RAPPORT D'ANALYSE DES DONNÉES MARKETING

17 MARS 2022

Maximebch
OpenClassrooms – Data Analyst
Projet 08

SOMMAIRE

Table des matières

INTRODUCTION

MON PARCOURS

Dans le cadre du parcours Data Analyst d'OpenClassrooms, il m'a été demandé de réaliser un projet libre dans mon domaine de prédilection.

Ayant étudié le commerce international et disposant d'un Master en Management et Marketing international, j'ai eu l'opportunité d'étudier tous les aspects théoriques et pratiques du marketing à l'université.

Ces connaissances m'ont préparé à mes expériences professionnelles en tant que salarié ou freelance. Je travaille principalement en tant que digital marketing manager pour des entreprises B2B.

Je travaille parfois avec des données marketing (notamment *web analytics* sur Google Analytics ou Excel) mais, pas à une grande échelle et, surtout, sans compétences en Python ou SQL, jusqu'à présent.

LES DONNÉES DANS LE MARKETING

L'intérêt de la *data analytics* dans le marketing d'une entreprise est désormais bien établi, mais son exploitation reste encore limitée (59% des entreprises l'utiliseraient aux USA selon MicroStrategy).

> ***« Le marketing sans données, c'est comme conduire les yeux fermés. » – Dan Zarella***

En effet, la profusion de données qui nous sont désormais accessibles permettent notamment d'aider à la prise de décisions marketing par :

- **L'analyse descriptive** : c'est l'utilisation de données historiques pour identifier des tendances et des comportements répétés, et ainsi repérer des potentiels problèmes ou opportunités (« Que s'est-il passé ? »).
- **L'analyse prédictive** : utilisation de données et de techniques statistiques pour déterminer la probabilité que quelque chose ait lieu ou non (« Que va-t-il se passer ? Quand cela va-t-il se passer ? »).
- **L'analyse prescriptive** : utilise les analyse descriptives et prédictives ainsi que le *testing* de données pour déterminer quelle action permettra d'obtenir les meilleurs résultats (« Pourquoi cela va-t-il se produire ? Que dois-je faire ? »).

OBJECTIFS DE LA MISSION

- **Analyse descriptive** : faire parler les données pour mieux connaître les clients et les relations entre leurs différentes caractéristiques.
- **Classification et segmentation** : classer les clients en leur attribuant un score et les regrouper pour identifier des profils types de client.

EXPLORATION DES DONNÉES

ORIGINE DES DONNÉES SOURCE

Kaggle.com, une société de Google, est une plateforme web spécialisée en *data science* qui est devenue une référence dans ce domaine. Elle est notamment connue pour l'organisation de compétitions.

Kaggle abrite également une des plus importantes « bases de données de bases de données ». J'ai ainsi exploré les *datasets* liés aux activités marketing des entreprises.

DESCRIPTION DES DONNÉES SOURCES

Les données sélectionnées sont celles d'une entreprise vendant des produits alimentaires et des produits à base d'or, aussi bien en ligne que sur le web et par catalogue et, ce, dans différents pays.

Les différentes variables sont :
- `ID` - Identifiant unique du client
- `Year_Birth` - Année de naissance du client
- `Education` - Niveau d'études du client
- `Marital_Status` - Situation familiale du client
- ` Income` - Revenu annuel du foyer du client
- `Kidhome` - Nombre d'enfants à la maison
- `Teenhome` - Nombre d'adolescents à la maison
- `Dt_Customer` - Date d'inscription du client
- `Recency` - Nombre de jours depuis le dernier achat du client
- `MntWines` - Montant dépensé en vins sur les 2 dernières années
- `MntFruits` - Montant dépensé en fruits sur les 2 dernières années
- `MntMeatProducts` - Montant dépensé en viandes sur les 2 dernières années
- `MntFishProducts` - Montant dépensé en poissons sur les 2 dernières années
- `MntSweetProducts` - Montant dépensé en sucreries sur les 2 dernières années
- `MntGoldProds` - Montant dépensé en or sur les 2 dernières années

