ANALYSE DE COMPORTEMENTS D'ACHAT DES CLIENTS.

Introduction :

Pour que vos clients continuent à acheter chez vous, vous devez obligatoirement les faire gagner eux aussi, les soutenir, les apporter de la vraie valeur. Il est impossible de vraiment bien apporter de la valeur à vos clients si vous ignorez totalement ce qu'ils veulent vraiment. Et pour découvrir ce qu'ils aiment et ce qu'ils veulent, vous devez absolument chercher à les comprendre. La méthode la plus scientifique de comprendre vos clients est de parvenir à cerner du coup leurs comportements d'achats, leurs interactions entre eux et vos produits et votre entreprise, votre magasin, votre commerce etc.

Ce document un est guide pour vous aider à appendre à faire de l'analyse marketing de vos données clients dans le but de bien pouvoir prendre des décision scientifiquement muries, susceptibles d'augmenter vos ventes, le chiffre d'affaire de votre entreprise et la minimisation des coûts, et d'éviter de faire fuir vos meilleurs clients.

Connaitre les causes qui poussent les clients à s'engager ou à effectuer des achats permettra à l'entreprise de mettre en place des stratégies marketing basées sur ces raisons, contribuant alors à augmenter les ventes lors de futures campagnes.

Outil d'analyse : Python.

Projet I:

Analyse descriptive, explicative et prédictive de l'engagement marketing des clients d'une entreprise.

Données :

I. Chargement de nos données.

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
%matplotlib inline

df = pd.read_csv('C:/Users/user/Downloads/WA_Fn-UseC_Marketing-Customer-Value-Analysis.csv',encoding='latin1')
df.head()
```

	Customer	State	Customer Lifetime Value	Response	Coverage	Education	Effective To Date	EmploymentStatus	Gender	Income	...	Months Since Policy Inception	Number of Open Complaints	Number of Policies	Policy Type	Policy	Renew Offer Type	Sales Channel	Total Claim Amount	Vehicle Class	Vehicle Size
0	BU79786	Washington	2763.519279	No	Basic	Bachelor	2/24/11	Employed	F	56274	...	5	0	1	Corporate Auto	Corporate L3	Offer1	Agent	384.811147	Two-Door Car	Medsize
1	QZ44356	Arizona	6979.535903	No	Extended	Bachelor	1/31/11	Unemployed	F	0	...	42	0	8	Personal Auto	Personal L3	Offer3	Agent	1131.464935	Four-Door Car	Medsize
2	AI49188	Nevada	12887.431650	No	Premium	Bachelor	2/19/11	Employed	F	48767	...	38	0	2	Personal Auto	Personal L3	Offer1	Agent	566.472247	Two-Door Car	Medsize
3	WW63253	California	7645.861827	No	Basic	Bachelor	1/20/11	Unemployed	M	0	...	65	0	7	Corporate Auto	Corporate L2	Offer1	Call Center	529.881344	SUV	Medsize
4	HB64268	Washington	2813.692575	No	Basic	Bachelor	2/3/11	Employed	M	43836	...	44	0	1	Personal Auto	Personal L1	Offer1	Agent	138.130879	Four-Door Car	Medsize

5 rows × 24 columns

I. ANALYSE DESCRIPTIVE DE L'ENGAGEMENT DES CLIENTS ET A/B TESTING.

I.1 Le Taux d'engagement global des clients.

	Modalité d'engagement	Pourcentage de Réponse
Non Engagés	0	85.6798 %
Engagés	1	14.32 %
		14.33

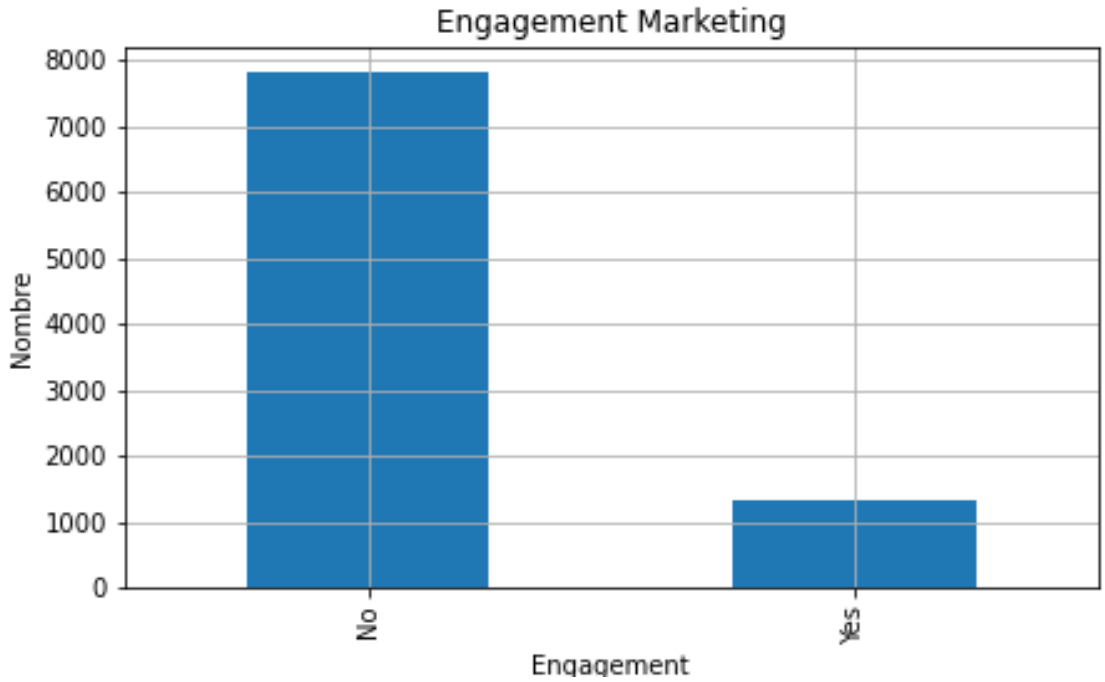
Tableau 1 : taux d'engagement global des clients.

Il y a eu globalement 14.32 % de clients engagés contre 85.68 % non engagés. Vérifions d'abord si ce 14% de clients engagés vaut la peine qu'on se penche dessus :

I.2 Nombre global des clients engagés :

	Modalité d'engagement	Nombre de clients
Non Engagés	0	7526
Engagés	1	1308

Tableau 2 : le nombre de clients engagés.



Clairement, un nombre de plus de 1308 clients engagés n'est de toute évidence pas négligeable. Nous avons donc intérêt à pousser notre analyse.

I.3 Engagement par canal de ventes.

Problème :

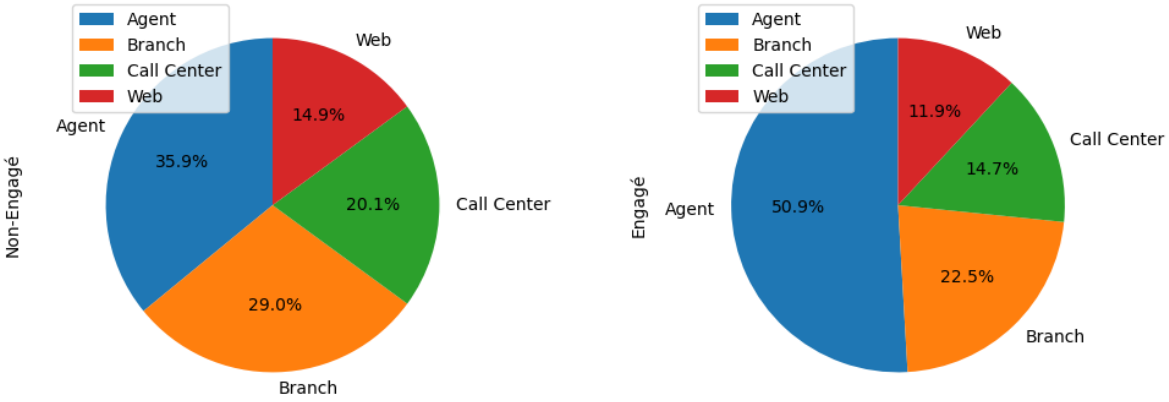
Là nous venons de connaitre le nombre et le taux d'engagement global de nos clients. Mais nous ne connaissons pas d'où ces clients engagés peuvent provenir. Connaitre le canal de provenance de ces clients nous permettra d'affiner notre ciblage et notre positionnement dans le marché. Cherchons donc le nombre et le taux de clients engagés par canaux de ventes :

I.3.1 Nombre de clients engagés par canal de vente :

	Nombre de Clients Non-Engagés	Nombre de clients Engagés
Canaux de ventes		
Agent	2811	666
Branch	2273	294
Call Center	1573	192
Web	1169	156

Tableau 3 : Nombre de clients par canal de vente.

I.3.2 Taux de clients engagés par canal de vente.



Interprétation :

La majorité des clients engagés, soit 666 clients, proviennent des agences de marketing

Ce qui, selon notre visuel, représente environ 50.9%, c-à-d la moitié des clients engagés.

Analyse de variance :

