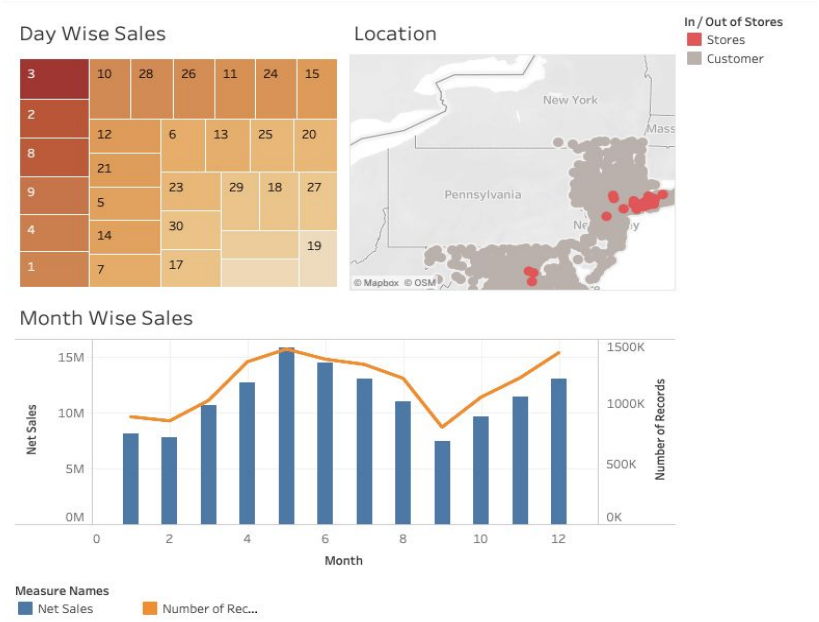
CSE 519 Final Report
Retail Sales Data Analysis

1. Introduction

Retail Sales has historically been a point of great analysis since time immemorial as every store cares about customer retention and addition. There are many different behavioural and buying patterns that can emerge from the analysis of such data. We have reviewed many classical and modern retail sales analysis techniques and have done an in-depth analysis of some such approaches.
The main aim of our project was to get insights about how we can increase the sales of Costello's ACE Hardware with the help of insights gained from data provided by them and make suggestions for the same. We're doing this by analyzing the data thoroughly to study which customers to focus on, how to stock goods and recommend products to specific customers..
Through our research about this topic, we came across many avenues which could be tackled but specifically we focused our approach on Customer Segmentation, Recommendation Systems and Store-Location and Demographic Analysis.

2. Data and EDA



The data that we were provided with had more than 32M rows for the years 2015 - 2018, which contained detailed information about the items, purchases and sales made in the Costello Stores. From our initial EDA, we found out that there were certain trends that were followed in the data which we had depicted using Tableau, to make an interactive dashboard. We plotted graphs for day-wise sales, month-wise sales as well as various locations where customers come from vs where Costello stores are there.

There were a couple of challenges that had to be encountered including data not in similar formats and the data set took a lot of time to analyze due to its size. Getting a sample that covered the details for the entire population demanded an intricate way of selecting samples. Proper data imputation was required because there were columns (like Return Code) which contained a lot of null values.

