**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 transections, and shopping patterns. As well as ratio of online and in-store transections, purchasing power of customers etc. These specifics help build effective promotional and sales tools to maximize business.

Growing businesses constantly need to understand and improve the health of customer communications. Identifying the factors that make or break loyal customers is 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 (Rating: 1 -5)
* Satisfied Selection (Rating: 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 Transections
* Store Transections
* Online Spend
* Store Spend

**Limitations of the Dataset:**

The dataset does not have information about the gender of the customers and neither give us details about the kind of products customers have purchased.

These two details could have provided us more insights for product analysis. With gender and product identification we would have been able to identify which product are performing better or poor among female and male customers.

Knowing the representation of both male and female customers in dataset could contributed towards better understanding of our target audience.

**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’.

**Preliminary Analysis of the Dataset:**

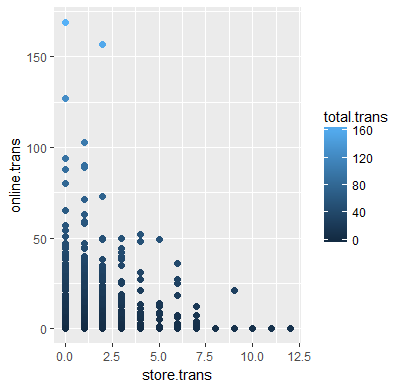
* The average age of our customer is 34.912 and the mean credit score of 725.5. Which means that our average customers are young adults who can manage their finances very well.
* Majority of our customers are internet savvy as 81.4 % of them have email. On average they visit the website 28 times.
* The mean distance to store is around 14 miles. But surprisingly distance is not a deciding factor for shopping online or at store. Whether a customer lives close to the store or not ,most prefer to shop online.

**Online and Store Transections distribution**

On average customers register **1.32 store transections** 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.

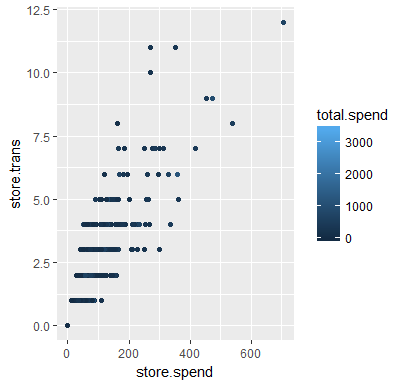
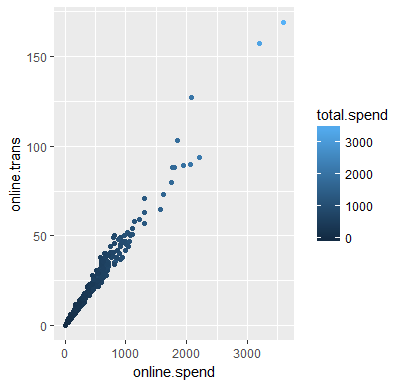
The above discovery aligns with our previous analysis that most customers prefer online shopping and return many times to shop.

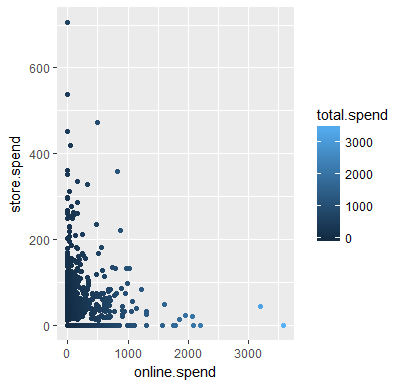


**Spending distribution of online and in-store purchases**

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 transections. 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.

