**Analyzing the Internet Data***Data Science with R Project*

As the web analytics team wants to understand the web activities of the site, which are the sources used to access the website. Based on the datasets and the given operational definition of the respective variables, following are the deduced inferences for the indicated problem statements of the dataset.

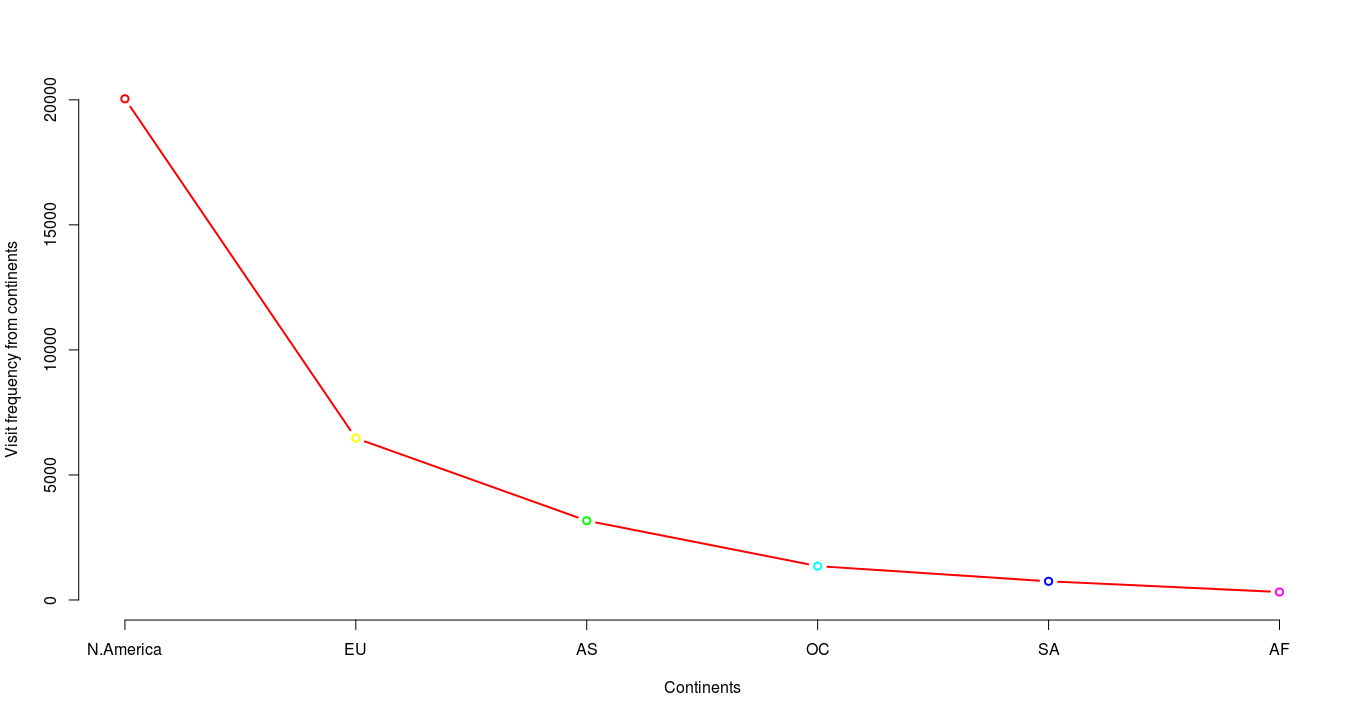
**NOTE: The successive screenshots of the project  
have also been attached for reference purpose.**

**1.**   
In order to understand the descriptive statistics the data is summarized mainly to obtain the min, max, median, mean etc. of the variables in the dataset and the R code is also mentioned below.  
**summary(id)**After the execution of the above command, the following are the insights on all the continuous variables:

* The maximum Bounces are known to be 30.
* The maximum Exits are known to be 36.
* The maximum Time on any give page of the site can be seen as 46745 units.
* The maximum and minimum number of sessions during which that page was viewed one or more times was known to 45 and 1 respectively.
* The maximum number of visits have been recorded to be 45.