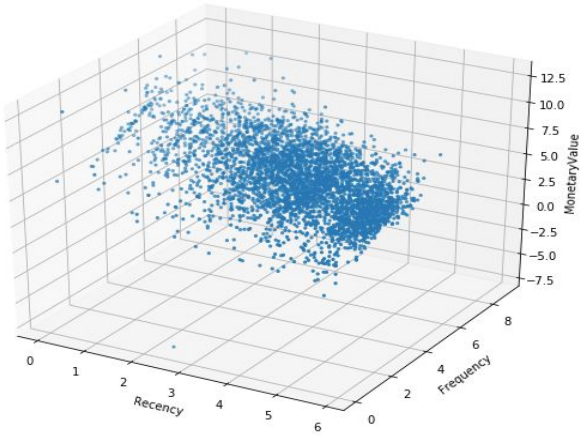
3. RFM Modelling and Analysis

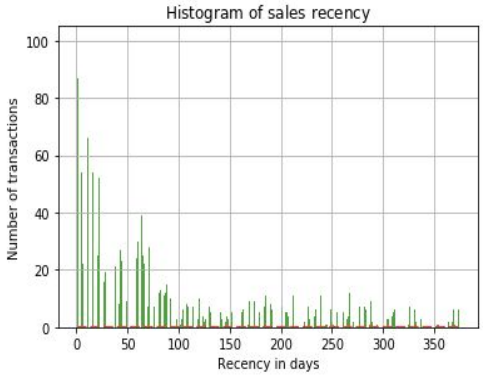
3.1 Customer Segmentation

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 they bought. The 3 metrics used for customer segmentation are:

RECENCY (R): Time since last purchase for each customer
FREQUENCY (F): Total number of purchases per customer
MONETARY VALUE (M): Total monetary value per customer

3.2 RFM Visualization :





Highlights about the Visualization -

3D Graph : The 3D graph represents the relation between recency, frequency and monetary value. It says customers who recently bought items are also the ones who buy more frequently and place high-value orders. Customers who transacted a long time ago, place low value orders. And, customers who buy occasionally, do not buy expensive items.

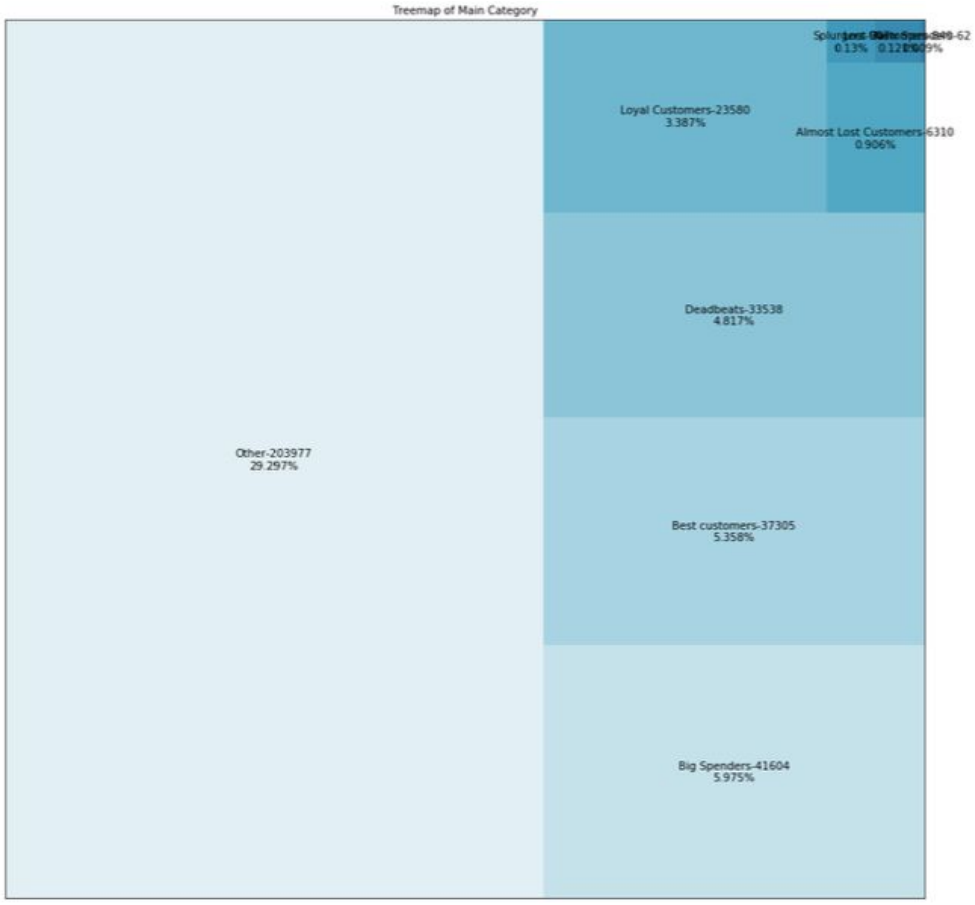
Histogram of Sales Recency : It represents the number of transactions vs the recency in days. We fix a reference date for finding the recent transactions. The reference date would be a day after the most recent transaction date in the dataset (1/1/2019). Then we calculate the difference(days) between the most recent transaction carried out by the customer and this reference date.

