

Dataset :

**Online Shoppers Purchasing  
Intention Dataset Data Set**

Objective :

**Identify key variables that best predict  
whether a user will buy or not ?**

# Introduction

## Source of the dataset:

<https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>

The dataset is composed of 12330 observations with 18 attributes each.

Each observation is a session of a user in a e-commercial website in a 1-year period of time.

The 'Revenue' attribute can be used as the label class. This attribute represent if the user for the session ended up buying something. The value is "True" if something has been bought during the session and "False" otherwise. The proportion of "False" value is around 84.5%.

## The goal are :

- To determine the most important variables and therefore the important parameters during a sale in order to increase the number of sales.
- Make a model that can predict user intention base on our available features with the most accuracy.

# Data processing

## Data Preprocessing

In our data, we have 3 categories of features and a target variable

User navigation data (red)

Webpage related info (green)

User profile (orange)

Revenue (Target variable)

The number of rows and columns are (12330, 18)

The columns are :

Administrative	int64
Administrative_Duration	float64
Informational	int64
Informational_Duration	float64
ProductRelated	int64
ProductRelated_Duration	float64
BounceRates	float64
ExitRates	float64
PageValues	float64
SpecialDay	float64
Month	object
OperatingSystems	int64
Browser	int64
Region	int64
TrafficType	int64
VisitorType	object
Weekend	bool
Revenue	bool


# Data processing

## User navigation data

Float and int type variables, they take a large amount of value.

“Administrative”, “Informational” and “ProductRelated” give us the information on the number of pages of this type visited by the user during his session.

The duration columns give us the information about the time spent on it, foreach page type.



Columns - number of unique value		
Administrative	27	int64
Administrative_Duration	3335	float64
Informational	17	int64
Informational_Duration	1258	float64
ProductRelated	311	int64
ProductRelated_Duration	9551	float64

## Data processing

Webpage related info

Information about the pages visited by the user during his session. Note that this information is the average of each pages visited by the user during his session.

The value of those three variables are harvested using Google Analytics.

The “BounceRate” for a page is the number of time a user enter the website though the page and then exit the web site divided by the number of time a user enter the website though the page and then go on a different page on the same web site. It's a percentage.

The “ExitRate” for a page is the number of times a user exit the web site though this page divided by the number of times the page was used. It’s also a percentage.

Link for a deeper explanation :

[https://support.google.com/analytics/answer/2525491?hl=en&ref\\_topic=6156780](https://support.google.com/analytics/answer/2525491?hl=en&ref_topic=6156780)

The “PageValues” for a page is the mean value of each session during which this page was visited. The value of a session is, for the owner of the web site, the value of the transaction if during this session the user bought something added to the value of each “Goal Page” visited by the user.

The value of a “Goal page” is defined by the owner of the web site. It’s a float type variable.

<https://support.google.com/analytics/answer/2695658?hl=en>

Columns - number of unique value

BounceRates	1872	float64
ExitRates	4777	float64
PageValues	2704	float64

# Data processing

## User profile (1/2)

Information about the user of the session.

“SpecialDay” represent the closeness to special day like Christmas for example. It’s a float type variable.

“Month” represent the month when the session take place. It’s a string type variable.

“OperatingSystems” tells us about the OS use by the user for the session. It’s an int type variable.

“Browser” tells us about the browser use for the session. It’s an int type variable.

Note that “Month” only have 10 unique values, we are missing January and April. We will take that as normal for the rest of our study.

For “SpecialDay”, it goes from 0 to 1 from 0.2 to 0.2.

Columns	number of unique value	
SpecialDay	6	float64
Month	10	object
OperatingSystems	8	int64
Browser	13	int64



# Data processing

## User profile (2/2)

“Region” tells us about place the user were during the session. It’s an int type variable.

“TrafficType” indicate the traffic use. There is 20 kind of possible traffic. It’s an int type variable.

“VisitorType” point out if the user is new or not on the web site. It’s a string type variable.

“Weekend” tells us if the session was during the weekend or not. It’s a Boolean variable.

For “VisitorType”, we have 3 unique values, but this does not seems normal because the feature is describe as a binary variable (the visitor is new or not new).

Columns - number of unique value

Region	9	int64
TrafficType	20	int64
VisitorType	3	object
Weekend	2	bool

# Data processing

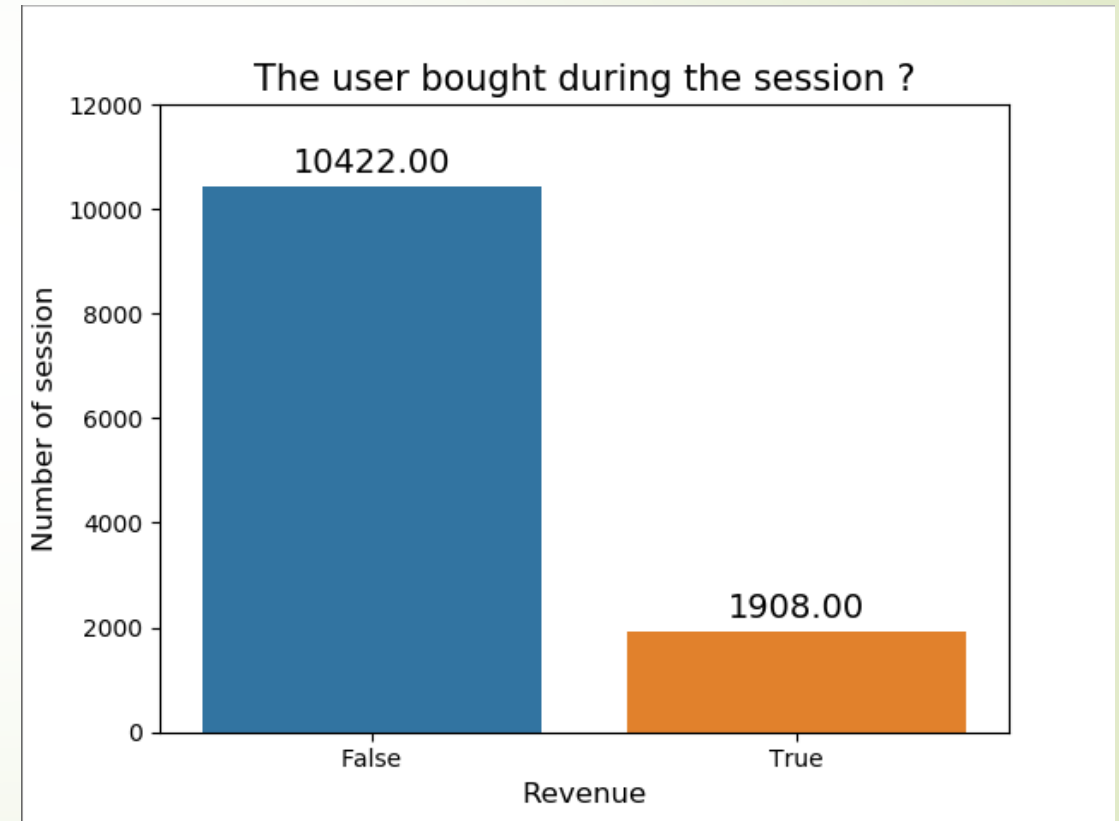
Target variable "Revenue"

"Revenue" tell us if the user bought something during the session.

We have the right amount of True/False type.

We can see that a large majority of user don't buy when using the web site. This is normal, like any other commercial web site or in any store, in most of the cases, users just want to see what is available and at what price.

It's a Boolean type variable.