```
# ANOVA
A = df[df['Sales Channel']=='Agent']['Engaged']
B = df[df['Sales Channel']=='Branch']['Engaged']
C = df[df['Sales Channel']=='Call Center']['Engaged']
W = df[df['Sales Channel']=='Web']['Engaged']

print(stat.f_oneway(A,B,C,W))
```

```
F_onewayResult(statistic=36.234804363809566, pvalue=2.857748658532577e-23)
```

D'après l'analyse de variance des moyennes (ANOVA), la différence des groupes de canaux de ventes est statistiquement significative (P-value =0.00000… < 5%) : **la différence de distribution n'est pas le fruit du hasard : les clients préfèrent les agences markétings que les autres canaux de vente. Mais nous ne savons cependant pas pourquoi !**

En effet, seule l'analyse explicative permettra de confirmer ou d'infirmer que **les agences marketing sont une cause réelle de la sensibilité des clients à l'engagement. En attendant, effectuons une analyse d'engagement par canal de vente selon la taille de voiture des clients.**

I.3.3 Engagement des clients selon le canal de vente et la taille de véhicule du client.

Problème :

Y aurait-il une différence du choix des canaux de vente en fonction de la taille des véhicules de clients ?

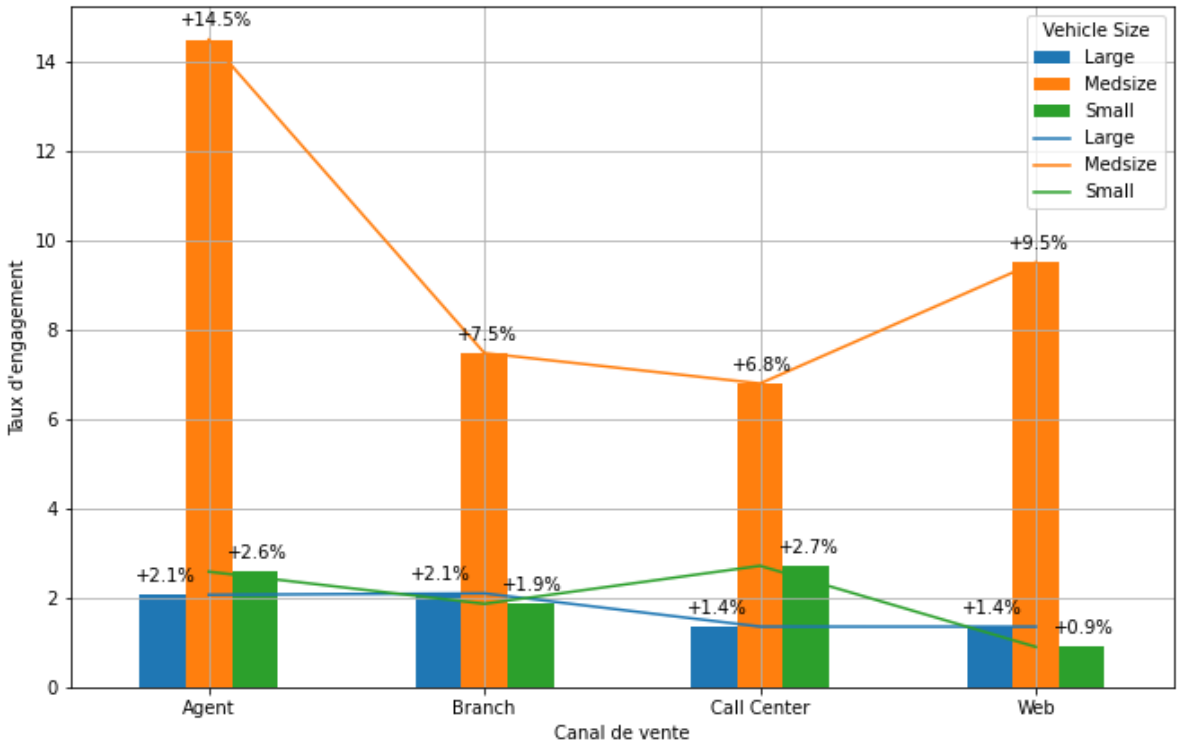
1.3.3.1 Taux d'engagement suivant les tailles de véhicules par canal de vente :

