Retail Sales Data Analysis

Extract more GMV out of customer's Pocket, Predicting Sales
(Learn the Causality and Ways to improve/increase Sales)

Computer Science Department
Stony Brook University
United States

Retail Sales Data Analysis

Abstract*— *The Costello's is a chain of retail hardware stores spread across 26 locations across Long Island, Brooklyn and New Jersey, more than 90% stores are on Long Island. The Costello's has graciously made its market basket data available to us for analysis. This data consists of records of several million sales, over several years of tens of thousands of different products.

Central Objective: *Our main focus is to understand customers behaviour and ways to extract more money out of a customer's pocket and to build a recommendation model that predicts customer's willingness to pay, customize products for consumers and forecasting sales.*

I. INTRODUCTION

The store owners generally face challenges while using the enormous amounts of data at the tip of their fingers to boost their business. Picture this: You and your competitor both operate in the same market, have about the same number of customers and similar product price points. You also have the same amount of money to spend on advertising. So how do you end up on top? The answer lies in using data-driven insights to make smarter decisions. Without the ability to track and analyze your decisions as well as your customers' purchases and behaviors, business risks are falling behind. This is why retail data analysis is a powerful tool for business. By prioritizing retail analytics basics that focus on the process and not exclusively on data itself, the stores can uncover stronger insights and be in a more advantageous position to succeed when attempting to predict business and consumer needs. The transaction data they possess can give them a competitive advantage over their peers in the market. Big companies representing diverse trade spheres seek to make use of the beneficial value of the data. In order to to take profitable decisions concerning business, data has become of primal importance. In addition to this, a thorough and comprehensive analysis of a vast amount of data helps in influencing or rather manipulating the customers' decisions. Numerous flows of information, along with channels of communication, are used for this purpose. A small, local chain of retail stores, the Costello's has graciously made its market basket data available to us for analysis. The Costello retail hardware stores founded by Vincent Costello in around 1976 are spread across 26 locations ranging across Long Island, Brooklyn and New Jersey. The Costello stores want to better understand their customers and business processes, in an effort to boost profit, reduce expenditure, gain competitive advantages and open more stores across the state. For this purpose, the Costello's has provided us with the retail transaction data for the past 4 years. We have been entrusted with the task of drawing valuable inferences from the data which can boost the sales and suggest effective strategies to raise the revenue generated. This is basically a problem of retail sales data analysis. The main aim of this data analysis is to extract useful insights from the data in order to assist the stores grow revenue and business' profitability, increase the return on investment while enhancing customer-centric experiences.

II. BACKGROUND

The sphere of the retail develops rapidly. The retailers manage to analyze data and develop a peculiar psychological portrait of a customer to learn his or her sore points. Thereby, a customer tends to be easily influenced by the tricks developed by the retailers. There are three main aims of any retail business - attract new customers, retaining existing customers, and sell more to each customer. This is only possible when the retail stores offer the products to the customers at the right prices. We did an extensive research on the prevalent systems and methods implemented in the past for predictive sales, market basket analysis, warehouse stocking, customer sentiment analysis and customer segmentation.

Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It analyses combinations of items that occur together frequently in transactions and helps retailers identify relationships between the items that people buy [1].

Association Rules are widely used to analyze retail basket or transaction data, and are intended to identify strong rules discovered in transaction data using measures of interest, based on the concept of strong rules. The association rule mining is an important area of research in Data mining which identifies useful associations or relationships among big set of data items. It can be decomposed into two sub problems, mining large item set (i.e. frequent item sets) and the generation of association rules. Two statistical measures that control the process of association rule mining are **support** and confidence. The two main algorithms for Association rule mining are **AIS Algorithm** and **Apriori algorithm** [2].

Interestingly, we can explore the arena of RFM (Recency, Frequency, Monetary) analysis. is a proven marketing model for behavior based customer segmentation. It groups customers based on their transaction history – how recently, how often and how much did they buy. RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services. One can do customer Segmentation by using **RFM model and Clustering methods**. RFM (recency, frequency and monetary) values have been used for detecting similarities and differences among customers,

predicting their behaviors, proposing better options and opportunities to customers for customer-company engagement. It helps in customer segmentation to identify customers valuable to the company and need promotional activities, etc. Clustering, which one of the tasks of data mining has been used to group people, objects, etc .
User experience and product strategy are important aspects which should be kept in mind while improving customer experiences. A Recommender System is capable of predicting the future preference of a set of items for a user, and recommend the top items. In practice four types of recommendation systems are used for data analysis: Content-based filtering, Collaborative filtering, Rule-based and Hybrid approaches. There is an interesting paper which leverages the category information of the transaction and very elegantly puts forward how recommendations can be given based on purchase patterns [3]. Sequential pattern analysis (SPA) is used in this approach, frequently occurring patterns are extracted from all sub sequences of the given set of sequences. In **contest-based filtering systems** an item profile consisting of a set of features is extracted for each item and user profiles are generated based on features of items that are purchased by each user. Then similarity scores between user profiles and item profiles are calculated to finally recommend items with top similarity scores [4]. **Collaborative filtering systems** are built on the assumption that a good way to find interesting content is to find other people who have similar interests and then recommend items rated highly by these similar users [5]. **Rule-based approach** is a simple but popular way of recommendations. Rules are usually derived from the database of previous transactions. Localized association rules among items that are purchased together, which are helpful for target marketing .
Following the background research comes the most important part which is - Exploratory data analysis. The most important aspect of this data analysis is - **Are we asking the right questions?**

- Which Costello stores are generating high revenues? What are the reasons for it?
- Do we have a loyal customer base? Are those customers more swayed by exclusive deals and discounts?
- What parameters affect our sales? Is there a geographical pattern in the data? Does the proximity of the store affect sales? Should the Costello's open a new store? If yes, where and why?

