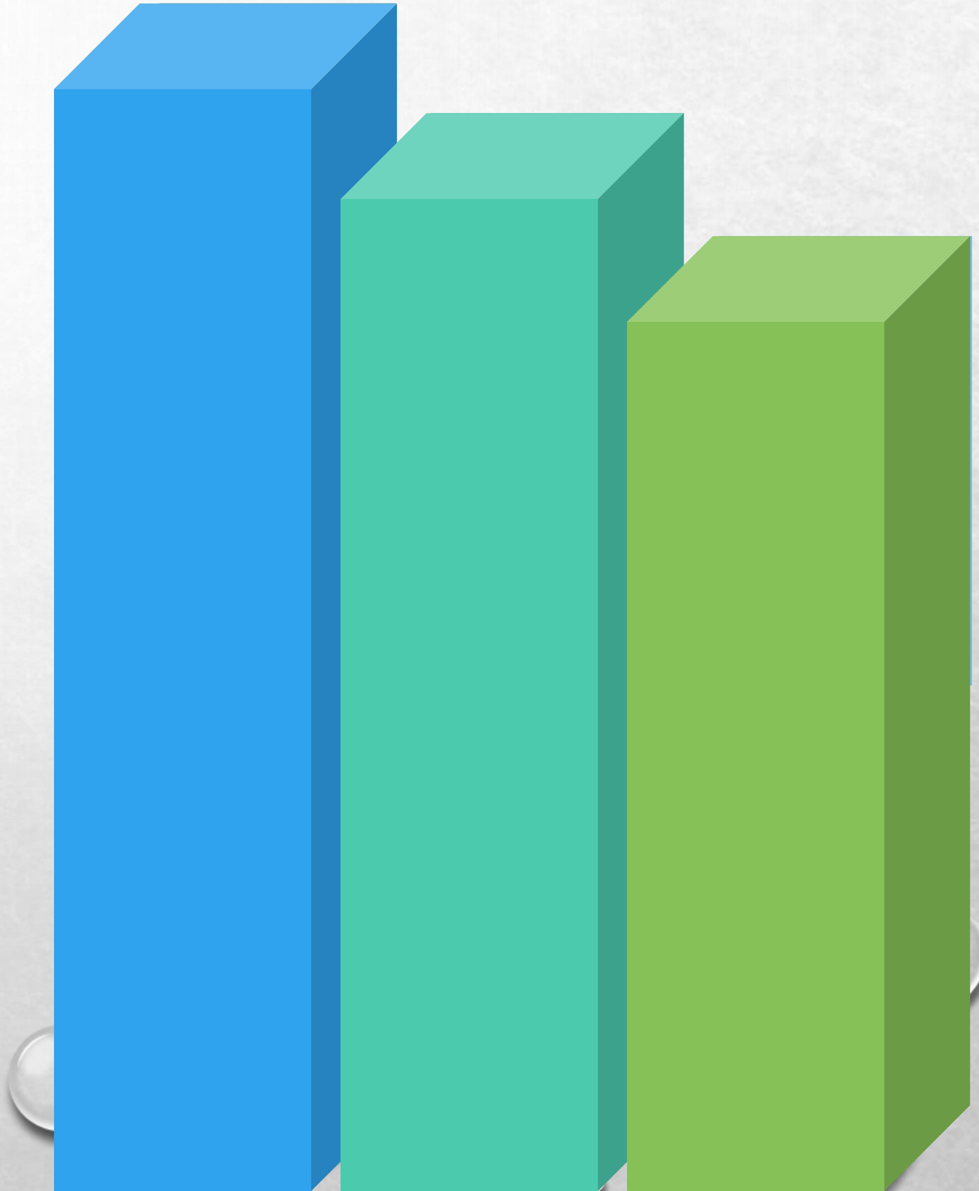
# REGRESSION – PREDICTION OF STORE SALES

PUNEETHKRISHN B

DSFT8



# INTRODUCTION:

- THE SUCCESS OF ANY RETAIL STORE DEPENDS UPON ITS SALES. MORE THE SALES MADE, MORE IS THE REVENUE. WITH A GOOD CUSTOMER SERVICE AND CARE, THE CUSTOMER TOO ENJOYS A GOOD SHOPPING EXPERIENCE.
- THIS WILL LEAD TO MORE IN-FLOW OF CUSTOMERS, OPENING MORE STORE BRANCHES ACROSS A CITY / COUNTRY.
- STORE OWNERS RELY HEAVILY ON PAST DATA TO PREDICT FUTURE SALES. MANY MEDIUM TO LARGE STORES IMPLEMENT THIS KIND OF ANALYTICS TO UNDERSTAND TRENDS

