By:- Mrinmay Saha

**Web Data Analysis**

**Statistical Analysis Using R**

**Business Scenario**

**DESCRIPTION**

**Background and Objective:**  
The web analytics team of www.datadb.com is interested to understand the web activities of the site, which are the sources used to access the website. They have a database that states the keywords of time in the page, source group, bounces, exits, unique page views, and visits.

**Domain:** Web

**Dataset Description:**  
The variables in the dataset are defined here for better understanding:

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Bounces | It represents the percentage of visitors who enter the site and "bounce" (leave the site) rather than continuing to view other pages within the same site. |
| Continent | It shows the continent from which the site has been accessed. |
| Source group | It shows how the visitor has accessed the site. |
| Time on page | It shows how long the user has spent on that particular page of the website. |
| Unique pageview | It represents the number of sessions during which that page was viewed one or more times. |
| Visits | A visit counts all visitors, no matter how many times the same visitor may have been to your site. |

**Analysis to be done:**

**Analysis Tasks:**  
The team is targeting the following issues:

* 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 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.
* 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.
* 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.
* 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:**

library(readxl)

web<-read\_xlsx('1555058318\_internet\_dataset.xlsx')

str(web)

**#converting the char value to categorical value**

web$Continent<-as.factor(web$Continent)

web$Sourcegroup<-as.factor(web$Sourcegroup)

**#summary of the data to get a basic understanding of the dataset and to prepare for further analysis**.

summary(web)

**#checking whether there is a relation between uniquepageviews and Visits**

cor(web$Uniquepageviews,web$Visits)

ano<-aov(Uniquepageviews~Visits,data=web)

summary(ano)

**#checking the factors thats affect the Exits**

anoe<-aov(Exits~.,data=web)

summary(anoe)

**#checking the factors that affects the timeinpage on the website**

anot<-aov(Timeinpage~.,data=web)

summary(anot)

