**Overview:**

Customers have always been the drivers of the retail business or any other market for that matter but in this digital age their influence has become more powerful than ever before.

Thanks to handy gadgets and social networks on the fingertips customers now actively engage with businesses to voice opinion and dictate the trends. With growing customer awareness, understanding customer behavior and expectation has become vital step to run a sustainable and profitable business.

**Problem:**

Marketing and sales teams of retail companies rely on key insights about their customers to design right communication tools. They need crucial details such as; targeted age group, frequency of Transactions, and shopping patterns to develop resourceful communication strategy. Specifics like, distribution of online and store sell, purchasing power and help build effective promotional and sales tools to maximize business.

As a routine practice, businesses now constantly comprehend and improve the health of customer engagement to see the business from a customer’s point of view. Identifying the factors that make or break loyal customers has become an essential tool to build long lasting customer relationship.

**About the Dataset:**

This customer survey dataset has observations from 1000 customers who have visited/purchased from the retailer at the store or online. The data is stored in 12 columns that display various observations and feedback from the customers, it can be divided into three segments.

* **Customer feedback**
* Satisfied Service
* Satisfied Selection

(Both rated on the scale of (1-5) where 1 means lowest and 5 means highest)

* **Personal Details**
* Customer ID
* Age
* Email
* Credit Score
* Distance to Store
* **Store Activity**
* Online Visits
* Online Transactions
* Store Transactions
* Online Spend
* Store Spend

**Data Wrangling Steps:**

* Examine the structure of the dataset
* Look for missing values and zero, decide how to deal with them
* Replace missing values in the satisfaction service/selection columns by mean values
* Turn email into a binary variable for easier analysis
* Create new column ‘total.spend that is the sum of ‘store.spend’ and ‘online.spend’
* Create new column ‘total.trans’ that is the sum of ‘store.trans’ and ‘online.trans’
* Create a new column ‘cust.category’ that divides customers into four categories; “Prospects”, “Regular”, “Premium” and “VIP” based on the ‘total.spend’.
* Create new column Returning\_cust that how the cust.id of customers who have more than one Transactions

**Preliminary Analysis of the Dataset:**

* The average age of our customer is 34.912 (Range 18-55 years) and the mean credit score of 725.5. (Range 600-950) Which means that our average customers are young adults who can manage their finances very well.
* Majority of our customers are internet savvy as 81.4 % of them have email. On average they visit the website 28 times (Range 0-150).
* The mean distance to store is around 14 miles. Apparently, customers living close to store do shop frequently at stores compare to those who live at 30 miles distance and more. But surprisingly distance is not a deciding factor for shopping online or at store, the analysis shows that customers who live close or far from store mostly prefer to shop online.
* 70.6% of the total customers return to shop again and in result register multiple transaction.

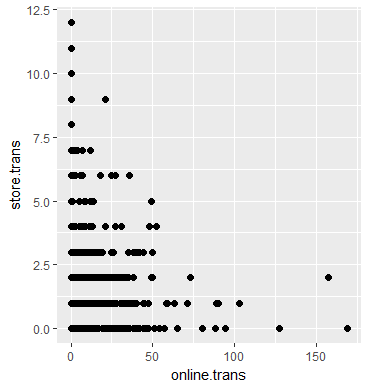
**Data Visualization and Statistical Analysis:**

**Online and Store Transactions distribution**

On average customers register **1.32 store Transactions** and **8.38 online transactions**. However, the spending amount per transaction for **store spend ($36**) is higher than **online spend ($20.32)**.

The maximum number for store transaction is 12 while the maximum number of online transaction is comparatively quite higher at 169.

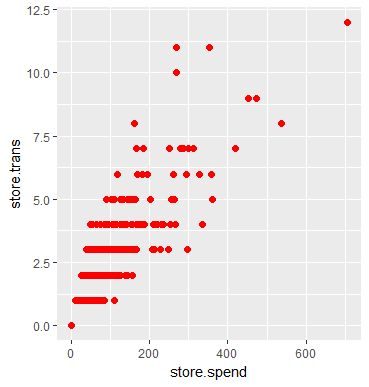
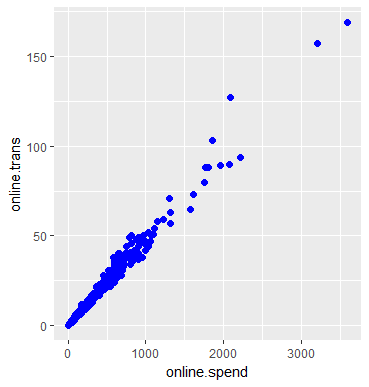
The above discovery aligns with our previous analysis that most customers prefer online shopping and return many times to shop.



