Zappos.com Inc Analysis Report

Introduction:

This is an analysis of the sample data set from Zappos.com Inc., It contains details about the various sites dealing with various purchases via online transactions. We are going to analyze how the purchases are done, what platform is used for the purchase and many other related information. Various graphs are plotted for finding the behavior of various fields. The concept of correlation, regression, mean and other statistical techniques are used to analyze this dataset.

Data Set Explanation:

In general, the given dataset consists of the orders placed by various sites by various users. When you closely look at the other fields in the dataset it gives additional information to the basic features. The various fields of the dataset are the 'date' when the order is placed or visit is made. There is a factor set for new users and returning users (previous users). It also consists of bounces in the visits, search pages and details of the visits that added product to the cart. The dataset consists of all the information but it is not possible to analyze and deduce from that dataset.

Analysis Language – R:

R is the language used for the analysis of this problem. R is a strong statistical tool that supports a variety of visualization packages. It also hold good for various statistical techniques like mean, correlation, scatter plots, regressions, etc., Here the given data is converted to a readable form for R, then it is loaded to R environment and then the analysis is done. Various features of R like **ggplots** and plots are used.

Apart from these graphs used, a special package called **shiny** is used. This shiny package helps in creating an application that has HTML code in the front end and runs an R script at the back end. The application has the controllers to handle the page in "ui.R" while the real calculations are done in "server.R".

Techniques:

Correlation is a technique used to find the dependence between two variables containing a set of values. The formula used for finding the correlation is

$$Cor(x,y) = 1/(n-1) \sum ((x - x`) / s_x) * ((y - y`) / s_y)$$

The value calculated from the correlation can be positive or negative. If it is a positive value then there is a strong correlation between these variables.

Regression is another statistical technique used for the analysis, here the relation between two variables in which one dependent and one independent variable. The regression line can be used to predict the deviation of the graph from the expected graph.

Analysis of the Dataset:

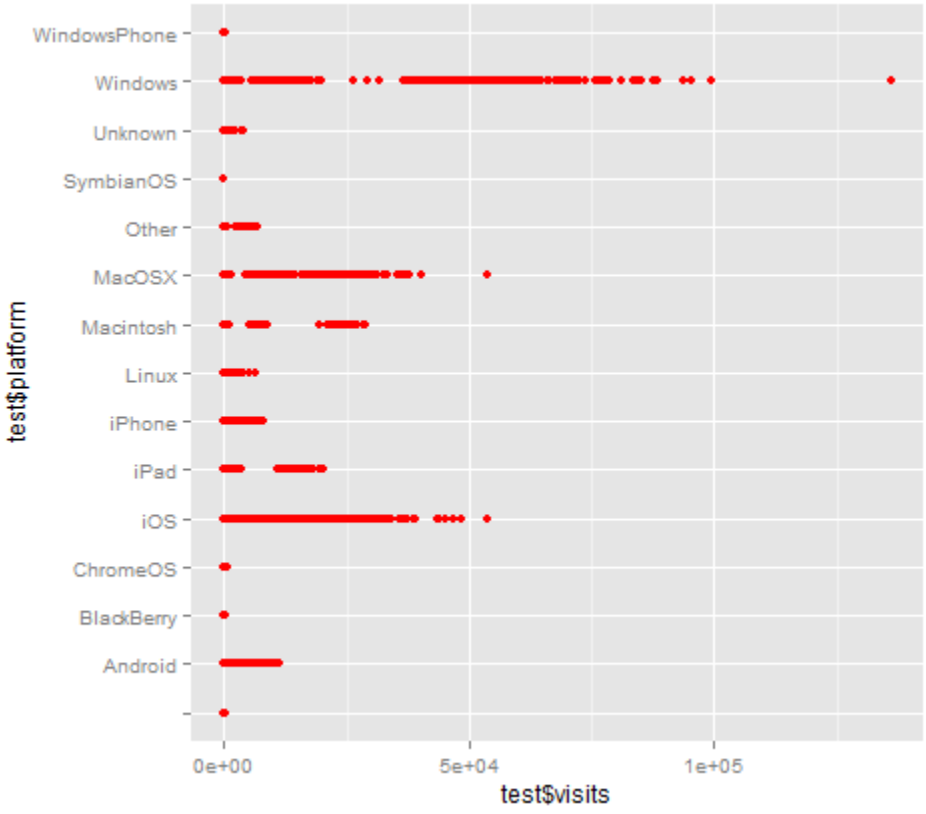
1. Basic Analysis

The basic analysis is first done on the dataset and based on the outcome of this analysis deeper analysis is done. The first step in analysis is loading the dataset in the R environment and finding the dimension of it.

```
> test <- read.csv("data1.csv",header = T,sep = ",")
> dim(test)
[1] 21061    12
```

