CSE 519 MID PROGRESS PROJECT REPORT

Retail Sales Data Analysis

Objective:

To come up with useful insights on the sales trend of Costello's Ace Hardware and predict future sales based on past sales.

Introduction:

Costello's Ace Hardware is a hardware retailer with 30 stores across Long Island. They are planning on launching more stores and are trying to find useful information in their sales to come up with better marketing strategies. The major expectation from the store representative is an investigation of how well the promotional offers performed and to find a correlation between weather conditions and sales on any given day. This analysis will be used to discover patterns in the sales and leverage them to better cater to the needs of customers.

We have started an investigation of the correlation between weather conditions and sales volume on a specific day. We have performed analysis on customers through the ZIP codes present in the datasets to see their distribution. As a part of Sales trend analysis, we have identified patterns of sales in different quarters as well.

Dataset:

- The initial dataset given by the retailer which contains the sales data for the years 2017 and 2018.
- Data on the weather conditions for the state of New York across multiple weather stations along with their geometric coordinates.[1]
- For all the 30 stores, we have taken the coordinates to map it with the closest weather station and customer location.
- Dataset consisting of geographical coordinates for different zip codes in the USA.

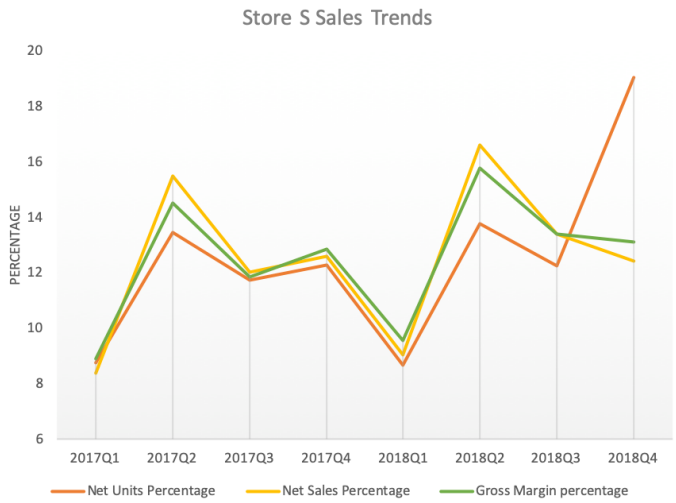
Data Preprocessing:

Some of the steps involved in preprocessing of data in the above datasets:

- To accommodate the different dynamics of the sales in each store we have broken the retail store data based on the store ID and created 30 smaller datasets. This aided in performing analysis faster and more conveniently.
- A lot of numerical columns were filled with commas and other special characters that had to be cleaned and brought into the right format.
- In location analysis, we removed the rows for which the customer zip codes were very far (> 200 miles) from the store since these were outliers that did not add any value.

Preliminary Analysis:

To evaluate the performance of different stores, we compared the Net sale units, Net sales and Gross margin obtained through all the transactions. These columns have a strong correlation (0.95 ~ 0.99) as expected.



The figure shows the quarterwise breakdown of Net sales units, Net sales and Gross margin. We observe a peak during the Quarter 2.
As an anomaly, in store S, net sale units have increased in 2018 Quarter 4 as compared to the Net Sales indicating larger number of items sold at a lesser price.
The departments within each store exhibit similar patterns of sale.

Seasonal and Weather-based Analysis

Seasonal Analysis:

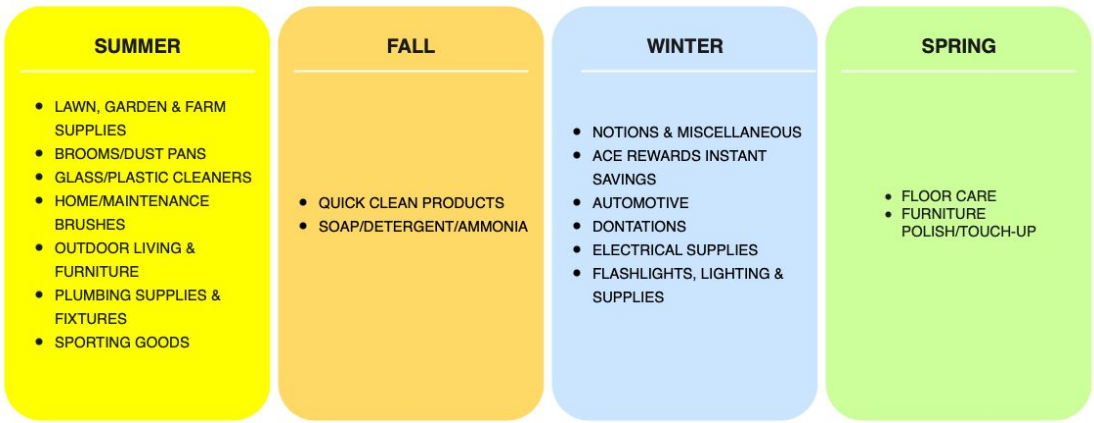
The transactions from the stores are aggregated into different seasons that are Summer (Jun-Aug), Fall (Sep-Nov), Winter(Dec-Feb) and Spring(Mar-May). A month before the season sets in is chosen as the first month of the aggregation group as people prefer to prepare for the upcoming season.

