Big Mart Sales Analysis

Project Report

A guide by Husanbano Shamlik

Summary

Industry: Retail Industry

Stores: Super Market 1,2,3,Grocery Store

City:Tier 1(Metropolitan City),Tier 2(Mediun-sized urban center),Tier 3 (smaller towns and urban area)

Items: Dairy', 'Soft Drinks', 'Meat', 'Fruits and Vegetables', 'Household', 'Baking Goods', 'Snack Foods', 'Frozen Foods','Breakfast', 'Health and Hygiene', 'Hard Drinks', 'Canned','Breads', 'Starchy Foods', 'Others', 'Seafood'

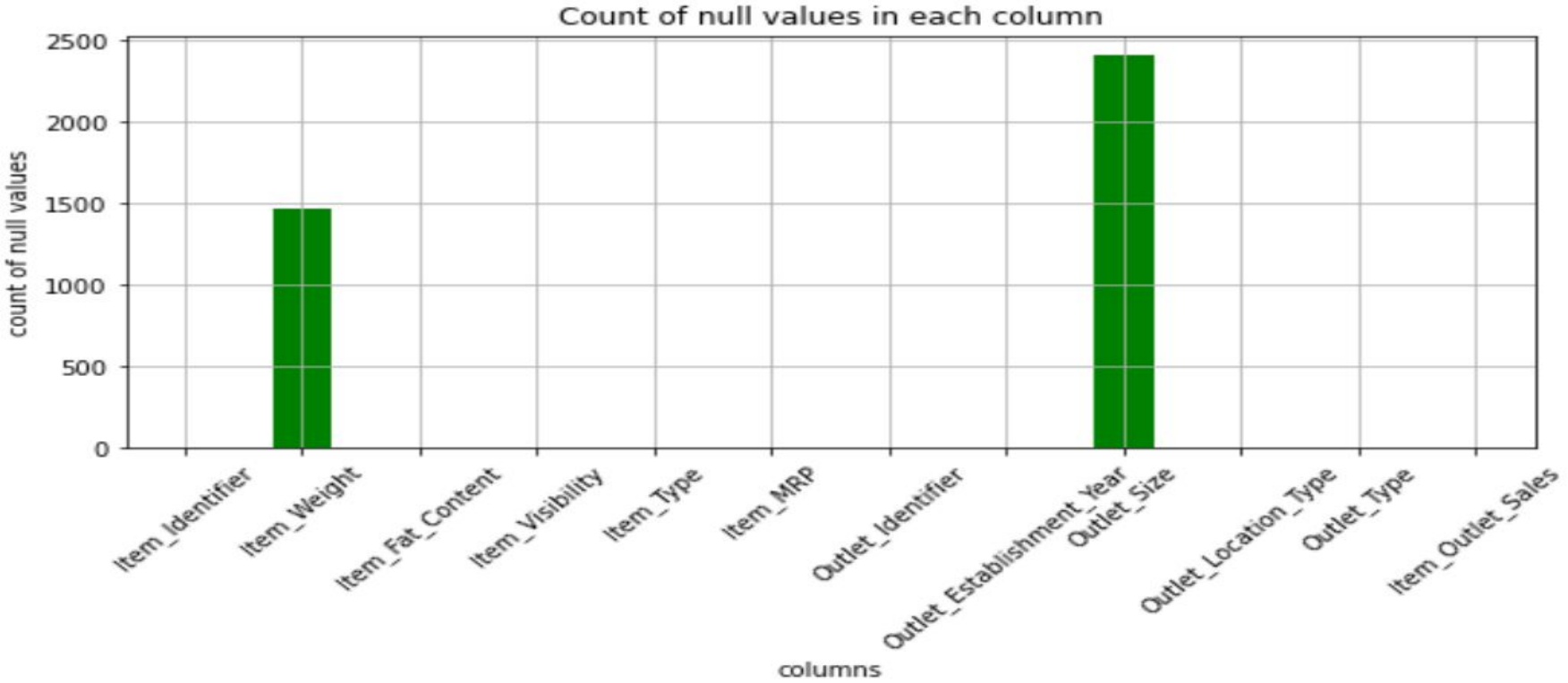
Data Description

- BigMart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities. Where the dataset consists of 12 attributes like Item Fat, Item Type, Item MRP, Outlet Type, Item Visibility, Item Weight, Outlet Identifier, Outlet Size, Outlet Establishment Year, Outlet Location Type, Item Identifier and Item Outlet Sales. Out of these attributes response variable is the Item Outlet Sales attribute and remaining attributes are used as the predictor variables.

- Item_Identifier -> Unique product ID
- tem_Weight -> Weight of product
- Item_Fat_Content -> Whether the product is low fat or not
- tem_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.

DATA CLEANING:

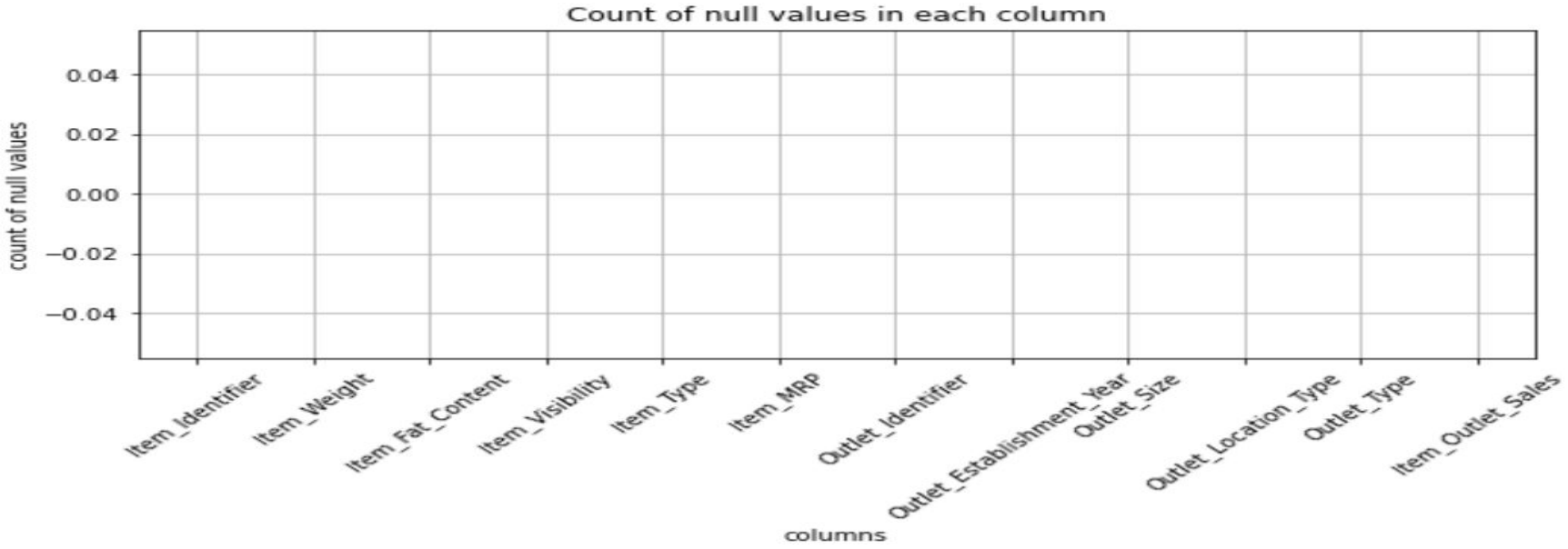
Before handling missing values using mode and statistical methods



Treatment of Null Values in Dataset through Mapping and Statistical Imputation.

There are null values present in item weight as count less than total rows, We have seen above that Item_Weight is a Numerical Feature So we can substitute by its mode value to fill the missing values. * Outlet_Size is a Categorical Feature so will use mode to impute the missing values in this column.

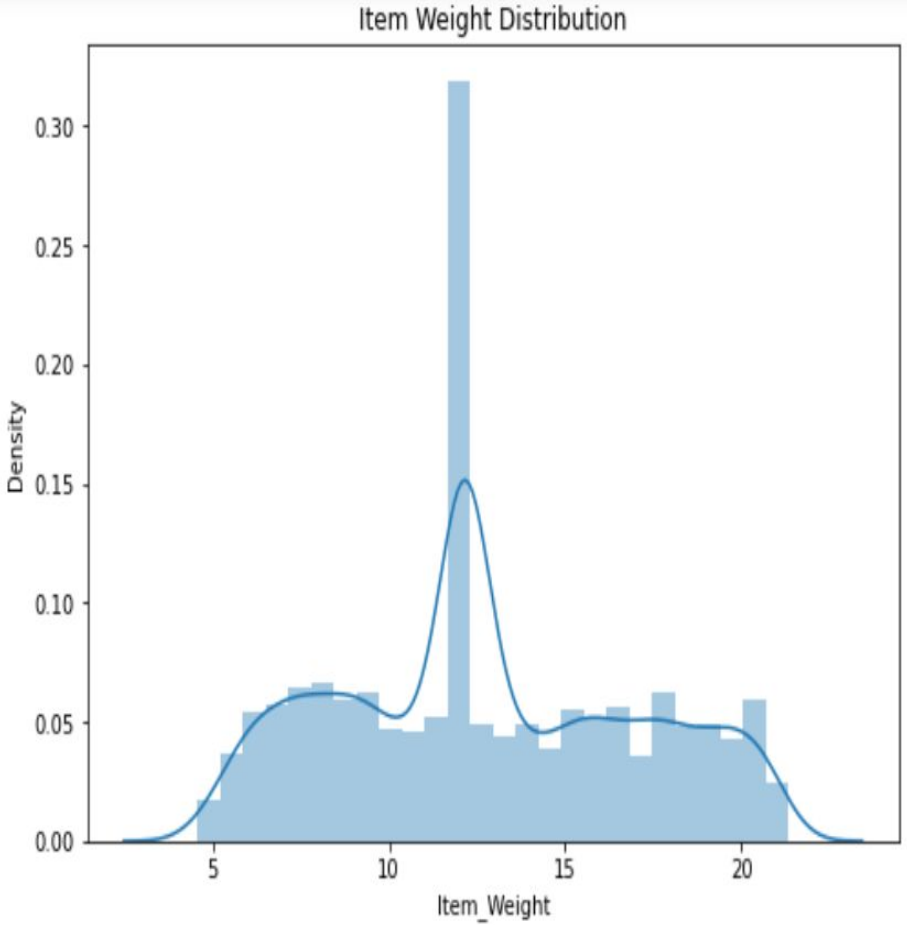
We know that, the Outlet_size and Outlet_Type are related to each other. So we are filling the missing values of the Outlet_Size by using the values in Outlet_Type.



