PREDICTING SALES OF DIFFERENT ITEMS IN BIG MART

CAPSTONE PROJECT (GROUP 2) | Batch: PGP-DSE (FT Online July 22)

Mentor: Siddharth Koshta
Team members: Shekhar Singh | Neha Mondal | Ritesh Singh | Azmath Ulla Baig | Aprajita Raj

INTRODUCTION TO THE PROBLEM STATEMENT

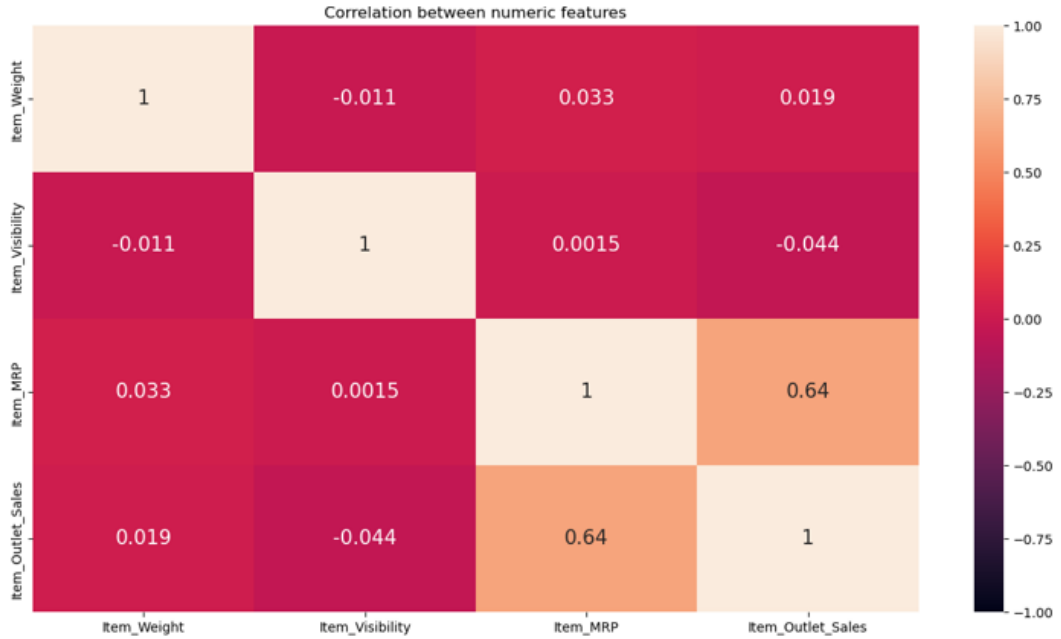
If in future, company want to expand its businesses to other location then which size of the mart would be more beneficial to open for them. This can be easily comprehended by analyzing the data of location and size of mart and how it is affecting the sale .

Business Problem Understanding:

- We are using big data of details about Bigmart products(items)
- The sale of the product that are sold in Bigmart depends on various attributes like, it's size , price , visibility etc .
- The attributes (properties of the items) are correlated with each other, which indicates that change in any one of the properties may effect the sales directly and also indirectly (as other properties are also effected).
- Bigmart items has visibility and also fat content so we can target the consumers who are health conscious.