Steps followed:

- For each store, the sum of transactions across different seasons are calculated and the season with maximum share of sales (greater than 30% of the sales) is identified.

Result:

- Following figure shows the different departments classified into seasons. This analysis can be further extended to identify the top selling classes or products in different seasons.

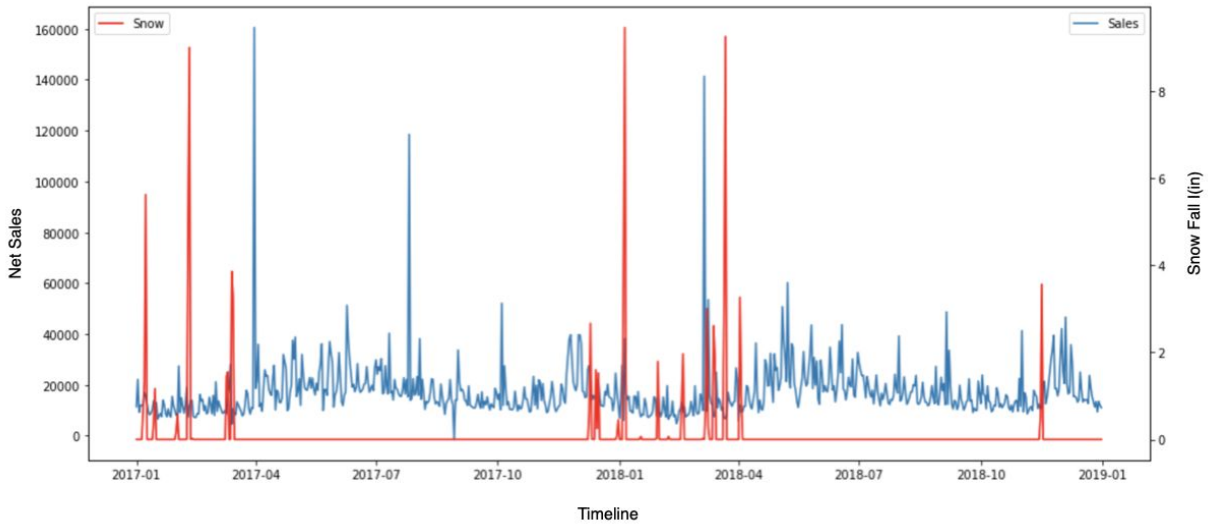


Weather based Analysis:

The aim of this analysis is to recognize the weather's impact embedded in the previous year's sales. Weather not only influences if a customer goes into a store, it also influences the items they place in their basket.

Steps followed:

- For the initial analysis, we have gathered data for 80 different weather stations across New York state from National Centres for Environmental Information.
- The prominent factors chosen from the dataset are Snow, Precipitation and Temperatures levels. These were extracted from 6~10 stations that are closest to the store and correlation with net sales was computed.



Result and future work:

- This exercise did not provide any conclusive evidence of the impact of weather on sales. However, There were a few spikes in sales around drastic weather changes. These will be investigated as a part of the project.
- When it comes to store traffic, "a rising tide lifts all boats." If favorable weather conditions bring customers into stores for a seasonal item, there is a good chance that they may pick up additional items on their shopping trip. Identifying and classifying a product as seasonal/non-seasonal will be taken up during the course of the project.

Market Basket Analysis:

Market Basket Analysis is one of the important techniques used by retailers with a variety of products to pinpoint associations between items. It works by looking for combinations that occur frequently in the sales data. In short, it helps retailers identify relationships between items that people buy.

One of the most effective ways of discovering frequent itemsets is done by leveraging the *Apriori algorithm* which not only finds the frequent itemsets and their relevant association rules

but also quantifies their importance in a given data. It is very important for effective Market Basket Analysis that helps customers in purchasing their items with more ease, increasing the sales of the corresponding store.

Apriori Algorithm breaks down the data and gives us the frequent itemsets but it is with the help of Association rules that we can visualize the importance of these itemsets. Consider the association rule X => Y which implies that whenever customers buy X, most of the time, they also buy Y. This association rule is regarded as an important one if it occurs frequently and is observed over multiple subsets.

To quantify the frequency on an association rule X => Y, we have the below metrics.