**#checking the factors thats affect the Bounce**

#data value should be between 0 to 1 so using BounsNew variable

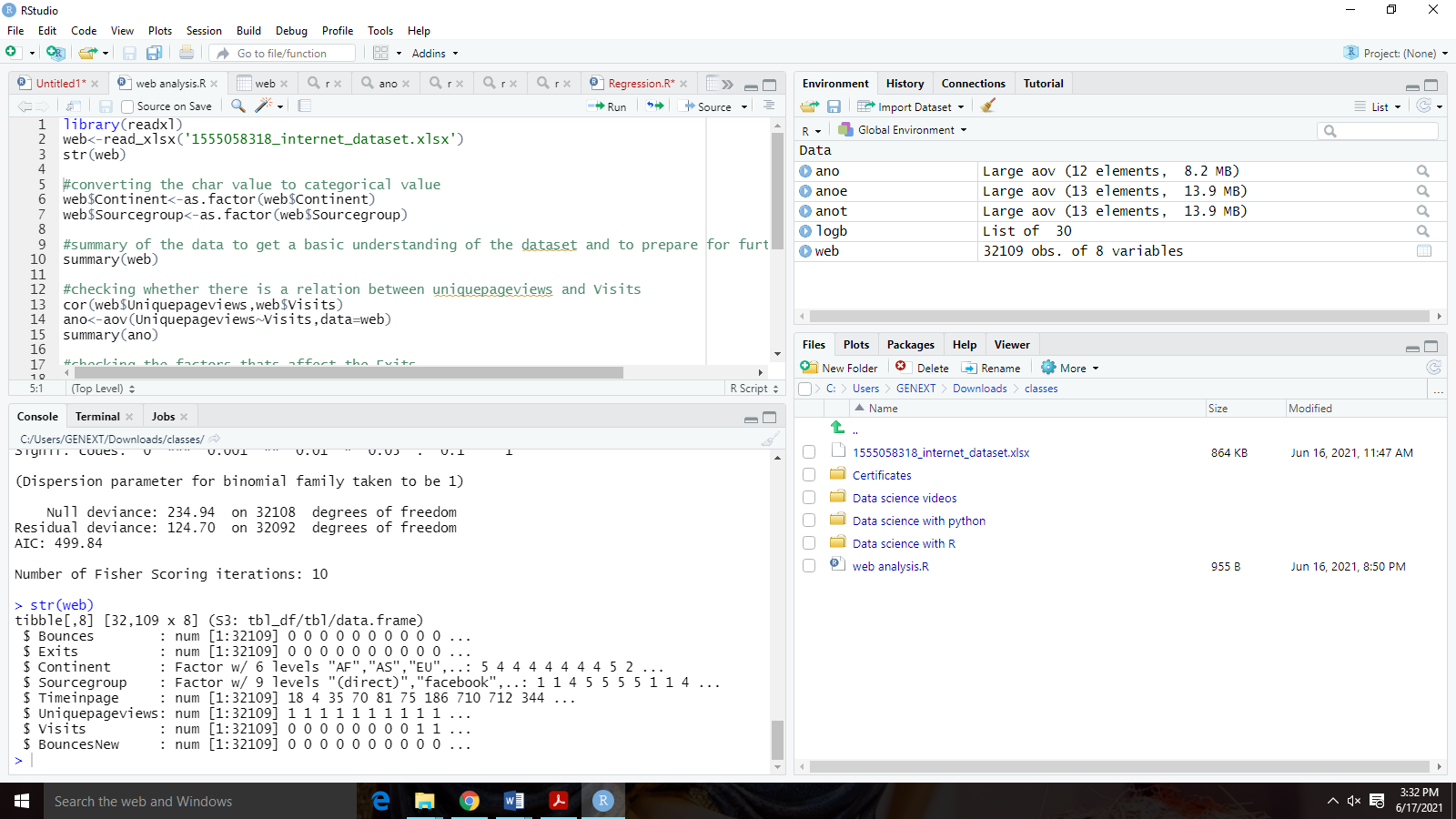
logb<-glm(BouncesNew~Timeinpage+Continent+Sourcegroup+Uniquepageviews+Visits,data = web,family = "binomial")

summary(logb)

