

Web Data Analysis

Andrei Enescu

10/15/2020

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:

Analysis Tasks:

The team is targeting the following issues:

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.

Used libraries:

```
library(rio)
```

Importing the data using RIO library

```
webDF <- import("1555058318_internet_dataset.xlsx")
```

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.

```
View(webDF)
str(webDF)
```

```
## 'data.frame':  32109 obs. of  8 variables:
## $ Bounces      : num  0 0 0 0 0 0 0 0 0 0 ...
## $ Exits        : num  0 0 0 0 0 0 0 0 0 0 ...
## $ Continent    : chr  "OC" "N.America" "N.America" "N.America" ...
## $ Sourcegroup  : chr  "(direct)" "(direct)" "Others" "public.tableausoftware.com" ...
## $ Timeinpage   : num  18 4 35 70 81 75 186 710 712 344 ...
## $ Uniquepageviews: num  1 1 1 1 1 1 1 1 1 1 ...
## $ Visits       : num  0 0 0 0 0 0 0 0 1 1 ...
## $ BouncesNew   : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
webDF$Continent <- as.factor(webDF$Continent)
webDF$Sourcegroup <- as.factor(webDF$Sourcegroup)
summary(webDF)
```

```
##      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
##
```

Observation

Based on the summary output we can conclude the following:

- * Maximum Bounce is 30
- * Maximum Exits is 36
- * The website was visited mostly by visitors from *North America*
- * Majority of the visits come from *Google*
- * The average time spent on page was 73.18 seconds

