AMSALLEM CÉCILE

ATTRAIT BASTIEN

PROJET PYTHON FOR DATA ANALYSIS

DATASET CHOISI

Dataset assigné à Bastien Attrait

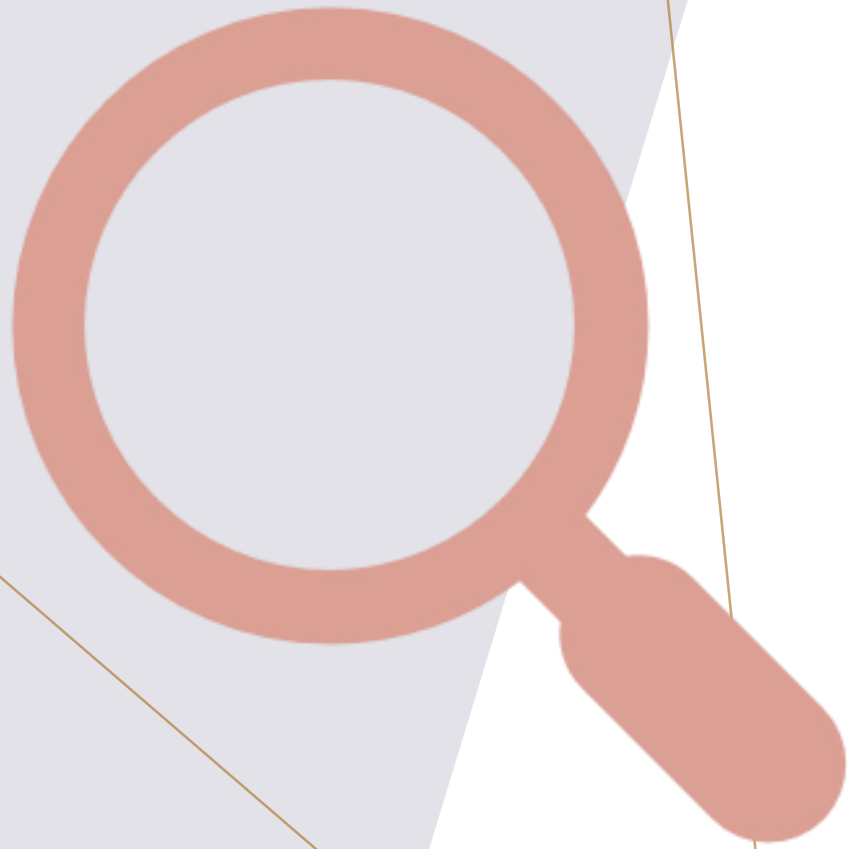
Online Shoppers Purchasing Intention Dataset

UCI MACHINE LEARNING REPOSITORY: ONLINE SHOPPERS PURCHASING INTENTION DATASET DATA SET

SOMMAIRE

DÉMARCHE SUIVIE

- Lecture de la description des deux datasets proposés et choix du dataset conservé.
- Importation et exploration des premières données (premières lignes, sommaire, type des colonnes, matrice de corrélation etc. …)
- Choix de questions pertinentes à poser sur ce jeu de données et étude des réponses associées.
- Réflexion sur le type d'algorithme de Machine Learning à utiliser en fonction du dataset (ici en l'occurrence, algorithme de classification de type clustering et un algorithme de prédiction en utilisant Revenue comme label).
- Mise en place de la méthode de KMeans. Variation des hyperparamètres.
- Etudes des résultats des clusters obtenus par l'application du KMeans avec les meilleurs paramètres.
- Mise en place d'un algorithme de prédiction. Variation des hyperparamètres.



EXPLORATION DE NOS DONNÉES

Description fournie par le dataset:

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another. The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to the 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. The value of "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. The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, 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. The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

A LA LECTURE DE CETTE DESCRIPTION NOUS AVONS DONC UN DATASET COMPOSÉ SUR UNE ANNÉE COMPRENANT LES INFORMATIONS RELATIVES AUX ACHATS EFFECTUÉS SUR UN SITE WEB.

NOTRE BUT PEUT DONC ÊTRE DE VOIR LES PARAMÈTRES AYANT UNE INFLUENCE SUR L'ACHAT OU NON D'UN PRODUIT ET DE CLASSER LES VISITEURS DU SITE INTERNET PAR UN ALGORITHME DE MACHINE LEARNING.

Administrative int64 : nombre de page administrative vues
Administrative_Duration float64 : temps passé sur les pages administratives
Informational int64 : nombre de page d'information vues
Informational_Duration float64 : temps passé sur les pages d'information
ProductRelated int64 : nombre de page produit vues
ProductRelated_Duration float64 : temps passé sur les pages produit
BounceRates float64 : pourcentage de visiteurs qui entrent sur le site à partir de cette page et la quittent sans déclencher d'autres demandes au serveur
ExitRates float64 : pourcentage de visiteurs qui quittent la session après avoir vu la page web
PageValues float64 : valeur moyenne d'une page web qu'un utilisateur a visitée avant de conclure une transaction**SpecialDay** float64 : indique la proximité d'un jour particulier (fête des mères, St Valentin etc. ..)
Month object : mois de l'année
OperatingSystems int64 : indique le système d'exploitation
Browser int64 : indique le navigateur utilisé
Region int64 : indique la région
TrafficType int64 : indique le type de trafic associé
VisitorType object : type de visiteur (nouveau, de retour ou autre)
Weekend bool : booléen indiquant si la visite a eu lieu un week-end
Revenue bool : indique si un achat a été effectué

EXPLORATION DU DATASET

LES COLONNES ET TYPES DES DONNÉES

	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	0.20	0.20	0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
1	0	0.0	0	0.0	2	64.000000	0.00	0.10	0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
2	0	0.0	0	0.0	1	0.000000	0.20	0.20	0.0	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
3	0	0.0	0	0.0	2	2.666667	0.05	0.14	0.0	0.0	Feb	3	2	2	4	Returning_Visitor	False	False
4	0	0.0	0	0.0	10	627.500000	0.02	0.05	0.0	0.0	Feb	3	3	1	4	Returning_Visitor	True	False

EXPLORATION DU DATASET

LES 5 PREMIÈRES LIGNES

	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

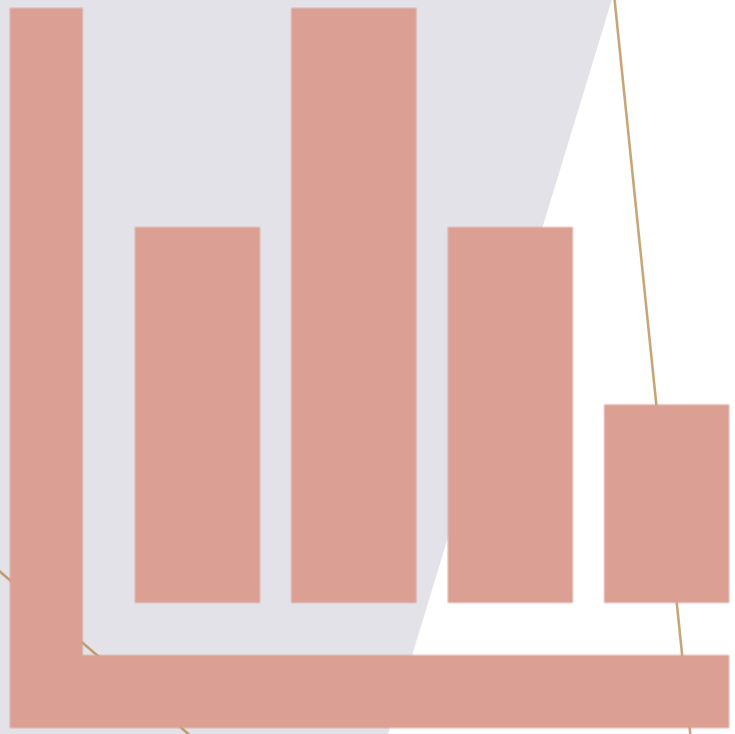
EXPLORATION DU DATASET

LES QUARTILES

AJOUT DE NOUVELLES COLONNES :

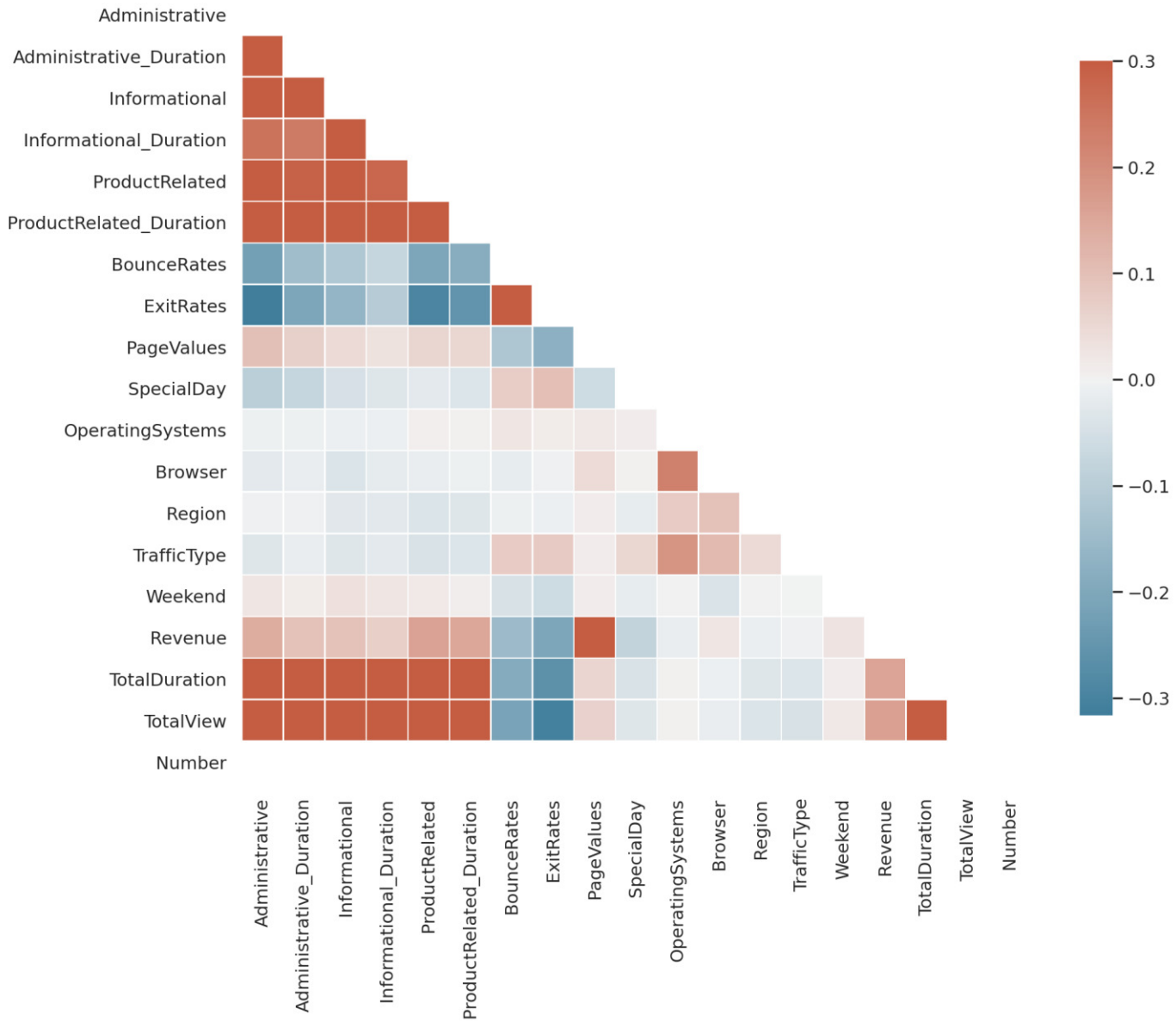
TOTALVIEW : INT LE NOMBRE DE VUES DE TOUS LES TYPES DE PAGES CONFONDUS

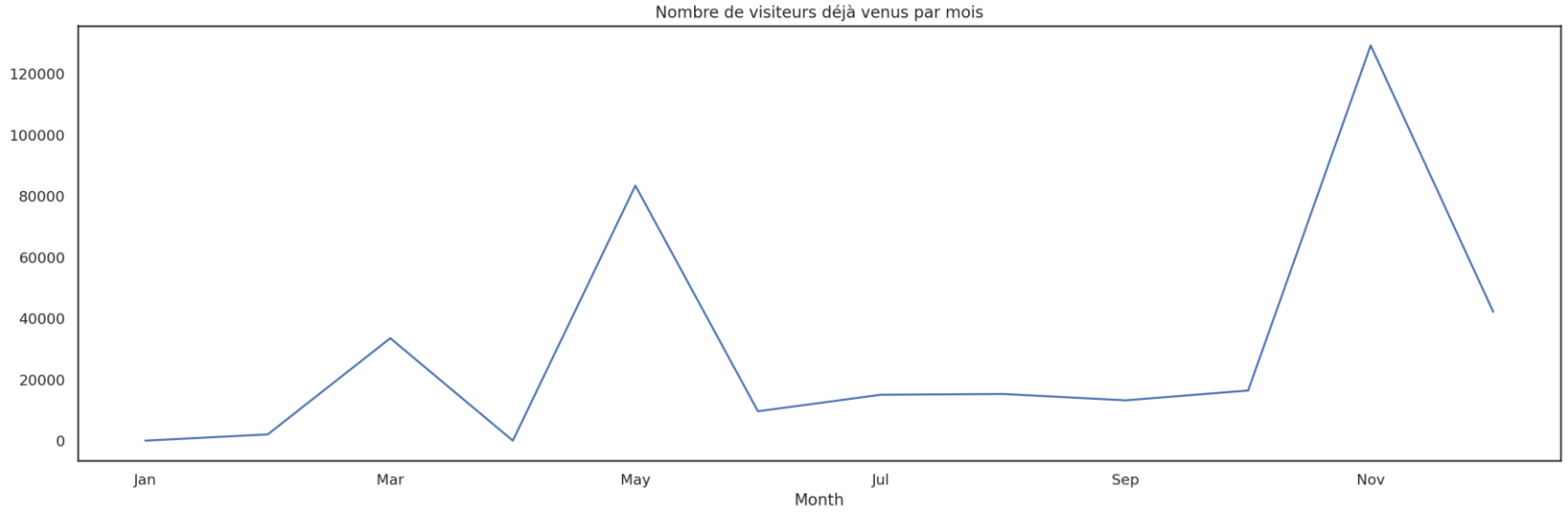
TOTALDURATION : FLOAT LE TEMPS TOTAL DÉPENSÉ SUR TOUS LES TYPES DE PAGES CONFONDUS