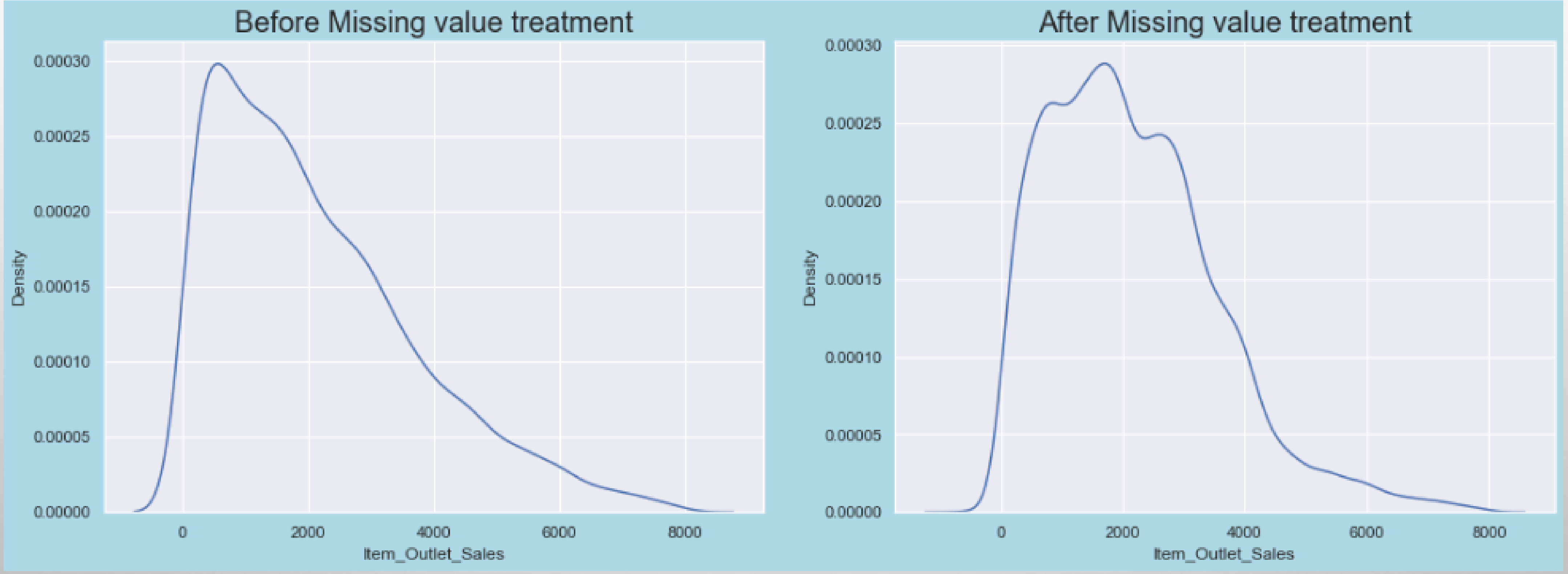
## OBJECTIVE:

- TO HELP THE STORE OWNERS BY ANALYSING PAST DATA OBSERVATIONS AND PROVIDING FUTURE SALES PREDICTIONS

| FEATURE | DATA TYPE | DESCRIPTION |
|---|---|---|
| Item_identifier | Character | Unique Product ID |
| Item_weight | Numeric | Weight of the product |
| Item_Fat_Content | Numeric | Total fat content in the product |
| Item_Visibility | Numeric | How visible is the product in the store |
| Item_Type | Categorical | Product category of the selected product |
| Item_MRP | Numeri | Product cost |
| Outlet_Establishment_Year | Numeric | The year when the store was opened |
| Outlet_Size | Categorical | Size of the store |
| Outlet_Location_Type | Categorical | Location type where the store is located |
| Outlet_Type | Categorical | The type of store |
| Item_Outlet_Sales | Numeric | Sales made by the store outlet |

