Zappos Analytics Challenge

Anirudh Pillai

July 14, 2016

Introduction

The first step to do before visualizing data is to understand our dataset.Lets check dimension of our dataset.

```
## [1] 21061    12
```

Our Dataset has 21061 rows and 12 columns.

I observed that dataset has lot of missing values and data for Months March, April and May are missing.Let's check each column name or feature name in our dataset.

```
##  [1] "day"                "site"               "new_customer"      
##  [4] "platform"           "visits"             "distinct_sessions" 
##  [7] "orders"             "gross_sales"        "bounces"           
## [10] "add_to_cart"        "product_page_views" "search_page_views"
```

Let us find datatype of each attribute.

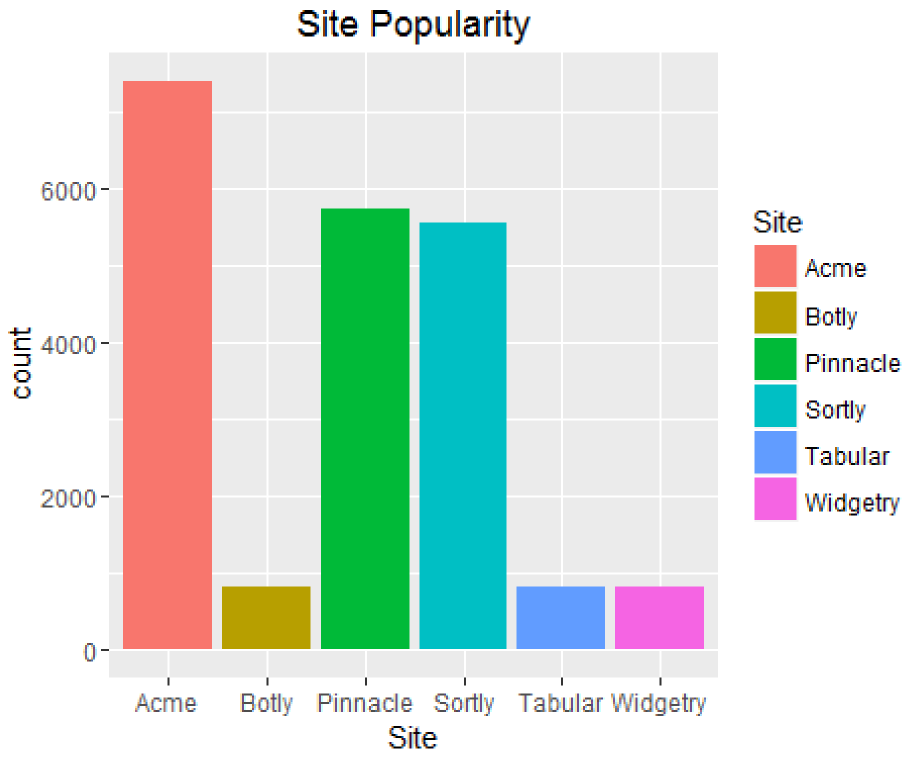
```
## 'data.frame':    21061 obs. of  12 variables:
##  $ day               : Factor w/ 268 levels "1/1/2013 0:00",..: 1 1 1 1 1
1 1 1 1 1 ...
##  $ site              : Factor w/ 6 levels "Acme","Botly",..: 1 1 4 1 2 1 4
4 1 1 ...
##  $ new_customer      : int  1 1 1 1 1 1 1 1 0 0 ...
##  $ platform          : Factor w/ 15 levels "","Android","BlackBerry",..: 2
3 6 14 2 9 2 14 8 7 ...
##  $ visits            : int  24 0 0 922 11 384 14 1 41 448 ...
##  $ distinct_sessions : int  16 0 0 520 10 214 10 0 27 368 ...
##  $ orders            : int  14 0 0 527 11 213 4 0 6 36 ...
##  $ gross_sales       : int  1287 13 98 60753 1090 28129 432 31 705 4637
...
##  $ bounces           : int  4 0 0 149 0 65 4 0 6 80 ...
##  $ add_to_cart       : int  16 0 0 610 11 245 7 0 12 79 ...
##  $ product_page_views: int  104 1 0 3914 4 1783 33 2 130 722 ...
##  $ search_page_views : int  192 0 0 7367 19 3255 52 2 272 1073 ...
```