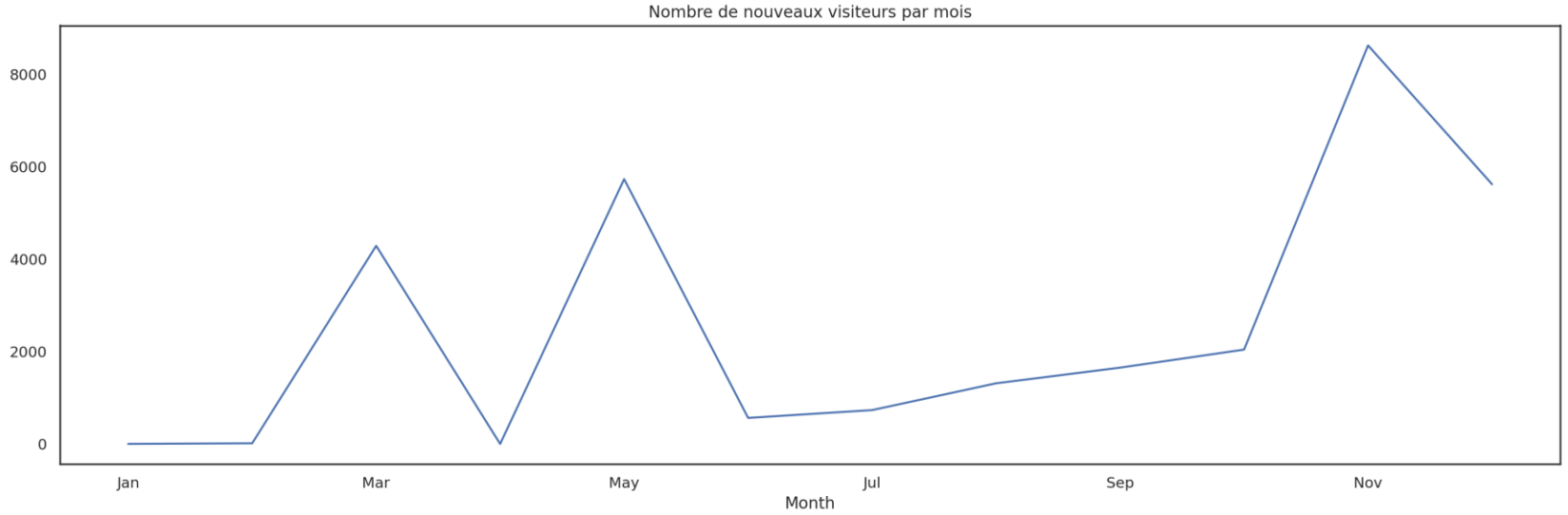
ANALYSE DES DONNÉES

MATRICE DE CORRÉLATION DE NOS DONNÉES



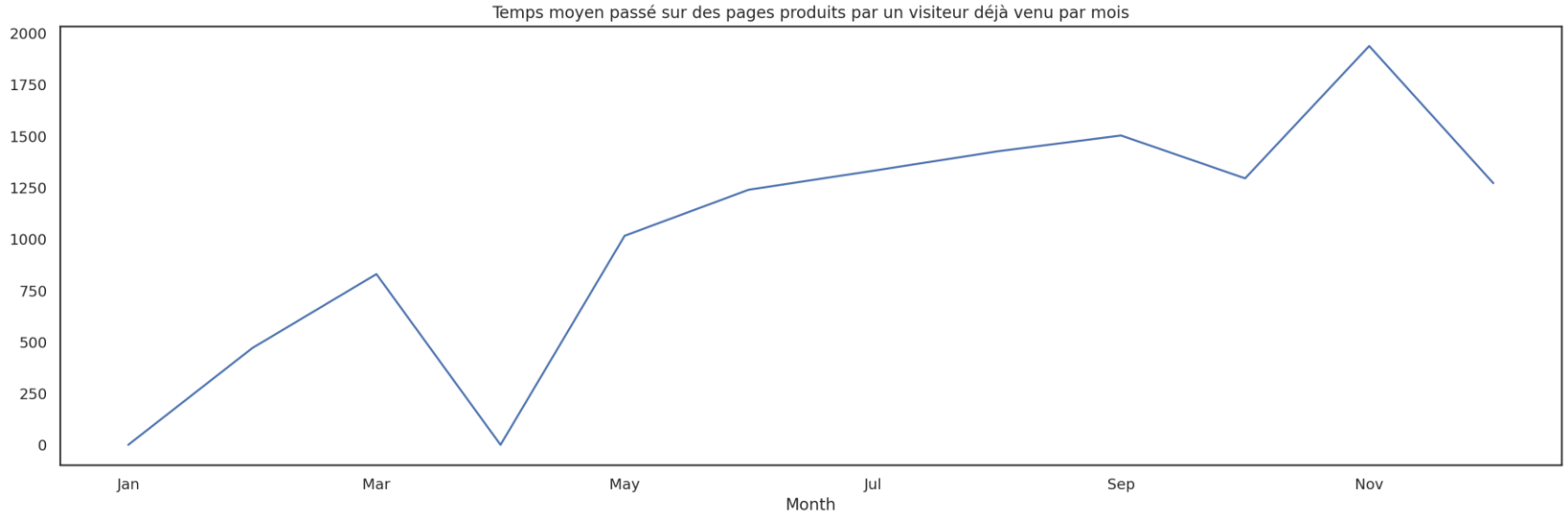


NOMBRE DE VISITEURS DÉJÀ VENUS PAR MOIS

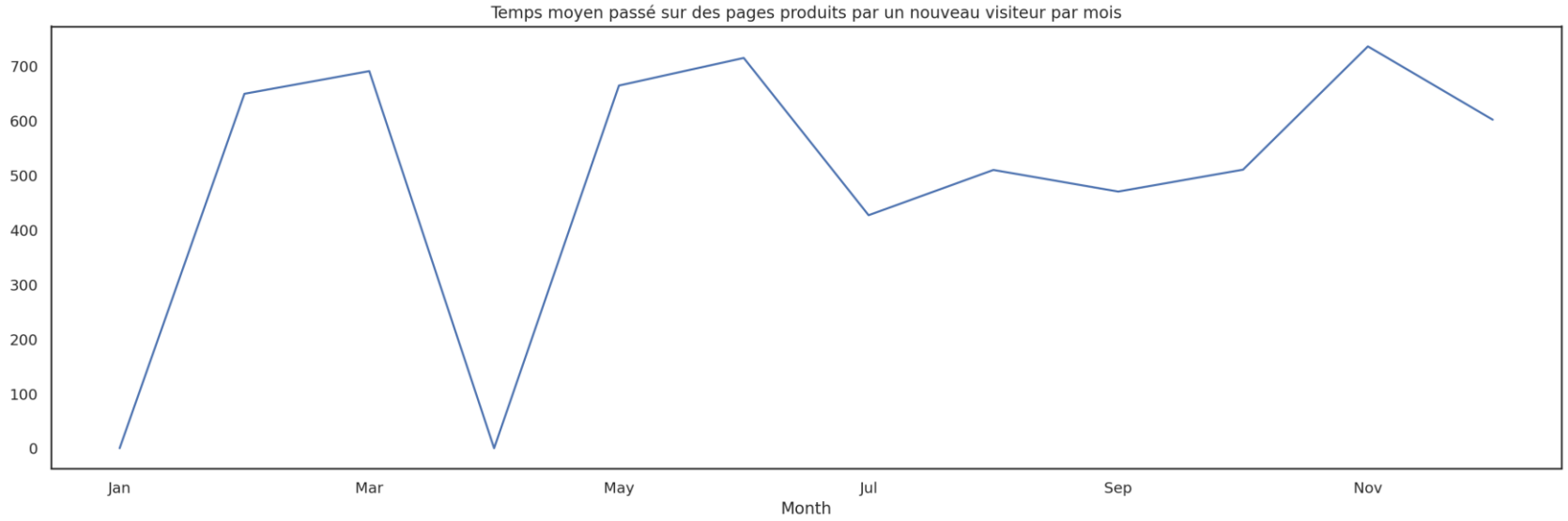


NOMBRE DE NOUVEAUX VISITEURS PAR MOIS

Les deux graphes sont similaires dans leur forme. Les nouveaux visiteurs restent dans les mêmes proportions.