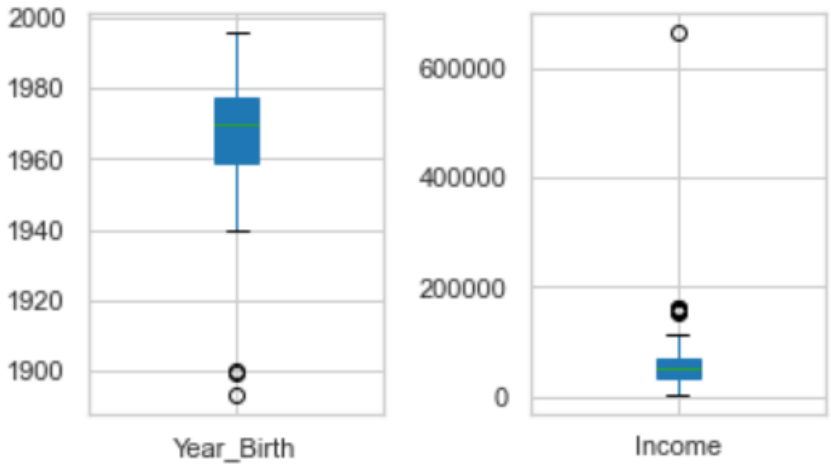
- `NumDealsPurchases` - Nombre d'achats réalisés avec une promotion
- `NumWebPurchases` - Nombre d'achats réalisés sur le site web
- `NumCatalogPurchases` - Nombre d'achats réalisés par catalogue
- `NumStorePurchases` - Nombre d'achats réalisés en magasin
- `NumWebVisitsMonth` - Nombre de visites sur le site web par mois
- `AcceptedCmp1` - 1 si le client a accepté l'offre de la 1ère campagne, 0 si non
- `AcceptedCmp2` - 1 si le client a accepté l'offre de la 2ème campagne, 0 si non
- `AcceptedCmp3` - 1 si le client a accepté l'offre de la 3ème campagne, 0 si non
- `AcceptedCmp4` - 1 si le client a accepté l'offre de la 4ème campagne, 0 si non
- `AcceptedCmp5` - 1 si le client a accepté l'offre de la 5ème campagne, 0 si non
- `Response` - 1 si le client a accepté l'offre de la dernière campagne, 0 si non
- `Complain` - 1 si le client a déposé une réclamation sur les 2 dernières années
- `Country` - Pays où est localisé le client (pas d'origine)

NETTOYAGE DES DONNÉES

```
df.isnull().sum()
```

```
ID                   0
Year_Birth           0
Education            0
Marital_Status       0
 Income             24
Kidhome              0
```

« Income » : 24 valeurs nulles à supprimer et deux espaces en trop dans le nom de la colonne



« Year_Birth », « Income » : Valeurs aberrantes à supprimer.

```
df['Marital_Status'].unique()
```

```
array(['Divorced', 'Single', 'Married', 'Together', 'Widow', 'YOLO',
       'Alone', 'Absurd'], dtype=object)
```

« Marital_Status » : certaines valeurs inattendues à changer pour garder 3 valeurs « Married », « Single » et « Together » (union libre).

CRÉATION DE NOUVELLES VARIABLES

1) **Total des dépenses** (« MntTotal ») = Dépenses en vins + fruits + viandes + poissons + produits sucrés + or

2) **Total des achats** (« TotalPurchases ») = Achats en magasin + web + catalogue

3) **Total de campagnes acceptées** (« TotalCampaignsAcc »)

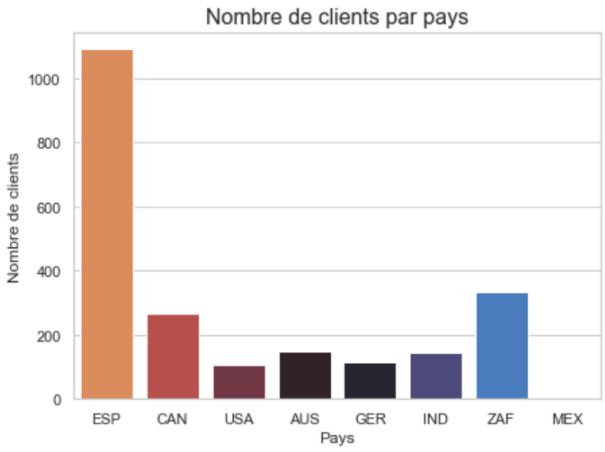
4) **Dépendants à la maison** (« DependentHome ») = enfants + adolescents

