CSE 519 Retail Sales Data Analysis

1 Introduction

In the world of retail, the biggest challenge is to aggregate the right data and make sense out of it. The retailers have a huge amount of data at their disposal. If used properly, we can bring them up to speed with their competitors. This report is about the retail sales analysis of Costello's data. Retail data analysis provides insights that are needed to make informed decisions to grow revenue and business' profitability. Data allows retailers to make better operational and behavioural decisions by delivering dynamically in real-time. This report describes the method used to achieve some of the major objectives of the whole project.

Forecasting plays an important role in various fields such as business, finance, weather et al. In retail business, a successful prediction of sales volume can help managers to make logical plans for the future. We have used ARIMA model **[4]** for time-series forecasting. Weather also plays an important role in the sales of products. For example: heavy precipitaion on the road tends to slow down the traffic and hence results in lesser in-store crowd. So we have made use of the weather dataset[Cite Weather Reference] and did its analysis with the sales data.

Determining association rules i.e. finding relationship between items which are generally bought together, is an important information which retailers can use in order to boost up their sales. This information is used in product placement strategy i.e. which products should be placed near to each other so as to boost up the sales by making it more convenient for customers to find items easily.

The report also contains analysis of some other objectives which we feel are interesting insights and some objectives which was posed by the representative from Costello. Some other important aspects of retail store chain is determining new spots for opening new stores. An attempt has been made in this regard in this project. The reports mentions about various data exploration and detailed analysis of results.

2 Objectives

In this final report we have worked on the following objectives which are some of the major objectives in this project.

- Which products to place near other products. (Market Basket Analysis)
- Which products to stock. (Time Series Analysis)
- Why Bethpage Store is performing low as compared to others?
- Prediction of possible new locations to open a store.

3 Time Series Analysis

In this part, we are going to perform a time series analysis to get some interesting insights from the data. Our goal is to find a pattern/trend in the sales of the data, make inferences from it and build a model to forecast the sales volume which will help us to find which items to stock during various time periods. The sales volume prediction is crucial to retailers, as it could help the retailers make correct decisions. Overestimated sales can result in excessive inventory or unhealthy cash flow, while the underestimated sales may lead to unfulfilled orders, decreased profit.

Time series data is a sequence of data points measured over time intervals. In other words, data is a function of time

$$f(t) = y$$

Researchers have developed models of time series forecasting. For a single variable prediction, only considering the sales volume of one product and predicting the future, we can use the classical statistical models such as autoregressive, and seasonal decomposition with any model **[4]**. In this project, we have attempted ARIMA model to do single variable prediction and being able to forecast the sales volume of various departments. We have also incorporated an external weather dataset to see if it has any ramifications on the outcome of sales.

3.1 External Data: Weather Analysis

Here we have tried to find a relationship between sales from a particular department and the weather. To start we incorporated hourly weather data of Long Isand from NOAA and merged it with the original sale data(2015 - 2018).

3.1.1 Weather Data Preparation

- Select the time range on NOAA's website and download the data.
- We made a dataframe of the downloaded data and removed the unnecessary columns and used columns "Temperature", "Precipitation" and "Snow Depth".
- Parsed the time into the format of POS data, to use it as key while merging with the POS data.
- Since, the weather dataset is hourly, we grouped it by date, taking the average on the above selected columns.
- Finally, merging the data with the POS data by using a left join.

3.1.2 Analysis with weather data

On plotting the Snow Depth with the sales of items in a Department, we some observed items in the department delineate a strong relationship with the weather. For instance sales in "Automotives" department peaks when the average Snow Depth increases, likely because cars require more care and maintenance when the mercury plummets. It's not a surprise then that people rush in to buy more "ANTIFREEZE" & "WIPER BLADES". Figure[1] is the plot that shows Snow Depth vs Sales in Automotive Department.

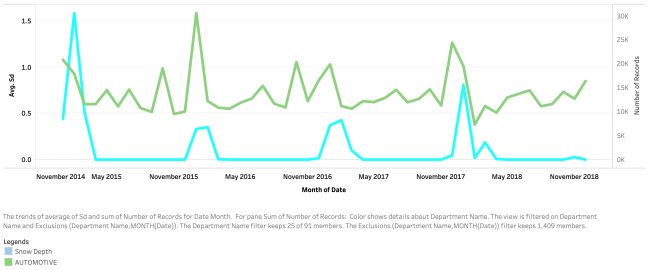


Figure 1: Snow depth vs Sales in "Automotives" Department

Similarly, we observed a similar pattern in "OUTDOOR EQUIPMENT & POWER TOOLS" department, that the sales in this department peaks when average Snow Depth increases. Figure[2] shows the same.

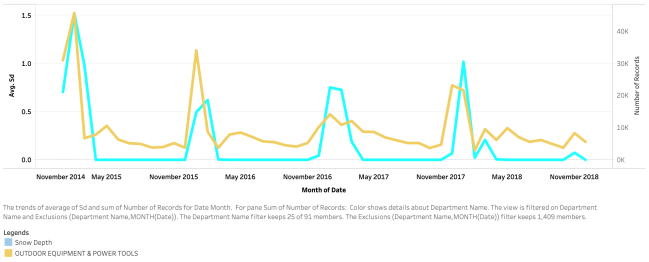


Figure 2: Snow depth vs Sales in "OUTDOOR EQUIPMENT & POWER TOOLS" Department

We reckon this happens because items like "SNOW SHOVEL" & "ICE MELT BAG". see [Figure 3] are heavily utilized when the snow starts to fall and so are in high demand.