Let's summarize our dataset.

```
##                day              site       new_customer      platform   
##  12/19/2013 0:00:   86   Acme    :7392   Min.   :0.000   iOS    :3435  
##  11/29/2013 0:00:   85   Botly   : 804   1st Qu.:0.000   Android:3172  
##  12/11/2013 0:00:   85   Pinnacle:5725   Median :0.000   Windows:2399  
##  12/7/2013 0:00 :   85   Sortly  :5532   Mean   :0.448   MacOSX :2054  
##  12/2/2013 0:00 :   84   Tabular : 804   3rd Qu.:1.000   Linux   :2036  
##  12/5/2013 0:00 :   84   Widgetry: 804   Max.   :1.000   Unknown:1641  
##  (Other)        :20552                   NA's   :8259    (Other):6324  
##      visits       distinct_sessions     orders          gross_sales    
##  Min.   :     0   Min.   :     0    Min.   :   0.00   Min.   :     1  
##  1st Qu.:     3   1st Qu.:     2    1st Qu.:   0.00   1st Qu.:    79  
##  Median :    24   Median :    19    Median :   0.00   Median :   851  
##  Mean   :  1935   Mean   :  1515    Mean   :  62.38   Mean   : 16473  
##  3rd Qu.:   360   3rd Qu.:   274    3rd Qu.:   7.00   3rd Qu.:  3145  
##  Max.   :136057   Max.   :107104    Max.   :4916.00   Max.   :707642  
##                                                       NA's   :9576    
##     bounces           add_to_cart     product_page_views search_page_views
##  Min.   :    0.0   Min.   :   0.0   Min.   :     0     Min.   :     0   
##  1st Qu.:    0.0   1st Qu.:   0.0   1st Qu.:     3     1st Qu.:     4   
##  Median :    5.0   Median :   4.0   Median :    53     Median :    82   
##  Mean   :  743.3   Mean   : 166.3   Mean   :  4358     Mean   :  8584   
##  3rd Qu.:   97.0   3rd Qu.:  43.0   3rd Qu.:   708     3rd Qu.:  1229   
##  Max.   :54512.0   Max.   :7924.0   Max.   :187601     Max.   :506629   
## 
```

I have added new column called month which contains initials of each Month.e.g:- Jan, Feb etc which will help when grouping attributes monthly.I have also calculated conversion rate, bounce rate and add to cart rate and added column for each in our dataset.

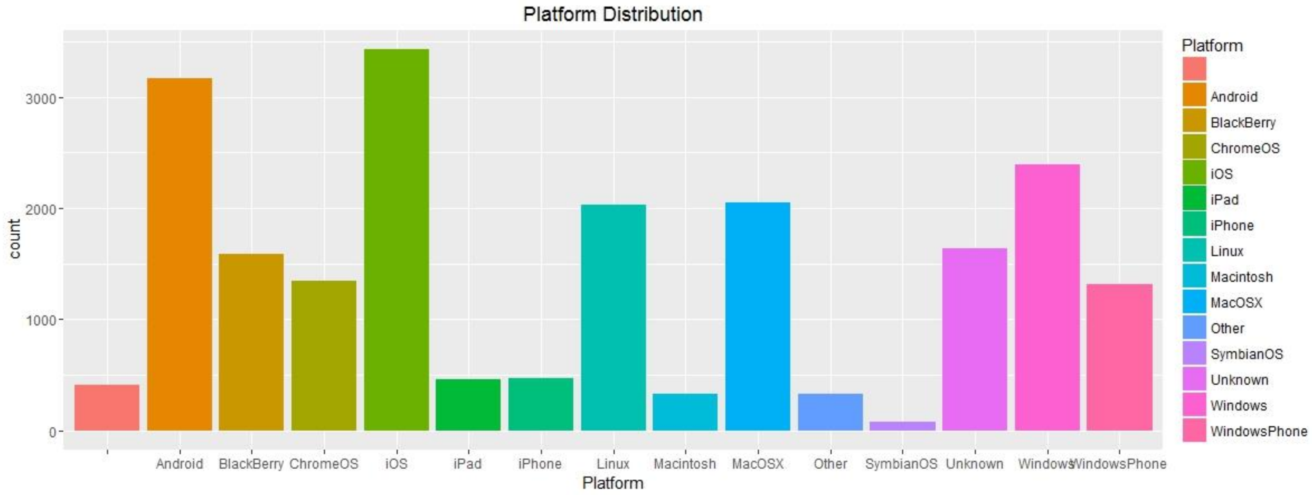
Let us now check which site is most visisted by our customers.



