**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")

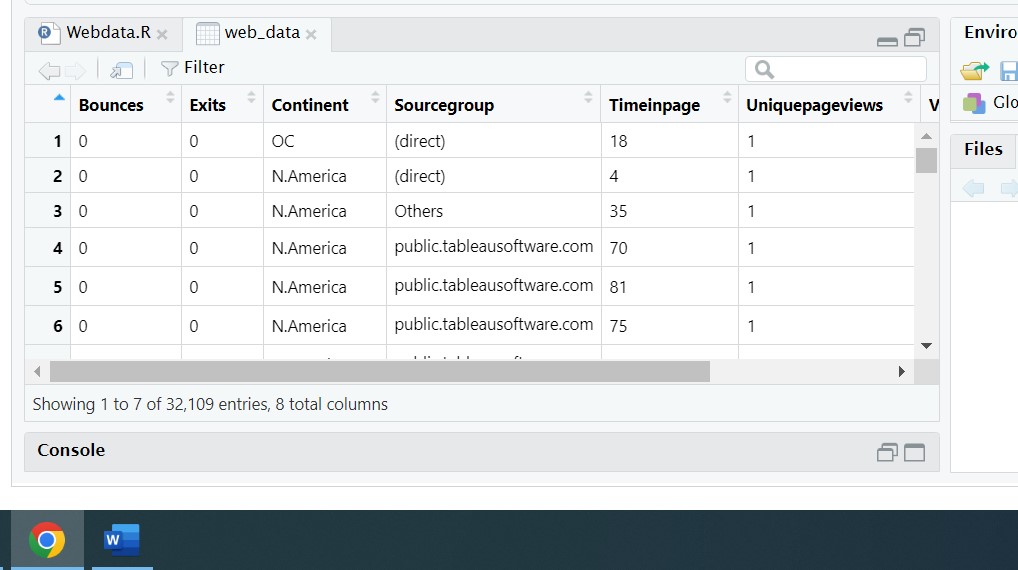
web\_data

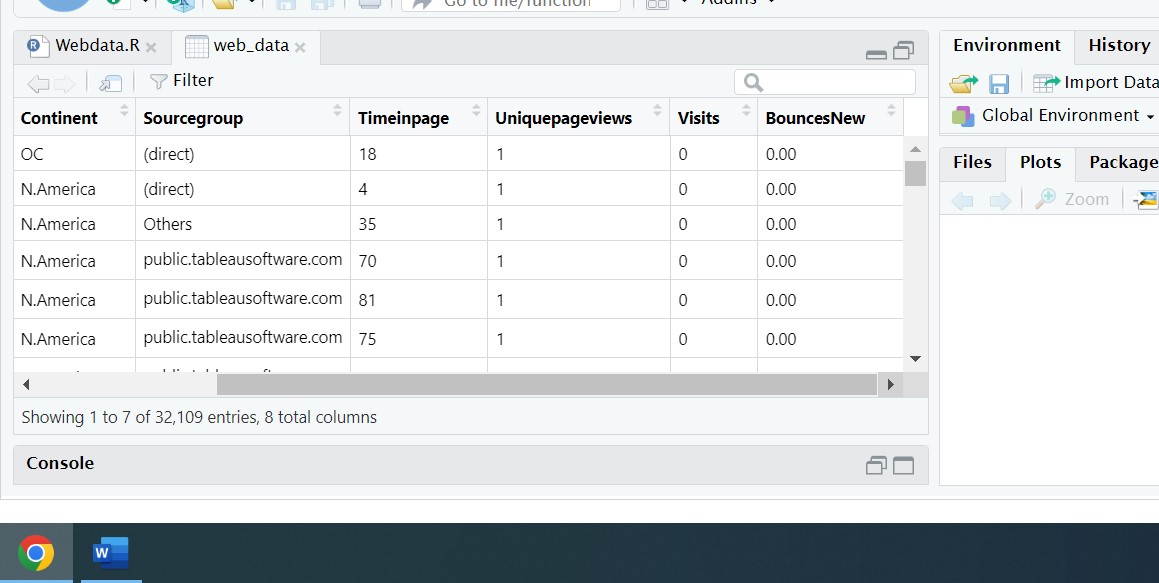
View(web\_data)

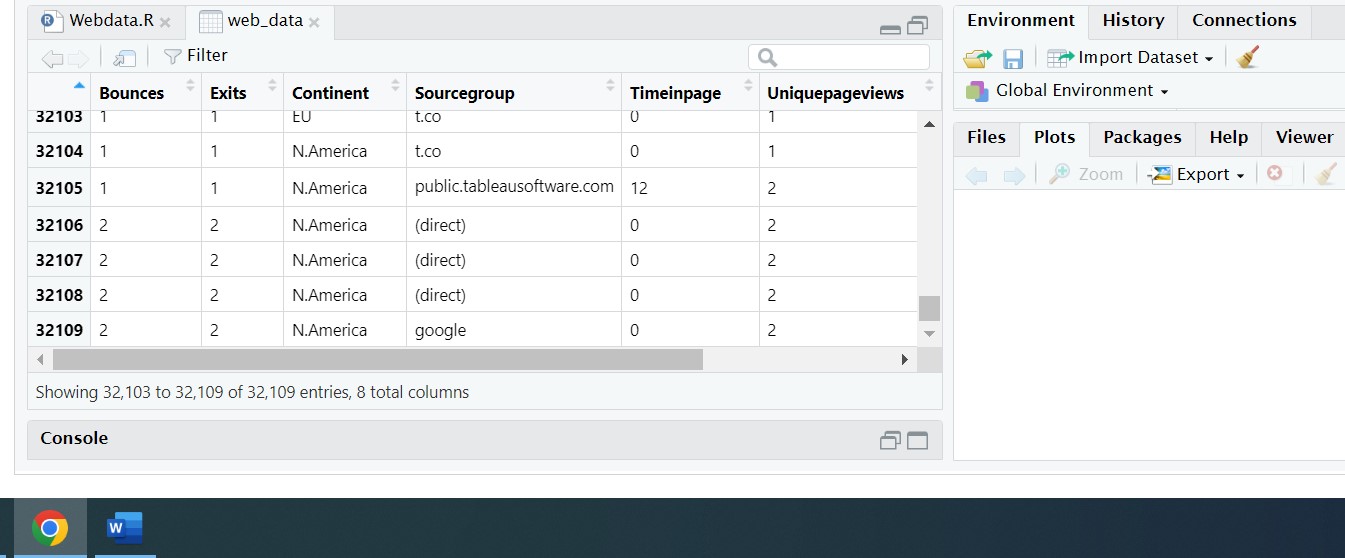
str(web\_data)

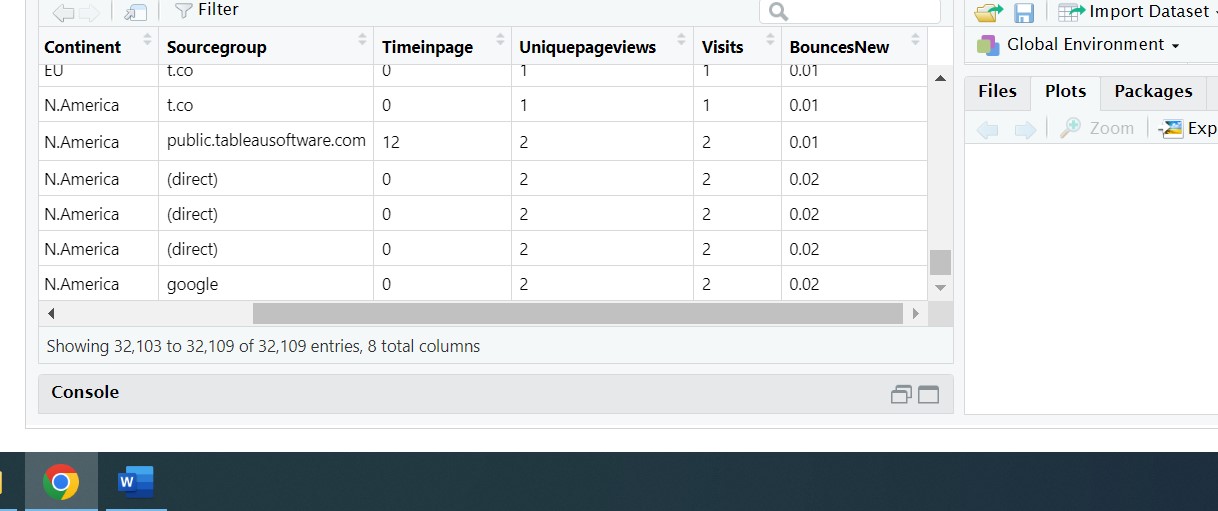
summary(web\_data)