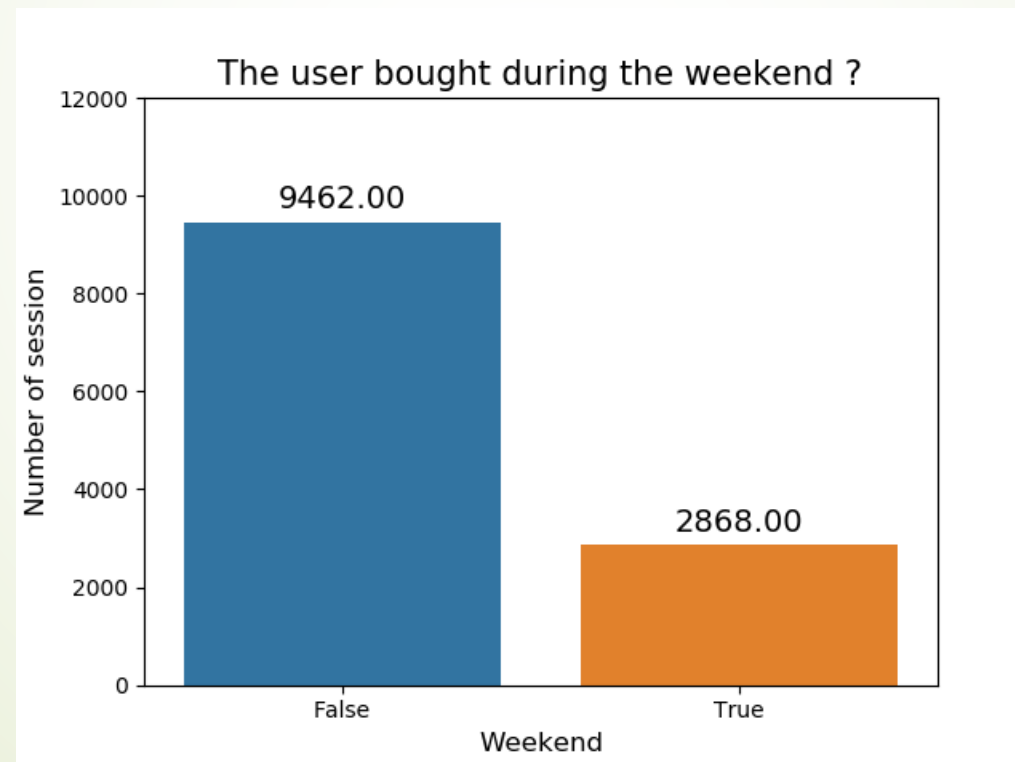


# Data processing

## “Weekend”

Show that a majority of session are during the week, and there is less sessions during the weekend

Considering that the weekend lasts 2 days, a majority of session are during the week, and there is slightly less sessions during the weekend (we miss around 700 sessions for a perfect balanced).



# Data processing

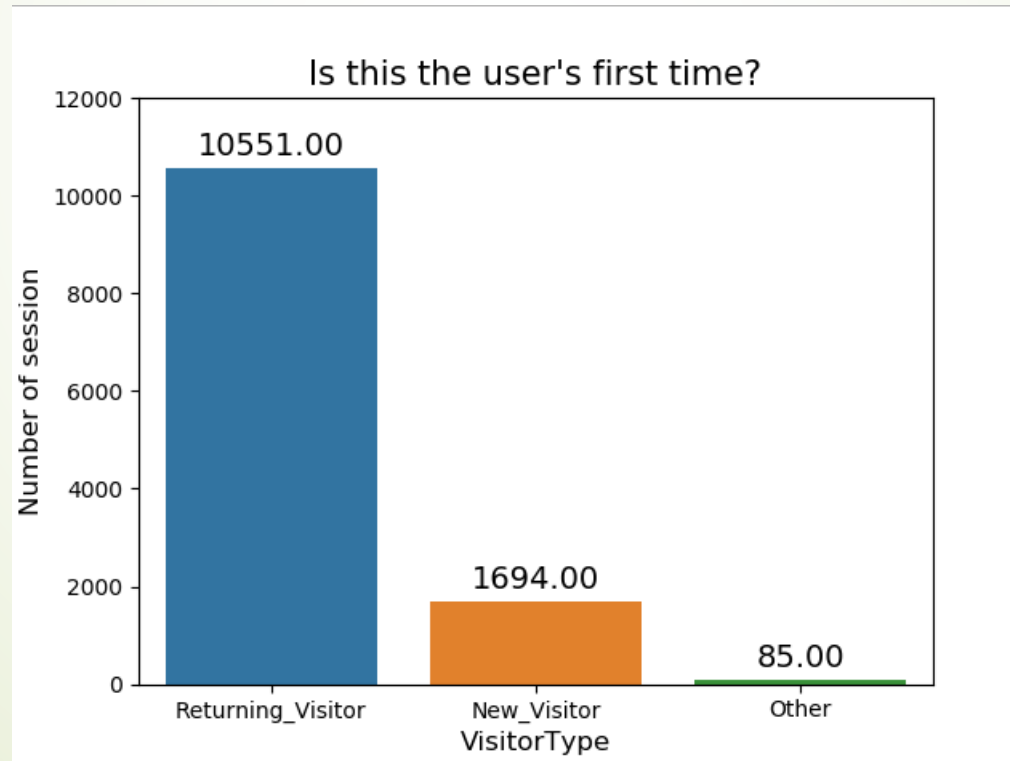
## “VisitorType”

Show that for a 1 year period, around 1700 unique users visited the web site.

We don't have the type of product this web site is selling, but based on that it may show a lack of visibility.

On the other hand, people seems willing to return on the web site.

We also have a third category, “Other”, for the rest of our study, we will consider that they are administrator or owner and don't really bring any value for our study.



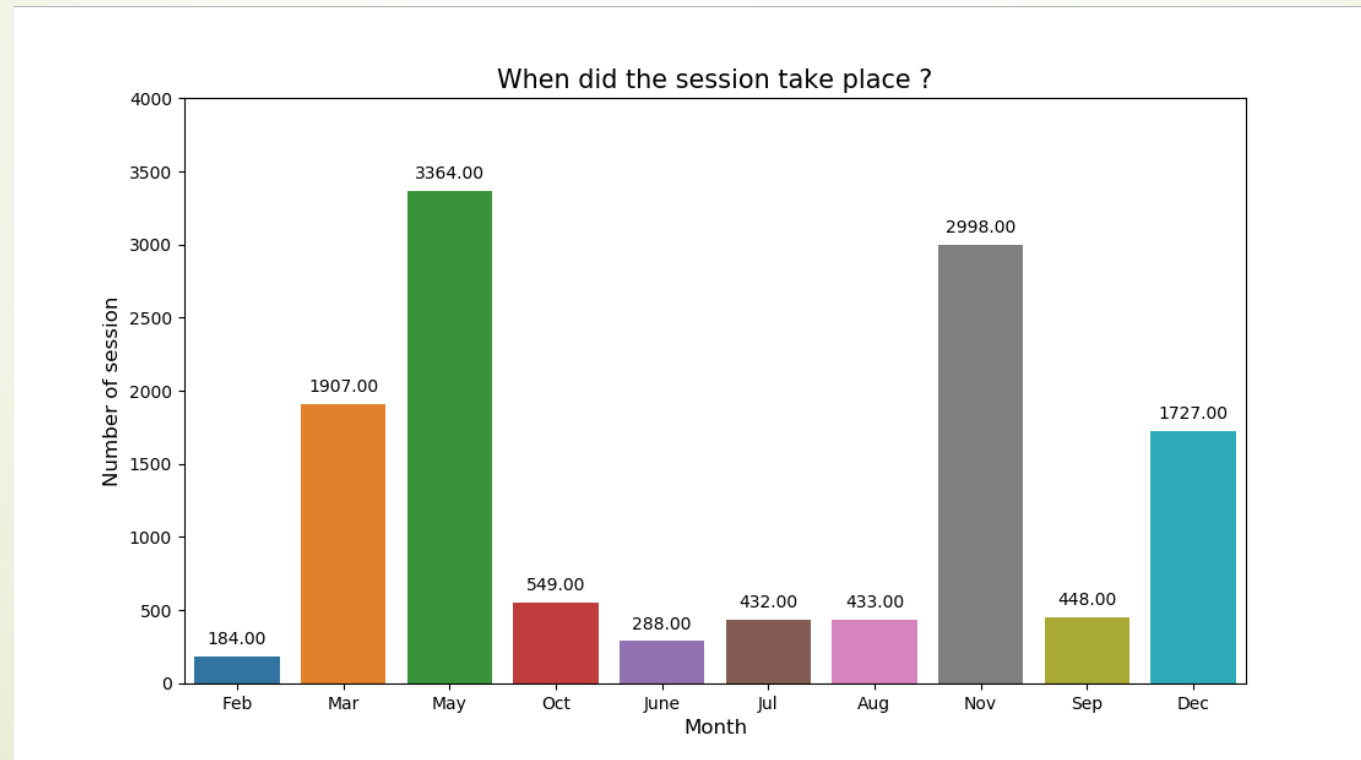
# Data processing

“Month”

Show us the number of sessions for each month. We can see that the majority of session take place during May, November, March and December.

