Web Analytics Insight

This document contains a brief analysis of the visitor trends and behavior on a various e-commerce websites. What I am trying to tell through this document is **a story** about the user trends and activities on the website.
(Note: Used R to analyze the document. The source code can be found at here).

Introduction to data:

Our data is about the user's activities on e-commerce websites and contains **12 attributes**:
1.day
2.site
3.new_customer
4.platform
5.visits
6.distinct_sessions
7.orders
8.gross_sales
9.bounces
10.add_to_cart
11.product_page_views
12.search_page_views

The number of **rows** in our data is **21061**

Here are the steps I used to analyze the data.

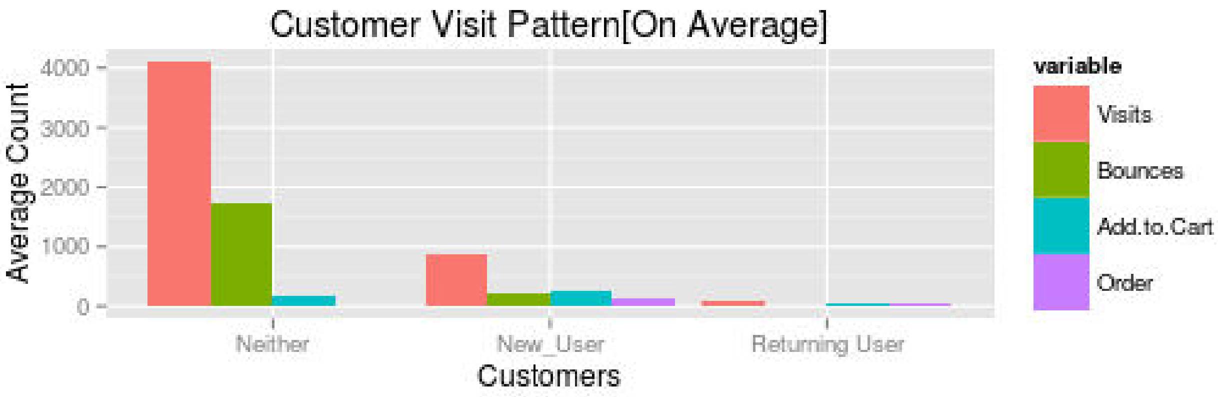
Step 1: Feature Engineering and cleansing of data

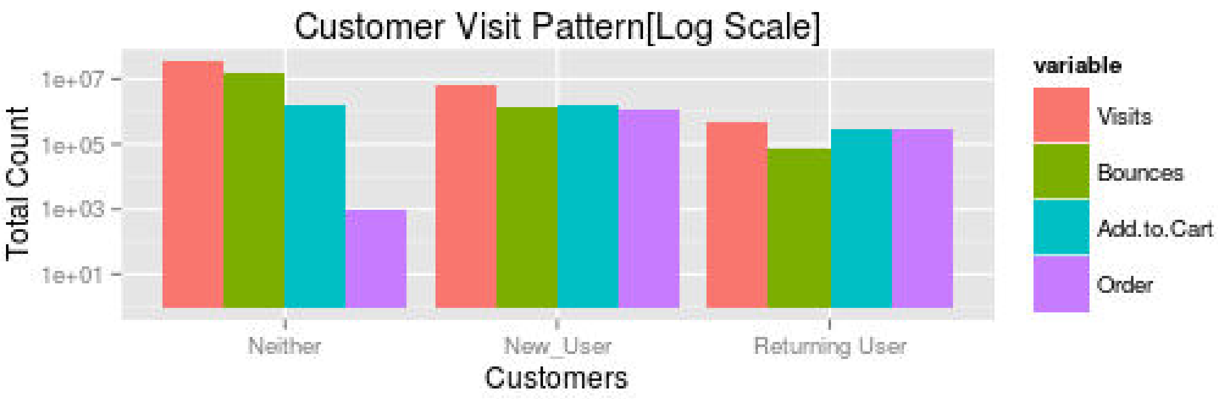
a. Data contains two unnecessary columns. Removed both “NA” and “NA..1” columns
b. Replaced the blank fields in **platform** column to “Unknown”
c. Replaced the blanks in **new_user** field to the number '2'.
d. There were blank fields in the **gross_sales** column. I am assuming that a person needs to create an account to buy something from the website. So replaced the missing values in the **gross_sales** column by 0 where the new_user was 2. For all the **other missing gross_sales** fields replaced them with the **average gross_sales value .**
e. Split up the date to day and month, in order to analyze the trends during weekdays and the months.
f. Added a attribute called sales that determines if a sale was made or not. Used this field in order to predict if a sale was made or not.

