Retail Sales Data Analysis

Abstract—**In this paper we have analyzed the data of Costello's ACE hardware retail store chain with the primary goal of providing the store with insights that will help them grow their business. It will provide some valuable and analytical insights like which product to stock and which product to place near other product to increase the overall sales. This analysis will help them improve their current business strategy. We have used Apriori and XGBoost algorithms to analyze the customers basket and predict the stock of a particular product in future. We also suggest certain stores to implement Loss Leader pricing method to boost their profits.**

I. Introduction

Retail analytics focuses on providing insights related to sales, inventory, customers, and other important aspects crucial for a retailers' decision-making process. Nowadays, the traditional approach of analysing the data is no more effective as the volume of data has increased from thousands to billions and trillions of records. In this competitive edge of marketing, retailers need 360-degree view of their business. Sometimes it becomes very difficult for retailers to comprehend the current market condition as their stores are located in various geographical locations. To compete with other market leaders and meet business goals, retailers provide varied offers and promotions to customers. Data science helps retailers to forecast their future sales by analyzing prior sales data.

In this project, we are analyzing retail sales data of Costello's Ace hardware which is a hardware store chain from Long Island. As it has a widely spread network of stores, it would be useful to suggest some insights which portray the customer needs and current market potential. In this project, we made use of sales dataset of 4 years (2015-2018) provided by the store chain to understand the factors affecting the sales. For example, time series analysis on dataset helps us identify seasonality trends across a given time frame. Using this, we can also observe the boost in sales on events like Black Fridays, Christmas and Easter holiday. This will help stores to manage the resources like inventory of products and decide upon offers on different products to increase the sales on such events.

In this project we have used Market Basket Analysis (MBA), also known as associative rule learning or affinity analysis. It is a modeling technique based upon the theory that if you buy a certain set of items, you are likely to buy another set of items. This will allow retailers to identify relationships between the items that people buy. We have built a XGBoost model using time series forecasting that predicts the future product demand that will help in inventory management. We have also performed Loss Leader analysis on various stores. It is an aggressive pricing strategy in which a store sells selected goods below actual cost in order to attract customers who may make up for the losses on highlighted products with additional purchases of profitable goods. Based on our analysis, we have suggested certain stores to implement this pricing method that may help them increase their profit.

II. Dataset

We have been provided retail sales data by Costello's Ace hardware. The dataset consists of 39 features. Following are some of the features which we have used for various tasks in this project:

- **Date**: Date of transaction (Format: mm/dd/yyyy)
- **Receipt Number**: Invoice number of the transaction.
- **Zip Code**: Area code.
- **Store #**: Unique identifier given to each store by Costello.
- **Item Number**: Unique ID given to each item in the hardware chain.
- **Item Description**: Unique item description.
- **Loyalty ID**: Unique membership ID given to regular customers.
- **Net sales units**: number of units/quantity purchased.
- **Actual price**: price for which shopkeeper has sold any item.
- **Retail Price**: price of an item in market.
- **Net Sales**: Net Sales Units* Actual Price.
- **Cost**: Actual cost price for the shopkeeper. It is the price for which shopkeeper has purchased items to sell in the shop.
- **Gross Margin**: Net sale - cost.
- **Gross Margin %**: Margin / Net sales.
- **$ Off Retail**: Price off shopkeeper is offering on retail price of product.
- **Actual-Retail**: Is same as $ Off Retail.

Interesting facts about data:

- Costello's Ace Hardware, even though being a hardware retail store, the store not only consists of the hardware items but also cookwares and food items like 'BIRD BLEND SUET' and 'Diamond Naturals Lamb and Rice Dog Food'. The same was observed while browsing through the dataset.
- For the rows where Loyalty ID is not null, the Zip Code represents the area code of the customer who is making the purchase. While in case if Loyalty ID is null, Zip Code will represent the area code of the store.
- Store # and Store Name both uniquely represent the store. Similarly, Item Number and Item Description both uniquely define the items in the stores. Hence, for the computational purposes, we will stick to Store # and Item

Number while for display purposes, we will use Store Name and Item Description wherever possible.

- An item may have different pricing across different stores. For some items across the stores, it may have different gross margin and different gross percentage.
- $ Off-retail and (Actual - Retail) shows same results. We can say that they are redundant.
- Records for item CMN donations shows the amount donated by the customer. So, there is no retail price for such records. Every time according to column values it calculates all margin and margin %. Another thing to be noted, pricing source is always manual overridden for CMN donation records.

Exploratory Analysis:

- We have plot the distribution of sales transactions based on hourly status:

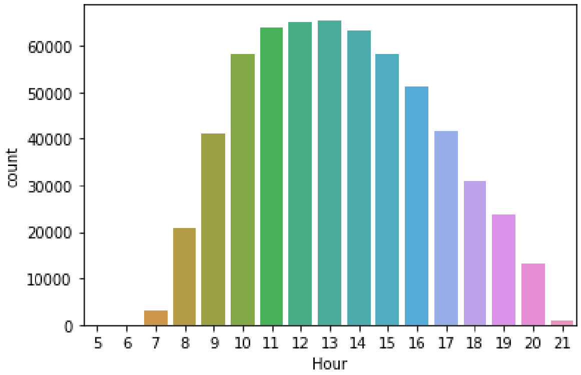


Fig. 1: Hourly Shopping Analysis

From the plot, we can see that the people tend to buy more during the time span of 9 am to 5 pm . This type of data can also help to in promotion of specific products based on the hourly trends of customers.