Based on the above questions and brainstorming further, we came forward with a set of following hypotheses which would be pursue for our further analysis, analyzing the essential patterns involved and building models for better prediction on the data. Broadly, we have based our approach from three angles **Customer Centric**, **Store Centric** and **Association of products**, upon which we have tried to build and test a series of below hypothesis:

- Does increase in Price of an item reduce our transaction count?
- What is the behavior of our regular/repeat customers compared to the non-regular/non-repeat customers?
- Cashier Sentiment analysis: Is there any Psychological trend in the way Cashier scans the items?
- What are the findings based on different stores? For example, what are the reasons for a particular store performing better than others?
- What are neck to neck comparison metrics, probably based on zip code, time period, type of products, pricing of the individual product or grouped in combo?
- Is there any Geographical Pattern in our data? Does proximity of stores matter?

III. EXPERIMETAL SETUP

The Costello dataset ranges over a period of 4 years from 2015-2018. The dataset consists of 39 columns. We will undergo **Data Prepossessing** and define **Methods** that we have used for our Analysis and prediction.

A. Data Preprocessing

The three columns with the highest percentage of missing values are:

- Return Code: Return codes specify if the product is either returned or defective. Return codes help investigating the issue why the product has been returned. This column has a significant number of NaN values for the transactions that are Return/Defective.
- MIP Promo Code: MIP stands for Moore's Ideal Products. It is an equipment manufacturing company based in the United States. This column has around 85 percent of the missing values. Since, it is an optional value per transaction, it can be believed that these values aren't missing at random.
- Promo/Discount: Promo/Discount specifies the discount code specified during the purchase of that particular product. It has around 71.83 percent Nan values. It has a total of 6 different values, namely, S - 15.12* - 12.32Q - 0.64D - 0.06P - 0.0002

We can infer from the above data that NULL or NaN values dominate these features. But, it can't be said that they do not mean anything. There might be a possibility that the customer did not have a promo code and hence that column was left blank. Thus, we need to treat the null values as a feature of their own and can replace them with a different code of its own. Therefore, we had to do the necessary imputations.

1. *External Datasources*

- GEOGRAPHIC VARIABLES: : We have extracted the US Zipcodes for Latitude/Longitude. we have mapped the Latitude, Longitude across the Zipcode of Stores and Customers address
- STORE LOCATION: : We have extracted the US Zipcodes of Stores
- EXTERNAL VARIBALES: : We have also tried to extract Population,Avg Household Size, GDP, Family Households

Portfolio of Costello Retail Stores		
Year	2017	2018
Total Unique Stores	29	31
% New Stores (1/2/3 Store #)	16%	17%
% Old Stores (4/5/6 Store #)	39%	40%
% Very Old Stores (7/8/9 Store #)	45%	43%
Total Sales (in Mn)	43.85	48.12
Unique Customer Base (in K)	160	197
%Senior Discount Accounts	**2.08%**	**1.83%**
#Revenue (in Mn USD)	36.60	40.22
Total Discounts Offered on Retail (in %)	12.40%	11.87%
Net Yearly Profit %	11.97%	11.47%
Top Selling Store (**14252 ISLAND PARK**) -%Share in Sales	14.53%	13.85%
Maximum Sold Item - **FASTENERS**	511K	608K
Costliest Item - **WESLEY DINING SET 7PC**	999$	999$

FIG. 1: **Portfolio of Costello Retail Stores.**

- Figure 1 defines the summary of statistics in the Costello Retail Stores.
- Number of stores have increased from 2017 to 2018.
- Total sales in millions have increased from 2017 to 2018.
- The stores have attracted new customers in 2018.
- Senior Discount Accounts have decreased in 2018.
- Revenue increased to 40.22 million dollars in 2018.
- '14252 Island Park' got reduction in shares sold in 2018.
- 'Fasteners' is the maximum sold item in 2017 and 2018.
- The costliest item in both the years is 'Wesley Dining Set'

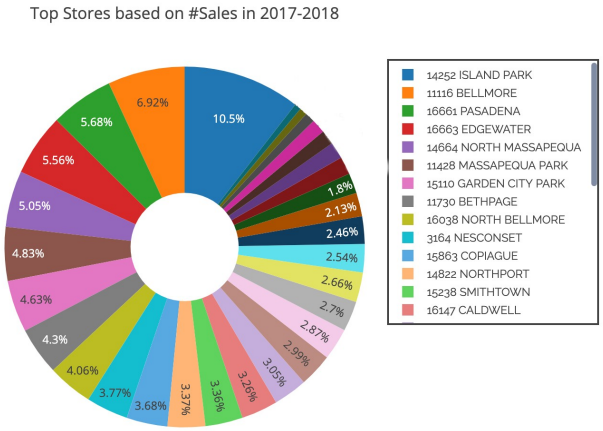

FIG. 2: **Top stores based on sales.**

- The pie chart shows the top stores based on sales.
- The store '14252 Island Park' has the maximum sales.
- There are 31 stores in all and 40% of all sales come from the top 5 stores.