3.3 RFM Quartiles -

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).
We have divided the customer base into different categories based on spend pattern, number of transactions, retention rate as described below:

Customer Segments -

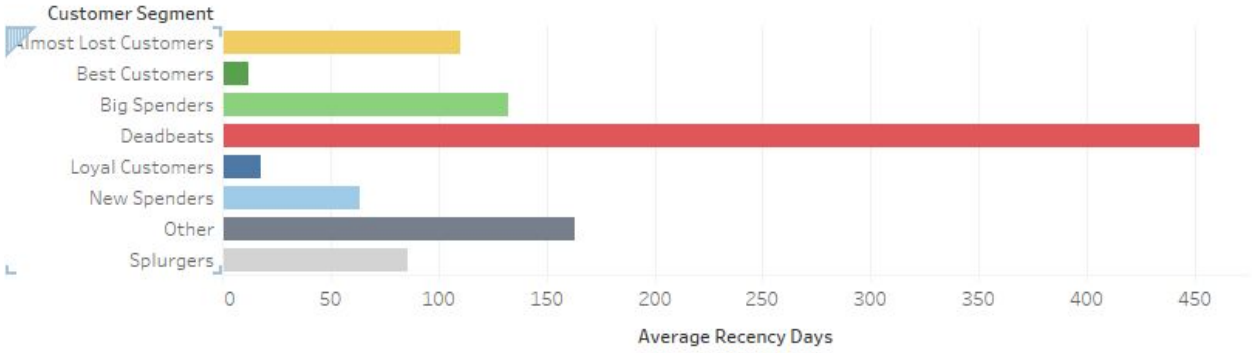
Segment	Users(%)	R	F	M	Activity
Best Customers	5.358	1	1	1	Brought most recently, most often and spend the most.
Loyal Customers	3.387	1	1	X	Irrespective of the Spends, they transact frequently.
Big Spenders	5.975	X	X	1	Spends the most but not that frequently.
Deadbeats	4.817	4	1	1	They used to spend a lot but their last purchase dates long back.
Almost Lost Customers	0.906	>=3	1	1	They used to purchase frequently and spend the most, but haven't purchased in a while.
Splurgers	0.13	<3	1	1	These customers spend freely within almost the same time span.



Highlights of TreeMap : It shows different segments and the proportion of customers lying in that particular segment. As we can see, 37305 customers are classified as best customers and 41604 are classified as big spenders. Our bulk order providers are the ones who do not visit the store frequently but provide bulk order. Thus, they are counted under Big spenders.

3.4 Inferences from RFM Analysis

Recency -
The recency of Deadbeats(lost customers) is 452 days. It can be said that these customers came long ago, purchased a costly items and never came back. Contrary to this, we can see that the recency of **Best Customers** and **Loyal Customers** is **12** days and **18** days respectively. The recency of **Big spenders** and **Almost Lost Customers** is **145** days and **120** days respectively. This shows that the time period elapsed since these customers, of both these categories came back to the store is approximately 4-5 months. Using this insight, one can examine these customers further and try to determine the causes affecting their visits to the store.



