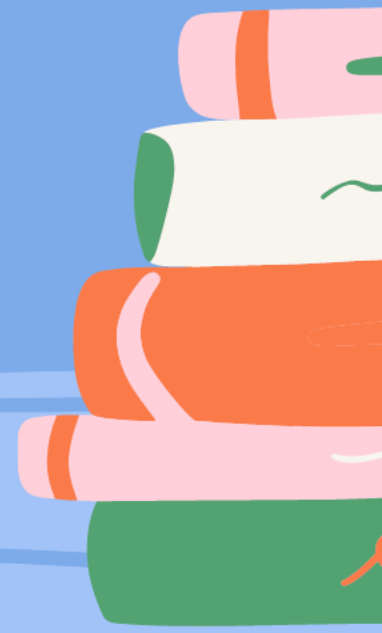
PYTHON FOR DATA ANALYSIS

Intentions d'achat en ligne

(Online Shoppers Purchasing Intention)

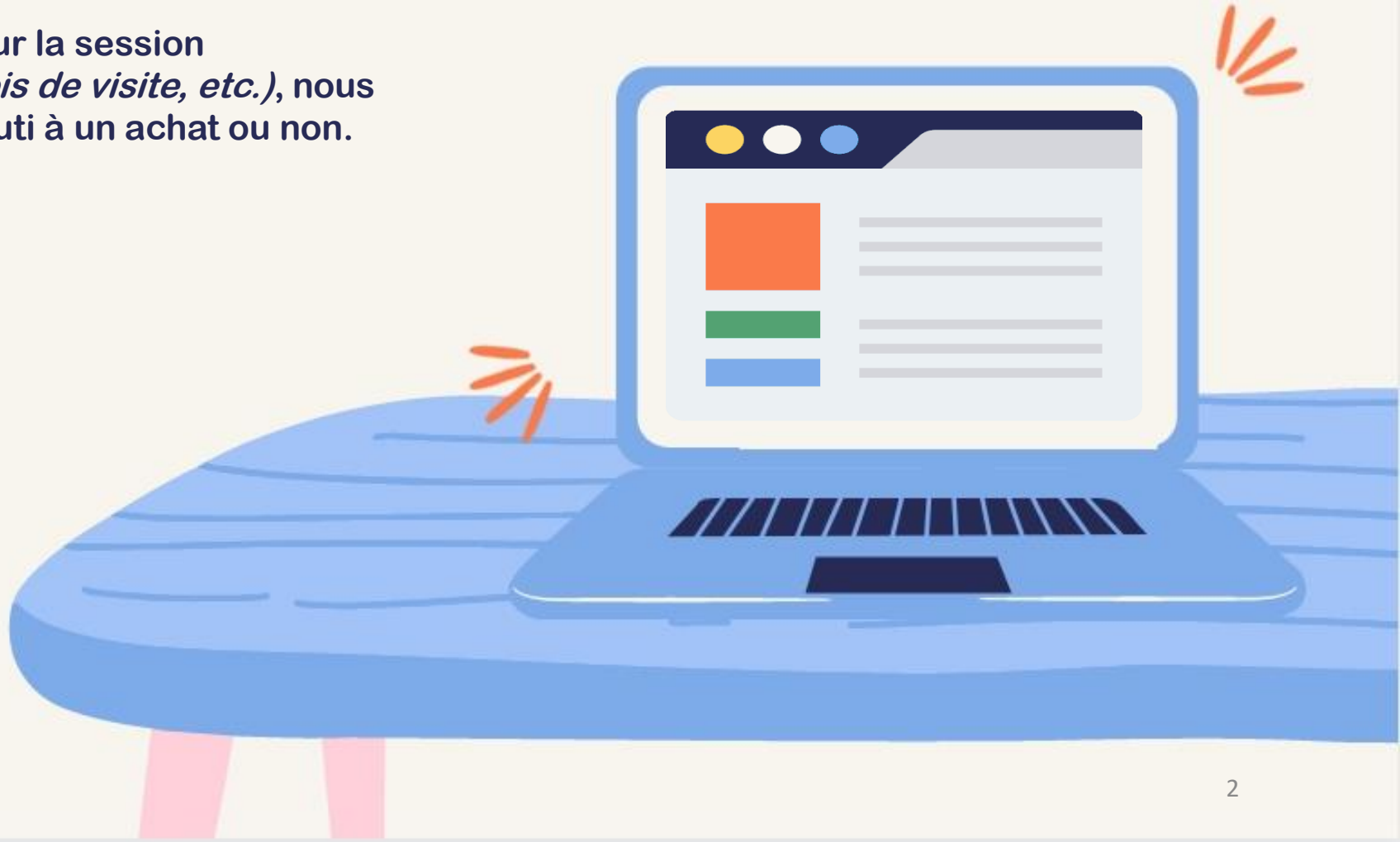
Khadidiatou BADJI – Vayshnavi SIVARAJAH



Présentation générale des données

Notre dataset* présente des sessions Internet (interaction entre un site ou une application et un visiteur ayant chargé au moins une page).

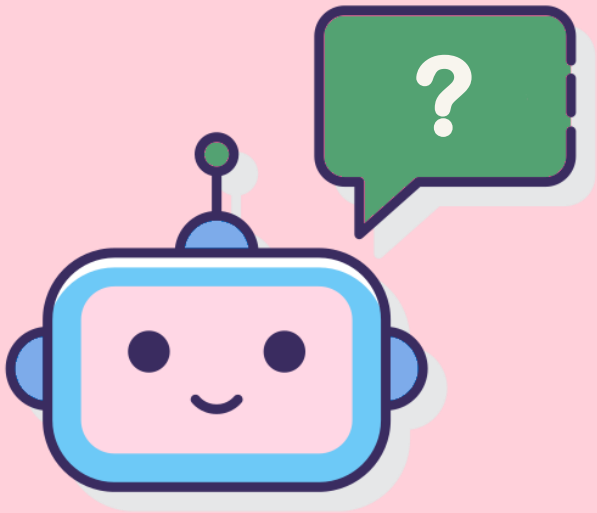
En plus de différentes informations sur la session *(exemples : région géographique, mois de visite, etc.)*, nous savons également si la session a abouti à un achat ou non.



* Dataset = ensemble de données

PROBLÉMATIQUE

A partir des informations sur une session Internet, nous cherchons à prévoir si le visiteur va effectuer ou non un achat au cours de la session ouverte.



