

Name: OLWETHU MATIWANE
Email: jimmyolwethu7@gmail.com

WEB DATA ANALYSIS - PROJECT 1

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.

CODE:

```
setwd(choose.dir())  
getwd()  
  
Web_Data <- read.csv("internet_dataset.csv")  
View(Web_Data)  
str(Web_Data)  
summary(Web_Data)
```

RESULTS:

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 for the website. This site was accessed maximum number of times by visitors from North America.

Screenshots:

```
> summary(web_Data)  
      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
```

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.

CODE:

```
Web_Data_AOV <- aov(Visits ~ Uniquepageviews, data = Web_Data)
summary(Web_Data_AOV)
```

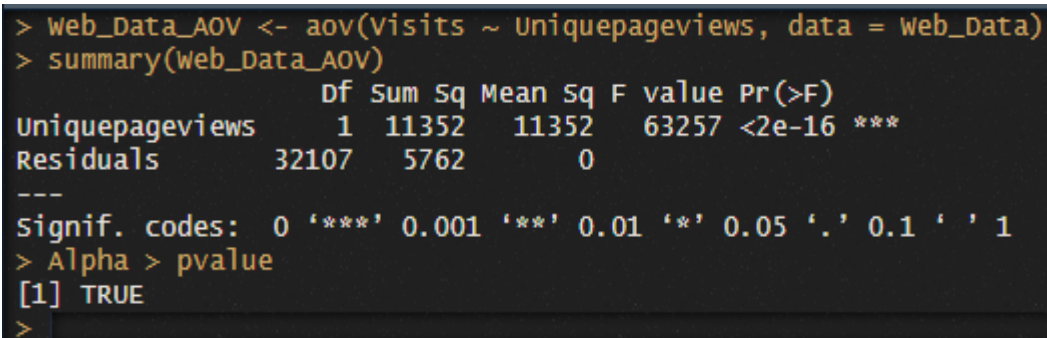
```
Alpha <- 0.05
pvalue <- 1.62e-05
```

```
Alpha > pvalue
```

RESULTS:

I can confirm that the Visits variable has a significant impact on UniquePageViews. In this case we reject the Null Hypothesis since P-value which is '2e-16' is less than Alpha '0.05'. This means that the team can conclude that unique page values depend on visits.

Screenshots:



```
> web_Data_AOV <- aov(visits ~ Uniquepageviews, data = web_Data)
> summary(web_Data_AOV)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Uniquepageviews	1	11352	11352	63257	<2e-16 ***
Residuals	32107	5762	0		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> Alpha > pvalue
[1] TRUE
>
```

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.

CODE:

```
Web_Exits_Data <- aov(Exits ~., data = Web_Data)
summary(Web_Exits_Data)
```

RESULTS:


Using ANOVA this code will print out the probable factors from the dataset affecting the Exits variable. By looking at the screen below we can tell which variable affect the Exits. It has all the information about each variable and in this case, this is what we are looking for to identify which variables affect **Exits** variable.

Screenshots:

```
> web_Exits_Data <- aov(Exits ~., data = web_Data)
> summary(web_Exits_Data)
```

	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
```



This is for Bounces, Timeinpage, and Uniquepageviews:

```
pvalue <- 2e-16
Alpha > pvalue
```

```
> pvalue <- 2e-16
> Alpha > pvalue
[1] TRUE
```

True means that Bounces, Timeinpage, and Uniquepageviews does affect Exits

This is for Continent:

```
pvalue <- 1.62e-05
Alpha > pvalue
```

```
> pvalue <- 1.62e-05
> Alpha > pvalue
[1] TRUE
```

True means that Continent does affect Exits

This is for Sourcegroup:

```
pvalue <- 4.89e-12
Alpha > pvalue
```

```
> pvalue <- 4.89e-12
> Alpha > pvalue
[1] TRUE
```

True means that Sourcegroup does affect Exits

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.

CODE:

```
Web_Time_In_Data <- aov(Timeinpage ~., data = Web_Data)
summary(Web_Time_In_Data)
```

RESULTS:

```
> Web_Time_In_Data <- aov(Timeinpage ~., data = Web_Data)
> summary(Web_Time_In_Data)
```

	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
```

This is for Bounces, Exits, Uniquepageviews, and Visits:

```
> pvalue <- 2e-16
> Alpha > pvalue
[1] TRUE
```

```
pvalue <- 2e-16
Alpha > pvalue
```

True means that Bounces, Exits, Uniquepageviews, and Visits influence Timeinpage

This is for Continent:

```
pvalue <- 2.51e-06
Alpha > pvalue
```

```
> pvalue <- 2.51e-06
> Alpha > pvalue
[1] TRUE
```

True means that Continent influence Timeinpage

This is for Sourcegroup:

```
pvalue <- 0.202
Alpha > pvalue
```

```
> pvalue <- 0.202
> Alpha > pvalue
[1] FALSE
```

In this case False means that Sourcegroup does not influence Timeinpage

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.

CODE:

```
Web_Bounce_Log <- glm(Bounces*0.01 ~.,
                      family = binomial(link = 'logit'),
                      data = Web_Data)
summary(Web_Bounce_Log)
```

RESULTS:

Using Generalized Linear Model (glm function) this code will determine the factors that are impacting the bounce. So, all the factors impacting Bounces will display in the short screen below.

```
> summary(Web_Bounce_Log)

Call:
glm(formula = Bounces * 0.01 ~ ., family = binomial(link = "logit"),
    data = web_Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.30605  -0.03435  -0.00133   0.00097   2.47635

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.028e+00  6.784e-01  -7.412 1.24e-13 ***
Exits        1.255e-01  4.879e-01   0.257  0.7970
ContinentAS  2.346e-02  6.934e-01   0.034  0.9730
ContinentEU  2.223e-02  6.790e-01   0.033  0.9739
ContinentN.America 2.779e-02  6.678e-01   0.042  0.9668
ContinentOC  2.812e-02  7.336e-01   0.038  0.9694
ContinentSA  4.073e-02  7.923e-01   0.051  0.9590
Sourcegroupfacebook 1.309e-02  1.105e+00   0.012  0.9905
Sourcegroupgoogle -2.002e-02  1.731e-01  -0.116  0.9079
Sourcegroupothers -4.256e-02  2.189e-01  -0.194  0.8458
Sourcegrouppublic.tableausoftware.com -6.676e-02  4.942e-01  -0.135  0.8925
Sourcegroupreddit.com -2.519e-03  4.713e-01  -0.005  0.9957
Sourcegroupt.co  2.346e-02  2.765e-01   0.085  0.9324
Sourcegrouptableausoftware.com -6.001e-02  3.196e-01  -0.188  0.8511
Sourcegroupvisualisingdata.com -5.000e-02  4.619e-01  -0.108  0.9138
Timeinpage  5.262e-05  1.353e-04   0.389  0.6973
Uniquepageviews -2.467e+00  5.784e-01  -4.266 1.99e-05 ***
Visits       1.167e+00  5.809e-01   2.010  0.0445 *
BouncesNew   1.596e+02  3.507e+01   4.551 5.34e-06 ***
---
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:  69.673  on 32090  degrees of freedom
AIC: 502.65
```

```
Number of Fisher Scoring iterations: 10
```