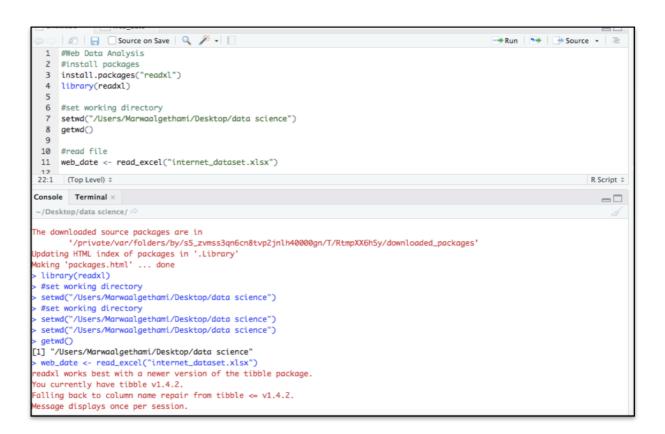
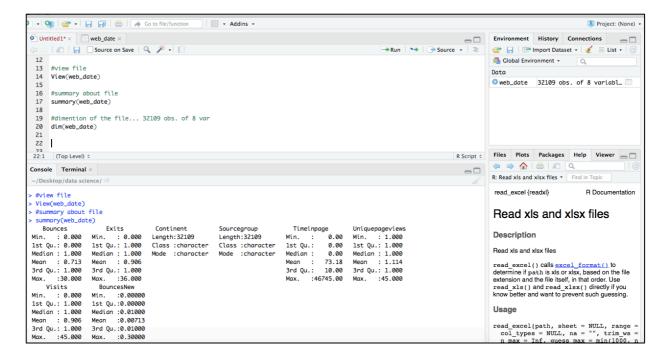
## **Web Data Analysis**

## Question 1:

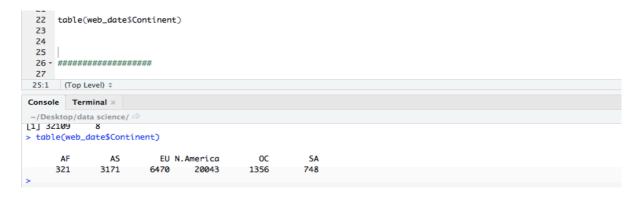
 The team wants to analyze each variable of the data collected through data summarization to get a basic understanding of the dataset and to prepare for further analysis.





As seen in summary function for:

- bounces min=0,max=30. there is a maximum value of 30 bounces for the website.
- exit min=o max=36



N.America has maximum number of times of visitors to the site.

### Question 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.

```
25 - #################
  # to find the correlation between Uniquepageviews and Visits
  27 cor(web_date$Uniquepageviews,web_date$Visits)
  28
  29
  30 anova_test<-aov(Uniquepageviews~Visits, data=web_date)
  31 summary(anova_test)
  32
  33
 32:1 (Untitled) $
Console Terminal ×
 ~/Desktop/data science/
> # to find the correlation between Uniquepageviews and Visits
> cor(web_date$Uniquepageviews,web_date$Visits)
[1] 0.8144457
> summary(anova_test
+ summary(anova_test)
Error: unexpected symbol in:
"summary(anova_test
summary"
> anova_test<-aov(Uniquepageviews~Visits, data=web_date)
> summary(anova_test)
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
```

from the results of anova test the visits variable has a significant impact on Unique Page views.

So the team can conclude that unique page values depend on visits.

### Question 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.

```
33 - ################
     # to find the varibles that affect exit variable
     anova_test2<-aov(Exits~.,data = web_date)</pre>
 36 summary(anova_test2)
 37
36:21 (Untitled) $
Console Terminal ×
~/Desktop/data science/
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '1
> # to find the varibles that affect exit variable
> anova_test2<-aov(Exits~.,data = web_date)</pre>
> summary(anova_test2)
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 32091 3254
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result of ANOVA test the exit from the site is affected by the factors of source group, bounces, and unique.pageviews. and not affected by visits .

#### Question 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.

From the result of ANOVA test all factors are afecting the time in page views except source group is not affecting the time in page views.

#### Question 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.

```
Call:
glm(formula = Bounces ~ Timeinpage + Continent + Exits + Sourcegroup +
   Uniquepageviews + Visits, family = "binomial", data = web_date)
Deviance Residuals:
                     Median
    Min 1Q
-2.26149 -0.02406 0.00206 0.00895 1.81288
Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
                                    -4.9667681 0.6784678 -7.321 2.47e-13 ***
(Intercept)
                                    -0.0010294 0.0005774 -1.783 0.0746 .
Timeinpaae
                                    0.0022768 0.6932044 0.003 0.9974
ContinentAS
                                    -0.0069240 0.6786600 -0.010 0.9919 0.0101334 0.6674188 0.015 0.9879
ContinentEU
ContinentN.America
                                    0.0201123 0.7333671 0.027 0.9781
ContinentOC
ContinentSA
                                    0.0237507 0.7914250 0.030
                                                                  0.9761
                                    1.3907608 0.3356504 4.143 3.42e-05 ***
Exits
Sourcegroupfacebook
                                   -0.0241949 1.1045171 -0.022 0.9825
                                   -0.0783631 0.1720157 -0.456
Sourcegroupgoogle
SourcearoupOthers
                                   -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
                                   0.0148690 0.2760157 0.054 0.9570
Sourcegroupt.co
Sourcegrouptableausoftware.com
Sourcegroupvisualisingdata.com
                                   -0.1129305 0.3190762 -0.354
                                                                  0.7234
                                   -0.0822525 0.4614866 -0.178
                                                                  0.8585
                                   -3.2363108 0.5791664 -5.588 2.30e-08 ***
Uniquepageviews
                                    2.1941121 0.5202216 4.218 2.47e-05 ***
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
```

from the result shown, the Unique.Pageviews and visits are the variables that impact the target variable bounces it has greater significance.



Smaller AIC values indicate the model is closer to the truth.

# Code

```
#Web Data Analysis
#install packages
install.packages("readxl")
library(readxl)
#set working directory
setwd("/Users/Marwaalgethami/Desktop/data science")
getwd()
#read file
web date <- read excel("internet dataset.xlsx")</pre>
#view file
View(web date)
#summary about file
summary(web date)
#dimention of the file... 32109 obs. of 8 var
dim(web date)
table(web date$Continent)
######################
# to find the correlation between Uniquepageviews and
Visits
cor(web date$Uniquepageviews,web date$Visits)
anova_test<-aov(Uniquepageviews~Visits, data=web_date)
summary(anova test)
#################
# to find the varibles that affect exit variable
anova_test2<-aov(Exits~.,data = web_date)
summary(anova test2)
#########
#to find the variables that effect on the time on page.
anova_test3<-aov(Timeinpage~.,data = web_date)
summary(anova test3)
```

genLinModel\$aic