# CSE 519 PROJECT PROGRESS REPORT:

# RETAIL SALES DATA ANALYSIS

## INTRO: WHAT IS RETAIL DATA ANALYSIS?

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. 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.

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

What we have is a very large dataset (17,328,044 rows/entries 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.
- 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.?
- Time-series predictions on seasonal sales - find out the best way to maximize profits and sales while minimizing costs varying for different seasons.
- Location and Demographic analysis - Finding out which stores have differential sales based on location and other demographic factors.
- Weather and Holidays -- Weather data is available on the internet on various weather websites such as Yahoo Weather, Google Weather etc., which will be used as an additional feature to see the trends of shoppers e.g. When it's damp season, the sale of Pesticides may increase due to increasing pest attacks.
- Future Stocking predictions – Monthly product wise sales seen so as to predict future sales in those months and stocking up accordingly beforehand.
- Finding lemon products i.e. the products being returned as larger percentage of their sales (in some cases maybe even more, thus negative sales!! )
- Some other trivial but Interesting insights : Does cashier scan bigger items earlier than smaller items?

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.

## DATA PREPROCESSING :

- 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.
- 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.

## DATA SETS :

For now, we generated a dataset using the data from:

- Costello Ace data (Given for 2015-16 and 2017-18)
- Stock Market Data
- Weather Data
- Demographic Data

## SOME INSIGHTS INTO DATA (Provided by Costello's Representative) :

The visit by Costello's representative was a great experience providing a more detailed look into the project Data and the Company's expectations from the analysis. Some useful insights that helped understanding the data better are:

- 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.

- 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

- 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 code (total 18 departments)
    - Dept assigned (products grouped)
    - Product Class – is less specific than department name
    - Fineline code - class broken down further

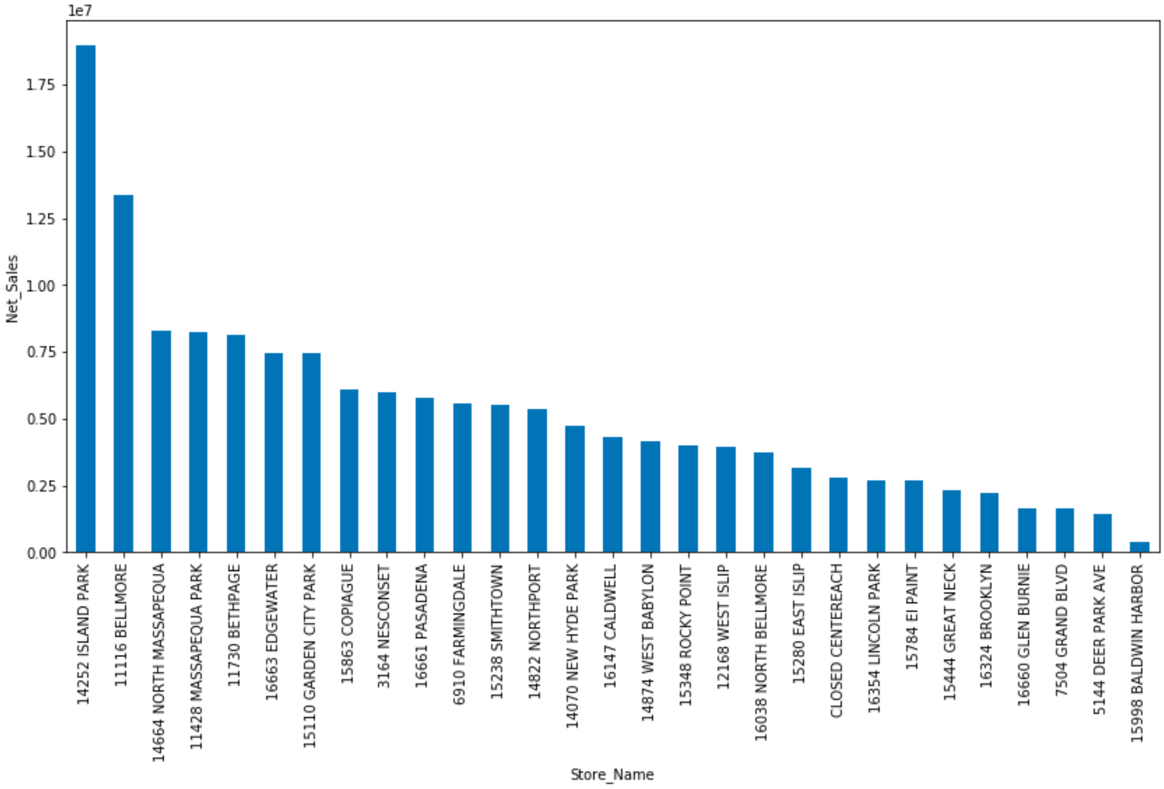
- Zip code – represents customers residence zip code if customer is registered, otherwise store's zip code where it is situated
- Loyalty ID - Account number that is unique to customer

