Data Science Capstone Project 1

Customer Loyalty and Recommender System

Keyur Patel

March 2019

Introduction:

When there was no recommender system, people tend to buy products through recommendation given by friends or people they have trust. This was the primary method of purchase when they had doubt about products. But in the digital age, that circle has expanded to include online sites that utilizes some sort of recommendation engine. The recommendation engine filters the data using different algorithms and recommends the most relevant items to the users. It first captures the past purchase behavior of a customer and based on that, recommends products which the users might like to buy.

Problem Case:

With e-commerce boom, we need competitive advantage and more personalized experience for our stakeholders. With increasing numbers of brands, products, online retail stores, growing number of users and changing environment, it is important for retail client to get insights into customers and product basis. Our business decisions influenced by analytics can drive our marketing efforts to increase customer retention, build loyal relationship with Users, and increase revenue and User engagement.

Users of the client system have a huge choice of products to purchase but limited time. The real challenge is to provide recommendations of products that are relevant to the users, help users accurately discover brands that they might be interested or brands them might not know they would like. Filtering brand from entire catalog of brands which are relevant to the users is basically the key focus.

Business Case:

With a recommender system, we can gain valuable insights into customer and product bases. In many ways, recommender system is beneficial for both customers and marketers. Customers get personalized recommendations on relevant products that are valued and helpful, and marketers can enhance offers in ways that proactively build better customer relationships, retention and sales. The result is increased loyalty, customer Satisfaction, frequency of visits, return purchases, customer lifetime value and reduced churn. The business strategy is to make relevant recommendations through personalization of products and convert the recommendations to the checkout process. The recommender system is expected to make a significant contribution in increasing sales revenue by 25%.

Business Challenges:

1) Cost – Cost of resources to maintain huge data
2) Quality - services and tools
3) Competition:
    a) From Market Segment:

        Selection, price, availability, convenience, information, discovery, brand recognition, personalized services, accessibility, customer service, reliability, speed of order fulfillment, ease of use, and ability to adapt to changing conditions, as well as customers' overall experience.

b) Others:

Vendors, distributors, manufactures, online, offline, producers, distributors, web search engines, social networks, web portals, companies that provide e-commerce services, shopping websites, advertising, companies that provide order fulfillment and logistic service for themselves and 3rd parties.

Stakeholders:

Shareholder, Employees, Suppliers, Customers, Government authorities, Financial Institution, Creditors, Contractors, Investors

Dataset:

The dataset is taken from Kaggle's Acquire Valued Shoppers Challenge. Consumer brands often offer discounts to attract new shoppers to buy their products. The most valuable customers are those who return after this initial incented purchase. With enough purchase history, it is possible to predict which shoppers, when presented an offer, will buy a new item.

Challenges:

There are some challenges associated with this data set. The entire dataset is huge and needs sufficient hardware resource to process. We have taken a subset of this data i.e. a month of data from January month. There are **1048574** rows and **11** features (columns).

Features:

The data column fields are described below:

id – unique customer id
chain - An integer representing a store chain
dept - An aggregate grouping of the Category (e.g. water)
category - The product category (e.g. sparkling water)
company - An id of the company that sells the item
brand - An id of the brand to which the item belongs
date - The date of purchase, productsize - The amount of the product purchase (e.g. 16 oz of water)
productmeasure - The units of the product purchase (e.g. ounces)
purchasequantity - The number of units purchased
purchaseamount - The dollar amount of the purchase

Data Wrangling:

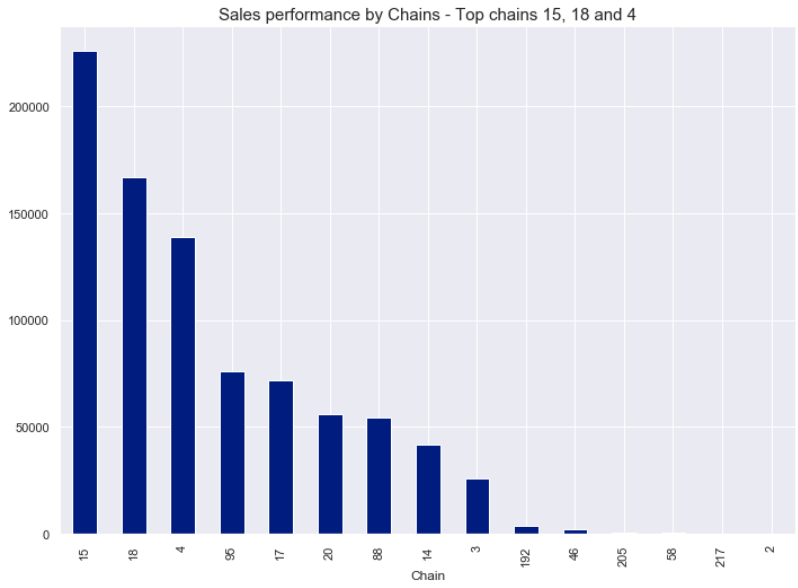
- **Missing Values:**
- We found that there are 11257 missing values only in column **productmeasure** and we dropped the rows with missing values.

- **Negative Values**
- There are 28417 negative values in Purchase amount column that indicates the return of products.

- **Data Pre-processing:**
- There are 14030 unique customer ids. Customers having less than 5 occurrences were removed as their buying pattern is not sufficient for recommendation. There were 13684 customers having less than 5 transaction instances.
- We also removed the Brands that were bought by customers for less than 5 times. There are 731 unique brands that are bought most frequently.
- For simplification, we removed the negative values from the dataset.

Exploring Chains:

