1. **Info function**
2. Null/ None and NaN are diff..NaN means not a number. Empty string and NaN is not same
3. Empty string (“ “)ke object hishebe treat korte hbe (j clmn a error dibe, oikhne ekta for loop chalate hbe) For example:

For i in range (...title…):

Impitution: suppose missing value “Nan” k value diye replace kora

So use mean/ median/ mod

Mean/median problem:

Interpolation:

20,nan,40

So number should be 30, means close to 20 and 40

Scalar:

Min Max standard

Suppose salary 20k\_\_\_\_150k

Exp (0—10) so 0 to 1

So ekhne 0 to 1 na hishab kore scale 5-10 o deowa jbe

Robust scaling:

Min max er diff onk beshi thakle tahole outliers a scaling process fix rakhar jonno

1. **How to solve problem:**

**#**  Jodi same value thake onk gulate tahole drop korbo, Dataframe.drop(“feature names ”,axis=1) clmn axis=1, row er jonno axis= 0

**#** Column drop is not always a good idea, j sob row te values nai sudhu oigula drop kora better (check how many values are missing in the category\_desc column.

Example: (volunteer\_subset= volunteer[volunteer[‘category\_desc’].notnull.sum()]

**#** column drop

**#**simple\_imputer….etake select kore dite hobe kon value replace korte chai. Imputer agee data sob nibe then oita dekhe NaN gula ke replace korbe

#Impute fit diye imputer function k basically ready kore

Column specify kora better.

1. **Preprocessing using zero mean**

So ekhne amra scaler.fit(x\_train) diye next dui line a direct transform kora hye jacche..ekhne testing lagtese na, karon we don’t need to learn new things from testing set.cz we already know.

**One hot encoding:**

**Level encoder= num of count**

**# get.dummies** use korle coplexicity bere jay

Nominal a non binary thakle binary kora jbe

There can be biasness issues

Jodi complexicity onk bere jay tokhon level encoder use kora jbe

**ordinal=** karom order bujhte pari easily

**Lamda raw** function use korle easily 5 ta row er kaj ekta row te kora possible (***To get time*** function use kore )

**Correlation matrix**

# Koto ta correlately kaj kore oita dekhe.

Function name: **corr**

#Basically duita diff topic er sathe relation khuje.

#High means two values are so close

#Weak correlation means difference is so high

Worst er moddhe best jodi na pai kichu

#But eta khuje ber kora tough, so import ***seaborn*** for color output so that we can understand and find out easily

Jokhon amra dekhbo multiple feature ek sathe tokhon sob gula consider kora important na.

Suppose we have 5 columns and we have one target, here all x vs y will be considered, input and output feature relation,

Here if A and B showcase high corr and B and C high corr then which one will be removed? My target is high corr features. But we want less corr between input. So j kono ekta rakhlei hbe in this case as both are giving me high corr.

**# categorical value agee handle korte hobe**