- Next, we plot the distribution of number of items purchased in a single basket. The frequency (i.e. the count) decreases as the number of items in a basket increase. Hence, we plot a separate graph for illustrating the higher number of purchased items in a basket.
From the figure 2, we can deduce that customers tend to buy fewer items more often while rarely they purchase higher number of items (i.e. they probably stock the items).

III. Methods

[A] Market Basket Analysis (MBA)

Few terms which are used while evaluating the association rules produced by the MBA are:

- **Antecedent:** It is the first half of a hypothetical proposition whenever the if clause precedes the then clause.
- **Consequent:** It is the second half of a hypothetical proposition whenever the if clause precedes the then clause.
- **Support (X):** Support of item is the number of times an item occurs in transactions in a database.
- **Confidence:** Confidence is a term associated with association rule. It is defined mathematically as: $Confidence = Support(XY)/Support(X)$
- **Lift:** It is the ratio of the observed support to that expected if the two rules were independent. The basic rule of thumb is that a lift value close to 1 means the rules were completely independent. Lift values $>$ 1 are generally more “interesting” and could be indicative of a useful rule pattern.
- **Leverage:** Leverage computes the difference between the observed frequency of A and C appearing together and the frequency that would be expected if A and C were independent. A leverage value of 0 indicates independence.
$leverage(A \rightarrow C) = support(A \rightarrow C) - support(A) * support(C)$
- **Conviction:** A high conviction value means that the consequent is highly depending on the antecedent. For instance, in the case of a perfect confidence score, the denominator becomes 0 (due to 1 - 1) for which the conviction score is defined as ’inf’. Similar to lift, if items are independent, the conviction is 1.
$conviction(A \rightarrow C) = (1 - support(C))/(1 - confidence(A \rightarrow C))$

1) **Data Preprocessing:**
We encountered certain irregularities in the dataset and resolved them as below:
   - Each month’s data is preceded by a header row that consists of column headers. We removed these redundant header row records.
   - Removing erroneous value ’\x1a’ in the ’Date’ feature. We then converted the column to DateTime format for further use and processing.
   - Label encoding for categorical features like ’Store #’ and ’Item Number’.
   - Treating null values from all the features.
   - Net Sales Units and Gross Margin columns consisted of values in different formats like ‘1000.00’, ‘1,000’ and ‘1000’. These values are converted to a numeric format.

2) **Model:**
Features used: ‘Receipt Number’, ‘Item Description’.

   Description: Each receipt number denotes one basket against which items are purchased. We have used Item Description to identify each item in a basket.

   First, we performed MBA using Apriori algorithm on all the stores. However, we observed that the item ‘CMN Donations’ (which is not really an item) is occurring almost in each association rule. Also, there are certain item descriptions that depict promotional codes containing strings like INST SAVINGS, COUPON SAVE and thus cannot be considered as actual items. Hence, we removed the entries for these records. We