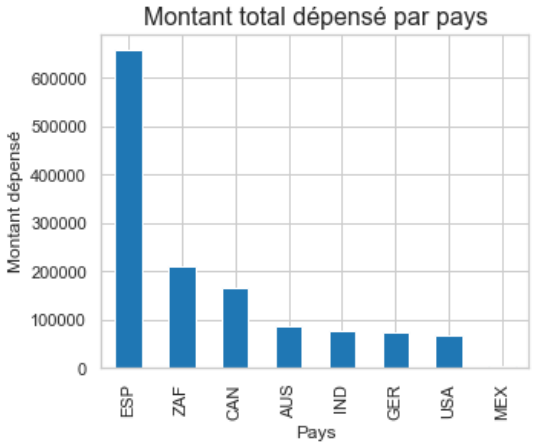
5) **Âge du client** (« Customer_age »)

6) **Âge du client au moment de l'inscription** (« Customer_age_enr »)

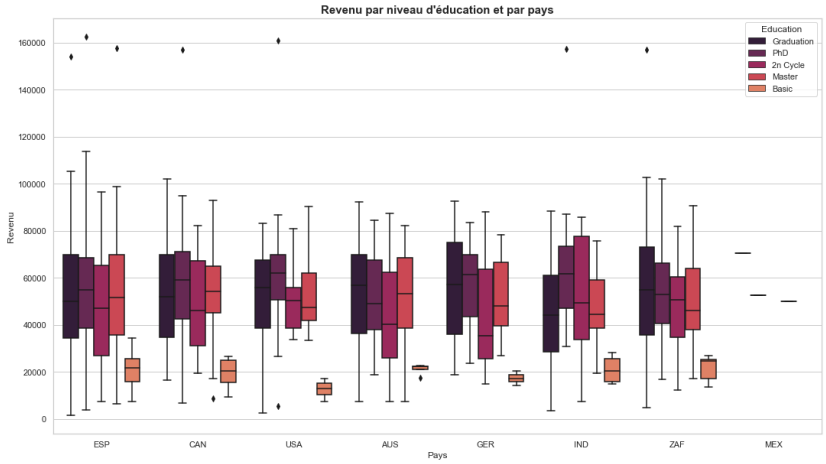
7) **Nombre de jours depuis l'inscription** (« Customer_duration_days »)

