RETAIL SALES DATA ANALYSIS

Department of Computer Science, Stony Brook University

***Abstract* -- The purpose of this project is to give the ability to effectively track customer actions, like their purchases and foot traffic in stores to business owners. Without this core insight, retailers are not as well equipped to make future decisions—therefore, risking the consequences of not being informed in their business planning efforts.**

I. INTRO: WHAT IS RETAIL DATA ANALYSIS?

Consumers' expectations of digitally engaging, personalized and customer-centric experiences are constantly being heightened by non-traditional competitors. Retail data analysis is exactly what the name suggests. It's an analysis of everything in retail sales business, from sales and inventory to customer data. It gives the ability to effectively track customer actions, like their purchases and foot traffic in stores. Lacking these capabilities will lead to missed expectations, unsatisfied customers and ultimately, lost sales. Data allows retailers to make better operational decisions by delivering real-time, relevant insights into their business management, inventory and, of course, customers. Without this core insight, retailers are not as well equipped to make future decisions—therefore, risking the consequences of not being informed in their business planning efforts.

II. BACKGROUND RESEARCH

Research shows that small businesses have yet to fully take advantage of the technologies on the market. Performing sales trend analysis gives valuable insight into the inner-workings of business. Merchants can use their data to make informed decisions like when to raise or lower prices on products. There has been some research done in the past to effectively analyse sales trends that can have a major impact on how to run the retail sales business:

In paper[1], the researchers collected sales data from a retailer of branded women's business wear in the Seoul-Kyunggi area in South Korea. Along with the sales data for seasonal basic styles, corresponding daily and weekly average temperature data were collected and evaluated. The analysis for the study was drawn using descriptive statistics including graphical evaluations, correlation analysis and paired samples t-test. Results of this study provide strong evidence that fluctuations in temperature can impact sales of seasonal products. During sales periods when drastic temperature changes occurred, more seasonal products were sold. "US aggregate retail sales have strong trend and seasonal patterns. How to best model and forecast these patterns has been a long-standing issue in time-series analysis" [4]. As stated in this paper, it would be interesting we feel to analyse the impact of weather on sales positively or negatively.

The Paper [2] talks about the rise of antibiotic consumption and the increase in use of last-resort antibiotic drugs which raises serious concerns for public health. Appropriate use of antibiotics in developing countries should be encouraged. However, to prevent a striking rise in resistance in low-income and middle-income countries with large populations and to preserve antibiotic efficacy worldwide, programmes that promote rational use through coordinated efforts by the international community should be a priority. This paper takes into account the average incomes of the targeted population which can act as a good feature for the modelling. Similarly, we found data related to average incomes, employment etc for each zip code (e.g. Stony Brook -11790)[7] which impacts the Purchasing Power of an individual and thus will help in better understanding of the retail sales of the stores at various locations depending upon the demographic trends of the location.

[3] is a US Patent by Google which gives us insights that can lead to better decisions on new product launches, sampling, merchandising, assortment, distribution, and other sales and marketing priorities. The techniques stated in the Patent allows the user to manipulate and extract information which is specific to the user's particular needs.

Paper [5] discusses the problem faced while modelling marketing analysis that is to select key variables from a large number of possibilities. In analysing high-dimensional marketing data, the problem faced is that valuable predictors of consumer behaviour are often hidden in a large number of useless noisy variables. When the dimensionality increases with the integration of intra- and inter-categorical information, the number of unreliable predictors which are correlated with valuable ones also increases rapidly. This paper develops a four steps methodological framework to overcome the problem. The method consists of the identification of

potentially influential categories, the building of the explanatory variable space, variable selection and model estimation by a multistage LASSO regression, and the use of a rolling scheme to generate forecasts.

III. OBJECTIVE: WHAT ARE WE ANALYSING IN THE COSTELLO'S DATA?

What we have is a very large dataset (~18million x 39 columns/features) in csv format, what can be done with a spreadsheet software on a machine is limited to simple analysis. To overcome this limitation and enlarge the spectrum, this project seeks to bring data science to enable analysis involving high dimensional data that is hard and even impossible to perform using conventional spreadsheet software.

