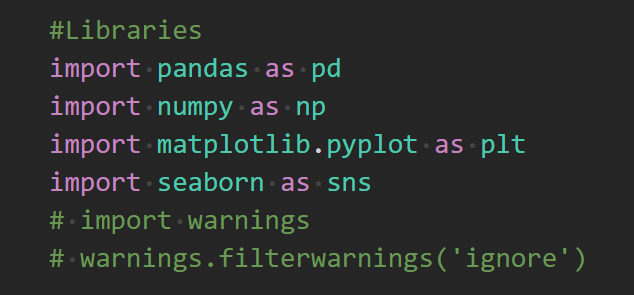
# **PART 0 : IMPORT AND FUNCTIONS**



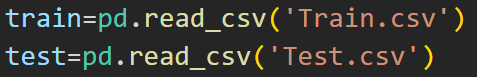
# **PART 1 : PROBLEM UNDERSTANDING - LOOK AT THE BIG PICTURE**

* End goal : "Item\_Outlet\_Sales"
* Type of MLs Approach : Regression
* Feature :

+) Feature of Item: 'Item\_Identifier', 'Item\_Weight', 'Item\_Fat\_Content', 'Item\_Visibility', 'Item\_Type', 'Item\_MRP',

+) Feature of Outlet : 'Outlet\_Identifier', 'Outlet\_Establishment\_Year', 'Outlet\_Size', 'Outlet\_Location\_Type', 'Outlet\_Type'

# **PART 2 : GET AND UNDERSTAND THE DATA  :**

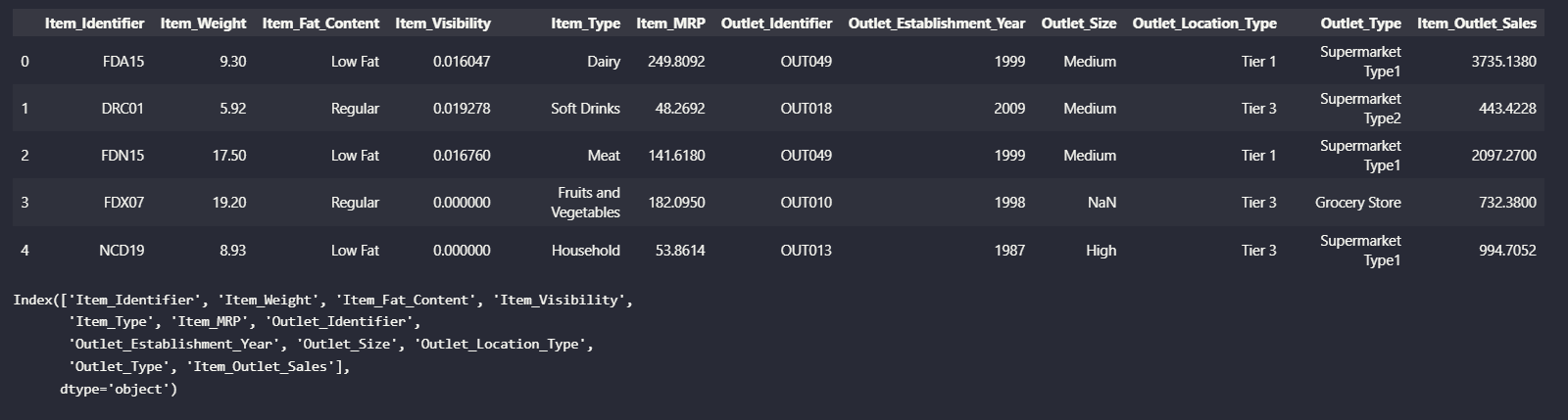


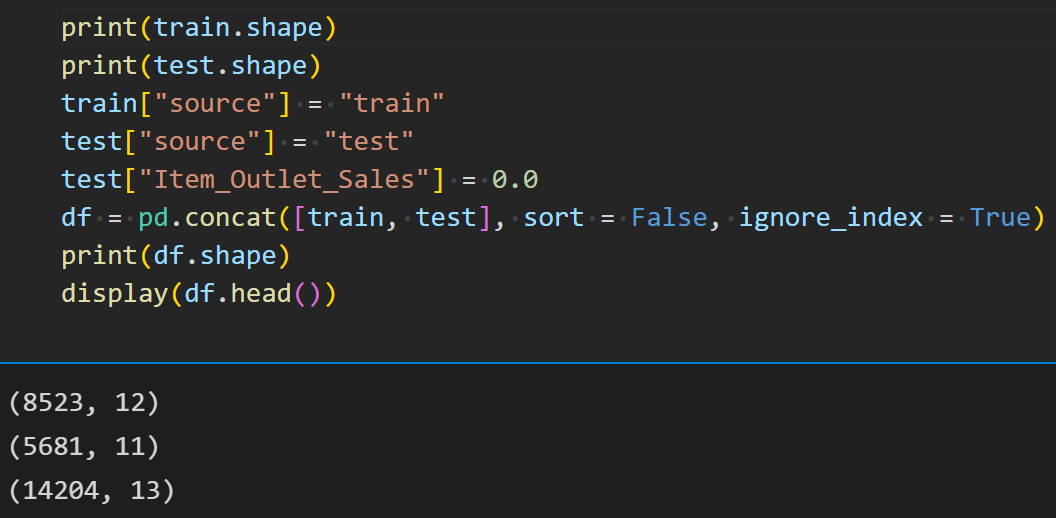
Text

Description automatically generated

# **PART 3. DISCOVER THE DATA TO GAIN INSIGHTS**

## 3.1. Quick view of the data





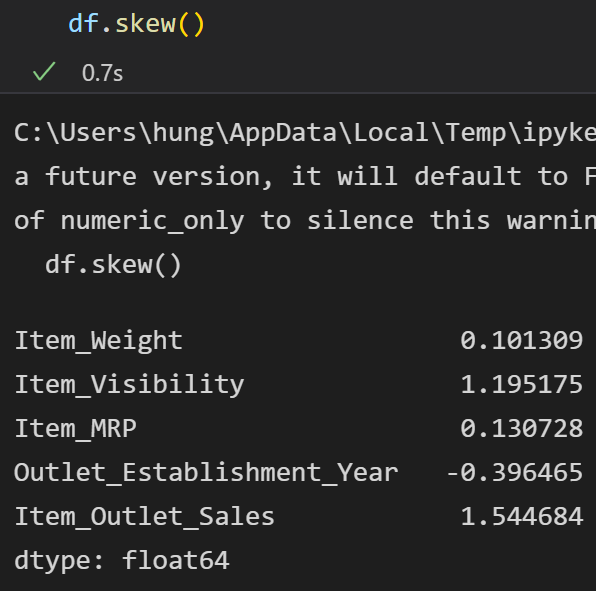
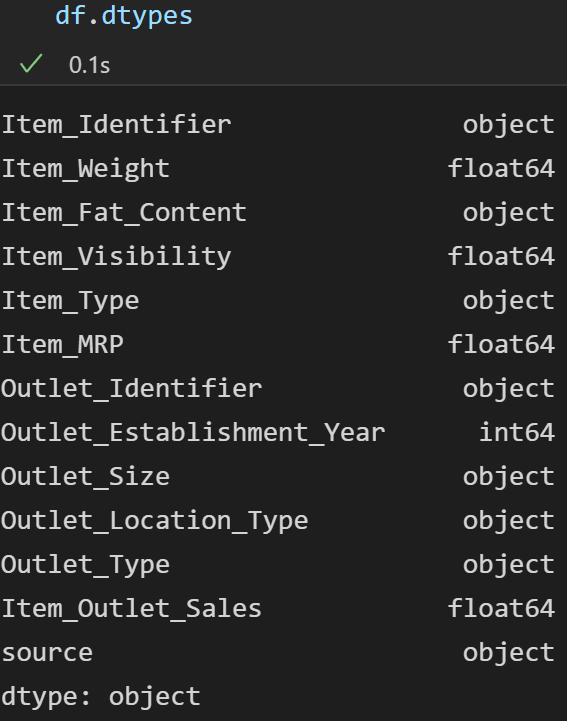
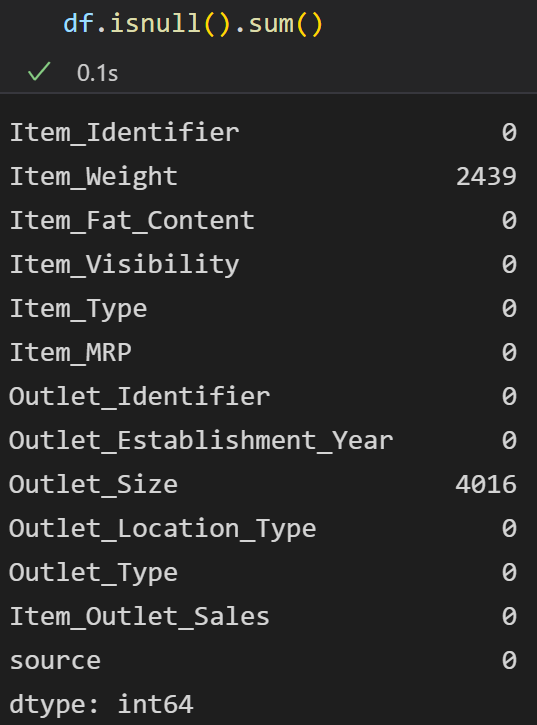
* Train set: 8532 samples, 12 features.
* Test set: 5681 samples, 11 features.
* We will combine train and test data for better Analysis: 14204 samples, 13 features.
* Data Describe: Data of Test does not have target 'Item\_Outlet\_Sales' ==> It should combine train and test into one to make the analysis and transformation. After that, we seperate back to train and test.