DATA DESCRIPTION

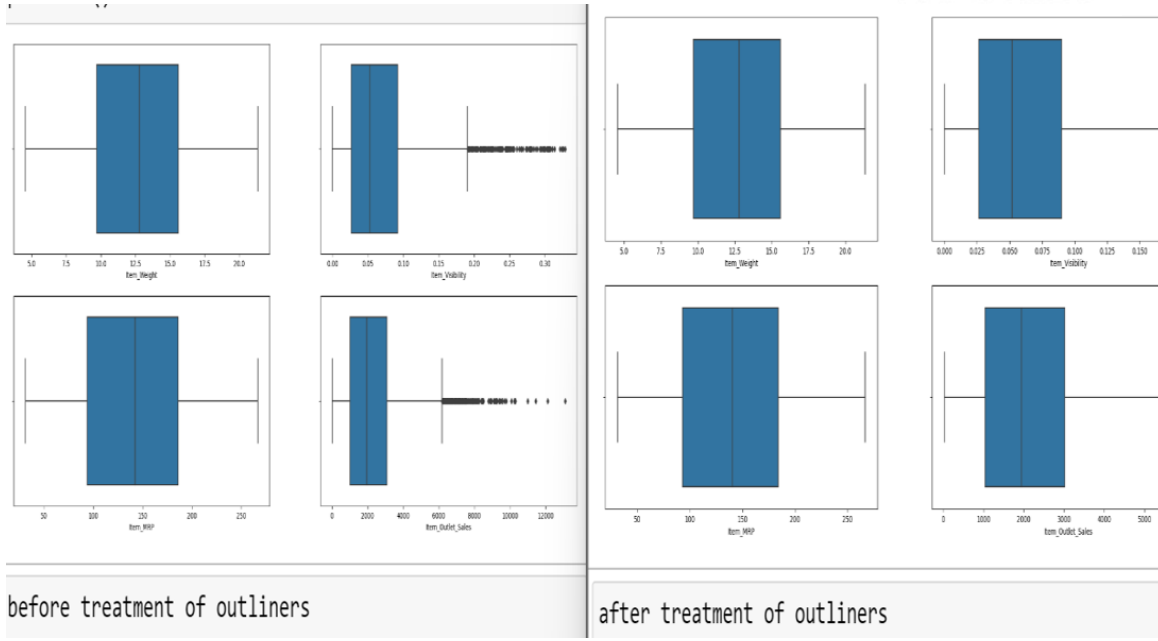
- Item details - Weight, Type, MRP, Outlet size, type and sales, Fat content, Visibility, etc.
- 14204 entries , 12 attributes
- Outliers – Item Outlet sales, Item MRP, Item weight, Item visibility.
- High cardinality – Item Identifier
- High correlation between columns- Heat Map
- Numerical, Categorical columns
- 28.273 maximum null percentage




```
[8]  (df.isnull().sum()/df.shape[0])*100

     Item_Identifier              0.000000
     Item_Weight                 17.171219
     Item_Fat_Content             0.000000
     Item_Visibility              0.000000
     Item_Type                    0.000000
     Item_MRP                     0.000000
     Outlet_Identifier            0.000000
     Outlet_Establishment_Year    0.000000
     Outlet_Size                 28.273726
     Outlet_Location_Type         0.000000
     Outlet_Type                  0.000000
     Item_Outlet_Sales            0.000000
     dtype: float64

[9]  df.shape

     (14204, 12)
```

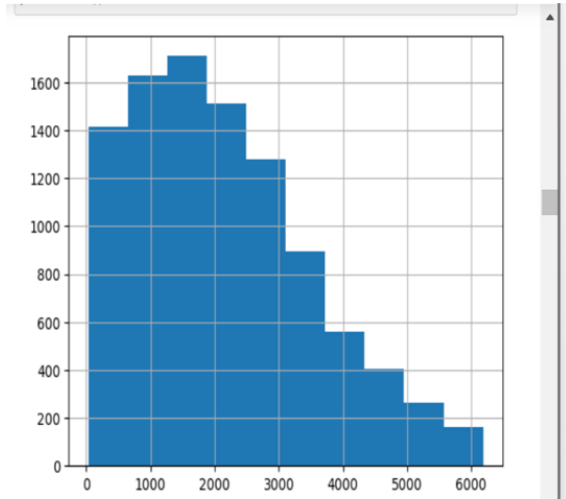
DATA CLEANING STEPS

- Converted categorical columns to numerical columns.
- Dropped some null values
- Imputed remaining
- Dropped duplicates
- Capped outliers of columns

