Online shoppers analysis

Stanislaw Czekalski

In this document I am going to analyze online shoppers data.

It describes 12,330 user sessions browsing online shop. Dataset comes from UCI Machine Learning Repository. Observations are associated with labels indicating whenever current session ended up with consumer buying something or not. Primary goal of this data is to perform classification of sessions, but the aim of my analysys is to explore the data and get knowledge about users' behaviour.

Raw data looks like this:

```
df = read.csv("online_shoppers_intention.csv")
head(df, 5)
```

```
##   Administrative Administrative_Duration Informational
## 1              0                       0             0
## 2              0                       0             0
## 3              0                       0             0
## 4              0                       0             0
## 5              0                       0             0
##   Informational_Duration ProductRelated ProductRelated_Duration
## 1                      0              1                0.000000
## 2                      0              2               64.000000
## 3                      0              1                0.000000
## 4                      0              2                2.666667
## 5                      0             10              627.500000
##   BounceRates ExitRates PageValues SpecialDay Month OperatingSystems
## 1        0.20      0.20          0          0   Feb                1
## 2        0.00      0.10          0          0   Feb                2
## 3        0.20      0.20          0          0   Feb                4
## 4        0.05      0.14          0          0   Feb                3
## 5        0.02      0.05          0          0   Feb                3
##   Browser Region TrafficType       VisitorType Weekend Revenue
## 1       1      1           1 Returning_Visitor   FALSE   FALSE
## 2       2      1           2 Returning_Visitor   FALSE   FALSE
## 3       1      9           3 Returning_Visitor   FALSE   FALSE
## 4       2      2           4 Returning_Visitor   FALSE   FALSE
## 5       3      1           4 Returning_Visitor    TRUE   FALSE
```

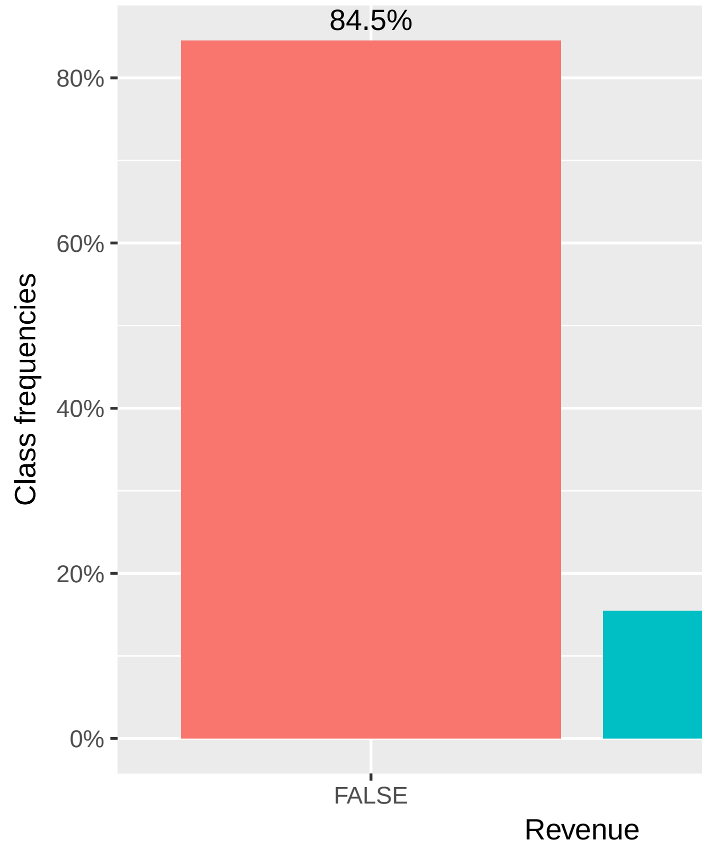
We can see that variables of different types, some of them are categorical. I will set their types to factors.

We should always check for missing values, executing any(is.na(df)) comes in handy.

```
## [1] FALSE
```

There is no missing values in this data.

First thing that must be done is checking class distribution. Most frequently, we do it using bar plot. As one



can see, we deal with a problem of imbalanced class distribution.

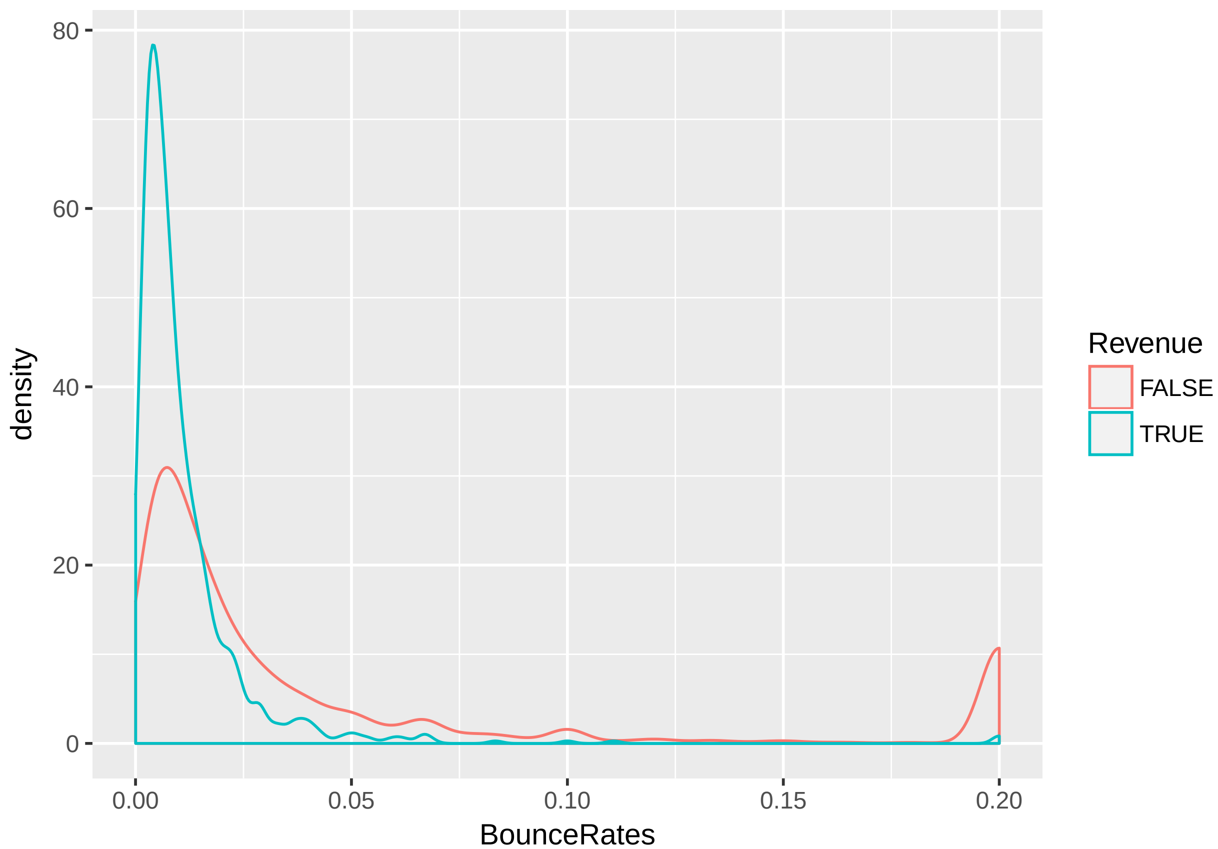
Let's print summary of variables' distributions.

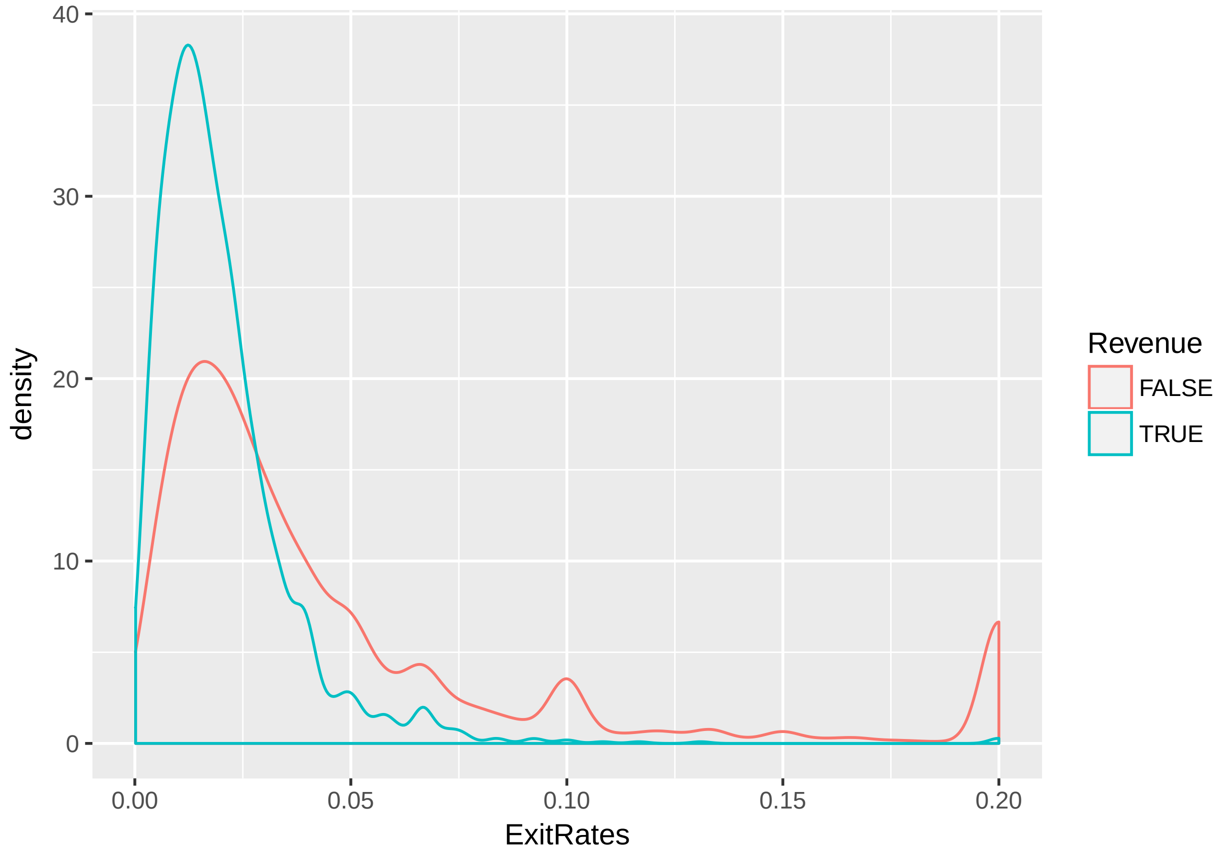
```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   7.50         Median : 0.0000
##  Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
##  3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.0000
##
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
##  Median :   0.00        Median : 18.00   Median :  598.9
##  Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
##  Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##
```

```
##      Month      OperatingSystems    Browser         Region
##  May    :3364   2      :6601     2      :7961   1      :4780
##  Nov    :2998   1      :2585     1      :2462   3      :2403
##  Mar    :1907   3      :2555     4      : 736   4      :1182
##  Dec    :1727   4      : 478     5      : 467   2      :1136
##  Oct    : 549   8      :  79     6      : 174   6      : 805
##  Sep    : 448   6      :  19     10     : 163   7      : 761
##  (Other):1337   (Other):  13     (Other): 367   (Other):1263
##   TrafficType              VisitorType     Weekend         Revenue
##  2      :3913   New_Visitor      : 1694   Mode :logical   Mode :logical
##  1      :2451   Other            :   85   FALSE:9462      FALSE:10422
##  3      :2052   Returning_Visitor:10551   TRUE :2868      TRUE :1908
##  4      :1069
##  13     : 738
##  10     : 450
##  (Other):1657
```

