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

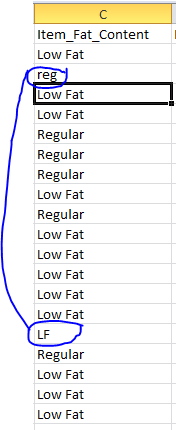
## Dataset 1 (big-mart-sales):

## 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:

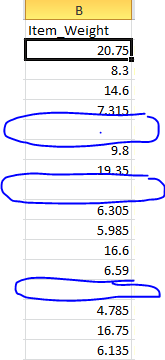
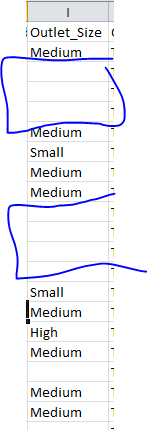
There are few fields in dataset which contain same data with different names, so we need to make the same. For example we have attribute item\_Fat\_Content which contain two labels, Low Fat and Regular, but for representing this same value “LF” and “reg” are used so we need to remove these shortcuts.



### 1.2Filling out missing values:

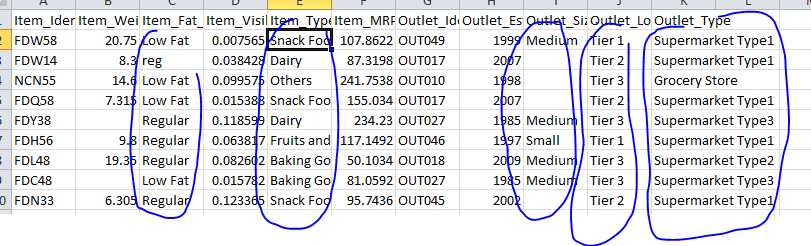
In our dataset we have a bunch of attributes having missing values. And we have to fill them by using well known pre-processing techniques. For example for numeric attribute we have methods like by using median, mean and mode and for ordinal attributes we will use some built in functions of python, like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

For example we have some attributes with missing values.

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

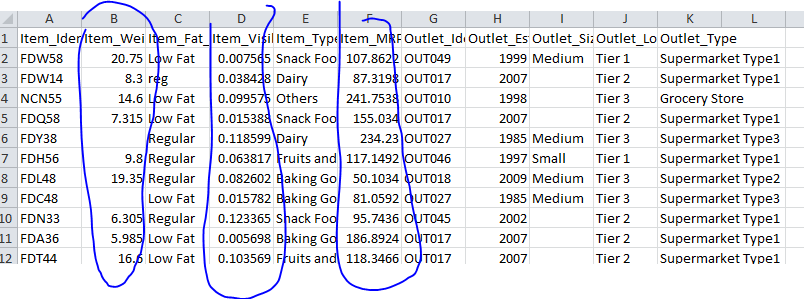
We have some values which are in string form so we need to transform them in numeric form. Since we have following values which need to be in numeric form.



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.



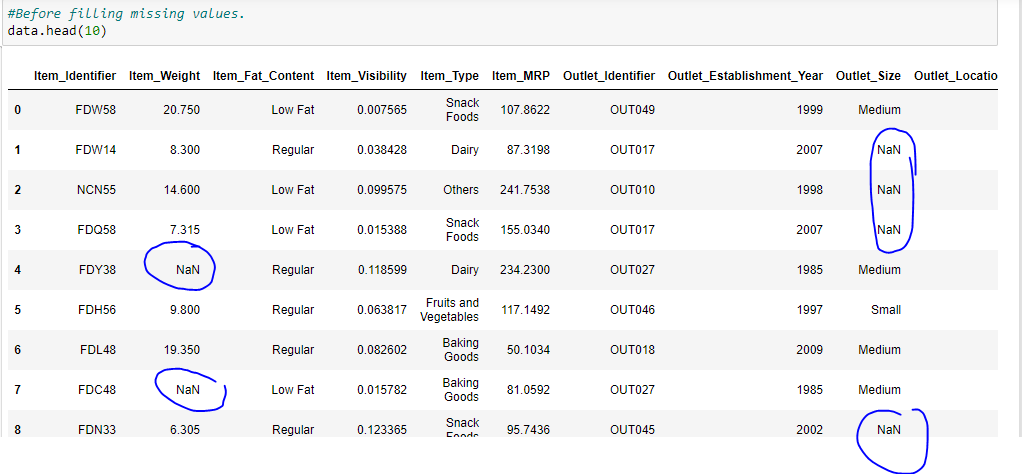
As highlighted values are those one which need to be standardized because this will affect the co-relation graph, which will be drawn later on.