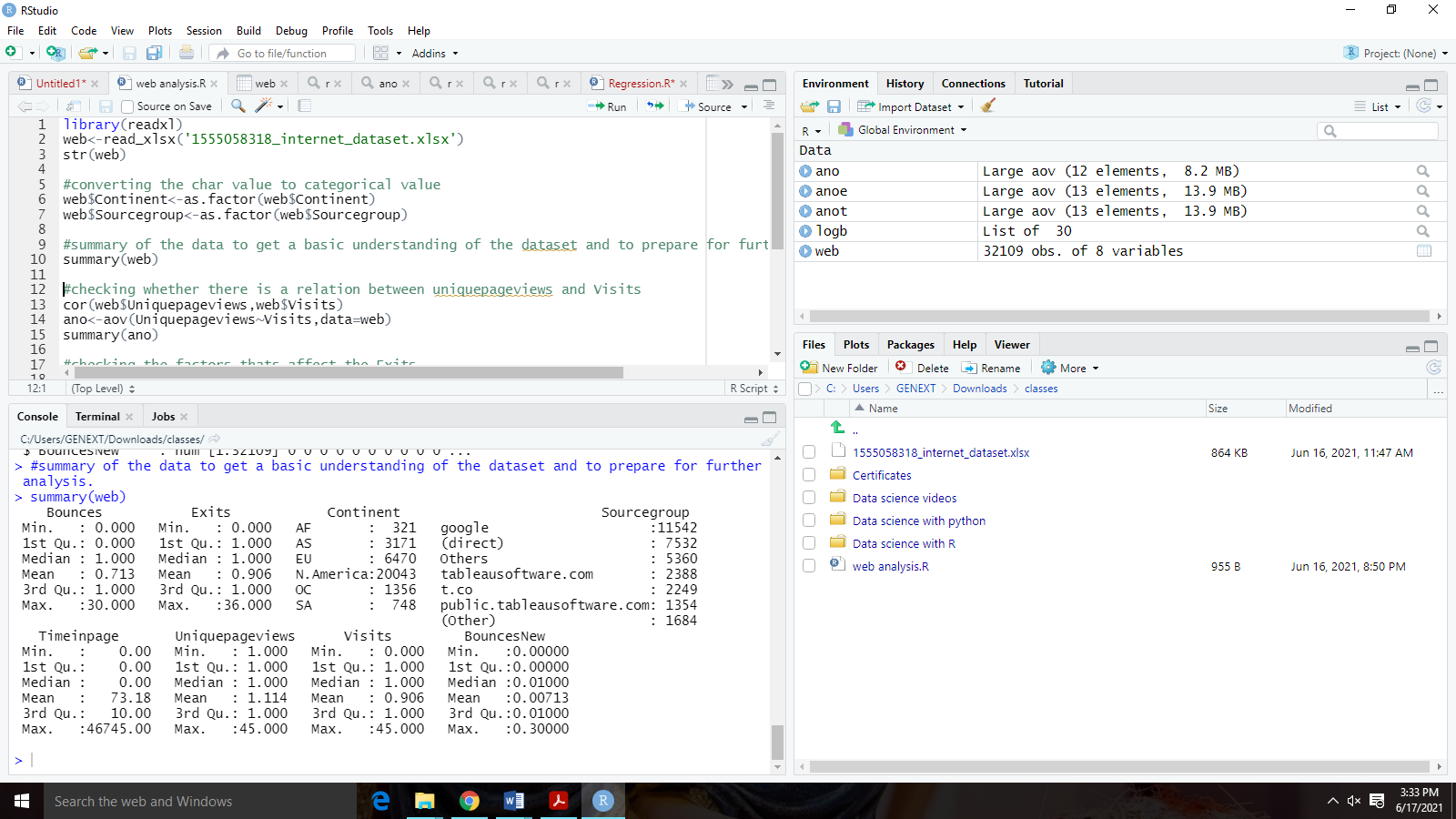
**Output**

library(readxl)

web<-rea d\_xlsx('1555058318\_internet\_dataset.xlsx')

str(web)

**#summary of the data to get a basic understanding of the dataset and to prepare for further analysis**.

summary(web)

**Insight:-** As we can see in the summary the min, max, mean, median, quartile range of the numerical values and the categorical values it shows number of times the value has appeared in the dataset.

As, we can see that the for Bounces min=1 &max=30 and for Exits is min=0 & max=36 respectively for other numerical values.