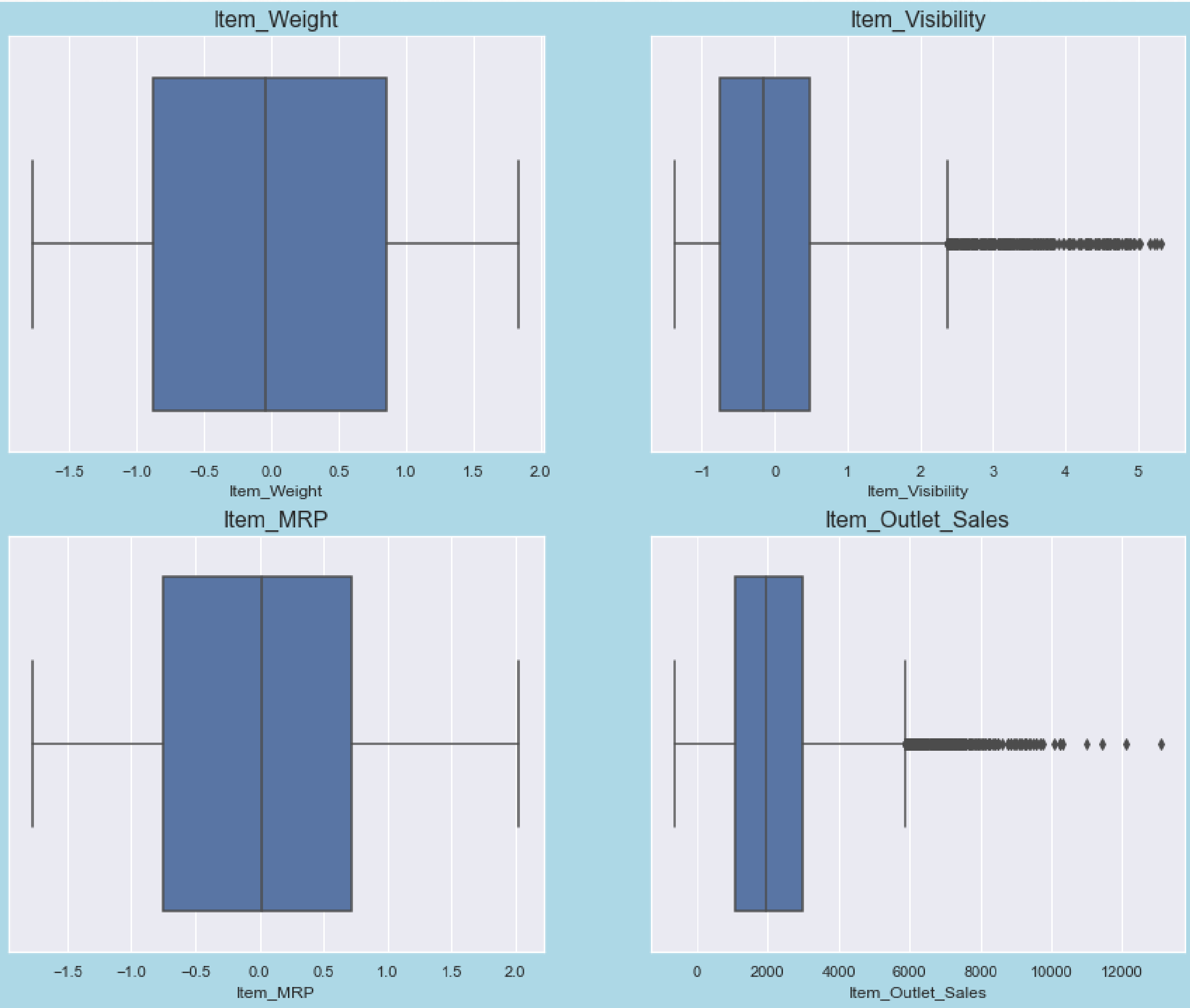
# ➢DATA PRE-PROCESSING

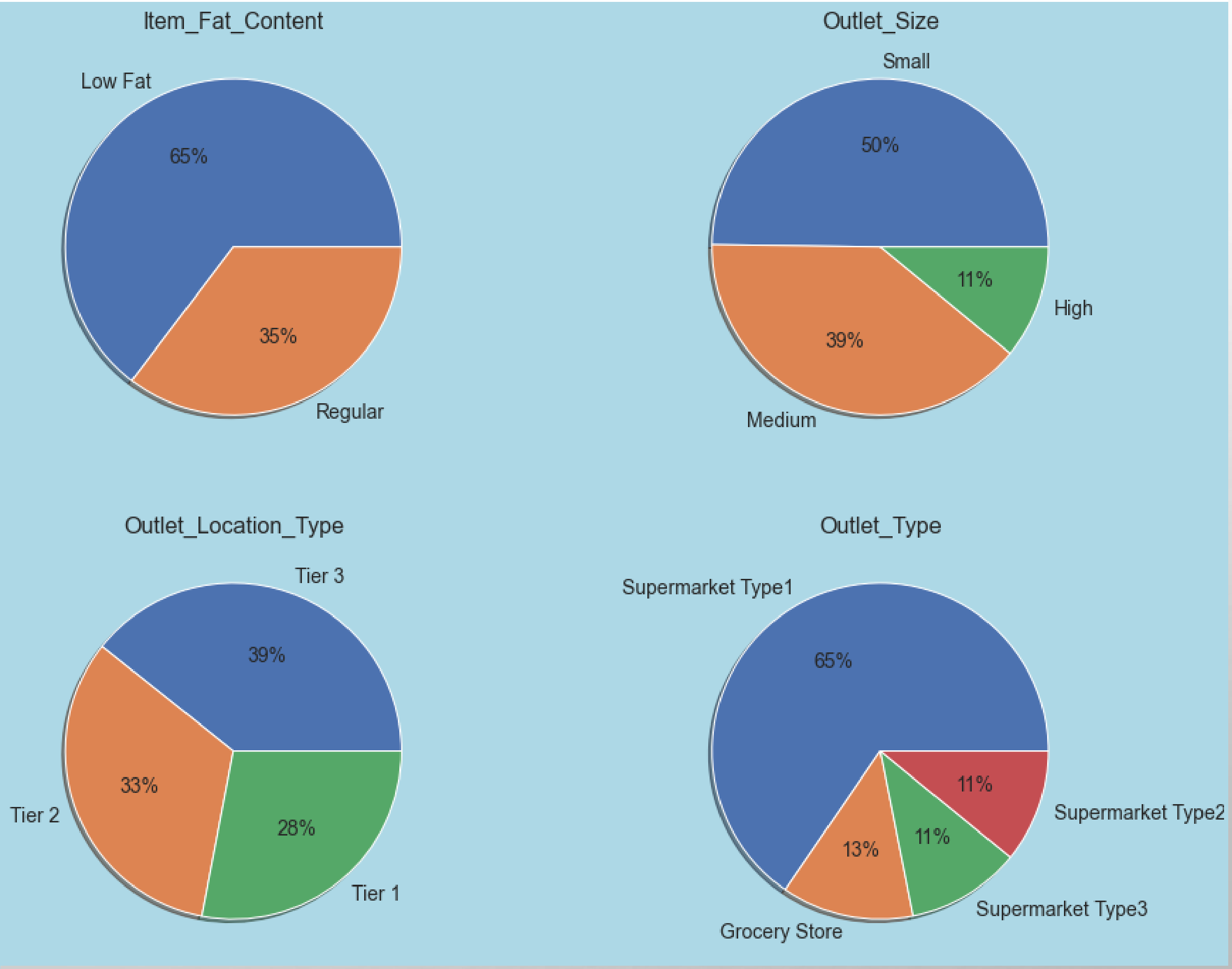
- 40% MISSING VALUES PRESENT IN SALES COLUMN
- THE IMPUTATION OF MISSING VALUE IS DONE BY MICE(MULTIPLE IMPUTATION BY CHAINED EQUATIONS)



