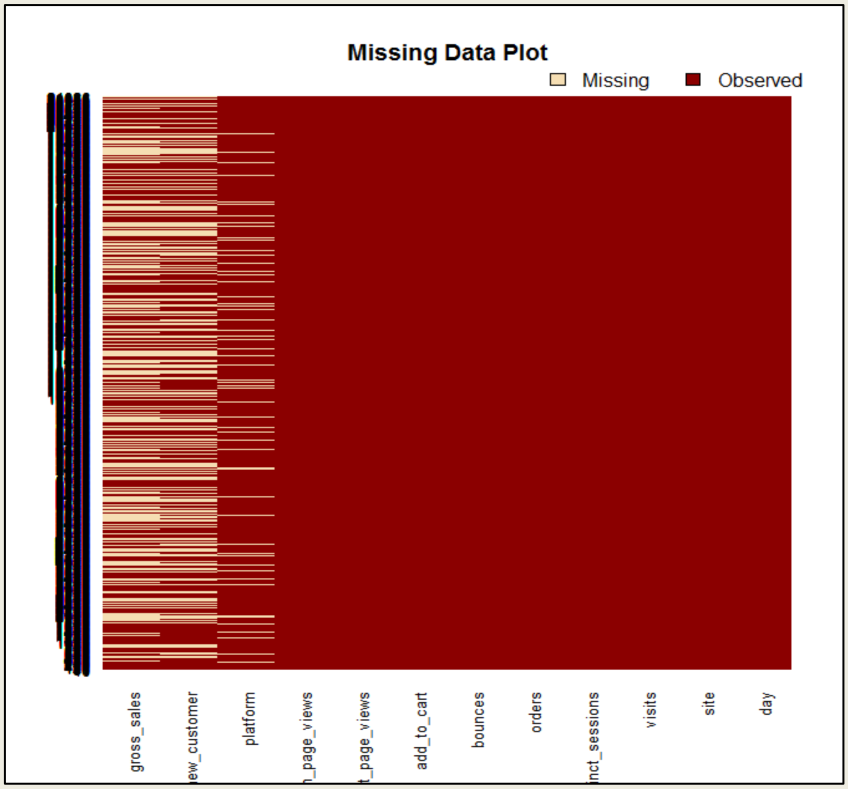
Zappos Advanced Analytics Summer Internship 2015 - Project Report

EXECUTIVE SUMMARY:

This project report analysis the dataset provided by Zappos as part of the 'Advanced Analytics Summer Internship 2015'. Entire analysis is done using an open source statistical language known as 'R'. The analysis of the data is divided into three major parts. The first part deals with reading the data and the missing value imputation in which a custom imputation technique and neural networks is used. This is followed by some exploratory graphs and analysis on the clean dataset. Further, the required metrics are calculated and supplementary graphs are made for additional exploration on these metrics. Furthermore, we perform a regression analysis to establish relationship between the orders placed and other variables. Finally, the report ends with a short conclusion summarizing all the findings. The lists of all the libraries used along with the user-defined functions are mentioned in the appendix.

GETTING AND CLEANING THE DATA:

The original data is in '.xlsx' format. For ease of analysis in R the excel file has been converted to comma-separated values format.



