Big Mart Sales Prediction

Lachandra Ash

October 29, 2022

Topic

Largest grocery chain in the nation, Big Mart can be found in most major cities. The Big Mart's management issued a call to all available data scientists, challenging them to develop a machine learning model that can forecast the sales of the items at each retail location (Dr.Bright**,** 2022**)**. During a certain time period, the retailer monitored product sales at 10 locations in various cities (Dr.Bright**,** 2022**).**

Business Problem

Data scientists at Big Mart have compiled sales information from 10 locations in various cities for 1559 goods in 2013 (Jain, 2016). There are also well defined characteristics of each product and retail outlet. The goal is to construct a predictive model that can be used to estimate how well each product will sell in a given retailer.

Background History

The Big Mart Brand's global expansion began in 2007 in the United States of America (Big Mart World, 2022). Located in both upscale and more casual shopping areas, Big Mart is a popular retail establishment. Big Mart has a wide selection of things that people use on a regular basis, such as food, beverages, snacks, desserts, ice cream, tobacco products, magazines, and newspapers. Big Mart provides supplementary services such as phone recharging and wire transfer in addition to the standard fare of goods and necessities. With humble beginnings as a single shop in April 2007, Big Mart is now a globally recognized retail powerhouse with over 40 locations.

Big Mart made a difference in the lives of young clients with our unparalleled services, affordable rates, and high-quality items.

Data Explanation

The dataset was obtained from the Kaggle website. The train dataset has twelve features and eight thousand five hundred twenty-three observations. The test dataset has five thousand six hundred eighty-one variables. The target label is item_outlet_sales. The dataset has one thousand forty-three unique values (Syed, n.d). The twelve attributes within the dataset are:

- 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 particular store. This is the outcome variable to be predicted (Syed, n.d.).

Methods

I used the Jupyter Python 3 Notebook to conduct the analysis of the Big Mart dataset. I imported the libraries, modules, and the train and test dataset into the notebook. I created the train dataframe and the test dataframe. I cleansed both datasets by using isnull(), isnull().sum(), treating the null values, seeking the total percentage of missing data, constructed a barplot to display the total percentage of the missing data, used fillna(0) to fill the missing values with a zero, and checked for null value treatment of the train dataset. The data visualization started with creating the countplots of the dataset features.

I explored the train and test dataframes using the info(), shape, sum(), describe(), corr(), cov(), dtypes(), and index methods.

View Appendix, Table 1: Countplots

The outlet_type countplot displayed four categories including supermarket type 1, supermarket type 2, grocery store, and supermarket type 3. The supermarket type 1 has more than five thousand outlets of its type. There are more supermarket type 1 outlets versus the other outlet types. The grocery store has the second highest number of outlets of its type. Supermarket type 2 and supermarket type 3 both have the least number of outlets of its type.

The outlet_location_type countplot displayed three location types including tier 1, tier 1, and tier 2. The Tier 3 is the most popular outlet location. The tier 3 has the highest number of outlet locations of its type. The tier 2 had the second highest number of outlet locations of its type. The tier 1 outlet location type had the lowest number of outlet locations of its type.

The outlet_size countplot displayed the outlets are constructed in three sizes. Those sizes are medium, high, and small. The medium size outlet has more than five thousand outlets that are medium size. The small size outlets has the second highest number of outlets of its size, and there are more than two thousand small sized outlets. The high size market has the lowest number of outlets of its size, which is less than a thousand outlets.

The item_type countplot possessed categories including dairy, soft drinks, meat, fruits and vegetables, household, baking goods, snack foods, frozen foods, breakfast, health and hygiene, hard drinks, canned foods, breads, starchy foods, other types of foods, and seafood. The fruits and vegetables category and snack foods had the highest number of items, which were twelve hundred or more. The products in fruits and vegetables and snack foods are more in demand versus the other item types. The household items, meats, frozen foods, dairy, baking goods, health and hygiene, and canned foods are the second highest item types that are in higher demand versus the breakfast, hard drinks, breads, starchy foods, other foods, and seafood items.

The item_fat_content countplot displayed five categories including three low fat content, and two regular fat content. There are more items with low fat content versus the items with regular fat content. The three low fat content categories can be combined into one category of low fat content. The regular fat content categories can be combined into one category. Food items with low fat content are the highest in demand.

View Appendix, Table 2: Distribution plots

The item_outlet_sales distribution plot displayed the distribution is right skewed. The item_outlet_sales possess a normal type of distribution.

The outlet_establishment_year distribution plot displayed there were more outlets established in 1985, and the least number of outlets were established in 1998. There were almost the same number of outlets established in 1987, 1997, 1999, 2022, 2004, 2007, and 2009.

The item_mrp distribution plot displayed the item_mrp has a symmetric distribution.

The item_weight distribution plot displayed the item_weight distribution is symmetric.

The item_visibility distribution plot revealed the distribution curve shows it is least susceptible to having outliers. The distribution has a positive type of skew which means the items are less visible.

View Appendix, Table 3: Scatterplots

The item_outlet_sales and item_mrp scatterplot displayed when the item's MRP was high a price, the sales were high as well.

The item_visibility scatterplot displayed if the visibility of the item is less than the number 0.100, the items have higher sales.

The item_outlet_sales and item_weight scatterplot displayed there is no relationship between the two.

View Appendix, Table 4: Barplots

The item_outlet_sales and outlet_establishment_year barplot displayed the very least outlet sales occurred in 1998. There were less than two hundred and fifty sales. The highest number of sales occurred in 2004. The second highest number of sales of sales occurred in 1987. The number of item outlet sales between one thousand and fifty and two thousand revealed they occurred in the outlet establishment years of 1985, 1997, 1999, and 2002.

The item_outlet_sales and outlet_type barplot displayed supermarket type 3 had the highest number of sales. The supermarket type 1 had the second highest number of sales, and supermarket type 2 had the second least number of item sales. The grocery store had the least

number of sales. The tier 2 location type is the best location for an outlet to be to make the most item sales.

The item_outlet_sales and outlet_location_type barplot displayed the tier 2 is the best location type because the outlet in tier 2 received the highest number of item sales. The tier 3 location would be the second best location because the outlet there received the second highest number of item sales. The outlet located in tier 3 location received the least number of item sales.

The item_outlet_sales and oulet_size barplot displayed the high size type of outlet received more item sales versus the other sizes of outlets. The medium sized outlet received almost the same number of sales as the high sized outlet. The small sized outlet received the least number of item sales.

The item_outlet_sales and the item_fat_content barplot showed that the items with regular fat content earned more sales versus the items with low fat content. The item with the regular fat content was in higher demand versus the food item with lower fat content.

The outlet_identifier and item_outlet_sales barplot displayed OUT27 earned the most number of sales versus the other outlets. The OUT010 and OUT019 received the least number of sales. The Out013, OUT017, Out035, Out046, and Out049 outlets made near or a little more than two thousand number of sales. The OUT016 and OUT045 received almost the same number of item sales.

View Appendix, Table 5: Pairplot

The pairplot displays the distinct types of graphs of the dataframe, that was automatically created.

View Appendix, Table 6: Heatmap

The substantial positive connection between item MRP and item outlet sales indicates that as the item's MRP rises, so do its sales. Similar to this, we can see that item visibility and item outlet sales are inversely correlated, meaning that if an item is less visible, its sale will be higher, and vice versa if it is more prominent, its sale would be lower.

The Gradient Booster Regressor Model

I used the gradient boosting regressor model to predict the item outlet sales at each outlet. The da taset was trained, tested, and split to provide a validated dataset. I used a standard scaler to fit the model. The root mean square error is 237.7863656123978.

View Appendix, Table 7: The Mean Test Score Results

The gridsearchcv was utilized to seek and obtain the best estimator and root mean square to the calculated score. The display results revealed the mean test score, standard deviation score, and params score.

View Appendix, Table 8: The Predicted Target Results

The predicted prices were sent to the submission_bigM file. The item outlet sales are the predicted sales of each item at each outlet.

Conclusion

The supermarket type 1 has the highest number of its outlet type, and tier 3 outlet location is the best city for outlet location type. The tier 3 location has the most outlets. Most of the outlet sizes are medium sized. The fruits and vegetables and snack foods items have the highest consumer demand. Big Mart should focus on increasing the demand for the breakfast, hard drinks, breads, starchy foods, other foods, and seafood items, to gain more item sales. There is a higher number of low fat content versus regular fat content food items.

The higher the item's mrp, the higher its sales will increase. The impact of item outlet sales on outlet establishment year is the outlet established in 2004 had the highest number of item outlet sales, but the least number of outlets were established in 1998. The supermarket type 3 gained the highest number of item sales, and the outlet located in tier 2 location gained the highest number of item sales, making it the best location to earn more profit from consumers.

The high sized outlets are the larger sized outlets that received the highest number of item sales. The food items with regular fat content are in higher demand than the foods with low fat content. The regular fat content foods earned the highest number of food item sales. The OUT027 outlet obtained the highest number of item sales, which makes it the most popular outlet with more items in demand.

