CEFET/RJ

Data Discretization & Smoothing

Transforming continuous variables into meaningful discrete intervals through proven statistical techniques

Eduardo Ogasawara
eduardo.ogasawara@cefet-rj.br
https://eic.cefet-rj.br/~eogasawara

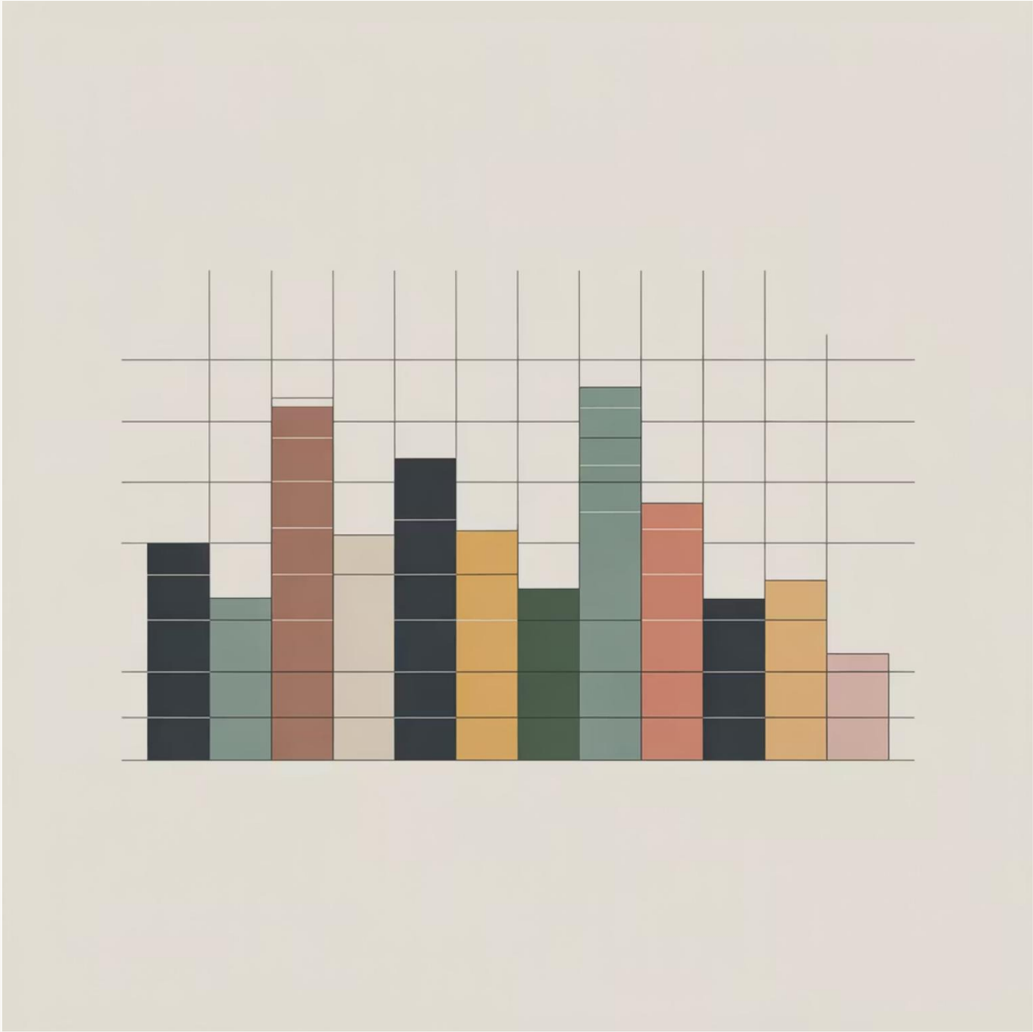
Interval-Based Discretization

How It Works

Divides the data range into equal-width bins, creating uniform intervals regardless of data distribution. This straightforward approach is highly interpretable and easy to implement.

Key Characteristics

- Simple and intuitive methodology
- Creates bins of equal width
- May ignore underlying data density
- Can be sensitive to outliers



Reference: Han, J., Kamber, M., & Pei, J. – Data Mining: Concepts and Techniques (3rd Ed.); Witten, I. H., Frank, E., Hall, M. A., & Pal, C. – Data Mining: Practical Machine Learning Tools and Techniques (4th Ed.)

Interval-Based Implementation

01

Initialize Smoother

Create a smoothing object specifying the number of bins

02

Fit to Data

Train the discretizer on your continuous variable

03

Transform Values

Apply the learned intervals to discretize the data

```
obj <- smoothing_inter(n=2)
obj <- fit(obj, datasets::iris$Sepal.Length)
bins <- transform(obj, datasets::iris$Sepal.Length)
```

Full code available at: https://github.com/cefet-rj-dal/daltoolbox/blob/main/transf/dal_smoothing_interval.md

Frequency-Based Discretization

Equal Frequencies

Each bin contains approximately the same number of observations, ensuring balanced representation

Density Adaptive

Automatically adapts to the underlying data distribution and density patterns

Prevents Sparsity

Avoids empty or sparse bins by maintaining consistent observation counts

Handles Skewness

Particularly effective for skewed distributions where interval-based fails

Reference: Han, J., Kamber, M., & Pei, J. – Data Mining: Concepts and Techniques (3rd Ed.); Witten, I. H., Frank, E., Hall, M. A., & Pal, C. – Data Mining: Practical Machine Learning Tools and Techniques (4th Ed.)

