Analyze the internet data of www.datadb.com

Business Analytic Foundation with R Tools- Question

Abstract

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 page, source group, bounces, exits, unique page views, and visits.

Presented by: Ankita Agarwal

Problem Statement:

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 page, source group, bounces, exits, unique page views, and visits.

Detailed description of the given dataset:

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.

Exits: It represents the percentage of visitors to a site who actively click away to a different site from a specific page, after possibly having visited any other page on the 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 page view: 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.

To Analyze:

1. 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.

2. 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.

3. 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.

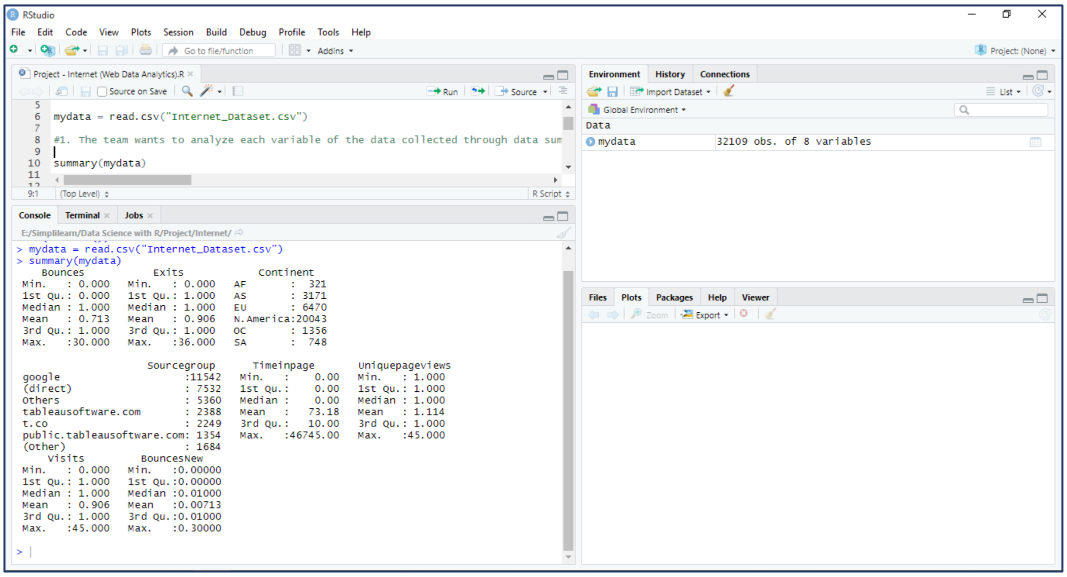
4. Every site wants to increase the time on page for a visitor. This increases the chances of the visitor understanding the site content better and hence there are more chances of a transaction taking place. Find the variables which possibly have an effect on the time on page.

5. 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.

Analysis and Interpretations:

1. 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.

To get a basic insight into the data, we do a 5-point summary analysis for the data set.



```
mydata = read.csv("Internet_Dataset.csv")

#1. The team wants to analyze each variable of the data collected through data sum

summary(mydata)
```

```
mydata          32109 obs. of 8 variables
```