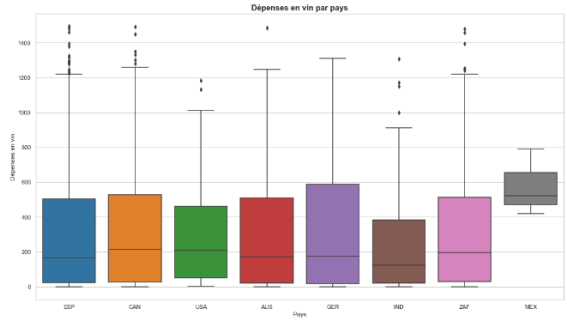
EXPLORATION DES VARIABLES



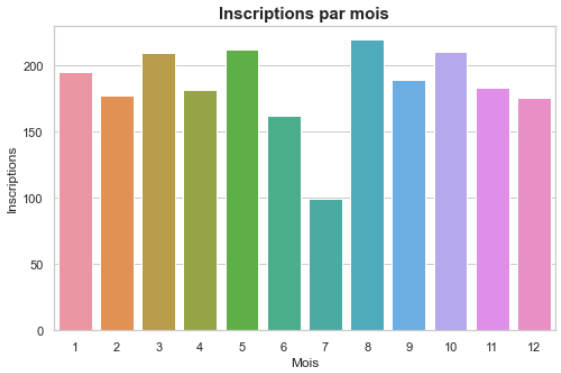


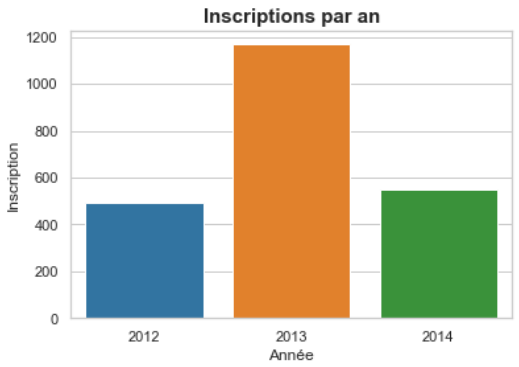
Pays : le premier marché est l'Espagne, encore plus important que tous les autres pays réunis. Le dernier marché est le Mexique avec 3 clients.





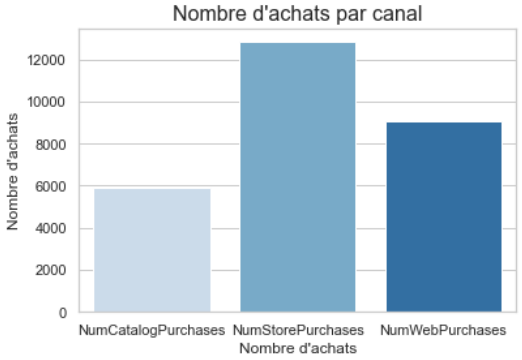
Revenus et éducation : les clients ont très majoritairement fait des études supérieures, mais ceux qui ont fait le moins d'étude gagnent et dépensent significativement moins.
Pays : L'analyse des profils et des comportements des clients ne révèle pas vraiment de différence entre pays, même pour des revenus et des dépenses en vin.

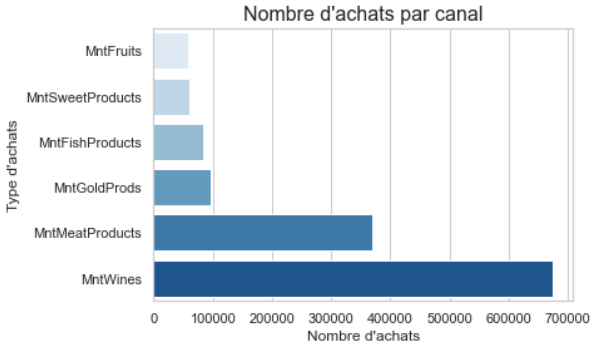




Nouvelles inscriptions : elles chutent en juillet, et la croissance s'est inversée en 2014.

Recommandation : consulter l'équipe growth hacking pour déterminer s'il y a un problème avec les données de 2014 ou avec la stratégie.





Canal d'achat : ce sont dans les magasins, puis sur le web, que le plus d'achats sont réalisés.

Types de produit : les vins et la viande représentent la grande majorité des achats. Une catégorie se distingue, les produits à base d'or, qui ne sont pas alimentaires.

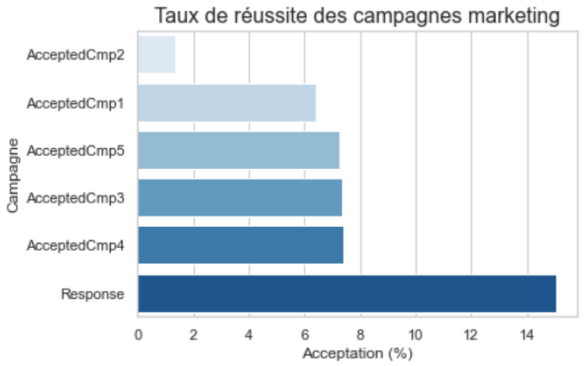
Promotions : 18,5 % des achats sont des promotions.