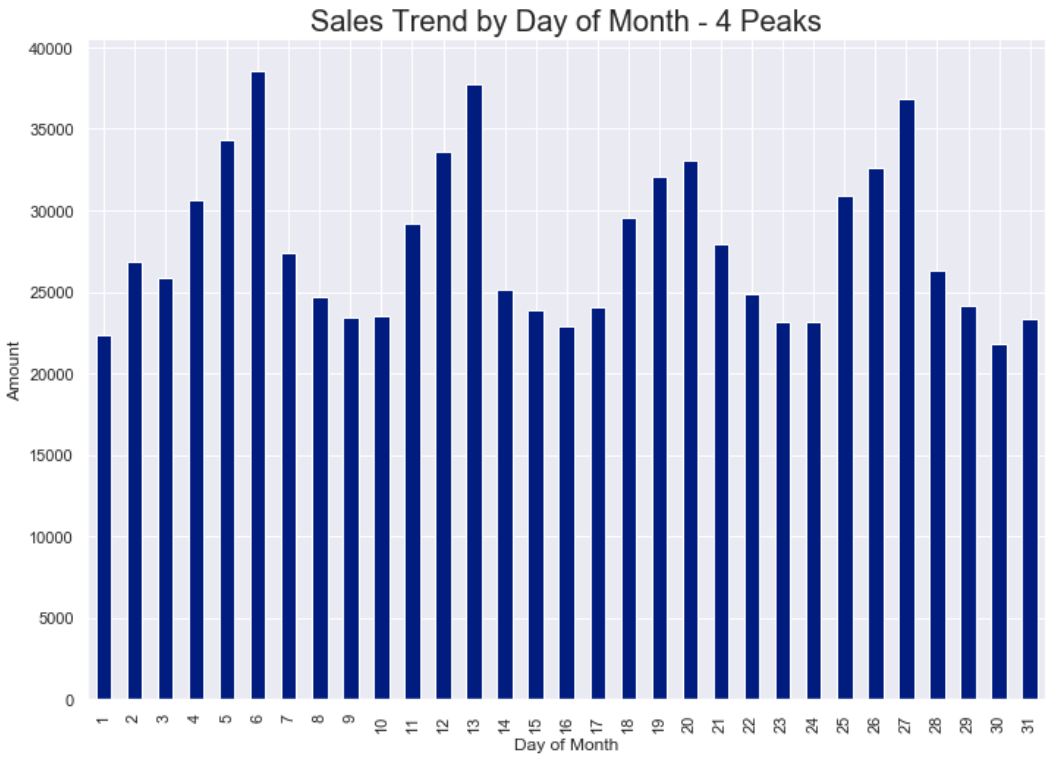
We explored the sales volume with respect to chains. Out of 15 unique chains, chain number 15 provides highest sales followed by 18 and 4. Chain number 2 provides least sales volume. The Client can think of investigating the reason of less sales at some chains with additional parameters and can provide additional offers to those chains.



Exploring Transactions by Day, Week and Weekdays:

Transactional Data Analysis – By Day of Month:

We explored the sales amount for each day of the month. 6th January is the day with maximum sales amount followed by 13th and 27th January.

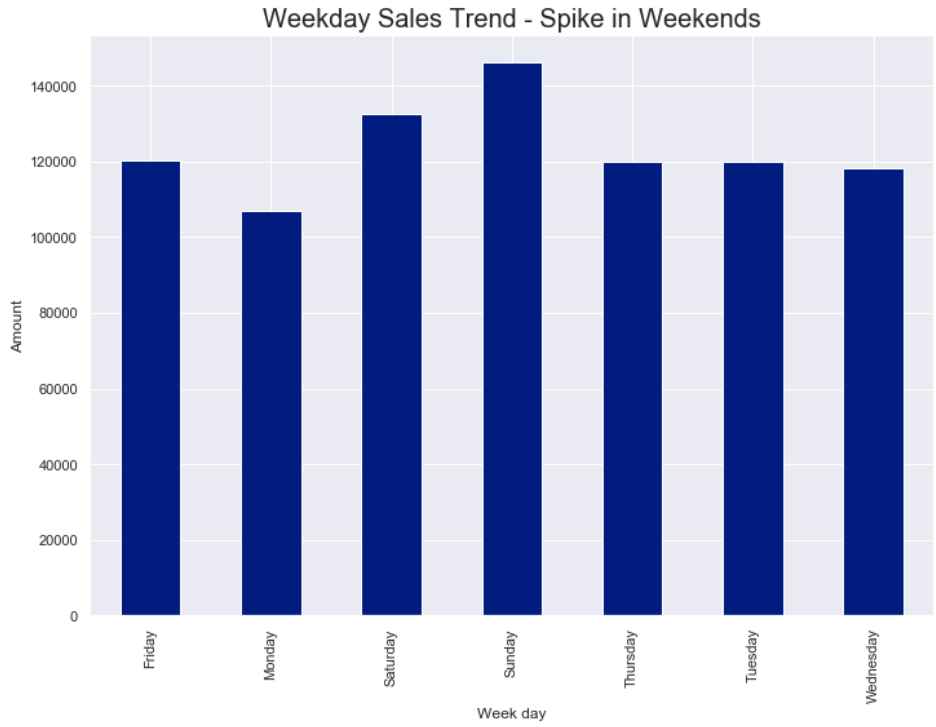


Transactional Data Analysis – By Week

For week wise transaction, **maximum sales amount** is noticed on the 4th week of the month, while **lowest on the 3rd week** of the month. The business owner can think of business strategy to promote sales during first 3 weeks of the month.

Transactional Data Analysis – By Week day

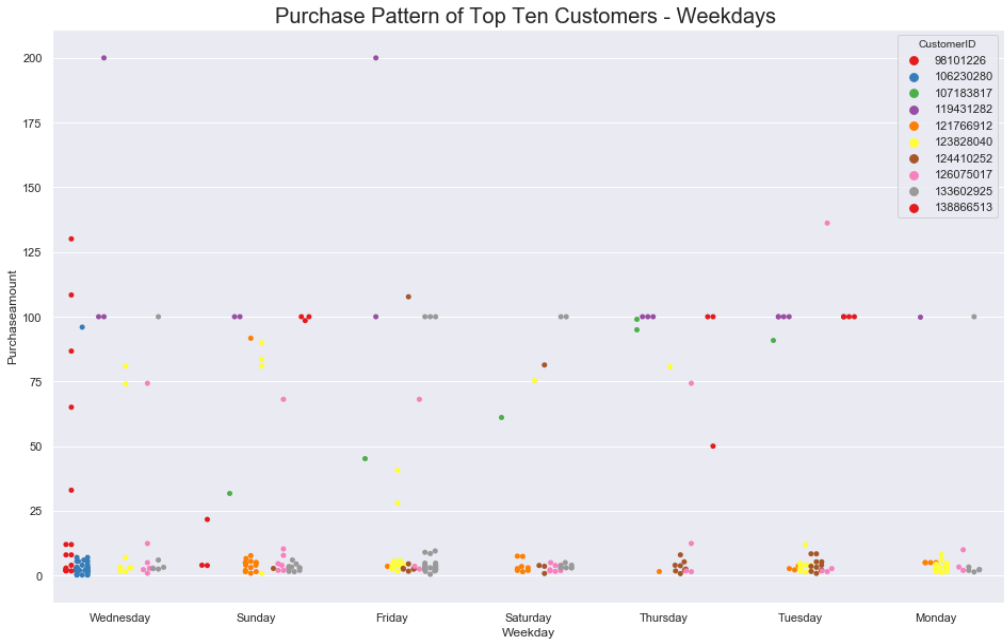
For day wise sales, trend shows **high sales during weekends** and lowest sales on Monday. The sales pattern is similar on rest of the weekdays-Tuesday through Friday.