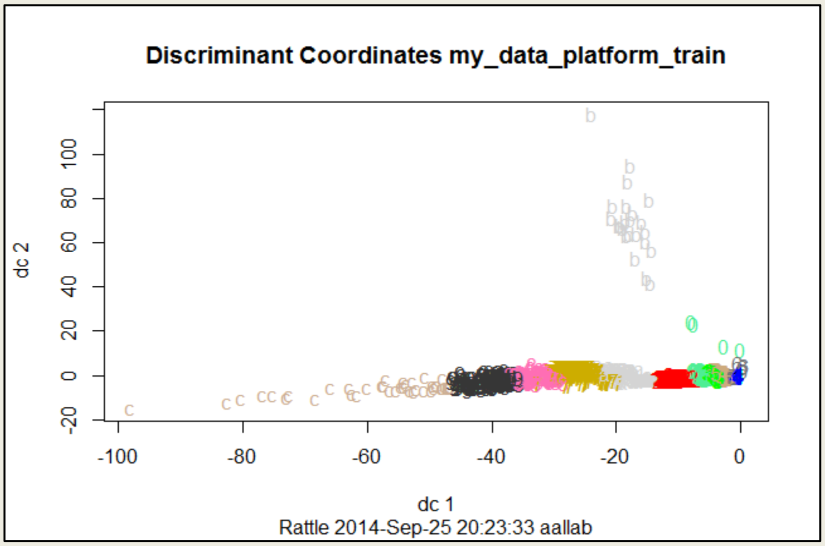
After reading the data into R we look for missing values in the data. We first try excluding the missing data. The amount of missing data, which gets excluded, is almost half of the total data available. So it is not advisable to waste that much data. Hence, we look to impute the missing data before diving deeper into our analysis. From the plot above we can see there is missing data only in columns named "new_customer" and "gross_sales". But on further exploration of the dataset it is also clear that there is some missing data in the "platform" variable. It is important to understand that garbage data will led to garbage analysis. Thus, great importance should be given to data imputation as they can lead to very erroneous results. First, we will impute the "gross_sales" column, followed by the "platform" variable and finally use neural network to impute "new_customer" column. For the ease of analysis the day column has been converted to a standardized date format which will be quiet useful in later parts of the analysis.

A custom made function has been used to impute missing data in the "gross_sales" column. This function takes in three arguments, a dataset, column_impute and a factor column. The values in the column_impute are averaged for each factor level and the missing values are replaced by one of these average values depending on the factor level they correspond to.

The mean table below lists the mean of the "gross_sales" from the original data and from the imputed tables. It is observed that when "gross_sales" is imputed using "day" as the imputing factor the mean is closest to that of the original. Thus, the data is imputed using day as the imputing factor.

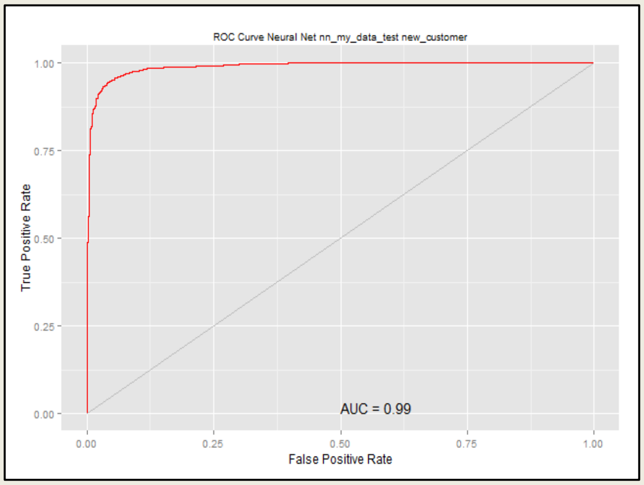
Original_mean	Mean_byDay	Mean_bySite
16473.40	16428.30	13886.19

The platform column has missing data in two forms. One is blank spaces and the other is categorized as unknown. Both these factor levels are converted to 'NA', which is the standard representation of missing values in 'R'. There are a total of 13 factor levels in the platform variable. Hence, we try to perform a cluster analysis using hierarchical clustering. A bunch of different clustering parameters are tried out the most successful of which is plotted below. But it is visible that the clusters overlap a lot and thus the error rate will be quiet high. Thus, missing values from the "platform" column are excluded.



Further, focus is shifted to the “new_customer” variable. The rattle package is used to quickly explore which technique can be used for its prediction. After looking at couple of different techniques we reach the conclusion that neural network gives the least error rate on the test data set.

