CSE 519 Data Science Fundamentals - Project Progress Report
Retail Store Data Analysis

Objective

To analyze and find out the reasons for the drop in the sales of the hardware store 'Castello's Ace and suggest ways to improve the sales by

- Placing similar or frequently bought products nearer, to improve the customer experience and also to suggest customers with related products incase the product goes out of stock.
- Stocking products before they run out of stock/Predicting when a product will be out of stock in a particular store.

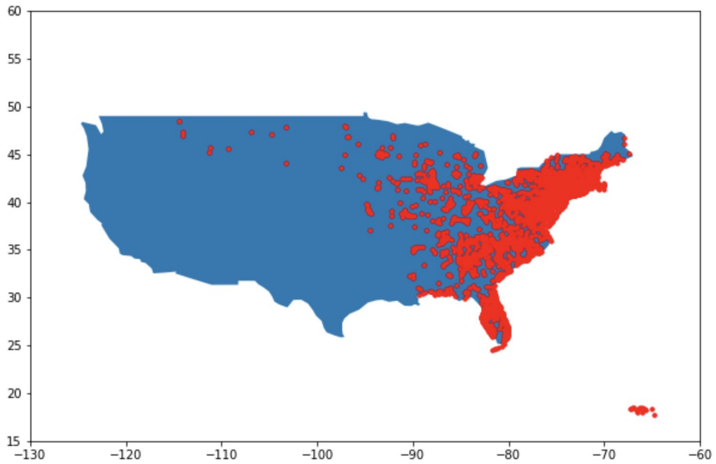
1. Introduction

In this work, we analyzed the Castello's Ace Hardware stores's last two years transaction data and help the company with providing insights into how their sales have fluctuated in the last two years. Along with stating the reasons for the reduction in Sales, we have also suggested ways and approaches to improve the Sales. Later, we have picked one of the ways and extended it into building a model that could implement the approach and come up with solutions for the problem stated. To gather additional information regarding the location, population and weather, we have linked the Stores Data with some external datasets. Our Data Analysis primarily revolves around the scenarios of *bad user experiences* like out of stock of products resulting in the user to visit some other stores(not necessarily Castello's Ace), not placing similar products nearer to each other, and understand how population and weather data could have affected the sales.

2. External Datasets

- **ZipCode data -** The US_Zip_Code Dataset[1] consists of geo-location information like Latitude, Longitude, border information, and total area for Zip codes within United States. We have merged store and customer data with this external dataset to get additional information like Latitude, Longitude of the stores and Customers based on the zip codes. This data is used to visualize the distribution of Customers across different states in the United States.

Distribution of Customers across the United States



- **Demographic Data -** We have used Demographic_Statistics_Data[2] based on zip code to get the information about the population density, total population and city details specific to various zip codes across the United States of America. Using this data, we have calculated the density of population around each Store and predicted for more openings of stores in those regions.

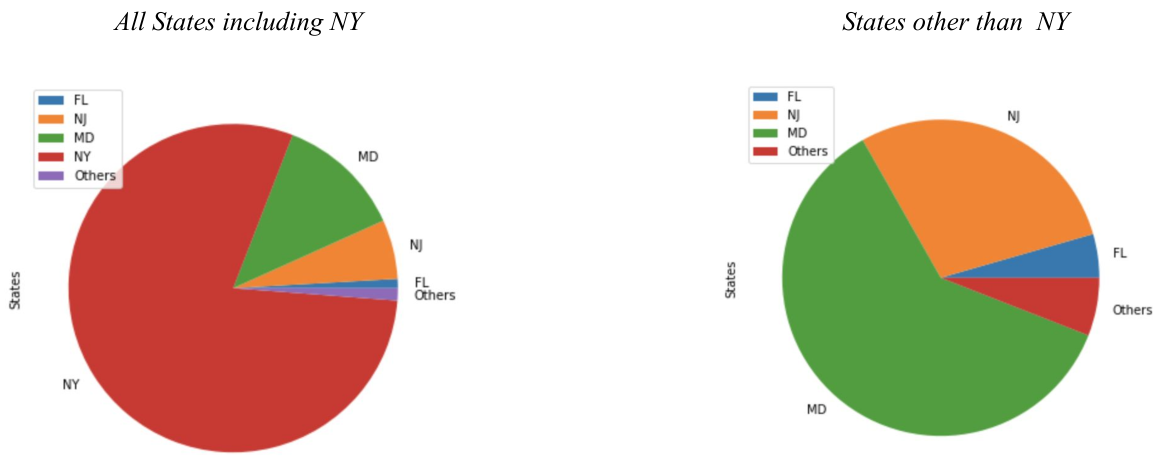
- **Google Maps API -** We have found the *crow_distance* using haversine function which unlike euclidean distance, considers Earth's radius while computing distance between two geographic location. We tried to find the *land_transport_distance* by using the Google_Maps_API[3].

