rm(list=ls())

getwd()

file <- read.csv('web.csv')

View(file)

1) The team wants to analyse each variable of the data collected through data summarization to get a basic understanding of the dataset and to prepare for further analysis.

To do this we must use the summary function. Apply it to our data to get summary statistics for all our variables.

Code:

summary(file)

str(file)

Output:



The output of summary(file) summarizes the data set, it gives values such as the mean, minimum, maximum, median and quartiles for the numeric variables in the data set. For categorical data it gives the frequency of occurrence of each category in the data set. From this output We can see that there is a maximum value of 30 bounces for the website. This site was accessed maximum number of times by visitors from North America. Google is the most used source group, and more.

2) The team needs to know whether the unique page view value depends on visits.

Here we trying to find out if page visits have an effect on unique page view. We used a one-way anova to determine the effect of visits on unique page view. First, we had to change the visits variable to a factor in order to run the anova model.

Ho= There is no significant effect of visits on Uniquepage view.

Ha= there is significant effect of visits on Uniquepage view.

Code:

av <- aov(Uniquepageviews ~ as.factor(Visits), data = file)

summary(av)

Output:

Df Sum Sq Mean Sq F value Pr(>F)

as.factor(Visits) 16 11476 717.3 34703 **<2e-16 \*\*\***

Residuals 32092 663 0.0

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The p value generated by the anova model(**2e-16)** is less than alpha(0.05) so the null hypothesis is rejected, this means that visits do have an effect on the Uniquepageview. Unique page view does depend on visits.

3) Find out the probable factors from the dataset, which could affect the exits.

Here we trying to see which other factors may have a significant effect on exits. We will run an anova model to determine the relationship between exit and all the other variables. We need to first convert all the numerical variables to factors.

H0= No variable has an effect on exit

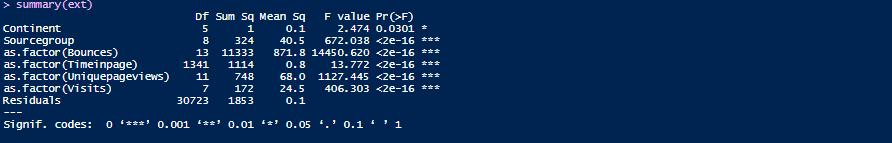
Ha=some variables may have an effect on exit

Code:

ext <- aov(Exits~Continent+Sourcegroup+as.factor(Bounces)+as.factor(Timeinpage)+as.factor(Uniquepageviews)+as.factor(Visits),data = data)

summary(ext)

Output:



From the output we can see that there exist significant relationship between exit and all the other variables expect Continent. This means that exit is not affected by the Continent of the vising person. Source groups, Bounces, Timeinpage, Uniquepagevies and visits all produced p values (2e-16 each). These p values are significant at alpha=0.05, these variables are the one that have a significant effect on exit. All the variables with p value<0.05, have an effect on exit.

4) Find the variables which possibly have an effect on the time on page.

Here we trying to see which of these variables have a significant effect on Timeonpage. We will run an anova model to depict the relationship of the other variables on Timeinpage.

H0=No variable has an effect on Timespent on the site

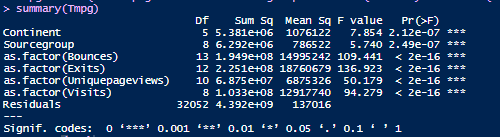
Ha=some variable(s) affect the time a visitor spends on the site.

Code:

Tmpg <- aov(Timeinpage~Continent+Sourcegroup+as.factor(Bounces)+as.factor(Exits)+as.factor(Uniquepageviews)+as.factor(Visits),data = data)

summary(Tmpg)

Output:



From the output we see that all the variables are significant at alpha=0.05. This means that we reject the null hypothesise since all the p values generated by the anova are less than 0.05. all these variables have a significant effect on Timeinpage. All these factors will be important when trying to increase the time visitors spend on the site.

5) Help the team in determining the factors that are impacting the bounce.

Here we will run a logistic regression model to check which factors influence bounce.

H0=No other factor has an impact on bounce

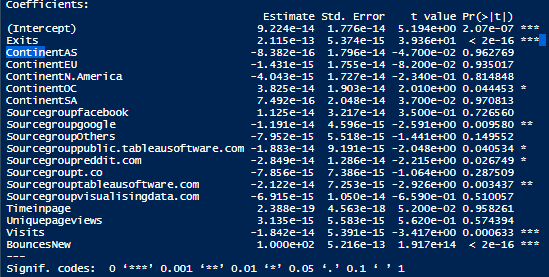
Ha=some factors may have an impact on Bounce

Code:

last <- lm(Bounces~., data = my)

summary(last)

Output:



The output shows that the Bounce is affected mostly by Visits and Timeinpage. These two variables returned p values that were less than 0.05. this means that these variables have an effect on Bounce. So we reject H0 and accept Ha. Bounce is affected by Exits and Visits.