Clearly Acme is favorite among users as seen from plot.Pinnacle and Sortly are not far behind and are doing ok. Botly, Tabular and Widgetry are on the same page with minimum customers visiting the website. In my opinion we shoud:

- promote Botly, Tabular and Widgetry by publishing their contents or offers as advertisments on Acme,Pinnacle and Sortly.
- provide huge discounts or deals initially to lure customers.

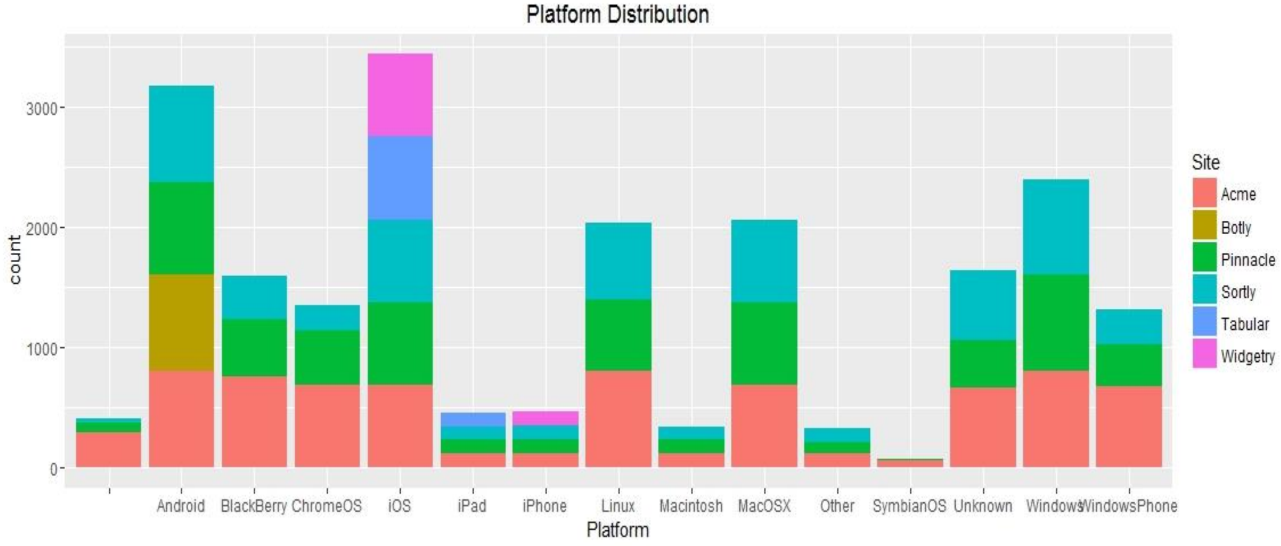
Let us check Platform Popularity or which platform is widely used by customers to visit our sites.



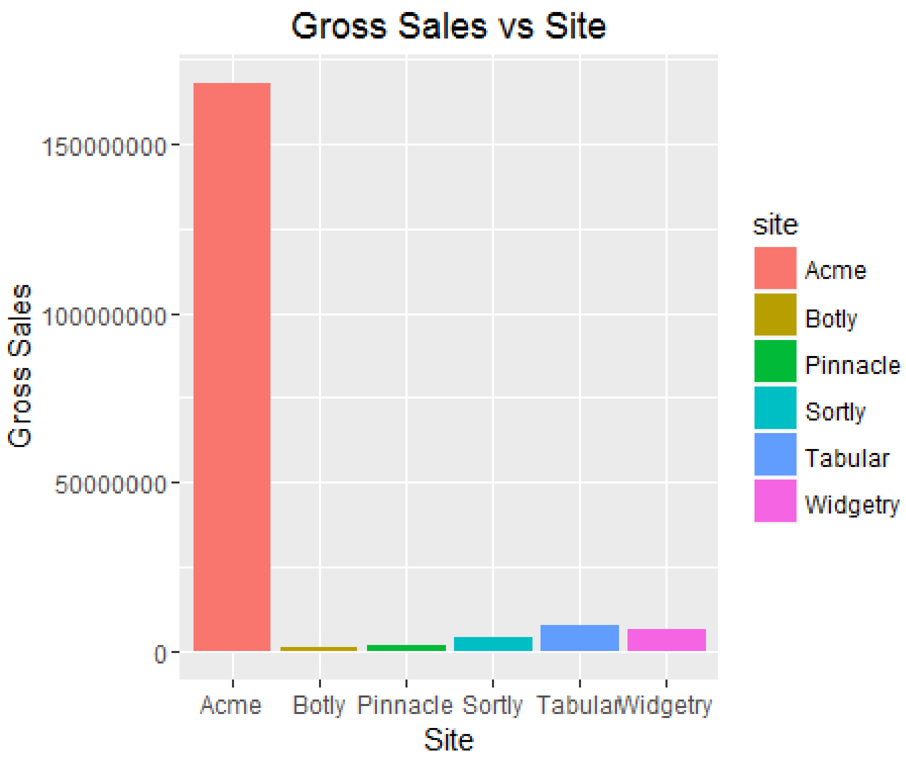
Both Android and iOS are seen to be widely popular platforms among customers for browsing our sites but iOS has slight edge compared to Android when we compare the exact count.