- Support : The support of an itemset X, supp(X) is the proportion of transaction in the database in which the item X appears. It signifies the popularity of an itemset.
- Confidence : It signifies the likelihood of an item Y being purchased when item X is purchased.
- Lift : It signifies the likelihood of an item Y being purchased when item X is purchased while taking into account the popularity of itemset Y.
- Conviction : It signifies the ratio of the expected frequency that X occurs without Y.

Steps Followed:

- We leveraged the Apriori algorithm to get the most frequent **Classes** of items distributed across the sales data.
- To be able to use the Apriori algorithm, we performed one hot encoding on the **Class** column and then merged all the rows having the same receipt number which gave us all the classes of products bought together.
- On these frequent itemsets, we quantified the frequency and their importance by using the Lift and Confidence metric.
- From the above table, the most valuable association rules are the ones with the combination of greatest lift and confidence values since lift accounts for consequent's popularity.

Antecedents	Consequents	Support	Confidence	Lift	Leverage	Conviction
LIQUID PAINT PAINTER TOOLS/PAILS/ACCS	PAINT ROLLERS	0.001288	0.657375	32.100395	0.001248	2.858874
SOAP/DETERGENT/AMMONIA BATHROOM CLEANERS	PAINT ROLLERS	0.000758	0.577123	18.169099	0.000716	2.28964
BAG PRODUCTS, FERTILIZERS	ACE COUPON/ CREDIT ITEMS	0.000685	0.568998	9.342261	0.000612	2.178863
GLASS/PLASTIC CLEANERS, SPONGES/CHAMOIS	HOUSEHOLD CLEANERS	0.00061	0.538153	16.942224	0.000574	2.096441
CAULKING GUNS	CAULK/SEALANTS/ GLAZING	0.001122	0.530108	22.828294	0.001073	2.078728
KEY RINGS/ACCESSORIES	KEYS	0.004882	0.524322	14.384435	0.004543	2.025632
PLASTIC PIPE	PLASTIC FITTINGS	0.001555	0.521772	40.94897	0.001517	2.06441
HAND/POWER SANDING ABRASI, KNIVES/SCRAPERS	WALL REPAIR MATERIALS	0.000662	0.514134	48.140807	0.000649	2.036201
CHRISTMAS DECOR - INDOOR, CORD SETS	CHRISTMAS LIGHTS/ ACCESS	0.000803	0.5	24.78752	0.000771	1.959657
STEEL PIPE	IRON PIPE FITTINGS	0.001104	0.498458	192.096239	0.001098	1.988679

Top rows for Market Basket Analysis for Class Codes in Store 6

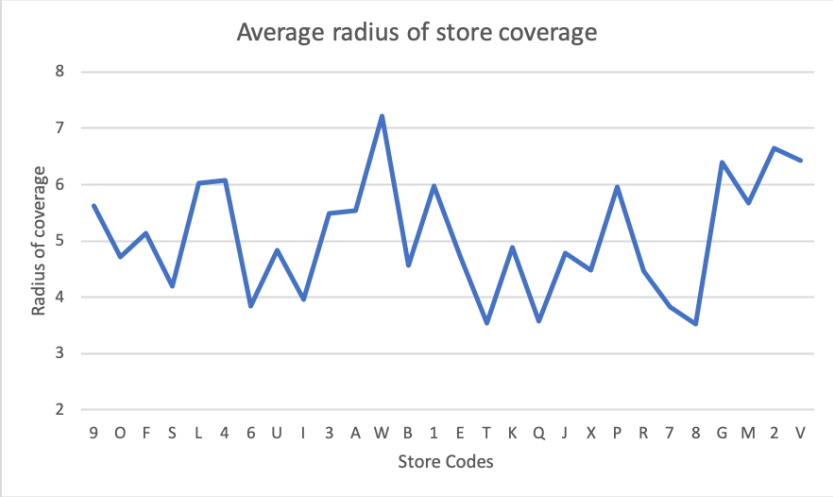
Result and future work:

- The results obtained through this analysis could help the retailer restructure their store to put these frequent classes together which would make it easier for the customers to purchase products.
- We were not able to perform this analysis on ProductID since we have more than 50,000 products. One Hot Encoding over this column was limited by the computational capacity. Going forward, we plan to use Seawolves clusters to run Apriori algorithm on products.

Customer Location Analysis:

Steps followed:

- We gathered the geographical coordinates for all the zip codes throughout the United States and calculated the average of distances between the customer location and store location for each store.
- The distance is calculated using Haversine formula.



Result and future work:

- The zip codes with large customer count and greater distance from the store can be identified and considered as a viable location to launch a new store. This would ease shopping for customers and maximize sales.

References

[1] https://www.noaa.gov/
[2] https://pdfs.semanticscholar.org/4dcf/50517e96401800a32cdd1c39392d8e4ccda8.pdf
[3] https://ieeexplore.ieee.org/document/7231468