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

> 
```

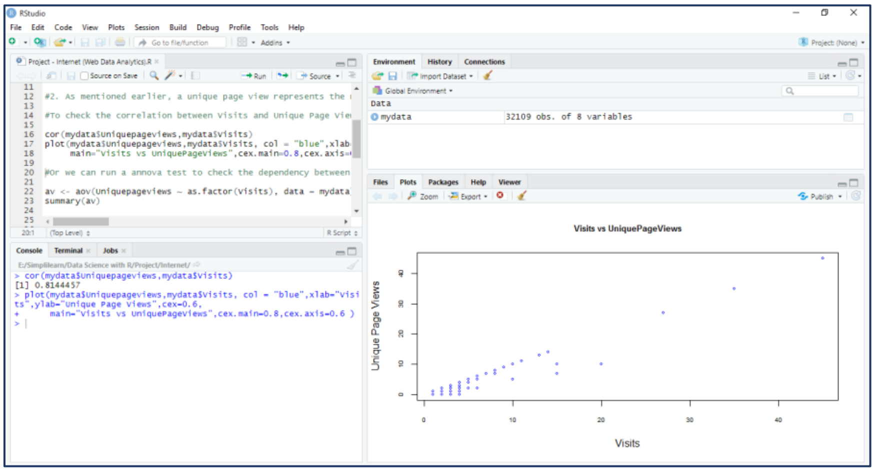
Interpretation:

By looking at the 5-point summary of the data, we can summarize that percentage of visitors that bounce or exit through the website ranges from 0 to 30 and 36 respectively, however, maximum visitors only do it once as shown by the median of the variables. Max number of visitors belong to North America – 20043 or accessed it using google – 11542. The average time spent on the page comes to be 73.2 seconds.

2. 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.

In order to check the dependencies of the variables to each other (unique page view depends on visits) we can either create check the correlation between the variables or do an ANOVA test to understand the dependency between the variables.

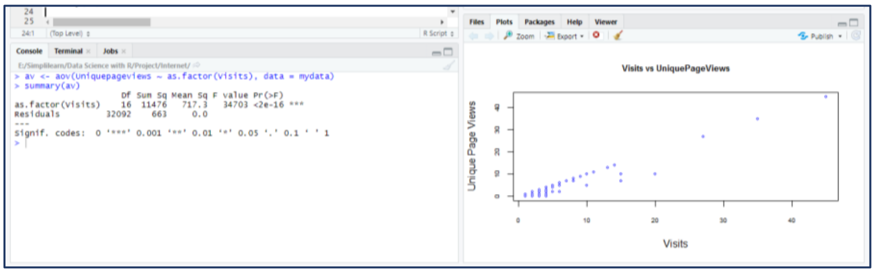
Defining Hypothesis:
Ho: The unique page view value does not depend on visits
Ha: The unique page view value depends on visits


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project - Internet (Web Data Analytics).R
Source on Save
Run
Source
Environment History Connections
Import Dataset
Global Environment
Files Plots Packages Help Viewer
Zoom Export
Publish
20:1 (Top Level) R Script
Console Terminal Jobs
Visits vs UniquePageViews
Unique Page Views
Visits
0
10
20
30
40

```
#2. As mentioned earlier, a unique page view represents the

#To check the correlation between Visits and Unique Page Vie

cor(mydata$Uniquepageviews,mydata$Visits)
plot(mydata$Uniquepageviews,mydata$Visits, col = "blue",xlab
     main="Visits vs UniquePageViews",cex.main=0.8,cex.axis=

#Or we can run a annova test to check the dependency between

av <- aov(Uniquepageviews ~ as.factor(Visits), data = mydata
summary(av)
```

Data
mydata 32109 obs. of 8 variables

```
E:/Simplilearn/Data Science with R/Project/Internet/
> cor(mydata$Uniquepageviews,mydata$Visits)
[1] 0.8144457
> plot(mydata$Uniquepageviews,mydata$Visits, col = "blue",xlab="Visi
ts",ylab="Unique Page Views",cex=0.6,
+      main="Visits vs UniquePageViews",cex.main=0.8,cex.axis=0.6 )
> |
```


24:1 (Top Level) R Script
Console Terminal Jobs
Files Plots Packages Help Viewer
Zoom Export
Publish
Visits vs UniquePageViews
Unique Page Views
Visits
0
10
20
30
40