Exploratory Data Analysis:

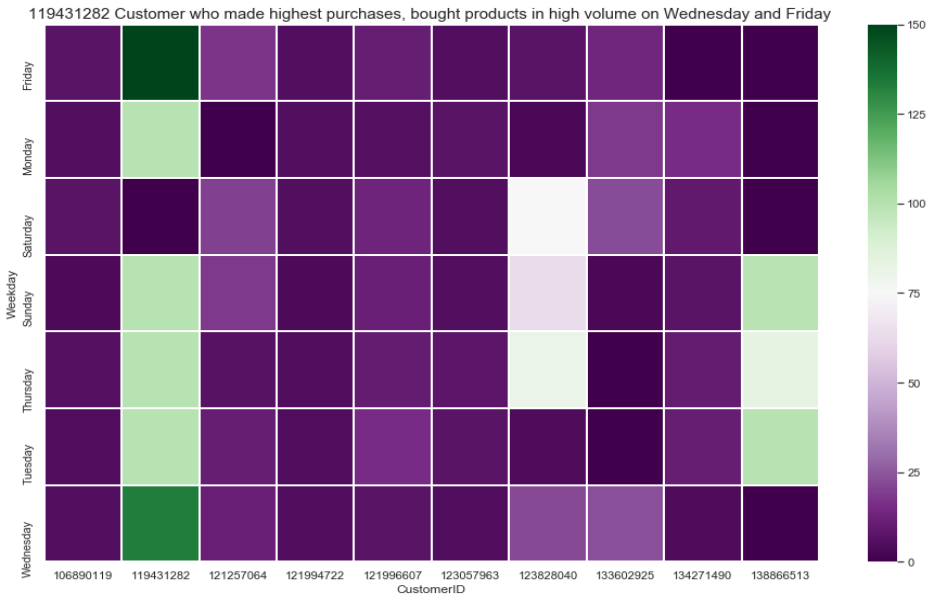
We made a set of ten customers who made **highest amount of purchase** in single transactions. We found, these ten customers made total 322 transactions throughout the month. We grouped them by the brands they bought.

We made a swarm plot between the purchase amount and day of the month for these ten customers. Their purchase is quite distributed across the month. Purchase amount is quite capped at 100.00. There are very few examples of purchase amount more than 100.

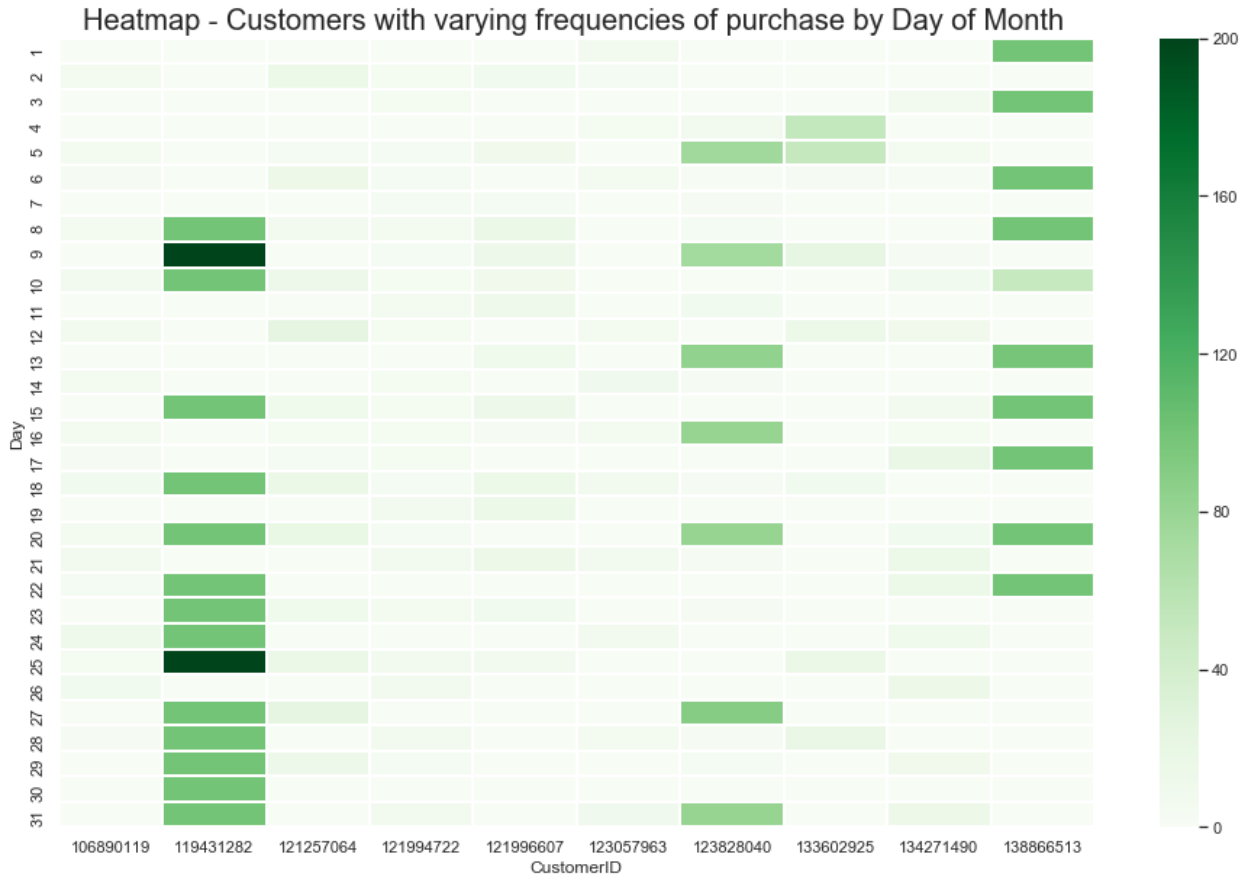


Weekly Purchase Patterns:

We made a heat map representation of the purchase amount and weekday to find their weekly purchase behavior.