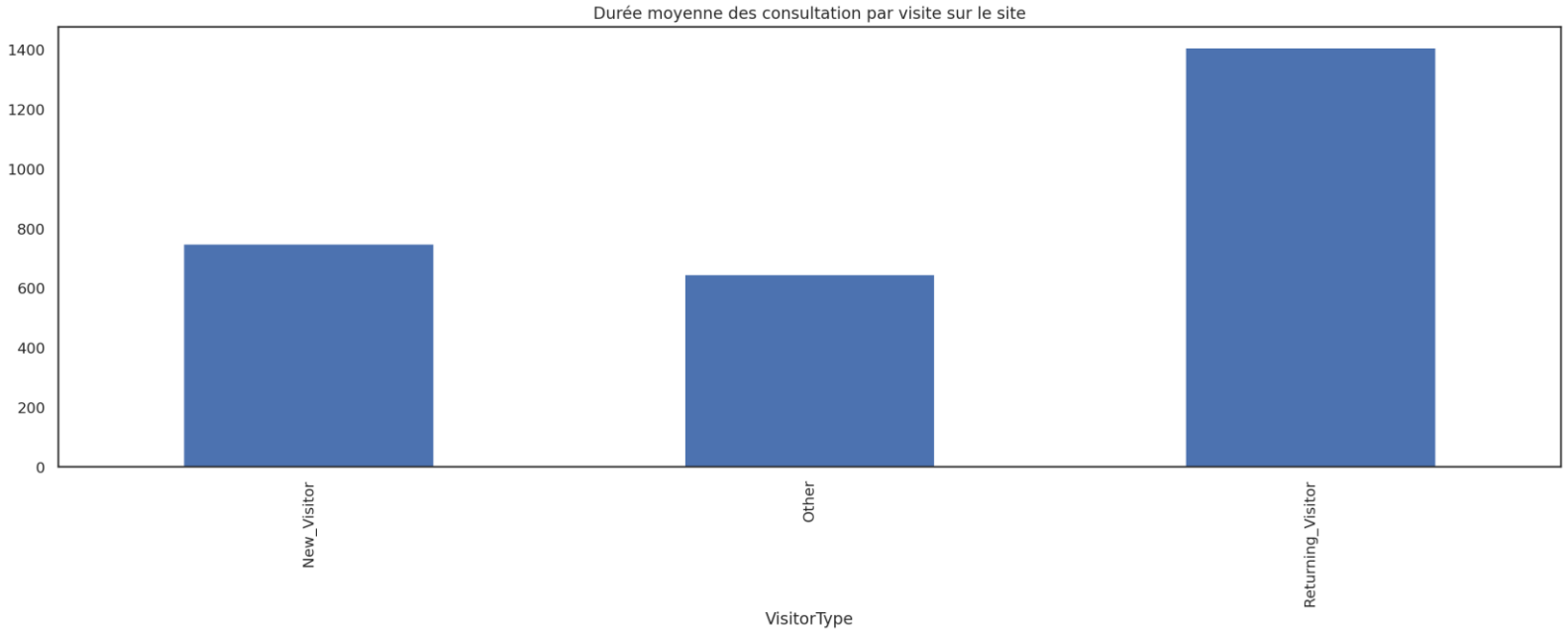


TEMPS MOYEN PASSE SUR DES PAGES PRODUITS PAR DES VISITEURS DÉJÀ VENUS PAR MOIS



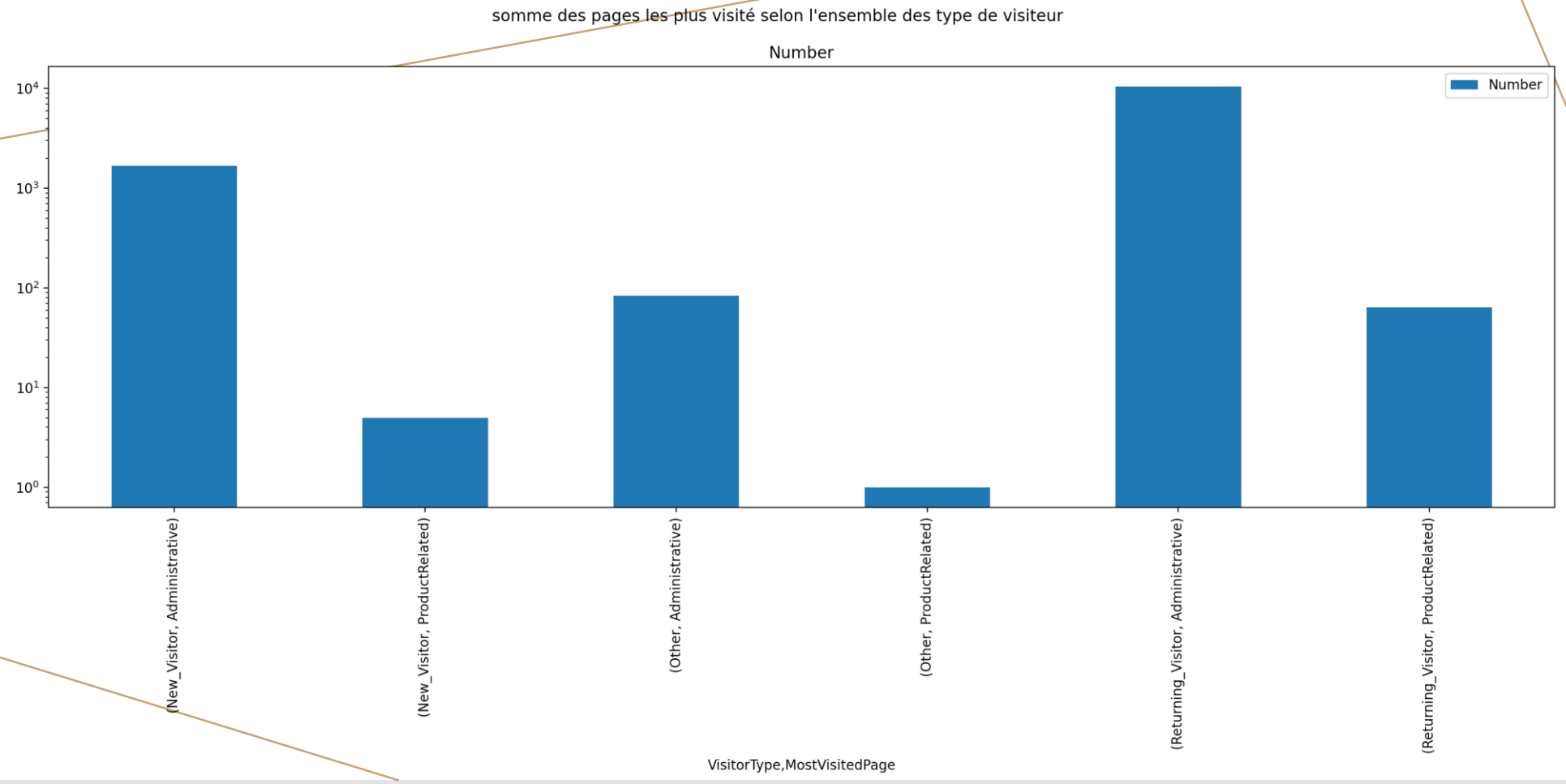
TEMPS MOYEN PASSE SUR DES PAGES PRODUITS PAR DE NOUVEAUX VISITEURS PAR MOIS

Les graphes et les échelles varient nettement plus sur ces données.



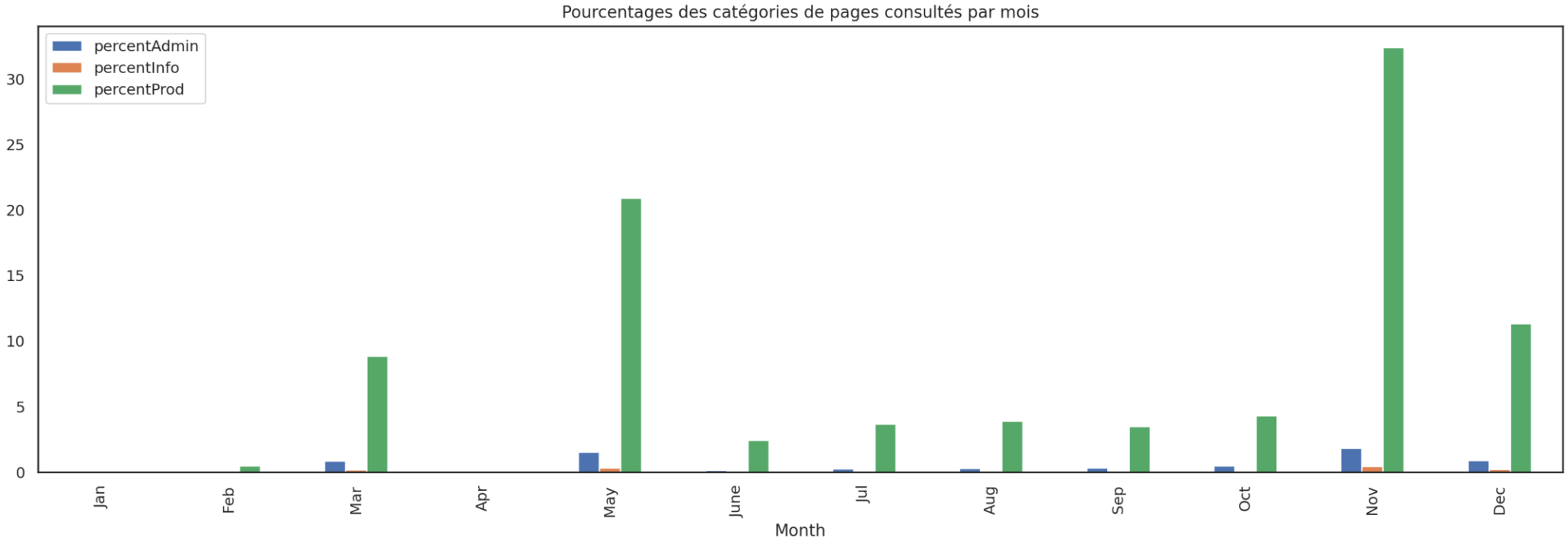
LES NOUVEAUX VISITEURS PASSENT EN MOYENNE MOINS DE TEMPS SUR LE SITE QUE LES VISITEURS DÉJÀ VENUS.

EST-CE QU'IL EXISTE UNE DIFFÉRENCE DE TEMPS PASSÉ ENTRE LES DIFFÉRENTS TYPES DE VISITEURS ?

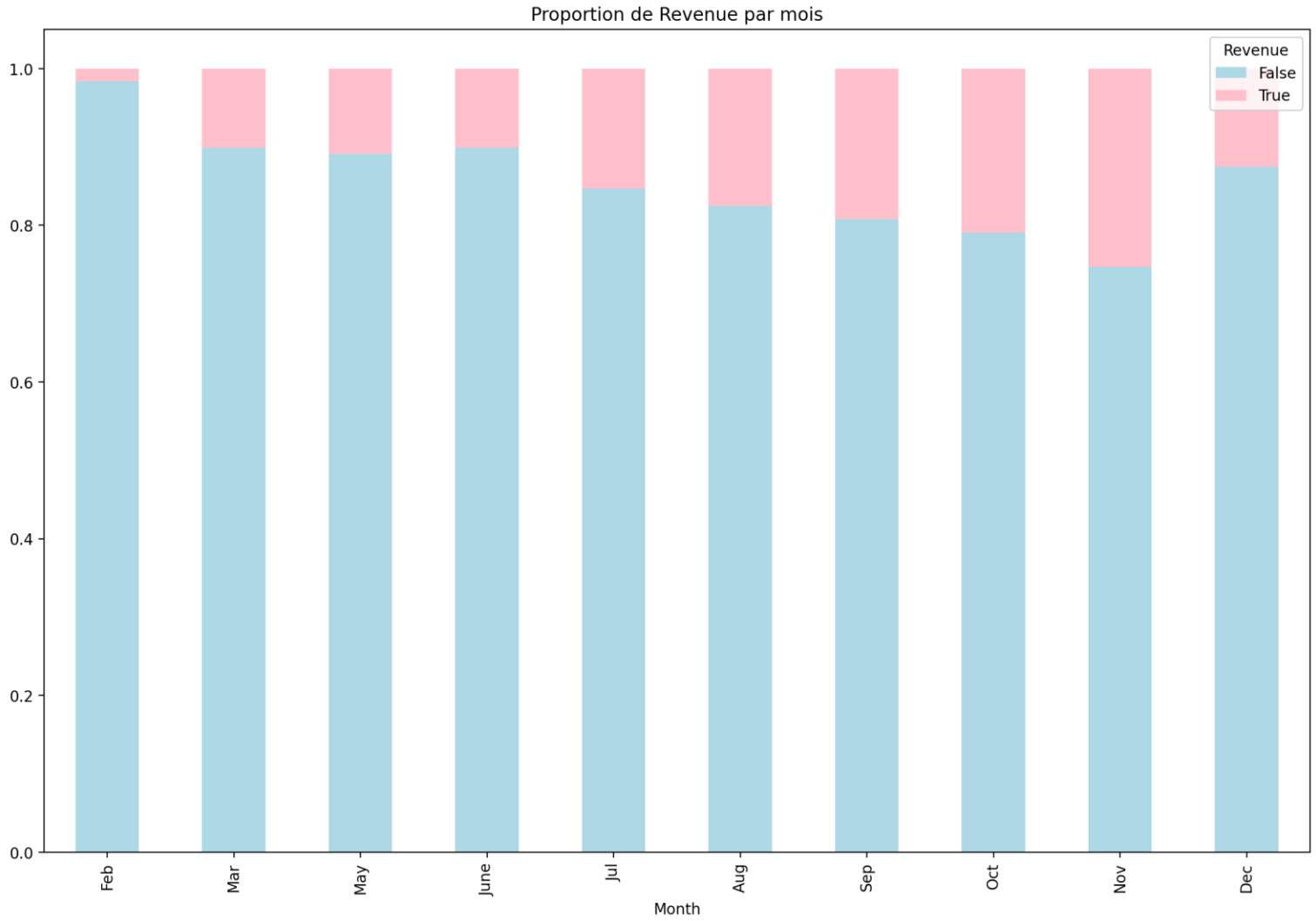


Quels sont les types de pages les plus vues par les différents types de visiteurs ?

Pourcentage de visite administrative parmi les Returning_Visitor : 6.1895918896774 22
Pourcentage de visite de produit parmi les Returning_Visitor : 92.3645946508241
Pourcentage de visite administrative parmi les New_Visitor : 12.186733572012516
Pourcentage de visite de produit parmi les New_Visitor : 86.22050573675753
Pourcentage de visite administrative parmi les Other : 10.416666666666668
Pourcentage de visite de produit parmi les Other : 88.33333333333333