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



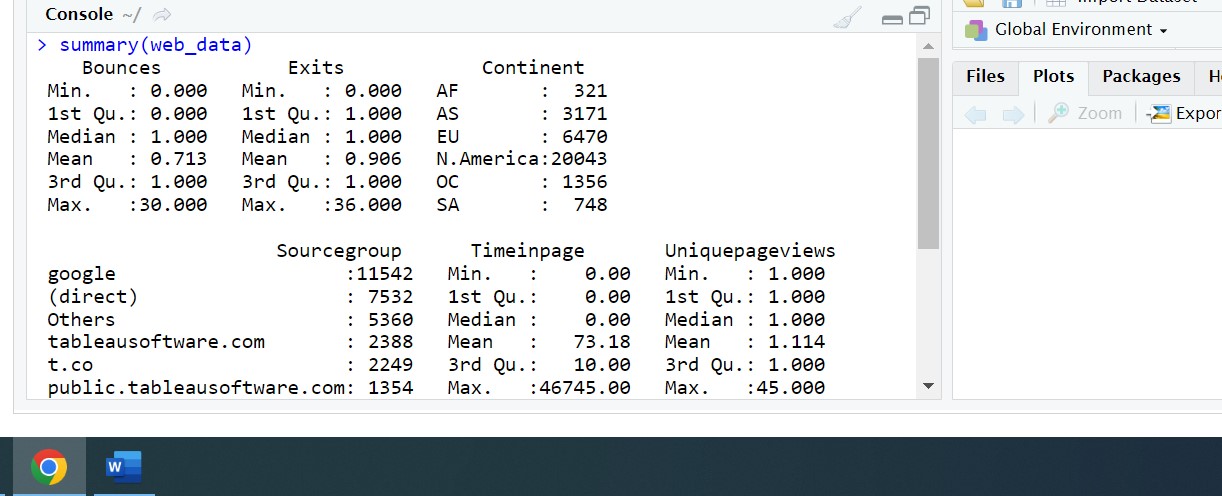


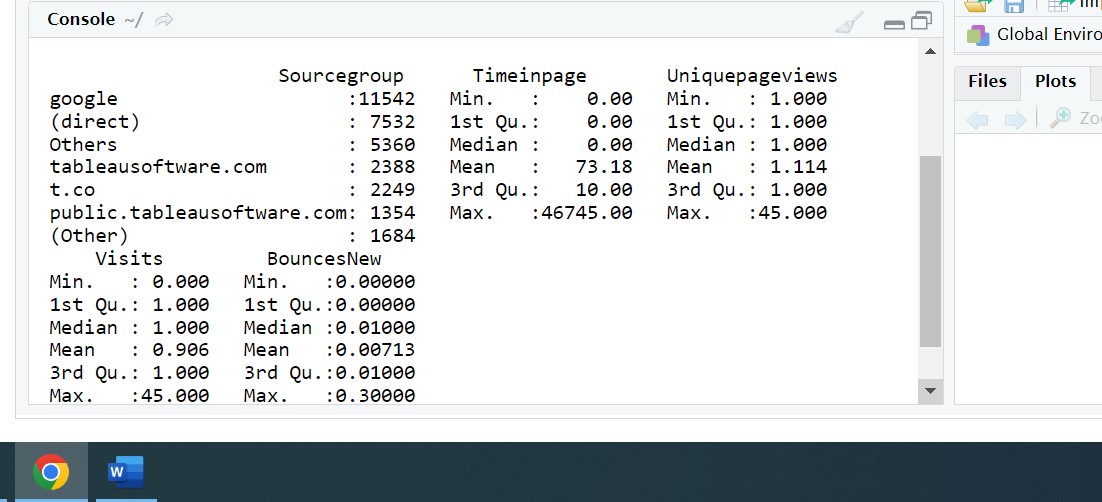




The above mention table of web data analysis has 32,109 entries of all 8 columns of web data set

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



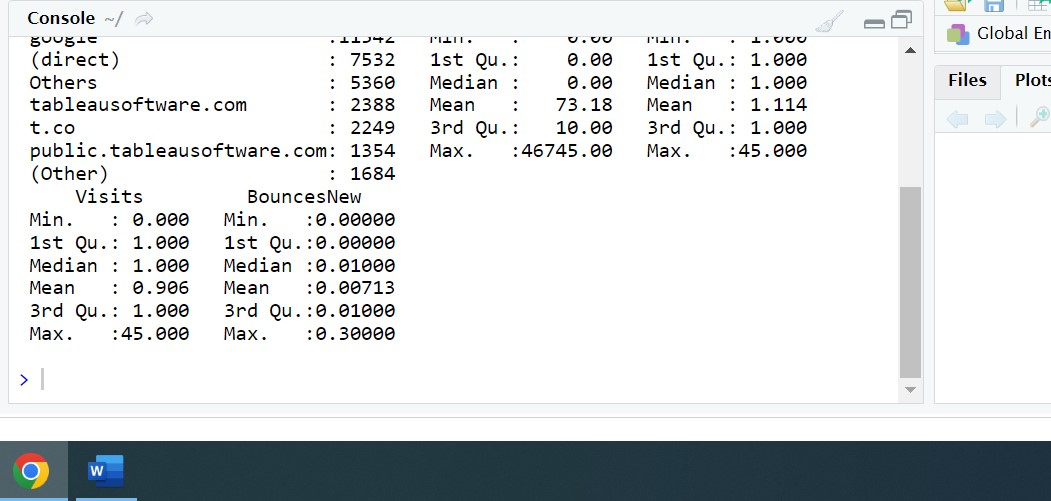


The internet dataset is an excel file with(.xlsx) extension. I have converted this file to internet.csv

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)) +

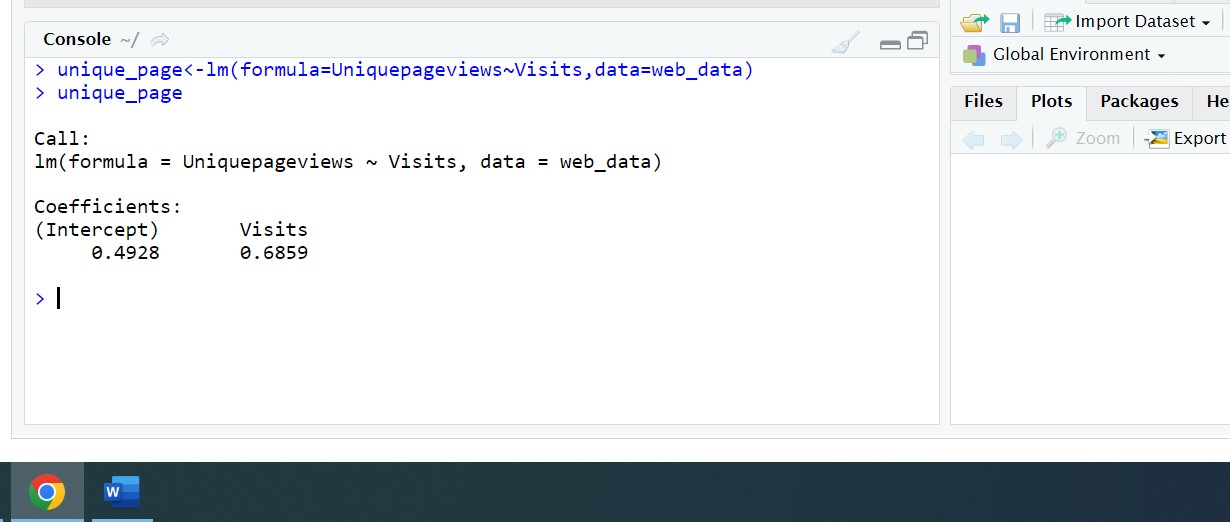
geom\_boxplot()

uds<-table(web\_data$Uniquepageviews,web\_data$Visits)