```
Sales Channel  Vehicle Size
Agent          Large           0.020708
               Medsize         0.144953
               Small           0.025884
Branch         Large           0.021036
               Medsize         0.074795
               Small           0.018699
Call Center    Large           0.013598
               Medsize         0.067989
               Small           0.027195
Web            Large           0.013585
               Medsize         0.095094
               Small           0.009057
Name: Customer, dtype: float64
```

1.3.3.2 Forme croisée de la table ci-dessus :

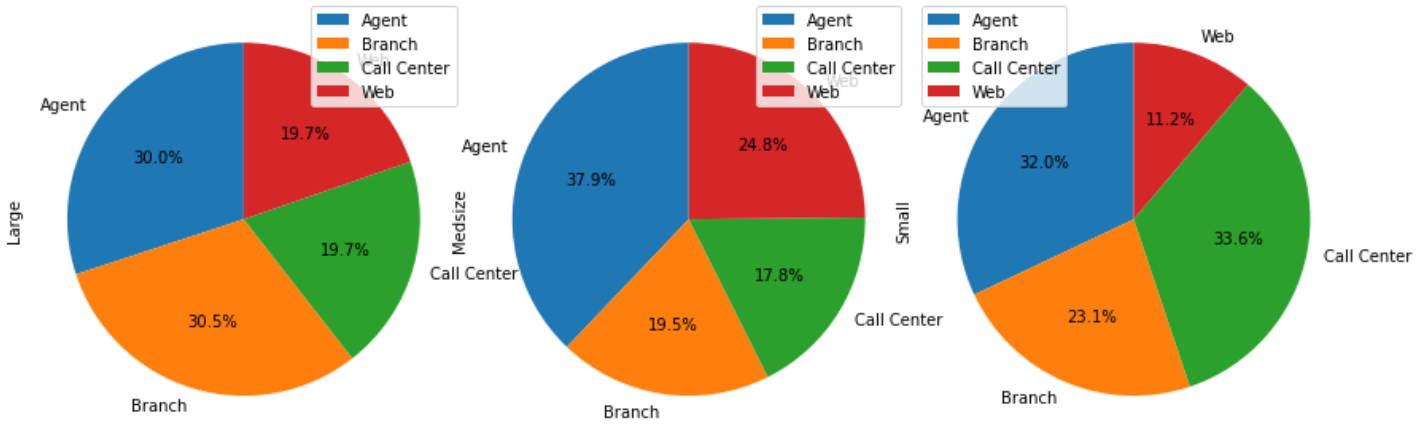
Vehicle Size	Large	Medsize	Small
Sales Channel			
Agent	0.020708	0.144953	0.025884
Branch	0.021036	0.074795	0.018699
Call Center	0.013598	0.067989	0.027195
Web	0.013585	0.095094	0.009057

On ne peut pas facilement interpréter ces deux tables. Visualisons ces résultats :



Clairement, les clients ayants des voitures de tailles moyennes ont des taux d'engagement beaucoup plus élevés que les clients ayants des voitures de grande et petite taille et cela dans tous les différents canaux de vente, surtout au travers des agences markéting.

D'ailleurs, nous avons vu que les clients s'engageaient plus via les agences que tout autre canal de vente. Et bien nous voyons ici que ces clients semblent différemment s'engager dans ces différents canaux de vente selon les différentes tailles de leurs voitures aussi**:**



Par exemple, les clients qui ont des **véhicules de petite taille** semblen**t autant** s'engager via les **agences** que via les **centres d'appel.** Tandis qu'ils semblent s'engager **plus** via les **agences** et les **centres d'appel** que via les **pages web** et les **«Branch».**

De même que les clients qui ont les **véhicules de grande taille** semblent moins engagés via les **agences** que via les **"Branch" (** 30.% de moins que 30.5%), mais plus engagés que les **centres d'appels** et **les pages web** (ces deux derniers semblent avoir le même taux d'engagement pour les clients proprio des voitures de grande taille). **Nous allons effectuer des tests d'indépendance pour confirmer ou infirmer ces différences que nous observons.**

1.3.3.3 Test d'Indépendance :

Cas 1: Engagement des clients ayant des véhicules de petite taille via les agences et les centres d'appel :

```
o1 = t[t['Sales Channel']=='Agent']['Engaged']
o2 = t[t['Sales Channel']=='Call Center']['Engaged']

import scipy.stats as stat
print(stat.ttest_ind(o1,o2, equal_var=True))
```

```
Ttest_indResult(statistic=0.12295707958445119, pvalue=0.9021651895669006)
```

Conclusion 1 :

La différence observée est statistiquement fausse :il n'y a pas de préférence spécifique entre agence et centre d'appel pour les clients ayant des véhicules de petites taille (P-Value> 5%) : le ciblage marketing aura le même résultat de réponse pour les clients qui auront à passer par ces deux types de canal de vente.

Cas 2: Engagement des clients ayant des véhicules de petite taille via les agences et les Branch :

