Bigmart Data Analytics Project

By:

Felipe M Duenas
Marcus Gleinser
Salman Bin Habib

Spring 2022

Executive Summary

Bigmart is a renowned supermarket which is currently facing issues with low product visibility for some of their items, which is resulting in decreased sales. We have collected the store dataset of Bigmart which contains information on the stores and items sold at various locations of their company. The dataset is labeled with our dependent variable being sales and visibility, and the independent variables being the store and product features to explain our store sales and visibility. To begin, first, we conduction exploratory data analysis (EDA), and created histogram and bar charts to understand and explain the relationships between the variables, and in the process look for outliers and missing values which may impede our analysis. Through EDA, we looked at the Item Outlet Sales counts across stores, MRP across products, and how the data is spread across different Supermarket types, outlets, and each product. Furthermore, we looked for distribution patterns among the product types, sales, and visibility. One interesting observation found through EDA is the relationship of Item Visibility and Item Outlet Sales, which is positively correlated.

Upon observation, we have developed the following research questions- "What factors affect outlet sales?", "How does item visibility affect sales", and "What factors affect item visibility?". Our methodology involves cluster analysis to group our data and determine if there is a statistically significant relationship between Item Outlet Sales and Item Visibility. Tableau is used to create clusters using the dependent variables Item Visibility, Item Outlet Sales, and the independent variables, Outlet Type, Outlet Location, Item Type, and Outlet Size. Our analysis led to a statistically significant relationship in Cluster 1 with Outlet Sales and Item Visibility, with a significance level at the 0.01 level. That means, if we focus on increasing the visibility of the Item Types Meat, Breads, Hard Drinks, Starchy Foods, Breakfast, and Seafoods, we can expect an increase in sales. Further analysis is done comparing Cluster 1 and Cluster 2 to look for patterns in Item Fat Content, Outlet Locations, and Item Weight to make recommendations.

Finally, we recommend Bigmart to focus on the Item Types in Cluster 1 to increase sales. Bigmart should prioritize products with Outlet Locations in Tier 2 and Tier 3 cities with Supermarket Type 1s, as well as items with large and small Outlet Sizes. Furthermore, Bigmart will discover that the outlets with the IDs OUT013, OUT035, and OUT046 are the best places to focus on low visibility products in order to improve sales. Bigmart's new approach for raising sales based on our advice should not include item fat content and weight. Furthermore, we recommend Bigmart to create designated areas for displaying and showcasing meat, bread, and seafood items to draw more. Shoppers usually begin in the produce department before moving on to the various aisles. We believe that by placing these lower visibility products next to the high visibility ones, people will gravitate toward these new displays. Furthermore, endcap displays can be utilized to highlight and offer coupons for hard drinks, starchy foods, and breakfast items.

Introduction

Problem Statement
Many of Bigmart's product categories with low visibility have low sales.

Dataset Details
There are 4,650 entries included in the Bigmart dataset. It has 12 different columns which include columns containing information on the items sold at the Bigmart locations and information on the different stores. Six columns have information on the products, while the last six have information on the stores.

Item Identifier is an alpha-numeric code associated with each of the different items. Item Weight gives the weight for each product. Item Fat Content divides the products into two categories, Low Fat and Regular. Item Visibility gives a numeric score based on how visible the product is to the customer. Item Type categorizes the products into one of sixteen different item types, which include categories such as Canned, Dairy, or Household. The final item column is Item MRP, or Item Maximum Retail Price, which is the highest price the product is sold at.

Outlet Identifier is another alpha-numeric code associated with each store. There are only five different IDs which are OUT013, OUT018, OUT035, OUT046, and OUT049. Outlet Establishment Year gives the year each store was established. OUT013 was the first to be established in 1987, with OUT046 coming ten years later in 1997. OUT049 was then created in 1999, followed by OUT035 in 2004, and OUT018 in 2009. The next store column is Outlet Size, which categorizes the stores into Small, Medium, or High sized stores. Outlet Location Type also has three categories, Tier 1, 2, or 3, which says where each store is located. Tier 1 stores are located in more urban areas/big cities. Tier 2 stores are in medium-sized towns, while Tier 3 stores are found in more rural areas/small towns. There is also Location Type which says what type of supermarket each store is. Supermarket Type 1 is your typical large supermarket, while Supermarket Type 2 is a smaller grocery store. The last column is Item Outlet Sales which gives a dollar amount for how much each store made.

