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.

```r
#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")
```

```
The downloaded source packages are in
        '/private/var/folders/by/s5_zvmss3qn6cn8tvp2jnlh40000gn/T/RtmpXX6hSy/downloaded_packages'
Updating HTML index of packages in '.Library'
Making 'packages.html' ... done
> library(readxl)
> #set working directory
> setwd("/Users/Marwaalgethami/Desktop/data science")
> #set working directory
> setwd("/Users/Marwaalgethami/Desktop/data science")
> setwd("/Users/Marwaalgethami/Desktop/data science")
> getwd()
[1] "/Users/Marwaalgethami/Desktop/data science"
> web_date <- read_excel("internet_dataset.xlsx")
readxl works best with a newer version of the tibble package.
You currently have tibble v1.4.2.
Falling back to column name repair from tibble <= v1.4.2.
Message displays once per session.
```

```
#view file
View(web_date)

#summary about file
summary(web_date)

#dimention of the file... 32109 obs. of 8 var
dim(web_date)
```

```
> #view file
> View(web_date)
> #summary about file
> summary(web_date)
    Bounces           Exits          Continent          Sourcegroup          Timeinpage        Uniquepageviews
 Min.   : 0.000   Min.   : 0.000   Length:32109       Length:32109       Min.   :     0.00   Min.   : 1.000
 1st Qu.: 0.000   1st Qu.: 1.000   Class :character   Class :character   1st Qu.:     0.00   1st Qu.: 1.000
 Median : 1.000   Median : 1.000   Mode  :character   Mode  :character   Median :     0.00   Median : 1.000
 Mean   : 0.713   Mean   : 0.906                                         Mean   :    73.18   Mean   : 1.114
 3rd Qu.: 1.000   3rd Qu.: 1.000                                         3rd Qu.:    10.00   3rd Qu.: 1.000
 Max.   :30.000   Max.   :36.000                                         Max.   :46745.00   Max.   :45.000
     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
```

As seen in summary function for :

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

```
table(web_date$Continent)



####################
```

```
[1] 32109     8
> table(web_date$Continent)

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

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.

```r
###################
 # 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)
```

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.

```
################
# to find the varibles that affect exit variable
anova_test2<-aov(Exits~.,data = web_date)
summary(anova_test2)
```

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)
> summary(anova_test2)
                   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 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.

```r
#########
#to find the variables that effect on the time on page.
anova_test3<-aov(Timeinpage~.,data = web_date)
summary(anova_test3)
```

42:21 (Untitled)

Console Terminal

~/Desktop/data science/

```
> #########
> #to find the variables that effect on the time on page.
> anova_test3<-aov(Timeinpage~.,data = web_date)
> summary(anova_test3)
                   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
>
```

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.

```r
##############
#Q5: determine the factors that are impacting the bounce.
#data for the variable bounces has to be between 0 and 1,
web_date$Bounces=web_date$Bounces*0.01
genLinModel<-glm(Bounces~Timeinpage+Continent+Exits+Sourcegroup+Uniquepageviews+Visits,data = web_date, family = "binomial"
summary(genLinModel)
```

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

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

>
```

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

```
51
52  genLinModel$aic
53
51:1   (Untitled)                                                    R Script
Console   Terminal
~/Desktop/data science/

> genLinModel$aic
[1] 506.5577
>
```

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")

#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)
```

```
##############
#Q5: determine the factors that are impacting the bounce.
#data for the variable bounces has to be between 0 and 1,
web_date$Bounces=web_date$Bounces*0.01
genLinModel<-
glm(Bounces~Timeinpage+Continent+Exits+Sourcegroup+Unique
pageviews+Visits,data = web_date, family = "binomial")
summary(genLinModel)

genLinModel$aic
```