```
As = df[(df['Sales Channel']=='Agent')&(df['Vehicle Size']=='Small')]['Engaged']
bs= df[(df['Sales Channel']=='Branch')&(df['Vehicle Size']=='Small')]['Engaged']

import scipy.stats as stat
print(stat.ttest_ind(As,bs, equal_var=True))
```

```
Ttest_indResult(statistic=2.0102757870853565, pvalue=0.044633340378917225)
```

Conclusion 2 :

Les clients de voitures de petite taille s'engagent plus via les agences et les centres d'appel que via les Branch (Taux: 32% Contre 23.1%). Rappelons-nous que Agences et Centres d'appel offrent le même taux d'engagement.

Cas 3: Engagement des clients ayant des véhicules de petite taille via les pages web et les Branch

```
ws = df[(df['Sales Channel']=='We')&(df['Vehicle Size']=='Small')]['Engaged']
bs= df[(df['Sales Channel']=='Branch')&(df['Vehicle Size']=='Small')]['Engaged']

import scipy.stats as stat
print(stat.ttest_ind(ws,bs, equal_var=True))
```

```
Ttest_indResult(statistic=-2.192131183527613, pvalue=0.02868294274324231)
```

Conclusion 3 :

Les clients de voitures de petites taille s'engagent plus via les «Branch» que via les pages web (engagement:23% contre 11.3%)

Cas 4: Engagement des clients ayant des véhicules de grand

```
a = df[(df['Sales Channel']=='Agent')&(df['Vehicle Size']=='Large')]['Engaged']
e = df[(df['Sales Channel']=='Branch')&(df['Vehicle Size']=='Large')]['Engaged']

import scipy.stats as stat
print(stat.ttest_ind(a,e, equal_var=True))
```

```
Ttest_indResult(statistic=0.8013068957920935, pvalue=0.4232561076307897)
```

Conclusion 4:

Ici, la P-Value est supérieur à 5%: la différence observée dans l'engagement via les agences et les «Branch» pour les clients ayant des voiture de grande taille est fausse: Ainsi, les clients propriétaires des véhicules de grande taille préfèrent donc autant passer par les agences de marketing que par les « Branch»

Conclusion :

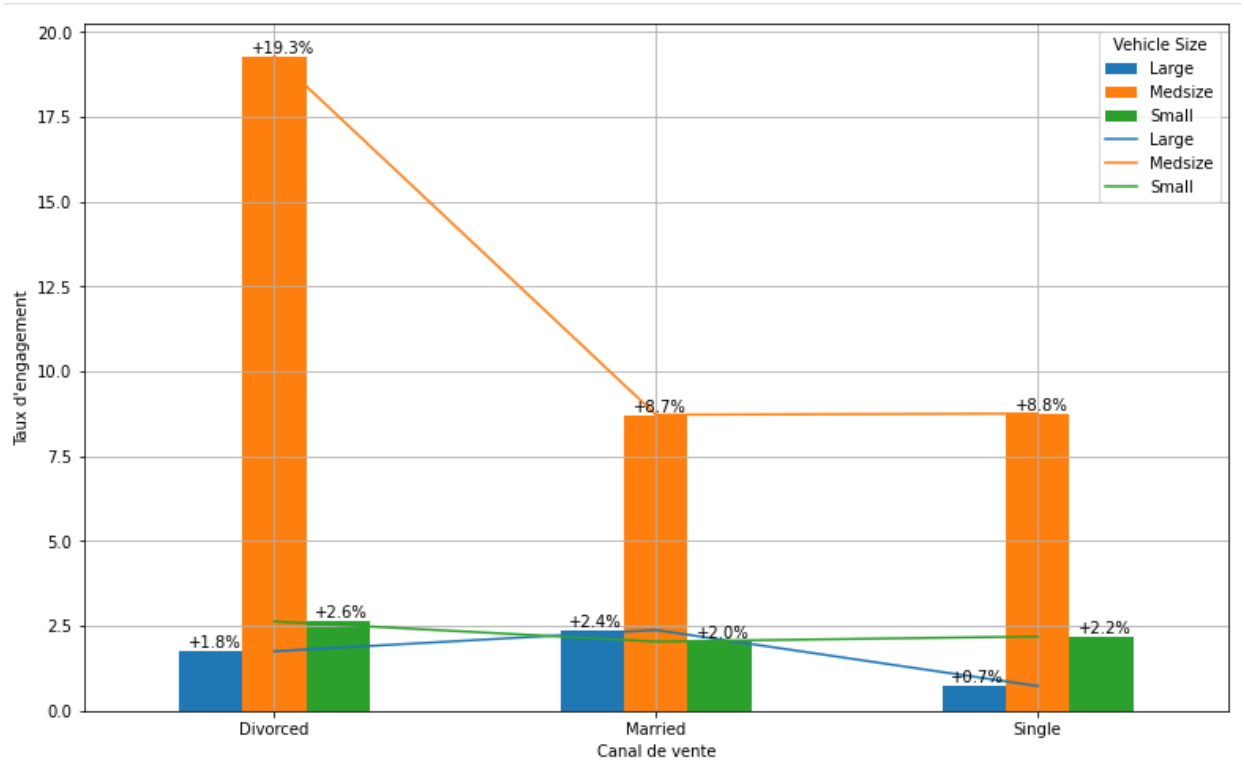
On peut bien évidemment ajouter d'autres tests mais l'objectif ici est surtout de montrer leurs importances dans l'analyse de données. Selon l'objectif marketing, vous choisirez quel type de tests vous aurez besoin pour valider ou infirmer les hypothèses déduites de vos observations au niveau de vos calcules et vos visuels.

1.3.3.4 Taux d'engagement suivant les tailles de véhicules par statut marital.