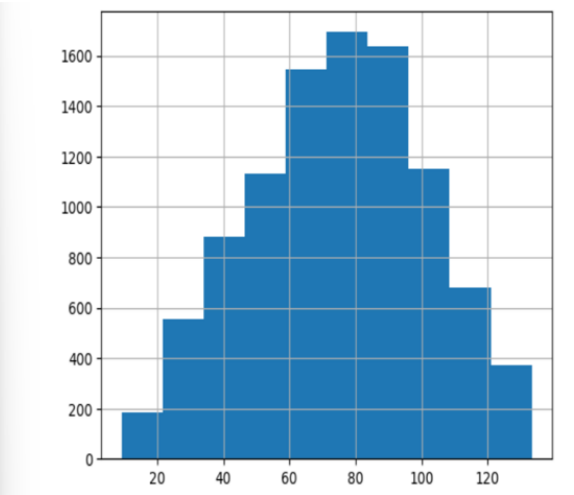


before treatment of outliners

after treatment of outliners



distribution plot of outlate sales



boxcox transforamtion on outlet sales(target column)

PROBLEM SOLVING STEPS

Checked datatypes → Analyzed summery stats → Univariate and bivariate analysis

Tested effect of features on target → Encoding → Split data into train-test

Fit base model → Fit final model

DETAILED ANALYSIS

As our target value is numerical we will be using linear regression method to build a suitable model.

Use of Computational techniques, analysis techniques and statistical techniques that can help in designing better store sale performance based on the previous trend performance.

For our first try model we build a MLR full model using the new data set that has been transformed using transformation of numerical and encoding of categorical data.

From the r square value, we can see the model is performing nice.

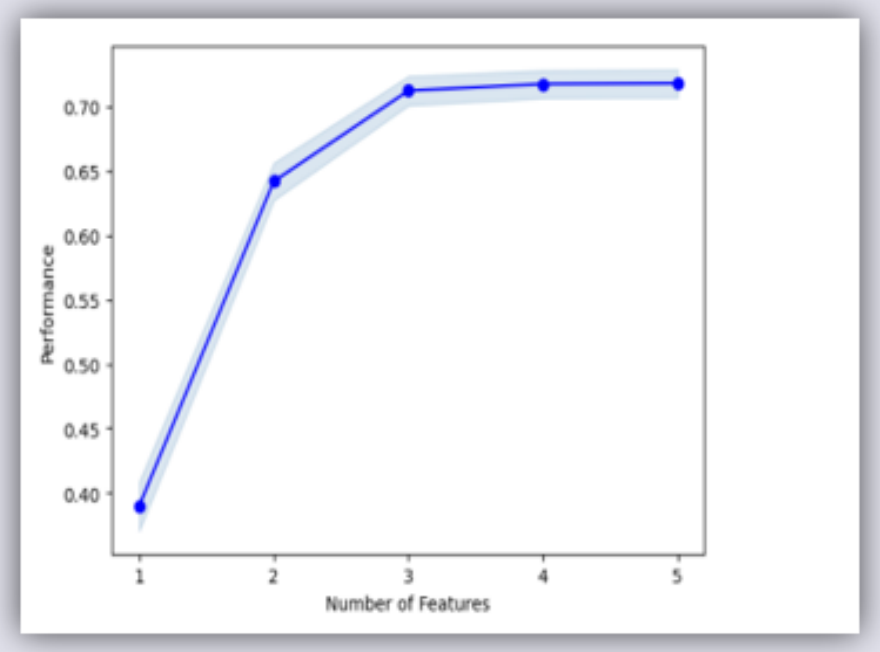
MODEL BUILDING

For numerical columns we checked the VIF values and found all the VIF values less than 10. hence, we can conclude that there is no multicollinearity among the numerical attributes.