DÉCOUVERTE DES DONNÉES

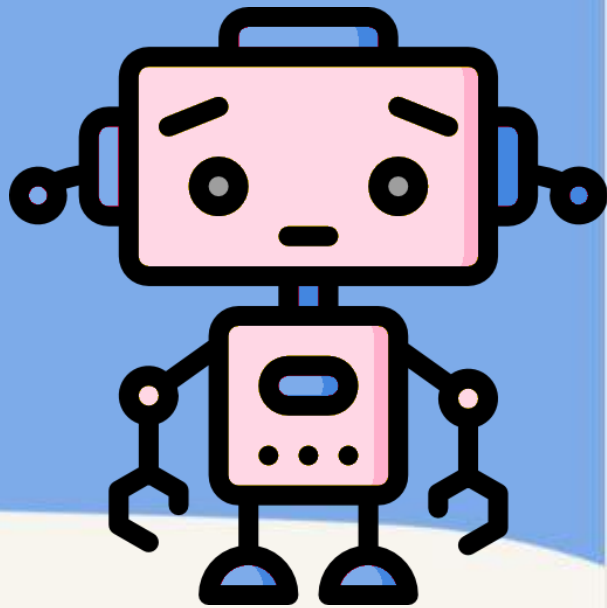
Certains champs de notre dataset n'étaient pas clairement explicités, nous avions donc quelques difficultés à comprendre le contenu mais aussi parfois l'utilité de ces champs.

Nous avions 3 grands axes d'interrogations :

1. Les champs "Bounce Rate", "Exit Rate" et "Page value" sont décrits comme étant des mesures de Google Analytics décrivant une page web. Or, lors d'une session Internet, le visiteur peut consulter plus d'une page (de différents types). A quoi correspondent donc exactement ces 3 champs ? Est-ce une moyenne des valeurs pour toutes les pages ? Ou bien concernent-ils une seule page, mais dans ce cas laquelle ?