Summary grouped by Continent

```
table(webDF$Continent)
```

```
##
##      AF      AS      EU N.America      OC      SA
##      321     3171     6470     20043     1356     748
```

```
by(webDF, webDF$Continent, summary)
```

```
## webDF$Continent: AF
##      Bounces      Exits      Continent      Sourcegroup
## Min.   :0.00000   Min.   :0.00000   AF      :321   google      :166
## 1st Qu.:0.00000   1st Qu.:1.00000   AS      :  0   (direct)    : 59
## Median :1.00000   Median :1.00000   EU      :  0   Others      : 57
## Mean   :0.7196   Mean    :0.8847   N.America:  0   tableausoftware.com: 16
## 3rd Qu.:1.00000   3rd Qu.:1.00000   OC      :  0   visualisingdata.com: 10
## Max.   :3.00000   Max.    :3.00000   SA      :  0   t.co       :  8
##                                     (Other)    :  5
##      Timeinpage      Uniquepageviews      Visits      BouncesNew
## Min.   : 0.00   Min.   :1.000   Min.   :0.00000   Min.   :0.000000
## 1st Qu.: 0.00   1st Qu.:1.000   1st Qu.:1.00000   1st Qu.:0.000000
## Median : 0.00   Median :1.000   Median :1.00000   Median :0.010000
## Mean   : 55.23   Mean    :1.053   Mean    :0.8847   Mean    :0.007196
## 3rd Qu.: 0.00   3rd Qu.:1.000   3rd Qu.:1.00000   3rd Qu.:0.010000
## Max.   :2613.00   Max.    :3.000   Max.    :3.00000   Max.    :0.030000
##
## -----
## webDF$Continent: AS
##      Bounces      Exits      Continent      Sourcegroup
## Min.   :0.000   Min.   :0.00000   AF      :  0   google      :1603
## 1st Qu.:0.000   1st Qu.:1.00000   AS      :3171   (direct)    : 612
## Median :1.000   Median :1.00000   EU      :  0   Others      : 506
## Mean   :0.731   Mean    :0.9085   N.America:  0   tableausoftware.com: 196
## 3rd Qu.:1.000   3rd Qu.:1.00000   OC      :  0   visualisingdata.com: 137
```

```

## Max. :5.000 Max. :5.0000 SA : 0 t.co : 51
## (Other) : 66
## Timeinpage Uniquepageviews Visits BouncesNew
## Min. : 0.00 Min. :1.000 Min. :0.0000 Min. :0.00000
## 1st Qu.: 0.00 1st Qu.:1.000 1st Qu.:1.0000 1st Qu.:0.00000
## Median : 0.00 Median :1.000 Median :1.0000 Median :0.01000
## Mean : 60.85 Mean :1.071 Mean :0.9085 Mean :0.00731
## 3rd Qu.: 2.00 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:0.01000
## Max. :5197.00 Max. :8.000 Max. :7.0000 Max. :0.05000
##
## -----
## webDF$Continent: EU
## Bounces Exits Continent Sourcegroup
## Min. :0.0000 Min. : 0.0000 AF : 0 google :1926
## 1st Qu.:0.0000 1st Qu.: 1.0000 AS : 0 Others :1489
## Median :1.0000 Median : 1.0000 EU :6470 (direct) :1432
## Mean :0.7134 Mean : 0.8981 N.America: 0 t.co : 630
## 3rd Qu.:1.0000 3rd Qu.: 1.0000 OC : 0 tableausoftware.com: 390
## Max. :9.0000 Max. :10.0000 SA : 0 visualisingdata.com: 342
## (Other) : 261
## Timeinpage Uniquepageviews Visits BouncesNew
## Min. : 0.00 Min. : 1.000 Min. : 0.0000 Min. :0.000000
## 1st Qu.: 0.00 1st Qu.: 1.000 1st Qu.: 1.0000 1st Qu.:0.000000
## Median : 0.00 Median : 1.000 Median : 1.0000 Median :0.010000
## Mean : 60.36 Mean : 1.075 Mean : 0.8981 Mean :0.007134
## 3rd Qu.: 4.00 3rd Qu.: 1.000 3rd Qu.: 1.0000 3rd Qu.:0.010000
## Max. :14392.00 Max. :11.000 Max. :11.0000 Max. :0.090000
##
## -----
## webDF$Continent: N.America
## Bounces Exits Continent
## Min. : 0.0000 Min. : 0.0000 AF : 0
## 1st Qu.: 0.0000 1st Qu.: 1.0000 AS : 0
## Median : 1.0000 Median : 1.0000 EU : 0
## Mean : 0.7079 Mean : 0.9089 N.America:20043
## 3rd Qu.: 1.0000 3rd Qu.: 1.0000 OC : 0
## Max. :30.0000 Max. :36.0000 SA : 0
##
## Sourcegroup Timeinpage Uniquepageviews
## google :7010 Min. : 0.00 Min. : 1.000
## (direct) :4945 1st Qu.: 0.00 1st Qu.: 1.000
## Others :2917 Median : 0.00 Median : 1.000
## tableausoftware.com :1652 Mean : 77.56 Mean : 1.138
## t.co :1448 3rd Qu.: 14.00 3rd Qu.: 1.000
## public.tableausoftware.com:1134 Max. :14108.00 Max. :45.000
## (Other) : 937
## Visits BouncesNew
## Min. : 0.0000 Min. :0.000000
## 1st Qu.: 1.0000 1st Qu.:0.000000
## Median : 1.0000 Median :0.010000
## Mean : 0.9089 Mean :0.007079
## 3rd Qu.: 1.0000 3rd Qu.:0.010000
## Max. :45.0000 Max. :0.300000
##

