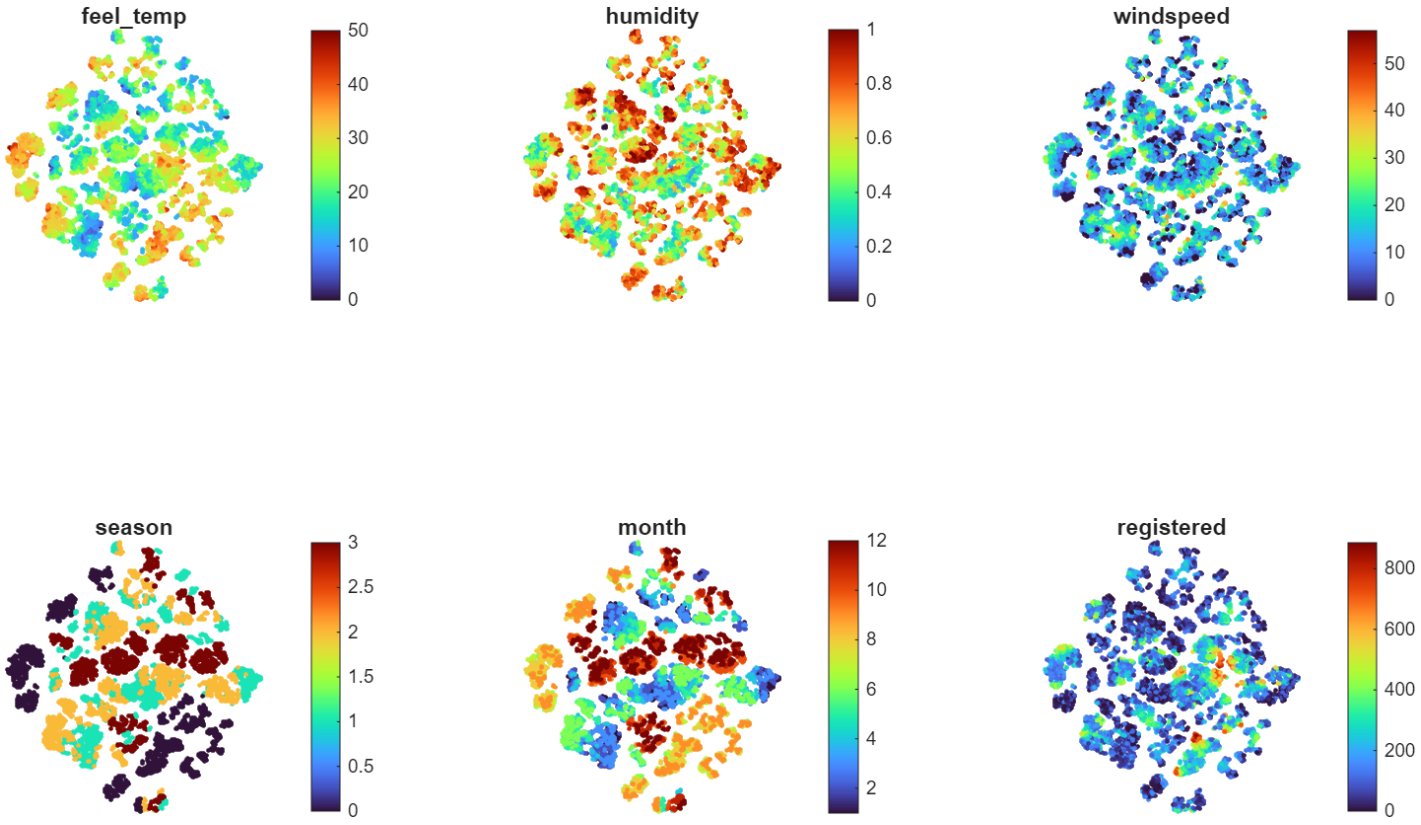

Figure 3. t-SNE colored by selected features

A Random Forest regression model was used to predict the total number of rented bikes. Three versions of the model were trained and compared: original (no dimensionality reduction), PCA components, and t-SNE embeddings.

Model performance was evaluated using the coefficient of determination (R^2) and the root mean squared error (RMSE). R^2 measures how well the model explains the data variance, while RMSE expresses the average prediction error.

Table 1 shows the results. The model trained on the original data performed the best ($R^2 \approx 0.99$), showing that the raw features contain all the necessary information for accurate prediction. Using PCA slightly reduced accuracy ($R^2 \approx 0.97$), which is expected because some information is lost when the data is linearly compressed into a smaller number of components. The model based on t-SNE features performed poorly ($R^2 < 0$), since t-SNE preserves local relationships for visualization but does not maintain the global structure required for regression.

Model	RMSE	R2
{'RF raw' }	6.5916	0.99873
{'RF PCA' }	27.535	0.97782
{'RF t-SNE'}	194.3	-0.10424

Table 1. Model comparisation

Overall, PCA can reduce dimensionality with minimal accuracy loss, but non-linear visualization methods like t-SNE should not be used directly for predictive modeling.

Task 2 - Visualizing with SOM

For the visualization task, a self-organizing map (SOM) was trained using a subset from the MNIST-784 dataset. The MNIST-784 dataset consists of 70 000 grayscale images of handwritten digits (0–9), each represented by 784 pixel values. For faster computation, a random subset of 5 000 samples was used.