**Spending distribution of online and store purchases**

Online business is the most popular way to shop amongst customers as 78% of the total purchase happen online while 22% of the total purchase take place in store.

The following graphs represent association between online and store spend and Transactions. Although both graphs are skewed to right, the graph on the right that displays online variables is much more precise compare to the with store variables.

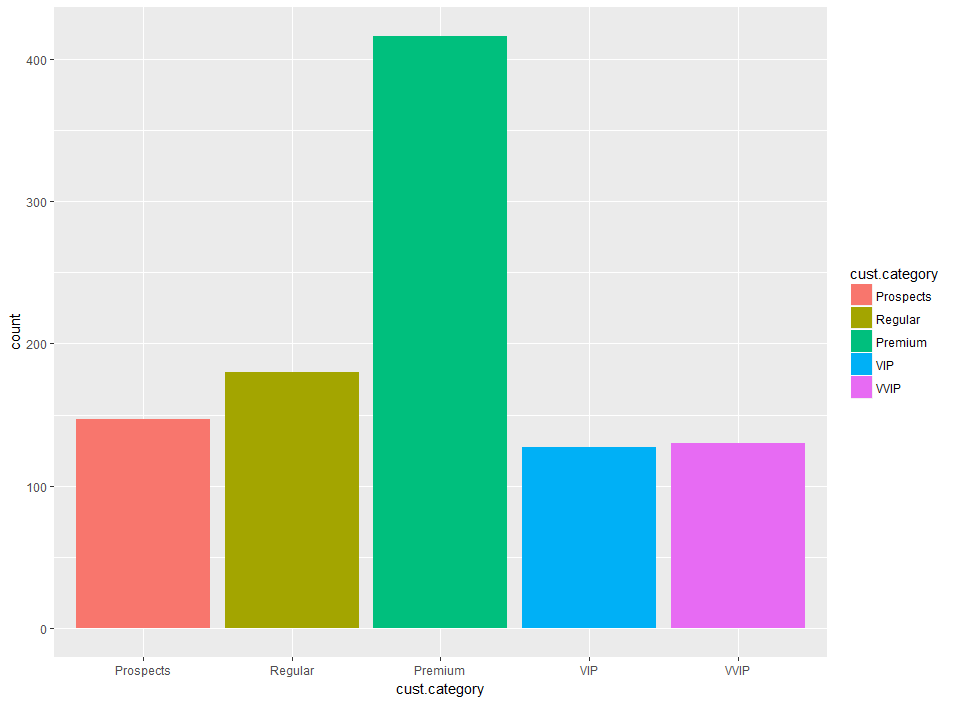


Customers spend a lot more money shopping online compare to what they do in store. Maximum value for store.spend is merely $706 while the highest value of online.spend is $3593. Clearly online is a preferred shopping choice where customers spend more money via multiple Transactions.

**Customer categorization based on total.spend**

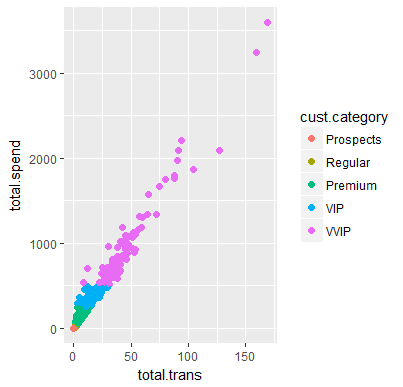
Customers can be divided into five categories according to the total amount they spend shopping. The following chart shows the customers distribution.

