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. In last we have select the features (will drop some columns on the basic of two assumptions that either they are unnecessary or there some other strongly co-related columns).
* 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, in case of classification algorithms, 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/DM-Project> )

Note: For the purpose of results used in this document I have run provided code in different ways, by running some specific parts sometimes whole code at a time. So you might need to un comment some code and comment some code to get those results.

# Pre-processing

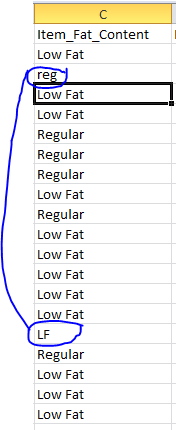
## Dataset 1 (big-mart-sales):

## 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:

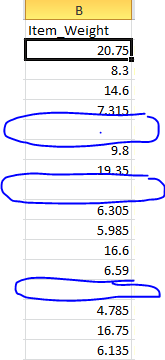
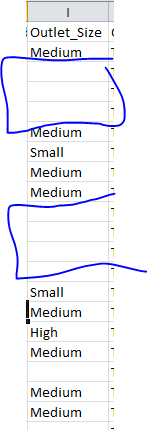
There are few fields in dataset which contain same data with different names, so we need to make the same. For example we have attribute item\_Fat\_Content which contain two labels, Low Fat and Regular, but for representing this same value “LF” and “reg” are used so we need to remove these shortcuts.



### 1.2Filling out missing values:

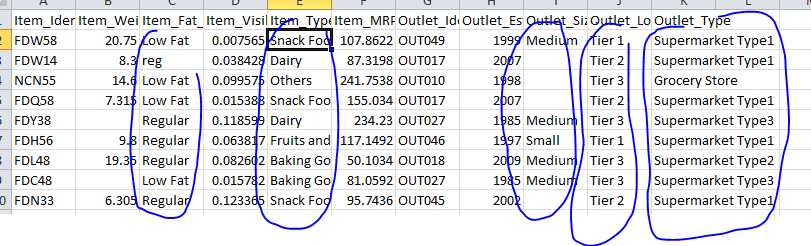
In our dataset we have a bunch of attributes having missing values. And we have to fill them by using well known pre-processing techniques. For example for numeric attribute we have methods like by using median, mean and mode and for ordinal attributes we will use some built in functions of python (which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

For example we have some attributes with missing values.

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

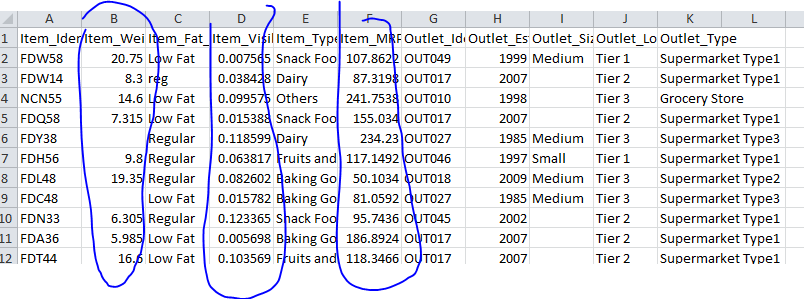
We have some values which are in string form so we need to transform them in numeric form. Since we have following values which need to be in numeric form.



As you can see clearly that these are some columns which are like normal values but some are showing that those can be used as class label. So we are assuming that feature named **Outlet\_Type will be used for classification purpose**. So it will be translated accordingly function.

### 1.5 Conversion of normal values to standardized values.

We have some values which are stated in 100s unit and some are 1s unit. What I meant is that as we have following values.