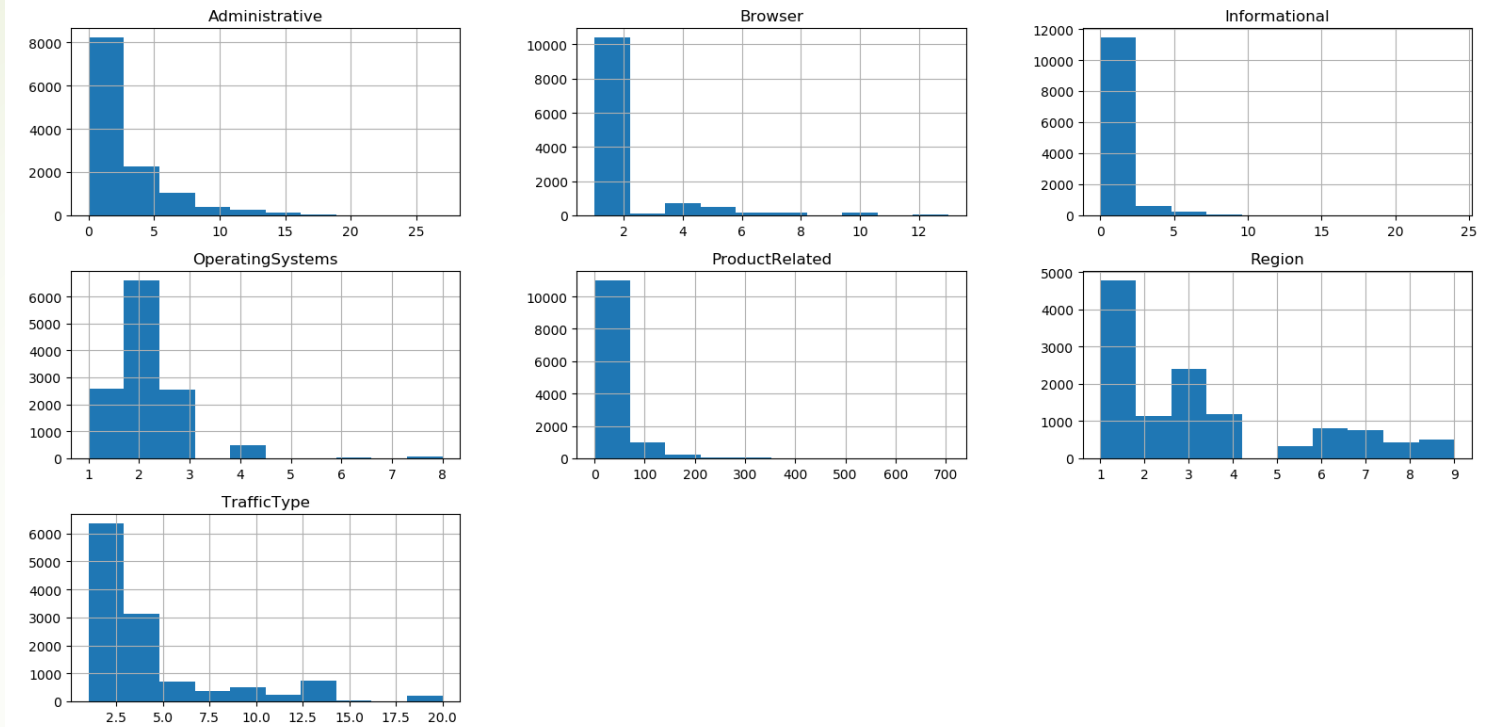
One way to explain those number are that the product sell here is seasonal.

And of course, we are missing 2 whole months, this can be a huge loss in information, but because we don't have any more information about that, we will consider that it is normal.



# Data processing

## Int-type variable

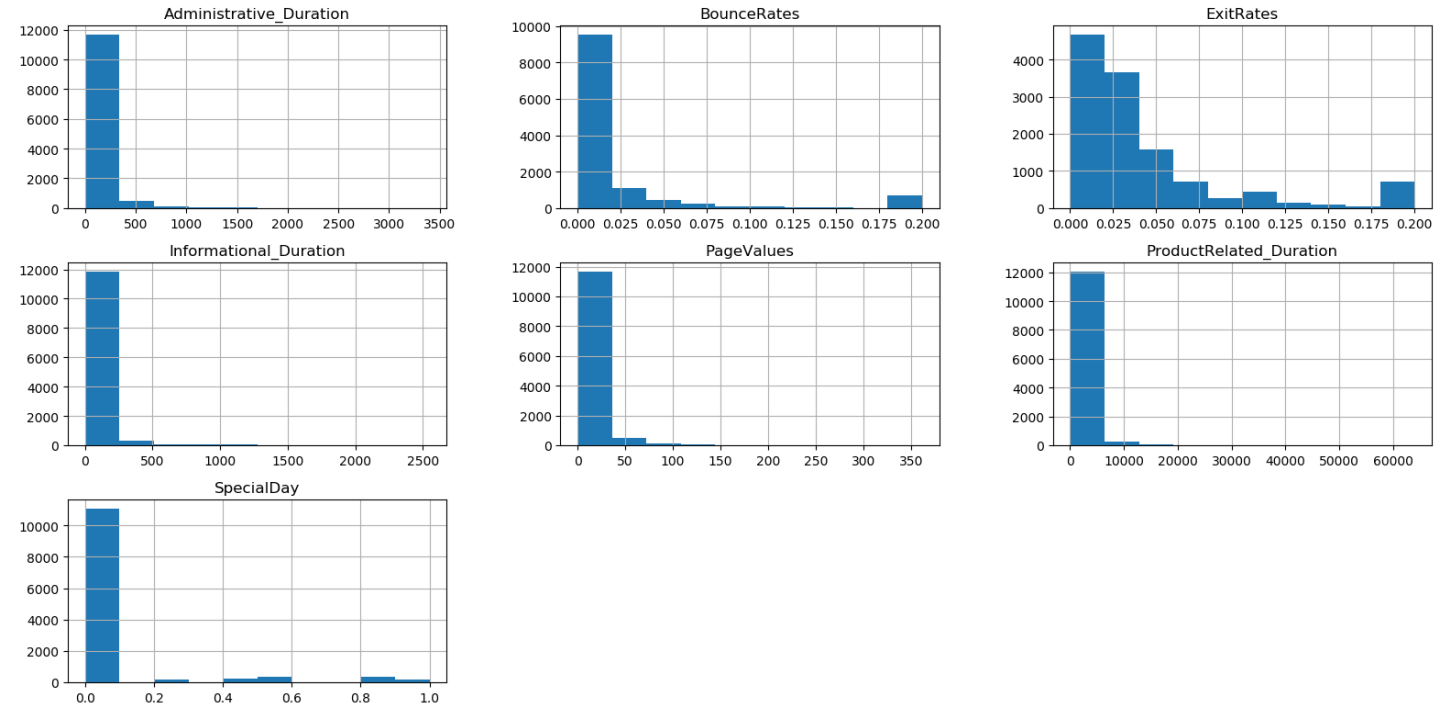


From pages variables, we can see that a majority of users use a very little number of pages, even 0 for most cases. Maybe because of pop-ups or other advertising method.

In majority, users focus on product related pages more than other pages. They, for the majority use the same browser, and same OS and traffic type.

# Data processing

Float-type variable



Same conclusion as before for duration type variable.