2. Nous voulions être sûres d'avoir bien compris le contenu du champ "SpecialDay". La Saint Valentin (14 février) était mentionnée en tant qu'exemple : « Ce champ prend une valeur non nulle entre le 2 février et le 12 février et prend une valeur nulle avant et après cet intervalle sauf si la date est proche d'un autre jour spécial et son maximum de 1 est atteint le 8 février." Est-ce que cela signifie que le champ "Special Day" prend toujours la valeur 1 une semaine avant un jour spécial ? Si c'est le cas, pendant les quelques jours avant et après ce jour (lorsque le "SpecialDay" est à 1), plus la date est proche de ce jour plus la valeur est proche de 1 (par exemple le 7 ou le 9 février). A l'opposé, la valeur sera proche de 0 pour le 2 et le 12 février par exemple.

3. Pour les champs "OperatingSystems", "Browser", "Region" et "TrafficType", nous n'avons pas trouvé de table de mapping. Nous aurions voulu savoir si lorsque le champ "OperatingSystems" prend la valeur 1 cela signifie que la session a été lancée depuis un OS "Windows" , par exemple.



Nous avons donc tenté de contacter les créateurs du dataset via le formulaire présent sur leur site mais aussi via LinkedIn.
Malheureusement, nous avons reçu une réponse tardivement et qui ne nous fournissait pas assez de détails…

Réponse du créateur du dataset

Dear Khadidiatou and Vayshnavi,

I hope it is not too late for answers.

1- For "Bounce Rate", "Exit Rate" and "Page value" attributes, it is written in the description that they are measures from Google Analytics. But it's about a single page, and in one row, the visitor may have visited more than one page, be it "administrative", "informational" or "product related". So, what are these 3 features about exactly ? Is it an aggregation of these vales for each page ? Or is it about one page only, but then which one ?

These are average values of the pages visited by the visitor during his/her session.

2- About the "SpecialDay" feature. We wanted to make sure that we understand well. You mentioned Valentine's day (February 14) as an example : "this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8."
Does this mean that "Special Day" always takes 1 one week before a special day ? If so, for a few days before and after that day (where "SpecialDay" is 1), the closer the day is to this day, it will take a value closer to 1 (February 7 or February 9 for example). And on the opposite, the closer to 0 for February 2 and 12 for example. Is that correct ?

This variable is 0 except the days close to a special day. It becomes getting non-zero values about 12 days before the special day. For example, for Valentina's day it takes value of 0.1 on February 2, increases up to 1 on February 8, then starts to decrease again and get 0 (zero) on February 13.

3 - As for the "OperatingSystems", "Browser", "Region" and "TrafficType" attributes, is there a mapping table or something similar available ? So that we know that "OperatingSystems" : 1 means "Windows" for instance.

The only info we can share is the one at the data repository due to the agreements we have made.

Best regards,
C. Okan Şakar

Finalement, nous avons dû supprimer quelques champs pour lesquels, sans informations complémentaires, nous ne pouvions pas travailler. Ci-dessous, les champs que nous avons supprimés avec la description assez floue du champ que nous avions :

BOUNCE RATE

"Percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session"

EXIT RATE

"Calculated as for all pageviews to the page, the percentage that were the last in the session."

PAGE VALUE

"Average value for a web page that a user visited before completing an e-commerce transaction."

TRAFFIC TYPE

Aucune information

CRÉATION DE NOUVELLES VARIABLES

MonthNumber : index de chaque mois

Afin de pouvoir trier nos données dans l'ordre chronologique des mois, nous avons créé une variable MonthNumber qui prend comme valeur l'index de chaque mois.

TotalPages : nombre total de pages visitées

Cette nouvelle variable correspond à la somme des champs Administrative, Informational et ProdcutRelated.

TotalDuration : temps total passé sur les pages visitées

Cette nouvelle variable correspond à la somme des champs Administrative_Duration, Informational_Duration et ProductRelated_Duration.

MAPPING DES CHAMPS

Pour les champs "OperatingSystems", "Browser" et "Region", nous n'avions pas de table de mapping. Nous ne pouvions pas savoir si, par exemple, lorsque le champ "OperatingSystems" prend la valeur 1 cela signifie que la session a été lancée depuis un OS "Windows".