Assumptions

I assumed the outlets that are located within the tier 2 city locations will continue to gain the highest number of sales because it is in an urban city, and there are many consumers who visit the outlets in the tier 2 locations. The larger stores may have a higher number of item sales versus the other sizes of the stores.

Limitations

A limitation within the dataset was the item_weight and outlet_size had many null values. The training and test datasets had a total percentage of 445.44 % of missing data. The missing data was treated with the fillna method.

Challenges | Issues

The missing values were treated, and the columns were lacking the missing values. The regular fat contents and low fat item contents were two categories that were already separated into more categories of the same type of contents. The categories can be combined.

Future Uses | Additional Applications

The big mart sales prediction can be used by businesses that need to predict their future item sales. The businesses can use the big mart sales prediction to predict their income against their competition.

Recommendations

I recommend more accurate data placed inside of the dataset.

Implementation Plan

The big mart sales prediction model can be useful for supermarkets, grocery stores, and other businesses that sale items within the store or online.

Ethical Issues

The stores need to ensure that they protect their consumers' account information with excellent quality security measures. These account security measures should be used for online and physical stores.

References

Big Mart (2022). Big Mart World. Retrieved from Big Mart World | LinkedIn, on October 29, 2022.

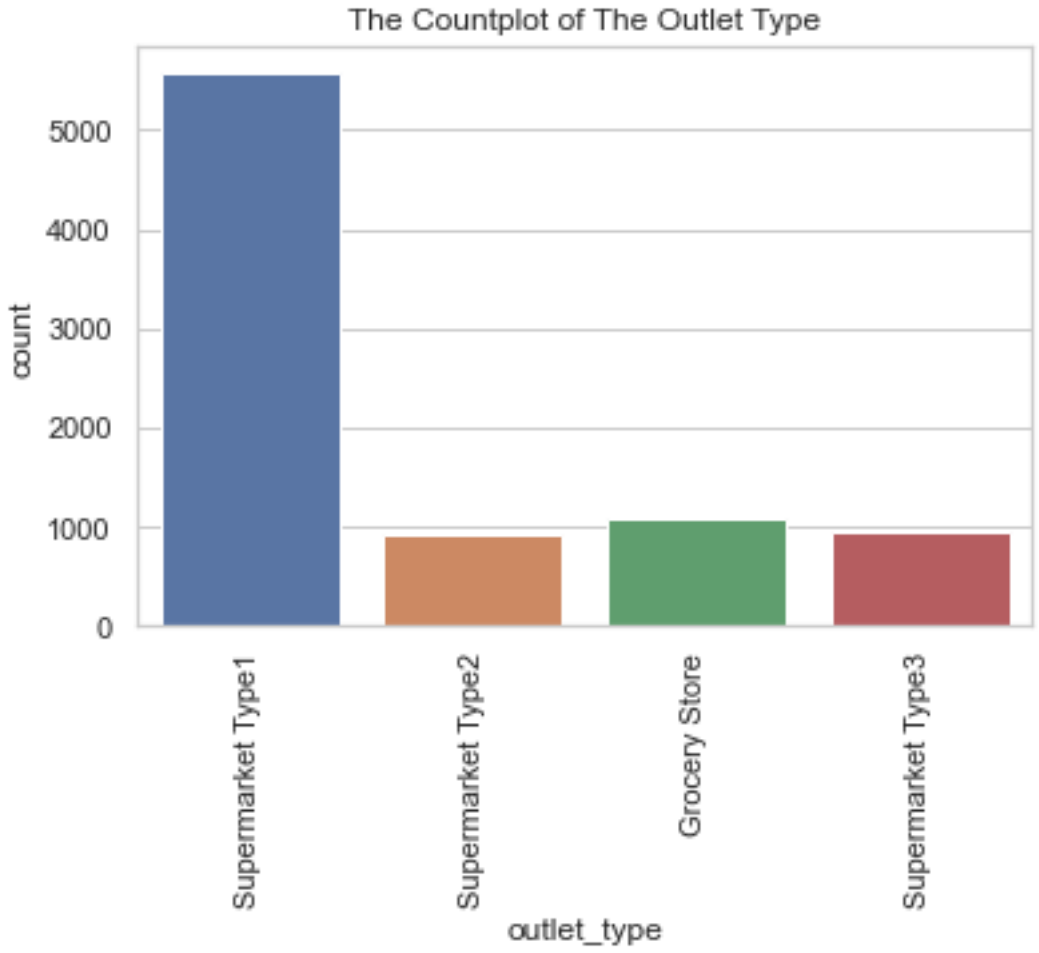
Dr. Bright (2022). Big Mart Sales-Data Science Projects. Retrieved from BIG MART SALES-Data Science Projects | by Dr Bright (PhD Data Science) | Total Data Science | Medium, on October 29, 2022.

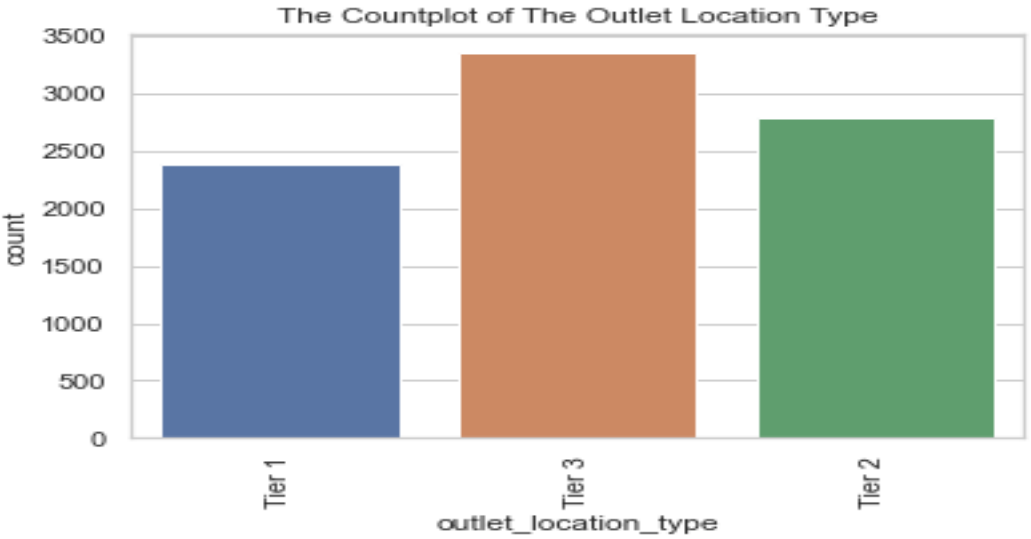
Jain, A. (2016). Approach and Solution to Break in Top 20 of Big Mart Sales Prediction. Retrieved from Approach and Solution to break in Top 20 of Big Mart Sales prediction (analyticsvidhya.com), on October 29, 2022.

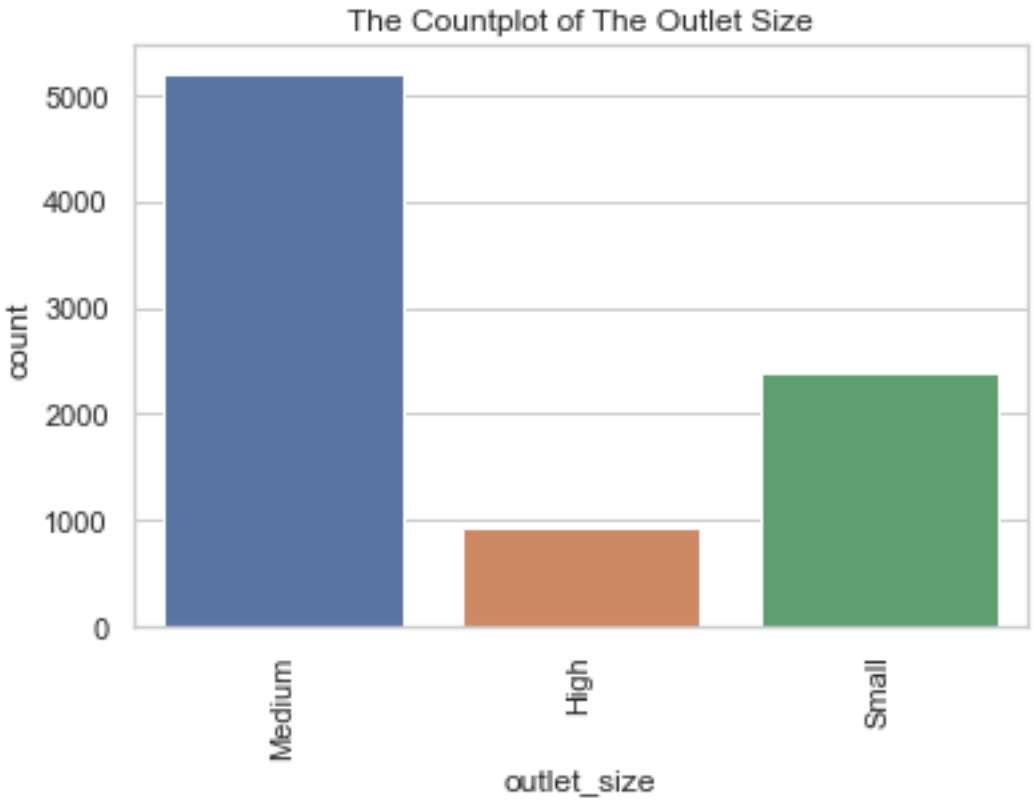
Syed, K (n.d.). Big Mart Sales Prediction Dataset. Retrieved from Big Mart Sales Prediction | Kaggle, on October 29, 2022.

