**Big Mart Sales Prediction**

**Sales Prediction for Big Mart Outlets**

The data scientists at BigMart 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. The aim is to build a predictive model and predict the sales of each product at a particular outlet.

Using this model, BigMart will try to understand the properties of products and outlets which play a key role in increasing sales.

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.

**Data Dictionary**

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.

**Train file: CSV containing the item outlet information with sales value**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| 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. |

**Test file: CSV containing item outlet combinations for which sales need to be forecasted**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| 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 |

We will explore the problem in following stages:

* **Hypothesis Generation** – understanding the problem better by brainstorming possible factors that can impact the outcome
* **Data Exploration** – looking at categorical and continuous feature summaries and making inferences about the data.
* **Data Cleaning** – imputing missing values in the data and checking for outliers
* **Feature Engineering** – modifying existing variables and creating new ones for analysis
* **Model Building** – making predictive models on the data

**The Hypotheses**

I came up with the following hypothesis while thinking about the problem. These are just my thoughts and you can come-up with many more of these. Since we’re talking about stores and products, lets make different sets for each.

**Store Level Hypotheses:**

* **City type:** Stores located in urban or Tier 1 cities should have higher sales because of the higher income levels of people there.
* **Population Density:** Stores located in densely populated areas should have higher sales because of more demand.
* **Store Capacity:** Stores which are very big in size should have higher sales as they act like one-stop-shops and people would prefer getting everything from one place
* **Competitors:** Stores having similar establishments nearby should have less sales because of more competition.
* **Marketing:** Stores which have a good marketing division should have higher sales as it will be able to attract customers through the right offers and advertising.
* **Location:** Stores located within popular marketplaces should have higher sales because of better access to customers.
* **Customer Behavior:** Stores keeping the right set of products to meet the local needs of customers will have higher sales.
* **Ambiance:** Stores which are well-maintained and managed by polite and humble people are expected to have higher footfall and thus higher sales.

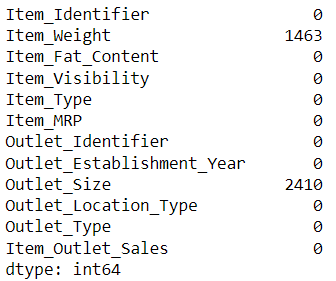
**Product Level Hypotheses:**

* **Brand:** Branded products should have higher sales because of higher trust in the customer.
* **Packaging:** Products with good packaging can attract customers and sell more.
* **Utility:** Daily use products should have a higher tendency to sell as compared to the specific use products.
* **Display Area:** Products which are given bigger shelves in the store are likely to catch attention first and sell more.
* **Visibility in Store:** The location of product in a store will impact sales. Ones which are right at entrance will catch the eye of customer first rather than the ones in back.
* **Advertising:** Better advertising of products in the store will should higher sales in most cases.
* **Promotional Offers:** Products accompanied with attractive offers and discounts will sell more.

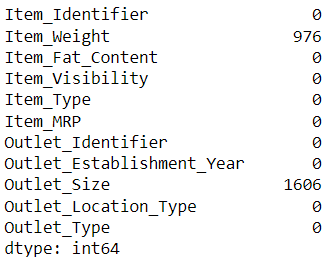
These are just some basic 15 hypothesis I have made, but you can think further and create some of your own. Remember that the data might not be sufficient to test all of these, but forming these gives us a better understanding of the problem and we can even look for open source information if available.