- Various levels of the data set are found and then the graphs are plotted based on their value s
- The plot between the visits on various platforms are plotted using ggplots and the output is as follows :

```
> ggplot(test,aes(test$visits))+geom_point(aes(y=test$platform),colour
="red")
```



This graph shows that there are more visits done by windows users and the second highest are MacOS X and iOS.

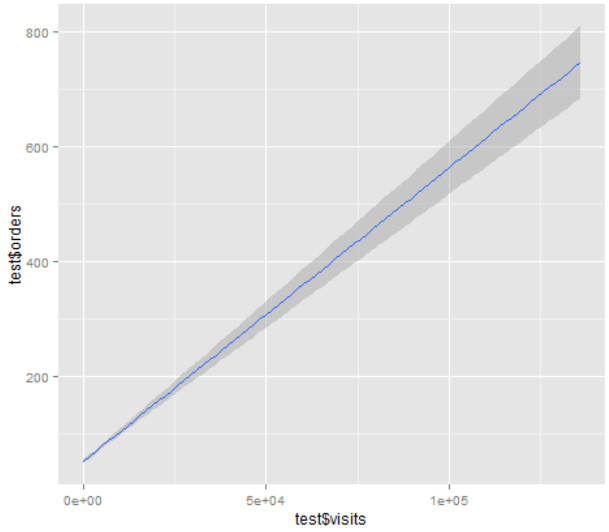
- The correlation between distinct_sessions and visits are found. This is done in order to get the relation between them

```
> cor(test$distinct_sessions,test$visits)
[1] 0.9987397
```

The value is 0.9987 which is closer to 1; this means that most of the visitors have visited the site only once.

- A graph is plotted between visits and orders; regression model is framed between them. The code for this can be like

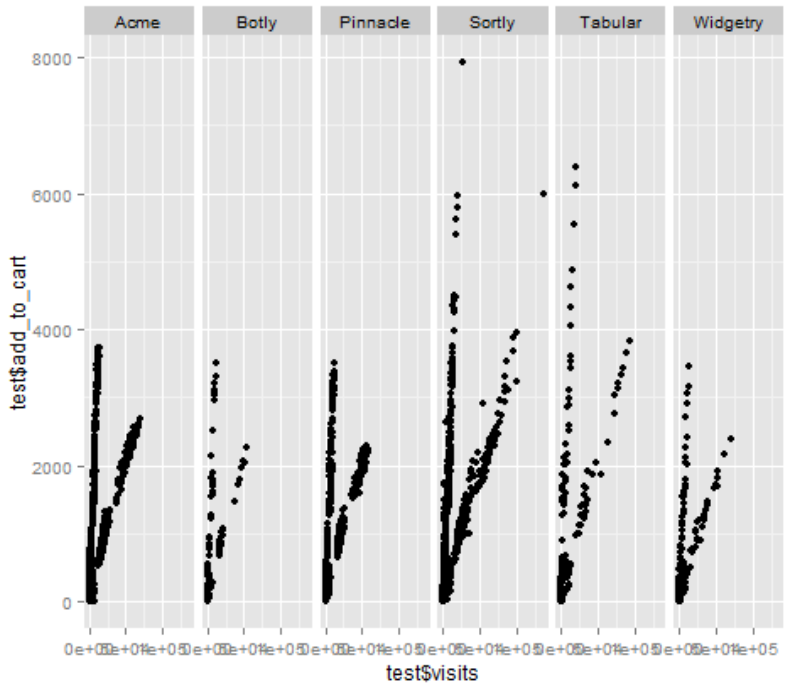
```
> qplot(data=test, x = test$visits,y = test$orders, method = "lm", geom = "smooth", formula = y~x)
```



This graph will show that there is a **linear** regression between these two elements

- Going to the business aspects, let us discuss the graph between visits to add_to_cart values of various sites separately, then you can observe the business characteristics

```
> qplot(test$visits,test$add_to_cart,facets= .~site,data = test)
```