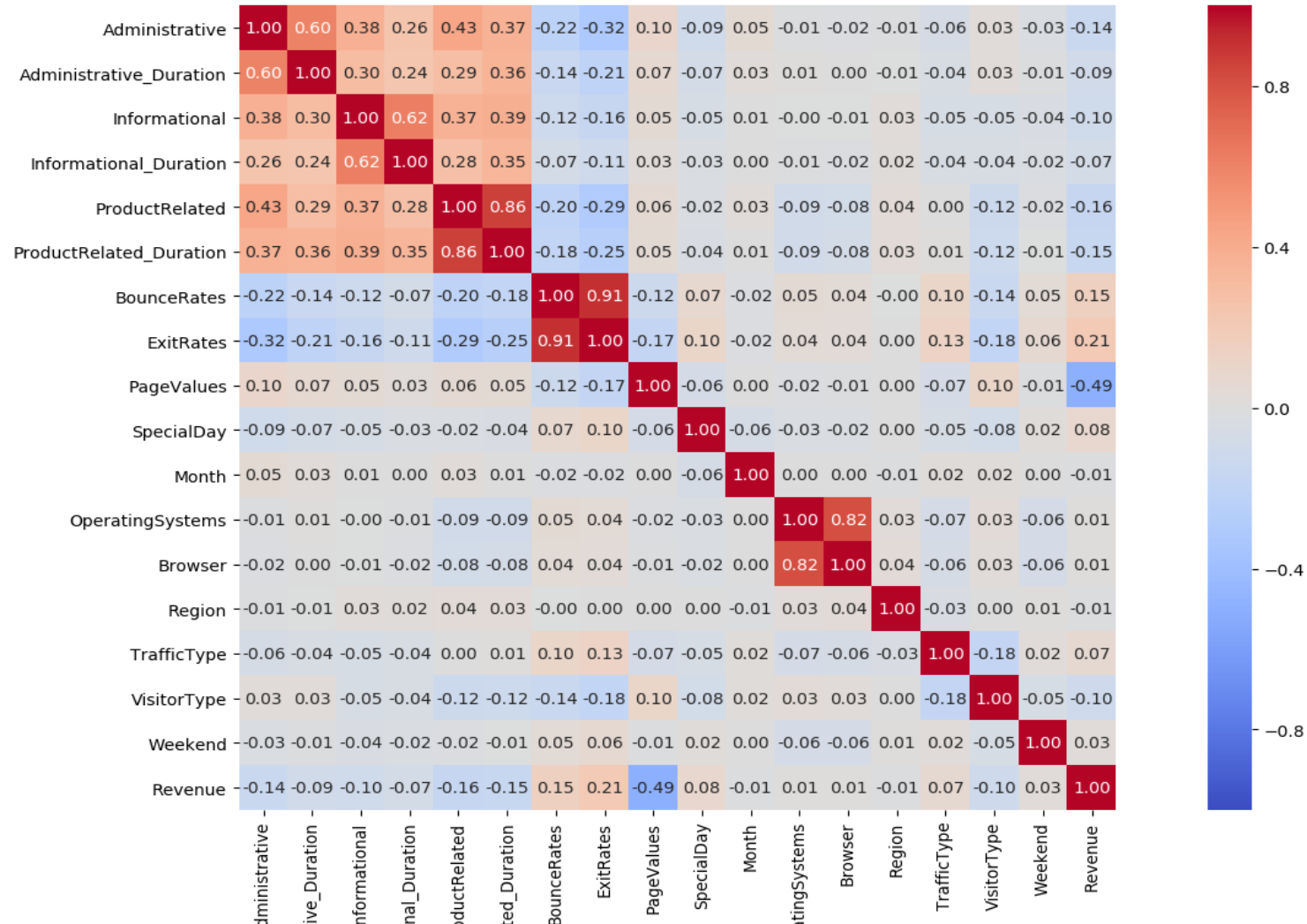
According to “SpecialDay”, the majority of sessions take place during normal day.

For the variables about the pages “BounceRates” and “ExitRates”. Considering that a user have to view multiple pages during a session, a bounce rate under 20% seems to be a good thing.

# Data processing

## Link between variables

Show us the linear relationship between our column's variable.



We can see that “BounceRates” and “ExitRate” are very correlate. As “Browser” and “OperatingSystem” variables.

We can also note that all the variable about the type of pages view by the user, their number and the time he spent on them are all correlate. This seems normal, because the more page you view, the more time you are spending on the web site.

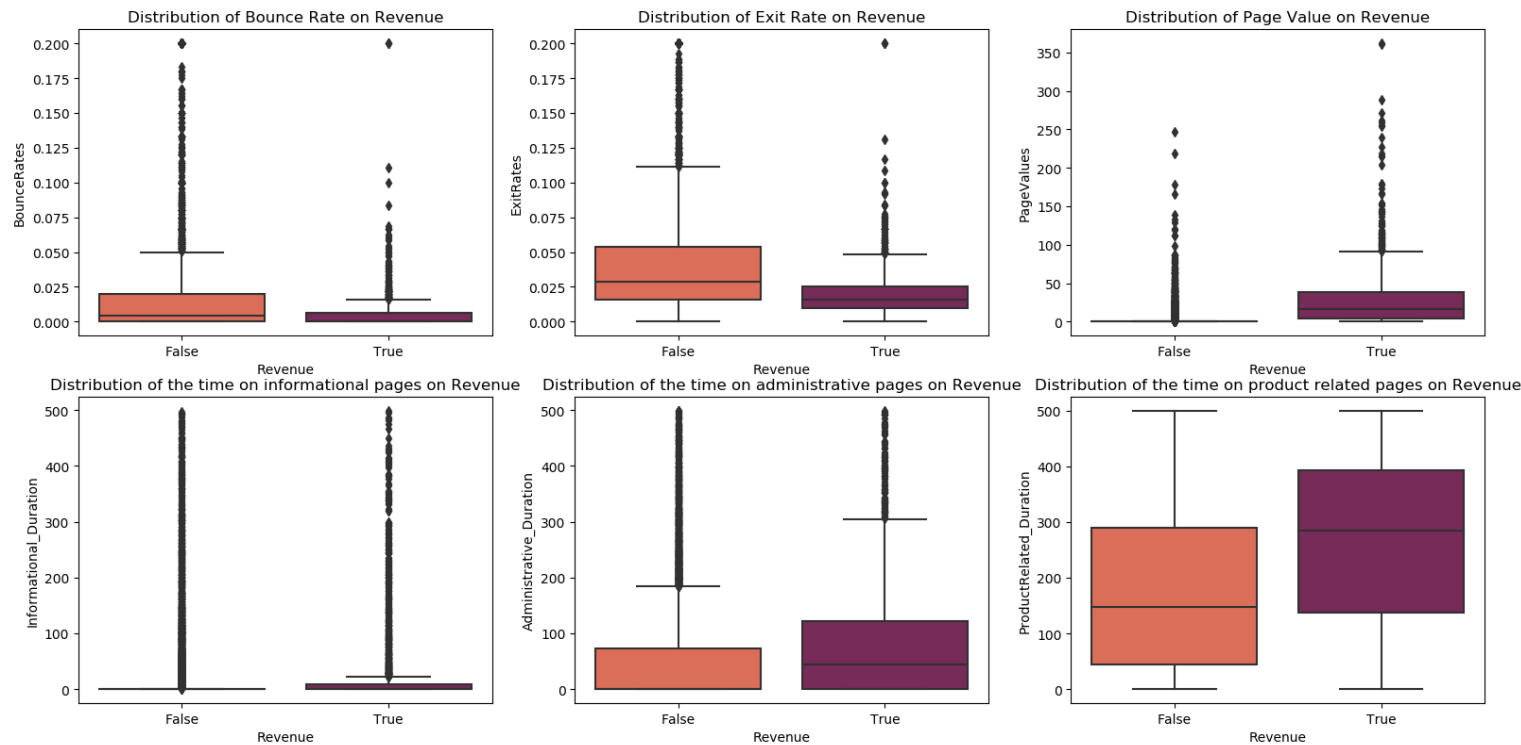
Finally, we can see that the target variable “Revenue” is the most correlate with the “ValuePages” variable.



# Data processing : link to “Revenue”

## Web related pages

For these graphs, we have removed the extreme values



These graphs show us that the page value for a user who has not purchased is very low and that it has a more significant value when the user has completed a transaction on the site.

