Getting to know about your data:

Well, Let's ask yourself a Question? How important is to know about the data you will work on?

To answer this Question, It is very important to learn about data before directly jumping into Data Preprocessing. It is essential to have an overall picture of your data. There are various attributes/ features that give a lot of information about the dataset. To find a missing value in the dataset or to smoothen noise we often need to do some statistical operations like finding the mean, median, mode, standard deviation etc. These are the measures of central tendency which gives us an idea of central distribution. Data sets are made of data objects that are described by attributes.

**Naming Conventions:**

* *Data objects can also be referred to as samples, examples, instances, data points, or objects.*
* *The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature.*
* *Fields, features, variables are used for columns.*
* *Samples, observations, entries are used for rows.*

There are 4 different types of attributes:

1. **Nominal attribute**

Nominal attributes are generally the name of things that do not have any impact on the target class. We can convert these attributes into numeric attributes by assigning a unique value to each, but should not perform any mathematical operations on Nomaial attributes. This is because let's say we are working on a College dataset where student\_id is one of the fields, We cannot add or subtract or apply ‘mean’ to the student\_id field. However, we can find the most occurrences by finding mode as a central tendency.

1. **Binary Attribute:**

Binary Attribute is also a nominal attribute with only two categories typically 0 or 1, Yes or No, True or False. By convention, We code the rarest one as ‘1’ and other by ‘0’.

1. **Ordinal Attribute:**

An ordinal attribute is an attribute with possible values that have a meaningful order or

ranking among them, but the magnitude between successive values is not known. Even these look like nominal attributes but not. For example, customer satisfaction has the following categories; *extremely satisfied, satisfied, neutral, dissatisfied, Very dissatisfied*.

1. **Numeric Attributes:**

Numeric attributes is a measurable quantity represented in real numbers. The numeric attribute is quantitative and has a high order to the target variable. These can be interval scaled or ratio scaled. This is also known as a continuous attribute.

**Statistical Operations on DaTa**:

Why do we need to perform statistical operations on data? Since the data we gather will be from diverse sources, there is much possible to have missing values in your dataset. There are two ways we can deal with missing values. We either have to drop all the missing values or fill all the null with some value. If you go for dropping all the Null values, we might end up with losing some valuable information in other fields. So, it is always encouraged to fill all the missing values.

**What should be my value to fill in?**

Best option researchers apply is “**finding the central tendency of the data**”. This could be **Mean, Median, Mode, Standard Deviation, Variance** etc,

The most effective measure is Arithmetic Mean.

Sometimes each value in x is associated with weights w. Arithmetic mean for this is given by:

Where i = 1,2,3,4,…………… N

This is called weighted **Arithmetic mean or Weighted Average**.

Of course, mean is used the most to fill missing values but it's not always the best. Let's take a small example of having 5 employees in a startup firm. The salary for the three high ranked employees are 100k$, 90k $ and 80K$ and the least ranked employee is 10k$. If you are asked to find the approximate salary for the top 4th employee, applying ‘mean’ to this gives outlier values, that is the mean will be close to the majority class. Instead, we may need to use the trimmed mean for this kind of problems.

**Steps to find trimmed mean:**

1. Sort out the values in ascending or descending order.
2. Chop out the top and bottom 5% values.
3. Now, find the mean value and fill all null values. (Python – fillna() can be used to fill all Null values of a field).

***Source: DATA MINING Concepts and techniques – Jiawei Han| Micheline Kamber | Jian Pei***