- After checking the summary of the sites which has made the highest visits,

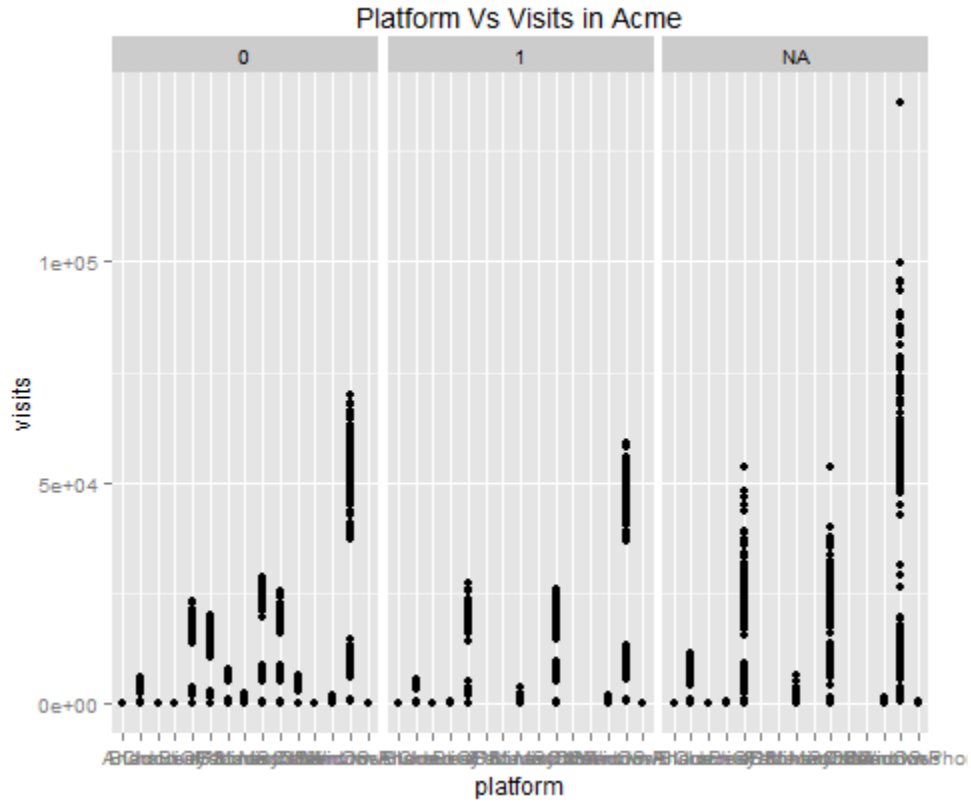
```
> summary(test$site)
  Acme    Botly Pinnacle   Sortly  Tabular Widgetry
  7392      804     5725     5532      804      804
```

From the result of the summary it is clear that Acme had made the highest number of visits. So it can be useful if we analyze the visits done by Acme alone. Hence we subset the data according to the visits done by Acme

```
> dataacme <- test[test$site == "Acme",]
```

- As we did the analysis between platform and visits, we will similarly analyze the 'dataacme ' dataset. In this case, we will divide based on the new_customers. The key says that 0 represents the existing users and 1 represents the existing users

```
> qplot(dataacme$platform,dataacme$visits,main = "Platform Vs
Visits  in Acme",xlab = "platform",ylab = "visits",data=dataacme, facet
s=.~new_customer, na.omit(dataacme$new_customer))
```



This is same as that of original dataset. Most of the existing users use windows platform, while the other customers use iOS and MacOS X equally.

- Similarly the dataset for the second highest used site, Pinnacle is formed. The analysis of this data set also shows that the platform used for this dataset is also same as that of the whole

```
> datapinnacle <- test[test$site == "Pinnacle",]
```

```
> head(datapinnacle)
```

- The correlation between the visits vs. distinct_sessions (visitors) is made to find the distinct views by various visitors and we found that it is close to 1. Now we can perform the same for the dataacme and datapinnacle datasets

```
> cor(datapinnacle$distinct_sessions,datapinnacle$visits)
[1] 0.992045
> cor(dataacme$distinct_sessions,dataacme$visits)
[1] 0.9986482
```

This is also same as that for the original dataset; it means that most of the visits are distinct.