```

```

## -----
## webDF$Continent: OC
##      Bounces      Exits      Continent
##  Min.   :0.0000   Min.   :0.0000   AF      :    0
##  1st Qu.:0.0000   1st Qu.:1.0000   AS      :    0
##  Median :1.0000   Median :1.0000   EU      :    0
##  Mean   :0.7404   Mean   :0.9027   N.America:    0
##  3rd Qu.:1.0000   3rd Qu.:1.0000   OC      :1356
##  Max.   :5.0000   Max.   :6.0000   SA      :    0
##
##      Sourcegroup      Timeinpage      Uniquepageviews
##  google              :611   Min.   :  0.00   Min.   :1.000
##  (direct)            :243   1st Qu.:  0.00   1st Qu.:1.000
##  Others              :227   Median :  0.00   Median :1.000
##  tableausoftware.com :103   Mean   : 66.23   Mean   :1.071
##  t.co                : 63   3rd Qu.:  1.00   3rd Qu.:1.000
##  public.tableausoftware.com: 52   Max.   :3246.00   Max.   :6.000
##  (Other)              : 57
##      Visits      BouncesNew
##  Min.   :0.0000   Min.   :0.000000
##  1st Qu.:1.0000   1st Qu.:0.000000
##  Median :1.0000   Median :0.010000
##  Mean   :0.9027   Mean   :0.007404
##  3rd Qu.:1.0000   3rd Qu.:0.010000
##  Max.   :6.0000   Max.   :0.050000
##
## -----
## webDF$Continent: SA
##      Bounces      Exits      Continent      Sourcegroup
##  Min.   :0.0000   Min.   :0.0000   AF      :    0   (direct)      :241
##  1st Qu.:0.0000   1st Qu.:1.0000   AS      :    0   google       :226
##  Median :1.0000   Median :1.0000   EU      :    0   Others       :164
##  Mean   :0.7179   Mean   :0.9037   N.America:    0   t.co        : 49
##  3rd Qu.:1.0000   3rd Qu.:1.0000   OC      :    0   tableausoftware.com: 31
##  Max.   :5.0000   Max.   :5.0000   SA      :748   visualisingdata.com: 16
##                                     (Other)      : 21
##      Timeinpage      Uniquepageviews      Visits      BouncesNew
##  Min.   :  0.00   Min.   :1.00   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:  0.00   1st Qu.:1.00   1st Qu.:1.0000   1st Qu.:0.000000
##  Median :  0.00   Median :1.00   Median :1.0000   Median :0.010000
##  Mean   : 139.42   Mean   :1.11   Mean   :0.9037   Mean   :0.007179
##  3rd Qu.:  7.25   3rd Qu.:1.00   3rd Qu.:1.0000   3rd Qu.:0.010000
##  Max.   :46745.00   Max.   :5.00   Max.   :5.0000   Max.   :0.050000
##

```

Observation

From the above data we can observe that:

- * *North America* is the continent with the most bounces, most unique page views and most visits
- * *SA* is the continent that has visitors with the most time spent on page.
- * *AF* is the continent that has visitors with the least time spent on page.