- **Location data -** The Location_Data[4] is specific to zip codes that gives information about the Latitudes and Longitude details, geopoint, city and state a particular zip code belongs to. We have used this location data,

  (i) to understand how many customers are located far away from the stores

  (ii) find the crow distance between two stores or a store and a customer location by using the latitude and longitude details of the zipcode.

Percentage of customers across different states of US



All States including NY

States other than NY

3. Data Preprocessing

- Dealing with the original dataset provided by one of the famous Retail Store Chain was quite a challenging task when 39 columns and 17328044 data points consisted errors and artifacts such as few characters (comma, '%', etc.) with numerical values, high amount of NaN values in few specific columns, column names as in data points (probably multiple datasets were not properly stacked while building this final dataset!), data points stored in higher data types resulting into consuming extreme memory even during simple processing, and many more.

- For faster execution, we corrected datatypes after applying few data cleaning techniques to raw data in the beginning. Based on our objective and data science concepts, we imputed nan values in important columns by considering dependent column values and logical ideas regarding to retail store domain, others were dropped. One of the challenges faced was to come up with the unique mappings between a few pairs of columns having code and name/descriptions. For easy execution, columns having categories were encoded to numerical figures. At the end of preprocessing, dataset size got reduced to 60% of the original data.

- Working with location based objective, more information from the Zip Codes of customers and store was required. As mentioned earlier, external data giving lat., long., values for particular zip codes was added which introduced new columns to our dataset such as Store Zip Code(derived from Store Name), latitude and longitude values of customers as well as stores, Crow distances between store region and customer region.

- To calculate Crow Distance **d** between two points on earth having radius r = 6371 km, we used the given formula of Haversine distance(1) and validated results with the function given by Google Maps API(2).

$$\operatorname{hav}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1-\cos(\theta)}{2}$$

$$\begin{aligned} d &= 2r \arcsin\left(\sqrt{\operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)}\right) \\ &= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \end{aligned}$$

where

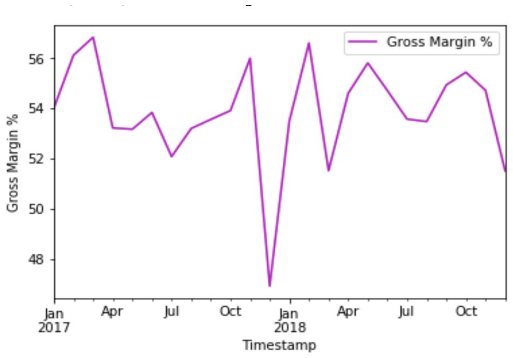
•φ_1, φ_2: latitude of point 1 and latitude of point 2 (in radians),

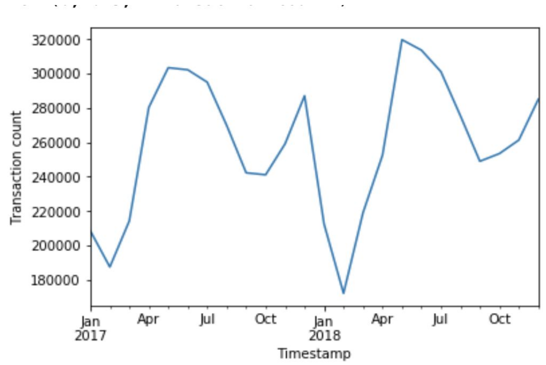
•λ_1, λ_2: longitude of point 1 and longitude of point 2 (in radians).