1. ***Prospects*** *= Spent $0 in purchase but visited online or at store*
2. ***Regular****: Spent less than $50 in total*
3. ***Premium****: Spent anything between $50 to less than $250 in total*
4. ***VIP****: Spent anything between $250 to less than $500 in total*
5. ***VVIP****: Spent anything between $500 to less than 4,000 in total*



**Total spend increases with total transaction**

As the number of transactions increase so does the total spend amount. When we map the findings on customer category we see a clear trend (except few outliers). Which means that most of our Premium, VIP and VVIP customers are loyal customers who come back to shop again and again.

****

Note: There is some overlapping in the beginning of the graph between premium and regular customer points so regular are not quite visible in the graph.

**Satisfaction Ratings Analysis**

Satisfaction Service and Satisfaction Selection charts indicate that customers are not very satisfied with either Service or Product Selection.

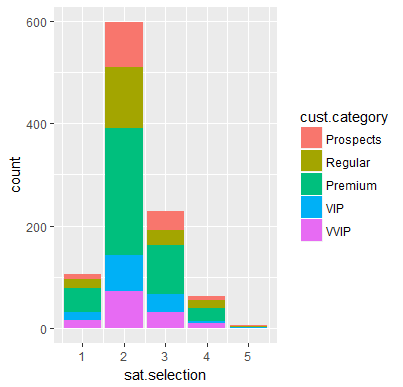
(Note: Satisfaction ratings are measured on the scale of 1 to 5 where 1 is the very poor and 5 is excellent)

**Selection Satisfaction:**

Analysis of the selection satisfaction data shows that 59.7% of customers have rated the product selection as poor (2). While 22.9% customers think of the selection as average (3). More importantly only 0.63% customers have rated it as Good (Rated 4) and just 0.05% are totally satisfied with the product selection (Rated 5). 10.6% of them rated it as very poor (1).

**Distribution of selection satisfaction with customer category**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rating** | **1** | **2** | **3** | **4** | **5** |
| **Prospects** | **7%** | **59%** | **26%** | **6%** | **1%** |
| **Regular** | **9%** | **67%** | **16%** | **7%** | **1%** |
| **Premium** | **11%** | **60%** | **23%** | **6%** | **0%** |
| **VIP** | **13%** | **55%** | **28%** | **3%** | **15%** |
| **VVIP** | **12%** | **55%** | **25%** | **8%** | **0%** |

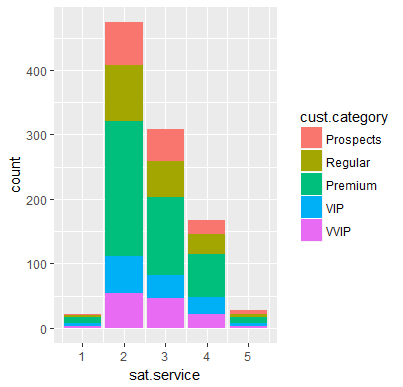


**Service Satisfaction:**

Service is rated a bit better compare to selection. If we look at the distribution of the selection ratings, majority (47.5%) of the customers have rated the service as 2 (poor). However, unlike sat.selection a good (30.9% )customers have rated the service as average(rated 3) and 16.7% have rated it as good (rated 4). And 0.22% have given the lowest rating(1) while 0.28% are fully satisfied with the service.

**Distribution of service satisfaction with customer category**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rating** | **1** | **2** | **3** | **4** | **5** |
| **Prospects** | **1 %** | **46%** | **34%** | **15%** | **4%** |
| **Regular** | **1 %** | **48%** | **31%** | **16%** | **3%** |
| **Premium** | **2 %** | **50 %** | **29 %** | **16%** | **3%** |
| **VIP** | **3 %** | **41%** | **28%** | **20%** | **4%** |
| **VVIP** | **2%** | **42%** | **36%** | **16%** | **2%** |

