

## **BigMart Sales Insights: A Data-Driven Approach**



# **BigMart**

# case story

**BigMart** is a large retail chain operating across multiple cities, selling thousands of products ranging from food and beverages to household items. The management has noticed significant variations in sales across different stores, with some products performing very well in certain stores while performing poorly in others.

The company aims to understand the factors affecting sales for each product and each store, in order to make informed business decisions. The goal is to improve overall sales and increase profitability by analyzing the available data on products, stores, and past sales.

# Business Questions

- Which products achieve the highest sales across all stores?
- Is there a relationship between product type (Item\_Type) and sales?
- Which stores (Outlet\_Identifier) achieve the highest and lowest sales?
- Does the store size (Outlet\_Size) affect sales performance?
- Does the store location type (Outlet\_Location\_Type) impact sales?
- Does the product price (Item\_MRP) influence sales volume?
- **Which products have low sales and might require further analysis?**
- **Which product categories require increased inventory in high-performing stores?**
- **Is there a relationship between product attributes (e.g., weight, fat content) and sales performance?**

# Dataset Overview

- This dataset contains historical sales records for products sold across multiple outlet stores.

It includes detailed information about items, store characteristics, and sales performance, making it suitable for retail analytics, demand forecasting, and store-level performance analysis.

- **◆ Number of records:**  
≈ 8,500 rows (after cleaning)
- **◆ Number of features (columns):**  
11 features after preprocessing  
(including engineered features such as Outlet\_Age and Weight\_Category)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8523 entries, 0 to 8522
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Item_Identifier    8523 non-null   object 
 1   Item_Weight         7060 non-null   float64
 2   Item_Fat_Content    8523 non-null   object 
 3   Item_Visibility     8523 non-null   float64
 4   Item_Type           8523 non-null   object 
 5   Item_MRP            8523 non-null   float64
 6   Outlet_Identifier   8523 non-null   object 
 7   Outlet_Establishment_Year 8523 non-null   int64  
 8   Outlet_Size          6113 non-null   object 
 9   Outlet_Location_Type 8523 non-null   object 
 10  Outlet_Type          8523 non-null   object 
 11  Item_Outlet_Sales    8523 non-null   float64
dtypes: float64(4), int64(1), object(7)
memory usage: 799.2+ KB
```

# Data Issues

The screenshot shows a Jupyter Notebook interface with three code cells:

- Item\_Weight:** A cell displaying the count of items in the 'Item\_Weight' column as 1463.
- Item\_Visibility:** A cell showing the minimum value of 'Item\_Visibility' as 0.0.
- Item\_Fat\_Content:** A cell displaying the unique values in the 'Item\_Fat\_Content' column, which are 'Low Fat', 'Regular', 'low fat', 'LF', and 'reg'.

## Missing Values in Item\_Weight

- The `Item_Weight` column had several missing values, which can distort statistical analysis and model training.

## Zero Values in Item\_Visibility

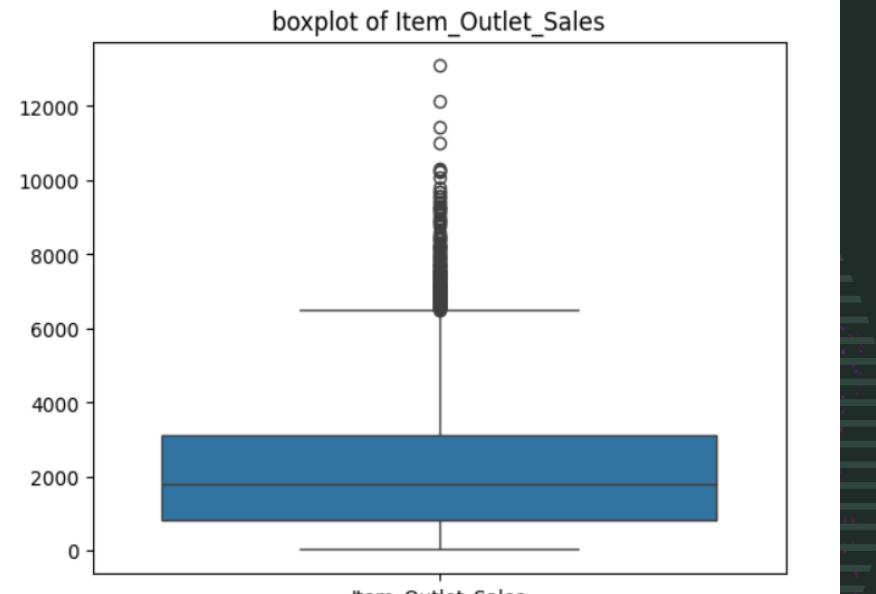
- `Item_Visibility` had many zeros, which is impossible in real-life retail data (an item always has some visibility). These zeros negatively impact correlation, sales relationships, and model results.

## Inconsistent Categorical

- The column contained duplicates written differently:
  - "Low Fat", "LF", "low fat"
  - "Regular", "reg"

# Data Issues

```
df['Outlet_Age'] = 2025 - df['Outlet_Establishment_Year']
```



Outlet\_Size

2410

## Missing Values in Outlet\_Size

- Outlet\_Size had missing values.

