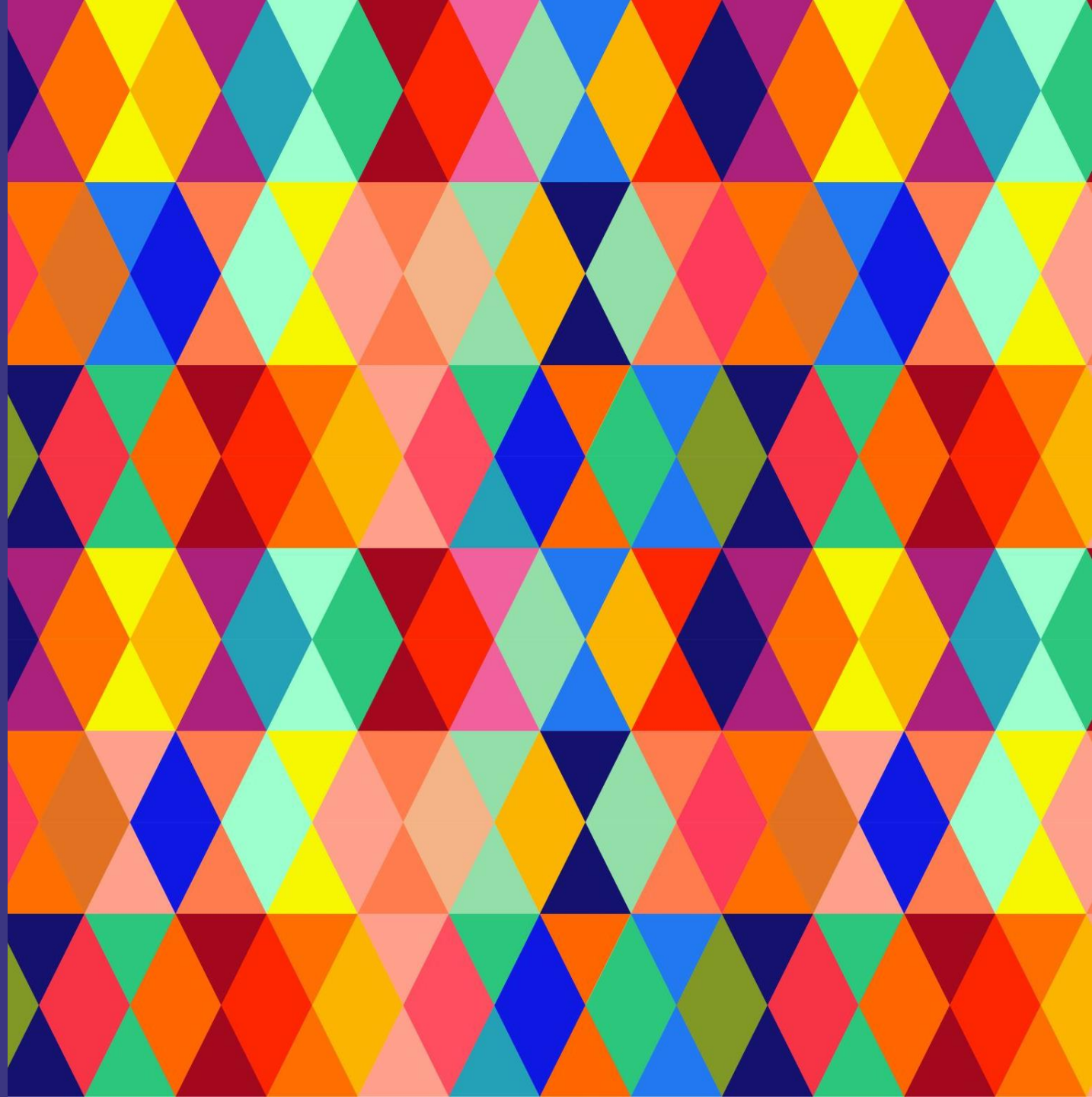
ONLINE SHOPPERS PURCHASING INTENTION DATASET STUDY

By Vincent DANIEL



This study by Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018) explores the statistics of whether or not a user buys a product on a website using different measures, some custom to this study, and others provided by google directly, most specifically the interactions between the browser and the site.