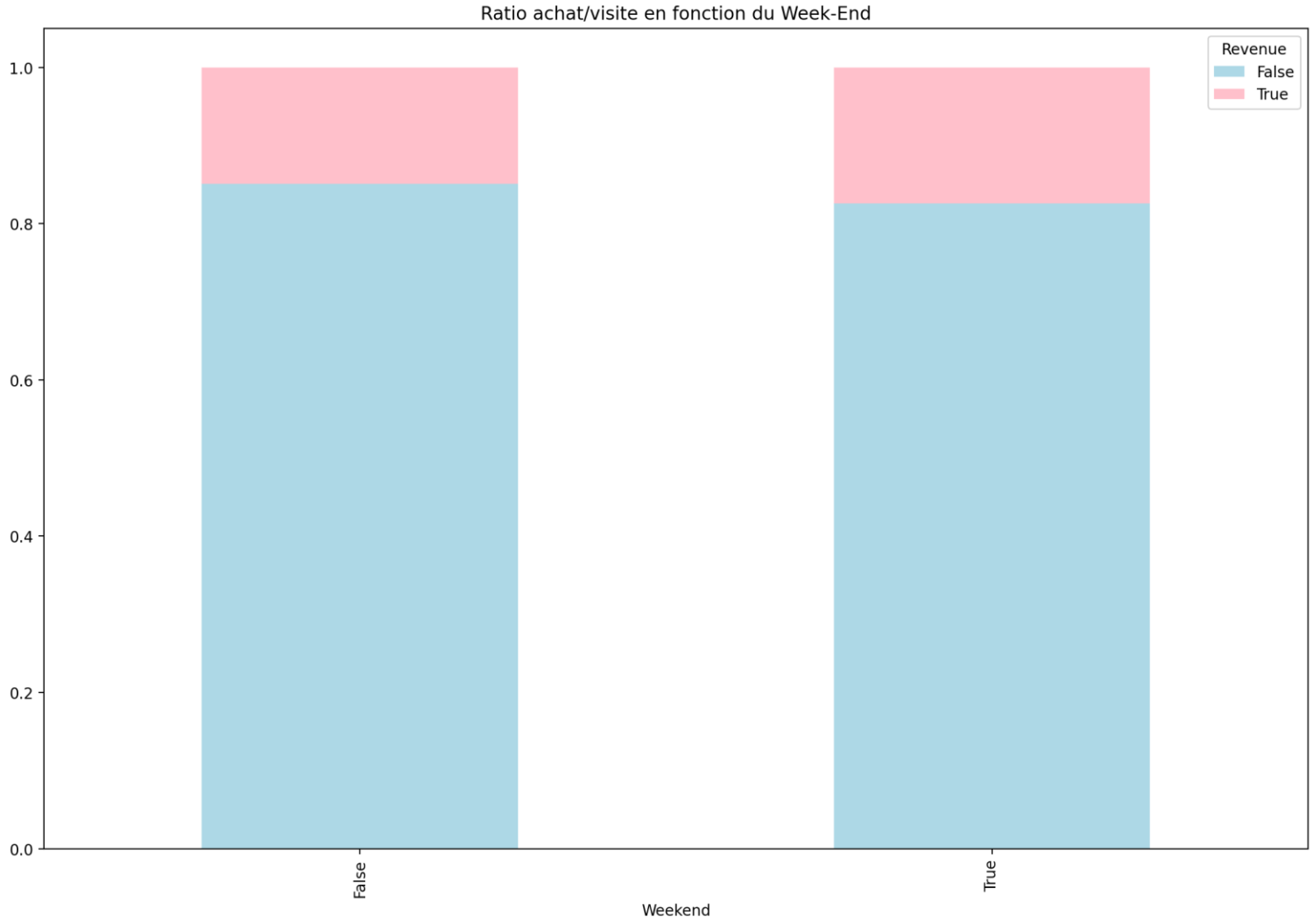


REGARDONS LES POURCENTAGES DE CATÉGORIES DE PAGES CONSULTÉES PAR MOIS



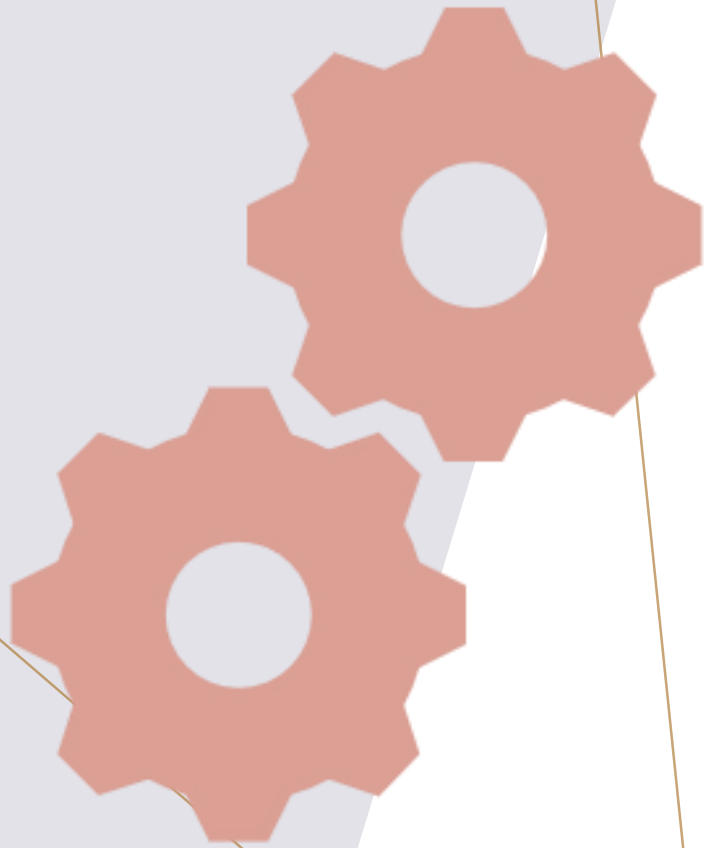
EST-CE QUE LES PROPORTIONS D'ACHAT / VISITE CHANGENT EN FONCTION DU MOIS ?

On constate que les mois de Août à Novembre ont de plus grande portion d'achat. Ne connaissant pas le site à l'origine de ce dataset, nous ne pouvons savoir quelles en sont les justifications.



EST-CE QUE LES PROPORTIONS D'ACHAT / VISITE CHANGENT EN FONCTION DE LA PÉRIODE DE LA SEMAINE?

On constate que les ratios ne changent pas en fonction que ce soit le week-end ou la semaine hors week-end.



MACHINE LEARNING

PARAMÈTRES DE L'ALGORITHME KMEANS

On transforme nos données non-numériques ou booléenne en données numériques par le one hot encoding.

Nos paramètres sont :
n_clusters : le nombre de cluster
n_init : le nombre de fois où l'algorithme se lance avec des centroïdes aléatoires
max_iter : le nombre maximal d'iteration de l'algorithme
tol : 0.01 : la tolerance, ou taux d'apprentissage

LA MÉTRIQUE CHOISIE EST LA SOMME DES DISTANCES CARRÉS AU CENTROÏDE DU CLUSTER.

K \ n	10	20	30	40	50	60	70	80	90	100	110	120	130
1	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10	9.703371 85e+10
2	4.592670 64e+10	4.592670 64e+10	4.592670 64e+10	4.592670 64e+10	4.592670 64e+10	4.592670 64e+10	4.592670 64e+10	4.592670 64e+10	4.592657 65e+10	4.592657 65e+10	4.592657 65e+10	4.592657 65e+10	4.592657 65e+10
3	2.825941 79e+10	2.825941 79e+10	2.825939 84e+10	2.825939 84e+10	2.825939 84e+10	2.825939 84e+10	2.825939 84e+10	2.825939 84e+10	2.825939 84e+10	2.825939 84e+10	2.825939 84e+10	2.825939 84e+10	2.825939 84e+10
4	1.844570 07e+10	1.844570 07e+10	1.844570 07e+10	1.844535 94e+10	1.844444 01e+10	1.844441 06e+10	1.844441 06e+10	1.844441 06e+10	1.844441 06e+10	1.844413 33e+10	1.844413 33e+10	1.844413 33e+10	1.844413 33e+10
5	1.260100 22e+10	1.260100 22e+10	1.260100 22e+10	1.260100 22e+10	1.260100 22e+10	1.260100 22e+10	1.260094 83e+10	1.260094 83e+10	1.260094 83e+10	1.260094 83e+10	1.260094 83e+10	1.260094 83e+10	1.260094 83e+10
6	9.226420 81e+09	9.226420 81e+09	9.225873 03e+09	9.225873 03e+09	9.225873 03e+09	9.225873 03e+09	9.225873 03e+09	9.225873 03e+09	9.225873 03e+09	9.225873 03e+09	9.225873 03e+09	9.225814 73e+09	9.225814 73e+09
7	6.757528 40e+09	6.757528 40e+09	6.757528 40e+09	6.757528 40e+09	6.757528 40e+09	6.757516 33e+09	6.757445 42e+09	6.757445 42e+09	6.757445 42e+09	6.757301 80e+09	6.757301 80e+09	6.757301 80e+09	6.757301 80e+09
8	5.172403 87e+09	5.172403 87e+09	5.172403 87e+09	5.172403 87e+09	5.172396 20e+09	5.172396 20e+09	5.172396 20e+09	5.172396 20e+09	5.172396 20e+09	5.172396 20e+09	5.172396 20e+09	5.172396 20e+09	5.172396 20e+09
9	4.230785 39e+09	4.230785 39e+09	4.230785 39e+09	4.227305 79e+09	4.226894 95e+09	4.226894 95e+09	4.226894 95e+09	4.226894 95e+09	4.226894 95e+09	4.226894 95e+09	4.225922 32e+09	4.225922 32e+09	4.225922 32e+09
10	3.565324 72e+09	3.562931 02e+09	3.562931 02e+09	3.562931 02e+09	3.562931 02e+09	3.562931 02e+09	3.562922 01e+09	3.562922 01e+09	3.562869 75e+09	3.562869 75e+09	3.562869 75e+09	3.562869 75e+09	3.562869 75e+09
11	3.103999 84e+09	3.103999 84e+09	3.103999 84e+09	3.102970 96e+09	3.102970 96e+09	3.102970 96e+09	3.102501 39e+09	3.102314 08e+09	3.102314 08e+09	3.102314 08e+09	3.102069 43e+09	3.102069 43e+09	3.102069 43e+09
12	2.728458 40e+09	2.720027 72e+09	2.720027 72e+09	2.720027 72e+09	2.720027 72e+09	2.714219 07e+09	2.714219 07e+09	2.714219 07e+09	2.714219 07e+09	2.714219 07e+09	2.714219 07e+09	2.713662 45e+09	2.713662 45e+09
13	2.400599 20e+09	2.394292 66e+09	2.394292 66e+09	2.394292 66e+09	2.394292 66e+09	2.394292 66e+09	2.394292 66e+09	2.394292 66e+09	2.394292 66e+09	2.392671 55e+09	2.392671 55e+09	2.392671 55e+09	2.392671 55e+09

Tableau des variations des valeurs de K le nombre de cluster et n le nombre minimal de fois où l'algorithme KMeans

t \ m	100	200	300	400	500	600	700	800	900	1000	1100	1200	1300
0.0001	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09
0.0002	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09
0,0003	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09	2.145300 18e+09
0,0004	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09
0,0005	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09	2.145343 68e+09
0,0006	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09	2.145486 72e+09
0,0007	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09
0,0008	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09
0,0009	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09
0,0010	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09
0,0011	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09
0,0012	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09
0,0013	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09	2.146564 63e+09

Tableau des variations des valeurs de t le taux d’apprentissage et m le nombre maximal d’itérations

COMPARAISON ET COMMENTAIRES DES VALEURS OBTENUES

Concernant K :
Quand K = 1 les résultats ne changent pas, puisqu'il ne peut y avoir qu'un seul cluster. Comme les meilleurs résultats sont pour K = 13, il n'y a rien de surprenant à cela, si nous augmentions encore K, nous aurions une somme de l'erreur moyenne toujours plus faible, cependant il n'est pas pertinent d'avoir un nombre de cluster trop grand. Visualisons nos données pour K = 13 puis jugeons si nous devons l'augmenter ou le diminuer.

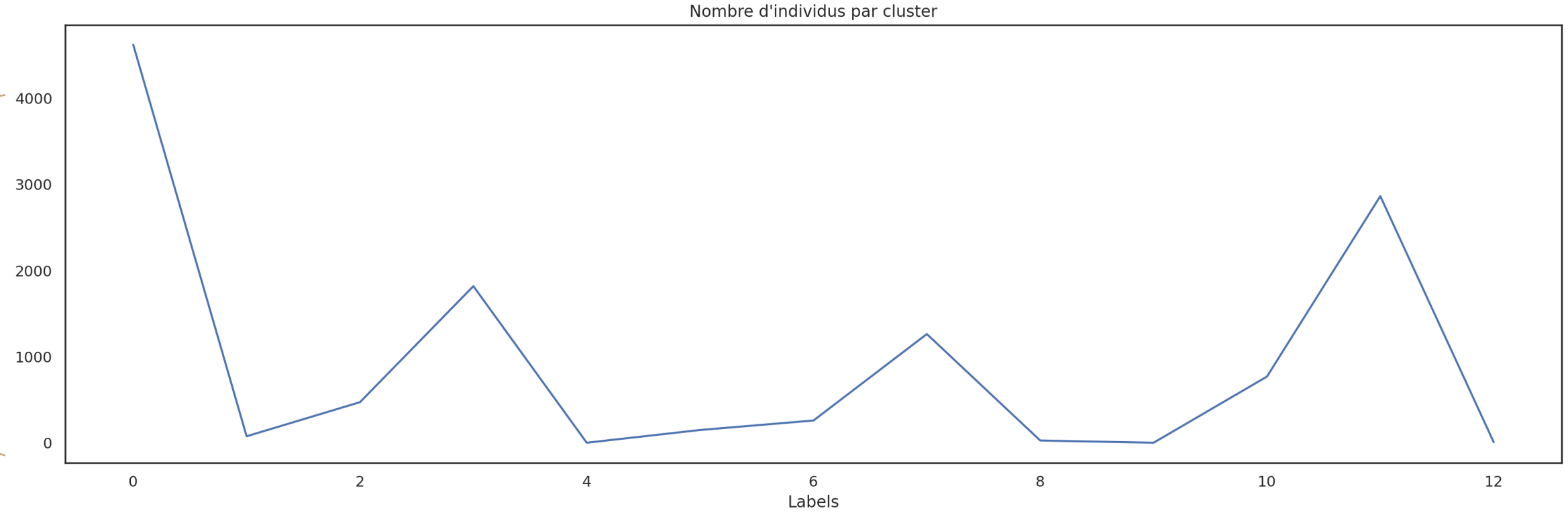
Concernant n:
Les meilleurs résultats sont pour n = 40.

Concernant t:
Les meilleurs résultats sont pour t = 0,001.

Concernant m:
Les meilleurs résultats sont pour m = 100.

