Web Data Analysis

Q1) 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.

Ans) For data analysis of each variable of the data,below attached is the code in R which gives an overview of the summary of web data analysis dataset

print("Web Data Analysis")

web_data<-
read.csv("https://raw.githubusercontent.com/shivanipriya89/WebData/main/Internet.csv")

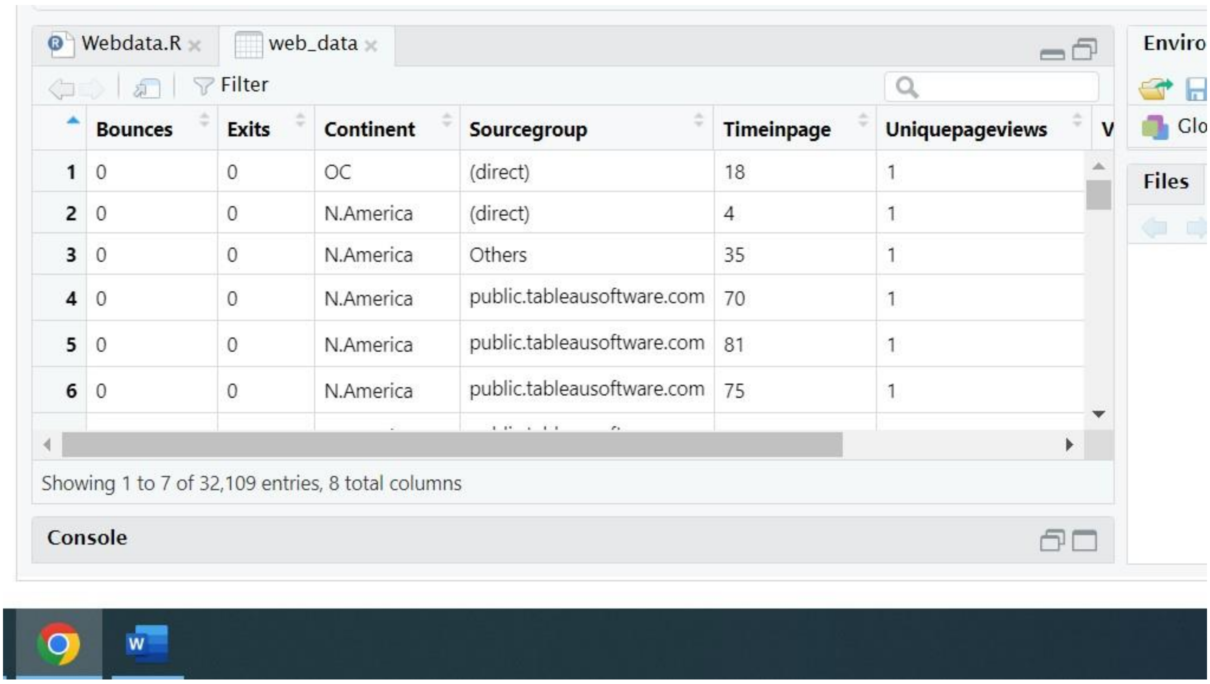
web_data

View(web_data)

str(web_data)

summary(web_data)

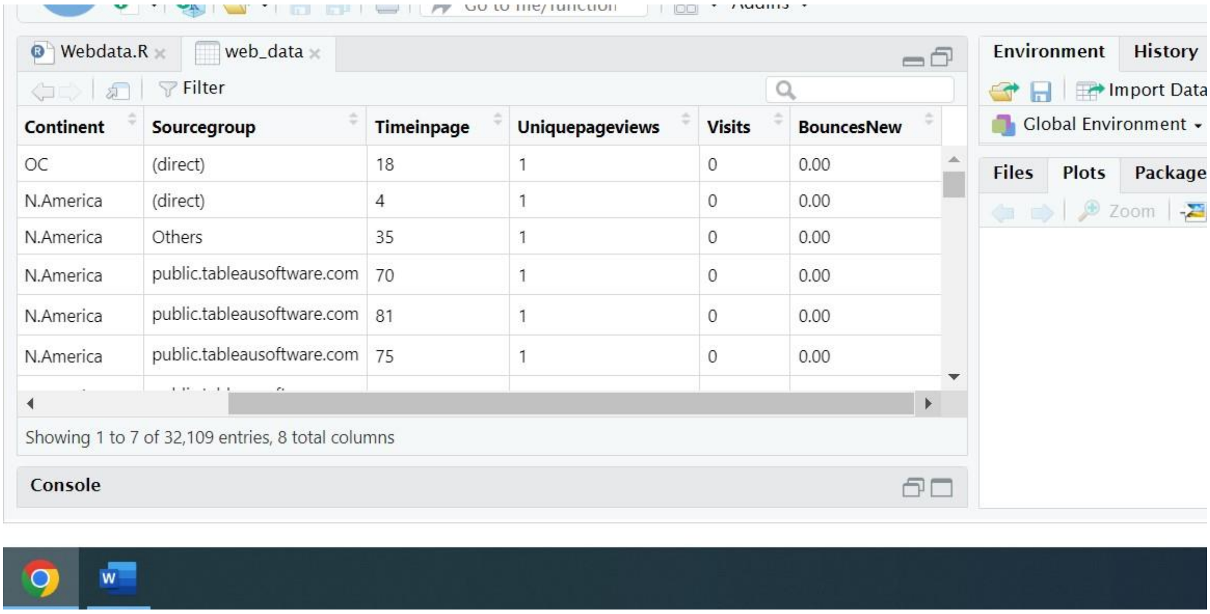
The below attached are the screenshots of tabular view of webdata page,datatypes of various column and the summary of web data analysis dataset



	Bounces	Exits	Continent	Sourcegroup	Timeinpage	Uniquepageviews	V
1	0	0	OC	(direct)	18	1	
2	0	0	N.America	(direct)	4	1	
3	0	0	N.America	Others	35	1	
4	0	0	N.America	public.tableausoftware.com	70	1	
5	0	0	N.America	public.tableausoftware.com	81	1	
6	0	0	N.America	public.tableausoftware.com	75	1	

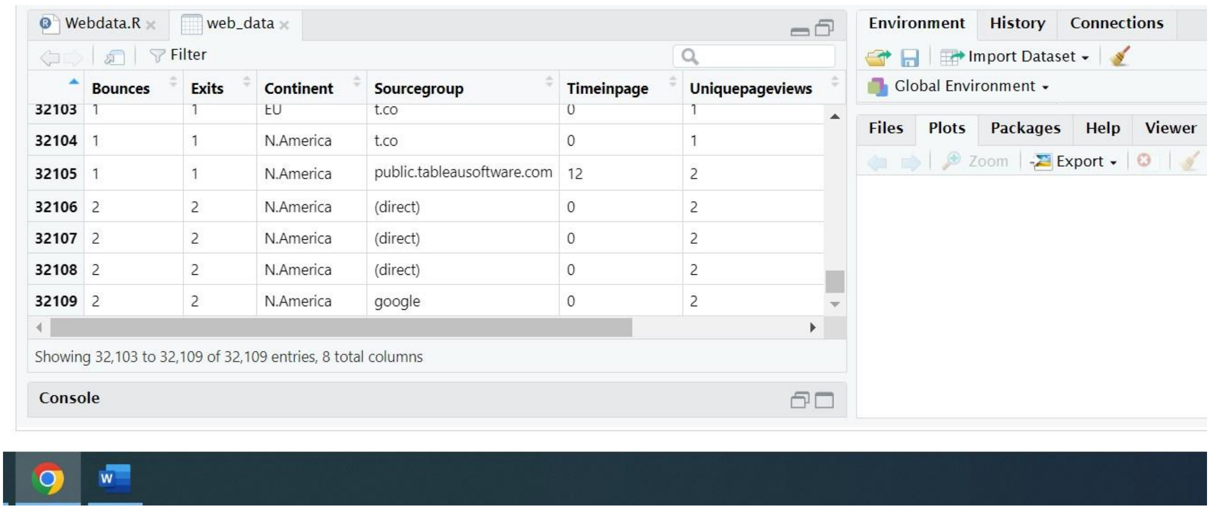
Showing 1 to 7 of 32,109 entries, 8 total columns

Console



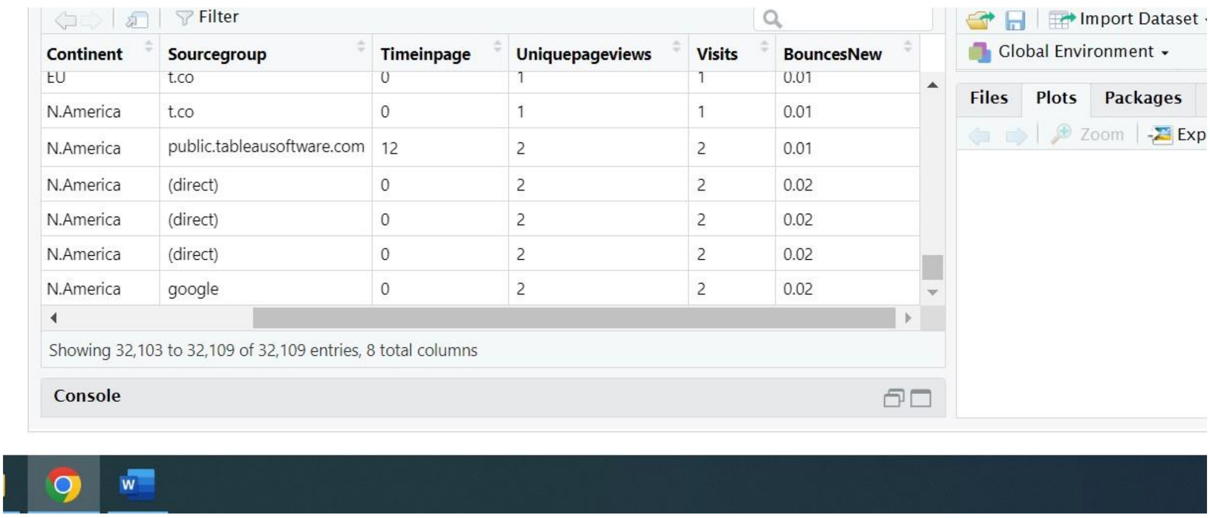
Continent	Sourcegroup	Timeinpage	Uniquepageviews	Visits	BouncesNew
OC	(direct)	18	1	0	0.00
N.America	(direct)	4	1	0	0.00
N.America	Others	35	1	0	0.00
N.America	public.tableausoftware.com	70	1	0	0.00
N.America	public.tableausoftware.com	81	1	0	0.00
N.America	public.tableausoftware.com	75	1	0	0.00