We aim to perform the following analysis/predictions with the data:

- **Stock Market correlation analysis** – Analysing the impact of stock market/economy ups and downs on the net sales.
- We endeavour to find the various customer segments using the data obtained using the well-known RFM model which helps analyse better marketing campaigns and improves CRM and customer's behaviours. Various segments based on R and F are made to categorise the customer behaviour.
- **Products bought together** – Analysing which products are often bought together so that they can be placed adjacent on the shelves in order to increase the combined sales of both these products. E.g. Whether Snow shovels and salt is bought together.? We aim to provide recommended actions based on the insights drawn from the data analysis, with priority given to the results based on terms of net sales.
- **Future Sales Predictions**: **Time series modelling** of sales - Based on the training and modelling, future predictions could be brought up to know which items to be stocked in higher quantity than others and vice versa.
- Some other trivial but **Interesting insights** : Future scope of Project and the various directions this analysis can lead the Costello market reach.!
- Analysing **Key performance indicators** like Net Sales Units, Net Sales, Cost, Gross Margin, Gross Margin % and sales growth will tell the story of store performance and help make profitable decisions. Understanding what customers aren't buying: poor displays or not enough sales reps could be among the many reasons why customers aren't buying what is being sold.
- **Product embeddings** - Build distributed representations of products (embeddings), using the network of co-purchased items. Doing these embeddings would provide insights into which products are similar or are bought together or not, so that the particular products could be targeted for improvement in sales by changing their shelf positions or otherwise promoting them i.e. **Promotional recommendations** in a tailor made fashion for usual Shoppers to make their user experience better.

IV. DATASETS

a. DATA DESCRIPTION:

For the project, we are using the following datasets:

Costello's Ace data (Given for 2015-16 and 2017-18)
Stock Market Data
Other Hardware Stores Geolocation Data

Costello's Ace data:
The dataset contains 17,328,044 rows/entries x 39 columns / features in csv format, out of which some important features are listed below:

Table 1

Date and Time	Date/Time of Transaction
Customer Number	Customer Account Number
Receipt Number	First letter shows how old are the transactions Letter number at last is store number Lower case letter indicates new stores Receipt number itself is not unique but the (receipt number + store) would be a unique key.
Store Name	Name of the store
Item Description	Description of item
Net Sales	Unit * Price of Product
Net Sales Unit	# of units bought
UPC	Scanned UPC barcode of product Different UPC for same products – different due to different sources of procurement QF (quick find) - entered item from directory on screen
ZipCode	Represents customers residence zip code if customer is registered, otherwise store's zip code where it is situated.
Product Categorization	Overall products are divided into 3 tier system of hierarchy and there is always 1 to 1 mapping

	i.e. no product lies in greater than 1 category.
Dept Name	Name of the dept to which the product belongs.
Dept code	Total 18 departments Dept assigned (products grouped)
Product Class	less specific than department name
Fineline code	class broken down further
Loyalty ID	Account number that is unique to customer

Stock Market Data

The stock market data downloaded from Yahoo finance is a day wise S&P 500 daily closing rate for the dates corresponding to the Costello's Sale data.

Other Hardware Stores Geolocation Data

Location and Zip Code of other competitive stores situated at Long Island.

b. DATA PREPROCESSING :

- Removing Null values from DataSet
- Common cleaning task – Many numerical value columns are of type object which are to be converted to Numerical(int, float etc) format
- Spaces between Column names is replaced with _ (underscore) so as to make it Python compatible.
- Date and transaction time records to be converted to a format that can be processed by Python (Format – DateTime MM:DD:YY HH:MM:SS).
- Customer_Number with *5 values (Cash transactions) have been replaced with 1 and *6(10% semi-discount) values with 2 and other String values into uniform integer format for processing further.
- Scanned UPC – other, auto, quick find etc have been assigned unique integer values.
- Promo/Discount and MIP Promo Id – NaN values have been assigned 0 value.
- Actual-Retail – Bracketed values are representing loss making items i.e. Selling Price is lower than Cost Price. This is actually the margin on a particular product that is being made by the company. Bracketed values are converted to -ve values for uniformity. $ sign is also being removed to convert into DataType float.

V. EXPLORATORY DATA ANALYSIS:

Plotting the Line graph for seeing which department has higher sales and which are on the lower end of the spectrum so as focus can be shifted to these in Fig 1.

In Fig 2 graph, we see that maximum sales occur during winter and summer breaks. We can infer from this that it might be useful to stock items during these months.