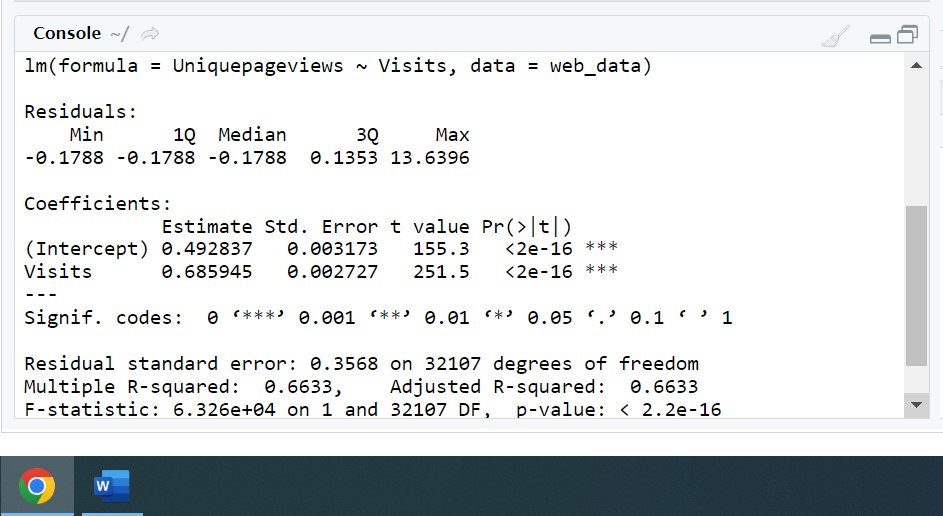
uds

View(uds)

The below attached are the screenshots

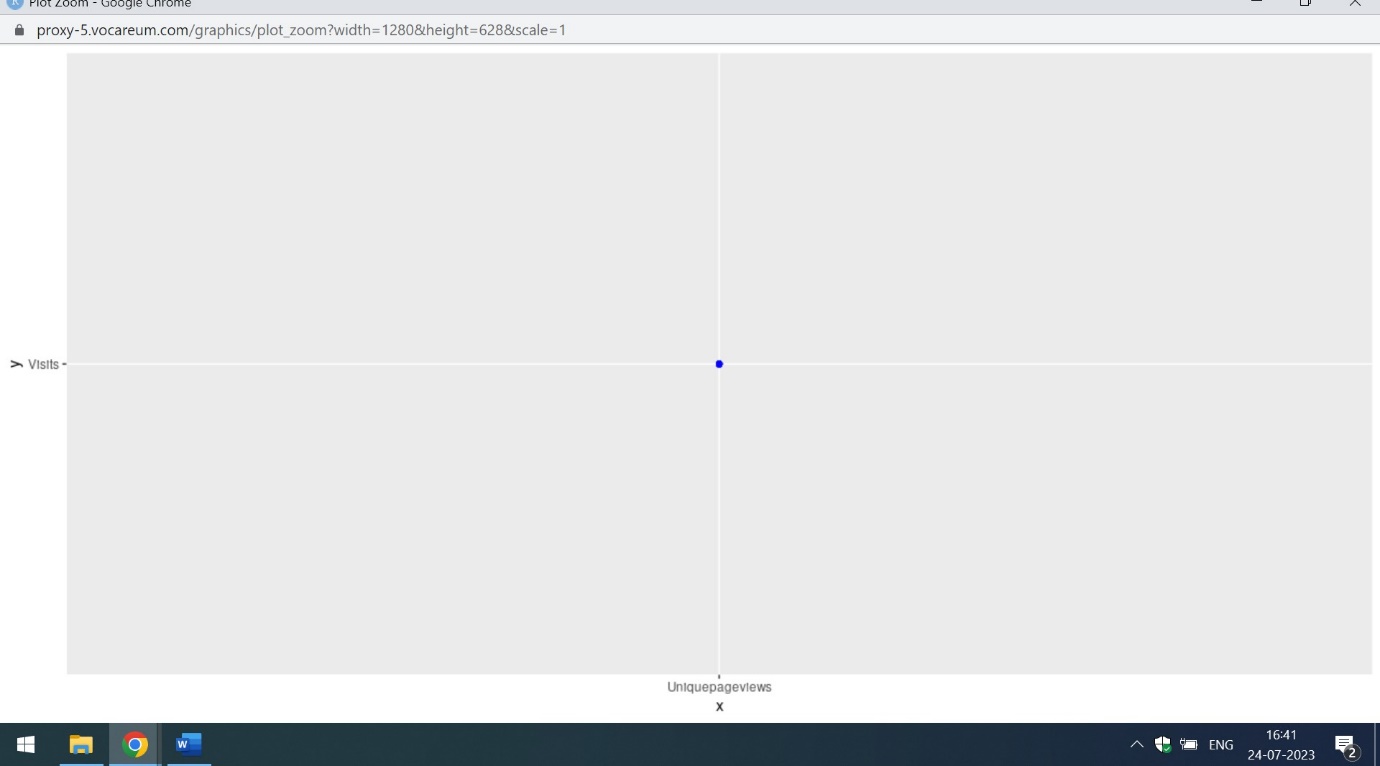


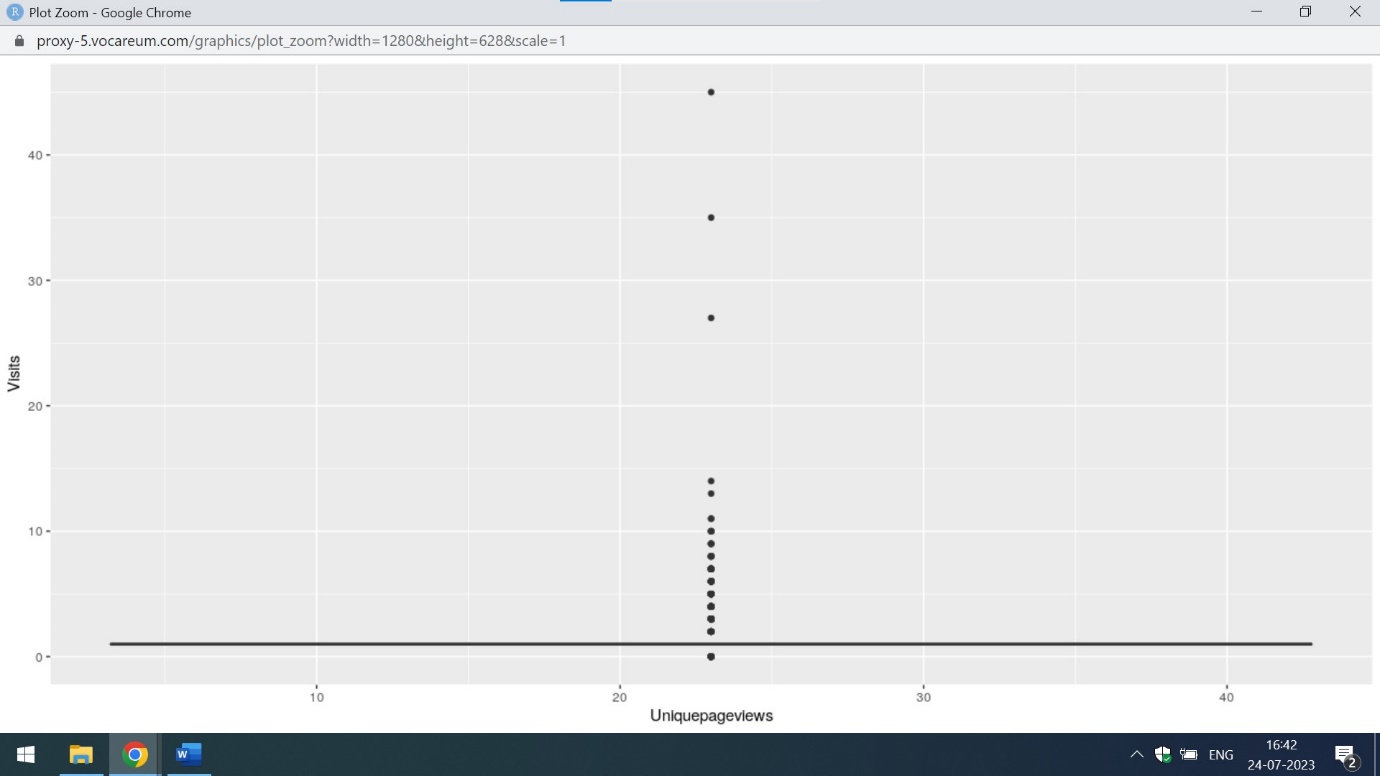
It is clear from the above mention screenshot that there is a positive Linear Regression between UniquepageViews and Visits as the value on the Y intercept is positive