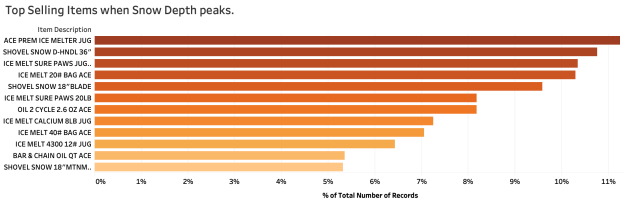


Figure 3: Top 10 items sold when snow peaks.

Based on our analysis here for the above departments we will now try to predict the sales in these departments.

3.2 Data Preparation

For the time series analysis, we considered the Date, Net Sales Units, Department Name from the dataset.

- Transactions were filtered based on Sales.
- We filtered data based on the required Department Name. For instance we have picked the Department "Automotive".
- We removed the unnecessary columns from the data.
- Date column was converted to DateTime format and Net Sales Units to numeric.
- The resulting data frame was sorted and grouped by Date, summing over the Net Sales Units and sampling out the mean(i.e. Taking the Net Sales Units in a month with a mean over the daily sales).
- To make the data linear and smoother we have log transformed the "Net Sales Unit".
- We have used Differencing technique to make the data stationary.

3.2.1 Stationary Test

Stationarizing a time series is the process to be able to obtain meaningful sample statistics such as means, variances, and correlations with other variables. Such statistics are useful as descriptors of future behavior only if the series is stationary. For example, if the series is consistently increasing over time, the sample mean and variance will grow with the size of the sample, and they will always underestimate the mean and variance in future periods rendering the current model with non stationary data useless and we have to in turn build a new model. We simply predict that its statistical properties will be the same in the future as they have been in the past. To check the whether the data is stationary we used the Augmented Dickey-Fuller (ADF) test. The **null hypothesis**, H_0 = data is not stationary. ADF test result provides test statistic and P value. **P value** $>= 0.5$ means the data is not stationary, otherwise, we reject the null hypothesis and say that the data is stationary.

3.2.2 Differencing

Data can be made stationary by transforming it using a method known as Differencing. The first difference of a time series is the series of changes from one period to the next. It is the difference between y at time = t and y at time = $t - x$, where x is the lag. $\text{diff}_1 = y_t - y_{t_x}$ Differencing makes the data stationary as it removes time series components, trends and seasonality from the data and you are left with changes between time periods.

3.3 Model

Time series forecasting is used to predict future values based on previously observed values. We will use ARIMA: stands for Autoregressive Integrated Moving Average, model which is denoted with the notation ARIMA(p, d, q). These three parameters account for seasonality, trend, and noise in data. Seasonal ARIMA(SARIMA) is used to understand the trend and forecast the results. There are four seasonal elements that are not part of ARIMA that must be configured; they are:
P: Seasonal autoregressive order.
D: Seasonal difference order.
Q: Seasonal moving average order.
m: The number of time steps for a single seasonal period. (12 is used since it's a yearly data)
The general $SARIMA(p,d,q)(P,D,Q)_m$ process X_t is the solution of the following equation, where Z_t is the white noise process. [7]

$$\Phi(B^m)\,\phi(B)\nabla_m^D\nabla^d X_t = \Theta(B^m)\,\theta(B)Z_t$$
$$\nabla_m X_t = X_t - X_{t-m}, \nabla X_t = X_t - X_{t-1}$$
$$\Phi(B) = 1 - \Phi_1 B^m - \cdots - \Phi_P B^{PM}$$
$$\phi(B) = 1 - \phi_1 B^m - \cdots - \Phi B$$
$$\Theta(B^m) = 1 - \phi_1 B^m - \cdots - \Theta_Q B^{Qm}$$
$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q \text{ and } B^n X_t = X_{t-n}$$

The data was filtered for Department "AUTOMOTIVE". After preparing the data as mentioned above, i.e log transformation and differencing, the ADF test gave us p-value of 0.001, which is less than 0.05 and hence significant. Thus, the Null hypothesis was rejected and the data under consideration is now stationary. We now need to find the optimal value for p, d, q to train the data which is accomplished using hyper parameter using grid search. Parameters (p, d, q) i.e (1, 0, 1) and (P, D, Q) i.e (1, 1, 0) were chosen as it had the lowest AIC value (64.03). A low AIC value indicates a better fit.

Using the above tuned values, the model was trained on data having date range from (2015 - 2017). The data was then tested on 2018 sales data. Unlike random sampling that is usually performed on the cross sectional data, we do the above to preserve the temporal nature of the time series of the data. We cannot do random sampling like we do for cross-sectional data. We have to keep the temporal behaviour (dependence on time) of time series data.

To help us understand the accuracy of our forecasts, we compared the predicted sales to real sales of the time series, and set forecasts to start from 2018–05–01 to the end of the data.