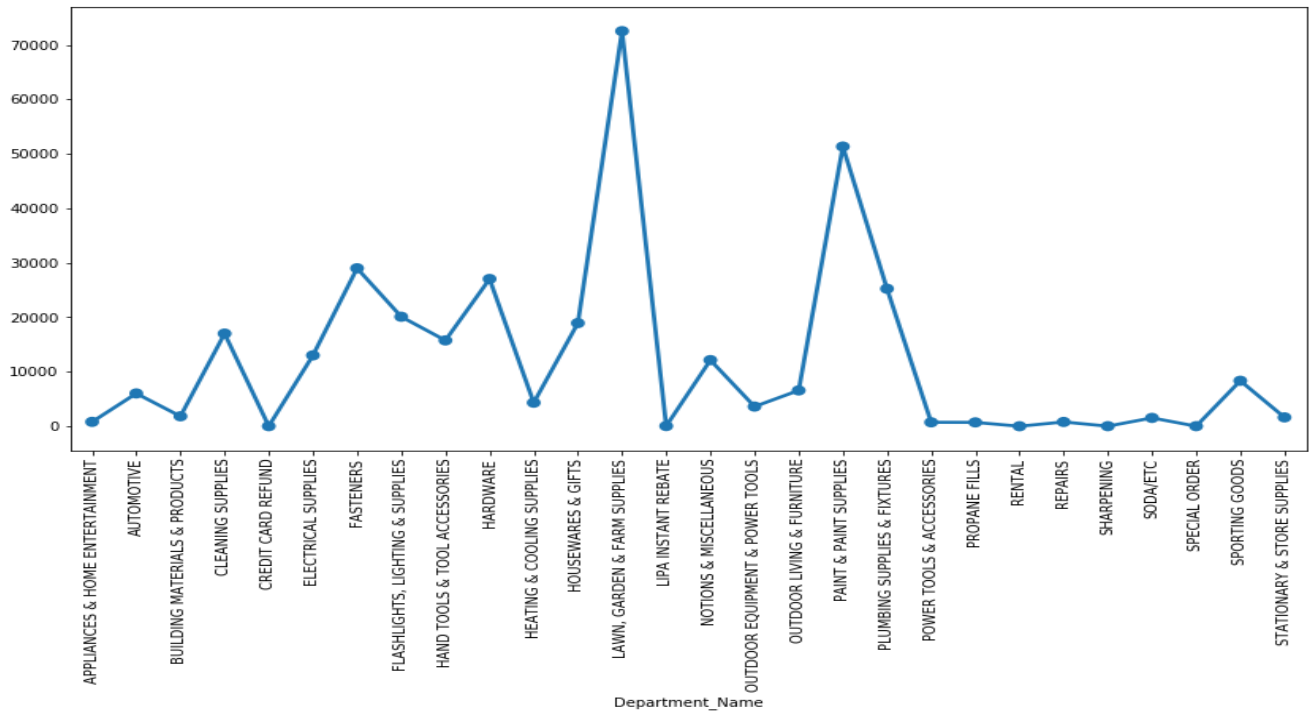

Fig 1. Department wise sales

Further analysis would be required to find which products are in demand during these periods. These can be stocked in advance to help increase the revenue.

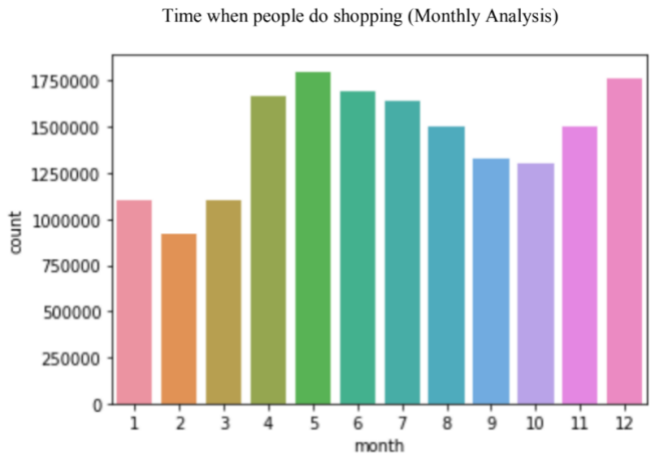

Fig 2. Monthly Sale

In Fig 3 graph, we see that maximum sales occur on Saturdays.

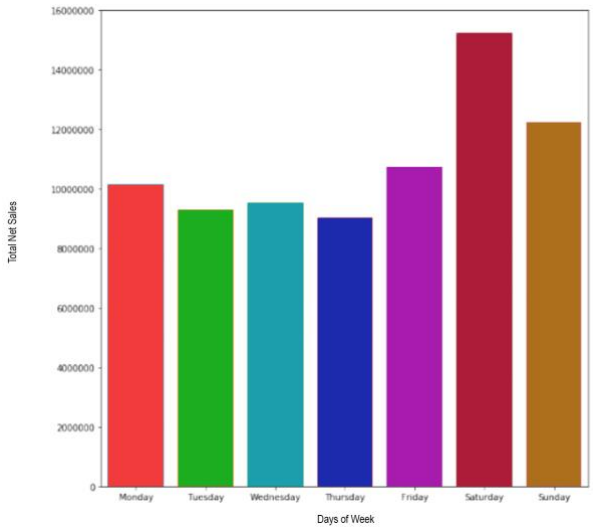

Fig 3. Days of the Week Sales

Fig 4 gives us the Store wise total sales, which clearly shows Store # 14252 Island Park is performing the best according to the given data. From this, we need to find reasons for other stores performing badly .