Showing 1 to 7 of 32,109 entries, 8 total columns



	Bounces	Exits	Continent	Sourcegroup	Timeinpage	Uniquepageviews
32103	1	1	EU	t.co	0	1
32104	1	1	N.America	t.co	0	1
32105	1	1	N.America	public.tableausoftware.com	12	2
32106	2	2	N.America	(direct)	0	2
32107	2	2	N.America	(direct)	0	2
32108	2	2	N.America	(direct)	0	2
32109	2	2	N.America	google	0	2

Showing 32,103 to 32,109 of 32,109 entries, 8 total columns

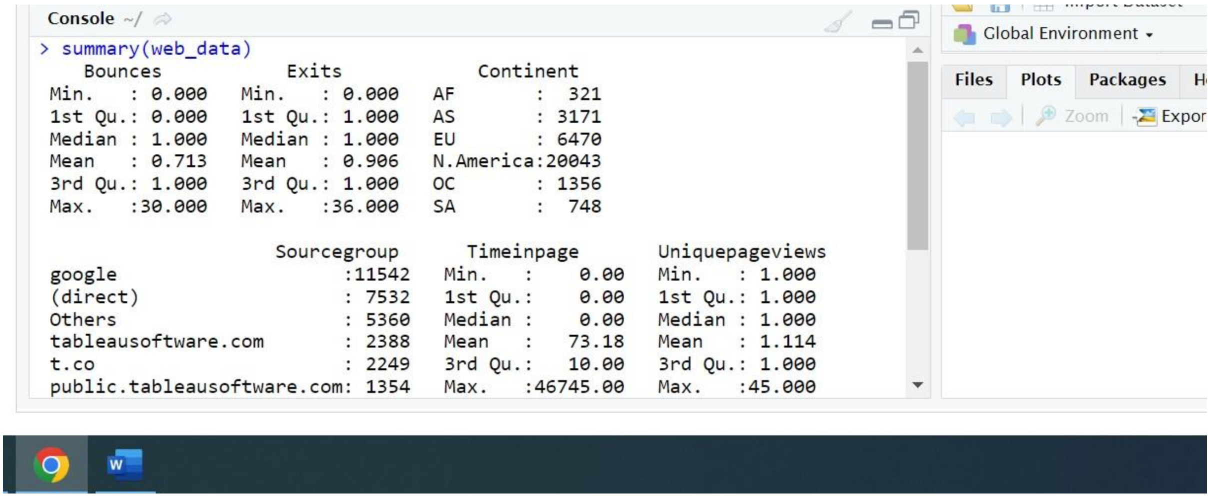


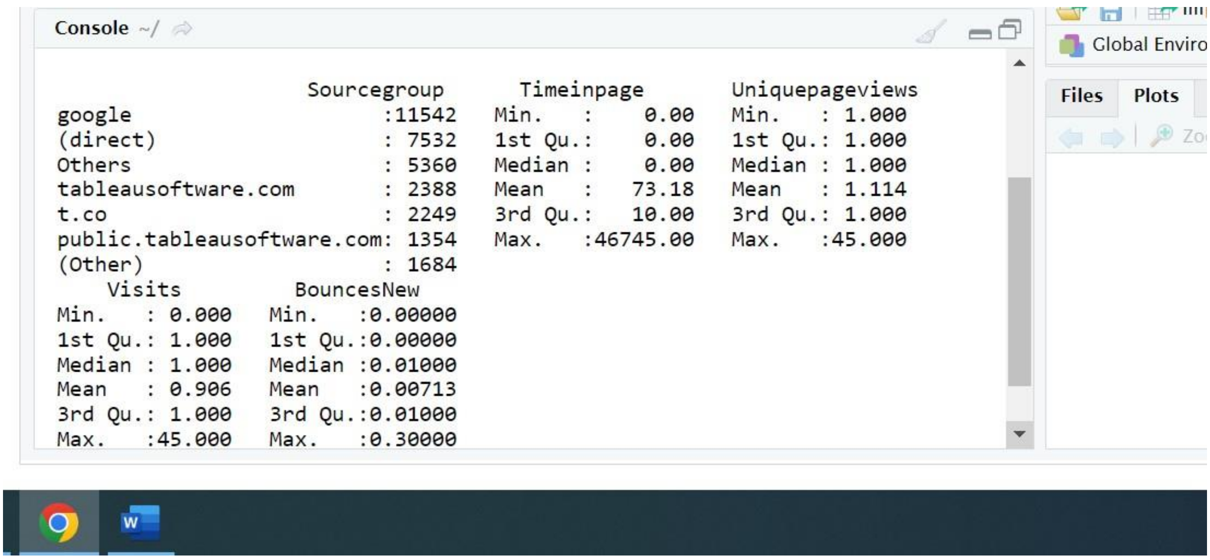
Continent	Sourcegroup	Timeinpage	Uniquepageviews	Visits	BouncesNew
EU	t.co	0	1	1	0.01
N.America	t.co	0	1	1	0.01
N.America	public.tableausoftware.com	12	2	2	0.01
N.America	(direct)	0	2	2	0.02
N.America	(direct)	0	2	2	0.02
N.America	(direct)	0	2	2	0.02
N.America	google	0	2	2	0.02

Showing 32,103 to 32,109 of 32,109 entries, 8 total columns

The above mention table of web data analysis has 32,109 entries of all 8 columns of web data set

The below attached is the summary of the web data analysis dataset which has min,median,1st and 3rd Quantiles of various columns of the webdata set



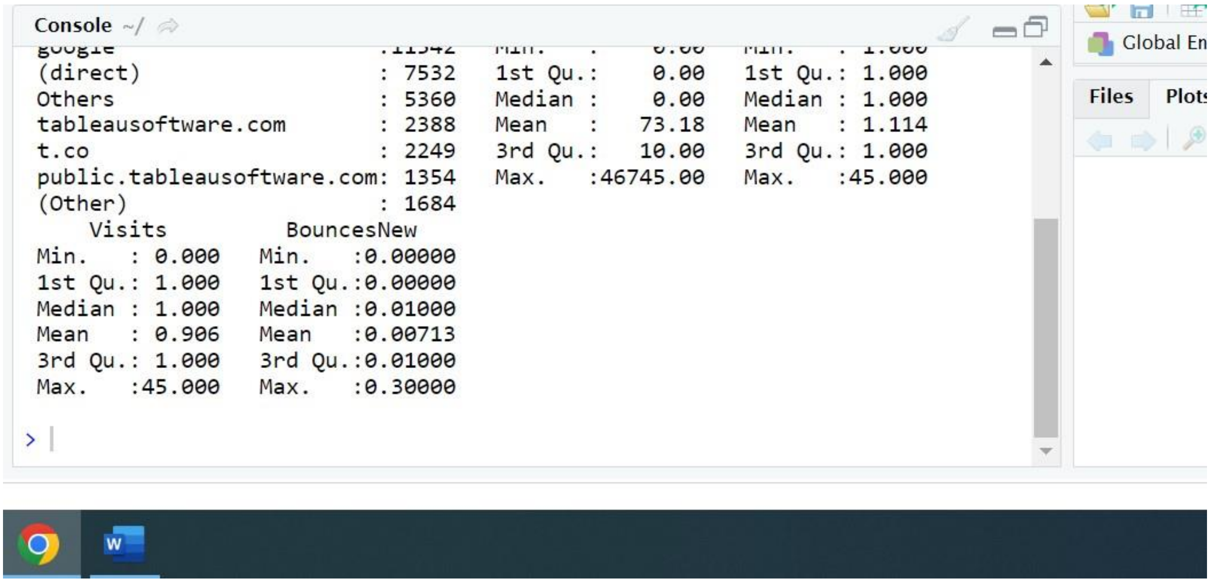


The internet dataset is an excel file with(.xlsx) extension. I have converted this file to internet.csv

Click on this link

https://raw.githubusercontent.com/shivanipriya89/WebData/main/Internet.csv

for viewing the csv file



Q2) 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.

Ans) For determining the relationship between the unique page value and visits, I am using the concept of Simple Linear Regression for determining the relationship between unique page value and visits. The below mention is the code in R which gives an overview of Simple Linear Regression

```
print("Web Data Analysis")

web_data<-
read.csv("https://raw.githubusercontent.com/shivanipriya89/WebData/main/Internet.csv")

web_data

View(web_data)

str(web_data)

unique_page<-lm(formula=Uniquepageviews~Visits,data=web_data)

unique_page

summary(unique_page) # Positive Linear Regression


# GGPlot


ggplot(data=web_data,mapping =
aes(x="Uniquepageviews",y="Visits"))+geom_point(alpha=0.1,color="blue")
```

ggplot(data = web_data, mapping = aes(x = Uniquepageviews, y = Visits)) +

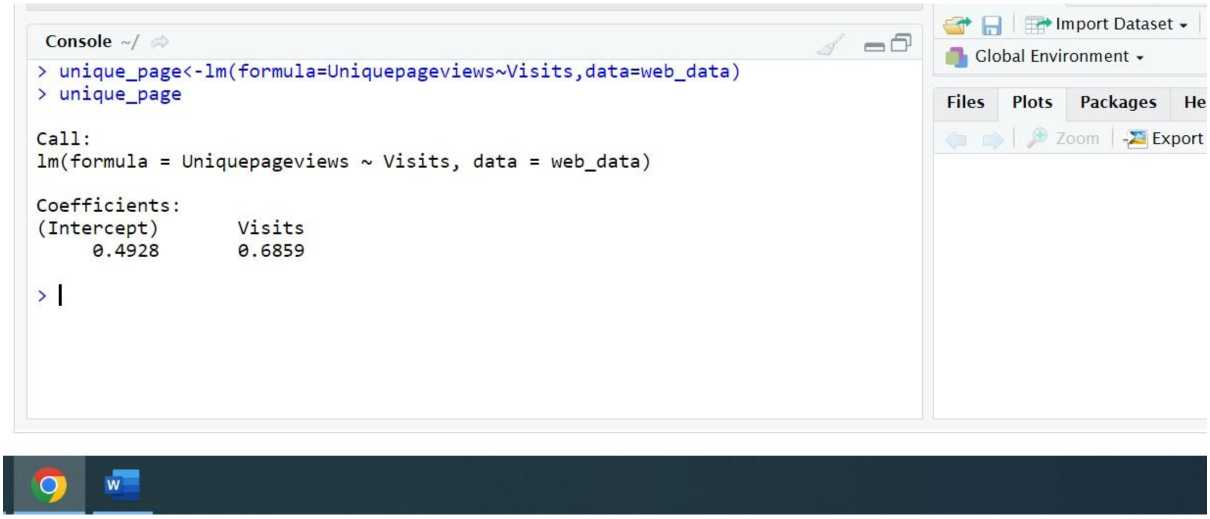
geom_boxplot()