Avec l'accord de notre professeure, madame Imen OULED DLALA, nous avons donc déterminé nous-mêmes un mapping pour chaque champ "OperatingSystems", "Browser" et "Region".

OPERATING SYSTEMS

1 → Windows
2 → Ubuntu
3 → Mac OS
4 → Fedora
5 → Solaris
6 → Free BSD
7 → Chrome OS
8 → CentOS

BROWSER

1 → Google Chrome
2 → Mozilla Firefox
3 → Safari
4 → Internet Explorer
5 → Arachne
6 → AWeb
7 → Dillo
8 → Dooble
9 → HighWire
10 → IBrowse
11 → iCab
12 → Lunascape
13 → Konqueror

REGION

1 → Africa
2 → North America
3 → South America
4 → Eastern Europe
5 → South Asia and Southeast Asia
6 → Oceania
7 → Europe
8 → Central Asia
9 → Middle East

DESCRIPTION DE L'ENSEMBLE DES CHAMPS

Administrative

Nombre de sites administratifs *(exemple : site du gouvernement)* visités au cours de la session

Administrative_Duration

Temps passé sur des sites administratifs *(exemple : site du gouvernement)* au cours de la session

Informational

Nombre de sites d'informations *(exemple : site de Franceinfo) visités* au cours de la session

Informational_Duration

Temps passé sur des sites d'informations *(exemple : site de Franceinfo)* au cours de la session

ProductRelated

Nombre de sites vendant des produits *(exemple : site de Aliexpress) visités* au cours de la session

ProductRelated_Duration

Temps passé sur des sites vendant des produits *(exemple : site de Aliexpress)* au cours de la session

SpecialDay

Pourcentage (0 à 1) de proximité de la date de la session d'un jour spécial susceptible de booster les ventes (*ex: la St-Valentin)*

Month

Mois au cours duquel la session a été ouverte

OperatingSytems

Système d'exploitation *(exemple : Windows)* depuis lequel la session a été ouverte

Browser

Navigateur *(exemple : Mozilla Firefox)* depuis lequel la session a été ouverte

Region

Région dans le monde *(exemple : Africa)* depuis laquelle la session a été ouverte

VisitorType

Champ indiquant si le visiteur est un nouveau visiteur, un visiteur de retour ou autre

Weekend

Booléen indiquant si la session a été ouverte lors d'un weekend ou non

Revenue

Booléen indiquant si la session a abouti à un achat ou non

MonthNumber*

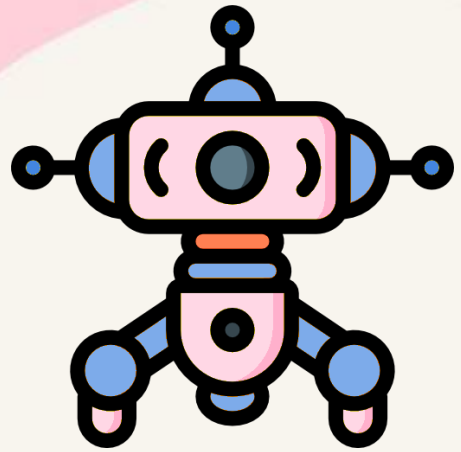
Index du mois au cours duquel la session a été ouverte *(exemple : février → 2)*

