# Web Data Analysis

# By

# Simeon Ifalore

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 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.

**Solution**

## 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.

```
setwd("C:/Users/Simeon/Desktop")

internet <- read_excel("internet.xlsx")

View(internet)

summary(internet)
```

```
> summary(internet)
    Bounces           Exits          Continent          Sourcegroup
 Min.   : 0.000   Min.   : 0.000   Length:32109       Length:32109
 1st Qu.: 0.000   1st Qu.: 1.000   Class :character   Class :character
 Median : 1.000   Median : 1.000   Mode  :character   Mode  :character
 Mean   : 0.713   Mean   : 0.906
 3rd Qu.: 1.000   3rd Qu.: 1.000
 Max.   :30.000   Max.   :36.000
   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
```

From the result of summarized dataset, it is observed that the numerical data includes information related to the maximum, minimum, and mean data. The categorical data like continent includes the data of the number of times the category has been repeated in the dataset. We can see that there is a maximum value of 30 bounces and 36 Exits for the website.

## 2. As mentioned earlier, a unique page view represents the number of sessions during which that page was viewed one or more times.

```
cor(internet$Uniquepageviews,internet$Visits)
```

```
> cor(internet$Uniquepageviews,internet$visits)
[1] 0.8144457
>
```

From the above code we can see that the correlation coefficient between Unique page views and Visits is a fairly strong positive relationship at **0.8144457**

## 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 find out if the Unique page view depends on visit, we can use the the linear model or ANOVA

```
UPV_Value <-aov(Uniquepageviews~Visits, data=internet)

summary(UPV_Value)
```

```
> UPV_value <-aov(Uniquepageviews~Visits, data=internet)
> summary(UPV_value)
              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
>
```

## OR

Using the linear model:

```
UPV <- lm(formula = internet$Uniquepageviews~internet$Visits, data = internet)

summary(UPV)
```

```
> UPV <- lm(formula = internet$Uniquepageviews~internet$visits, data = internet)
> summary(UPV)

Call:
lm(formula = internet$Uniquepageviews ~ internet$Visits, data = internet)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1788 -0.1788 -0.1788  0.1353 13.6396

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.492837   0.003173   155.3   <2e-16 ***
internet$visits 0.685945   0.002727   251.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3568 on 32107 degrees of freedom
Multiple R-squared:  0.6633,    Adjusted R-squared:  0.6633
F-statistic: 6.326e+04 on 1 and 32107 DF,  p-value: < 2.2e-16
```

From the analysis above we can see that the P value is **p-value: < 2.2e-16** which is less than 0.5 and we can infer from the results that the visits variable has a significant impact on Unique Page views therefore the team can conclude that unique page view values depend 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.

To determine this, we can use ANOVA or linear model as used previously.

```
Exit_Model <-aov(Exits ~.,data = internet)

summary(Exit_Model)
```

```
> Exit_Model <-aov(Exits~.,data = internet)
> summary(Exit_Model)
                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
>
```

From the result of ANOVA given here:

we can see that Bounces, Source group and Unique Page views have more significance. we can therefore say that exit from the site is affected by the above-mentioned factors. Although, Visits also affect the Exit Model but it has comparatively less significance.

### 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.

Using ANOVA to see which variable affects the Time in Page the most

```
Time_Model<-aov(Timeinpage~.,data = internet)
```

summary(Time_Model)

```
> Time_Model<-aov(Timeinpage~.,data = internet)
> summary(Time_Model)
                  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
```

We can see from the above that only the source group is not affecting the

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.**

BounceModel <- lm(Bounces~., data = internet)

summary(BounceModel)

```
Call:
lm(formula = Bounces ~ ., data = internet)

Residuals:
       Min         1Q     Median         3Q        Max
-2.635e-11 -1.000e-15  1.000e-15  3.000e-15  4.813e-11

Coefficients:
                                    Estimate Std. Error    t value Pr(>|t|)
(Intercept)                        9.224e-14  1.776e-14  5.194e+00 2.07e-07 ***
Exits                              2.115e-13  5.374e-15  3.936e+01  < 2e-16 ***
ContinentAS                       -8.382e-16  1.796e-14 -4.700e-02 0.962769
ContinentEU                       -1.431e-15  1.755e-14 -8.200e-02 0.935017
ContinentN.America                -4.043e-15  1.727e-14 -2.340e-01 0.814848
ContinentOC                        3.825e-14  1.903e-14  2.010e+00 0.044453 *
ContinentSA                        7.492e-16  2.048e-14  3.700e-02 0.970813
Sourcegroupfacebook                1.125e-14  3.217e-14  3.500e-01 0.726560
Sourcegroupgoogle                 -1.191e-14  4.596e-15 -2.591e+00 0.009580 **
SourcegroupOthers                 -7.952e-15  5.518e-15 -1.441e+00 0.149552
Sourcegrouppublic.tableausoftware.com -1.883e-14  9.191e-15 -2.048e+00 0.040534 *
Sourcegroupreddit.com             -2.849e-14  1.286e-14 -2.215e+00 0.026749 *
Sourcegroupt.co                   -7.856e-15  7.386e-15 -1.064e+00 0.287509
Sourcegrouptableausoftware.com    -2.122e-14  7.253e-15 -2.926e+00 0.003437 **
Sourcegroupvisualisingdata.com    -6.915e-15  1.050e-14 -6.590e-01 0.510057
Timeinpage                         2.388e-19  4.563e-18  5.200e-02 0.958261
Uniquepageviews                    3.135e-15  5.583e-15  5.620e-01 0.574394
Visits                            -1.842e-14  5.391e-15 -3.417e+00 0.000633 ***
BouncesNew                         1.000e+02  5.216e-13  1.917e+14  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.065e-13 on 32090 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 9.521e+27 on 18 and 32090 DF,  p-value: < 2.2e-16
```

From the linear model if we assume that the data set is normally distributed we can see that the Bounces are affected by BouncesNew, Visits and Exits.