CSE 519 PROJECT PROPOSAL:

RETAIL SALES DATA ANALYSIS

INTRO: WHAT IS RETAIL DATA ANALYSIS?

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

BACKGROUND RESEARCH

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

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

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

Purchasing Power of an individual and thus will help in better understanding of the retail sales of the stores at various locations depending upon the demographic trends of the location.

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

Paper [5] discusses the problem faced while modelling marketing analysis that is to select key variables from a large number of possibilities. In analysing high-dimensional marketing data, the problem faced is that valuable predictors of consumer behaviour are often hidden in a large number of useless noisy variables. When the dimensionality increases with the integration of intra- and inter-categorical information, the number of unreliable predictors which are correlated with valuable ones also increases rapidly. This paper develops a four steps methodological framework to overcome the problem. The method consists of the identification of potentially influential categories, the building of the explanatory variable space, variable selection and model estimation by a multistage LASSO regression, and the use of a rolling scheme to generate forecasts.

GOAL AND APPROACH

Certain directions which our team plans to explore into are:

- We will be doing a **Location and Demographic** Analysis -- Finding out which stores have differential sales based on location and other demographic factors.

“Population change is a key factor influencing local retail sales performance. From one year to the next, area population gains or losses alter the number of potential shoppers in the region. In the longer term, population trends reflect the general economic climate of the region”.[6]
With preliminary exploration, we were able to find demographic data about the locations where these retail stores are situated based on the Store Address given in the Data. e.g. Snippet of the type of information available for each Zip-Code on City Data Website[7] gives us information about Population, Median Household Income, Age, Gender Ratio, Median House Value etc. **Gender ratio** could also help in finding whether the store products are gender biased i.e. Men buying more hardware tools and women having less products to be related with thus having lower footsteps which can be analysed to target specific genders.

- **Future Sales Predictions**: **Time series modelling** of sales - Based on the training and modelling, future predictions could be brought up to know which items to be stocked in higher quantity than others and vice versa.

- **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. Similarly, we hope to see increasing sale trends during holiday periods or such occasions like Thanksgiving when sales would increase due to the Gifts people are buying for their family and friends.

● **Time of the Day Analysis:** _Analysing which times in the day are the rush hours for the shoppers as knowing this would help know when the Stores should run Promotions so as to effectively target more people with higher cost-benefit ratio.

● Analysing **Key performance indicators** like Net Sales Units, Net Sales, Cost, Gross Margin, Gross Margin % and sales growth will tell the story of store performance and help make profitable decisions. Understanding what customers aren't buying: poor displays or not enough sales reps could be among the many reasons why customers aren't buying what is being sold.

● **Product embeddings** - Build distributed representations of products (embeddings),
using the network of co-purchased items. Doing these embeddings would provide insights into which products are similar or are bought together or not, so that the particular products could be targeted for improvement in sales by changing their shelf positions or otherwise promoting them i.e. **Promotional recommendations** in a tailor made fashion for usual Shoppers to make their user experience better.

GETTING STARTED:

● Pre-processing and cleaning of the data - solving anomalies

● Perform data analysis to identify trends in shopping over the whole year

● Analyse the shifting preferences of shoppers from certain products categories to other

● Ranking function to analyse the popularity of some products over the other ones.

● **Problems**: As mentioned above, the spelling errors in String type datasets can be a hindrance while trying to corporate it into the final dataset. E.g. the amounts in the dataset when exceed 3 digit have used a comma (,) to represent 4th digit such as 1,270. This needs to interpreted as 1270 for the analysis.

EXPLORATORY DATA ANALYSIS

Fields in Data include:

- Date
- Transaction Time
- Customer Number
- Sales Header
- STR
- Item Number
- Item Description
- Class Code
- Class Name
- Department Code
- Department Name.

Apart from the given data we plan on scraping and collecting some other data from the web such as:

- **Demographic data**: Population, Median Household Income, Age, Gender Ratio, Median House Value etc. for each store.
- **Weather** information for the sales period.

For preliminary data exploration, we took about **450,000 records** from the **~18million records** we have, and plotted some charts to have a better understanding of the data at hand:

1. Totalling all the sales by summing all transactions from the records shows a concentration of sale in some hours of the store than other hours.