Exploratory Data Analysis(EDA)

Analysis of Item_Weight

From the plot ,We can see that Item_Weight with range 13 is having the highest distribution.



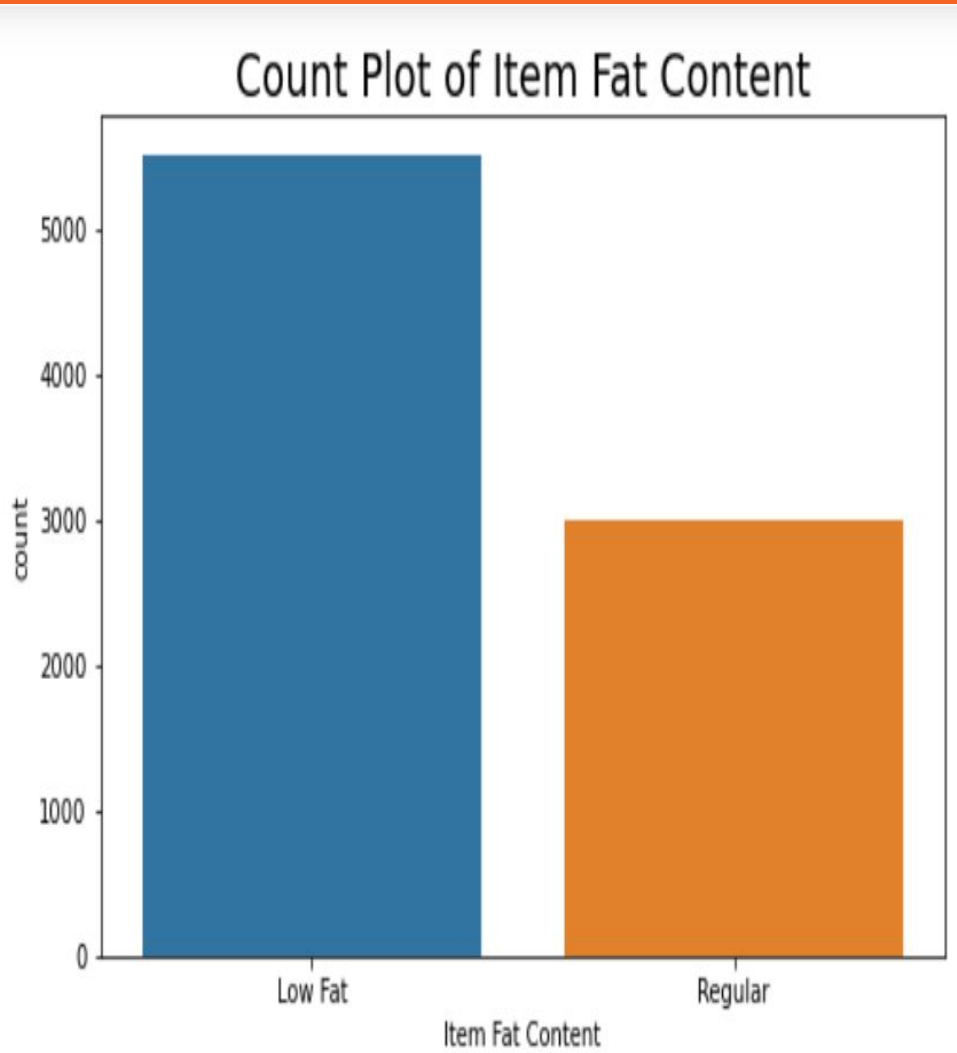
Story for illustration purposes only

Analysis of Item Fat

As you can see the low fat and regular in item_fat_content is written differently so first we need to correct it for further analysis because it represents the same thing.

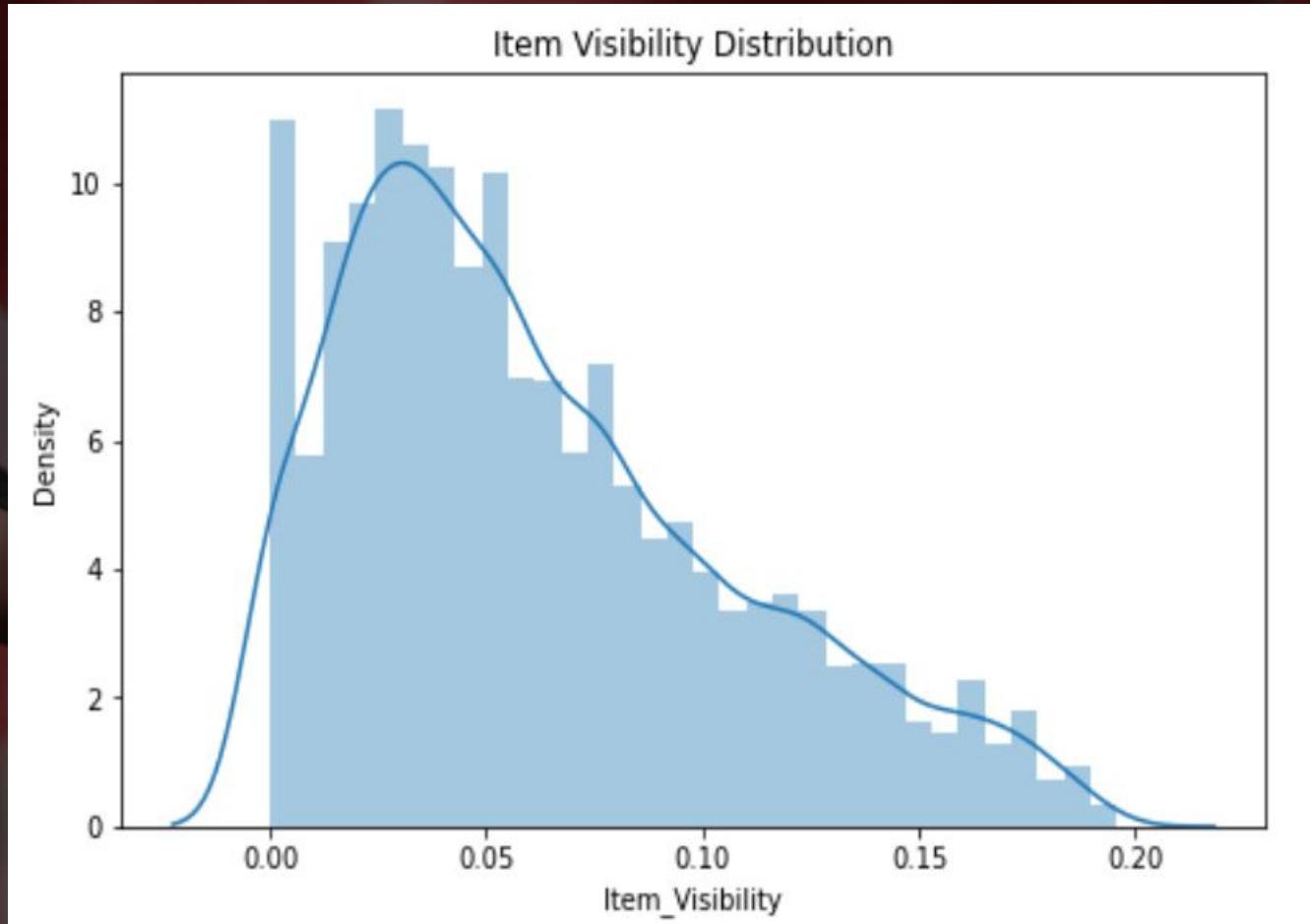
```
1 df['Item_Fat_Content'].value_counts()
```

```
Low Fat    5517
Regular    3006
Name: Item_Fat_Content, dtype: int64
```



Analysis of Item Visibility

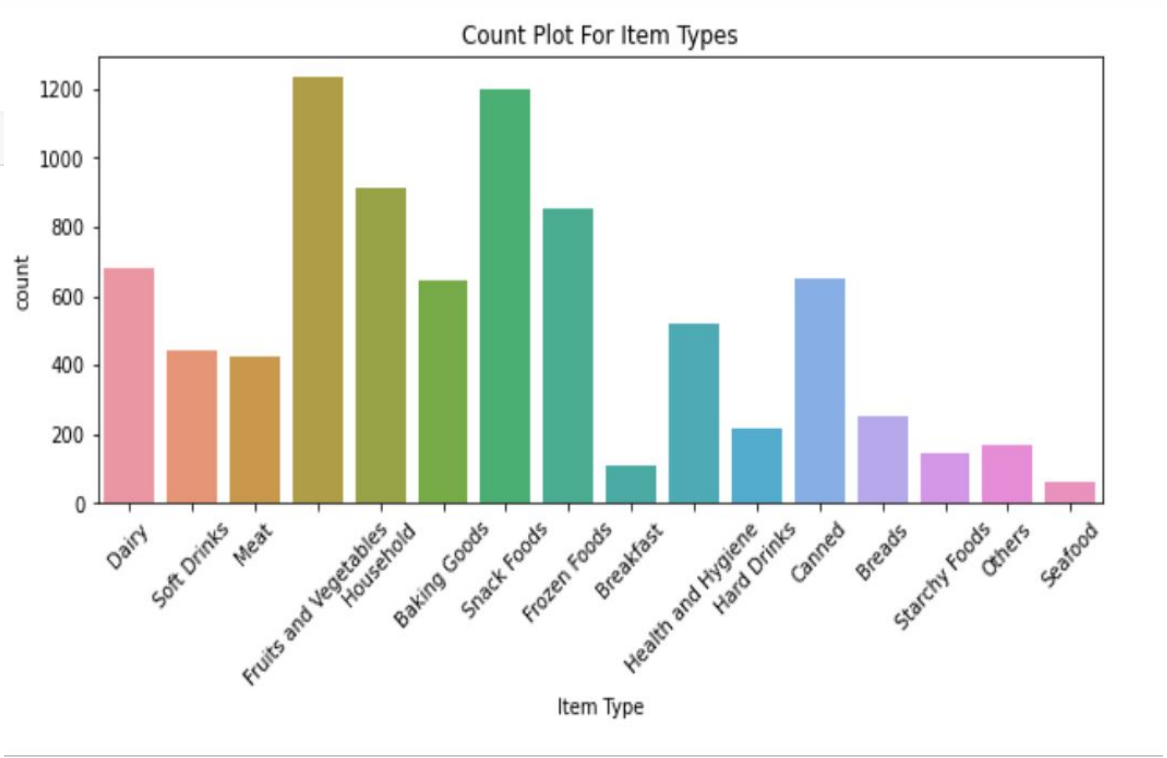
The plot suggest that its not a normal distribution or we can say that the distribution is kind of right skewed or positive skewed