```
Vehicle Size  Sales Channel
Large         Agent            0.076110
              Branch           0.057082
              Call Center      0.025370
              Web              0.019027
Medsize       Agent            0.078456
              Branch           0.029888
              Call Center      0.018680
              Web              0.019614
Small         Agent            0.051020
              Branch           0.027211
              Call Center      0.027211
              Web              0.006803
Name: Customer, dtype: float64
```

Sales Channel	Agent	Branch	Call Center	Web
Vehicle Size				
Large	0.076110	0.057082	0.025370	0.019027
Medsize	0.078456	0.029888	0.018680	0.019614
Small	0.051020	0.027211	0.027211	0.006803

Visualisation :



Suivant le rapport entre tailles de voitures et statuts marital, les clients qui ont les voitures de tailles moyennes sont les plus engagés que les propriétaires des voitures d'autres tailles, et cela de façon beaucoup plus considérable lorsqu'ils sont divorcés que lorsqu'ils sont mariés ou célibataires.

I.4 Engagement par type d'offre.

Problème :

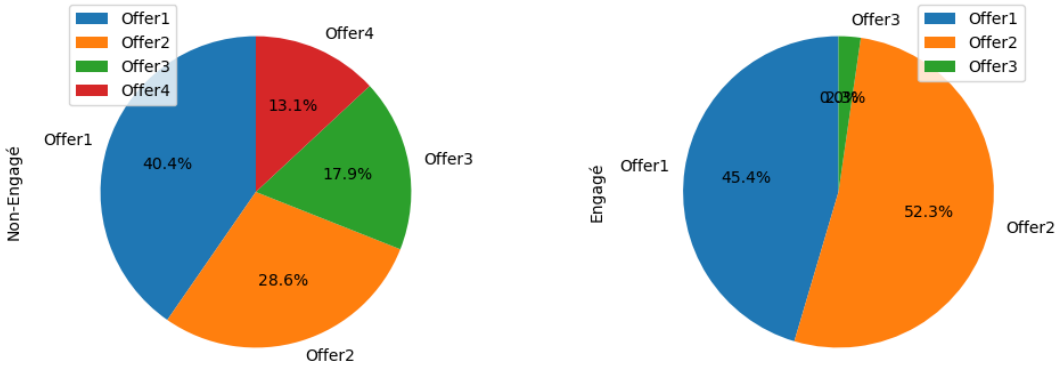
Souvent les clients achètent de produits ou services parce que telles offres semblent plus correspondre à leurs besoins plutôt que d'autres. Analyser l'engagement par type d'offre proposée permet donc de découvrir le type de produits ou d'offre le plus apprécié. Faisons cela tout de suite :

I.4.1 Nombre de clients engagés par offres :

	Nombre de Clients Non-Engagés	Nombre de Clients Engagés
Types d'offres		
Offer1	3158.0	594.0
Offer2	2242.0	684.0
Offer3	1402.0	30.0
Offer4	1024.0	0.0

Tableau 4 : nombre clients par type d'offres.

I.4.2 Taux d'engagement par type d'offres



Interprétation :

Nous remarquons qu'aucun clients engagé ne s'intéresse à l'offre numéro 4. Tandis que 52.3% et 45.4% d'entre eux éprouvent un intérêt respectif aux offres 2 et 1. Il faut aussi noter que la majorité des clients non-engagés semble désintéressés de ces deux offres pourtant très appréciées par les clients répondants à l'appel.

I.4.3. **Pouvons-nous dire qu'il y a réellement une différence de préférence à l'offre 2 qu'à l'offre 1 chez nos clients engagés?**

Problème :

En effet, on peut se demander s'il existe réellement une différence de préférence significative entre les offres 2 et 1. Car nous observons une différence de nombre de clients entre ceux qui apprécient l'offre2 et ceux qui apprécient l'offre1 : une confirmation de cette hypothèse nous aiderait à prendre une décision réaliste sur le type d'offre à proposer à nos clients. Nous allons donc effectuer un test d'indépendance statistique pour le savoir.

Résultat du test

Après un t-test d'indépendance des deux offres 1 et 2 appréciées par les clients engagés, nous obstinons le résultat suivant :

Statistiques : t-test = -7.81 et P-value = $6.54.10^{-15}$ % < 5%

Interprétation :

La P-value a une valeur très inférieur 5 % : nous rejetons l'hypothèse nulle et considérons l'hypothèse alternative : il y a bel et bien une différence de préférence statistiquement significative entre ces deux types d'offres par nos clients répondants à l'appel marketing.

Conclusion :

L'offre 2 est bel et bien appréciée que l'offre1. **La stratégie marketing consistera donc à proposer plus l'offre 2 que l'offre 1 à nos clients.**

Mais avant de proposer l'offre2, il faut s'assurer que **cet intérêt à l'offre2 constitue une cause assez suffisante d'engagement :** ce n'est pas parce que les clients s'intéressent à l'offre 2 que leur intérêt soit la raison de leur engagement . Nous étudierons ce point dans l'analyse explicative de l'engagement (Partie «II»). Ce que nous allons faire tout de suite c'est plutôt de comprendre le renouvellement du type d'offre de chaque client en fonction de la classe de sa véhicule.

I.4.4. Renouvellement de type d'offre par classe de véhicule du client

Problème :

Nous avons vu que les clients s'engagent généralement pour les offres de type 2 que pour les offres de type 1. Mais n'y a-t-il pas de différence significatives de taux de réponses en fonction des classes de véhicule de ces clients ? Effectuons quelque analyse là-dessus pour avoir nos idées claires.

1.4.4.1 Taux d'engagement par type d'offre et par classe de véhicule