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

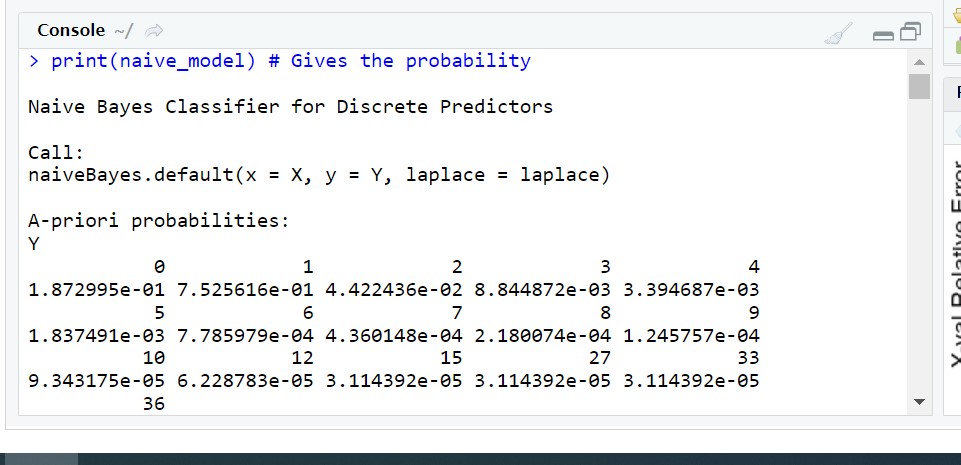
naive\_decision<-rpart(Exits~.,data=web\_data,method="class")

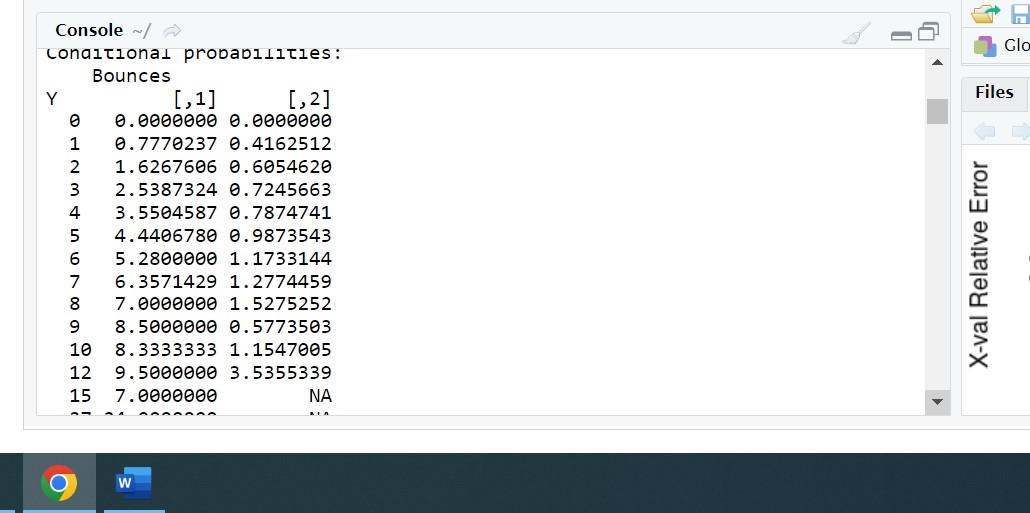
naive\_decision

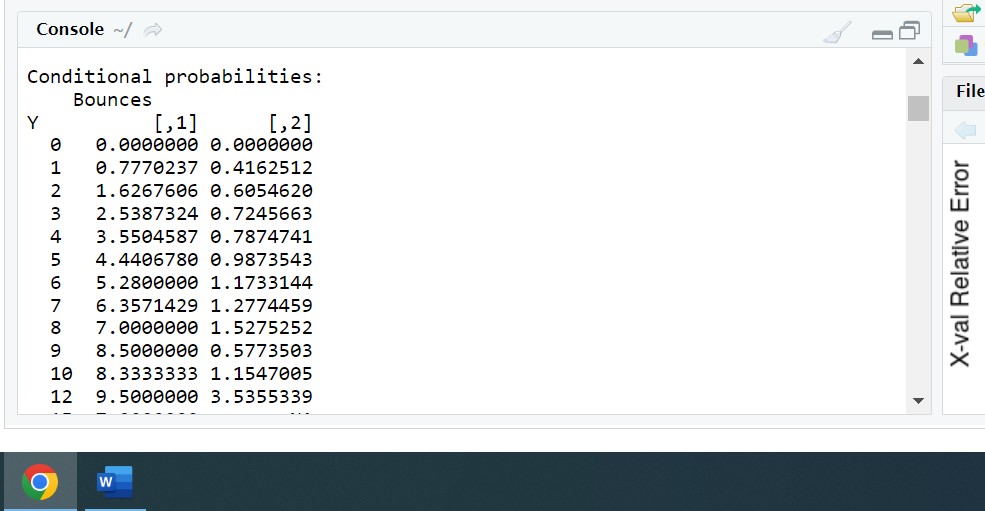
printcp(naive\_decision)

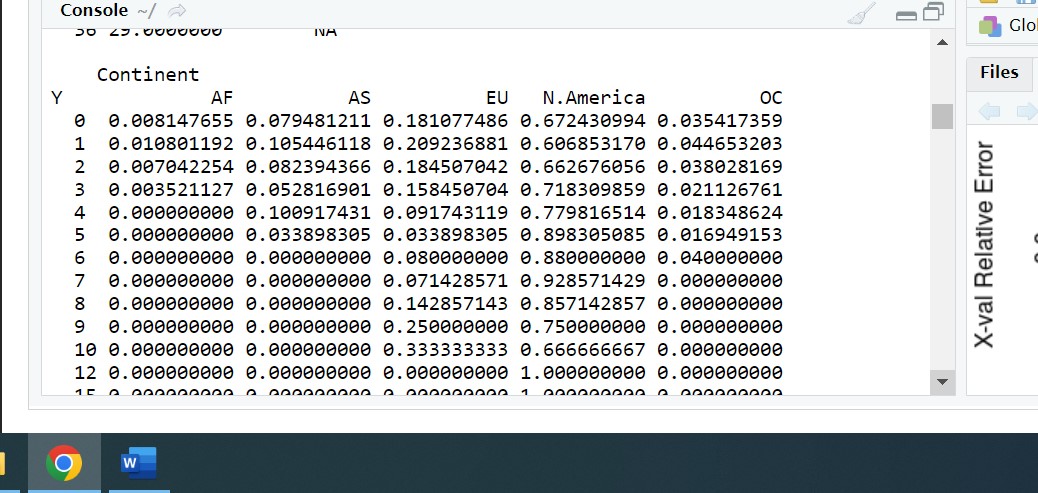
plotcp(naive\_decision)

