**RESEARCH PAPER: BINNING**

* **By Sajith Vellappillil**

***DATA BINNING: Meaning and Uses***

Data Binning (also known as bucketing, discretization, categorization or quantization) simplifies and compresses a column of data by reducing the number of possible values or levels represented in the data*.* It is an essential tool in preparing numerical data for machine learning and is helpful in scenarios like these:

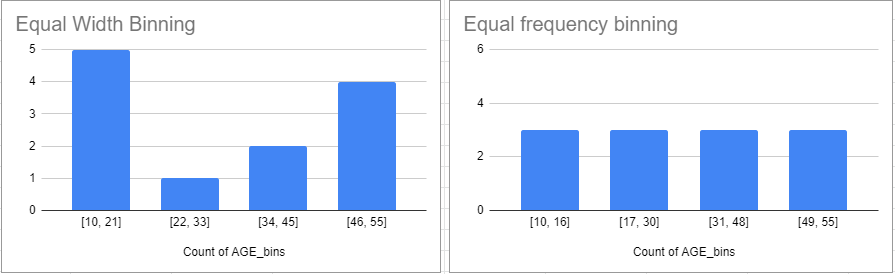
* A column of continuous numbers has too many unique values to model effectively, so you automatically or manually assign the values to groups to create a smaller set of discrete ranges.
* Replace a column of numbers with categorical values that represent specific ranges. For example, you might want to group values in an age column by specifying custom ranges, such as 1-15, 16-22, 23-30, and so forth for user demographics.
* A dataset has a few extreme values, all well outside the expected range, and these values have an outsized influence on the trained model. To mitigate the bias in the model, you might transform the data to a uniform distribution, using the quantiles (or equal-height) method.
* Vehicle for handling missing values (missings are assigned a particular bin)
* Simplification, summarization, and reporting

***TYPES OF BINNING***

* **Unsupervised Binning:** Equal width binning, Equal frequency binning

Unsupervised binning is a category of binning that transforms a numerical or continuous variable into categorical bins without considering the target class label. Unsupervised binning are of two types: Equal Width Binning & Equal Frequency Binning

Example:



* **Supervised Binning:** Entropy-based binning

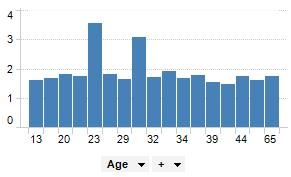
Supervised binning is a binning that transforms a numerical or continuous variable into a categorical variable considering the target class label into account. It refers to the target class label when selecting discretization cut points.

Entropy-based binning is a supervised binning that categorizes the continuous or numerical variable majority of values in a bin or category that belongs to the same class label. It calculates entropy for target class labels, and it categorizes the split based on maximum information gain.

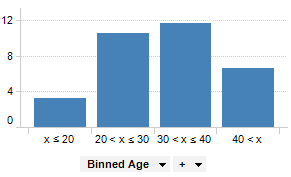
***EXAMPLES***

**Example of binning: continuous data:**

The data table contains information about many persons.

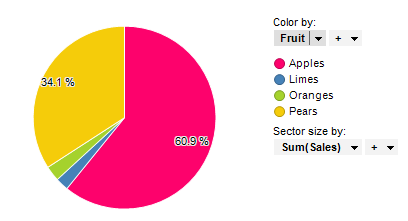


By binning the age of people into a new column, data can be visualized for the different age groups instead of for each individual.

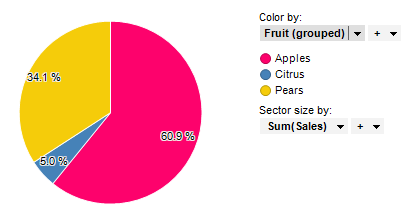


**Example of binning: categorical data**

The pie chart shows sales per apples, limes, oranges and pears.



Below, oranges and limes have been grouped into a bin called “Citrus.”



***REFERENCES***

<https://www.minitab.com/content/dam/www/en/uploadedfiles/content/products/spm/IntroDataBinning.pdf>

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/group-data-into-bins>

<https://docs.tibco.com/pub/spotfire/7.0.1/doc/html/bin/bin_what_is_binning.htm>