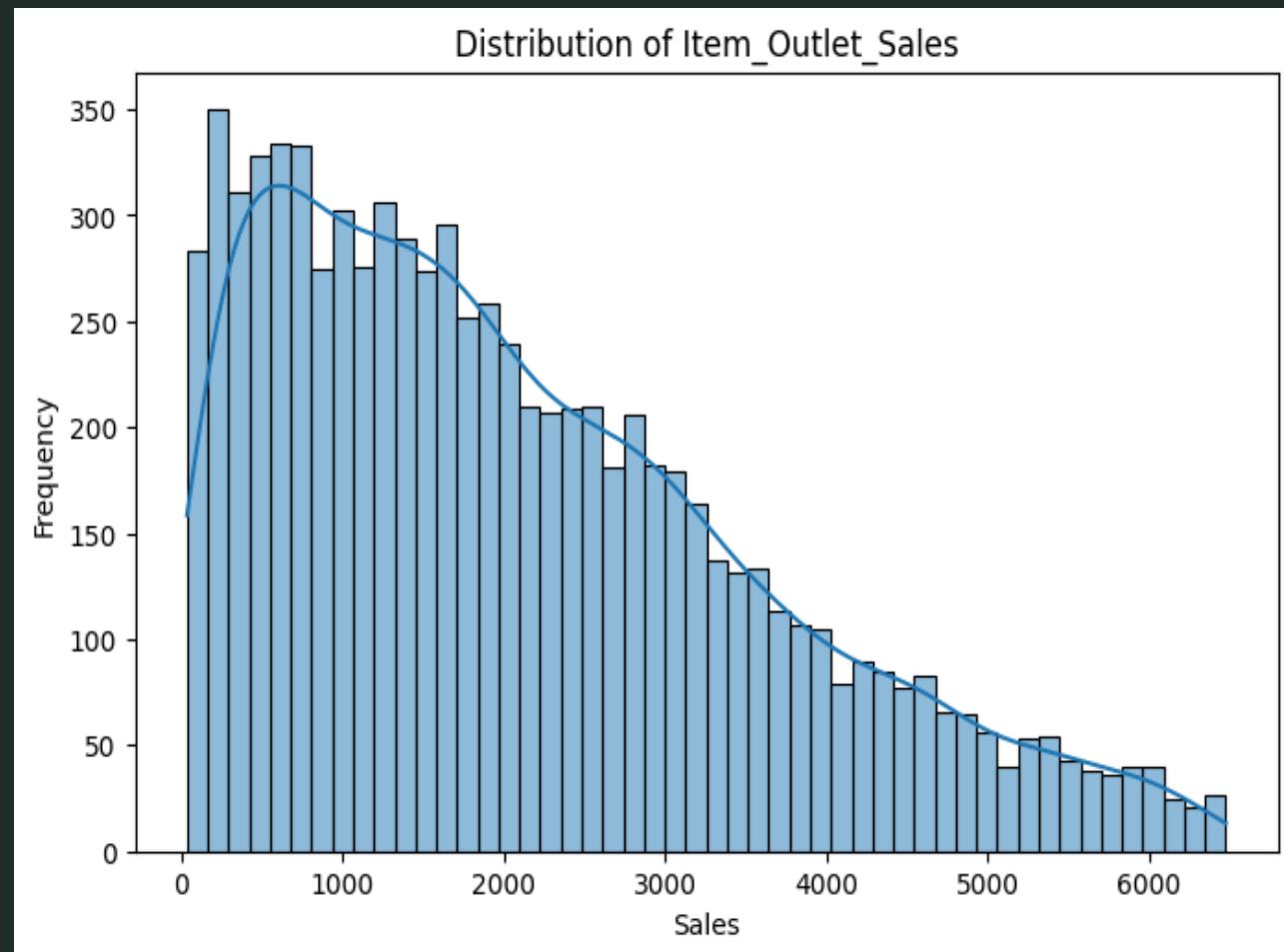
## Outliers in Item\_Outlet\_Sales

- Sales had extreme outliers → skew results, distort visualization, affect ML models.

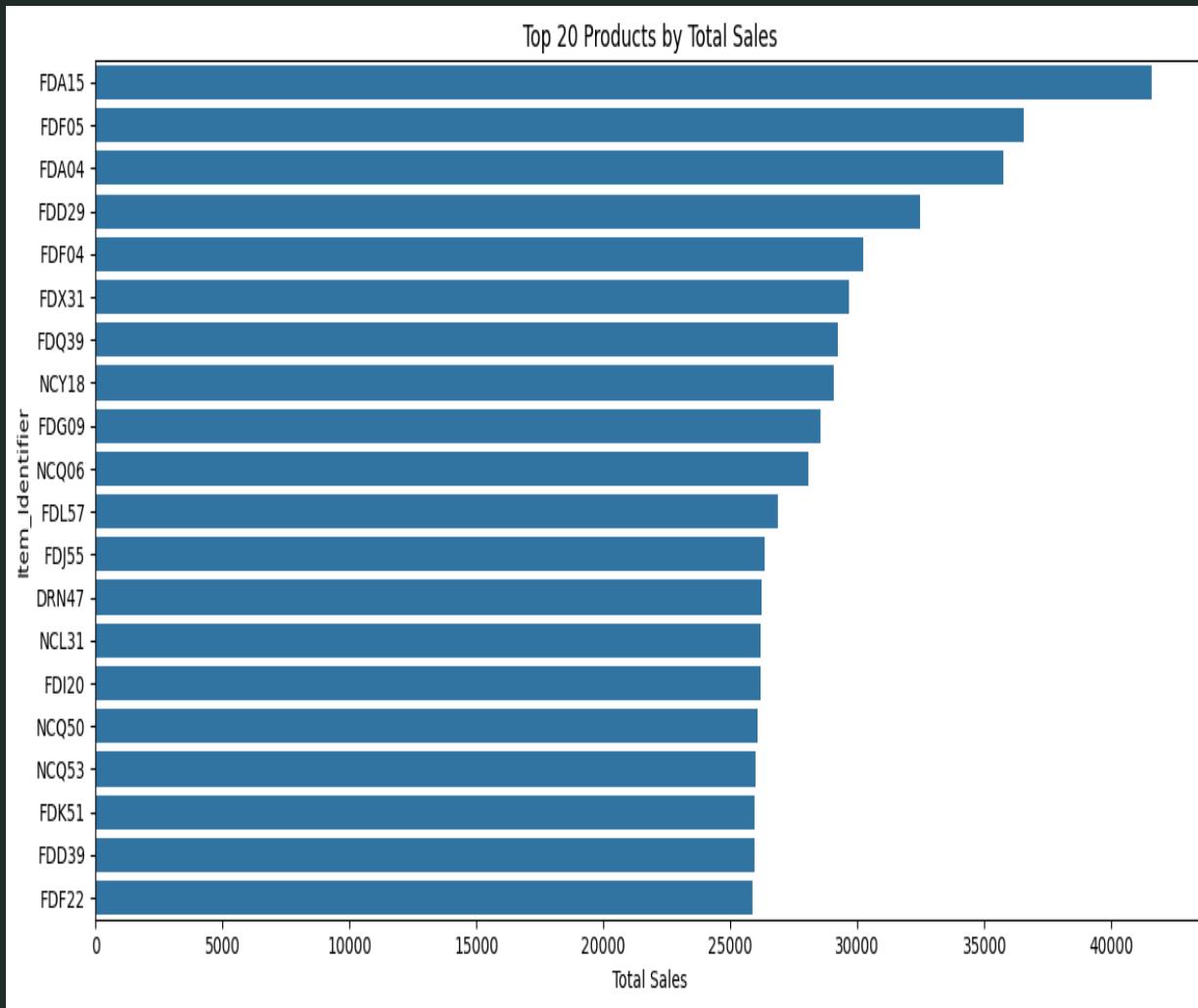
## Creation of Engineered Features

- Raw data lacked some useful business variables.

The plot is a **Histogram** showing the distribution of the `Item_Outlet_Sales` variable. The distribution is **highly right-skewed (positively skewed)**. The **majority** of sales transactions are concentrated in the lower range, peaking near the \$500 - \$1000 mark. The frequency gradually decreases as sales values increase, indicating that high sales figures (above \$4000) are relatively **rare**.



