**Encoding Numerical Features (FE)**

Techniques of Encoding Numerical features to categorical data:

1. Discretization (Binning)
2. Binarization
3. **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 of interval.

**Why uses Discretization?**

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

e.g: age = {10,23,25,………80} => {10-20:4, 21-40:5, 50-80:3} this is Discretization

Types of Discretization (Binning):

1. Unsupervised Binning
   * + - 1. Equal width (uniform)
         2. Equal Frequency (quantile)
         3. K-means
2. Supervised
   * + - 1. Decision tree binning
3. Custom Binning
4. **Equal width**: you decide the number of bins in this technique. Let say we want 10 bins

E.g: Bins = 10, formula = max – min / bins , if max = 100 and min = 0 => 100-0/10 = 10

Now we have 10 bins => 0-10,10-20,20-30 ……… .

It is useful for handling the outliers and it doesn’t change the spread of data.

1. **Equal Frequency:** you decide the number of bins let say 10 bins(interval), in this technique Each bin contains 10% of the total observation.

**e.**g 0-10 contains the 10% of the total population means 10th percentile and 10-20 contains 20% of the population means 20th percentile and so on.

it is use useful for handling outliers and it convert the spread of data uniform.

1. **K-means**: this technique we do clustering. First of all, we will choose the **number of (bins) clusters (k)**(In this case **k=2**). What this means is that now we will assume **2 points** randomly, they will act as our **cluster centroids**. (**Cluster centroid** is the **center point**of a **cluster**).

There are two main steps in **K Means Clustering**:

* **Cluster Assignment Step**: In this step, the data points that are close to the centroids will fall in those centroids’ clusters respectively.
* **Move Centroid Step**: In this step, we will compute the mean of all data points in a cluster and move the centroid of that cluster to that mean position.

We will **repeat**the **above two steps** once one of the **following** conditions is **true:**

1. Our centroids stop changing their positions.
2. Maximum number of iterations are reached.

Our **data** is now arranged into**clusters.**

This technique is used when your data is in form of clusters.

**Note**: Scikit library contain the class called

**KBinDiscretizer**(bins = ?, strategy = (uniform|quantile|k-means), encoding = (ordinal | onehotencoding))

**Handling Mixed Variables (FE)**

**Mixed Data**: your column may contain numerical and categorical data like in cabin column of titanic data. E.g (B45,C23). To handle this type of data we can create two columns one for numerical data and one for categorical data.

**Handling Date and Time (FE)**

Normally when you see the data type of data or time column in pandas it shows object data type. Its your duty to first convert it to date type.

Now Extract year , month and day for date column and put them into separate columns.