# ➢ UNIVARIANT ANALYSIS

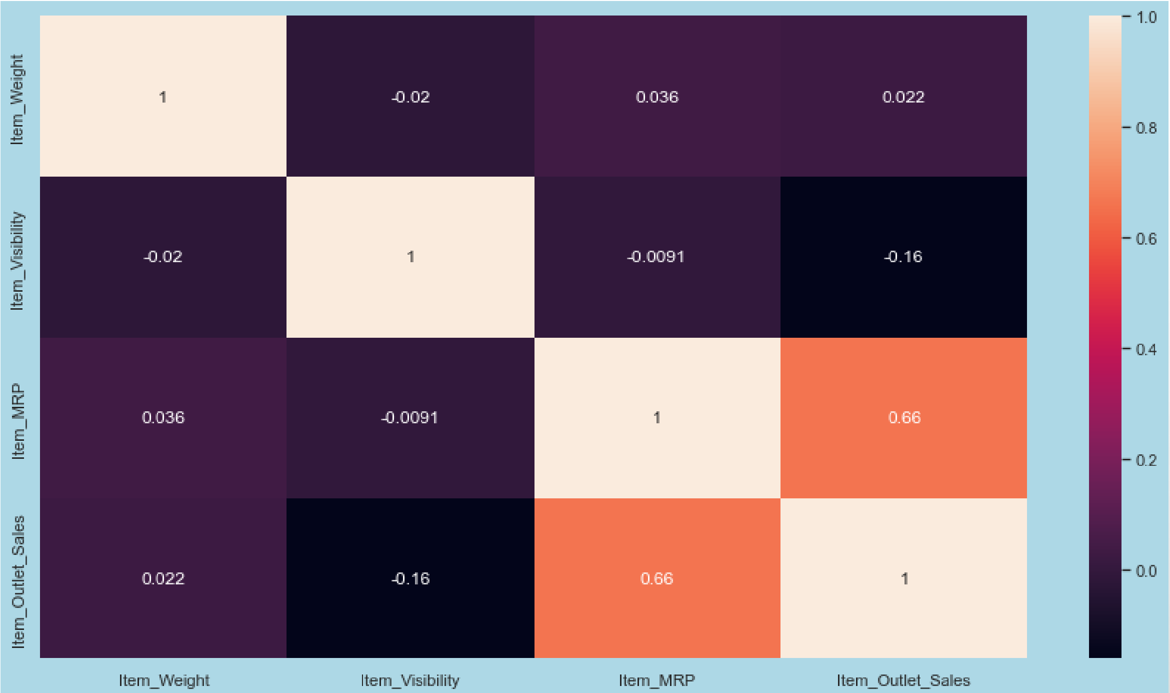
BELOW PLOTS SHOWS THE DISTRIBUTION OF NUMERICAL AND CATEGORICAL FEATURES IN THE DATASET