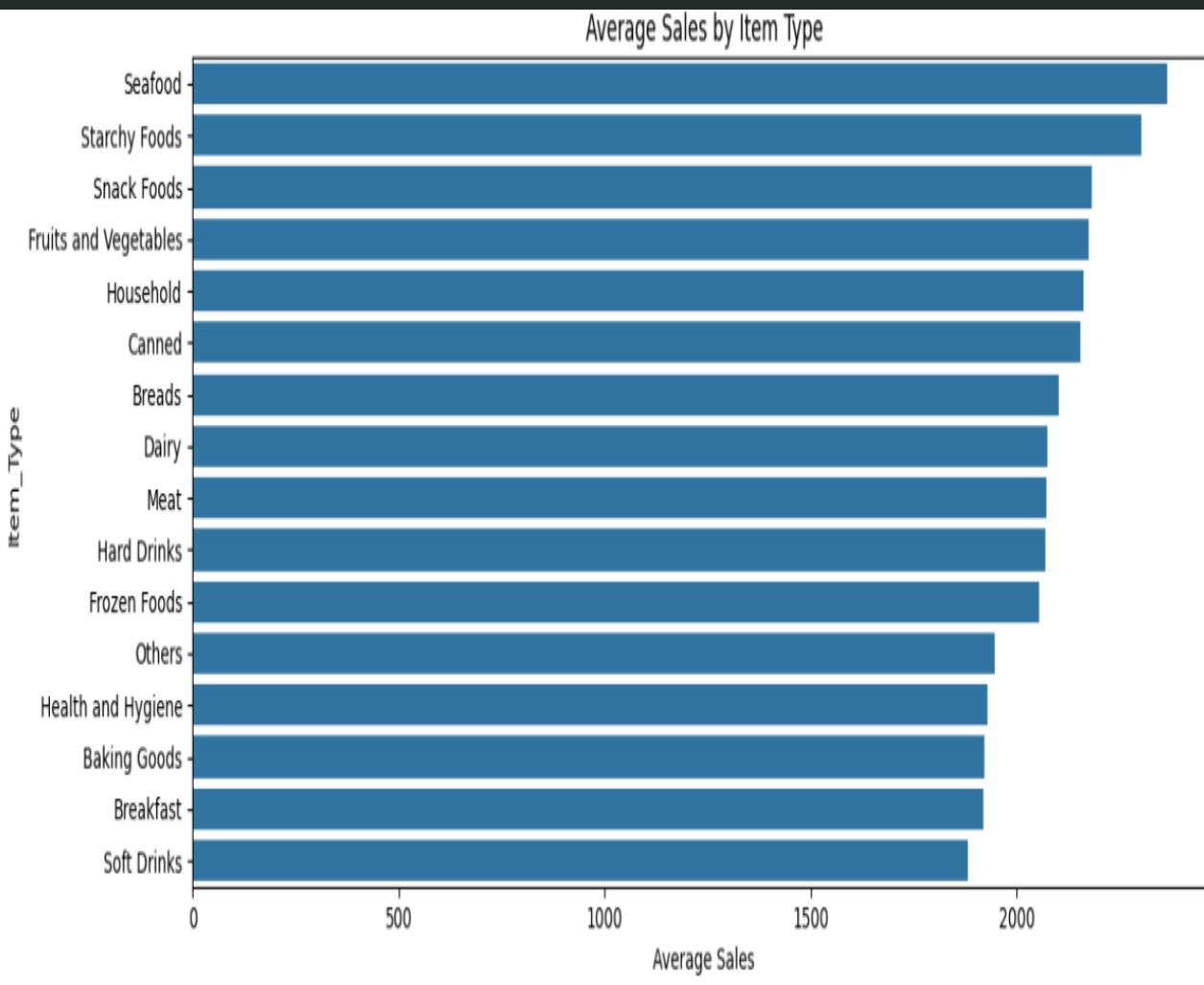
## Which products achieve the highest sales across all stores?



The chart displays the **Top 20 Products by Total Sales** across the entire BigMart chain.

- **Answer to the Question:** The product that achieves the highest total sales is **FDA15**, followed closely by **FDF05** and **FDA04**.
- **Interpretation:** The bars clearly show a **steep drop-off** in total sales after the very top products. This indicates that a small number of items (like FDA15) are responsible for a disproportionately large share of the company's total revenue, highlighting their importance as **key revenue generators**.

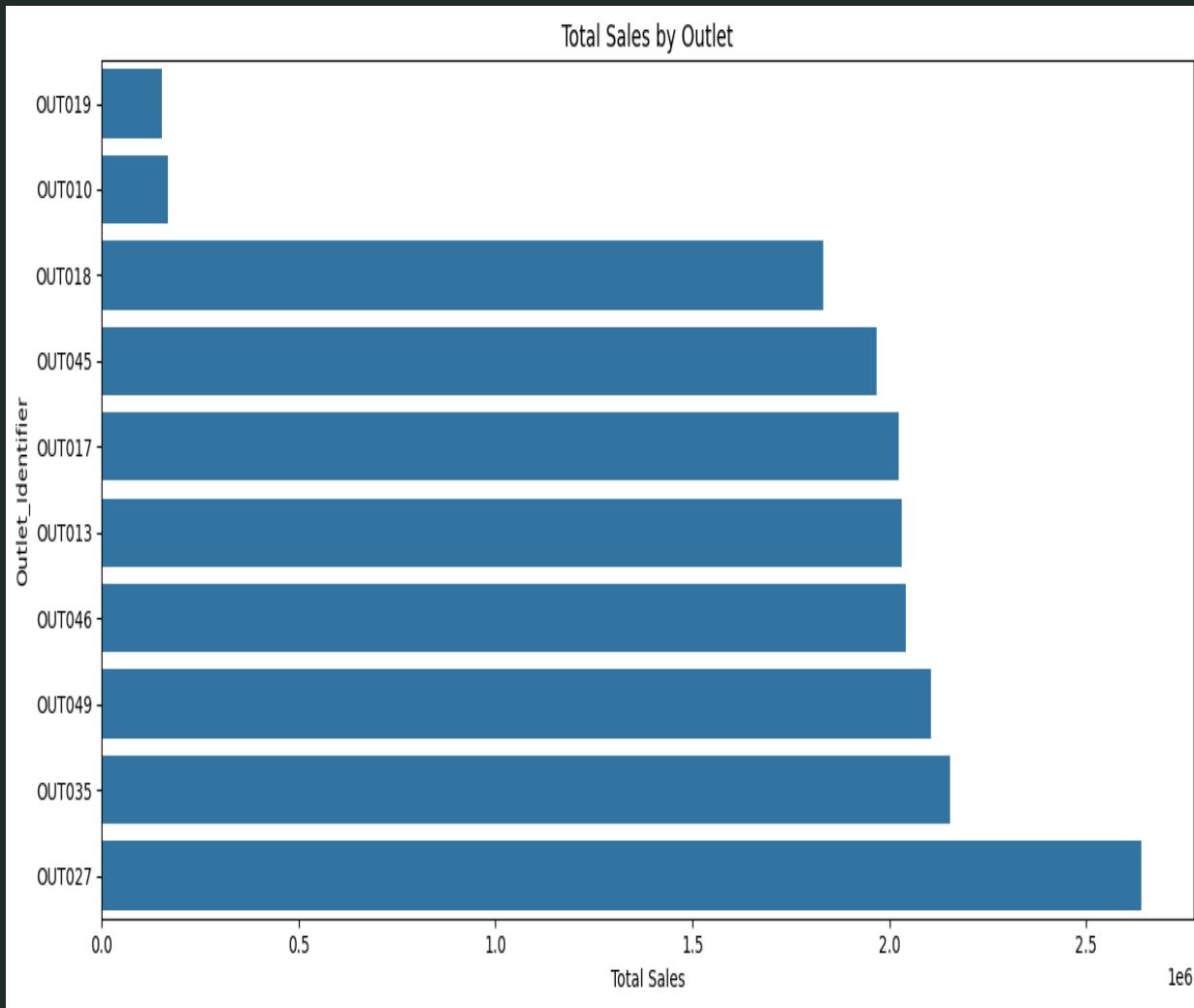
Is there a relationship between product type (Item\_Type) and sales?