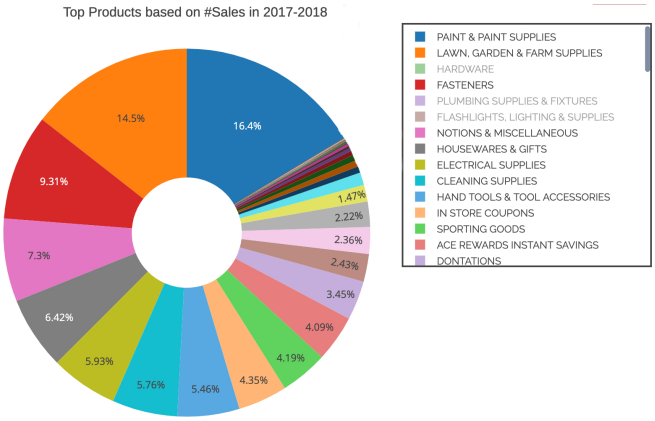

FIG. 3: **Top products based upon sales.**

- The pie chart shows the top product department names based upon sales in 2017-18.
- The department "Paint Paint Supplies" had the most sold products.
- There are 48 departments in which top 5 capture around 54% of the total sales.

FIG. 4: **Results of department wise product classification.**

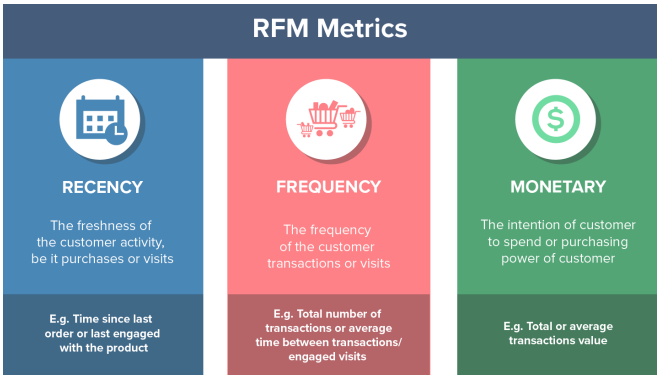

FIG. 5: RFM Metrics

- The tree map shows the classification of products based upon their department names.
- Each department shows products with the statistics with reference to its parent department.
- For instance, "Liquid Paint - National" falls under the category of "Paint Paint Supplies". It is 22,843 in quantity which equals 13% of total quantity of its parent department.

B. Methods

1. *RFM Analysis*

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits. RFM (Recency, Frequency, Monetary) analysis is a proven marketing model for behavior based customer segmentation. RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement. RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.

Why RFM? RFM is an easy way to answer the following questions - Who are my best customers? Which customers are at the verge of churning? Who has the potential to be converted in more profitable customers? Who are lost customers that you don't need to pay much attention to? Which customers you must retain? Who are your loyal customers? Which group of customers is most likely to respond to your current campaign?

Shown below is the R/F/M Values of our segmented Customers.

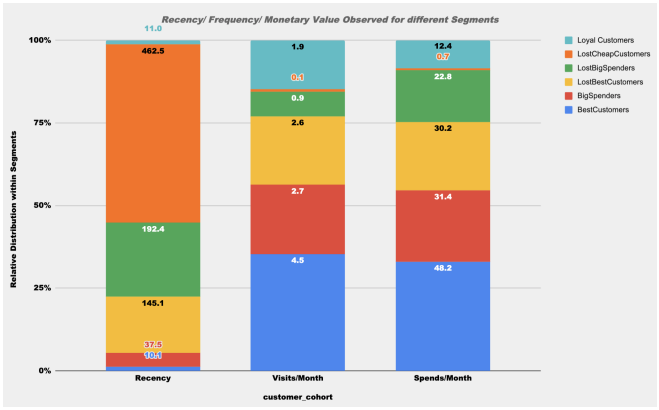

FIG. 6: **Results of R/F/M for different segments.**

- **Recency** - The recency of least cheap customers is 462 days. It can be said that these customers came long ago, purchased a few cheap items and never came back. Contrary to this, we can see the recency of loyal customers and best customers is 11.0 days and 10.1 days respectively. The recency of lost best customers and lost big spenders is 145 days and 192 days respectively. This shows the time period elapsed since these customers have come back to the store is approximately 5-6 months. Using this insight one can examine these customers further and try to determine the causes affecting their visits to the store.
- **Visits/Month** - We can observe that the visits/-month is inversely proportional to recency which actually makes sense. The best customers have at least 4-5 visits per month whereas a loyal customer has 2 visits per month. The lost cheap customer hardly visits the store. The lost big spender used to visit the store at least once a month and lost

best customer used to visit the score 2-3 times a month on an average.

- **Spends/Month** - The major chunk of spends/-month is attributed to best customers followed by big spenders having values 48.2 USD and 31.4 USD respectively. We can say that these two customer segments account for more than 50% of the store's sales. The lost best customers and lost big spenders have 30.2 and 22.8 spends/month. This shows that these two segments attributed a large amount of the store's sales and efforts should be made to bring them back in order to increase their profit margins.