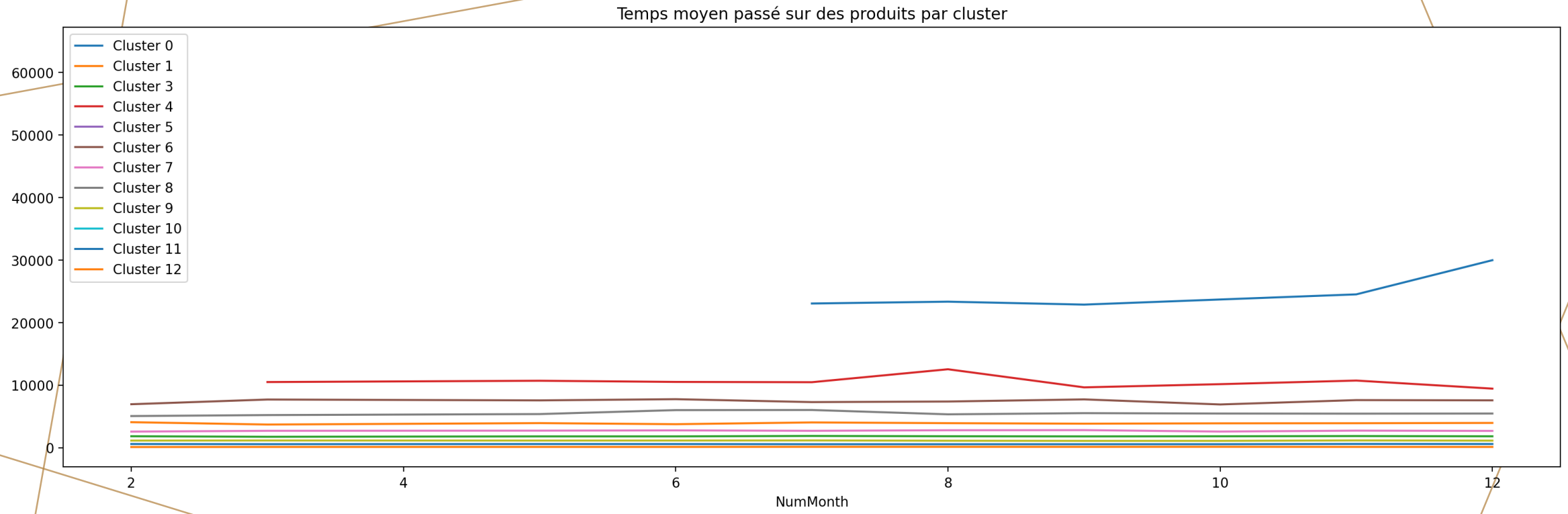
NOTONS QUE LES GRAPHIQUES SUIVANTS NE SONT PAS CEUX DU NOTEBOOK PUISQUE LORSQUE L'ON RELANCE LE NOTEBOOK LA RÉPARTITION DES INDIVIDUS DANS LES CLUSTERS PEUT VARIER AINSI QUE LE NUMÉRO DU CLUSTER ASSOCIÉ.

VISUALISATION DES RESULTATS POUR K = 13



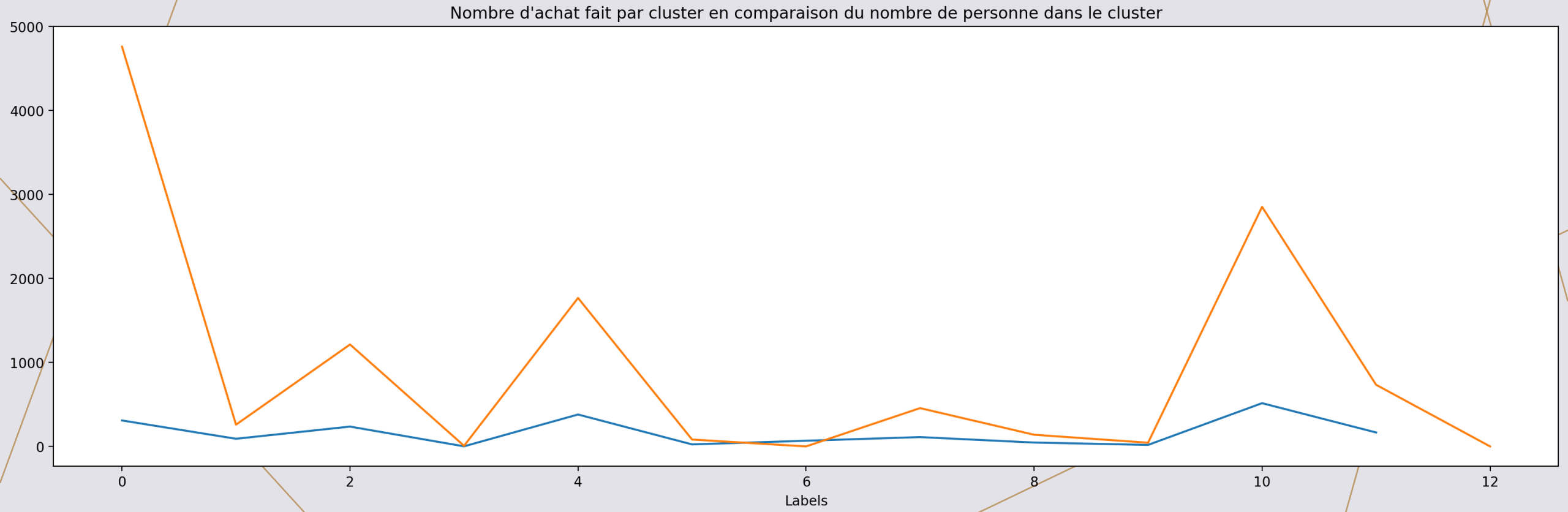
NOMBRE D'INDIVIDUS PAR CLUSTER

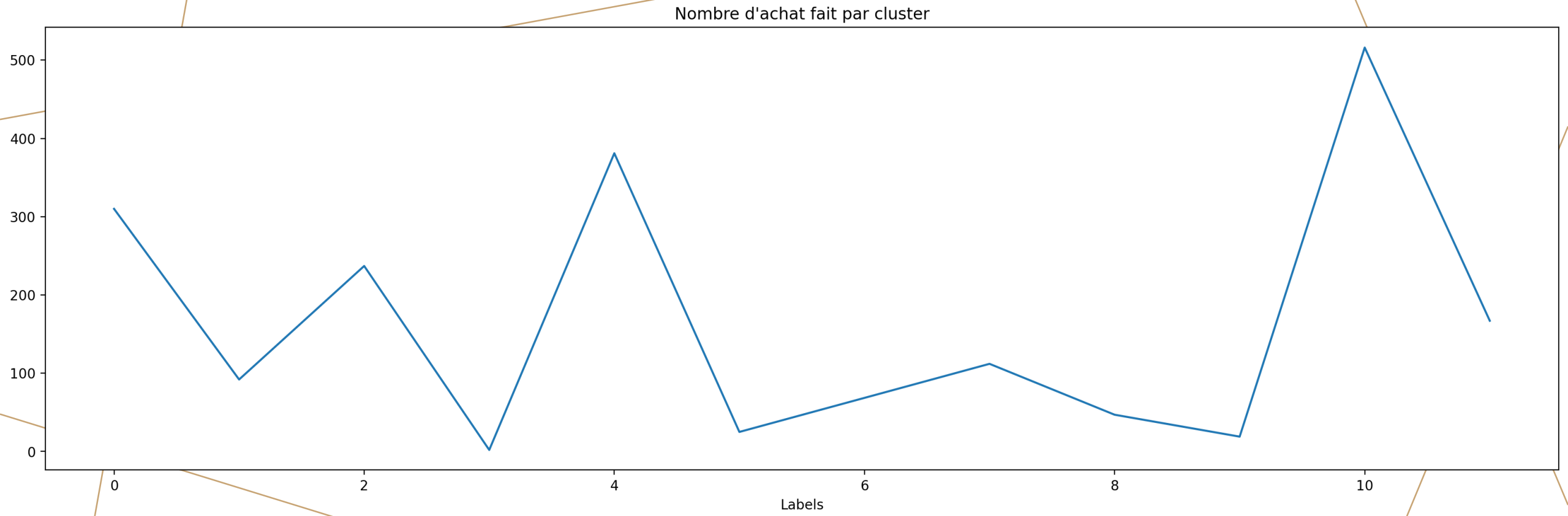
On constate que certains clusters sont presque vides, d'autres sont très remplis. Il sera donc intéressant de tester avec un K plus petit, afin de voir les différences qui peuvent apparaître.



ON REMARQUE QUE CERTAINS CLUSTERS NE CONTIENNENT PAS D'INFORMATION POUR CERTAINS MOIS

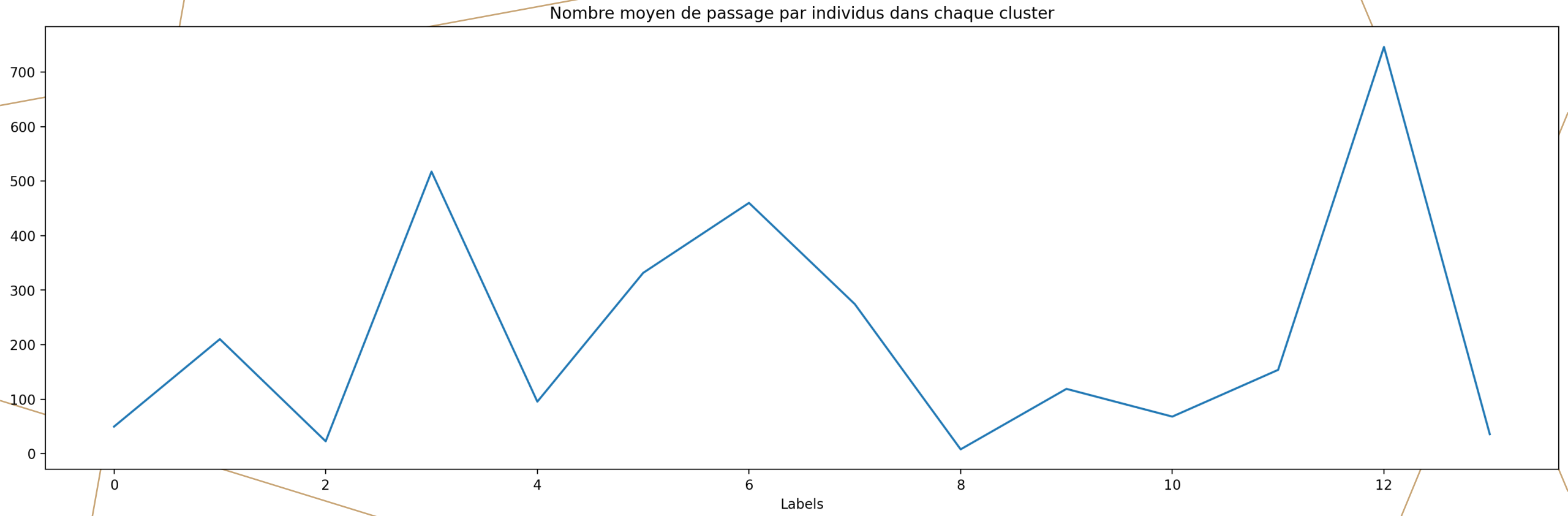
NOMBRE D'ACHAT ET NOMBRE D'INDIVIDUS PAR CLUSTER



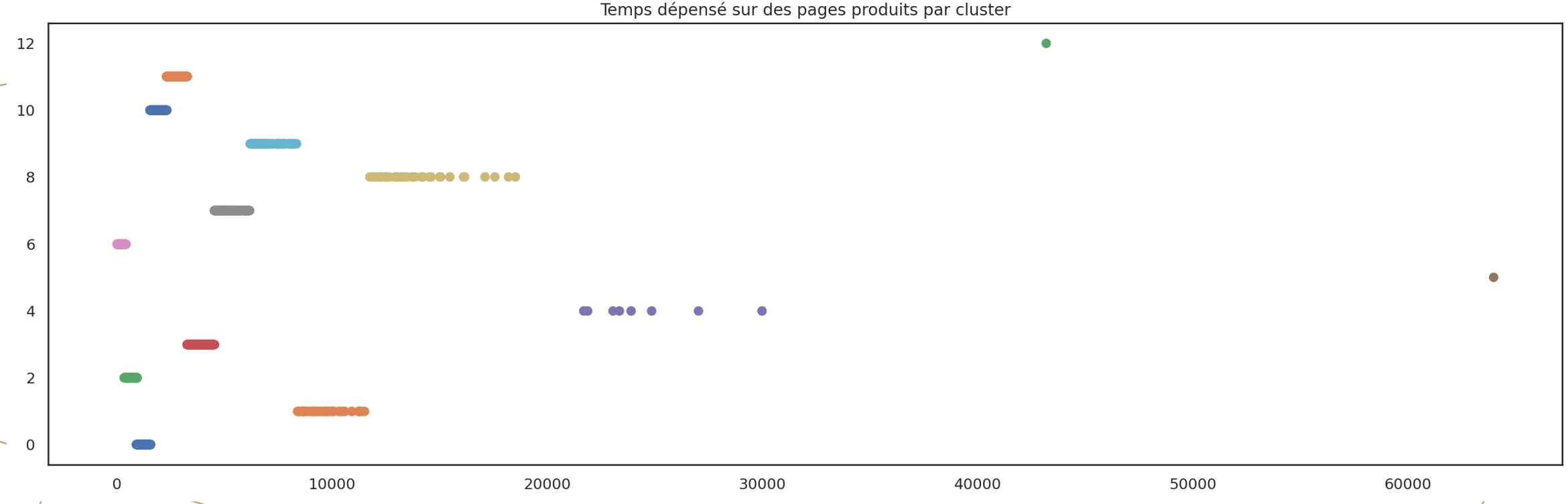


Quels sont les pourcentages d’achat par cluster ?

