Categorical Data

How to deal with Categorical data?

- machine learning algorithms cannot work directly with categorical data(discontinuous data) and you do need to do some amount of engineering and transformations on this data before you can start modelling on your data.
- It's very important to convert the categorical data to not miss important information.
- I will explain advanced techniques on dealing with categorical data when I start Kaggle series.
- Most used techniques are line label encoding, one-hot encoding, converting numerical categorical data to strings and again to numerical(Zipcode), Replace Values, dummy encoding.

Categorical Data

- **Replace values:** This is the basic method like replacing the categorical values with some matching numerical like A->1, B->2 etc....
- **Label Encoding:** Another approach is to encode categorical values with a technique called "label encoding", which allows you to convert each value in a column to a number. Numerical labels are always between 0 and n_categories-1. it also has a disadvantage that the numerical values can be misinterpreted by the algorithm as a Ranking.
- **One-Hot encoding:** One problem with the Label encoding is they weight the categories like ranking. so here what we do is like we create the columns equal to the categories and if that category belongs to that column we give 1 and all other columns have 0 value. which also leads to a curse of dimensionality problem
- **Let's discuss more advanced in kaggle series.**