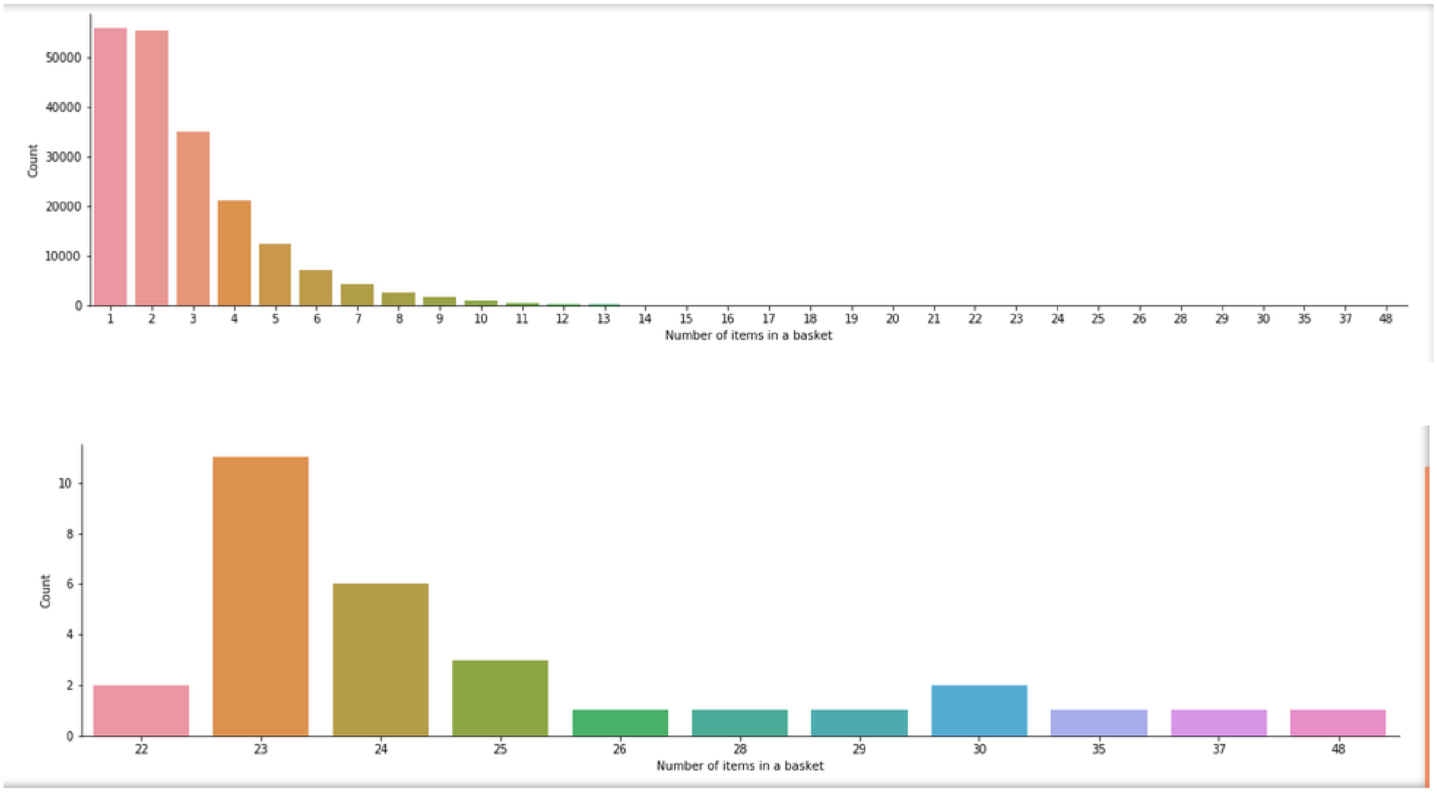


Fig. 2: Frequency of number of items in basket

observed that ‘Fasteners’ is an item that is frequently bought, almost with all the items. Hence, to identify significant associations, we removed ’Fasteners’ records.

To perform MBA, we consolidated the items into one transaction per row, ie, all items with the same receipt number form a single row. Now, ’Receipt Number’ is our index and all the items form the columns. We then performed one hot encoding on the columns, ie, when the item is present in the basket (any positive net sales units value), the value of that column is one otherwise, zero.

We determine association between Item 1 and Item 2 by taking into account the frequency of both being in the same basket. Hence, baskets that have a single item cannot be used to determine any association rules. So, while performing MBA we have excluded such baskets. Now that the data is structured properly, we can generate frequent item sets that have a support of at least 1%. Then we generated the rules with their corresponding support, confidence and lift, using MBA.

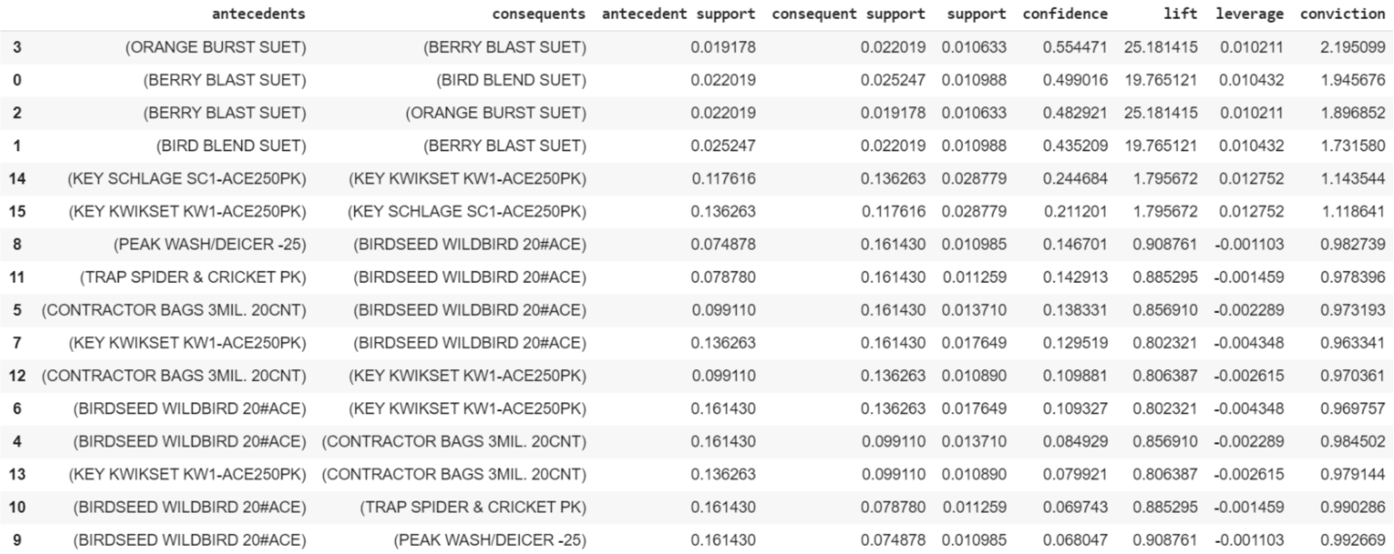
[Figure 3] depicts the association rules generated by MBA. From the output, the lift of the association rule *if ’Orange Burst Suet’ then ’Berry Blast Suet’* is 25.18 and the confidence is 55%. This means that consumers who purchase *Orange Burst Suet* are 25.18 times more likely to purchase *Berry Blast Suet* than randomly chosen customers. Larger lift means more interesting rules. Association rules with high support are potentially interesting rules.
Similarly, rules with high confidence are interesting rules as well. For example, there is a definite association between ORANGE BURST SUET and BERRY BLAST SUET. Which means there is a high chance that the two things will co-occur in a basket. We can also see some strong associations between other items which are bought together like KEY SCHLAGE and KEY KWIKSET. Hence, optimally placing the associated items in the store can help increase the revenue.

[B] Product Inventory Management - Time Series Analysis:

Time series analysis is done with a purpose of identifying trends, cycles and seasonal variances to aid in the forecasting of the future events based on previously observed values. In this project, we build a model to predict the stock of products in future.