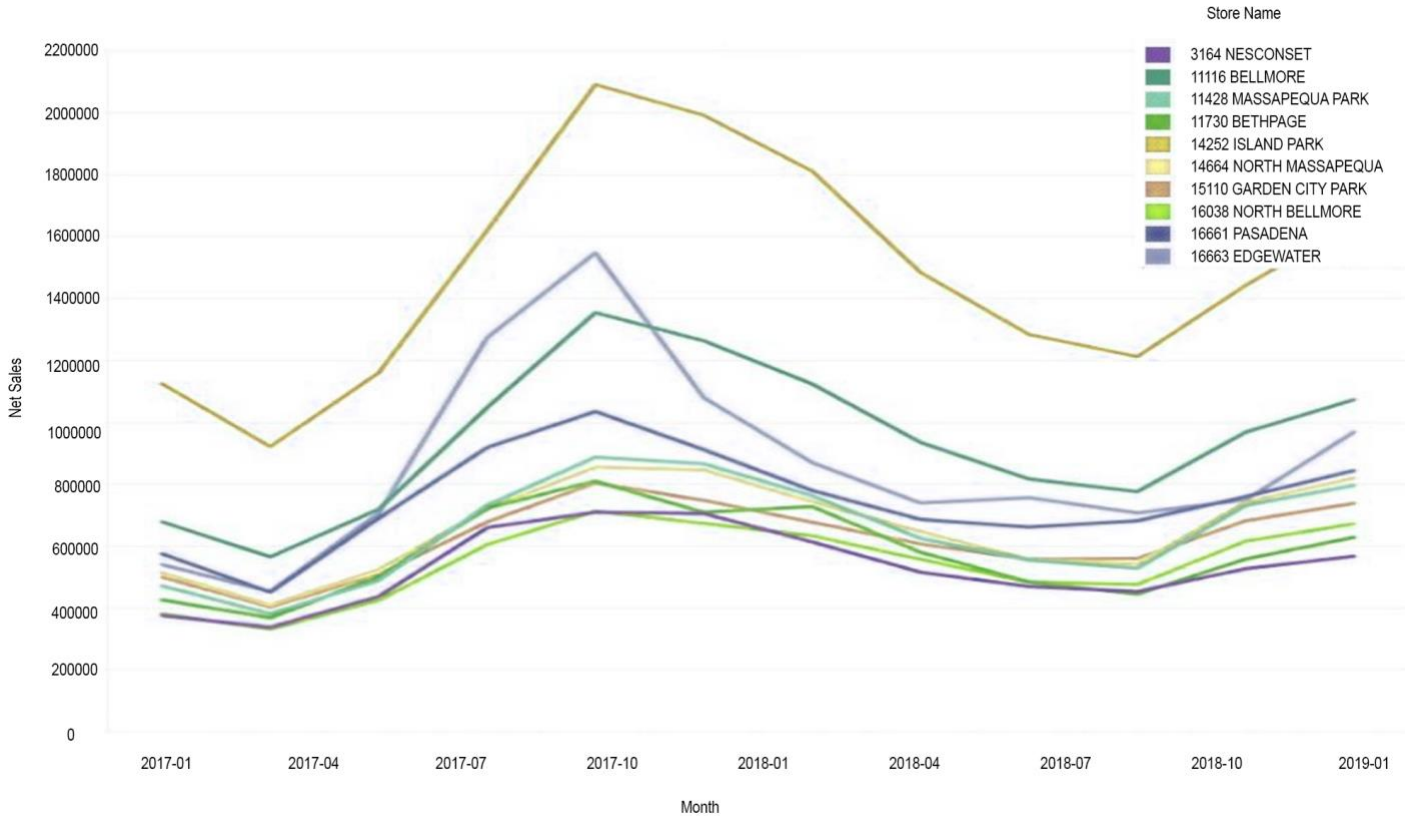

Fig 4. Store wise Total Sale

VI. IMPLEMENTATION:

1. STOCK MARKET ANALYSIS

The S&P 500, or just the S&P, is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States. It is one of the most commonly followed equity indices, and many consider it to be one of the best representations of the U.S. stock market.

Our target here is to compare the economic conditions prevailing at the time of sale which would be a good estimate to know the buying trends of people according to their disposable income.