### 1.5Implementation of above stated issues

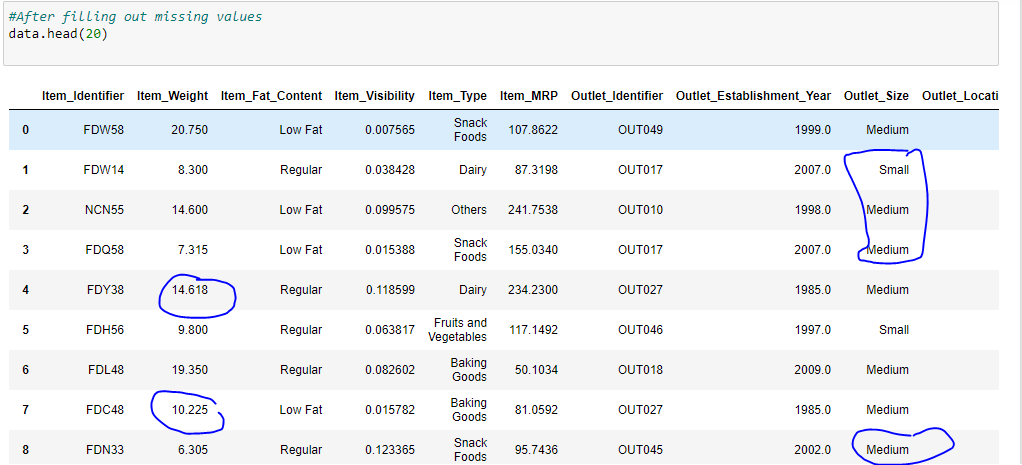
So we have a real dataset having some missing values and some other issues like **conversion of ordinal values to numeric form** so we need to perform above stated steps of pre-processing.

Find Code at <https://github.com/ZeeWING-Projects/DM-Project/blob/main/Preprocessing-Code/Dataset-1_pre-processing.py> )