Ranking Analysis -

Best Customer

Segment	Customer Number	Monetary Value	RFM Class
Best Customers	10000	244,014	111
	100020	104,764	111
	792541	50,864	111
	*5	8,138,712	111
	*6	132,069	111

Big Spender

Segment	Customer Number	Monetary Value	RFM Class
Big Spenders	2061	2,622	311
	20085	7,673	311
	88808	2,889	321
	99996	107,064	421
	792523	3,946	311

Loyal Customer

Segment	Customer Number	Monetary Value	RFM Class
Loyal Customers	*10294	27.06	112.00
	*10457	91.86	114.00
	*10538	82.10	113.00
	*11497	25.65	112.00
	*13302	78.91	112.00

Deadbeats

Segment	Customer Number	Monetary Value	RFM Class
Deadbeats	146668	655	411
	274158	947	411
	401769	17,264	411
	904140	422	411
	*32225	1,673	411

Insight about Ranking analysis : We know *5, *6, 10000 are the bulk order customers. If we remove these, customer number 792541 is the most valuable customer. This customer visits very frequently, more often and gives a huge amount of business. Customers who visit more often and frequently are our loyal customers. We have their customer ID which can be used for improving upon the existing loyalty program. There are certain customers who spend a lot but do not happen to visit the stores frequently. We can promote costliest new products to these customers.

Loyalty Program Revamp -
The existing loyalty program of the stores has some flaws. Customers who are now lost or are not rated as loyal customers have been assigned a loyalty ID. This shows the existing system has some flaws which needs to be improved. Our RFM model proposes the new loyalty program which can be used to assign loyalty ID to actual loyal customers.
Let's look at the current scenario:

Customer Number	Loyalty ID
146668	1.90425e+09
146668	1.90425e+09

Customer Number	Loyalty ID
904140	1.94682e+09

The above customer Number comes under Deadbeats segment means these are the lost customers. There is no point in assigning the Loyalty ID to these customers. We should instead focus on higher ranked customers in the RFM analysis who would be better suited for promotions/deals.
Based upon our loyal customer segment the program can be revamped.

Segment wise Marketing -

Segment	Marketing
Best Customers	No price incentives, new products, and loyalty programs
Loyal Customers	Promoting new products to loyal customers is a great way for getting initial traction and feedback
Big Spenders	Market your most expensive products
Almost Lost Customers	Aggressive price incentives

RFM analysis can be used to plan the market strategies according to different customer segments. Here we suggest some marketing tactics based upon the customer segments.

- **Best customers** are the ones who purchase frequently, with high monetary value. We could further drill down into their requirements and also make personalized recommendation systems for these people as we would not want such people to go to others at any cost .
- Customers who are classified as **loyal customers** can be used to market new products, their feedback would be beneficial. They are still not the "Best customers". We would like to convert these into our Best Customers. For this, we can try to further drill down into what are the reasons they are still not Best Customers. They might not be getting some products that they would want or the quality of material they need is not present with us.
- **Big Spenders** have a big purse and hence we should make the most of it. Promote the costliest product to them. These are the people we have to target to increase our profit margin.
- **Almost lost customers** have little faith in us. We would like to gain their trust back. Aggressive discounts and pricing strategies can be used for making things affordable for them. Tracking these customers overtime would be helpful.

4. <u>Product Recommendation System and Store Orientation</u>

4.1 Word2Vec based Recommendation system

One of the most essential marketing techniques that can help augment customer sales, can be a way to sell products to customers which they require, but they didn't know of. This can be accomplished by creating a recommendation system which uses the previous history of each of the customers to get an idea of what kind of products a customer buys and then suggest similar items to them.

We first started off by creating a Buffer - **customer purchase history** - which finds all the distinct customers and finds all the items bought by each of them in the 4 years of data. We had also kept 10 percent of customers purchase history for testing.

We'd be happy to share our trained model with Costello as well as others who would like to test the performance. After some hyperparameter tuning, we found the best results coming with a Word2Vec model with a **sliding window of '10' and negative sampling of '10', learning rate of '0.03' and 10 epochs.** We then built 'word2vec' embeddings for our vocabulary and trained our model using the list of customer's purchase history.

There are **two types of recommendations** that we are proposing. First, we can recommend products which are similar to the current item that the customer would like to purchase. This is a fairly trivial approach. So we do one more analysis wherein we use the complete history as well as recent few purchases.