1) **Data Preprocessing:**
We encountered certain irregularities in the dataset and resolved them as below:

- Header rows and Date feature were treated the same way as done for MBA in section [A].
- Converted ’Net Sales Units’ feature to numeric value.
- Label encoding for categorical features like ’Store #’ and ’Item Number’.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
3	(ORANGE BURST SUET)	(BERRY BLAST SUET)	0.019178	0.022019	0.010633	0.554471	25.181415	0.010211	2.195099
0	(BERRY BLAST SUET)	(BIRD BLEND SUET)	0.022019	0.025247	0.010988	0.499016	19.765121	0.010432	1.945676
2	(BERRY BLAST SUET)	(ORANGE BURST SUET)	0.022019	0.019178	0.010633	0.482921	25.181415	0.010211	1.896852
1	(BIRD BLEND SUET)	(BERRY BLAST SUET)	0.025247	0.022019	0.010988	0.435209	19.765121	0.010432	1.731580
14	(KEY SCHLAGE SC1-ACE250PK)	(KEY KWIKSET KW1-ACE250PK)	0.117616	0.136263	0.028779	0.244684	1.795672	0.012752	1.143544
15	(KEY KWIKSET KW1-ACE250PK)	(KEY SCHLAGE SC1-ACE250PK)	0.136263	0.117616	0.028779	0.211201	1.795672	0.012752	1.118641
8	(PEAK WASH/DEICER -25)	(BIRDSEED WILDBIRD 20#ACE)	0.074878	0.161430	0.010985	0.146701	0.908761	-0.001103	0.982739
11	(TRAP SPIDER & CRICKET PK)	(BIRDSEED WILDBIRD 20#ACE)	0.078780	0.161430	0.011259	0.142913	0.885295	-0.001459	0.978396
5	(CONTRACTOR BAGS 3MIL. 20CNT)	(BIRDSEED WILDBIRD 20#ACE)	0.099110	0.161430	0.013710	0.138331	0.856910	-0.002289	0.973193
7	(KEY KWIKSET KW1-ACE250PK)	(BIRDSEED WILDBIRD 20#ACE)	0.136263	0.161430	0.017649	0.129519	0.802321	-0.004348	0.963341
12	(CONTRACTOR BAGS 3MIL. 20CNT)	(KEY KWIKSET KW1-ACE250PK)	0.099110	0.136263	0.010890	0.109881	0.806387	-0.002615	0.970361
6	(BIRDSEED WILDBIRD 20#ACE)	(KEY KWIKSET KW1-ACE250PK)	0.161430	0.136263	0.017649	0.109327	0.802321	-0.004348	0.969757
4	(BIRDSEED WILDBIRD 20#ACE)	(CONTRACTOR BAGS 3MIL. 20CNT)	0.161430	0.099110	0.013710	0.084929	0.856910	-0.002289	0.984502
13	(KEY KWIKSET KW1-ACE250PK)	(CONTRACTOR BAGS 3MIL. 20CNT)	0.136263	0.099110	0.010890	0.079921	0.806387	-0.002615	0.979144
10	(BIRDSEED WILDBIRD 20#ACE)	(TRAP SPIDER & CRICKET PK)	0.161430	0.078780	0.011259	0.069743	0.885295	-0.001459	0.990286
9	(BIRDSEED WILDBIRD 20#ACE)	(PEAK WASH/DEICER -25)	0.161430	0.074878	0.010985	0.068047	0.908761	-0.001103	0.992669

Fig. 3: Market Basket Analysis - Association Rules

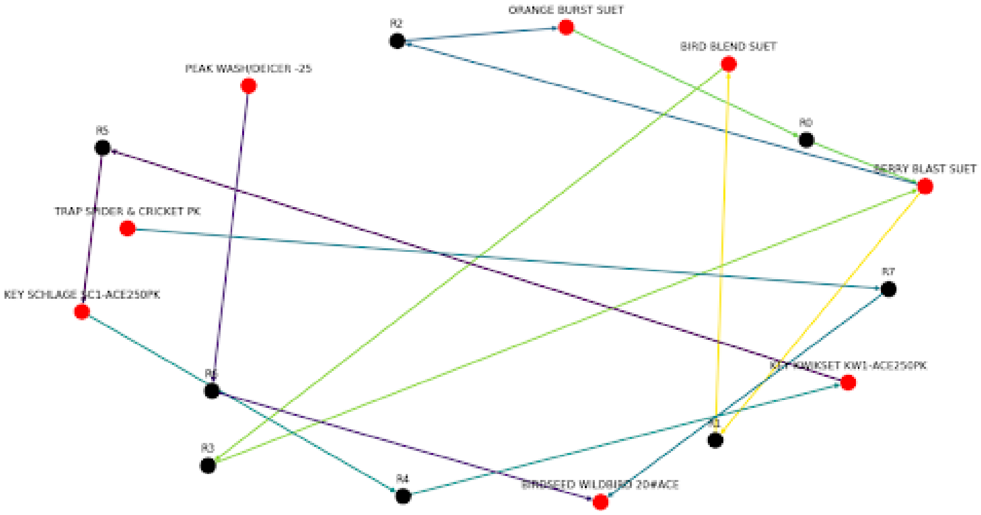


Fig. 4: Visualization of Association Rules

- Removed the records with ’Item Type’ as ’Defective’ and ’Return’. This is because the stock prediction should only be done based on the actual sales of items.

2) **Model:**
Features used: ’Date’, ’Store #’, ’Item Number’, ’Net Sales Units’.

Description:
- We grouped the data based on Date, Store #, Item number features and did summation of Net Sales Units for the combination.