Although meeting with the Costello's representative helped us understand the data more, there is still some confusion around some of the features and their values.

## IMPLEMENTATION:

Part 1: Store Wise net sales analysis - Finding out which stores have differential sales based on location.

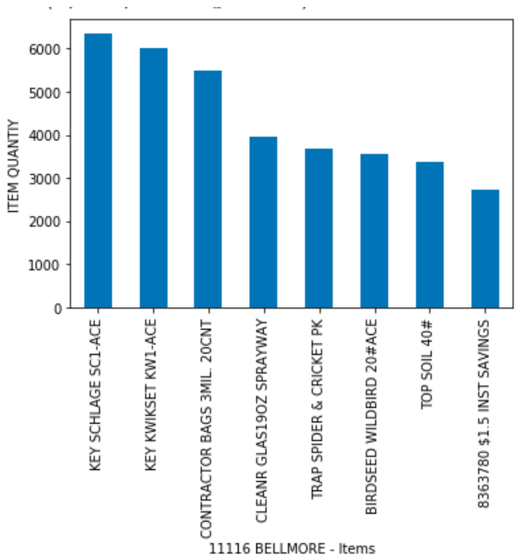
- Converted Net sales feature from Object type to float type for summation and grouped Store wise to calculate net sales for individual store. Also, sorted the net sales in non-decreasing order to show which stores are performing the best vs the least performing ones in terms of Revenue generated.