For our dummy implementation, we tried to do an in-depth analysis for the 100th customer in our unique customer list.

```
products_dict['10461']
```

```
['SCRUBR KTCH 5X7/8X2-7/8']
```

```
similar_products(model['10461'])
```

```
[('KITCHENBAG CITRUS 13GAL', 0.6271333694458008),
 ('K-CUP ENGLISH TEA', 0.6119390726089478),
 ('TIDE LIQ CL&FR REF BRZE 100 OZ', 0.6061972379684448),
 ('KLIP IT JUICE JUG 67.6OZ', 0.6050052642822266),
 ('GREENWORKS TOILET', 0.6027259826660156),
 ('LYSOL SPONGE MOP', 0.6022356748580933)]
```

- We found out that the customer #100 's last purchase was Kitchen Scrubber, so our model suggested other Kitchen items like Tea, Juice, Sponge etc.
- For the second part of the analysis, we found out that the customer had bought 108 products from the period of 2015-18. (Fig1)
- Customer#100 was suggested items like Wrench, Hammer, Firestarter etc when all the items were passed to our model for suggestions. (Fig2)
- There is also a case of recommendations depending upon recent purchases, so we also tried to recommend items based on last 10 purchases(Fig3). We can also see the influence of last bought item(Scrub) as more Kitchen related products have been suggested.

Fig1

```
len(purchases_testing[100])
```

```
108
```

Fig2

```
similar_products(aggregate_vectors(purchases_testing[100]))
```

```
[('WALLDOG CONTRACTOR KIT', 0.9199569225311279),
 ('WRENCH COMB 7/16MR ACE', 0.906193196773529),
 ('BRUSH PELLET 3" W/10\'ROD', 0.905761182308197),
 ('RESPIRATOR VFLEX N95 2PK', 0.9002688527107239),
 ('TUMBLEWEEDS FIRESTARTER', 0.8997049927711487),
 ('HANDLE HAMMR 14"TUFSTUF', 0.8964860439300537)]
```

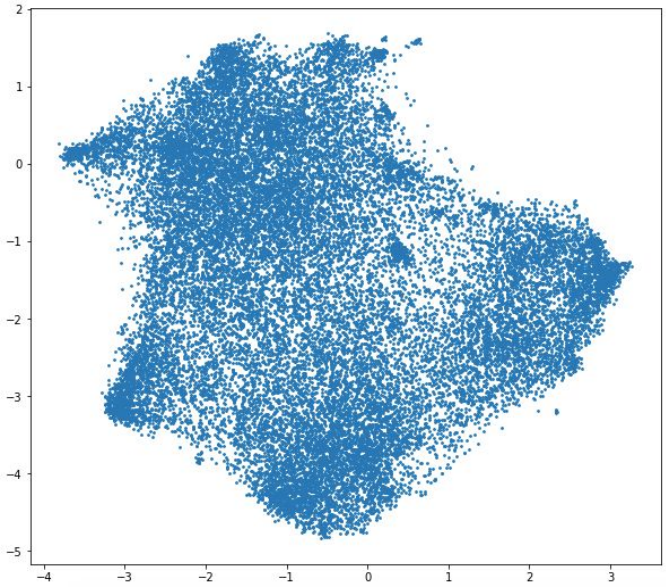
Fig3

```
similar_products(aggregate_vectors(purchases_testing[0][-10:]))
```

```
[('BRUSH PELLET 3" W/10\'ROD', 0.8217720985412598),
 ('COMPOSITE MATES LARGE EGG', 0.8205544948577881),
 ('GASKET STOVE 3/8"X132\'', 0.8186802268028259),
 ('FP SHOVEL30"TWISTD STEEL', 0.8137106895446777),
 ('REPLACEMENT TRIMMERSPOOL', 0.8116650581359863),
 ('#40 CONCRETE MIX SAKRETE', 0.8092637062072754)]
```