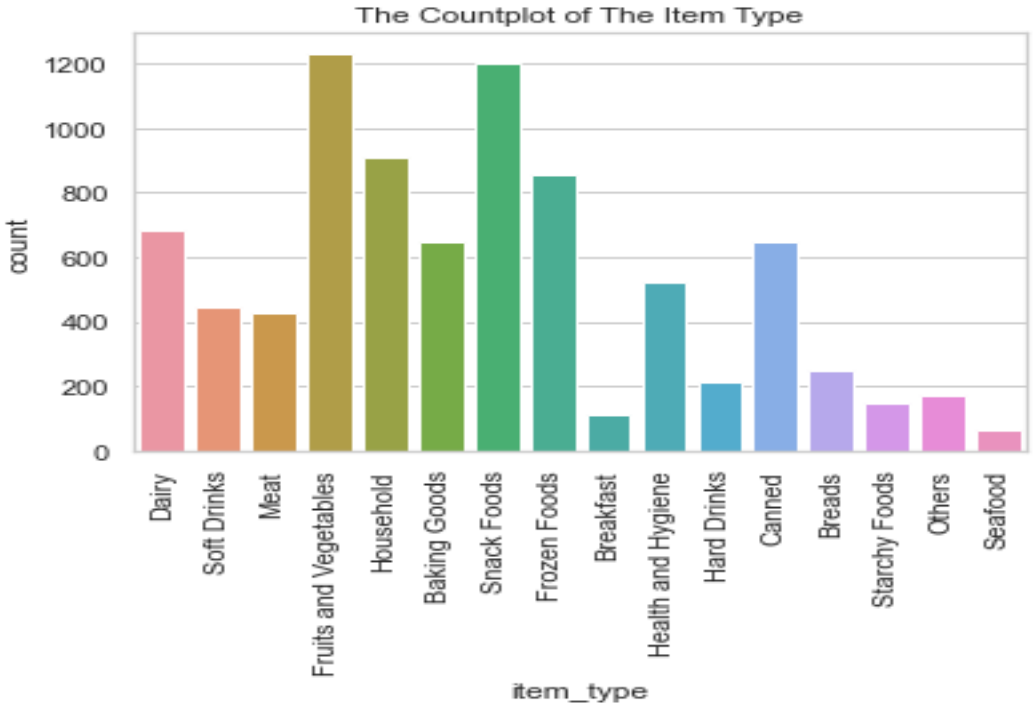
Appendix:

Table 1: Countplots






The Countplot of The Outlet Size
count
5000
4000
3000
2000
1000
0
Medium
High
Small
outlet_size


The Countplot of The Item Type
count
1200
1000
800
600
400
200
0
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
item_type

