Handling Numerical Data

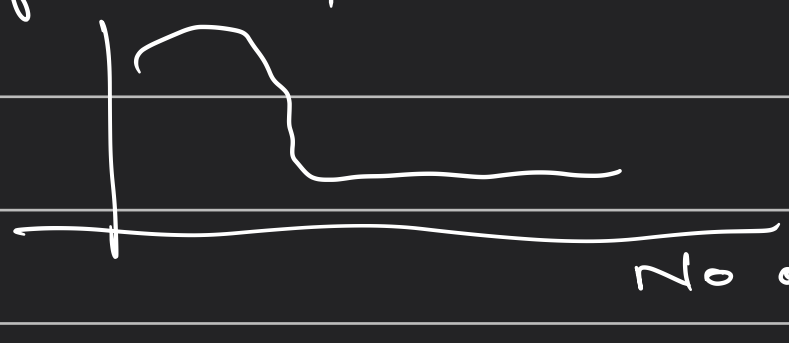
→ Encoding Numerical Data

Why Encode Numerical to Categorical?

Say For Ex you have columns like No of downloads in playstore.

few Apps has 100+ downloads, few have 1000+, few have 10000+. 100000 etc

So now if we plot this we will have graph



Now this is sometimes troublsome so we can create Bins where based upon Bin condition we can add our data accordingly

1st Bin → 100+ 2nd → 1000+

3rd → 100000+ etc etc.

Two famous Techniques to convert Numerical data to Categorical

1) Discretization which is called Binning

2) Binarization

Discretization or Binning

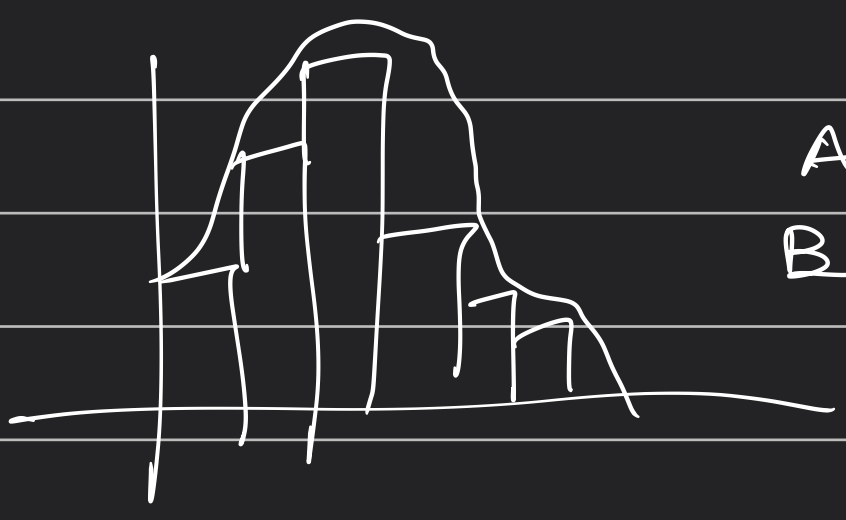
process of transforming continuous value

into discrete variables by creating set of continous intervals that span on range of variable value.