4.2 Visualizing Product Embeddings using PCA

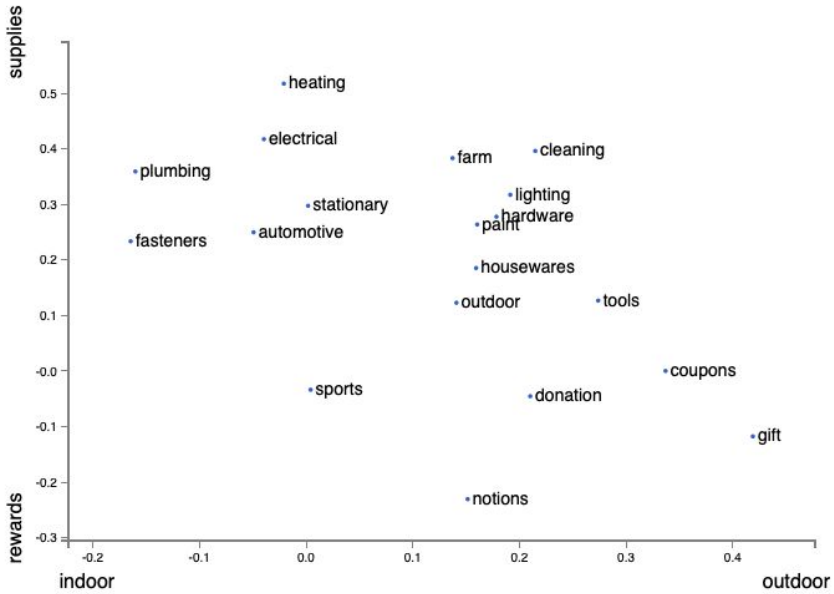
To get the overview of all the items available in the store, we had to come up with a metric to not only visualize but also group all the available items in our data. For this we used Word2Vec embeddings, where we tried to plot the embeddings for all the products. But this was a challenge as the embeddings created were in high dimensional space and to visualize them for better analysis, we used Principal Component Analysis to make the embeddings plottable in 2-D graph.



From the above graph we can see that there aren't many distinct clusters formed which. As Costello is a speciality store which has items of a certain type, there is an issue which can be witnessed when there are so many types of products that are available in the store. But eventually it becomes easier for our model to suggest similar models when a particular product might not be in stock, a future implementation that could be worked on.

4.3 Word Embeddings to determine nearby departments

We used pre-trained word embeddings to find out which all departments are close to each other in the word embeddings space. This can be an accurate measure as to how we can place the various items together physically. There were a couple of challenges faced with this approach as the department names were often not in the libraries and there were some spelling errors in the department names as well(Donation was named as Donation).



We plotted these embeddings with respect to their relation to intrinsic characteristics like whether these departments are closer to indoor type of products or outdoors. Similarly in Y axis we did this for whether they belong to rewards type category or normal supplies.

From our plot, we could find that Departments like Plumbing, Heating, Electrical etc and Farm, Cleaning, Paint etc should be close to each other in the physical stores.

5. <u>Demographics Analysis</u>

We wanted to find out whether some sort of demographic relation exists between the stores and the net sales of those particular stores. We also wanted to find out whether these demographic features could be used to find where new stores could be opened. For this, we started with scraping certain income and demographic related features from the locations where Costello stores exists. We also did an in-depth analysis about how a certain demographic of people could help augment the footfall of the stores. To facilitate the traction of customers we have proposed certain locations where opening stores would be beneficial for the overall sales and profit. The process for the same has been explained below.

Customer Number *6(aged people) showed the presence of demography in existing dataset.

	Store Name	Population	Population density	Median Age	Married (15yrs & older)	Families w/ Kids under 18	Income per capita	Median household income
0	14252 ISLAND PARK	4718.0	11008.0	42.7	51.0	37.0	37652.0	85679.0
1	11116 BELLMORE	15722.0	6666.0	43.3	57.0	41.0	48843.0	124048.0
2	14664 NORTH MASSAPEQUA	18493.0	6175.0	43.0	66.0	43.0	41523.0	107328.0
3	11428 MASSAPEQUA PARK	17176.0	7773.0	43.1	67.0	45.0	42338.0	117934.0
4	11730 BETHPAGE	16197.0	4530.0	46.3	67.0	34.0	41887.0	107240.0
5	16663 EDGEWATER	11998.0	12356.0	38.2	64.0	48.0	61172.0	102355.0

We scraped data from https://www.areavibes.com and got the demographics of each and every store location. The features we scraped were:- Population, population density, Median Age, Income per capita, Income per household. A summary of the new dataset made is attached above.