uds<-table(web_data$Uniquepageviews,web_data$Visits)

uds

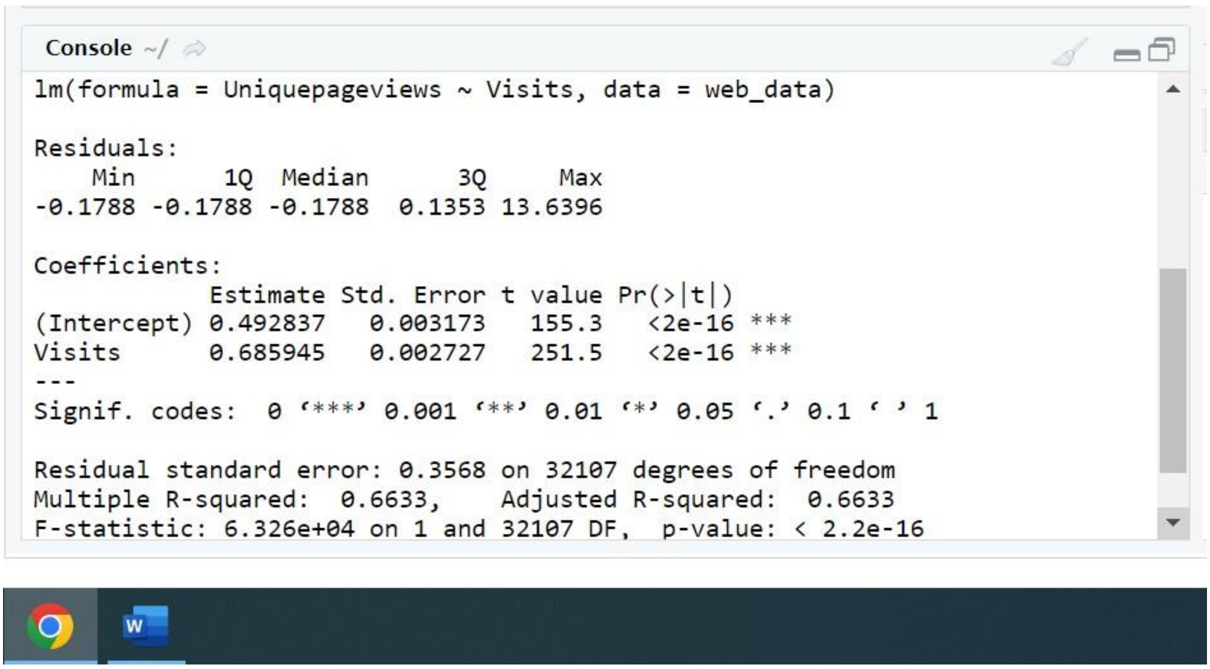
View(uds)

The below attached are the screenshots



```
Console ~/
> unique_page<-lm(formula=Uniquepageviews~Visits,data=web_data)
> unique_page

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

Coefficients:
(Intercept)       Visits
     0.4928       0.6859

> |
```

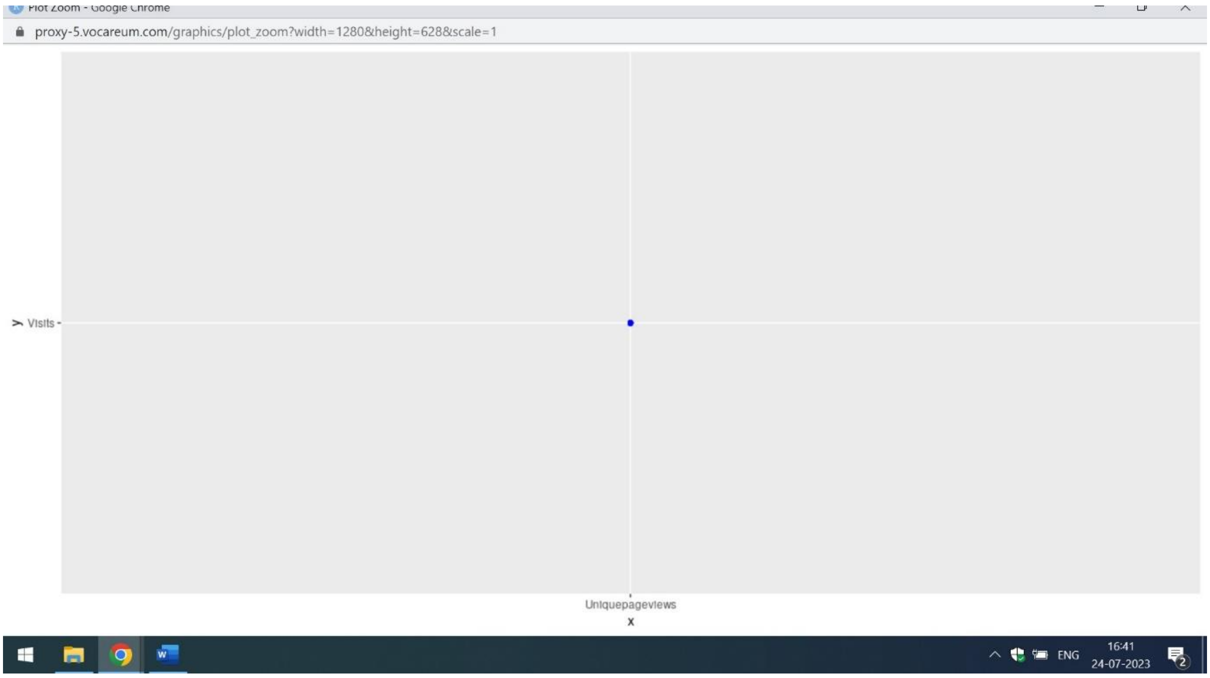

It is clear from the above mention screenshot that there is a positive Linear Regression between UniquepageViews and Visits as the value on the Y intercept is positive

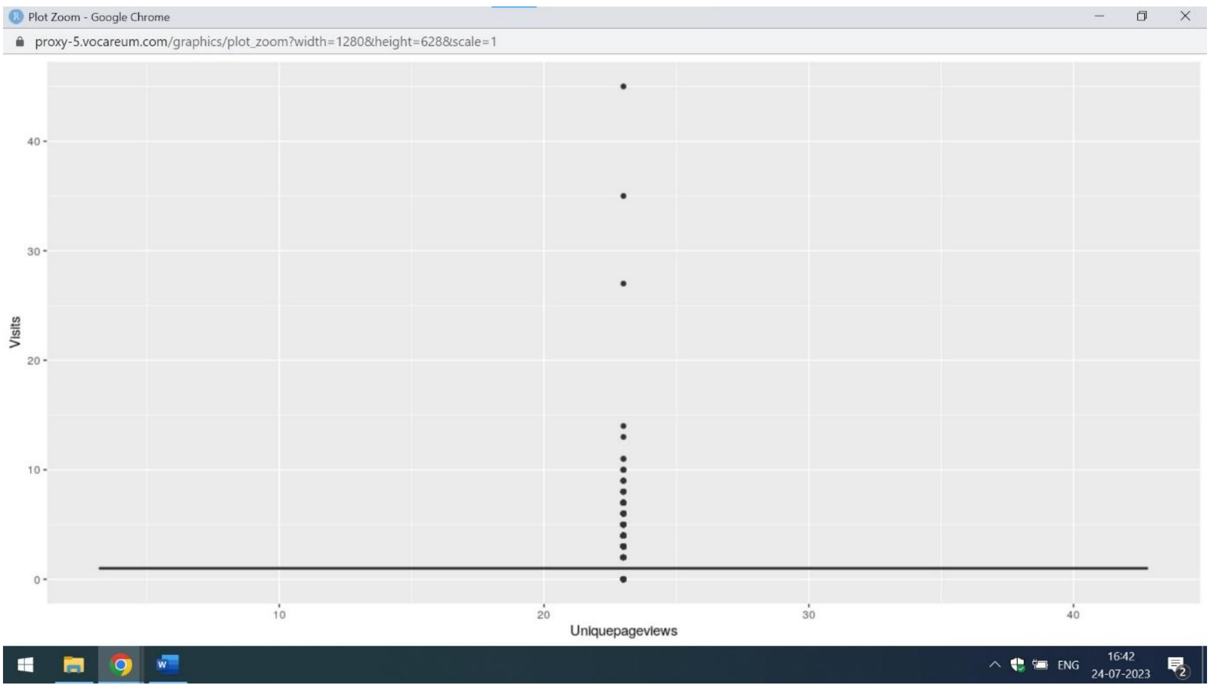


```
Console ~/
lm(formula = Uniquepageviews ~ Visits, data = web_data)

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 ***
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
```

The above mentioned are the residuals which is basically the difference between the dependent variable and predicted variable. Here the dependent variable is UniquepageViews and the independent variable is Visits. The maximum value of the residual is 14(approx.)