Step 2: Data Analysis

a. Exploratory analysis:

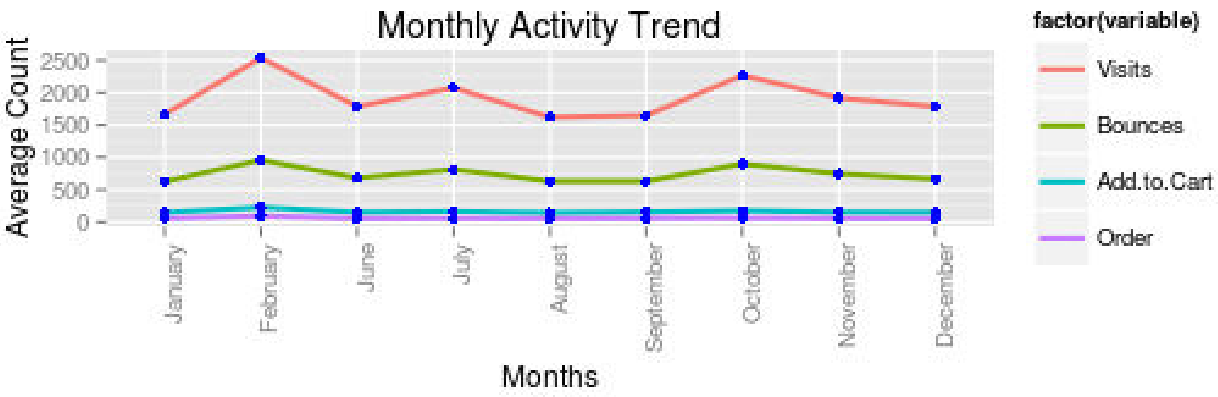
Let us first try to answer the question of what do people do on the website and who are they?

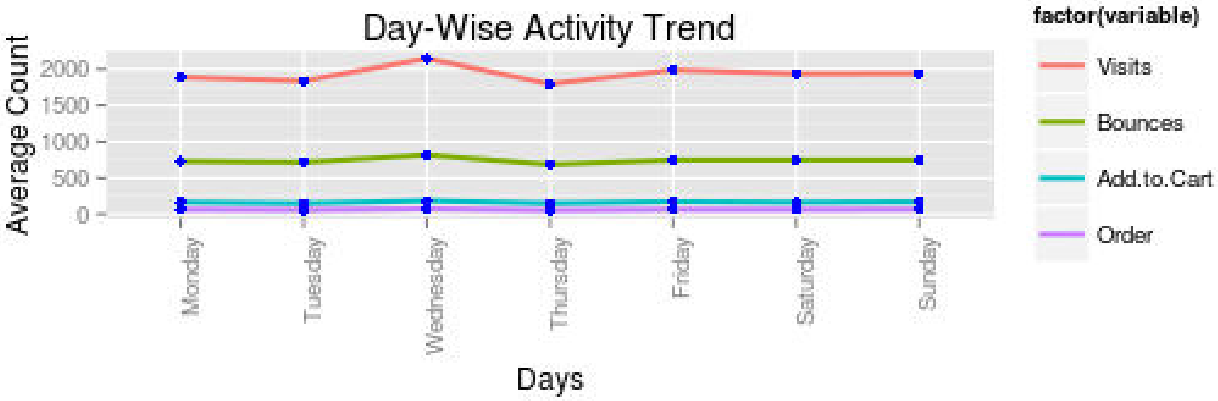




Inferences:
1. The first thing that seem obvious from the graph is that the number of customers that belong to the neither category is high as compared to the new and returning users. There might be a problem in logging in the details of the users on the website
2. The visits of the neither category is also high on average and in total. But the order rate is pretty low on average suggesting that there might be crawlers extracting information from the website.
3. The average and the total orders of the new customers is low as compared to the returning customers indicating that there might be some problem with the customer service or the usability of the websites
4. The negligibly zero bounce rate of the new customers gives an insight into the customer service or the usability issue with the websites.