Analysis of Item Type

```
1 df['Item_Type'].value_counts()
```

```
Fruits and Vegetables     1232
Snack Foods               1200
Household                  910
Frozen Foods               856
Dairy                      682
Canned                     649
Baking Goods               648
Health and Hygiene         520
Soft Drinks                445
Meat                       425
Breads                     251
Hard Drinks                214
Others                     169
Starchy Foods              148
Breakfast                  110
Seafood                     64
Name: Item_Type, dtype: int64
```

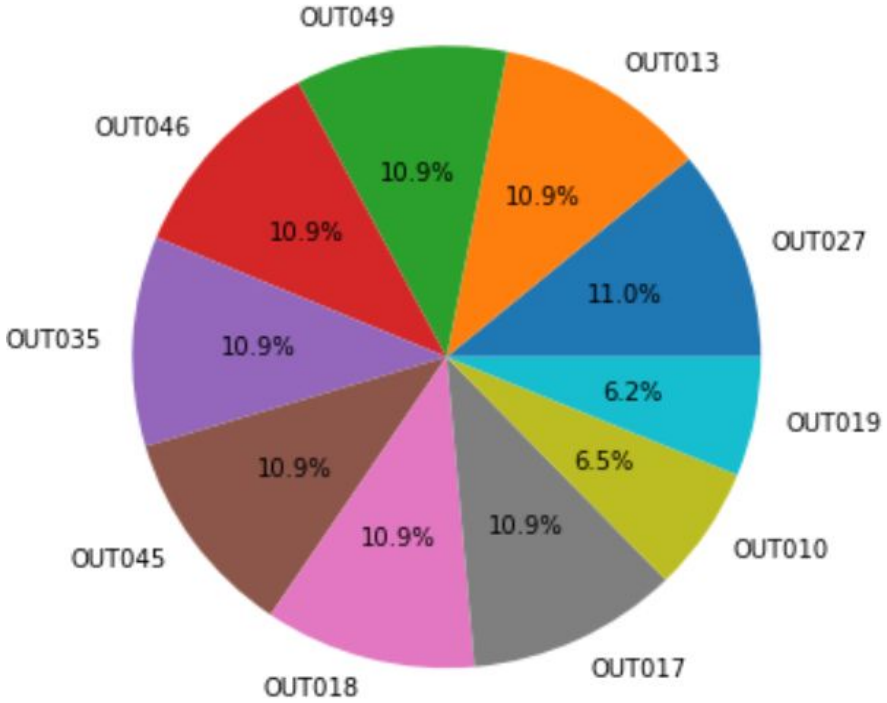


Fruits and Vegetables and Snack Foods are the two categories in which most of the items fall whereas Seafood is the least

Analysis of Outlet Identifier

```
In [68]:  1 df['Outlet_Identifier'].value_counts()
```

```
Out[68]: OUT027    935
         OUT013    932
         OUT049    930
         OUT046    930
         OUT035    930
         OUT045    929
         OUT018    928
         OUT017    926
         OUT010    555
         OUT019    528
         Name: Outlet_Identifier, dtype: int64
```

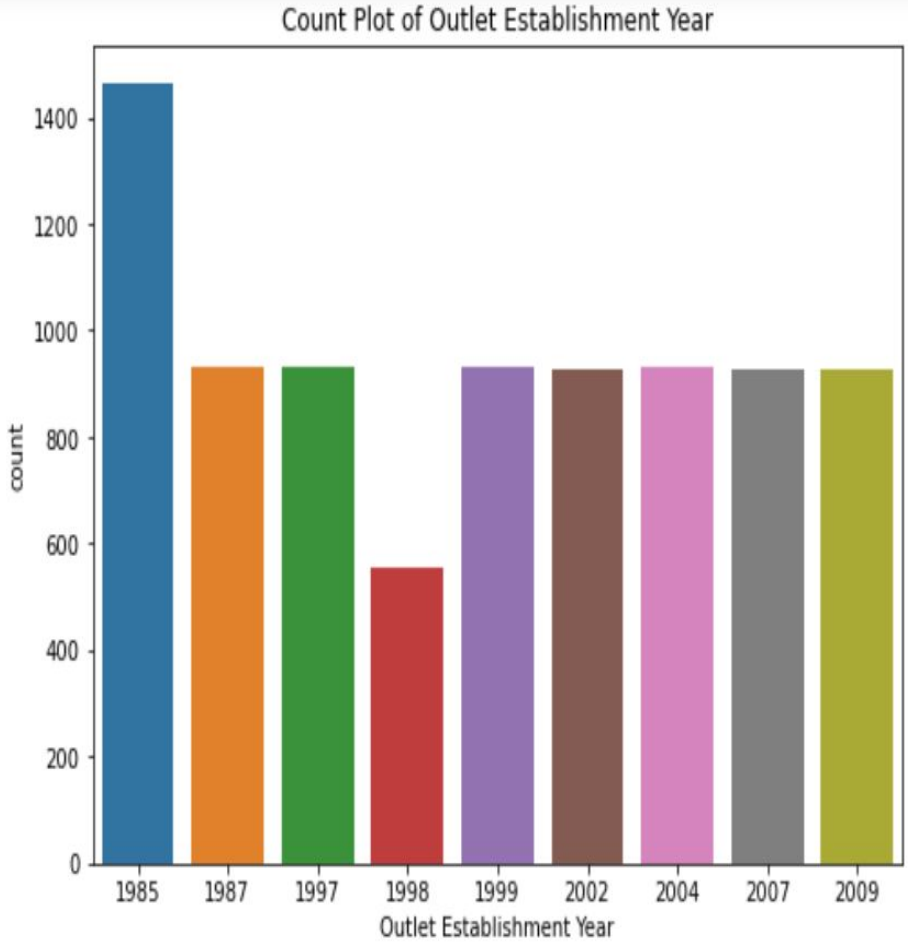


There are 10 outlets which are almost balanced except for two outlets.

Analysis of EstablishmentYear

```
1985    1463
1987     932
1999     930
1997     930
2004     930
2002     929
2009     928
2007     926
1998     555
Name: Outlet_Establishment_Year,
```

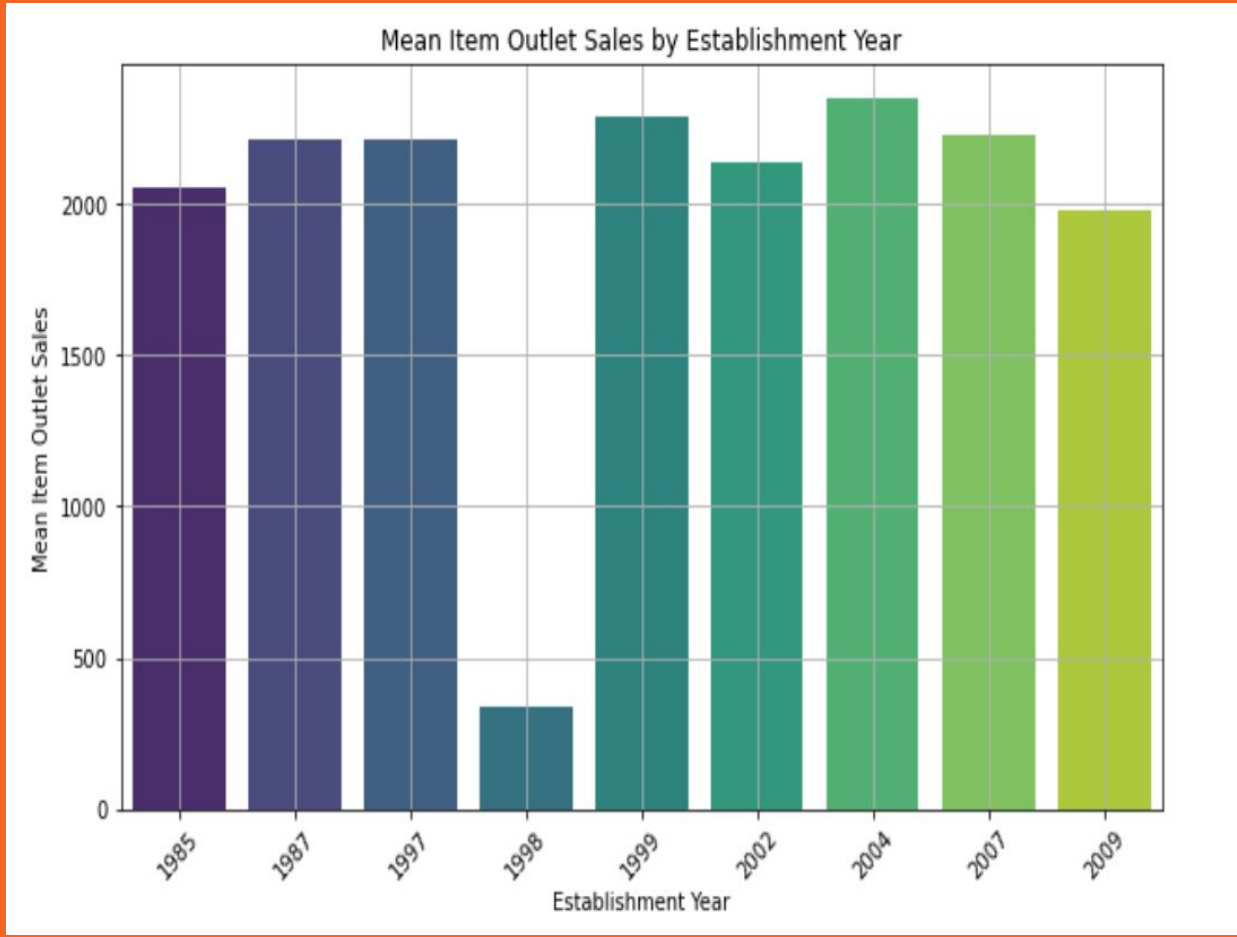
Most of the outlets were established in the year of 1985 and least in 1998.