## 3.2. Check Missing Values

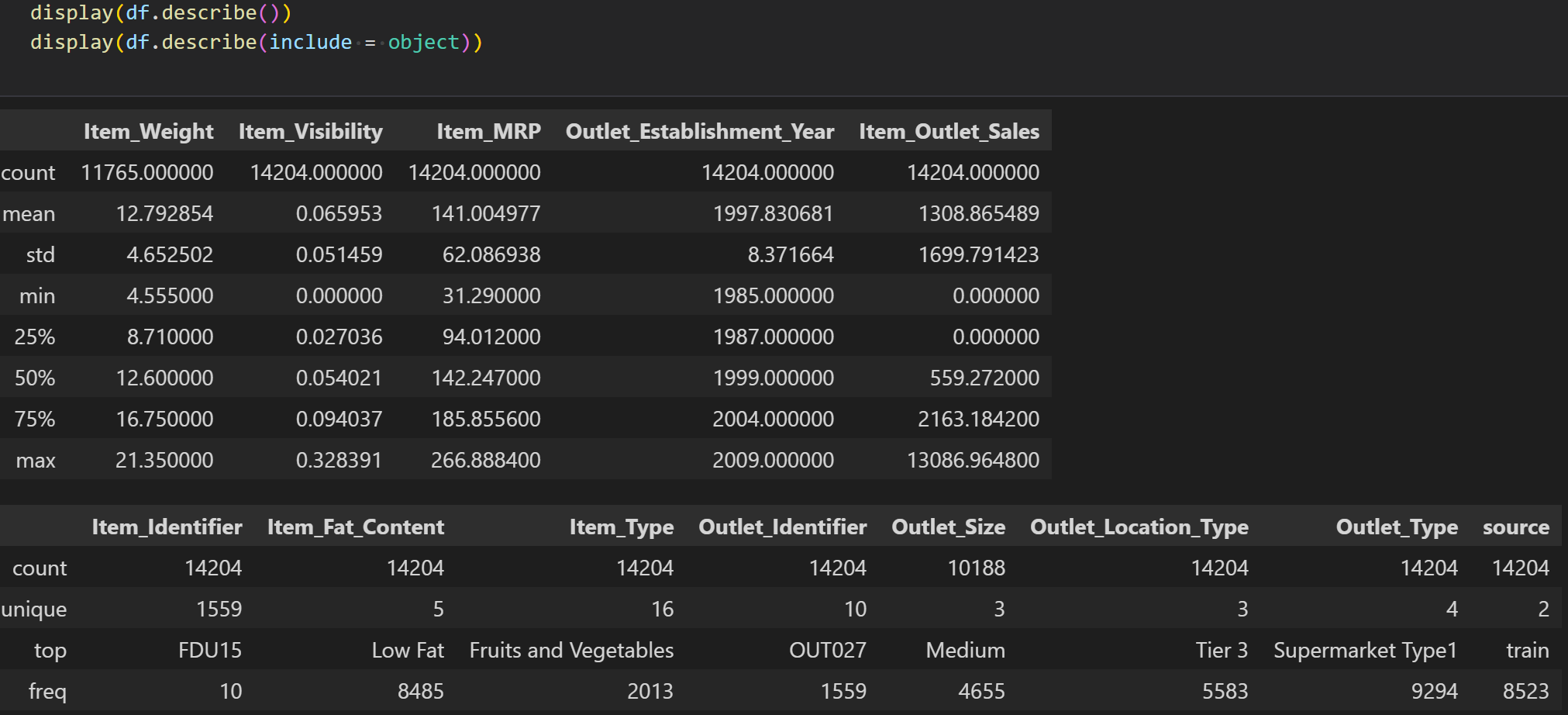
- Item weight column has about 2439 missing values and outlet size column has 4016 NaN values.

- Datset has 8 rows with object type data out of which we have created one and one is an identifier column. others we will have to encode them.

- There is skewness present in the item visibility which need to be handled.



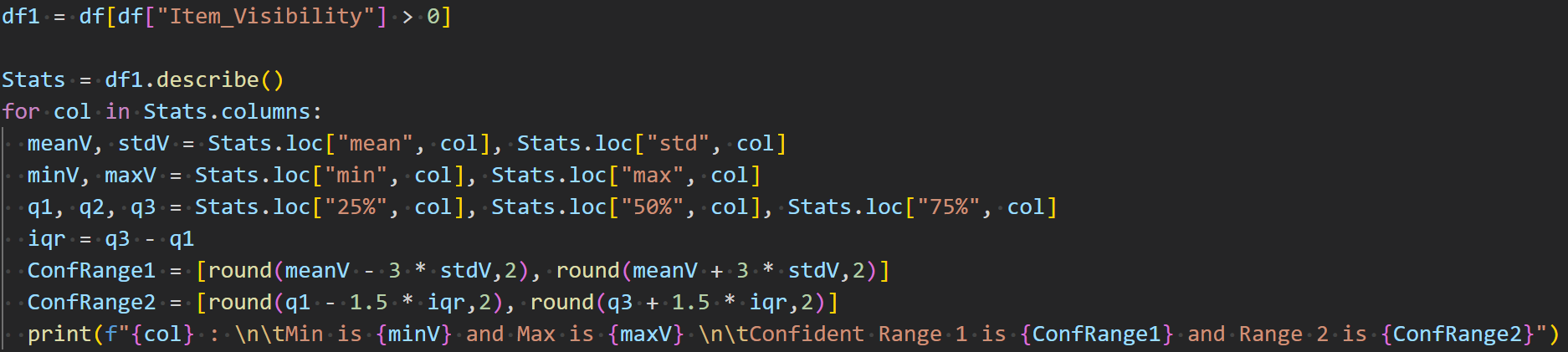
* There are null values present in item weight as count less than total rows, mean is less than median in item MRP and outlet establishment year, rest of the features have median greater than median. Variance is almost zero in item visibility column and very high in item MRP. There are some outliers present as difference between min, max and interquartile range is unequal. Minimum value of visibility is zero that can not be possible so we will treat it as a null value.



* **Missing Values:** Look at count in describe : Missing Value occur at features: Item\_Weight(17.17%) and Outlet\_Size(28.27%)**.** Investigate some unreasonable value at feature : Item\_Visibility with value 0 (6.18%)

## 3.3. Check Outlier(Noise)

- Code : analyze the each number feature



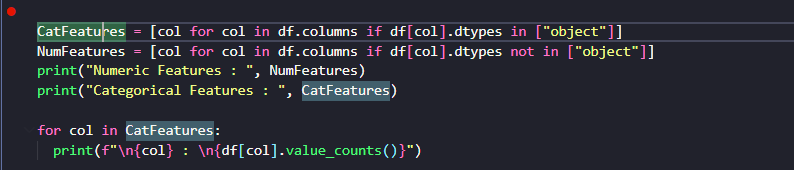
- Result :

Text

Description automatically generated

🡺 Conclusion : Feature having outlier values is Item\_Visibility because the Max value of this feature is out of both Confident Range 1 and Confident range 2 .

## ***3.4. Check Inconsistent***



Text

Description automatically generated

🡺 Inconsistent: Low Fat & LF & low fat is the same meaning, Regular & reg is the same meaning.

## 3.5. Check Imbalanced

Text

Description automatically generated

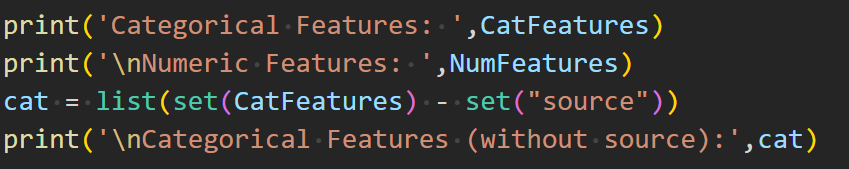
🡺Conclusion : It now skew so much because the skewness is less than 4

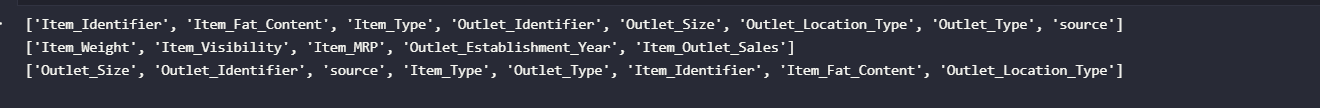
Text

Description automatically generated with low confidence

🡺Conclusion : There is no skewed feature because all their skewness are less than 4

## 3.6. EDA with Visualization on Univariate, BiVariate and MultiVariate Analysis





- Categorical Features : Bar Chart (Comparation), Pie (Percentage Measurements) , [TreeMap, ClusterMap]

- Numeric Features : Histogram (Distribution Shape), Boxplot (Range and Outlier) , [TreeMap, ClusterMap]

## 3.7. Exploring Data Analysis

### 1. Univariate Analysis

**a) Categorical Features**

***Pie (Percentage Measurements) and Bar Chart (Comparation )***

* *“Item\_fat\_content” features :*  
  Chart, bar chart

  Description automatically generated
* *”Outlet\_Identifier” feature :*

Chart, pie chart

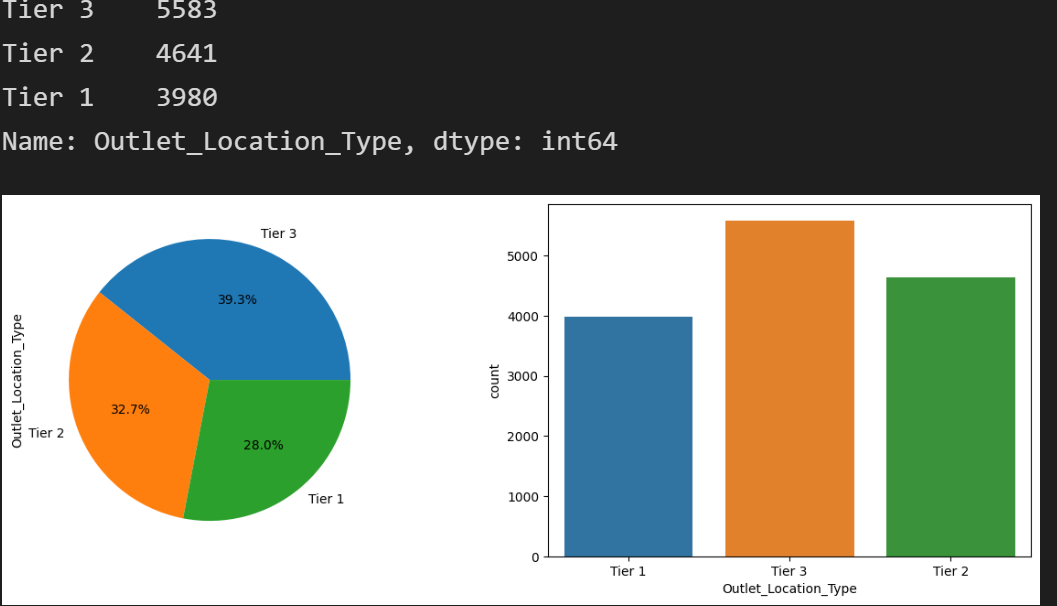
Description automatically generated

* *“Outlet\_size” :* There are few outlets with high size, most of the outlets are medium size.