	Average non-complainer	Average complainer
ID	5574.8	6667.4
Year_Birth	1968.9	1968.4
Customer_age	53.1	53.6
Income	52016.6	45672.4
DependentHome	0.9	1.2
Customer_duration_days	3168.0	3244.7
Customer_Age_Enr	44.1	44.5
Recency	49.0	50.8
MntWines	306.5	176.7
MntFruits	26.4	25.1
MntMeatProducts	167.6	117.7
MntFishProducts	37.8	26.7
MntSweetProducts	27.1	18.2
MntGoldProds	44.1	27.6
MntTotal	609.4	392.0

Réclamations : Le taux de réclamation est de 0,9 % (20 sur 2211 clients). En comparant le profil de ceux ayant déposé une réclamation et celui du reste, on remarque que ces premiers ont tendance à : *avoir un revenu plus faible et plus de dépendants à la maison, réaliser presque autant d'achats mais moins dépenser que les autres (presque 2 fois moins en vins), davantage visiter le site web, acheter davantage de promotions, moins accepter les campagnes.*

Recommandation : le taux de réclamation est un KPI important à surveiller.

Recommandation : obtenir plus de données sur ces réclamations afin de déterminer quels sont les problèmes les plus courants.

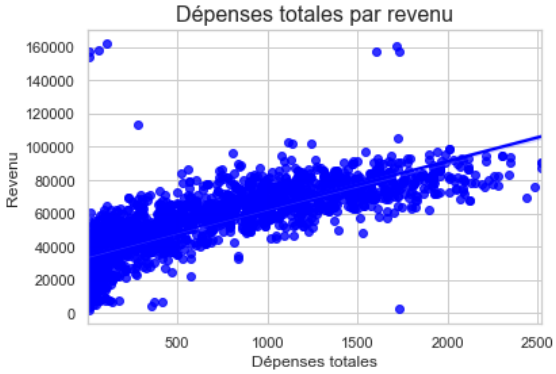


Campagnes marketing : La 2ème campagne a moins bien fonctionné mais l'offre de la dernière campagne ('Response') a été acceptée deux fois plus que les autres.

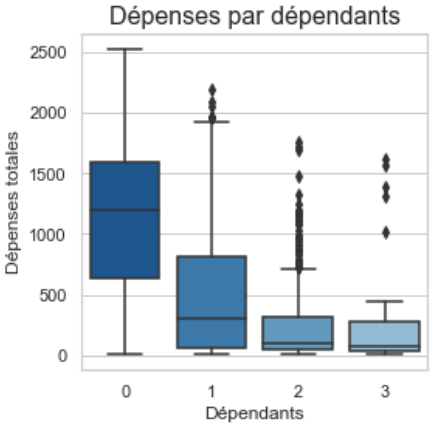
Recommandation : demander à l'équipe marketing les données permettant de déterminer quels sont les éléments qui ont fait le succès de la dernière campagne.

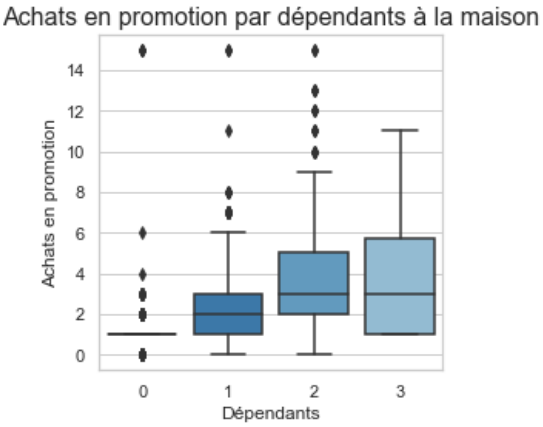
CORRÉLATIONS

Observations :



- **Revenu** : plus le revenu est important, plus les dépenses (surtout en vin et viande) et le nombre d'achats sont importants, sauf pour les produits or. Forte corrélation positive entre revenu et achats par catalogue. Forte corrélation négative entre revenu et visites du site web.