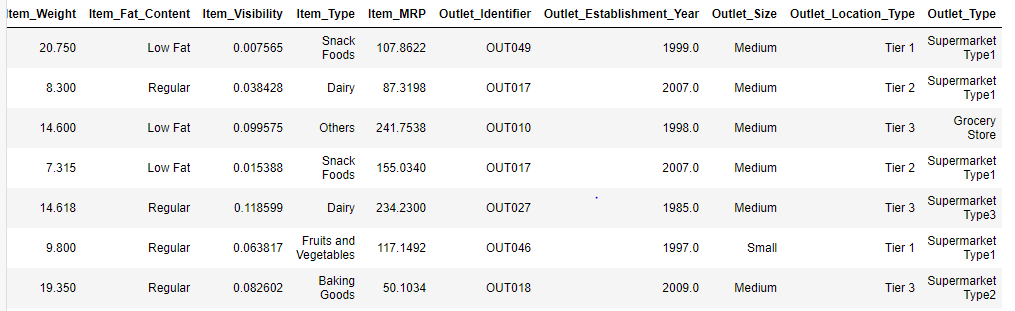
#### Before filling missing values



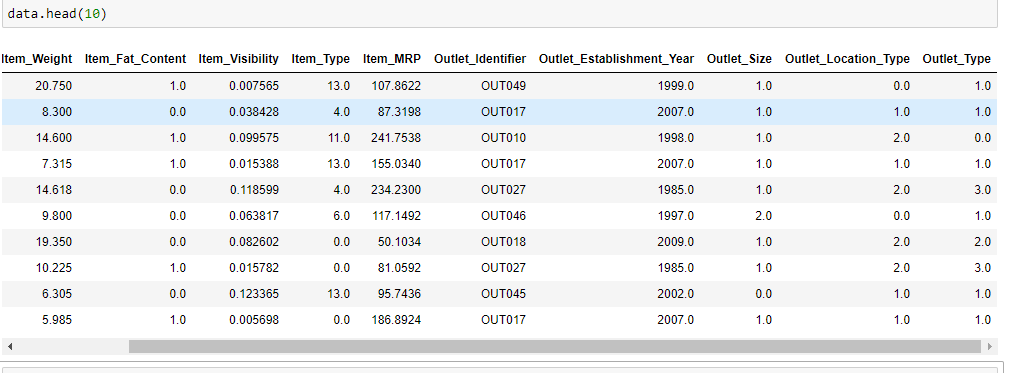
#### After filling missing values



#### Before converting string values to numeric form

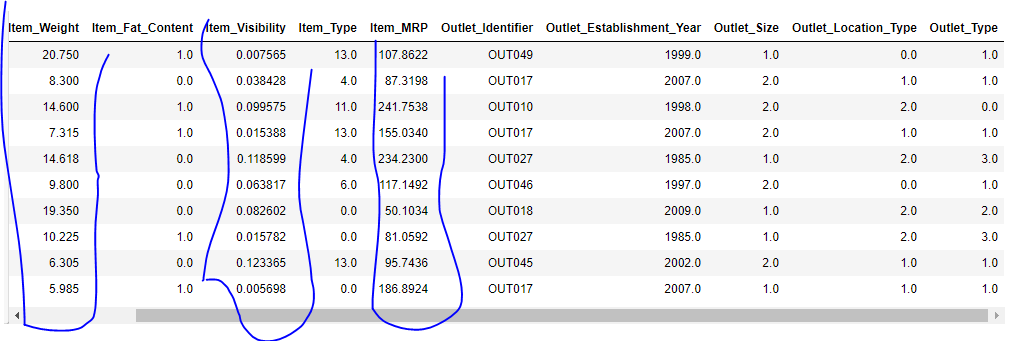


#### After converting string values to numeric form

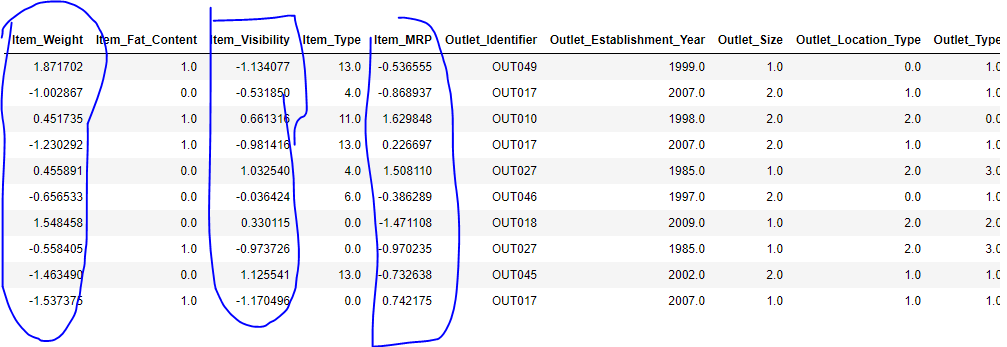


#### Before standardization of values

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> )

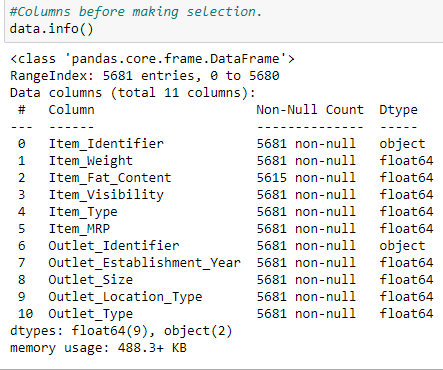


#### After standardization of values



#### 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 (standardized values) :** [**https://github.com/ZeeWING-Projects/DM-Project/blob/Pre-procssing/Dataset-1%20Pre-processed/Dataset\_01\_standarized\_.csv**](https://github.com/ZeeWING-Projects/DM-Project/blob/Pre-procssing/Dataset-1%20Pre-processed/Dataset_01_standarized_.csv)

**Dataset (non-standardized values) :** [**https://github.com/ZeeWING-Projects/DM-Project/blob/Pre-procssing/Dataset-1%20Pre-processed/Dataset\_01\_non\_standarized\_.csv**](https://github.com/ZeeWING-Projects/DM-Project/blob/Pre-procssing/Dataset-1%20Pre-processed/Dataset_01_non_standarized_.csv)

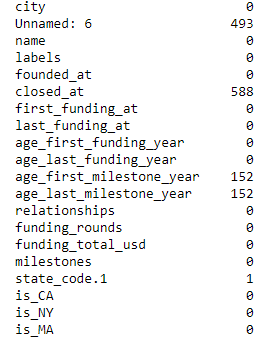
## 2. Applying some pre-processing steps (Data set 2):

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

### 2.2Filling out missing values:

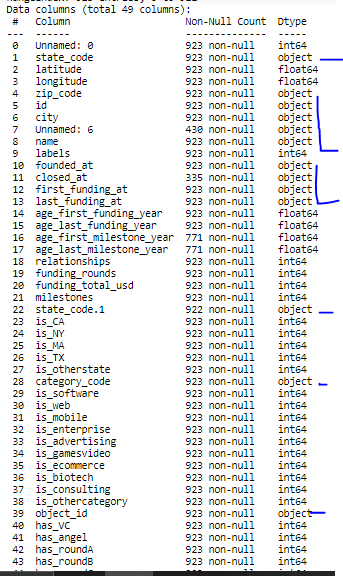
In our dataset we have a bunch of attributes having missing values. And we have to fill them by using well known pre-processing techniques. For example for numeric attribute we have methods like by using median, mean and mode and for ordinal attributes we will use some built in functions of python, like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

For example we have some attributes with missing values. As there 49 columns so I am just showing by using this form of result.