4. Analyzing the reasons for the drop in Sales

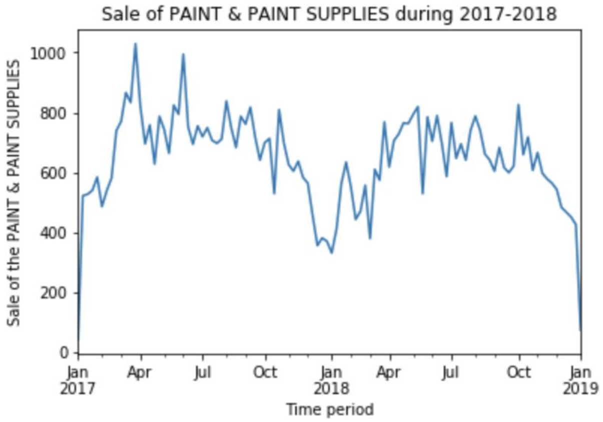
1. Sales affected during the Winter (December - January)

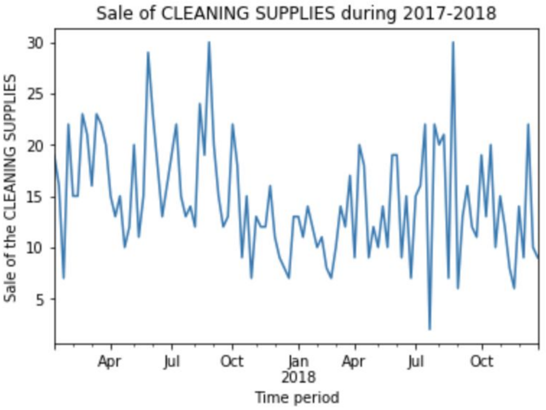
There has been a significant reduction in the Gross Margin% of sales and number of transactions during the *winter months of December-January*. The time series analysis of no of transactions and Gross Margin% is shown below



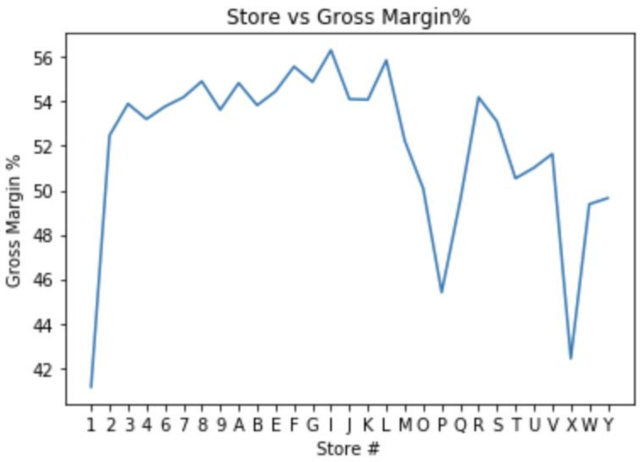


This can be attributed to various reasons, one such is **weather**. *For example*, *heavy snow did not encourage customer to buy products like paints and brushes*. From the below two figures, we can notice that the sales of the items in Department 'Paint & Paint Supplies' during the winter is less, whereas in 'Cleaning Supplies', there is always a steady sale of the items across all the months. Since the department 'Paint & Paint Supplies' accounts for more than 50% of the total transactions of the Store, there is a significant drop in the sales during this period.



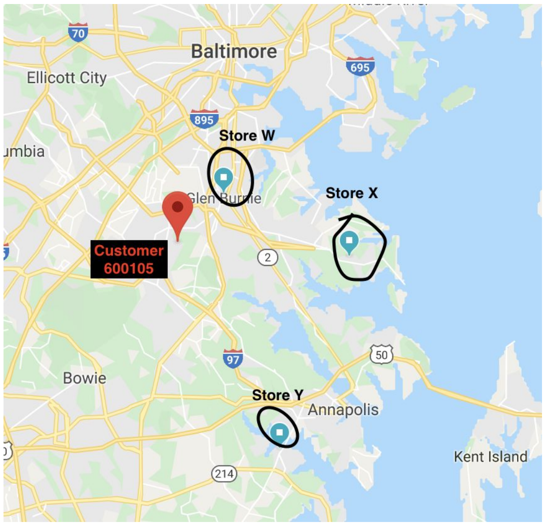


2. Stocking not done after product out of stock



All the store locations do not contribute equally to the profit of the hardware company. From the graph shown, it can be seen that the store X and store P contribute relatively poor towards the Gross Margin when compared to other stores.