ERROR MATRIX	Predicted Values		
Actual Values		0	1
	0	1748	52
	1	118	1405

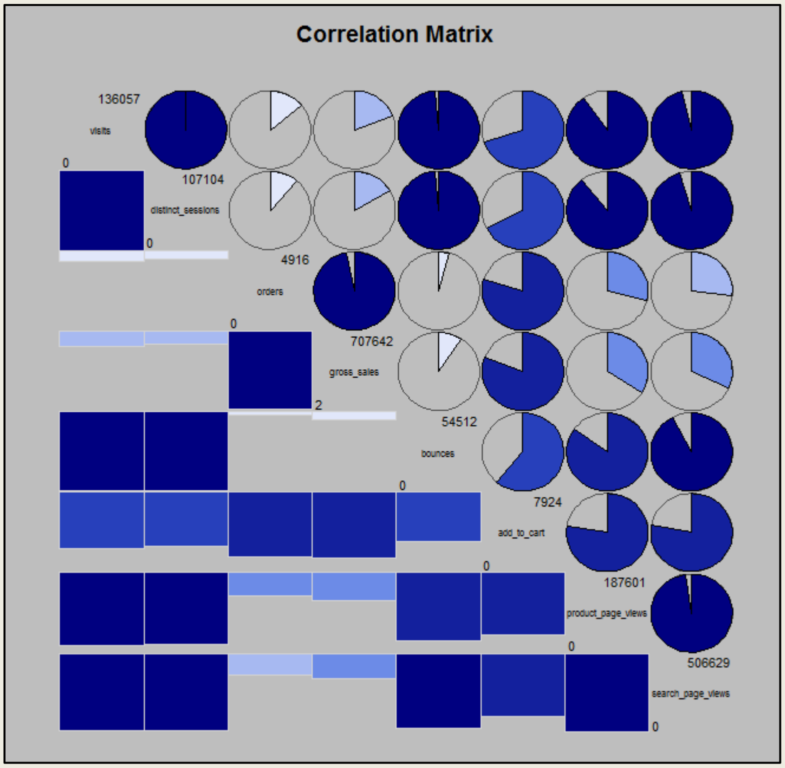


For purpose of analysis two data sets are created. The my_data_ml dataset contains the rows with "new customer" variable values present in it. The my_data_predict contains the rows with missing values in "new customer" variable. The my_data_ml dataset is used to train our neural network and then from that the missing values in my_data_predict dataset are predicted. The error matrix generated along with the ROC curve confirms the accuracy of the neural network model. The accuracy of the neural network model is nearly 93%.

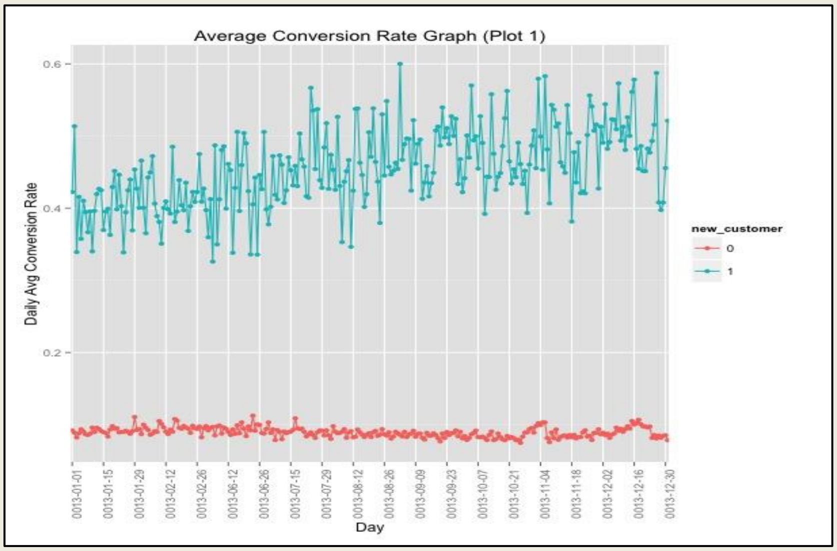
The “conversion_rate”, “bounce_rate” and “add_to_cart_rate” are calculated as required. As some of the entries in the visits variable are zeros we simply replace those corresponding values in the above metrics with zeros to avoid 'division by zero' error. The required metrics are attached in the final dataset that is generated.