**Exploratory Data Analysis**

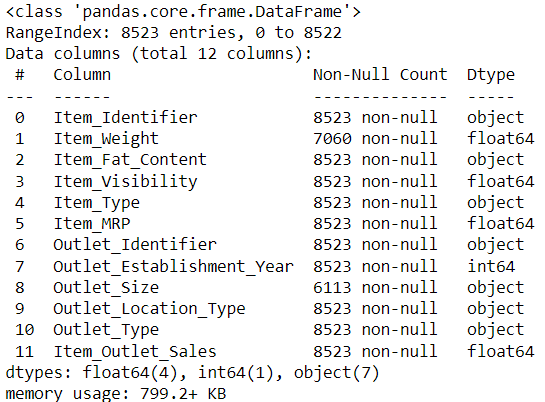
**Exploring Train data null values**



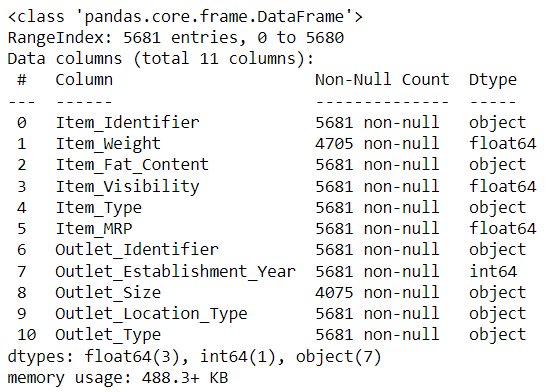
**Exploring Test data null values**

****

**Train Data types**

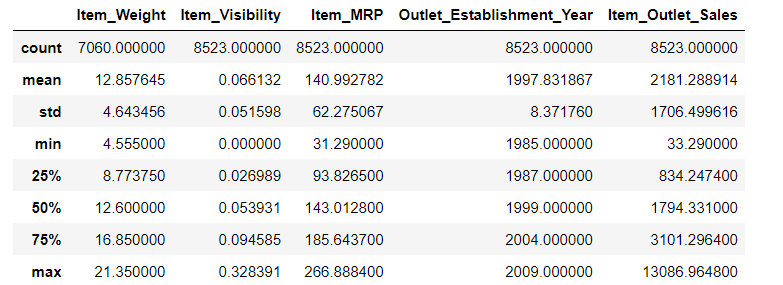
****

**Test Data types**

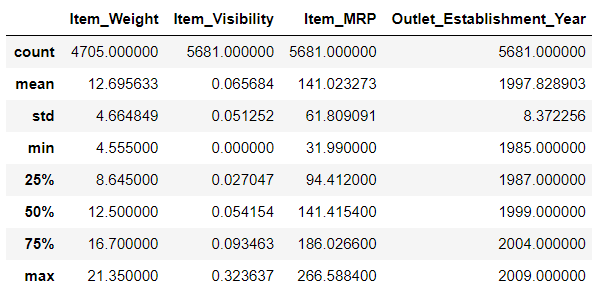
****

**Basic Statistics**

**Train:**

****

**Test:**

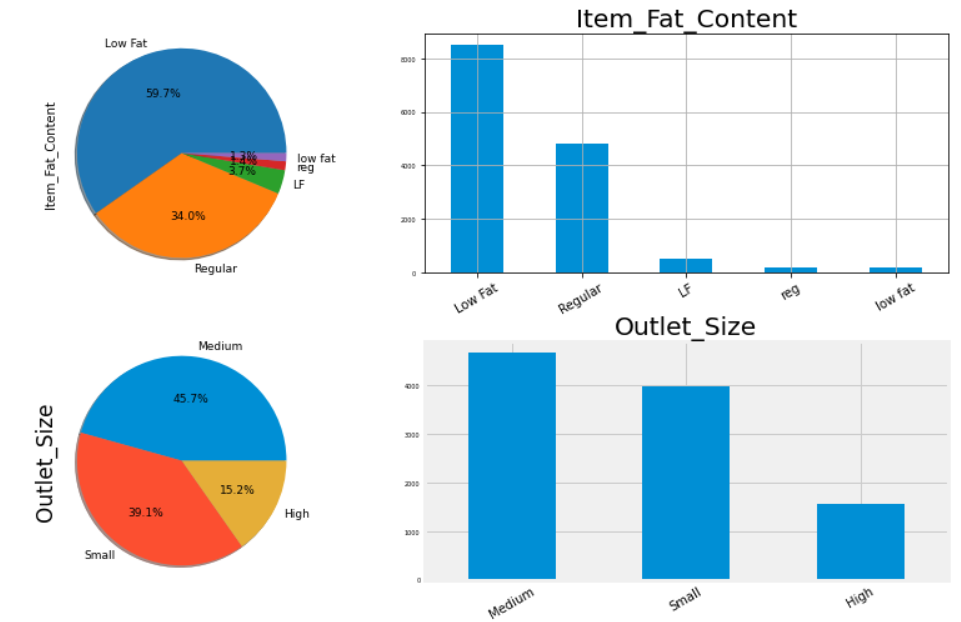
****

**Unique items**

After concatenating train and test data, there are 1559 unique values of Item\_Identifier with max frequency of 10 and min is 7