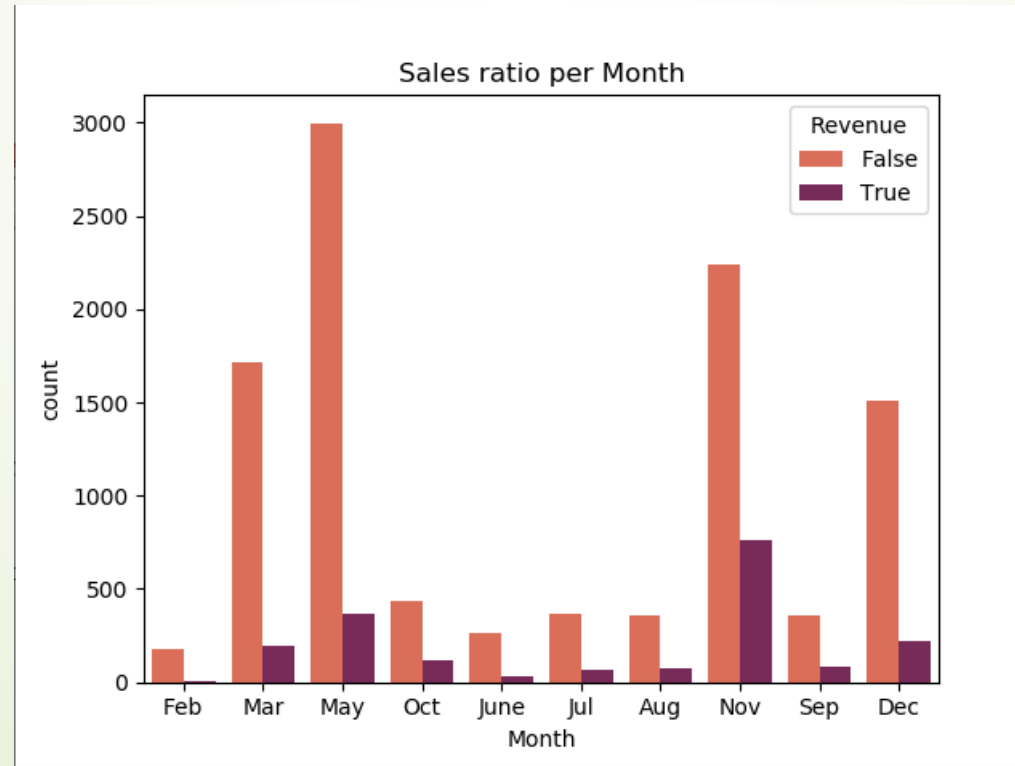
Note that the time a user spend on product related pages is way greater when the user actually purchased something.

Next, a user who don't bought something have a bigger “ExitRate”.

## Data processing : link to “Revenue”

### “Month”

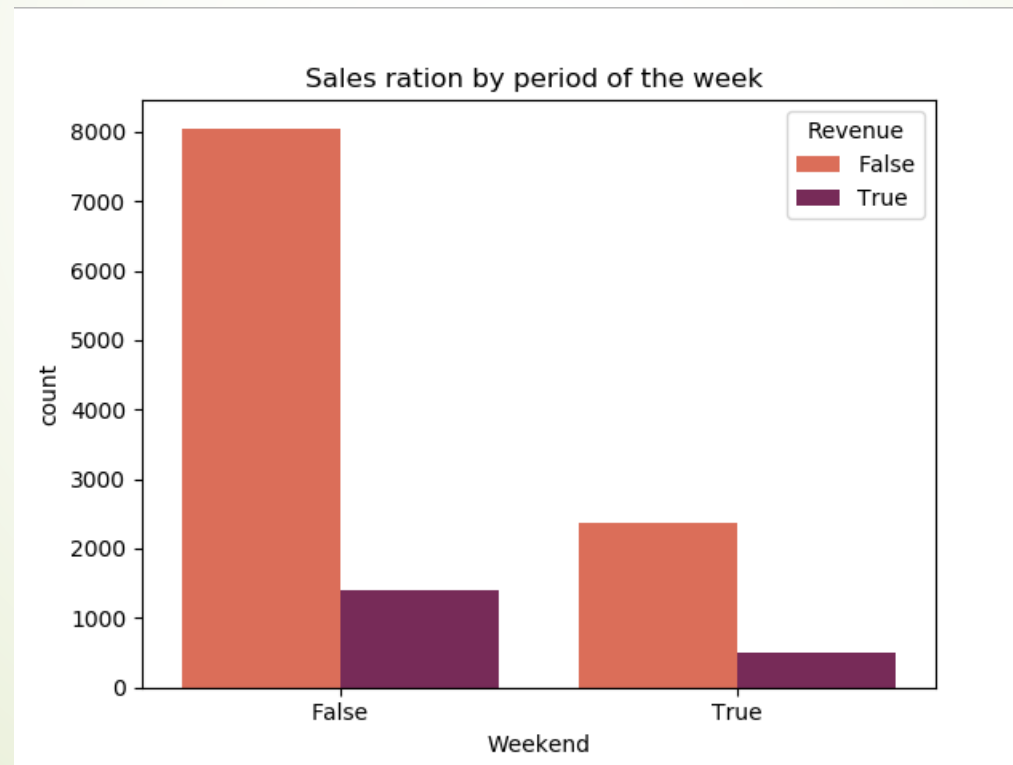
The most part of sales take place during the month of November. Even though the month with the bigger amount of session is May.



# Data processing : link to “Revenue”

## “Weekend”

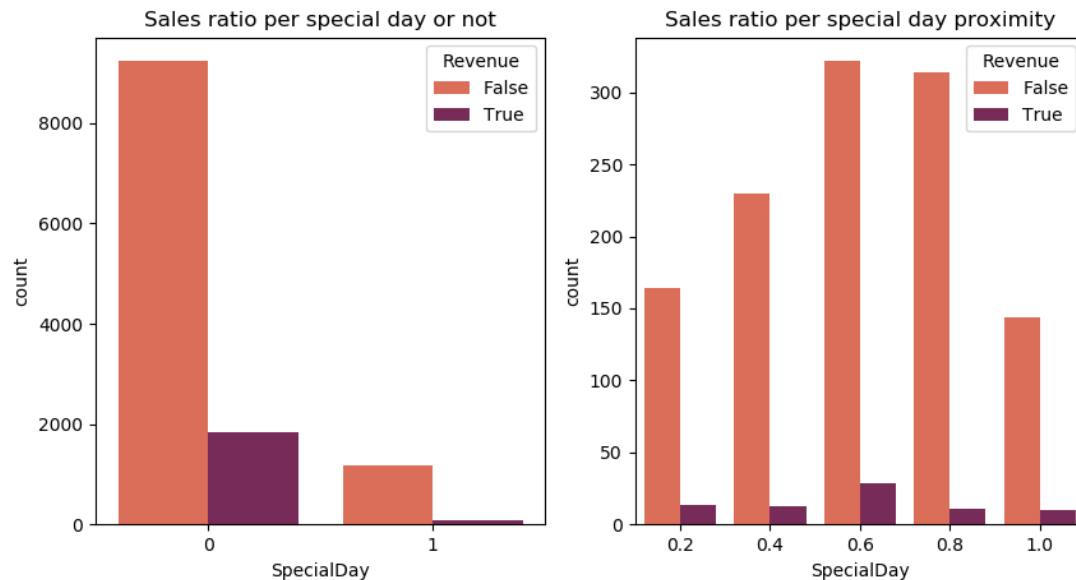
More transactions are carried out during the week.  
But otherwise, it's well balanced.