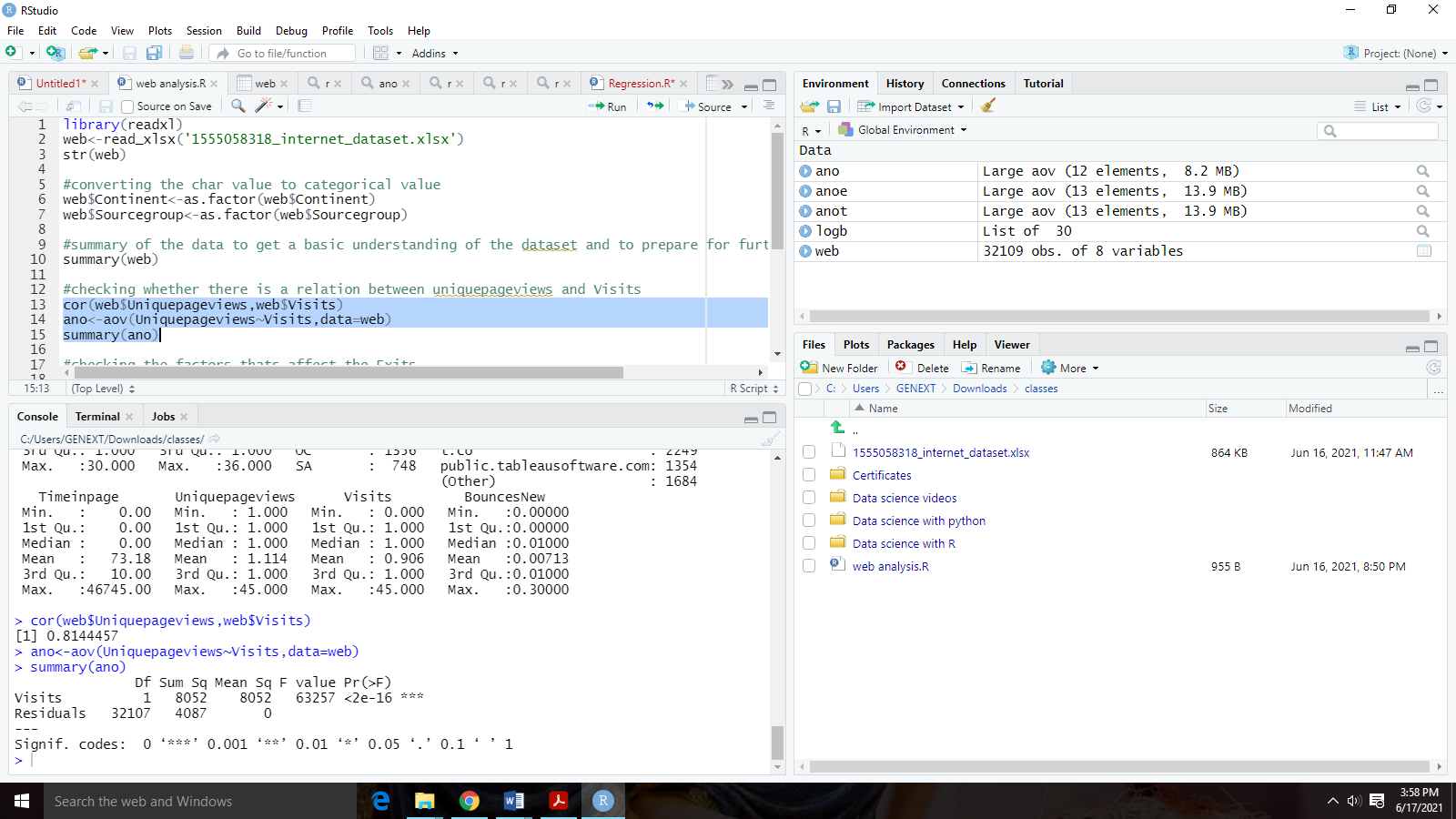
We can also see that maximum no. of visit was from North America.

**#checking whether there is a relation between uniquepageviews and Visits**

cor(web$Uniquepageviews,web$Visits)

ano<-aov(Uniquepageviews~Visits,data=web)

summary(ano)