Exploratory Data Analysis

Item Outlet Sales/Count

Item Outlet Sales (bin)
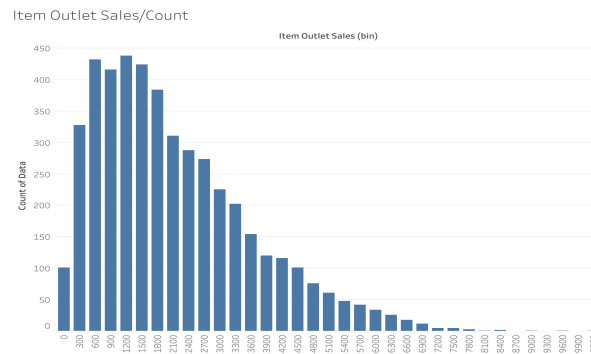
This histogram shows the distribution of Item Outlet Sales counts across all the Bigmart stores' data. Item Outlet Sales with totals between $1200 and $1500 are the highest, and have a data count of 438. We can see that many of the total Item Outlet Sales are within the range of $600 to $900 or $1500 to $1800. After the $1800 to $2100 range, we see a drop, indicating that there is a high count of data with Item Outlet Sales that are significantly lower compared to the rest of the data.

Item MRP/Count

Item MRP (bin)
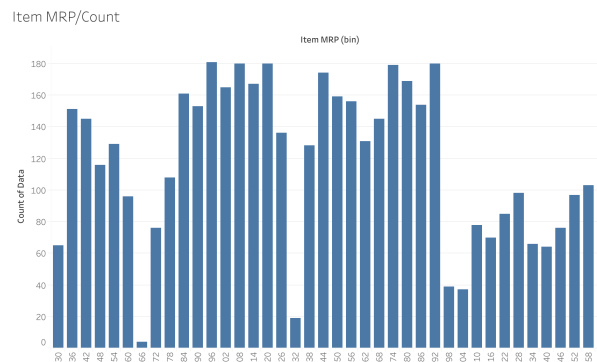
This histogram shows the distribution of MRP across products. The $96 bin is the highest with 181 products that fall in that bin. The $108, $120, and the $192 bins are the next highest with 180 products included within them. It seems MRP count fluctuates up and down as MRP increases. It does look like the frequency of anything higher than $192 decreases, never getting higher than 103.

Outlet Type/Count
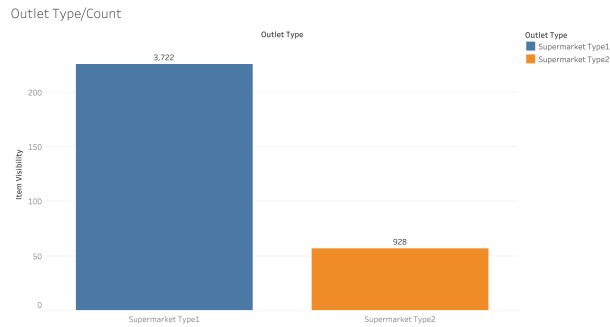
Here is a bar chart showing how much of the data falls into each of the two types of supermarkets. 3,722 data counts belong to Type 1 Supermarkets, which is much greater than the 928 data counts that are Type 2 Supermarkets.


Outlet Size/Count
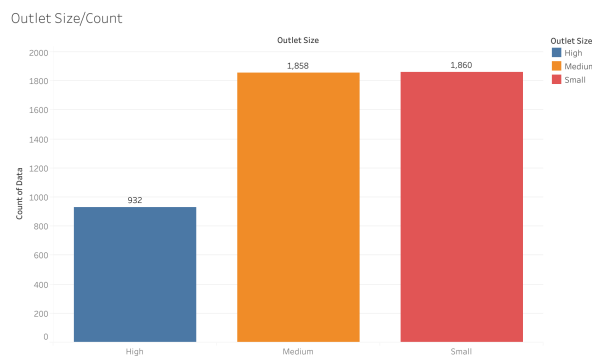
This bar chart represents how much data falls into each size of outlet stores. We can see that most of that data comes from Small size outlets, with 1,860 of the data being categorized as such. 1,858 data counts are considered to be Medium size, while High sized data counts are the least common with only 932.