The below attached are the boxplot and ggplot view of the UniquepageView wrt to Visits





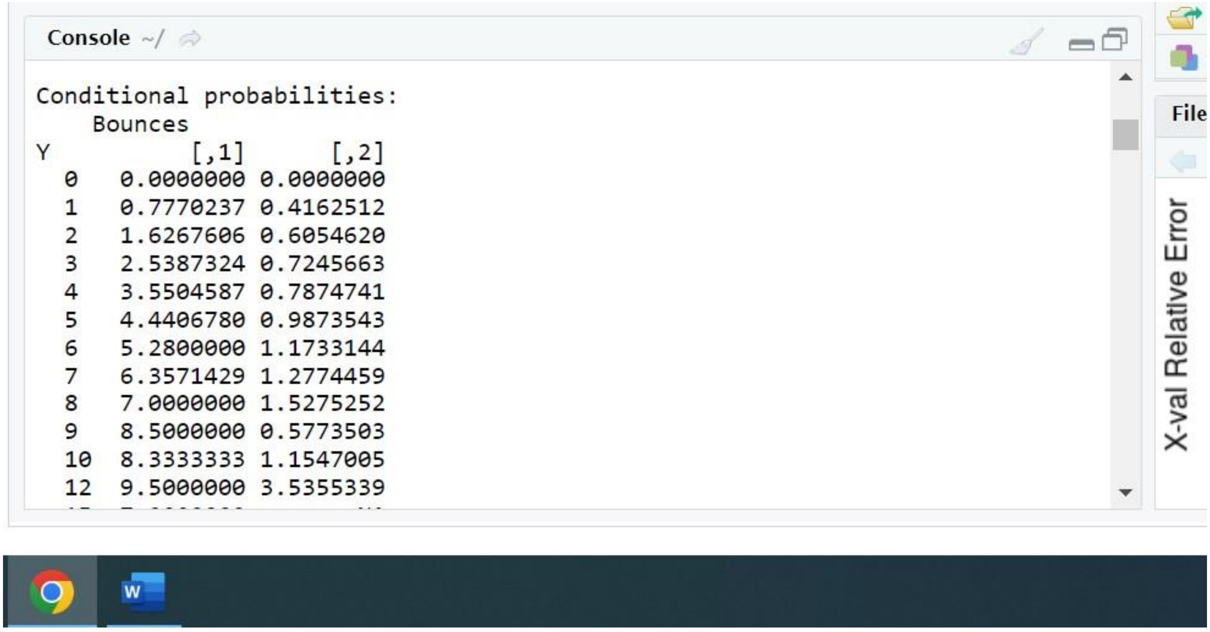
Q3) Find out the probable factors from the dataset, which could affect the exits. Exit Page Analysis is usually required to get an idea about why a user leaves the website for a session and moves on to another one. Please keep in mind that exits should not be confused with bounces.

Ans) The probable factors from the dataset which could affects the exits of a page are listed below. The below attached is the code in R

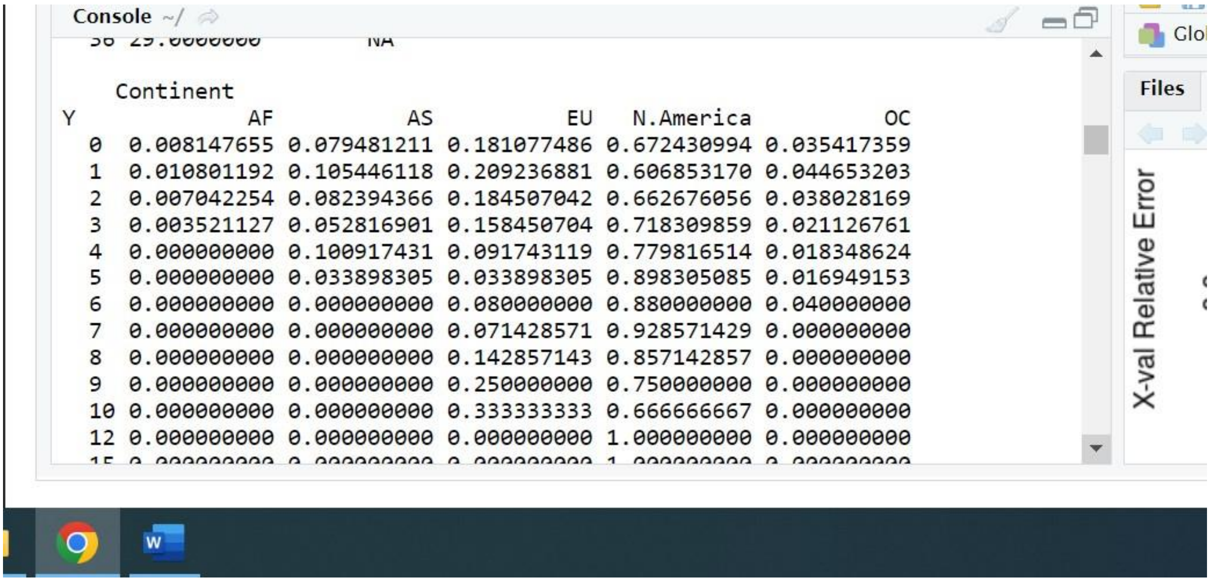
```
print("Web Data Analysis")

web_data<-
read.csv("https://raw.githubusercontent.com/shivanipriya89/WebData/main/Internet.csv")

web_data

View(web_data)

str(web_data)


web_data$exits<-sapply(web_data$Exits,factor)

# Build the model

naive_model<-naiveBayes(Exits~.,data=web_data)

print(naive_model) # Gives the probability


# Prediction

naive_predict<-predict(naive_model,web_data)

naive_predict


# Decision-Tree


naive_decision<-rpart(Exits~.,data=web_data,method="class")

naive_decision

printcp(naive_decision)

plotcp(naive_decision)

summary(naive_decision)
```

```
> print(naive_model) # Gives the probability

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
           0            1            2            3            4 
1.872995e-01 7.525616e-01 4.422436e-02 8.844872e-03 3.394687e-03 
           5            6            7            8            9 
1.837491e-03 7.785979e-04 4.360148e-04 2.180074e-04 1.245757e-04 
          10           12           15           27           33 
9.343175e-05 6.228783e-05 3.114392e-05 3.114392e-05 3.114392e-05 
          36
```

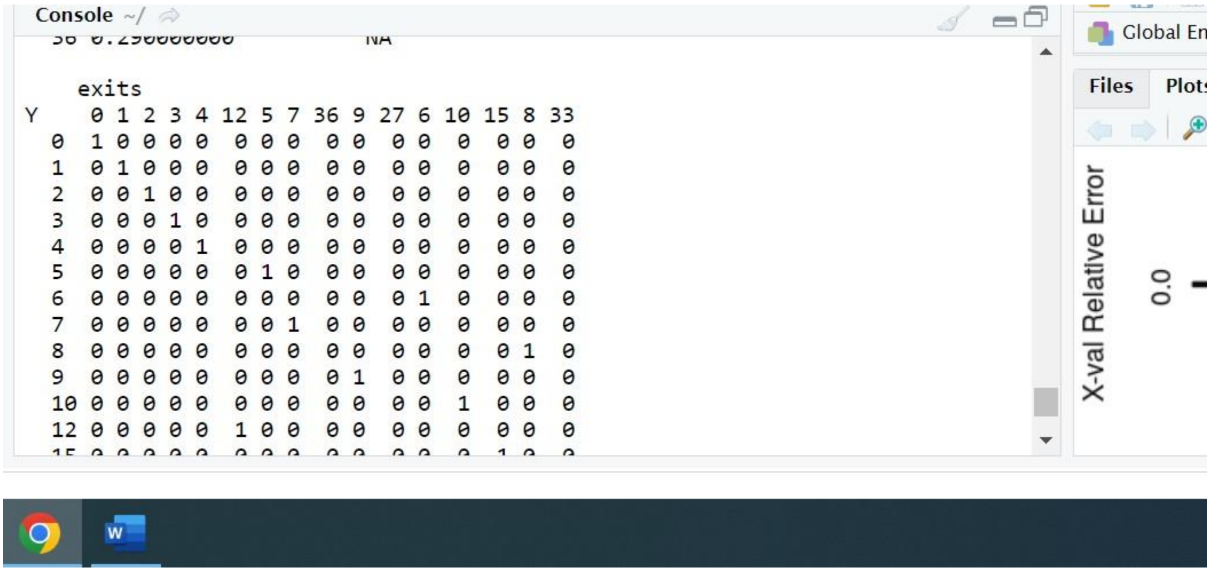
```
Conditional probabilities:
    Bounces
Y         [,1]      [,2]
  0  0.0000000 0.0000000
  1  0.7770237 0.4162512
  2  1.6267606 0.6054620
  3  2.5387324 0.7245663
  4  3.5504587 0.7874741
  5  4.4406780 0.9873543
  6  5.2800000 1.1733144
  7  6.3571429 1.2774459
  8  7.0000000 1.5275252
  9  8.5000000 0.5773503
  10 8.3333333 1.1547005
  12 9.5000000 3.5355339
  15 7.0000000        NA
```



```
Console ~/
Conditional probabilities:
    Bounces
Y          [,1]      [,2]
  0   0.0000000 0.0000000
  1   0.7770237 0.4162512
  2   1.6267606 0.6054620
  3   2.5387324 0.7245663
  4   3.5504587 0.7874741
  5   4.4406780 0.9873543
  6   5.2800000 1.1733144
  7   6.3571429 1.2774459
  8   7.0000000 1.5275252
  9   8.5000000 0.5773503
  10  8.3333333 1.1547005
  12  9.5000000 3.5355339
```