TotalPages*

Nombre total de pages visitées au cours de la session

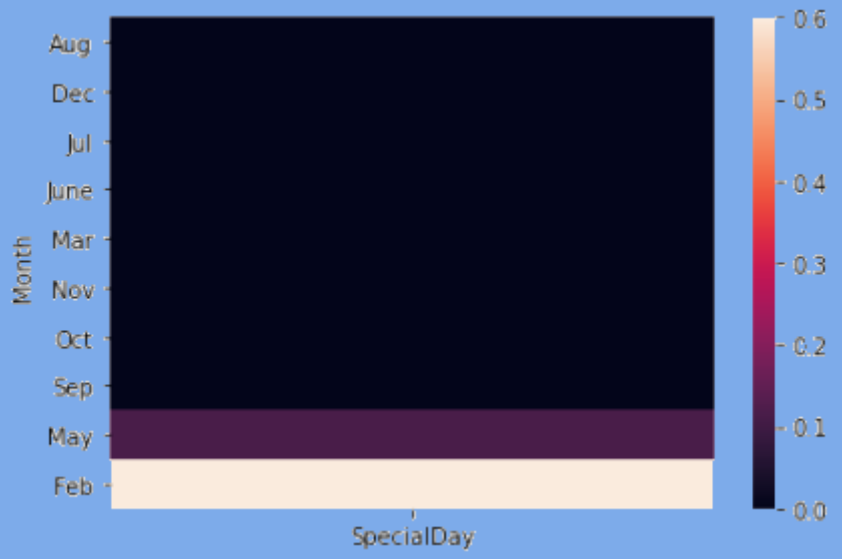
TotalDuration*

Temps total passé sur les pages visitées au cours de la session



*** Les nouveaux champs que nous avons ajoutés**

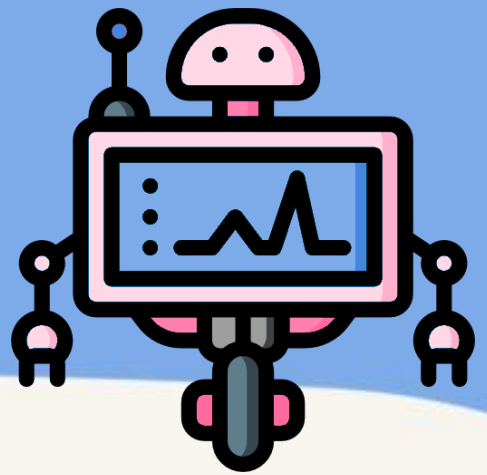
DATA VISUALISATION

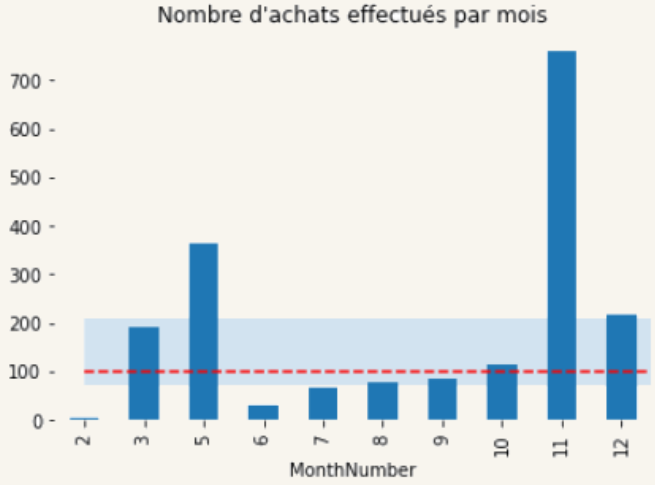


Nous avons créé un heatmap affichant le nombre de SpecialDay moyen par mois lors desquels des achats ont aboutis.

Nous remarquons qu'il n'y a qu'en mai et en février qu'il y a eu des Special Day (lors desquels des achats ont été effectués).

En mai, les achats ayant abouti étaient à environ 11% proche d'un Special Day.
En février, les achats ayant abouti étaient à environ 60% proche d'un Special Day.