Segment	%Users	R	F	M	Characteristics
BestCustomers	11.8%	1	1	1	Brought most recently, most often and spend the most.
LostCheapCustomers	9.1%	4	4	4	Their last purchase dates long back, purchased a few items and spent little
LostBestCustomers	2.6%	>=3	1	1	They used to purchase frequently and spend the most, but haven't purchased in a while.
LostBigSpenders	1.5%	>=3	>1	1	They used to spend a lot but their last purchase dates long back
Loyal Customers	12.2%	1	1	X	Irrespective of the Spends, they transact frequently
BigSpenders	1.5%	X	X	1	Spends the Most

- We have divided the R/F/M into quartiles namely 1(0.75-1), 2(0.5-0.75), 3(0.25-0.5), 4(0-0.25)

FIG. 7: **Customer Segmentation Definition**

- Figure 7 shows that based on the Recency, Frequency and Monetary Value, the customers are segmented.
- The quartiles' range show how the R/F/M of a customer segmentation are divided.

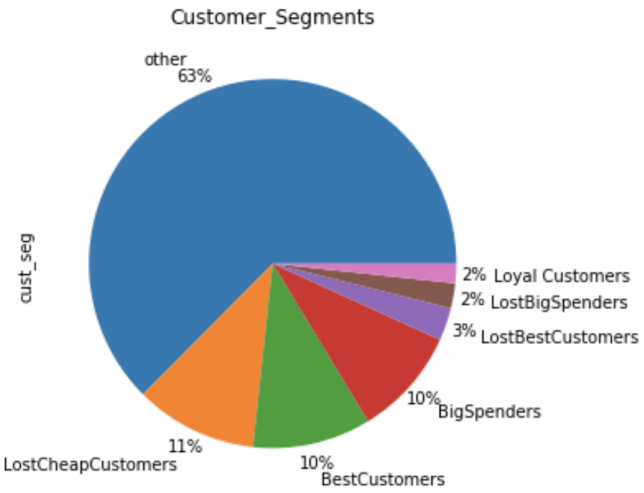

FIG. 8: **Results of customer segmentation statistics.**

- The pie chart shows customer segmentation as - 'Loyal Customers', 'Lost Big Spenders', 'Lost Best Customers', 'Big Spenders', 'Best Customers', 'Lost Cheap Customers' and others.
- 'Lost Cheap Customers' have the highest share 11% apart from others.
- 'Big Spenders' and 'Best Customers' equal at 10% of the share.

2. *Apriori and Association Rule Mining*

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store. It is very important for effective Market Basket Analysis and it helps the customers in purchasing their items with more ease which increases the sales of the markets. Association rule learning is a prominent and a well-explored method for determining relations among variables in large databases.

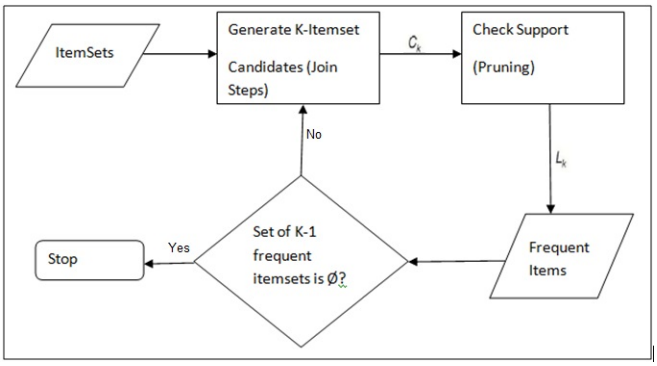

FIG. 9: Apriori Algorithm

Algorithm 1 Apriori algorithm

```
1: begin
2:     L_1 ← Frequent1 − itemset
3:     k ← 2
4:     while L_{k−1} ≠ φ do
5:         Temp ← candidateItemSet(L_{k−1})
6:         C_k ← frequencyOfItemSet(Temp)
7:         L_k ← compareSetWithMinSupport(C_k, minsup)
8:         k ← k + 1
9:     end while
10:    return L
11: end
```

Why Apriori? Each shopper has a distinctive list, depending on one's needs and preferences. Understanding these buying patterns can help to increase sales in several ways. If there is a pair of items, X and Y, that are

frequently bought together: Both X and Y can be placed on the same shelf, so that buyers of one item would be prompted to buy the other. Promotional discounts could be applied to just one out of the two items. Advertisements on X could be targeted at buyers who purchase Y. X and Y could be combined into a new product, such as having Y in flavors of X.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(PAINT ROLLERS)	(PAINT BRUSHES)	0.025375	0.025678	0.007734	0.304791	11.869615	0.007082	1.401480
(PAINT BRUSHES)	(PAINT ROLLERS)	0.025678	0.025375	0.007734	0.301189	11.869615	0.007082	1.394691
(PAINT ROLLERS)	(LIQUID PAINT- NATIONAL)	0.025375	0.035395	0.009572	0.377227	10.657605	0.008674	1.548888
(LIQUID PAINT- NATIONAL)	(PAINT ROLLERS)	0.035395	0.025375	0.009572	0.270435	10.657605	0.008674	1.335898
(LIQUID PAINT- NATIONAL)	(PAINT BRUSHES)	0.035395	0.025678	0.006679	0.188696	7.348483	0.005770	1.200933
(PAINT BRUSHES)	(LIQUID PAINT- NATIONAL)	0.025678	0.035395	0.006679	0.260101	7.348483	0.005770	1.303697
(DONATIONS)	(CHRISTMAS LIGHTS/ACCESS)	0.071127	0.021368	0.005937	0.083471	3.906380	0.004417	1.067759
(CHRISTMAS LIGHTS/ACCESS)	(DONATIONS)	0.021368	0.071127	0.005937	0.277849	3.906380	0.004417	1.286259

FIG. 10: **Results of product basket analysis.**

- The table shows the pairing of product best fit to be bought (consequent) when a given product (antecedent) is bought.
- If 'Paint Rollers' are bought, then 'Paint Brushes' are fit to be recommended to the customers.