On scrutinizing the customers buying in the Store X, we got an interesting insight that *there is no frequent stocking of the items (after they become out of stock) in store X*, which is forcing even the customers in that same location to buy in a store which is very far from their location



For our analysis, we have considered a single customer 600105, who buys items from only three stores W,X and Y of Costello's Ace Hardware. The Customer 600105

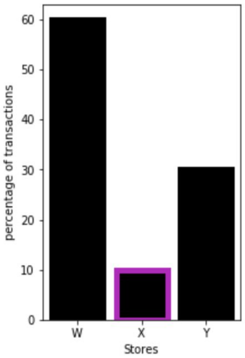
- buys 60% of the products from Store W as it is nearer
- buys only 10% of the products from Store X
- buys 30% of the products from Store Y even though it's far

The reason for this is that items in Store X are not stocked frequently.

Store #	ZipCode	Distance from Customer Location
W	21061	6.4 kms
X	21122	13.8 kms
Y	21037	21.21 kms

- The total no of transactions in Store X is very less compared to Store W and Y
- From the Customer 600105 table, we can say that total 81 types of times were bought from store X whereas 292 times of items were bought from Store Y. *Out of those 292 items, 285 items were not available in the Store X*
- Similarly for Customer 600607, 50 of the 55 items bought at Store Y were not stocked at Store X
- The Customer in our example has been loyal to Castello's Hardware Store by buying from a farther Castello's Ace store. But not all customers could be expected to do the same. There could be some customers who would *leave Castello's Ace Hardware store and buy from a different hardware store* which could have caused loss in their sales.

No of transactions at each stores of Customer 600105



Customer 600105 [zip code - 21144]

Store #	Zip Code	Distance from customer	No of transactions	No of items bought	No of items which were not there in X but in Y
W	21061	6.4 km	170	538	-
X	21122	13.8 km	29	81	**285**/292
Y	21037	21.21 km	82	292	

Customer 600607 [zip code - 21122]

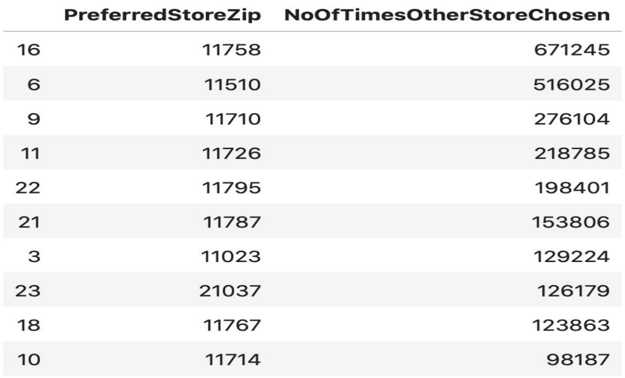
Store #	Zip Code	Distance from customer	No of transactions	No of items bought	No of items which were not there in X but in Y
W	21061	9.346 km	99	252	-
X	21122	0 km	18	39	**50**/55
Y	21037	21.37 km	20	55	

3. Some Customers prefer farther store nearer to store because of out of stock in nearer store

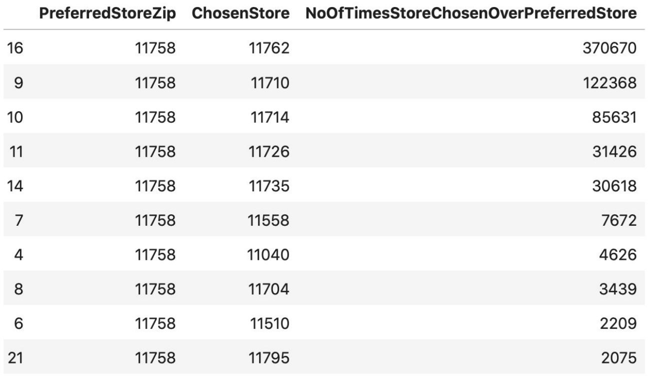
Not all customers are preferring to buy the product from the store that is located close to their location.

- By using ZipCode data and crow distance function, we first figured out the store that is located close to the customer location.
- Left side table shows the number of times the customer choose a farther store over the nearby preferred store.