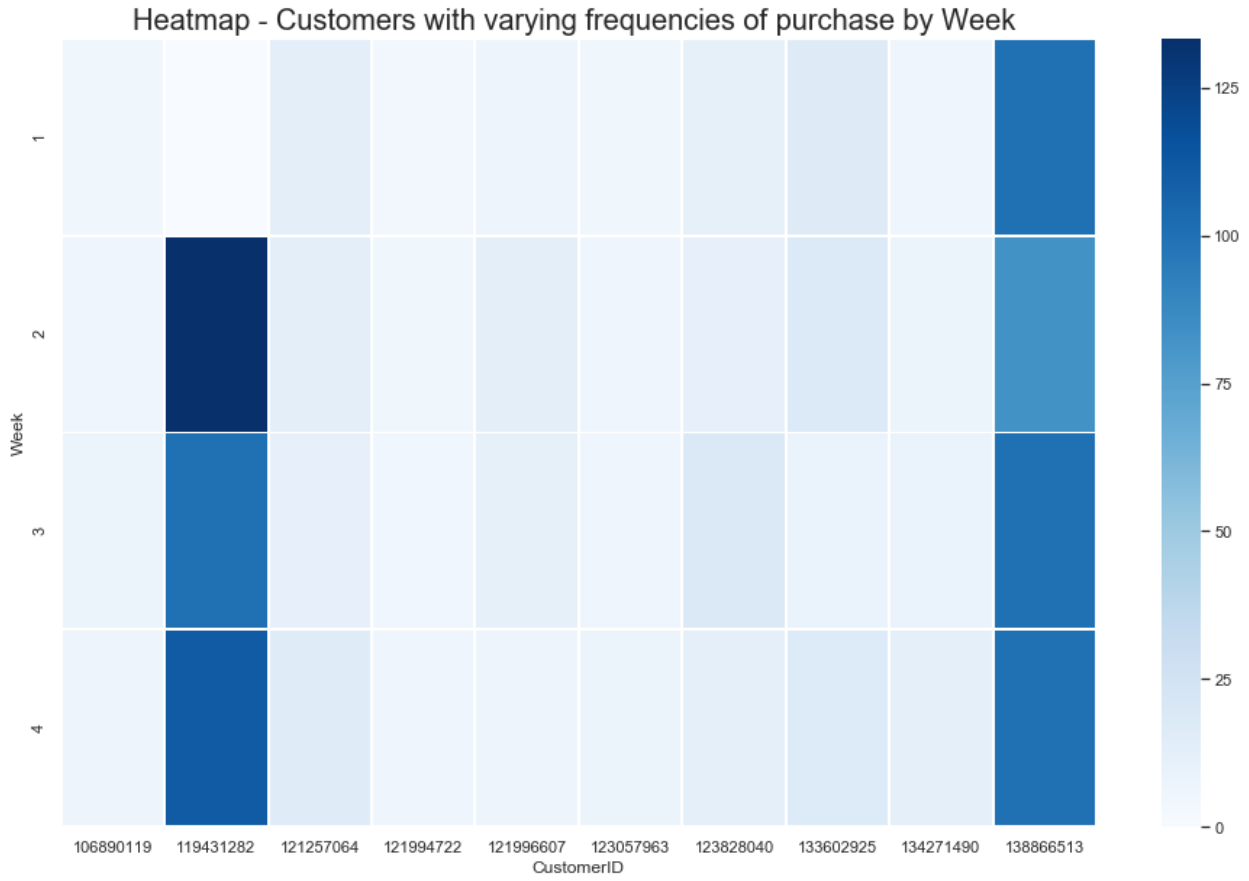


We made a heatmap plot with the purchase amount and day of the month for these ten customers. The color density allows easy exploration of their buying pattern throughout the month. Customer id **119431282** made **highest amount of purchase** on 9th and 25th of this month. For this customer, purchase amount was quite steady throughout the month. He can be a gold customer for the business.