In times of boom, the disposable income increases and thus should reflect in the increase in sales. This is the expected trend but it may vary according to some other extraneous factors not linked to the economy. So, the business leaders can decide what information to be used and when.
The data is analyzed with reference to the Yahoo Finance's stock market price of the S&P Index for the time period of Jan 2017 to Dec 2019.The total sales is plotted against the average closing price for that time period to see the correlation between the economic conditions w.r.t to the stores' sales.

In Fig 5, it is seen although the trend is not unidirectional which can be explained by the role of extraneous factors that have an impact on the sales:

- Holiday season
- Salary Day/Starting days of every month
- Sale Period
- Promotions by the manufacturers
- Season – Summer/Winter
- Other Natural factors
- Certain viral trends of products which may increase/decrease the sales for some product categories.
- Geological location

Some of the correlations are positive as seen in the later part of the plot where there seems to be some kind of relation that exists between stock prices and sales.

Future Work : This idea may further be explored to get the real time notifications on the stock prices of the manufacturers whose products are being sold by Costello to somewhat hedge the price variations and thus control the losses arising out of volatility of the market.

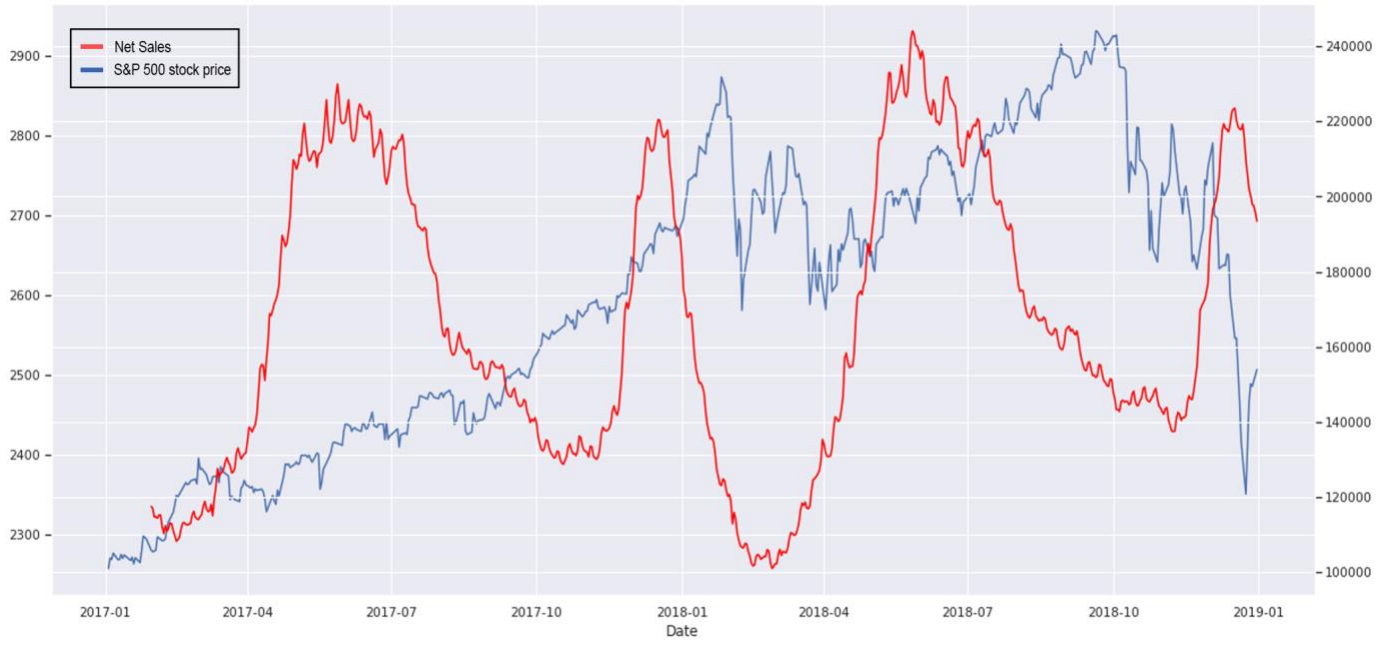

Fig 5. S&P 500 stock price vs Total Sale

2. RFM (Recency, Frequency, Monetary)

To get the RFM score of a customer, we need to first calculate the R, F and M scores on a scale from 1 (worst) to 5 (best).