- From further analysis, data reveals that the store located at the zip code 11758 is ignored by the customer for nearly 671245 transactions i.e customer have chosen different store over this store even though they stay close to this store.
- Deep diving into the data reveals that customers mainly tend to choose store 11762 and 11710 over the store at 11758.



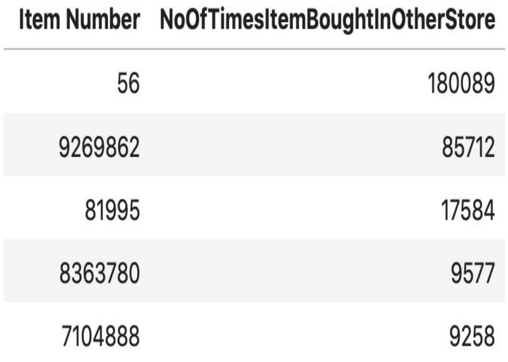
	PreferredStoreZip	NoOfTimesOtherStoreChosen
16	11758	671245
6	11510	516025
9	11710	276104
11	11726	218785
22	11795	198401
21	11787	153806
3	11023	129224
23	21037	126179
18	11767	123863
10	11714	98187



	PreferredStoreZip	ChosenStore	NoOfTimesStoreChosenOverPreferredStore
16	11758	11762	370670
9	11758	11710	122368
10	11758	11714	85631
11	11758	11726	31426
14	11758	11735	30618
7	11758	11558	7672
4	11758	11040	4626
8	11758	11704	3439
6	11758	11510	2209
21	11758	11795	2075

- We continued our exploration further to identity which item is bought by the customer from the farther stores than the nearby store.

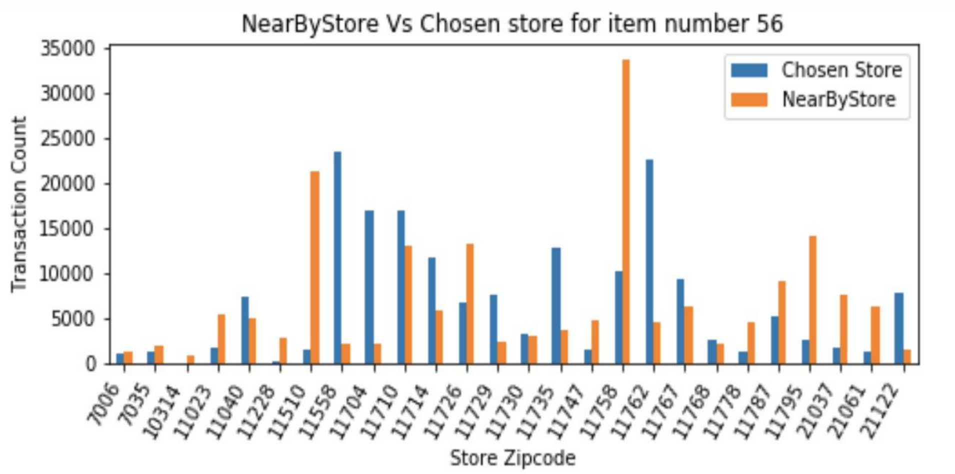
- From the table, it looks like the customer is willing to buy the item ('item number' with value 56) from store 11762 and 11710 instead of the nearby store 11758.



Item Number	NoOfTimesItemBoughtInOtherStore
56	180089
9269862	85712
81995	17584
8363780	9577
7104888	9258

- One probable cause for the customer to choose store 11758 may be due to the out of stock. Stocking the item could attract the customer to prefer this customer and the ones who can’t commute easily to prefer this nearby store over other stores.

- If we could get stock data, then it would be more helpful to suggest a product that can be stocked to improve the sales in the store.