```
offreParClasse
```

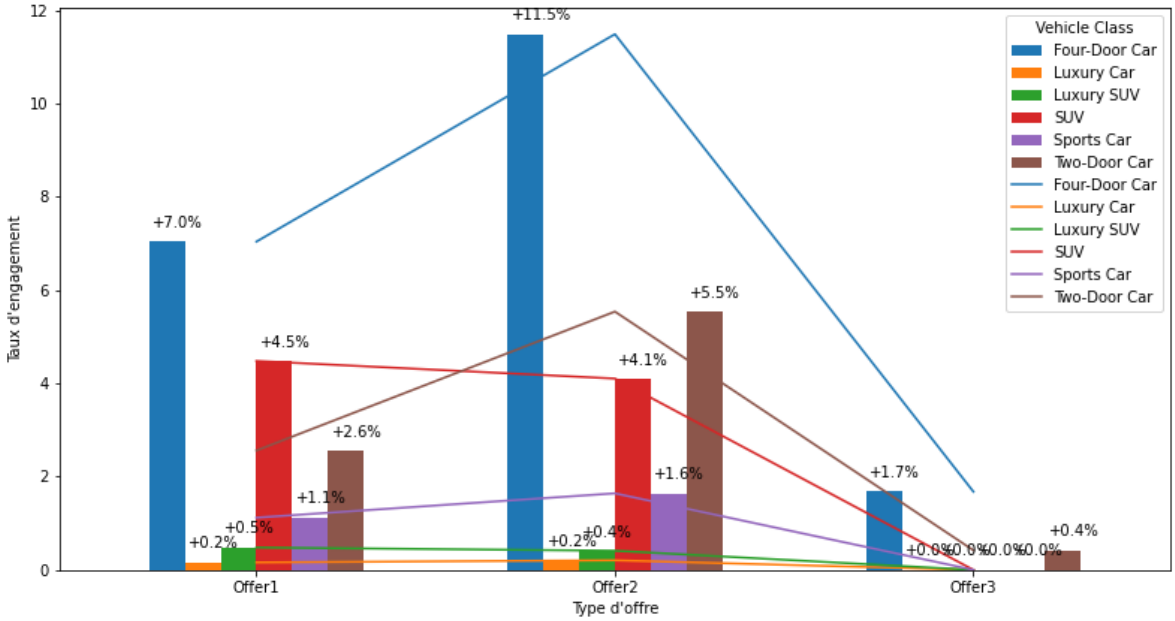
```
Renew Offer Type  Vehicle Class
Offer1            Four-Door Car    0.070362
                  Luxury Car       0.001599
                  Luxury SUV       0.004797
                  SUV              0.044776
                  Sports Car       0.011194
                  Two-Door Car     0.025586
Offer2            Four-Door Car    0.114833
                  Luxury Car       0.002051
                  Luxury SUV       0.004101
                  SUV              0.041012
                  Sports Car       0.016405
                  Two-Door Car     0.055366
Offer3            Four-Door Car    0.016760
                  Two-Door Car     0.004190
Name: Customer, dtype: float64
```

14.4.2 Tables croisée :

```
offreParClasse
```

Vehicle Class	Four-Door Car	Luxury Car	Luxury SUV	SUV	Sports Car	Two-Door Car
Renew Offer Type						
Offer1	0.070362	0.001599	0.004797	0.044776	0.011194	0.025586
Offer2	0.114833	0.002051	0.004101	0.041012	0.016405	0.055366
Offer3	0.016760	0.000000	0.000000	0.000000	0.000000	0.004190

1.4.4.3 Visualisation d'engagement par type d'offre et par classe de véhicule



Interprétation :

Nous constatons que les clients possédants les véhicules à 4 portes et à deux portes ont un taux de réponse élevé dans les 3 types d'offre, avec une grande prédominance au niveau de l'offre 2 puis à l'offre 1. **Il faut cependant noter que les clients qui ont des voiture de classe SUV semblent attiré plus pour l'offre 1 que l'offre 2**. Effectuons un test pour nous rassurer sur ce dernier cas.

1.4.4.4 Test d'indépendance

```
f1 = df[(df['Renew Offer Type']=='Offer1')&(df['Vehicle Class']=='SUV')]['Engaged']
f2 = df[(df['Renew Offer Type']=='Offer2')&(df['Vehicle Class']=='SUV')]['Engaged']

import scipy.stats as stat
print(stat.ttest_ind(f1,f2, equal_var=True))
```

```
Ttest_indResult(statistic=-2.6796956469481352, pvalue=0.0074572002999812965)
```

Conclusion :

Il y a bien une indépendance de préférence pour l'offre 1 que pour l'offre 2 pour les clients ayants les véhicules de classe SUV.

En définitive, nous pouvons donc cibler les clients ayants des voitures à 4 et à 2 portes pour l'offre 2 tandis que le ciblage va se faire pour l'offre 1 dans le cas des clients ayants les voitures de classe SUV.

I.5 Engagement selon la valeur à vie des clients et leurs durées d'assurance par l'entreprise.

Problème :

1- Considérons que les clients ayant la valeur à vie supérieure à la médiane soient ceux qui rapportent le plus d'argent à l'entreprise et que les clients dont la valeur à vie est inférieure à la médiane soient ceux qui rapportent le moins d'argent à l'entreprise.

-2 Considérons aussi que ces clients ont un âge élevé de fidélité à mesure que le nombre de mois, depuis que la police veille pour eux, est supérieur à la médiane et qu'ils ont un âge de fidélité moindre à mesure que ce nombre de mois d'assurance soit inférieur à la médiane.

Notre Hypothèses est:

1- **Serait-ce les clients qui rapportent le plus de chiffre d'affaire (la valeur à vie de clients) à l'entreprise et qui bénéficient une durée d'assurance élevée par cette entreprise qui s'engagent le plus?**

Avouons que devant de telles questions on est souvent démunie et cela est très dangereux pour votre businesse si vous ne savez pas quoi faire. Cherchons à apporter des réponses à ces questions :

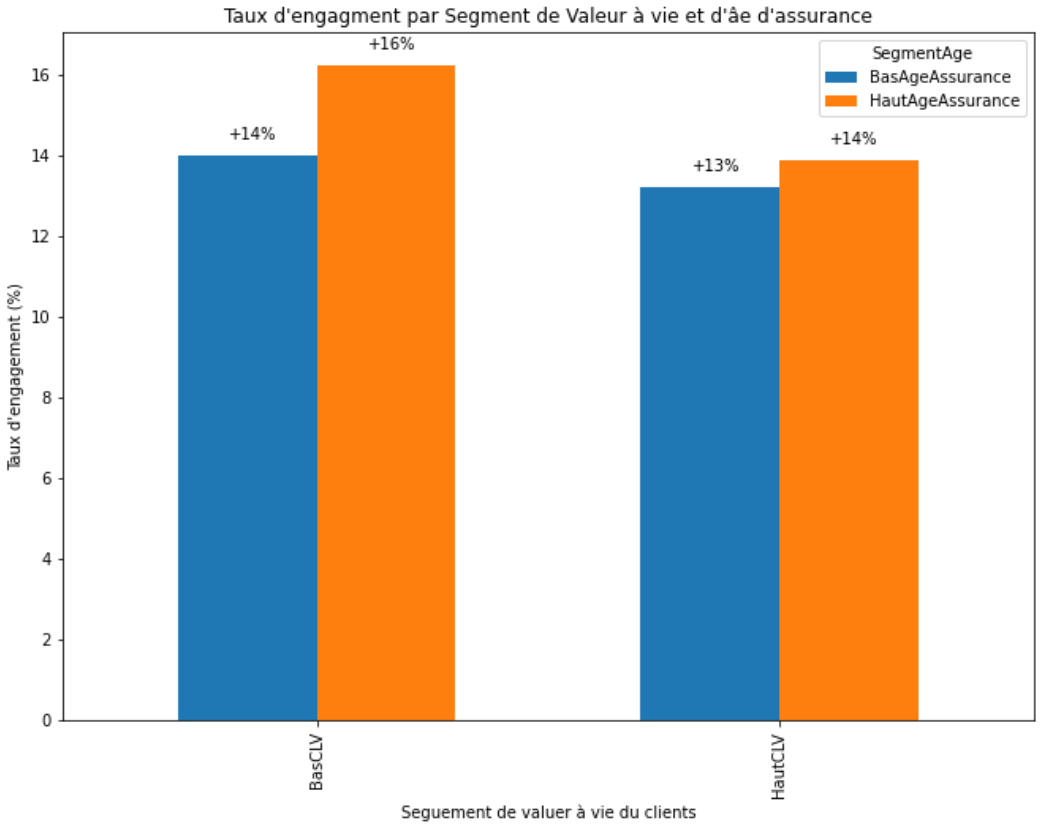
I.5.1 Taux d'engagement par valeur à vie et par âge d'assurance.