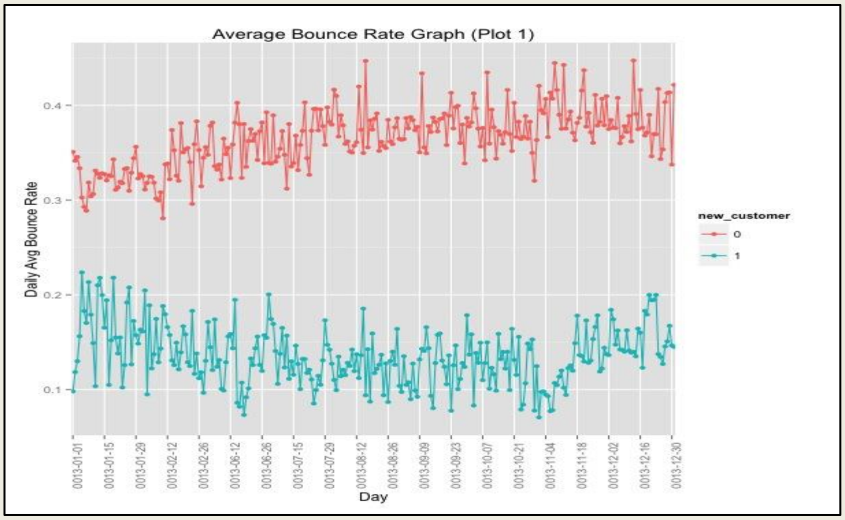
EXPLORATORY DATA ANALYSIS:

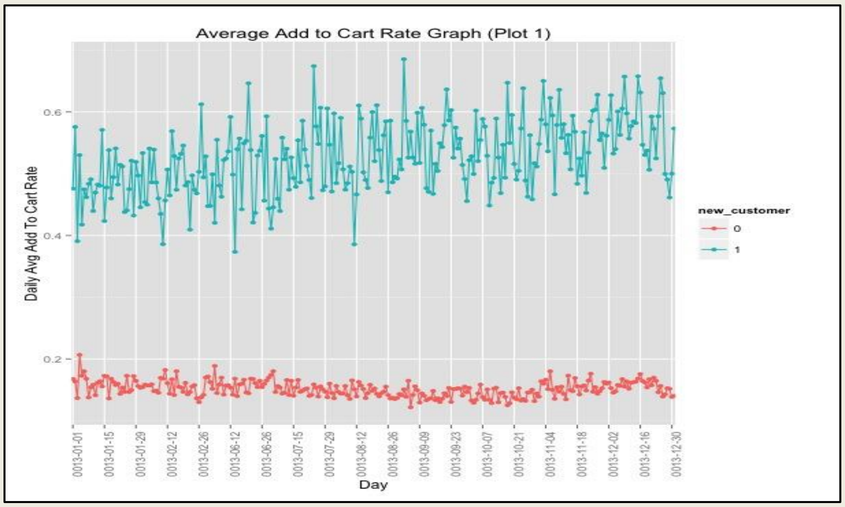
We first start by taking a look at the correlation matrix, which is shown below. The above correlation matrix shows some interesting results. The amount of filled circles symbolizes the degree of correlation. For example “visits” variable is very highly correlated to “distinct_sessions”, which was quite expected. “Product_page_views” is highly correlated to “search_page_views”. The number of visits and the number of bounces are also highly correlated. The lower part of the matrix also represents the same thing in a different manner. Additionally, the diagonal column displays the name of the variable along with the maximum and minimum value of that variable in the dataset. But it should be kept in mind that sometimes the correlation matrix could be misleading, as correlation doesn’t necessarily mean causality.



Now, we plot the change in daily average “conversion_rate”, “bounce_rate” and “add_to_cart_rate” for all the factors namely new_customer, platform and site. There are three main plots. Each main plot contains three sub graphs. Plot 1 represents the change in all the three rates on a daily basis both for old customer and new customers. Plot 2 represents the daily change in the above-mentioned rates for various sites and plot 3 displays the daily change in the above-mentioned rates for various platforms.