**Explore categorical variables**

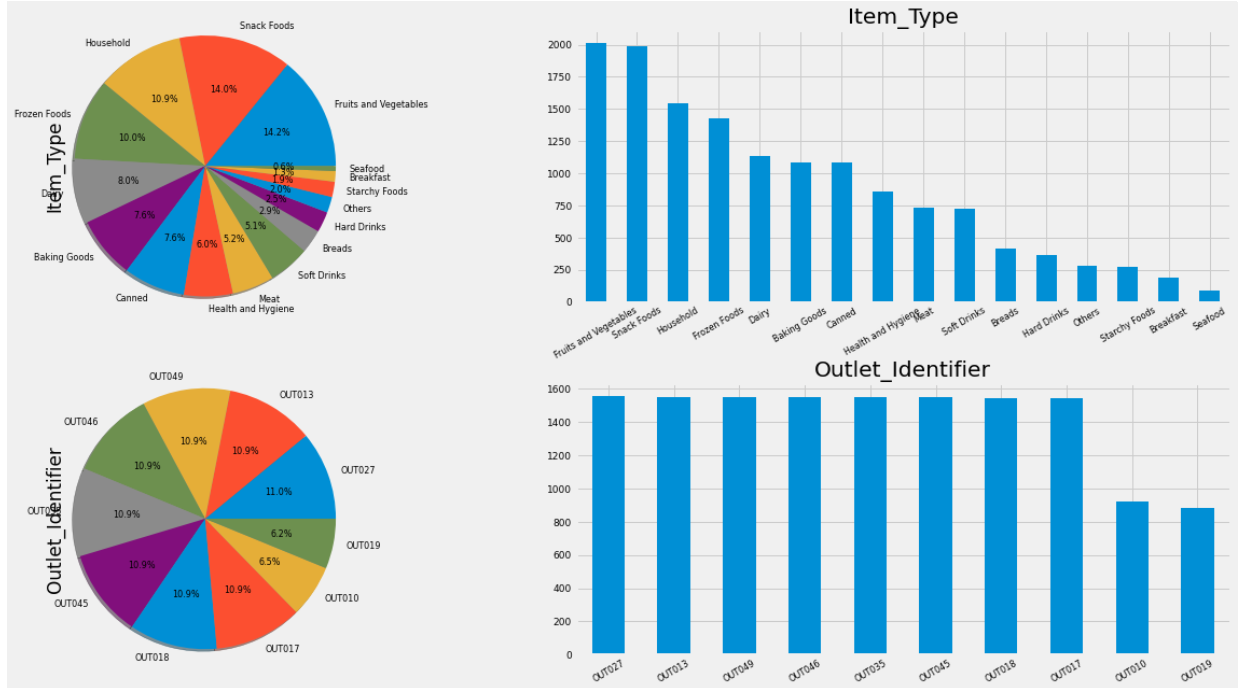
**'Item\_Fat\_Content', 'Outlet\_Size'**

****

**Observation:**

* **Item fat content:** Low Fat is having maximum (59.7%) count
* **outlet size:** Medium size outlet (45.7%) is having max count

