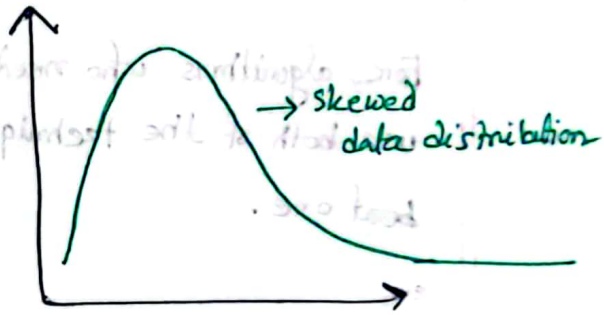
Encoding Numerical Features: (Binning and Binarization)

There are some scenarios where we need to encode our numerical features also.

Suppose, the download feature values of google play store. Let's look at a sample →

Downloads
100
18001012
23
8134
8
49
49000000

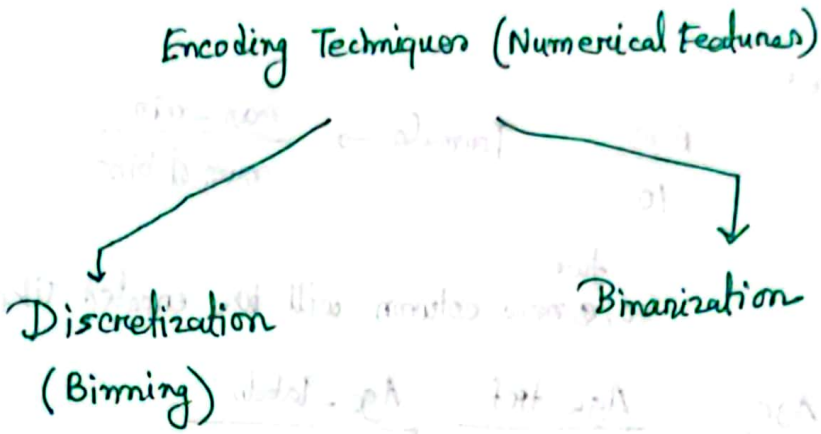
this leads →



We can convert that values into bins to ease our work.

Downloads	Bins
100	100
18001012	10M+
23	23
8134	8K+
8	8
49	49
49000000	100M+

There are two techniques to encode numerical columns →

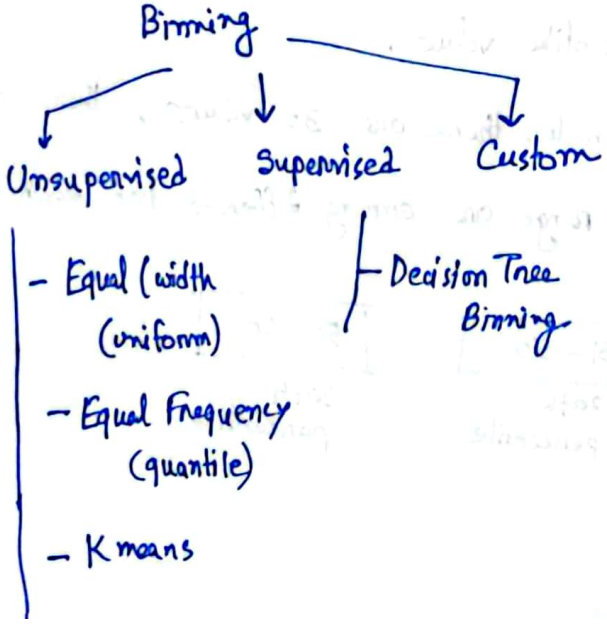


Discretization: The process of transforming continuous variables to discrete variables by creating a contiguous intervals that span the range of the variable values. Discretization is also called binning, where bin is an alternative name for interval.

Why use Discretization?

1. To handle outliers
2. To improve the value spread.

Types:



Equal width/ Uniform Binning:

Age:
27
32
84
56

Bins → 10

Formula → $\frac{\max - \min}{\text{num of bins}}$

Let say, Max = 110, min = 0 } Age

So, two new column will be created like this →

Age	Age_bcf	Age_labels
5	21	0-10
4	21	10-20
21	23	20-30
30	24	30-40

↳ The number of bin (Bin number)

Advantage:

1) Handle outliers 2) No change in spread.

Equal Frequency / Quantile Binning:

Previously the bin range was same for every bin. But here the bin range is dependendent on percentile value.

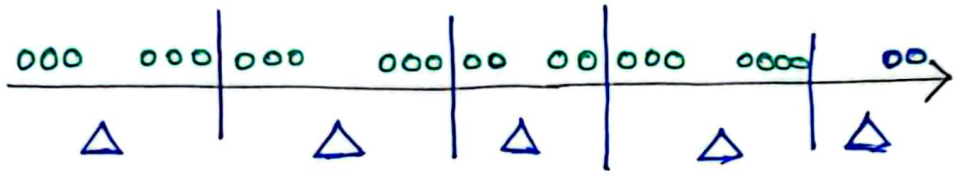
Suppose, for 10th percentile, there are 30 values, then 30 - 46, 20th percentile be like this. So, range are coming different for each bin

0-30	31-50	50-60
10th percentile	20th percentile	30th percentile

Advantage: 1) Handle outliers 2) Make value spread uniform

K-Means Binning: It comes from the clustering algorithm K-means concept.

When you can make clusters in your data, K-means can be used than.



(please refer to the video be it's a bit complex to explain here)

Custom / Domain Based binning: When you have the data and domain knowledge, you can set the number and range for binning by yourself which becomes more useful sometime, That's called Domain Based Binning.

Scikit learn doesn't provide that, we need to create it by ourself using pandas.

Binarization: He we convert a continuous numerical feature to a Binary feature (0,1).

Suppose, we need to find if someone's income is taxable or not.

Amount (LPA)	Taxable
6LPA →	1
5LPA →	1
3.2LPA →	0
9LPA →	1

Threshold = 3.5 LPA