3. Arima and Sarima

Autoregressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting. ARIMA is an Ideology that captures autocorrelation in the series by modelling it directly. Lags of the stationarized series are called "autoregressive" that refers to (AR) terms Lags of the forecast errors are called "moving average" which refers to (MA) terms.ARIMA models are also capable of modelling a wide range of seasonal data. A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA models we have seen so far.
A pure Auto Regressive (AR only) model is one where Yt is a function of the 'lags of Yt'.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \epsilon_1$$

A pure Moving Average (MA only) model is one where Yt depends only on the lagged forecast errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q}$$

The ARIMA model is :
predicted Yt= Constant + Linear combination Lags of Y(upto p lags) + Linear Combination of Lagged forecast errors (upto q lags)

The three parameters (p,d,q) that are used to help model the major aspects of a time series are : seasonality, trend, and noise.

- P: is the order of the AR model
- Q: is the order of the MA model
- D: is the differencing order (how often we difference the data)

The ARMA and ARIMA combination is defined as

$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

$$\underbrace{(p,d,q)}_{\text{Non - seasonal part}} \qquad \underbrace{(P,D,Q)_m}_{\text{Seasonal part}}$$

It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1139	0.391	0.291	0.771	-0.653	0.881
ma.L1	-0.6933	0.161	-4.310	0.000	-1.009	-0.378
ar.S.L12	0.6528	0.120	5.425	0.000	0.417	0.889
sigma2	1.671e+07	3.26e-09	5.13e+15	0.000	1.67e+07	1.67e+07

FIG. 11: **Results of ARIMA Algorithm**

4. Heatmaps

Heat maps represent the density of data within different areas of a map. They are effective visualization tools for representing different values of data over a specific geographical area. Data is aggregated based on its location and given a radius of influence, which you can adjust. As the density of data increases in that area the heat map will display a color indicating higher intensity. Geographic heat maps can help you identify trends in data that would otherwise be hard to see. You can quickly see areas that might already be saturated and other areas where there is still a market opportunity. Heatmaps when combined with sales data can help in following ways -

- View Market availability: It gives us a very good sense of which customers have access to your stores or products.We can add a location based heat map to your map. The gradient of the heat map will show what areas of the map are close to your stores and which might be too far for your customers to travel to.

- SALES DENSITY: Combine the data with the geographic heat map to weight your locations. On doing this with sales data, one can see what areas have the highest sales density. This can help focus on the marketing budget where it will do the most good. The heat mapping software can be used to create heat map by zip codes to see the demographic data of the areas having most sales.

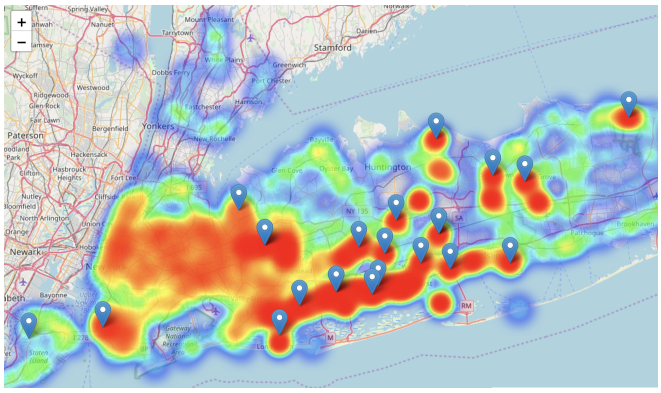

FIG. 12: **Map of demand from population and the supplying stores.**

- The map shows in red the places in New York with maximum demands, yellow with lesser demands and the blue pin shows the stores.
- The southern part of the Long Island shows maximum demand and we have a number of stores fulfilling the demand.
- The least number of demands is from the west part of New York which also shows the least number of stores in that area.

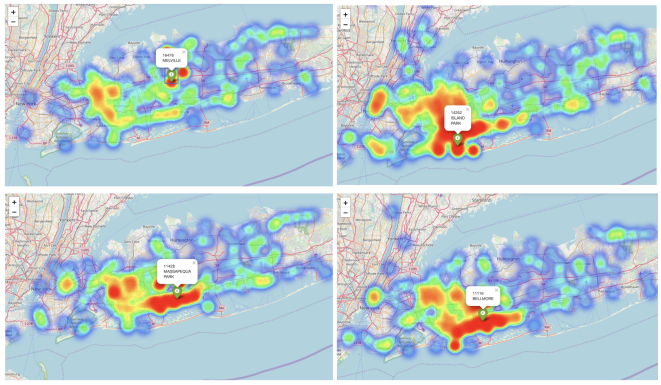

FIG. 13: **Map of stores and population attraction.**

- The map shows top 4 stores by sales and the population clusters they attract.
- We can see that the densest population cluster they attract are shown in red and the lesser density is shown in orange, yellow and green.
- Stores like '14252 Island Park', '11428 Massapequa Park' and '11116 Bellmore' have outreach more spread at distances more than their adjacent locality.
- '16476 Melville' store does not follow this progression and has the adjacent locality with maximum outreach.