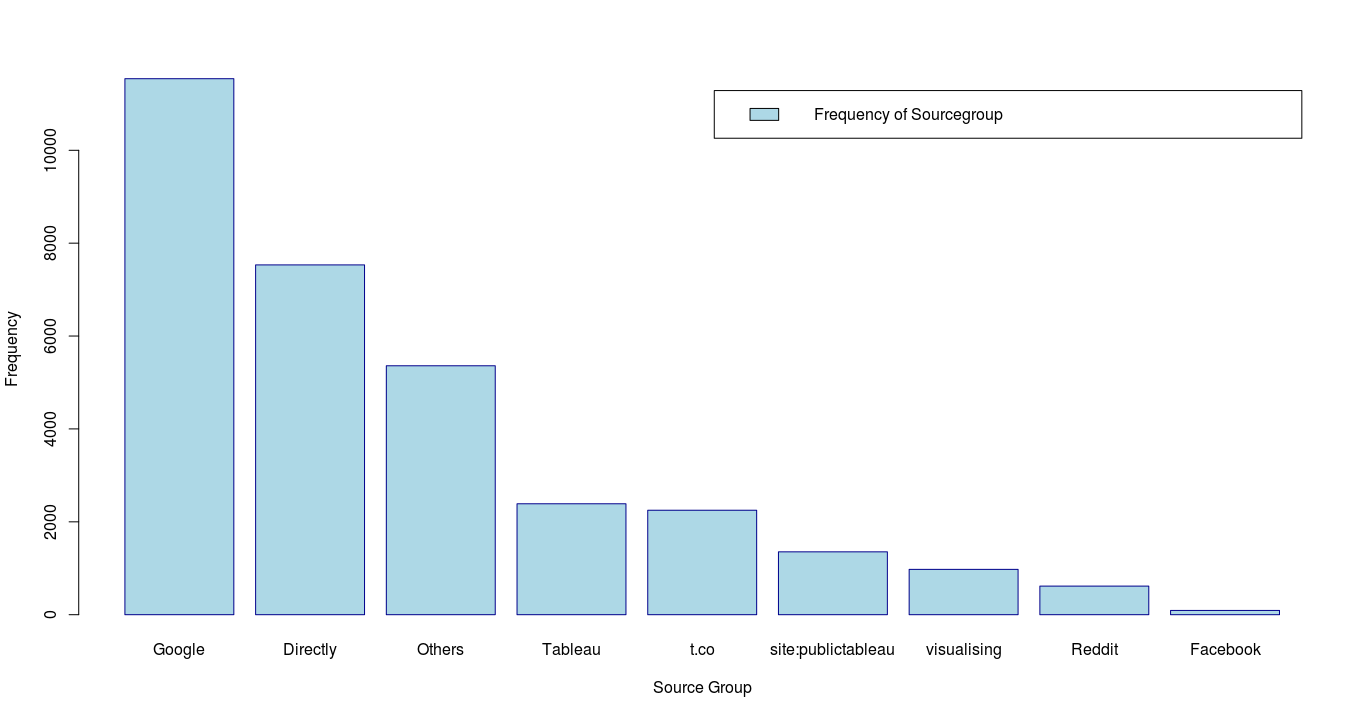
For deducing the descriptive statistics of the categorical variables, the relevant data has been plotted.

The visiting frequency to website from different continents can be seen in the graph below with the used R code.

**contfreq = sort(table(id$Continent), T)  
plot(contfreq, col= rainbow(6), type = "b", xlab = "Continents", ylab = "Visit frequency from continents")**  


The frequency of visits of individual users from various different source groups to the website can be seen in the bar plot given below and the respective R code is also mentioned.

**sg=sort(table(id$Sourcegroup),T)  
lab = c("Google", "Directly", "Others", "Tableau","t.co", "site:publictableau","visualising","Reddit", "Facebook")  
barplot(sg, col = "lightblue", xlab="Source Group", ylab = "Frequency", width = 3 ,legend.text = "Frequency of Sourcegroup",space = .2, border = "darkblue", names.arg = lab)**



**2.**  
As the web analytics wants to know whether there is a relation between the number of sessions during which a specific page was viewed one or more times and the visit counting all instances.  
Here we run the ANOVA function to the discern the relation of the concerned variables, here the null hypothesis is considered as a state where there is no relation between the variables. The correlation function in R is also used to understand the extent to which these variables are related. The R code used is mentioned below.  
**prob2 = aov(Uniquepageviews ~ Visits, data = id)  
prob2  
summary(prob2)  
cor(Uniquepageviews, Visits)**

After executing the code, the insights can be drawn on the basis of p value which is less than 0.05, thus we can reject the null hypothesis to state there is relation between the two variables. The extent to which these variables are correlated can be known by the correlation coefficient which is found to be 0.8144, which affirms that the variables are positively correlated.

**3.**The probable factors from the dataset which have an effect on occurring exits from the website can be found by doing the Exit Page Analysis. Here the ANOVA function is used again, to understand the factors influencing the exit.  
**prob3 = aov(Exits~., data = id)  
summary(prob3)**

It becomes clearly evident after running the code that the influencing variables are Uniquepageviews, Timeinpage, Bounces and Visits. Note that the categorical variables (Sourcegroup , Continent) have not been considered as they have no effect on the model.

**4.**As the website wants to increase the time on page by a visitor, a multiple linear regression model is used to understand the extent of influence by the independent variables.   
**prob4 = lm(Timeinpage~.,data = id)  
summary(prob4)**

After execution of the above mentioned R code, we can reject the null hypothesis of no effect for the continuous variables which have p value less than 0.05. The independent variables in this model, namely Bounces, Exits, Uniquepageviews and Visits have a significant effect on the Timeinpage dependent variable.

**5.**The team is interested in reducing the bounce rate of the website, to tackle this situation we can run the logistic regression model taking Bouncesnew as the dependent variable. Bouncesnew is calculated by the multiply the whole column of Bounces with 0.01. This makes the value of Bouncesnew fluctuate between 0 and 0.3.

**table(Bounces)  
Bouncesnew = Bounces \* 0.01  
table(Bouncesnew)  
prob5 = glm(Bouncesnew~Timeinpage+Visits+Exits+Uniquepageviews+Sourcegroup+Continent, family = "binomial", data = id)  
summary(prob5)**

After the execution of the mentioned R code, we can reject the null hypothesis [which states there is no relation between the variables] it becomes evident that the high bounce rate can be calibrated by keeping a keen eye on the website’s factors, namely Visits, Exits, Uniquepagereviews and Timeinpage.   
***NOTE: The significance level for this model was assumed to be 10%*.**