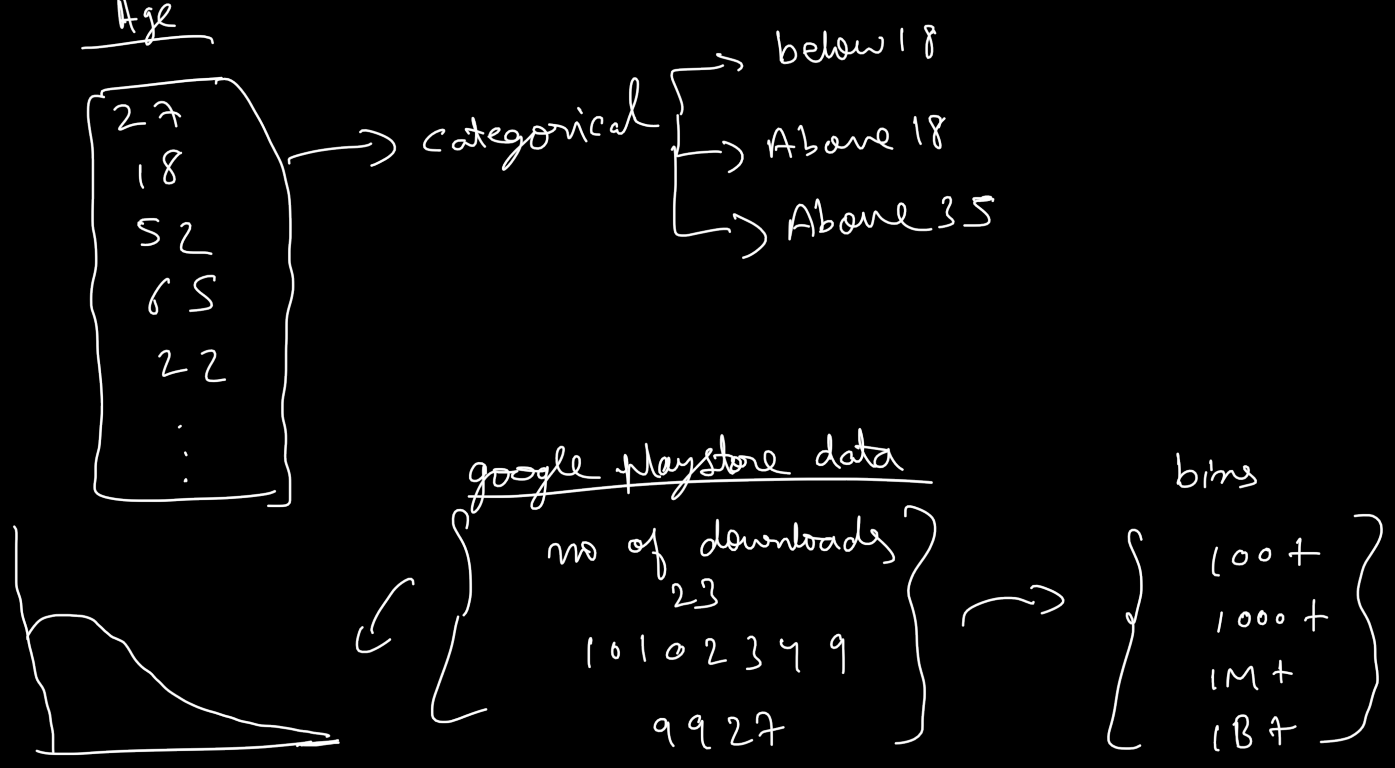
Binning and Binarization
Discretization | Quantile Binning | KMeans Binning

Encoding Numerical features ⇒ categorical



Discretization (Binning)

Discretization is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals that span the range of the variable's values. Discretization is also called, where bin is an alternative name for interval.

→ Why use Discretization

(1) To handle outliers

(11) To improve the value Spread

Types of Discretization

- Binning
  - Unsupervised
    - Equal width (uniform)
    - Equal frequency (quantile)
    - KMeans binning
  - Supervised
    - Decision Tree Binning
  - Custom Binning

Equal width / uniform Binning

Age

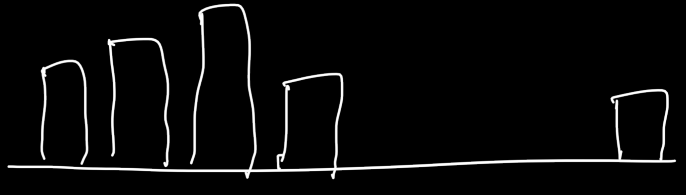
27, 32, 84, 56, ...

max 100
min 0

Bins = 10

$$\frac{max - min}{bins} = \frac{100 - 0}{10} = 10$$

(0 - 10), (10 - 20], (20 - 30] . . . . (90 - 100)



(i) helps in handling outliers

(ii) No change in Spread of data

Equal Frequency / quantile Binning

Intervals = 10

Each interval Contains 10% of total observations

Interval:

0-16; 16-20; 20-22; 22-25; 50-74

* This is default value in SKlearn class.

(i) used mostly (better than Previous one)

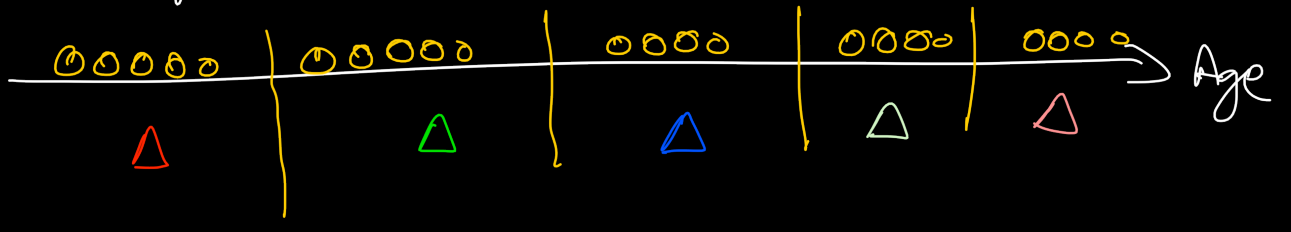
(ii) helps in handling outliers

(iii) Make value Spread uniform

KMeans Binning

It is used when data is distributed in clusters.

1) Calculate distance of each point from each centroids and make cluster with nearest centroid.



2) Now again centroid of each cluster change according to mean of values inside it and then again repeat Step 1 and make cluster.

3) we will repeat this until there is no change in previous and Current Cluster.

4) And finally After getting centroid from final cluster we take centroid value, which is the bin value for our binning.

Encoding the discretized variable

SK Learn
↳ KBinsDiscretizer()

- bins
- Strategy
  - uniform
  - quantile
  - Kmeans
- encoding
  - ordinal
  - one hot encoding

Custom / Domain Based Binning

* According to domain knowledge, we decide our bins or category like

[0 - 18] — Kidy
[18 - 60] — Adult
[60 - 80] — old

Binarization

SKlearn Binarizer class

- Threshold
- Copy = True / False