Pourcentage d'achat dans le cluster 0 : 6.509869802603947
Pourcentage d'achat dans le cluster 1 : 35.38461538461539
Pourcentage d'achat dans le cluster 2 : 19.50617283950617
Pourcentage d'achat dans le cluster 3 : 25.0
Pourcentage d'achat dans le cluster 4 : 21.5375918598078
Pourcentage d'achat dans le cluster 5 : 30.120481927710845
Pourcentage d'achat dans le cluster 6 : 0.0
Pourcentage d'achat dans le cluster 7 : 24.50765864332604
Pourcentage d'achat dans le cluster 8 : 33.57142857142857
Pourcentage d'achat dans le cluster 9 : 42.22222222222222
Pourcentage d'achat dans le cluster 10 : 18.07988787666433
Pourcentage d'achat dans le cluster 11 : 22.721088435374153
Pourcentage d'achat dans le cluster 12 : 0.0

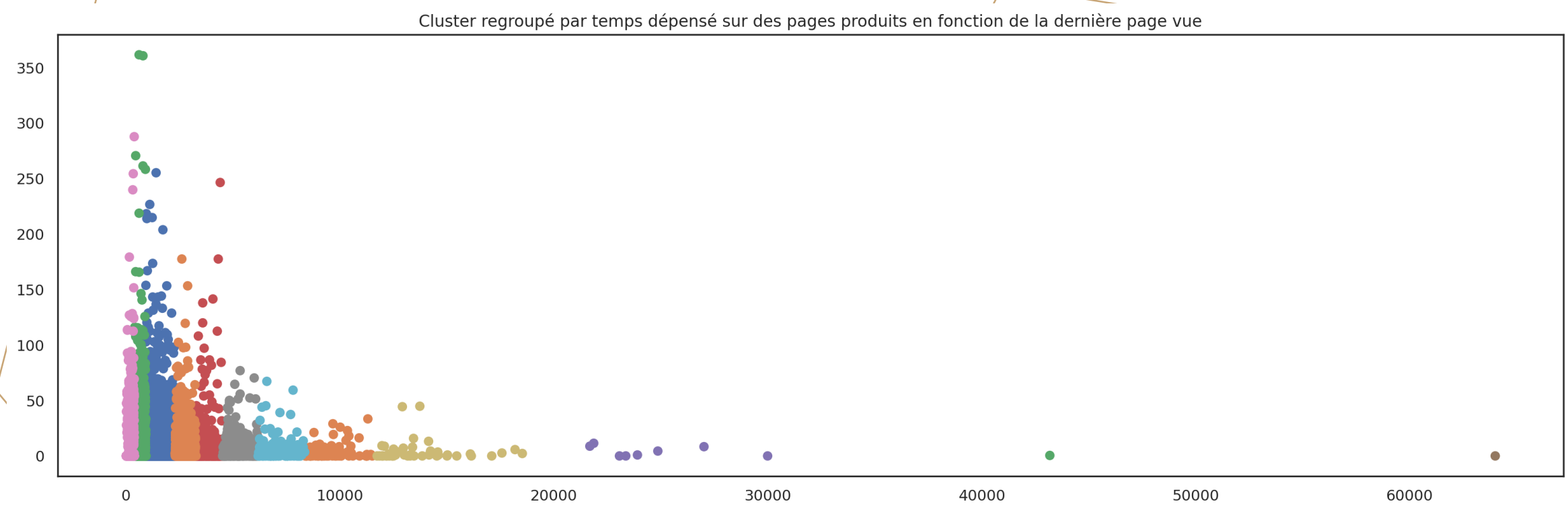


NOMBRE MOYEN DE PAGES VUES PAR LES INDIVIDUS DE CHAQUE CLUSTER



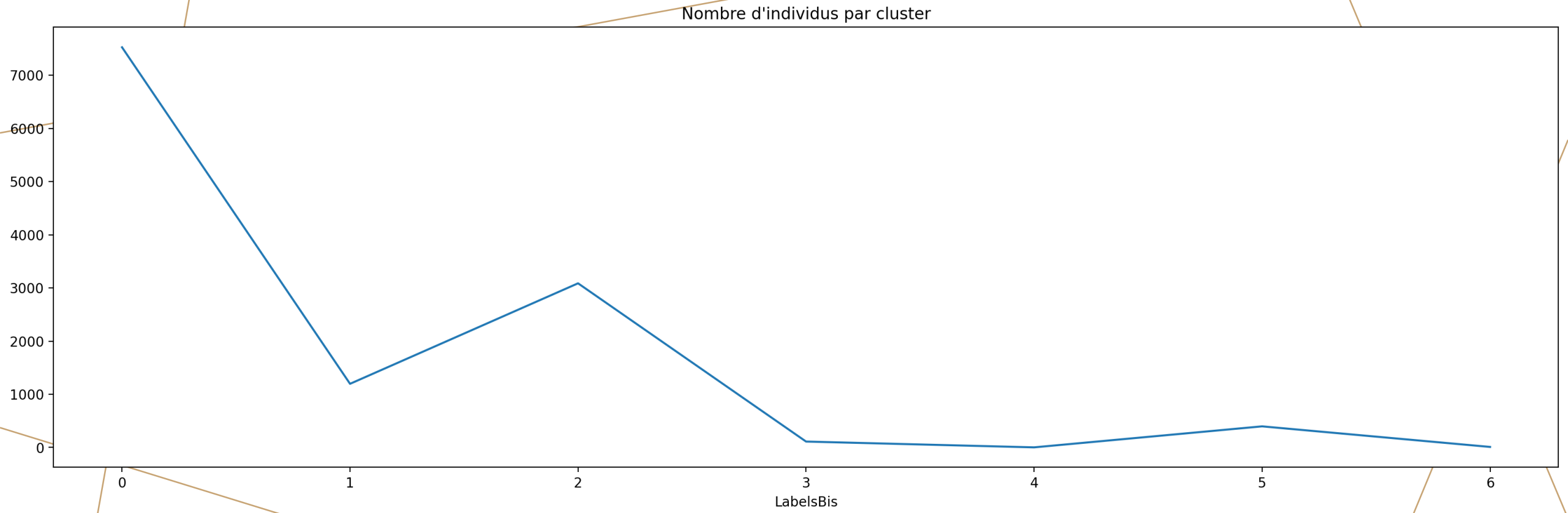
KMEANS

TEMPS PASSÉ SUR DES PAGES PRODUITS PAR LES INDIVIDUS DE CHAQUE CLUSTER



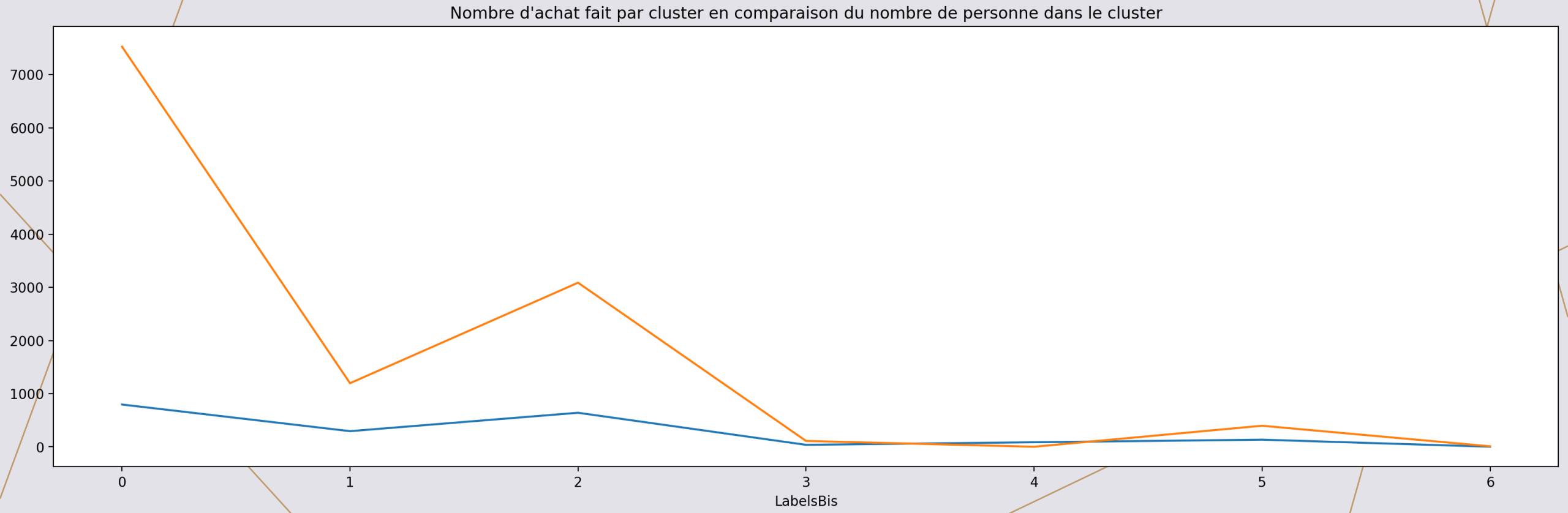
INDIVIDUS REGROUPÉ PAR CLUSTER AFFICHÉ PAR DERNIÈRE PAGE VUES EN FONCTION DU TEMPS DÉPENSÉ SUR LES PAGES PRODUITS