IV. RESULTS

Top5 Stores	Yearly Sales	#Avg Monthly Sales	#Monthly Unique Users	Load/Cashier	Profit %	Purchase/Cust/ Month	Population	%Coverage
14252 ISLAND PARK	97,745	8,145	4,599	10.16	71%	1.77	4,655	98.80%
11116 BELLMORE	51,175	4,264	3,342	6.87	75%	1.28	16,218	20.61%
14664 NORTH MASSAPEQUA	46,735	3,894	3,051	7.28	71%	1.28	17,886	17.06%
15238 SMITHTOWN	45,503	3,792	2,675	6.87	70%	1.42	117,801	2.27%
15110 GARDEN CITY PARK	43,749	3,645	2,982	6.5	68%	1.22	7,806	38.20%

Bottom5 Stores	Yearly Sales	#Avg Monthly Sales	#Monthly Unique Users	Load/Cashier	Profit %	Purchase/Cust/ Month	Population	%Coverage
7504 GRAND BLVD	7,677	639	300	2.75	53%	2.13	22,603	1.33%
15784 EI PAINT	7,094	591	100	1.67	63%	0.62	15,784	0.63%
16663 EDGEWATER	5,899	491	303	1.33	58%	1.11	9,023	3.36%
16661 PASADENA	4,475	372	334	1.33	71%	1.62	24,287	1.38%
16660 GLEN BURNIE	704	58	93	1.2	52%	5.91	67,639	0.14%

FIG. 14: **Statistics of Top 5 and Bottom 5 stores.**

- The store '14252 Island Park' is the one with maximum yearly sale and the maximum number of unique customers with a 98.8% coverage.
- The store '16660 Glen Burnie' has the lowest yearly sales, with the least number of unique users and the least coverage of 0.14%.

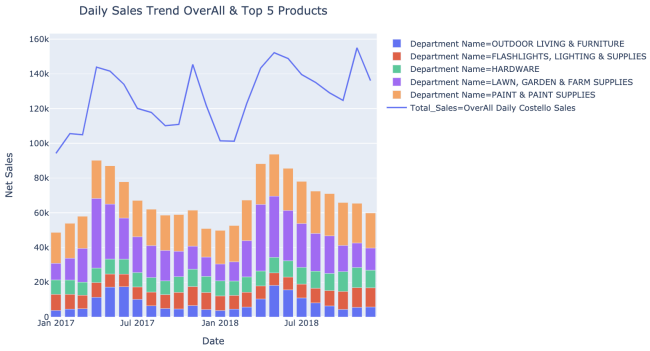

FIG. 15: **Results of daily sales trend.**

- The pie chart shows the top product department names based upon sales in 2017-18.
- The department "Paint Paint Supplies" had the most sold products.

- There are 48 departments in which top 5 capture around 54% of the total sales.

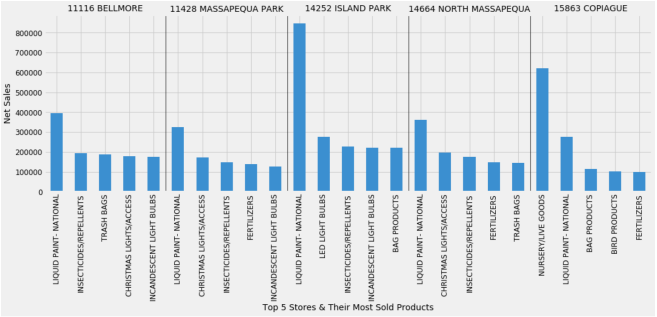

FIG. 16: **Results of top products in top stores by sales.**

- The bar graph shows the "Top 5 Stores" by sales of their "Top 5 Products"
- "14252 Island Park" is the top store by sales with "Liquid Paint-National" as the top selling product department.
- "Insecticides/Repellents", "Christmas Lights/Access" and "Fertilizers" are common top products sold in the top stores.

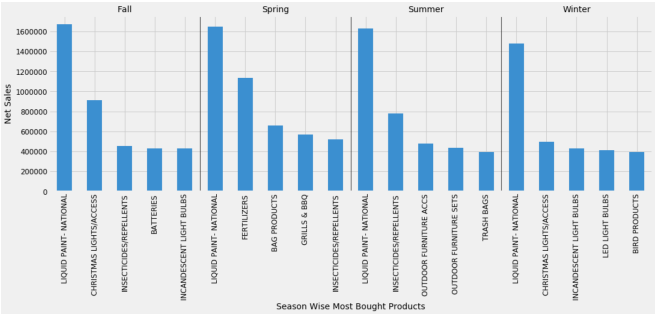

FIG. 17: **Results of season wise top products.**

- The bar graph portrays the "Top 5 Products" by sales in the 4 seasons - Fall, Spring, Summer and Winter.
- "Liquid Paint - National" is the top product by sales in all of the 4 seasons.
- "Insecticides/Repellents" is the next product that is most sold in Fall, Spring and Summer.