# Data processing : link to “Revenue”

## “SpecialDay”

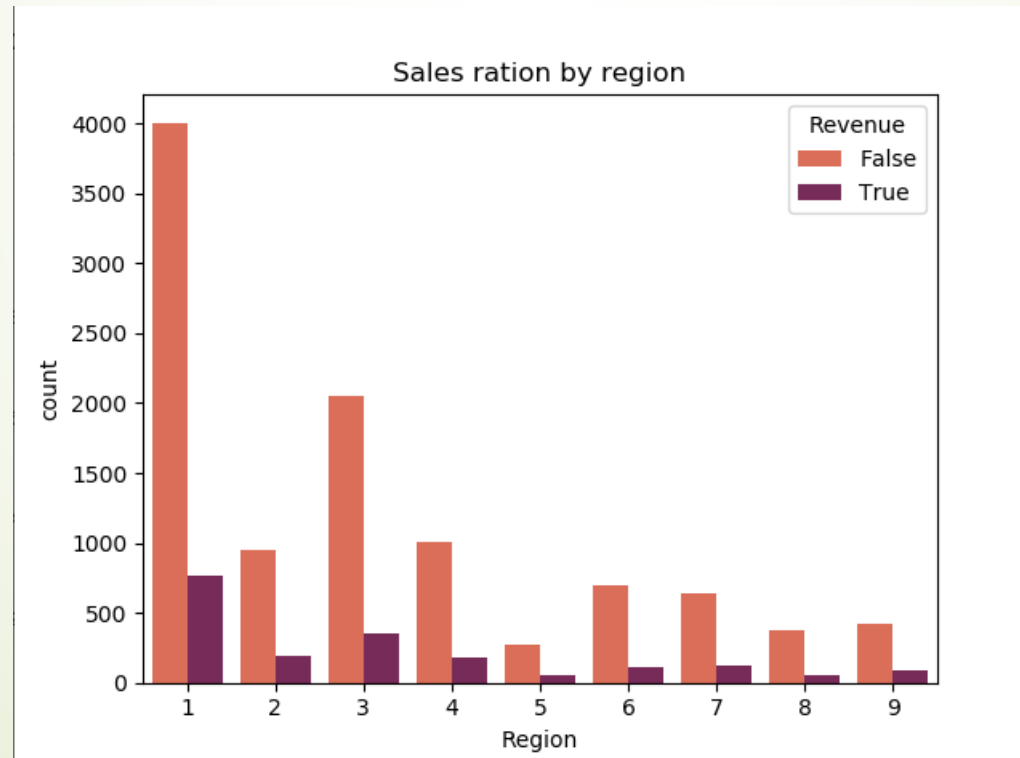
Special days don't seem to have an impact on sales.  
Moreover, when you consider only special days,  
proximity doesn't have a major impact either.



# Data processing : link to “Revenue”

## “Region”

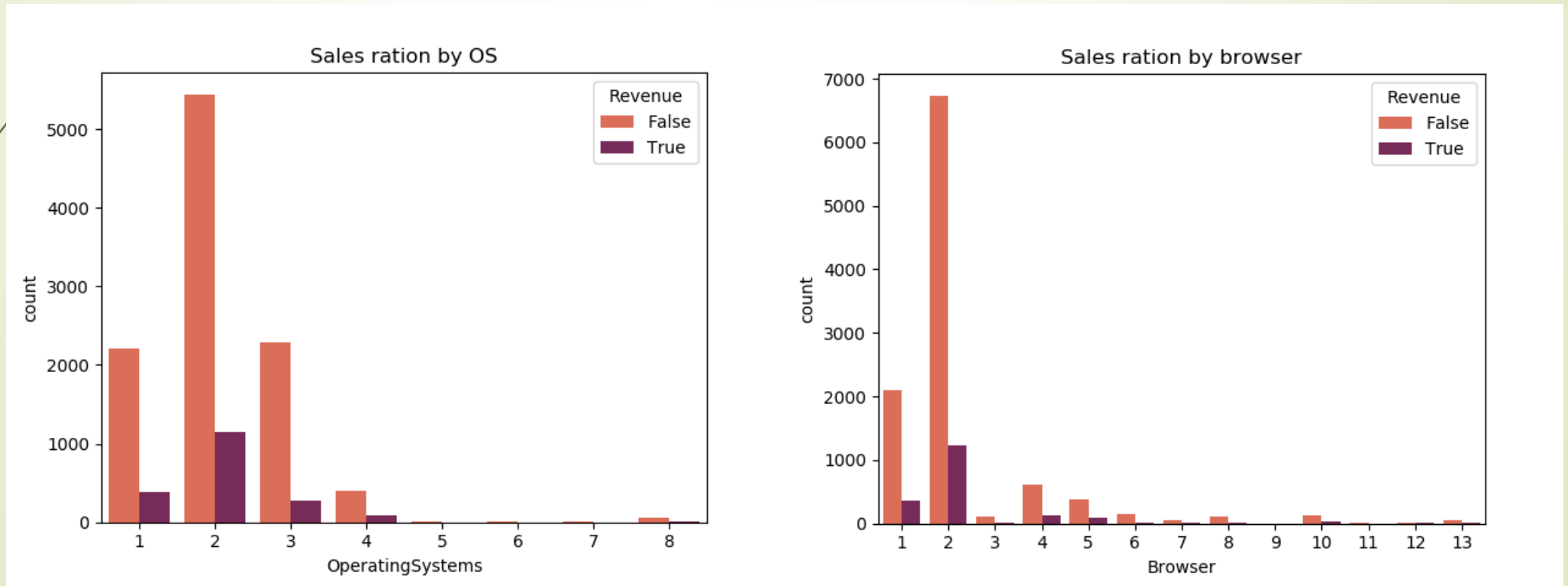
Most sales are made in region 1, which is also the place with the most sessions. There are sales in all regions and no great disparity between the number of sessions and the number of sales



# Data processing : link to “Revenue”

“OperatingSystems” and “Browser”

Most people use one of the three major OS and the same Browser. OS and Browser doesn't seem to have a huge impact on whether the user buys or not.