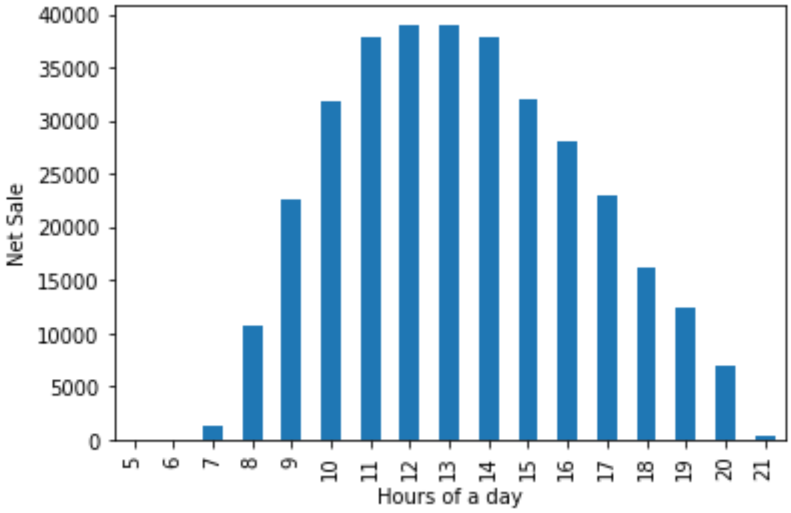

Fig 1. Net Sale during the hours of the day.