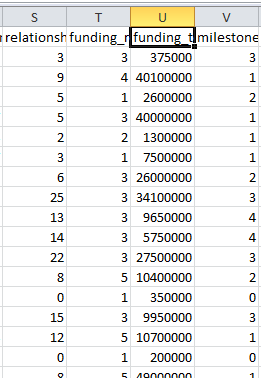
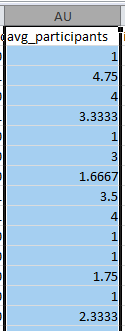
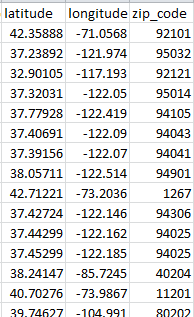
### 2.2 Conversion of string values to numeric values.

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.



### 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.

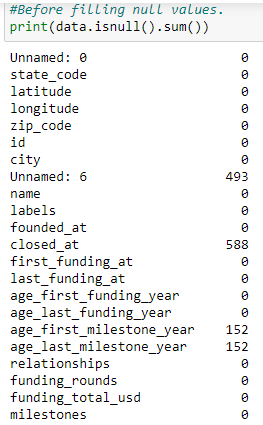
As highlighted values are those one which need to be standardized because this will affect the co-relation graph, which will be drawn later on. And we will have some other columns aswell which will be in numeric form after conversion.

### 1.5Implementation of above stated issues

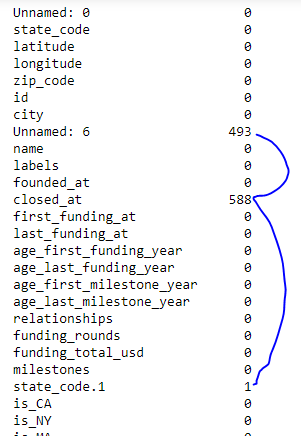
So we have a real dataset having some missing values and some other issues like **conversion of ordinal values to numeric form** so we need to perform above stated steps of pre-processing.

Find Code at :( <https://github.com/ZeeWING-Projects/DM-Project/blob/main/Preprocessing-Code/Dataset-2_pre-processing.py> )

#### Before filling missing values

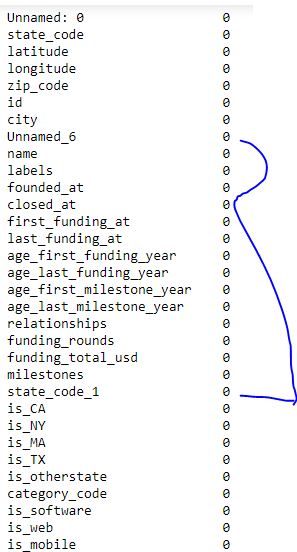


#### After filling missing values



Since you can observe un named: 6 and closed\_at are showing still null values. Reason is that I have used KNN Inputter which works for numbers. So we will try an other method for this.

#### After filling missing values (after 2nd try)



**Now we have successfully removed all null values.**

#### Before converting string values to numeric form

Data columns (total 49 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 923 non-null float64