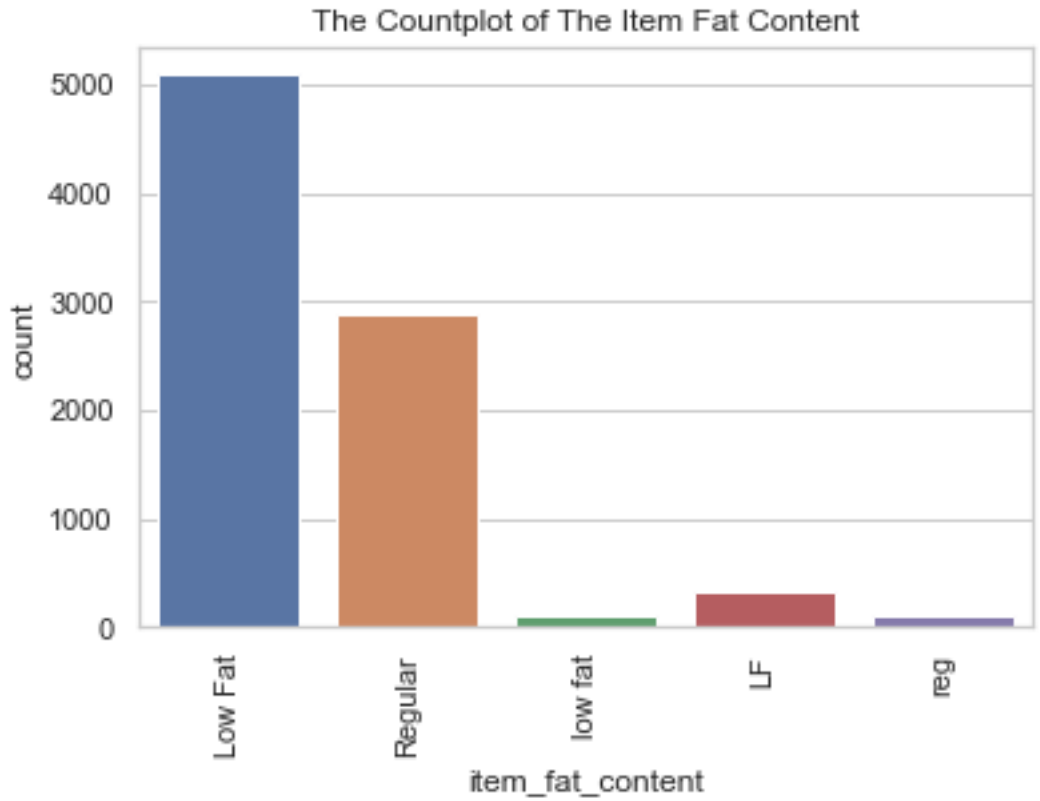
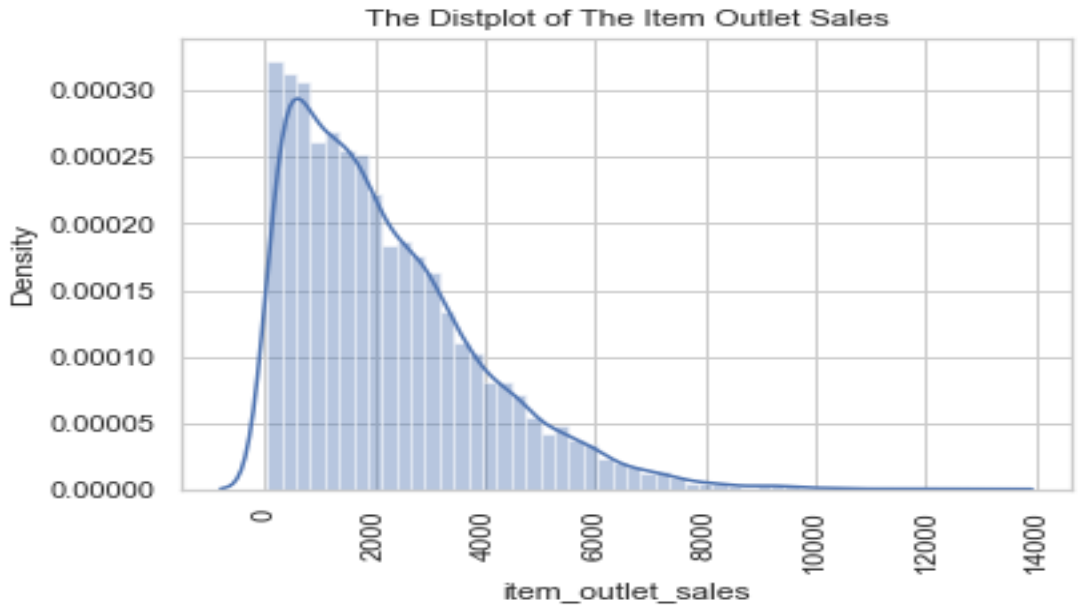
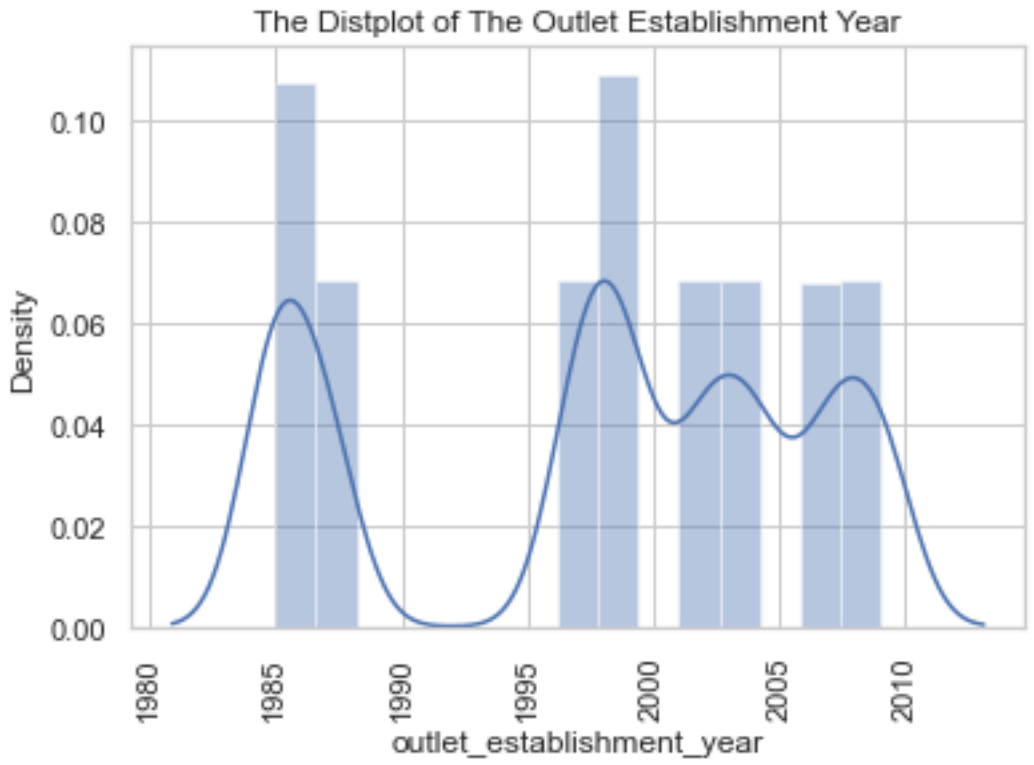
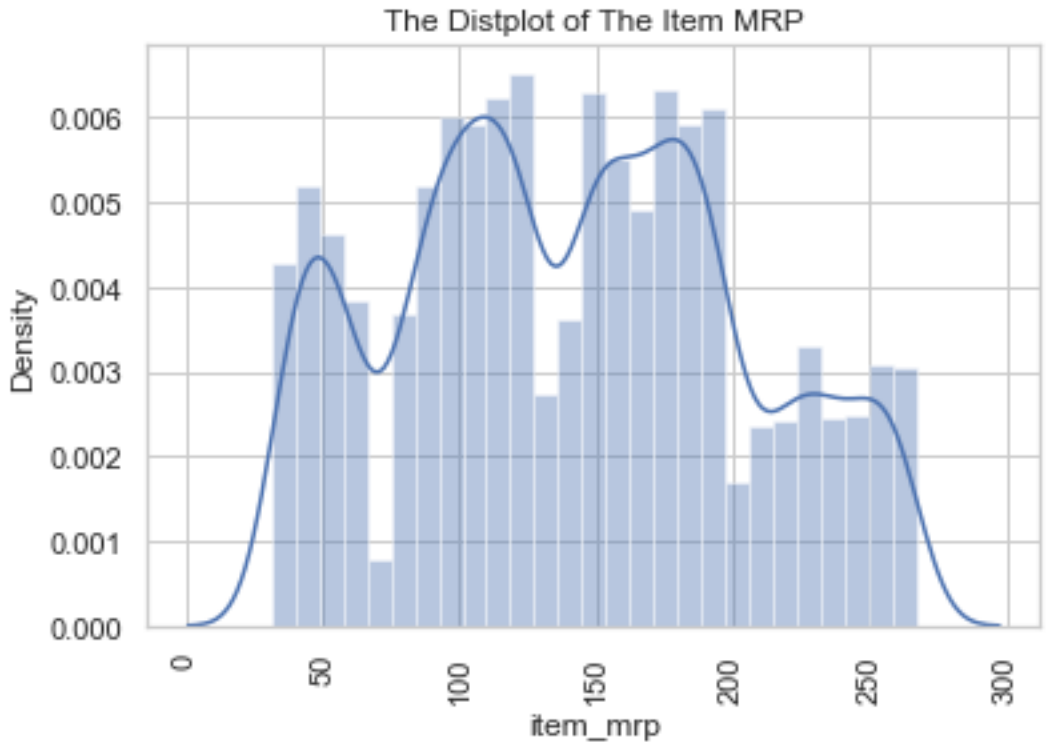
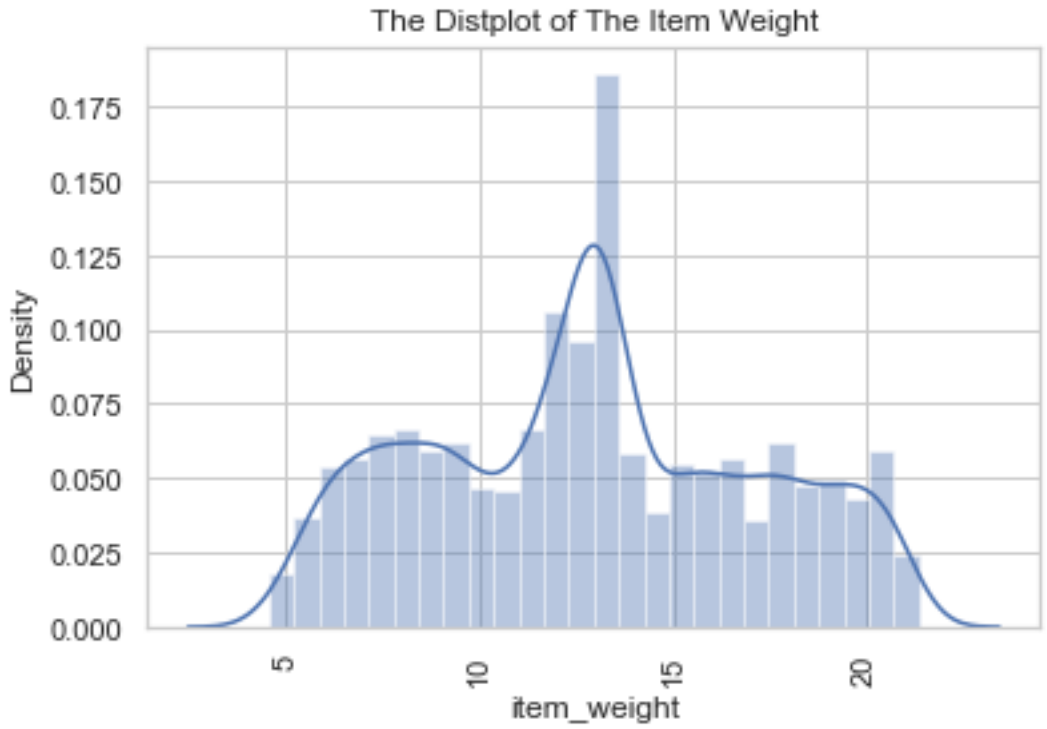


Table 2: Distribution Plots




The Distplot of The Outlet Establishment Year
Density
0.10
0.08
0.06
0.04
0.02
0.00
1980
1985
1990
1995
2000
2005
2010
outlet_establishment_year


The Distplot of The Item MRP
Density
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
50
100
150
200
250
300
item_mrp


The Distplot of The Item Weight
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Density
5
10
15
20
item_weight