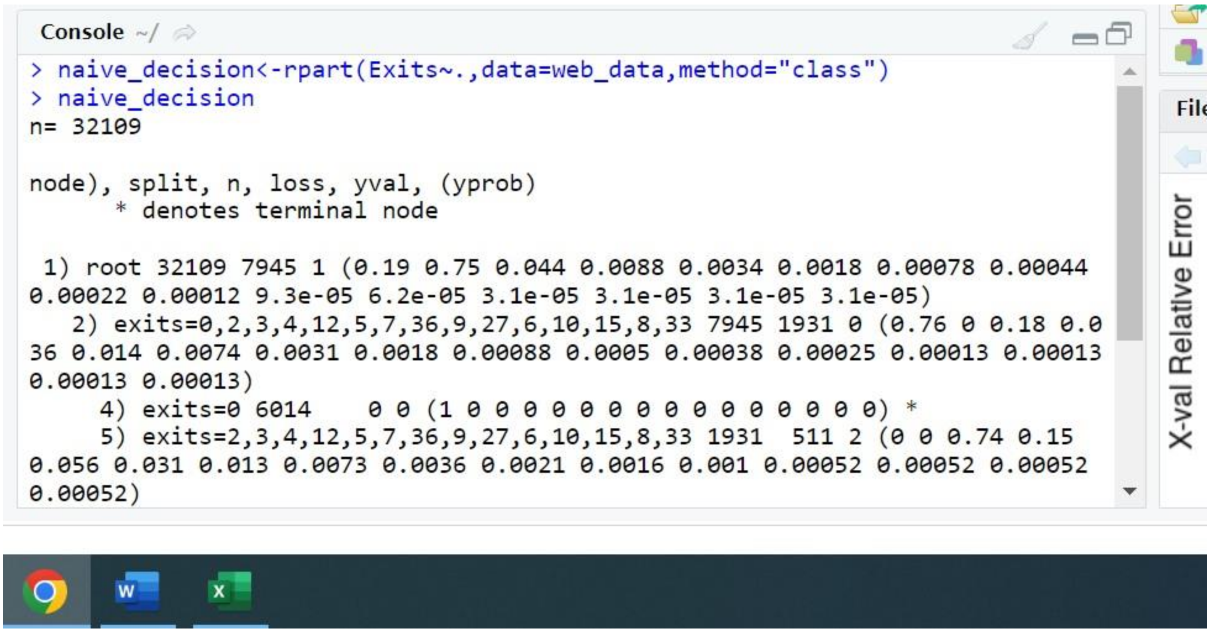


```
Console ~/

   Continent
Y            AF          AS          EU   N.America          OC
  0  0.008147655 0.079481211 0.181077486 0.672430994 0.035417359
  1  0.010801192 0.105446118 0.209236881 0.606853170 0.044653203
  2  0.007042254 0.082394366 0.184507042 0.662676056 0.038028169
  3  0.003521127 0.052816901 0.158450704 0.718309859 0.021126761
  4  0.000000000 0.100917431 0.091743119 0.779816514 0.018348624
  5  0.000000000 0.033898305 0.033898305 0.898305085 0.016949153
  6  0.000000000 0.000000000 0.080000000 0.880000000 0.040000000
  7  0.000000000 0.000000000 0.071428571 0.928571429 0.000000000
  8  0.000000000 0.000000000 0.142857143 0.857142857 0.000000000
  9  0.000000000 0.000000000 0.250000000 0.750000000 0.000000000
  10 0.000000000 0.000000000 0.333333333 0.666666667 0.000000000
  12 0.000000000 0.000000000 0.000000000 1.000000000 0.000000000
```



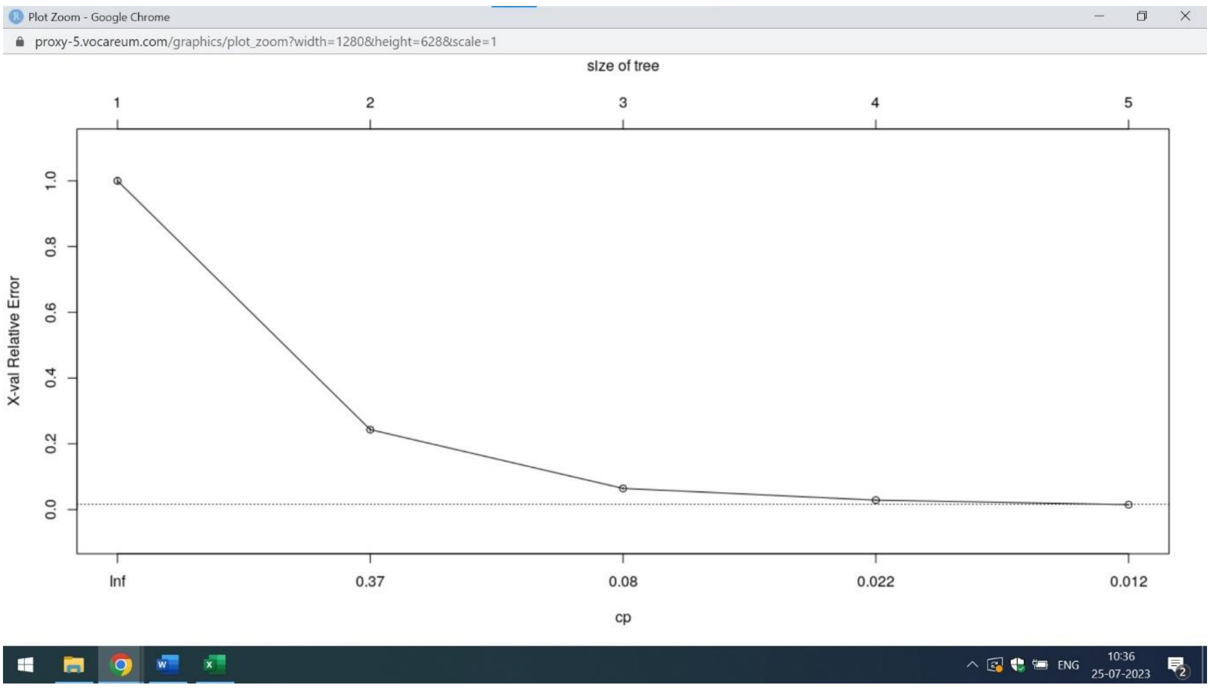
It is clear from the above mention screenshots,that Naïve Bayes Algorithm gives an overview of Apriori probabilities and conditional probabilities of the various factors of exits. The factors affecting the exits of the page are the Bounces,Continent,SourceGroup,TimeinPage and UniquePage Views

From the decision tree algorithm,one can also find the yval,probability,loss and split ends of the exits attribute which is on the Y-intercept



```
Console ~/
> printcp(naive_decision)

Classification tree:
rpart(formula = Exits ~ ., data = web_data, method = "class")

Variables actually used in tree construction:
[1] exits

Root node error: 7945/32109 = 0.24744

n= 32109

        CP nsplit rel error   xerror      xstd
1 0.756954      0  1.000000 1.000000 0.0097325
2 0.178729      1  0.243046 0.243046 0.0053620
3 0.035746      2  0.064317 0.064317 0.0028225
```

The above mentioned are the root node error of the exits variable

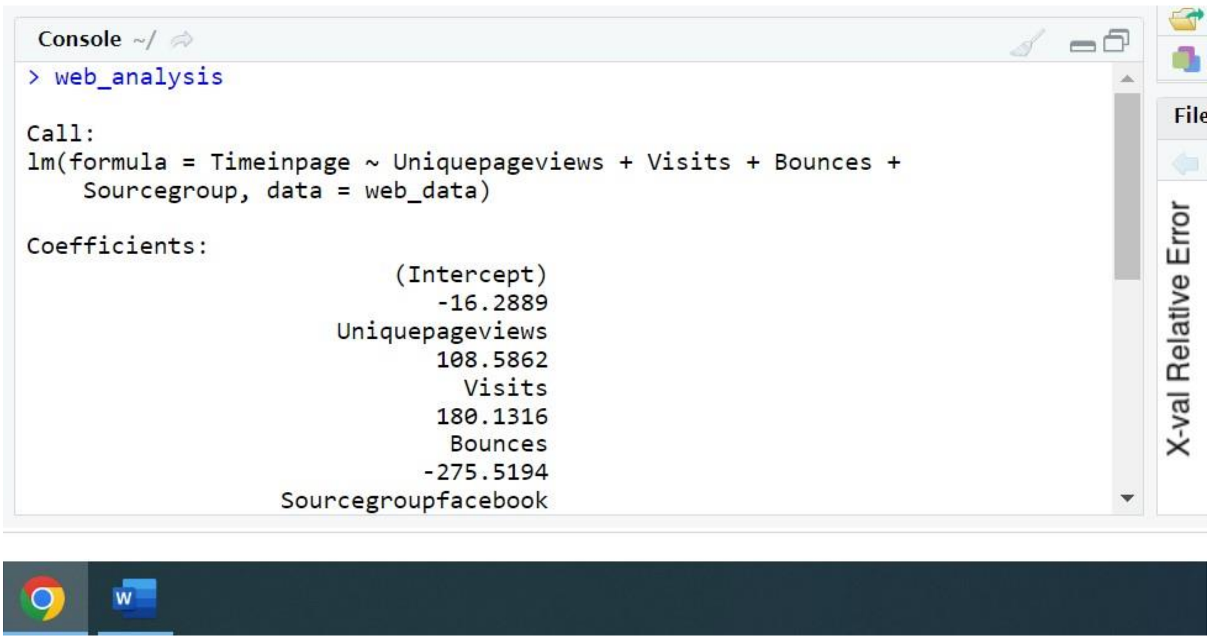


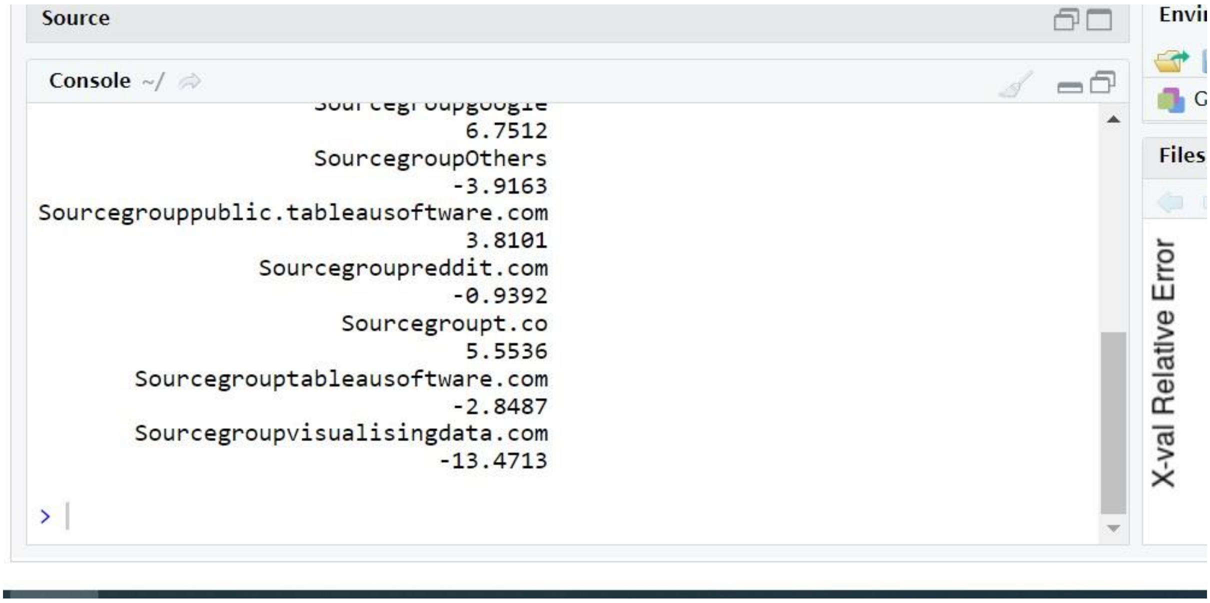
The above mention plot represents that relative error decreases when the size of tree increases

Q4) 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.

