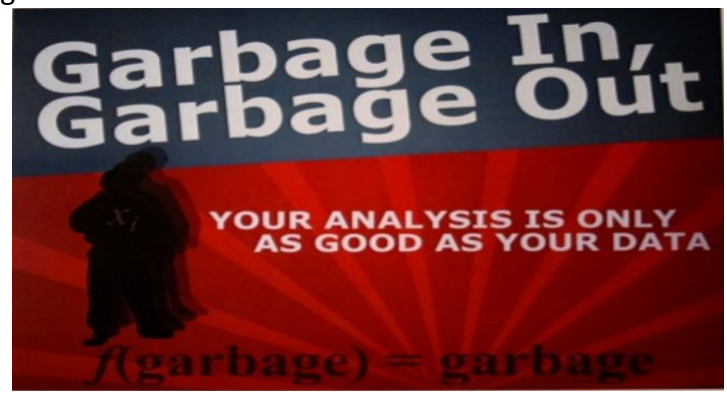
Data Preprocessing for Non-Techies:

"**Data is the new gold**" and everybody wants data in its purest form. But in the real-world scenario it is almost impossible to attain clean data as the sources for data collection are different, sometimes data-set has missing data, and at other times data is inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Also the convention used while collecting data is different for different databases. Such data is called ***Dirty.***



Using dirty data to resolve problems leads to misleading results. As it is rightly said, ***"garbage in garbage out".*** So to tackle this situation of ***Dirty*** data we require data pre-processing.

So, what is Data Preprocessing ?

- **Data preprocessing** is a **data** mining technique that involves transforming raw **data** into an understandable format.

And what are some of the basic strategies that data scientists use to clean their data and improve the amount of information they get from it?

- The type of cleaning and engineering strategies used usually depend on the business problem and type of target variable, since this will influence the algorithm and data preparation requirements.
- The most important part of data cleaning is the experimentation and checking how applying one or many of these strategies affects your ability to predict or classify in the model.
- Given below is a list of practices widely used for data exploration and data cleaning:

Data Exploration

A. Variable Identification:

- ✓ Context of Target Variable (logical connection)
- ✓ Data Type per Feature (character, numeric, etc)
- ✓ Variable Category (Continuous, Categorical, etc.)

B. Uni-variate Analysis:

- ✓ Central Tendency & Spread for Continuous
- ✓ Distribution(levels) for Categorical

C. Bi-variate Analysis:

- ✓ Correlation of Continuous Variables
- ✓ Two-Way Table or Stacked Columns for Categorical

Data Cleaning

A. Remove Noise:

- ✓ Duplicates
- ✓ Paragraph Columns
- ✓ Erroneous Values
- ✓ Contradictions
- ✓ Mislabels

B. Missing Values:

- ✓ Delete
- ✓ Mean/Mode/Median Imputation
- ✓ Prediction Model
- ✓ KNN Imputation

C. Outliers:

- ✓ Assign Weights
- ✓ Cut-Off or Delete
- ✓ Natural Log
- ✓ Binning
- ✓ Mean/Mode/Median Imputation
- ✓ Build Predictive Model
- ✓ Treat them separately

By now, you must have got a rough idea of what methods of preprocessing data are used in the industry. Generally, a combination of above techniques are used in iterative way. Also, there are many more strategies that at times give fruitful results. But don't worry, you are going to learn everything step by step.