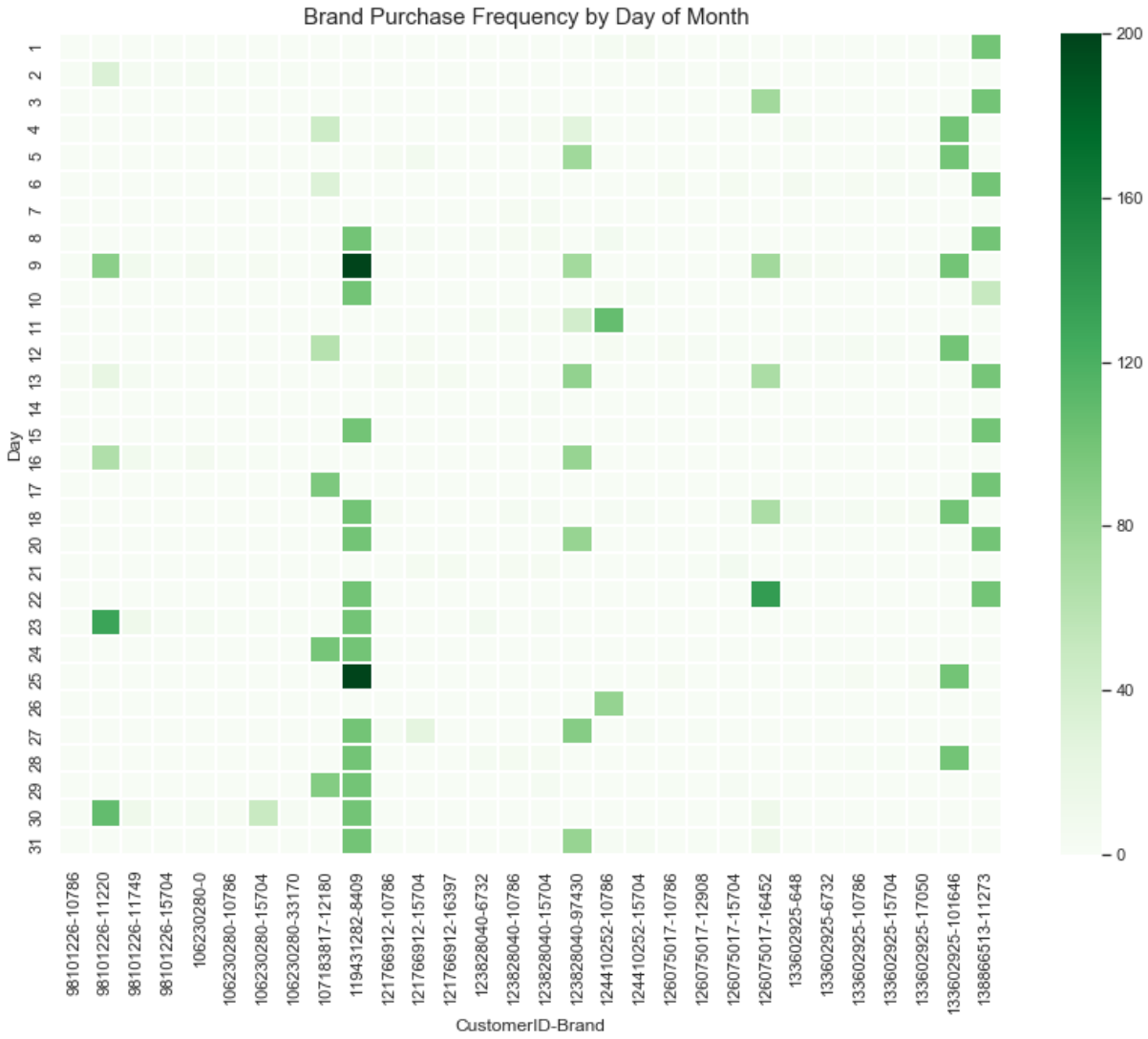


Some customers prefer buying during the first half of the month. Some prefers the second half. Customer id **138866513** bought products by 22nd day and there is no transaction made after that. Customer id **107183817** prefers second half.

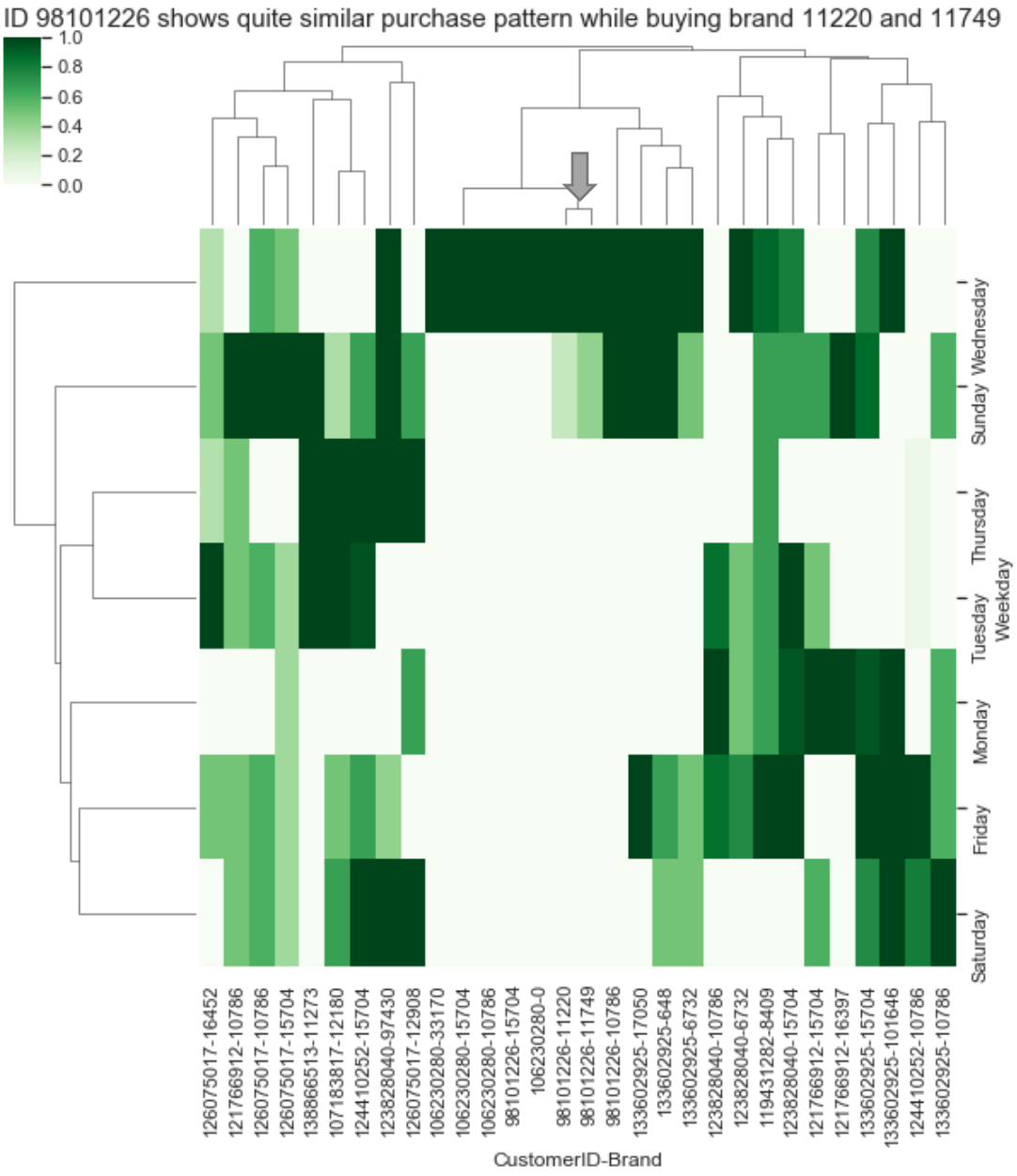
We made heatmap with the purchase amount and week for these ten customers. The weekly buying pattern of these customers are shown here. Customer id **119431282** made highest transactions in 2nd week. It looks like 2nd and 4th weeks are most preferable.



