Rapport projet Python

Online Shoppers Purchasing Intention

Simon Pongan - Damien RHENY

Data Exploration

Our dataset untitled shopper online intentions is composed of 18 features.
This dataset is a trimestrial rapport that we can get thanks to Google analytics par exemple. Those features give us some information about all the visite on the online shop. We can classify our features like so.

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	0	0.0	0	0.0	1	0.000000	0.200000	0.200000	0.000000	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
1	0	0.0	0	0.0	2	64.000000	0.000000	0.100000	0.000000	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
2	0	0.0	0	0.0	1	0.000000	0.200000	0.200000	0.000000	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
3	0	0.0	0	0.0	2	2.666667	0.050000	0.140000	0.000000	0.0	Feb	3	2	2	4	Returning_Visitor	False	False
4	0	0.0	0	0.0	10	627.500000	0.020000	0.050000	0.000000	0.0	Feb	3	3	1	4	Returning_Visitor	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
12325	3	145.0	0	0.0	53	1783.791667	0.007143	0.029031	12.241717	0.0	Dec	4	6	1	1	Returning_Visitor	True	False
12326	0	0.0	0	0.0	5	465.750000	0.000000	0.021333	0.000000	0.0	Nov	3	2	1	8	Returning_Visitor	True	False
12327	0	0.0	0	0.0	6	184.250000	0.083333	0.086667	0.000000	0.0	Nov	3	2	1	13	Returning_Visitor	True	False
12328	4	75.0	0	0.0	15	346.000000	0.000000	0.021053	0.000000	0.0	Nov	2	2	3	11	Returning_Visitor	False	False
12329	0	0.0	0	0.0	3	21.250000	0.000000	0.066667	0.000000	0.0	Nov	3	2	1	2	New_Visitor	True	False

	Page visited
	Duration of the visitor on the page
	User behaviour
	Period of visit
	Acquisition of user
	Revenue generated

```
(12330, 18)
```

Problematic

After the explanation of our dataset, we can define our Problematic. Indeed the idea is to make a prediction of the user comportment. In another word it’s about using several features to predict if the user will generate some revenue for the online shop or not.