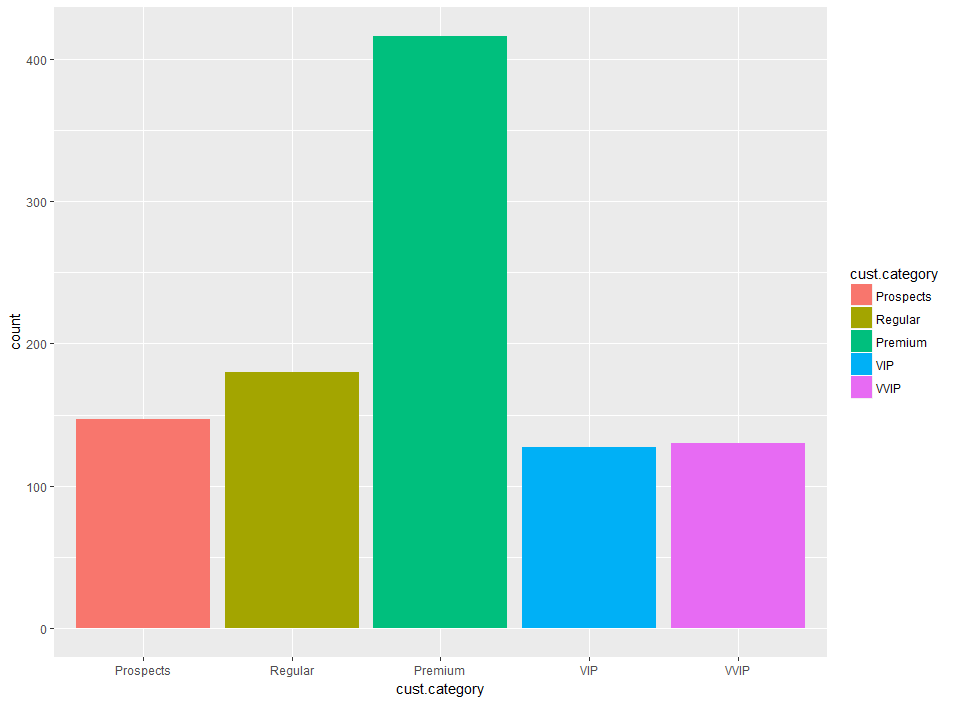


Customers spend a lot more shopping online compare to what they do in store. Maximum value for store.spend is merely $706 while the highest value of online.spend is $3593. Clearly online is a preferred shopping choice where customers spend more money via multiple transections. 

**Customer categorization based on total.spend**

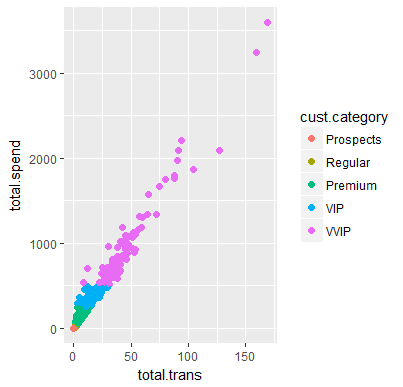
Customers can be divided into five categories according to the total amount they spend shopping. The following chart shows the customers distribution.

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 depends on total transection**

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**

A quick look at the 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 poorest 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).

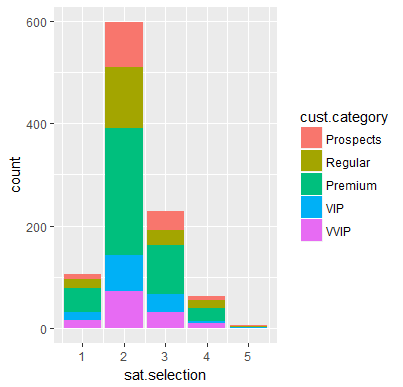
**1 – 10.6%**

**2 – 59.7%**

**3 – 22.9%**

**4 – 0.63%**

**5 – 0.05%**



**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).

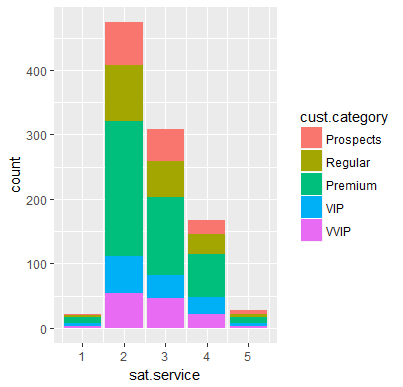
**1 – 0.22 %**

**2 – 47.4 %**

**3 – 30.9 %**

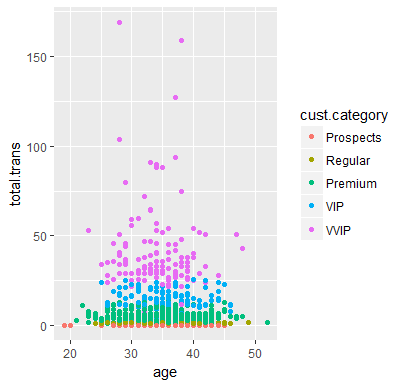
**4 – 16.7 %**

**5 – 0.28 %**

****

**Age and Purchase Pattern:**

The average age of our customer is 34.91 and a sizable number of frequent shoppers (repeat customer) belong to the age group of 25-45. There are very few customers out of this age range, who make multiple transactions and belong to VVIP category.



**Further Analysis:**

* Classify the individuals marked as “prospects” who can be potential sales and make good targets for regular promotions
* Identify our VIP and VVIP customers who could be targeted with extra special offers and regular promotions.
* Determine the prime factors that affect total.spend and build a model that can predict total.spend
* Experiment with the variables to determine whether a regression or a clustering model is the right predictable model for this dataset