1 state\_code 923 non-null object

2 latitude 923 non-null float64

3 longitude 923 non-null float64

4 zip\_code 923 non-null object

5 id 923 non-null object

6 city 923 non-null object

7 Unnamed\_6 923 non-null object

8 name 923 non-null object

9 labels 923 non-null float64

10 founded\_at 923 non-null object

11 closed\_at 923 non-null object

12 first\_funding\_at 923 non-null object

13 last\_funding\_at 923 non-null object

14 age\_first\_funding\_year 923 non-null float64

15 age\_last\_funding\_year 923 non-null float64

16 age\_first\_milestone\_year 923 non-null float64

17 age\_last\_milestone\_year 923 non-null float64

18 relationships 923 non-null float64

19 funding\_rounds 923 non-null float64

20 funding\_total\_usd 923 non-null float64

21 milestones 923 non-null float64

22 state\_code\_1 923 non-null object

23 is\_CA 923 non-null float64

24 is\_NY 923 non-null float64

25 is\_MA 923 non-null float64

26 is\_TX 923 non-null float64

27 is\_otherstate 923 non-null float64

28 category\_code 923 non-null object

29 is\_software 923 non-null float64

30 is\_web 923 non-null float64

31 is\_mobile 923 non-null float64

32 is\_enterprise 923 non-null float64

33 is\_advertising 923 non-null float64

34 is\_gamesvideo 923 non-null float64

35 is\_ecommerce 923 non-null float64

36 is\_biotech 923 non-null float64

37 is\_consulting 923 non-null float64

38 is\_othercategory 923 non-null float64

39 object\_id 923 non-null object

40 has\_VC 923 non-null float64

41 has\_angel 923 non-null float64

42 has\_roundA 923 non-null float64

43 has\_roundB 923 non-null float64

44 has\_roundC 923 non-null float64

45 has\_roundD 923 non-null float64

46 avg\_participants 923 non-null float64

47 is\_top500 923 non-null float64

48 status 923 non-null object

#### After converting string values to numeric form

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 923 non-null float64

1 state\_code 923 non-null float64

2 latitude 923 non-null float64

3 longitude 923 non-null float64

4 zip\_code 923 non-null float64

5 id 923 non-null float64

6 city 923 non-null float64

7 Unnamed\_6 923 non-null float64

8 name 923 non-null float64

9 labels 923 non-null float64

10 founded\_at 923 non-null float64

11 closed\_at 923 non-null float64

12 first\_funding\_at 923 non-null float64

13 last\_funding\_at 923 non-null float64

14 age\_first\_funding\_year 923 non-null float64

15 age\_last\_funding\_year 923 non-null float64

16 age\_first\_milestone\_year 923 non-null float64

17 age\_last\_milestone\_year 923 non-null float64

18 relationships 923 non-null float64

19 funding\_rounds 923 non-null float64

20 funding\_total\_usd 923 non-null float64

21 milestones 923 non-null float64

22 state\_code\_1 923 non-null float64

23 is\_CA 923 non-null float64

24 is\_NY 923 non-null float64

25 is\_MA 923 non-null float64

26 is\_TX 923 non-null float64

27 is\_otherstate 923 non-null float64

28 category\_code 923 non-null float64

29 is\_software 923 non-null float64

30 is\_web 923 non-null float64

31 is\_mobile 923 non-null float64

32 is\_enterprise 923 non-null float64

33 is\_advertising 923 non-null float64

34 is\_gamesvideo 923 non-null float64

35 is\_ecommerce 923 non-null float64

36 is\_biotech 923 non-null float64

37 is\_consulting 923 non-null float64

38 is\_othercategory 923 non-null float64

39 object\_id 923 non-null float64

40 has\_VC 923 non-null float64

41 has\_angel 923 non-null float64

42 has\_roundA 923 non-null float64

43 has\_roundB 923 non-null float64

44 has\_roundC 923 non-null float64

45 has\_roundD 923 non-null float64

46 avg\_participants 923 non-null float64

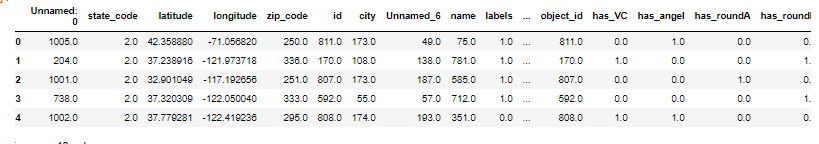
47 is\_top500 923 non-null float64

48 is\_acquired 923 non-null int64

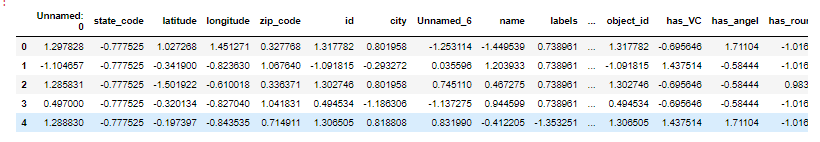
**Now we have successfully converted the string value to numeric form.**

