How will be the flow of this project?

* We have two datasets; choose so that I may cover all the concepts and issues. So both will be used for different purposes.
* we know that for achieving the requirements of this project like graphing techniques along with usage and understanding the usage of those graph’s results **we need numeric data.** So for this we will convert columns into numeric data by using python.
* So 1st of all we will perform some pre-processing on both datasets separately and will try to identify the different types of pre-processing issues and their solution buy using python. While doing pre-processing we will also perform some steps which make data standardized. In the result of pre-processing we will get the two new CSV files, one with non-standardized values and other with standardized values, which will be processed further for rest of operations.
* For the purpose of graph to visualize and understanding of the data using boxplot, histogram, qq-plot, scatter plot and etc, we will use the dataset 1. We will also analyze the effects of using standardized data and non-standardized data.
* For the purpose of Apply machine learning algorithm for finding frequent patterns, classification, clustering we will make use of both pre-processed dataset files.
* In last we will make the confusion matrix which will show the accuracy and cross validation scores, incase of clasfication algorithems, and other interesting findings which will be gathered by changing the parameters for algorithems.

**NOTE: I have used two datasets in this assignment. The reason for choosing different datasets is to cover as many as concepts of data mining.**

**Dataset 1 (big-mart-sales) source:** <https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data>

**Dataset 2 (startup-success-prediction) source:** <https://www.kaggle.com/manishkc06/startup-success-prediction>

**Whole code used for this project and other resources:** ( <https://github.com/ZeeWING-Projects/Data-Mining-Operations> )

**Tools used:** Spider IDE (For using python)

# Pre-processing

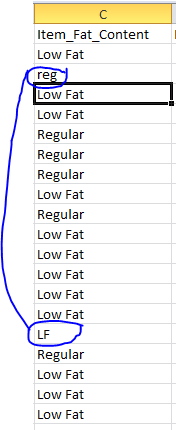
## Dataset 1 (big-mart-sales):

## 1. Applying some pre-processing steps (Data set 1):

First we need to apply some pre-processing techniques before we process it.

### 1.1 Matching field’s values:

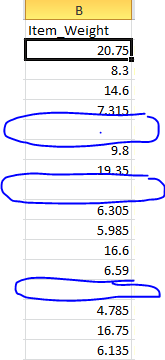
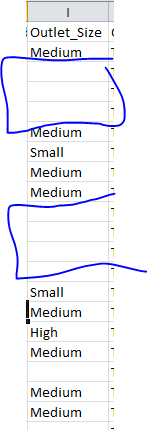
There are few fields in dataset which contain same data with different names, so we need to make the same. For example we have attribute item\_Fat\_Content which contain two labels, Low Fat and Regular, but for representing this same value “LF” and “reg” are used so we need to remove these shortcuts.



### 1.2Filling out missing values:

In our dataset we have a bunch of attributes having missing values. And we have to fill them by using well known pre-processing techniques. For example for numeric attribute we have methods like by using median, mean and mode and for ordinal attributes we will use some built in functions of python (which use this same mode technique).

For example we have some attributes with missing values.

### 1.4Implementation of above stated issues

So we have a real dataset having some missing values and some other issues so we need to perform above stated steps of pre-processing

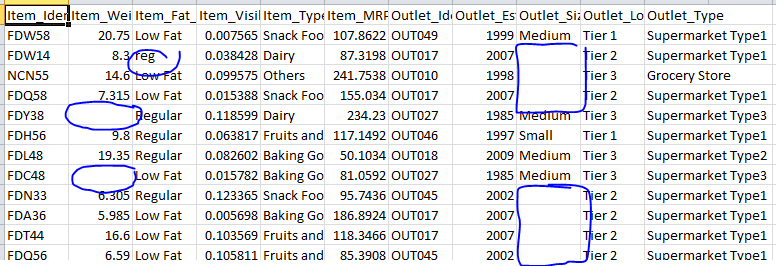
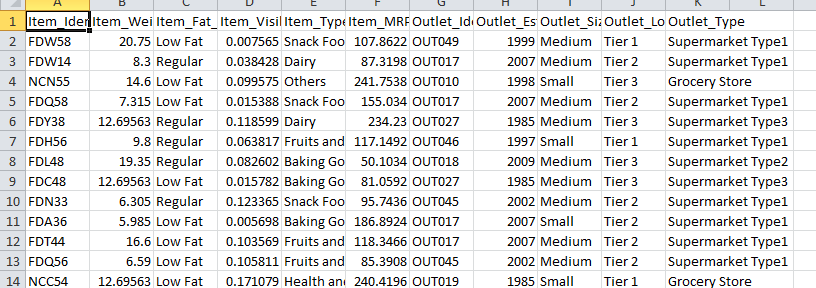
Realdataset:(<https://github.com/ZeeWING-Projects/Data-Mining-Operations/blob/main/RealDataSet/RealDataSet.csv> )

Pre-processed-dataset:(<https://github.com/ZeeWING-Projects/Data-Mining-Operations/tree/main/Pre-processed%20DataSet> )

Pre-processing-code:(<https://github.com/ZeeWING-Projects/Data-Mining-Operations/blob/main/Preprocessing-Code/Preprocessing.py>)



Dataset comparison

### 1.5 Conversion of string values to numeric values.

## Dataset 2 (startup-success-prediction):