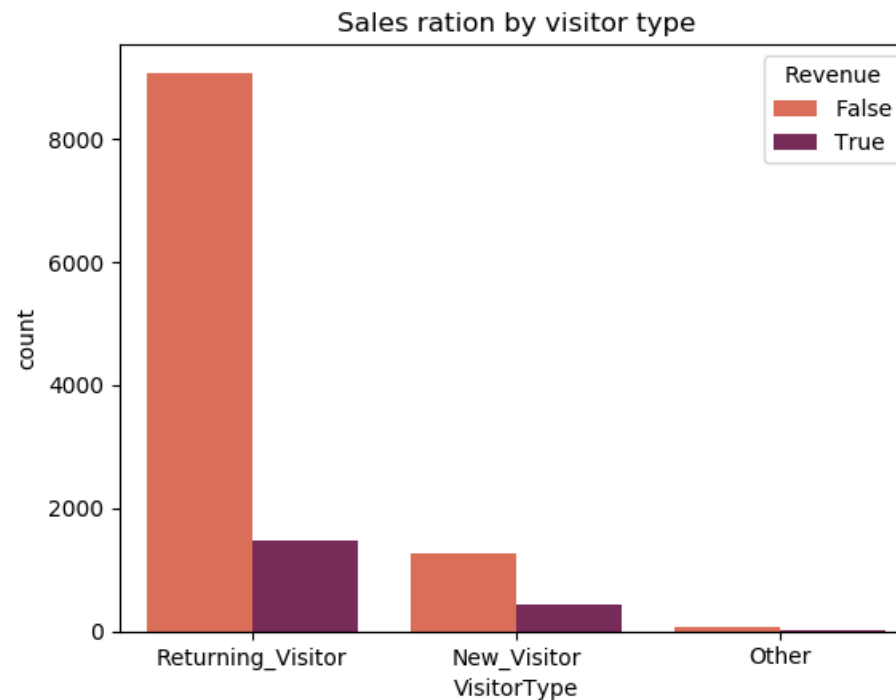




## Data processing : link to “Revenue”

### “VisitorType”

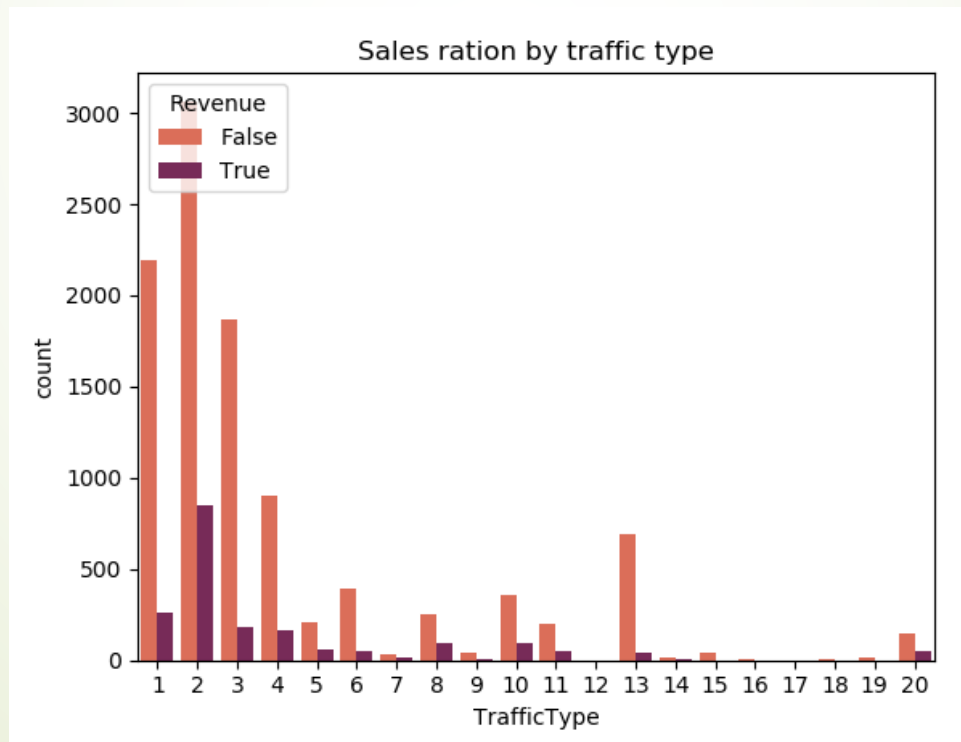
There are more users who have already come to the web site willing to transact on the site but the number of new users who complete a transaction is quite high considering the total number of new users.



# Data processing : link to “Revenue”

## “TrafficType”

More users using the traffic type corresponding to 2 finalize a transaction.



## Data processing : Conclusion

There is a possible problem for users using traffic other than 2 based on the number of sales. Maybe check if the site works correctly with other traffic.

A large proportion of returning customers do not buy on the web site, maybe the prices are lower on other site and most users come to this site to gather information. It may be a good thing to check if the prices are competitive.

Most of the audience is based in region 1 and 3, increasing the ad share in this region can be a good thing to increase sales.

Sales and visits to the site mostly happen in November and April, if the product is seasonal, maybe increase advertising and better offer during these months, otherwise put more advertising during other months can improve sales.

# Shoppers' intention prediction model

## Cleaning data

“Other” value from “VisitedType” were remove from the dataset.

The sharing of dataset data is 75% training and 25% test.

Due to the important part of “False” in “Revenue”, the split is stratified to keep the same percentage of “True” and “False” values in our two sub-set.

Non-numerical data type and variable where the hierarchy does not make sense (“Month”, “OperatingSystems”, “Browser”, “Region”, “TrafficType”, “VisitorType”) were transformed into dummies variable.

# Shoppers' intention prediction model

## Gaussian Naïve Bayes Classifier model

First of all, we implement a basic model whose results we will serve as a reference when the precision of the other models

Here are the results:

```
Gaussian Naive Bayes model accuracy : 84.62 %  
MSE : 0.15382103200522534
```

With this model we have an accuracy of about 85%

We also have a mean square error of 0,15.

We will try to improve those value with our next model.

# Shoppers' intention prediction model

Logistic regression model

Next, we implement a logistic regression model

Here are the results:

```
Logistic regression model accuracy : 88.37 %  
MSE : 0.116263879817113
```

With this model we have an accuracy of about 88%

We also have a mean square error of 0,11.



# Shoppers' intention prediction model

## Random forest model

Next, we implement a Random forest model.

According to the result of the GridSearch function and to keep a relatively low processing time for the computer, the best hyperparameters for this model are :

```
Best parameters : {'max_depth': 7, 'n_estimators': 300}
```

Here are the results with those parameters:

```
Random Forest Classifier model accuracy : 89.97 %  
MSE : 0.10026126714565643
```

With this model we have an accuracy of about 90%

We also have a mean square error of 0,10.

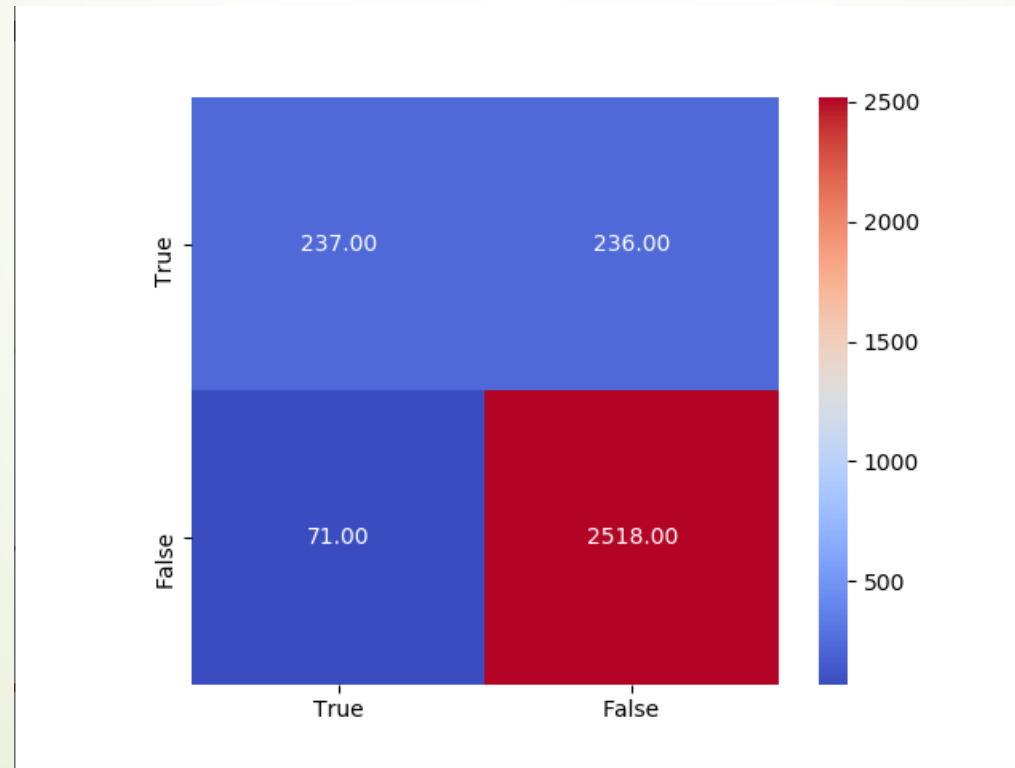
# Shoppers' intention prediction model

## Random Forest – Confusion matrix

From this confusion matrix, We can see that our model is good at predicting that a buyer is not going to buy from the site.

But, even with 90% accuracy, that doesn't seem like enough to be able to effectively predict whether a user is going to buy or not.

Among the people who bought according to the model, 50% did not actually do so.