Outlet Location Type/Count
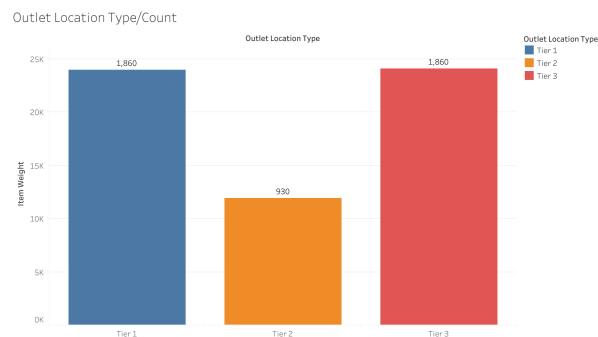
This bar chart shows how much of the data is located in Tier 1, 2, or 3. Tier 1 and 3 have an equal amount of counts, at 1,860. This means there are an equal number of data counts in both big cities and small towns. Tier 2 has the least amount of data with 932 counts, meaning that less of the data is from medium-sized towns' sales data.
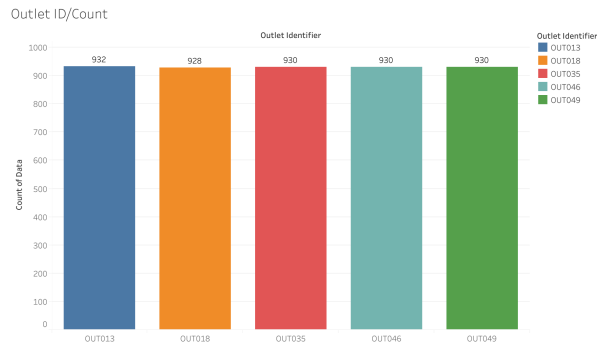
Outlet ID/Count

This bar chart shows the data counts associated with each outlet identifier. OUT035, OUT046, and OUT049 each have 930 data counts with those IDs. OUT013 has the data with 932 counts, while OUT018 has the least with 928.


Item Fat Content/Count
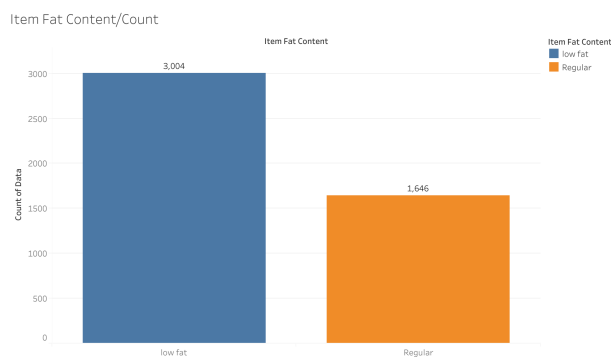
This bar chart shows how many of each product is categorized as Low Fat or Regular. Low Fat products are much more common with 3,004 counts of the data set being categorized as such. Only 1,646 of the data entries are Regular.

Item Type/Count
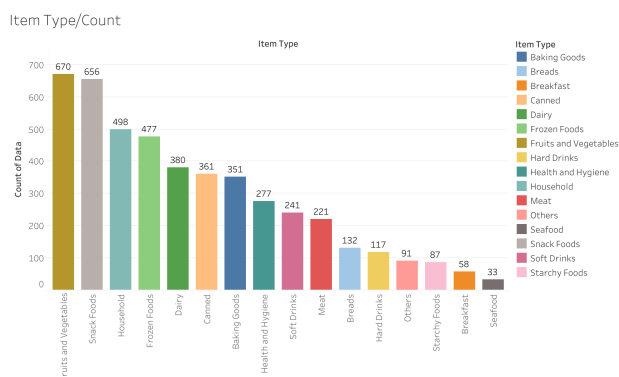
This bar chart represents the distribution of product types in the data. Fruits and Vegetables are the most common product type sold with 670 entries in our dataset, followed by Snack Food with 656 entries. From here we see a sharp decrease to 498 Household product entries until ending with only 33 entries for Seafood.
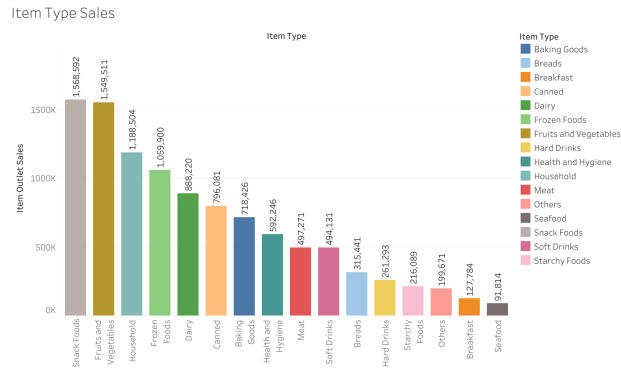
Item Type Sales

