PROJECT 3

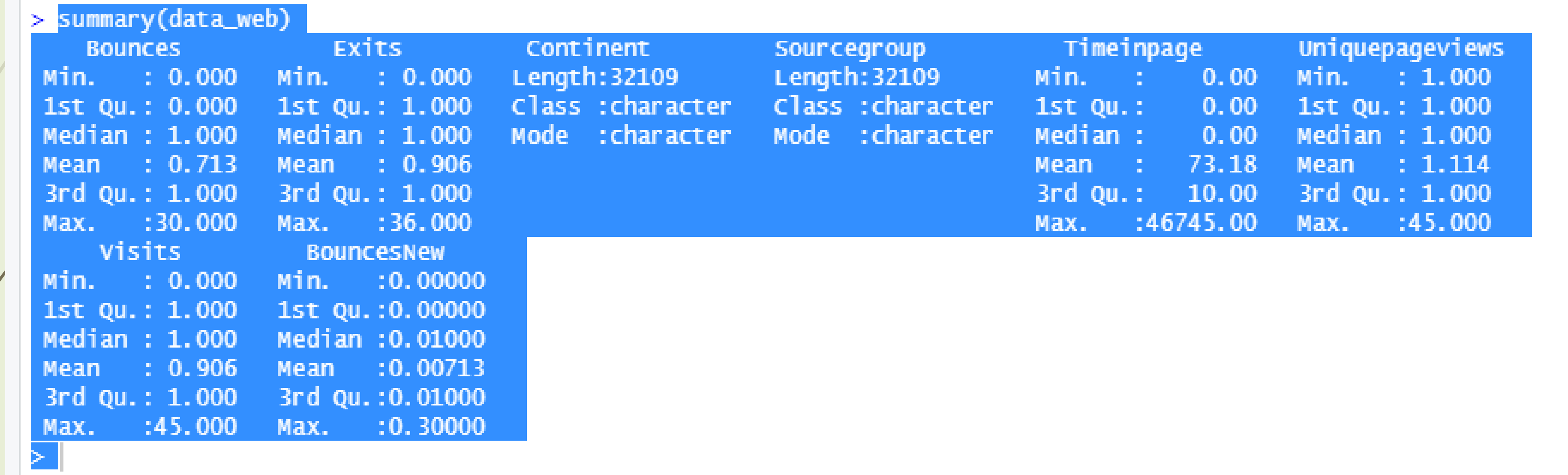
WEB DATA ANALYSIS

USING R PROGRAMMING

EXECUTED BY : NEETHU KRISHNA

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.

```
data_web<-read.csv(file.choose())
install.packages("dplyr")
library("dplyr")
summary(data_web)
```



```
> summary(data_web)
    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
>
```

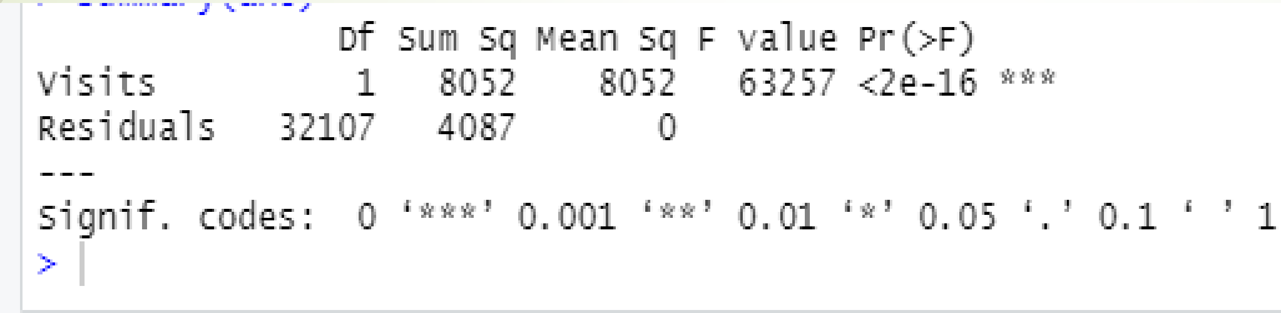
BOUNCES: Min:0   Max: 30   VISITS: Min: 0   Max:45   UNIQUEPAGEVIEW: MIN:0
MAX:45
EXIT : Min: Min:0   Max:36   TIME IN PAGE : MIN:0 MAX=46745