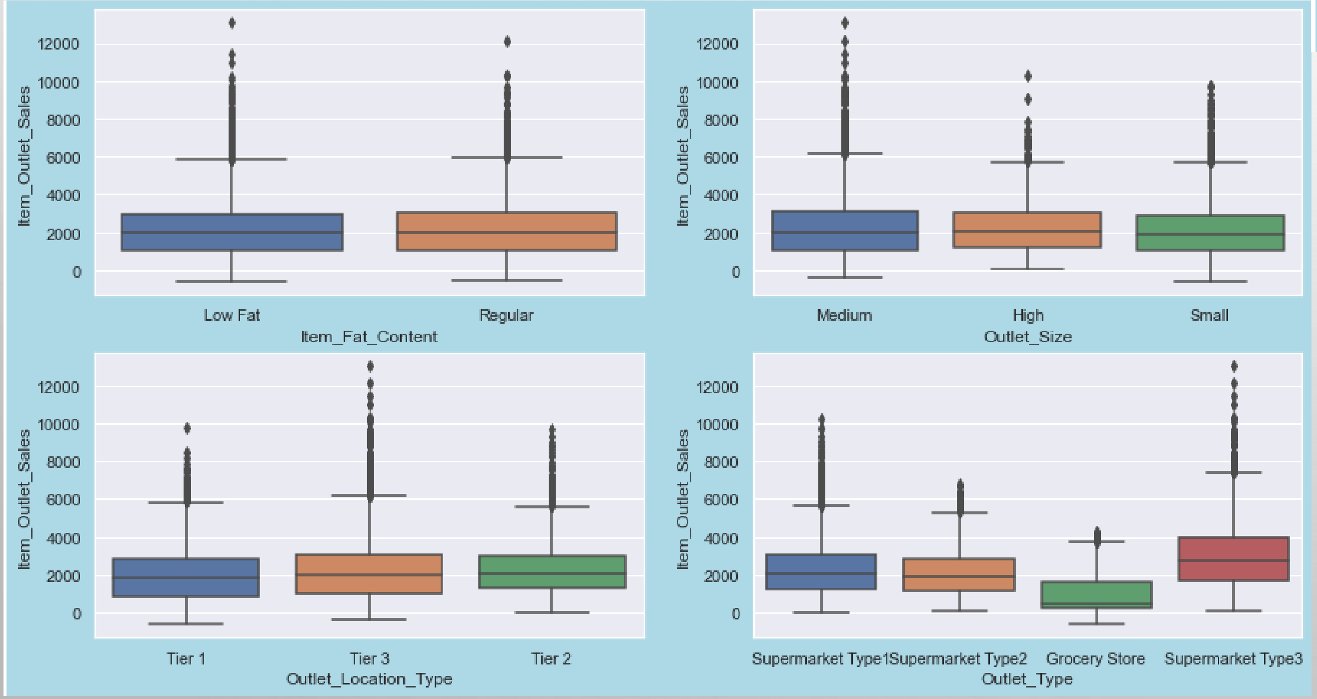




# ➢ BIVARIANT ANALYSIS

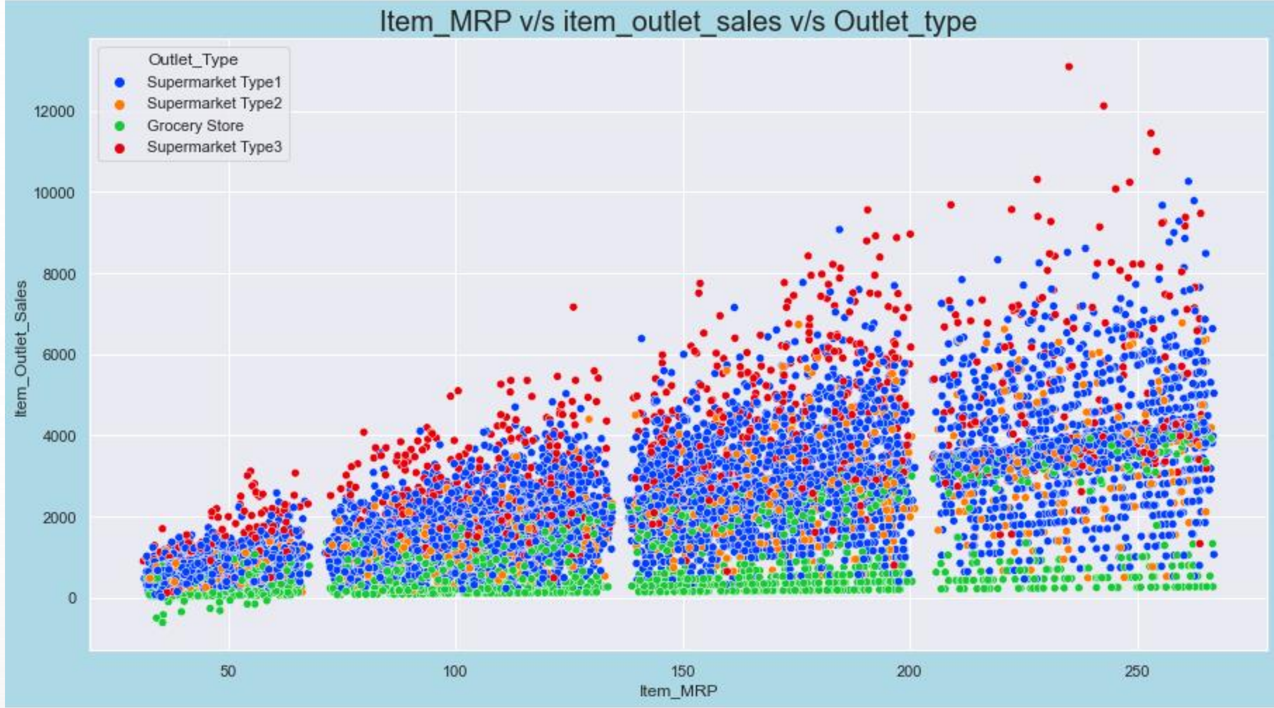
- ➢ CORRELATION WITH RESPECT TO EACH FEATURE