- calculate Recency = number of days since last purchase
- calculate Frequency = number of purchases during the studied period (usually one year)
- calculate Monetary = total amount of purchases made during the studied period
- find quintiles for each of these dimensions
- give a grade to each dimension depending in which quintiles it stands
- combine R, F and M scores to get the RFM score

- map RF scores to segments

Calculate the R, F and M scores

- At this point, we have the values for Recency, Frequency and Monetary parameters. Each customer will get a note between 1 and 5 for each parameter.
- We can do this by setting ranges based on expected behaviour.
- For example, to rate Recency, we could use this scale:

Table 2

Segment	Description
Champions	Bought recently, buy often and spend the most
Loyal Customers	Buy on a regular basis. Responsive to promotions.
Potential Loyalists	Recent customers with average frequency.
Recent Customers	Bought most recently, but not often.
Promising	Recent shoppers, but haven't spent much.
Customers needing attention	Above average recency, frequency and monetary values. May not have bought very recently though.
About to sleep	Below average recency and frequency. Will lose them if not reactivated.
At Risk	Purchased often but a long time ago. Need to bring them back!
Cant Lose them	Used to purchase frequently but haven't returned for a long time.
Hibernating	Last purchase was long back and low number of orders. May be lost.

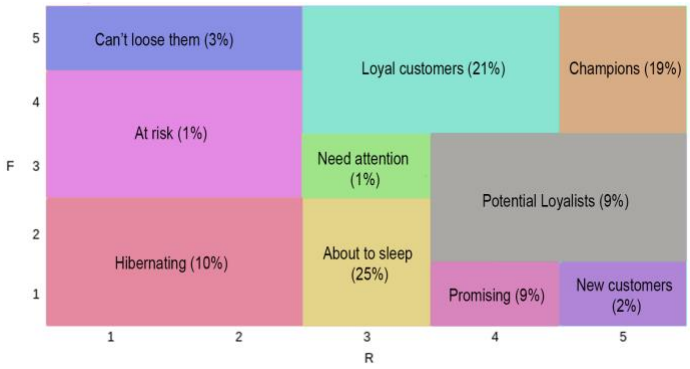

Fig 6

1: 0-30 days
2: 31-60 days
3: 61-90 days
4: 91-180 days
5: 181-365 days

We could also use quintiles. Each quintiles contains 20% of the population. Using quintiles is more flexible as the ranges will adapt to the data and would work across different industries or if there's any change in expected customer behaviour.

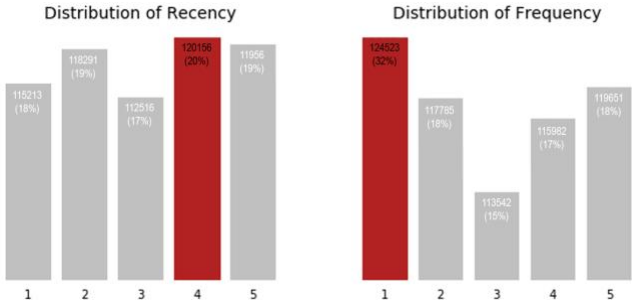

Fig 7

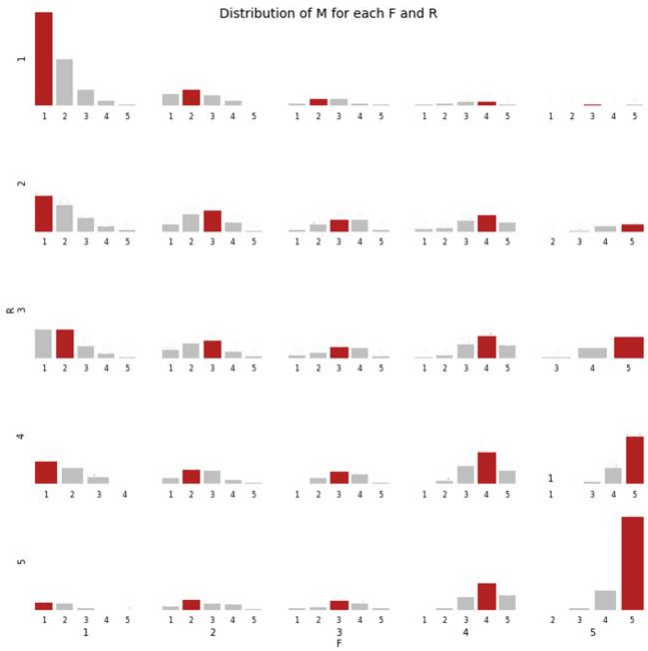

Fig 8

We can see in Fig 7 and 8 that if recency seems almost evenly distributed, almost half of the customers don't purchase very often (50% of customers have a frequency of 1 or 2). When looking at the monetary value, we see that the customers spending the most are those with the highest activity (R and F of 4-5).

In Fig 6, we have a lot of customers who don't buy frequently from Costello (25% are about to Sleep), however, 40% of customers are either Champions or Loyal Customers.

