**BIG MART SALES ANALYSIS**

**Executive:** The aim is to perform a data analysis on the sales of Big Mart

**Background:** Given the data, Big Mart is a grocery sales company .

**Purpose:** The main purpose is to check whether there is a statistically significant relation between the features of items like its establishment or location of selling with the sale of the product.

**Questions:**

1. Import Train & Test DataSet from BigMart Dataset folder.

2. Check dimensions (number of row & columns) & Structure in dataset.

3. Find Missing Values in the dataset.

4. Find Missing Values according to Columns.

5. Find Summary of DataSet & Draw Conclusions from it.

6. ScatterPlots

A. Plot a ScatterPlot using ggplot for Item\_Visibility vs Item\_Outlet\_Sales & draw conclusion from which products visibility is more sales.

B. Plot a Barplot using ggplot for Outlet\_Identifier vs Item\_Outlet\_Sales & Draw conclusion who has contributed to majority of sales.

C. Plot a Barplot using ggplot for Item\_Type vs Item\_Outlet\_Sales also draw conclusion which items are sold more.

D. Plot a Boxplot using ggplot for Item\_Type vs Item\_Outlet\_Sales also draw conclusion which items are sold more.

7. Manipulating Dataset to make it consistent

A. Add Item\_Outlet\_Sales Column to test dataset which is’nt available & assign integer 1. Also Combine Both Train + Test Datasets.

B. Impute missing value in Item\_Weight using median because it is highly robust to Outliers.

C. We saw item visibility has zero value also, which is practically not feasible. Impute median value where item\_visibility 0.

D. Rename level in Outlet\_Size to since mis-matched levels in variables needs to be corrected..

E. Rename levels of Item\_Fat\_Content since value are “LF” / “low fat”, so make them consistent

F. Create a new column 2013 – Year ( For Prediction ).

G. Drop variables not required in modelling i.e. Item\_Identifier, Outlet\_Identifier, Outlet\_Establishment\_Year as they aren’t needed for prediction. H. Divide data set into Train and Test.

I. Perform a Regression testing on training dataset

J. Plot Summary and Predict sales for Testing Dataset.

**Methods:**

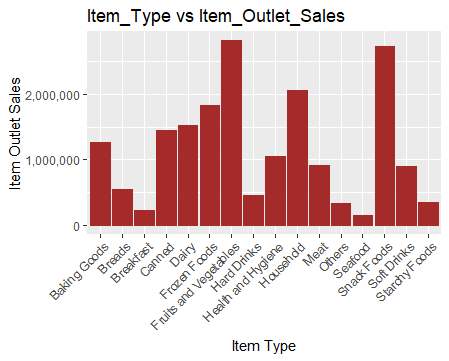
Read.csv() was used to read and import the dataset from the system to do the analysis.

Exploratory analysis was done using dim(), table(is.na()), colSums(), summary(), rbind(), median(), levels() etc. A statistical analysis plots were visualized with the help of ggplot(), geom\_point() boxplot()etc.

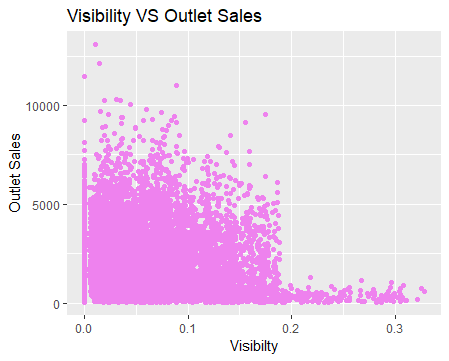
Regression testing was performed on the dataset.

**Instrumentation:** The analysis was done on RStudio with R-programming

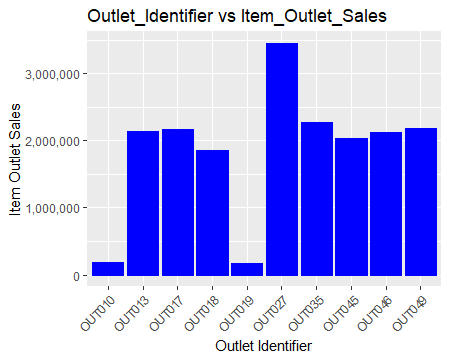
**Results:** The sales were successfully analyzed by their features. Following are the findings and the plots.

Item type Vs Item outlet sales plot

Visibility VS Outlet Sales



Outlet Identifier VS Item Outlet sales



**Summary:**

The Big Mart dataset was analyzed. There were two datasets with 8523 x 12 rows and columns in one and 5681 x 12 rows and columns in other. The dataset contained features like item identifier, item weight, item fat content, item visibility, item type, item MRP, outlet identifier, outlet size, outlet established year, outlet type and outlet location type The 1st dataset had 1463 values missing(NA) in the weight column and the 2nd had 976 missing weight values. There were no other missing values in any section. NAs were replaced by the median value of the weight for better prediction. The two datasets were then combined and only relative features were taken for prediction. Graphs were plotted for analysis and Regression testing was done on the dataset.