Then, we performed some analysis:
- We analysed weekly Store Sales which for each store over the course of a few months. Here we wanted to see if there are any seasonality trends in the total store sales.

We visualised our data using time series decomposition that broke down the time series into trend, seasonality, and noise. Refer figure 5 for this. The trend of sale is varied and there is a pattern in seasonality wherein the sale increases from the month of May to June as well as in December every year.

- There’s definitely a seasonality in the store’s sales across the year.
- Day of week plays a role in sales. However, all the stores have similar distributions i.e. it follows a general weekly trend.

Model building:
- Date being a non-numeric field, we have engineered the dataframe by adding features such as Day of

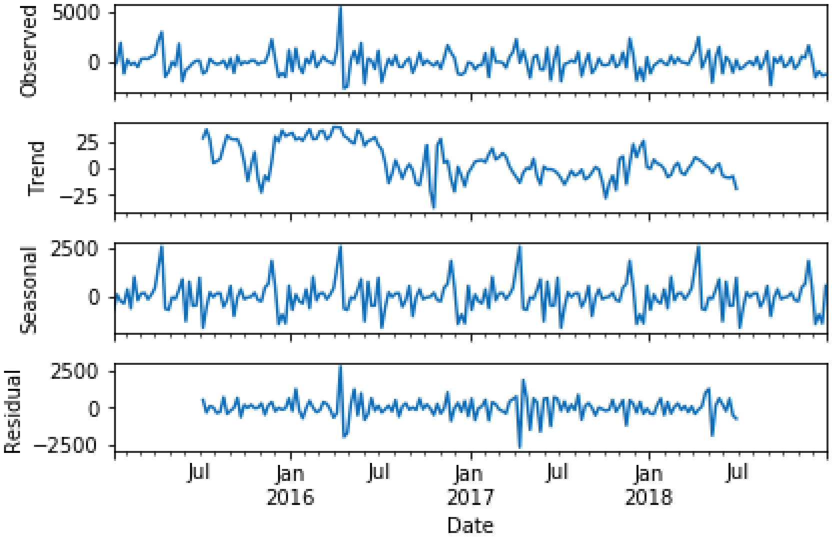


Fig. 5: Time Series Analysis: Seasonality trends

Year, Day of week, Month and whether it is a weekday/weekend.
- We have also added prior year sales.
- We have split the dataset into train and test where test data consists of last 5 months.
- We have used XGBoost model to predict the sales units and gauge the accuracy of the model using RMSE.

3) **Results:**
We tested our model against the test dataset and got the root mean square error (RMSE) value as 1.06006. From this, we can deduce that our model has performed quite well. Hence, the predicted values by our model could be helpful for item-wise demand prediction for various stores.

[C] Loss Leader Analysis

Loss leader is a product offered at cheaper price in expectation of selling other products having higher margin and in turn increase customer base in future. For an instance, giving offers and discounts on some product which is related to other product such that purchasing both products at a time will increase overall profit.

1) **Data Preprocessing:**
We encountered certain irregularities in the dataset and resolved them as below:

- Header rows and Date feature were treated the same way as done for MBA in section [A].
- Converted ’Gross Margin’ feature to numeric value.
- Removed the records belonging to items containing Item Description such as ’Coupons’, ’Rewards’, ’INST SAVINGS’, etc.. This is done because these promotional items are not actual items and hence cannot be considered for this analysis.

2) **Analysis:**
Features used: ’Store Name’, ’Item Description’, ’Gross Margin’.

Description:

For this analysis, we calculated ’Gross Margin’ per store by grouping dataset per store and summing over the margin for each store. Figure 6 shows the aggregate gross margin of each store.

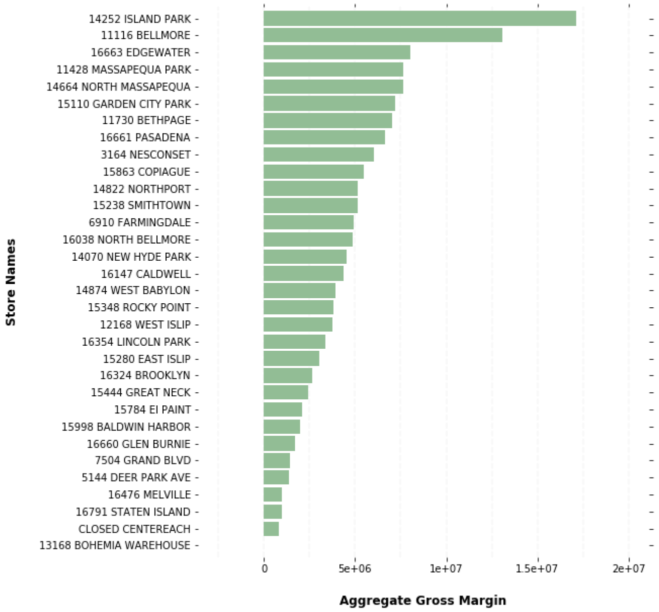


Fig. 6: Aggregated Gross Margin

From the figure, we can observe that Island Park store has the maximum gross margin among other stores. We also created a plot for only negative gross margin that each store has incurred. This is done by grouping store items of each store that have negative gross margin. From this plot, we observed that Island Park store has second highest negative loss margin among all the stores. Figure 7 shows the negative loss margin.