The team needs to know whether the unique page view value depends on visit

cor(data_web$Uniquepageviews, data_web$Visits)

0.8144457

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



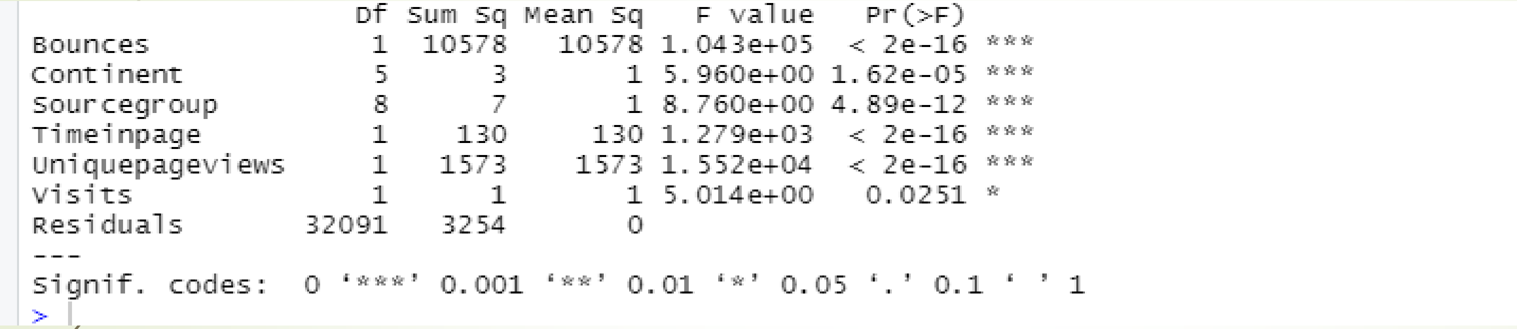
```
               Df Sum Sq Mean Sq F value Pr(>F)    
Visits          1   8052    8052   63257 <2e-16 ***
Residuals   32107   4087       0                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

The visit variable has significant impact on unique page views .

Find out the probable factors from the dataset, which could affect the exits.

anoo<-aov(Exits~.,data = data_web)

summary(anoo)



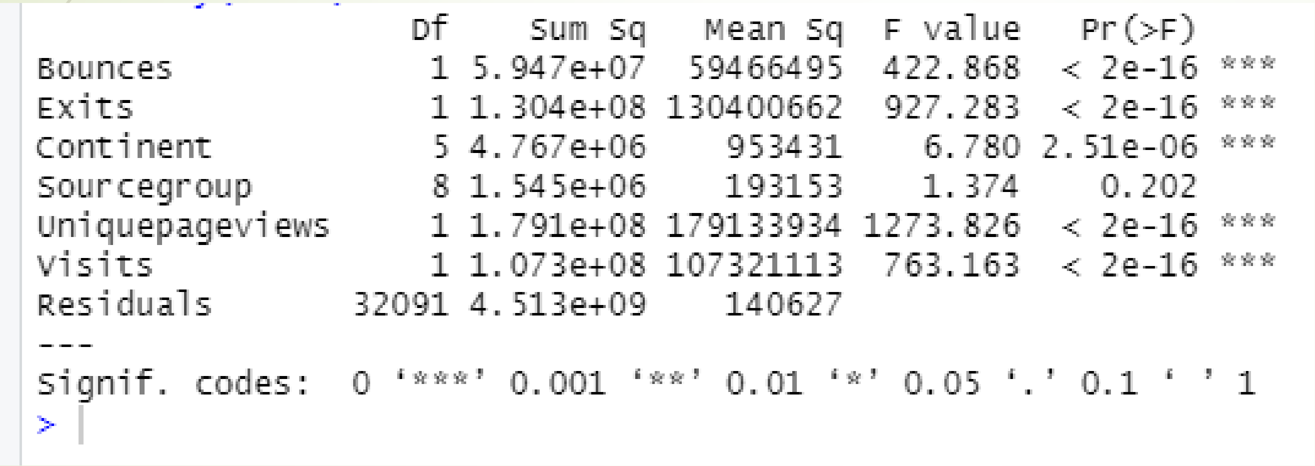
```
                   Df Sum Sq Mean Sq   F value   Pr(>F)    
Bounces             1  10578   10578 1.043e+05  < 2e-16 ***
Continent           5      3       1 5.960e+00 1.62e-05 ***
Sourcegroup         8      7       1 8.760e+00 4.89e-12 ***
Timeinpage          1    130     130 1.279e+03  < 2e-16 ***
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
> 
```

Bounces,sourcegroup, unique page views all has significant impact on exits . Visits have comparatively less significance.

Find the variables which possibly have an effect on the time on page.

anooo<-aov(Timeinpage~.,data = data_web)

summary(anooo)