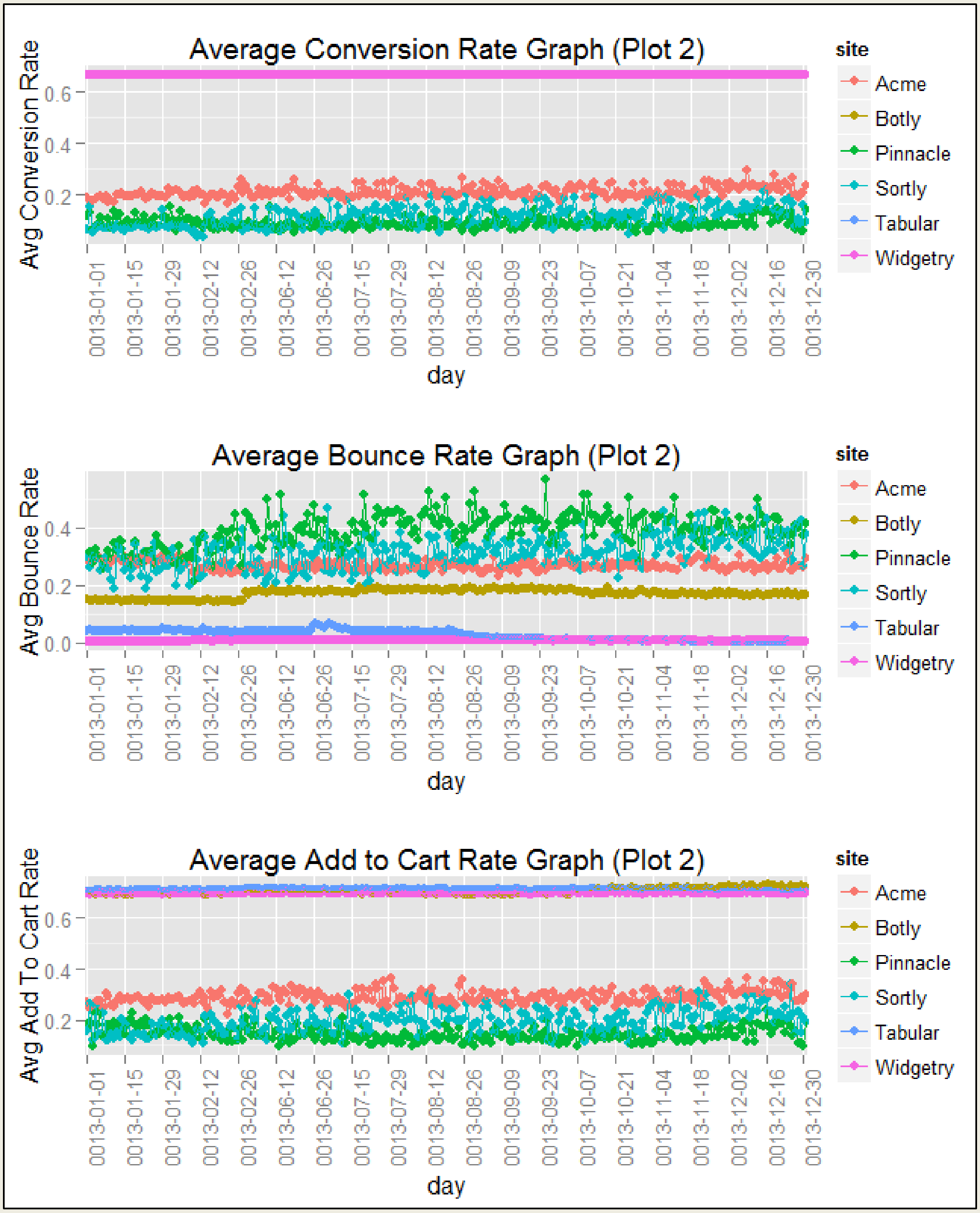




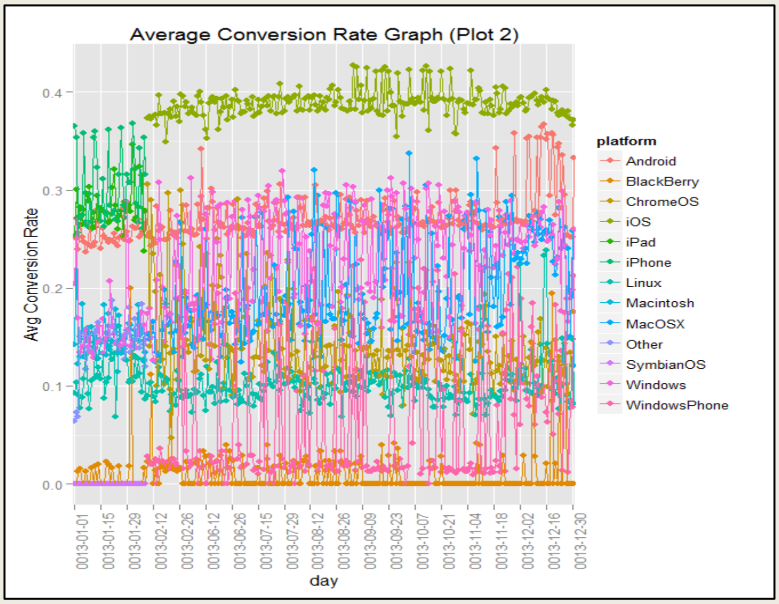


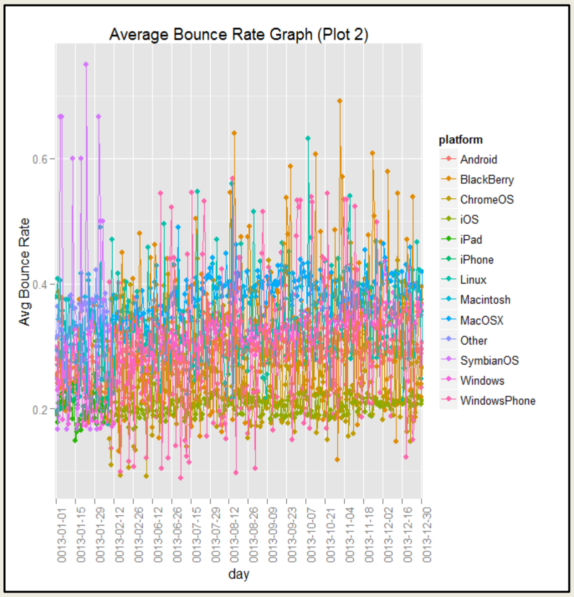
Plot 1 results: The above plots are the daily average change of rate plots. On observing the plots we observe that the daily average "conversion_rate" and daily average "add_to_cart_rate" is more for "new_customers" while the bounce rate is more for returning customers. Moreover, the daily average "conversion_rate" and the daily average "add_to_cart_rate" have a slightly increasing trend as the day's progress. Also, the daily average "bounce_rate" has a slightly increasing trend for returning customers as the day's