The relationship between **Product Type (Item\_Type)** and **Sales** is **Confirmed**.

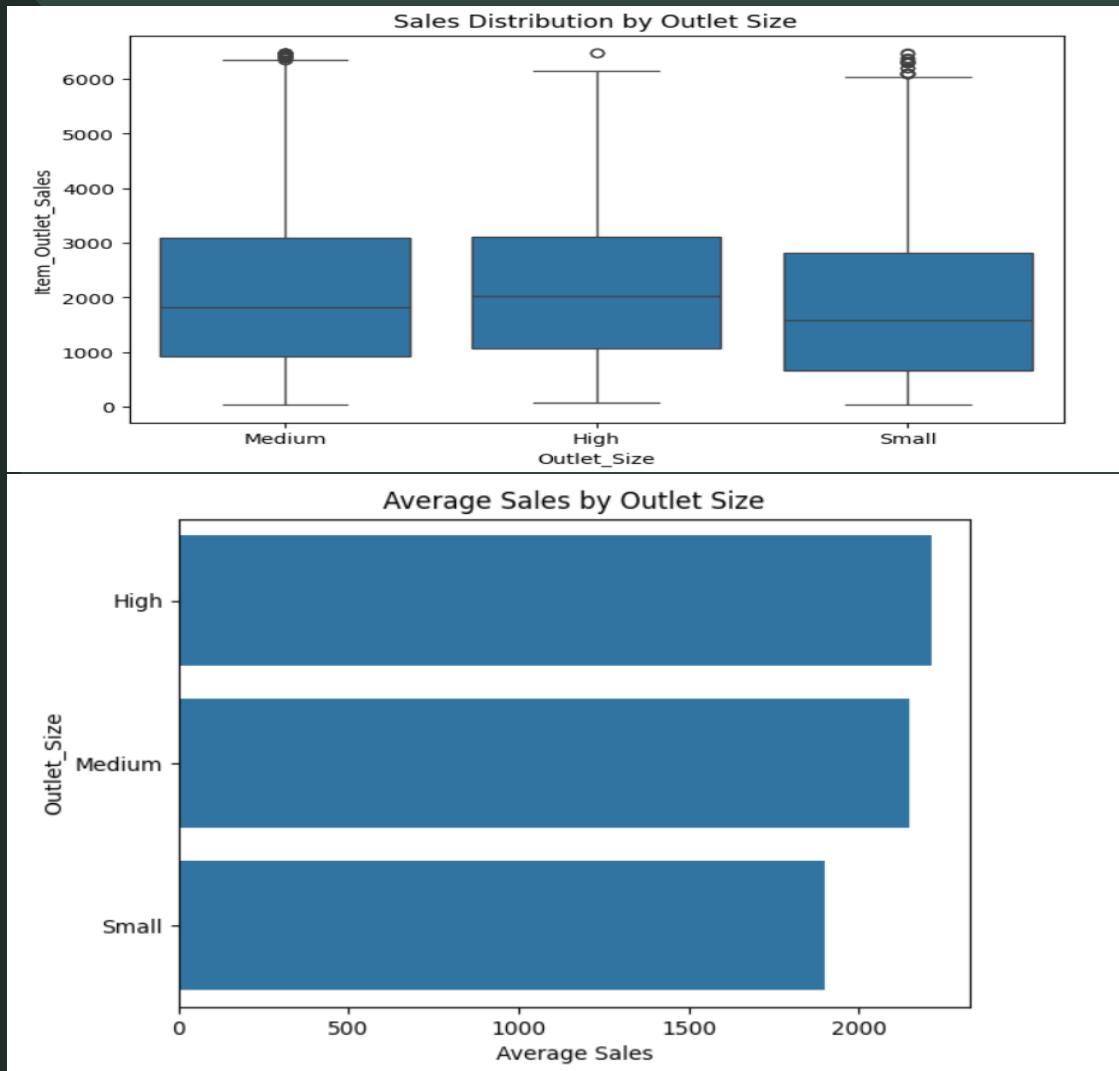
- **Box Plot:** Shows that while most product categories share a similar **median** sales range, categories like **Starchy Foods** and **Seafood** have a higher central tendency (median/mean) and potential for higher quartile performance than others. The wide array of **outliers** confirms that high sales are possible across **all** product types.
- **Bar Plot:** Quantitatively confirms that **Starchy Foods, Seafood, and Fruits and Vegetables** achieve the highest **average sales**, while **Others** and **Breakfast** show the lowest averages.

Which stores (Outlet\_Identifier) achieve the highest and lowest sales?