Analysis of Establishment Year with item Sales

	Outlet_Establishment_Year	Item_Outlet_Sales
0	1985	2054.684740
1	1987	2210.295979
2	1997	2215.219692
3	1998	339.351662
4	1999	2286.007118
5	2002	2134.445147
6	2004	2346.946432
7	2007	2224.100586
8	2009	1979.629310

Avg Outlet Sales in the 2004 year are more and least sales in 1998

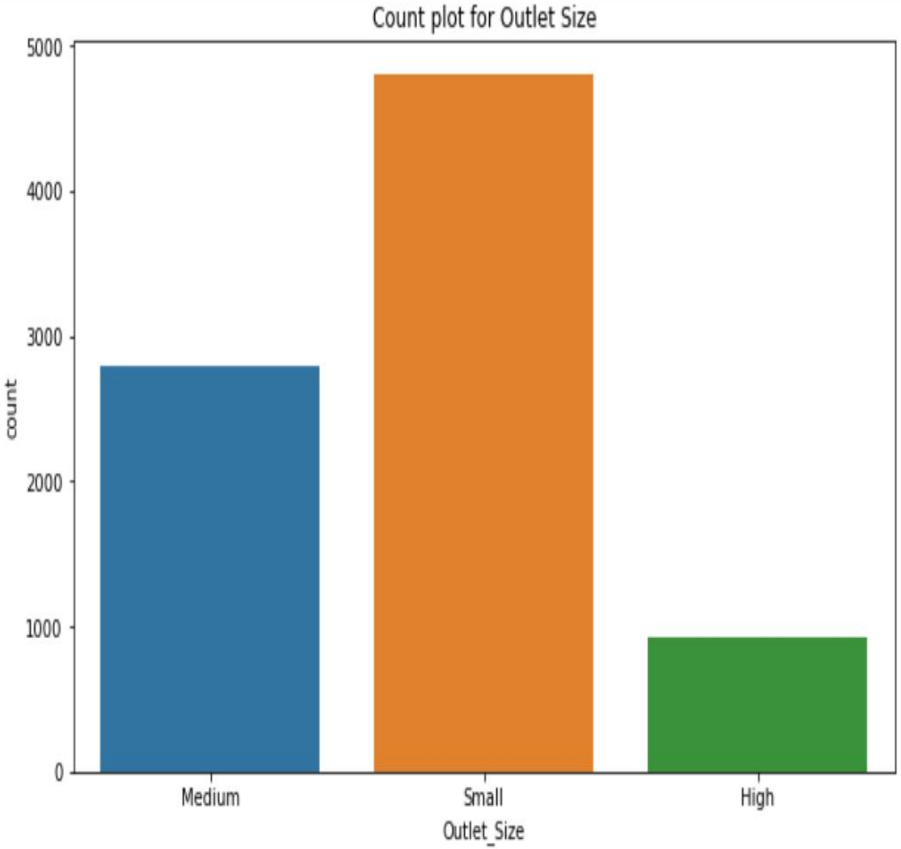


Analysis of Outlet Size

```
Small     4798
Medium    2793
High       932
Name: Outlet_Size, dtype: int64
```

There are few outlets with high size. Most of the outlets are of small size.

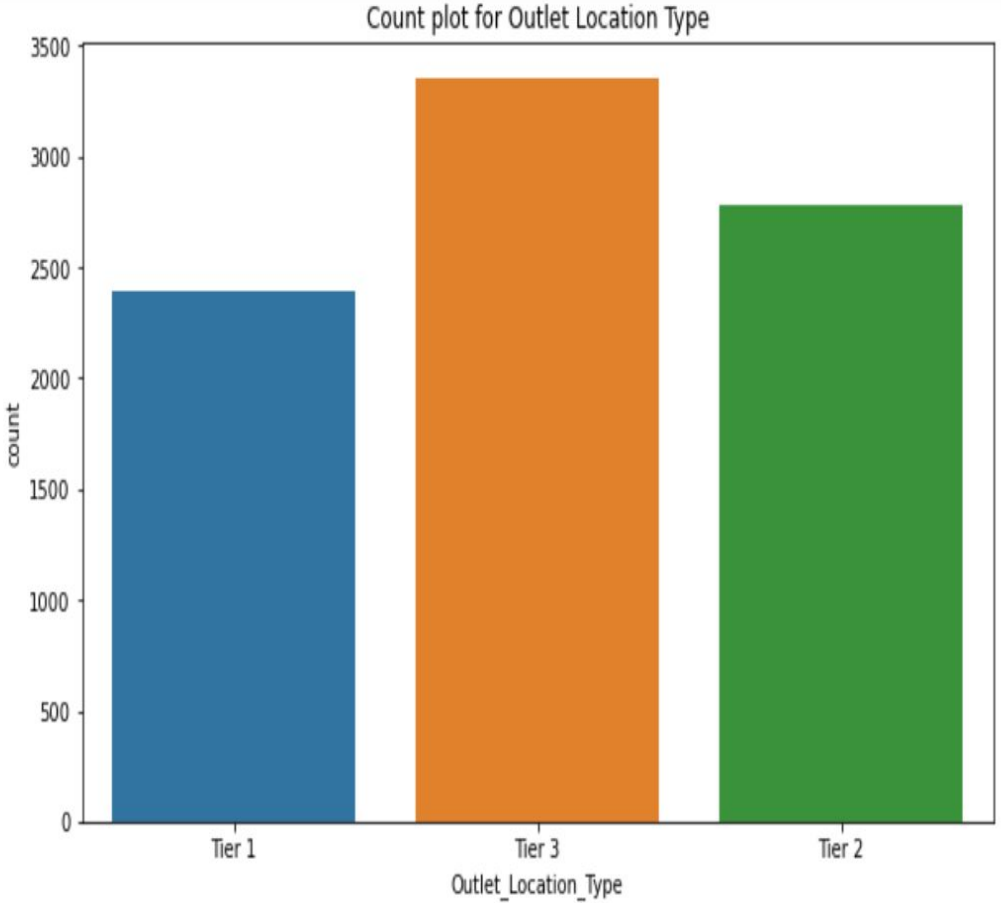
Source: theguardian.com



Analysis of Outlet Location Type

Tier 3	3350
Tier 2	2785
Tier 1	2388

Most of the stores are located in Tier 3 cities.