	VIF_Factor	Features
0	1.001219	Item_Weight
1	1.000115	Item_Visibility
2	1.001109	Item_MRP

But our model shows there is a strong multicollinearity. So, we will drop unrelated or less significant attributes and then build a model.

MODEL BUILDING



We did build a model on Linear Regression SGD and our R^2 value decreased than the value we got from the MLR full model. It shows signs of under fit.

And when we used the forward feature selection our R^2 value was above 90 but the RMSE value where around 20

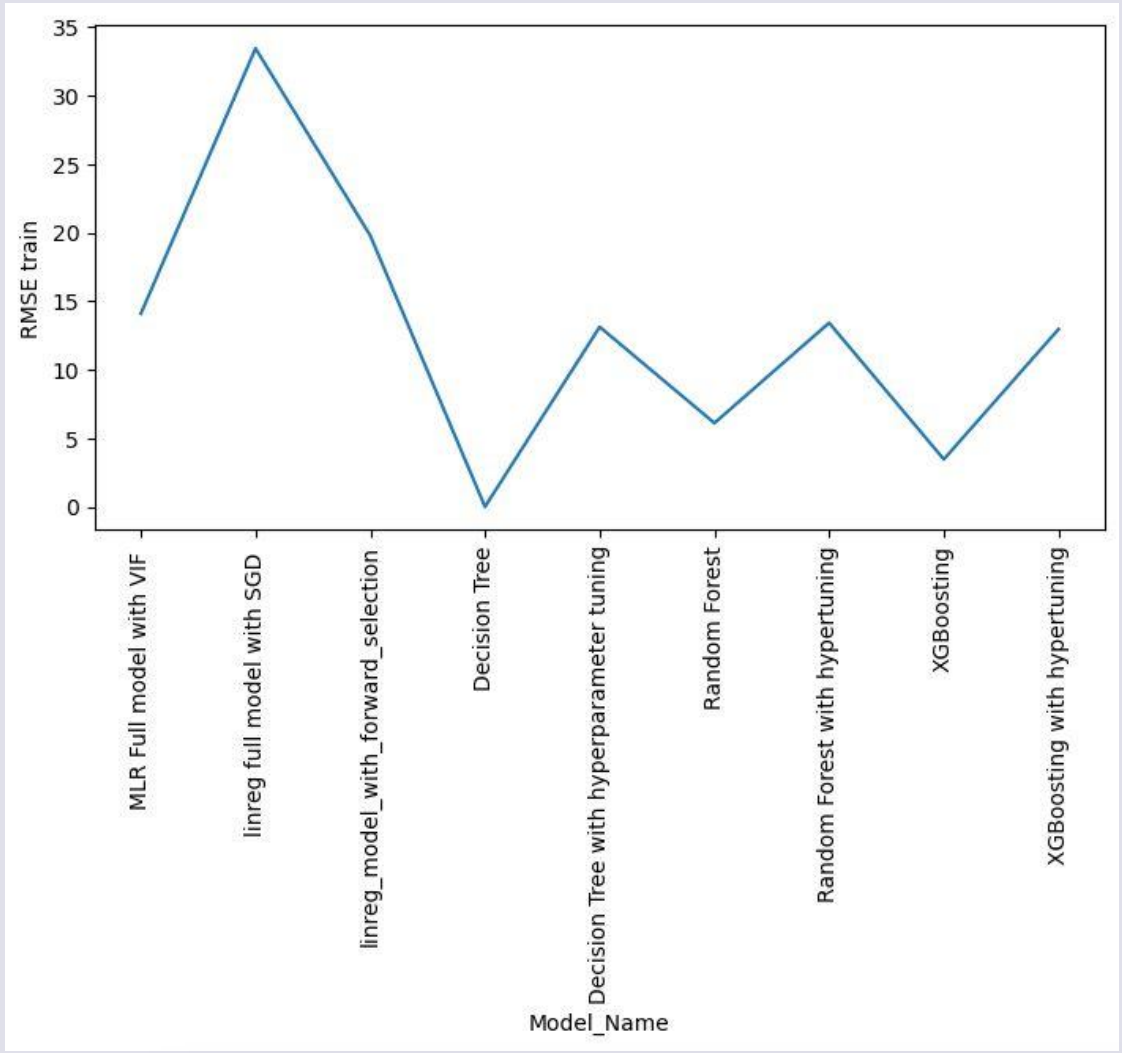
Then we build models with decision tree, Random Forest and XGboosting , all these models
shown overfitting conditions.