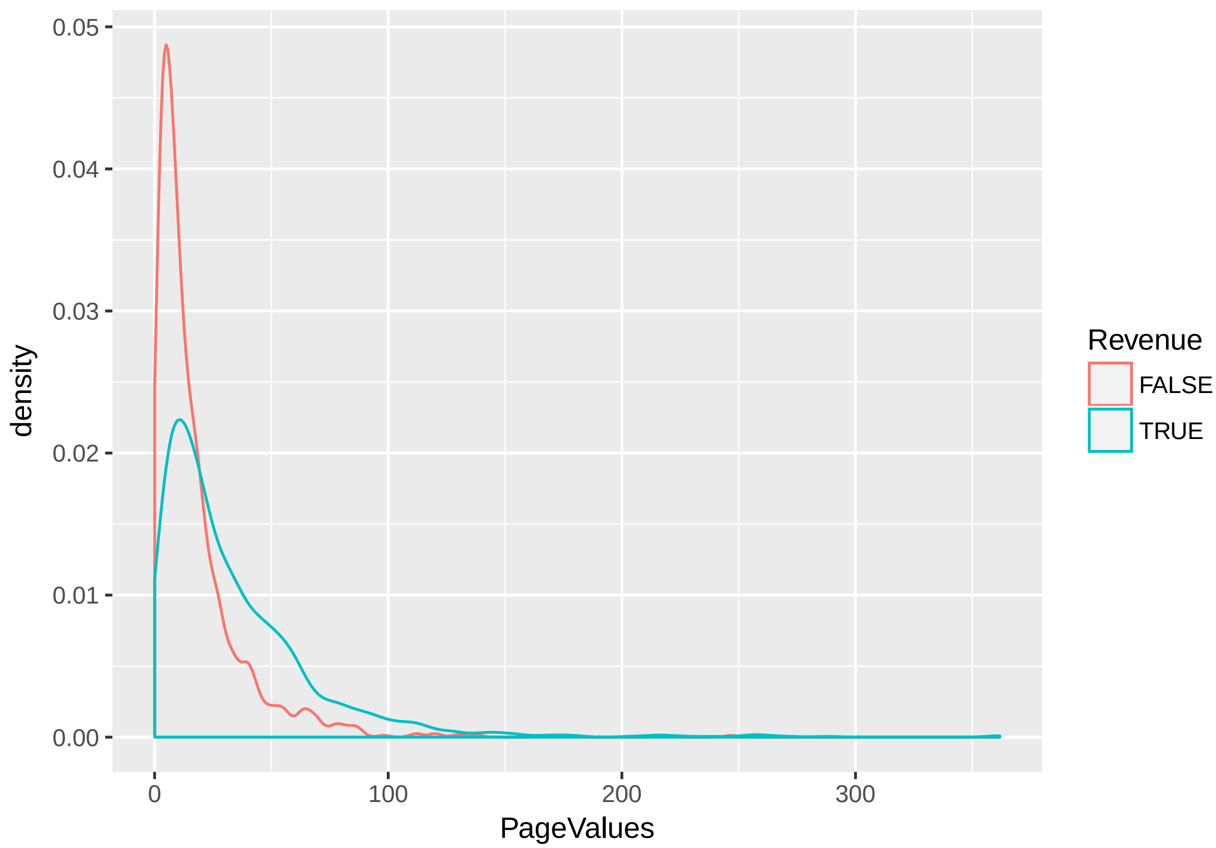
We can see that most user visit parts of the website that are product related. They also spend the most time on them. “Bounce Rate”, “Exit Rate” and “Page Value” are somehow misleading names, after looking up we may discover that they are related to Google Analytics names. “Bounce Rate” describe percentage of visitors that come from Google Analytics, enter the site and then leave (“bounce”) without triggering any other requests to the analytics server during that session. “Exit Rate” feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The “Page Value” feature represents the average value for a web page that a user visited before completing an e-commerce transaction. Values of all “Bounde Rate” and “Exit Rate” are quite low. To furhter investigate this features we can plot their distributions.

Most of values for all three of them are zeros, so I will only plot distribution of non zero values to have a closer look. Additionaly, I decided to split each distribution to two, depending on classes.


80
60
40
20
0
density
0.00
0.05
0.10
0.15
0.20
BounceRates
Revenue
FALSE
TRUE

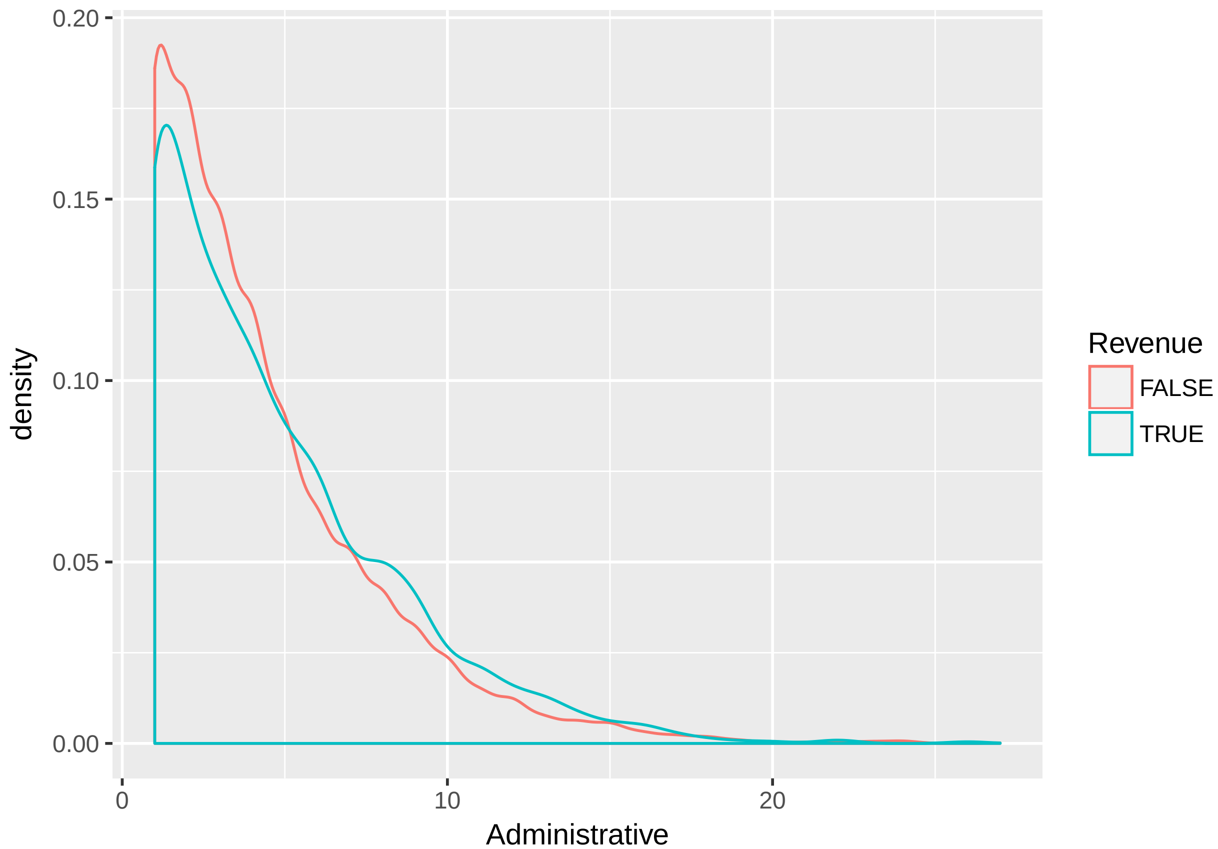


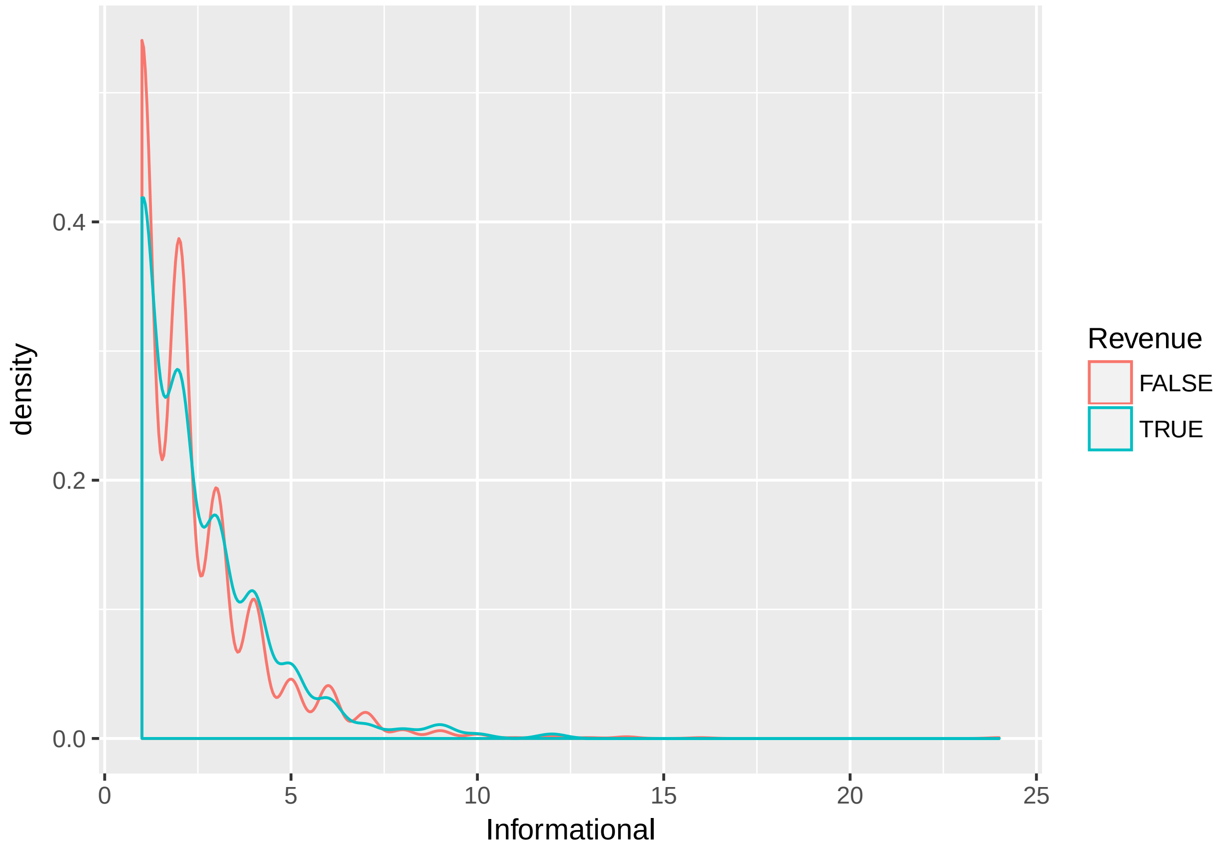
40
30
20
10
0
density
0.00
0.05
0.10
0.15
0.20
ExitRates
Revenue
FALSE
TRUE