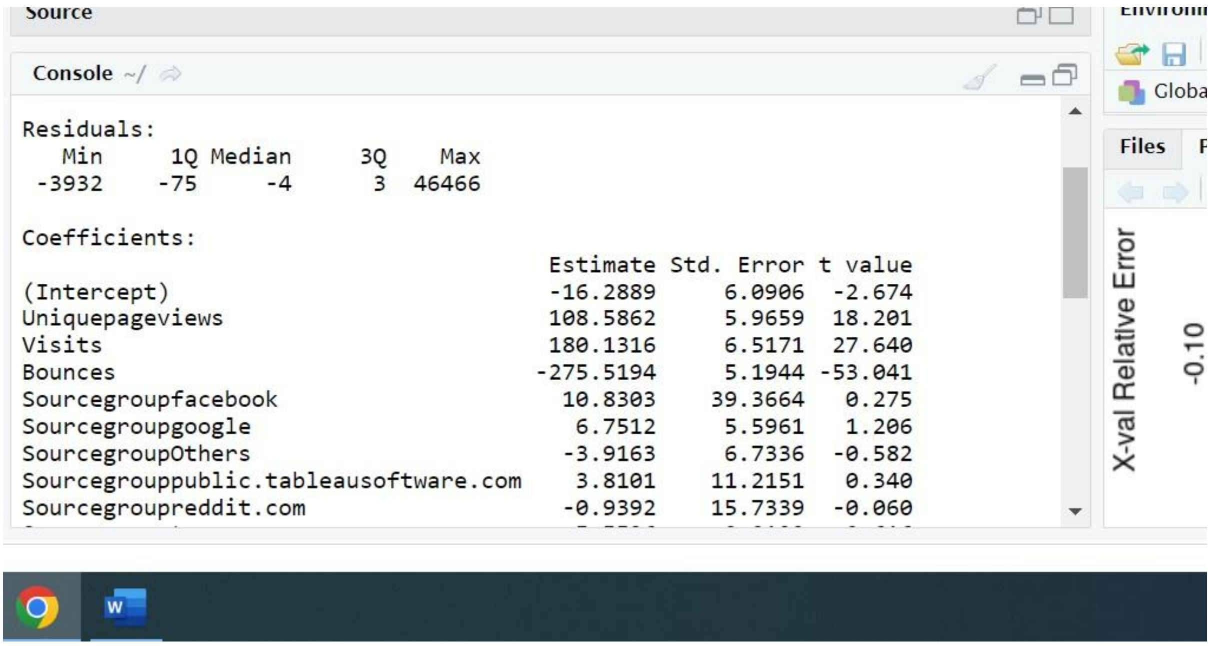
Ans) Through Regression Analysis with Multiple Variables,one can analyse the variables having an effect on time on page. The below attached is the code in R which analyses the various factors effecting on time on page

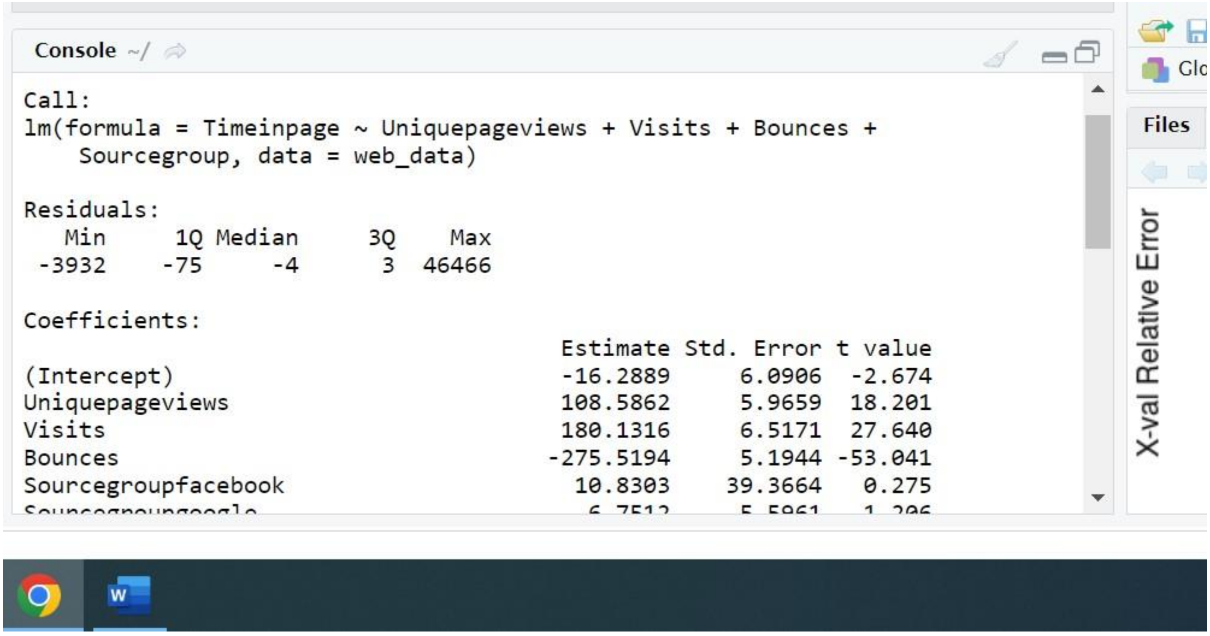
```
print("Web Data Analysis")

web_data<-
read.csv("https://raw.githubusercontent.com/shivanipriya89/WebData/main/Internet.csv")

web_data

View(web_data)

web_data$Timeinpage<-as.integer(web_data$Timeinpage)

str(web_data)

web_analysis<-
lm(formula=Timeinpage~Uniquepageviews+Visits+Bounces+Sourcegroup,data=web_data)

web_analysis

summary(web_analysis)
```

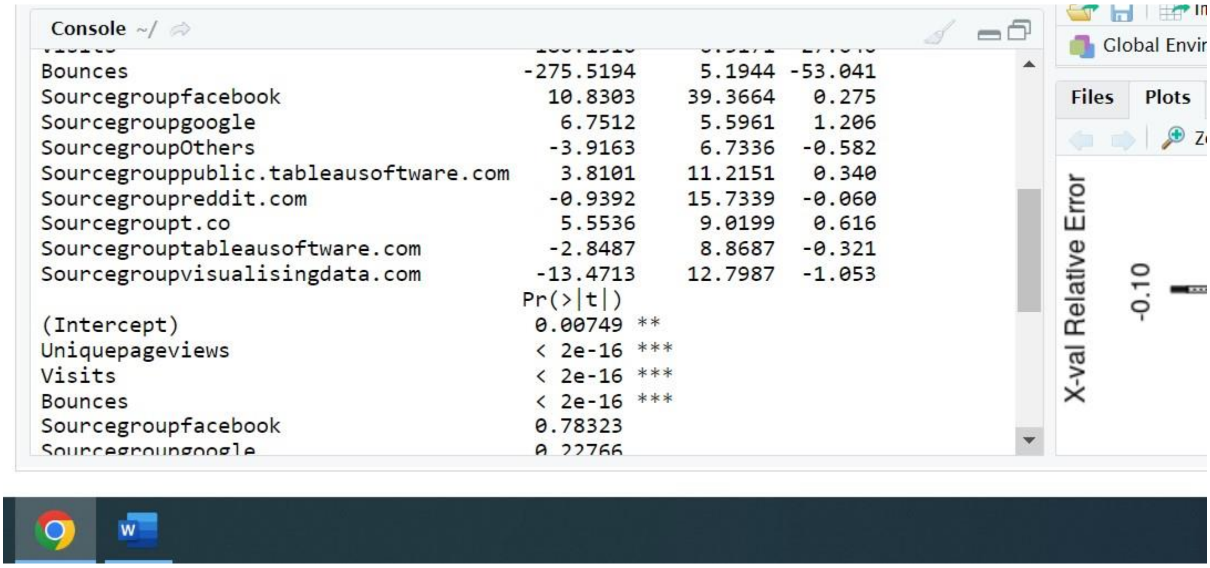




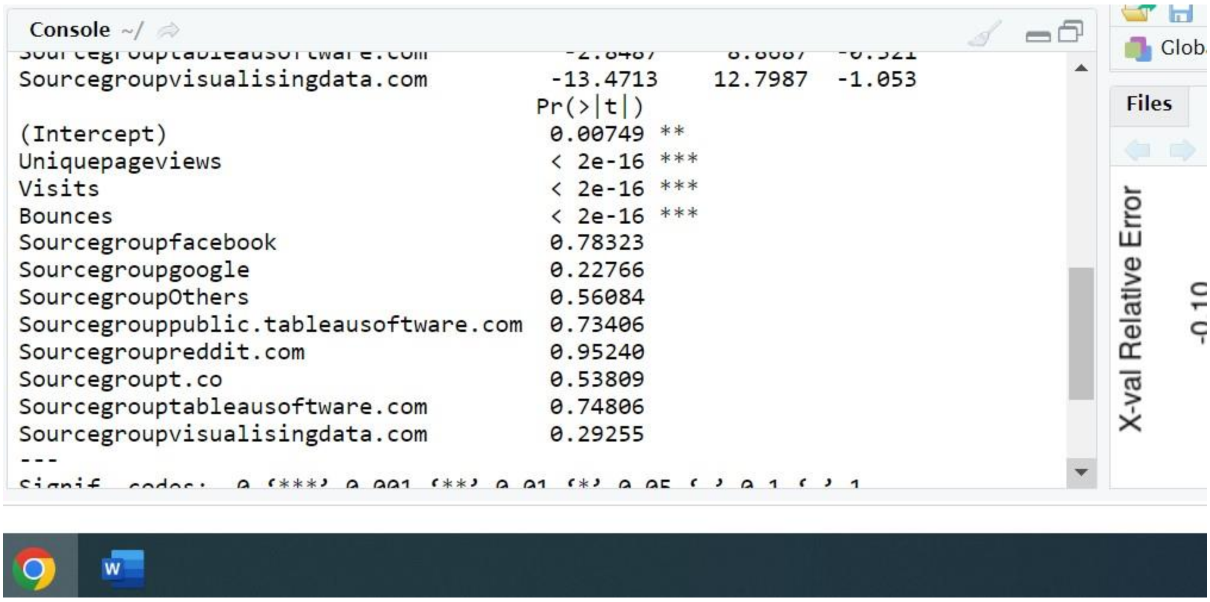
From the above mentioned screenshots,it is clear that Timeinpage is highly dependent upon UniquePage Views and Visits.