Summary grouped by Source group

```
table(webDF$Sourcegroup)
```

```
##
##          (direct)          facebook
##          7532             92
##          google           Others
##          11542            5360
## public.tableausoftware.com reddit.com
##          1354             616
##          t.co             tableausoftware.com
##          2249             2388
##          visualisingdata.com
##          976
```

```
by(webDF, webDF$Sourcegroup, summary)
```

```
## webDF$Sourcegroup: (direct)
##      Bounces      Exits      Continent
##  Min.   : 0.0000  Min.   : 0.000  AF       : 59
## 1st Qu.: 0.0000  1st Qu.: 1.000  AS       : 612
## Median : 1.0000  Median : 1.000  EU       :1432
## Mean   : 0.8323  Mean   : 1.017  N.America:4945
## 3rd Qu.: 1.0000  3rd Qu.: 1.000  OC       : 243
## Max.   :30.0000  Max.   :36.000  SA       : 241
##
##      Sourcegroup      Timeinpage      Uniquepageviews
## (direct)             :7532  Min.   : 0.00  Min.   : 1.000
## facebook              : 0  1st Qu.: 0.00  1st Qu.: 1.000
## google                 : 0  Median : 0.00  Median : 1.000
## Others                  : 0  Mean   : 70.71  Mean   : 1.226
## public.tableausoftware.com: 0  3rd Qu.: 6.00  3rd Qu.: 1.000
## reddit.com              : 0  Max.   :6062.00  Max.   :45.000
## (Other)                  : 0
##      Visits      BouncesNew
##  Min.   : 0.000  Min.   :0.000000
## 1st Qu.: 1.000  1st Qu.:0.000000
## Median : 1.000  Median :0.010000
## Mean   : 1.017  Mean   :0.008323
## 3rd Qu.: 1.000  3rd Qu.:0.010000
## Max.   :45.000  Max.   :0.300000
##
## -----
## webDF$Sourcegroup: facebook
##      Bounces      Exits      Continent      Sourcegroup
##  Min.   :0.000  Min.   :0.000  AF       : 0  facebook      :92
## 1st Qu.:1.000  1st Qu.:1.000  AS       : 8  (direct)      : 0
## Median :1.000  Median :1.000  EU       :16  google        : 0
## Mean   :0.913  Mean   :1.076  N.America:65  Others        : 0
## 3rd Qu.:1.000  3rd Qu.:1.000  OC       : 1  public.tableausoftware.com: 0
## Max.   :2.000  Max.   :2.000  SA       : 2  reddit.com     : 0
##                                     (Other)      : 0
##      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.:0.01000
## Median : 0.00   Median :1.000   Median :1.000   Median :0.01000
## Mean    : 57.21   Mean    :1.109   Mean    :1.076   Mean    :0.00913
## 3rd Qu.: 0.00   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.01000
## Max.    :1195.00   Max.    :2.000   Max.    :2.000   Max.    :0.02000
##
## -----
## webDF$Sourcegroup: google
##      Bounces      Exits      Continent
## Min.      :0.0000   Min.      :0.0000   AF       : 166
## 1st Qu.:0.0000   1st Qu.:1.0000   AS       :1603
## Median :1.0000   Median :1.0000   EU       :1926
## Mean     :0.7565   Mean     :0.9392   N.America:7010
## 3rd Qu.:1.0000   3rd Qu.:1.0000   OC       : 611
## Max.     :5.0000   Max.     :6.0000   SA       : 226
##
##      Sourcegroup      Timeinpage      Uniquepageviews
## google                :11542   Min.      : 0.00   Min.      :1.00
## (direct)              : 0     1st Qu.: 0.00   1st Qu.:1.00
## facebook              : 0     Median : 0.00   Median :1.00
## Others                : 0     Mean    : 67.31   Mean    :1.07
## public.tableausoftware.com: 0     3rd Qu.: 0.00   3rd Qu.:1.00
## reddit.com            : 0     Max.    :46745.00   Max.    :6.00
## (Other)               : 0
##      Visits      BouncesNew
## Min.      :0.0000   Min.      :0.000000
## 1st Qu.:1.0000   1st Qu.:0.000000
## Median :1.0000   Median :0.010000
## Mean     :0.9389   Mean     :0.007565
## 3rd Qu.:1.0000   3rd Qu.:0.010000
## Max.     :6.0000   Max.     :0.050000
##
## -----
## webDF$Sourcegroup: Others
##      Bounces      Exits      Continent
## Min.      :0.0000   Min.      :0.000   AF       : 57
## 1st Qu.:0.0000   1st Qu.:1.000   AS       : 506
## Median :1.0000   Median :1.000   EU       :1489
## Mean     :0.6368   Mean     :0.833   N.America:2917
## 3rd Qu.:1.0000   3rd Qu.:1.000   OC       : 227
## Max.     :8.0000   Max.     :8.000   SA       : 164
##
##      Sourcegroup      Timeinpage      Uniquepageviews
## Others                :5360   Min.      : 0.00   Min.      :1.000
## (direct)              : 0     1st Qu.: 0.00   1st Qu.:1.000
## facebook              : 0     Median : 0.00   Median :1.000
## google                : 0     Mean    : 70.38   Mean    :1.067
## public.tableausoftware.com: 0     3rd Qu.: 12.00   3rd Qu.:1.000
## reddit.com            : 0     Max.    :4190.00   Max.    :8.000
## (Other)               : 0
##      Visits      BouncesNew
## Min.      :0.0000   Min.      :0.000000
## 1st Qu.:1.0000   1st Qu.:0.000000