3.4 Analysis and Results

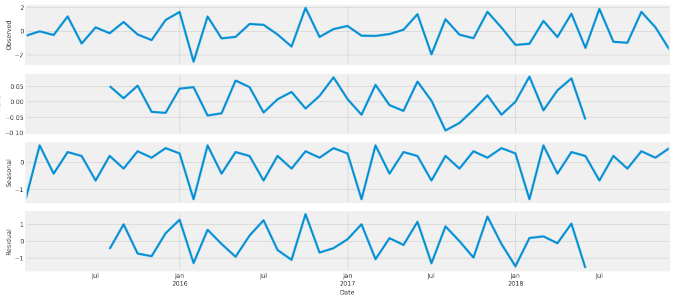


Figure 4: Trend in Sales and Seasonal Pattern

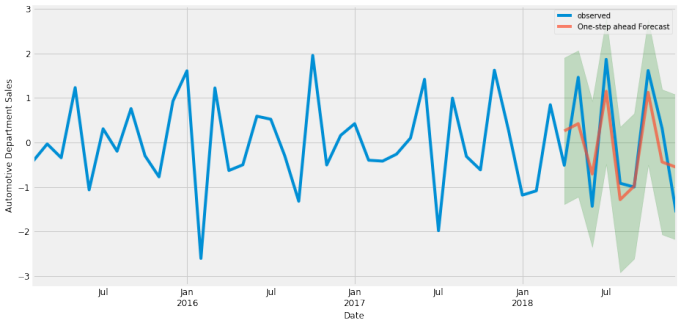


Figure 5: Validation of our prediction (Units vs. Time)

In Figure 4 we visualised our data using time series decomposition that broke down the time series into trend, seasonality, and noise. The trend of sale is variable and varies in a particular year and there is a pattern in seasonality wherein the sale dips in the month of January but increases in March every year. We also observe that this is overall seasonality but if we compare with the snow data, the sales increases before january which was what we observed in snow data from Figure[1].

In Figure 5 the line plot depicts the observed values compared to the rolling forecast predictions. Overall, our forecasts align with the true values very well which validates our forecasts, showing an upward trend that starts from the beginning of march and captured the seasonality toward the end of the year. The mean squared error of our forecast is 0.52. The root mean squared error of our forecast is 0.72.

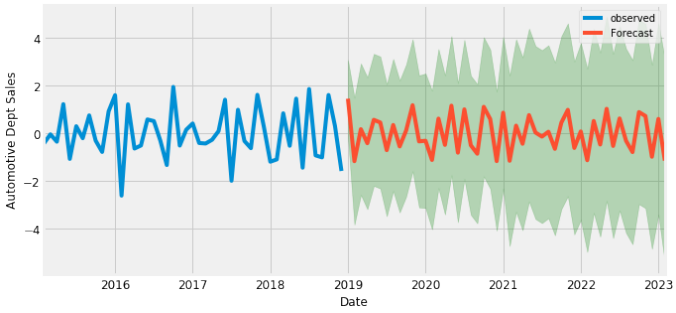


Figure 6: Future Sales Prediction (Units vs. Time)

Our model clearly captured AUTOMATIVE sales seasonality. As we forecast further out into the future in Figure 6, it portrays an upward trend in sales. It is natural for us to become less confident in our values. This is reflected by the confidence intervals generated by our model, which grow larger as we move further out into the future.

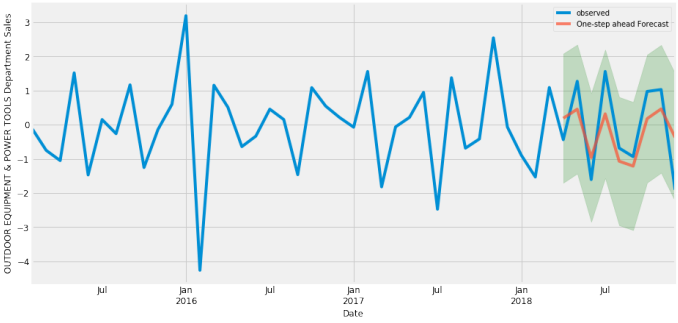


Figure 7: Future Sales Prediction for OUTDOOR EQUIPMENT & POWER TOOLS (Units vs. Time)

We also ran this model on Department: OUTDOOR EQUIPMENT & POWER TOOLS. Predictions can be observed in Figure[7] (sales peak during August & November) with ADF test resulting in p value of $3 * 10^{-5}$, Parameters (p, d, q) i.e (1, 0, 1) and (P, D, Q) i.e (1, 1, 0) were chosen as it had the lowest AIC value (70.05). The mean squared error of our forecast is 0.73. The root mean squared error of our forecast is 0.85. So based on the seasonality and forecasting we recommend that these items be stocked.

4 Market Basket Analysis

This is one of the major objectives of this project. Market Basket Analysis (MBA) is performed to find items that are brought together based on the popularity of the items that were purchased together by the previous customers. This information will be useful to understand the items that can be placed together, making it convenient for customers to find and buy them, which will aid in boosting sales of the store.

