Statistical Analysis Using R

Web Data Analysis

By:- Mrinmay Saha

Business Scenario

DESCRIPTION

Background and Objective:
The web analytics team of www.datadb.com is interested to understand the web activities of the site, which are the sources used to access the website. They have a database that states the keywords of time in the page, source group, bounces, exits, unique page views, and visits.

Domain: Web

Dataset Description:
The variables in the dataset are defined here for better understanding:

Attribute	Description
Bounces	It represents the percentage of visitors who enter the site and "bounce" (leave the site) rather than continuing to view other pages within the same site.
Continent	It shows the continent from which the site has been accessed.
Source group	It shows how the visitor has accessed the site.
Time on page	It shows how long the user has spent on that particular page of the website.
Unique pageview	It represents the number of sessions during which that page was viewed one or more times.
Visits	A visit counts all visitors, no matter how many times the same visitor may have been to your site.

Analysis to be done:

Analysis Tasks:

The team is targeting the following issues:

- The team wants to analyze each variable of the data collected through data summarization to get a basic understanding of the dataset and to prepare for further analysis.
- As mentioned earlier, a unique page view represents the number of sessions during which that page was viewed one or more times. A visit counts all instances, no matter how many times the same visitor may have been to your site. So the team needs to know whether the unique page view value depends on visits.
- Find out the probable factors from the dataset, which could affect the exits. Exit Page Analysis is usually required to get an idea about why a user leaves the website for a session and moves on to another one. Please keep in mind that exits should not be confused with bounces.
- Every site wants to increase the time on page for a visitor. This increases the chances of the visitor understanding the site content better and hence there are more chances of a transaction taking place. Find the variables which possibly have an effect on the time on page.
- A high bounce rate is a cause of alarm for websites which depend on visitor engagement. Help the team in determining the factors that are impacting the bounce.

Code:

```
library(readxl)
web<-read_xlsx('1555058318_internet_dataset.xlsx')
str(web)

#converting the char value to categorical value
web$Continent<-as.factor(web$Continent)
web$Sourcegroup<-as.factor(web$Sourcegroup)

#summary of the data to get a basic understanding of the dataset and to prepare for further analysis.
summary(web)

#checking whether there is a relation between uniquepageviews and Visits
cor(web$Uniquepageviews,web$Visits)
ano<-aov(Uniquepageviews~Visits,data=web)
summary(ano)

#checking the factors thats affect the Exits
anoe<-aov(Exits~.,data=web)
summary(anoe)

#checking the factors that affects the timeinpage on the website
anot<-aov(Timeinpage~.,data=web)
summary(anot)

#checking the factors thats affect the Bounce
#data value should be between 0 to 1 so using BounsNew variable
logb<-glm(BouncesNew~Timeinpage+Continent+Sourcegroup+Uniquepageviews+Visits,data = web,family = "binomial")
summary(logb)
```

Output

library(readxl)

web<-rea d_xlsx('1555058318_internet_dataset.xlsx')

str(web)

```
> str(web)
tibble[,8] [32,109 x 8] (S3: tbl_df/tbl/data.frame)
 $ Bounces        : num [1:32109] 0 0 0 0 0 0 0 0 0 0 ...
 $ Exits          : num [1:32109] 0 0 0 0 0 0 0 0 0 0 ...
 $ Continent      : Factor w/ 6 levels "AF","AS","EU",..: 5 4 4 4 4 4 4 4 5 2 ...
 $ Sourcegroup    : Factor w/ 9 levels "(direct)","facebook",..: 1 1 4 5 5 5 5 1 1 4 ...
 $ Timeinpage     : num [1:32109] 18 4 35 70 81 75 186 710 712 344 ...
 $ Uniquepageviews: num [1:32109] 1 1 1 1 1 1 1 1 1 1 ...
 $ Visits         : num [1:32109] 0 0 0 0 0 0 0 0 1 1 ...
 $ BouncesNew     : num [1:32109] 0 0 0 0 0 0 0 0 0 0 ...
>
```

#summary of the data to get a basic understanding of the dataset and to prepare for further analysis.

summary(web)

```
> summary(web)
    Bounces            Exits            Continent                    Sourcegroup
 Min.   : 0.000   Min.   : 0.000   AF       :  321   google                   :11542
 1st Qu.: 0.000   1st Qu.: 1.000   AS       : 3171   (direct)                 : 7532
 Median : 1.000   Median : 1.000   EU       : 6470   Others                   : 5360
 Mean   : 0.713   Mean   : 0.906   N.America:20043   tableausoftware.com      : 2388
 3rd Qu.: 1.000   3rd Qu.: 1.000   OC       : 1356   t.co                     : 2249
 Max.   :30.000   Max.   :36.000   SA       :  748   public.tableausoftware.com: 1354
                                                     (Other)                  : 1684
   Timeinpage       Uniquepageviews      Visits          BouncesNew
 Min.   :    0.00   Min.   : 1.000   Min.   : 0.000   Min.   :0.00000
 1st Qu.:    0.00   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:0.00000
 Median :    0.00   Median : 1.000   Median : 1.000   Median :0.01000
 Mean   :   73.18   Mean   : 1.114   Mean   : 0.906   Mean   :0.00713
 3rd Qu.:   10.00   3rd Qu.: 1.000   3rd Qu.: 1.000   3rd Qu.:0.01000
 Max.   :46745.00   Max.   :45.000   Max.   :45.000   Max.   :0.30000
```

Insight:- As we can see in the summary the min, max, mean, median, quartile range of the numerical values and the categorical values it shows number of times the value has appeared in the dataset.