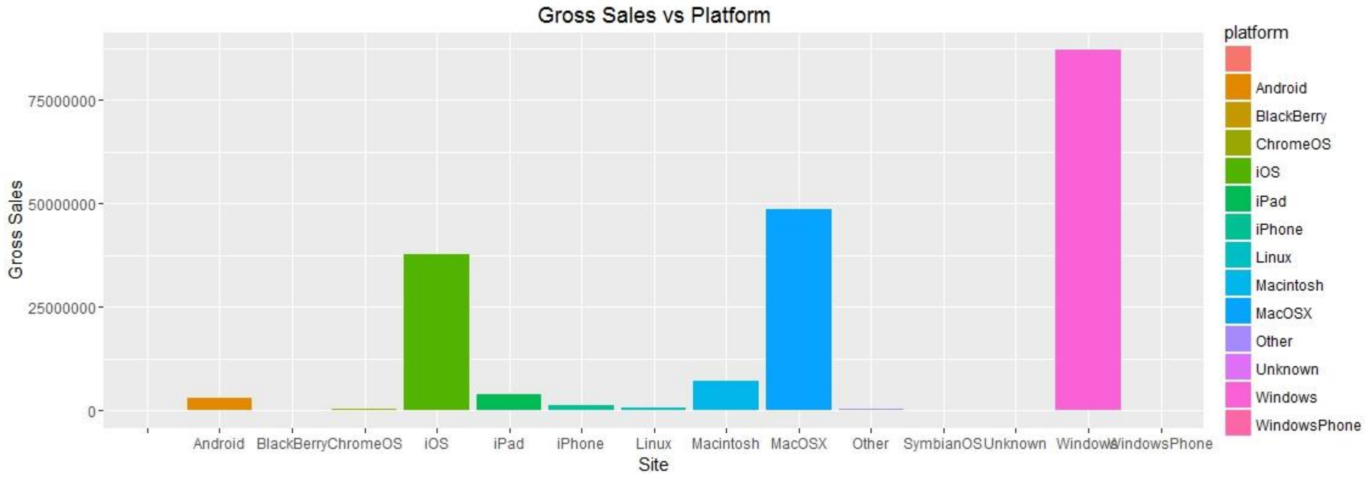
If we divide Platform into sub classes like Mobile OS and System OS. We can see that iOS tops the chart among smartphones and tablets whereas Windows slightly edges past all the other Operating systems.

We can further plot stacked bar chart indicating popularity of sites in each platform.



Let's find out our highest revenue generating website and platform.