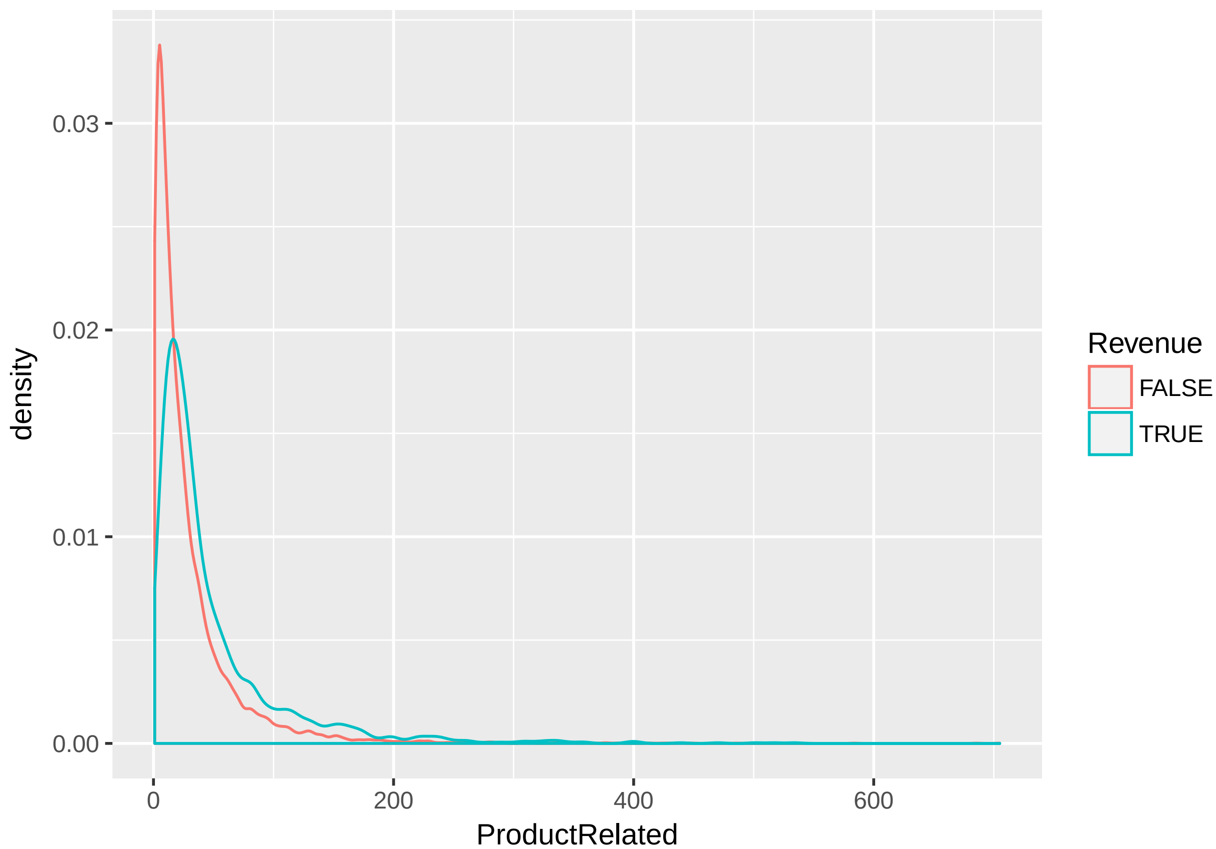
We can see that distributions of all three features differ depending on classes. Especially, all observations having bounce rates and exit rates 0.20 belong to negative class. Distribution of page values is more skewed towards high values for positive class. It means that all three variables might be usefull for building classification model.

Next, we will take a look at times the user spend on different parts of the website.

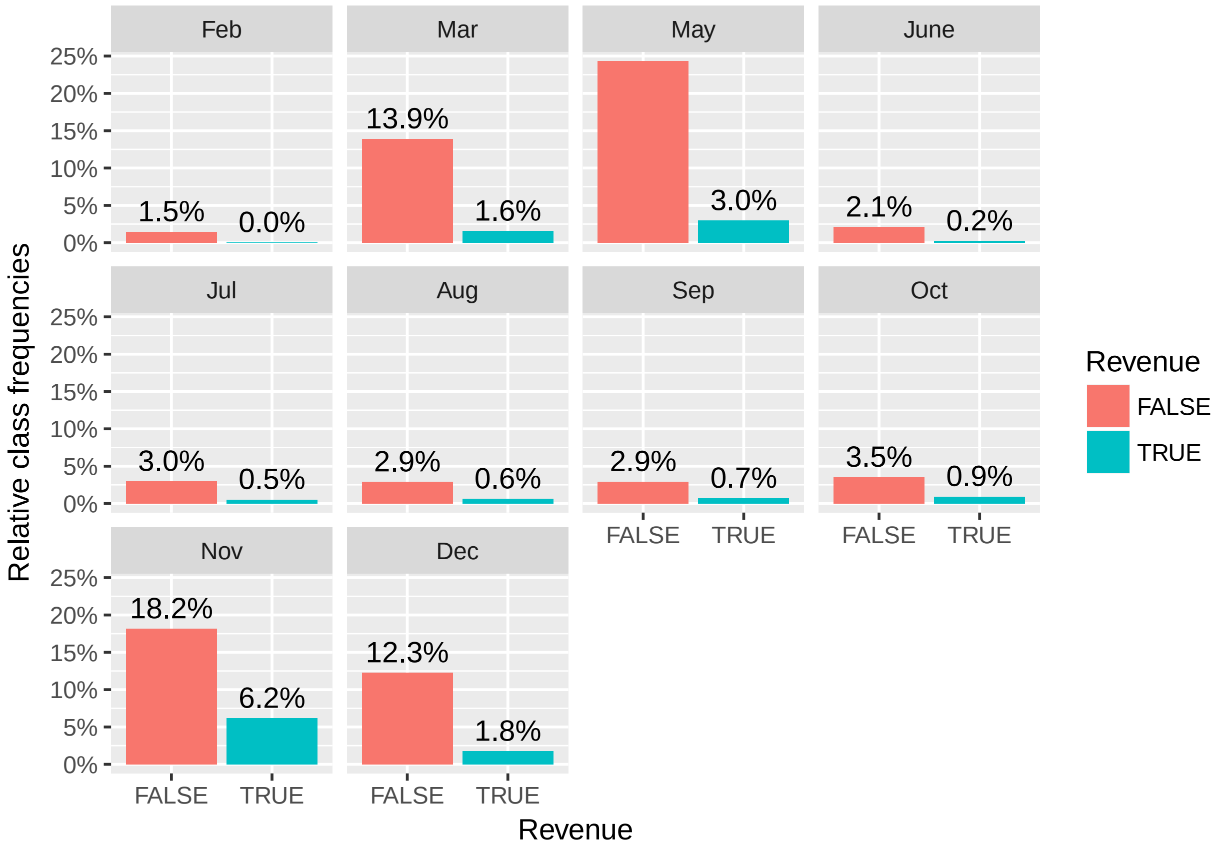

0.20
0.15
0.10
0.05
0.00
density
0
10
20
Administrative
Revenue
FALSE
TRUE



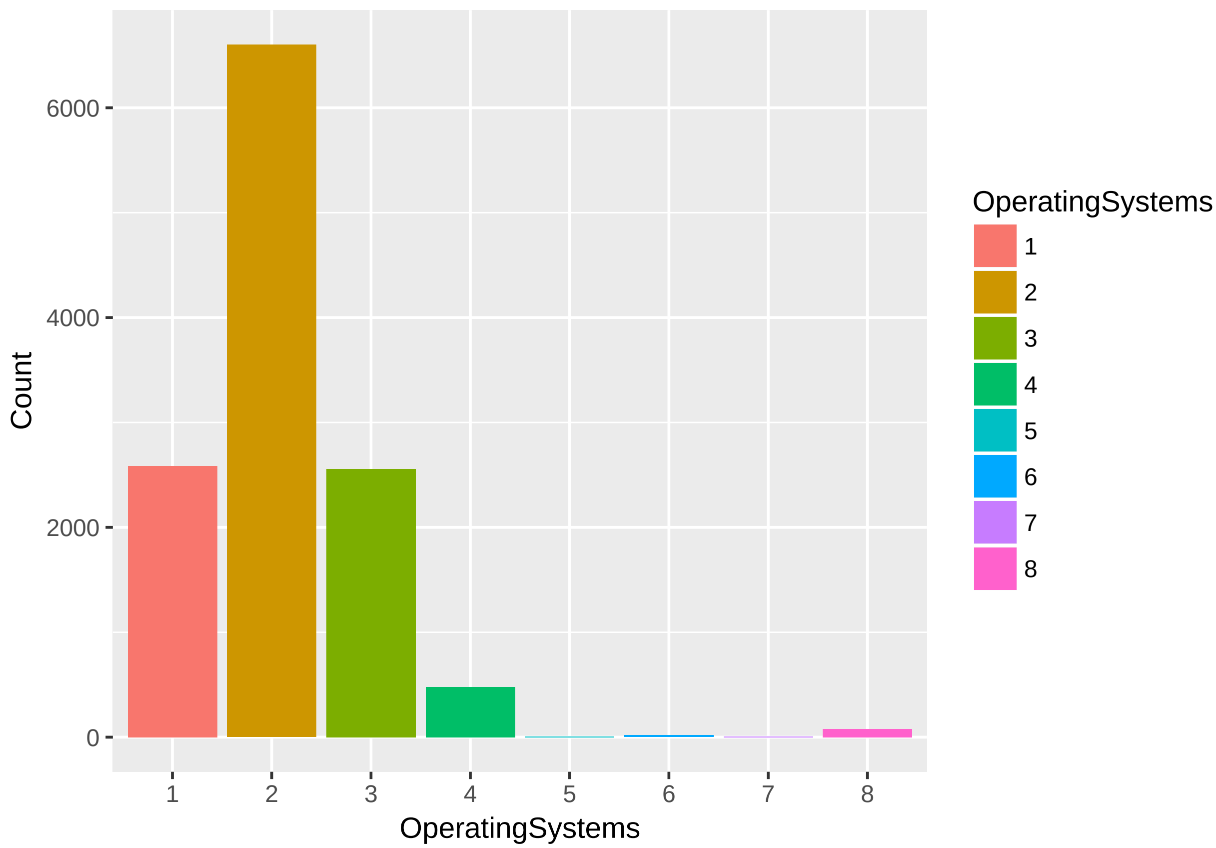
density
0.4
0.2
0.0
0
5
10
15
20
25
Informational
Revenue
FALSE
TRUE