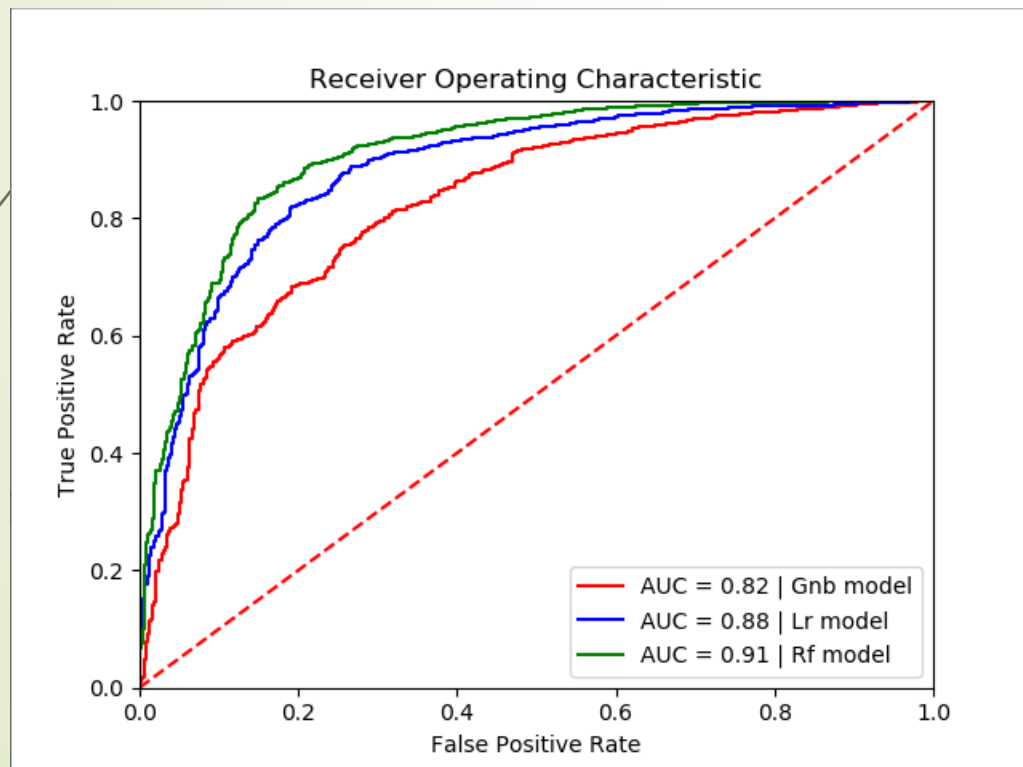


# Shoppers' intention prediction model

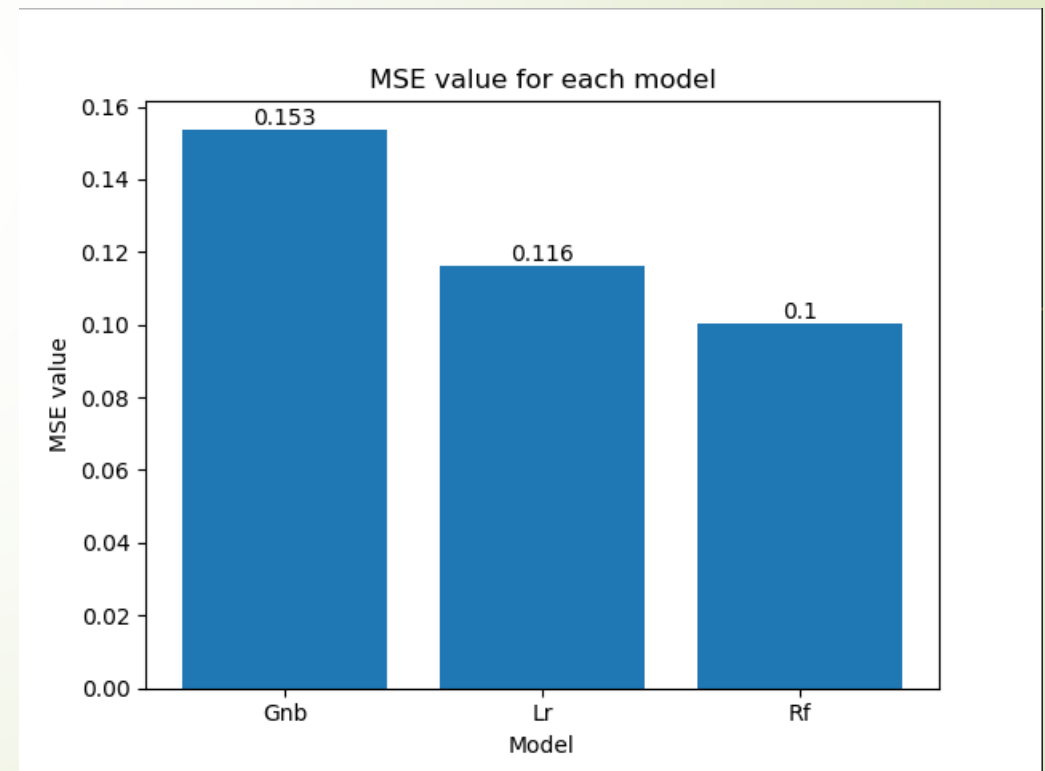
## Model comparison

Using Roc Curve and the MSE value :

The area under the ROC curve is: 0.82 % for the Gnb model.  
The area under the ROC curve is: 0.88 % for the logistic regression model.  
The area under the ROC curve is: 0.91 % for the random model model.



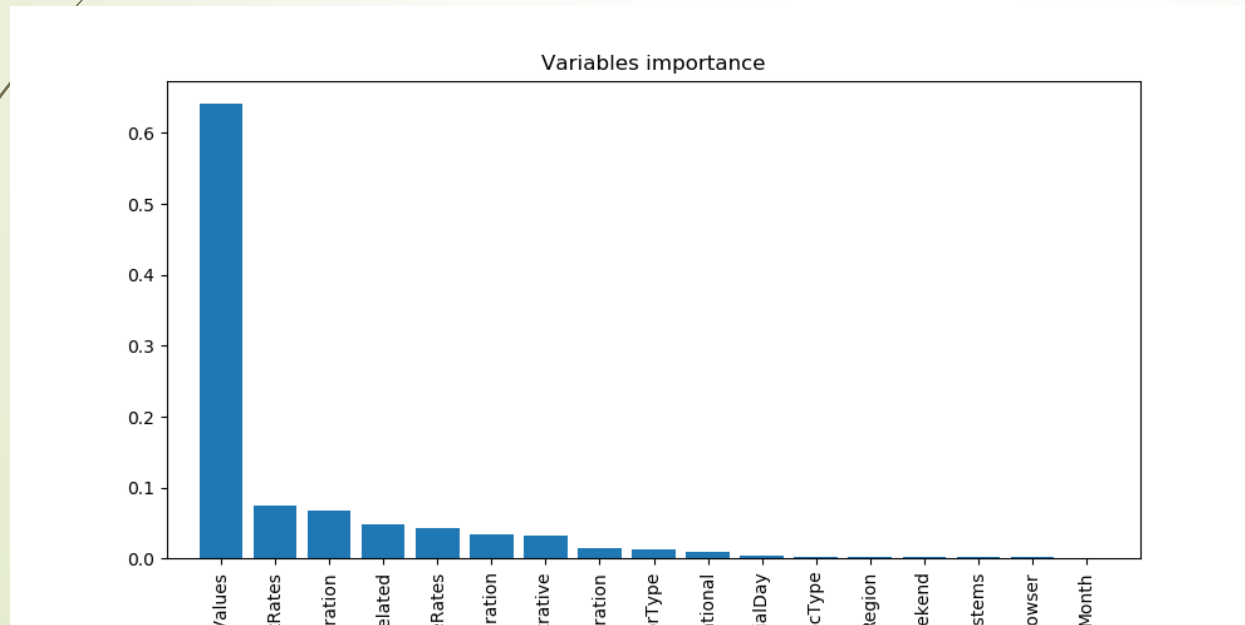
All three model seem to have good performance, but the best model here is the random forest model according to those graph.



# Shoppers' intention prediction model

## Variables importance

The most important feature here is the "PageValues" variable  
Then we have "ExitRates" and  
"ProductRelated" variables.



	Importance
PageValues	0.641934
ExitRates	0.074967
ProductRelated_Duration	0.067849
ProductRelated	0.048934
BounceRates	0.044010
Administrative_Duration	0.034663
Administrative	0.031958
Informational_Duration	0.014067
VisitorType	0.013625
Informational	0.009147
SpecialDay	0.004048
TrafficType	0.003209
Region	0.003076
Weekend	0.002898
OperatingSystems	0.002124
Browser	0.002105
Month	0.001387

After removing all the variable after "BouncesRate", we do not gain in accuracy for our model.

Random Forest Classifier model accuracy : 89.81 %  
MSE : 0.10189418680600915

# Conclusion

## **The important features:**

The model perform well with only few features witch are “PageValues”, “ExitRates”, “ProductRelated\_Duration”, “BouncesRate”, “ProductedRelated”.

Because “PageValues” is a very important features, finding good values for pages seems as important.

## **The model :**

By using a random forest classifier, we are able to achieve approximately 90% accuracy with a low sensibility.

Even with this accuracy, the error is still too high to properly identified real buyer.

## **Recommendation :**

It therefore seems important to pay attention to how define the value of each page, to have a good page for the products and control the Exit and Bounce rate of our pages.

Make great product pages seems also important