Assignment #2

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

Kmart wants to better understand the properties of products and stores which play a key role in increasing sales.

**Goal**: The aim is to build a predictive model to find out the sales of each product at a particular store.

**Questions to answer**:

1. What steps should be taken for pre-processing the data?
2. Explain your modeling approach:
   1. Choice of ML algorithm
   2. Feature Selection
3. How would you evaluate the performance of your model? Explain the choice of metrics used.
4. Show plots for visualizing your model performance.

**Data**:

Please note that the data may have missing values as some stores might not report all the data due to technical glitches. Hence, it will be required to treat them accordingly.

We have train (8523) and test (5681) data set, train data set has both input and output variable(s). You need to predict the sales for test data set.

Datafiles: test\_kmart.csv, train\_kmart.csv

Item\_Identifier: Unique product ID

Item\_Weight: Weight of product

Item\_Fat\_Content: Whether the product is low fat or not

Item\_Visibility: The % of 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 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. This is the outcome variable to be predicted.