Note: The source code of the base analysis is found in "Appendix-I" and the name of the file will be baseanalysis.R

2. Ratio Analysis

It is necessary to find the ratio functions for each variable as it would give the productive value of the dataset. The various ratios that should be analyzed in this dataset are

1. Conversion rate
2. Bounce rate
3. Add to cart rate

- Conversion rate is the ratio between the orders and visits in the dataset. But it is necessary to exclude the na values from the data set. The conversion rate for the given dataset is as follows

```
> conversion_rate(test)
[1] 0.2201141
```

This shows that 22.01% of the visits have been converted into orders. That is for every 100 visits 22 visits have resulted in productive business. It is not good for the company to have a lesser conversion rate.

Let us compare the conversion rate of dataacme and datapinnacle dataset and analyze the result.

```
> conversion_rate(dataacme)
[1] 0.203613
> conversion_rate(datapinnacle)
[1] 0.101094
```

This shows that the conversion rate of dataacme is similar to that of the original dataset, but the conversion rate of datapinnacle is 10% lesser than that of the original

- Bounce rate is the ratio between the bounces and the visits in the dataset. Bounce specifies that a visit is done for only one page. It is better to have less bounce rate, it means that every visit includes traversal of several pages or at least more than one.

```
> bounce_rate(test)
[1] 0.3396112
```

The bounce rate of dataset is 33%. As the bounce rate is less, it is assumed that most of the visits traverse through more than one page. This is a success to the company. Let us compare the bounce rate of dataacme and datapinnacle datasets.

```
> bounce_rate(dataacme)
[1] 0.3132728
> bounce_rate(datapinnacle)
[1] 0.4786508
```

These results are also similar to that of conversion rate. The bounce rate of dataacme is approximately equivalent to that of the original, while the datapinnacle is 10% less efficient.

- Add to cart rate is the ratio between the add_to_cart field and visits. The field add_to_cart represent the number of visits that added product to the cart. It is good to have a higher add_to_cart rate, which means that most of the visits are productive.

```
> add_to_cart_rate(test)
[1] 0.293506
```

The percentage of add_to_cart ratio is 29%, this means that out of 100 only 29 visits have added products to the cart. The add to cart rate of dataacme gives the same rate as of the original dataset, while the add to cart rate of datapinnacle is 10% lesser than other.

```
> add_to_cart_rate(dataacme)
[1] 0.2813376
> add_to_cart_rate(datapinnacle)
[1] 0.1620457
```

These results shows that the dataacme has the majority of values, therefore the result of dataachme is similar to that of whole dataset.

Note: The source code for the ratio analysis is also found in Appendix I and the name of the file will be 'conversion.R'

3. Analysis App using shiny

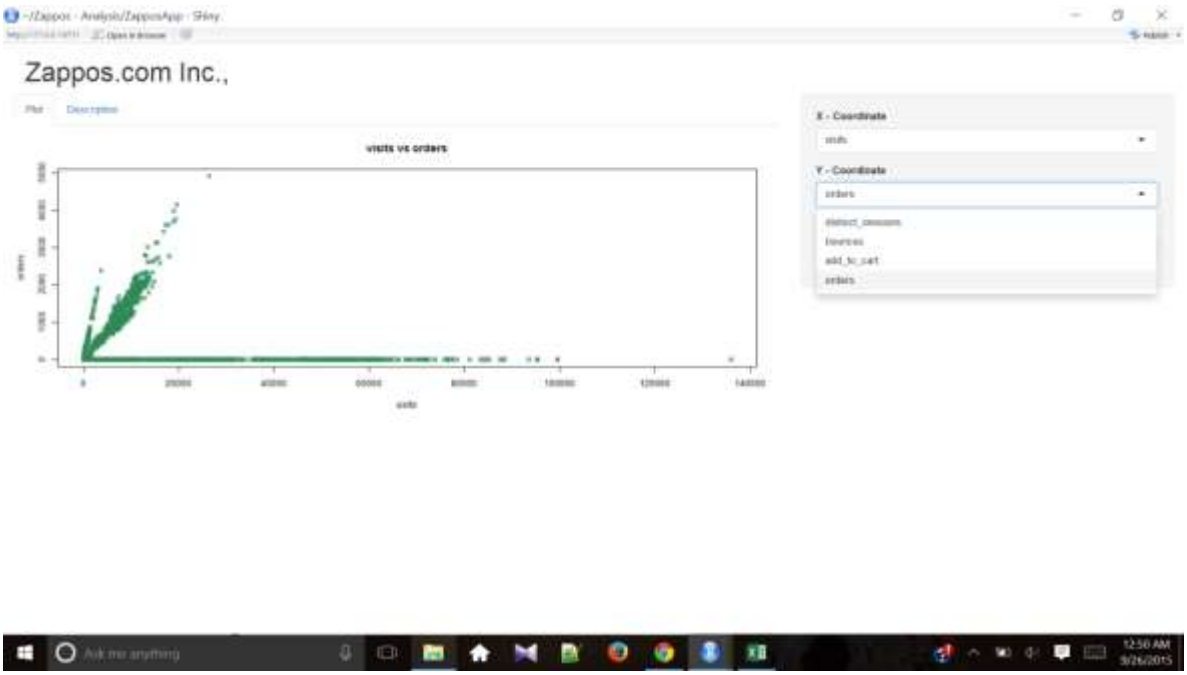
The analysis options and the output can be embedded in the form of an application. There is a package in R called **shiny**, which is used to write HTML code using R functions. There are basically two scripts called the "ui.R" and "server.R" that comprises the application. The ui.R is the interface for the program, it usually contains the construction of input buttons, panels, collections, etc.,