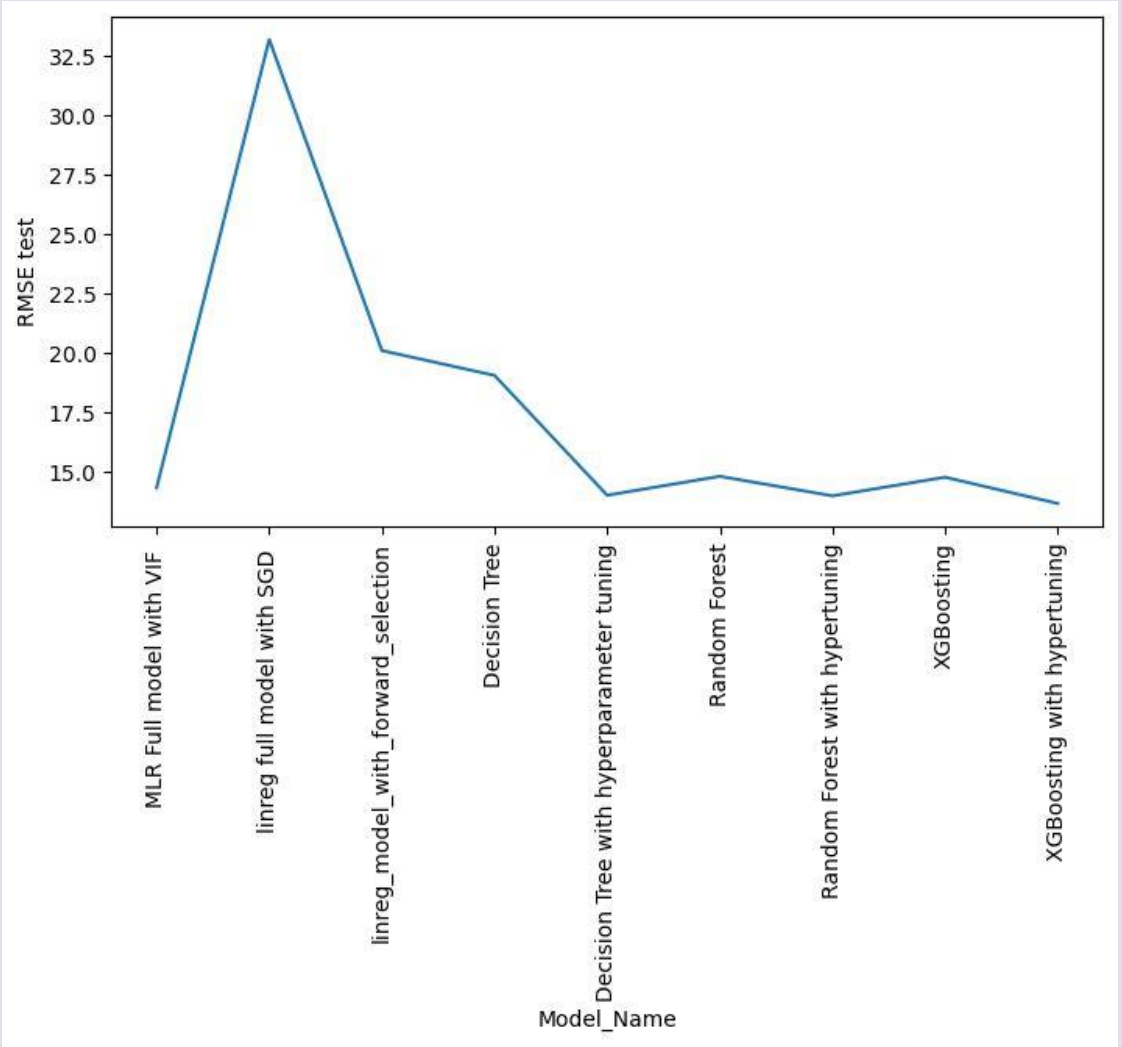
we applied hyperparameter tunning for all three models and after comparing the R2 , adj. R2 , RMSE train and RMSE test, we took XGboosting with hyperparameter tunning as the final model.