This graph gives shows the total Item Outlet Sales for each Item Type. We can see that Snack Food has the highest sales total with $1,568,592 followed by Fruits and Vegetables whose sales total $1,549,511. These two item types are making nearly $400,000 more than the next highest item type. Again the Seafood Item Type is the lowest with only $91,814 in sales.
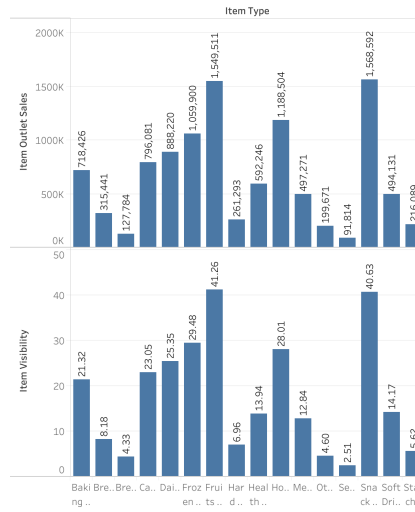

Item Type Visibility
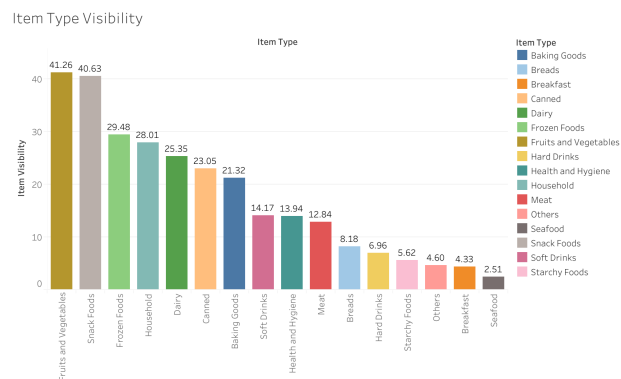
This chart gives item visibility for each item type. Fruits and Vegetables have the highest total item visibility with 41.26, followed by 40.63 associated with Snack Foods. Again the Fruits and Vegetables and Snack Foods categories are much higher than the other item types. Seafood has the lowest visibility with a score of 2.51.

Research Questions

- What factors affect outlet sales?
- How does item visibility affect sales?
- What factors affect item visibility?

Methodology

Sales & Visibility Relationship

When displaying the Item Outlet Sales of Bigmart's Item Type's, and then comparing each Item Types' Visibility, we found that the relationship of Item Visibility and Item Outlet Sales seemed to have a mirrored relationship. Item Types with low Item Visibility had low Item Outlet Sales and vice versa, Item Types with high Item Visibility had high Item Outlet Sales.


Clusters Visibility/Sales
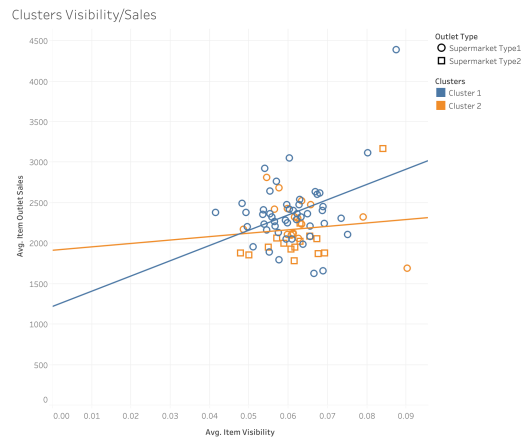
We decided to utilize cluster analysis to group our data and determine if there was a relationship between Item Outlet Sales and Item Visibility. Clustering our data using Item Visibility and Item Outlet Sales, and the categorical variables of Outlet Type, Outlet Location, Item Type and Outlet Size, we found a positive relationship in Cluster 1 of two clusters. This Cluster had a significance at the 0.01 significance level. This signified that there was a relationship in the Cluster 1 data, between Item Outlet Sales and Item Visibility. Cluster 2 had a P-value of 0.52 which means that the data in Cluster 2 could not prove a relationship. After this determination, we then focused on the data in Cluster 1 and used the variables of Item Type, Outlet Location Type, Outlet Type, Outlet Size, Outlet Identifier, Item Weight and Item Fat Content to examine the data in Cluster 1 compared to Cluster 2, to make recommendations for Bigmart.