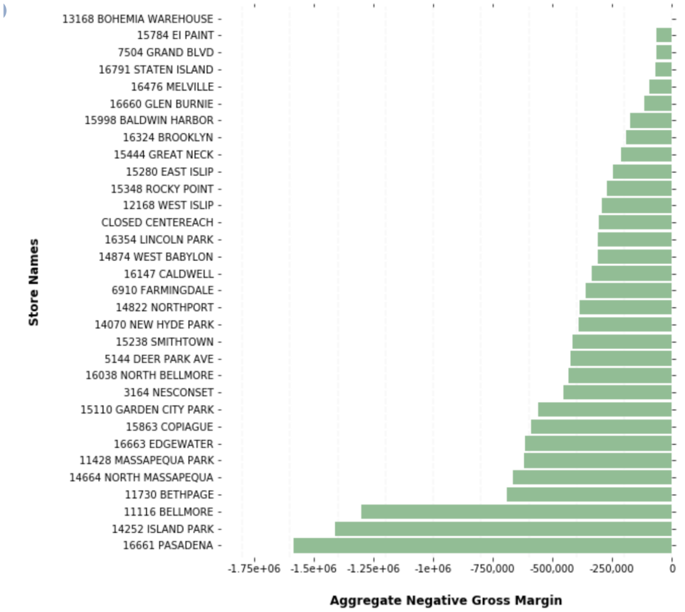


Fig. 7: Aggregated Negative Gross Margin

Hence, it could be noted from above two figures that even after bearing a major negative gross margin, the Island Park store achieves the maximum total gross margin. Thus, from this, it could be deduced that some loss leader pricing has been done by Island Park store. To further analyze, we curated the list of items for Island Park store which have negative gross margin. We selected top 15 products having highest negative loss margin. Figure 8 shows that there are some items like Fasteners, Lawn Food, LED Feit and Top Soil etc. are offered at lower prices (with negative gross margin) to the customers in expectation of purchasing other products which are relatively expensive. Thus we can say that, the items like Fasteners, Lawn Food, LED Feit and Top Soil etc. are used as Loss Leader products which gives us an indication that giving away these products at a loss, will lead to customers having to buy other products, in order to make use of them.

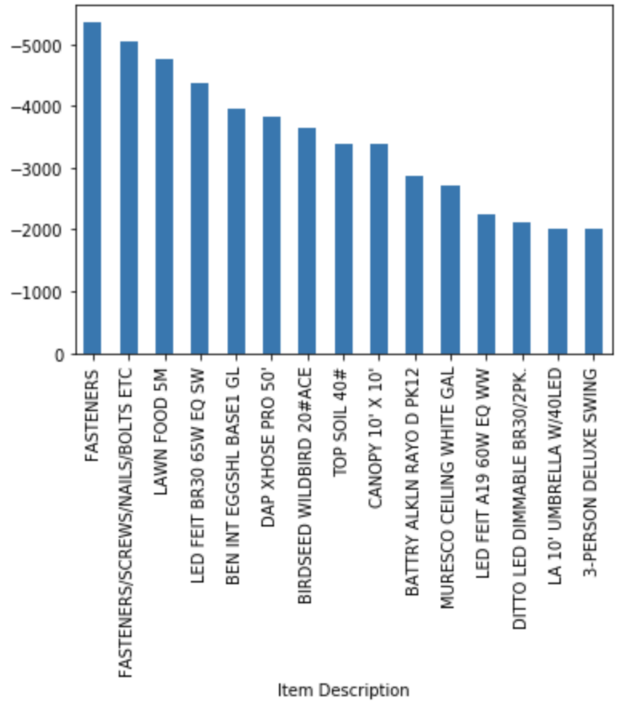


Fig. 8: Negative Gross Margin per Item (Island Park store)

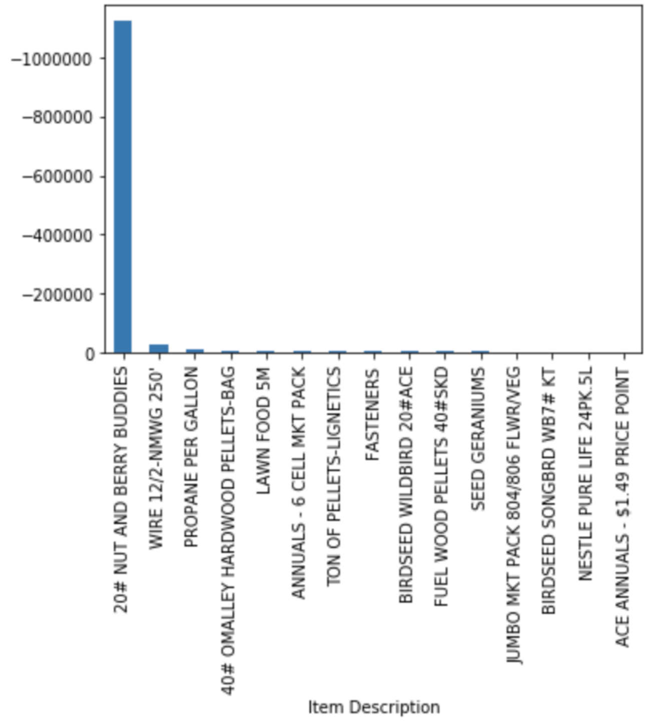


Fig. 9: Negative Gross Margin per Item (Pasadena store)

Furthermore, we did similar Loss Leader analysis for PASADENA store having store number 16661. We grouped each item in PASADENA store having negative gross margin and selected top 15 items having most negative gross margin. Figure 9 shows that NUT AND BERRY BUDDIES has a huge negative gross margin, most among all items in PASADENA store. However, upon further analysis, it was found out that the positive margin for NUT AND BERRY BUDDIES is 1139074.12 while the negative margin is 1125783.58. The resultant difference between Positive and Negative margin is very low i.e. 13290.54. This large negative gross margin shows that either the Leader Loss analysis is not done properly for item NUT AND BERRY BUDDIES or that the item itself is not profitable for the PASADENA store. The figure 9 shows negative gross margin of NUT AND BERRY BUDDIES is significantly large than other items. Hence, PASADENA store should revisit its negative margin items and increase margin on NUT AND BERRY BUDDIES in order to increase profits. Also, interesting insights could be taken from Island Park store which has implemented the Loss Leader strategy successfully.