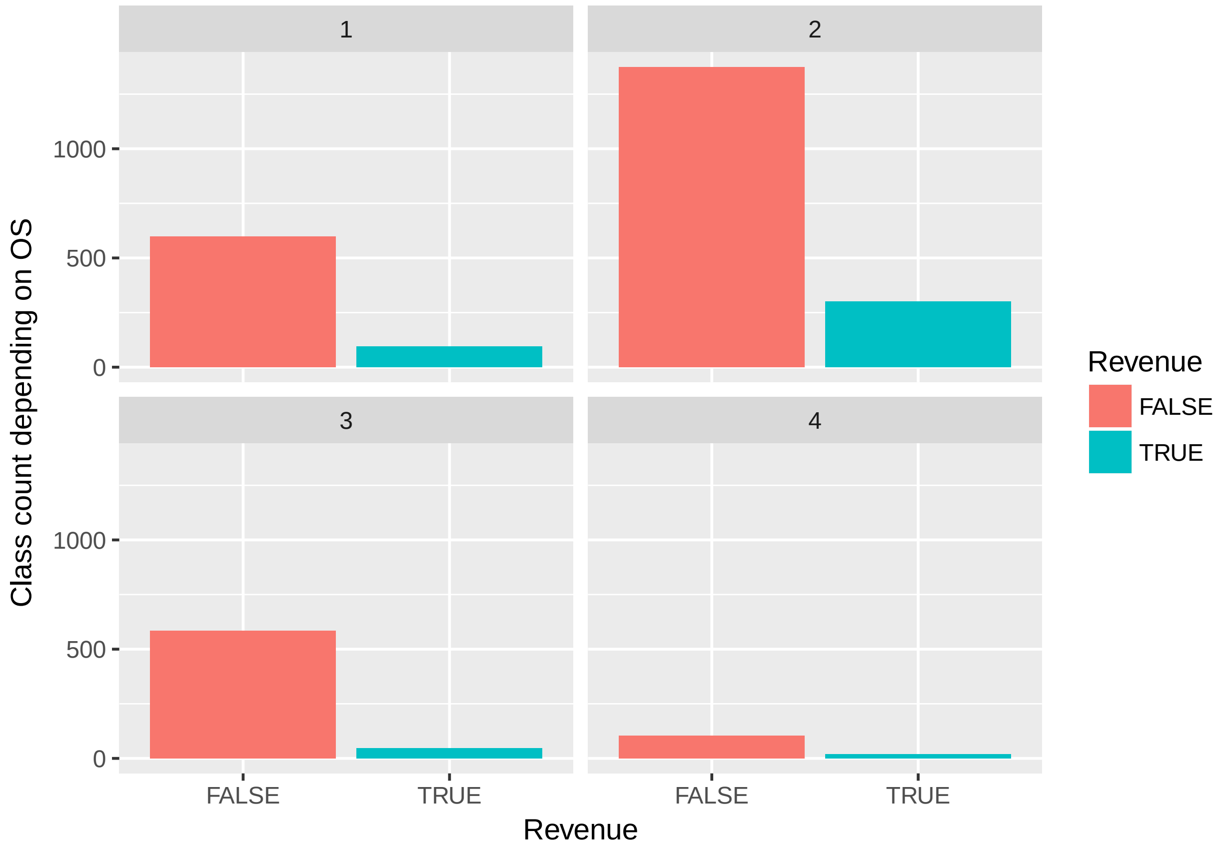


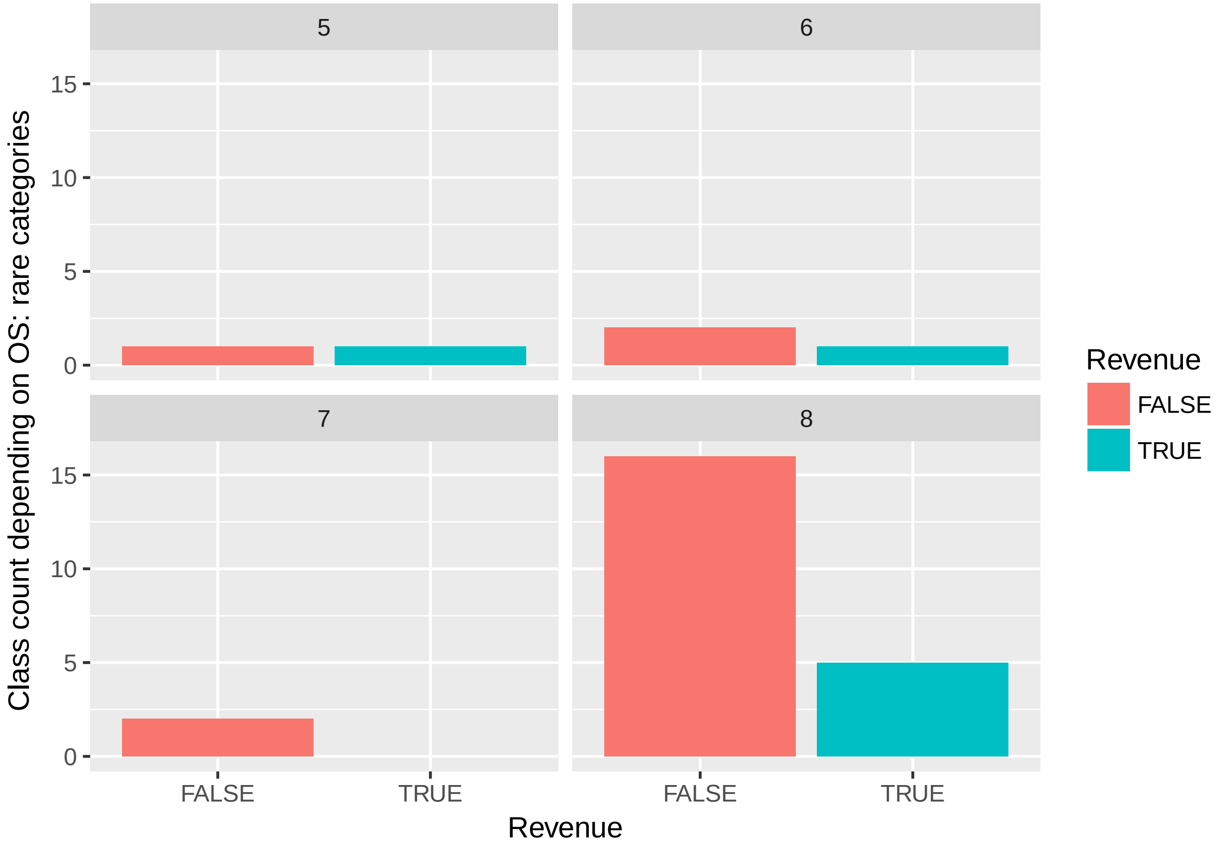
As one can see, there is no much difference bitween times spend on parts of website depending on class. This raw features might not be useful for prediction.



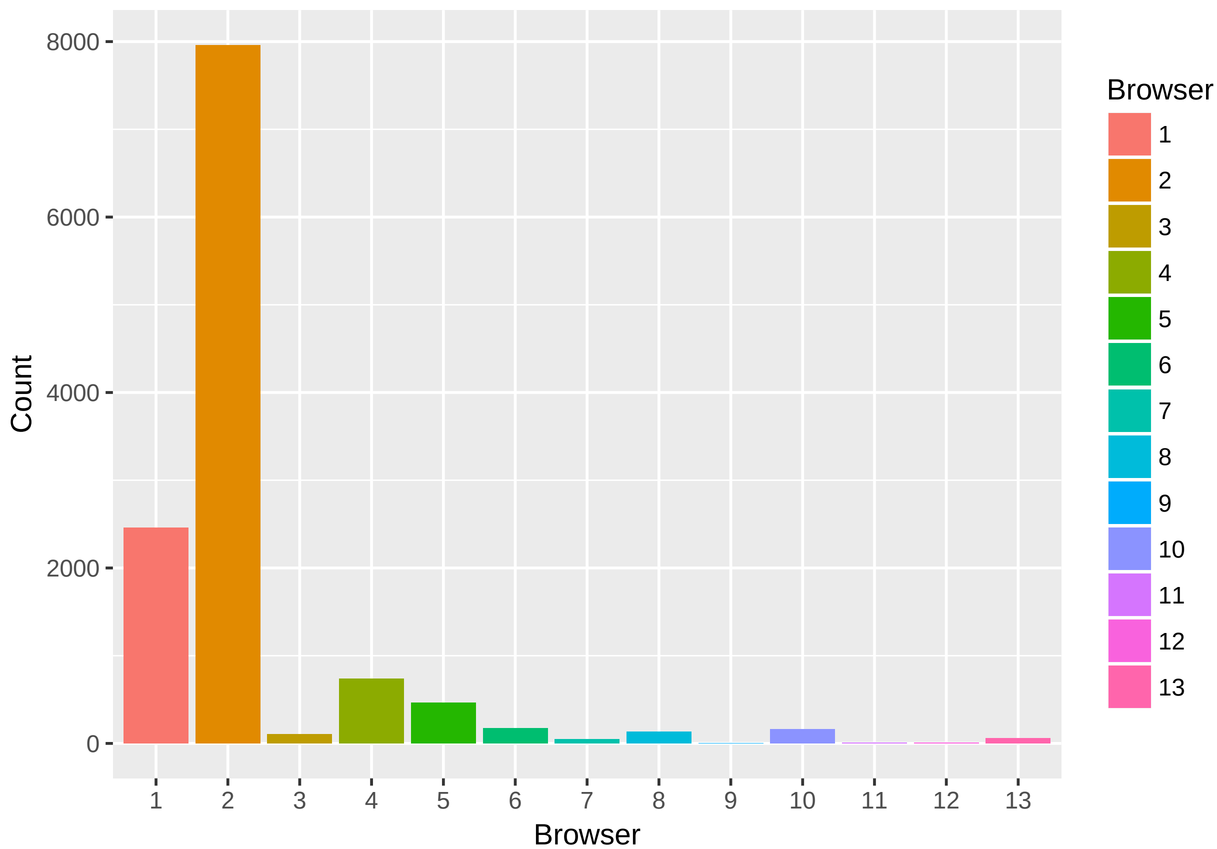
Let's look at month information. Data is not evenly distributed between month, for some of them we have very little information. For instance, classification model might learn, that there is no point predicting a purchase for session in June, because chances are that no session from this month in training set will be part of our training set. Using this feature might cause overfitting.