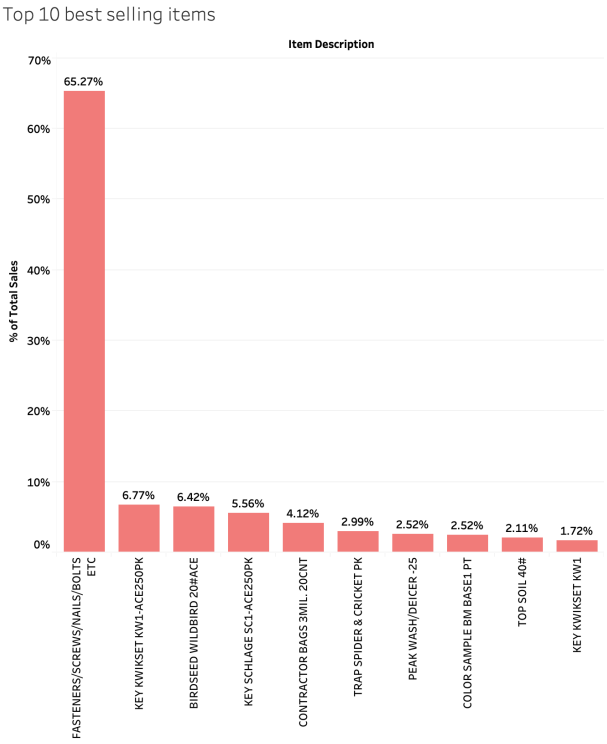


Figure 8: Top 10 Selling items in 11730 Bethpage

Figure[8] shows top 10 best selling items in the Bethpage store. Let's see if these top items show any strong associations in our results.

4.1 Data Preparation

- We started with converting data type from object to numeric types for the appropriate columns such as "Net Sales".
- For the analysis we used columns such as Receipt Number, Item Description and Net Sales Unit.
- Then we filtered the transactions based on Sales i.e (Line Item Transaction Type = Sale) AND a store in consideration.
- Item Descriptions such as 'CMN Donation', 'In-Store Coupon', '50% OFF 1 ITEM UNDER $30' and 'THANKSGIVING SALE 15%OFF' was dropped. This step is important to avoid associations that include aforementioned products with the actual items.
- Under Item Descriptions there two entries for FASTENERS i.e (FASTENERS & FASTENERS/SCREWS/NAILS/BOLTS ETC) with substantial sale frequency, so we clubbed them together.
- We then decided upon a threshold to filter out low selling items which varied with stores. For example for the store "11730 BETHPAGE", we set a threshold of 500, which

means we leave out the items which have been bought less than 500 times in the span of four years. Setting a threshold allows us to not include low selling items in our basket as it would not form a good association with any other item.

- Associations are formed for more than one item in a receipt or in a basket, so it makes sense to remove receipts that had less than two items. This step will also help reduce the processing. Below [Figure 9] shows the number of items in the Basket for bethpage, more than 70% sales involve a purchase of one item. We are left with a roughly 30% of data on which we will run our model.

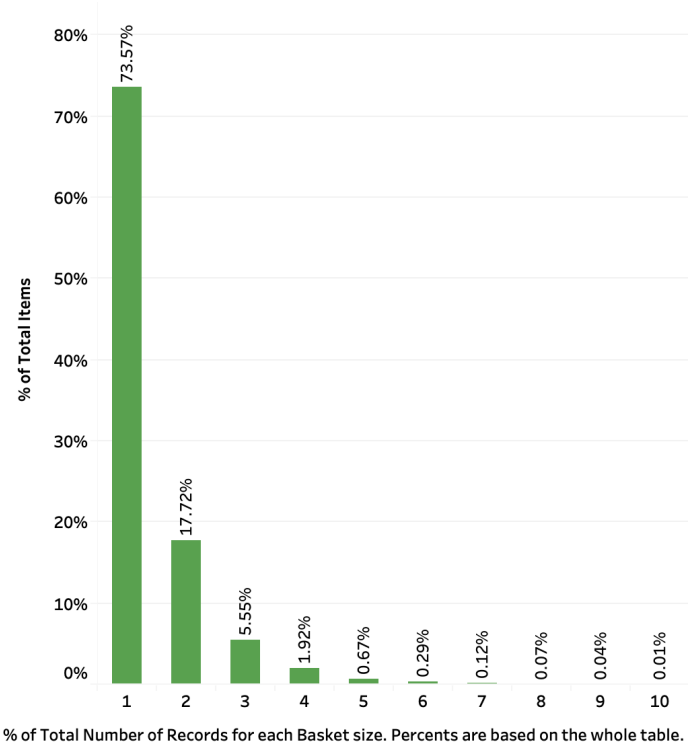


Figure 9: Number of items in basket for 11730 Bethpage before Step[6] in Data Preparation

4.2 Model

Market Basket Analysis is one of the key techniques to uncover associations between items. We have used association rules to analyze retail transaction data, to identify strong rules in transaction data using measures of interestingness, based on the concept of strong rules.

Understanding confidence, support and lift.

- **Support** is the number of transactions that include items in the $\{A\}$ and $\{B\}$ parts of the rule as a percentage of the total number of transactions (N). It is a measure of how frequently the collection of items occur together as a percentage of all transactions.

$$\text{Support} = \frac{frq(A,B)}{N}$$

- **Confidence** of the rule is the ratio of the number of transactions that include all items in $\{B\}$ as well as the number of transactions that include all items in $\{A\}$ to the number of transactions that include all items in $\{A\}$.

$$\text{Confidence} = \frac{frq(A,B)}{frq(A)}$$

- **Lift** is the ratio of confidence to expected confidence. Expected confidence is the confidence divided by the frequency of B. The Lift tells us how much better a rule is at predicting the result than just assuming the result in the first place. Greater lift values indicate stronger associations. [9]

$$\text{Lift} = \frac{\left(\frac{frq(A,B)}{frq(A)}\right)}{\left(\frac{frq(B)}{N}\right)}$$

After Pre-processing, we grouped the data by the 'Receipt number' and 'Item Descriptions' by summing up the 'Net sales unit'. After grouping them up, we encoded the information in the grouped-matrix for any quantity greater than 1 as 1 and rest as 0. This matrix was fed into the Apriori algorithm from the mlxtend library which resulted in the Item Description and its support value.