```

```

## Median :1.0000    Median :0.010000
## Mean   :0.8334    Mean    :0.006368
## 3rd Qu.:1.0000    3rd Qu.:0.010000
## Max.   :8.0000    Max.    :0.080000
##
## -----
## webDF$Sourcegroup: public.tableausoftware.com
##      Bounces      Exits      Continent
## Min.   :0.0000    Min.   :0.0000    AF      : 4
## 1st Qu.:0.0000    1st Qu.:0.0000    AS      : 33
## Median :0.0000    Median :1.0000    EU      : 118
## Mean   :0.3331    Mean   :0.6477    N.America:1134
## 3rd Qu.:1.0000    3rd Qu.:1.0000    OC      : 52
## Max.   :3.0000    Max.   :4.0000    SA      : 13
##
##      Sourcegroup      Timeinpage      Uniquepageviews
## public.tableausoftware.com:1354    Min.   : 0.00    Min.   :1.000
## (direct)      : 0    1st Qu.: 0.00    1st Qu.:1.000
## facebook      : 0    Median : 12.00    Median :1.000
## google        : 0    Mean   : 134.32    Mean   :1.123
## Others        : 0    3rd Qu.: 72.75    3rd Qu.:1.000
## reddit.com    : 0    Max.   :4561.00    Max.   :6.000
## (Other)      : 0
##      Visits      BouncesNew
## Min.   :0.0000    Min.   :0.000000
## 1st Qu.:0.0000    1st Qu.:0.000000
## Median :1.0000    Median :0.000000
## Mean   :0.6477    Mean   :0.003331
## 3rd Qu.:1.0000    3rd Qu.:0.010000
## Max.   :6.0000    Max.   :0.030000
##
## -----
## webDF$Sourcegroup: reddit.com
##      Bounces      Exits      Continent
## Min.   :0.0000    Min.   :0.0000    AF      : 1
## 1st Qu.:0.7500    1st Qu.:1.0000    AS      : 25
## Median :1.0000    Median :1.0000    EU      :127
## Mean   :0.8036    Mean   :0.9432    N.America:433
## 3rd Qu.:1.0000    3rd Qu.:1.0000    OC      : 24
## Max.   :4.0000    Max.   :4.0000    SA      : 6
##
##      Sourcegroup      Timeinpage      Uniquepageviews
## reddit.com      :616    Min.   : 0.0    Min.   :1.000
## (direct)      : 0    1st Qu.: 0.0    1st Qu.:1.000
## facebook      : 0    Median : 0.0    Median :1.000
## google        : 0    Mean   : 49.2    Mean   :1.086
## Others        : 0    3rd Qu.: 0.0    3rd Qu.:1.000
## public.tableausoftware.com: 0    Max.   :2047.0    Max.   :5.000
## (Other)      : 0
##      Visits      BouncesNew
## Min.   :0.0000    Min.   :0.000000
## 1st Qu.:1.0000    1st Qu.:0.007500
## Median :1.0000    Median :0.010000
## Mean   :0.9432    Mean   :0.008036

```



```

## 3rd Qu.:1.0000 3rd Qu.:0.010000
## Max. :5.0000 Max. :0.040000
##
## -----
## webDF$Sourcegroup: t.co
## Bounces Exits Continent
## Min. :0.0000 Min. : 0.0000 AF : 8
## 1st Qu.:0.0000 1st Qu.: 1.0000 AS : 51
## Median :1.0000 Median : 1.0000 EU : 630
## Mean :0.7661 Mean : 0.9289 N.America:1448
## 3rd Qu.:1.0000 3rd Qu.: 1.0000 OC : 63
## Max. :9.0000 Max. :10.0000 SA : 49
##
## Sourcegroup Timeinpage Uniquepageviews
## t.co :2249 Min. : 0.00 Min. : 1.000
## (direct) : 0 1st Qu.: 0.00 1st Qu.: 1.000
## facebook : 0 Median : 0.00 Median : 1.000
## google : 0 Mean : 69.92 Mean : 1.146
## Others : 0 3rd Qu.: 5.00 3rd Qu.: 1.000
## public.tableausoftware.com: 0 Max. :14108.00 Max. :11.000
## (Other) : 0
## Visits BouncesNew
## Min. : 0.0000 Min. :0.000000
## 1st Qu.: 1.0000 1st Qu.:0.000000
## Median : 1.0000 Median :0.010000
## Mean : 0.9289 Mean :0.007661
## 3rd Qu.: 1.0000 3rd Qu.:0.010000
## Max. :11.0000 Max. :0.090000
##
## -----
## webDF$Sourcegroup: tableausoftware.com
## Bounces Exits Continent
## Min. :0.0000 Min. :0.0000 AF : 16
## 1st Qu.:0.0000 1st Qu.:0.0000 AS : 196
## Median :0.0000 Median :1.0000 EU : 390
## Mean :0.5054 Mean :0.7374 N.America:1652
## 3rd Qu.:1.0000 3rd Qu.:1.0000 OC : 103
## Max. :4.0000 Max. :4.0000 SA : 31
##
## Sourcegroup Timeinpage Uniquepageviews
## tableausoftware.com :2388 Min. : 0.00 Min. :1.000
## (direct) : 0 1st Qu.: 0.00 1st Qu.:1.000
## facebook : 0 Median : 0.00 Median :1.000
## google : 0 Mean : 90.85 Mean :1.072
## Others : 0 3rd Qu.: 32.00 3rd Qu.:1.000
## public.tableausoftware.com: 0 Max. :4707.00 Max. :5.000
## (Other) : 0
## Visits BouncesNew
## Min. :0.0000 Min. :0.000000
## 1st Qu.:0.0000 1st Qu.:0.000000
## Median :1.0000 Median :0.000000
## Mean :0.7374 Mean :0.005054
## 3rd Qu.:1.0000 3rd Qu.:0.010000
## Max. :5.0000 Max. :0.040000