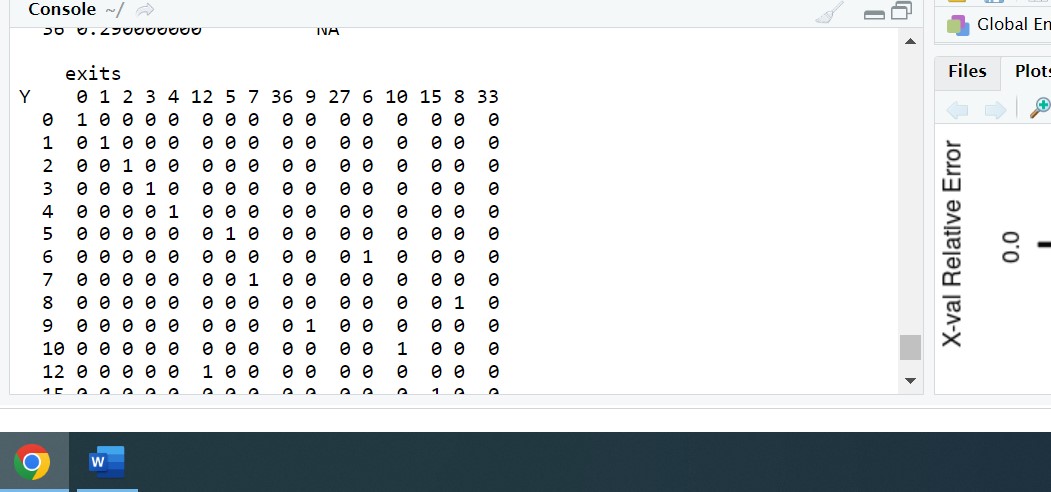
summary(naive\_decision)





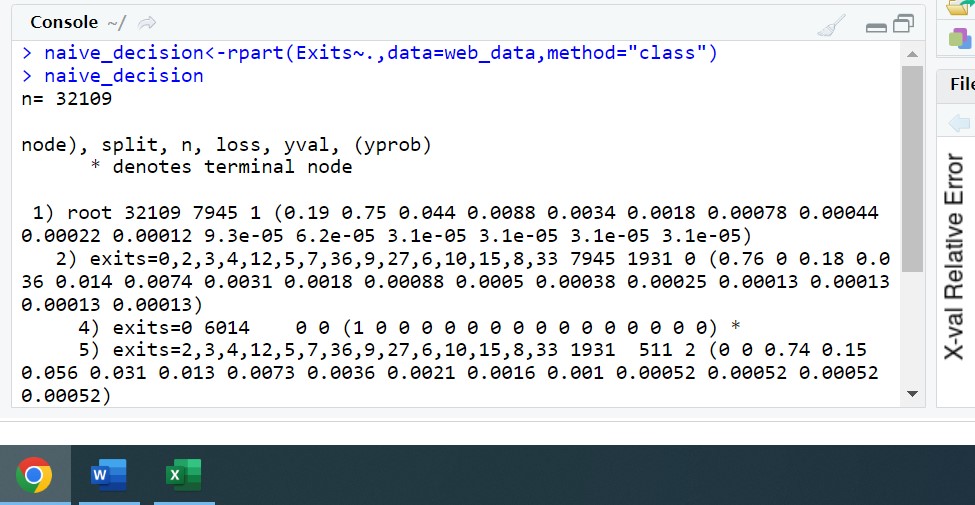


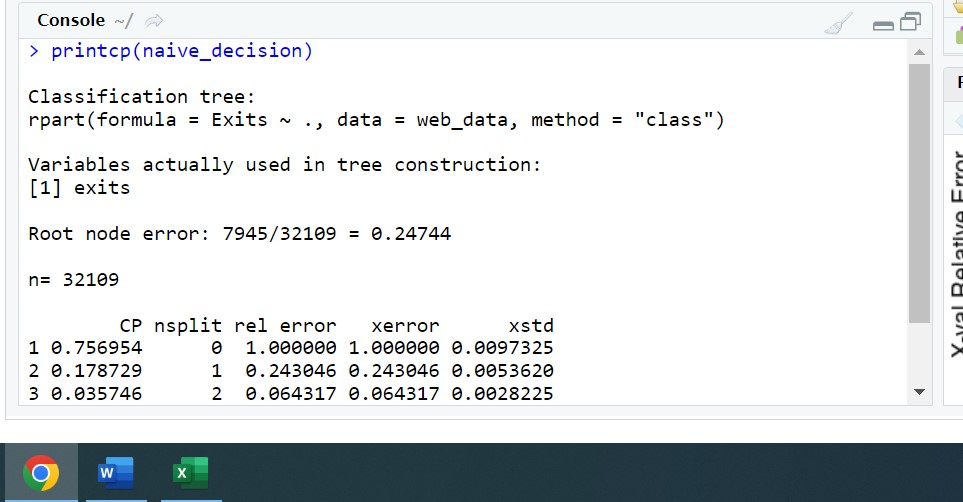




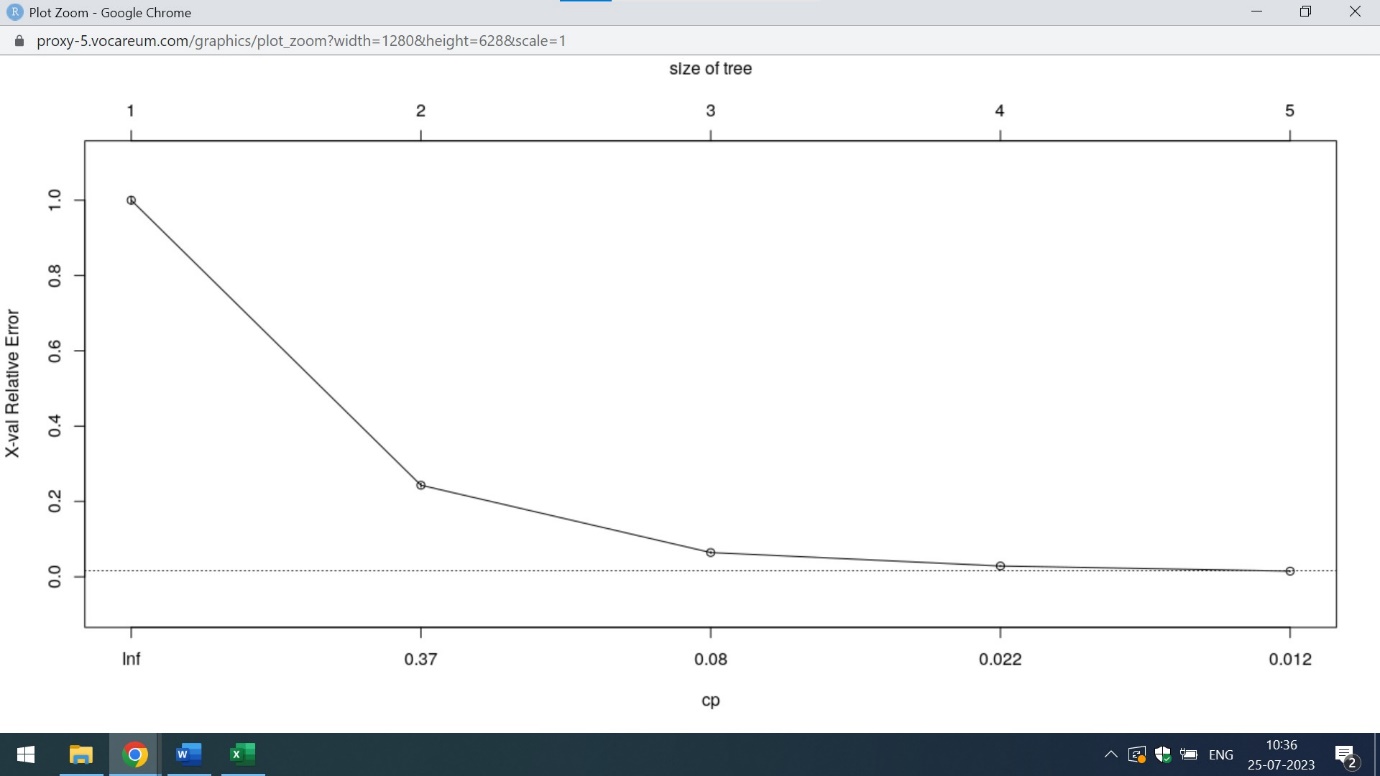
It is clear from the above mention screenshots,that Naïve Bayes Algorithm gives an overview of Apriori probabilities and conditional probabilities of the various factors of exits. The factors affecting the exits of the page are the Bounces,Continent,SourceGroup,TimeinPage and UniquePage Views

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





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)