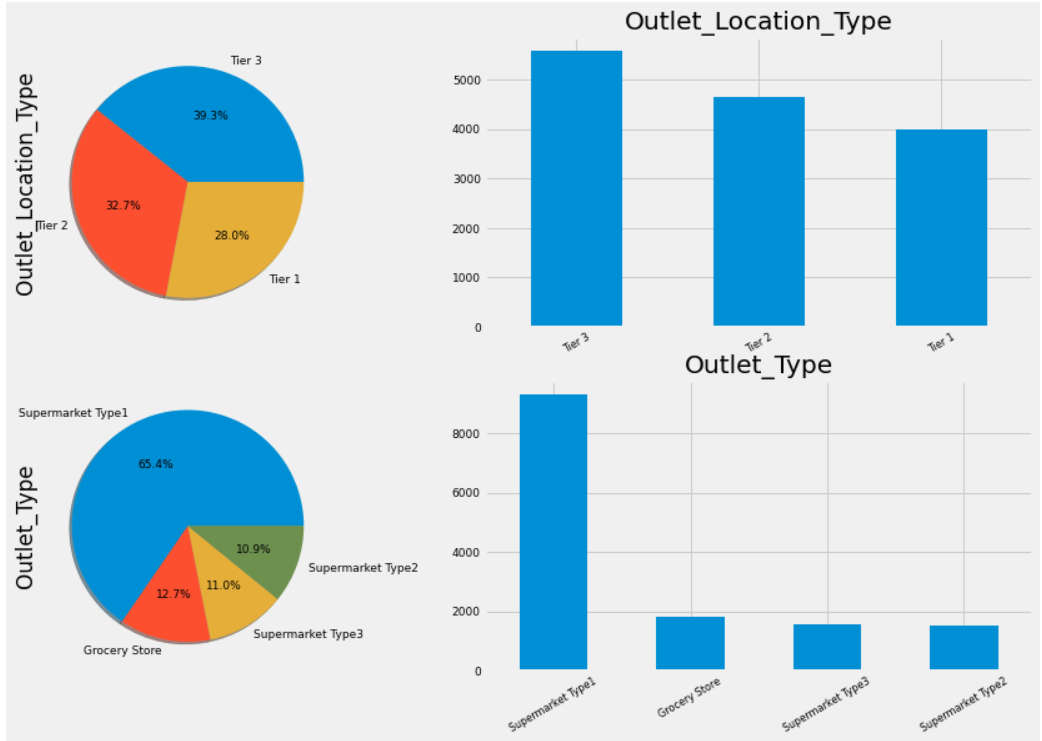
**'Item\_Type', 'Outlet\_Identifier'**

****

**Observation:**

* **Item type:** Fruits and vegetables has max count of 2k and seafood is min
* **outlet identifier:** outl027 is having max count and out019 is having min count

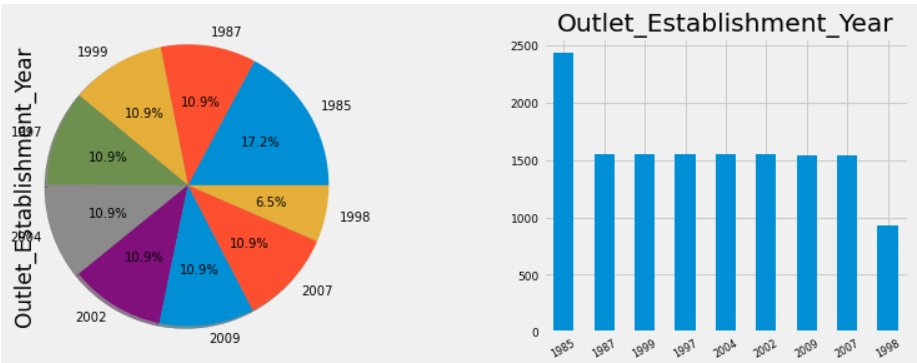
**'Outlet\_Location\_Type', 'Outlet\_Type'**

****

**Observation:**

* **Outlet\_Location Type**: Max count for Tier3 (39.3%) and Tier 1 is having lowest count
* **Outlet\_Type:** Supermarket Type1 has mx count (65.4%)