3. New Store Recommendation

In Fig Blue markers are those Costello stores whose sales are on the high end of the spectrum, black ones are performing not quite good and the red ones are the competitors. So, we can deduce and recommend the following:

- Stores(Black) having high competition in their affinity are performing badly. E.g. 15348 Rocky Point, 15280 East Islip, 16324 Brooklyn, 5144 Deer Park Avenue.
- Areas with less number of store (either Costello's or other hardware stores) are hot points for new locations. E.g N-E of the Long Island as seen in the Fig 9.
- Black points should try to bring in more customers by having more promotions compared to other stores. Parallelly, they can analyse the items being sold by the competitors and accordingly adjust price so as to be more competitive.
- Blue points are already performing above average and thus needs no special attention such as 14252 Island Park, 16663 Edgewater, 11116 Bellmore.

Fig 9 Costello vs Competitors

4. MBA (Market Basket Analysis)

When considering physical store, if similar or products which are frequently bought together are placed near each other, it helps to boost sales. It is also helpful for customers to easily find the products they need. From this motivation, we intend to find the sets of items which are frequently bought together. We used Apriori algorithm - an Association Rule Mining based technique for this purpose.

Since different stores have same receipt numbers in our Dataset, we combined "Store #" with "Receipt Number" to get unique receipt number for each transaction. Then we "grouped by" the rows according to newly updated receipt number, thus got the list of items for each transaction. Upon these transactions, where each transaction has one or more items, we applied Apriori algorithm to find the association rules. We used three following metrics for this algorithm:

Support: This rule gives the popularity in terms of its frequency of occurrence in the dataset, as measured by the ratio of transactions in which a product appears. As we have a very huge dataset with varied products, we selected a lower value of threshold support which is 0.0001.

Confidence: Association rule states that if item Y is also purchased when item X is purchased shown as {X -> Y}. This is calculated by the ratio of transactions with item X, in which item Y also is seen. These rules with high ratios represent strong relation/association. We selected ratios >0.3.

Lift: This rule states that if item Y is purchased when item X has already been purchased, while taking in control for how popular item Y is. To put it simply, many people bought X and Y alongside each other, therefore the confidence ratio would be high. However if also many customers bought only Y , the lift value will be low. Lift of 1 implies no association between items and items are independent of each other. A lift value greater than 1 means that item Y have more chances to be bought if

item X is bought, while a value less than 1 means that item Y is less chance to be bought if item X is bought.

	Antecedants	Consequents	Antecedent support	Consequent support	Support	Confidence	Lift	Leverage	Conviction
0	(BIRDSEED WILDBIRD 20#ACE, CMN Donations)	(FASTENERS)	0.29406	0.56255	0.15235	0.84251	1.32564	0.03012	1.84528
1	(TRAP SPIDER & CRICKET PK)	(FASTENERS)	0.38824	0.56385	0.14985	0.82564	1.29841	0.02954	1.75420
2	(SPORTING GOODS, PAINT & PAINT SUPPLIES)	(LAWN, GARDEN & FARM SUPPLIES)	0.20153	0.56248	0.15236	0.80526	1.27598	0.02958	1.71257
3	(FLASHLIGHTS, LIGHTING & SUPPLIES)	(PAINT & PAINT SUPPLIES)	0.18542	0.56489	0.15698	0.80198	1.23680	0.02831	1.67234
4	(FASTENERS)	(PAINT & PAINT SUPPLIES)	0.19326	0.56126	0.14854	0.79546	1.12679	0.02563	1.41238
5	(PLUMBING SUPPLIES & FIXTURES)	(FASTENERS)	0.19562	0.56845	0.15148	0.78452	1.12254	0.02298	1.25845

Fig 10 Market Basket Analysis

We ran the MBA model on departments rather than fineline code or class code to get better results. After running the MBA model on the department code, we observe that if the antecedent is (BIRDSEED WILDBIRD 20#ACE, CMN Donations) then the consequent are (FASTENERS). It is very likely that if the customer purchases anything from the above antecedent, they purchase item from the consequent as well as they high confidence and lift value.

5. ARIMA

AutoRegressive Integrated Moving Average (ARIMA) is most commonly used for forecasting in time series analysis. AutoRegressive means the model using a linear combination of past values of the variable. The p stands for the order of the model and "t is the white noise. Moving Average model use the linear combination of past errors. ARIMA combines these two models and requires the time series stationary.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

The three parameters in notation of ARIMA account for seasonality, trend, and noise in data.
The data was filtered for item. Bayesian optimisation was performed to find the optimal parameters for obtaining the best results of the model. After performing optimisation, the model was trained on data having date range from Jan-2017 to Dec-2018.

