This is my contribution to the Big Mart Sales prediction competition at:

<http://datahack.analyticsvidhya.com/contest/practice-problem-bigmart-sales-prediction>

Given sales data for 1559 products across 10 stores of the Big Mart chain in various cities the task is to build a model to predict sales for each particular product in different stores.

The train and test data, which can be found at the link given above, contain the following variables:

|  |  |
| --- | --- |
| **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 particulat store. This is the outcome variable to be predicted. |

**Imputing missing value:**

For Continuous variable Item weight:

Imputing with mean

For categorical variable Outlet size:

KNN Imputation(used K=6)

**Data Exploration and Preparation**

For categorical variable:

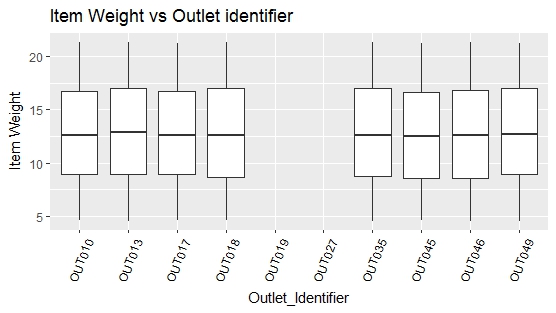
Item type: Data is distributed with balanced one, with each variable is having certain amount of explanation. But Fruits and vegetables, snacks, household, frozen, diary products having majority.

Item\_Fat\_content: Low\_fat and regular is widely used and it is skewed one with 92% in combination of both it explains. Remaining fat contents are transformed to others since it has only less than 10 % of contents.

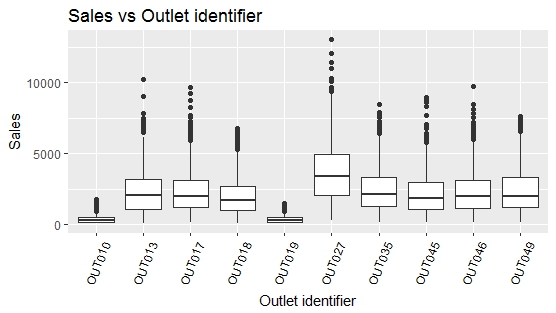
Outlet\_identifier: Small and medium contributes the more, with medium is having most observations.

Outlet Location\_type: It has 3 tier with Tier-3 having the most identified location type.

Outlet-Type: Super-market-1 contributes with 65% of outlet type. Other type only contributes remaining 35%.



Looking at those plots one notices that all the medians, boxes and whiskers are identical to each other. Have those practice data been faked by any chance?



we see that the two grocery stores OUT010 and OUT019 have reported far fewer sales than the supermarkets. This is neatly illustrated by a boxplot:

## Predictive Modelling

Used Logistic regression to predict the model with output parameter as Item outlet sales and putting all variables as input parameter.

Adjusted R-square: 35.76%

RMSE: 1366.88.