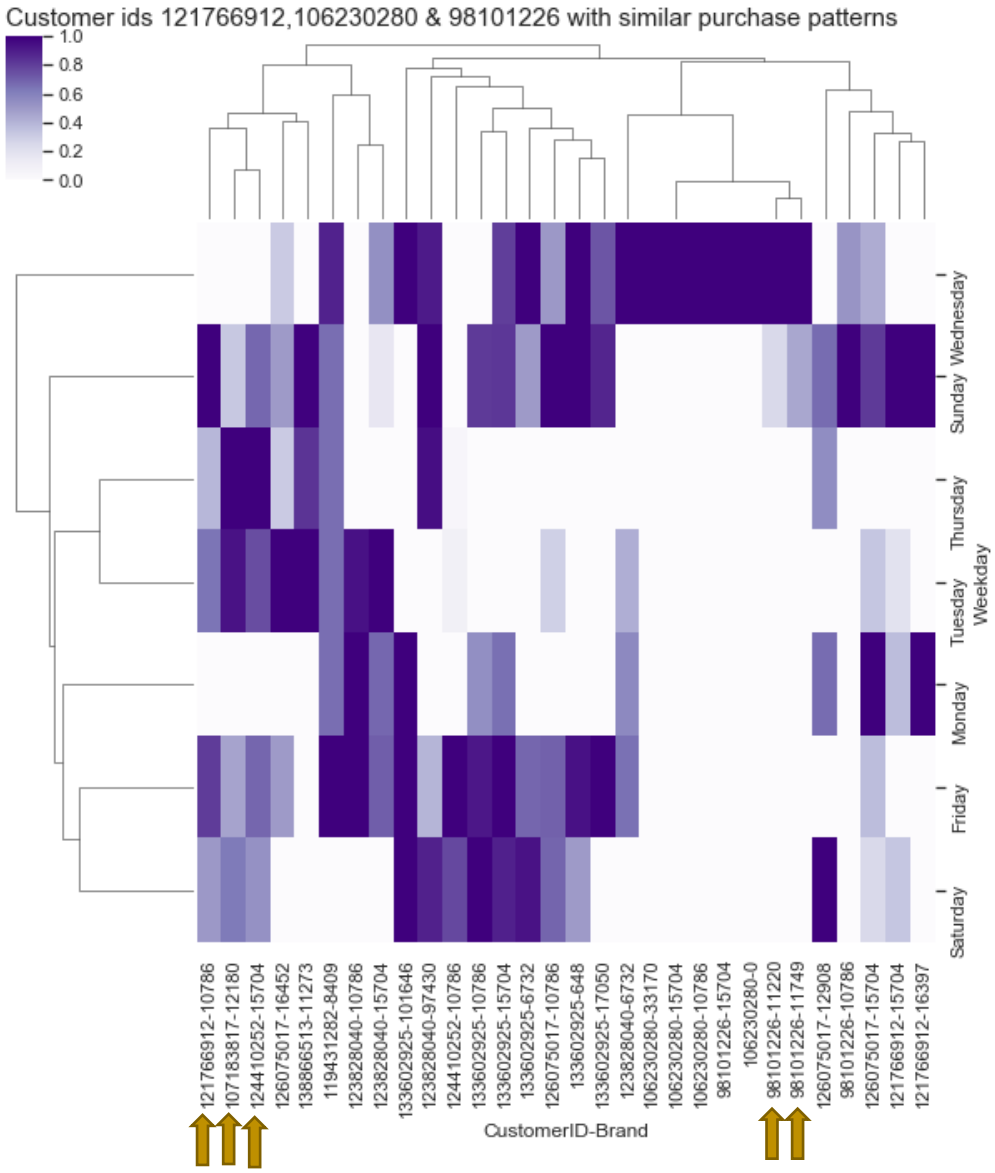
We made heatmap with the purchase amount and day for these ten customers along with brands purchased by them. It shows the pattern of brands purchased by the customers throughout the month. Customer id **119431282** purchased brand 8409 with highest purchase amount on 9th and 25th days of the month. Customer id **133602925** prefers brand **101646** most and purchased it with good amount.



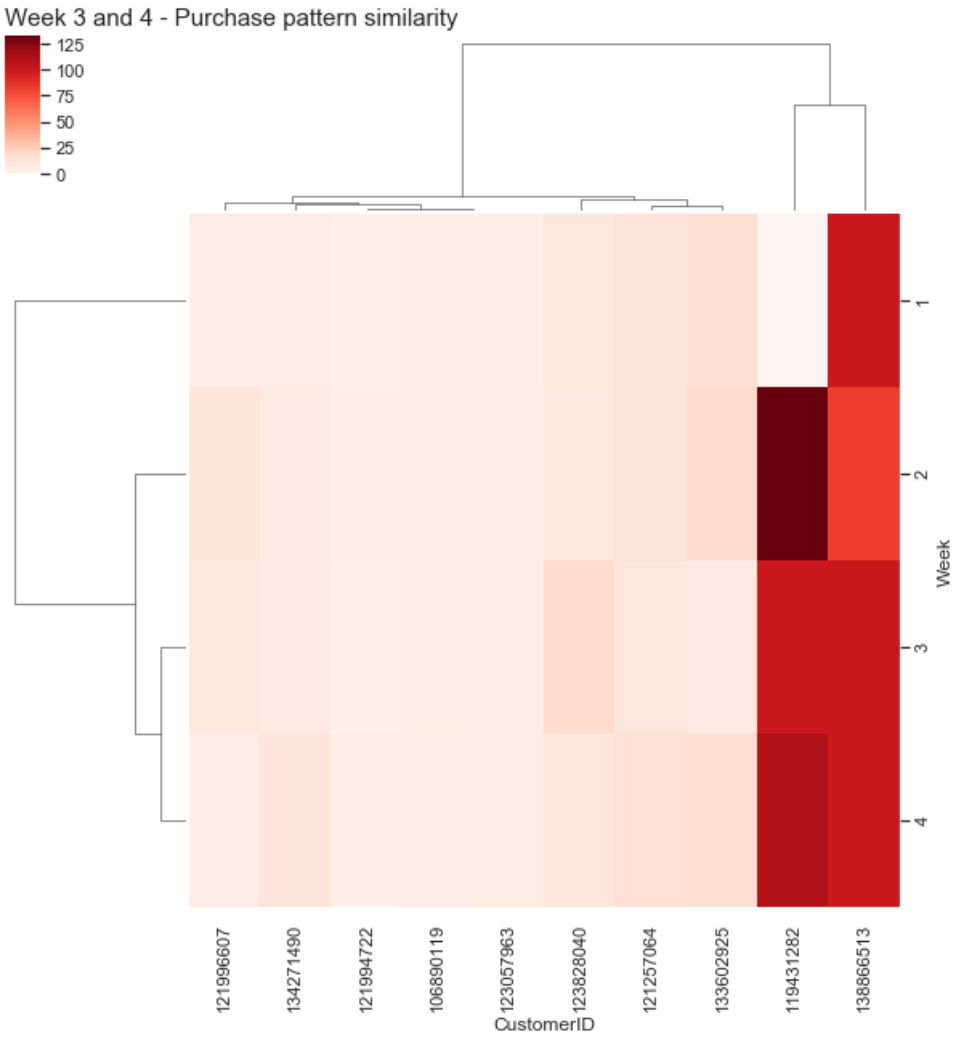
We made a pivot table of purchase quantity and weekday for customer and brands. We also plotted a Dendrogram (cluster map) for it. It shows similar buying pattern of weekdays. Customers purchase habit on Thursday and Tuesday are quite similar. Same goes for Friday and Saturday. Customer id **98101226** shows quite similar purchase pattern while buying brand 11220 and 11749.