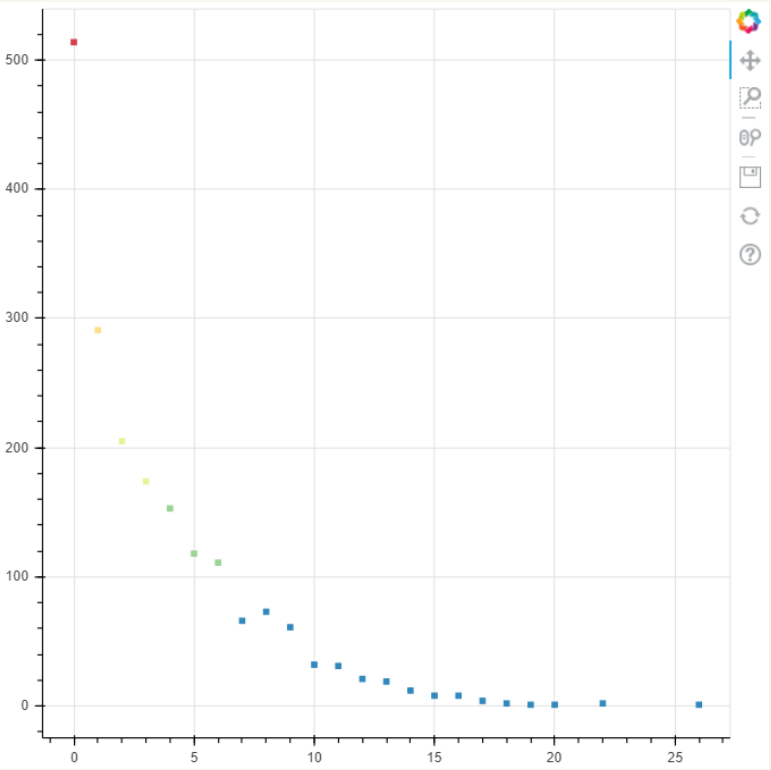


Nous avons ensuite créé un histogramme affichant le nombre d'achats effectués par mois.

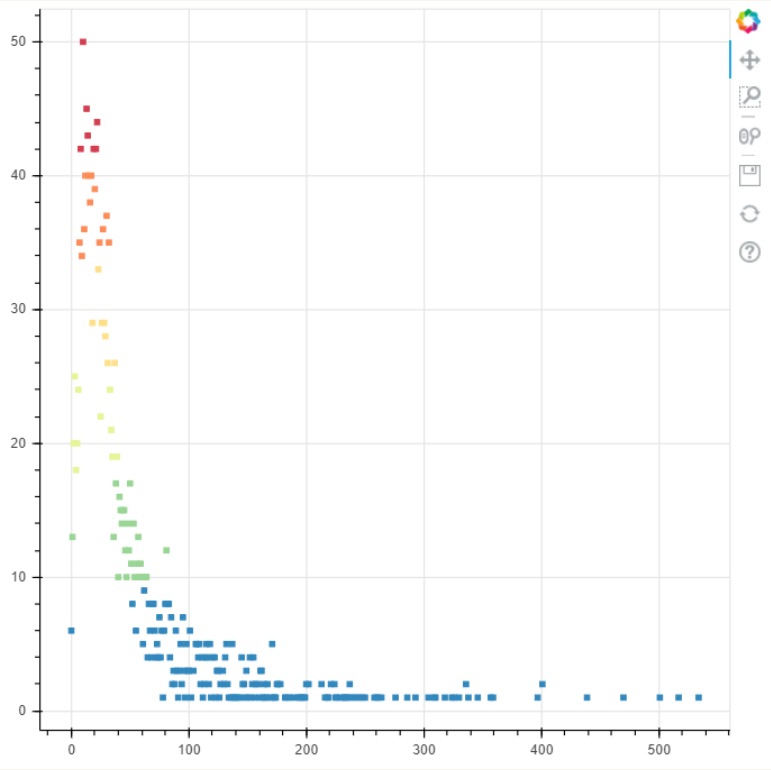
Nous remarquons que contrairement à ce que nous aurions pu penser, il n'y a pas eu plus d'achats durant les mois ayant le plus de Special Days (février et mai).