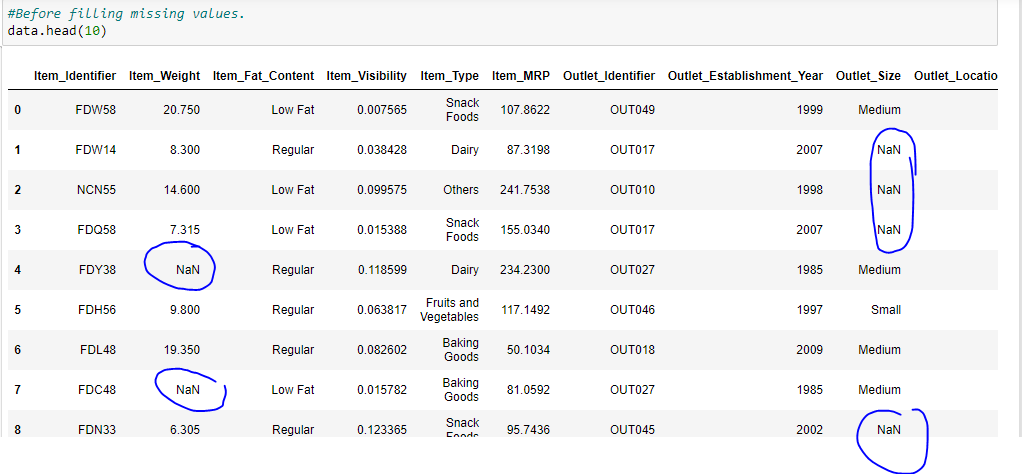
As highlighted values are those one which need to be standardized because this will affect the co-relation graph, which will be drawn later on.

### 1.5Implementation of above stated issues

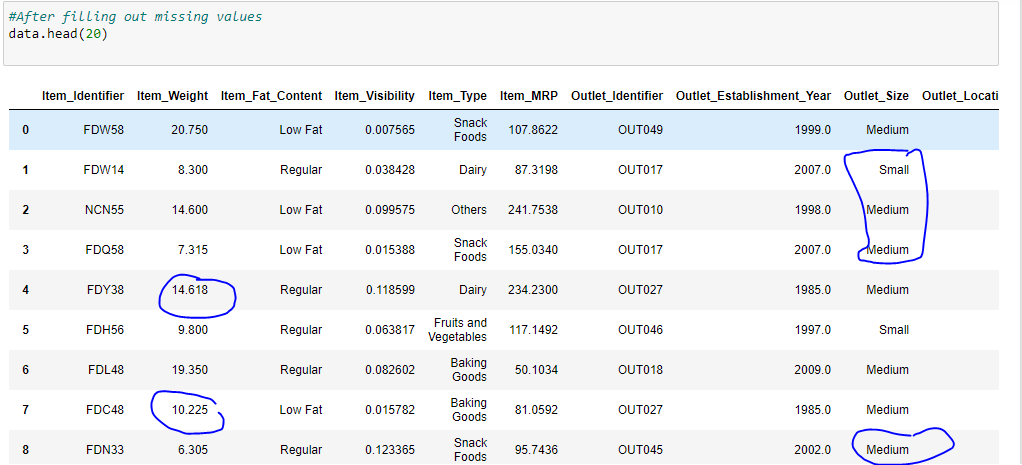
So we have a real dataset having some missing values and some other issues like **conversion of ordinal values to numeric form** so we need to perform above stated steps of pre-processing.

Find Code at :(<https://github.com/ZeeWING-Projects/DM-Project/blob/main/Preprocessing-Code/Preprocessing.py> )