So, at first, I didn’t really have any idea for the most important variable in this study : most of what I thought about wasn’t a known : the age of the person, his bank account, his interest in the product… Some of these we can have a hint towards via some of the variables we have, such as the “Returning_Visitor” Boolean, or the “Special day” Boolean (proximity to a special buying day, like Christmas or valentines' day), which gives us information respectively about the interest, and the likelihood of the person buying the product.

The first operation I did on the dataset was a corr(method ='pearson’), to check what variable was the most representative.

We only look at the part of this function regarding the “ Revenue ” Variable, which represents what we are interested in. We can see that the most representative variables is “Page Values”.

The dataset information tells us that “This variable is supposed to represent the average value for a web page that a user visited before completing an e-commerce transaction. ”

I found this explanation unclear, so I looked it up, and basically, this value is intended to give an idea of which page in a site contributed more to this site's revenue. In our case, this variable represents how much the studied case visited the website.

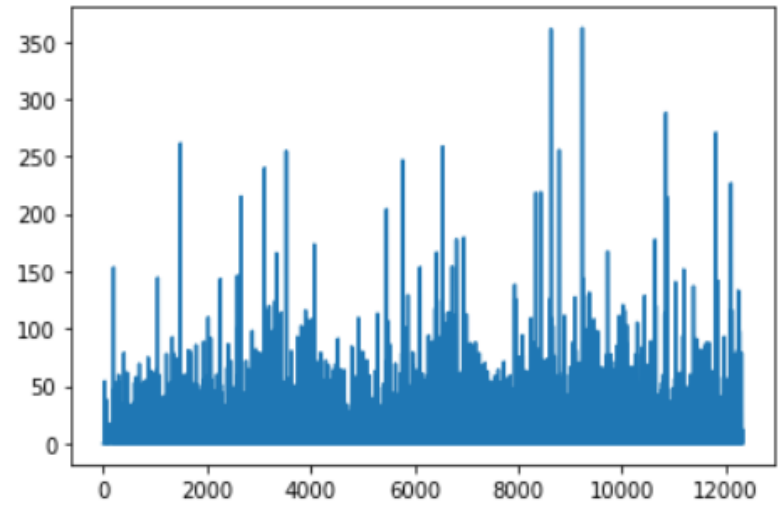
```
In [16]: dataset.corr(method ='pearson')["Revenue"]

Out[16]: Administrative             0.138917
         Administrative_Duration    0.093587
         Informational              0.095200
         Informational_Duration     0.070345
         ProductRelated             0.158538
         ProductRelated_Duration    0.152373
         BounceRates               -0.150673
         ExitRates                 -0.207071
         PageValues                 0.492569
         SpecialDay                -0.082305
         OperatingSystems          -0.014668
         Browser                    0.023984
         Region                    -0.011595
         TrafficType               -0.005113
         Weekend                    0.029295
         Revenue                    1.000000
         Name: Revenue, dtype: float64
```

Knowing that “Page Values” is the most representative variable, I tried to plot this variable for all studied cases, and for only the ones that interested us, but the result doesn’t seem very clear :