Summary

1 Data vizualisation

2 Preprocessing

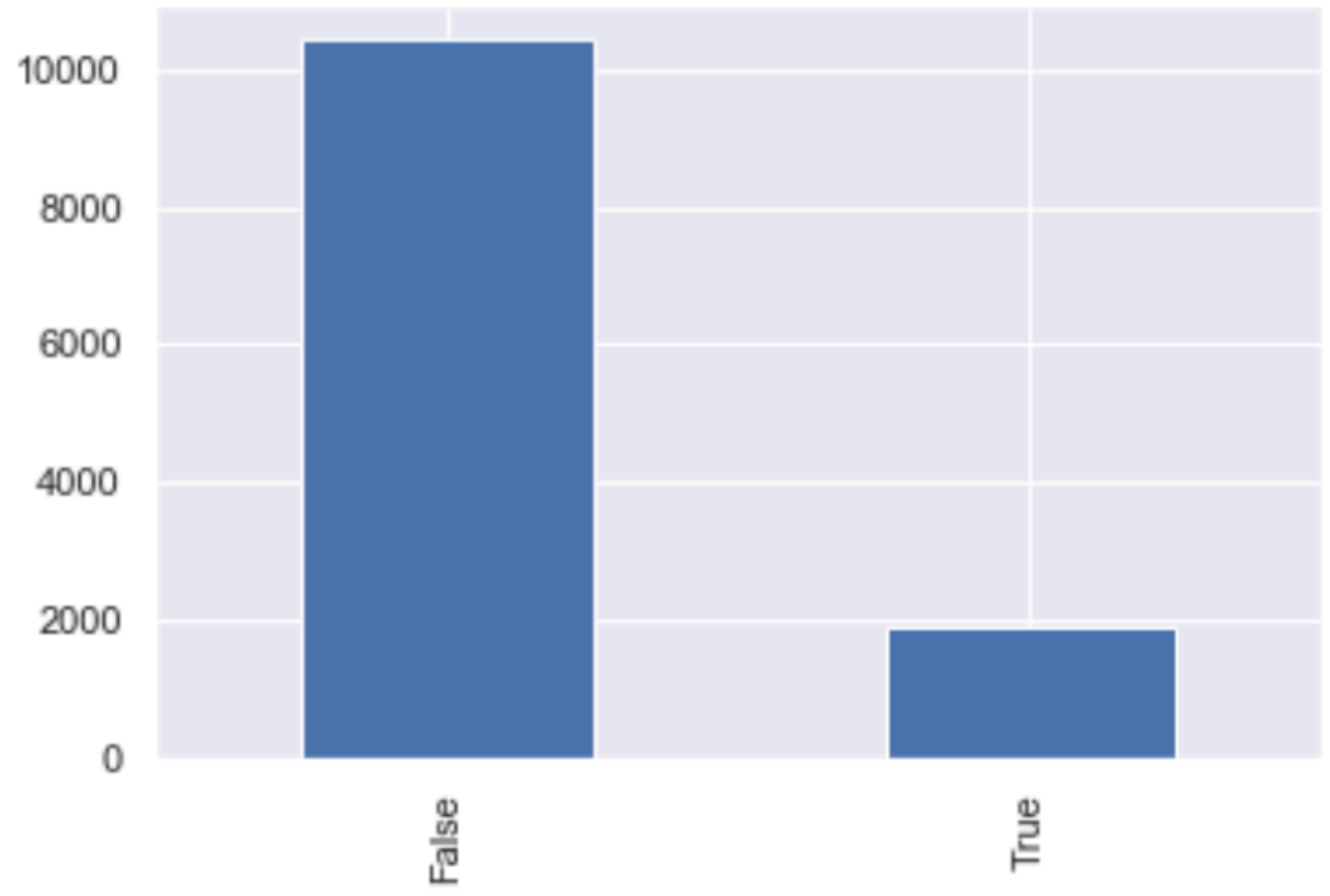
3 Modeling

4 Final work

Data vizualisation

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.315166	80.818611	0.503569	34.472398	31.731468	1194.746220	0.022191	0.043073	5.889258	0.061427	2.124006	2.357097	3.147364	4.069586
std	3.321784	176.779107	1.270156	140.749294	44.475503	1913.669288	0.048488	0.048597	18.568437	0.198917	0.911325	1.717277	2.401591	4.025169
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000	184.137500	0.000000	0.014286	0.000000	0.000000	2.000000	2.000000	1.000000	2.000000
50%	1.000000	7.500000	0.000000	0.000000	18.000000	598.936905	0.003112	0.025156	0.000000	0.000000	2.000000	2.000000	3.000000	2.000000
75%	4.000000	93.256250	0.000000	0.000000	38.000000	1464.157213	0.016813	0.050000	0.000000	0.000000	3.000000	2.000000	4.000000	4.000000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	0.200000	0.200000	361.763742	1.000000	8.000000	13.000000	9.000000	20.000000

After the visualization of the quantitative data we analyze the target value : Revenue.




```
False    10422
True      1908
Name: Revenue, dtype: int64
```