We also plotted a dendrogram (hierarchical cluster map) for customer ids and days. It shows that buying pattern of customer ids 121766912 and 106230280 are quite similar. Then purchase pattern of customer id 98101226 finds similarity with them.



We also plotted a dendrogram (cluster map) for customer ids and weeks. It shows that buying pattern of these customers on week 3 and 4 are quite similar followed by week 2.



Sales Insights:

Statistics on Weekend Purchases:

There are total 9279 unique customers. Out of that, total 7587 number of customers bought products in weekends. That means **81.76% customers prefer buying during weekends**.

Weekend – Top 3 transactions:

Out of 7587 unique customers who bought during weekends, we found out the top Sales. Customer ID # 133990544 purchased the most i.e. total purchase amount of 437.34 only on weekends followed by customer IDs 121257064 and 125449807 who spent amount of 416.42 and 345.14 respectively.

Average Purchase Amount per Transaction:

We also found out **average purchase amount spent by customers on each transaction**. Customer Id 119431282 **spent highest average amount per transaction** i.e. 113.33.

Purchase Trend by Weekend and Weekdays:

There are 353 **customers who shop exclusively on weekends**. Also, there are 1692 **customers who exclusively shops on weekdays**.

Unique Transaction and Product ID's:

Assumption on Transactions:

The dataset contains transactions of each product by every customer. It is assumed that all the product purchases by a customer on a single day was a part of single transaction. That means, user visited the shop once in a single day and purchased the products once. Hence a unique transaction id for each transaction was highly required.

Unique Transaction ID:

A subset of data frame was developed by taking the variables such as - 'CustomerID', 'Chain', 'Dept', 'Category', 'Company', 'Brand', 'Productsize', 'Purchasequantity' and 'Day'. Unique transaction IDs were made by concatenating 'CustomerID','Chain' and 'Day' together.

Assumption on Brand:

It is assumed that each Brand of a category from a specific department and developed by a company is a unique product. Hence a unique Product ID for each product was highly required.

Unique Product ID:

Unique Product IDs were made by concatenating 'Dept', 'Category', 'Company' and 'Brand' together separated by dash.

The Products (Product IDs) were then grouped by each Transaction ID. Hence, we found the list of products bought in every transaction by the customers.

Clustering:

Clustering is an unsupervised learning approach to deal with finding a structure in a collection of unlabeled data. It organizes objects into groups whose members are similar in some way.

Here we attempt to cluster the products based on Chain, Dept, category and Brand. The aim is to find similar products in each cluster so that we can design better upselling and cross-selling offers for the customers based on the products they are purchasing. The customers buying high-end products would be interested in purchasing similar high-end products. The customers purchasing low-end products would be interested in similar kind of products and so on. Hence clustering is done here on the transaction data to segment the similar type of products for better offer management.

Clustering with Categorical Data:

Here we can see that the chosen parameter, such as - Chain, Dept, category and Brand are all categorical variables. Each categorical attribute is represented with a small set of unique categorical values. Unlike numeric data, categorical values are discrete and unordered. Therefore, the clustering algorithms for numeric data (like: K-Means clustering) cannot be used to cluster categorical data that exists in many real-world applications.

K-Modes Clustering:

K-modes is an extension of k-means. It modifies the standard k-means process for clustering categorical data by replacing the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent cluster centers and updating modes with the most frequent categorical values in each of iterations of the clustering process. Dissimilarity measure is the quantification of the total mismatches between two objects: the smaller this number, the more similar the two objects. These modifications guarantee that the clustering process converges to a local minimal result. A mode is a vector of elements that minimizes the dissimilarities between the vector itself and each object of the data. We will have as many modes as the number of clusters we required, since they act as centroids.

Here we implemented the K-Modes algorithm for clustering. The best value of K is chosen by Elbow Method. K-Modes was executed for multiple values of K and the within cluster mean distance is plotted with changing values of K. The point where elbow is formed is considered as best value of K. We found that 5 is the best value for K.

We applied K-Modes algorithm for 5 clusters. It was found that number of members in each clu ster is 158307, 39822, 3712, 8545, 3817 respectively. We added the cluster value as a new vari able in the data frame. The data frame can be grouped by cluster value on a particular input vari able (for example: dept) and value counts can be calculated.

How does the recommendation system work?

If a customer id is given, top three product recommendation can be found from the clusters. First the clusters are found which the customer id belongs to. The cluster where maximum number of times this customer id appears is the ideal cluster. Then top three brand from that cluster for that customer id are found out for recommendation.

For completely new customer, where the data do not exists in the system, we want the system to recommend top 3 selling brands.

Project Deliverables: Strategies to Increase Average order value

The Recommender system delivered by this project provides a marketing tool on which following strategies can be applied on website to increase the average order value

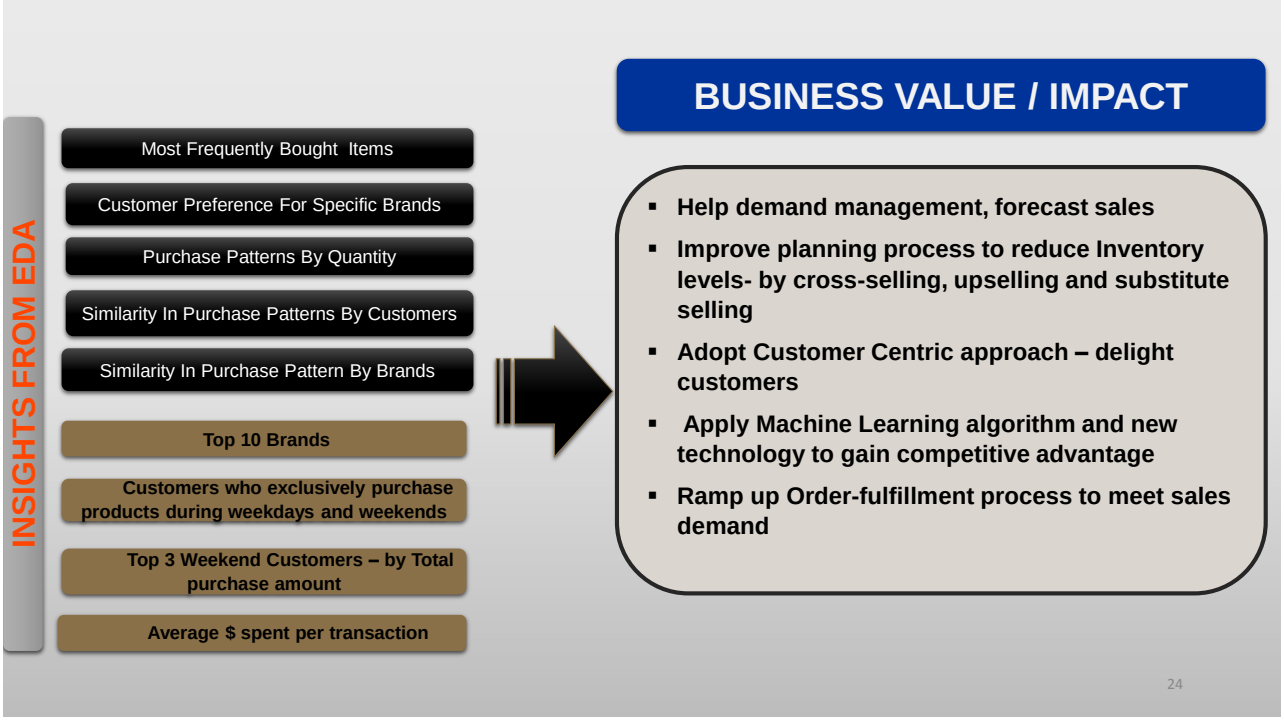
- Personalization
- Customer who bought this item also bought
- Recommendation to you based on your previous purchase
- Best-selling item in brand or Category

