Big Mart Sales

Based on the various analyses and the visualizations from the data (the code can be found on my Github page), we can make the following observations for the different comparisons:

**Independent Variables (numeric variables)**

1. There seems to be no clear-cut pattern in Item\_Weight.
2. Item\_Visibility is right-skewed and should be transformed to curb its skewness.
3. We can clearly see 4 different distributions for Item\_MRP. It is an interesting insight.

### ****Independent Variables (categorical variables)****

1. Lesser number of observations in the data for the outlets established in the year 1998 as compared to the other years.
2. Supermarket Type 1 seems to be the most popular category of Outlet\_Type.

### ****Target Variable Vs Independent Numerical Variables****

1. Item\_Outlet\_Sales is spread well across the entire range of the Item\_Weight without any obvious pattern.
2. In Item\_Visibility vs Item\_Outlet\_Sales, there is a string of points at Item\_Visibility = 0.0 which seems strange as item visibility cannot be completely zero. We will take note of this issue and deal with it in the later stages.
3. In the third plot of Item\_MRP vs Item\_Outlet\_Sales, we can clearly see 4 segments of prices that can be used in feature engineering to create a new variable.

### ****Target Variable Vs Independent Categorical Variables****

1. Distribution of Item\_Outlet\_Sales across the categories of Item\_Type is not very distinct and same is the case with Item\_Fat\_Content.
2. The distribution for OUT010 and OUT019 categories of Outlet\_Identifier are quite similar and very much different from the rest of the categories of Outlet\_Identifier.
3. Tier 1 and Tier 3 locations of Outlet\_Location\_Type look similar.
4. In the Outlet\_Type plot, Grocery Store has most of its data points around the lower sales values as compared to the other categories.