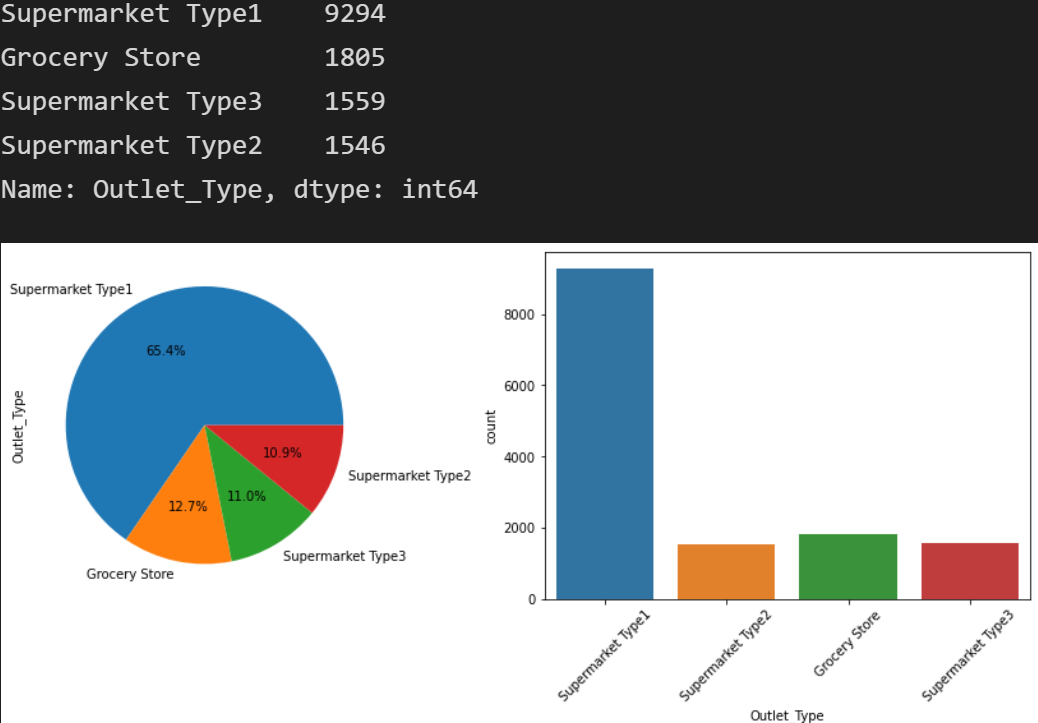
Chart, bar chart

Description automatically generated

* *“Outlet\_Location\_Type”:* Most of the stores are located in tier 3 cities



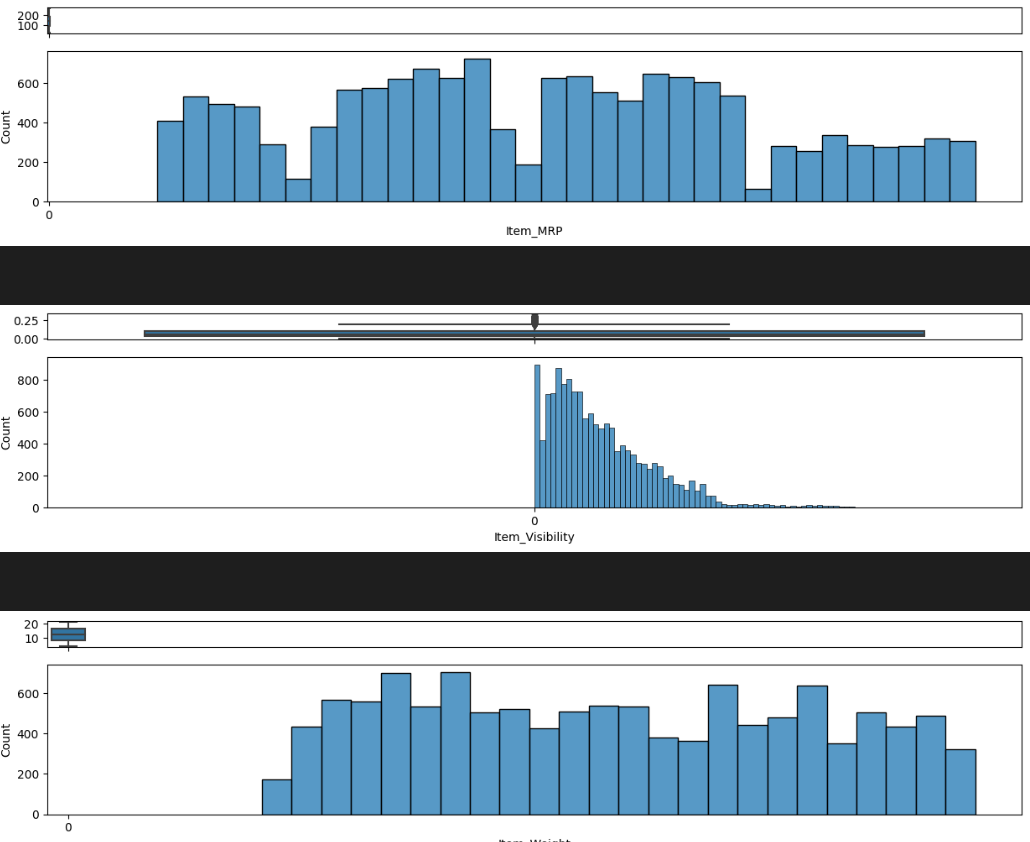
* *“Outlet\_Type”:* 65% of the outlets are of supermarket type 1



1. ***Numeric Features :***

List all the features except “Outlet\_Establishment\_Year” and “Item\_Outlet\_Sales”

Histogram (Distribution Shape) and Boxplot (Range and Outlier) virtualization: Virtualize and show boxplot and Histogram for each feature in the continue array that we filtered out in the previous step :

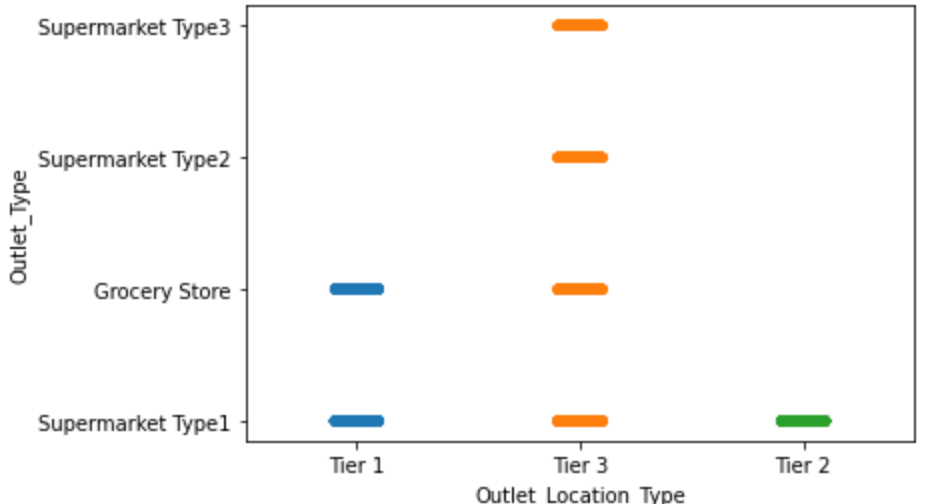


🡺Conclusion : Only Item Visibility has large no. of outliers and it is skewed to the right while others almost follow the gaussian distribution.

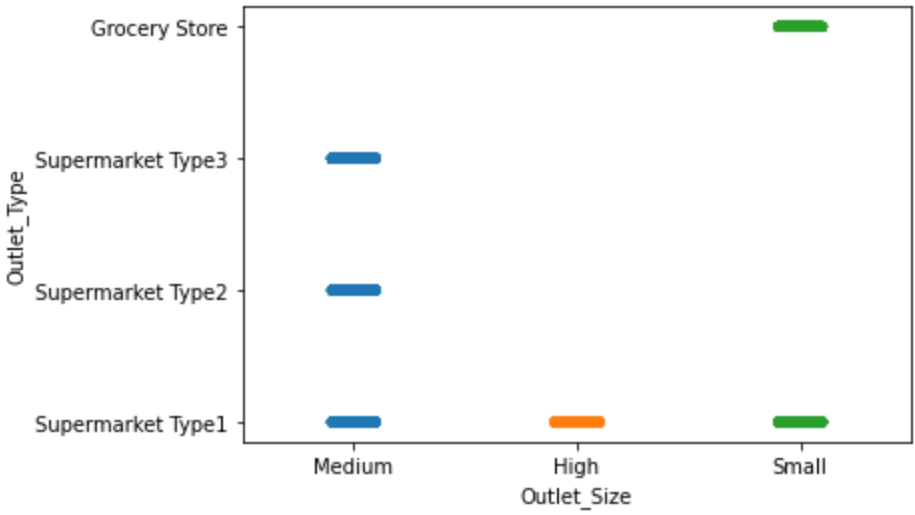
### 2. Bivariate Analysis (Scatter plot b/w 2 features)

***a) Categorical vs Categorical******(stripplot)***

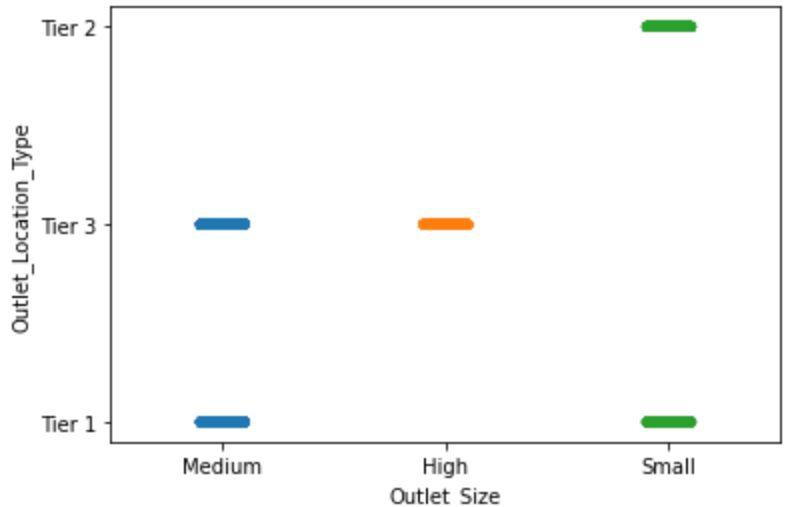
* *Outlet\_Location\_Type vs Outlet\_Type :* Tier 2 cities have only Supermarket type 1 whereas Tier 1 cities have only supermarket 1 and grocery stores



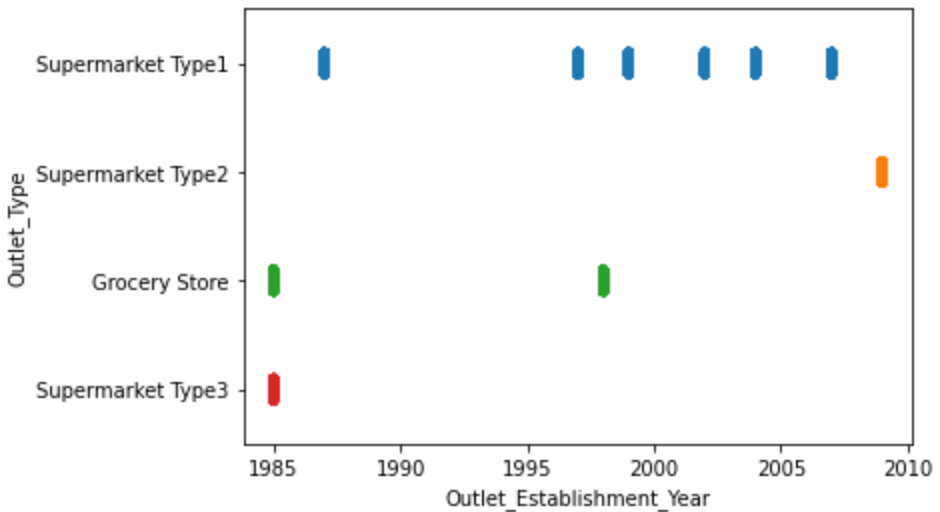
* *Outlet\_Size and Outlet\_Type:* Supermarket type 1 are of all sizes whereas grocery stores are only small and Supermarket type 2, Supermarket type 3 are of medium size only.