Findings & Interpretations

Using Cluster 1's significant relationship between Item Outlet Sales and Item Visibility, we then visualized the relationship of Item Visibility and Item Type with Item Outlet Sales in Cluster 1'2 data and received the chart below.
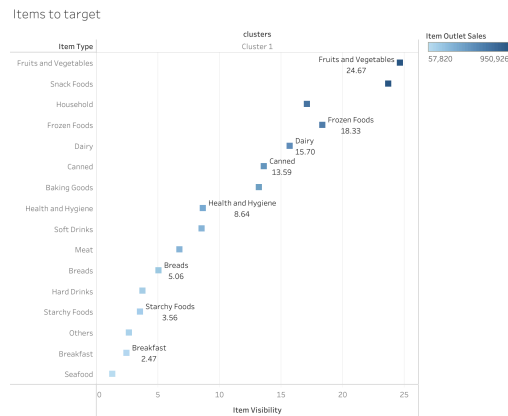
Item Types with higher Item Visibility had the highest Item Outlet Sales. We then decided to focus on all items with lower Item Visibility as they were the items with lower Item Outlet Sales. These Item Types included Meat, Breads, Hard Drinks, Starchy Foods, Others, Breakfast, and Seafood. Because of the significant relationship between Item Visibility and Item Outlet Sales, we could infer that if we focused on increasing the visibility of these Item Types in Cluster 1, we could expect an increase in sales. The next questions we needed to answer consisted of what Cluster 1 Data consisted of regarding Outlet Location, Outlet Type, Outlet Size, which Outlet Identifiers, Item Weight and Item Fat Content.

When comparing Cluster 1 and Cluster 2, we found that Cluster 1 consisted of Outlet Locations in Tier 2 and Tier 3 cities, which leveraged Supermarket Type 1s in those city tiers. The Outlet Size of Bigmarts in Cluster 1 consisted of only High and Small. When reviewing Cluster 1 Outlet Identifiers, we learned that the three identifiers of OUT013, OUT035, and OUT046 out of the possible five, made up our target cluster. Lastly, when comparing average Item Weight and Item Fat Content in both clusters, we found that these variables didn't have different results depending on the cluster they were in. The average item weights were similar for both clusters at 12.90. Item Fat Content also wasn't affected as the data showed both Regular and Low fat variables behaved the same way in Cluster 1 when comparing to Cluster 2. We could not form a recommendation from Item Fat Content.

Using this data, our recommendation is that Bigmart focus on the Item Types in Cluster 1 to increase sales for Bigmart. Bigmart should focus on the items with Outlet Locations in Tier 2 and Tier 3 cities with Supermarket Type 1s, that are of high and small Outlet Size. Furthermore, Bigmart will find that the outlets with identifiers of OUT013, OUT035, and OUT046 are the locations to focus on the low visibility items to increase sales. Item Fat Content and weight shouldn't be part of Bigmart's new strategy for increasing sales using our recommendation.

Recommendations/Strategies

Business Recommendations

We have seen that in Cluster 1, item visibility and sales are correlated with each other. There are some item types in Cluster 1 that have low visibility, thus low sales. These include the meat, breads, hard drinks, starchy foods, breakfast, and seafood item types. Our recommendations will deal with ways to increase the visibility of these item types, which will lead to an increase in sales.

For the meat, bread, and seafood item types, we suggest having designated areas with large signs to draw attention to the products. Having a meat counter to showcase the different cuts of meat could be a way for customers to more easily see the product. A similar situation could be done with seafood, where you have different fish and shellfish on display in freezers for customers to see and judge the quality. Including a bakery in the store could help draw more attention to the bread products. Both the smell of baked goods, along with having them on display can draw people in to see what types of bread products the store is selling. The layout of these additions is also important. We suggest having these meat counters, seafood displays, and bakeries near the fruits and vegetables. Fruits and vegetables have the highest visibility and sales out of the item types. Shoppers typically start in the produce section before going into the different aisles. By having these lower visibility items closer to these highly visible items, we think customers will gravitate towards these new displays.

For hard drinks, starchy foods, and breakfast item types we suggest the stores start utilizing their endcaps more. Endcap displays are a great way for products to be highlighted as they draw more attention. If the store starts to use colorful signage in tandem with the endcaps, it can better draw attention to these products. Another option would be running deals or offering coupons for these item types. Having promotional signage or coupons in a very visible place could lead customers to actively look for these products.