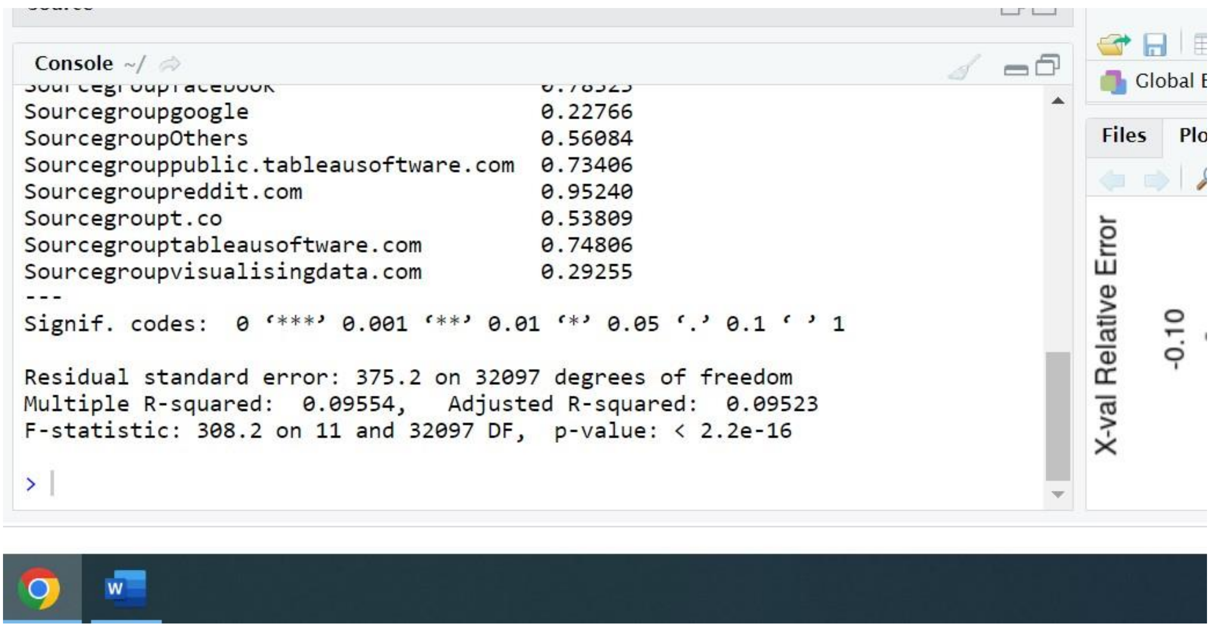




It is clear from the above mentioned screenshots that Unique Page Views,Visits,Sourcegroupfacebook,Sourcegroupgoogle ,Sourcegrouppublic.tableausoftware.com and sourcegroupt.co are the major variables which possibly have an effect on the time on page.



```
Sourcegroupvisualisingdata.com          -13.4713     12.7987   -1.053
                                       Pr(>|t|)
(Intercept)                            0.00749 **
Uniquepageviews                        < 2e-16 ***
Visits                                 < 2e-16 ***
Bounces                                < 2e-16 ***
Sourcegroupfacebook                    0.78323
Sourcegroupgoogle                      0.22766
SourcegroupOthers                      0.56084
Sourcegrouppublic.tableausoftware.com  0.73406
Sourcegroupreddit.com                  0.95240
Sourcegroupt.co                        0.53809
Sourcegrouptableausoftware.com         0.74806
Sourcegroupvisualisingdata.com         0.29255
---
```



```
Sourcegroupgoogle                      0.22766
SourcegroupOthers                      0.56084
Sourcegrouppublic.tableausoftware.com  0.73406
Sourcegroupreddit.com                  0.95240
Sourcegroupt.co                        0.53809
Sourcegrouptableausoftware.com         0.74806
Sourcegroupvisualisingdata.com         0.29255
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 375.2 on 32097 degrees of freedom
Multiple R-squared:  0.09554,    Adjusted R-squared:  0.09523
F-statistic: 308.2 on 11 and 32097 DF,  p-value: < 2.2e-16

>
```

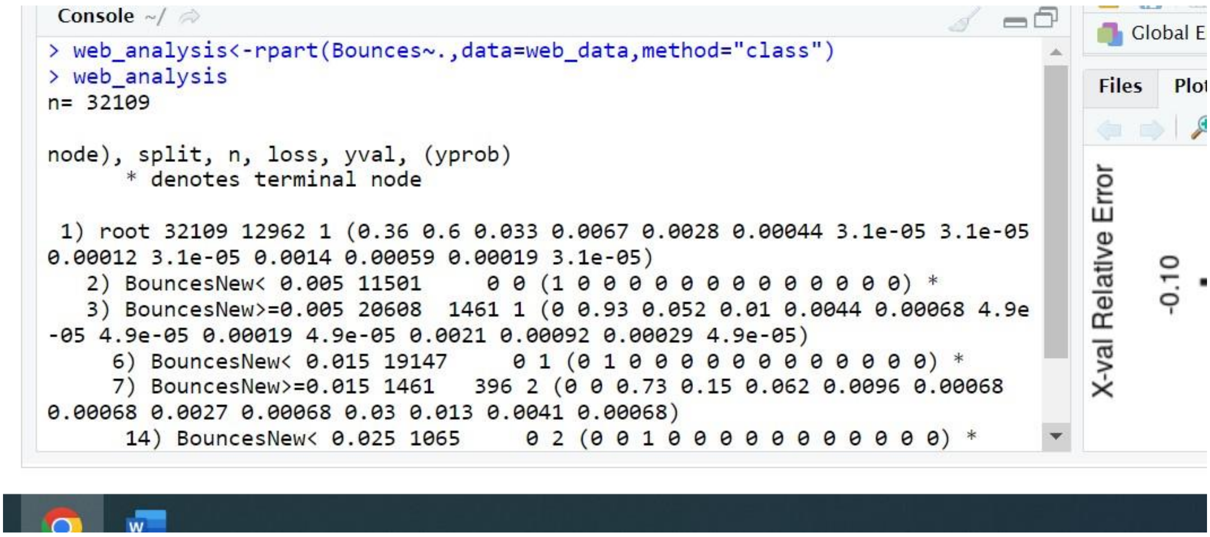
Q5) A high bounce rate is a cause of alarm for websites which depend on visitor engagement. Help the team in determining the factors that are impacting the bounce.

Ans) Decision Tree Algorithm helps in determining the factors that are impacting the bounce.The below attached is the code in R which analyses the various factors of Bounces attribute of the website

print("Web Data Analysis")

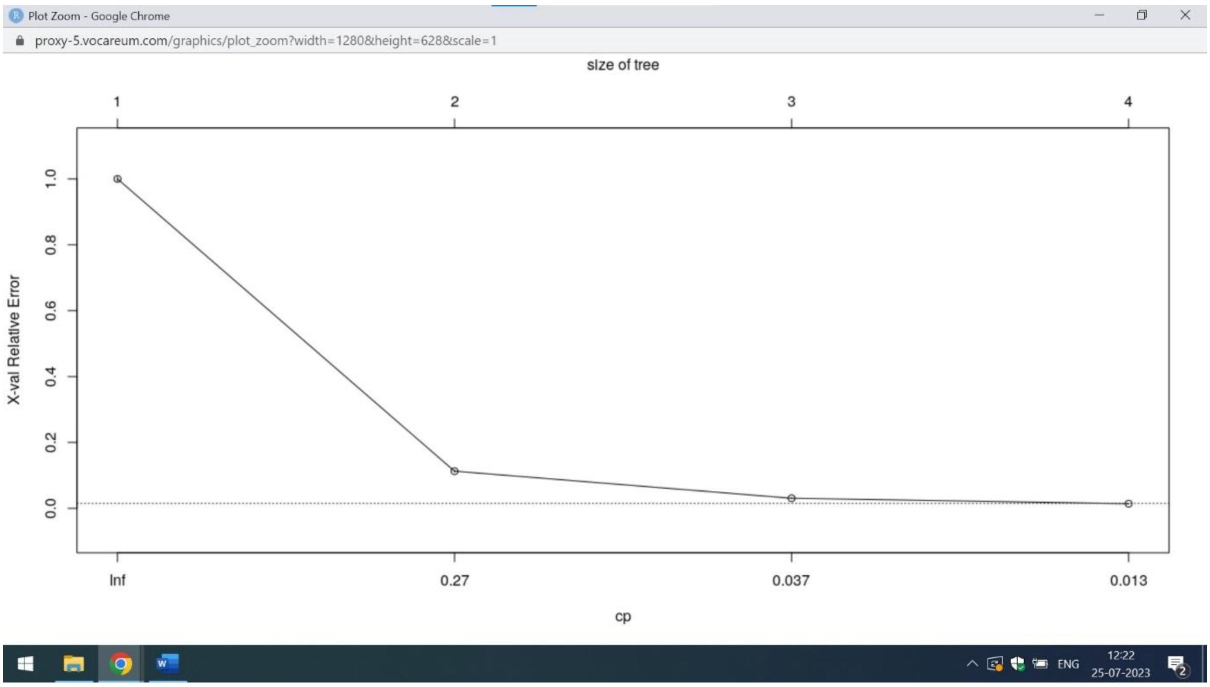
web_data<-
read.csv("https://raw.githubusercontent.com/shivanipriya89/WebData/main/Internet.csv")

```
web_data
View(web_data)
web_data$Bounces<-sapply(web_data$Bounces,factor)
str(web_data)

web_analysis<-rpart(Bounces~.,data=web_data,method="class")
web_analysis
printcp(web_analysis)
plotcp(web_analysis)
summary(web_analysis)
plot(web_analysis)
```