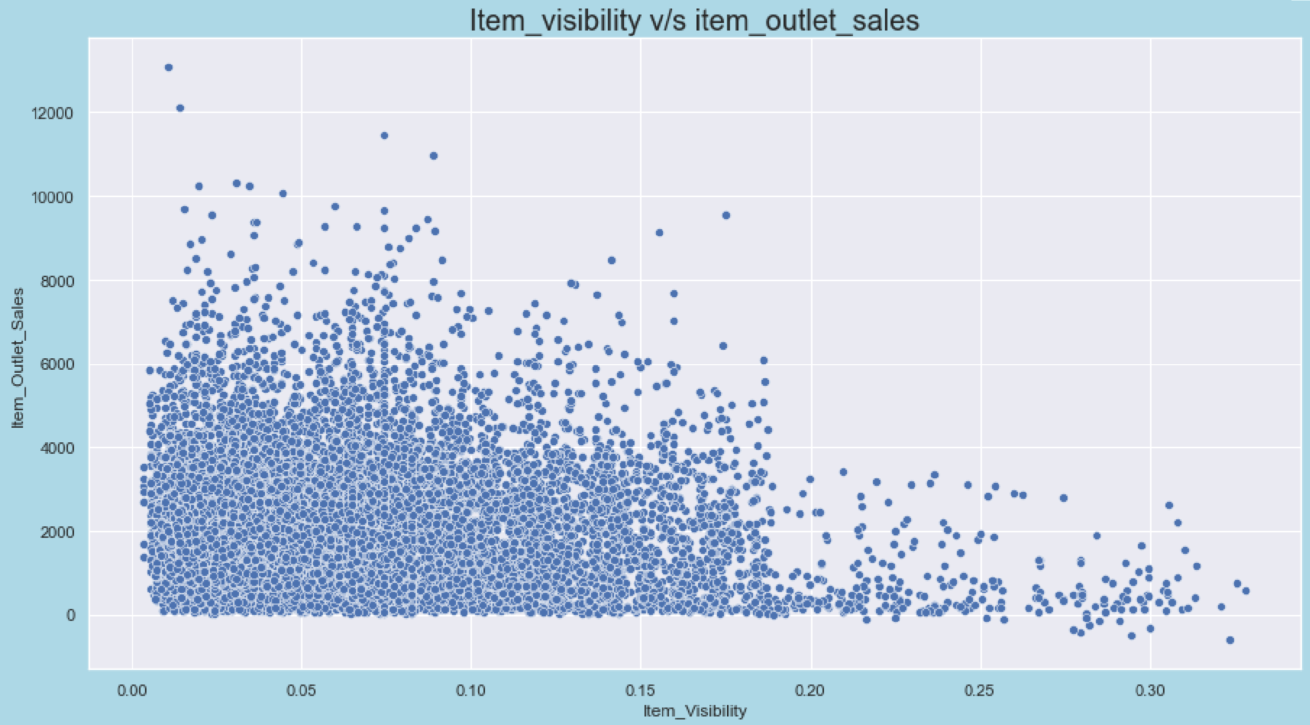




- ➢ DISTRIBUTION OF SALES IN DIFFERENT ASPECTS LIKE FAT CONTENT PRODUCTS, OUTLET SIZE, OUTLET LOCATION TYPE AND OUTLET TYPE