```
                   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
> 
```

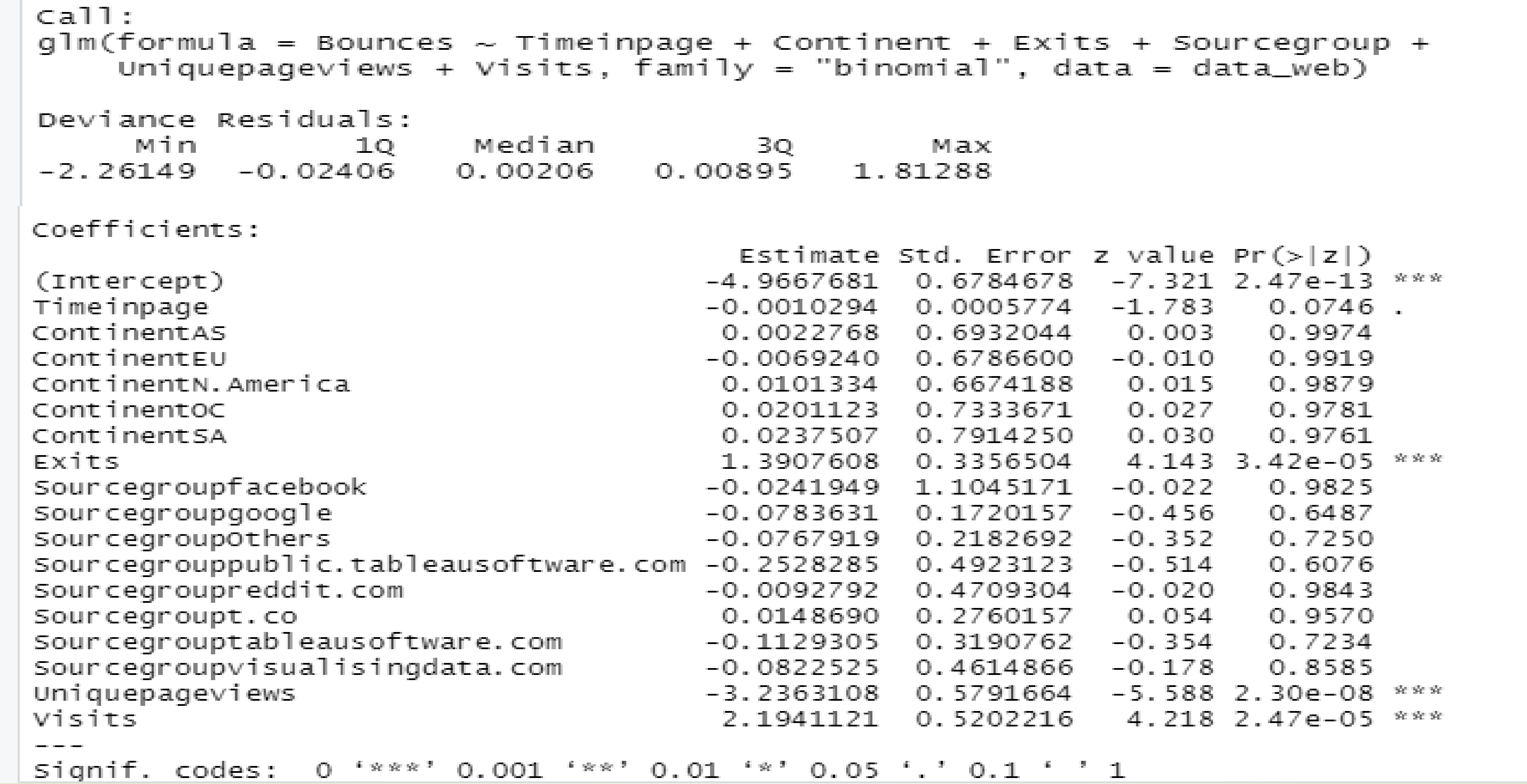
Source group has less significance rest all affect the time on page.

Help the team in determining the factors that are impacting the bounce.

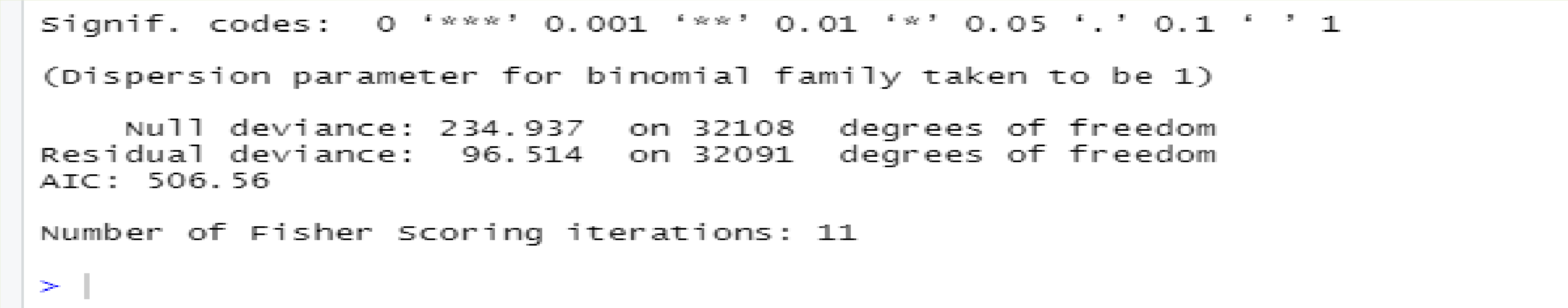
data_web$Bounces=data_web$Bounces*0.01

rmm<-glm(Bounces~Timeinpage+Continent+Exits+Sourcegroup+Uniquepageviews+Visits,data = data_web,family = "binomial")

summary(rmm)

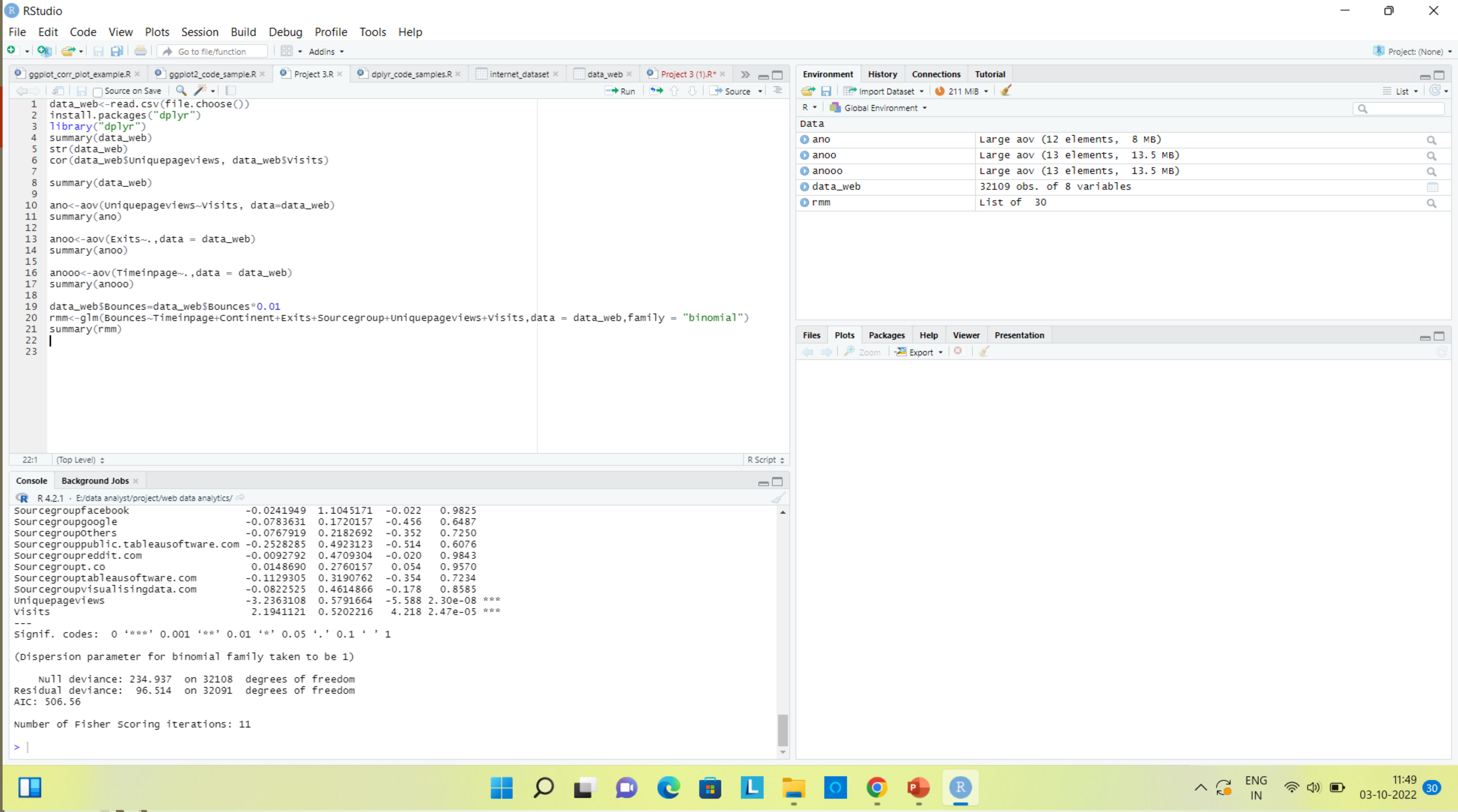


```
Call:
glm(formula = Bounces ~ Timeinpage + Continent + Exits + Sourcegroup + 
    Uniquepageviews + Visits, family = "binomial", data = data_web)

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
```



```
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

> 
```

Unique pageviews, visits, Exists have significance and influence the bounce back.



```
data_web<-read.csv(file.choose())
install.packages("dplyr")
library("dplyr")
summary(data_web)
str(data_web)
cor(data_web$Uniquepageviews, data_web$Visits)

summary(data_web)

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

anoo<-aov(Exits~.,data = data_web)
summary(anoo)

anooo<-aov(Timeinpage~.,data = data_web)
summary(anooo)

data_web$Bounces=data_web$Bounces*0.01
rmm<-glm(Bounces~Timeinpage+Continent+Exits+Sourcegroup+Uniquepageviews+Visits,data = data_web,family = "binomial")
summary(rmm)
```

```
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

> 
```

Data

ano	Large aov (12 elements, 8 MB)
anoo	Large aov (13 elements, 13.5 MB)
anooo	Large aov (13 elements, 13.5 MB)
data_web	32109 obs. of 8 variables
rmm	List of 30