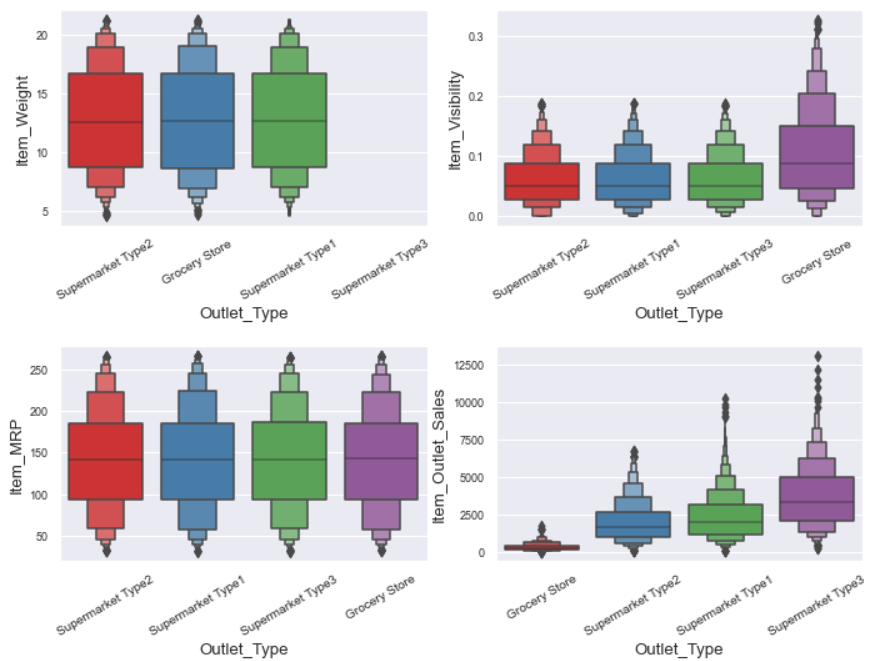
**'Outlet\_Establishment\_Year'**

****

**Observation:**

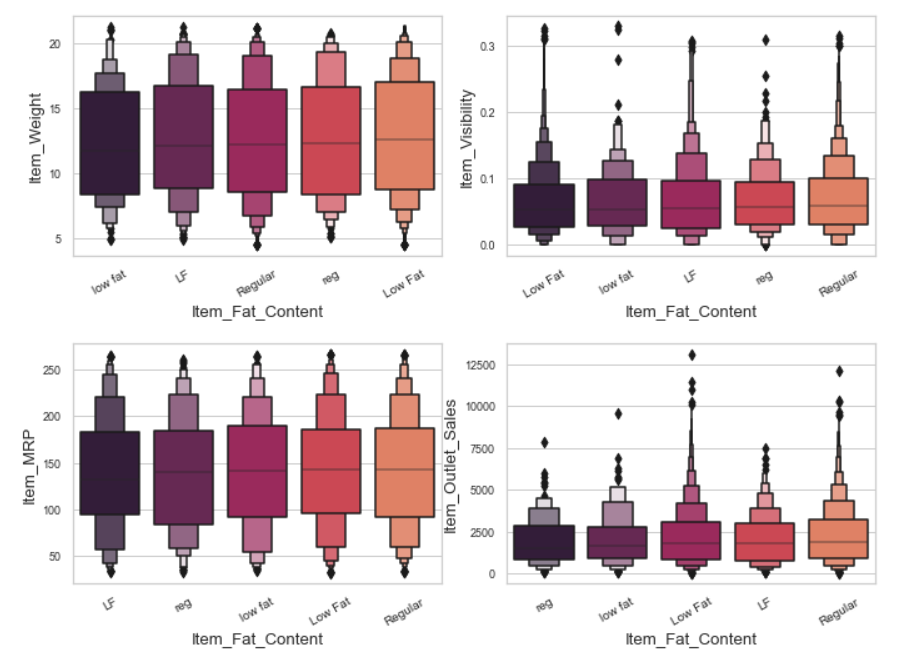
* 1985: Maximum count of 2300+ can be seen
* 1998: Minimum count of 900 can be seen
* All other years have same count which is surprising