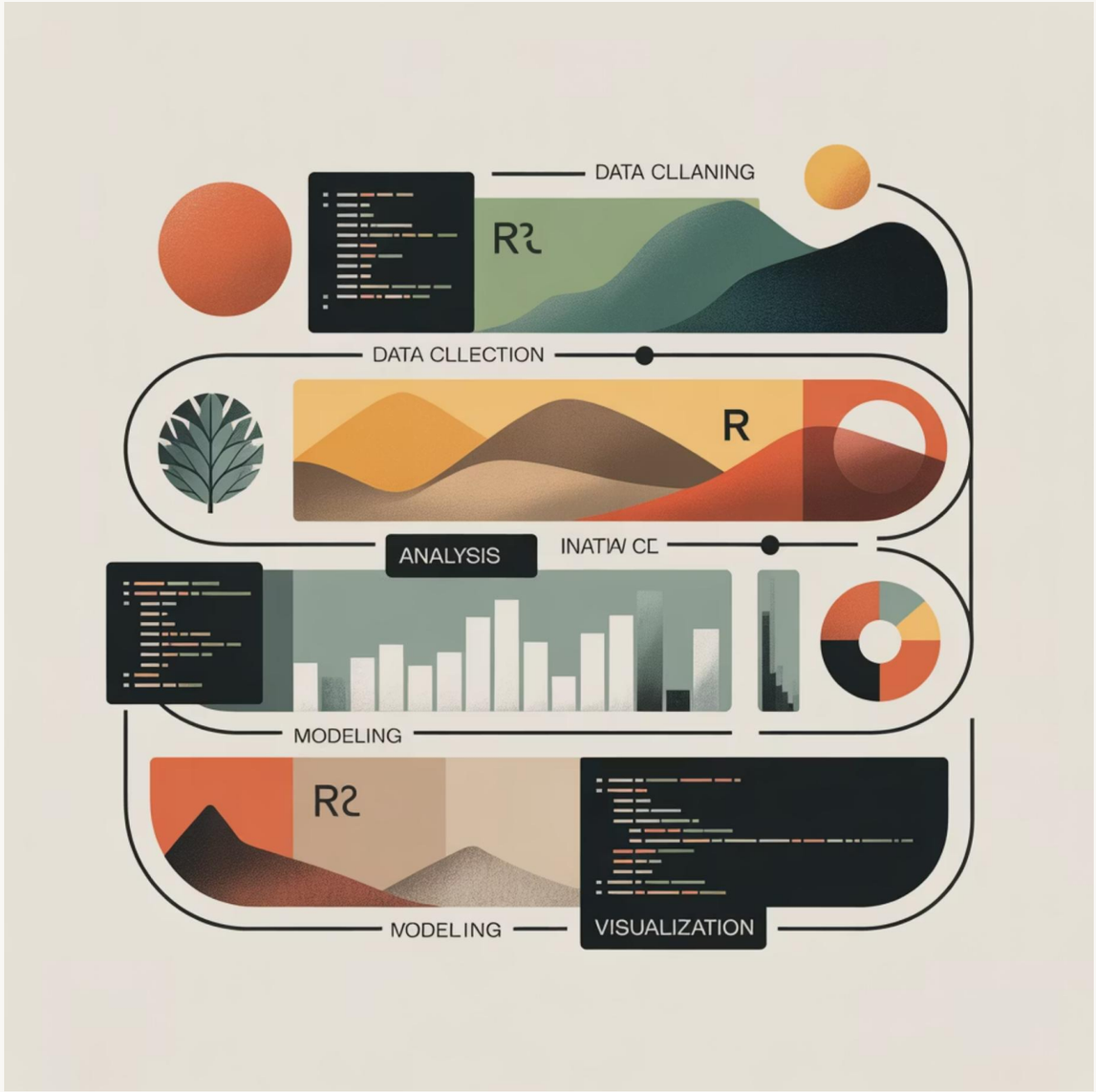
Frequency-Based Implementation

DAL Toolbox Workflow

The frequency-based approach in DAL Toolbox follows the same intuitive fit-transform pattern but creates bins with equal observation counts rather than equal widths.

This method is particularly useful when working with real-world datasets that often exhibit skewed distributions.

```
obj <- smoothing_freq(n=2)
obj <- fit(obj, datasets::iris$Sepal.Length)
bins <- transform(obj, datasets::iris$Sepal.Length)
```



Full code available at: https://github.com/cefet-rj-dal/daltoolbox/blob/main/transf/dal_smoothing_frequency.md

Clustering-Based Discretization

Natural Structure

Uses clustering algorithms to identify natural groupings in the data, capturing inherent patterns and relationships

Highly Adaptive

Adjusts bin boundaries based on actual data distribution rather than predetermined rules or counts

Added Complexity

More sophisticated than simpler methods, requiring additional computational resources and parameter tuning

Initialization Sensitive

Results may vary depending on initial cluster centers, though typically converges to stable solutions

Reference: Han, J., Kamber, M., & Pei, J. – Data Mining: Concepts and Techniques (3rd Ed.); Witten, I. H., Frank, E., Hall, M. A., & Pal, C. – Data Mining: Practical Machine Learning Tools and Techniques (4th Ed.)

Clustering-Based Implementation

Adaptive Discretization in Practice

The clustering-based smoother discovers natural groupings in your continuous data, creating bins that reflect the underlying structure rather than arbitrary divisions.

```
obj <- smoothing_cluster(n=2)
obj <- fit(obj, datasets::iris$Sepal.Length)
bins <- transform(obj, datasets::iris$Sepal.Length)
```

Full code available at: https://github.com/cefet-rj-dal/daltoolbox/blob/main/transf/dal_smoothing_clustering.md

Choosing the Right Method

Interval-Based

Best for uniform distributions and when simplicity is paramount

Frequency-Based

Ideal for skewed data requiring balanced bin populations

Clustering-Based

Perfect when data has natural groupings and structure

DAL Toolbox makes discretization accessible: All three methods share a consistent API with fit and transform functions, enabling seamless experimentation to find the optimal approach for your specific dataset.