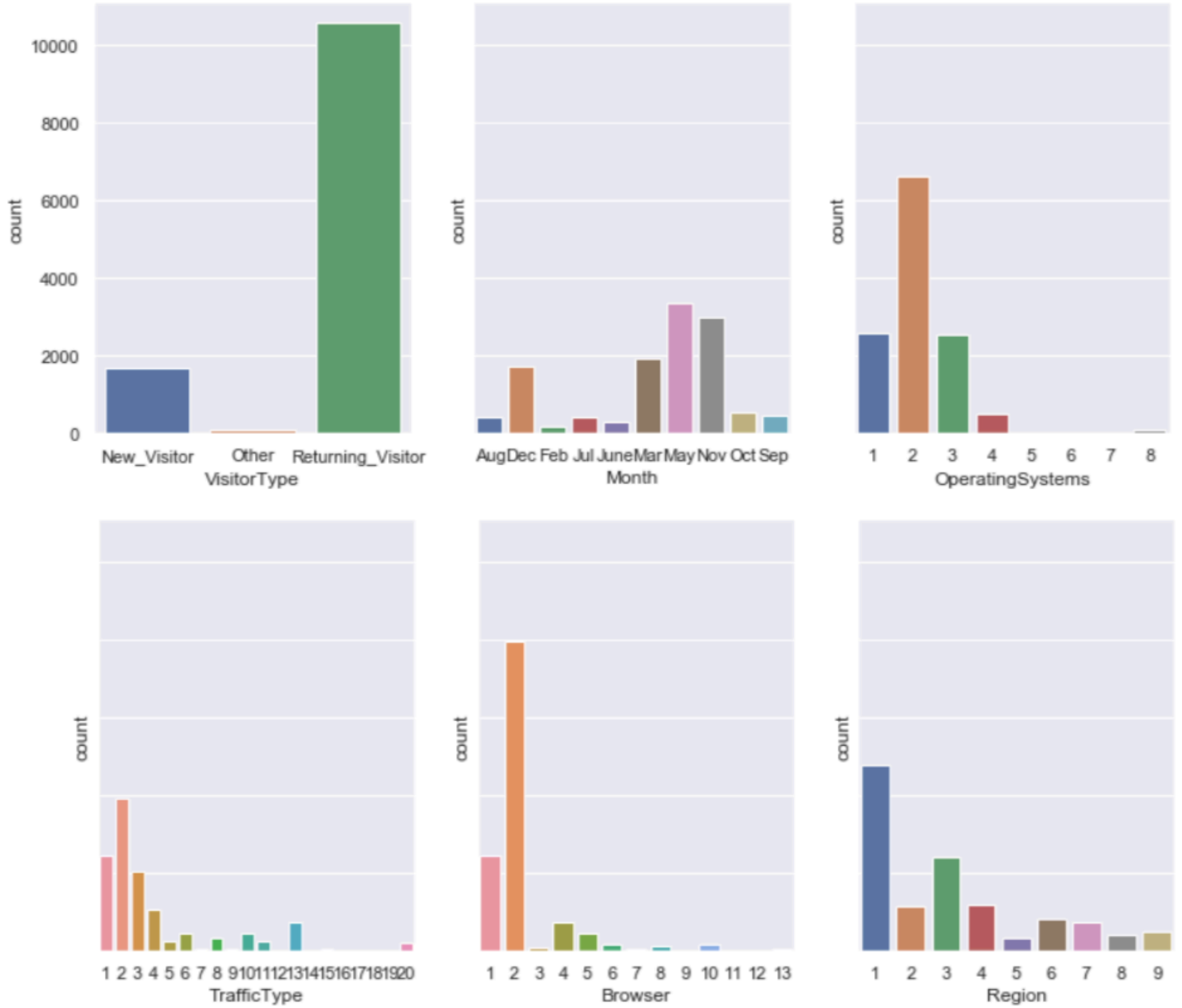
Data vizualisation

Categorical, Numerical and Boolean Features

Data set before conversion	Data set after conversion	number of unique values for each column
Administrative int64	Administrative int64	Administrative 27
Administrative_Duration float64	Administrative_Duration float64	Administrative_Duration 3335
Informational int64	Informational int64	Informational 17
Informational_Duration float64	Informational_Duration float64	Informational_Duration 1258
ProductRelated int64	ProductRelated int64	ProductRelated 311
ProductRelated_Duration float64	ProductRelated_Duration float64	ProductRelated_Duration 9551
BounceRates float64	BounceRates float64	BounceRates 1872
ExitRates float64	ExitRates float64	ExitRates 4777
PageValues float64	PageValues float64	PageValues 2704
SpecialDay float64	SpecialDay float64	SpecialDay 6
Month object	Month category	Month 10
OperatingSystems int64	OperatingSystems category	OperatingSystems 8
Browser int64	Browser category	Browser 13
Region int64	Region category	Region 9
TrafficType int64	TrafficType category	TrafficType 20
VisitorType object	VisitorType category	VisitorType 3
Weekend bool	Weekend bool	Weekend 2
Revenue bool	Revenue bool	Revenue 2
dtype: object	dtype: object	dtype: int64