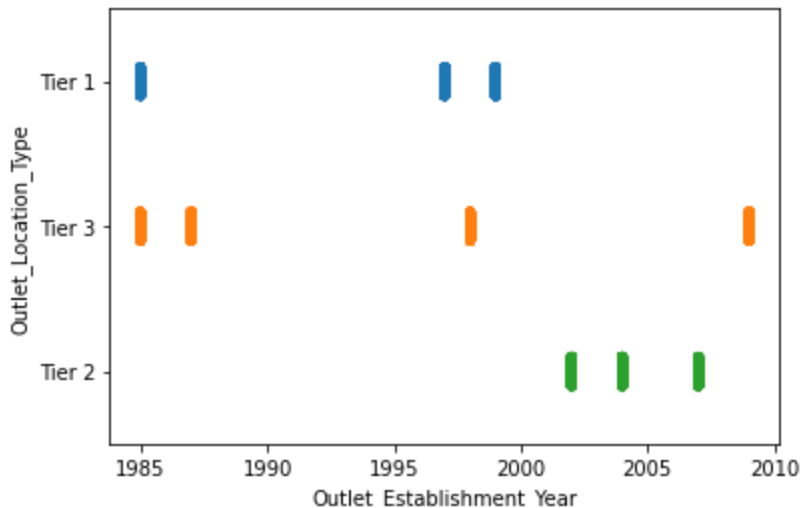
* *Outlet\_Size and Outlet\_Location\_Type: Tier 2 citis have only small outlet size and high outlet size is only found in tier 3 cities*

******

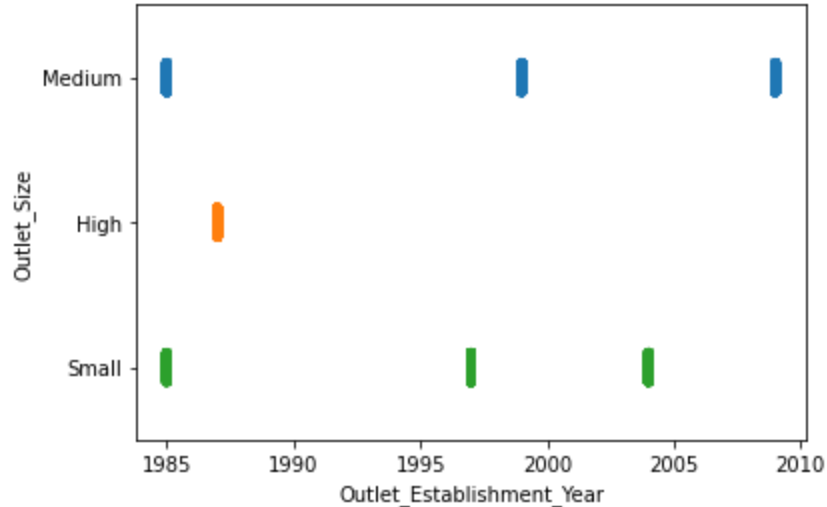
* *Outlet\_Establishment\_Year and Outlet\_Type:* Supermarket type 2 was build much later while grocery stores and supermarket are the oldest outlet type.

**

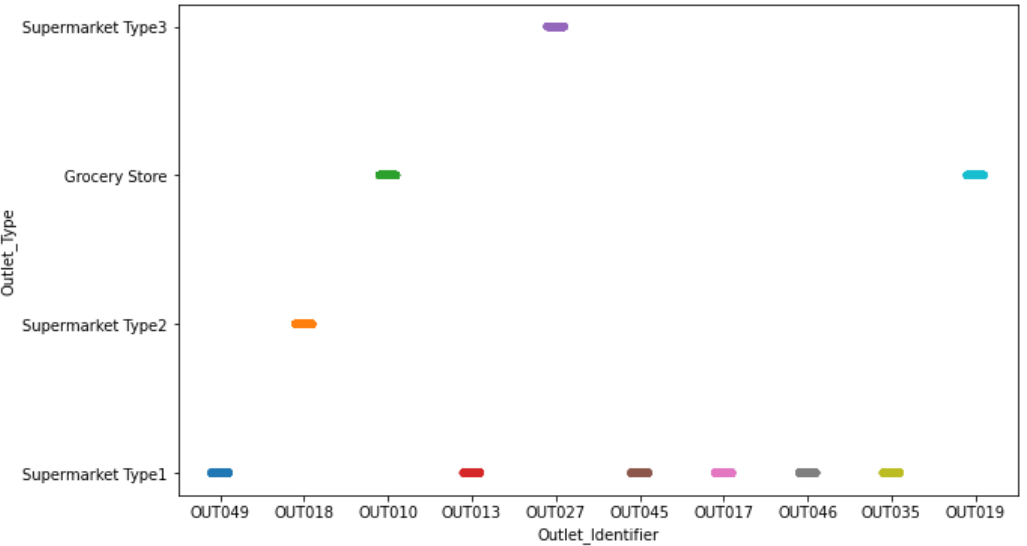
* *Outlet\_Establishment\_Year and Outlet\_Location\_Type:* In Tier1 and tier2 cities outlets were established in 1985 whereas tier2 got outlets after 2000

**

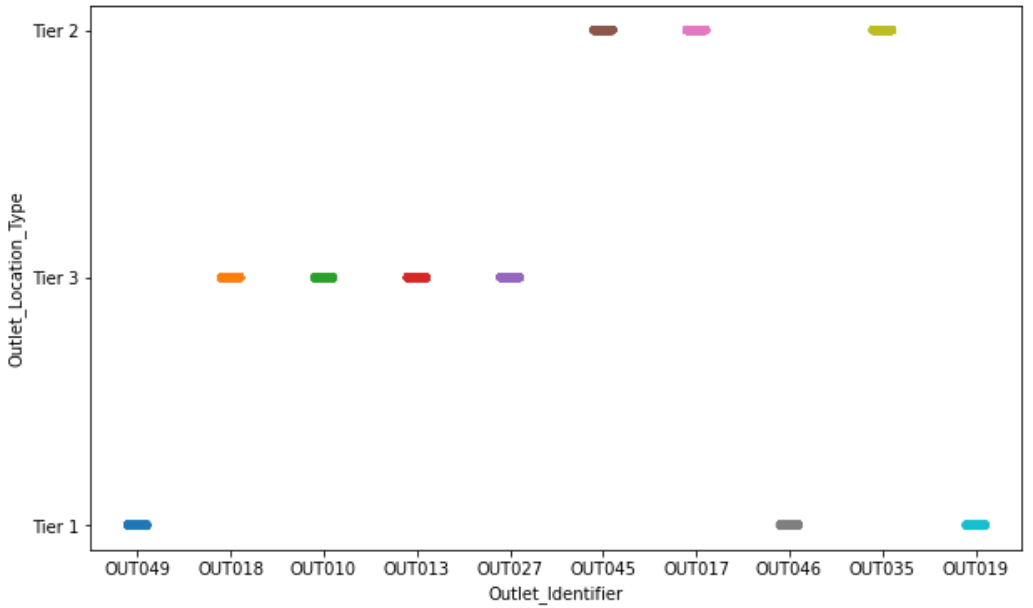
* *Outlet\_Establishment\_Year and Outlet\_Size: After 1990 no outlet of high size was established.*



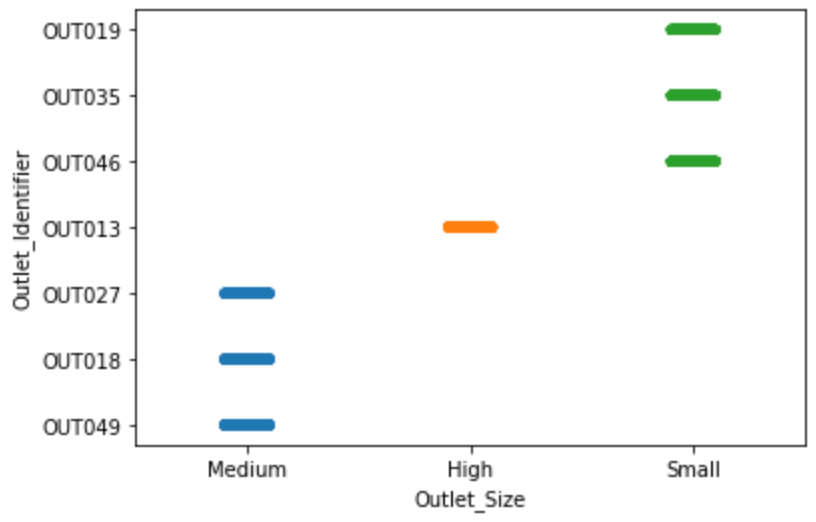
* *Outlet\_Identifier and Outlet\_Type:* There is only one outlet identifier for supermarket 2 and 3 while most of the outlet identifiers belong to supermarket1