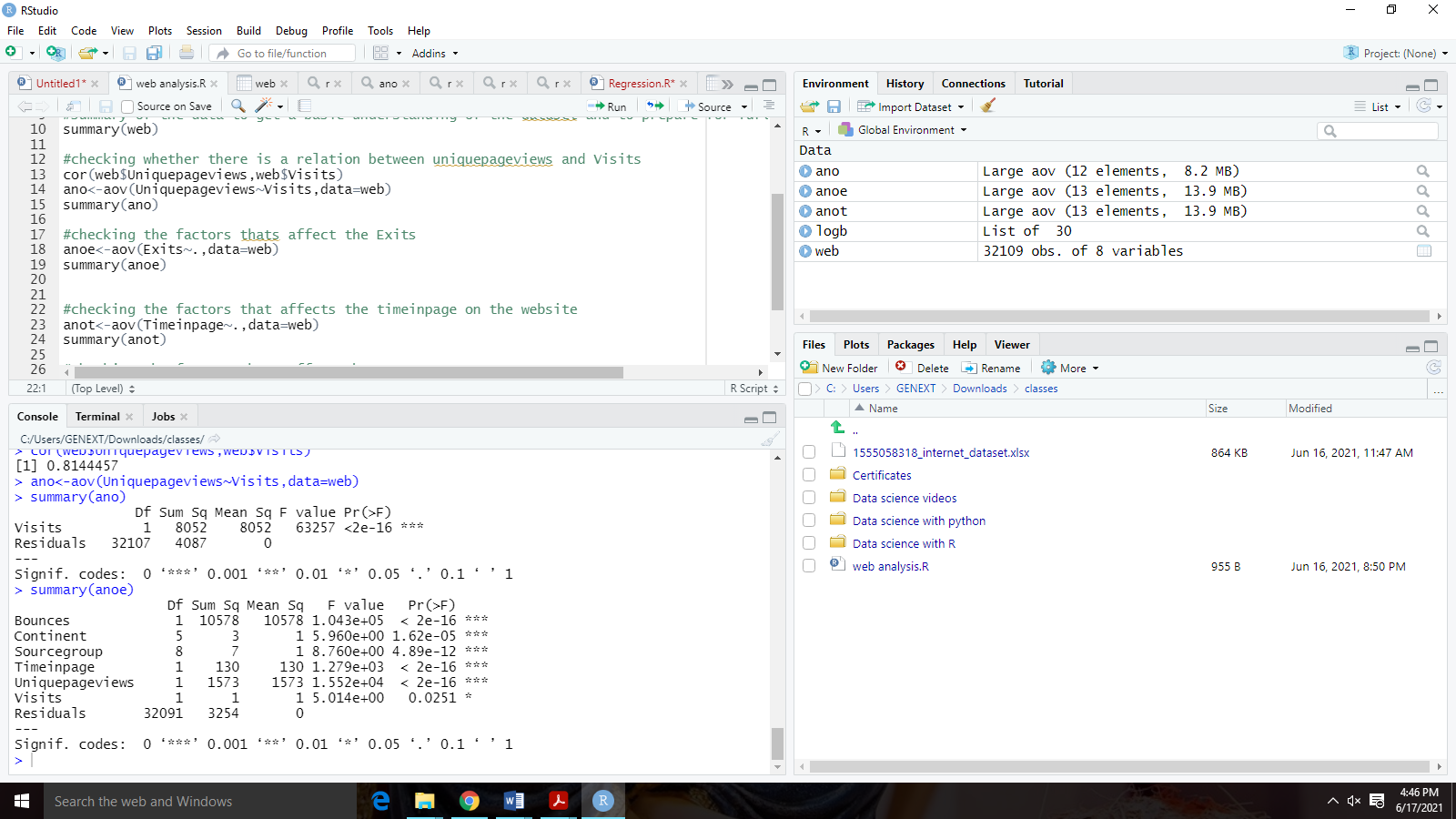


**Insight:** As we can see that from both the test conducted we can infer from the results that the visits variable has a significant impact on Unique.Pageviews. So the team can conclude that unique page values depend on visits.

**#checking the factors thats affect the Exits**

anoe<-aov(Exits~.,data=web)

summary(anoe)