This was then used for association mining with a minimum support value of 1%. The outcome of the above association mining is the following antecedent, consequent departments, support, confidence and lift. We used the confidence and the lift values to determine the optimal values having a high confidence and lift value greater than 5.

4.3 Analysis, Results and Validation

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
8	(VIGOROUS VEGGIES 3 1/4" POTS)	(HERBS 3 1/2" POTS)	0.030971	0.044488	0.011183	0.361091	8.116521	0.009805	1.495537
9	(HERBS 3 1/2" POTS)	(VIGOROUS VEGGIES 3 1/4" POTS)	0.044488	0.030971	0.011183	0.251374	8.116521	0.009805	1.294411
10	(SPRING ANNUALS 4 1/2"POTS)	(SPRING ANNUALS 6" POT)	0.105763	0.033238	0.015741	0.148833	4.477742	0.012226	1.135808
11	(SPRING ANNUALS 6" POT)	(SPRING ANNUALS 4 1/2"POTS)	0.033238	0.105763	0.015741	0.473579	4.477742	0.012226	1.698710
4	(VIGOROUS VEGGIES 3 1/4" POTS)	(ASSORTED CELL PACKS $2.49)	0.030971	0.116235	0.012784	0.412778	3.551250	0.009184	1.504994
5	(ASSORTED CELL PACKS $2.49)	(VIGOROUS VEGGIES 3 1/4" POTS)	0.116235	0.030971	0.012784	0.109985	3.551250	0.009184	1.088778
3	(ASSORTED CELL PACKS $2.49)	(SPRING ANNUALS 4 1/2"POTS)	0.116235	0.105763	0.042132	0.362471	3.427209	0.029838	1.402662
2	(SPRING ANNUALS 4 1/2"POTS)	(ASSORTED CELL PACKS $2.49)	0.105763	0.116235	0.042132	0.398360	3.427209	0.029838	1.468928
0	(HERBS 3 1/2" POTS)	(ASSORTED CELL PACKS $2.49)	0.044488	0.116235	0.016497	0.370815	3.190225	0.011326	1.404618
1	(ASSORTED CELL PACKS $2.49)	(HERBS 3 1/2" POTS)	0.116235	0.044488	0.016497	0.141928	3.190225	0.011326	1.113557
6	(SPRING ANNUALS 4 1/2"POTS)	(HERBS 3 1/2" POTS)	0.105763	0.044488	0.013362	0.126340	2.839843	0.008657	1.093688
7	(HERBS 3 1/2" POTS)	(SPRING ANNUALS 4 1/2"POTS)	0.044488	0.105763	0.013362	0.300350	2.839843	0.008657	1.278120

Figure 10: Market Basket Analysis Result for 15863 COPIAGUE

Above is the result of running our model on data for the '15863 COPIAGUE' store.

From the output above, we see that the top associations are not surprising, with one item being purchased with another from the same item family (eg: Veggie Pots with Herb Pots, Sprint annuals with herb pots, etc). As mentioned, one common application of association rules mining is in the domain of recommender systems. Once item pairs have been identified as having positive relationship, recommendations can be made to customers in order to increase sales. And hopefully, along the way, also introduce customers to items they never would have tried before or even imagined existed!

Higher lift value means strong association. Earlier we were getting low lift values, we figured it was because of incorrect handling/filtering of data, for instance we did not incorporate Step[6] i.e(filtering low selling items form the basket) in our implementation earlier. Also the order of steps to be followed in the processing data matters as it produces different results. As a result of correcting and learning from our own mistakes we were able to get higher lift values.

We also ran the model on data from different stores & found various other strong associations. Some of them are shown below.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
5	(KEY KWIKSET KW1)	(KEY SCHLAGE SC1250PK)	0.017116	0.020898	0.010307	0.602210	28.816155	0.009950	2.461353
4	(KEY SCHLAGE SC1250PK)	(KEY KWIKSET KW1)	0.020898	0.017116	0.010307	0.493213	28.816155	0.009950	1.939441
6	(KEY KWIKSET KW1-ACE250PK)	(KEY SCHLAGE SC1-ACE250PK)	0.057400	0.074610	0.030827	0.537068	7.198339	0.026545	1.998974
7	(KEY SCHLAGE SC1-ACE250PK)	(KEY KWIKSET KW1-ACE250PK)	0.074610	0.057400	0.030827	0.413181	7.198339	0.026545	1.606289
3	(KEY SCHLAGE SC1-ACE250PK)	(KEY ARROW AR1-ACE)	0.074610	0.029504	0.015792	0.211660	7.174064	0.013591	1.231064
2	(KEY ARROW AR1-ACE)	(KEY SCHLAGE SC1-ACE250PK)	0.029504	0.074610	0.015792	0.535256	7.174064	0.013591	1.991184
8	(KEY SEGAL SE1-ACE)	(KEY SCHLAGE SC1-ACE250PK)	0.029693	0.074610	0.013712	0.461783	6.189303	0.011496	1.719364
9	(KEY SCHLAGE SC1-ACE250PK)	(KEY SEGAL SE1-ACE)	0.074610	0.029693	0.013712	0.183777	6.189303	0.011496	1.188777
1	(KEY SCHLAGE SC1-ACE250PK)	(FASTENERS/SCREWS/NAILS/BOLTS ETC)	0.074610	0.778440	0.010118	0.135615	0.174213	-0.047961	0.256321
0	(FASTENERS/SCREWS/NAILS/BOLTS ETC)	(KEY SCHLAGE SC1-ACE250PK)	0.778440	0.074610	0.010118	0.012998	0.174213	-0.047961	0.937577