#### 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

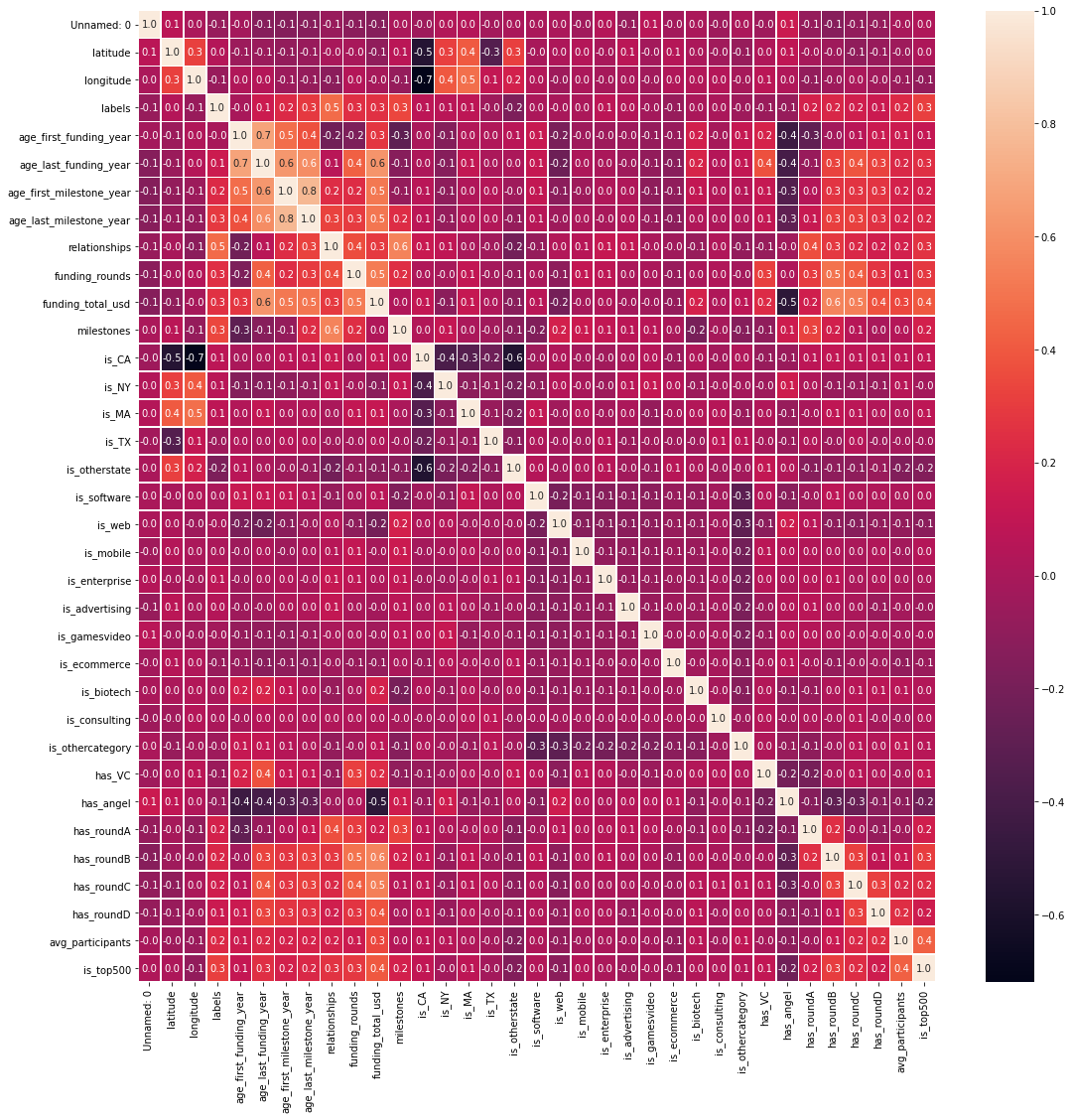


#### Now making selection of features

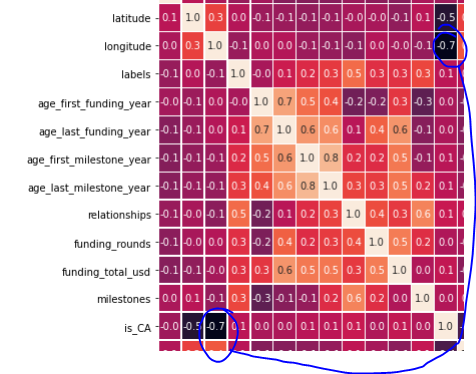
#### Columns before selection

So we have two reasons to drop a column from dataset. Either that is un necessary like it may be the id of some column and other reason can be that there exist some other columns which are highly co-related to it, due to that we can remove all and can keep only one column.

Here we have that graph of co-relation.



we have summed that for strong relation threshold is 0.7>= so we have will have few columns which need to be remove due to strong relation. Following will be removed.

* 
  + See there is strong co-relation between age\_first\_funding\_year and last\_funding\_year so we need to remove either of them. I am removing first\_funding\_year.
* 
  + See there is strong co-relation between longitude and is\_CA so we need to remove either of them. I am removing longitude.
* 
  + See there is strong co-relation between age\_first\_milestone\_year and age\_last\_milestone\_ so we need to remove either of them. I am choosing age\_last\_milestone\_