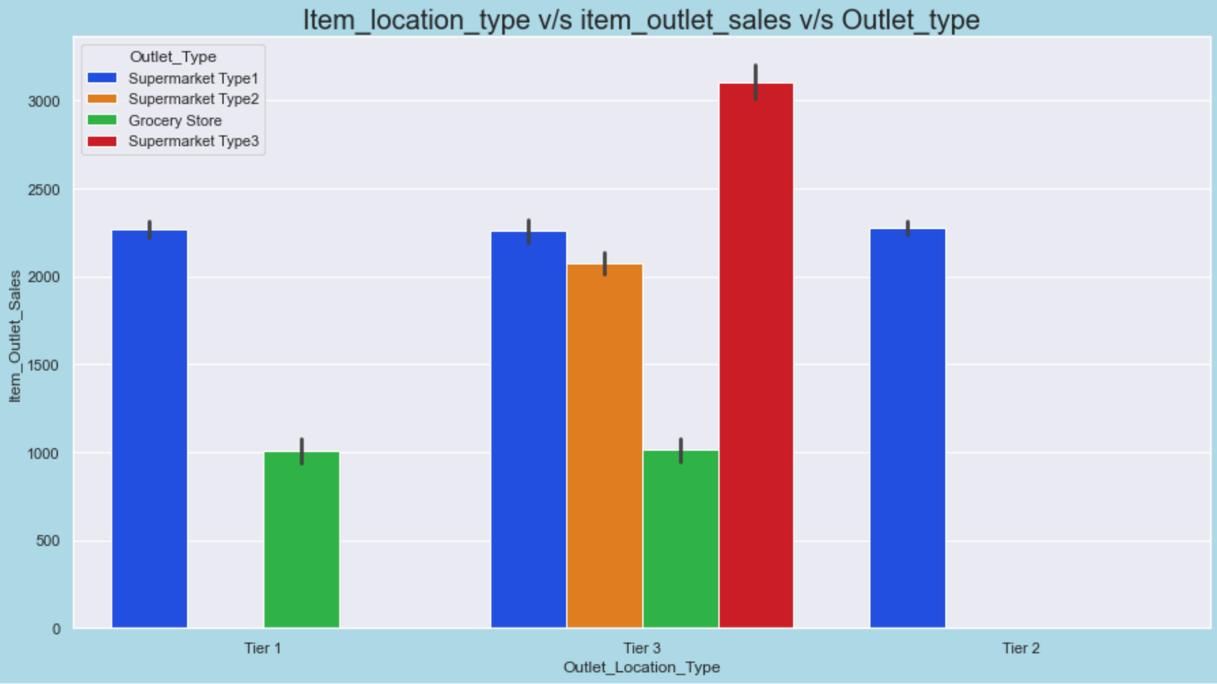
- HERE WE CAN SEE HOW SALES RELATED WRT MRP AND OUTLET TYPE

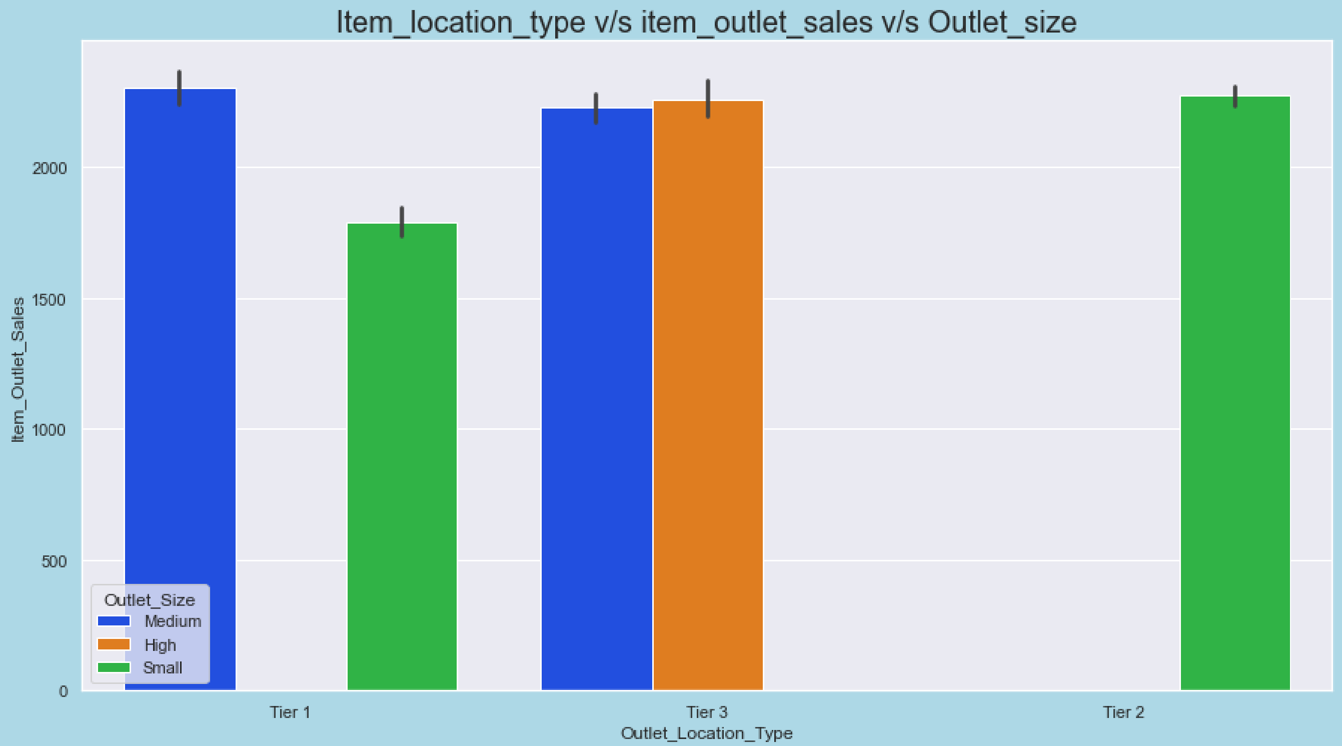




- HERE WE CAN SEE HOW SALES RELATED WITH VISIBILITY OF THE PRODUCTS

- HERE WE CAN SEE HOW SALES RELATED WITH RESPECT TO OUTLET LOCATION TYPE AND OUTLET TYPE





- HERE WE CAN SEE HOW SALES RELATED WITH RESPECT TO OUTLET LOCATION TYPE AND OUTLET SIZE

# FEATURE ENGINEERING

## FEATURE TRANFORMATION

- TO TRANSFORM ALL THE FEATURES INTO NUMERICAL DATATYPE
- LABEL ENCODING TECHNIQUE IS USED FOR FEATURE TRANSFORMATION

## FEATURE SCALING

- TO GET ALL THE FEATURE INTO SIMILAR RANGE
- IN THIS PROJECT THE STANDARDIZATION SCALING TECHNIQUE IS USED BECAUSE THE DATA HAS OUTLIERS AND THE NORMALIZATION IS SENSITIVE TO THE OUTLIERS

## FEATURE SELECTION

- LESS FEATURES ARE AVAILABLE IN THE DATASET AND ALL ARE RELEVANT FEATURES SO NO NEED TO PERFORM FEATURE SELECTION

## SPLITTING TECHNIQUE

- KFOLD TECHNIQUE IS USED IN THIS PROJECT

# ➢MODEL BUILDING

## ➢ EDA OBSERVATIONS

- THE DEPENDENT COLUMN IS CONTINUOUS
- THE OUTLIERS ARE PRESENT IN THE INDEPENDENT FEATURES AS WELL AS DEPENDENT FEATURE
- THE BIAS IS LESS AND VARIANCE IS MORE