Email Campaign can also be run in parallel for off-site recommendations to include above scenarios.

Launch promotions to increase order value for Customers and Brands with low transactions (less than 5), and for the customers who exclusively shops during weekends or weekdays. Client can think of introducing free shipping offers to encourage increase in basket size since customers have to spend over a certain amount to receive free shipping.

Business Value/ Recommendations:

Through EDA we were able to derive statistics on transactional data and get valuable insights on purchasing behavior of customers. This information can help client make proactive business decisions and bring value to the business in the following segments:

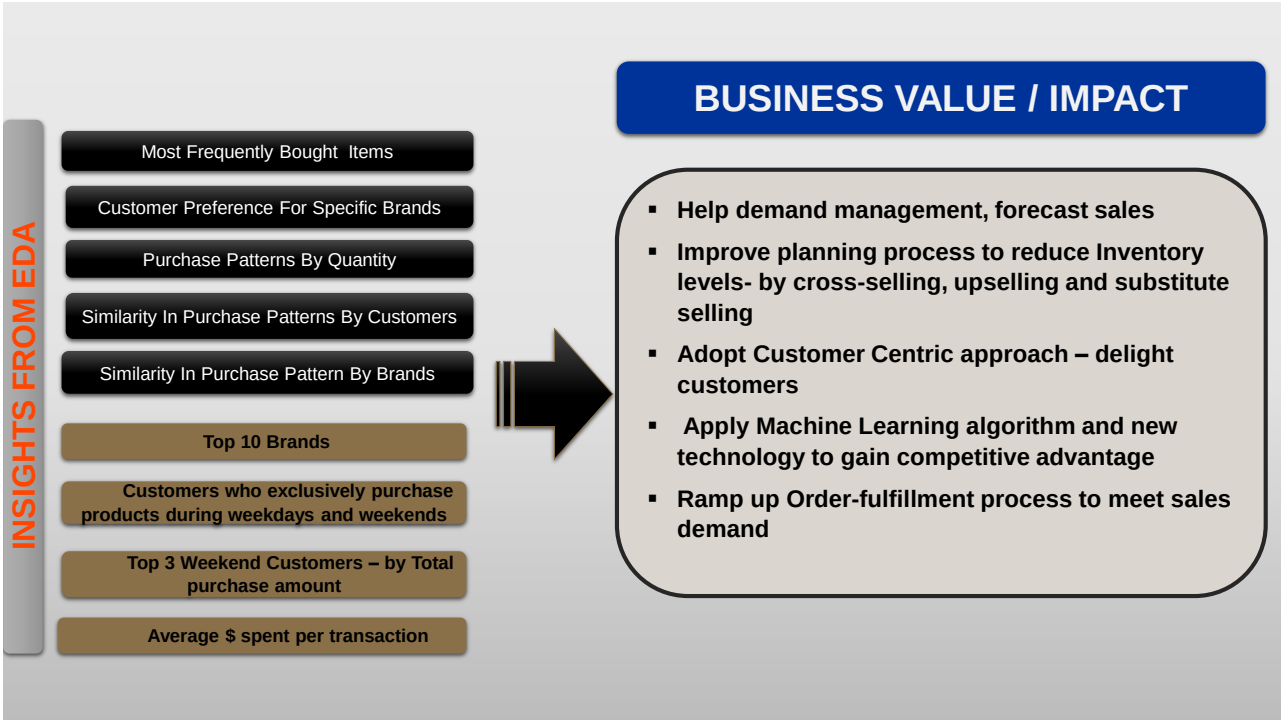


Demand management and Forecasting

- Provide precise forecast for market and sales.
- Help business platform to understand customer and market, figure out the preferences of the market, so establish effective marketing strategies.

Improve planning process to reduce Inventory levels

- This can be done by enhancing cross-selling, upselling and substitute selling capacity. According to the purchase data of users, the systems can figure out an addition product and make recommendation to them. Therefore, the client can make plans to reduce the inventory.



Adopt Customer Centric approach

- Take customer centric approach to understand customer requirements and desire to delight customers through better offer management.

Apply Machine Learning algorithm and new technology

- To gain competitive advantage, use machine learning capabilities for demand forecasting, product search ranking, deals recommendations, merchandising placement, and much more.
- Increase customer loyalty program, member rewards, work on initiatives to offer free shipping and much more.

Ramp up Supply Chain:

- As a result of recommendation system, the sales revenue in expected to increase up to 25%. In order to meet the Sales demand, supply chain process has to ramp up in order to make fast and accurate order fulfillment.

Future Analysis:

- Due to the complexity in dataset presented by sparse data, categorical and anonymous data types, we were unable to dive deeper into the analysis. Analysis such as 'frequently bought items together" that depends on browsing history, could not be performed due to lack of data i.e. transaction time.

- Customer Id, Transaction date, Brand Id, Purchase amount and purchase quantity were among the important features that played a significant role in extracting useful information from the dataset. A full fledge recommender system can be built , if we have more information on transaction time, user comments, feedbacks, browsing history, ratings and more features.
- Collaborative filtering technique can be further applied to the data set to build a recommender system. As we had a very small subset of data for analysis (only January month), we were unable to discover seasonality in the purchase behavior.