******

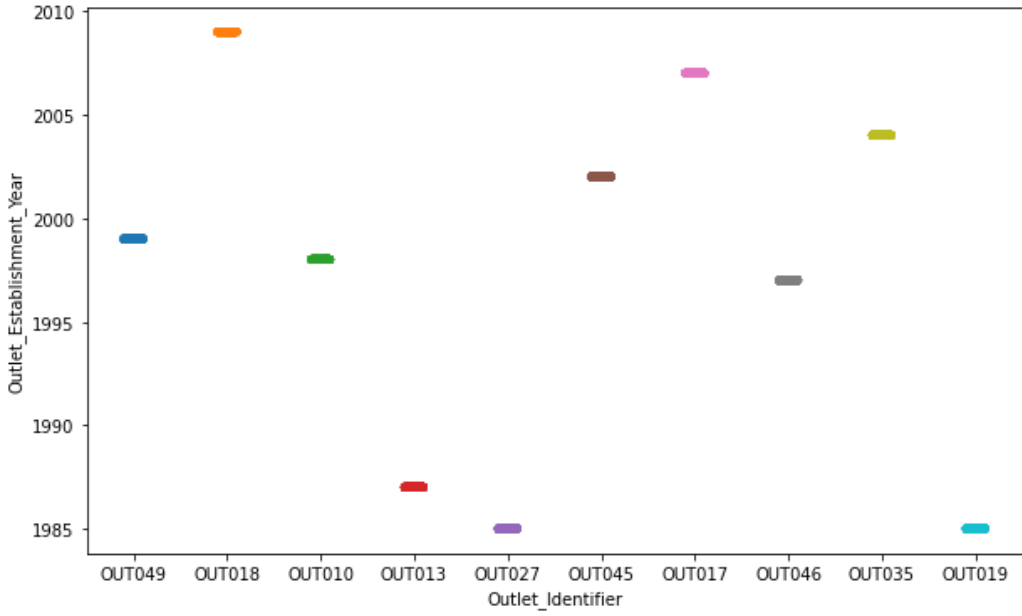
* *Outlet\_Identifier and Outlet\_Location\_Type:* Tier 3 has the most different types of outlets, but they are almost balanced.



* *Outlet\_Size and Outlet\_Identifier: There is only Outlet13 with high outlet size while medium and small of outlets each have 3 outlet identifiers.*

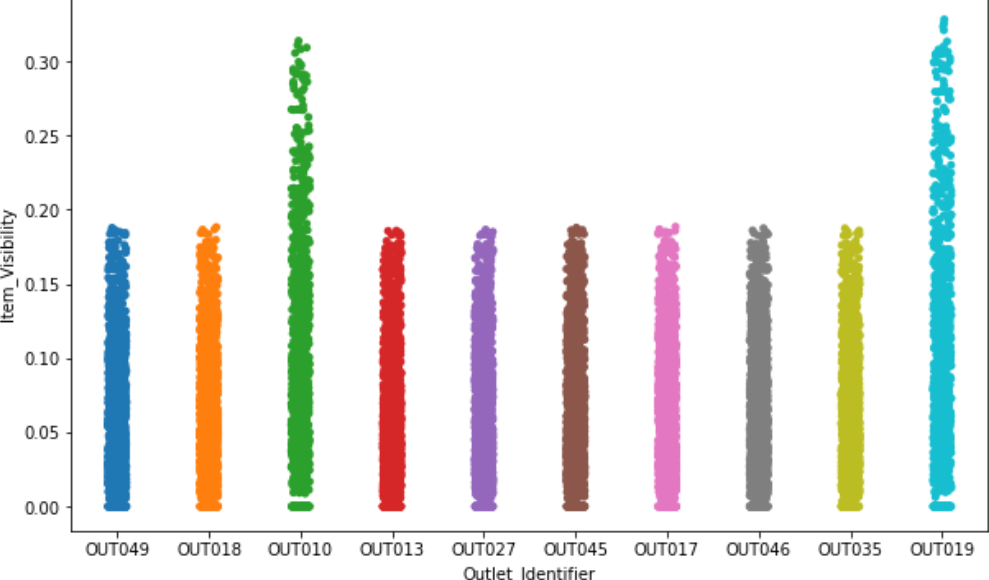
**

* *Outlet\_Identifier and Outlet\_Establishment\_Year:* Outlet 27 and 19 are the oldest outlet identifier and outlet 18 is the newest, outlet no. does not hold any order towards year of establishment.

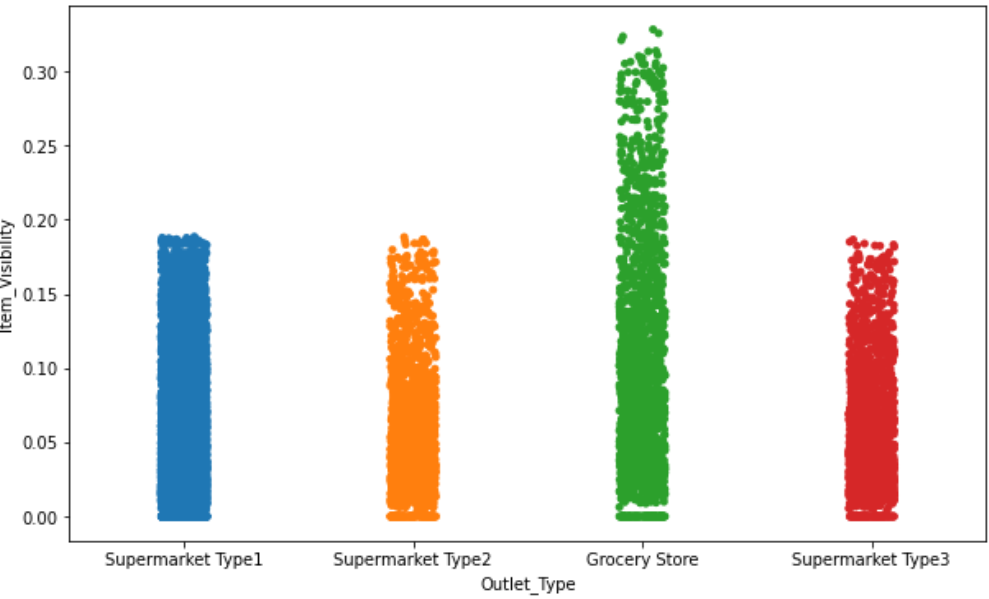
**

***b) Categorical vs Numeric :***

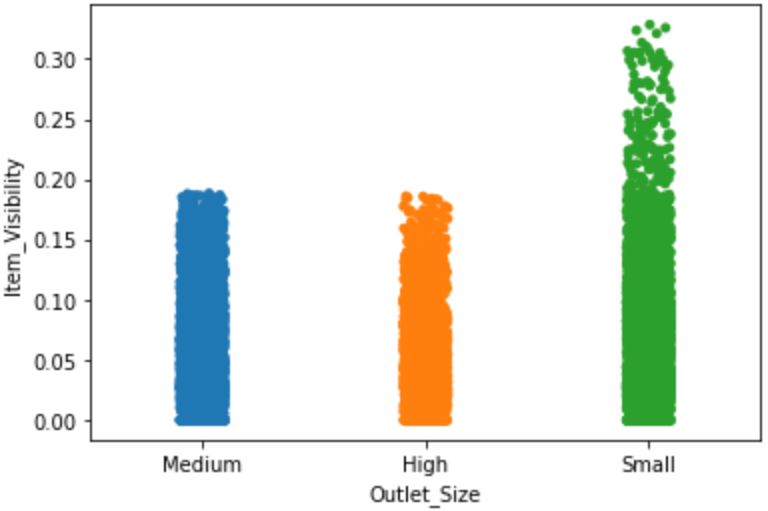
* *Outlet\_Identifier vs Item\_Visibility : Outlet 10 and 19 have give the highest visibilty to products. Other outlets provide almost equal visibility.*

******

* *Outlet\_Type vs Item\_Visibility :* Products are most visible in grocery stores rather than any other super markets



* *Outlet\_Size and Outlet\_SizeItem\_Visibility:* Smallest Stores provide the most visibility to products



1. ***Numeric vs Numeric:***

* *Item\_Visibility vs Item\_Outlet\_Sales*: Item outlet sales sharply decreases for the most visible items

Chart, scatter chart

Description automatically generated

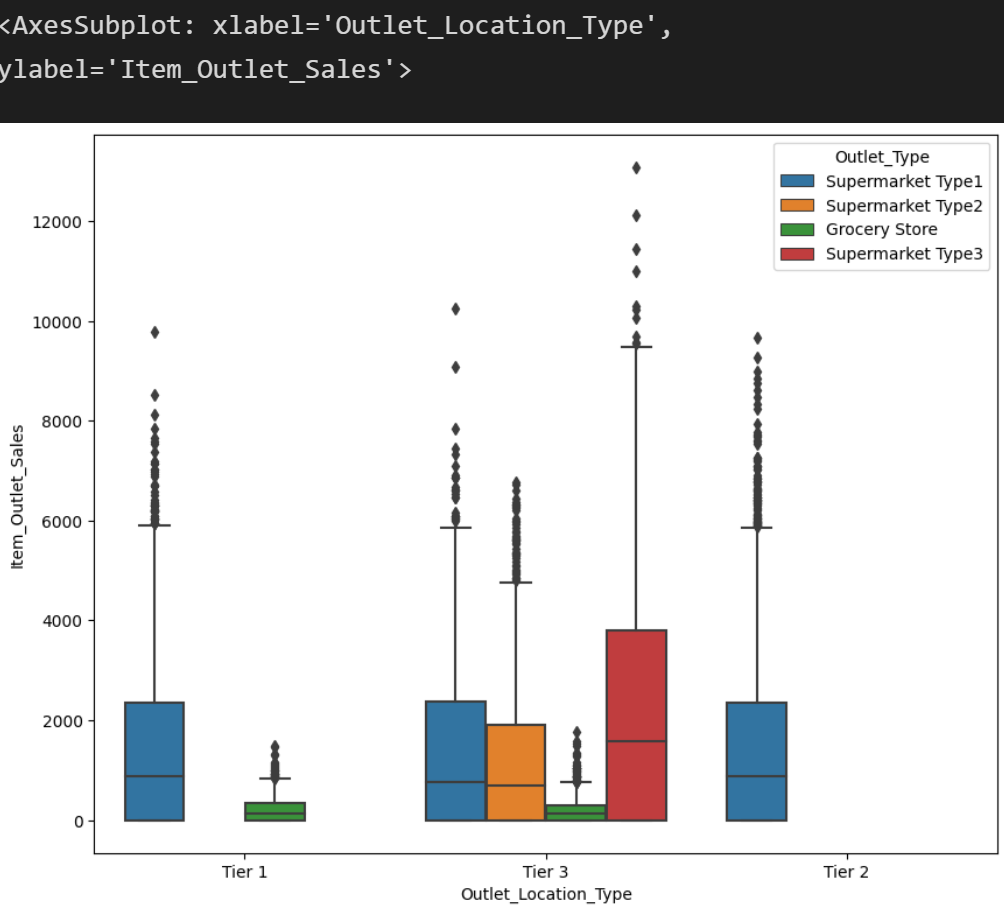
* *Item\_MRP vs Item\_Outlet\_Sales :* As the mrp of an item increases item outlet sales also increases