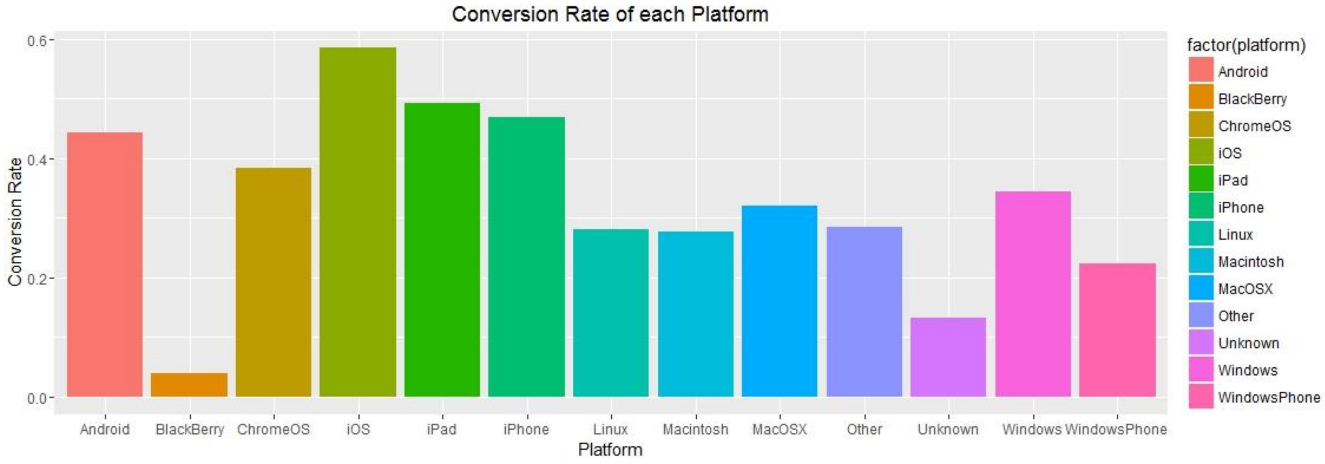




Clearly, Acme generates way more revenue than any other website.We should take steps to improve revenue from other websites.

** Windows Desktops, MacOSX, iOS Devices generate majority of revenue with Windows being top performer.

Let us check conversion rate for each platform.