```
E:/Simplilearn/Data Science with R/Project/Internet/
> av <- aov(Uniquepageviews ~ as.factor(Visits), data = mydata)
> summary(av)
                     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
> |
```

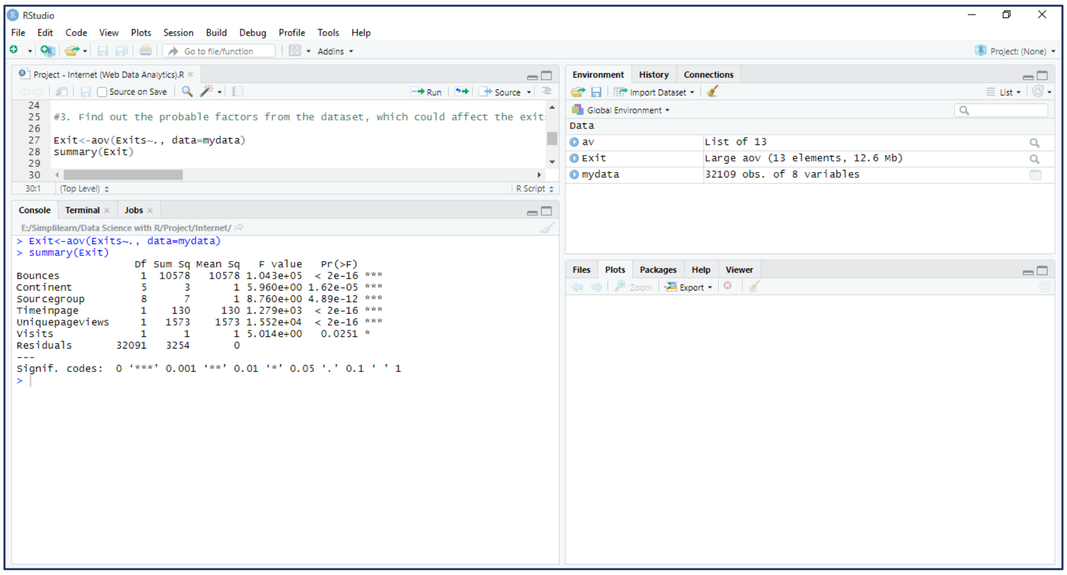
Interpretation:

As seen in the above plot, there is a positive linear correlation between the two variables Unique page views and visits.

Also, we know, when p-value < alpha, p-value is less than alpha; we reject the null hypothesis. We take alpha value as 0.05 at 95% confidence level. Here, p-value significantly less than the speculated 0.05, concluding that the variable visits have a very high impact on Unique page views.

3. 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.

To check the dependencies of other variables on Exits we do an ANOVA test to understand the dependencies between the variables.



```
#3. Find out the probable factors from the dataset, which could affect the exit:

Exit<-aov(Exits~., data=mydata)
summary(Exit)
```

```
> Exit<-aov(Exits~., data=mydata)
> summary(Exit)
                   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
> 
```

Data	
av	List of 13
Exit	Large aov (13 elements, 12.6 Mb)
mydata	32109 obs. of 8 variables

Interpretation:

As we know, when p-value is less than 0.05, which we assume while creating regression model, we consider that variable to be significant in the model.

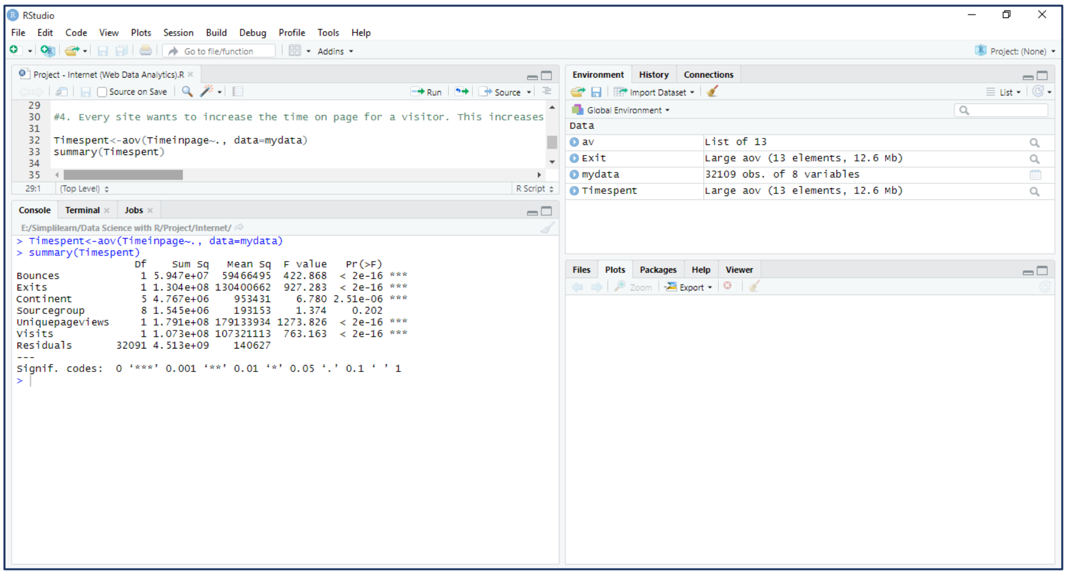
Here, we can see that Bounces, Unique page views and Time in page have significantly low p-values, implying that these variables are extremely significant to exit. Although Continent and Source group have low p-values, they are less significant as compared to bounces, time in page and unique page views.

Also, number of visits do not have as much significance as other variables.

.

4. 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.

To check the dependencies of other variables on Time spent in page we do an ANOVA test to understand the dependencies between the variables.