Chart, scatter chart

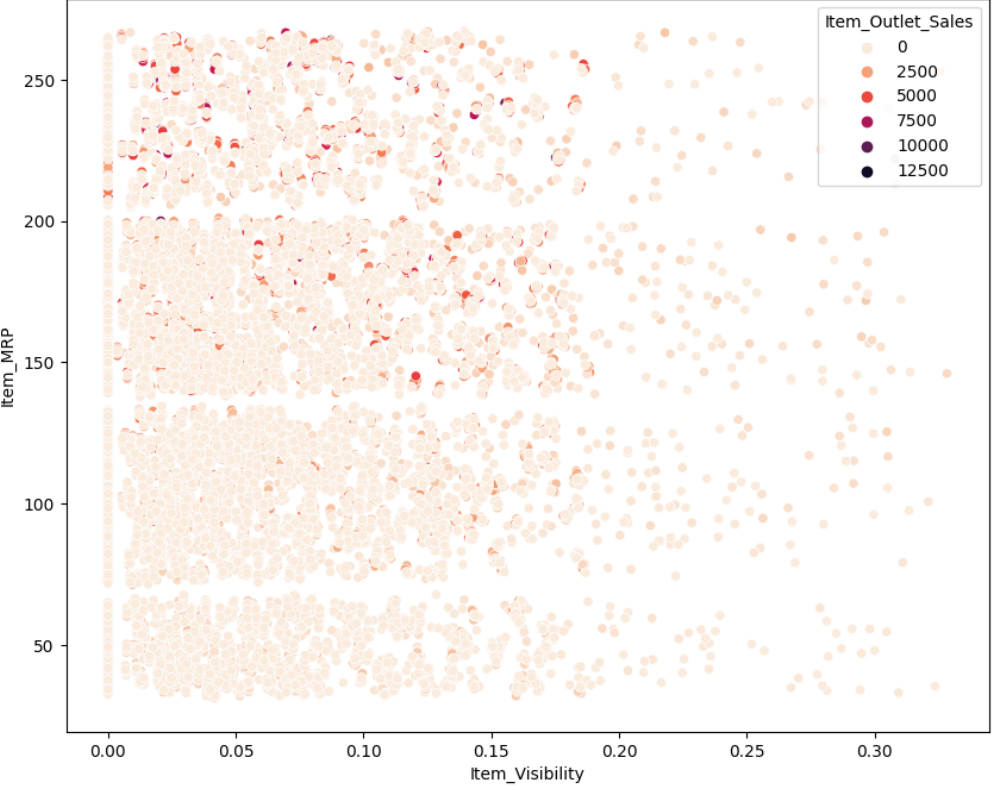
Description automatically generated

### 3. Multivariate Analysis

* *Outlet\_Size vs Item\_Outlet\_Sales vs Outlet\_Type :* It can be seen that medium size outlets sell the most and also they have the most type of outlets hence the sales increases even more



* *Item\_Visibility vs Item\_MRP:* Item\_Outlet\_Sales increases with low item visibility and high item price.

******

* *All of features of Numerical features set ([TreeMap, ClusterMap]):* MRP has the strongest positive correlation with Sales, while cisibility and establishment show a little negative correlation. There is not much correlation between the independent features.

**Chart, treemap chart

Description automatically generated**

* *Item\_Fat\_Content vs Outlet\_Establishment\_Year vs Outlet\_Type vs Item\_Fat\_Content (Boxenplot):*

***Chart, box and whisker chart

Description automatically generated***

**3.8. Compute correlations b/w features**

***\*Correlation Between Numeric with Numeric***

A picture containing table

Description automatically generated

***\*Correlation Between Categorical with Categorical***

A black screen with white text

Description automatically generated with low confidence

***\*Correlation Between Numeric with Categorical***

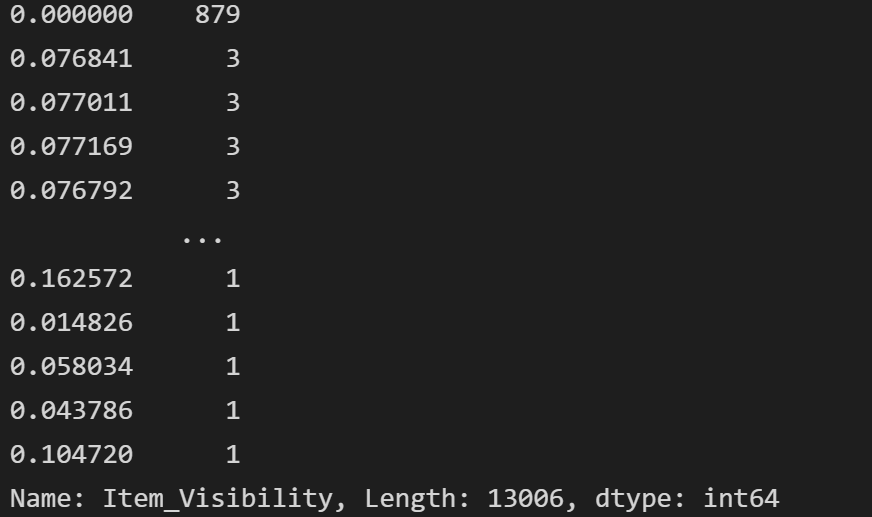
Text

Description automatically generated with low confidence

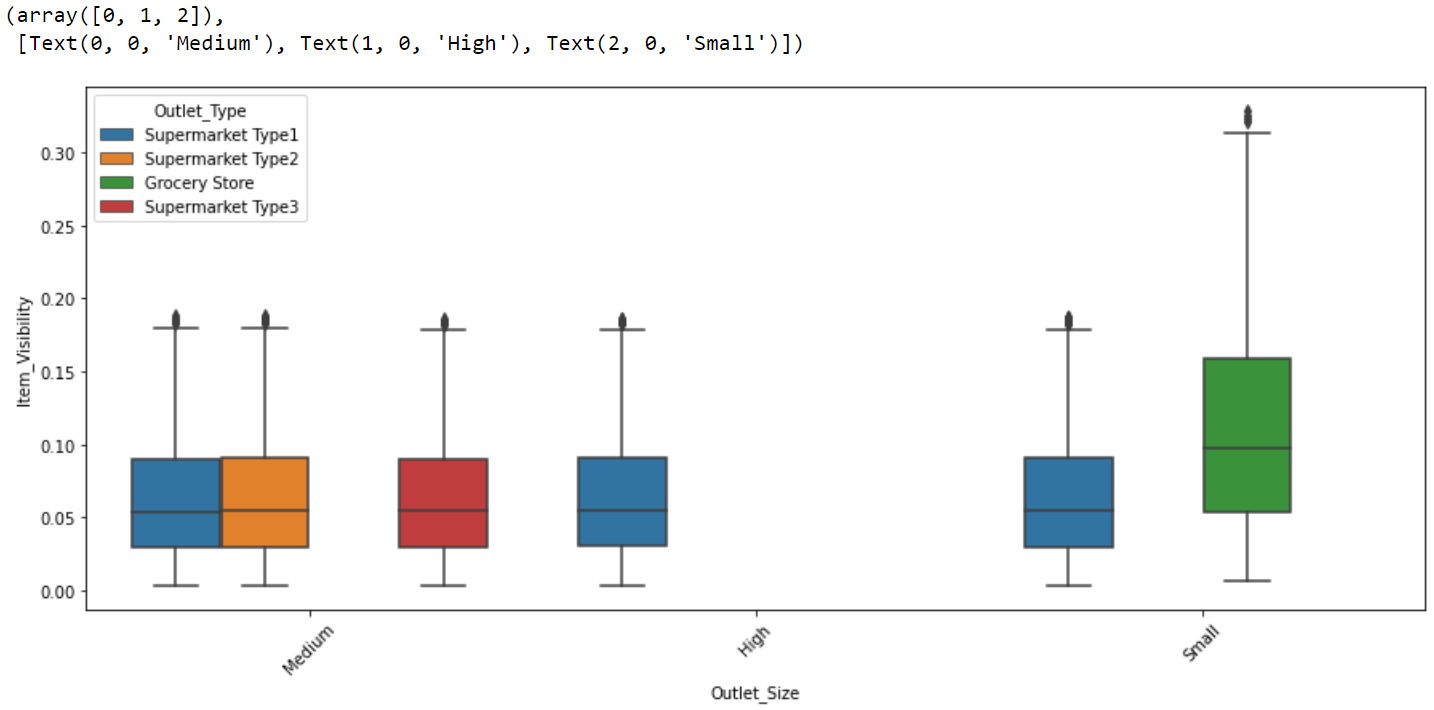
# **PART 4 : PREPARE THE DATA**

## 4.1. Skewness/Inconsistent/Missing/Outlier Handling

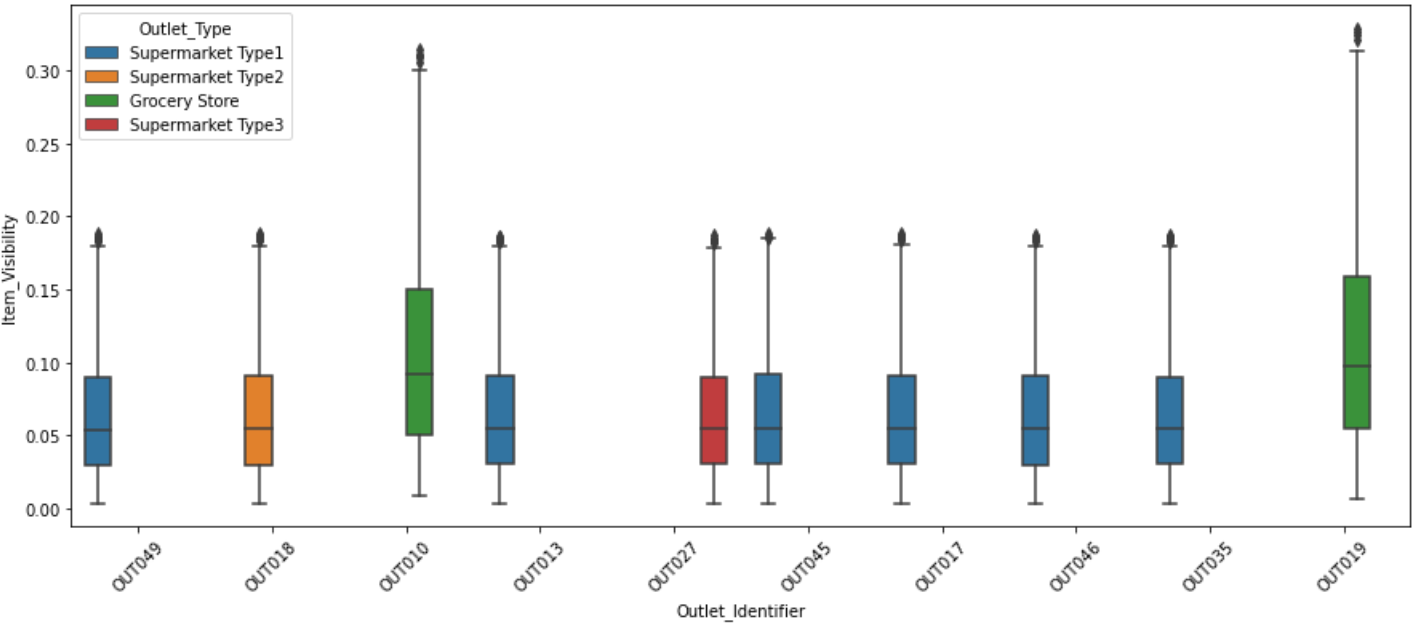
* *Item\_Visibility :* As seen earlier visibilty has 0 values which is not possible as a product will have some visibility in marts