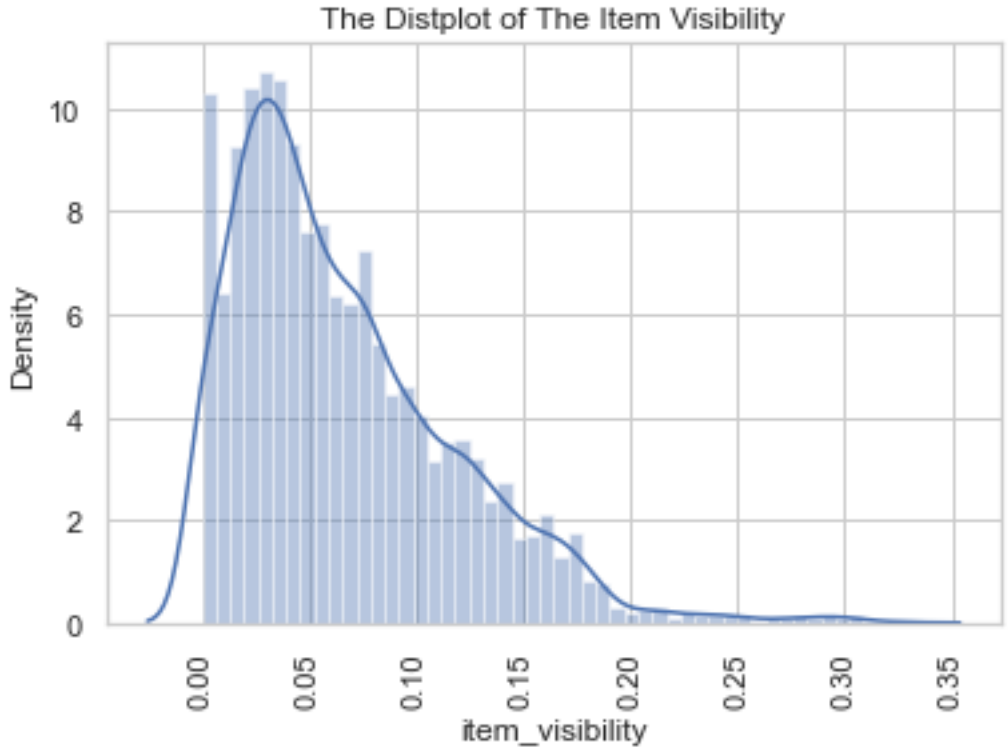
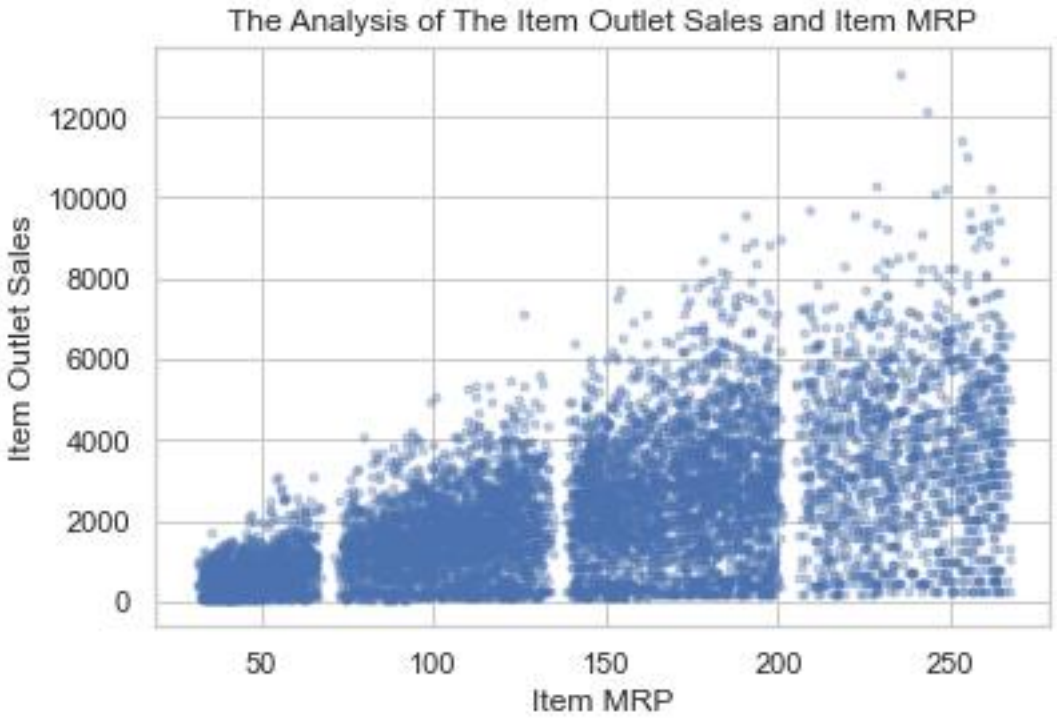
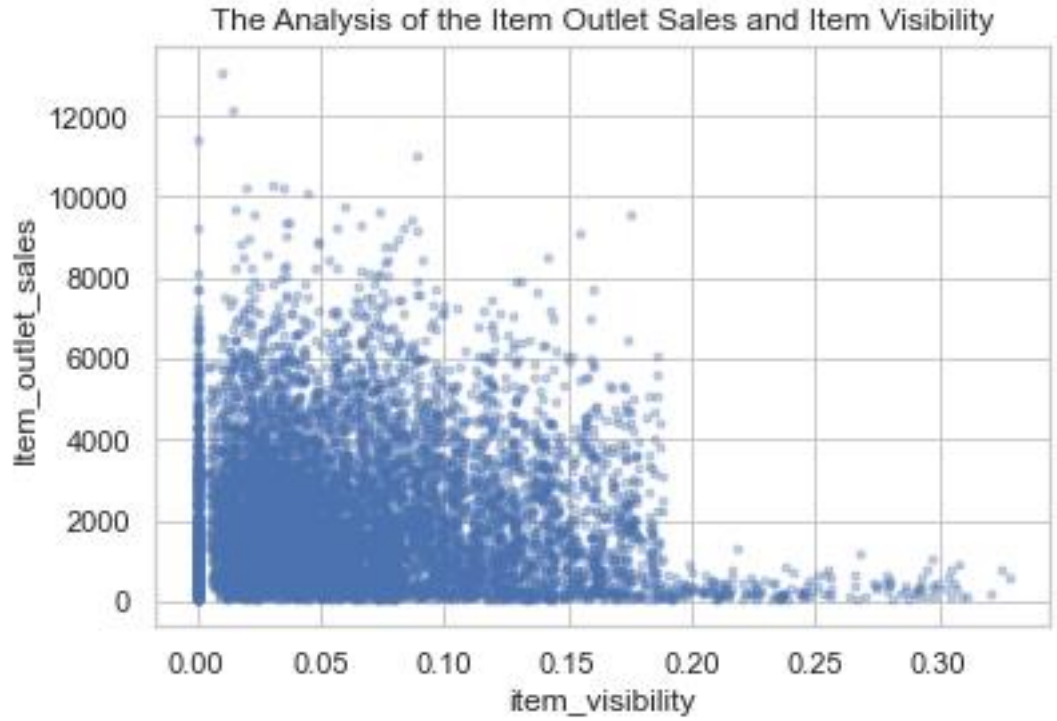

The Distplot of The Item Visibility
10
8
6
4
2
0
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
item_visibility