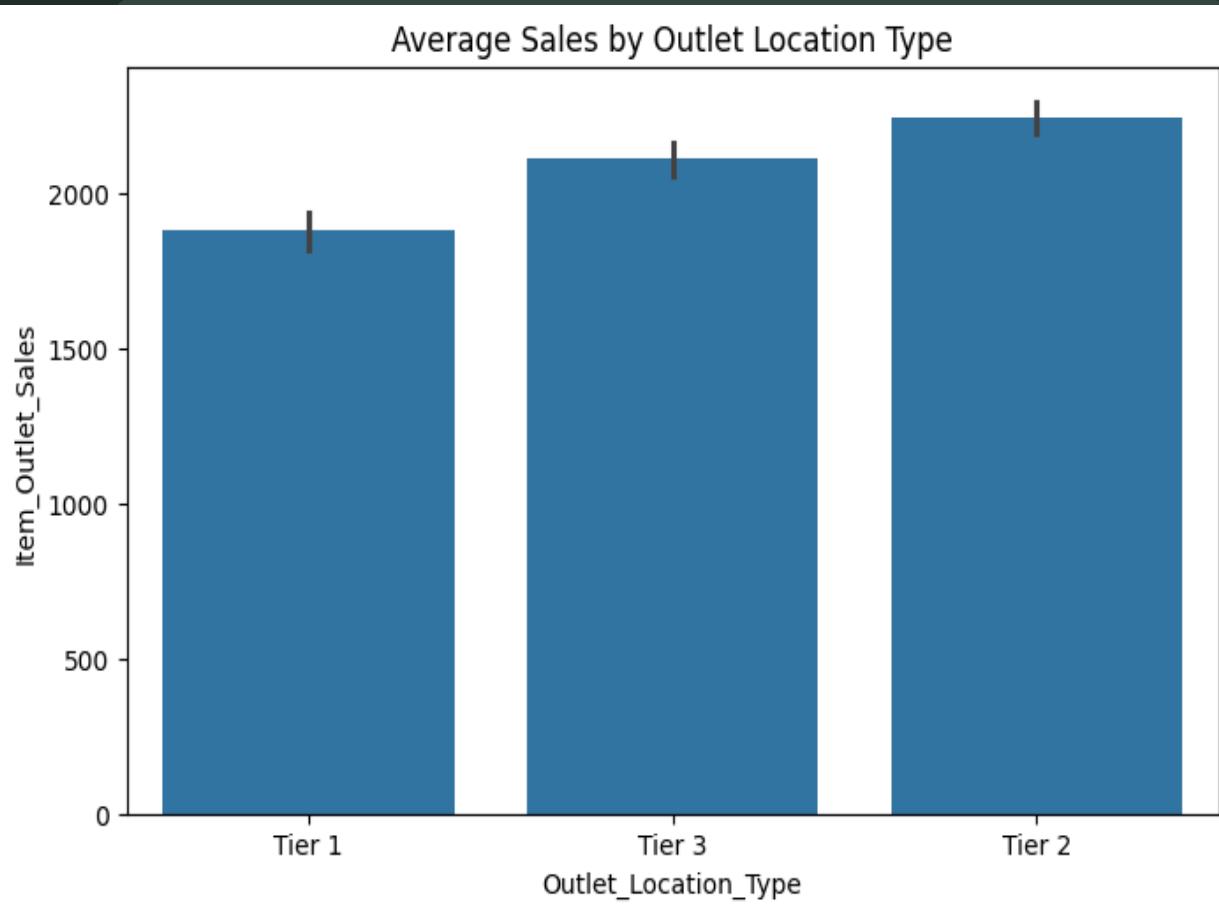
- **Highest Sales:** Outlet **OUT027** achieved the highest total sales (over 2.5 million).
- **Lowest Sales:** Outlet **OUT019** achieved the lowest total sales.
- **Interpretation:** There is a significant performance gap between the top-performing outlet (OUT027) and the bottom-performing ones (OUT019 and OUT010). This indicates that store-specific characteristics (like type, location, or size) are **major drivers** of overall sales performance.

## Does the store size (Outlet\_Size) affect sales performance?



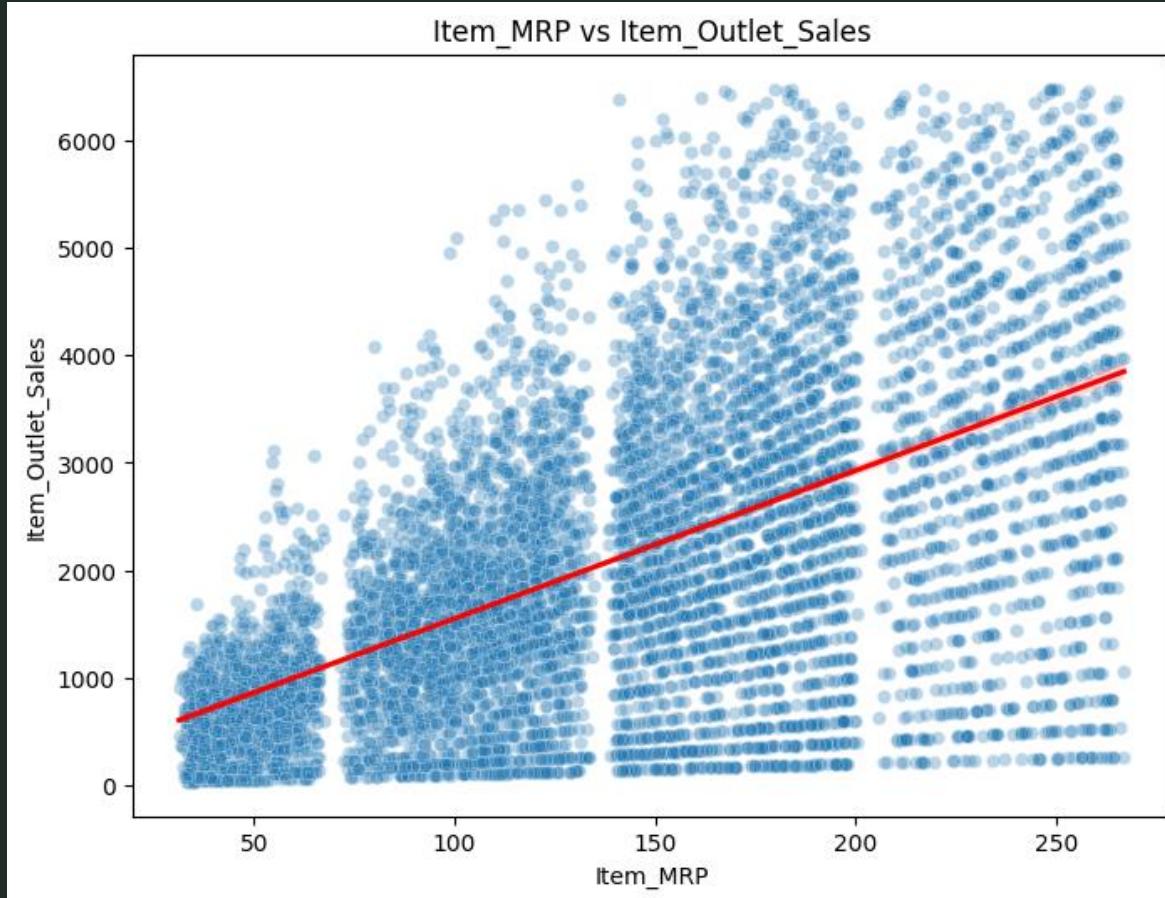
- **higher median sales** (the line inside the box, around \$2000) and a wider Interquartile Range (IQR) than Small outlets.