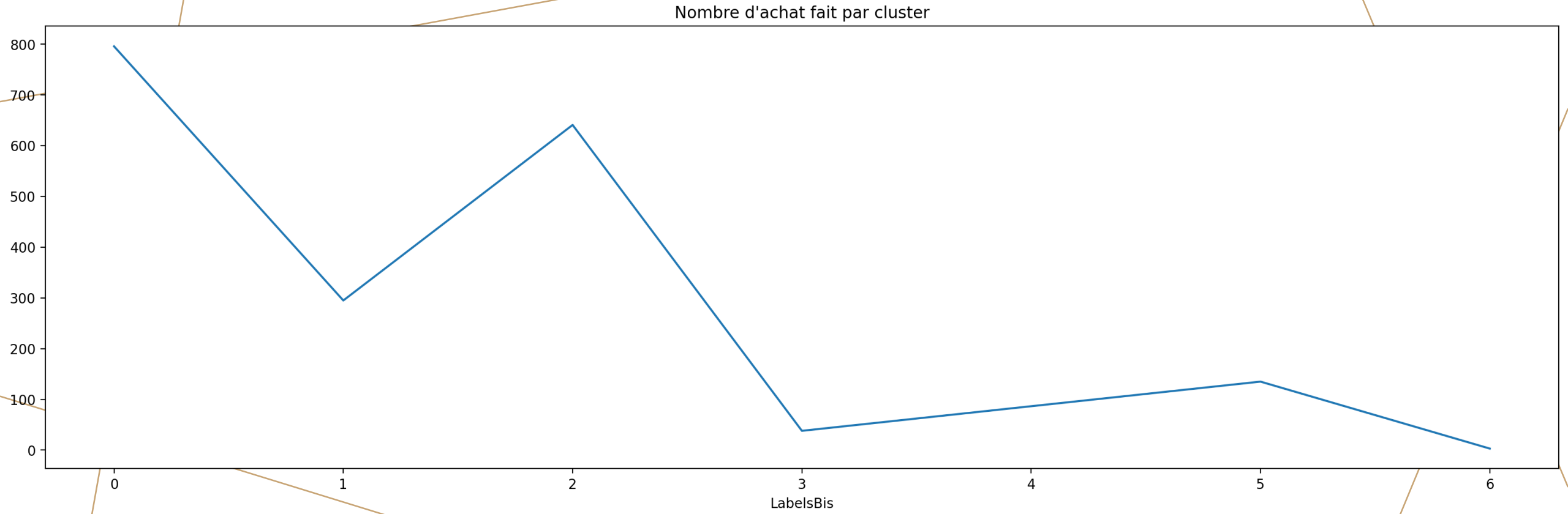
VISUALISATION DES RESULTATS POUR K = 7



NOMBRE D’INDIVIDUS PAR CLUSTER

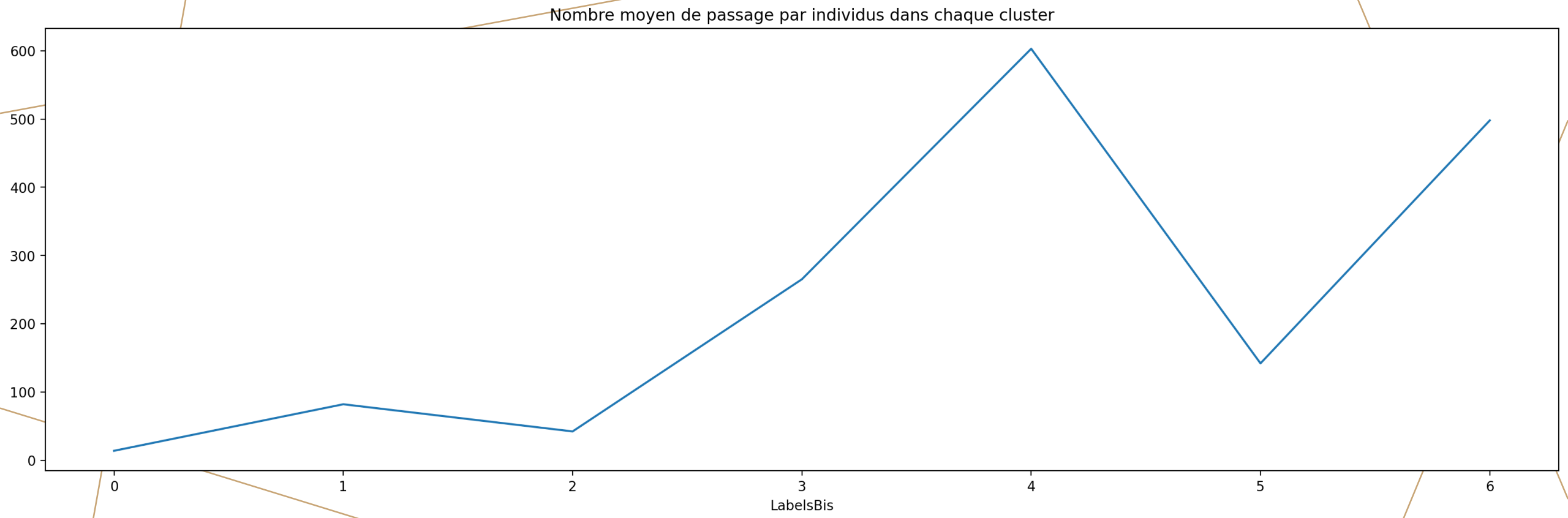
NOMBRE D'ACHAT ET NOMBRE D'INDIVIDUS PAR CLUSTER



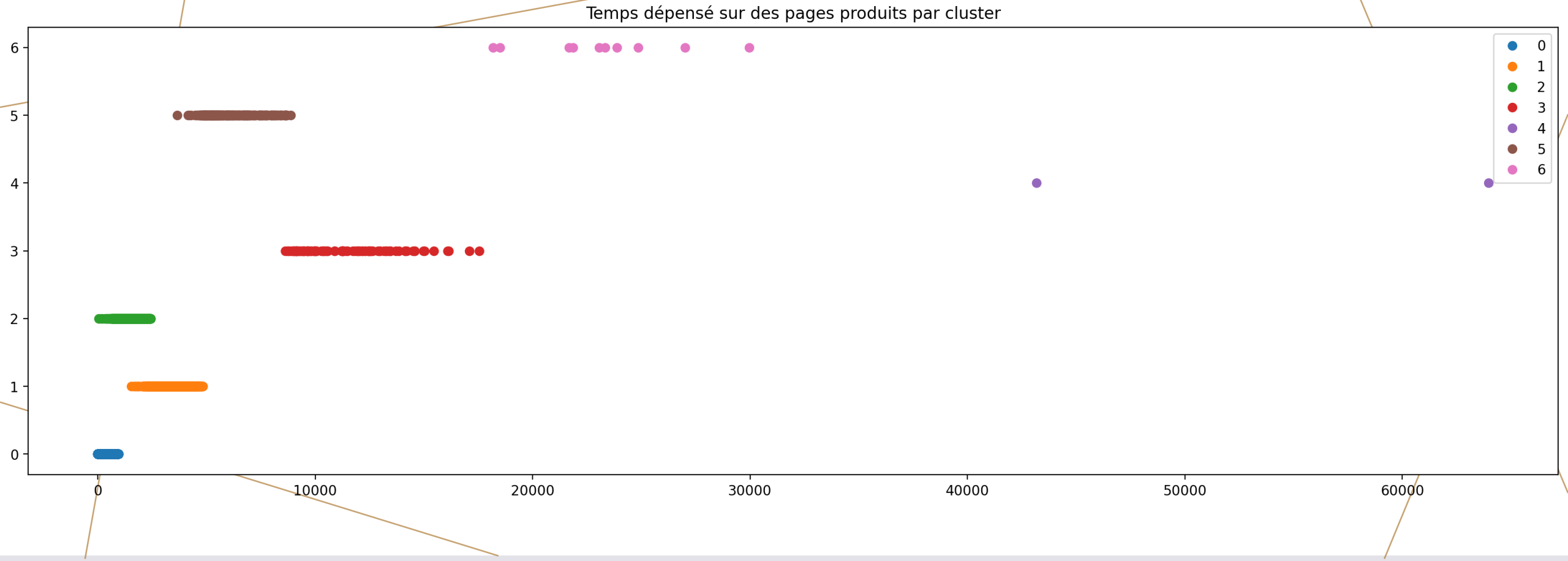


Quels sont les pourcentages d’achat par cluster ?

Pourcentage d'achat dans le cluster 0 : 10.576667552484 72
Pourcentage d'achat dans le cluster 1 : 24.644945697577274
Pourcentage d'achat dans le cluster 2 : 20.764496274700356
Pourcentage d'achat dans le cluster 3 : 34.2342342342342 36
Pourcentage d'achat dans le cluster 4 : 0.0
Pourcentage d'achat dans le cluster 5 : 34.00503778337531 5
Pourcentage d'achat dans le cluster 6 : 30.0



NOMBRE MOYEN DE PAGES VUES PAR LES INDIVIDUS DE CHAQUE CLUSTER



KMEANS

TEMPS PASSÉ SUR DES PAGES PRODUITS PAR LES INDIVIDUS DE CHAQUE CLUSTER

LA PERTINENCE DU CHOIX DU NOMBRE DE CLUSTER DÉPEND AVANT TOUT DE L'OBJECTIF FIXÉ.

PAR EXEMPLE DANS LE CAS D'UNE OPÉRATION COMMERCIALE, NOUS POUVONS NOUS DEMANDER COMBIEN DE CATÉGORIE DE « CLIENT » LE CORPS MÉTIER POURRAIT SOUHAITER.

EN L'OCCURRENCE, LES RÉSULTATS DES DEUX APPLICATIONS DU KMEANS SONT JUSTE, L'UNE SERA PRÉFÉRÉE À L'AUTRE EN FONCTION DE L'OBJECTIF FIXÉ.

PARAMÈTRES DE L'ALGORITHME RANDOM FOREST

Voici les paramètres que nous pouvons modifier :

n_estimators : nombre d’arbre dans la forêt
max_depth : profondeur maximale des arbres
min_samples_split : le nombre minimum de données requis pour créer un noeud
min_samples_leaf : le nombre minimum de données requis pour créer une feuille

LA MÉTRIQUE UTILISÉE EST L’ACCURACY ET LA MATRICE DE CONFUSION

m \ n	10	20	30	40	50	60	70	80	90	100	110	120	130
5	0.990917 93707427 83	0.997729 48426856 96	0.998702 56243918 27	0.999026 92182938 7	0.999351 28121959 13	0.999351 28121959 13	0.999351 28121959 13	0.998702 56243918 27	0.999351 28121959 13	0.999351 28121959 13	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13
10	0.996756 40609795 66	0.995783 32792734 35	0.995134 60914693 49	0.997405 12487836 52	0.997729 48426856 96	0.998053 84365877 39	0.997080 76548816 09	0.997080 76548816 09	0.996756 40609795 66	0.996432 04670775 22	0.995783 32792734 35	0.997729 48426856 96	0.996432 04670775 22
15	0.997729 48426856 96	0.999026 92182938 7	0.998378 20304897 83	0.999351 28121959 13	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999026 92182938 7
20	0.997080 76548816 09	0.998378 20304897 83	0.998702 56243918 27	0.998702 56243918 27	0.999026 92182938 7	0.999026 92182938 7	0.999026 92182938 7	0.999026 92182938 7	0.999026 92182938 7	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999351 28121959 13
25	0.998053 84365877 39	0.999026 92182938 7	0.999351 28121959 13	0.998702 56243918 27	0.999351 28121959 13	0.999351 28121959 13	0.999026 92182938 7	0.999026 92182938 7	0.999026 92182938 7	0.999351 28121959 13	0.999351 28121959 13	0.999026 92182938 7	0.999026 92182938 7
30	0.998053 84365877 39	0.997729 48426856 96	0.998702 56243918 27	0.999026 92182938 7	0.999026 92182938 7	0.999351 28121959 13	0.999351 28121959 13	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999026 92182938 7	0.999026 92182938 7	0.999351 28121959 13
35	0.996756 40609795 66	0.998378 20304897 83	0.998702 56243918 27	0.999351 28121959 13	0.999351 28121959 13	0.999026 92182938 7	0.999026 92182938 7	0.999026 92182938 7	0.999026 92182938 7	0.999026 92182938 7	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13
40	0.997729 48426856 96	0.998702 56243918 27	0.999026 92182938 7	0.999026 92182938 7	0.998702 56243918 27	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999026 92182938 7	0.999026 92182938 7	0.999351 28121959 13
45	0.997729 48426856 96	0.998378 20304897 83	0.998702 56243918 27	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999026 92182938 7	0.999351 28121959 13	0.999351 28121959 13	0.999026 92182938 7	0.999026 92182938 7

Tableau des variations de l’accuracy sur le test pour de n le nombre d’arbre et m la profondeur maximale

ON PEUT NOTER QUE LES VALEURS VARIENT PEU, EN EFFET, LE DATASET ÉTANT ASSEZ REDUIT, IL N'Y A PAS BEAUCOUP DE POSSIBILITÉ D'ACCURACY DIFFÉRENTES.

CEPENDANT L'ACCURACY AUGMENTE AVEC LES PREMIÈRES VALEURS DE N ET DE M, CE QUI N'EST PAS SURPRENANT. EN EFFET, SI ON LIMITE LA PROFONDEUR À DE TRÈS FAIBLES VALEURS, NOUS NE POUVONS PAS AVOIR UNE BONNE ACCURACY.

PARAMÈTRES DE L'ALGORITHME SUPPORT VECTOR CLASSIFIER

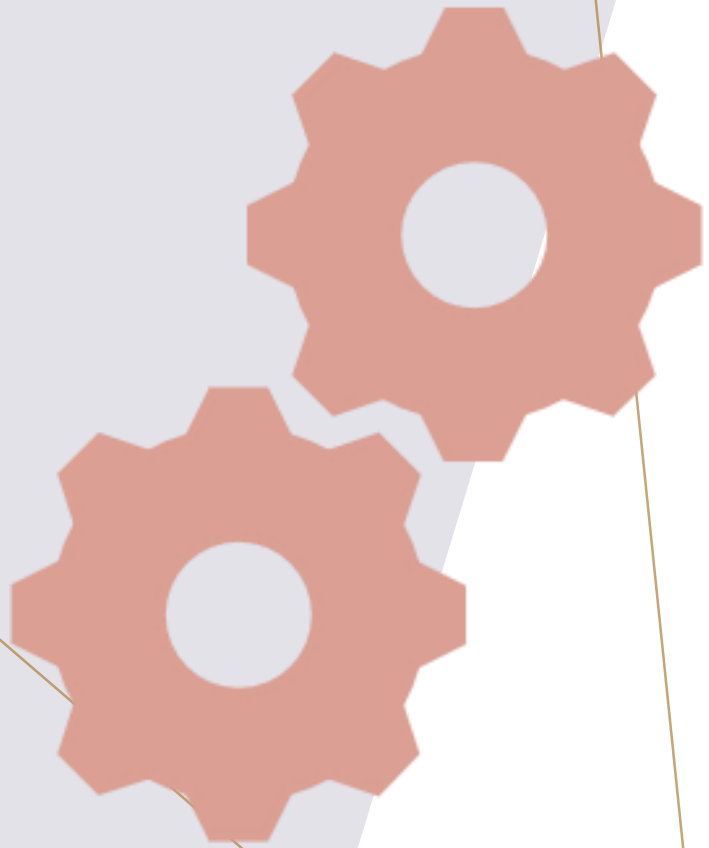
Voici les paramètres que nous pouvons modifier :

Kernel : le type de kernel
degree : le degré du polynôme si le kernel est polynomial
tol : tolérance au critère d'arrêt

NOUS UTILISONS LES MÊMES MÉTRIQUES QUE POUR LE RANDOM FOREST

NOUS AVONS TESTÉ L'ALGORITHME AVEC LES DIFFÉRENTS KERNEL SANS CONSTATER UN CHANGEMENT D'ACCURACY.

DE MÊME LORSQUE L'ON A CHANGÉ LE SEUIL DE TOLÉRANCE L'ACCURACY N'A PAS CHANGÉ.



MISE EN PLACE DE L'API

NOUS AVONS UTILISÉ PYTHON AVEC LA LIBRAIRIE FLASK.

EXÉCUTER LE FICHIER PYTHON.PY POUR LANCER LA DÉMONSTRATION.