Table 3: Scatterplots





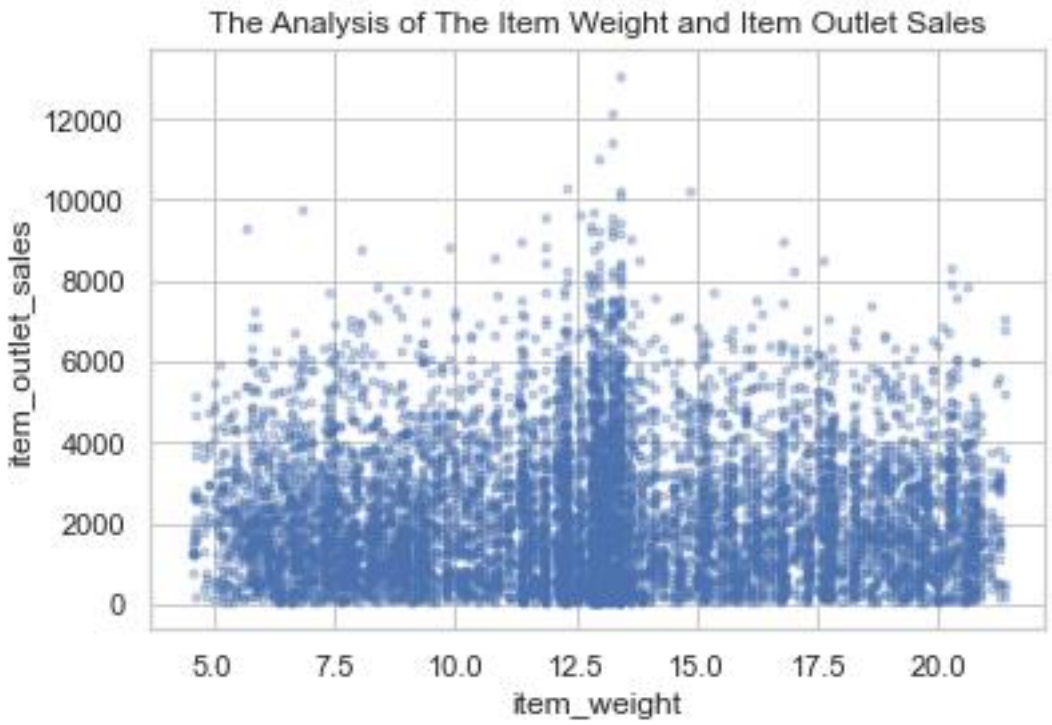
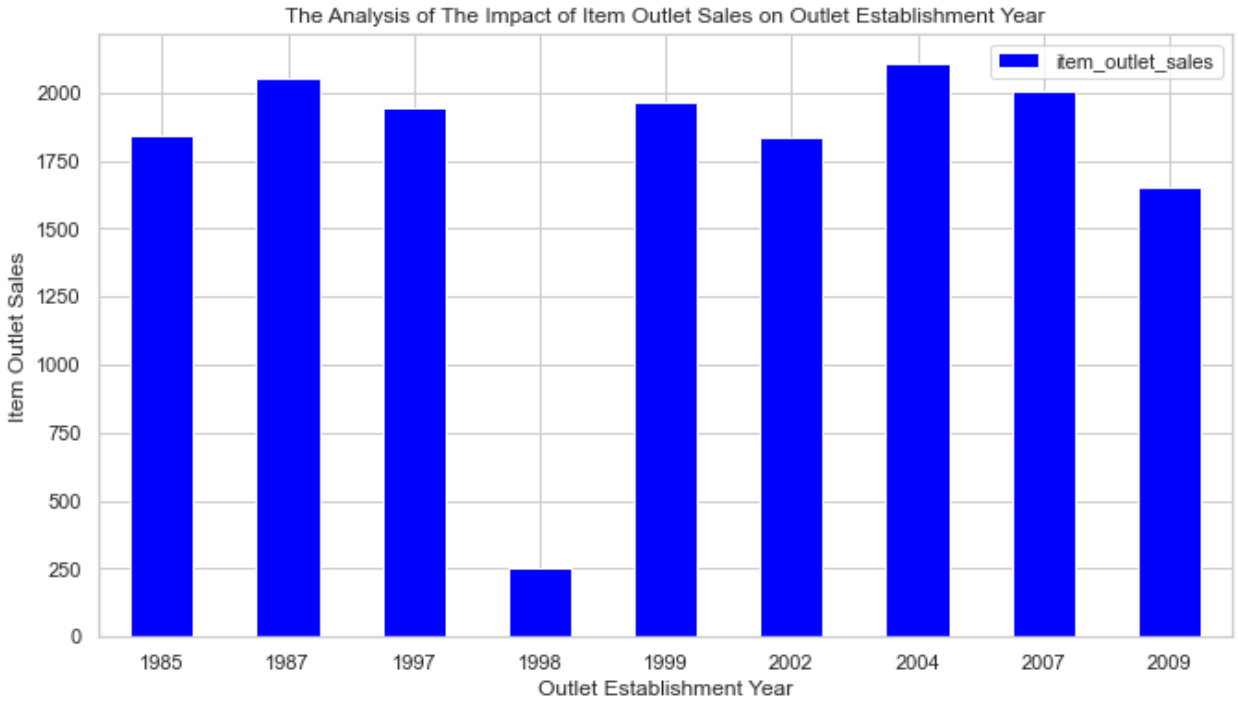
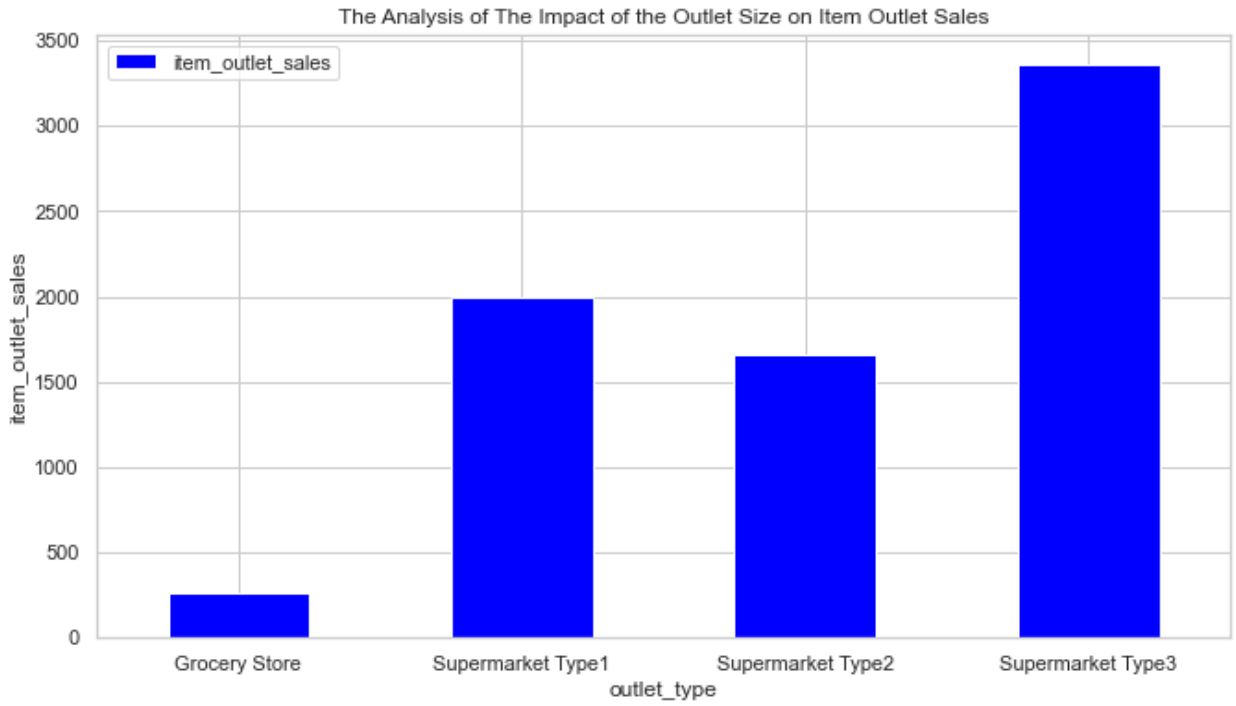
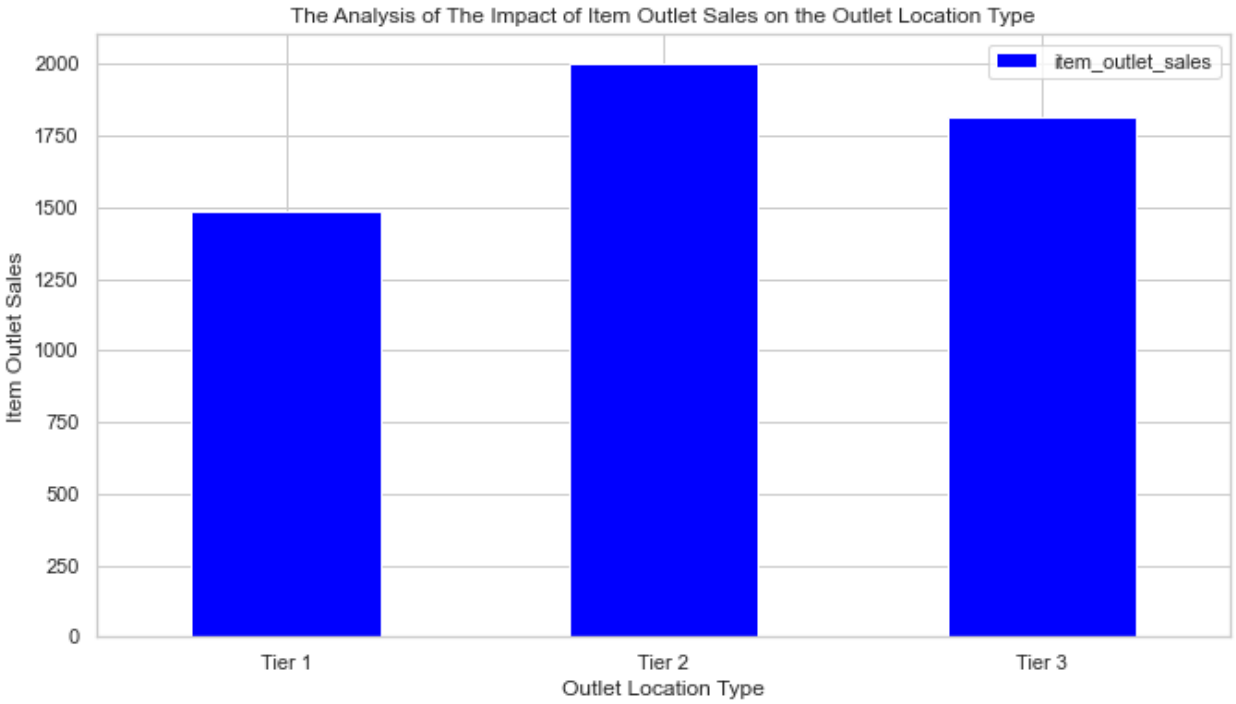
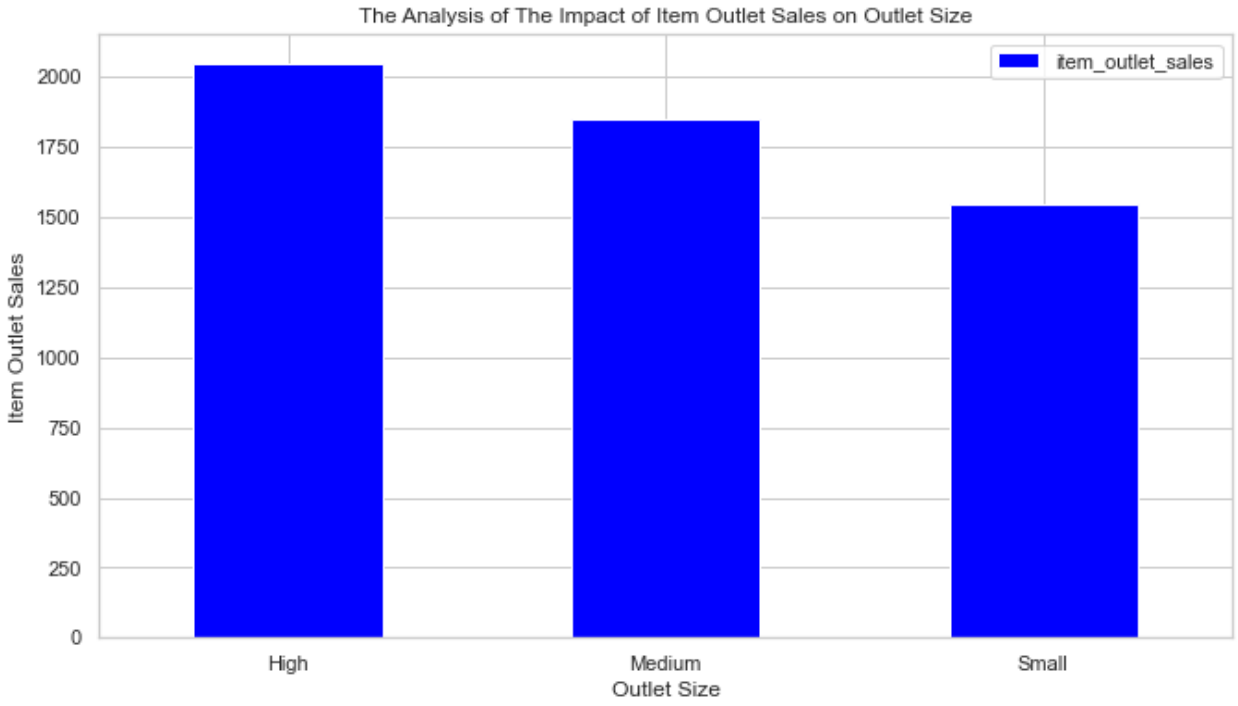