Figure 11: Market Basket Analysis for 15444 GREAT NECK, threshold used here was 1000 to filter out low selling items

Out[243]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
6	(LAWN FOOD 5M)	(WEED & FEED 5M)	0.015595	0.014633	0.012070	0.773973	52.890953	0.011842	4.359501
7	(WEED & FEED 5M)	(LAWN FOOD 5M)	0.014633	0.015595	0.012070	0.824818	52.890953	0.011842	5.619314
4	(KEY KWIKSET KW1-ACE250PK)	(KEY SCHLAGE SC1-ACE250PK)	0.052534	0.046161	0.019867	0.378177	8.192569	0.017442	1.533940
5	(KEY SCHLAGE SC1-ACE250PK)	(KEY KWIKSET KW1-ACE250PK)	0.046161	0.052534	0.019867	0.430390	8.192569	0.017442	1.663357
0	(KEY KWIKSET KW1-ACE250PK)	(FASTENERS/SCREWS/NAILS/BOLTS ETC)	0.052534	0.475887	0.014402	0.274144	0.576070	-0.010598	0.722062
1	(FASTENERS/SCREWS/NAILS/BOLTS ETC)	(KEY KWIKSET KW1-ACE250PK)	0.475887	0.052534	0.014402	0.030263	0.576070	-0.010598	0.977034
2	(KEY SCHLAGE SC1-ACE250PK)	(FASTENERS/SCREWS/NAILS/BOLTS ETC)	0.046161	0.475887	0.011233	0.243347	0.511356	-0.010734	0.692673
3	(FASTENERS/SCREWS/NAILS/BOLTS ETC)	(KEY SCHLAGE SC1-ACE250PK)	0.475887	0.046161	0.011233	0.023605	0.511356	-0.010734	0.976898

Figure 12: Market Basket Analysis for 11730 BETHPAGE, threshold used here was 600 to filter out low selling items

Validation with Sniff Test:

After running the MBA model on 11730 BETHPAGE, we observe that if the antecedent is (LAWN FOOD) then the consequent are (WEED & FEED). With a high confidence and lift values greater than 5 we can say that if the customer purchases anything from the above antecedent they are very likely to purchase from the consequent. We can ignore high-confidence values which had lift values greater 1, because it doesn't indicate strong association. Similar inferences can be made from other rows in the table above.

The total number of times when A = LAWN FOOD, B = WEED & FEED were bought together are 676. Total times when A occurs in the list is 876. Therefore we can say that when ever A is purchased, B is bought along with it 77% of the time. This is in accordance with the results of our model see Figure[12]. Similar is the case with the other results we have shown so far.