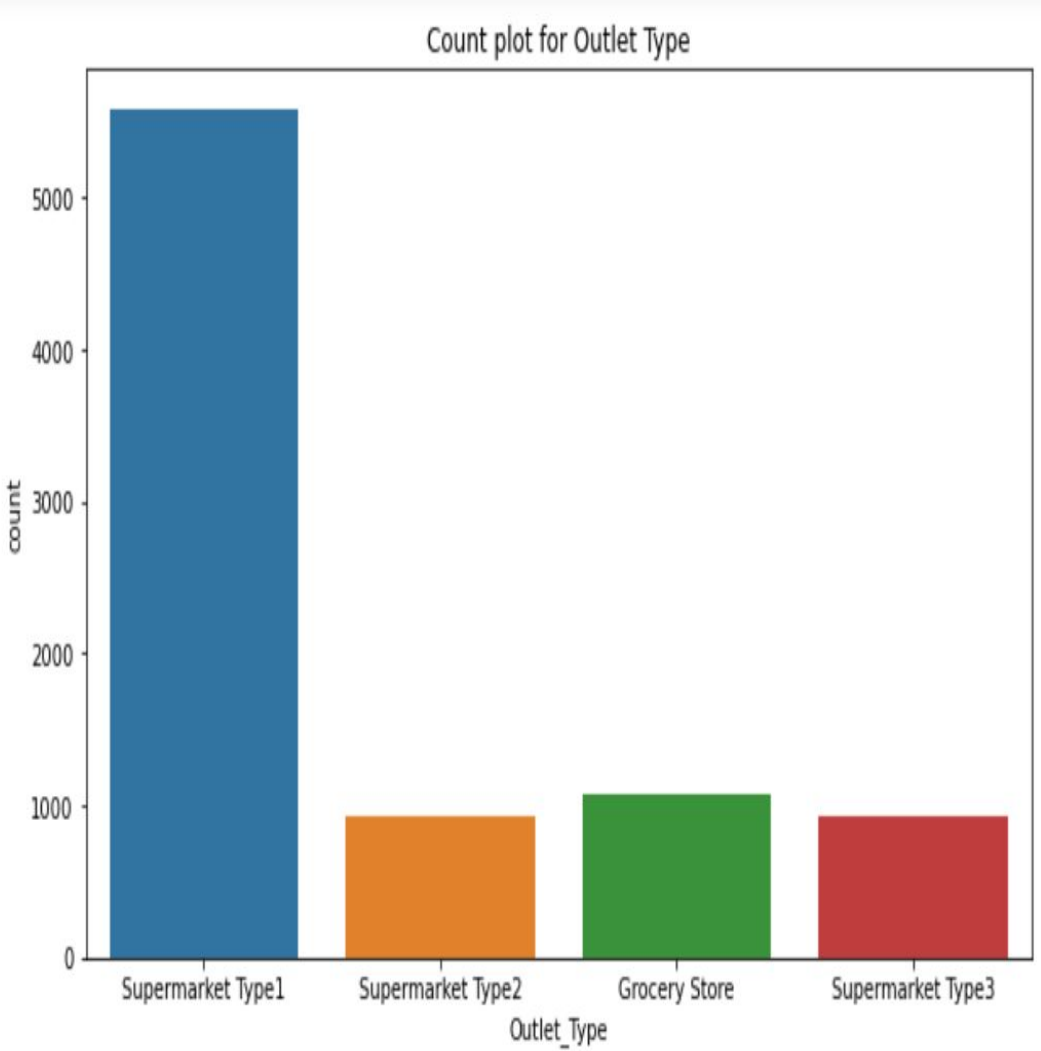


Source: travel.trade.gov

Analysis of Outlet Type

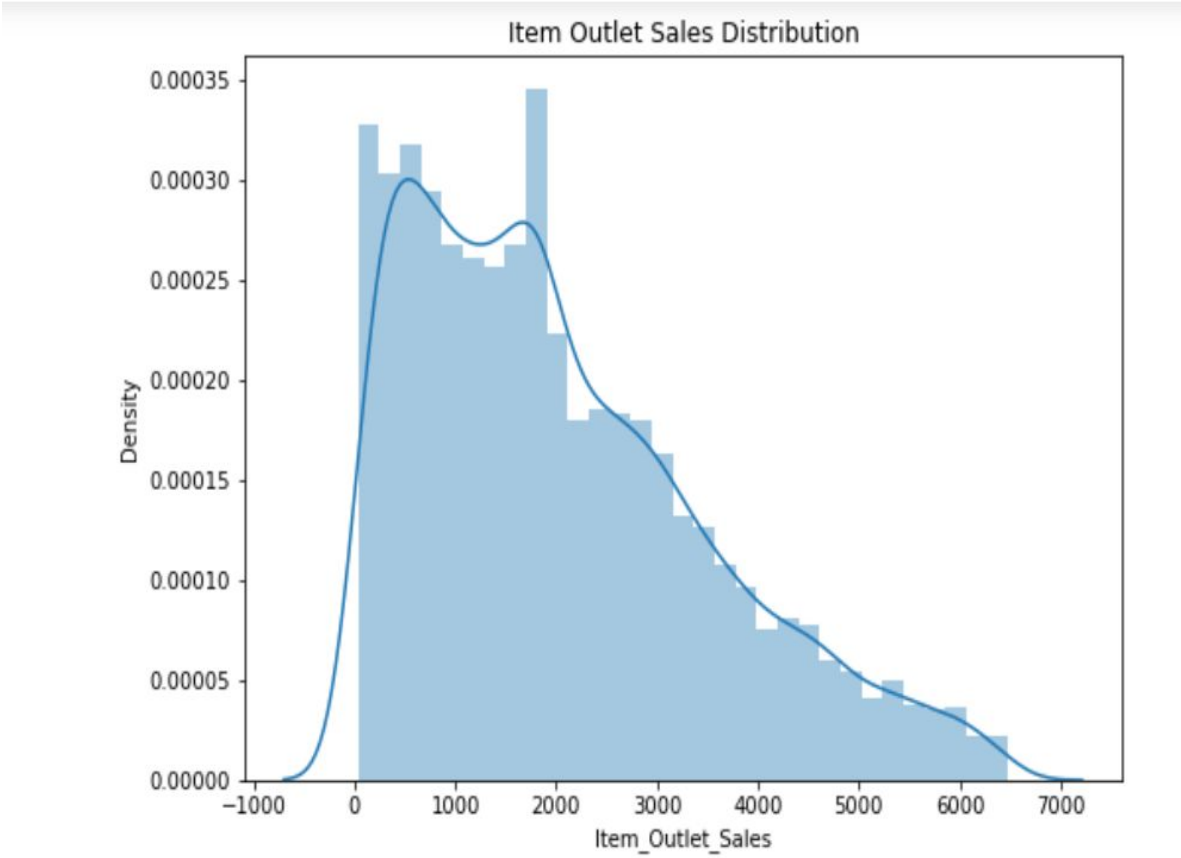
```
Supermarket Type1    5577
Grocery Store        1083
Supermarket Type3     935
Supermarket Type2     928
Name: Outlet_Type, dtype:
```

supermarket Type 1 has more no of outlet type category



Analysis of Item outlet sales

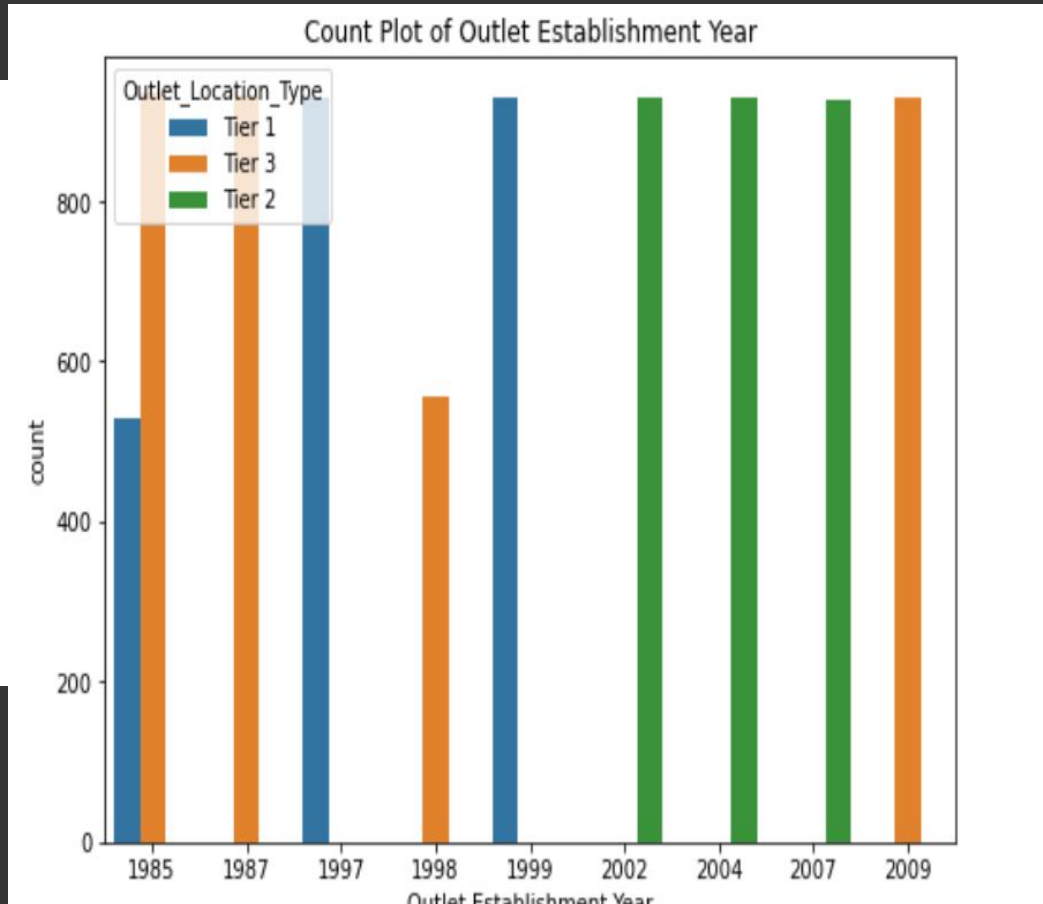
Most Outlet sales in range of 1000 to 2000 and its a right skewed distribution.



Analysis of Outlet Establishment Year and Outlet Location Type

Outlet_Location_Type	Outlet_Establishment_Year	
Tier 1	1997	930
	1999	930
	1985	528
Tier 2	2004	930
	2002	929
	2007	926
Tier 3	1985	935
	1987	932
	2009	928
	1998	555

Name: Outlet_Establishment_Year, dtype: int64

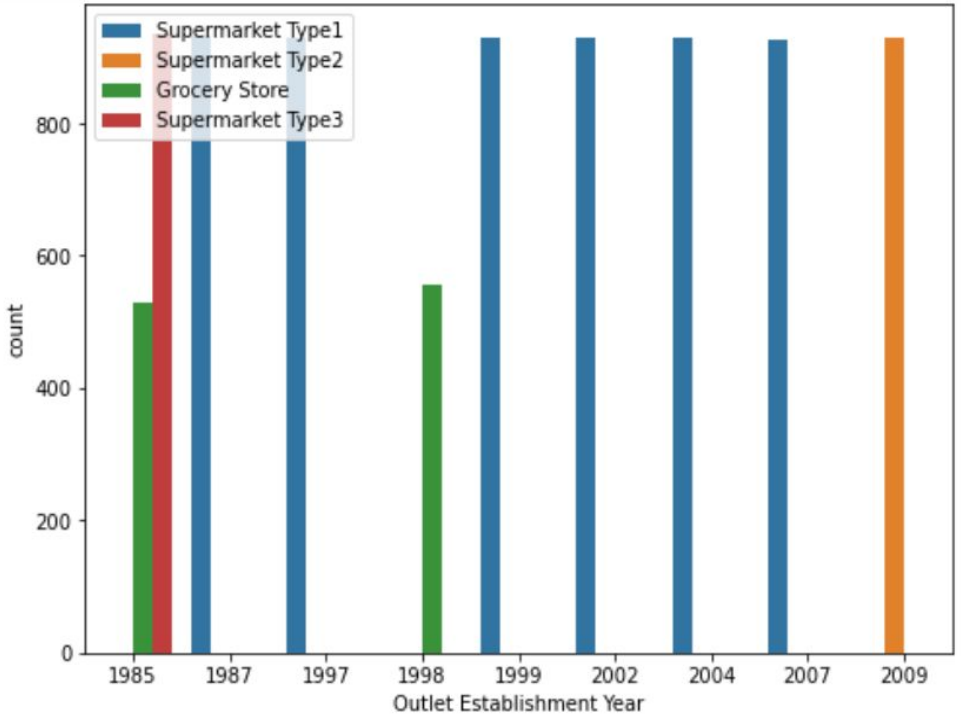


In Tier1 and Tier3 cities outlets were established in 1985 whereas tier2 got outlets after 2000

Analysis of Outlet Establishment year and Outlet Type

Outlet_Establishment_Year	Outlet_Type	
1985	Supermarket Type3	935
	Grocery Store	528
1987	Supermarket Type1	932
1997	Supermarket Type1	930
1998	Grocery Store	555
1999	Supermarket Type1	930
2002	Supermarket Type1	929
2004	Supermarket Type1	930
2007	Supermarket Type1	926
2009	Supermarket Type2	928

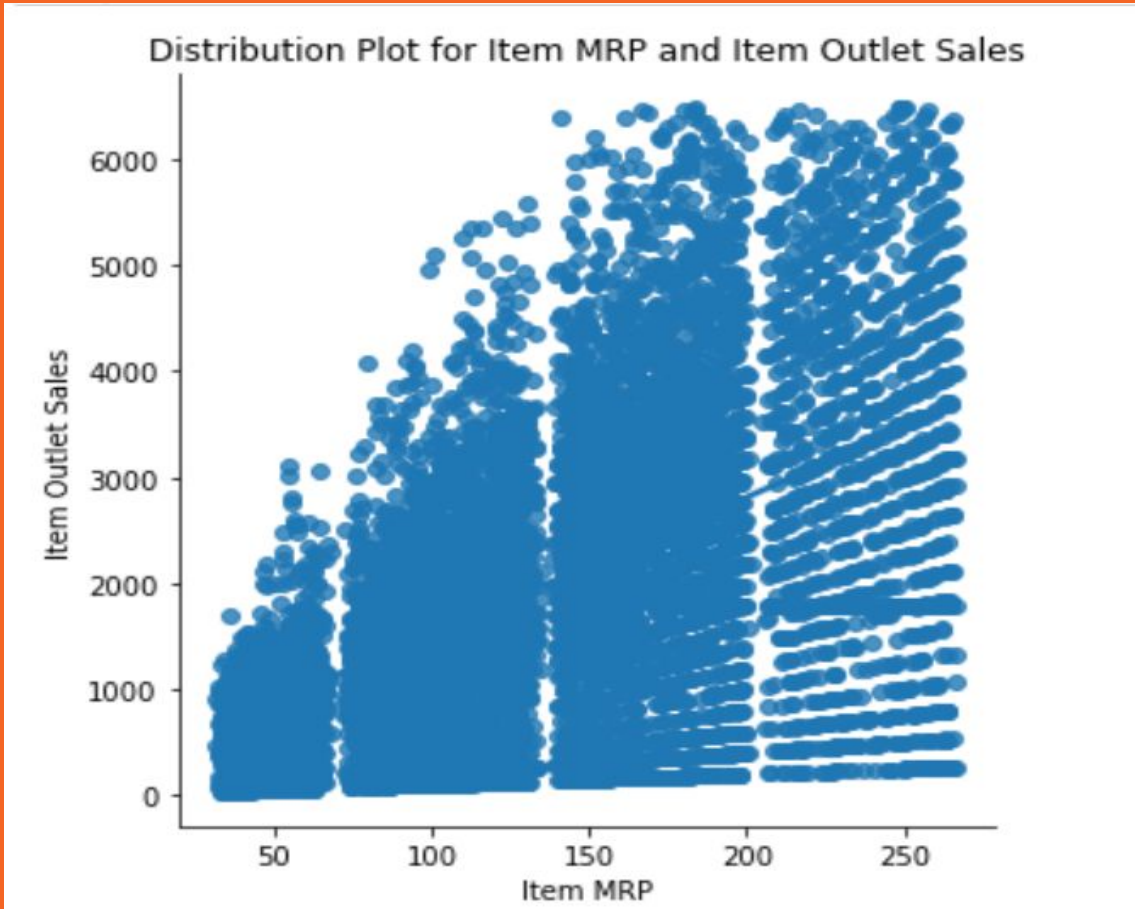
Name: Outlet_Type, dtype: int64



Supermarket type 2 was build much later while grocery stores and supermarket Type1 and Type3 are the oldest outlet type.