```

```
##
## -----
## webDF$Sourcegroup: visualisingdata.com
##      Bounces      Exits      Continent
## Min.   :0.0000   Min.   :0.0000   AF      : 10
## 1st Qu.:0.0000   1st Qu.:0.0000   AS      :137
## Median :1.0000   Median :1.0000   EU      :342
## Mean   :0.5338   Mean   :0.7387   N.America:439
## 3rd Qu.:1.0000   3rd Qu.:1.0000   OC      : 32
## Max.   :3.0000   Max.   :3.0000   SA      : 16
##
##      Sourcegroup      Timeinpage      Uniquepageviews
## visualisingdata.com      :976   Min.   : 0.00   Min.   :1.000
## (direct)                  : 0   1st Qu.: 0.00   1st Qu.:1.000
## facebook                  : 0   Median : 0.00   Median :1.000
## google                    : 0   Mean   : 73.27   Mean   :1.078
## Others                    : 0   3rd Qu.: 31.00   3rd Qu.:1.000
## public.tableausoftware.com: 0   Max.   :2954.00   Max.   :5.000
## (Other)                   : 0
##      Visits      BouncesNew
## Min.   :0.0000   Min.   :0.000000
## 1st Qu.:0.0000   1st Qu.:0.000000
## Median :1.0000   Median :0.010000
## Mean   :0.7387   Mean   :0.005338
## 3rd Qu.:1.0000   3rd Qu.:0.010000
## Max.   :5.0000   Max.   :0.030000
##
```

Observation

From the above data we can observe that:

- * *Google* source group has the most visitors and they spent the most time on the website
- * *(direct)* source group has the most unique page views, most exits and most bounces.
- * *Facebook* source group has the least amount of visits and unique page views

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.

Option 1: ANOVA

```
UnqPg_View <- aov(Uniquepageviews ~ Visits, webDF)
summary(UnqPg_View)
```

```
##      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
```

Option 2: Correlation function

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

```
## [1] 0.8144457
```

Observation

Based on the ANOVA and correlation function output we can see that there is a strong positive relationship between the *unique page views* and the *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.

```
Ext <- aov(Exits ~ ., webDF)
summary(Ext)
```

```
##              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
```

Observation

From the above result of ANOVA we can observe that *Bounces*, *Continent*, *Sourcegroup*, *Timeinpage* and *Uniquepageviews* have a higher significance over *Exits*. The *Exits* are less affected by *Visits*.

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.

```
Time_Page <- aov(Timeinpage ~ ., webDF)
summary(Time_Page)
```

```
##              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
```

Observation

From the above result of ANOVA we can observe that *Bounces*, *Continent*, *Sourcegroup*, *Uniquepageviews* and *Visits* have a higher significance over the time spent on page. The *Sourcegroup* doesn't have any impact on the time spent 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.

```
fit1 <- glm(
  BouncesNew ~ Timeinpage + Continent + Exits + Sourcegroup +
    Uniquepageviews + Visits,
  webDF,
  family = "binomial"
)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit1)
```

```
##
## Call:
## glm(formula = BouncesNew ~ Timeinpage + Continent + Exits + Sourcegroup +
##       Uniquepageviews + Visits, family = "binomial", data = webDF)
##
## 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
##
## (Intercept)          ***
## Timeinpage           .
## ContinentAS
## ContinentEU
## ContinentN.America
## ContinentOC
## ContinentSA
## Exits                ***
## Sourcegroupfacebook
## Sourcegroupgoogle
## SourcegroupOthers
## Sourcegrouppublic.tableausoftware.com
## Sourcegroupreddit.com
## Sourcegroupt.co
## Sourcegrouptableausoftware.com
## Sourcegroupvisualisingdata.com
## Uniquepageviews      ***
## Visits                ***
## ---
## 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
```

Observation

As it can be observed from the output shown above, the *Exits*, *Uniquepageviews* and *Visits* are the variables that have the most impact on the bounce.