DATA VISUALISATION

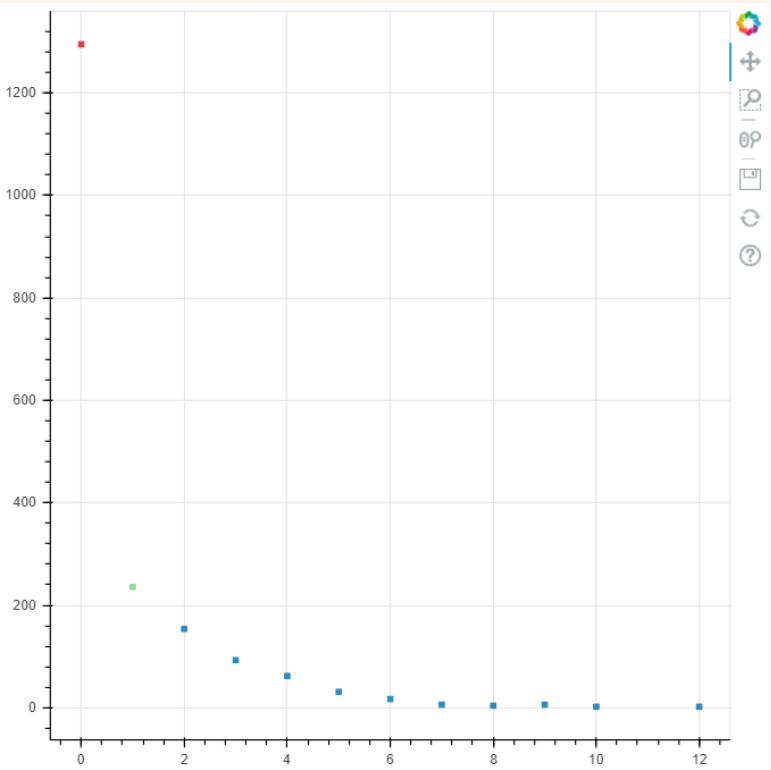
Nous avons créé 3 graphiques affichant le nombre d'achats effectués par nombre de site de produits/d'information/d'administration visités.



Nombre d'achats effectués par nombre de sites administratifs visités

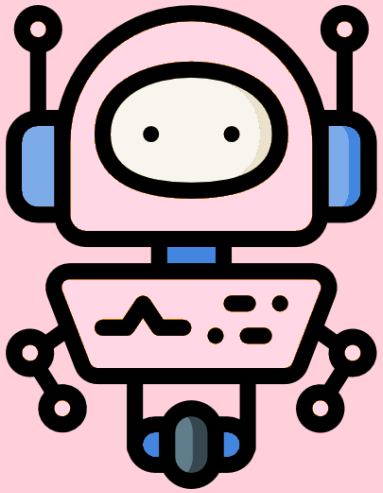


Nombre d'achats effectués par nombre de sites de produits visités



Nombre d'achats effectués par nombre de sites d'informations visités

Le maximum d'achats effectués est atteint vers 20 sites de produits visités. A partir de 50 sites de produits visités, il y a de moins en moins d'achats effectués. Et aux alentours de 170 sites de produits visités, le nombre d'achats effectués est proche de 0. On remarque pour les sites d'informations (à droite) et administratifs (à gauche) on a bien moins de points. Cela indique, comme on pouvait s'y attendre, que les personnes qui ont visité un site de produits sont bien plus susceptibles de faire un achat.

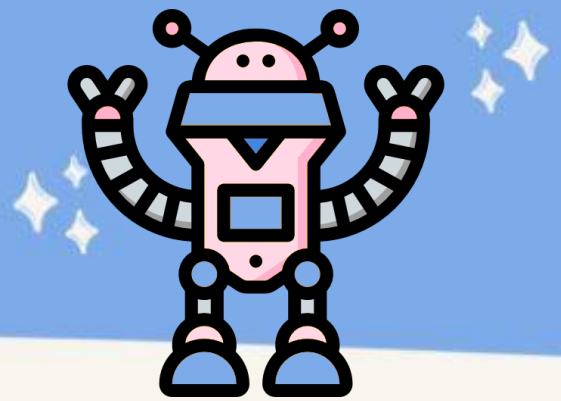


Sur les 12 330 sessions présentées dans le dataset, 84,5% (10 422) étaient des échantillons de classe négative c'est-à-dire qu'ils n'aboutissaient pas à un achat. Seulement 15,5% (1 908) étaient des échantillons de classe positive, aboutissant à un achat.

Ainsi, il se peut que notre modèle aie plus de difficultés à prédire un achat qu'une absence d'achat. Le nombre d'indices manque pour une précision optimale.

DES QUESTIONS ?
DES CLARIFICATIONS ?

N’hésitez pas à nous contacter par mail.



Khadidiatou BADJI

khadidiatou.badji@edu.devinci.fr

Vayshnavi SIVARAJAH

vayshnavi.sivarajah@edu.devinci.fr