Analysis of Outlet Sales and Item MRP

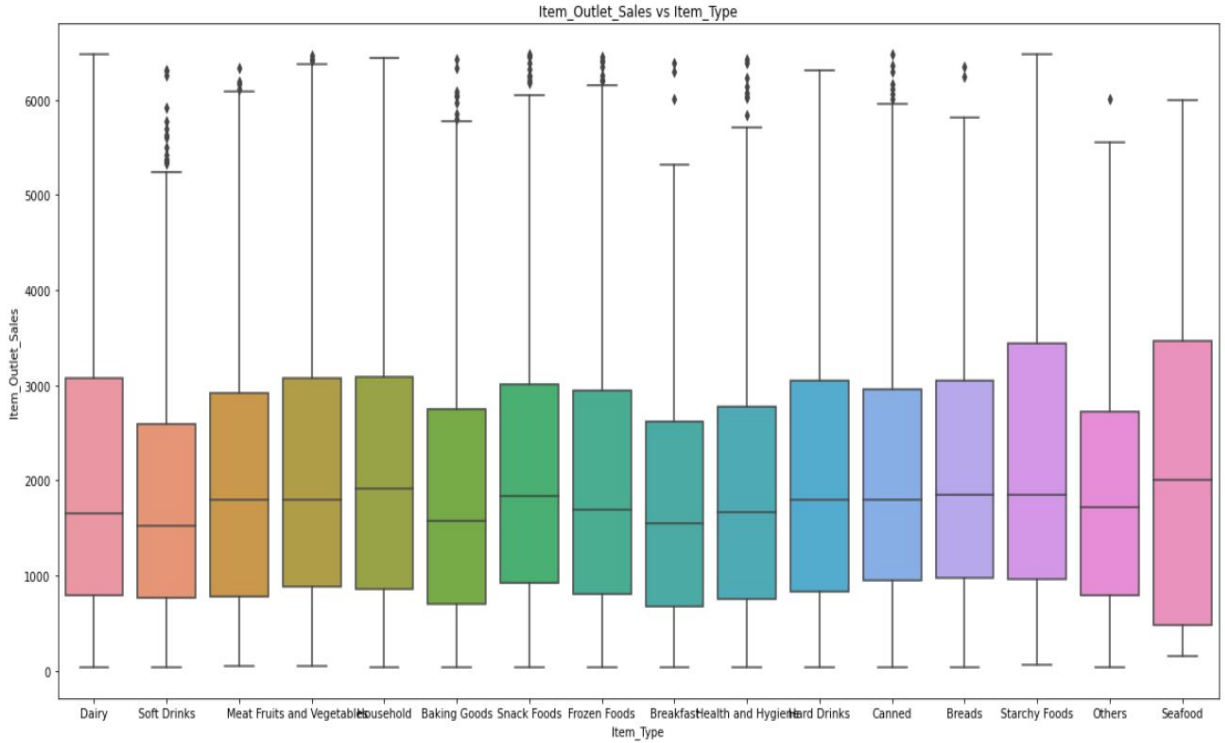
We can see the correlation between these two variables as the mrp of an item increases item outlet sales also increases.



Analysis of Outlet Sales and Item Type

Item_Type	Item_Outlet_Sales	
Baking Goods	1794.3310	11
	539.2980	5
	1518.0240	5
	1211.7560	4
	126.5020	3
		..
Starchy Foods	5452.9020	1
	5642.6550	1
	5712.5640	1
	6301.1312	1
	6478.2340	1

Name: Item_Outlet_Sales, Length: 6949, dtype: int6

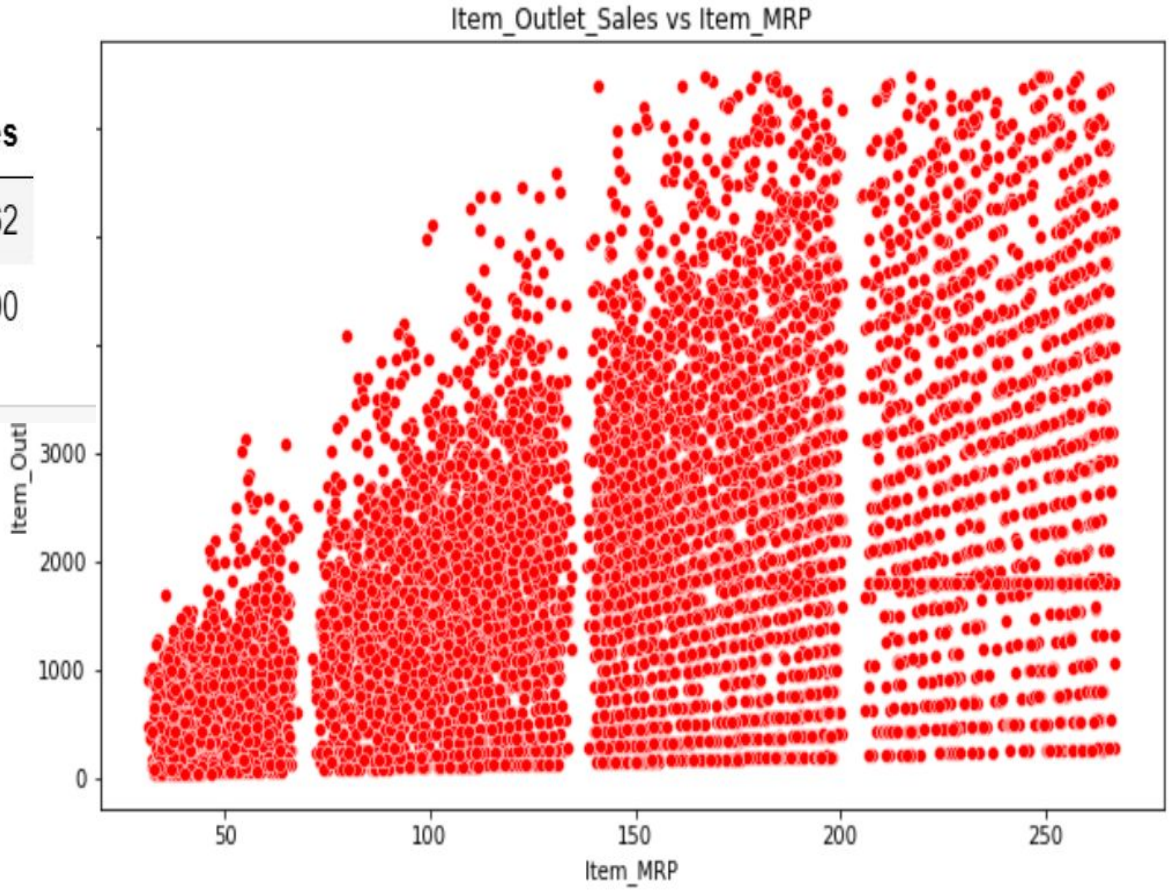


Baking Goods, Starchy Foods contribute towards the item outlet sales.

Analysis of Outlet Sales and Item MRP

	Item_MRP	Item_Outlet_Sales
Item_MRP	1.00000	0.53562
Item_Outlet_Sales	0.53562	1.00000

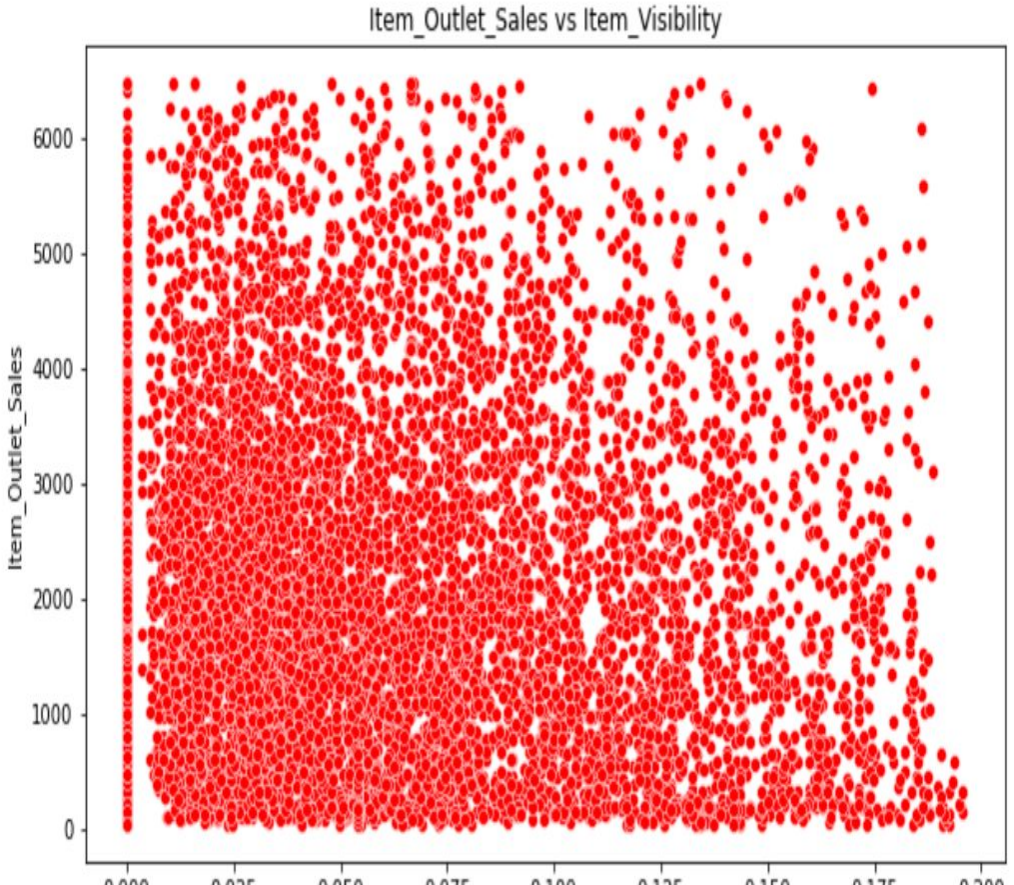
There is moderate positive correlation between these two variables.