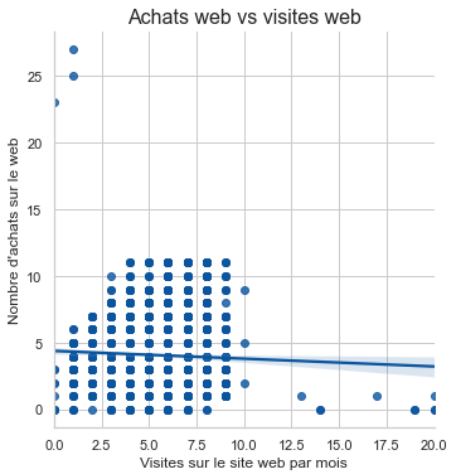


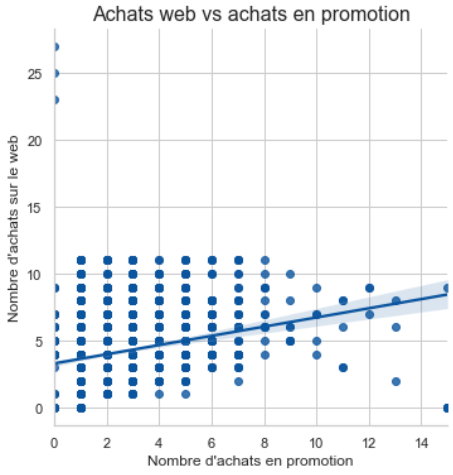


- **Dépendants** : plus le client a d'enfants à la maison, moins il dépense, réalise des achats (dans toutes les catégories), répond à des campagnes, commande par catalogue, et a un fort revenu ; mais plus il visite le site web et achète des promotions.

- **Produits** : corrélation positive entre les différentes catégories de produits.

- **Achats sur le web** : corrélation positive entre achats sur le web et revenu, vin et or.

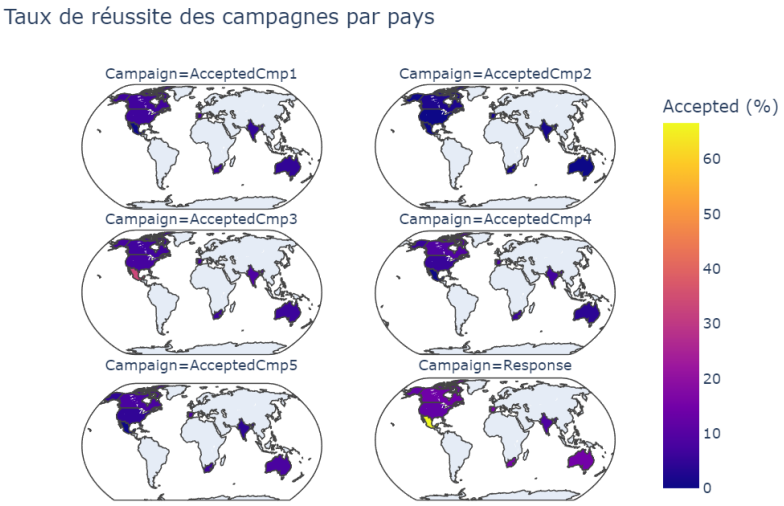




- **Visites sur le site web** : le nombre de visites sur le site web n'est pas corrélé au nombre de ventes sur le site web, mais il l'est positivement au nombre d'achats de promotion.

- **Acceptation des campagnes marketing** : le nombre de campagnes acceptées sont corrélées entre elles ; mais le nombre total de campagne acceptées est faiblement négativement corrélé au nombre de dépendants à la maison, et faiblement positivement corrélé aux revenus.

- **Acceptation campagne 5 et 1** : corrélation positive entre elles, ainsi qu'avec les achats en vin, revenus, et montant total dépensé (hypothèse : elles ont ciblé les mêmes clients et/ou partagent des éléments qui les rendent attractives).



- **Acceptation des campagnes par pays** : la dernière campagne a été la plus populaire de toutes les campagnes et de tous les pays, sauf en Inde (hypothèse : elle n'était pas adaptée à la culture ou au marché indien).