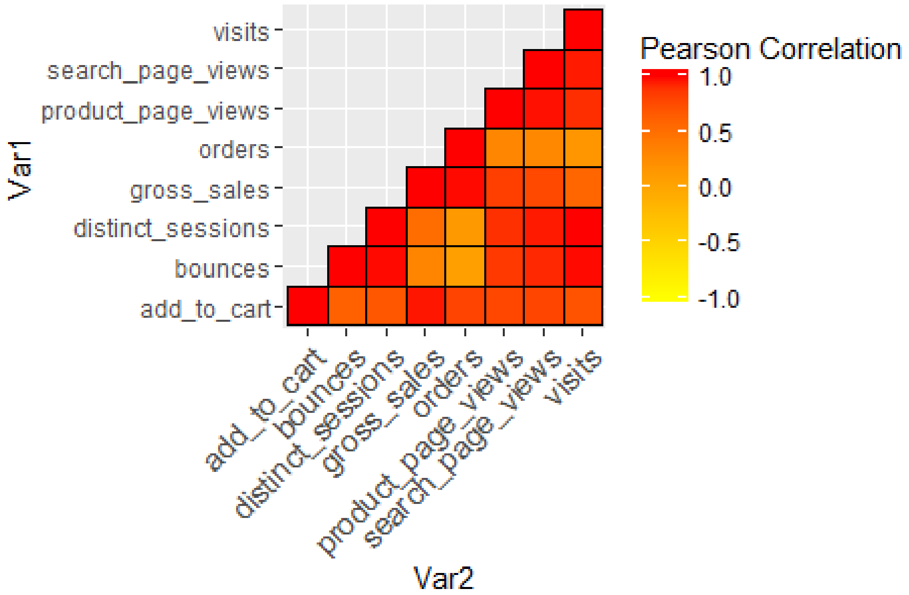


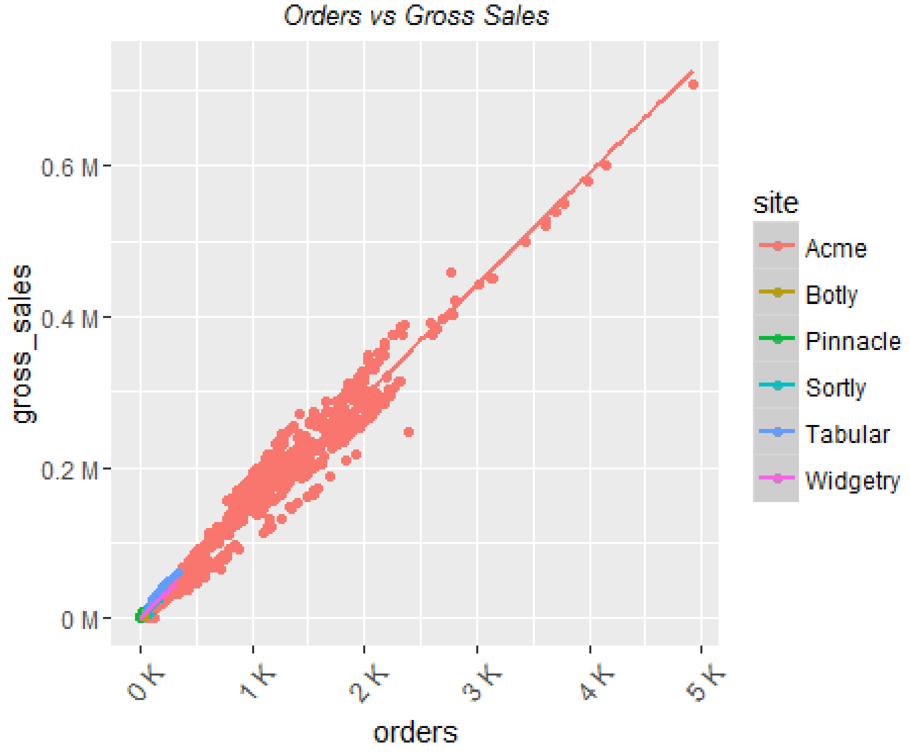
Conversion rates for Blackberry and Symbian OS is very low. We can ignore Symbian OS since there are only 74 users who have used symbian device for browsing. Moreover Symbion is no longer popular or widely used among users.

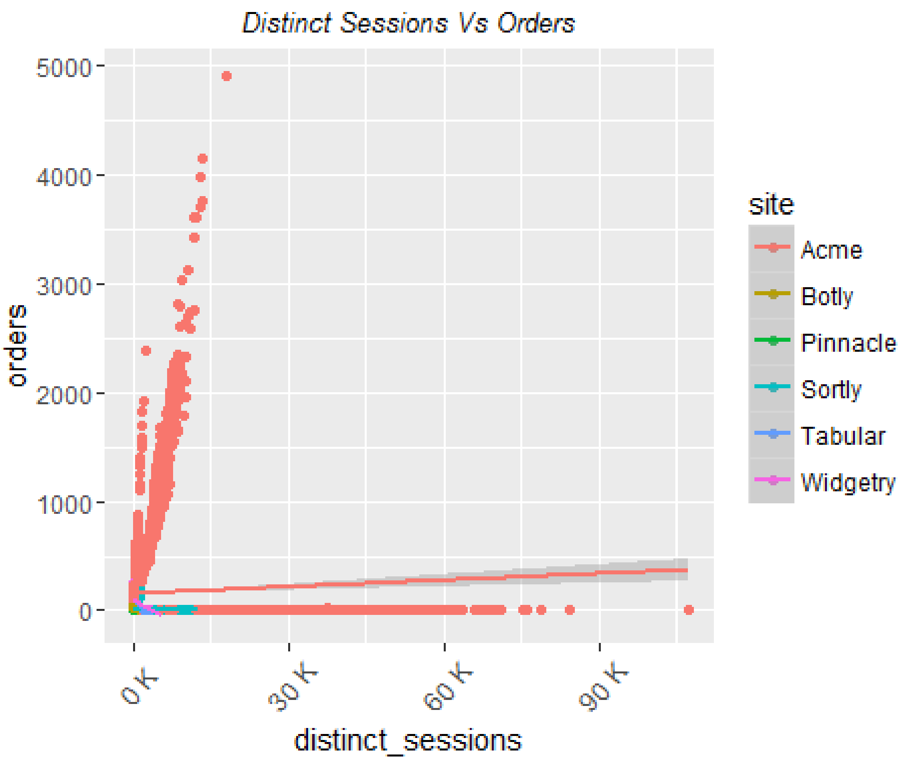
We have 1574 customers who use Blackberry device for browsing our websites and we should take steps to improve conversion rate on this platform by providing some Blackberry specific deal or offers.