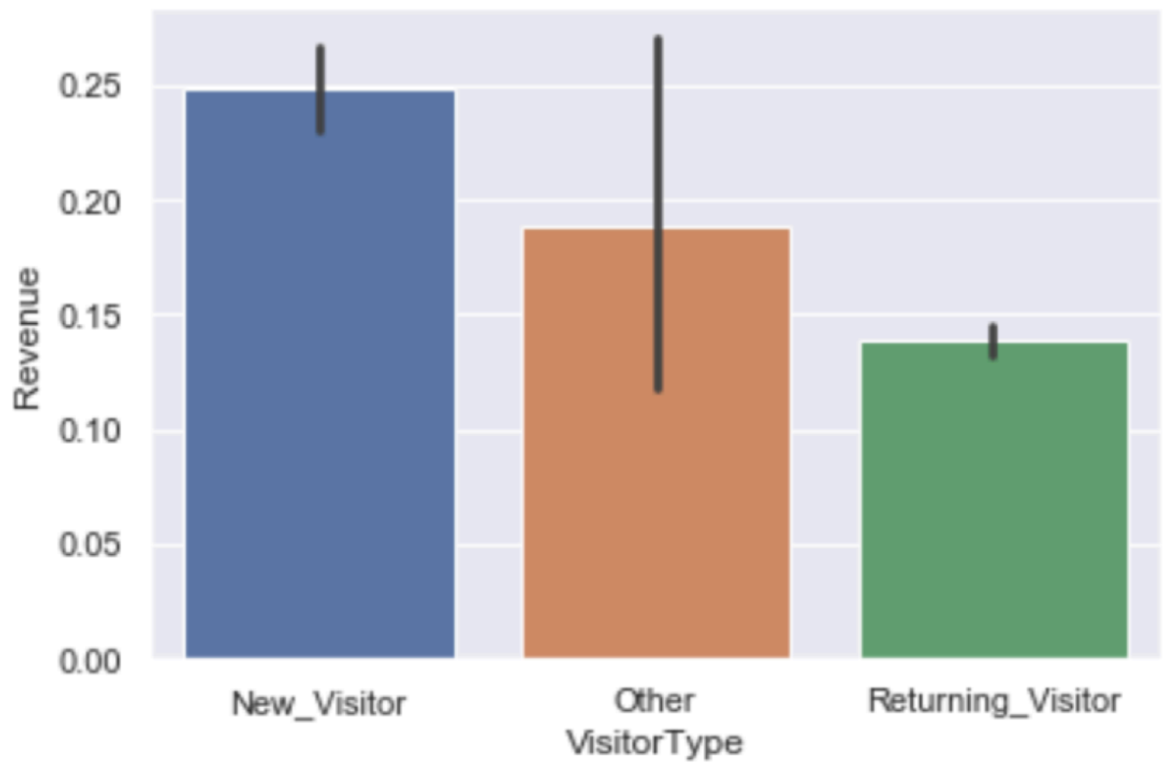