IV. Interesting Insights

[A] While calculating Aggregated gross margin for Loss Leader Analysis we analysed the sales of the stores having very less sales. We found that three of the stores namely Melville, Staten Island and Bohemia Warehouse are newly opened in the year 2017-2018. The figure 10 shows weekly store sales of these stores.

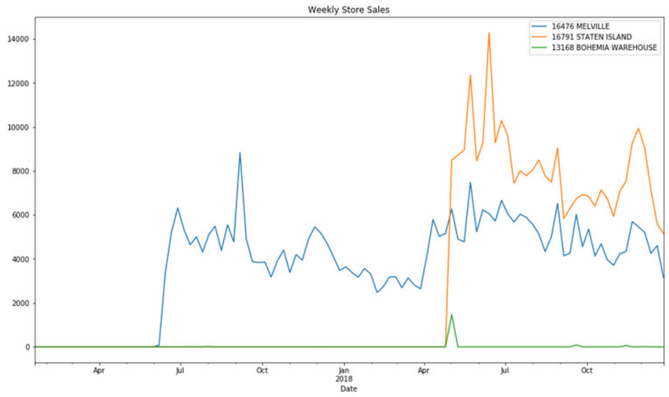
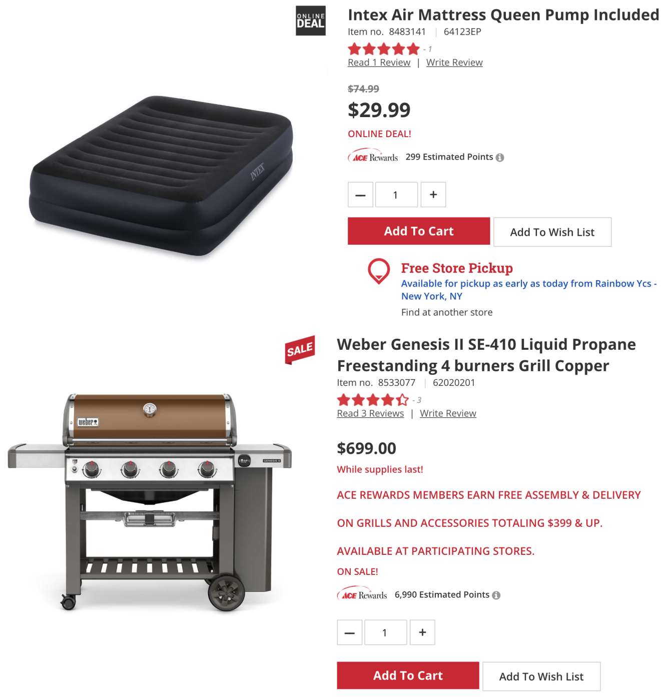


Fig. 10: Sales for newly opened store

When we checked on these stores on ACE hardware web-site, Bohemia Warehouse is a company warehouse instead of ACE Hardware store. On web-site we can see that, some of the items on sale can be picked up directly from the ACE Hardware Warehouse. This is the reason behind the the transaction entries of Bohemia Warehouse.
The Melville and Staten Island stores weekly sales graph in above figure shows that even though Staten Island Store is opened after an year it has gained more sales as compared to Melville store.

[B] A few items are in high demand only during certain seasons. Such items get stocked up in inventory when the season is over and can be sold again only in the next season. The cost of storage of such items can be more and hence retailers try to clear up these stocks by offering promotional deals during off season time. Such deals can be offered on festival sales like Labour day sales and Black Friday sales. To explore on this baseline, we analyzed the items by their Date of purchase. If it shows that the item is sold only in specific months, it suggests that the item is a season item. These seasonal items may be eligible for discount in off season sales. To validate whether the item is seasonal or not, we calculated season wise mode for each item by date. For example, people tend to buy Liquid Propane Freestanding 4 burners Grill Copper (Item Number: 8533077) in summers when there are holidays. So, in winter sales Costello can offer discount on such item. We found more examples on this trend like Traeger Pro Series 22 Wood Pellet Freestanding Grill Bronze(Item Number: 8474793) and Intex Air Mattress (Item Number: 8483141). We validated this prediction on Black Friday Sale (Fig. 11) where item number 8533077 and item number 8483141 was on sale which are seasonal.