### **Explore numeric variables –**



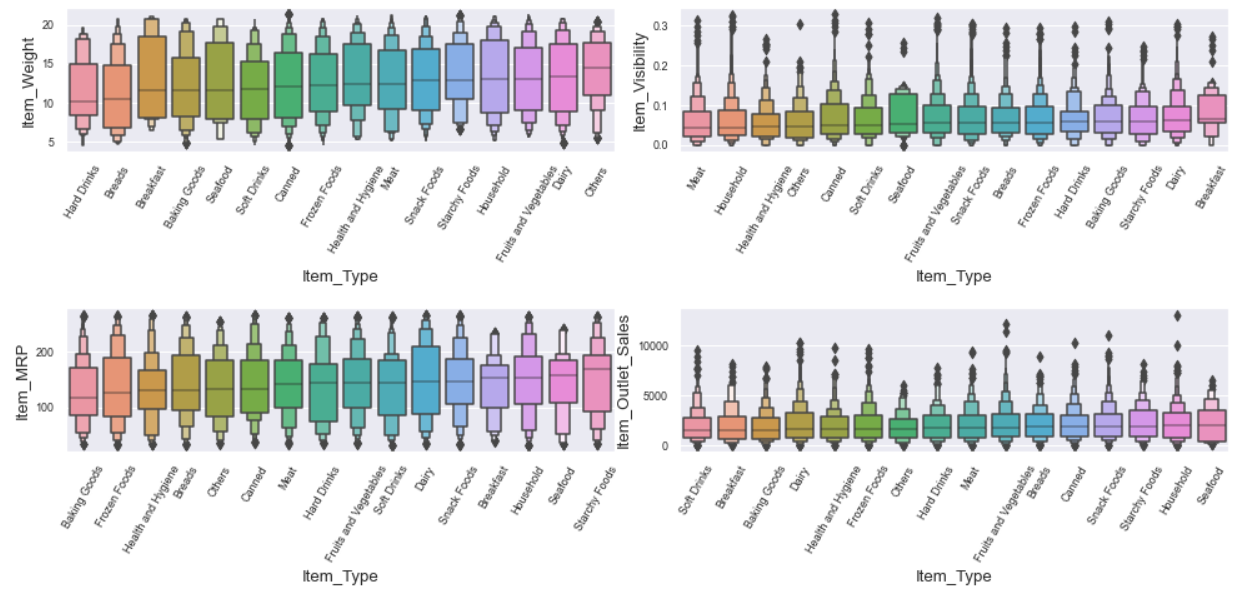
**Observation:**

* **Item weight**: Its same across diff outlets. Item weight is missing for super mkt 3 and grocery as per below stats
* **Item\_Visibility:** Grocery stores have highest visibility, rest all super markets have same visibility
* **Item\_MRP:** MRP is same across all outlets which is obvious
* **Item\_Outlet\_Sales:** Median Sales volume are highest at super mkt type 3 > super mkt 1 > super mkt 2 > grocery