```
#4. Every site wants to increase the time on page for a visitor. This increases

Timespent<-aov(Timeinpage~., data=mydata)
summary(Timespent)
```

E:/Simplilearn/Data Science with R/Project/Internet/

```
> Timespent<-aov(Timeinpage~., data=mydata)
> summary(Timespent)
                   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
> 
```

Data

av	List of 13
Exit	Large aov (13 elements, 12.6 Mb)
mydata	32109 obs. of 8 variables
Timespent	Large aov (13 elements, 12.6 Mb)

Interpretation:

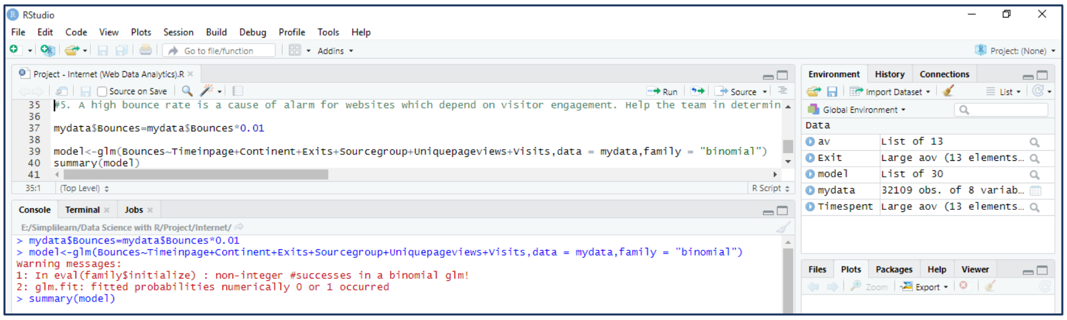
As we know, when p-value is less than 0.05, which we assume while creating regression model, we consider that variable to be significant in the model.

Here, we can see that Bounces, Exits, Unique page views and Visits have significantly low p-values, implying that these variables are extremely significant to Time in page. Although Continent has a low p-value, it is less significant.

As noticed only source group with high p-value does not have much significance to Time spent in the page.

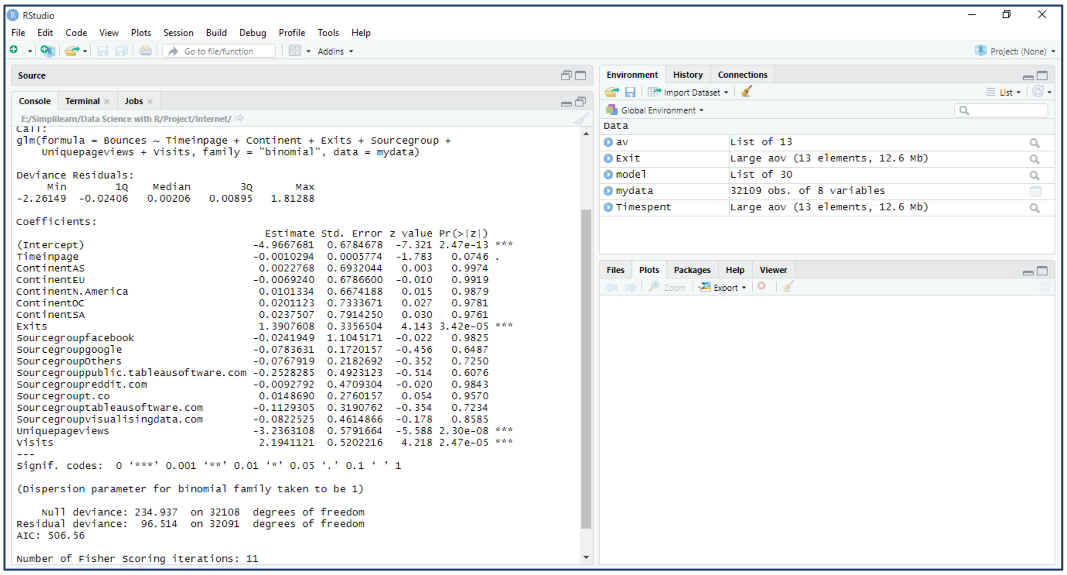
5. A high bounce rate is a cause of alarm for websites which depend on visitor engagement. Help the team in determining the factors that are impacting the bounce.

To understand factors impacting the bounce we create a logistic regression model. The data variables for bounce should be between 0 and 1 as logistic regression shows probability which is always between 0-1. Hence, we multiply bounce with 0.01 to create a regression model.