- **Small Outlets:** The Small outlets have the **lowest median** and the narrowest IQR, indicating that most Small stores consistently achieve lower sales volumes compared to the larger formats.
- **Average Sales Plot (Implied):** The mean sales plot (not pictured, but the calculation is implied) would confirm that the average sales are highest in Medium and High stores, and lowest in Small stores.
- **Outliers:** All sizes have a similar number of high-value outliers, meaning any store size *can* achieve very high sales, but Medium and High stores do so more frequently and consistently in their regular operations.

Does the store location type  
(Outlet\_Location\_Type) impact sales?



- **Bar Plot Interpretation:** Sales performance increases from Tier 1 to Tier 2. **Tier 2** locations have the **highest average sales**, followed by **Tier 3**, and finally **Tier 1** locations have the lowest average sales. This suggests that the highest volume of transactions or the most valuable items are being sold in Tier 2 cities/areas.

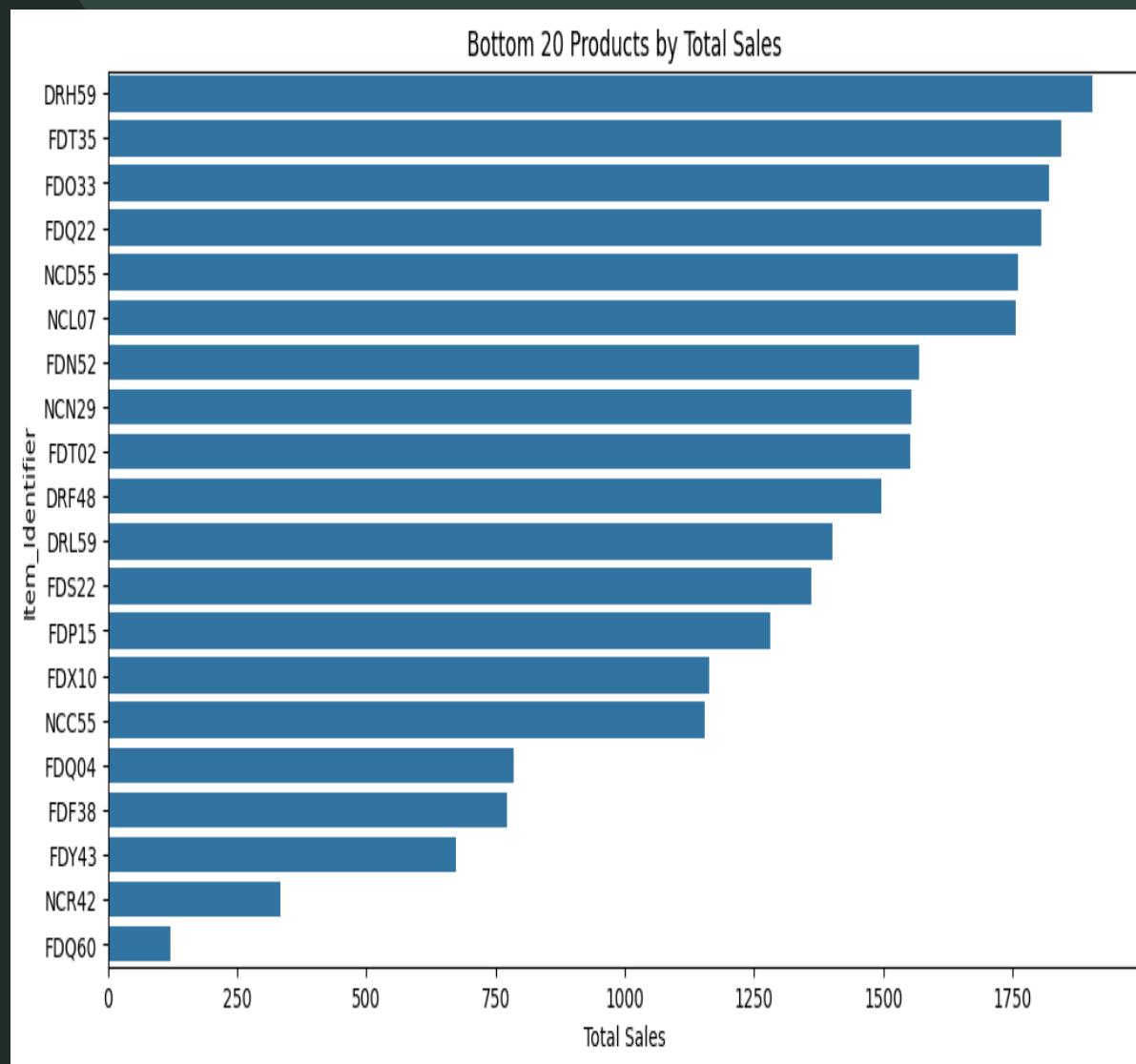
Does the product price (Item\_MRP) influence sales volume?



**Scatter Plot (Top):** The red regression line shows a clear **upward trend**. This indicates that **higher priced items (Item\_MRP) achieve higher average sales (Item\_Outlet\_Sales)**. The vertical banding shows that prices are clustered around specific price points.