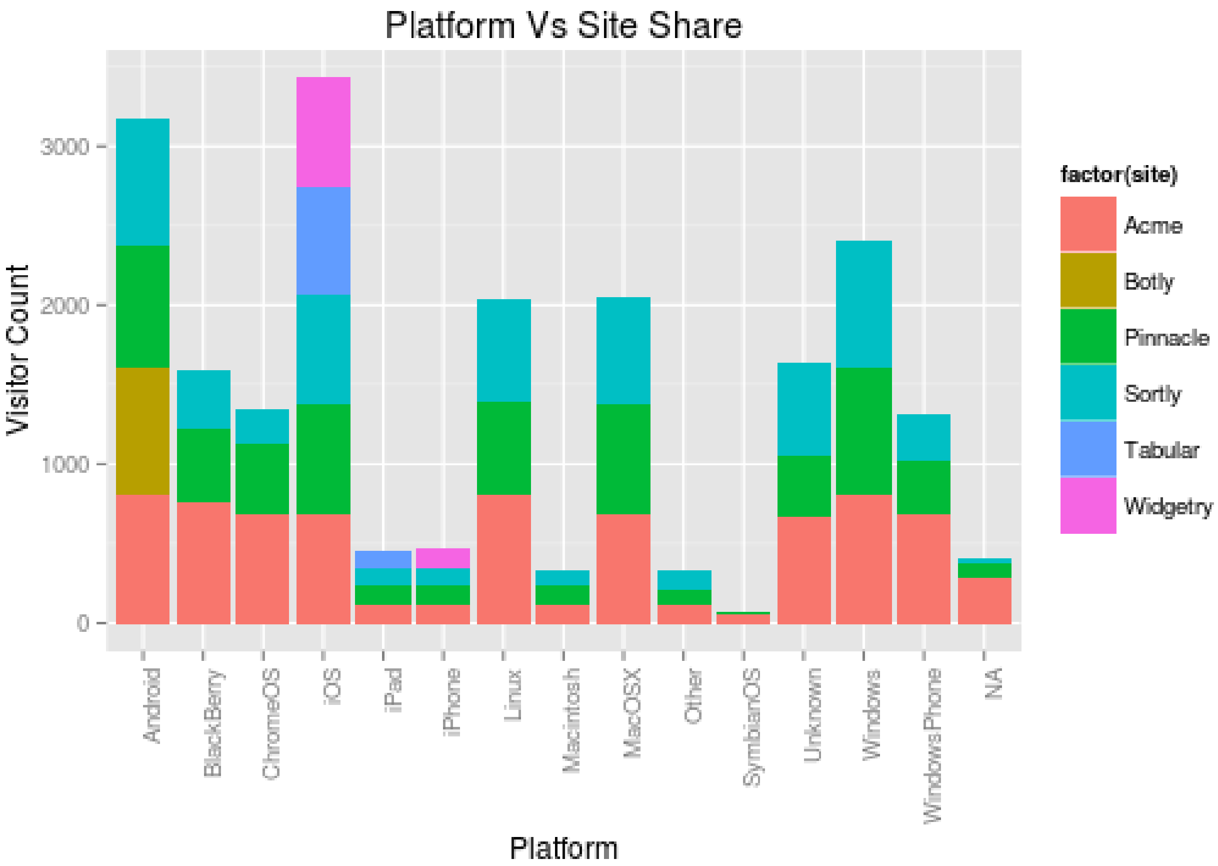
Now let us check the trend along various months and weekdays.





Inference:
As we can see from the plots the visits usually increase in the month of February and October. Similar trend can be seen in the visits on the weekdays.