```
#5. A high bounce rate is a cause of alarm for websites which depend on visitor engagement. Help the team in determin

mydata$Bounces=mydata$Bounces*0.01

model<-glm(Bounces~Timeinpage+Continent+Exits+Sourcegroup+Uniquepageviews+Visits,data = mydata,family = "binomial")
summary(model)
```

```
> mydata$Bounces=mydata$Bounces*0.01
> model<-glm(Bounces~Timeinpage+Continent+Exits+Sourcegroup+Uniquepageviews+Visits,data = mydata,family = "binomial")
Warning messages:
1: In eval(family$initialize) : non-integer #successes in a binomial glm!
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(model)
```



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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.26149  -0.02406   0.00206   0.00895   1.81288

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -4.9667681  0.6784678  -7.321 2.47e-13 ***
Timeinpage                      -0.0010294  0.0005774  -1.783   0.0746 .
ContinentAS                      0.0022768  0.6932044   0.003   0.9974
ContinentEU                     -0.0069240  0.6786600  -0.010   0.9919
ContinentN.America               0.0101334  0.6674188   0.015   0.9879
ContinentOC                      0.0201123  0.7333671   0.027   0.9781
ContinentSA                      0.0237507  0.7914250   0.030   0.9761
Exits                            1.3907608  0.3356504   4.143 3.42e-05 ***
Sourcegroupfacebook             -0.0241949  1.1045171  -0.022   0.9825
Sourcegroupgoogle               -0.0783631  0.1720157  -0.456   0.6487
SourcegroupOthers               -0.0767919  0.2182692  -0.352   0.7250
Sourcegrouppublic.tableausoftware.com -0.2528285  0.4923123  -0.514   0.6076
Sourcegroupreddit.com           -0.0092792  0.4709304  -0.020   0.9843
Sourcegroupt.co                  0.0148690  0.2760157   0.054   0.9570
Sourcegrouptableausoftware.com  -0.1129305  0.3190762  -0.354   0.7234
Sourcegroupvisualisingdata.com  -0.0822525  0.4614866  -0.178   0.8585
Uniquepageviews                 -3.2363108  0.5791664  -5.588 2.30e-08 ***
Visits                           2.1941121  0.5202216   4.218 2.47e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.937  on 32108  degrees of freedom
Residual deviance:  96.514  on 32091  degrees of freedom
AIC: 506.56

Number of Fisher Scoring iterations: 11
```

```
Data
av          List of 13
Exit        Large aov (13 elements, 12.6 Mb)
model       List of 30
mydata      32109 obs. of 8 variables
Timespent   Large aov (13 elements, 12.6 Mb)
```

Interpretation:

As we know, when p-value is less than 0.05, which we assume while creating regression model, we consider that variable to be significant in the model.

Here, we can see Exists, Unique page views and Visits have lesser p-value, implying that these variables are extremely significant to bounce.

Programming Codes:

```
#Reading Internet Data

mydata = read.csv("Internet_Dataset.csv")

#1.	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.

summary(mydata)

#2.	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.

#To check the correlation between Visits and Unique Page Views

cor(mydata$Uniquepageviews,mydata$Visits)
plot(mydata$Uniquepageviews,mydata$Visits, col = "blue",xlab="Visits",ylab="Unique Page Views",cex=0.6,
    main="Visits vs UniquePageViews",cex.main=0.8,cex.axis=0.6 )

#Or we can run a annova test to check the dependency between Visits and Unique Page Views

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

#3.	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.

Exit<-aov(Exits~., data=mydata)
summary(Exit)

#4.	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.

Timespent<-aov(Timeinpage~., data=mydata)
summary(Timespent)
```

#5. 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.

```
mydata$Bounces=mydata$Bounces*0.01

model<-glm(Bounces~Timeinpage+Continent+Exits+Sourcegroup+Uniquepageviews+Visits,data =
mydata,family = "binomial")
summary(model)
```

---The End---