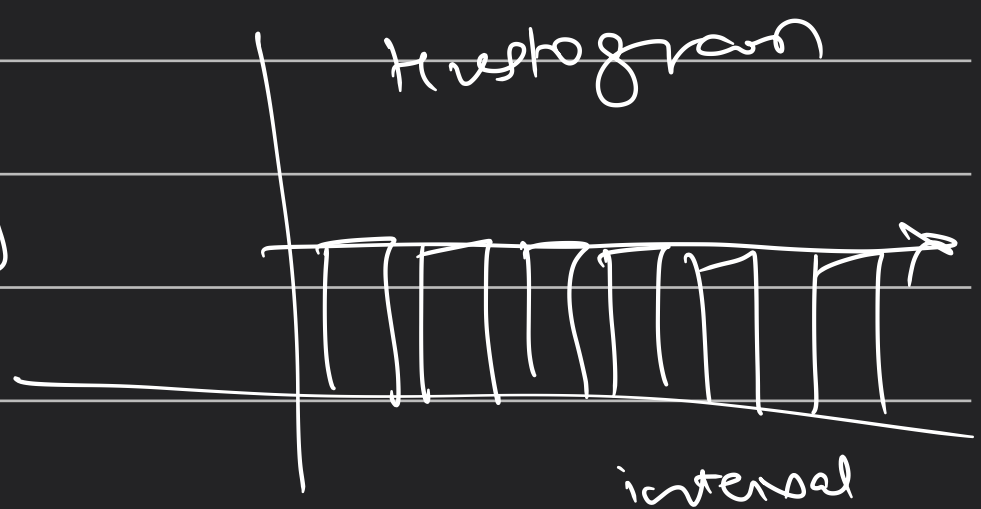
Why Discretization is called Binning?
→ when intervals can be called as Bins for storing continues variable we call it binning.

Advantage of Binning

⇒ Say we have a age distibution 0 - 100 years & for that our histogram is like this



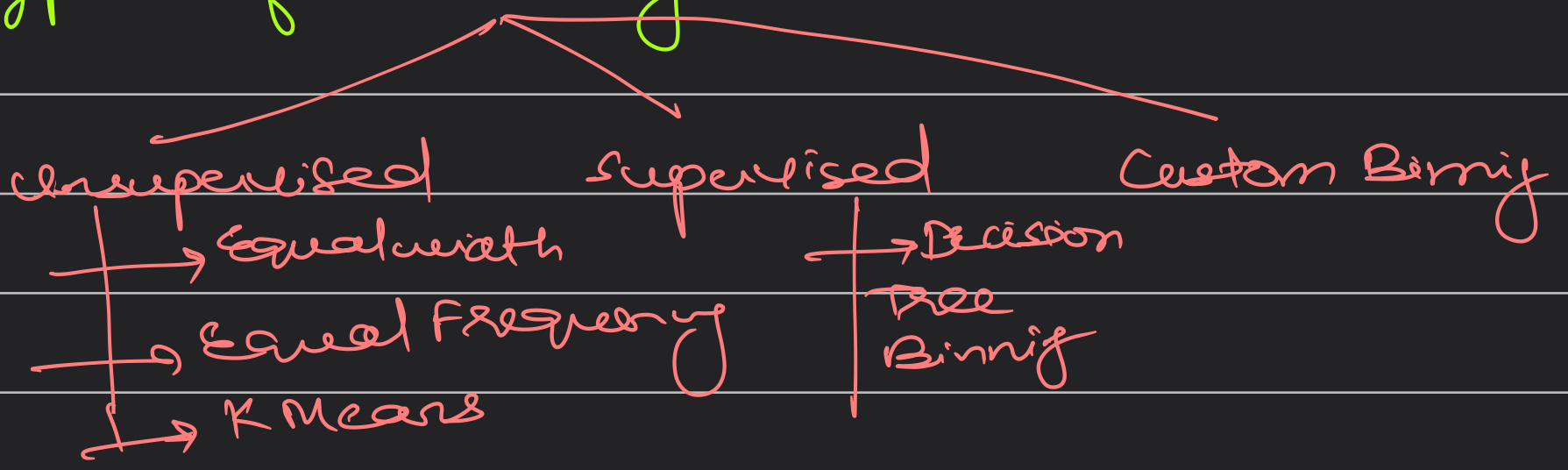
After Binning



Adv ⇒ 1) we handle outliers
2) Improve value spread

outliers will come under 1st or last bin accordingly

- Types of Binning



Unsupervised Binning

- Equal width or Uniform Binning

We give No of Bins upfront

$$\text{Interval range} = \frac{\text{Max} - \text{Min}}{\text{No of Bins}}$$

$$\text{Ex} : \frac{100 - 0}{10} = 10$$

Intervals would be

1 2 3 4 5
0-10, 10-20, 20-30, 30-40, 40-50,
6 7 8 9 10
50-60, 60-70, 70-80, 80-90, 90-100

total 10 bins

So for a particular variable/column values we will put in related bin accordingly.