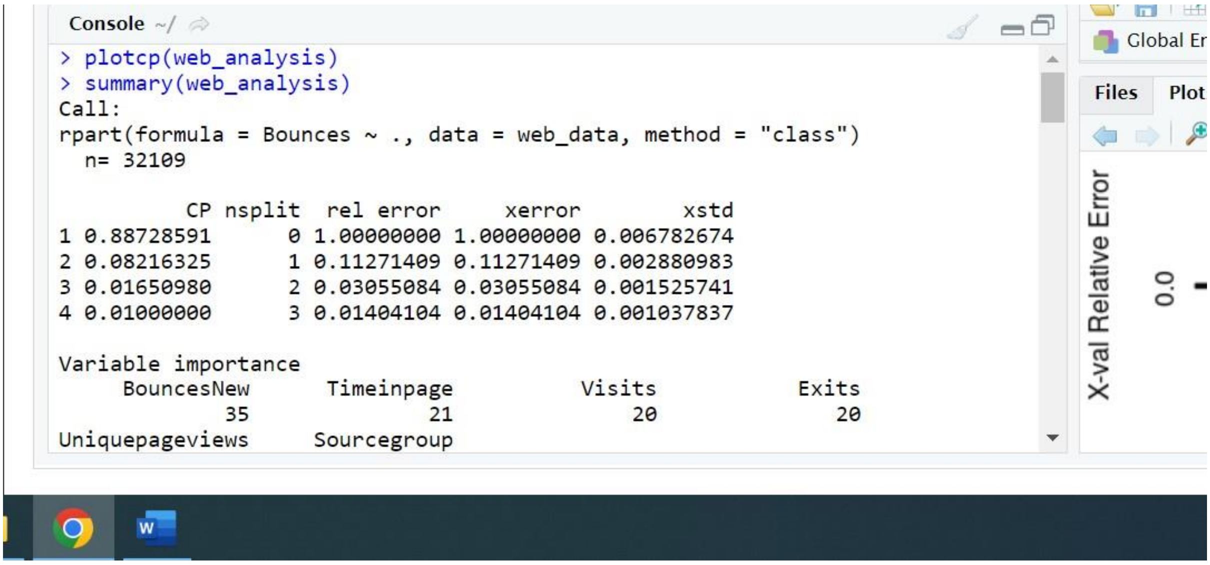


```
> web_analysis<-rpart(Bounces~.,data=web_data,method="class")
> web_analysis
n= 32109

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 32109 12962 1 (0.36 0.6 0.033 0.0067 0.0028 0.00044 3.1e-05 3.1e-05
0.00012 3.1e-05 0.0014 0.00059 0.00019 3.1e-05)
   2) BouncesNew< 0.005 11501     0 0 (1 0 0 0 0 0 0 0 0 0 0 0 0 0) *
   3) BouncesNew>=0.005 20608  1461 1 (0 0.93 0.052 0.01 0.0044 0.00068 4.9e
-05 4.9e-05 0.00019 4.9e-05 0.0021 0.00092 0.00029 4.9e-05)
     6) BouncesNew< 0.015 19147     0 1 (0 1 0 0 0 0 0 0 0 0 0 0 0 0) *
     7) BouncesNew>=0.015 1461   396 2 (0 0 0.73 0.15 0.062 0.0096 0.00068
0.00068 0.0027 0.00068 0.03 0.013 0.0041 0.00068)
      14) BouncesNew< 0.025 1065     0 2 (0 0 1 0 0 0 0 0 0 0 0 0 0 0) *
```

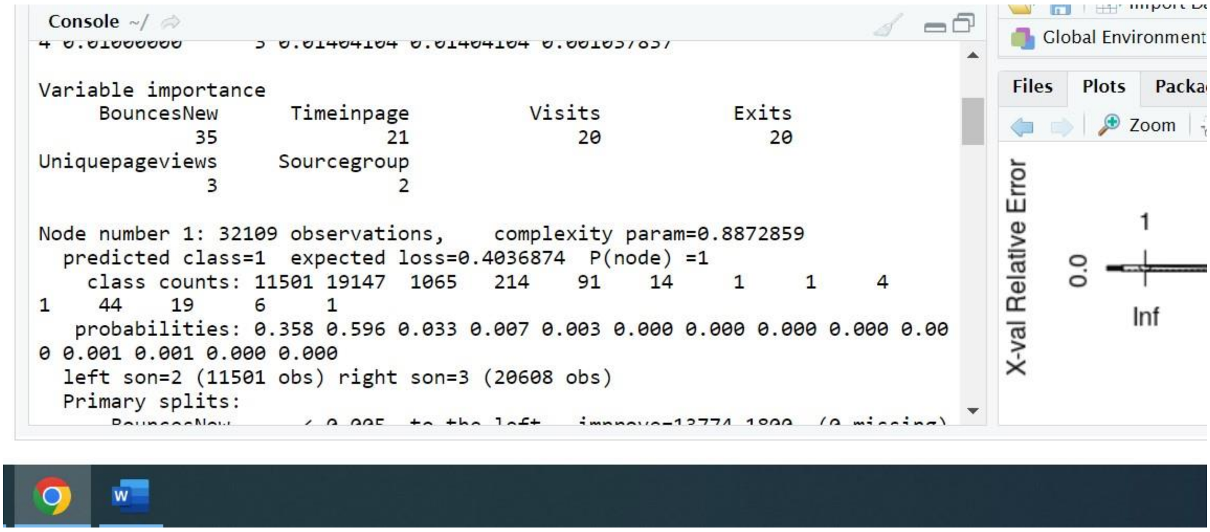
```
> printcp(web_analysis)

Classification tree:
rpart(formula = Bounces ~ ., data = web_data, method = "class")

Variables actually used in tree construction:
[1] BouncesNew

Root node error: 12962/32109 = 0.40369

n= 32109

        CP nsplit rel error   xerror      xstd
1 0.887286      0  1.000000 1.000000 0.0067827
2 0.082163      1  0.112714 0.112714 0.0028810
3 0.016510      2  0.030551 0.030551 0.0015257
```



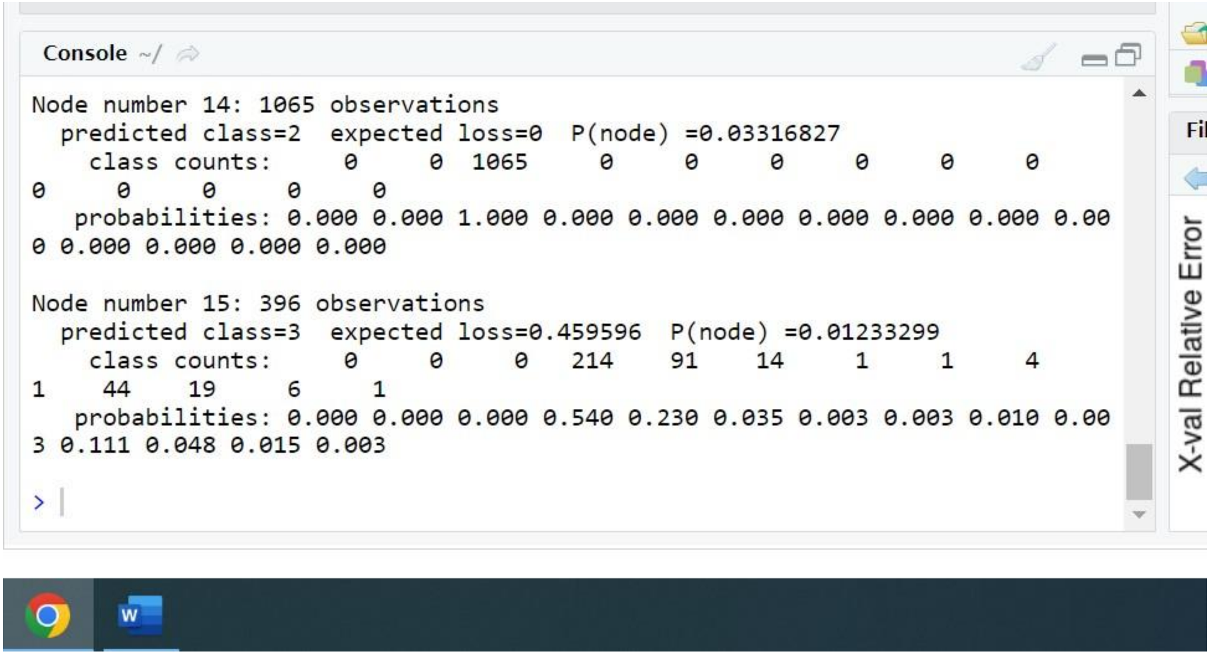
This tree represents that when the size of tree increases the relative error decreases




```
> plotcp(web_analysis)
> summary(web_analysis)
Call:
rpart(formula = Bounces ~ ., data = web_data, method = "class")
  n= 32109

          CP nsplit  rel error     xerror        xstd
1 0.88728591      0 1.00000000 1.00000000 0.006782674
2 0.08216325      1 0.11271409 0.11271409 0.002880983
3 0.01650980      2 0.03055084 0.03055084 0.001525741
4 0.01000000      3 0.01404104 0.01404104 0.001037837

Variable importance
     BouncesNew      Timeinpage           Visits           Exits
             35              21               20              20
Uniquepageviews     Sourcegroup
```




```
4 0.01000000      3 0.01404104 0.01404104 0.001037837

Variable importance
     BouncesNew      Timeinpage           Visits           Exits
             35              21               20              20
Uniquepageviews     Sourcegroup
              3               2

Node number 1: 32109 observations,    complexity param=0.8872859
  predicted class=1  expected loss=0.4036874  P(node) =1
    class counts: 11501 19147  1065   214    91    14     1     1     4
1    44    19     6     1
   probabilities: 0.358 0.596 0.033 0.007 0.003 0.000 0.000 0.000 0.000 0.00
0 0.001 0.001 0.000 0.000
  left son=2 (11501 obs) right son=3 (20608 obs)
  Primary splits:
      BouncesNew      < 0.005    to the left,  improve=13774.1800 (0 missing)
```



```
Node number 14: 1065 observations
  predicted class=2  expected loss=0  P(node) =0.03316827
    class counts:     0     0  1065     0     0     0     0     0     0
0     0     0     0     0
   probabilities: 0.000 0.000 1.000 0.000 0.000 0.000 0.000 0.000 0.000 0.00
0 0.000 0.000 0.000 0.000

Node number 15: 396 observations
  predicted class=3  expected loss=0.459596  P(node) =0.01233299
    class counts:     0     0     0   214    91    14     1     1     4
1    44    19     6     1
   probabilities: 0.000 0.000 0.000 0.540 0.230 0.035 0.003 0.003 0.010 0.00
3 0.111 0.048 0.015 0.003

> 
```


From the above mentioned screenshots,it is clear that BouncesNew,TimeinPage,Visits,Exits,uniquepageviews and Sourcegroups are the major factors that are impacting the bounce. Out of these factors BouncesNew are the variable with high importance