# Now we will again remove

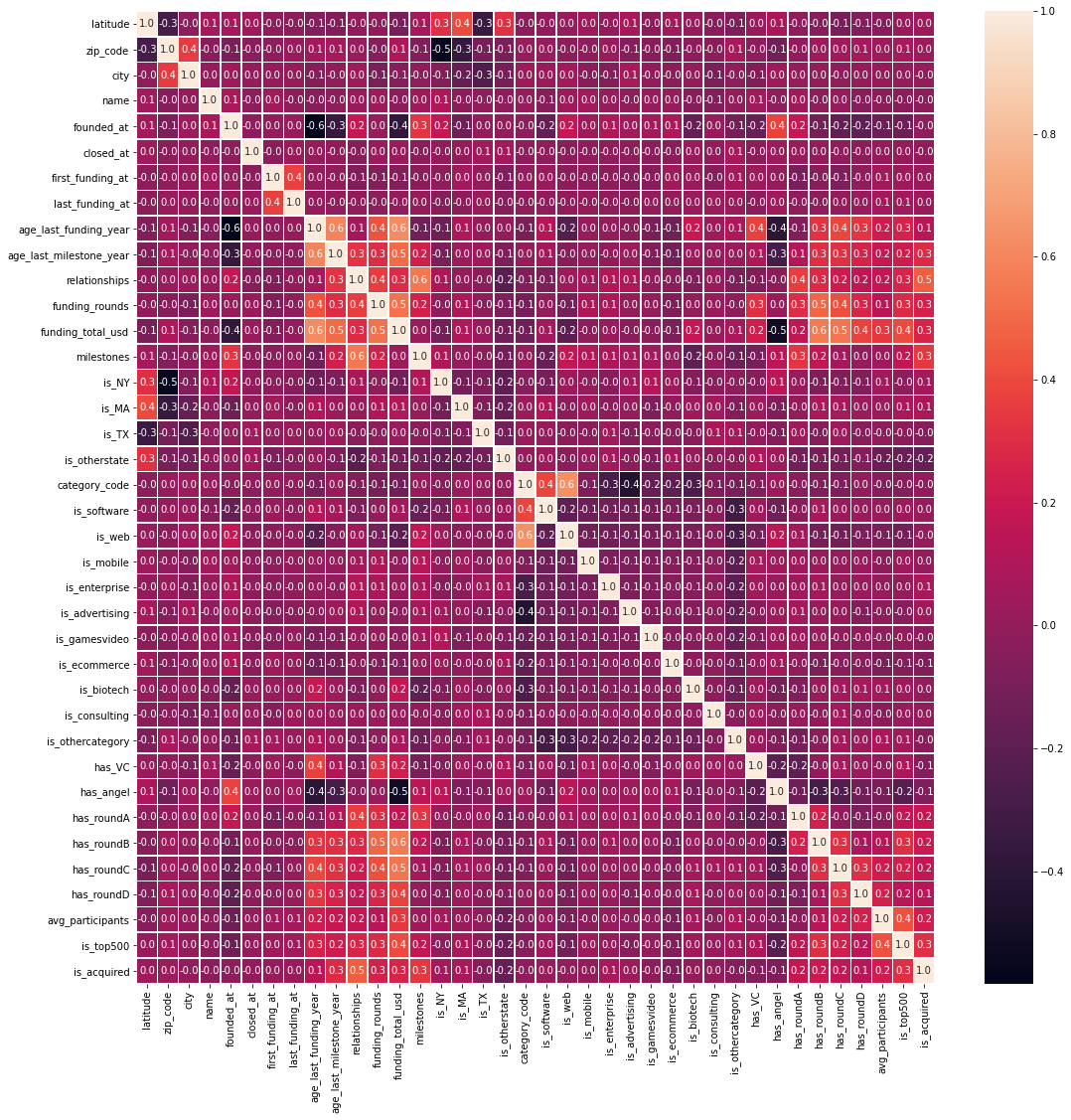
# • labels

# • state\_code

# • is\_CA

After removing these we have following graph

So



Now we have successfully made a selection of features.

**Dataset (standardized values):** <https://github.com/ZeeWING-Projects/DM-Project/blob/main/Dataset-2%20Pre-processed/Dataset_02_standarized_.csv>

**Dataset (non-standardized values):** <https://github.com/ZeeWING-Projects/DM-Project/blob/main/Dataset-2%20Pre-processed/Dataset_02_non_standarized_.csv>

# Data visualization

NOTE: For this purpose I am using dataset 1. Reason for this is the meaning of features required for understanding of data visualization and scattered plot.

For understating the data graphically we can use different types of graphs. For this dataset analysis we will use some well know graphs to analyze few features of dataset.

## 2.1 Box-Plot

Def: A Box Plot is also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum. In the box plot, a box is created from the first quartile to the third quartile; a vertical line is also there which goes through the box at the median. Here x-axis denotes the data to be plotted while the y-axis shows the frequency distribution. (geeksforgeeks). So this is used for quick summary of data. we have following points to observe by using this box-plot.

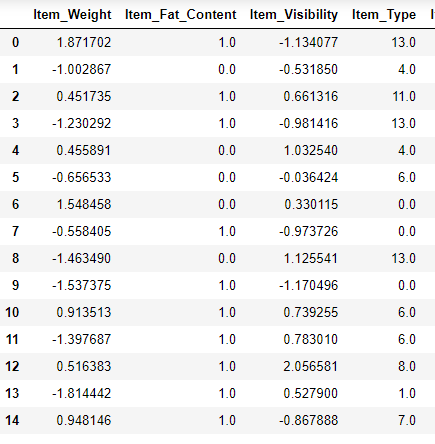
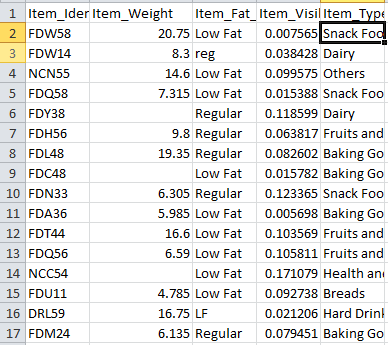
* Detect outlier values.
* Mean tendency of values.
* Symmetry of data.