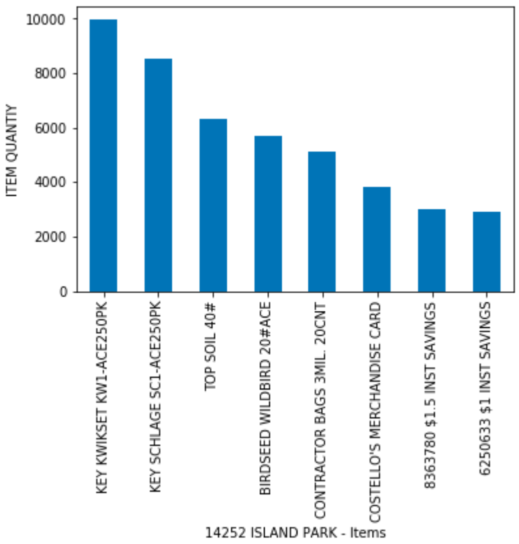


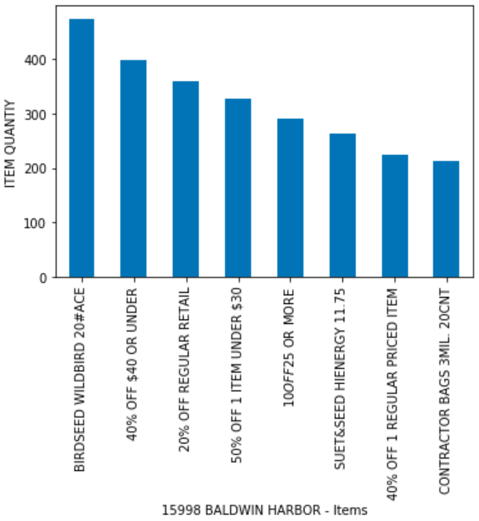
The bar chart above shows that the store at Island Park is performing the best with respect to other stores which can be verified from the fact that the number of transactions also are highest for the said store which is approx. 1.8 Million transaction records. Following this are the stores at Bellmore and North Massapequa. Interestingly, the store at Pasadena has higher number of transactions than store at North Massapequa. From this, it can be inferred that although number of transactions were high for Pasadena store, it didn’t amount to higher Dollar value net sales which maybe the result of selling high number of cheaper items than at store North Massapequa. The least performing store is not actually a store but the company’s harbor where it is logical to have few transactions.

Part 2: Store Wise trend analysis: Predicting stores top 10 product items sold and it’s inferences

- For this, Store Name has been grouped to get items being sold in that particular store and filtering out the Most sold top 10 items by their sold quantity. This analysis will enlist the products which are hot cakes in the respective stores. Also, to keep results practical, Fasteners and Donations have been excluded out of the plot because Fasteners are the most sold item at each store and the Donations also doesn’t help in our analysis of the items sold to general customers.





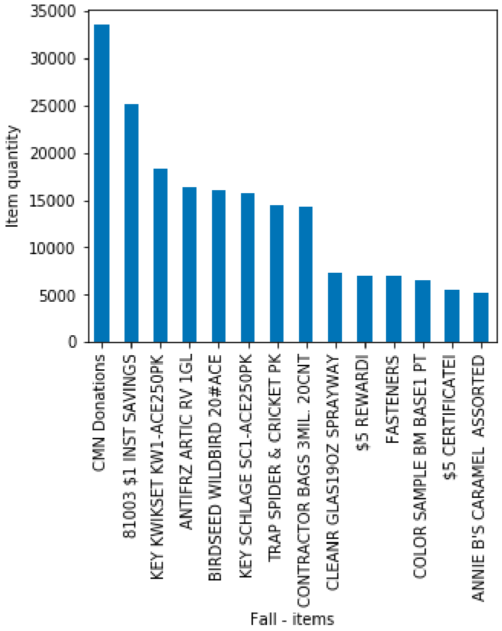


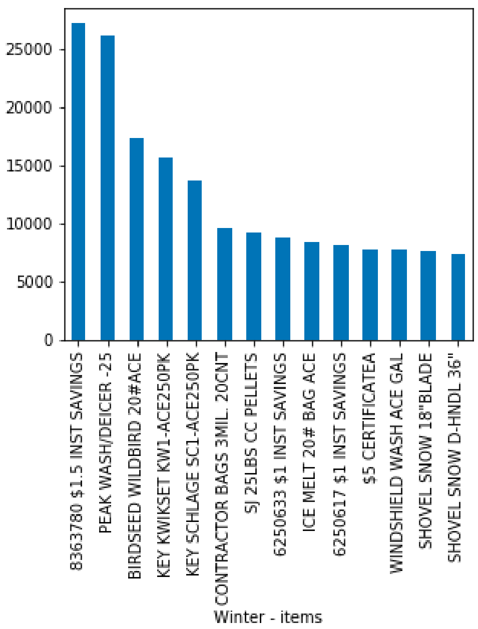
The above sample of 3 plots are of stores out of all the locations namely Bellmore, Island Park and Baldwin Harbor. Similar plots can be generated for all stores. The trends in above plots can help us infer the top selling items and this analysis can be further correlated with demographic data to predict location wise buying trends of the population according to the various demographic features of that place.