SEGMENTATION RFM DES CLIENTS

MÉTHODE DE SEGMENTATION RFM

« La segmentation RFM consiste à segmenter les clients en fonction de leur comportement d'achat. C'est une segmentation 100% comportementale. On ne segmente pas les clients en fonction de ce qu'ils sont (genre, sexe, ville, CSP…), ni en fonction de ce qu'ils aiment (centres d'intérêt, goûts), mais en fonction de ce qu'ils achètent. Le principe sous-jacent est simple : un client qui a acheté récemment, qui achète fréquemment et qui génère beaucoup de chiffre d'affaires commandera à nouveau, à coup quasiment sûr. Les chances de se tromper sont très faibles.

La segmentation RFM se construit en se focalisant sur trois critères uniquement, qui sont :

- **La Récence**. Combien de temps s'est-il écoulé depuis la dernière activité du client ? Dans la plupart des cas, moins cela fait longtemps qu'un client a interagi avec la marque ou acheté un produit, plus il y a de chances qu'il réagisse favorablement aux sollicitations marketing qu'on lui soumet.
- **La Fréquence**. Au cours d'une période donnée, combien de fois un client a-t-il acheté ou interagi avec la marque ? Les clients qui achètent le plus souvent ou qui interagissent le plus souvent avec la marque sont par définition plus engagés avec elle, donc probablement plus fidèles aussi que les autres.
- **Le Montant** (en anglais « monetary »). Combien un client a-t-il dépensé d'argent au cours d'une période donnée ? Les gros acheteurs doivent être traités différemment de ceux qui n'achètent presque jamais, c'est normal. »

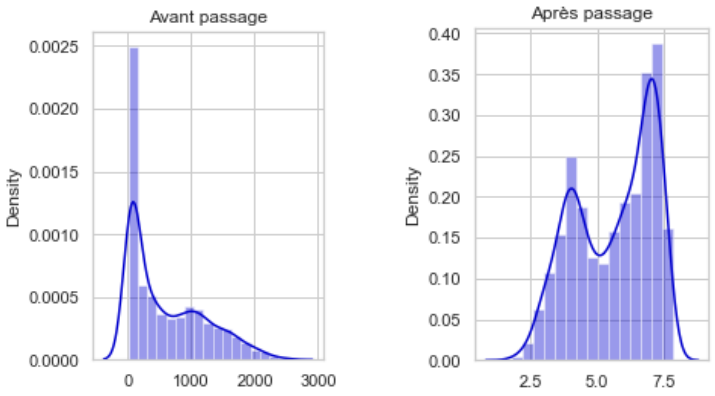
Segmentation RFM – Définition & Exemple de calcul de scores , Cartelis, 2019

ATTRIBUTION DU SCORE RFM À NOS CLIENTS

	ID	Recency	Frequency	Monetary	R_score	F_score	M_score	rfm_score
0	1826	0	14	1190	3	2	1	321
1	1	0	17	577	3	2	2	322
2	10476	0	10	251	3	2	2	322
3	1386	0	3	11	3	3	3	333

Création des 3 variables :

- **Recency** = variables « Recency » (nombre de jours depuis le dernier achat) déjà existante dans le dataset.
- **Frequency** = nombre total d'achats en magasin + sur le site web + par catalogue.
- **Monetary** = variable de dépenses totales déjà existante.