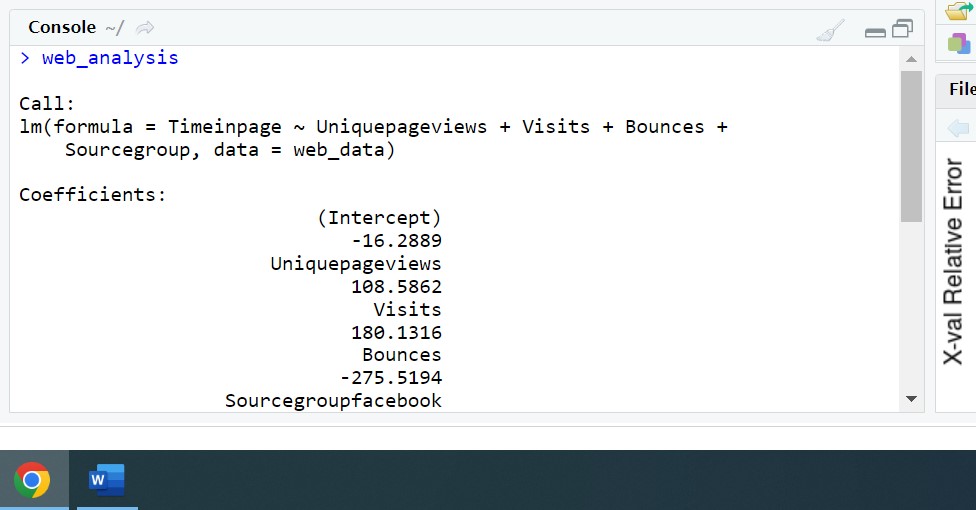
web\_data$Timeinpage<-as.integer(web\_data$Timeinpage)

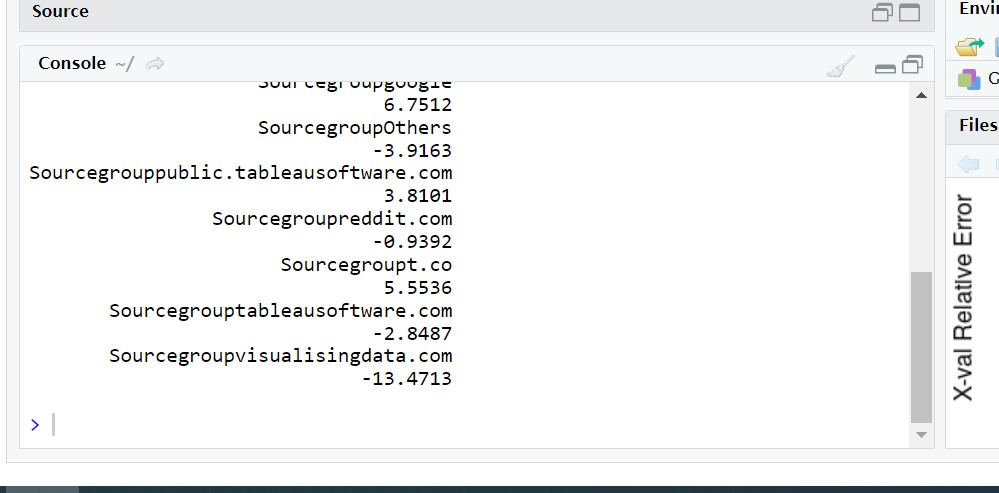
str(web\_data)

web\_analysis<-lm(formula=Timeinpage~Uniquepageviews+Visits+Bounces+Sourcegroup,data=web\_data)

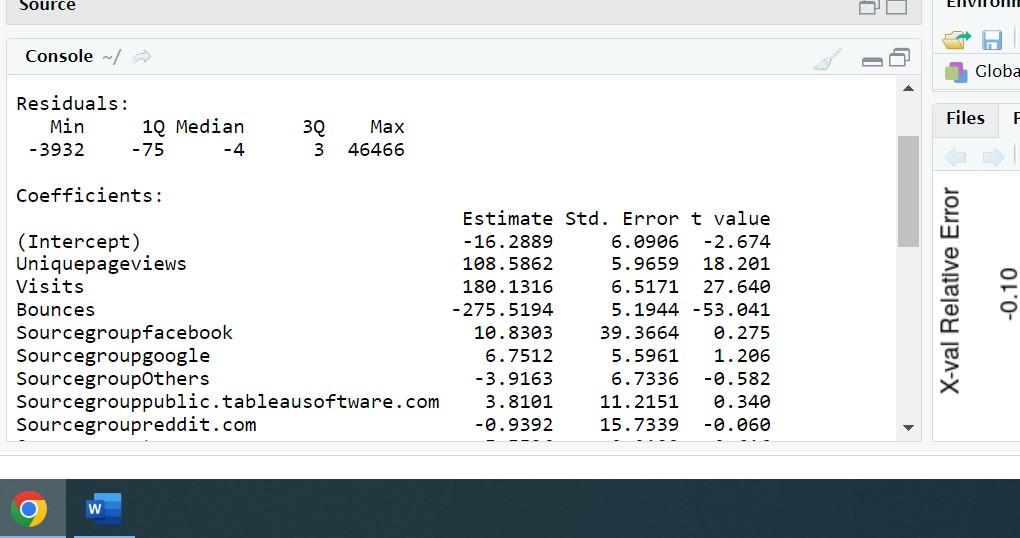
web\_analysis

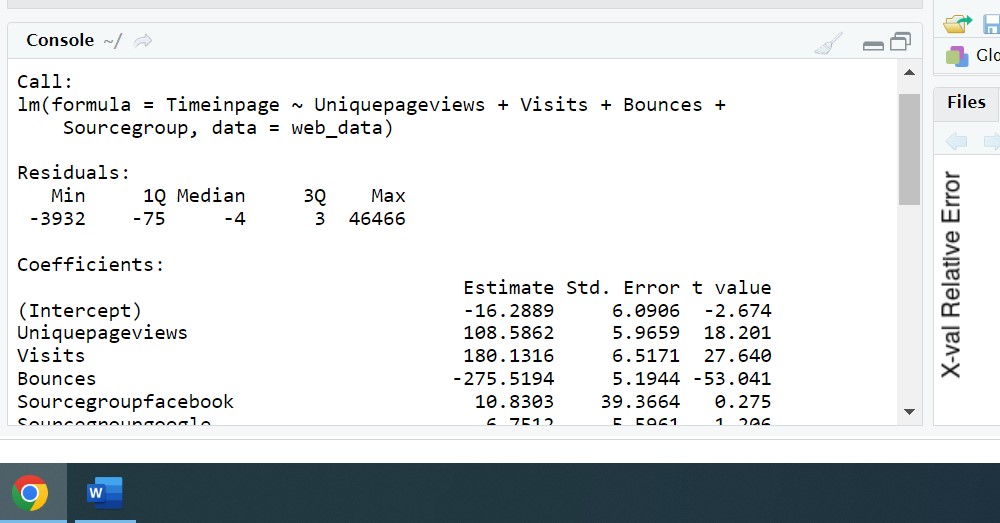
summary(web\_analysis)