2. Plotting the various products being sold and their contribution to the unit sales.

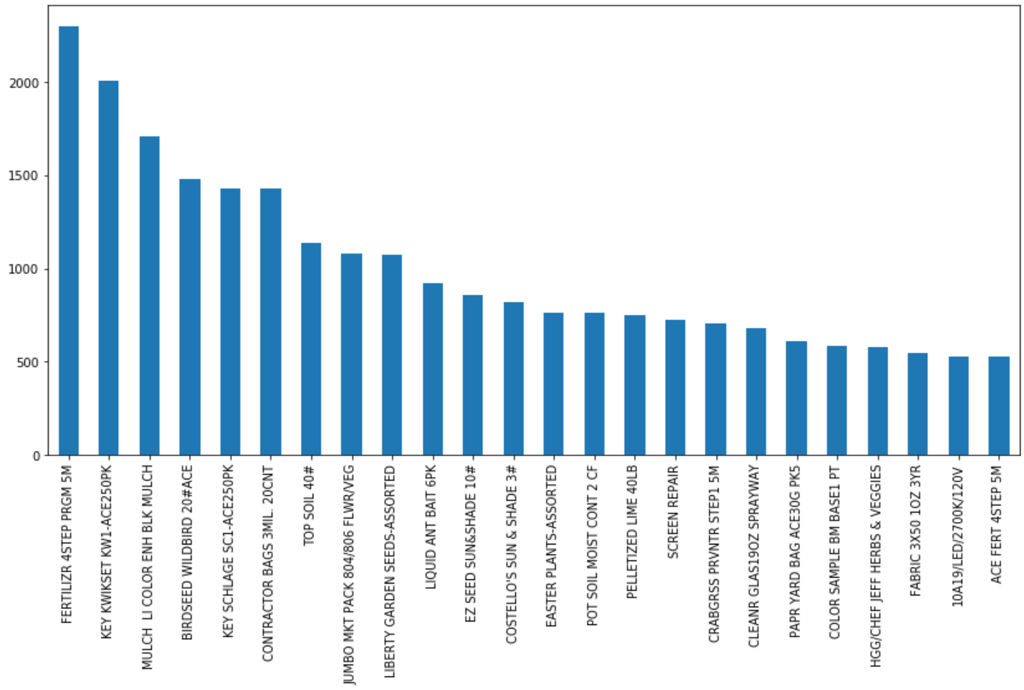

Fig 2. Top 25 Products being sold

3. Store wise contribution to Net Revenue of the Costello chain of Retail stores.

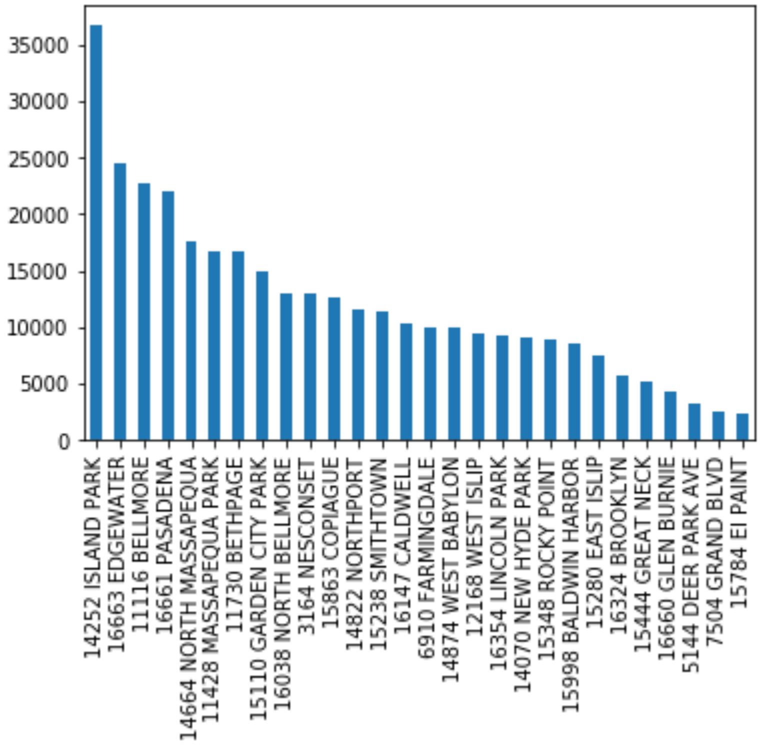

Fig 3. Store specific sales

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

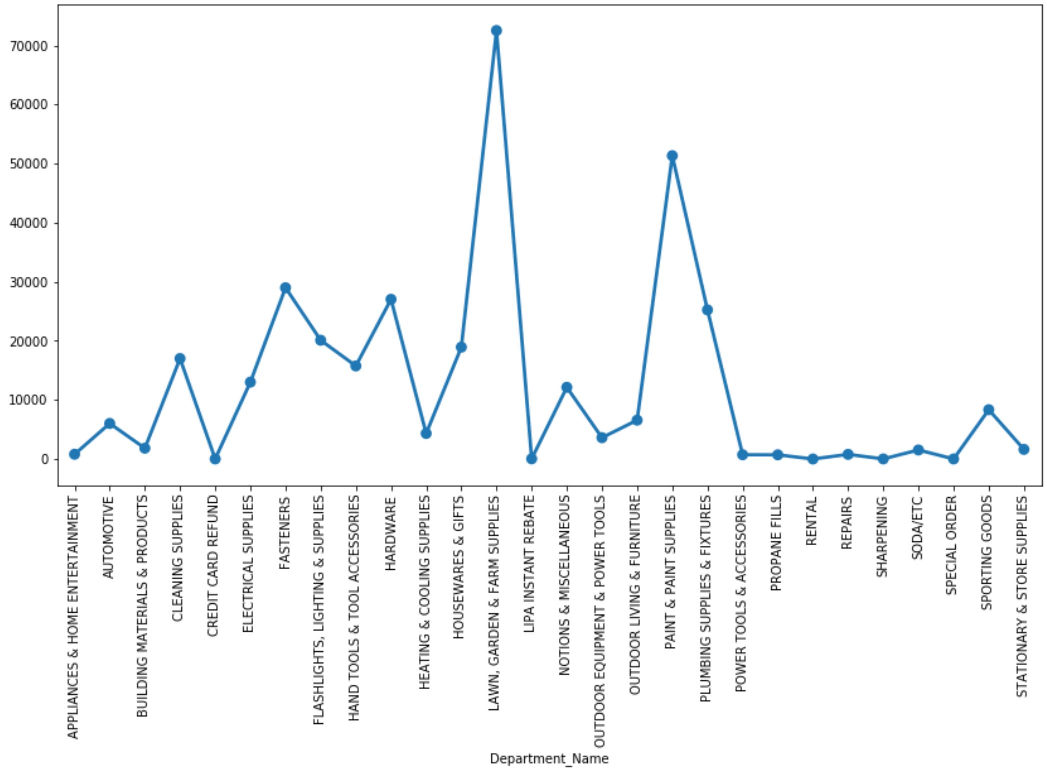

Fig 4. Department wise sales

5. Seeing which top products dominate the sale spectrum.

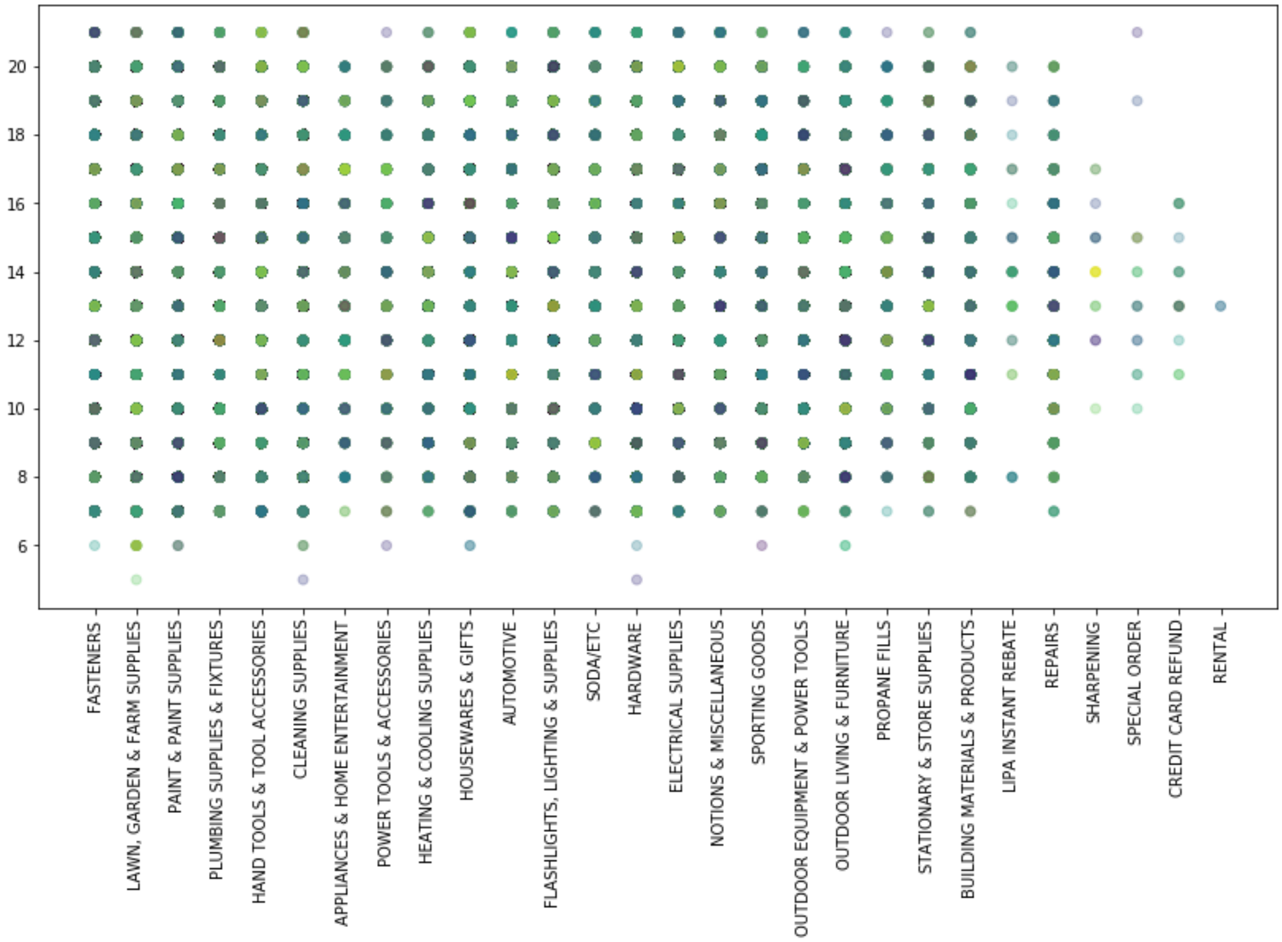

Fig 5. Top 25 Products and their contribution to sales.

Moving Forward

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

Some questions we hope to some put light on:

- How are the customers responding to promotional signages?
- Are they struggling to find information that is being provided and promoted ?
- Establish how consumers are interacting with store and identify ways to improve their experience.
- Develop employees and set goals: Train employees to always put customers first by setting goals and behaviours that reward that.

REFERENCES

1. Youngjin_Bahng,_Doris H._Kincade "The relationship between temperature and sales" International Journal of Retail & Distribution Management", 2012
2. Thomas PVan Boeckel PhD, Sumanth Gandra MD, Ashvin Ashok MPP, Quentin Caudron PhD , Prof Bryan T Grenfell PhD, Prof Simon A Levin PhD, Prof Ramanan Laxminarayan: "Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data",2014.
3. Paul Springfield, Edwin Blake, David Stern : "Method for performing retail sales analysis", Patent US8214246B2 United States
4. Ilan Alon ,Min Qi , Robert J. Sadowski , "Forecasting aggregate retail sales :: a comparison of artificial neural networks and traditional methods", Journal of Retailing and Consumer Services , 2001
5. Shaohui Ma, Robert Fildes, Tao Huang, "Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information", European Journal of Operational Research, 2016
6. Retail Trade Analysis Report Fiscal Year 2018 of Winnebago County by Iowa State University Department of Economics.
7. Website, city-data.com ("http://www.city-data.com/city/Stony-Brook-New-York.html")