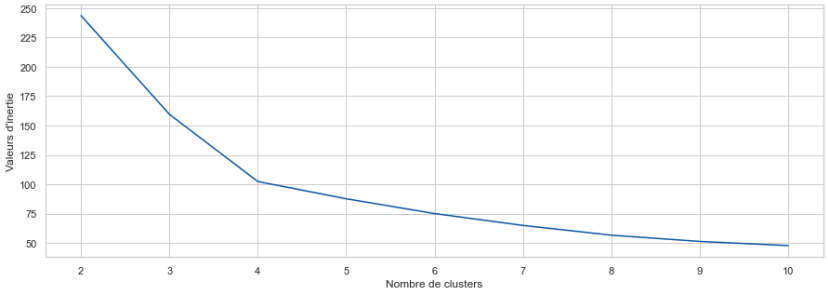


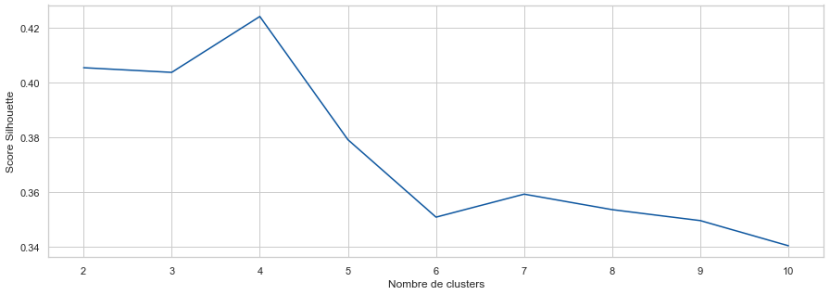
Monetary : la variable a une asymétrie vers la droite, après passage au logarithme, sa distribution semble plus normale.

ALGORITHME DE CLUSTERING K-MEANS

Le clustering est une méthode d'analyse statistique utilisée pour organiser des données brutes en silos homogènes. À l'intérieur de chaque grappe (« cluster »), les données sont regroupées selon une caractéristique commune. L' algorithme de clustering mesure la proximité entre chaque élément à partir de critères définis.

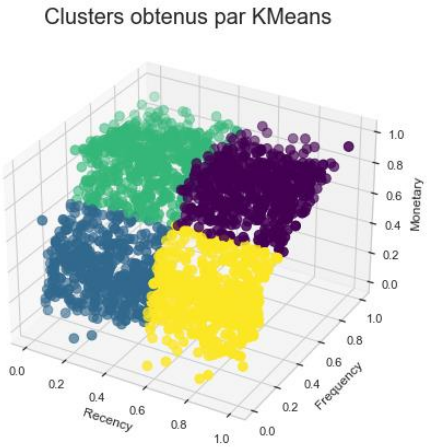
Le clustering k-means va identifier un nombre K de centroïdes (point d'équilibre pour le cluster) et répartir les valeurs dans leur cluster le plus proche.





Deux méthodes pour identifier le nombre de clusters à retenir :

- **Méthode du coude** : à partir de 4 clusters, la distorsion / inertie commence à diminuer de façon linéaire.
- **Score silhouette** : à 4 clusters, le score est à son maximum.



	Average C1	Average C2	Average C3	Average C4
Recency	0.7	0.2	0.2	0.8
Frequency	0.6	0.2	0.6	0.2
Monetary	0.8	0.4	0.8	0.4
clusters	0.0	1.0	2.0	3.0

À gauche : distribution des individus de chaque cluster.

À droite : moyenne du score dans chaque catégorie de chaque cluster.

CONCLUSIONS ET RECOMMANDATIONS

Les caractéristiques de chaque cluster permettent de dresser des portraits types de clients :

- **Cluster 4 "champions/whales"** : top clients dans toutes les catégories.

- **Cluster 2 "at risk/lapsed"** : top dépenses, top fréquence mais pas récent, risque de perdre ce client.

- **Cluster 1 "new customers"** : récent mais pas fréquent.

- **Cluster 3 "hibernation"** : score bas dans toutes les catégories.

Recommandations par cluster :

- **Cluster 4 "champions/whales"** : il est impératif de surveiller les baisses de récence, c’est notre groupe de clients le plus important.

- **Cluster 2 "at risk/lapsed"** : il faut les faire repasser à l'achat à l'aide de campagnes marketing : publicités personnalisées, promotions, demande de feedback.

- **Cluster 1 "new customers"** : fidéliser ces nouveaux clients, par exemple avec un e-mail personnalisé envoyé suite à leur récent achat.

- **Cluster 3 "hibernation"** : priorité plus faible car il est possible que rien ne permettra d'améliorer la performance de ces clients, demande davantage d'analyse (par exemple, demande de feedback).