IL Y A UN TOTAL DE 7 FICHIERS HTLM DANS UN DOSSIER TEMPLATES, AINSI QU'UN FICHIER PYTHON À LA RACINE ET UN DOSSIER STATIC OÙ LES IMAGES SERONT ENREGISTRÉES.

IL FAUT AJOUTER DANS LE DOSSIER RACINE LE CSV DE NOTRE DATASET.

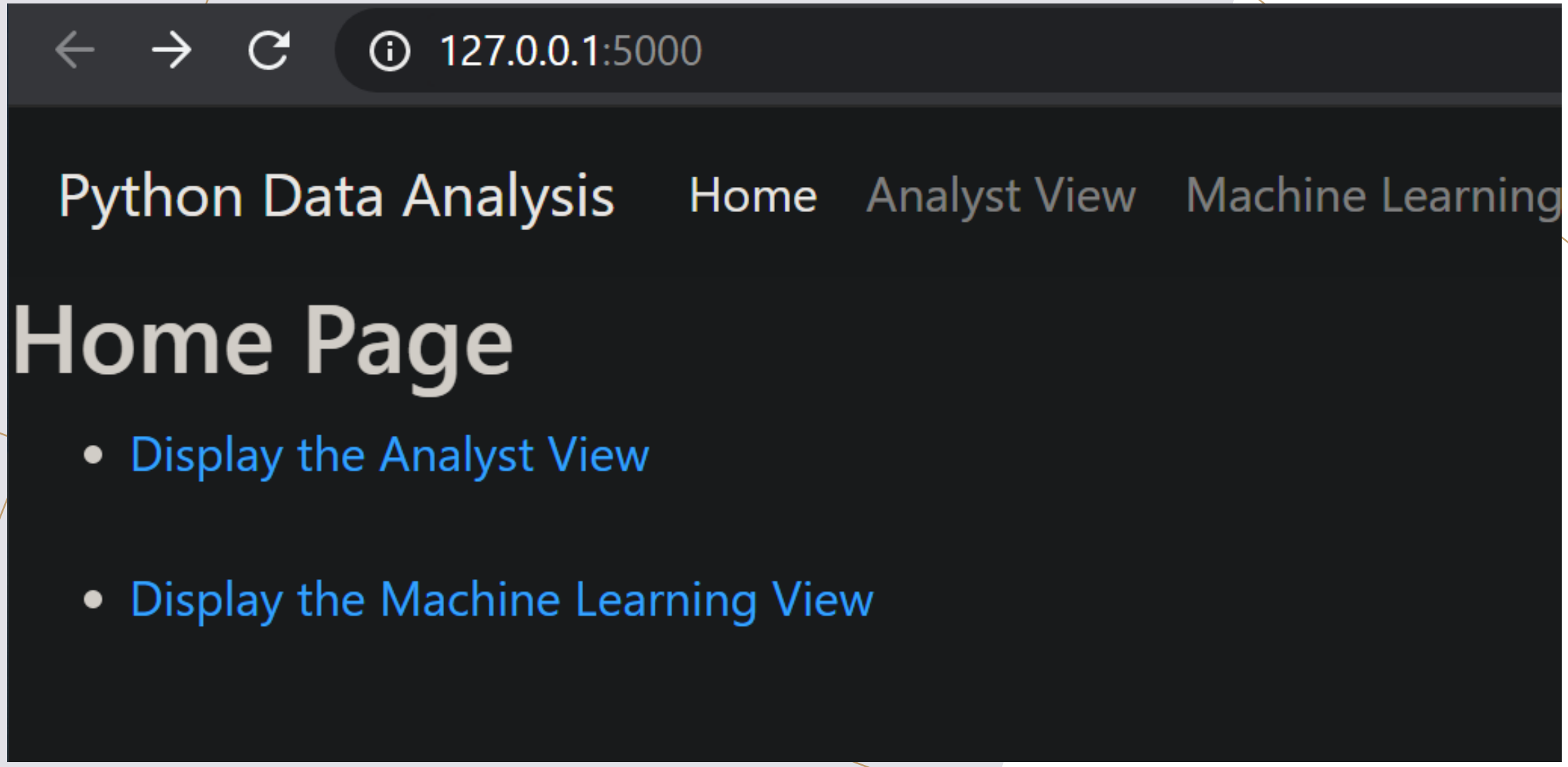
NOUS AVONS IMPLEMENTÉ LA POSSIBILITÉ D'ENTRER DES QUERIES PERSONNALISÉE CEPENDANT CELA A DES LIMITES.

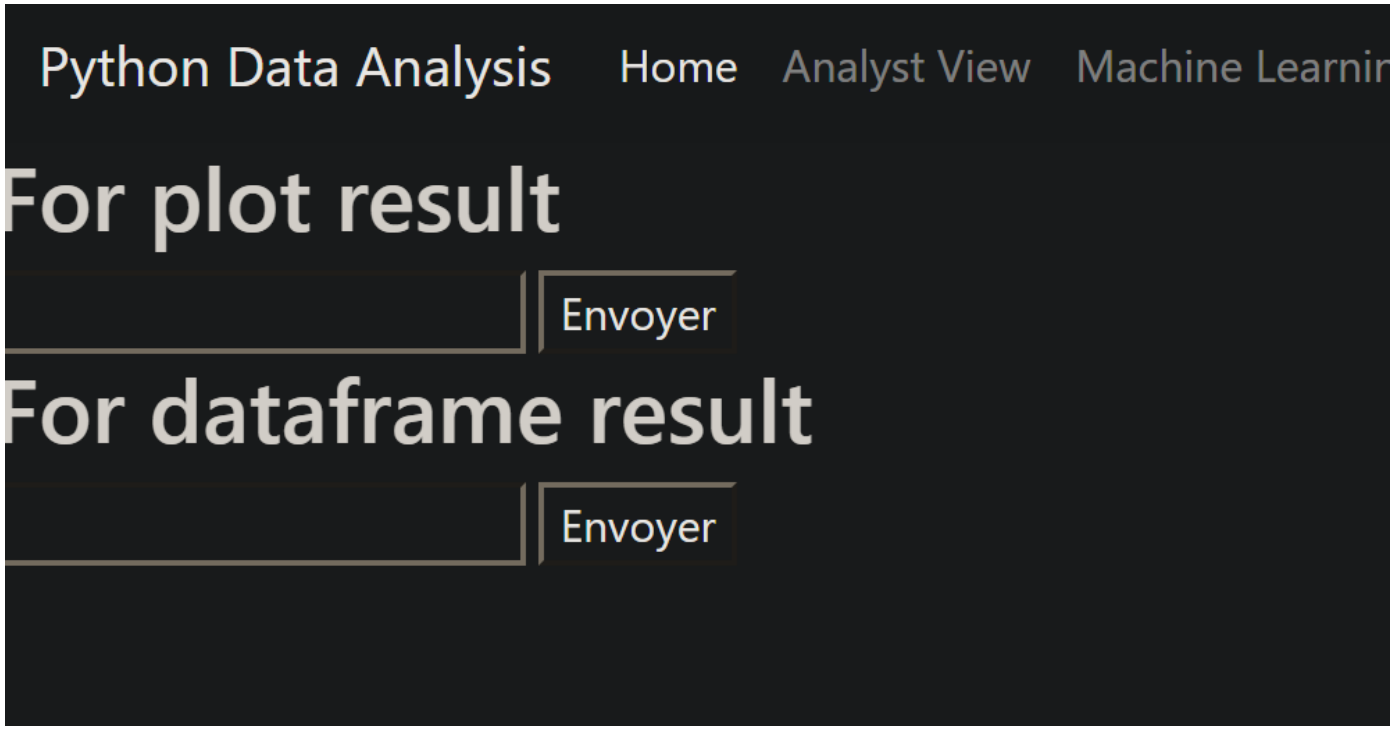
EN EFFET LES QUERYS SE DOIVENT D'ÊTRE AUTOSUFFISANTES, C'EST-À-DIRE QU'ELLES NE NÉCÉSSITENT PAS DE CALCUL EN AMONT, TEL QUE L'AJOUT D'UNE COLONNE AU PRÉALABLE. SACHANT QUE CERTAINS AJOUTS ONT D'ORES ET DÉJÀ ÉTÉ EFFECTUÉS DIRECTEMENT DANS LE CODE PYTHON.

PUIS NOUS AVONS IMPLÉMENTÉ LA NAVIGATION GRÂCE AUX STRUCTURES FLASK (CF IMAGE CI-DESSOUS

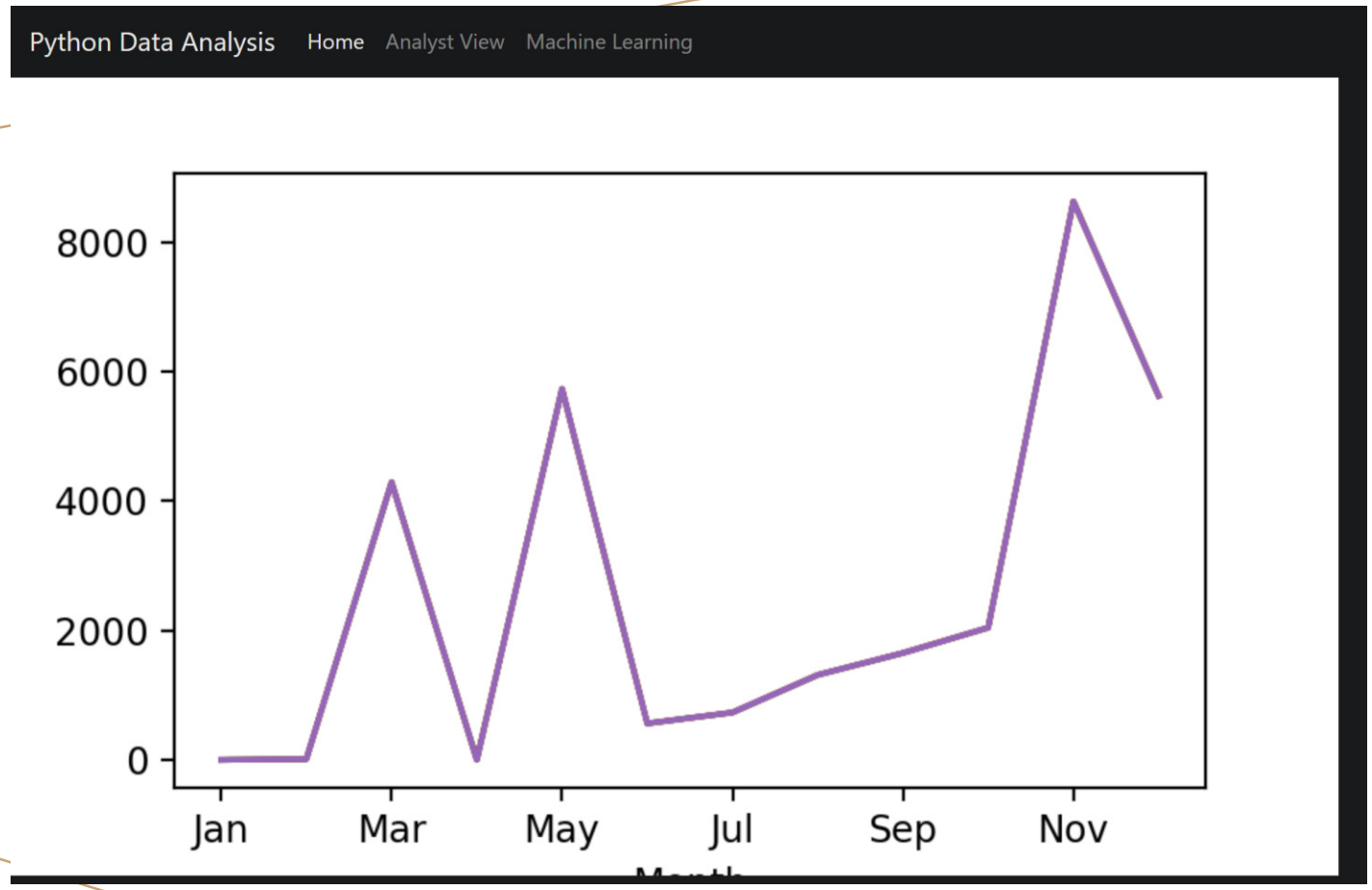
LE FICHIER HEADER.HTLM CONTIENT TOUT LE NÉCÉSSAIRE POUR AFFICHER LE BANDEAU SUPÉRIEUR DE NOTRE PAGE. ET AINSI NOUS POUVONS SIMPLEMENT UTILISER {% EXTENDS "HEADER.HTML" %} POUR LE METTRE À CHACUNE DE NOS PAGES.

```python
@app.route('/AnalystView', methods=['GET', 'POST'])
def AnalystView():
    return render_template('AnalystView.html')
```


127.0.0.1:5000
Python Data Analysis
Home
Analyst View
Machine Learning
Home Page
Display the Analyst View
Display the Machine Learning View



NOUS AVONS DIVISÉ LES CATEGORIES D'AFFICHAGE POUR UNE GESTION PLUS SIMPLE



EXEMPLE AVEC UNE DEMANDE DE PLOT AVEC :SHOPPERS[SHOPPERS.VISITORTYPE=="NEW_VISITOR"].GROUPBY(["MONTH"]).PRODUCTRELATED.SUM().PLOT()