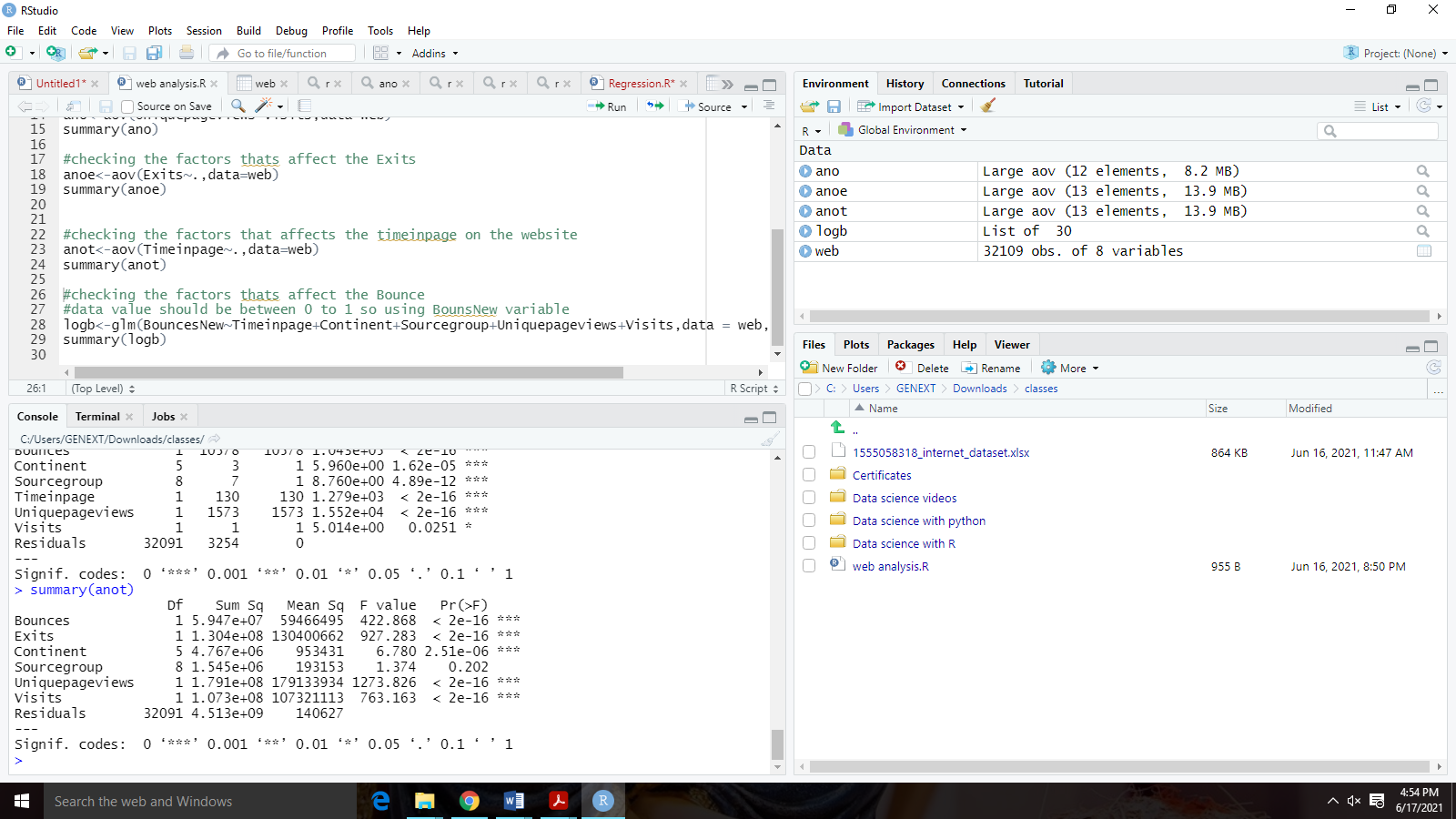
**Insight:**From the result of ANOVA given here, we can see that source.group, bounces, and unique.pageviews have more significance. Visits have comparatively less significance.

Hence we can say that exit from the site is affected by the factors of source group, bounces, and unique.pageviews.

**#checking the factors that affects the timeinpage on the website**

anot<-aov(Timeinpage~.,data=web)

summary(anot)



**Insight:**From the result of ANOVA given here, we can see that bounces,Exits,Visits and unique.pageviews have more significance. source.group have comparatively less significance.