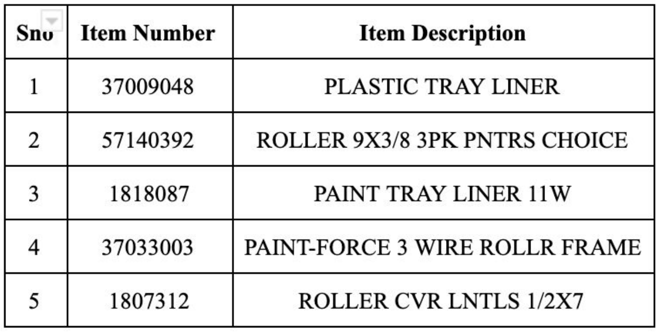


5. Suggesting ways to improve the Sales

- **Place similar or frequently bought together products nearer in the stores**

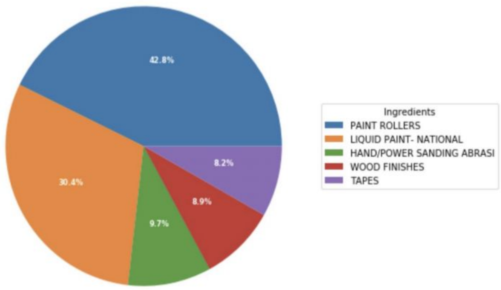
  Place similar products together in the stores so that in case of out of stock of the products, the customer can pick the closely related products without having to search for it. Frequently bought together products should also be placed together to improve the customer experience.

Fig 5a. Similar Products of PAINT ROLLERS



Sno	Item Number	Item Description
1	37009048	PLASTIC TRAY LINER
2	57140392	ROLLER 9X3/8 3PK PNTRS CHOICE
3	1818087	PAINT TRAY LINER 11W
4	37033003	PAINT-FORCE 3 WIRE ROLLR FRAME
5	1807312	ROLLER CVR LNTLS 1/2X7

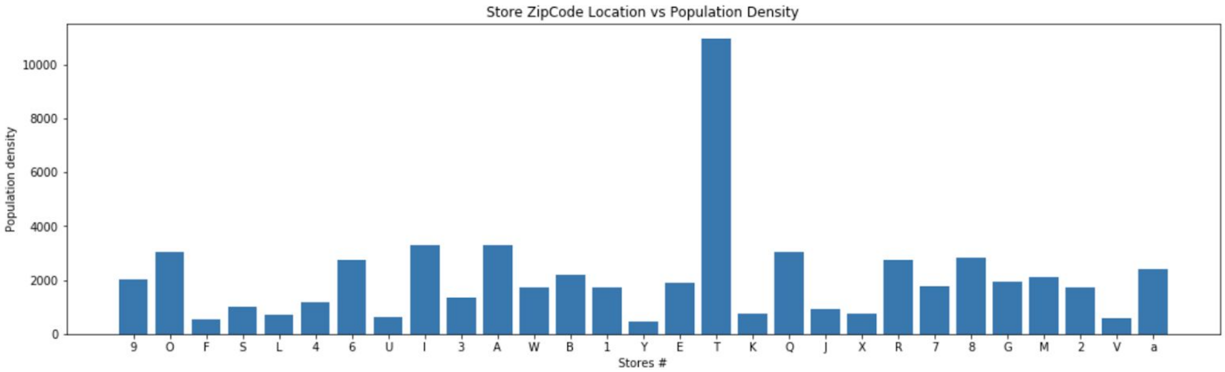
Fig 5b.Frequently bought together products with PAINT ROLLERS



From Fig 5a, we can notice that if one of the PAINT ROLLERS is out of stock, then the customer can actually pick from the other paint rollers available in the store, if they are placed nearer. From Fig 5b, we can notice that PAINT ROLLERS are bought together with Liquid Paints, Tapes etc and shows a high correlation between them

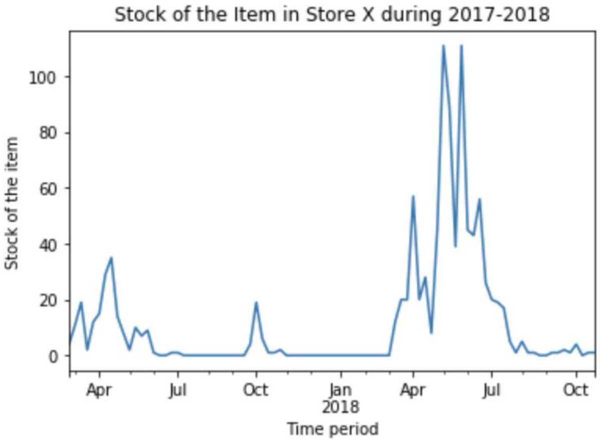
- **Expanding the stores in the area of high population density**