	Population	Population density	Median Age	Married (15yrs & older)
Population	1.000000	-0.022955	-0.422195	-0.312365
Population density	-0.022955	1.000000	-0.057852	-0.073840
Median Age	-0.422195	-0.057852	1.000000	0.353042
Married (15yrs & older)	-0.312365	-0.073840	0.353042	1.000000
Families w/ Kids under 18	0.234900	-0.048506	-0.527726	0.021791
Income per capita	-0.495303	-0.097428	0.358952	0.421533
Median household income	-0.469294	-0.263765	0.383951	0.648520
Net Sales	-0.344857	0.226324	0.224797	-0.042855

We then found that a correlation of 0.22 exists between the Median Age and the Net Sales when we consider stores from all the regions where Costello stores are present. But we realized that analyzing the demographic traits from different states might not be a wise idea as people from New York and their demographics vary widely to those of Maryland. So, we only considered the demographic information from Costello Store regions in NY and we found that the correlation increased to more than 0.33.

	Population	Population density	Median Age	Married (15yrs & older)
Population	1.000000	0.001494	-0.730360	-0.481085
Population density	0.001494	1.000000	-0.075762	-0.228441
Median Age	-0.730360	-0.075762	1.000000	0.529702
Married (15yrs & older)	-0.481085	-0.228441	0.529702	1.000000
Families w/ Kids under 18	0.411978	0.030934	-0.491066	-0.079869
Income per capita	-0.624586	-0.426655	0.696960	0.460047
Median household income	-0.661477	-0.370755	0.452280	0.578495
Net Sales	-0.425169	0.202444	0.332161	-0.082180

There seem to be a huge difference in the spending habits of the NJ and NY people. On further diving into the per capita income and also the mean age of the people in NJ and NY, we found that the average age was higher in case of the NJ with respect to their counterparts in NY. It was also seen that the per capita income of the people in NY is higher than those of NJ. This is exactly the reason why we wanted to do an in-depth analysis of the highest performing store in NJ.

Lincoln Park - Location of highest earning store in NJ

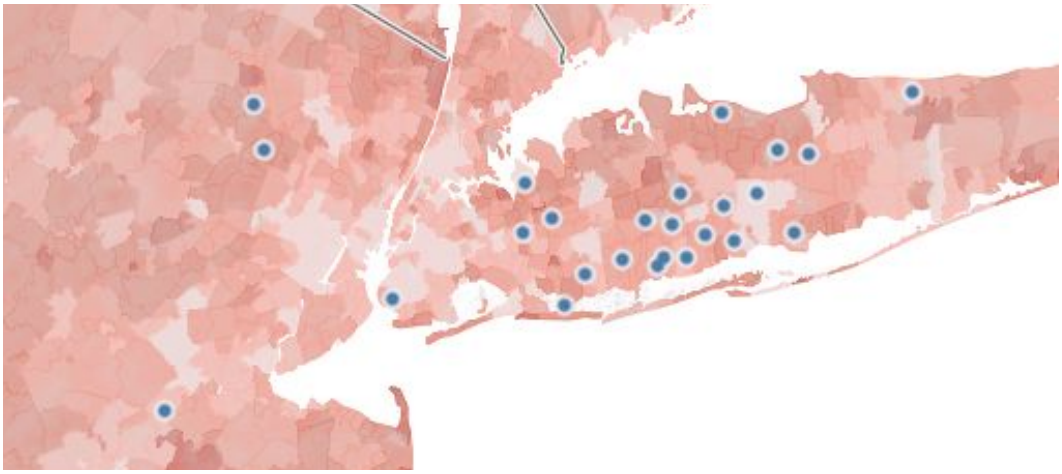
When we delved into the amount and type of items being sold in this store, we found that Power Tools and Accessories were the most sold items. After trying to find out why that may be the case, we stumbled upon a resource[3] from where we got to know that Lincoln Park is one of the '**Manufacturing Hotspots**' in NJ region. This might have a direct effect in the sales of power tools and accessories as Costello, Lincoln Park is one of the only prominent stores there in the region. Building and manufacturing has been the primary source of income for the Lincoln Park people and hence the presence of hardware store there seems a better fit.