**conclusion:** BigMart's most expensive products (by MRP) tend to be the highest performing

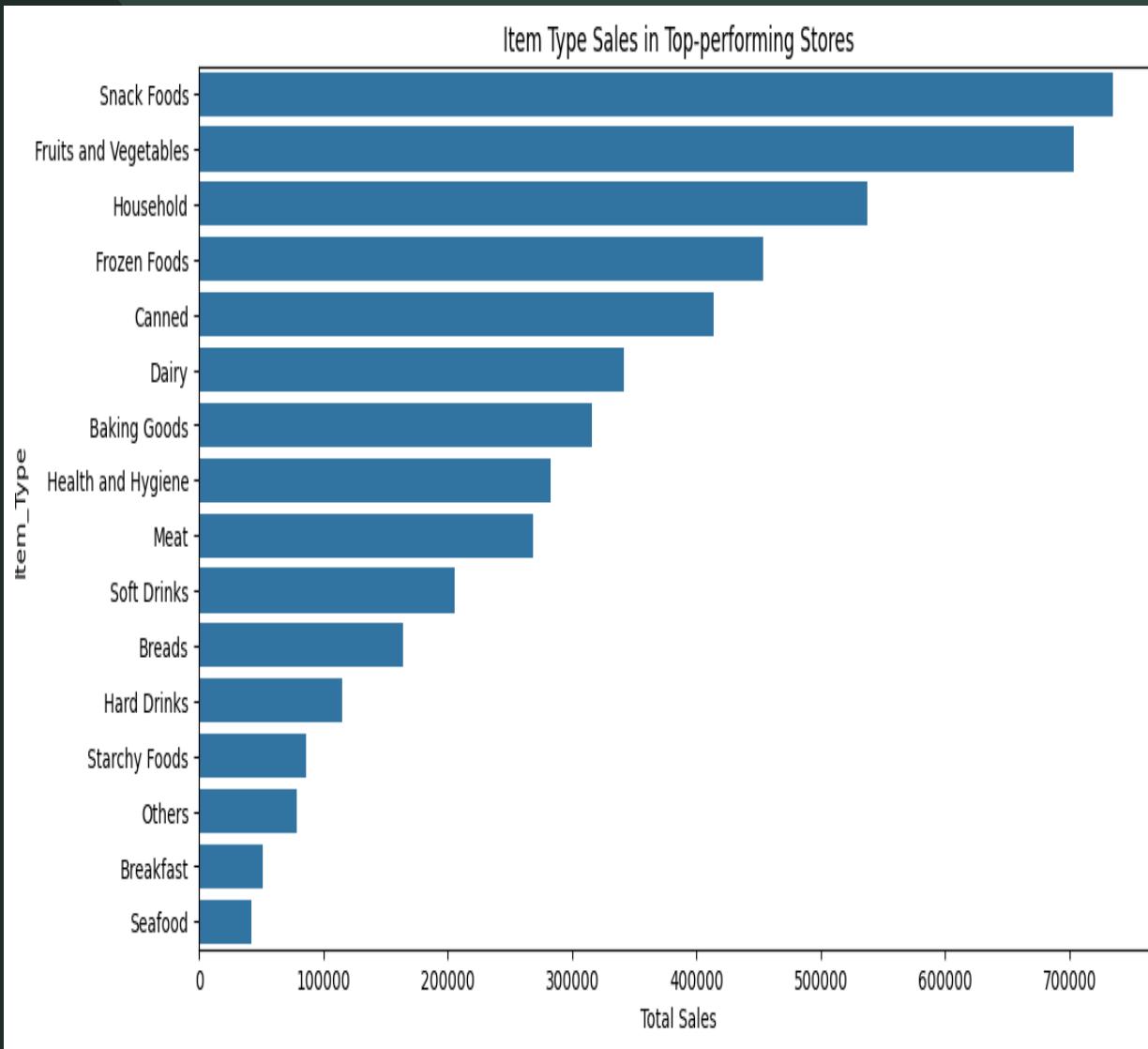
## Which products achieve the highest sales across all stores?



The chart displays the **Bottom 20 Products by Total Sales** across the entire BigMart chain.

- **Answer to the Question:** The product with the lowest total sales is **FDQ60**, followed by **NCR42** and **FDY43**. These products, along with the entire bottom 20 list, require immediate further analysis.
- **Interpretation:** The bars clearly show a group of items (especially the bottom 5 or 6) with extremely low total sales, indicating that they are either very unpopular or are poorly managed/displayed. These items are prime candidates for **diagnostic analysis** (e.g., checking visibility, pricing strategy, or whether they should be delisted from certain stores).

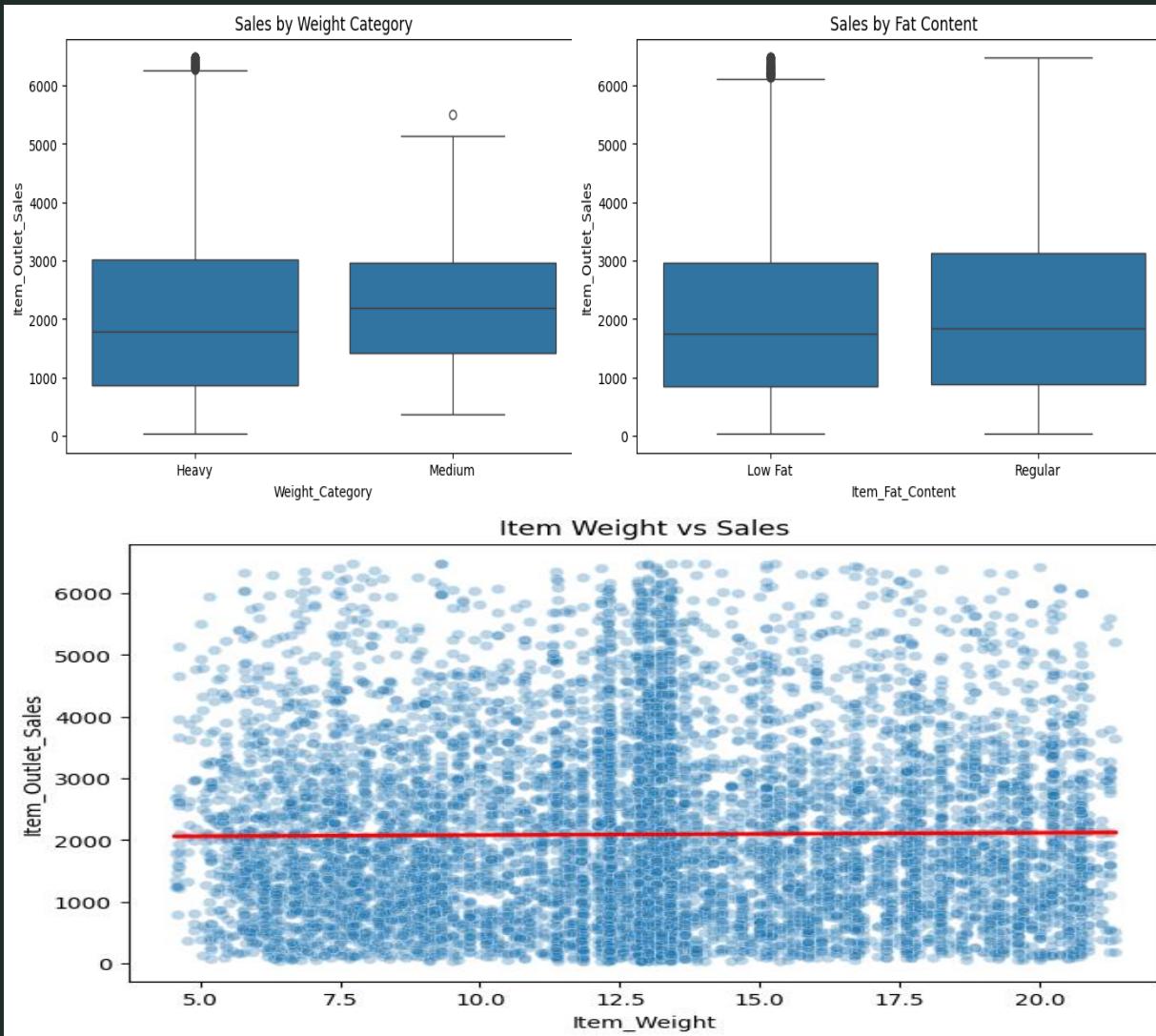
Which product categories require increased inventory in high-performing stores?