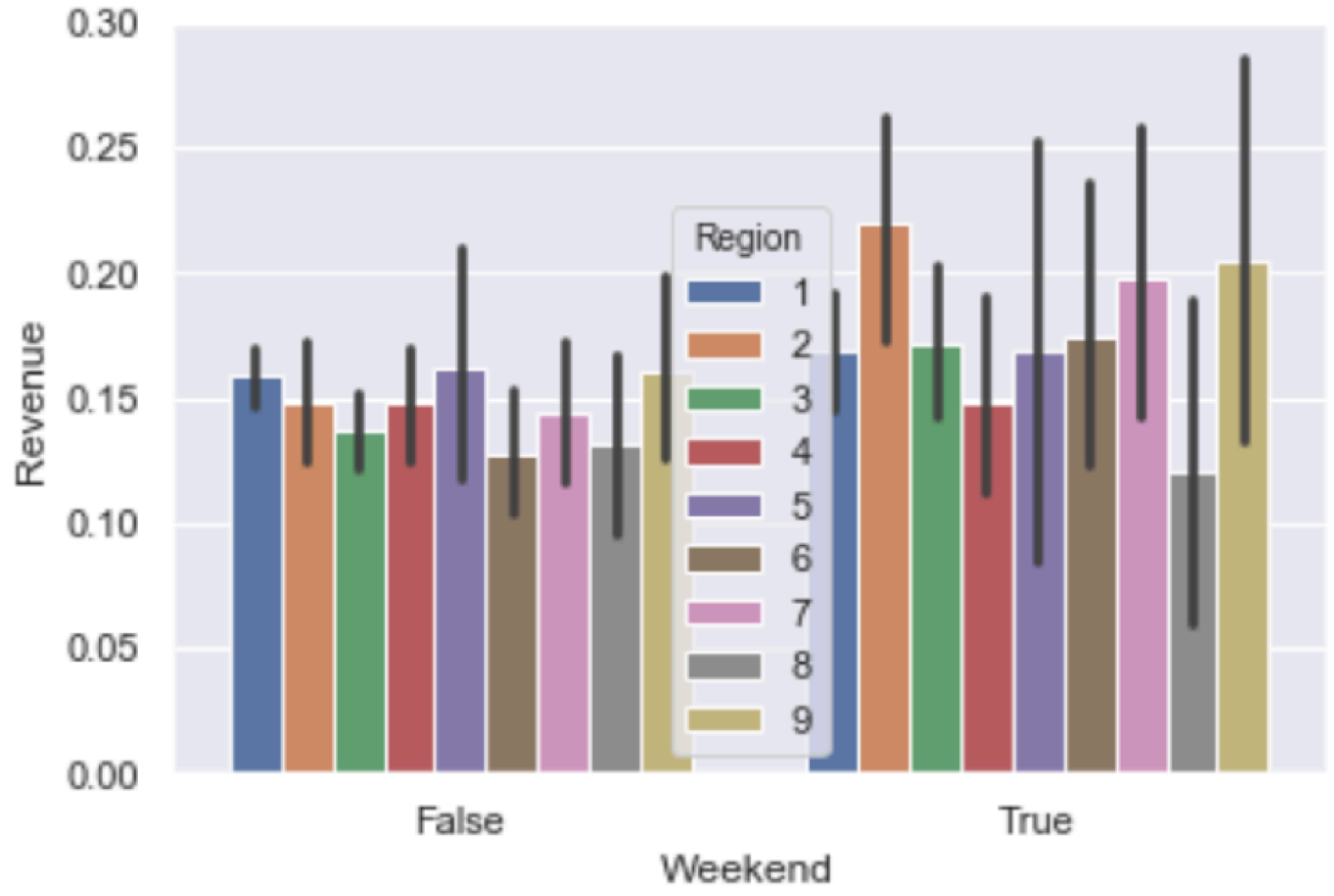
Data vizualisation