5 Predicting New Location

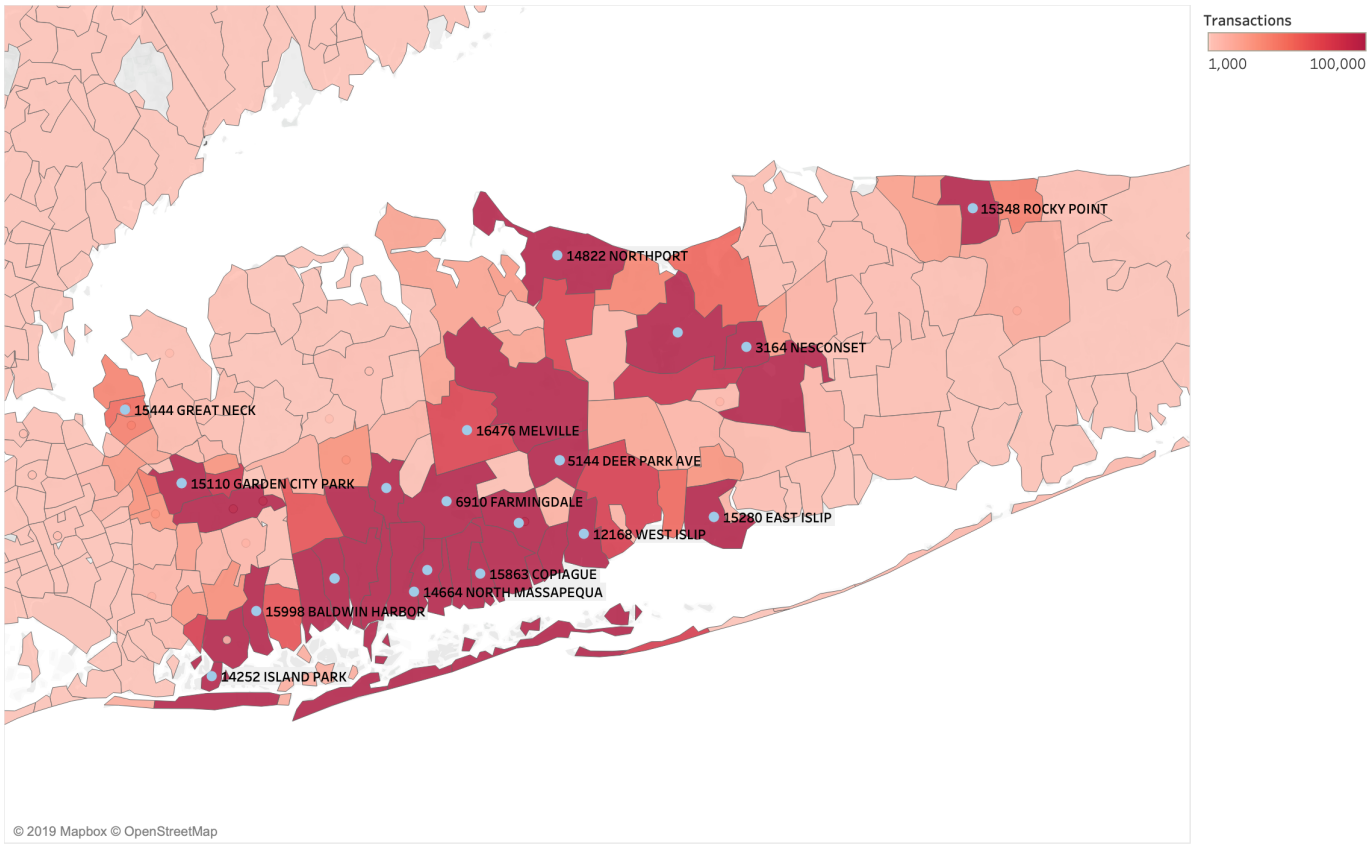
The store locations can tell us about their customer base and what could be the attractive locations to serve this base? We can infer at a high level where they should open a new store or explore new areas.

5.1 Incorporating External Dataset

We used uszipcode which is a programmable zipcode database in python to get the consensus data. It gives various fields like population, population_by_age, timezone, directional_bounds, etc. We used population_by_age and median_household_income.

5.2 Method

By analyzing the data and how it maps to Census demographics [8], we can create a high-level profile of where the new stores should or could be opened. In the given dataset we have 31 store names, we crawled the web to get their zipcodes. From the zipcode of each store located in NY(26 Stores), we calculated the median of the population above 19 years of age and the median household income. We used this median as a threshold while iterating over other zipcodes in NY. If any zipcode surpassed the threshold then we made it a prospective candidate for suggesting the new store location. Further, we mapped Figure[13] all the transactions based on zip codes on a map to see which locations have considerable movements, i.e (people purchasing from Costello's in other town) and if that location didn't have a store already and if it was present in our prospect candidate list then we finalized the selection of that zipcode. This way we finalized 10 zipcodes which are '11743', '11725', '11803','11566', '11731', '11001', '11050', '10016', '10023', '11793', '10011'. Inference from Figure[13] Near-magenta regions in the graph which doesn't have a blue point(Store) are the prospective locations of the stores.



Map based on registered customer's ZIP Code and the Transactions. For pane: Color shows sum of Number of transctions. Details are shown for Zip Code. Blue Dots are the store locations. The data is filtered on Loyalty ID, which excludes Null.

Figure 13: Transaction based on zipcodes

ZIP	Malls nearby	Competitors
11793	Willow wood, Cherry wood & Wantagh wood	WANTAGH 5&10
11566	Merrick common, Baldwin harbor shopping center	NA
11731	Huntington, Melville mall, Turnpike plaza	Karp's Hardware
11001	Glen oaks	HillSide Hardware

Table 1: Prospective ZIP codes exploration

Conclusion: We explored the prospective zipcodes suggested above, to see whether it was a residential area or it contained any commercial space where the new stores could be opened.This was done using google search and maps. We suggest that new store be opened in ZIPCODE 11566 since there is no other hardware store in the vicinity. Or in ZIPCODE 11001 where through google reviews we see that it's competitor HillSide Hardware isn't rated well, so it could be an ideal location to open the new store here.

Our suggestion for new stores is to promote as well as place the items next to each other based on the results from the Market Basket Analysis above such as "VIGOROUS VEGGIES POTS" , "HERB POTS" and "KEY KWIKSET", "KEY SCHLAGE" and others. Apart from this, the time series analysis would help in predicting which item to be stocked up periodically. This would help in the inventory management, boosting sales and driving profits.