Count
6000
4000
2000
0
1
2
3
4
5
6
7
8
OperatingSystems
OperatingSystems
1
2
3
4
5
6
7
8

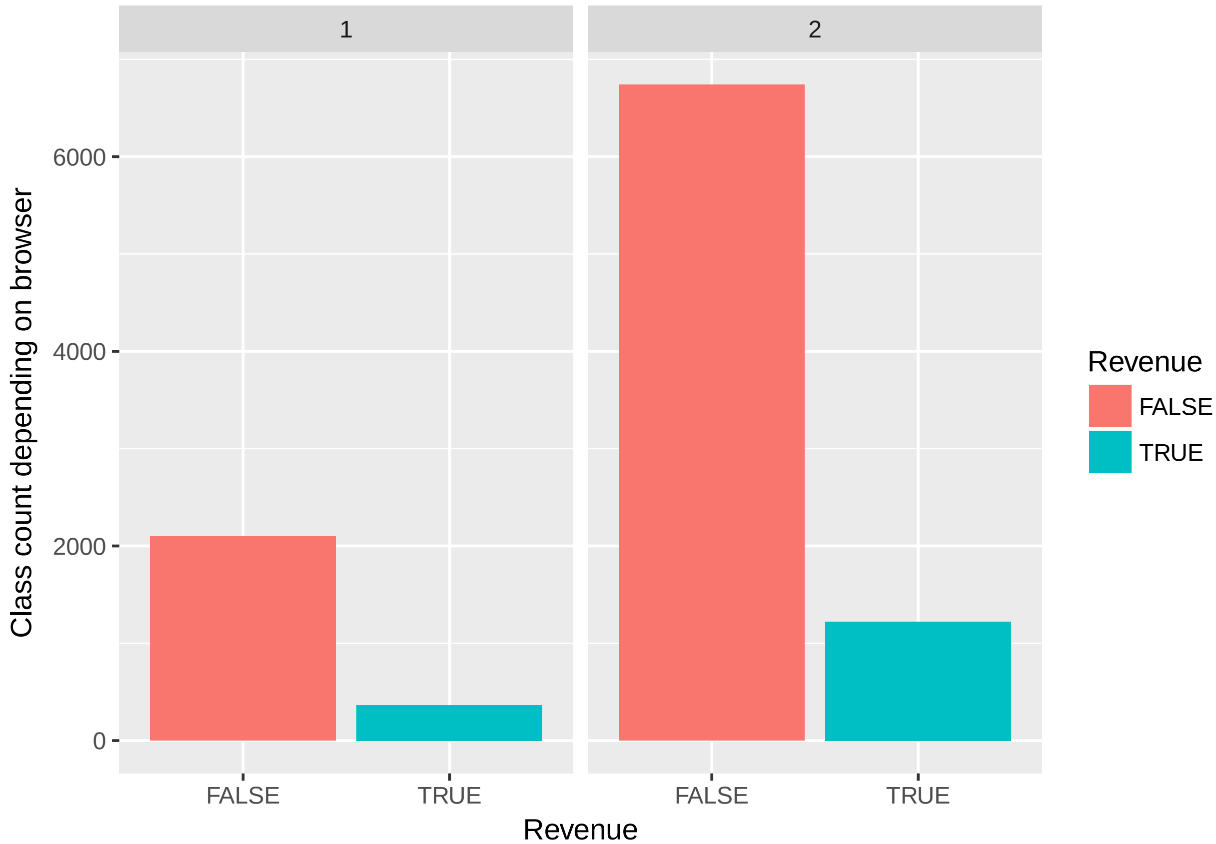

1
2
3
4
1000
500
0
Class count depending on OS
FALSE
TRUE
Revenue
Revenue
FALSE
TRUE

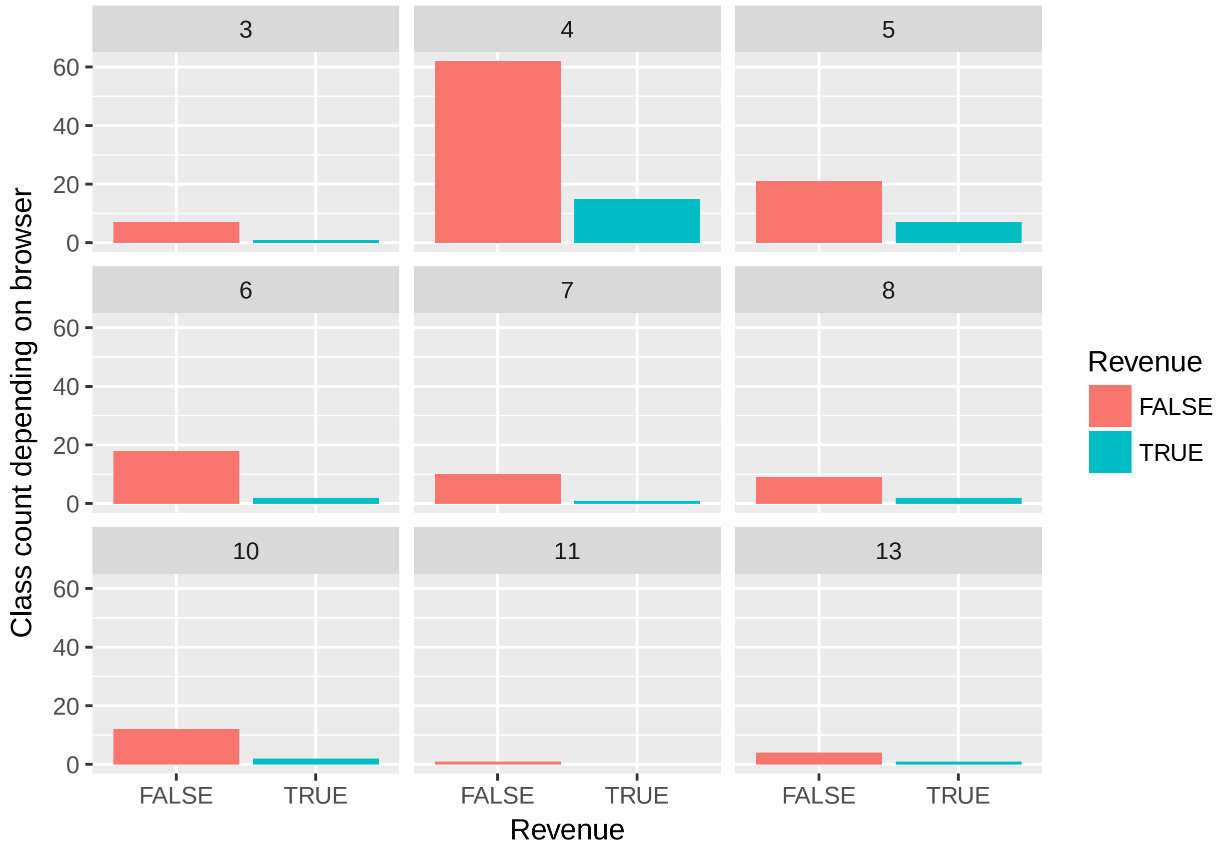


There is 8 kinds of operating systems used by our users. Some of them are much more popular, and some of them are very rare. It may make sense to group all rare categories into one, because alone they are not very informative.