Repartition of Categorical Features



Data vizualisation

Focus on Revenue

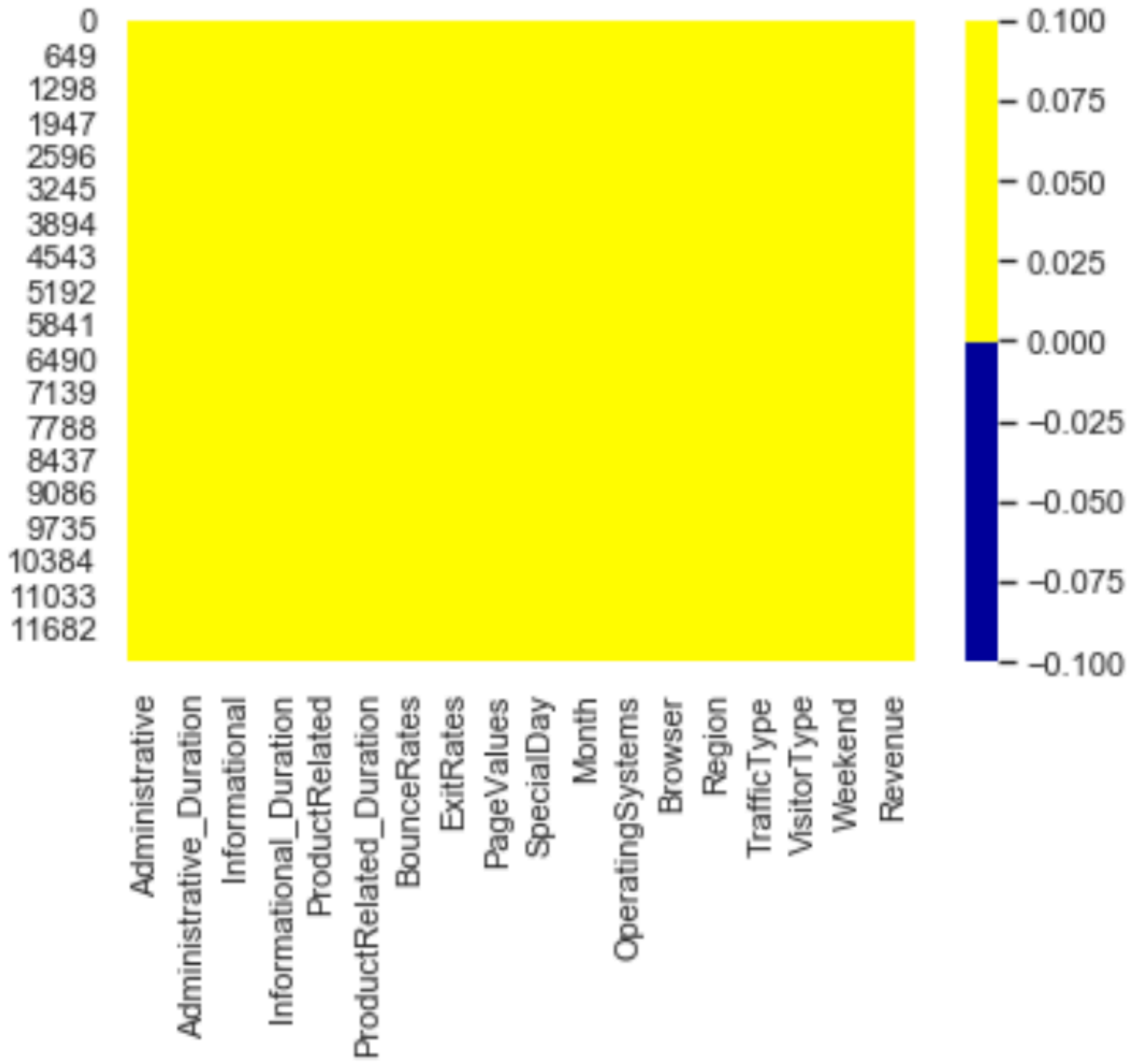
<table>
<tr><th>Revenue / VisitorType</th><th>Revenue / Weekend</th><th>Revenue / Weekend /Region</th></tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
</tr>
<tr><td>We can see that new visitors generate more revenue than the others</td><td>More revenue are generate during the week end</td><td>The Région 2 generate the most part of the revenue during the week end</td></tr>
</table>

Preprocessing

Check the missing value

To make our models the most efficient as possible, We fist check if there is some missing value.