This would be an interesting avenue to look upon while opening new stores - by looking into various industrial hotspots, so that there is a guaranteed stream of income readily available.

	Store Name	Net Sales
16	15998 BALDWIN HARBOR	3.671718e+05
25	5144 DEER PARK AVE	1.405439e+06
27	7504 GRAND BLVD	1.625430e+06
21	16660 GLEN BURNIE	1.632684e+06
19	16324 BROOKLYN	2.218065e+06
2	11730 BETHPAGE	8.128507e+06
1	11428 MASSAPEQUA PARK	8.252226e+06
6	14664 NORTH MASSAPEQUA	8.301779e+06
0	11116 BELLMORE	1.337950e+07
5	14252 ISLAND PARK	1.897247e+07

Baldwin Harbour vs Island Park

Baldwin Harbor store in particular seems to have a bad reputation. Google reviews show that 50% of its reviews were negative whereas that of Island Park were only 25%. This might also have an affect on the sales of the store. Island Park earns the most of all the stores while Baldwin earns the least. Baldwin Harbour has 2 other hardware stores besides Costello, Do It Hardware" and "East Coast Sprinkler", which might sway away the people living there. The reviews of these stores as found on the Google Reviews were much better than those of Costello's. There might be a case that these stores have certain items that Costello does not possess or generally provide better prices. Selective discounts for such stores can be provided so as to remain competitive in business.



After going through data from Bureau of Labor Statistics(BLS)[3] data, we tried to understand which places might potentially be the manufacturing-hotspots but do not have Costello stores. It was found that White Plains in NY and Wayne in NJ had the highest number of manufacturing jobs and is a manufacturing hotspot and hence opening a store at this place would be fruitful for Costello.

Highlights of Map -

New Store Prediction based upon demography: The graph shows the existing stores present with blue labels. The highlighted area(red) shows the places which are coherent with our metrics of demography of the region ideal for opening new stores. Zip Code **07474** i.e. Wayne, NJ is a place where a new store can be opened. This place, as confirmed with data from BLS, has high manufacturing output and hence it can be useful for Costellos to be there. Similarly, **11001 Floral Park, NY** can also be a place where a new store can be opened.

From our RFM analysis, we had found out that Customer# *6,i.e. the people who got an elderly discount had one of the highest RFM values, they generally provide good amount of business to the stores.

11428	MASSAPEQUA PARK	7283
16791	STATEN ISLAND	5712
16354	LINCOLN PARK	5673
11730	BETHPAGE	4668
11116	BELLMORE	4625
16147	CALDWELL	4482

Therefore, we analysed the footfall of Customer number *6. The stores with the highest footfall of such elderly people are shown below in snippet. We found that the stores we have proposed (Wayne and Floral Park) had a similar average median age as compared to the demography of existing stores with high elderly footfall. This shows that it might be possible that opening a store at these particular locations will have a 2 fold effect. It will target the manufacturing areas coupled with the elderly population of that and the neighbouring areas.

6. References -

[1] RFM Analysis https://www.blastam.com/blog/rfm-analysis-boosts-sales
[2] Bureau of Labour Statistics https://www.bls.gov/eag/eag.ny_newyork_md.htm
[3] Lincoln Park info http://www.newarkhistory.com/lincolnpark.html
[4] Tableau https://community.tableau.com/thread/272452
[5] Word2Vec https://arxiv.org/pdf/1402.3722.pdf