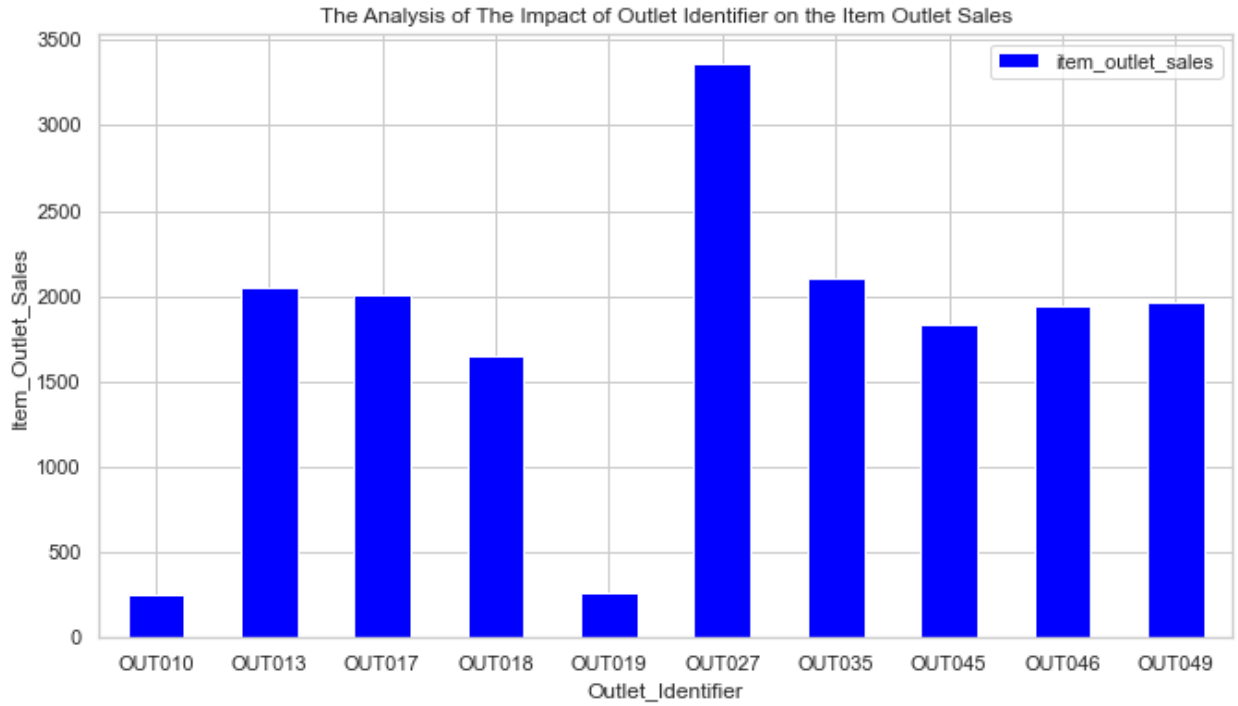
Table 4: Barplots

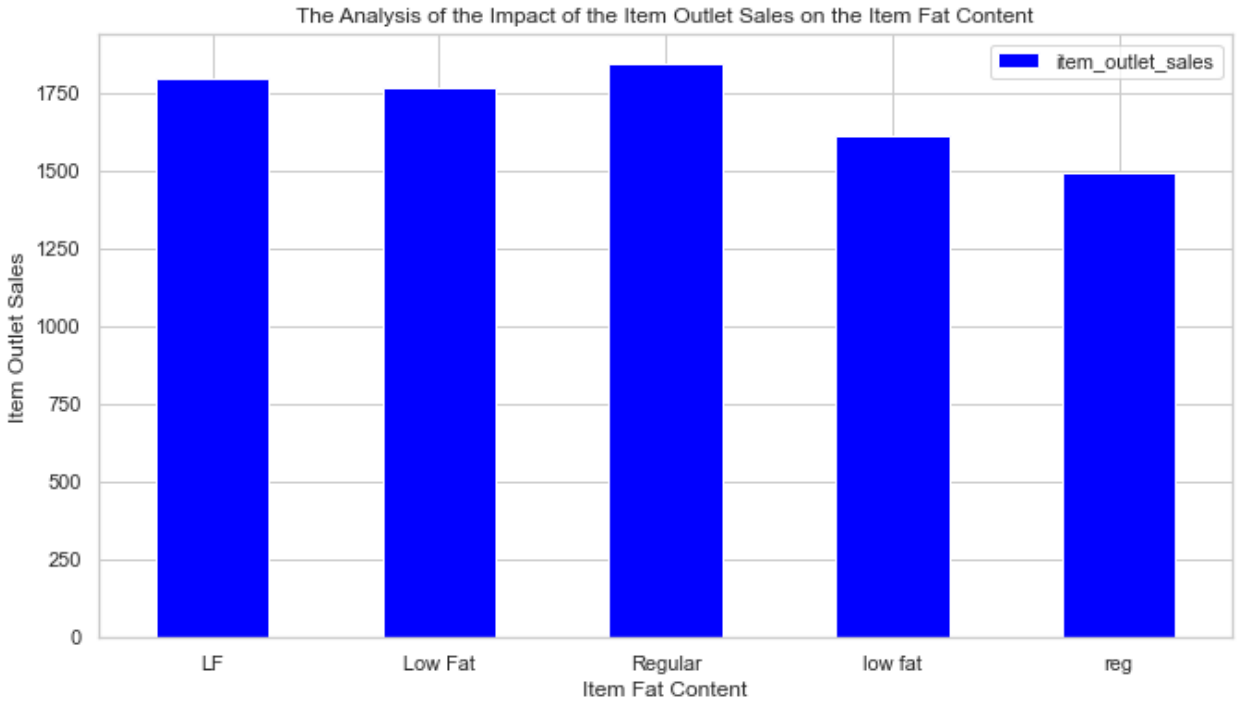



The Analysis of The Impact of the Outlet Size on Item Outlet Sales
item_outlet_sales
3500
3000
2500
2000
1500
1000
500
0
item_outlet_sales
Grocery Store
Supermarket Type1
Supermarket Type2
Supermarket Type3
outlet_type


The Analysis of The Impact of Item Outlet Sales on the Outlet Location Type
item_outlet_sales
2000
1750
1500
1250
1000
750
500
250
0
Item Outlet Sales
Tier 1
Tier 2
Tier 3
Outlet Location Type


The Analysis of The Impact of Item Outlet Sales on Outlet Size
item_outlet_sales
Item Outlet Sales
2000
1750
1500
1250
1000
750
500
250
0
High
Medium
Small
Outlet Size


The Analysis of The Impact of Outlet Identifier on the Item Outlet Sales
item_outlet_sales
Item_Outlet_Sales
3500
3000
2500
2000
1500
1000
500
0
OUT010
OUT013
OUT017
OUT018
OUT019
OUT027
OUT035
OUT045
OUT046
OUT049
Outlet_Identifier