Final Model

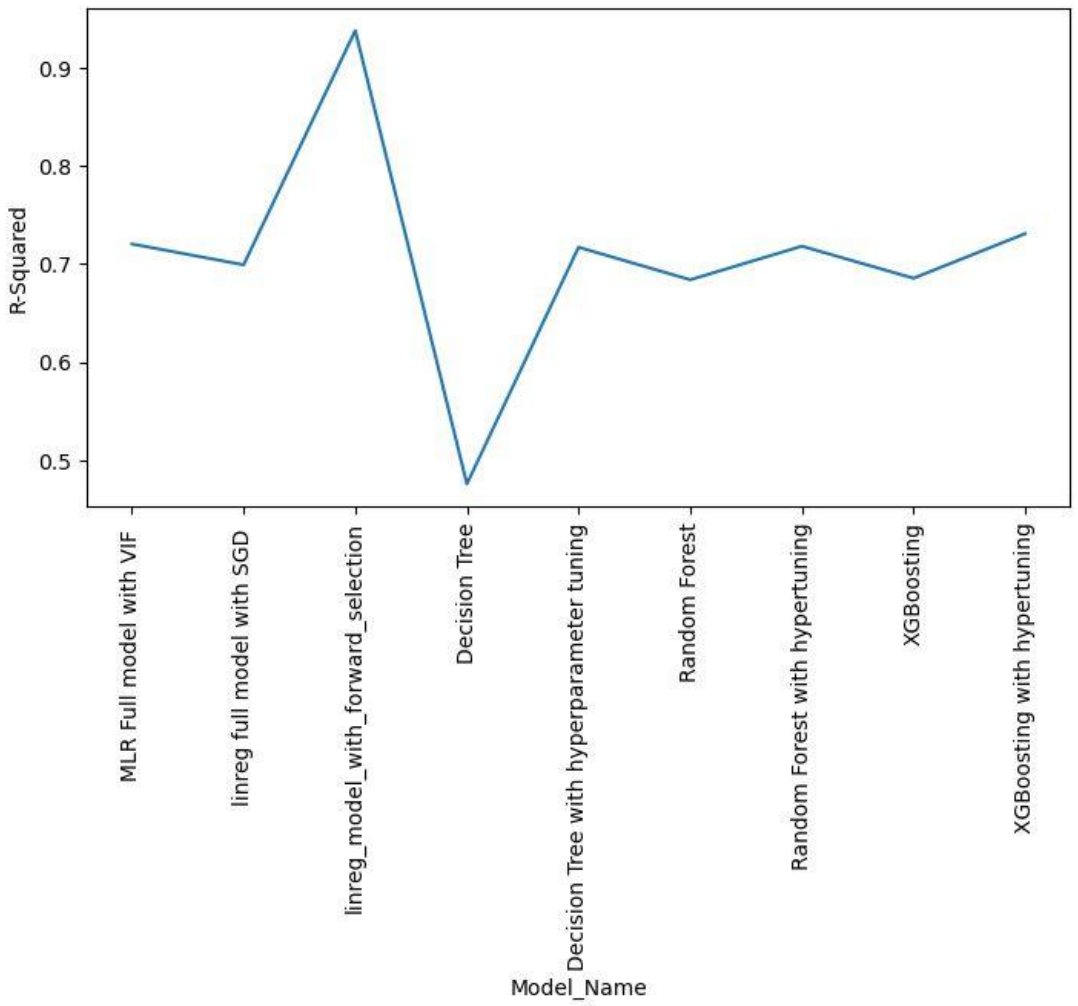
Our final model has R2 0.73 that means our model can explain 73% of the variation in the data.