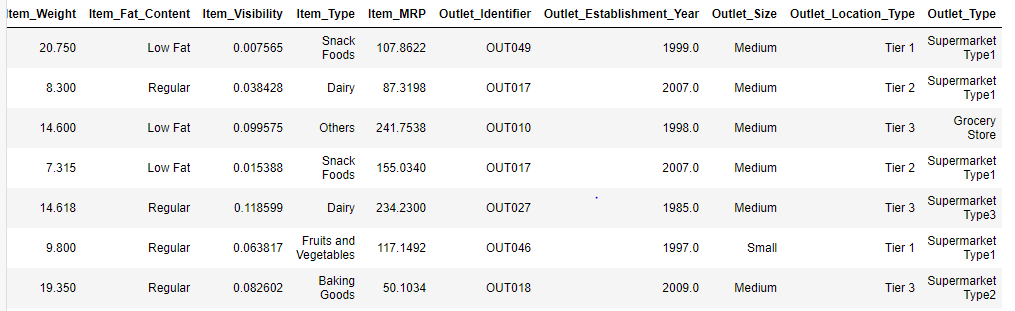
#### Before filling missing values



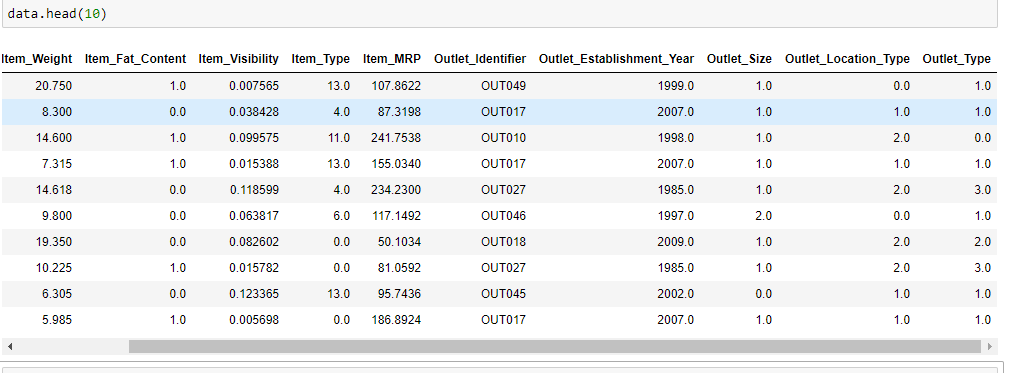
#### After filling missing values



#### Before converting string values to numeric form

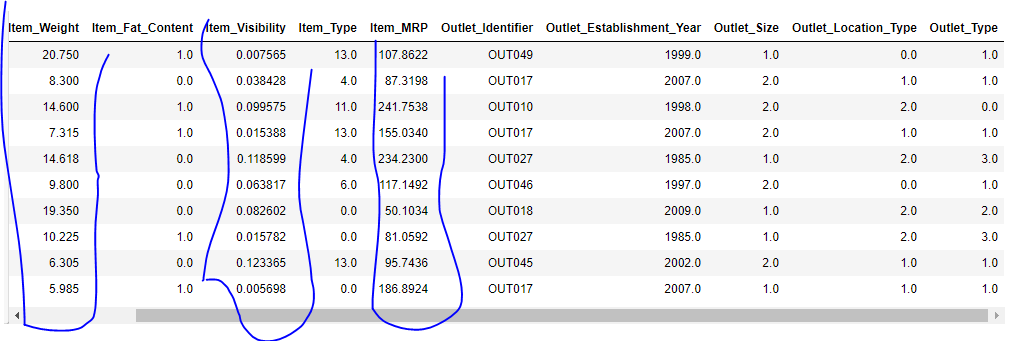


#### After converting string values to numeric form

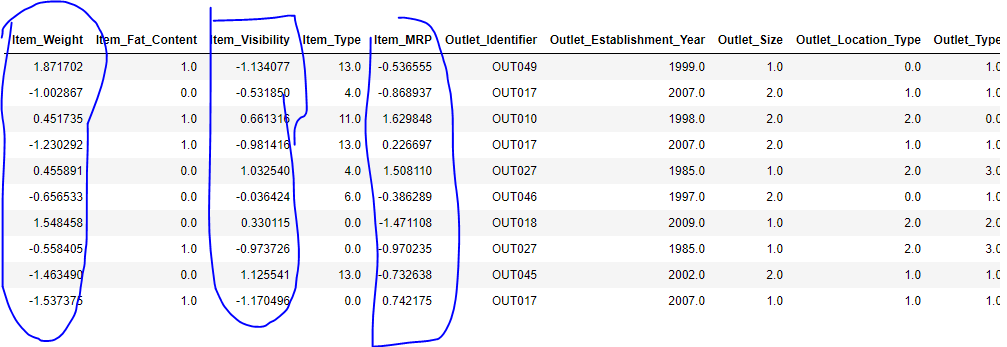


#### Before standardization of values

Since we don’t standardize the labels: (<https://www.google.com/search?q=should+we+standarize+the+lablel+values&oq=should+we+standarize+the+lablel+values+&aqs=chrome..69i57j33i10i160l4.16858j0j4&sourceid=chrome&ie=UTF-8> )

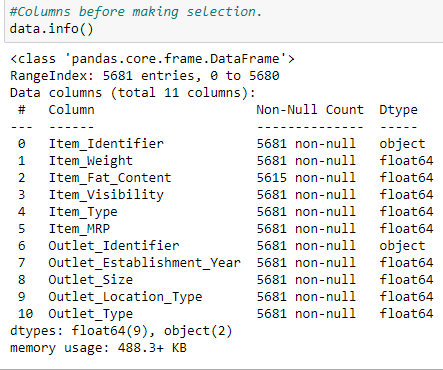


#### After standardization of values



#### Now making selection of features

#### Columns before selection



#### Columns after selection

We will drop two columns which are just showing the IDs.



**So from this we can say that there is no column is strongly related with any other so we don’t need to remove any column.** Because these will perform important role In classification.

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