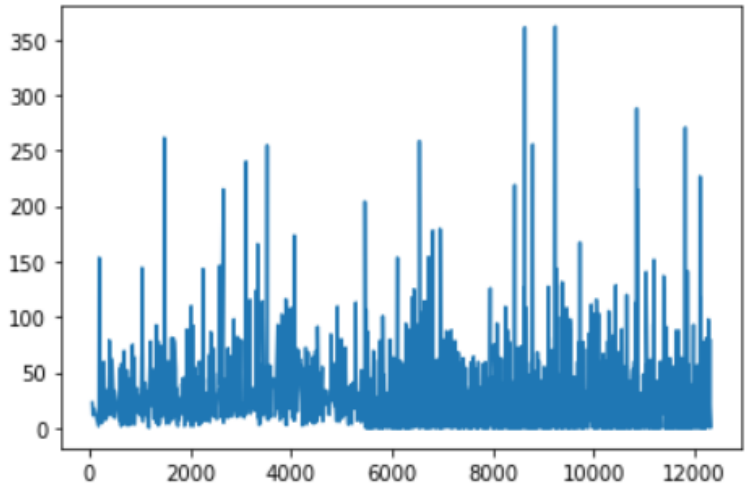
This is normal since there was only about 0.49 correlation, however we can see that when the values are getting higher on the X axis, the more the two graphs look alike.

Because of this, we can assume that when we have a high Page Values value, the user is most likely to buy the product, but below a certain point we cannot tell.

```
In [9]: dataset["PageValues"].plot()
Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x1f0b2226630>
```



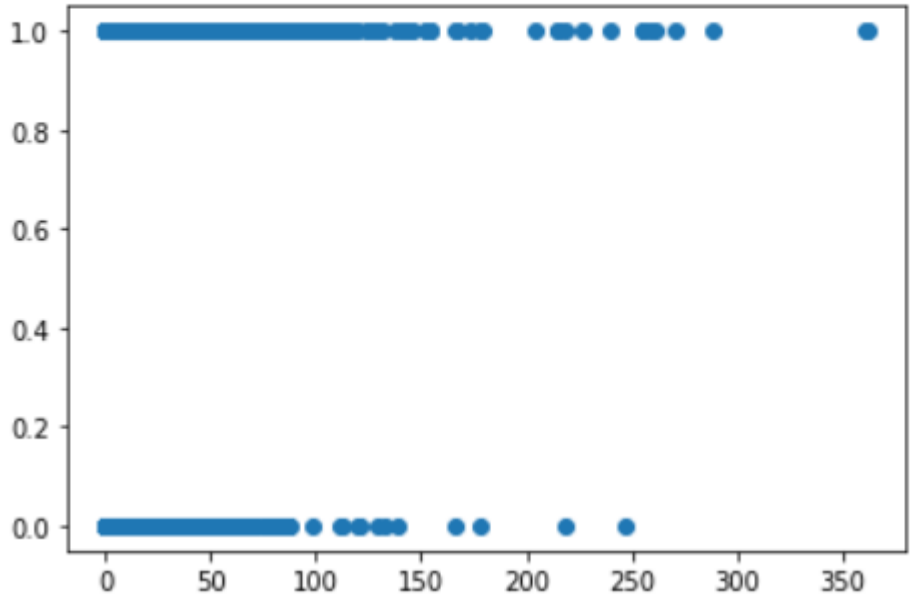

```
In [7]: dataset[(dataset.Revenue == True)]["PageValues"].plot()
Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x1f0e3568668>
```



To try to confirm this, I represented on a graph how the page values influences the Revenue, and we find out that at low values, this doesn’t really matter, but indeed, at higher values, most positive studied tests have a high page value.

We can even say that above 150 or 100 page value, the data is most likely positive. Below that, we can’t draw any conclusion.

```
In [26]: plt.scatter(dataset["PageValues"],dataset["Revenue"])
Out[26]: <matplotlib.collections.PathCollection at 0x1f13181c7f0>
```



After having tested out all other variables, the only other interesting one I thought would be the “Special Day” variable.

This variable represents the proximity of a special day with the case, such as Christmas, valentine's day, etc…

This variable seems like a good idea, however it wasn’t representative, because there was too much positive studied cases with negative “Special Days”.

The final result of this study is that the final result is hard to predict, however it can still be estimated is some specific cases (high page value).