While the server is used for the R logic part of the application. It will contain the models commands to construct the graph, its characteristics, etc.,

Demonstration

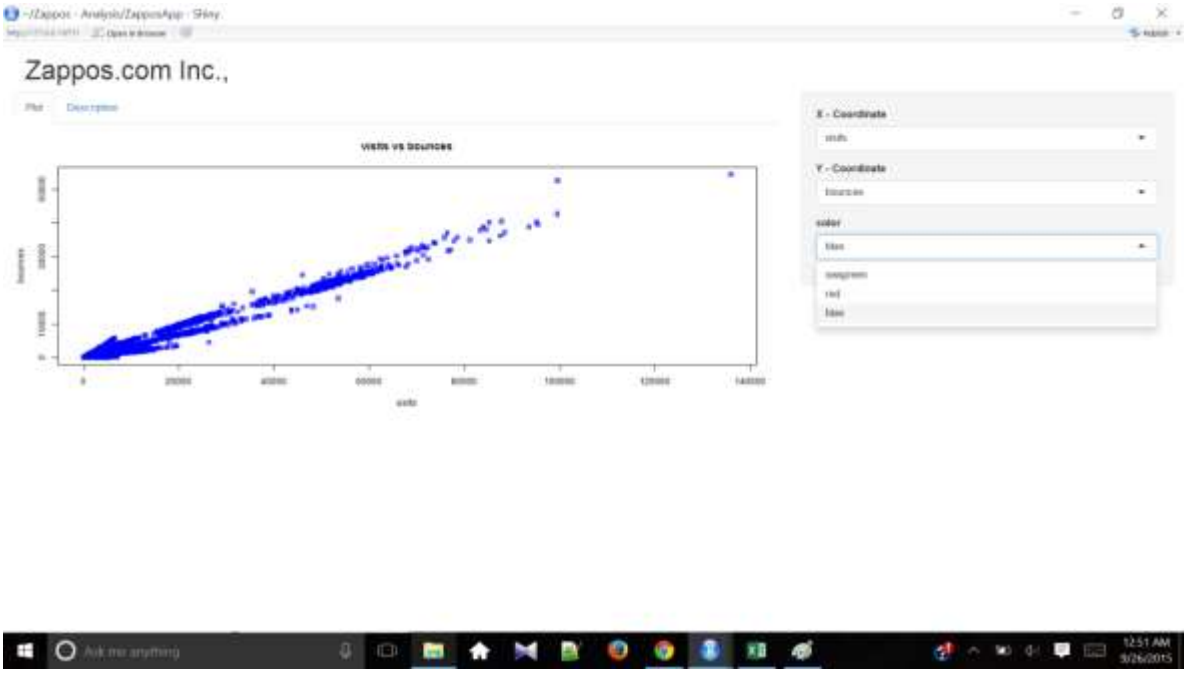
Let us look at the demonstration of the application created for the analysis of various factors against the visits column. The application has an option to select the y-coordinate and color according to the user's preference, the code is constructed in a responsive way that changes the output according to the input selected.

Here is a demonstration of how the application works and how the graph can be analyzed. For better understanding I am including the screenshots of the application.



Here is the look of the application. The side Panel in the right side of the window provides the control tool to change the patterns and properties of the graph according to the requirements. The description tab in the left part, when selected gives the details of the graph in a single line.