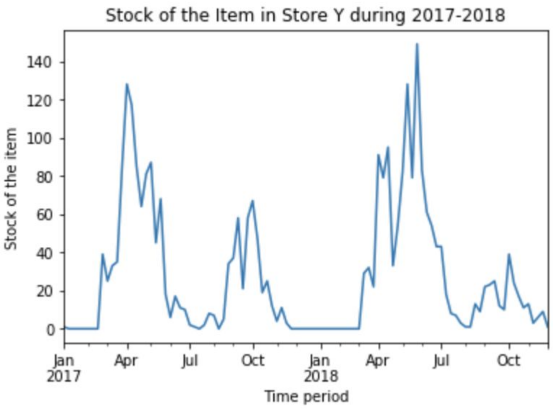
  We have used the demographic data with the Castello's Ace stores data to get the number of customers living in the surrounding area of each of the stores. Though the total population near '*Store a'* is more, the population density of the location of '*Store a'* is less. The population density near the ***Store T*** is more which means there should be more opening of branches of the stores in that particular Zip Code



- **Stock products before they run out of stock**

  Below are the graphs stating the stock of a random product in *Store X* and *Store Y*. In *Store X*, the product is stocked very less frequently, whereas in Store Y, the product is stocked at regular intervals.



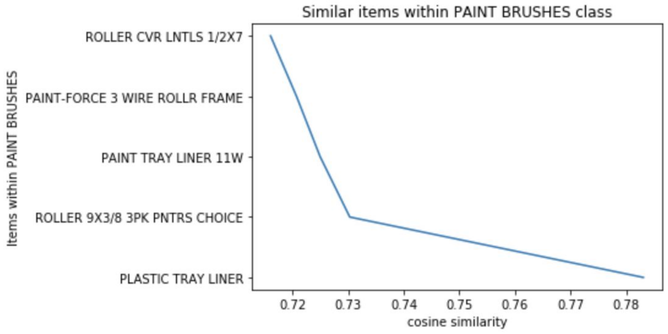


6. Implementation & Results

We have implemented a model for one of our suggested ways to improve sales - **Suggest similar products when out of stock by placing then together** by using Word2Vec model. We have trained our Word2Vec model with a vocabulary of 147066 words compared to 2156 words during our project proposal.

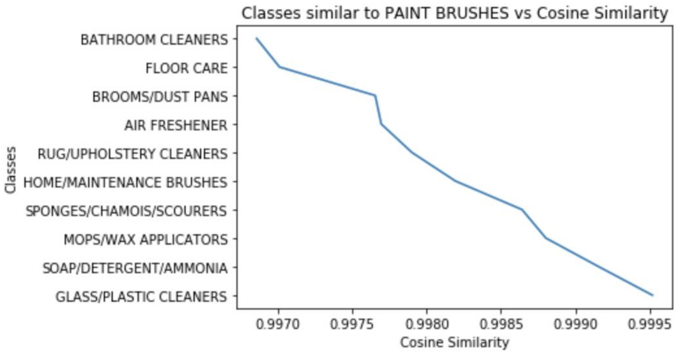
Model 1 (Similar items within a class)

- In this approach, we have tried to predict the items within a class which are similar to each other.
- To build this word2Vec model, we have considered many items features specific to class like item department name, class name, Actual Price, etc.
- The trained model has more words to the vocabulary than the model 2
- From the graph, we can notice that the item *'Paint Roller'* of PAINT BRUSH class has items like '*Rollers', 'Paint Trays' and 'Liner*'(which are used for the same purpose) similar with it.



Model 2 (Similar classes to a class)

- In this approach, we have predicted how the classes within a department are related.
- This is a high level prediction of the Model 1, using which classes can be grouped together.
- By this grouping we can shift the locations of similar classes within a department together.
- From the graph we can notice that the product with greater value of cosine similarity are more close to the class PAINT BRUSH.



7. Next steps

- Implement a model to perform our second objective (ie) to predict when an item will be out of stock.
- Decide an evaluation metrics for the model and evaluate the model more effectively.
- Analyse and include the newly provided data(last 4 years data) to train and improve the accuracy of the model.

8. References

[1] https://developers.google.com/maps/documentation/javascript/reference/geometry#spherical.computeDistanceBetween
[2] https://stackoverflow.com/a/1502821
[3] https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
[4] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In EMNLP, volume 14, pages 1532–1543, 2014.
[5] https://arxiv.org/pdf/1607.07326.pdf