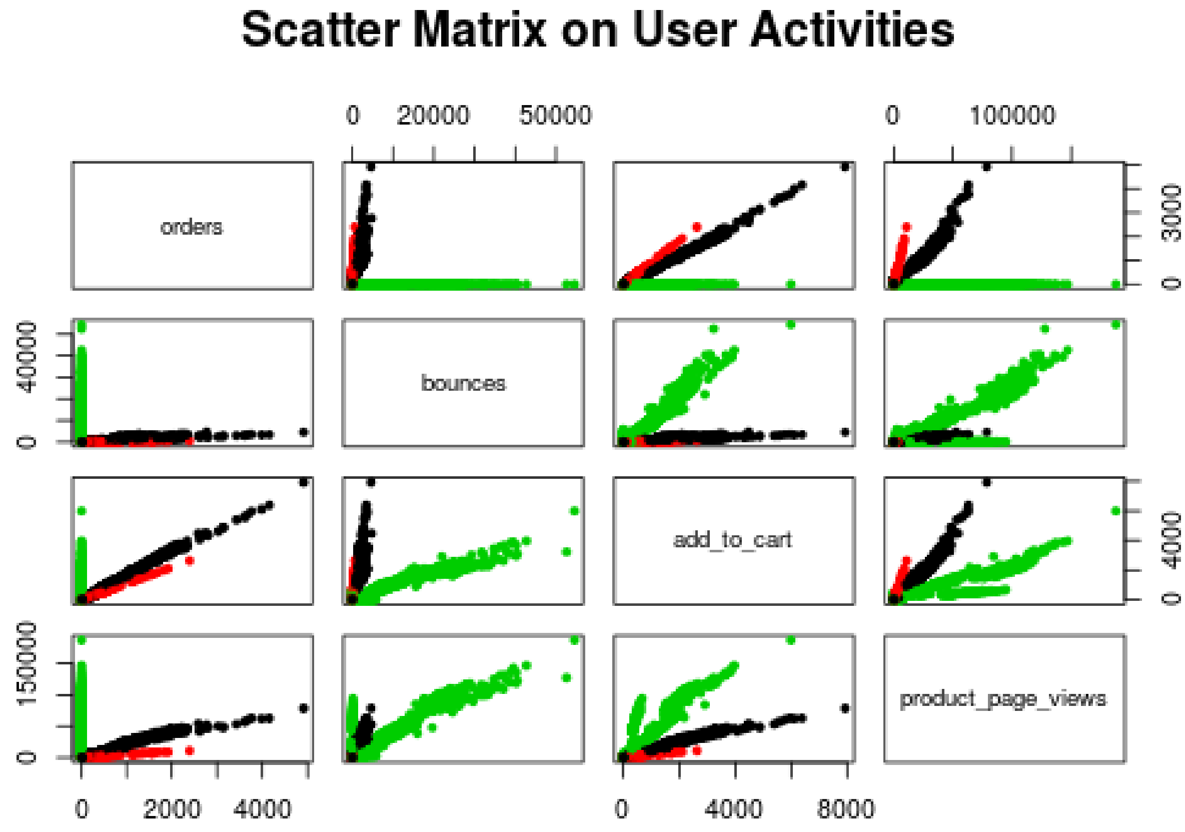
What kind of platform do the people use the most and what is the percentage share of each platform in the sales?



Inferences:

1. As we can see from the graph that **Android and iOS** are the most used platforms.
2. Acme,Botly and pinnacle are only three websites that are used on almost all of the platforms.
3. Websites like widgterly,sortly and tabular aren't that much on all other platforms. Work needs to be done on these websites inorder to increase their user base.

Are there any clusters in the data, and what do they indicate?



Inferences:
1. As we can see there are 3 natural clusters in the data[Red -> “New User”, Black -> “Returning User”, Green -> “Neither”]
2. We can also observe that three clusters are clearly separated from each other, thus we can easily identify a user's based on it's patterns and make separate plans to increase the traffic for each type of customers.

Step 3: Regression

What factors are responsible in making a sale on the website?

Let us try to answer this question. We added a binary attribute called sales to the data,that represents if a sale has been made or not.

Applying **logistic regression** and using **stepwise forward and backward selection** we get the following predictors for the sale attribute(followed by coefficient):
1. Search page views(positive less than 0.1)
2. Product page views(negative less than 0.1)
3. Bounce rate(negative ,-0.74)
4. Platform(Blackberry(N,-0.33); ChromeOS(N,-0.82); iOS(N,-0.26); Linux(N,-0.61); MacOSX(P,0.67) Other(N,-0.75); SymbianOS(N,-1.4); Unknown(P,1.16); Windows(P,0.46); WinPhone(P,-0.1))
5. Add to cart rate(positive,115.3)
6. Returning customer(negative, 1.43)

So those customers with **high product page views,low search page view,low bounce rate,explore through Windows or MacOSX devices have high add to rate and returning customers are likely to create more sales.**

Let us now try to get the factors affecting sales amount using **Multiple Linear Regression.**

In the final model the sales amount is related to:

1.Weekday(Positive on weekends,1.3)
2. Distinct Sessions(Negative, -0.6)
3. Visits(Positive, 0.5)
5. Platform Blackberry(N,-315); ChromeOS(N,-187); iOS(P,522); Linux(N,-113); MacOSX(P,1248) Other(P,380); SymbianOS(P,26); Unknown(N,-391); Windows(N,-3262); WinPhone(N,-220)
6. Add to cart rate (Negative, -0.39)
7. Orders (Positive, 369)
8. Returning customer (Negative, -143)

AIC value of the model-60.9
R squared-0.92

Strategy that can be used to increase the total sales:
1. Discounts/Coupons for returning customers
2. Better customer service and user experience
3. Regular emails about deals
4. Make separate adds for new customers