- The store age is defined as - **New (1,2,3), Old (4,5,6), Very Old (7,8,9)** of the 'Store #' feature

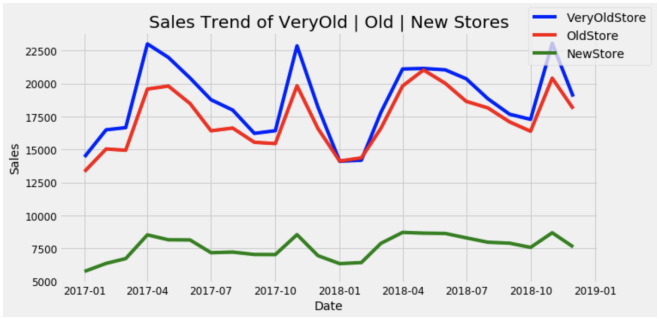

FIG. 18: **Results of monthly sales trend as per store age.**

- The graph shows that the Very Old Stores have the maximum sales.
- The New Stores have the least sales and the Old Stores have sales lying in the middle.

FIG. 19: **Results of customer cohorts purchase trends.**

- 'Best Customers' buy most items per month throughout the year.
- 'Big Spenders' buy items per month even if the average price is high.
- 'Loyal Customers' buy items per month irrespective of the average price.
- 'Lost Best Customers' frequency of buying items per month is shown getting reduced at the end of year 2018.
- 'Lost Big Spenders' frequency of buying items per month decreased when the average product prices increased.
- 'Lost Cheap Customers' frequency of buying items per month decreased on decrease in the average price.

- 'Liquid Paint - National' is the most sold product in all the customer segments.

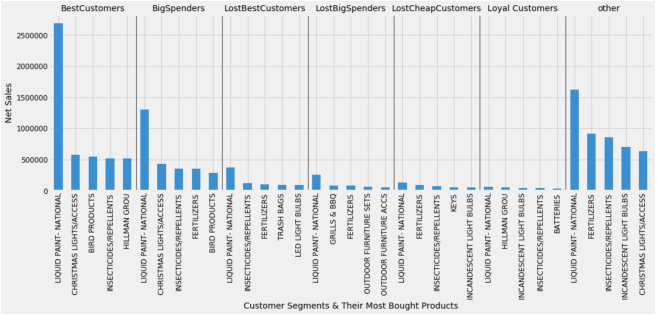

FIG. 20: **Results of most products bought by the customer segments.**

- 'Best Customers' are the ones who buy the most number of products.
- 'Loyal Customers' buy the least number of products yet are loyal to visit the store regularly.

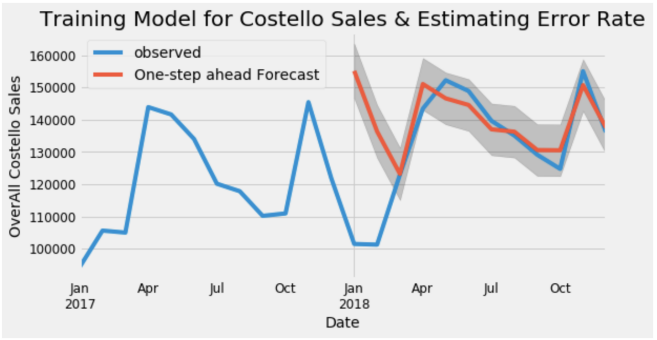

FIG. 21: **Results of training Costello Sales.**

- The 'Auto Regressive Integrated Moving Average' (ARIMA) is used for modelling.
- The training data is 70% and test data is 30% of the entire data set.
- As visible in the graph, the plot corresponding to the test data (red) corresponds to the actual data in year 2018.
- Hence, the validity of ARIMA model can be seen.

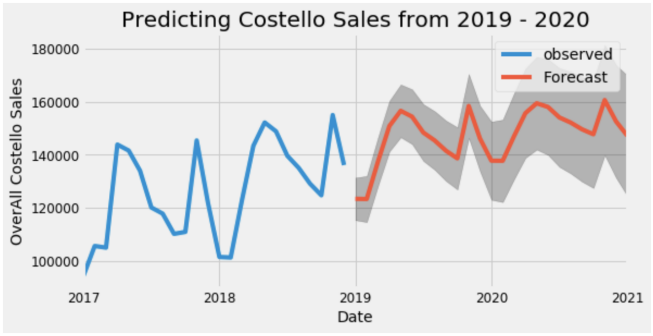

FIG. 22: **Results of predicting Costello sales.**

- Based upon the 'Auto Regressive Integrated Moving Average', the future sales are predicted.
- The ARIMA model helps us to forecast the sales (in red) in the period 2019-21.

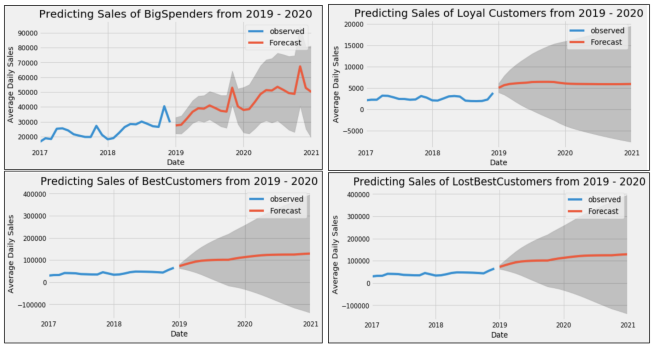

FIG. 23: **Results of future sales forecast for customer segments.**

- Based upon the forecast, the 'Big Spenders' will be spending more in the future.
- The 'Loyal Customers' forecast shows a minimum change in their spending habits.
- The 'Best Customers' and the 'Lost Big Customers' forecast also show a negligible change in their spending.

