* Binning :

Binning (also called bucketing) is a feature engineering technique that groups different numerical subchanges into bins (or) buckets.

→ In many cases, binning turns numerical data into categorical data.

considering an example where the feature X has lowest value 15 and highest value 425. with binning we can represent X as 5 bins :

Bin1 : 15 to 34

Bin 2 : 35 to 117

Bin 3 : 118 to 279

Bin4 : 280 to 392

Bin5 : 393 to 425

Bin1 spans the range 15 to 34, so all value blw them endup in Bin1. A model which is trained on these bins will react no differently to X values of 17 and 29

* Feature Vector :

Bin number	Range	Feature Vector
1	15-34	[1.0, 0.0, 0.0, 0.0, 0.0]
2	35-117	[0.0, 1.0, 0.0, 0.0, 0.0]
3	118-279	[0.0, 0.0, 1.0, 0.0, 0.0]
4	280-392	[0.0, 0.0, 0.0, 1.0, 0.0]
5	393-425	[0.0, 0.0, 0.0, 0.0, 1.0]

Even though X is a single coloumn, binning causes a model to treat X as 5 seperate features there the model learns seperate weights for each bin.

→ It is a good alternative for scaling (or) clipping when:

* The overall linear relationship b/w feature and the label is weak.

* When the feature values are clustered.

* Quantile Bucketing:

Quantile Bucketing creates bucketing boundaries such that the number of examples in each bucket is exactly or nearly equal. It mostly hides outliers.

Note: Bucketing with equal intervals works for many data distributions. For skewed data, try quantile bucketing. Equal intervals give extra information space to the long tail while compacting the large torso into a single bucket.

* Scrubbing : As a ML engineer, the supremacy lies in the data we feed to the model Even a few bad data examples can ruin a large model let us look at some of them :

Problem category	Example
Omitted Values	[A census taken fails to record a resident's age]
Duplicate examples	[A server uploads same logs twice]
out-of range feature Values	[A human accidentally types an extra digit]
Bad labels	[A mislabel of pictures]

When labels are generated by multiple People, we recommend statistically determining whether each rater generated equivalent set of labels.

* Qualities of good numerical features :

Each feature should have a clear, sensible and obvious meaning to any human on the Project.

For ex : house_age : 851472oo ✗

house_age : 27 ✓

Similarly check for values in the dataset before training so that we can remove most of the wrong data and outliers.