As, we can see that the for Bounces min=1 &max=30 and for Exits is min=0 & max=36 respectively for other numerical values.

We can also see that maximum no. of visit was from North America.

#checking whether there is a relation between uniquepageviews and Visits

cor(web$Uniquepageviews,web$Visits)

ano<-aov(Uniquepageviews~Visits,data=web)

summary(ano)

```
> cor(web$Uniquepageviews,web$Visits)
[1] 0.8144457
> ano<-aov(Uniquepageviews~Visits,data=web)
> summary(ano)
              Df Sum Sq Mean Sq F value Pr(>F)
Visits         1   8052    8052   63257 <2e-16 ***
Residuals  32107   4087       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Insight: As we can see that from both the test conducted we can infer from the results t hat the visits variable has a significant impact on Unique.Pageviews. So the team can conclude that unique page values depend on visits.

#checking the factors thats affect the Exits

anoe<-aov(Exits~.,data=web)

summary(anoe)

```
> summary(anoe)
                Df Sum Sq Mean Sq    F value    Pr(>F)
Bounces          1  10578   10578 1.043e+05   < 2e-16 ***
Continent        5      3       1 5.960e+00 1.62e-05 ***
Sourcegroup      8      7       1 8.760e+00 4.89e-12 ***
Timeinpage       1    130     130 1.279e+03   < 2e-16 ***
Uniquepageviews  1   1573    1573 1.552e+04   < 2e-16 ***
Visits           1      1       1 5.014e+00    0.0251 *
Residuals    32091   3254       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Insight: From the result of ANOVA given here, we can see that source.group, bounces, and unique.p ageviews have more significance. Visits have comparatively less significance.

Hence we can say that exit from the site is affected by the factors of source group, bounces, and uniq ue.pageviews.

#checking the factors that affects the timeinpage on the website

anot<-aov(Timeinpage~.,data=web)

summary(anot)

```
> summary(anot)
                Df     Sum Sq    Mean Sq  F value    Pr(>F)
Bounces          1 5.947e+07   59466495  422.868   < 2e-16 ***
Exits            1 1.304e+08  130400662  927.283   < 2e-16 ***
Continent        5 4.767e+06     953431    6.780 2.51e-06 ***
Sourcegroup      8 1.545e+06     193153    1.374     0.202
Uniquepageviews  1 1.791e+08  179133934 1273.826   < 2e-16 ***
Visits           1 1.073e+08  107321113  763.163   < 2e-16 ***
Residuals    32091 4.513e+09     140627
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Insight: From the result of ANOVA given here, we can see that bounces,Exits,Visits and unique.pag eviews have more significance. source.group have comparatively less significance.

Hence we can say that Timeinpage from the site is affected by the factors of bounces,Exits,Visits and unique.pageviews.

#checking the factors thats affect the Bounce

#data value should be between 0 to 1 so using BounsNew variable

logb<-glm(BouncesNew~Timeinpage+Continent+Sourcegroup+Uniquepageviews+Visits,data = web,family = "binomial")

summary(logb)

```
Call:
glm(formula = BouncesNew ~ Timeinpage + Continent + Sourcegroup +
    Uniquepageviews + Visits, family = "binomial", data = web)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.86800  -0.03579   0.00337   0.01256   1.79722

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -4.8466054  0.6755127  -7.175 7.25e-13 ***
Timeinpage                            -0.0017540  0.0006931  -2.530   0.0114 *
ContinentAS                            0.0142202  0.6930422   0.021   0.9836
ContinentEU                           -0.0031592  0.6784051  -0.005   0.9963
ContinentN.America                     0.0190728  0.6671541   0.029   0.9772
ContinentOC                            0.0360241  0.7331752   0.049   0.9608
ContinentSA                            0.0302026  0.7911968   0.038   0.9695
Sourcegroupfacebook                   -0.0076519  1.1042651  -0.007   0.9945
Sourcegroupgoogle                     -0.0980354  0.1703840  -0.575   0.5650
SourcegroupOthers                     -0.1484919  0.2168259  -0.685   0.4934
Sourcegrouppublic.tableausoftware.com -0.4370338  0.4911158  -0.890   0.3735
Sourcegroupreddit.com                 -0.0512534  0.4702490  -0.109   0.9132
Sourcegroupt.co                       -0.0176847  0.2753879  -0.064   0.9488
Sourcegrouptableausoftware.com        -0.2436464  0.3175223  -0.767   0.4429
Sourcegroupvisualisingdata.com        -0.2024044  0.4602673  -0.440   0.6601
Uniquepageviews                       -2.3701652  0.5201423  -4.557 5.19e-06 ***
Visits                                 2.5907917  0.5169444   5.012 5.39e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.94  on 32108  degrees of freedom
Residual deviance: 124.70  on 32092  degrees of freedom
AIC: 499.84

Number of Fisher Scoring iterations: 10
```

Insight: As can be inferred from the result shown, the Unique.Pageviews and visits are the variables that imp act the target variable bounces it has greater significance.