Intex Air Mattress Queen Pump Included
Item no. 8483141 | 64123EP
★★★★★ · 1
Read 1 Review | Write Review
$74.99
$29.99
ONLINE DEAL!
ACE Rewards 299 Estimated Points
– 1 +
Add To Cart | Add To Wish List
Free Store Pickup
Available for pickup as early as today from Rainbow Ycs - New York, NY
Find at another store

Weber Genesis II SE-410 Liquid Propane Freestanding 4 burners Grill Copper
Item no. 8533077 | 62020201
★★★★☆ · 3
Read 3 Reviews | Write Review
$699.00
While supplies last!
ACE REWARDS MEMBERS EARN FREE ASSEMBLY & DELIVERY
ON GRILLS AND ACCESSORIES TOTALING $399 & UP.
AVAILABLE AT PARTICIPATING STORES.
ON SALE!
ACE Rewards 6,990 Estimated Points
– 1 +
Add To Cart | Add To Wish List

Fig. 11: Result Validation

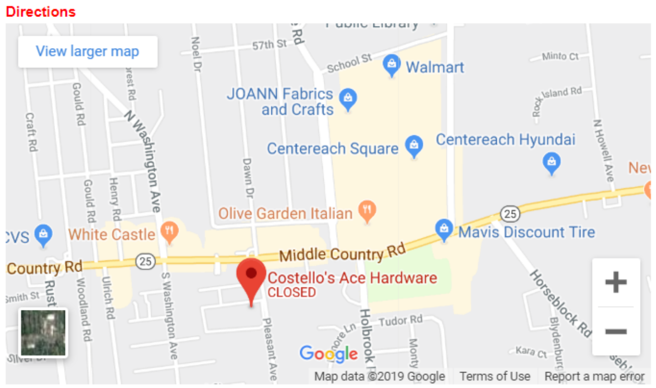


Fig. 12: CLOSED CENTEREACH store

[C] Each shop in costello dataset is represented in format (STORE_NUMBER STORE_NAME). For example, '16476 MELVILLE'. While calculating Gross Margin aggregate of all stores during Loss Leader Analysis, we found that one of the store in Costello's dataset is named as 'CLOSED CENTEREACH'. While validating this on Costello's website we found that this store was closed in 2016 (Figure 12). On further investigation, it was found that one of the probable reasons for this can be 'Harbor Freight Tools', another Retail chain store carrying a wide range of tools, hardware other products for the home, garden car, that is situated in the

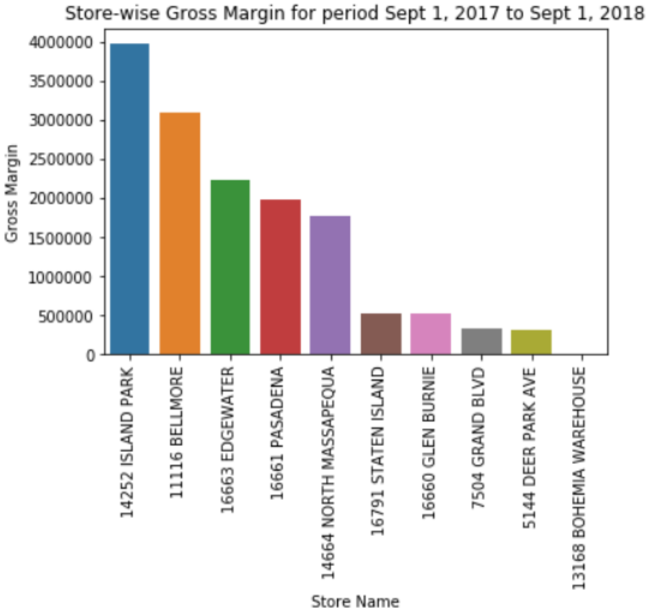


Fig. 13: Store-wise Gross Margin

area nearby. This store has google rating of 4.2 which is quite a good rating for any store. So it is possible that people could have preferred to buy items from Harbor Freight Tools than Costello's ACE hardware store which resulted in their decreased sales.

[D] While doing a deeper analysis of the sales data, we came up with an idea to provide the company with a script that could be weekly run by them to generate a sales report. This report will consist of the statistics and graphs showing the weekly (or for a span the user wants) gross margin for top 'k' stores and top-selling products among all the stores. Though the script is relatively simple and prima facie the report seems pretty basic, it would be a good actionable insight for the store owner. We think this could be a potential deliverable to the store owner. Using such a report, store owner would be able to gauge the performance of different stores and could also suggest specific item sale to specific stores based on their relative performance. This report could also help predict the sales of specific items based on the weekly sales and stock up the products in advance if necessary.
Figure 13 shows a sample graph from such a report.

V. Challenges

1) Working with huge dataset consisting of 32955543 (around 32 million) records in total.
2) Many of the items like 'INST SAVINGS', 'COUPONS' were not suitable for the analysis and needed to be detected and removed.
3) During Market Basket Analysis, many items needed to be filtered before the analysis. In addition to the promotional records, we also needed to filter out the baskets/transactions which consisted of only one item as such records would not be useful for our association analysis.

VI. References

[1] Jain, A.K., Menon, M.N., Chandra, S. (2015). Sales Forecasting for Retail Chains.

[2] Beheshti-Kashi, Samaneh Karimi, Hamid Thoben, Klaus-Dieter Lütjen, Michael Teucke, Michael. (2015). A survey on retail sales forecasting and prediction in fashion markets.

[3] Manpreet Kaura, Shivani Kanga (CMS 2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining.

[4] Introduction to Market Basket Analysis: https://pbpython.com/market-basket-analysis.html

[5] Frequent Itemsets via Apriori Algorithm: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori

[6] Andrej Trnka (2010). Market Basket Analysis with Data Mining methods.

[7] Soheila Mehrmolaei ; Mohammad Reza Keyvanpour (2016). Time series forecasting using improved ARIMA.

[8] http://airccse.org/journal/ijcsea/papers/4214ijcsea02.pdf

[9] https://www.kaggle.com/enolac5/time-series-arima-dnn-xgboost-comparison

[10] https://mode.com/example-gallery/python_horizontal_bar/

[11] https://intelligentonlinetools.com/blog/2018/02/10/how-to-create-data-visualization-for-association-rules-in-data-mining/