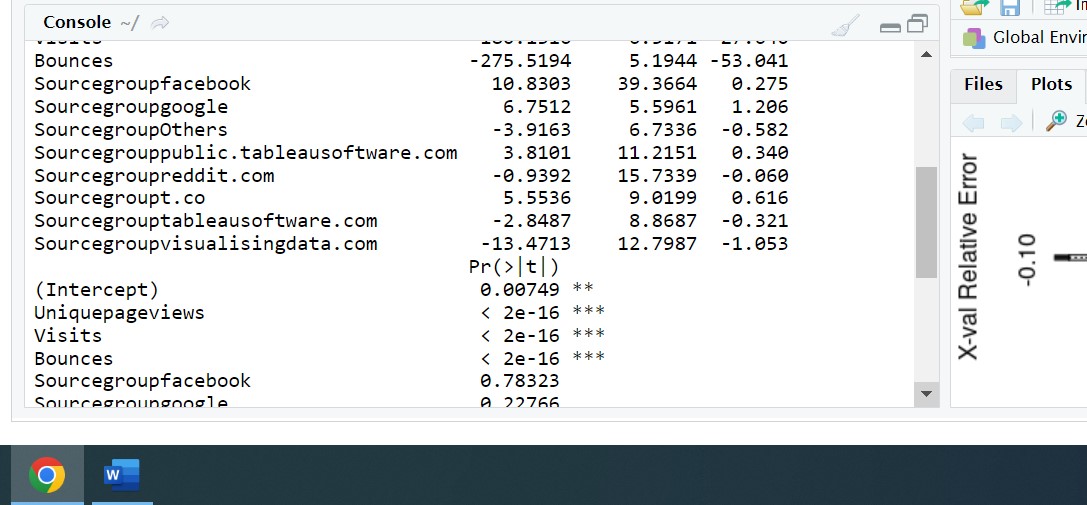




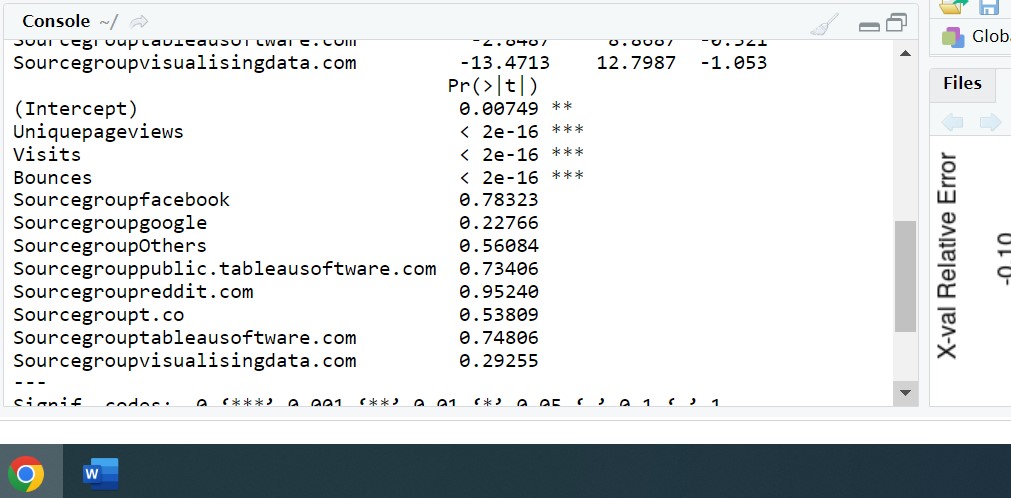
From the above mentioned screenshots,it is clear that Timeinpage is highly dependent upon UniquePage Views and Visits.

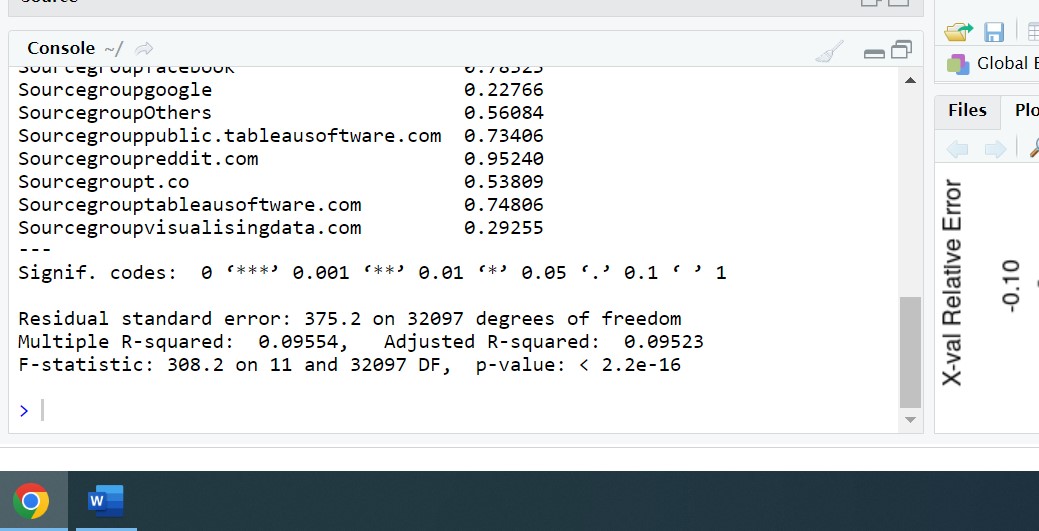






It is clear from the above mentioned screenshots that Unique Page Views,Visits,Sourcegroupfacebook,Sourcegroupgoogle ,Sourcegrouppublic.tableausoftware.com and sourcegroupt.co are the major variables which possibly have an effect on the time on page.





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")

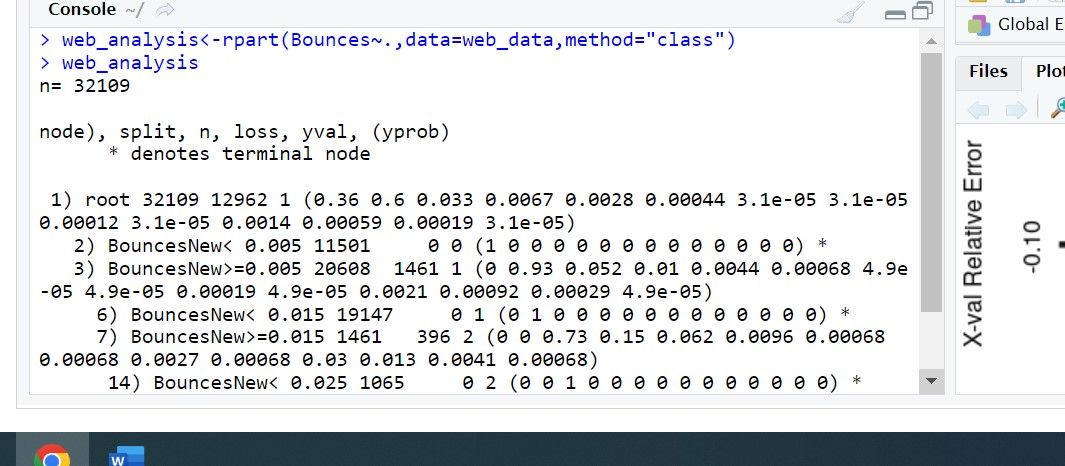
web\_analysis

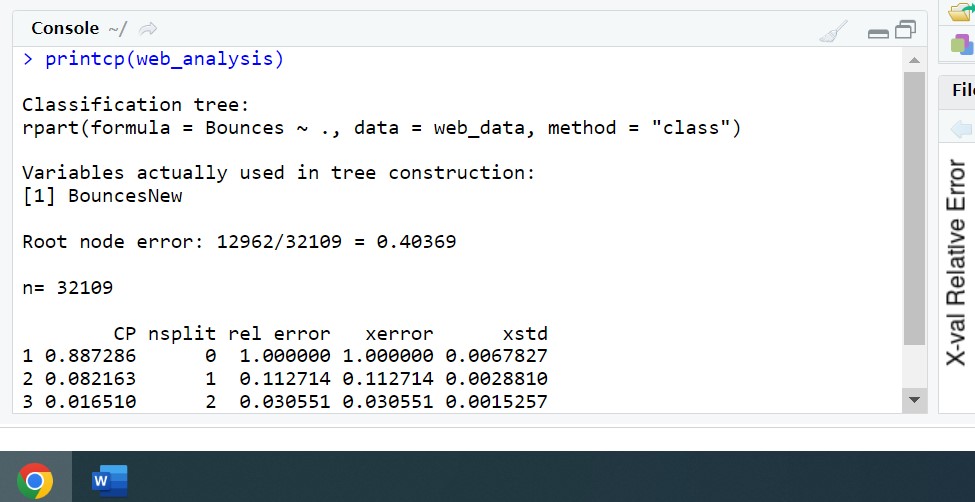
printcp(web\_analysis)