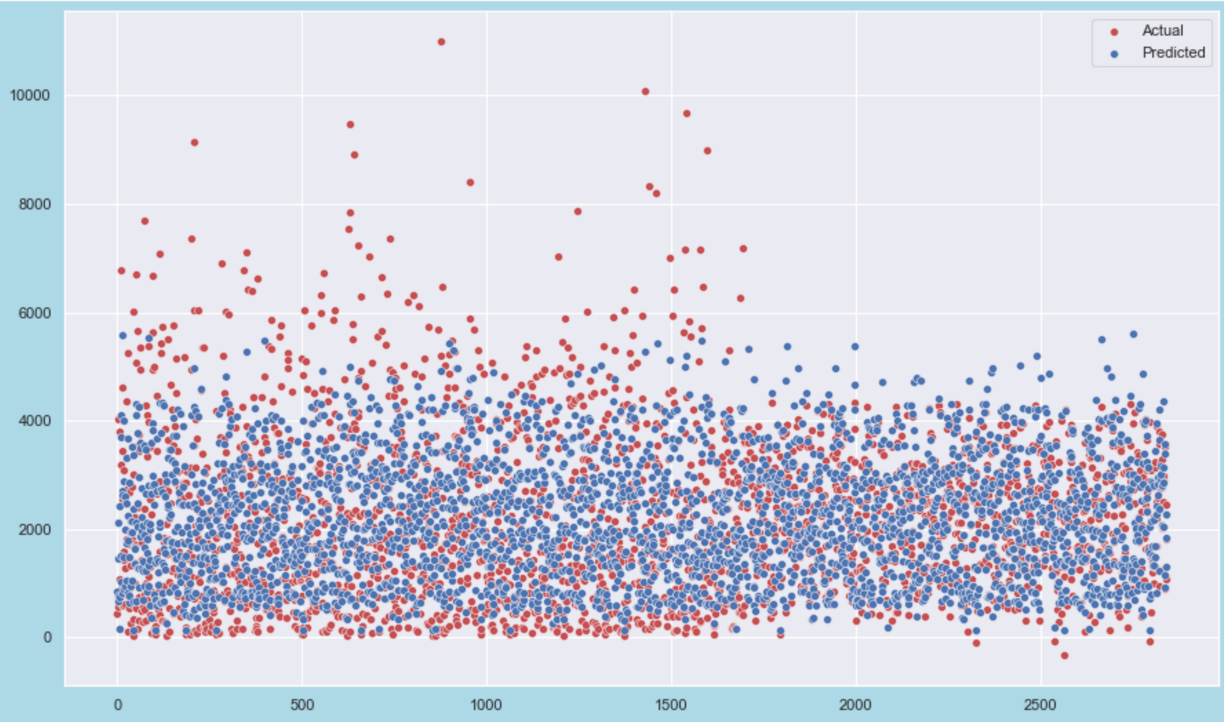
## ➢ ALGORITHM SELECTION

- THE ALGORITHM SHOULD BE REGRESSOR
- CAPABLE OF HANDLING OUTLIERS
- CAPABLE OF MINIMIZING THE VARIANCE
- WE CAN TAKE RANDOM FOREST ALGORITHM WHICH HAS ALL THE CAPABILITIES TO GET BETTER PERFORMANCE TO COMPARE WITH OTHER ALGORITHMS LINEAR REGRESSION AND DECISION TREE REGRESSOR ALGORITHMS ARE SELECTED

# ➢ MODEL PERFORMANCE AND EVALUATION RESULTS

| ALGORITHMS | TRAIN | TEST | MSE | RMSE |
|---|---|---|---|---|
| LINEAR REGRESSION | 0.54 | 0.54 | 975617.63 | 987.73 |
| DECISION TREE REGRESSOR | 0.52 | 0.53 | 975617.63 | 987.73 |
| RANDON FOREST REGRESSOR | 0.61 | 0.59 | 864243.28 | 929.65 |

- FROM THE TABLE WE CAN CONCLUDE THAT RANDOM FOREST REGRESSOR IS SHOWING GOOD PERFORMANCE COMPARED TO ALL OTHER ALGORITHMS



- THE PLOT SHOWS THE ACTUAL AND PREDICTED DATA POINTS WE CAN SEE THE ERROR IS MORE BETWEEN THE DATA POINTS

# ➢ CONCLUSION

- THE LOW FAT PRODUCTS SHOULD BE AVAILABLE IN THE STOCK
- THE VISIBILITY OF THE PRODUCT SHOULD BE LESS
- MORE SALES ARE IN THE LOCATION TYPE OF TIER 3
- MORE SALES ARE IN THE HIGH OUTLET SIZE
- MORE SALES ARE IN THE SUPERMARKET TYPE 3 OUTLET TYPE
- THE RANDOMFORESTREGRESSOR IS GIVING GOOD PERFORMANCE WITHOUT OVERFITTING AND UNDERFITTING
- THE PREDICTIONS ARE GOOD BUT NOT HIGHLY ACCURATE

# THANK YOU