progress. This also tells us that there is conclusive evidence that the regression model that will be built later should include "new_customer" as a significant input variable.

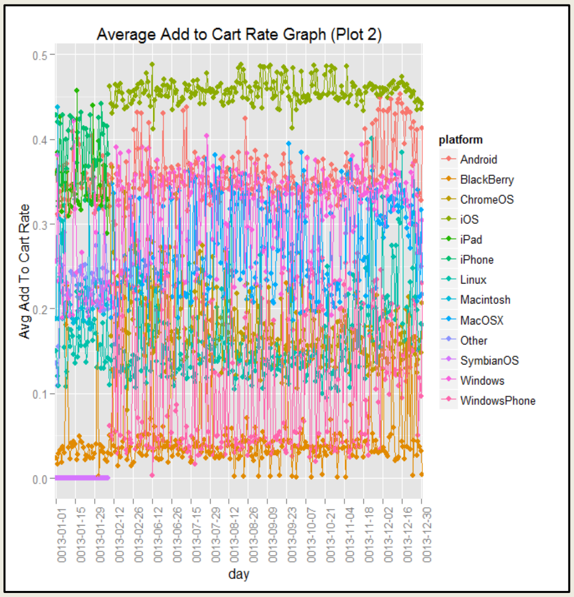
Plot 2 results: Plot 2 represents the daily change in all the three rates for different sites. We can see that for the "conversion_rate" and the "add_to_cart_rate" is low for Pinnacle, Sortly and Acme. Amongst these the lowest "conversion_rate" and "add_to_cart_rate" is lowest for Pinnacle. While the Widgetry, Tabular and Botly have a "conversion_rate" of more than 0.6. The bounce_rate graph is exact opposite of the "conversion_rate".