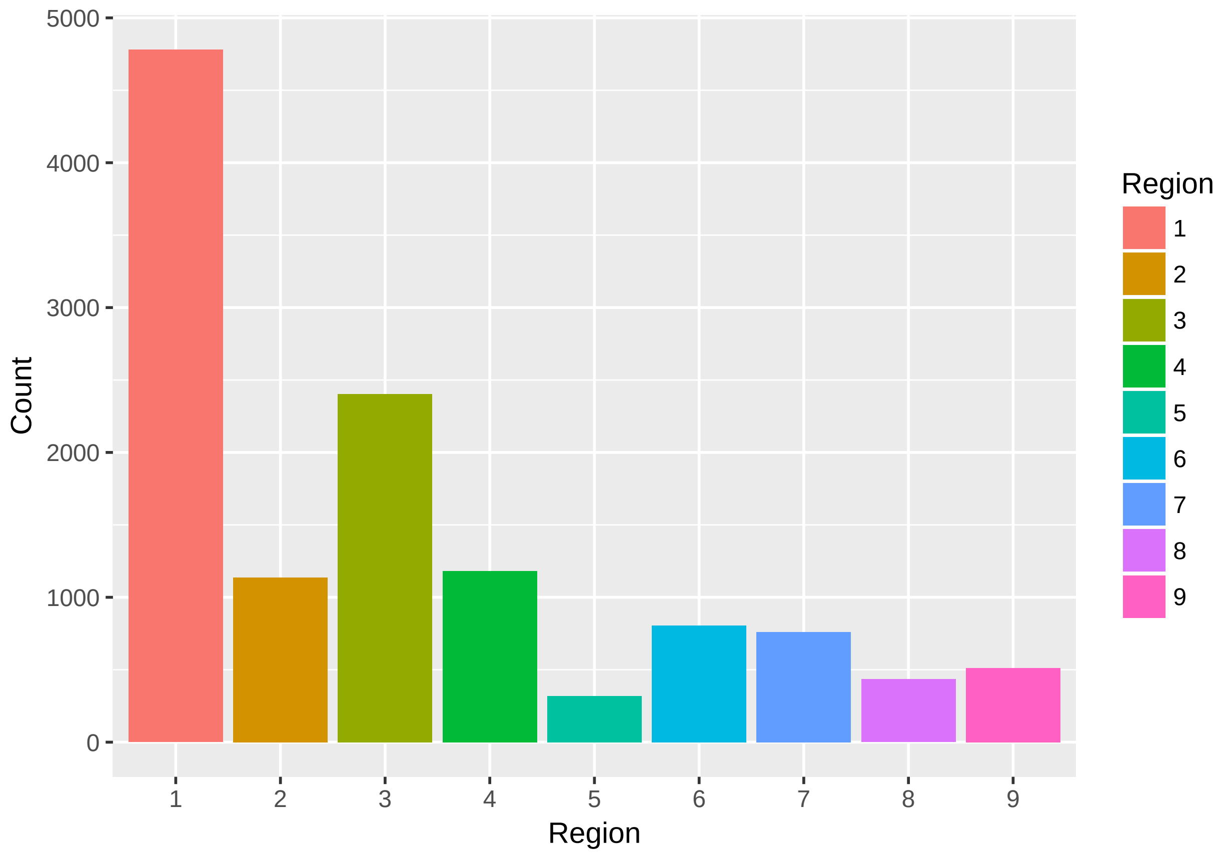


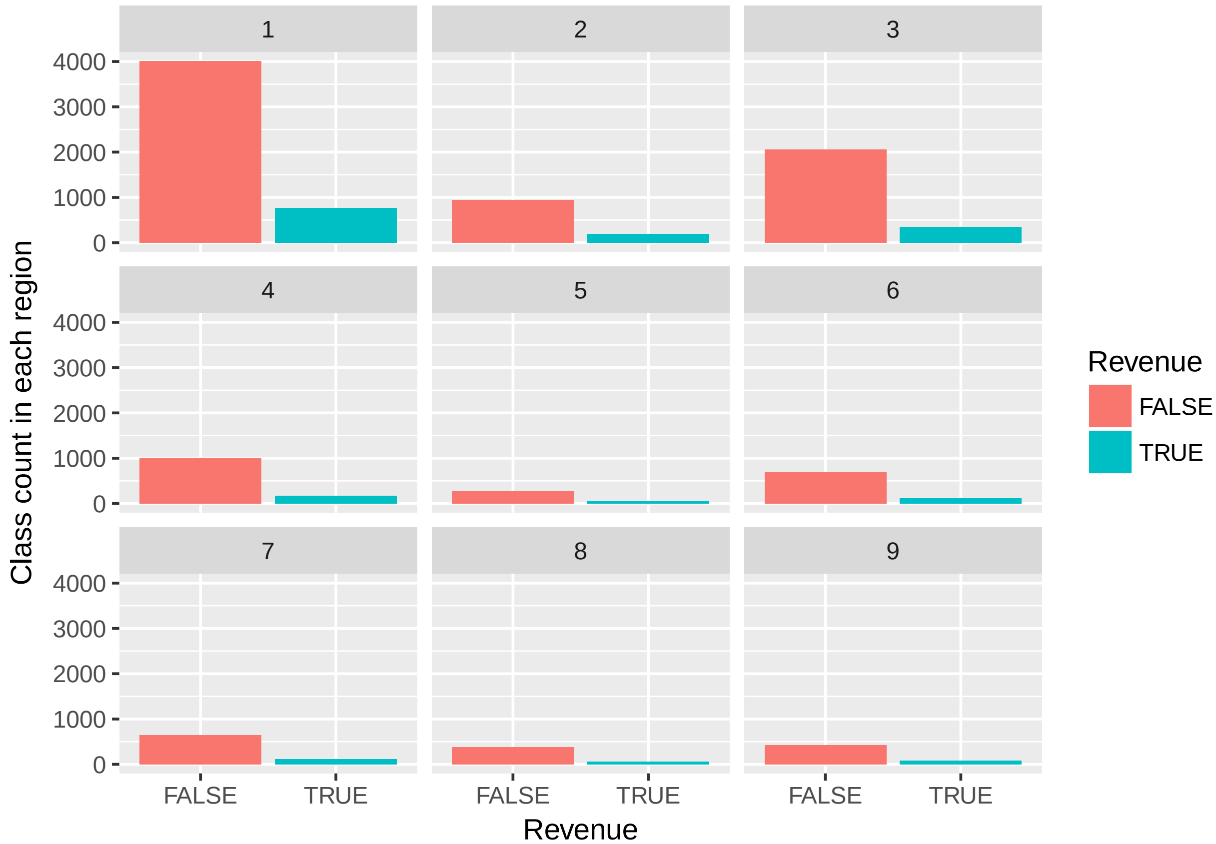
8000
6000
4000
2000
0
Count
1
2
3
4
5
6
7
8
9
10
11
12
13
Browser
Browser
1
2
3
4
5
6
7
8
9
10
11
12
13


1
2
6000
4000
2000
0
Class count depending on browser
FALSE
TRUE
FALSE
TRUE
Revenue
Revenue
FALSE
TRUE

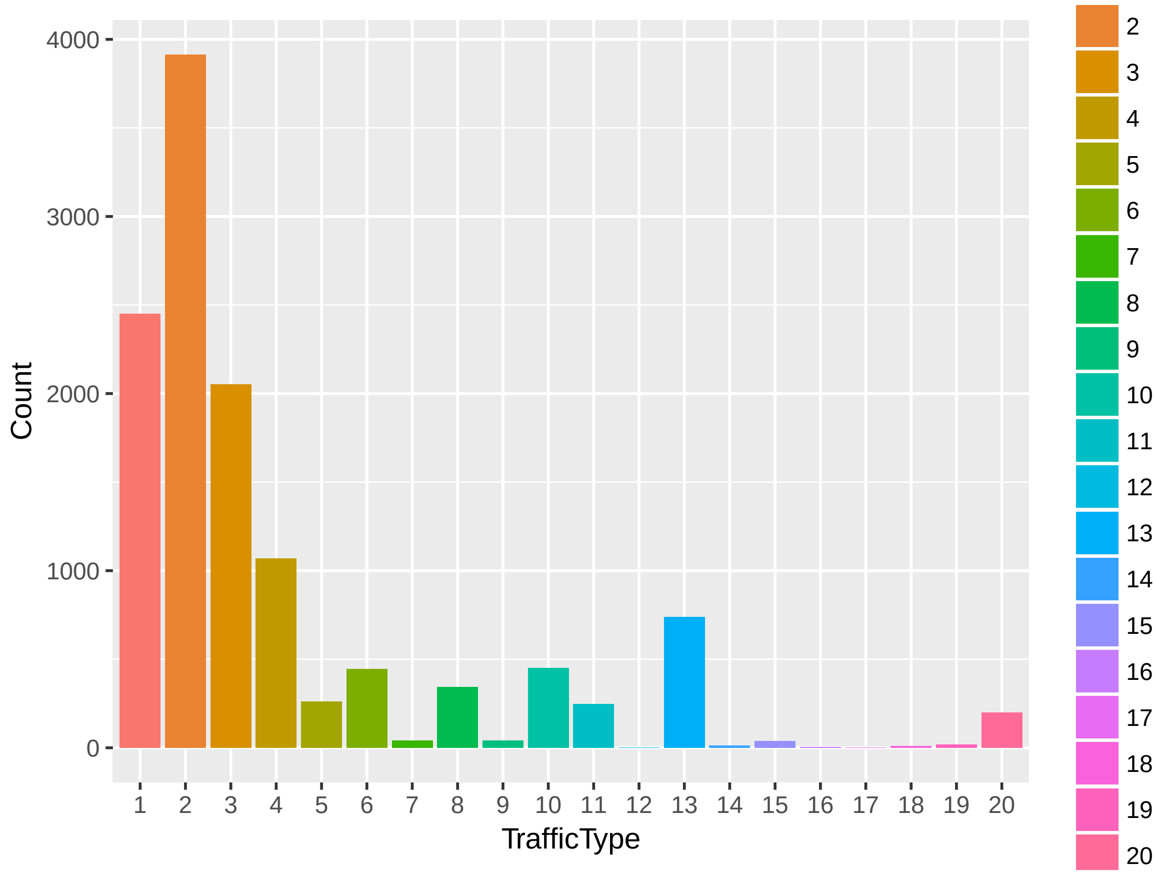


In case of browsers, we have two main categories, and quite a few less popular ones. We draw class distribution in each category to check if we may infer some class information depending on browser type. Once again, grouping rare categories might be desireble.

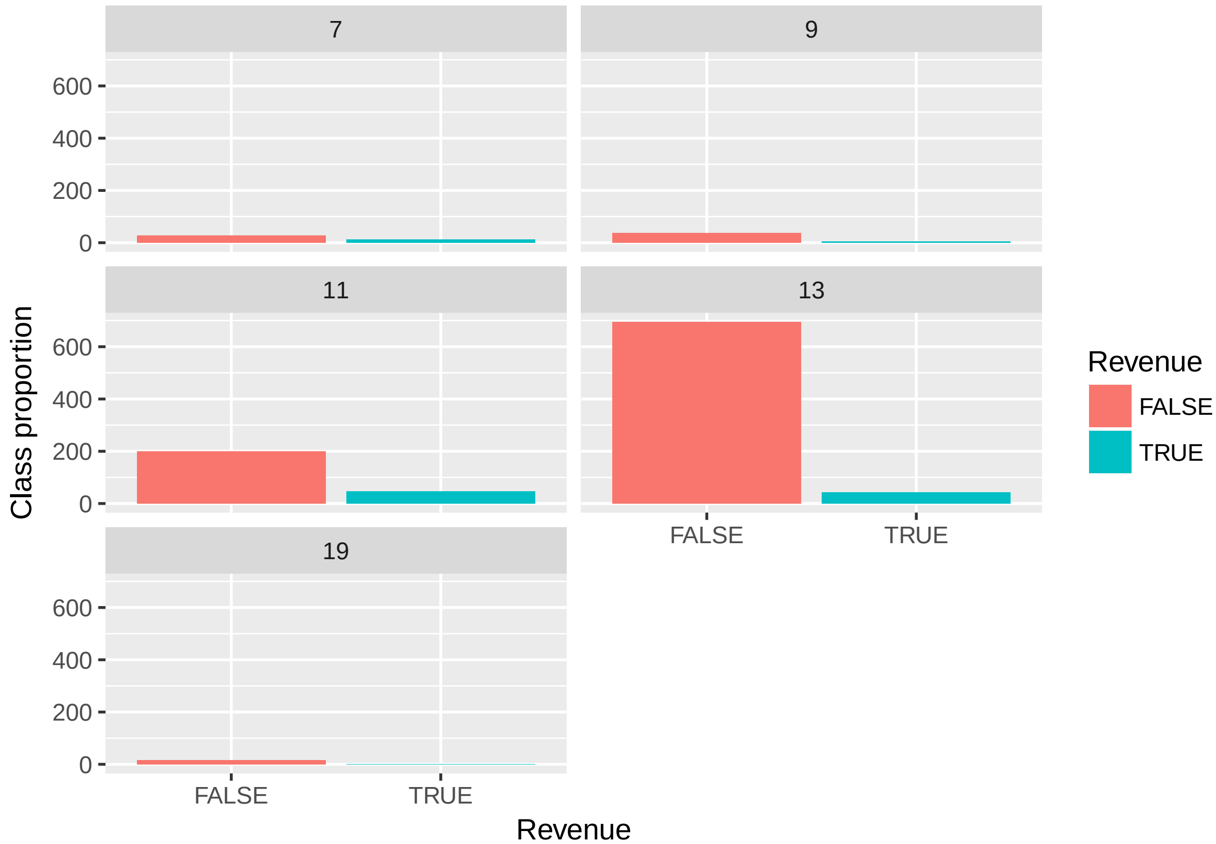

5000
4000
3000
2000
1000
0
Count
1
2
3
4
5
6
7
8
9
Region
Region
1
2
3
4
5
6
7
8
9



