Sometimes, we may need to convert our data from numerical to categorical. I know this sounds very strange, but it is possible that our data would be better represented when it is categorical. For example, if we have a dataset of app downloads from the Google Play Store, we might encounter some data where values are in millions, while others are in two digits. To handle such scenarios, we categorize the data into bins. For instance, we might transform values into categories such as "100+ downloads," "200+ downloads," "500K downloads," and "1M downloads." This process is called binning.

There are 2 techniques which we can use on our numerical data to convert it into categorical:

- **Discretization/Binning**
- **Binarization**

Discretization or 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 binning, where bin is an alternative name for interval.

For Example: It means we have a column representing ages with multiple age values, and I want to convert it into discrete categories. What we would do is we'll create intervals like [0-10], [10-20], and so on, and this is what discretization is.

Why use Discretization:

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

Types of Discretization

1. **Supervised Discretization**
    a. Equal Width (Uniform)
    b. Equal Frequency (Quantile Binning)
    c. K-Means Binning
2. **Unsupervised Discretization**
    a. Decision Tree Binning
3. **Custom Discretization**

1 - Supervised Discretization

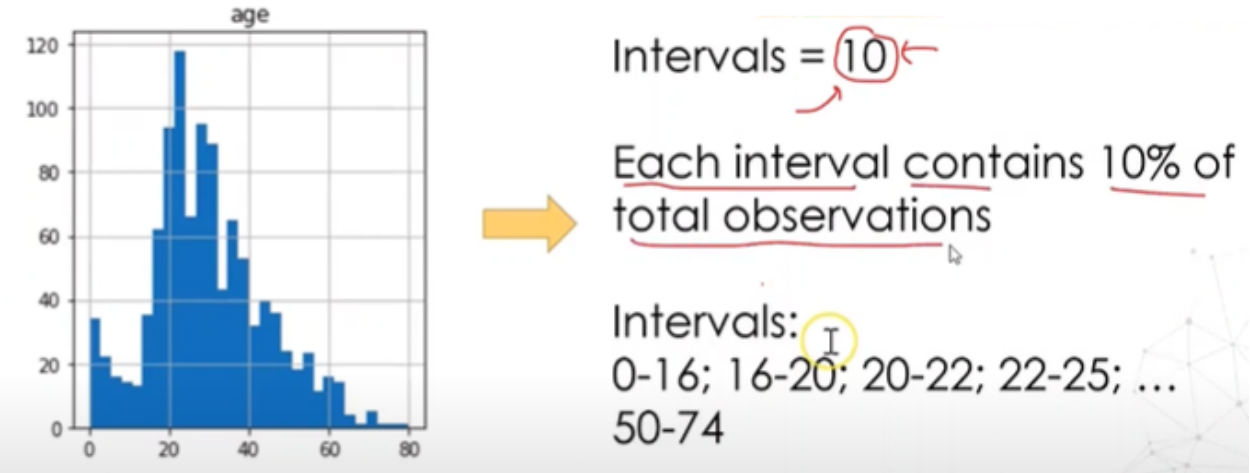
a. **Equal Width (Uniform):** we have a column representing ages with multiple age values, and I want to convert it into discrete categories. What we would do is we'll create intervals like [0-10], [10-20], and these intervals are made by the certain formula which is = max-min/bins so that's how this works on, and this is what Equal Width discretization is.

For Example:

	age	age_trf	age_labels
314	43.0	5.0	(40.21, 48.168]
523	44.0	5.0	(40.21, 48.168]
352	15.0	1.0	(8.378, 16.336]
534	30.0	3.0	(24.294, 32.252]
211	35.0	4.0	(32.252, 40.21]
530	2.0	0.0	(0.42, 8.378]
786	18.0	2.0	(16.336, 24.294]
827	1.0	0.0	(0.42, 8.378]
372	19.0	2.0	(16.336, 24.294]

The Benefits we get by using this technique is we can tackle outliers mostly and second the spread of data would not get changed it will remain the same as it was before.

b. **Equal Frequency or Quantile Binning**



10 Bins = 10 Quantiles, this means if we are making 10 intervals that means we have to make 10 quantiles as well, means we will go from 0 to that value in where our 10th percentile is lying and goes on. It is being used most because it works on outliers as well and it does not change the spread of data.

C. K-Means Binning

In this, we use a K-means clustering algorithm. Basically, this algorithm shifts our data into clusters regardless of the dimensions of our data. And we use this technique when our data is in the form of clusters.

3 - Custom Discretization

- Custom discretization, in the context of data analysis and machine learning, refers to the process of dividing continuous data into discrete intervals or bins based on custom criteria or domain-specific knowledge.

Note: I have skipped the 2nd Type will update it later!