Assume for example we are plotting item\_weight of different types of items (item\_type)

For this we need item weights of each individual item\_type. we have following types of unique items

**NOTE: We are using non-standardized dataset (the one which is already pre-processed see above section)**

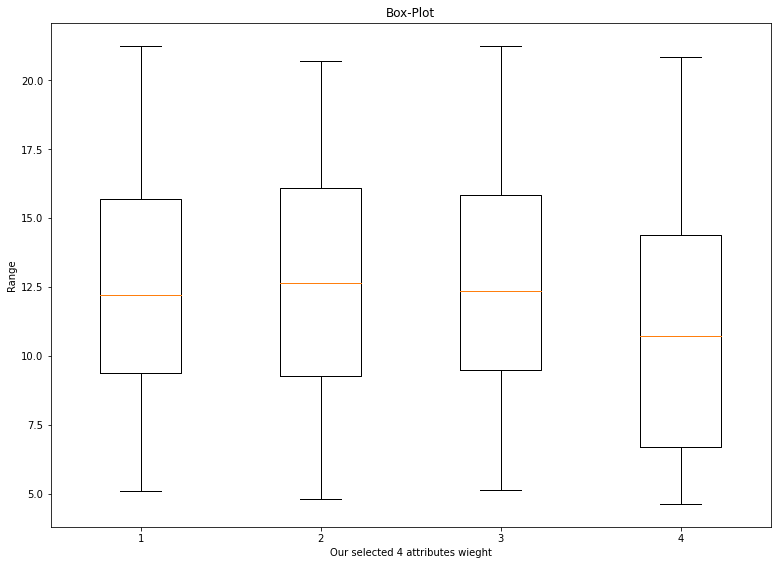
but as we have converted the all values in numeric form then now we need to match columns so that we be able to get the names against numbers.

**13.0 = Snack Food, 4.0 = Diary, 1.0 = bread and meat=10.0**

For example we are taking 4 items type.

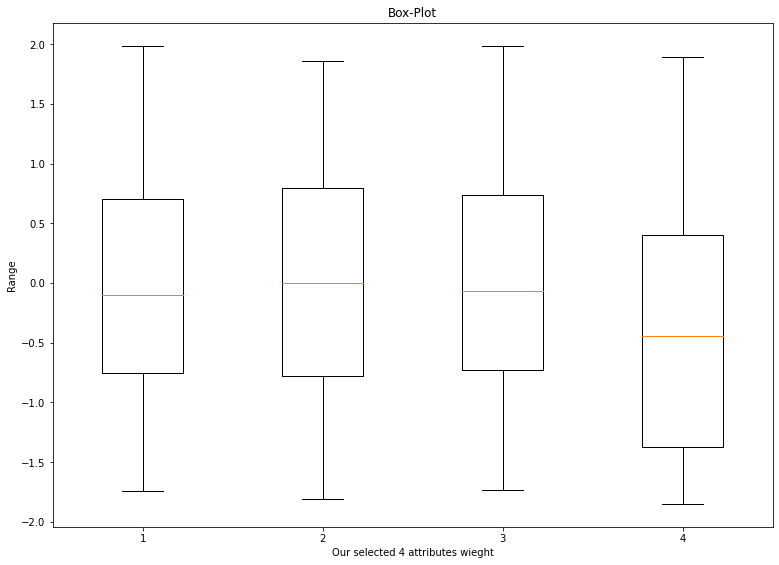
1. Snack foods
2. Dairy
3. Meat
4. Breads



### Analysis of box-plot

1. Snack foods:
   1. Average weight of this item is around 12.5 or a bit greater.
   2. Its max average value is around 15.5 (approximately)
   3. Its min average value is around 9.7 (approximately)
   4. its symmetric line shows that there is no skewness, which means no there are approximately equal number of values greater than mean and less than mean.
   5. There is no out liar value.
2. Dairy:
   1. Average weight of this item is around 12.
   2. Its max average value is around 16.25 (approximately)
   3. Its min average value is around 10 (approximately)
   4. its symmetric line shows that there is right skewned because it is toward lower half, the **number of items is greater** whose value is less than than mean.
   5. There is no out liar value.
3. Meat:
   1. Average weight of this item is around 12.5 or a bit greater.
   2. Its max average value is around 15.5 (approximately)
   3. Its min average value is around 9.8 (approximately)
   4. its symmetric line shows that there is no skewness, which means no there are approximately equal number of values greater than mean and less than mean.
   5. There is no out liar value.
4. Breads:
   1. Average weight of this item is around 11.25 or a bit greater.
   2. Its max average value is around 13.25 (approximately)
   3. Its min average value is around 6.5 (approximately)
   4. its symmetric line shows that there is left skewned because it is toward upper half, the **number of items is greater** whose value is greater than mean.
   5. There is no out liar value.

**Now we are using standardized values.**



## Summary of using standardized and non-standardized values

There is no different in results except the range.