RMSE train = 13.6 and RMSE test = 12.9 which implies there is no overfitting conditions





- MAPE and RMSE – We have done the hyperparameter tuning so that we can remove overfitted model and found the following results.
- According to RMSE test values our best model is XGBoosting with hypertuning , so we will use this as our final model to predict the sales of BIG MART.



	Model_Name	R-Squared	Adj. R-Squared	RMSE test	RMSE train
0	MLR Full model with VIF	0.719920	0.718853	14.387779	14.068925
1	linreg full model with SGD	0.698507	0.694152	33.305946	33.546567
2	linreg_model_with_forward_selection	0.937778	0.937736	20.087764	19.77997
3	Decision Tree	0.730998	1.054197	14.787682	6.107449
4	Decision Tree with hyperparameter tuning	0.730998	0.707689	13.990758	13.11933
5	Random Forest	0.730998	0.725981	14.787682	6.107449
6	Random Forest with hypertuning	0.730998	0.702561	13.972219	13.416964
7	XGBoosting	0.730998	0.727112	14.750682	3.460161
8	XGBoosting with hypertuning	0.730998	0.727112	13.644038	12.952878

TAKEAWAY

The highest item sale is 13086, lowest item sale is 33 and average sale is 2099 which is comparatively very low than the highest.

Highest cost of item is 266.88 and average is 141.

Visibility of item is not more than 40%

Most item has low fat

Sales of any item is not hugely affected by the fat content.

Dairy and seafood has slightly more sales than other items even though no.of dairy item sold is very much higher than seafood.