Analysis of Outlet Sales and Item Visivility

	Item_Visibility	Item_Outlet_Sales
Item_Visibility	1.000000	-0.065923
Item_Outlet_Sales	-0.065923	1.000000

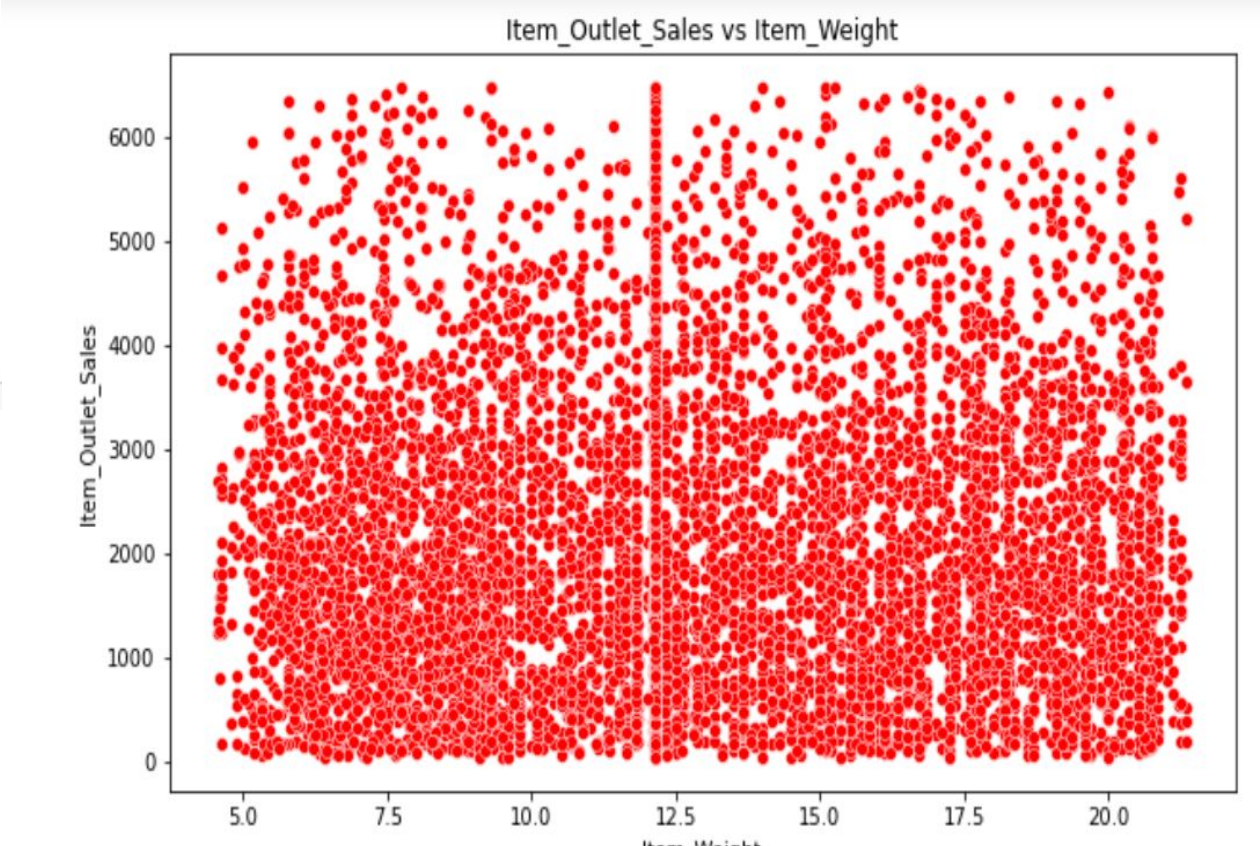
There is no correlation between these two variables.



Analysis of Outlet Sales and Item_Weight

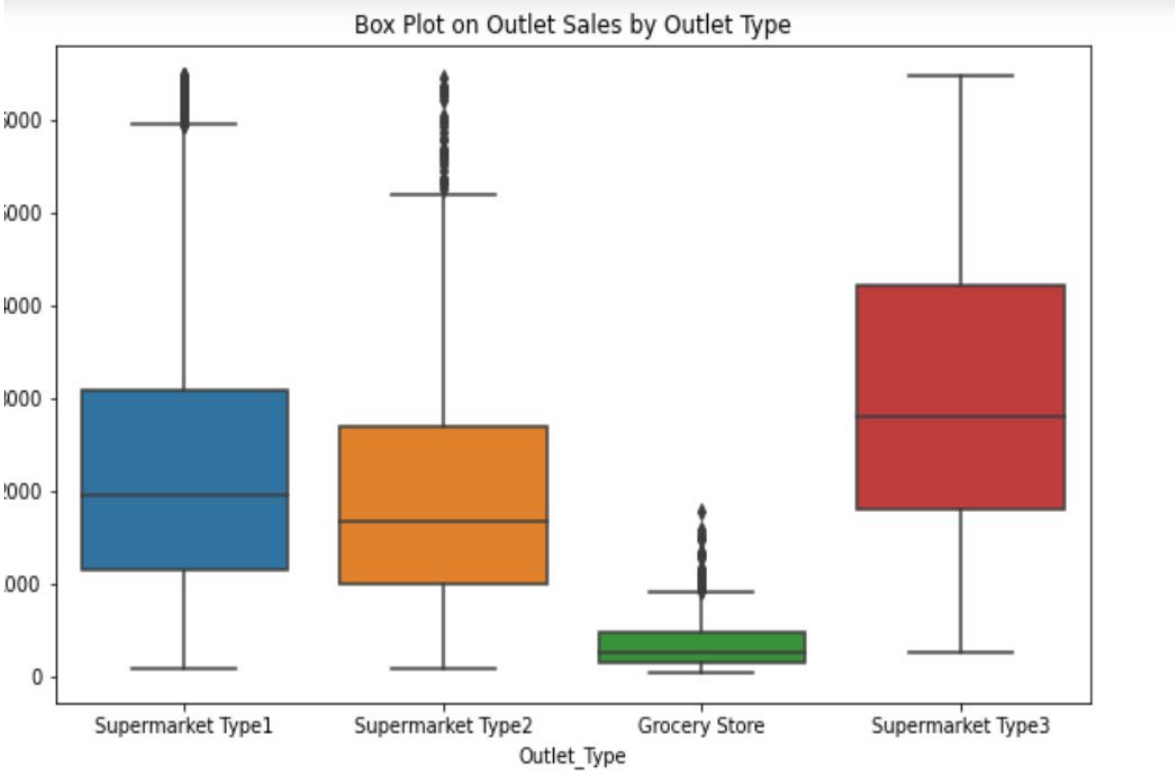
	Item_Weight	Item_Outlet_Sales
Item_Weight	1.000000	0.012705
Item_Outlet_Sales	0.012705	1.000000

There is very low correlation between these two variables.



Analysis of Outlet Sales and Outlet Type

```
Outlet_Type        Item_Outlet_Sales
Grocery Store      223.7088           8
                   123.8388           6
                   175.7712           6
                   280.9676           6
                   311.5944           6
                                      ..
Supermarket Type3  6391.6800          1
                   6431.6280          1
                   6454.9310          1
                   6465.5838          1
                   6478.2340          1
Name: Item_Outlet_Sales, Length: 4672, dtype: int64
```

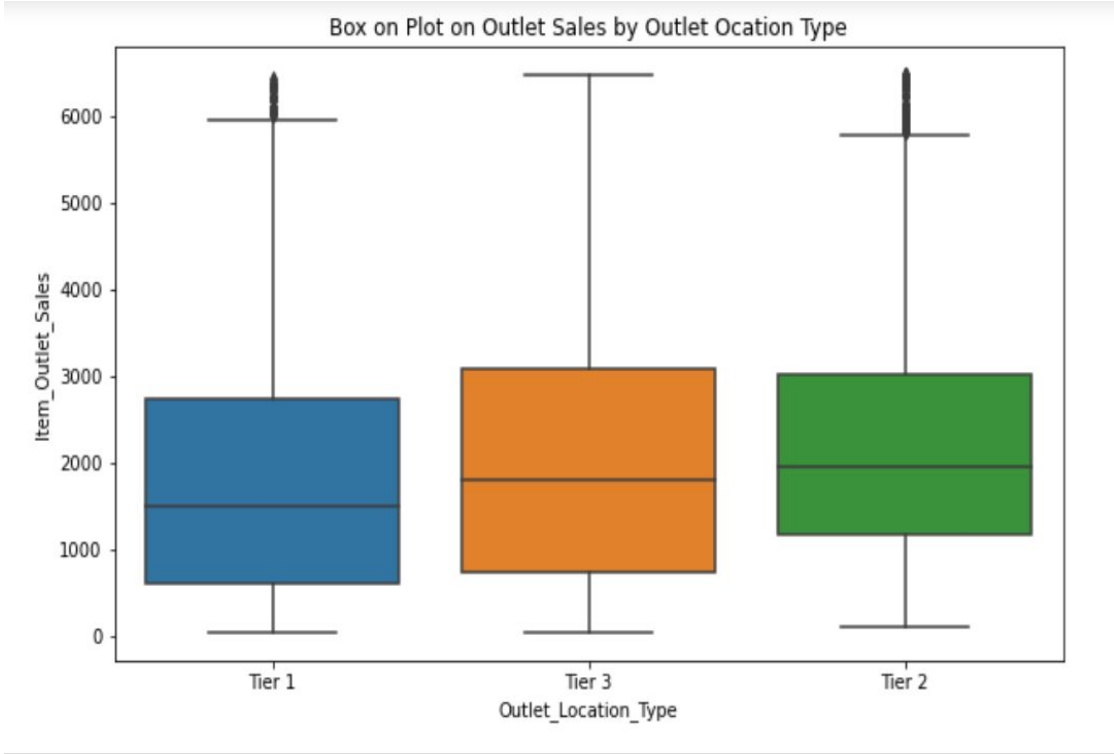


SuperMarket Type 3 has given the most item outlet sales

Analysis of Outlet Sales and Outlate Location Type

Outlet_Location_Type	Item_Outlet_Sales	
Tier 1	1794.3310	22
	1342.2528	8
	1438.1280	7
	1416.8224	6
	123.8388	5
		..
Tier 3	6431.6280	1
	6439.6176	1
	6454.9310	1
	6465.5838	1
	6478.2340	1