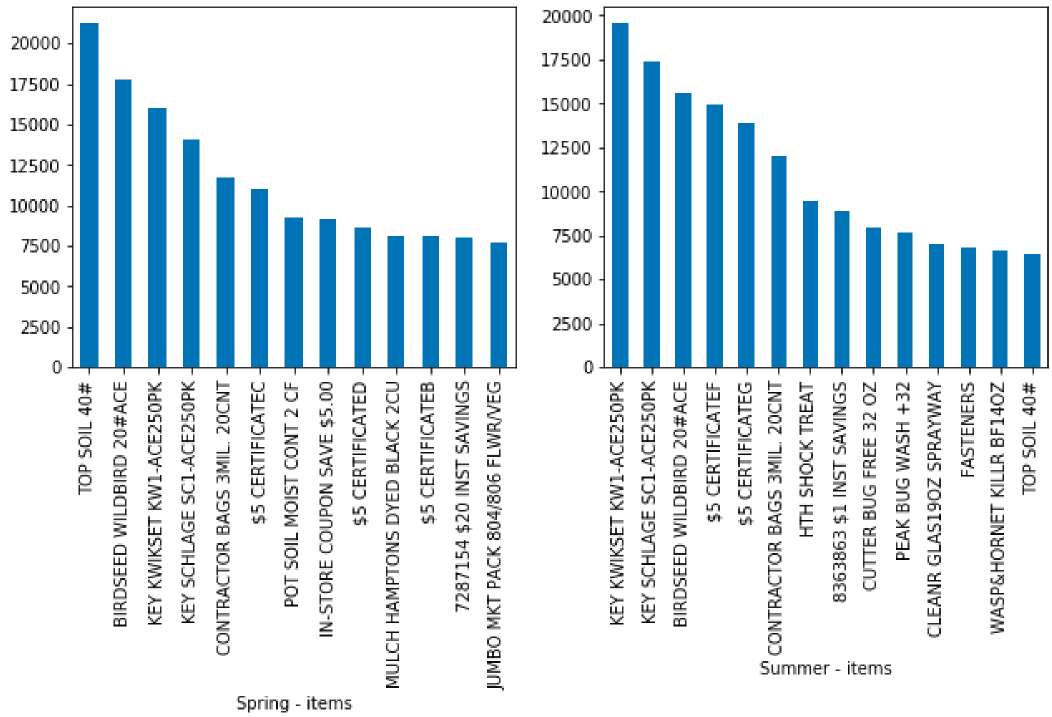
<u>Part 3: Seasonal Sale analysis:</u> Analysing sales for different seasons:

- Fall – September to November
- Winter – December to February
- Spring – March to May
- Summer – June to August

Grouping data firstly on monthly basis and then Season wise as depicted above to get the top items being sold in that particular season.







The above plots show the season wise number of items sold by the stores in different seasons. The top items of different seasons show the common as well as the exclusive items being demanded in that season only. So, this information can help in estimating the stocking needs of the Stores and thus help in maintaining the warehouses accordingly.

## Moving Forward:

We have determined some opportunities and markers to work on furthering the retail analysis. Some of them are listed below:

### Next Steps :

1. Further Sales Analysis: If sales are declining, timely decisions such as to cut prices, market more, or discontinue the product can be made. If an item is selling off the shelves, one can be sure to stock inventory accurately across channels. Through this data, one can gain valuable insight into the operations. Through key capabilities, like real-time updates and data visualization, one can make better informed business decisions.
2. Stock Market Analysis : Seeing the economic conditions prevailing at the time of sale can 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.

3. Demographic analysis : Finding out which stores have differential sales based on location (plots made) and other demographic factors (yet to be implemented). Taking the Part 2 of the Implementation further with the demographic data would be the next step to achieve this.
4. Weather Correlation : Diving a bit deeper into the seasonal analysis, daily weather reports can be correlated to the net sales for that day to see if and how the weather affects the footfall and thus the sale of the merchandise.
5. Interesting Insights: Categorisation of products size wise to analyse whether large or small products are being scanned earlier than the other by the clerk. Also, finding products being returned more than the actual sales to know the most stolen and/or returned products.