Outlet 019 has very less sales compared to other outlet

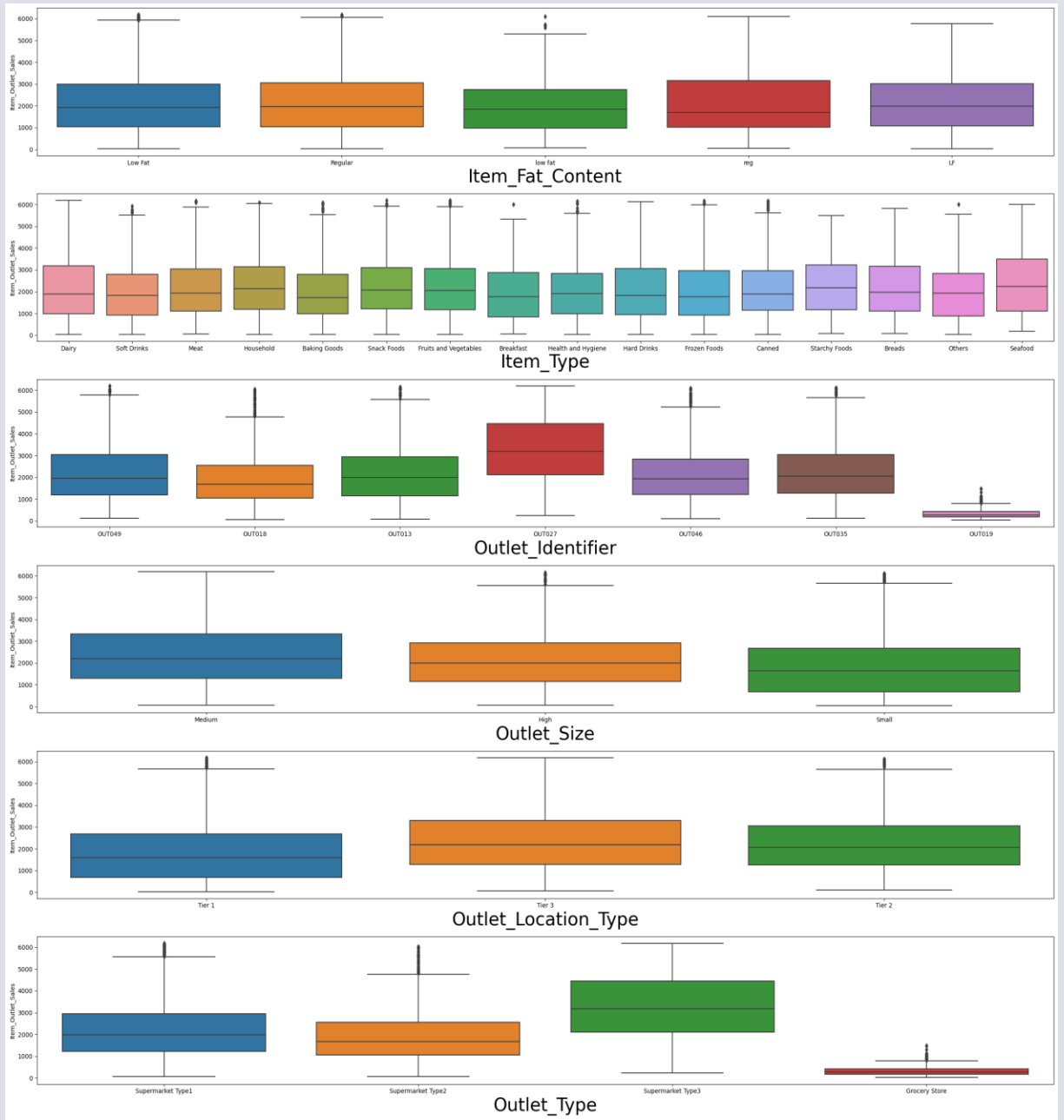
Teir 3 has more sales.

Grocery store has very less sales

Even though the supermarket type 3 and type 2 have less no.of item sold still its contribution in sale is not less than other type

ANALYSIS


Item_Fat_Content
Low Fat
Regular
low fat
reg
LF
Item_Type
Dairy
Soft Drinks
Meat
Household
Baking Goods
Snack Foods
Fruits and Vegetables
Breakfast
Health and Hygiene
Hard Drinks
Frozen Foods
Canned
Starchy Foods
Breads
Others
Seafood
Outlet_Identifier
OUT049
OUT018
OUT013
OUT027
OUT046
OUT035
OUT019
Outlet_Size
Medium
High
Small
Outlet_Location_Type
Tier 1
Tier 3
Tier 2
Outlet_Type
Supermarket Type1
Supermarket Type2
Supermarket Type3
Grocery Store


Item_Outlet_Sales
Item_Fat_Content
Item_Type
Outlet_Identifier
Outlet_Size
Outlet_Location_Type
Outlet_Type

CONCLUSION AND FUTURE STEPS

- **Conclusion:**
  - The sales hugely differ for different locations.
  - There are items with low cost, low sales.
  - Items like dairy, fruits and vegetables counts the most. Seafood has very less count but has huge sales, hence we can conclude that the overall sales of fresh product is good.
  - Grocery product need more attention. We should filter the grocery products based on different location.

- **Future steps:**
  - As we concluded sales dependence on the variables are different in different locations hence, we can make necessary modification in our model for different location for better sales
  - We can include more type of fresh product in the mart as its sales are seen to be more.
  - Taking product feedback from the customers will help to improve in item selection
  - Once the sales of the existing store are increased, stores can also we opened in other locations

THANK YOU!!