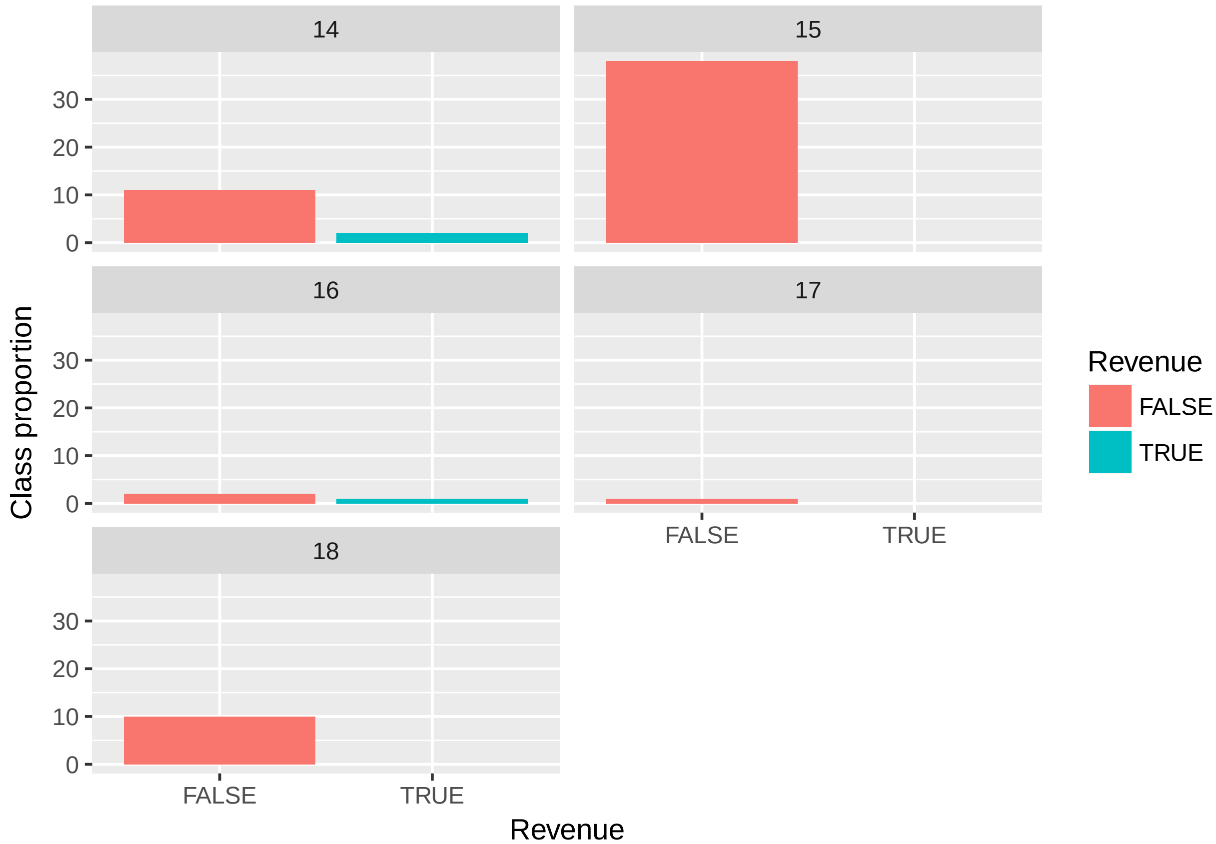
Region distribution is more even compared to OS and browser. Still, class distribution in each region category doesn't seem to be very helpful in classification.


4000
3000
2000
1000
0
Count
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
TrafficType
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

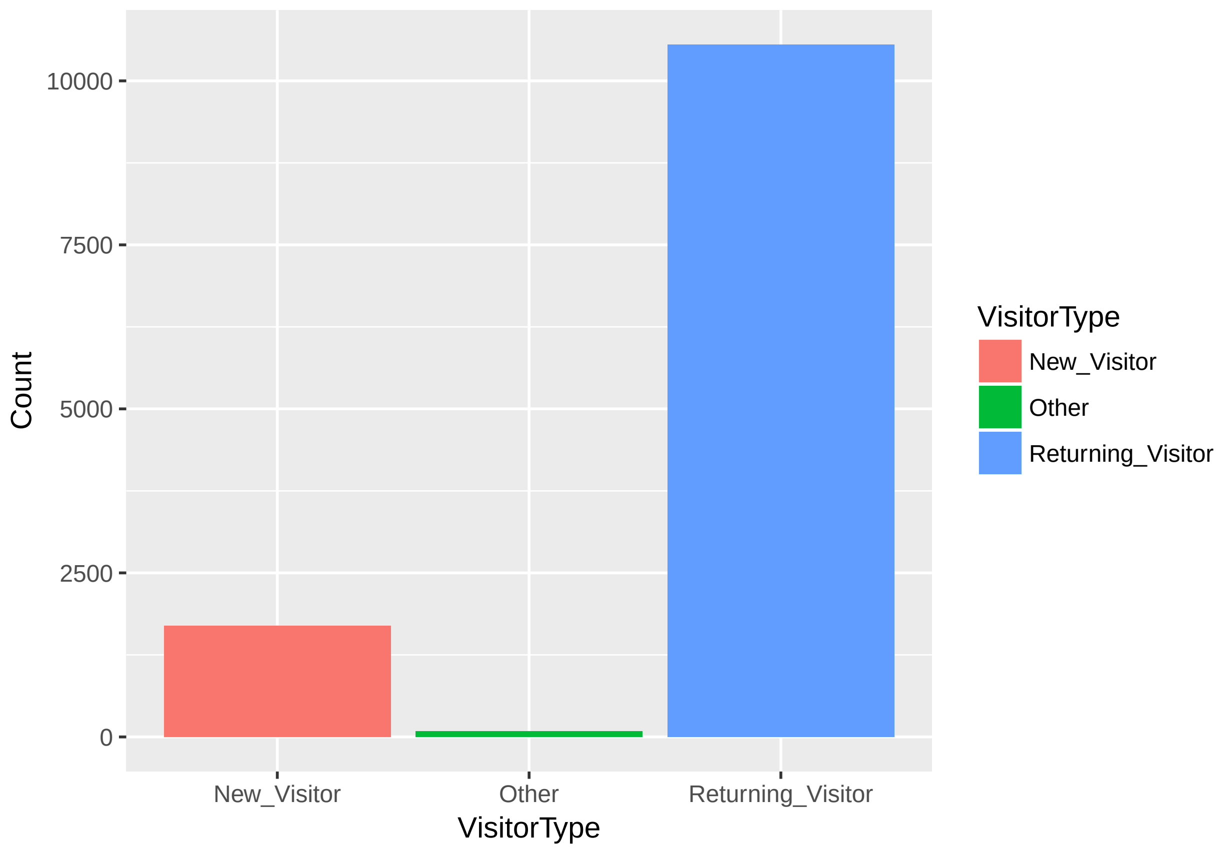
Traffic type distribution is very uneven. This variable might be interesting, especially rare categories might represent untypical user, i.e. administrator or developer. We take a closer look below.



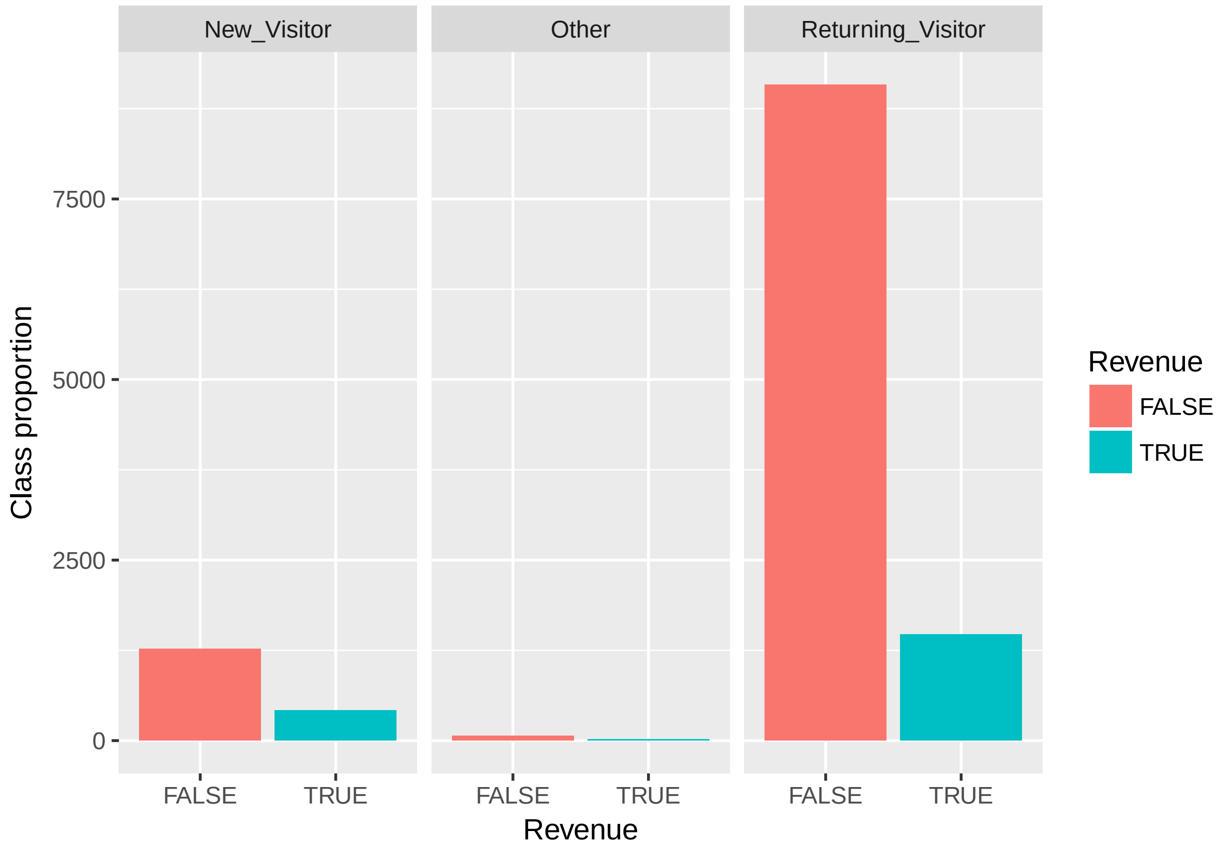
7
9
11
13
19
600
400
200
0
Class proportion
FALSE
TRUE
Revenue
Revenue
FALSE
TRUE