- Based upon the forecast for 2019-20, the average daily sales for Old Stores is the maximum.
- There is a slight increase in the average daily sales of New Stores.
- The average daily sales of Very Old Stores remains almost the same.
- The top selling product - 'Paint Paint Supplies' is forecast to be sold at a decreasing rate.

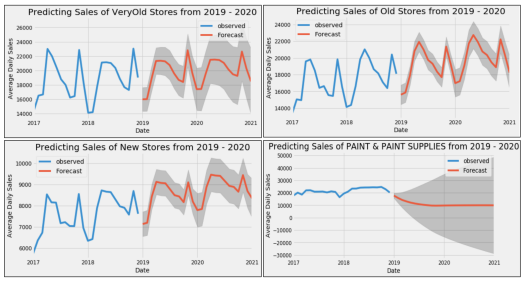

FIG. 24: **Results of store segments and top selling product forecast.**

	Rank – % Times a product is scanned in the invoice at the given Rank				
Top Products * Scanned Rank	**1**	**2**	**3**	**4**	**5**
$5 Costello's Cash	0%	26%	32%	14%	8%
BIRDSEED WILDBIRD 20#ACE	68%	16%	21%	7%	4%
CLEANR GLAS19OZ SPRAYWAY	31%	20%	16%	8%	7%
CMN Donations	0%	28%	20%	18%	11%
CONTRACTOR BAGS 3MIL. 20CNT	34%	28%	12%	12%	3%
FASTENERS	40%	56%	38%	26%	15%
KEY KWIKSET KW1	40%	31%	20%	13%	3%
KEY KWIKSET KW1-ACE250PK	35%	39%	18%	5%	5%
KEY SCHLAGE SC1-ACE250PK	35%	35%	16%	5%	4%
PEAK WASH/DEICER -25	50%	27%	14%	7%	2%

FIG. 25: **Ranking of products based upon scan rate.**

- The ranking 1-5 signifies the position of the product in the invoice.
- 'Birdseed Wildbird 20Ace' appeared 68% (maximum) at rank 1 in the invoice.
- '$5 Costello's Cash' and 'CMN Donations' did not appear at rank 1 in the invoice.
- 'Fasteners' was the second most scanned item by the cashier.
- 'Peak Wash/Deicer - 25' appeared the least number of times (2%) at rank 5.

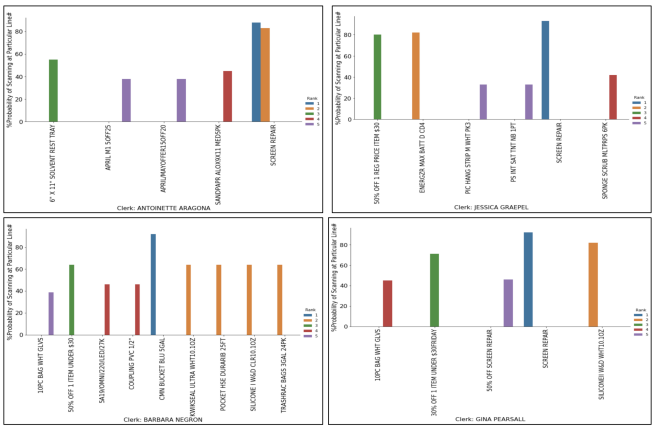

FIG. 26: **Results of cashier behavior while scanning.**

- The bar graph shows the products being scanned in the top 5 ranking in the invoice per clerk in the store. It hence, shows the **Cashier Behaviour** while scanning the product.
- 'Antoinette Aragona' scanned 'Screen Repair' at the top most ranking. She scanned 'April M1 5OFF25' and 'April/MayOffer150Off20' at the 5th rank.
- 'Jessica Graepel' scanned 'Screen Repair'at the top most priority. She scanned 'Pic Hang Strip M Wht PK3' and 'PS Int Sat TNT NB 1PT' at the 5th rank.
- 'Barbara Negron' scanned 'CMN Bucket BLU 5GAL.' first and '10 PC Bag Wht GLVS' the last.
- 'Gina Pearsall' scanned 'Screen Repair' first and '50% Off Screen Repair' the last.

V. FUTURE SCOPE

- NET PROFIT: It is difficult to calculate the net profit due to lack of information on transportation, logistics, employee salaries, maintenance costs, other costs, etc. All the above mentioned factors play an important role in determining the profits of a store and can be taken into account for future work.
- CASHIER EFFICIENCY: We cannot the efficiency of cashiers now as working hours for cashier are not given. But, it can be taken into account for future work. It is an important factor since sales of a store depend a great deal on the cashier efficiency.
- WAREHOUSE STOCKING: Predicting items to be stocked is challenging because we do not have the information about the size of the store (per square unit area).

[1] "Market basket analysis: Identify the changing trends of market data using association rule mining," https://core.ac.uk/download/pdf/82094674.pdf.

[2] "Recommender system based on customer behaviour for retail stores," http://www.iosrjournals.org/iosr-jce/papers/Vol19-issue3/Version-1/B1903010617.pdf.

[3] "Recommender systems and consumer product search (full paper, word count: 7773) vidyanand choudhary zhe (james) zhang university of california, irvine university of texas, dallas," http://misrc.umn.edu/wise/2014_Papers/Strategic20Product20Recommendation20WISE202014.pdf.

[4] "Rfm analysis," https://clevertap.com/blog/rfm-analysis/.

[5] "Apriori algorithm," http://dwgeek.com/mining-frequent-itemsets-apriori-algorithm.html/.