****

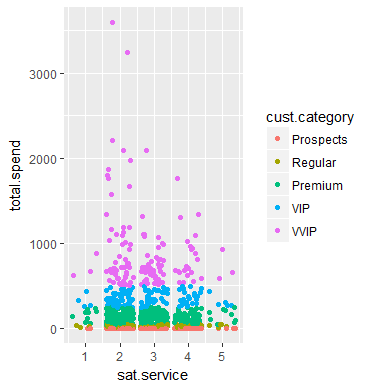
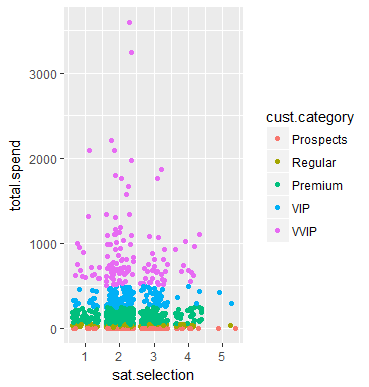
**Customer Feedback Mapped on Total Spend**

**Satisfaction Service**

* Most of our loyal customers rate the service as poor, average or good.
* Very few customers have voted service as very(rated 1) poor or excellent(rated 5)
* Majority of the VIP and VVIP customers rate the service as poor to average

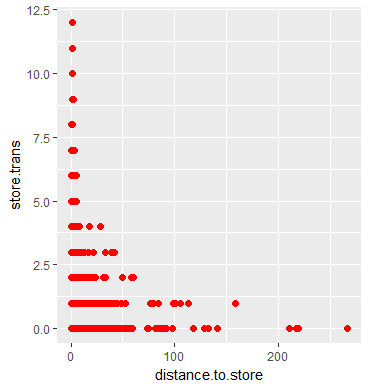
**Satisfaction Selection**

* Shows poor rating percentage compare to service
* Sizeable number of premium/VIP/VVIP have voted the selection as one and very few are fully satisfied(rated 5).
* Majority of Premium/VIP and VVIP rate the service as very poor to average

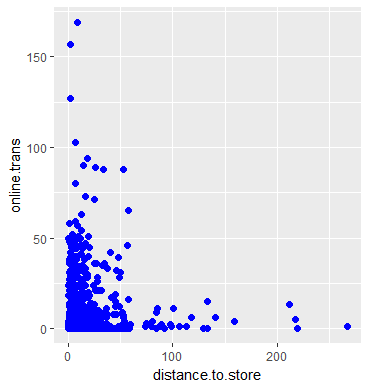
**** ****

**Distance to Store Correlation with Transactions**

* Customer who live closure to store are more likely to shop at store than those who live 50 miles away.
* Customers living within 50 miles of store are more likely to make online purchase.
* However, comparatively much denser online transaction graph shows that whether a customer lives close or far from store he/she is most likely to purchase online than go to store

****

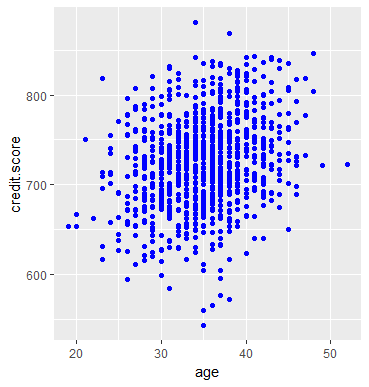
Distance Range: 0 – 275 miles , Store Transaction Range: 0 – 12



Distance Range: 0 - 275miles , Store Transaction Range: 0 – 167

**Age and Credit Card Distribution**

As stated in the preliminary analysis the graph shows that credit score gradually get better with age but the plot is a scattered and there is no defined pattern.



Age: Range from 18-55 , Credit Score: Range from 600-950

**Applying Machine Learning to Build Prediction Model**

The variables of the cust.df dataset are not distributed in a manner that allows opportunity for explore multiple prediction opportunity.

Furthermore, some data points like time and date, gender of the customer and type of product purchased are also missing from the dataset. These elements would have provided the opportunity to predict multiple scenarios that could be useful to the client.

Therefore, due to the limitation of the dataset we are going to focus on building prediction models for some key factors.

**Maximum Sell:**

Total Spend column represent the total amount spent by each customer via multiple store and online transactions. As highest total sell would mean maximum profit we would try to build a model to achieve maximum total.spend.

After multiple trials we have a model that has R squared value of 9.871 which is considered very good, so we can conclude that this is going to be our final prediction model to predict maximum sell.

We have also examined the model objects which are in align with our conclusion that this is going to be a good model to predict max sell.

