About Dataset

The data were collected in 2013 for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined.

Attributes

- **Item_Identifier** -> Unique Product ID
- **Item_Weight** -> Weight of the product
- **Item_Fat_Content** -> Whether the product is low fat or not
- **Item_Visibility** -> The % of the total display area of all products in a store allocated to the particular product
- **Item_Type** -> The category to which the product belongs
- **Item_MRP** -> Maximum Retail Price (list price) of the product
- **Outlet_Identifier** -> Unique store ID
- **Outlet_Establishment_Year** -> The year in which the store was established
- **Outlet_Size** -> The size of the store in terms of ground area covered
- **Outlet_Location_Type** -> The type of city in which the store is located
- **Outlet_Type** -> Whether the outlet is just a grocery store or some sort of supermarket
- **Item_Outlet_Sales** -> Sales of the product in the particular store.

Data cleaning

1. Using the count blank function to count the null values for each column, there are 1463 missing values in the item weight column and 2410 in the outlet size column.
2. Use the countif function to count the number of rows in the item visibility column that has a 0 value.
3. Fill null values in the item weight column with the item weight value for each item identifier.
4. Fill null values in the item visibility column with the mean.
5. Remove the outlet size column because it has many null values, and the column is less important for analysis.
6. Replace the “lf” value with ‘Low Fat’ in item fat content.
7. Replace the “reg” value with “Regular” in the item fat content column.
8. Convert the type of item_mrp and item_outlet_sales columns to currency type.