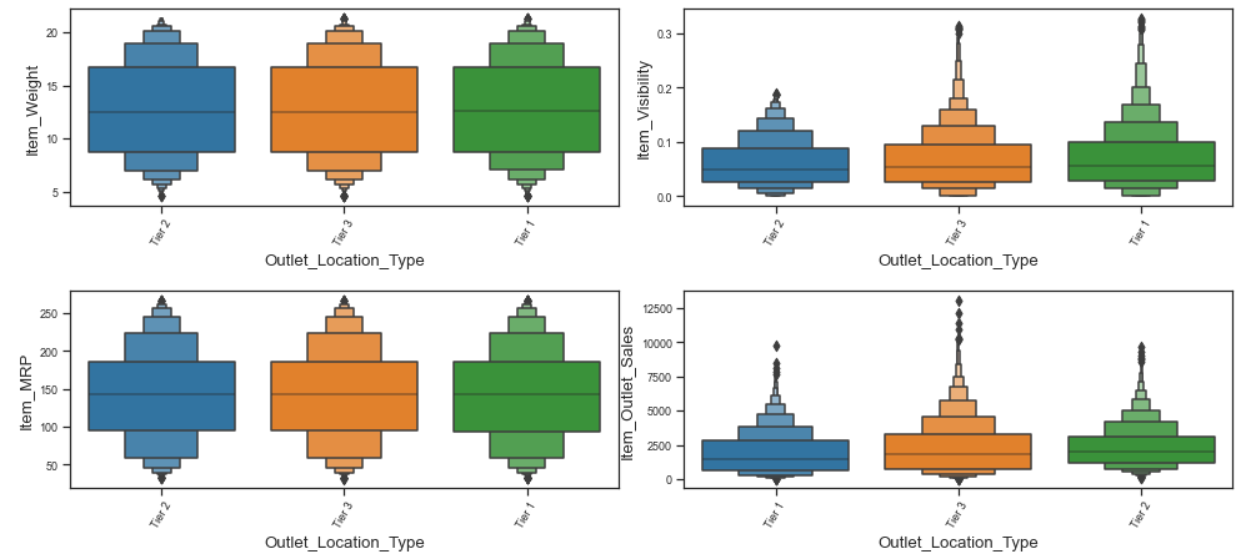
**Observation:**

* **Item weight:** "low fat" is having min median item weight < LF < Regular < reg < Low Fat
* **Item\_Visibility:** same across all categories
* **Item\_MRP:** Lf < reg & low fat < Low Fat < Regular
* **Item\_Outlet\_Sales:** Median Sales volume reg < low fat < Low Fat < LF < Regular (But not much variance among diff categories)

****

**Observation:**

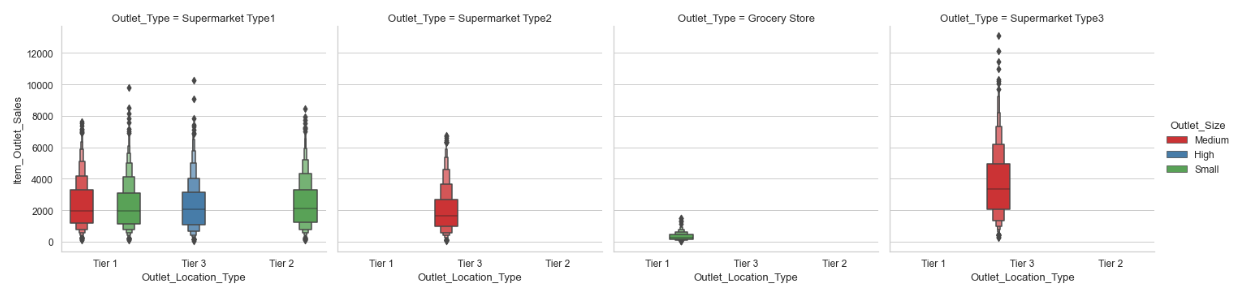
* **Item weight:** hard drinks is having min median item weight & others having max median value
* **Item\_Visibility:** meat is lowest & breakfast is highest in median values
* **Item\_MRP:** baking goods is lowest and starchy foods is highest in median values
* **Item\_Outlet\_Sales:** soft drinks have min median Vs sea food is having maximum median sales value



**Observation:**

* **Item weight:** same across all
* **Item\_Visibility:** Tier 2 < 3 & 1
* **Item\_MRP**: same across all
* **Item\_Outlet\_Sales:** Tier 1 < 3 < 2 (i.e. Tier 2 generates highest sales median value)

**Multivariate Analysis**



**Observation:**

* Tier 3 in type super mkt 3 is having highest median sales

**Missing Value Imputation:**

**Item Weight**

**Item Weight -** Item weight should depend upon the item type and Item\_Fat\_Content a per above numerical analysis with boxen plots.

Hence, we can impute the missing value based on median of **Item\_Type** and **Item\_Fat\_Content**

**Outlet Size**

There are three types of outlet sizes: small/medium/large

The mussing values exist for the below combination of outlet type & outlet location type

* Grocery & Tier 3
* Super market 1 & Tier 2

As per the above boxen plot we can see that for

* outlet type = grocery generally the outlet size is small
* outlet type = super market generally the outlet size is small & medium.

Hence, I decided to impute small value for all missing values of Outlet Size

**Model Building**

* **Model:** After trying out many models, I decided to use GBM to predict, as it was showing lowest RMSE (GradientBoostingRegressor(random\_state=seed))
* **Train Test Split:** 0.20
* **Scaling:** StandardScaler
* **param\_grid:** dict(n\_estimators=np.array([80,82,84,85,90,91,92,94,95,150,200]))
* **seed:** 7
* **best estimators:** 80

### **GridSearchCV**