A 15×15 self-organizing map was trained to visualize the data structure. The map size was chosen as a compromise between resolution and training time — large enough to separate the main digit clusters but still efficient to compute. Around 20 training epochs were done.

The U-matrix visualization (Figure 1) shows the distance between neighbouring SOM nodes. Bright areas correspond to clusters of similar digits, whereas dark boundaries mark transitions between different classes.

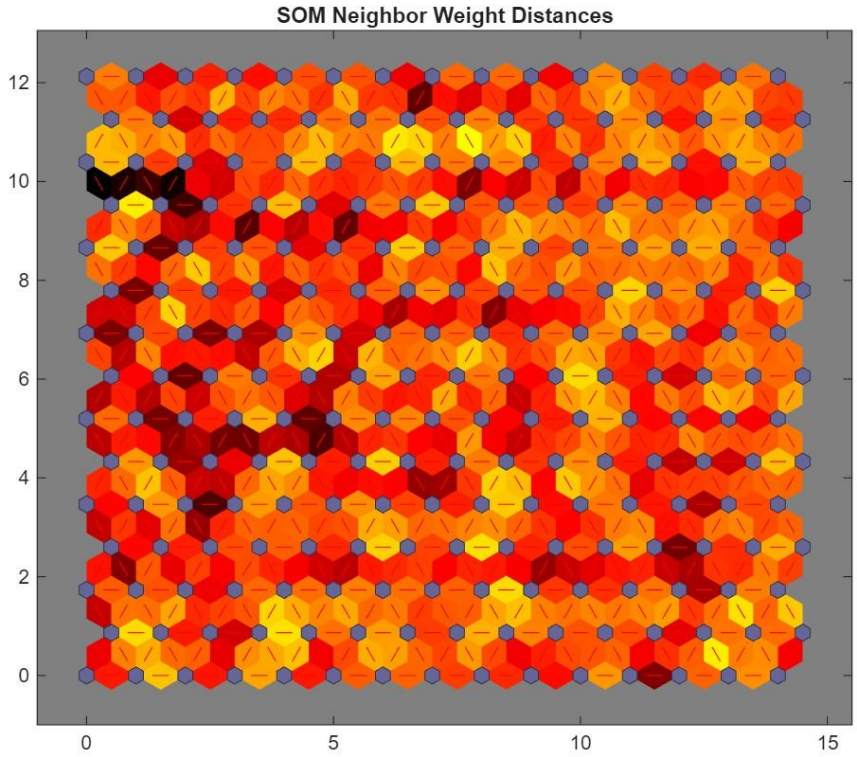

Figure 1. U-matrix

The even colouring may result from the high dimensionality of the MNIST data and the limited map size. Increasing the map size or training epochs would likely produce clearer separation lines between clusters.

The hits map (Figure 2) shows how many samples were assigned to each SOM node. Nodes with many hits represent frequently occurring digit styles. Nodes with only a few hits correspond to rare or ambiguous handwriting styles, and empty nodes indicate areas of the map that were not used. The distribution of hits is fairly even, which means that the map covers the data space well and that no single region dominates the representation.

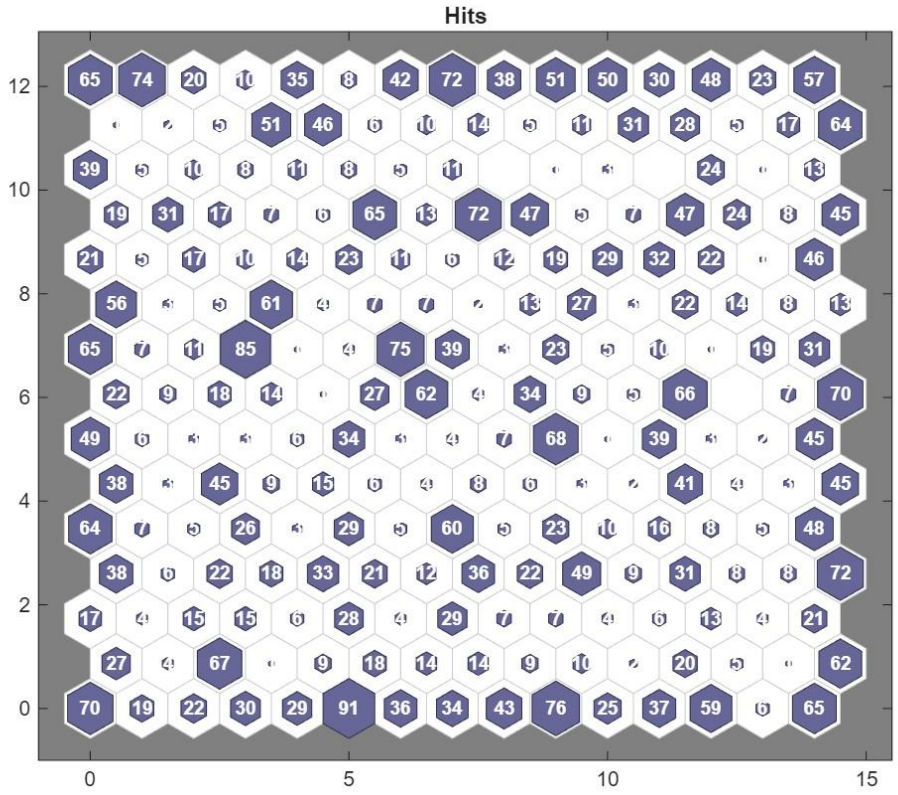

Figure 2. SOM hits map