Plot 3 results: From the plot 3 we see that platform variable is not a very good separator for the various rates. The only definite result that can be predicted from this graph is that “conversion_rate” and “add_to_cart_rate” is highest for the iOS platform.







REGRESSION ANALYSIS:

Furthermore, regression analysis is performed to explore the effects of other variables on the number of orders individually. Regression modeling techniques are preferred over others as the regression coefficients are easy to interpret. Keeping that in mind orders is taken as the response variable and rest as the input variables. As order is a count variable that is it can never take negative values we use Poisson regression technique for modeling purposes. We decide to include "new_customer" into to model and exclude site and platform factor variables from the regression model owing to the results from the above plots. Additionally, highly correlated variables are also eliminated from the models. The variable with greater effect on the response variable is retained while the variable with less effect is eliminated. The variables "distinct_session" and "bounces" are eliminated as they are very highly correlated to “visits” variable. The variable “gross_sales” is eliminated as it is very highly correlated to orders and it will be fruitless for our analysis.

Additionally, we use the *Akaike Information Criterion (AIC)* and ANOVA testing for model selection. AIC is a good measure which accounts for the tradeoff between the model complexity and goodness of fit. Using ANOVA we eliminate “product_page_views” from our regression model as removing it doesn’t affect the response significantly. ANOVA also strengthens our claim that the variables “bounces” and “distinct_sessions” are insignificant. Also it is important to note that we have taken order as our response variable and not conversion_rate so that a direct analysis can be performed on the data that is collected without any transformation on it. This will make the regression coefficients directly usable for interpretation. The regression coefficients along with their confidence intervals for the final model are as shown below:

Variable Name	**Regression coefficient**	**2.5%**	**97.5%**
Intercept	3.7	3.699677	3.705071
New_customer1(level 1)	0.092	0.08707347	0.09639833
Visits	-0.0002	-0.0001941475	-0.0001916218
Add_to_cart	0.0011	0.001125126	0.001131645
Search_page_views	0.00002260652	0.00002242716	0.00002278554

One important thing to be noted is that the Poisson regression coefficients should not be interpreted like logistics regression or linear regression coefficients. Thus, if there is one unit increase in the input variable then the expected value of the response variable is multiplied by exp(b1), where b1 = Poisson regression coefficient of that input variable.

CONCLUSION:

Prior to the analysis the data was completely cleaned by imputation of missing values and elimination. Further exploratory data analysis was performed on the clean data set. During the exploratory analysis it was found that there is a significant difference between the daily change in rates for new customers and returning customers. Likewise, to conclusively prove this analysis and establish the effect of other numeric input variables on the number of orders placed regression analysis was performed. The regression analysis supports the claim that orders placed are significantly higher for new customers when compared to returning customers as the limits of the confidence intervals do not overlap each other. Moreover, the effect of visits, add_to_cart variable and the search_page_views variable can also be inferred from the Poisson regression coefficients.

APPENDIX:

1] **imputeByFactor:** Function which takes in arguments a dataset, column to be imputed and a factor column. The values in the column_impute are averaged for each factor and the missing values are replaced by one of these average values depending on which factor they correspond to.

2] **Multiplot:** This function is copied from *"The cookbook for R website"* entirely. It draws multiple plots in the same figure. The code for both the functions is at the end of this document.