```
SegmentCLV  SegmentAge
BasCLV      BasAgeAssurance     0.139957
            HautAgeAssurance    0.162450
HautCLV     BasAgeAssurance     0.132067
            HautAgeAssurance    0.138728
Name: Customer, dtype: float64
```

I.5.2 Forme croisée de la table ci-haute :

SegmentAge	BasAgeAssurance	HautAgeAssurance
SegmentCLV		
BasCLV	0.139957	0.162450
HautCLV	0.132067	0.138728

15.2 Visualisation :



Interprétation :

1-Notre hypothèse était fausse : ce ne sont pas les clients qui rapportent le plus d'argent à l'entreprise et qui bénéficient le plus de durée d'assurance par elle qui s'engagement le plus mais bien ce sont plutôt nos clients qui rapportent le moins de valeur monétaire à l'entreprise et qui bénéficient le plus de durée d'assurance par l'entreprise qui s'engagent le plus à l'entreprise**.**

Conclusion :

Ainsi, pour avoir un taux d'engagement élevé il faut cibler les clients dont la valeur monétaire qu'ils rapportent à l'entreprise est inférieurs à la médiane et dont l'âge d'assurance qu'ils bénéficient par l'entreprise soit la plus élevé.

I.5.2 Taux d'engagement par valeur à vie et par âge d'assurance selon le statut marital

Problème :

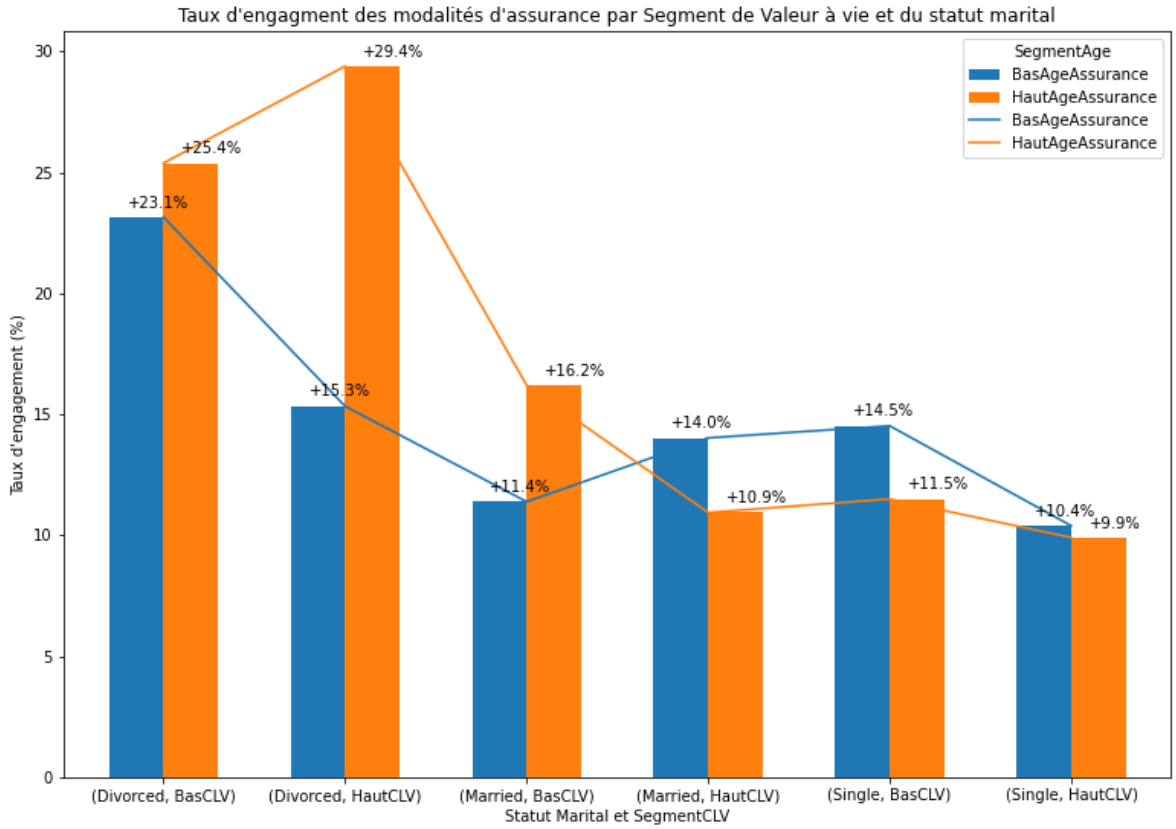
Il est évident qu'en fonction du nombre des caractéristiques (attributs) des clients qu'on considère dans l'analyse, les résultats changent. Nous venons de voir que les clients longtemps assurée et

peu valeureux aux yeux de l'entreprise sont les plus engagés : est-ce la même vérité si l'on ajoute le statut marital (les couples mariés, les divorcés ou les célibataires)? Étudions cela tout de suite :

I.5.2.1 Taux d'engagement des modalités d'assurances des clients selon leurs statuts marital et leurs valeurs à vie :