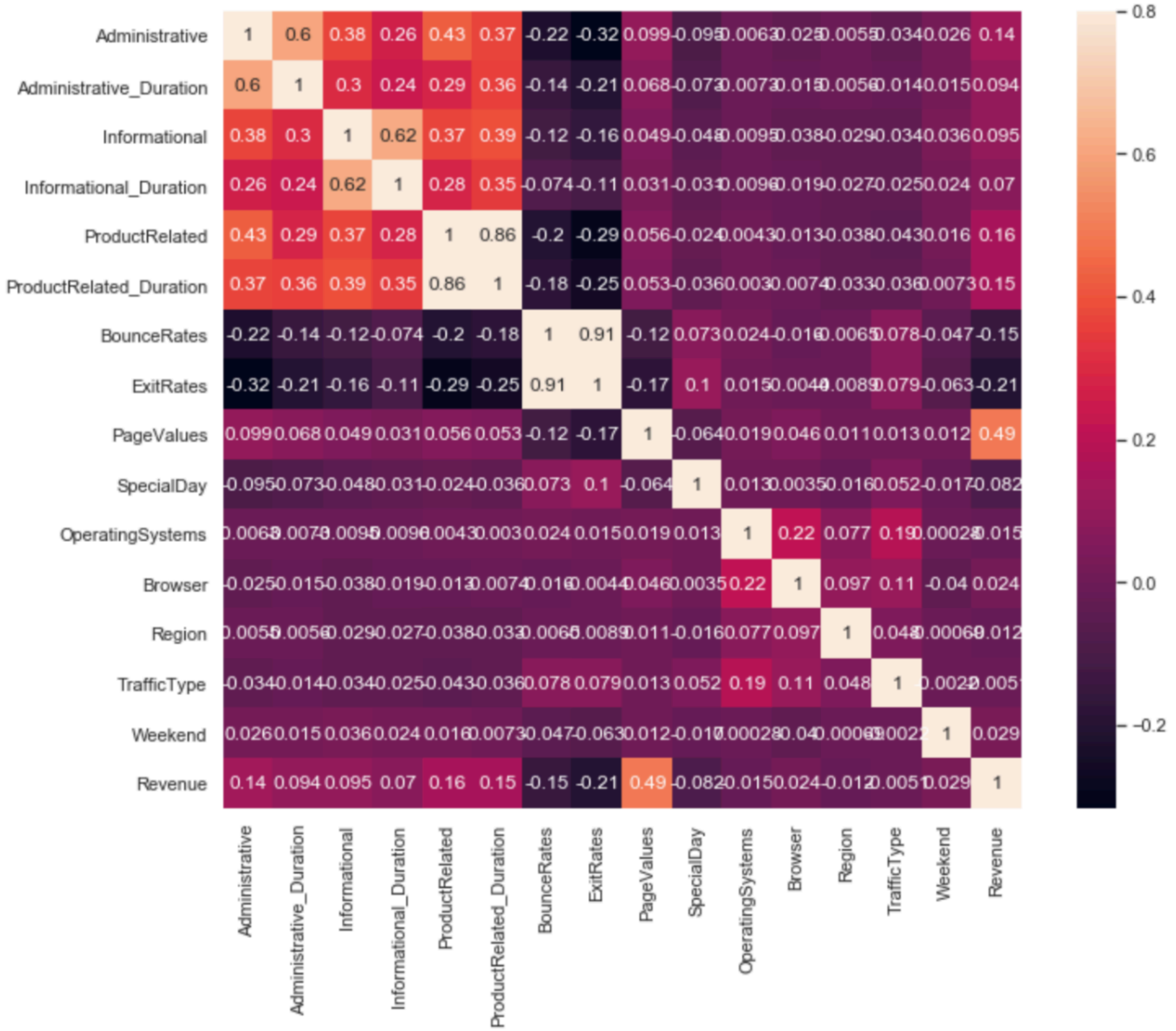
```
Administrative             0
Administrative_Duration    0
Informational              0
Informational_Duration     0
ProductRelated             0
ProductRelated_Duration    0
BounceRates                0
ExitRates                  0
PageValues                 0
SpecialDay                 0
Month                      0
OperatingSystems           0
Browser                    0
Region                     0
TrafficType                0
VisitorType                0
Weekend                    0
Revenue                    0
dtype: int64
```



As we can see there is any missing value.

Preprocessing
Data cleaning

Still in order to improve the accuracy of our models, we check the relevance of each features thanks to an heat map.



As we can see there is few features that we can avoid from the dataset to improve the following. Among this data we can quote :

- Month
- TrafficType
- Informational_Duration
- BounceRates
- Region

Obviously we also avoid our target data « Revenue »

Modeling

	Logistic regression	Random Forest	Bagging	Boosting
Accuracy	88,97 %	90,79 %	90,02 %	89,21 %

Final work

As we saw with the previous results the random forest model seems more accurate than the other models. Therefore we will implement this model in our API