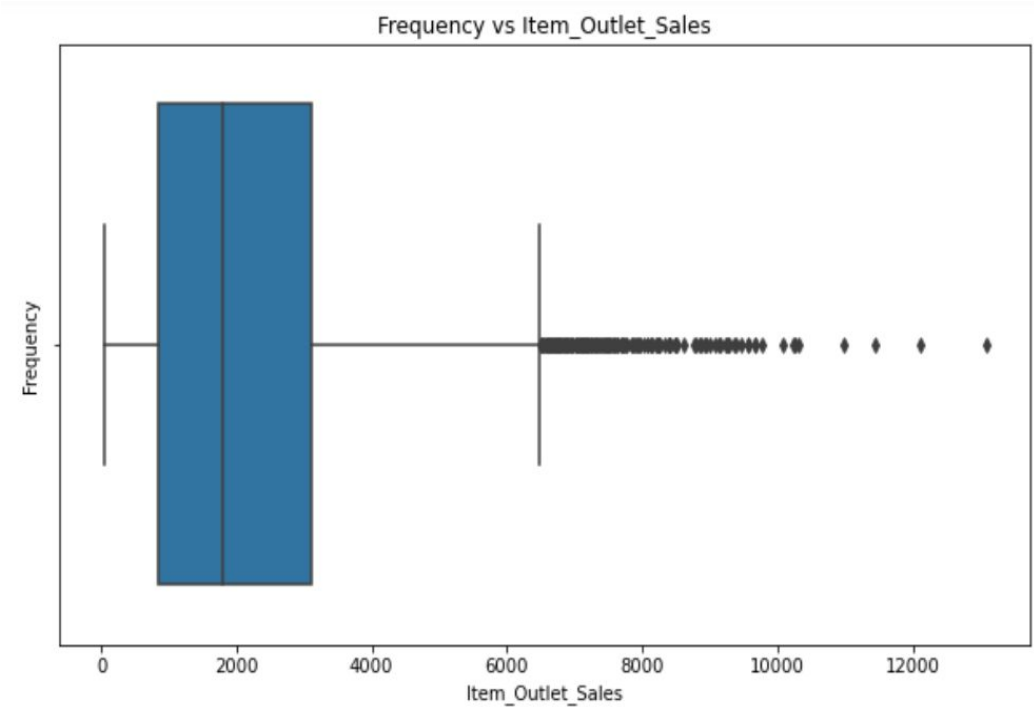
Name: Item_Outlet_Sales, Length: 5436, dtype: int64

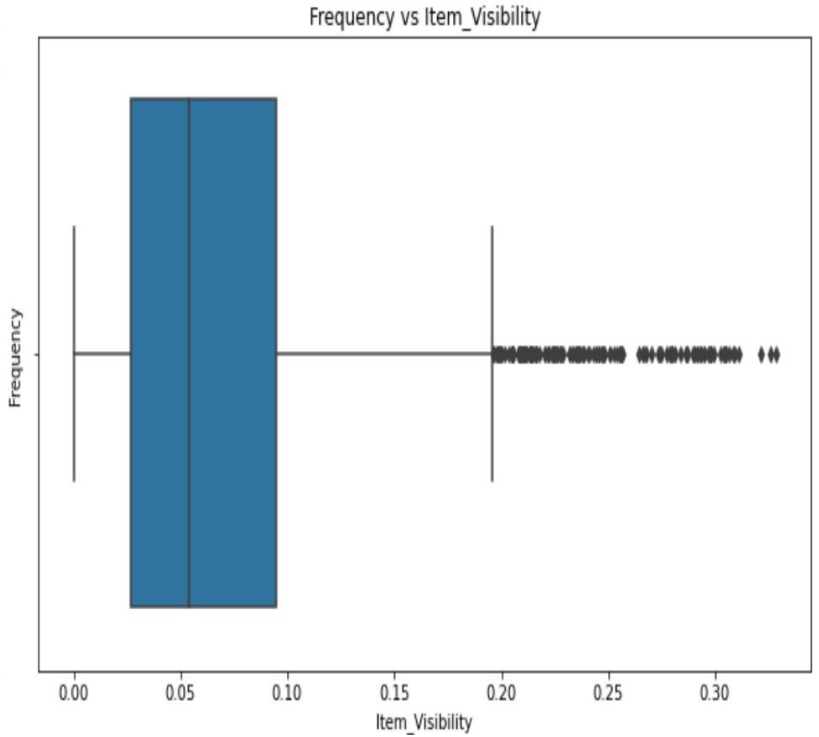


Tier 3 cities produced highest item outlet sales compared to other types of cities.

Outliers

Before removing outliers by its median values





Before

Item_Outlet_Sales
6768.5228
7968.2944
6976.2524
7370.4060
6704.6060
...
7549.5062
6630.0364
7240.5750

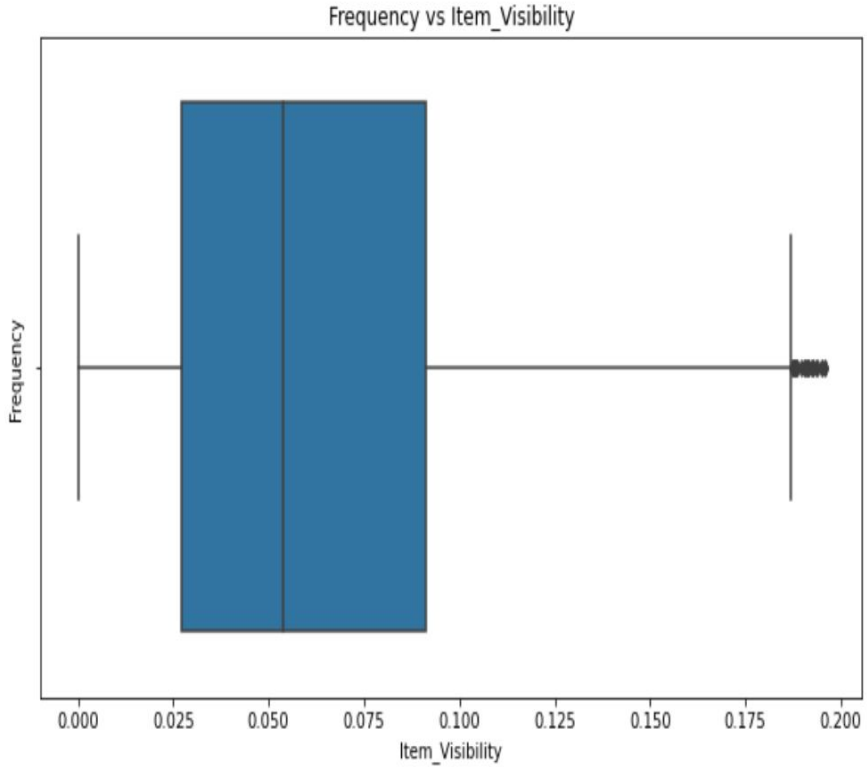
Item_Visibility	It
0.255395	
0.293418	H
0.278974	S
0.291865	V
0.204700	
...	
0.209163	V
0.266397	
0.214125	S
0.227261	

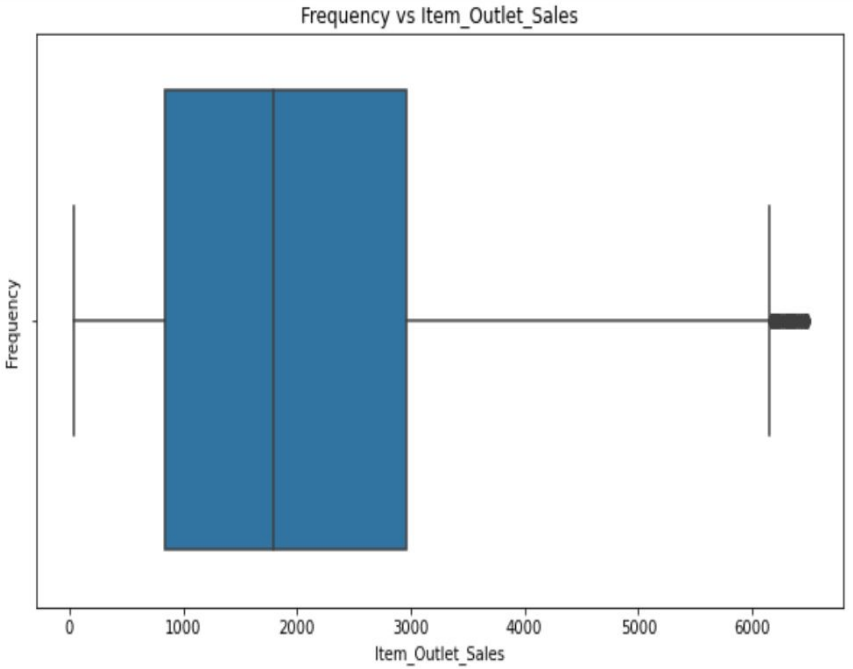
After

Item_Outlet_Sales
3735.1380
443.4228
2097.2700
732.3800
994.7052
...
2778.3834
549.2850
1193.1136
1845.5976

Item_Visibility	It
0.016047	
0.019278	S
0.016760	
0.000000	Ve
0.000000	H
...	
0.056783	
0.046982	
0.035186	H
0.145221	

After removing outliers by its median values





Statistical Analysis

1.Conduct Statistical test to determine the significance of factor such as lItem Type and product attribute on sales.
2.Used the technique like ANOVA to quantify the impact of item categories on sales.
3. a.F-Stistics: 2.602491065669314
b.P-value: 0.00064576257943147 1
4.A p-value below a chosen threshold (e.g., 0.05) indicates strong evidence against the null hypothesis.
5.Since our p-value (0.0006457625) is less than our chosen significance level of 0.05, we reject the null hypothesis. This indicates that at least one item category exhibits a statistically significant difference in outlet sales compared to others.
6.The significant differences observed in sales across item categories suggest that certain types of items may have a stronger impact on sales performance than others. This information can guide strategic decision-making related to inventory management, marketing, and sales promotions.

Thank You