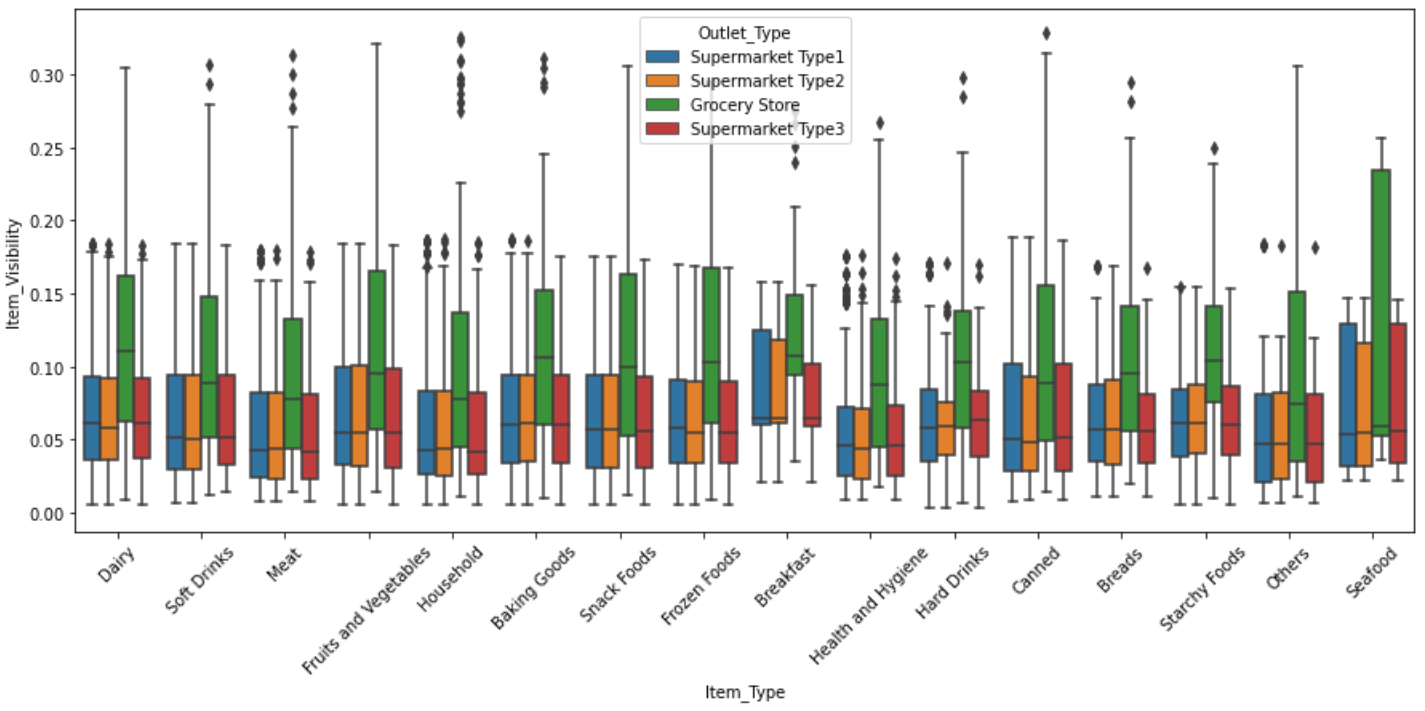
* *Checking visibility in Outlet size with Outlet type*: Here visibility contains missing values



* *Checking visibility in Outlet identifier with Outlet type:* This also contain missing values

**

* *Checking visibility in Item type with Outlet type:* Here thee are no missing values for missing visibility hence this can be used to fill nan values in visibility column

**

* *Creating pivot table to help fill nan values of visibility from here:*

### 1. Item\_Weight:

- We try to fill the nan values of weight by using values from item identifier

A screenshot of a computer

Description automatically generated with low confidence

### 2. Outlet\_Size

- Grocery stores are usually smaller than super markets so we will replace NaN values by smallA picture containing text

Description automatically generatedGraphical user interface, application

Description automatically generated

*-*  Filling nan values with mode : No null values remain

***Text

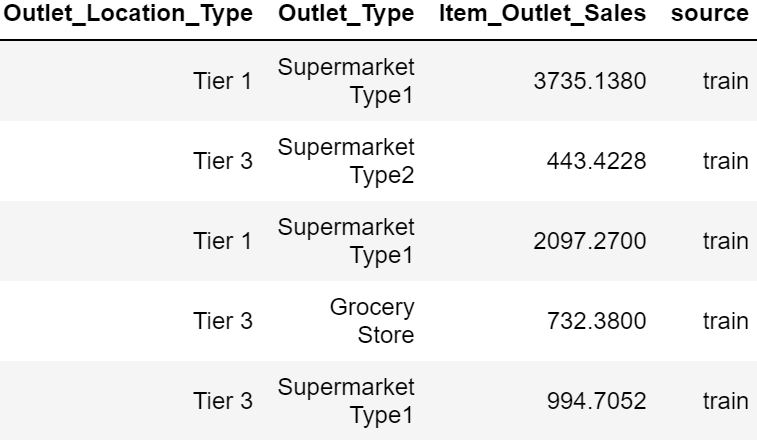
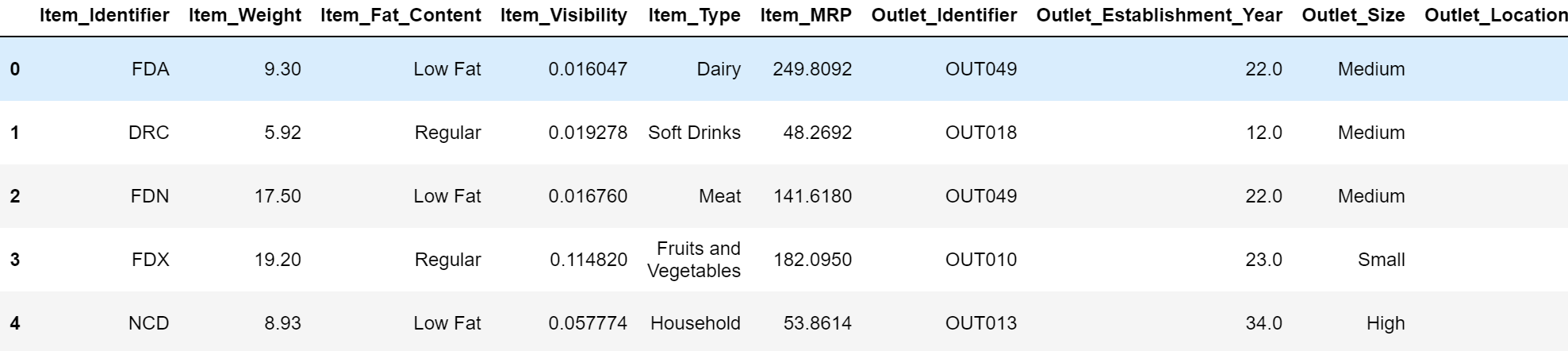
Description automatically generated***

### 3. Item\_Identifier:

- Now the item identifier column looks more meaningful so we will keep it.

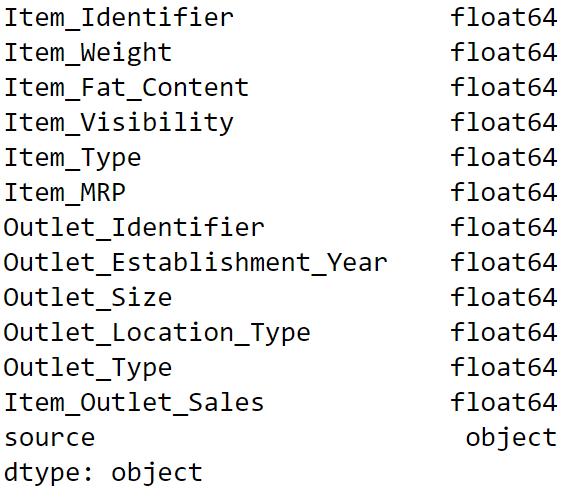


* Correcting year column by subtracting it from 2021



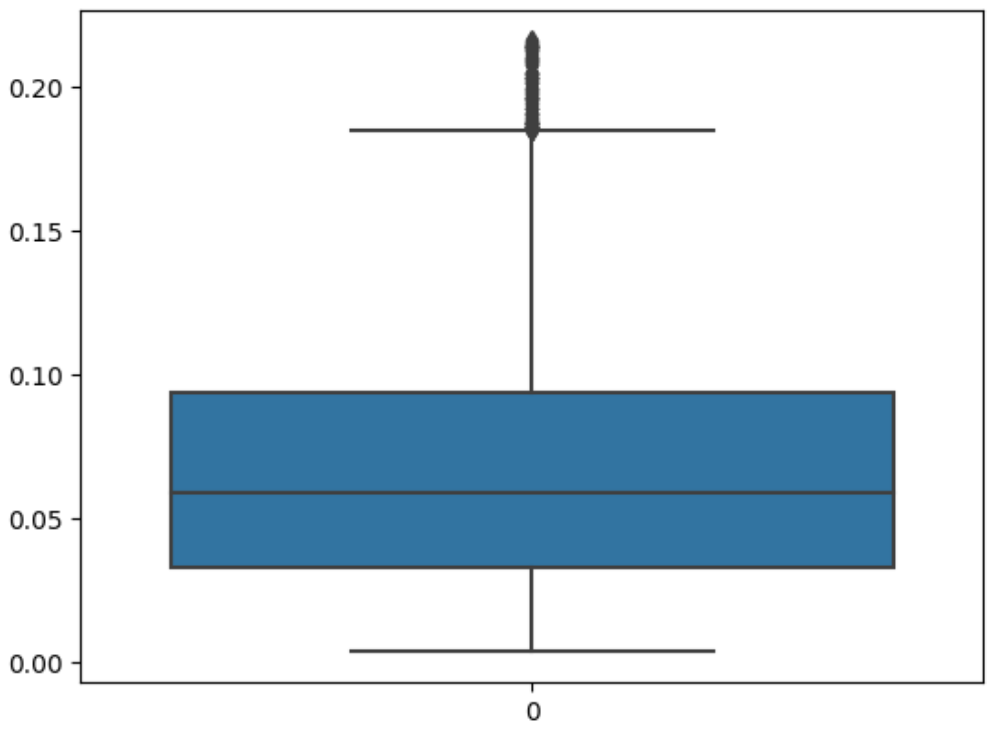
- Merging all the low fat categories to Low fat and regular categories to Regular