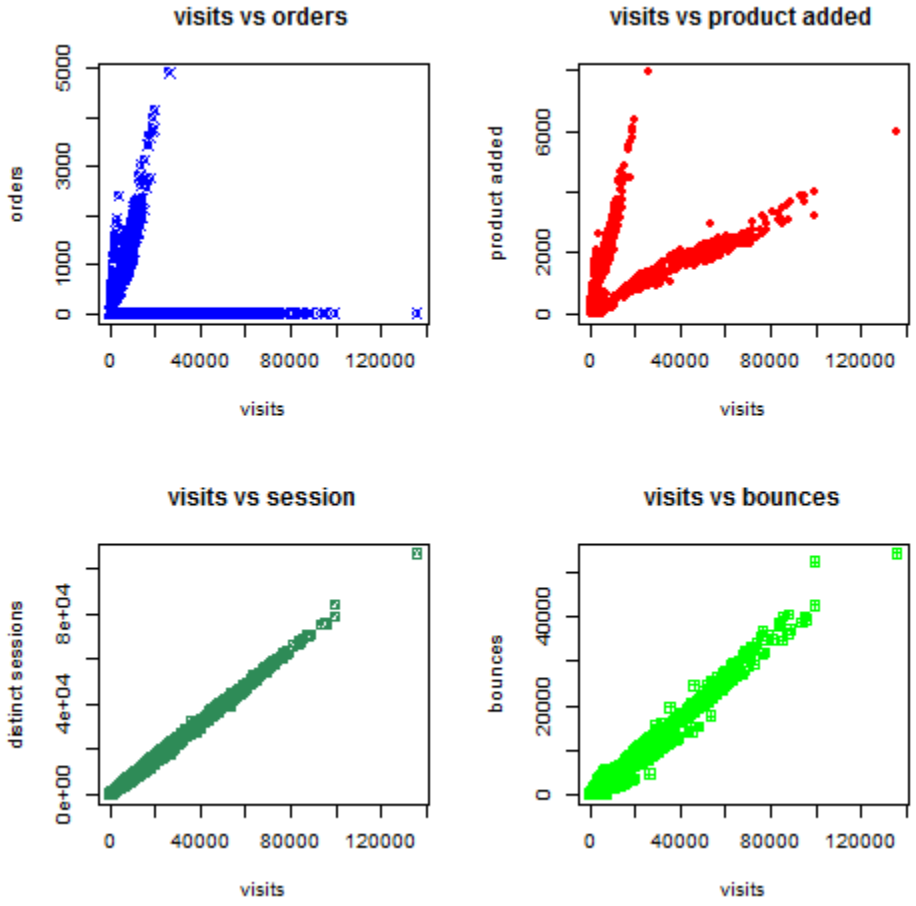
After changing the inputs of the application the output will change immediately. Here is another screenshot of the application exhibiting the responsive feature of the application



Thus the application is used to find the analysis of the given dataset. The parent data set can be inserted into the application folder in the name of file ”data.csv”.

Note: The app can be found in Appendix II and the file has a folder containing two R scripts.

Let us display all the four graphs from the application and then analyze them one by one:



The graph (top-left) between order and visits shows that there are many values of orders nearing zero for various values of the visits. This shows that the many visits do not result in orders. The other graph (top-right) shows the relationship between visits and products added. This is not a linear graph, but we can say that the graph has two branches. The third graph (bottom-left) is a linear one which shows that the visits mostly made are distinct visits. The last graph is constructed between the visits and the bounces. It is not a linear graph. It deviated below the linear regression line, this shows that the ratio will be less than 50% (<50). This graph shows that the percentage of visits that have bounced are less than 50%

Inference :

After performing a series of analysis we can infer few points about the given dataset. They are

- Most of the contribution comes from the site – ‘acme’
- When a visitor visits a site only distinct visits are made, there is less duplication. i.e. there is a linear relationship between the visits and distinct_visits
- The conversion rate of the given data set is not convincing. Only 22% of the visits have been converted to orders.
- Bouncing rate of the data set is 33%, which is a good sign for the website.
- Add to cart ratio of the data set is 30%, this means that more than one quarter of the visits have ended in adding products to cart.
- There is a positive correlation between the distinct_visits and visits.
- Windows OS is the most frequently used platform, MacOSX comes in the second place.
- There are large number of customers who are neither returning customers nor new customers.