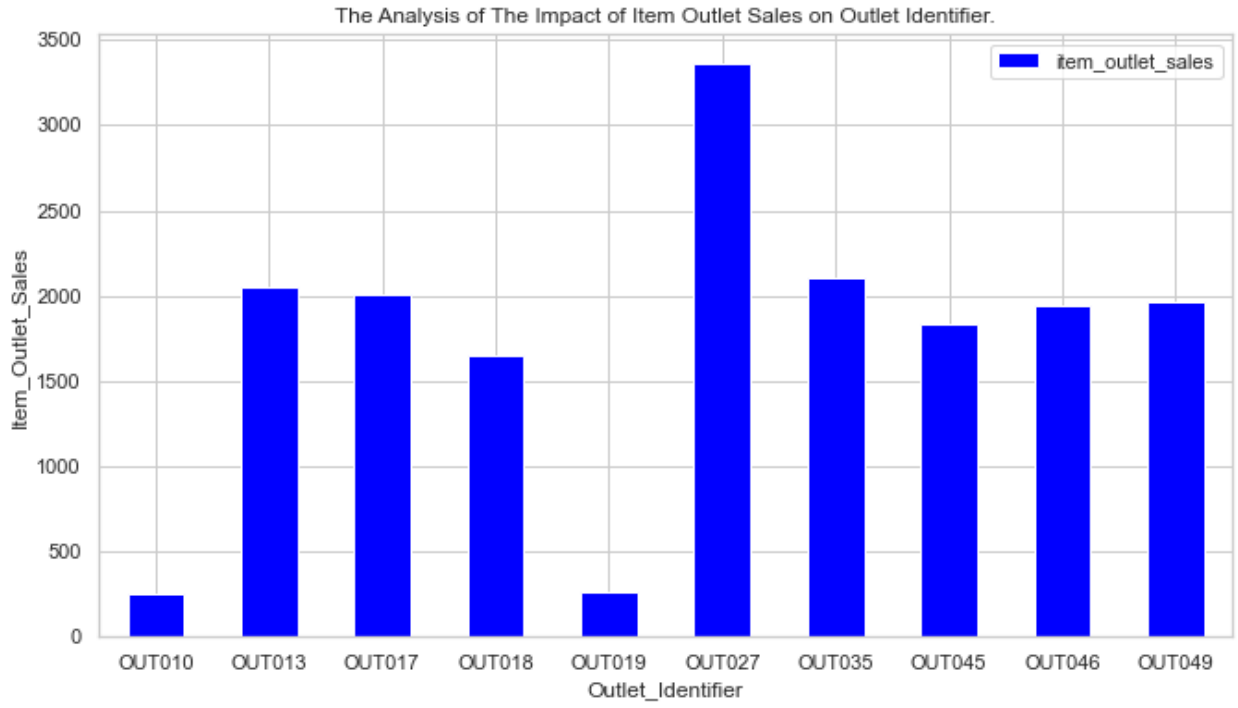


Table 5: Pairplot

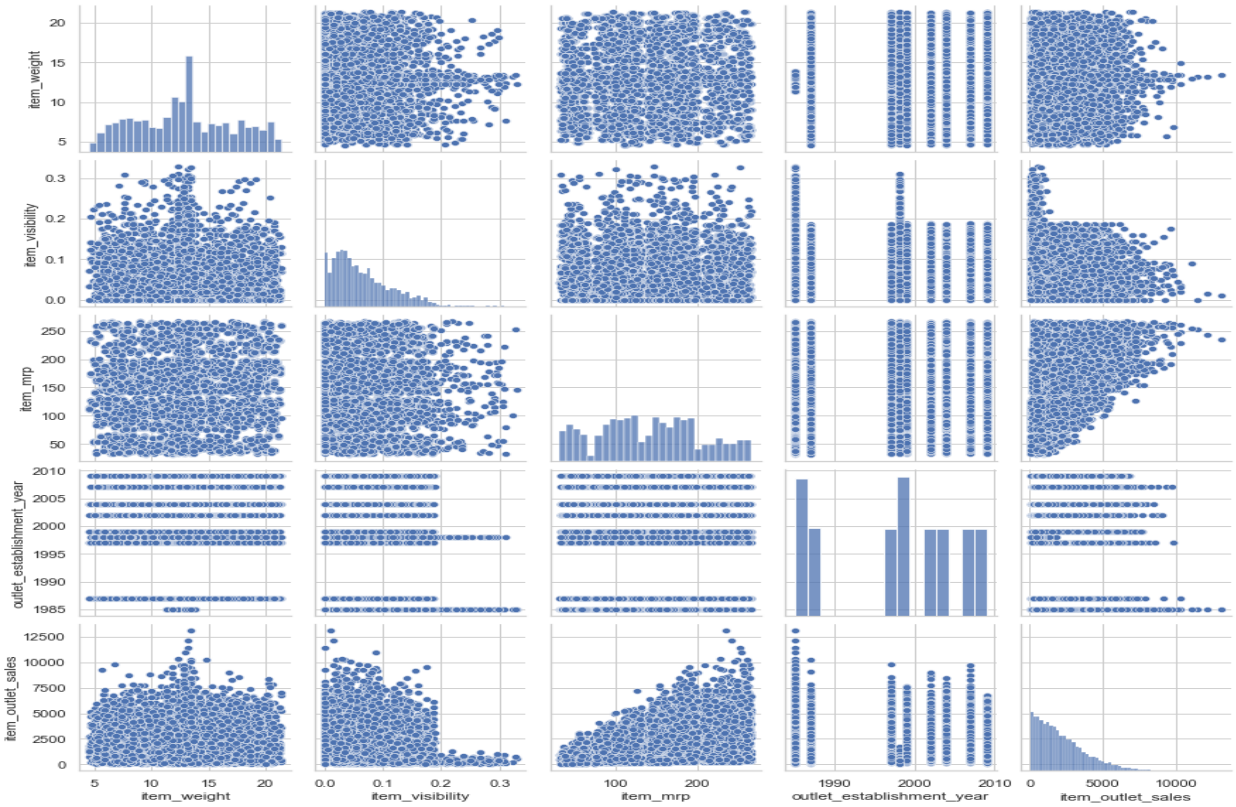


Table 6: Heatmap

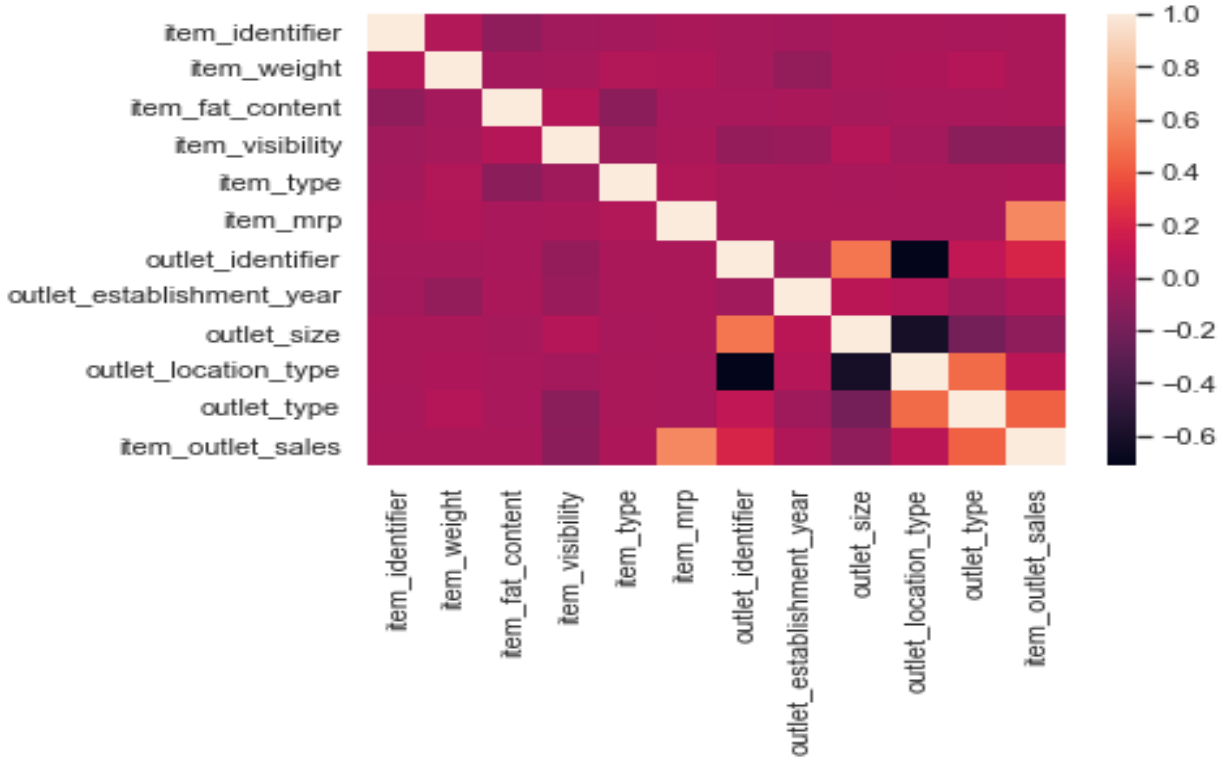


Table 7: The Mean Test Score Results

```
Best: -43.323055 using {'n_estimators': 200}
-51.234186 (4.141885) with: {'n_estimators': 80}
-50.784895 (4.080098) with: {'n_estimators': 82}
-50.344291 (4.025789) with: {'n_estimators': 84}
-50.155851 (3.969136) with: {'n_estimators': 85}
-49.405760 (3.755032) with: {'n_estimators': 90}
-49.330617 (3.851460) with: {'n_estimators': 91}
-49.170391 (3.783962) with: {'n_estimators': 92}
-49.019193 (3.902931) with: {'n_estimators': 94}
-48.990893 (3.920689) with: {'n_estimators': 95}
-46.808444 (3.854575) with: {'n_estimators': 150}
-43.323055 (3.754444) with: {'n_estimators': 200}
```

Table 8: The Predicted Target Results

	item_identifier	outlet_identifier	item_outlet_sales
0	156	9	2522.776312
1	8	3	439.177044
2	662	9	1638.403723
3	1121	0	684.291626
4	1297	1	876.601458
5	758	3	539.370855
6	696	1	352.334185
7	738	5	2640.259917
8	440	7	941.288598
9	990	2	2889.805184