In Fig 11, we visualised the data using Time Series Decomposition that broke down the time-series into trend, seasonality and noise. The trend of sale is increasing and there is a pattern in seasonality. The sale increases from the month of June to August as well as from Nov to Dec every year.

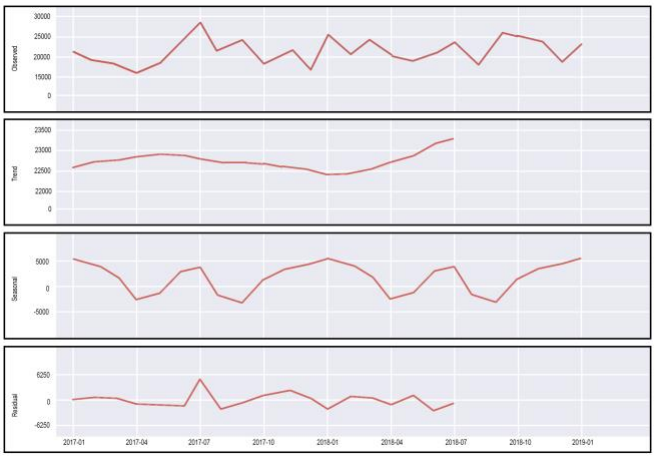

Fig 11

Fig 12 shows the comparison of predicted sales to net sales of the Costello's Ace. This gives us better understanding of the correctness of forecast. It can be observed that our forecast aligns well with the real value of the net sales showing an upward trend that increases and also captured the seasonality in the month of June to July and Nov to Dec.
The RMS (Root Mean Square) error of a forecast is 20.72.

Fig 12

VII. SUMMARY/RECOMMENDATIONS

- For stock Market Analysis, Some of the correlations are positive as seen in the later part of the plot where there seems to be some kind of relation that exists between stock prices and sales.
- Almost half of the customers don't purchase very often (50% of customers have a frequency of 1 or 2).
- Areas with less number of store (either Costello's or other hardware stores) are hot points for new locations. E.g N-E of the Long Island
- the customer purchases anything from the above antecedent, they purchase item from the consequent as well as they high confidence and lift value. E.g. the antecedent is (BIRDSEED WILDBIRD 20#ACE, CMN Donations) then the consequent are (FASTENERS).
- ARIMA forecast aligns well with the real value of the net sales showing an upward trend that increases and also captured the seasonality in the month of June to July and Nov to Dec.

VIII. MOVING FORWARD:

After looking for trends or patterns in sales data, one can determine both opportunities and potential problems, and can track if a particular product is increasing or decreasing in sales.

If it's declining, it may be due to some other alternative product or lower price available for that which is amounting to the decreasing sales. This can be observed by comparing the highs and lows of the price of the product and compare them with other sources where such product/alternative is being sold (most probable place could be an e-Commerce website).

Setting prices to the threshold values would be the key to retain customers in the long run even if that means loss on a few items. This analysis needs various other datasets and would need data to be scraped from these websites periodically to have the real-time data pipeline which feeds it to the analysis engine and provide the desired results.

IX. REFERENCES

1. Youngjin Bahng, Doris H. Kincade "The relationship between temperature and sales" International Journal of Retail & Distribution Management", 2012
2. Thomas PVan Boeckel PhD, Sumanth Gandra MD, Ashvin Ashok MPP, Quentin Caudron PhD , Prof Bryan T Grenfell PhD, Prof Simon A Levin PhD, Prof Ramanan Laxminarayan: "Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data",2014.
3. Paul Springfield, Edwin Blake, David Stern : "Method for performing retail sales analysis", Patent US8214246B2 United States
4. Ilan Alon ,Min Qi , Robert J. Sadowski , "Forecasting aggregate retail sales :: a comparison of artificial neural networks and traditional methods", Journal of Retailing and Consumer Services , 2001
5. Shaohui Ma, Robert Fildes, Tao Huang, "Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information", European Journal of Operational Research, 2016
6. Retail Trade Analysis Report Fiscal Year 2018 of Winnebago County by Iowa State University Department of Economics.
7. Website, city-data.com ("http://www.city-data.com/city/Stony-Brook-New-York.html")
8. Customer segmentation by using rfm model and clustering methods: a case study in retail industry
9. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, LlionJones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems,pages 5998–6008, 2017.
10. Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. arXiv preprint arXiv:1704.02971, 2017.
11. LUO Jian, HONG Tao, and YUE Meng. Real-time anomaly detection for very short-term load forecasting. Journal of Modern Power Systems and Clean Energy, 6(2):235–243, 2018.