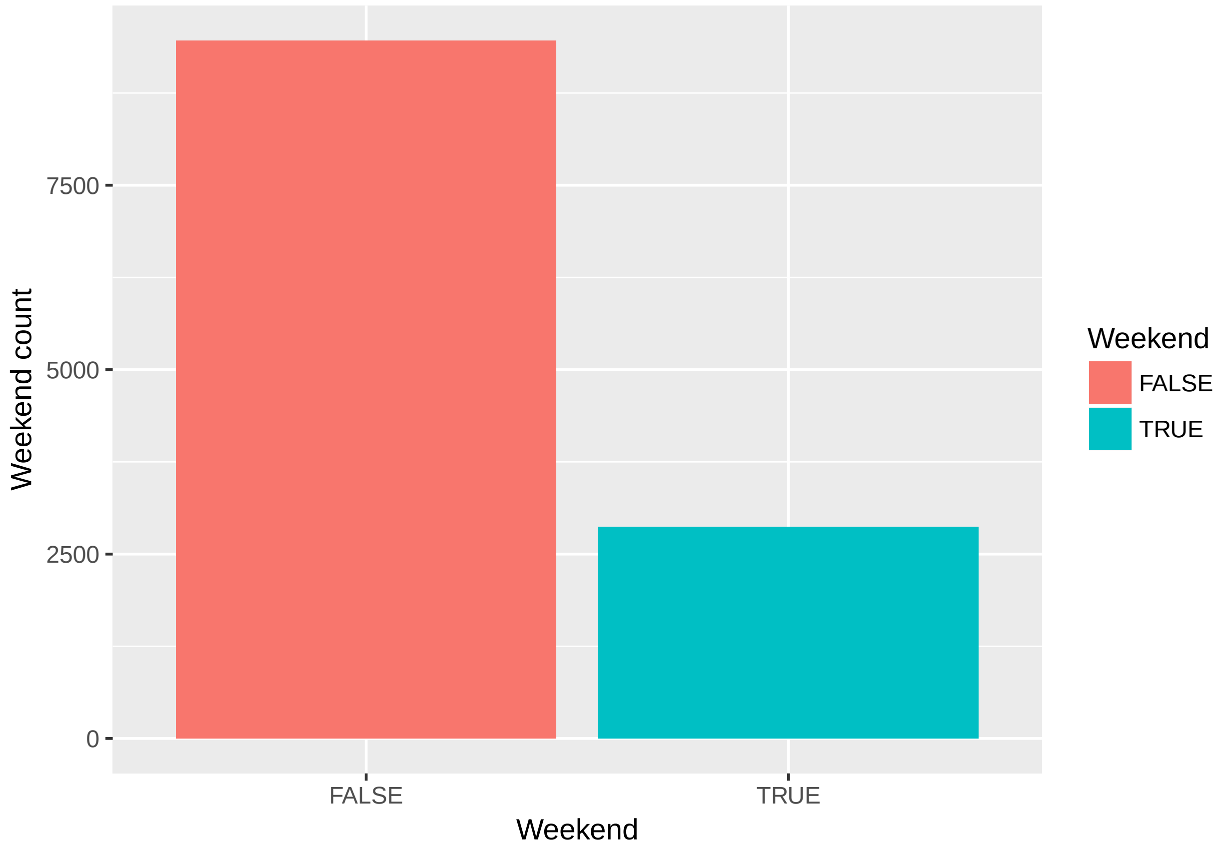
There is hard to reason what some traffic types might represent due to values annonymization.

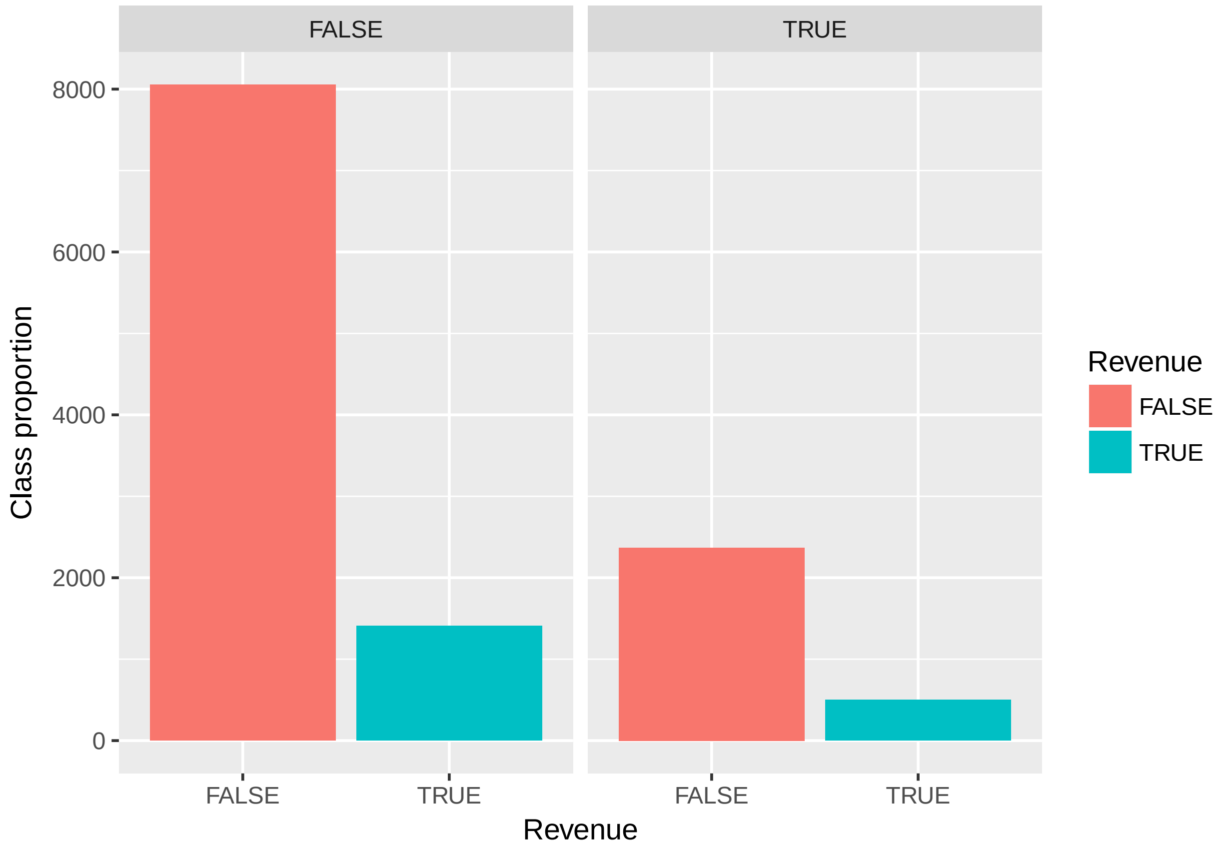


Count
10000
7500
5000
2500
0
New_Visitor
Other
Returning_Visitor
VisitorType
VisitorType
New_Visitor
Other
Returning_Visitor

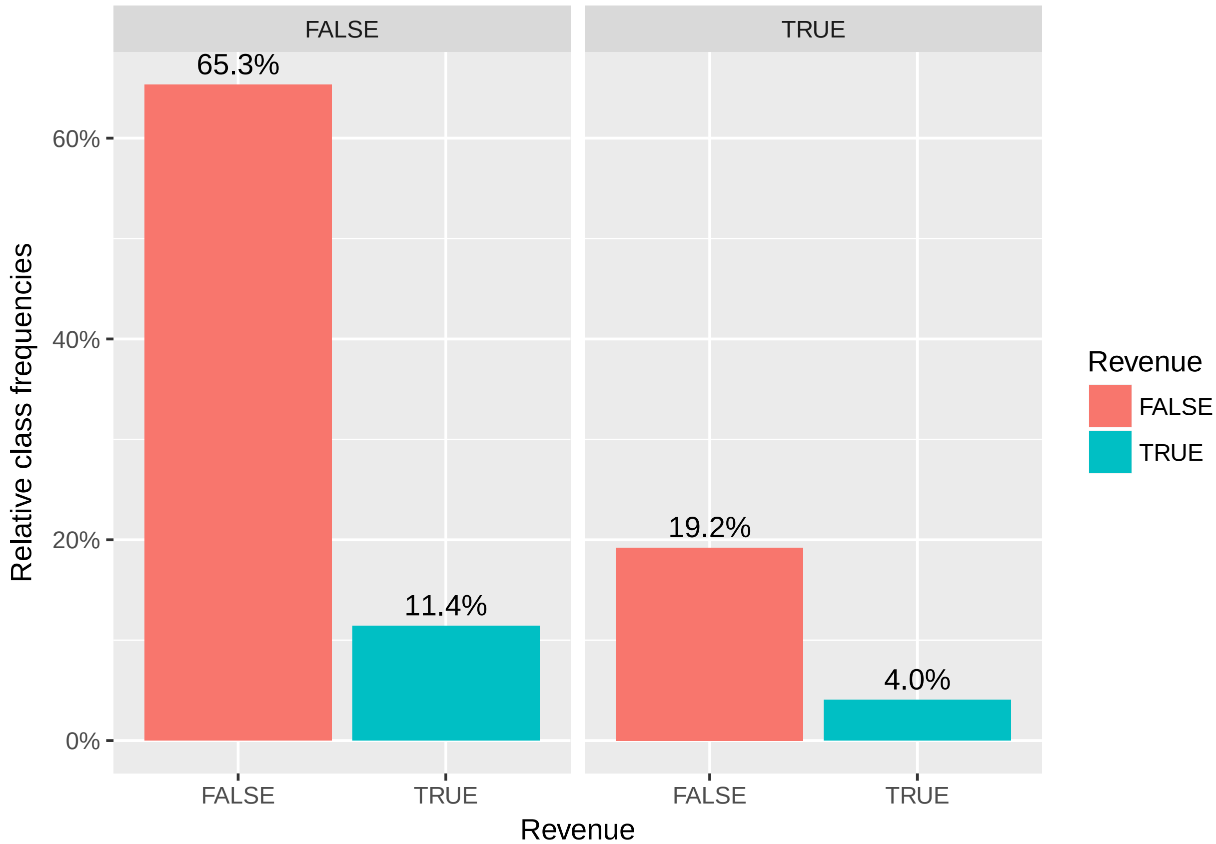


Most of the visitors are returning ones. Category "Other" looks suspicious, in this case we can be pretty sure it represents some abnormal type of users, like administrator or Google crowler, that is not likely to buy something!


Weekend count
7500
5000
2500
0
FALSE
TRUE
Weekend
Weekend
FALSE
TRUE



FALSE
TRUE
8000
6000
4000
2000
0
Class proportion
FALSE
TRUE
FALSE
TRUE
Revenue
Revenue
FALSE
TRUE


FALSE
TRUE
65.3%
11.4%
19.2%
4.0%
60%
40%
20%
0%
Relative class frequencies
FALSE
TRUE
FALSE
TRUE
Revenue
Revenue
FALSE
TRUE