6 Bethpage Store Performance Analysis

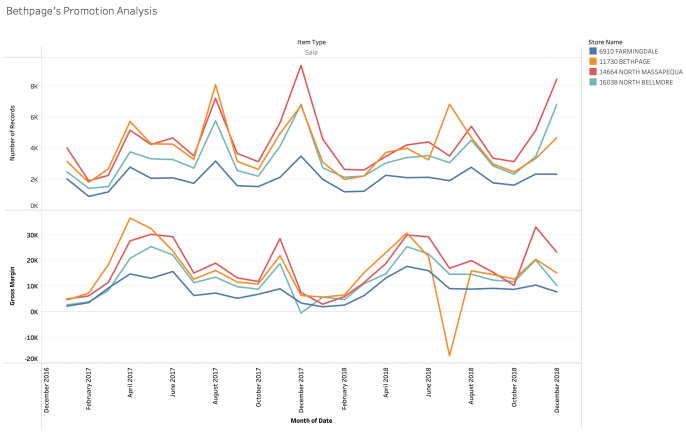


Figure 14: Bethpage's promotion Analysis

The Costello's store located in Bethpage is adjacent to stores located in FARMINGDALE, NORTH MASSAPEQUA and NORTH BELLMORE. The performance of Bethpage comparatively takes a hit due to the following possible reasons.

ZIP	STORE	M.Inc	Popu
11758	14664 NORTH MASSAPEQUA	104986	40123
11710	16038 NORTH BELLMORE	103686	25572
11735	6910 FARMINGDALE	91196	24182
11714	11730 BETHPAGE	87327	17304

Table 2: Data from uszipcode.

- Adjacent stores are located in areas where the median household income and population above 19+ years of age is greater, which tells us why not many people are spending in Bethpage.
- In addition, another hardware giant store, LOWES' is a few minutes away from the store in Bethpage which toughens competition for Costello.
- Further, from the POS data, plotting Figure[14] transactions filtering sales & returns that occurred during promotions we came across an interesting observation. There is a plunge in gross margin when the sale rose between June'18 - September'18. This suggests that the store took a drastic step to boost sales during that period i.e. in order to increase the sales they heavily discounted the products which lead to a negative gross-margin. This dip in gross-margin contributes heavily to its performance.

The top items that were promoted [as shown in Figure 15] with very high discount that brought the gross margin down, for instance take "DRILL & IMPACT", 53 of these products were sold during this period, it's Retail Price (R): $9539, but it was sold for $5532 with almost 40% off. More such promotions can been seen in the Figure 15.

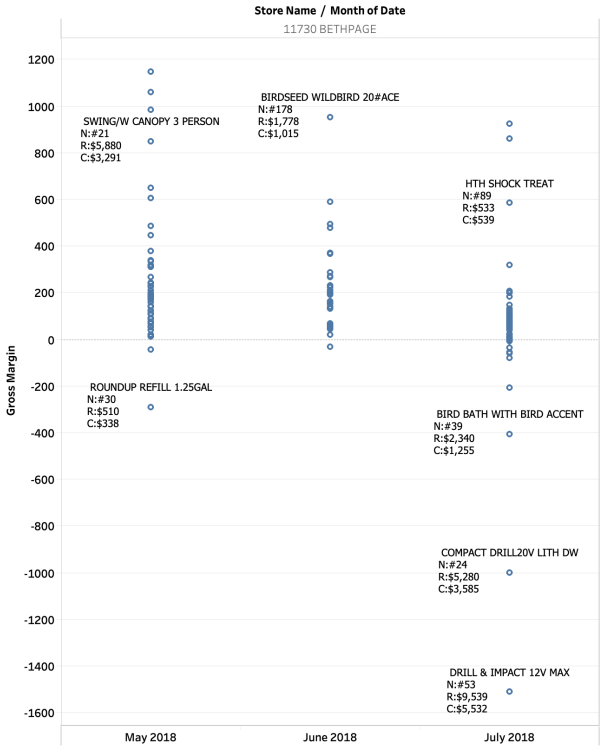


Figure 15: Heavily Discounted items between May & July

Conclusion: The nearby aforementioned stores such as 16038 NORTH BELLMORE, promoted items like "TRAP SPIDER CRICKET PK" & "BIRDSEED WILDBIRD", etc. We suggest that these items should have been promoted at BETHPAGE since those stores fared well during that period. However, we suggest that items from Figure [12] such as "LAWN FOOD 5M", could have been promoted as they are very strongly associated with "WEED & FEED 5M". Similarly, "KEY KWIKSET KW1" is strongly associated with "KEY SCHLAGE SC1". Our intuition is that if the former items are promoted, people would also buy the items associated with it, thus resulting in increase of sales.

7 Future Direction

Due to time constraint, we could not look into other interesting areas such incorporating the weather data to our time series model to perform multivariate analysis such as VARMAX or LSTM. This would have been beneficial in making our sales prediction more accurate and sensitive to weather.

The new store location prediction could be implement using a suitable model with more parameters other than Age & Median Income such as Occupation, Percent of Households with certain income, etc. This would help us pin-point prospective locations with even more confidence.

References

[1] R. D. Lawrence, Personalization of supermarket product recommendations.

[2] S. Li, A gentle introduction on market basket analysis association rules.

[3] D. Waring, Which products should you stock, Harvard Business Review November 2012.

[4] Rob J Hyndman and George Athanasopoulos. *Forecasting:principles and practice.* OTexts, 2018.

[5] Demographic Data **USZipcode**

[6] Weather Data: National Oceanic and Atmospheric Administration **NOAA** Data.

[7] Data Transformation & Time series analysis **SARIMA**

[8] Location prediction strategy for **Starbucks**

[9] Article on **Market Basket Analysis**