## 4.2. Encoding categorical features

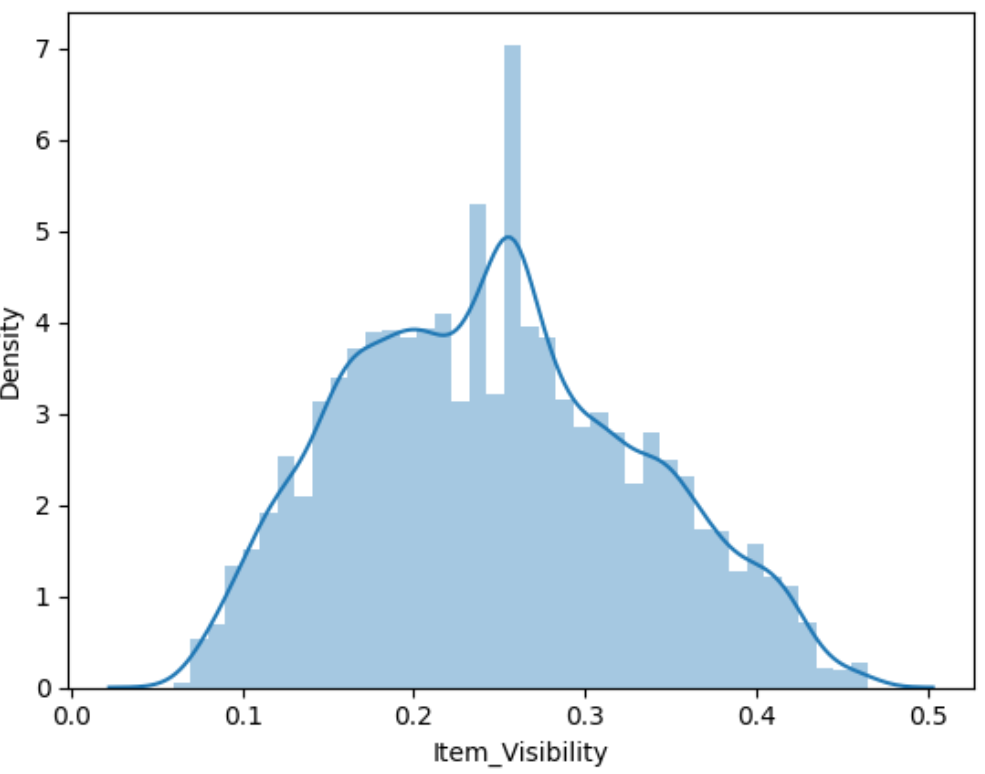


All columns are converted into float type except for source.

## 4.3. Removing outliers from Visibilitiy column

******

## 4.4. Removing skewness from visibility column

****

## 4.5. Feature scaling (Normalization & Standardization)

## 4.6. Separating the data into train and test

A screen shot of a computer

Description automatically generated with low confidence

Graphical user interface, text

Description automatically generated

# **PART 5. TRAIN AND EVALUATE MODELS**

## 5.1. Modelling Phase

1. Importing necessary modules
2. Importing models
3. Choosing the best random state using Logistic regression
4. Creating list of models and another list mapped to their names
5. Creating models
6. Calculating scores of the model and appending them to a list
7. Creating Dataframe

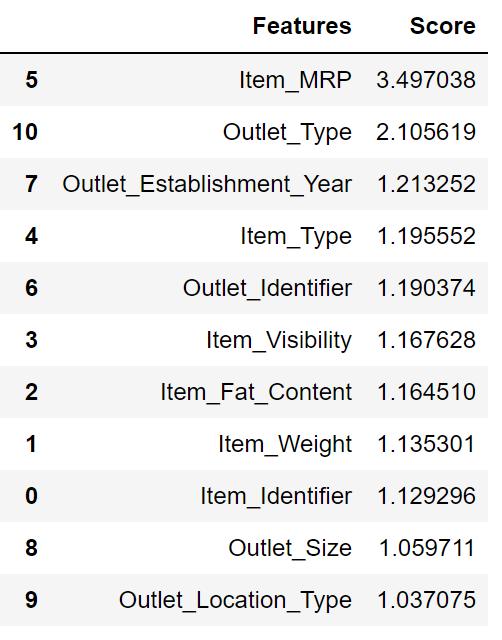
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Mean Absolute Error** | **Mean Squared Error** | **Root Mean Squared Error** | **R2 Score** | **Mean of Cross validaton Score** |
| **Kneighbors Regressor** | 809.6622 | 1.301844e+06 | 1140.9836 | 26.14 | 50.28 |
| **Linear Regression** | 886.4671 | 1.370332e+06 | 1170.6120 | 13.49 | 49.79 |
| **Lasso** | 885.4656 | 1.368904e+06 | 1170.0017 | 12.70 | 49.82 |
| **Ridge** | 886.2023 | 1.369934e+06 | 1170.4417 | 13.16 | 49.79 |
| **Elastic Net** | 1212.1723 | 2.344823e+06 | 1531.2815 | -7167.95 | 12.60 |
| **Decision Tree Regressor** | 1099.7594 | 2.518209e+06 | 1586.8865 | 20.70 | 16.26 |
| **Random Forest Regressor** | 779.8128 | 1.238802e+06 | 1113.0150 | 35.70 | 54.70 |
| **AdaBoost Regressor** | 1054.1472 | 1.693790e+06 | 1301.4571 | -27.04 | 45.14 |
| **Gradient Boosting Regressor** | 742.6823 | 1.119748e+06 | 1058.1814 | 37.77 | 58.73 |
| **XGB Regressor** | 808.1595 | 1.307496e+06 | 1143.4578 | 33.33 | 51.57 |

From above analysis only Random Forest, Gradient Boost and Xgboost perform well with r2 score more than 51 and mean absolute error less than 810. Though the results are not as good therefore we further try to increase the scores by Feature Selection.

## 5.2. Feature Selection

### **1. Using ANOVA test**

- Below are the best features with there score in decreasing order after going through anova test



### 2. Using feature importance of Extra trees regressor

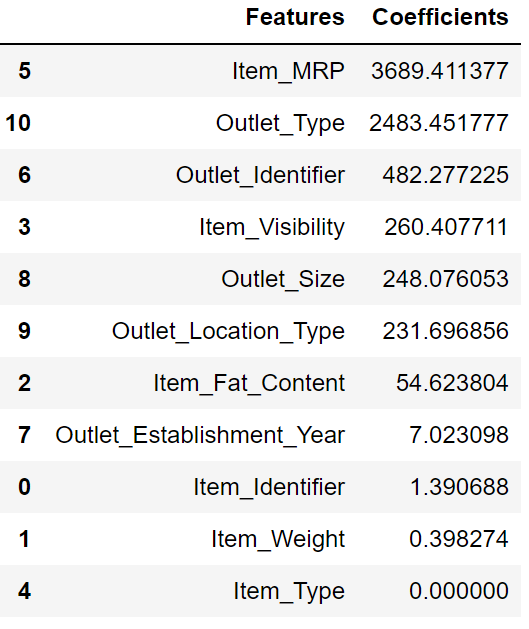
- Plot graph of feature importance for better visualization

Chart, bar chart

Description automatically generated

* MRP, Outlet Type are the most important features and the list follows. Above list contains features in order with most important feature on the top and least important feature below.

### 3. Using Lasso Coefficient



* Above dataframe shows features with their coefficients values. Item Type has coefficient 0 signifying that it is least important feature according to Lasso

## 5.3. Conclusion

- Anova test and feature importance tell us that Outlet\_Location\_Type is the least significant feature. Feature selection tells us that Item\_Type is the least significant feature. So we create 3 set of data set:

* x1 - After removing Outlet\_Location\_Type
* x2 - After removing Item\_Type
* x3 - After removing both Outlet\_Location\_Type and Item\_Type.

From above analysis we see that previous models which were performance well are the one performing well even this time and the dataset providing least Root errors and highest mean cross validation score is Dataset x2 which we get after removing Item\_Type which we got using feature selection of Lasso, even though results do not vary much even after doing feature selection so we will keep the original dataset for training purposes without loosing any data.

Table

Description automatically generated

Table

Description automatically generated

A picture containing text

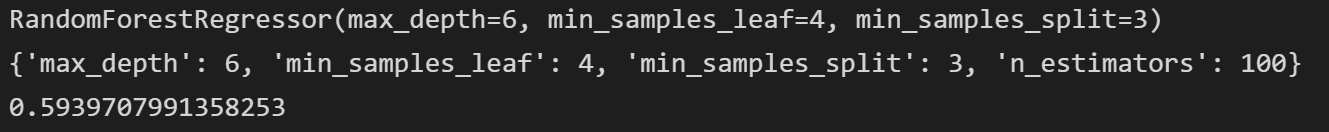
Description automatically generated

# **PART 6. FINE-TUNE MODELS**

## Hyperparameter Tuning

* After Hyperparameter tunning the best model with least error and highest r2 score and cross validation score is Gradient Boost

## 6.1. Random Forest



Text

Description automatically generated

## 6.2. Gradient Boost

Text

Description automatically generated

Text

Description automatically generated

## 6.3. Xtreme Gradient Boost

Text

Description automatically generated

Text

Description automatically generated

* Conclusion After Hyperparameter tunning the best model with least error and highest r2 score and cross validation score is Gradient Boost

# **PART 7. ANALYZE AND TEST YOUR SOLUTION**

## 7.1. Finalizing the best model

## ***7.2. Evaluation Metrics***

Text

Description automatically generated

- Plot image

Chart, scatter chart

Description automatically generated

* There are still a lot of outliers in our output
* Saving Model

Graphical user interface

Description automatically generated

* Predicting model

Text

Description automatically generated