Key Performance Index

Since there are so many parameters in our dataset which can be used to judge performance of webistes so let's make a correlation matrix of all key performance index.


visits
search_page_views
product_page_views
orders
gross_sales
distinct_sessions
bounces
add_to_cart
Var1
add_to_cart
bounces
distinct_sessions
gross_sales
orders
product_page_views
search_page_views
visits
Var2
Pearson Correlation
1.0
0.5
0.0
-0.5
-1.0


Orders vs Gross Sales
gross_sales
0.6 M
0.4 M
0.2 M
0 M
0 K
1 K
2 K
3 K
4 K
5 K
orders
site
Acme
Botly
Pinnacle
Sortly
Tabular
Widgetry



My observations from above is

- All the Key Performance Indicators have positive correlations
- Low correlation between Vists and Orders.
- Correlation between Vists and Bounces is very high and we need some kind of load balancer to handle traffic specially when there are lot of customers or during festive season. I also recommend load testing and stress testing from my personal experience.

Let's check average bounce rate per site.

```
##       site  Average Bounce Rate
## 1     Acme          0.313272750
## 2    Botly          0.171117865
## 3 Pinnacle          0.478650826
## 4   Sortly          0.382362938
## 5  Tabular          0.027045166
## 6 Widgetry          0.007752247

##       site  Number of Visits
## 1     Acme          35593107
## 2    Botly            156598
## 3 Pinnacle            636269
## 4   Sortly           2449196
## 5  Tabular            713049
## 6 Widgetry           1198667
```

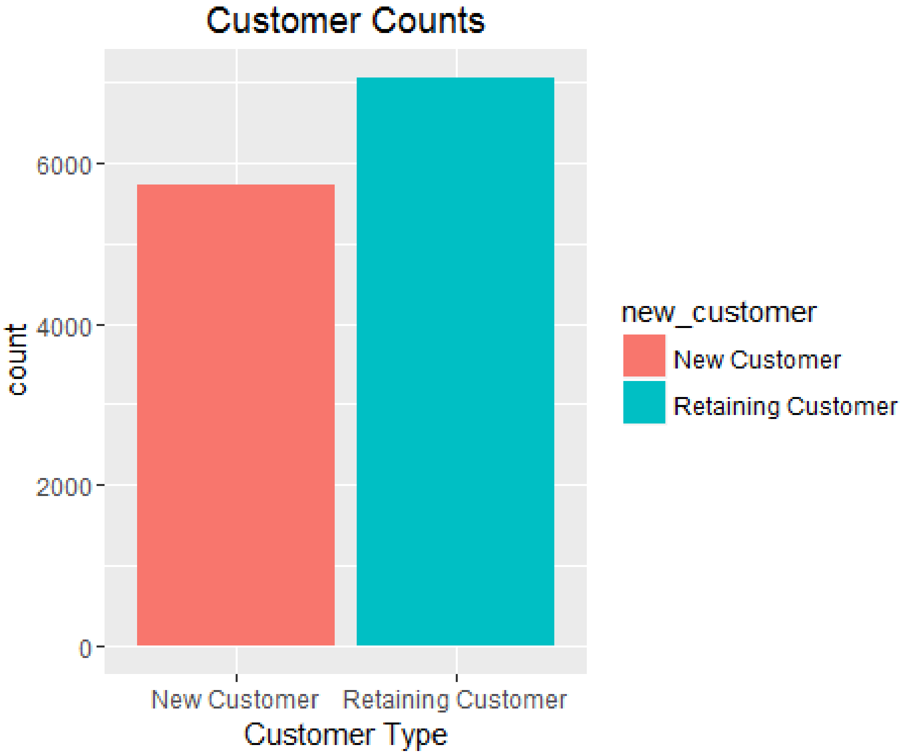
Bounce Rate represents the percentage of visitors who enter the site and then leave ("bounce") rather than continuing on to view other pages within the same site.

Bounce rate looks pretty decent and shows that customers do navigate to other pages in our webistes.

As seen from the results:

- Acme is our most visited website
- Botly is our least visited website.

These results help us to understand customer behavior and we can improve their experiences in areas we lack and get better results.



We have 7066 returning customer which shows that our customers had positive experience in past and are loyal to the brand. We have 5736 new customers which is equally good.

Future Implementations Possible

- Add Heatmap indicating busiest months. (I tried to implement but was getting error related to package version for which I could not find any solution).
- We can also have visualization related to new customers and retaining customers who add items to carts and their sessions.
- We can Implement Regression to predict sales for next year or Time series analysis to predict sales for each month next year.