The chart displays the total sales contribution of each product category specifically within the high-performing outlets (those above the 80th percentile of total sales).

- **Answer to the Question:** The product categories that require increased inventory in high-performing stores are **Snack Foods**, **Fruits and Vegetables**, and **Household**. These are the top three categories driving revenue in the best stores.
- **Interpretation:** The visualization clearly identifies the **core customer demand** in successful store environments. **Snack Foods** and **Fruits and Vegetables** dominate the list, suggesting that these high-performing stores excel at selling **convenience and fresh/essential goods**. Management should ensure these top categories are always fully stocked in their best outlets to maximize revenue.

Is there a relationship between product attributes (e.g., weight, fat content) and sales performance?



- **No Relationship (Weight & Fat Content):** Both **Item Weight** and **Item Fat Content** show **no significant impact** on sales revenue. The box plots and regression line are essentially flat, confirming these attributes are **not predictors** of sales performance.
- **Strong Positive Relationship (Price):** **Item MRP (Price)** shows a **strong positive linear relationship** with sales, meaning that more expensive items generate higher revenue per transaction on average. **Price** is the primary product attribute driving revenue performance.

# Summary

- Which products achieve the highest sales across all stores?

Product **FDA15** achieves the highest total sales across the chain, followed by **FDF05**.

- Is there a relationship between product type (**Item\_Type**) and sales?

**Yes**, there is a relationship; categories like **Starchy Foods** and **Seafood** achieve the highest average sales.

- Does the product price (**Item\_MRP**) influence sales volume?

**Yes**, price has a **strong positive influence**; higher-priced products generate higher revenue on average.

- Which products have low sales and might require further analysis?

Product **FDQ60** records the lowest total sales, followed by **NCR42**, making them primary candidates for review or potential delisting.

- Which product categories require increased inventory in high-performing stores?

**Snack Foods** and **Fruits and Vegetables** require increased inventory in top-performing stores as they are the core revenue drivers there.

- Is there a relationship between product attributes (e.g., weight, fat content) and sales performance?

**No** relationship exists with **Weight** or **Fat Content**, but there is a **strong positive relationship** with **Price** (**Item\_MRP**).

# Summary

- Which stores (**Outlet\_Identifier**) achieve the highest and lowest sales?

Outlet **OUT027** achieves the highest total sales, while Outlet **OUT019** achieves the lowest total sales.

- Does the store size (**Outlet\_Size**) affect sales performance?

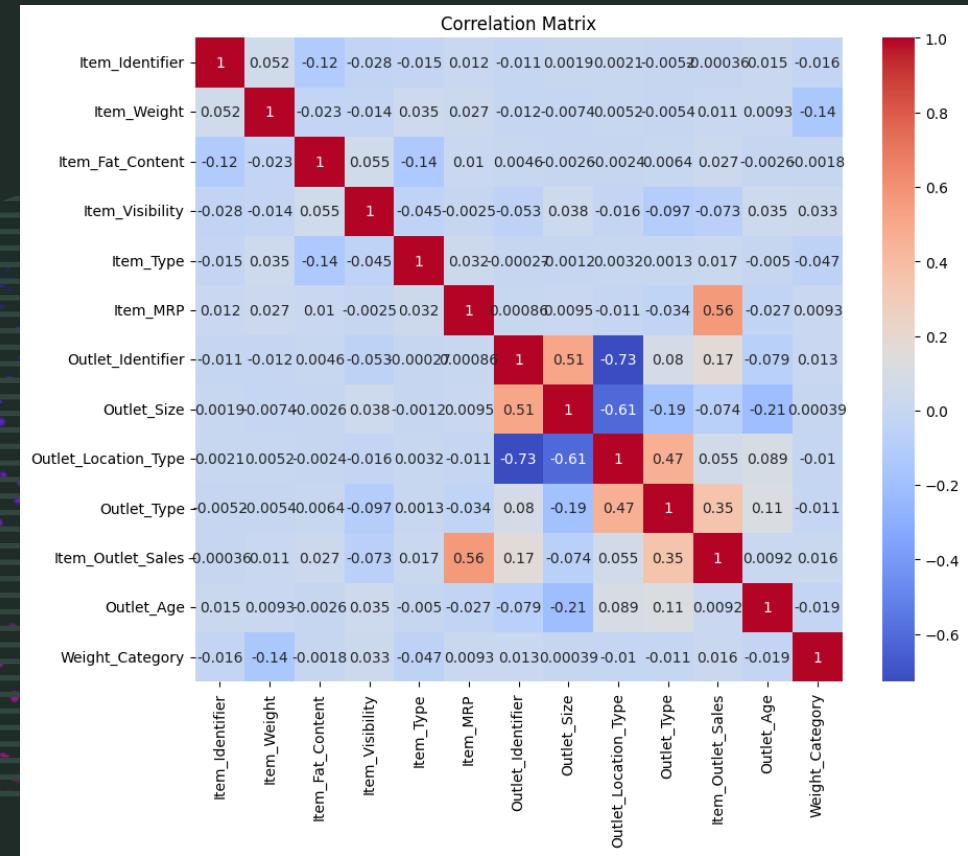
**Yes**, significantly; **Medium** and **High** size stores consistently outperform **Small** stores in average sales.

- Does the store location type (**Outlet\_Location\_Type**) impact sales?

**Yes**, it impacts sales notably, with **Tier 2** locations achieving the highest average sales.

# Correlation

- **Key Revenue Drivers:** Price (**Item\_MRP**) and **Outlet Type/Size** are the strongest sales drivers.
- **Top Performers:** OUT027 and **Tier 2 locations** generate the highest revenue.
- **Winning Products:** FDA15 leads in revenue; **Snack Foods/Vegetables** are crucial for top stores.
- **Risk Areas:** OUT019 (weakest store) and FDQ60 (weakest product) need immediate review.
- **Non-Factors:** **Item Weight, Fat Content, and Item Visibility** do not effectively influence sales revenue.





# Thank you

Name: Kyrillos Ayman

Role: Data Analyst | ML Practitioner

Email: keromihhv@gmail.com

Phone: +20 1229360705

LinkedIn: [www.linkedin.com/in/kyrilloazer-3344a7353](https://www.linkedin.com/in/kyrilloazer-3344a7353)

GitHub: [KerlessAyman](https://github.com/KerlessAyman)

Location: Alexandria, Egypt