**Conclusion:**

After studying all the outcomes of our multifaced analysis we have found answer to most of our questions. Below is the list of our conclusions.

**Key Findings**

Meet the Customer

* Young adults, majority of whom belong to (25-45 years) age group but overall ranges between 18 to 55 years.
* Prefer online shopping as 78% shop online and only 22% in-store.
* Mostly financially stable, credit score ranges from (600-950)
* 71% of our total customers return to shop. While 85% of total online customers return and 55% first time store buyers return to store
* Customer have significantly higher number of transactions online(169) compare to store(12) and spend a lot more online

**Levels of Customer Satisfaction**

* Happy customers make successful business but analysis of the customer feedback in the dataset shows that most customer are neither satisfied with service nor selection.
* Service and Selection both have opportunities for enhancements, but selection is rated lower than service and may need extra attention**.**

**Utilization of Customer Segmentation**

* Based on the average total spending amount we divided our customers in to a group of five. (Prospects, Regular, Premium, VIP, VVIP).
* With the precise target group in mind marketing, sales and customer communication teams of the business can develop unique communication tools and strategies to lure new customers and retain our most valuable loyal base.

**Client Proposals for Further Analysis**

**Overview**

With the help of the existing findings from the presented analysis our marketing and sales team should be able to chart a plan to target potential sales. Although, there are many other data points that we can combine with existing model to increase accuracy and extensive analysis of the business model.

**Customer Segmentation by Purchase Time Frame**

Purchase history time frame is one of the key elements missing in the ‘cust.df’ dataset. With the help of information like date of each transection, number of purchases on sale and non-sale days plus clearance we can build a time plot for more accurate predictive analysis and segmentation of customers.

The purchase graph could look dramatically different for sale and non -sale days. By adding the time frame and discount date details, we can efficiently segment our customer base into four groups and plan appropriate marketing strategy for each of them.

* **Potential/New customer**

Potential and new customers would record minimum to no transection. They are a good target for introductory customer friendly communication. A good marketing strategy for this group would be to reach them with timely communication to make them familiar with business but not overwhelm them with frequent follow ups.

* **Discount lover**

Discount lovers often look for a good bargain and return to shop during discount seasons. These customers could bring in a great profit if targeted with right promotional tool at the right time.

* **Loyal customer**

Unlike discount buyers, loyal customers would show a pattern of regular transections throughout the month/year not just during discount season.

These customers should be special focus of the customer communication team as their opinion on product line services weighs a lot more than new or seasonal customers who may know little about the business.

Loyal customers can also be divided in two subcategories

* **Regular loyal shoppers**

General shoppers are usually the largest group in the business who frequently

register multiple transections throughout the timeframe. However, the spending

amount of these transactions would be in the range on average to below average.

* **Impulsive loyal shoppers**

Impulsive shoppers may seem like a smaller group in the customer category, but they mostly provide huge contribution to the growth of the business. They mostly show patterns of multiple above average transection spread across the entire timeframe.

**Product Performance Analysis:**

Knowing the pulse of the customer is essential for businesses to thrive and succeed and satisfaction survey is a great way to look at the business through a customer’s point of view.

Detailed breakdown of the product selection by customers can give us insights about product approval ratings. By pairing the available satisfaction rating analysis with the product detail and transaction frequency we can determine the performance of the product.

Depending on the performance of the product, it can be decided to restock the popular items or drop the not-so loved products.

**Predict Product Preference and Offer Suggestion**

**Product Preference**

A retail store offers a range of products that can be divided into several categories and sub categories. For example; apparel, accessories, home are just a few types of items we usually find at a retail store.

If we can find product purchase details of unique categories and map them to customer preference we can find very precise target audience.

Now whenever we have special offers or fresh stocks available for a special category we can reach out to a selected group of potential customers who have previously shown interest in the category.

**Offer Suggestion**

An in-depth analysis of customer’s browsing and purchasing history could give us great insights about what brands, categories and styles a customer prefers to buy. Based on the findings we can predict the items customer is most likely to buy.

The ‘cust.df’ dataset does not contain information about customer’s gender and product preference, knowing such details could be crucial to predict suggestions for customers.