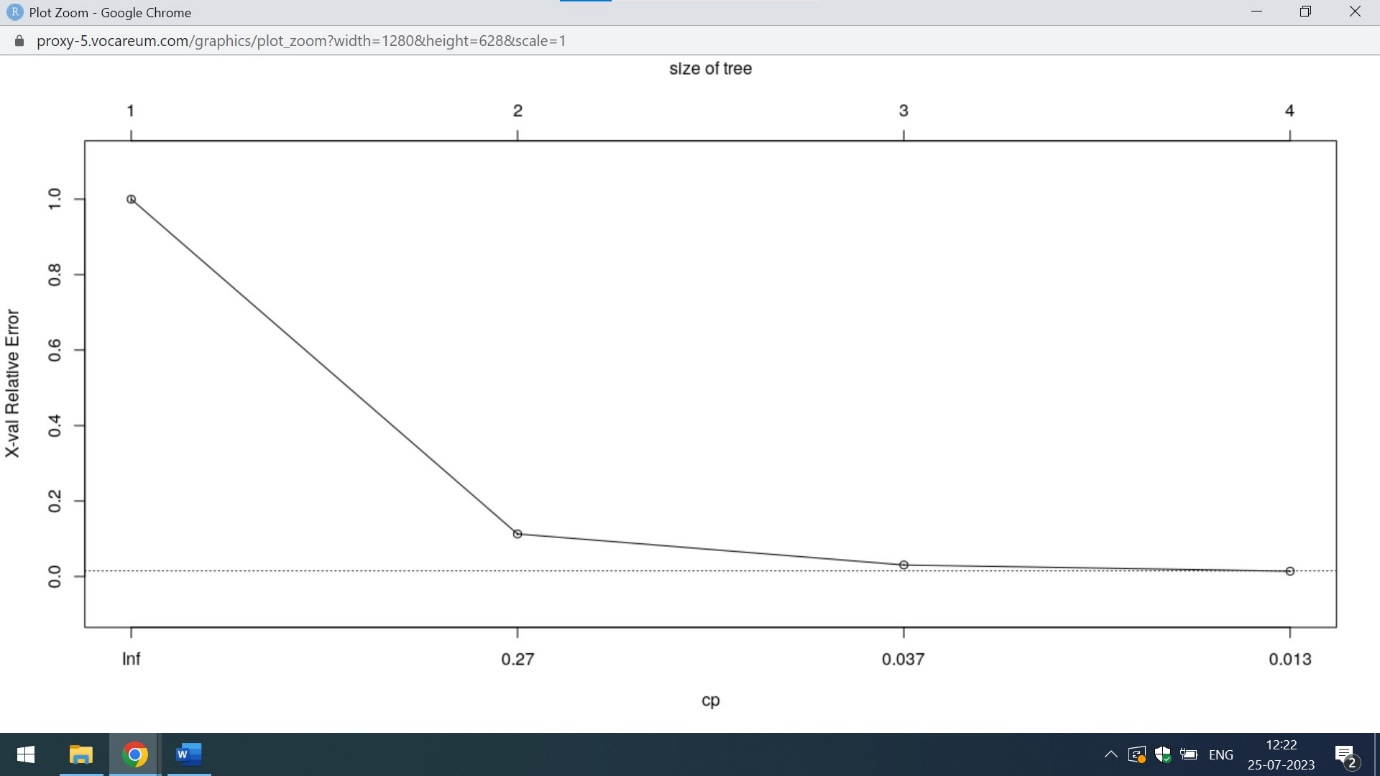
plotcp(web\_analysis)

summary(web\_analysis)

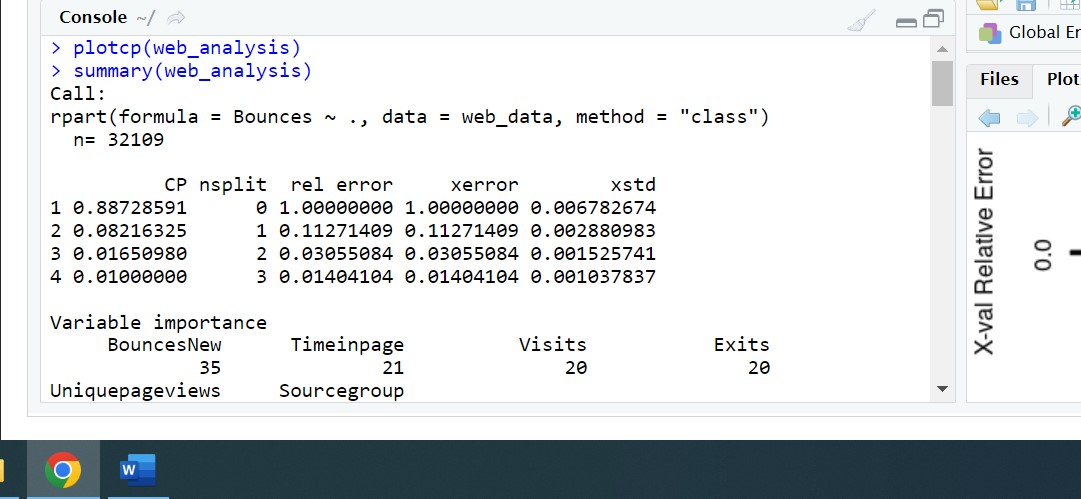
plot(web\_analysis)

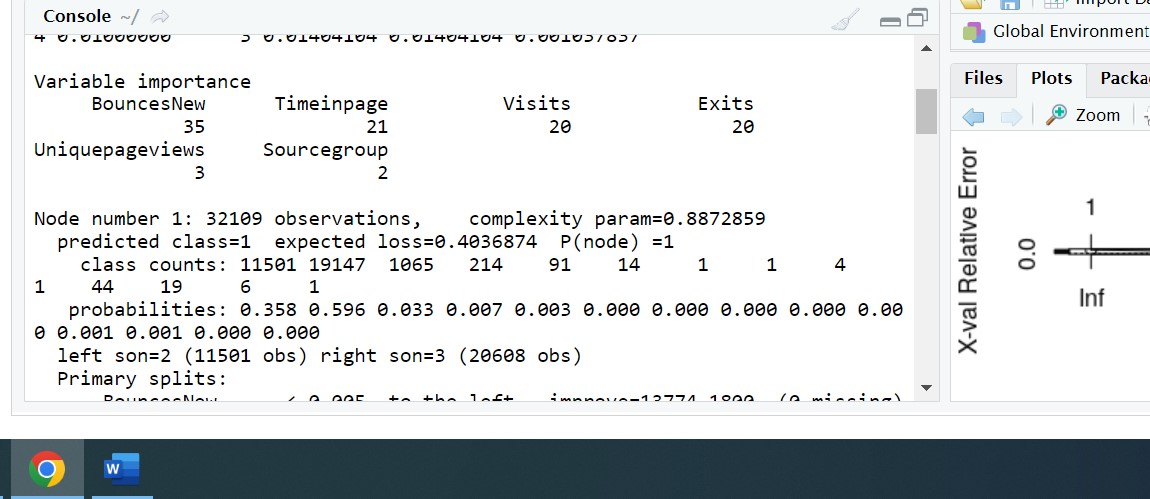


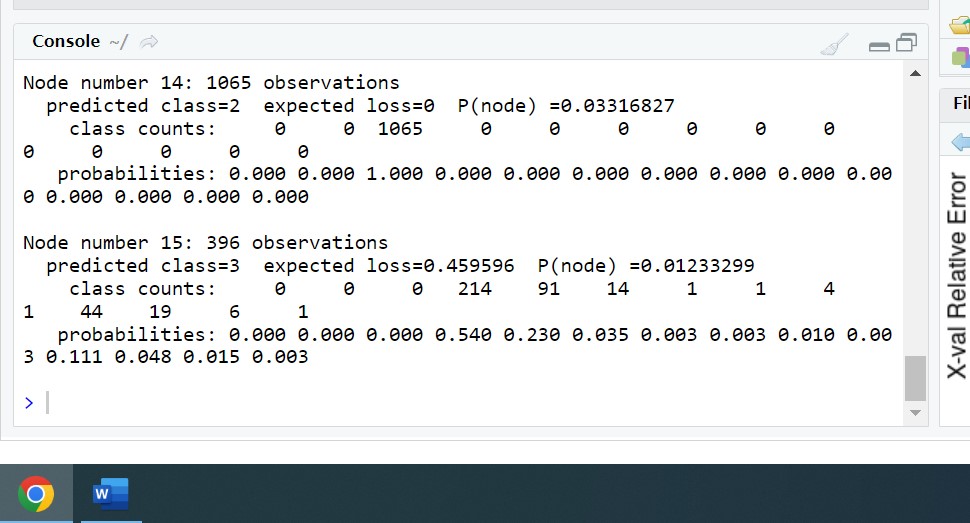




This tree represents that when the size of tree increases the relative error decreases







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