Somewhat

Now outliers gets handled easily, histogram remains same as no change in spread.

- Equal Frequency / Quantile Binning

We give no of Bins upfront.

Say 10

Now each interval or bin will contain 10% of total observation

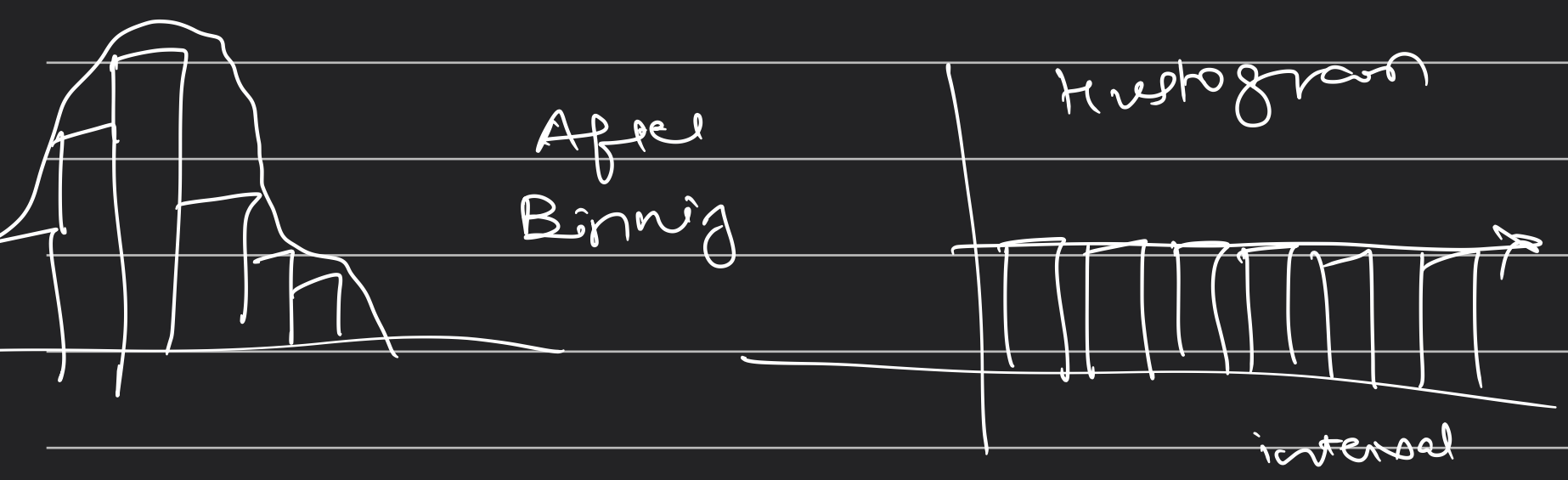
Value

Interval would be 0 to 10th percentile
But 10% data would be near
(is not percent)

0 - 16, 16 - 20, 20 - 22 etc
(10th, 20th percentile value)

Here the spread gets changed, performs well on outliers, here the spread becomes uniform

Most used Binning Method.



- K Means Binning

used on clustered data & works on the fundamentals of K-means Clustering algorithm.

Read more when you study K Means Clustering Algorithm.

What after this?

After the process we do Encoding using ordinal or nominal based upon relation in variable values

- Custom Based Binning

we create bins on the basis of our domain knowledge.

Ex → Age we created bins like

[0 - 18] → Young

[18 - 60] → Adult

[60 +] → Senior

- Binarization

Is a special case of binning, here we convert continous value to Binary [0 or 1]

Ex Annual Income

If Income < 100,000K → 0

Income > 100,000K → 1

At this value no tax

Taxing

Mostly used in Image Processing.