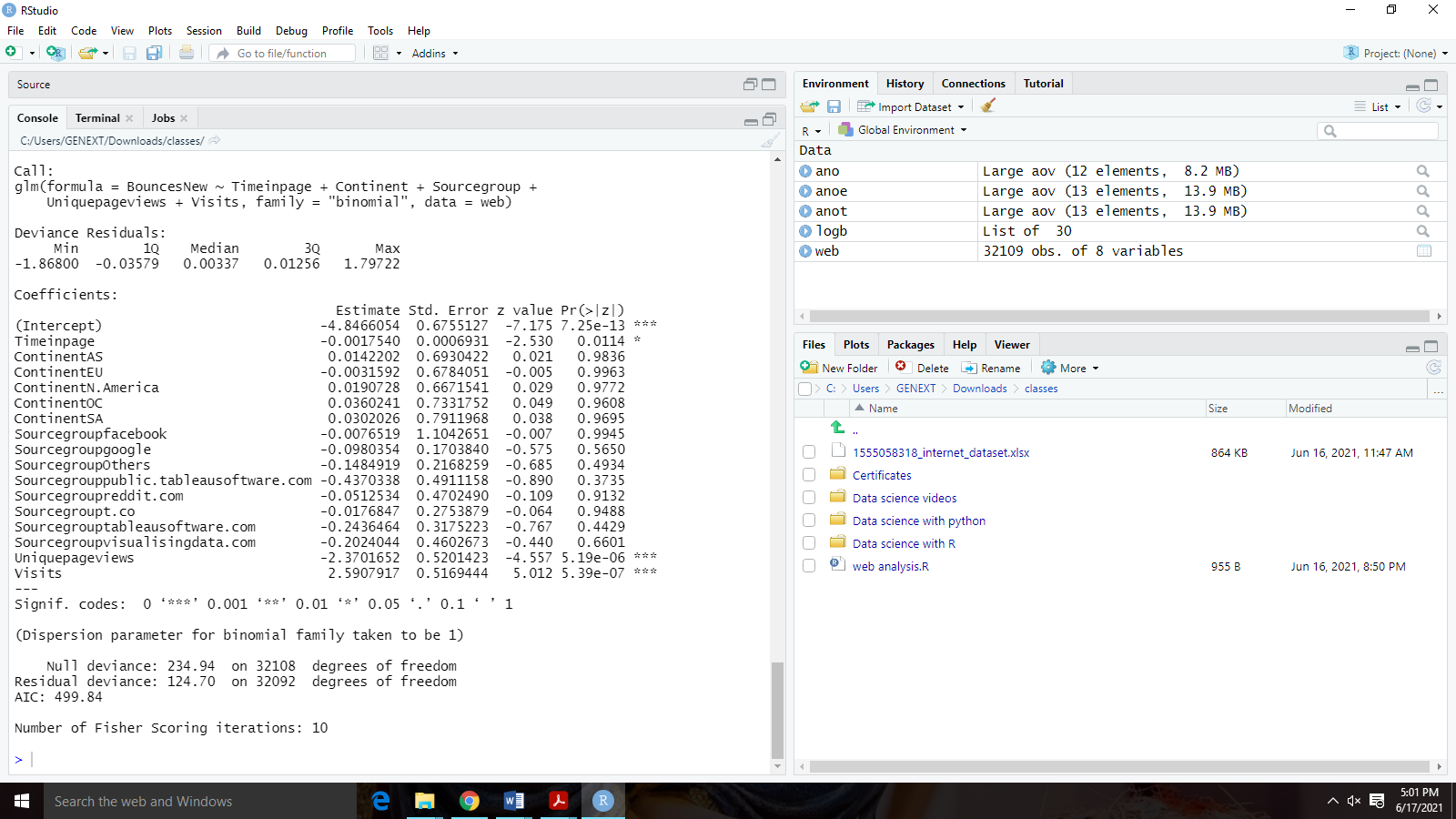
Hence we can say that Timeinpage from the site is affected by the factors of bounces,Exits,Visits and unique.pageviews.

**#checking the factors thats affect the Bounce**

#data value should be between 0 to 1 so using BounsNew variable

logb<-glm(BouncesNew~Timeinpage+Continent+Sourcegroup+Uniquepageviews+Visits,data = web,family = "binomial")

summary(logb)



**Insight:** As can be inferred from the result shown, the Unique.Pageviews and visits are the variables that impact the target variable bounces it has greater significance.