```
Marital Status  SegmentCLV  SegmentAge
Divorced        BasCLV      BasAgeAssurance     0.231454
                            HautAgeAssurance    0.253776
                HautCLV     BasAgeAssurance     0.153355
                            HautAgeAssurance    0.293814
Married         BasCLV      BasAgeAssurance     0.113895
                            HautAgeAssurance    0.162037
                HautCLV     BasAgeAssurance     0.140248
                            HautAgeAssurance    0.109422
Single          BasCLV      BasAgeAssurance     0.145234
                            HautAgeAssurance    0.115016
                HautCLV     BasAgeAssurance     0.103937
                            HautAgeAssurance    0.099083
Name: Customer, dtype: float64
```

I.5.2.2 Table croisée .



Interprétation

Cas 1 : Les clients qui bénéficient une durée d'assurance élevée et qui rapportent le plus d'argent à l'entreprise ont le taux d'engagement le plus élevé s'ils sont divorcés que lorsqu'ils sont mariés ou célibataires.

Cas 2 : De même, les clients qui bénéficient une durée d'assurance basse et qui rapportent le moins d'argent à l'entreprise ont, en deuxième position par rapport au cas 1, un taux d'engagement plus élevé s'ils sont divorcés que s'ils sont mariés ou célibataires.

Plusieurs cas sont à déduire ici ; et il faudra le faire selon l'objectif marketing. Il faut cependant ne jamais oublier les A/B Testing pour valider ou invalider les différences ou ressemblances observées entre les attributs des clients...

I.6 Engagement selon le genre.

Problème :

Il arrive qu'on ait des produits spécialement pour femmes que pour hommes, ou vis-versa, mais qu'on constate que des hommes achètent. On a alors du mal à décider si l'on doit cibler uniquement les femmes ou pas. Nous allons donc étudier l'engagement selon le genre et effectuer ensuite un test d'indépendance pour :

1- Découvrir le sexe qui s'intéresse le plus à nos produits ;
2- Déduire si la différence d'intérêt qu'on aura découverte entre hommes et femmes est bien réelle.

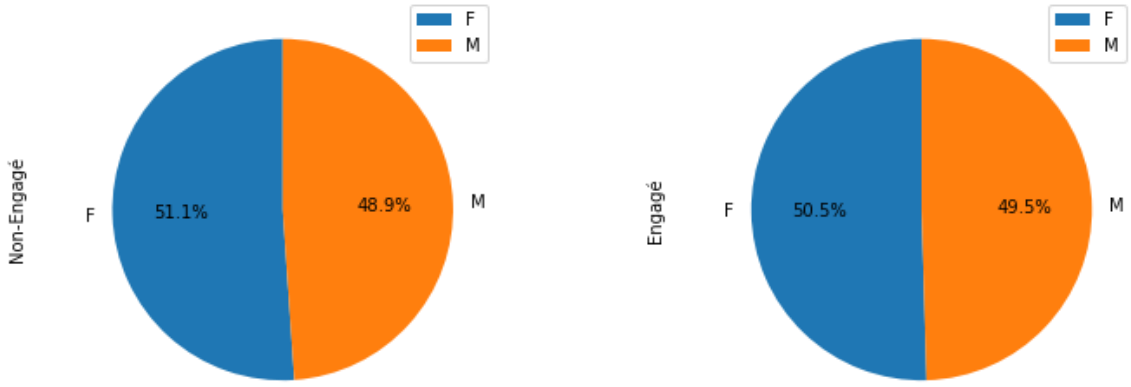
Ce qui nous permettra de décider si notre ciblage marketing doit viser uniquement la femme, l'homme, ou les deux sexes à la fois.

I.6.1 Nombre de clients engagés par genre :

	Nombre de Clients Non-Engagés	Nombre de Clients Engagée
Genre		
F	3998	660
M	3828	648

Tableau 5 : Nombres d'engagements par sexe.

I.6.2 Taux d'engagement par genre :



Interprétation :

Une petite différence d'engagements chez les femmes que chez les hommes existe : 660 femmes (soit 50.5 %) répondent à l'appel marketing contre 648 hommes (soit 49.5%) qui le font.

I.6.2.1 Test d'indépendance statistique de l'engagement par genre.

Problème :
Comme nous venons de le dire, la différence d'engagement observée peut conduire à penser que les femmes s'engagent (s'intéressent) plus aux offres que les hommes et obliger l'entreprise à viser plus les femmes que les hommes. Pour connaitre que cette différence soit vraie ou fausse, effectuons un test d'indépendance... Nous allons effectuer le test t.

Résultats du test d'indépendance:

1. **T-test :**
   t = - 0.42 et P-value = 67.4 % > 5%

Interprétation :

Nous remarquons que notre statistique t-test est faibles qui signifient que l'intensité de la différence est faible. Puis, et surtout, notre P-value vaut environ 67%, c'est-à-dire supérieurs à 5%. Ainsi, l'hypothèse nulle est à considérer contre l'hypothèse alternative:

Conclusion 1 :
La différence d'intérêt sur les offres entre femmes et hommes parmi nos clients engagés est un pur hasard: elle n'est pas statistiquement significative. Nous pouvons donc cibler autant les femmes que les hommes. Il faut toutefois s'assurer, dans une analyse explicative, si le genre constitue une raison suffisante d'engagement de ces clients.

Nous allons effectuer des analyses d'engagement des clients selon le genre et le statut d'emploi pour voir s'il y a une différence d'engagement pour les hommes ou femmes avec un job que ceux sans job , retraités ou autres etc. analysons d'abord l'engagement selon les modalité du statut d'emploi tout court d'abord.

I.6.Engagement par statut d'emploi

Problème :

Est-ce parce qu'on gagne de l'argent qu'on achète n'importe quoi, ou faut-il encore que l'offre réponde à un besoin du potentiel acheteur ? Quels clients sont susceptibles de répondre à un appel marketing selon qu'ils possèdent ou pas un emploi ?

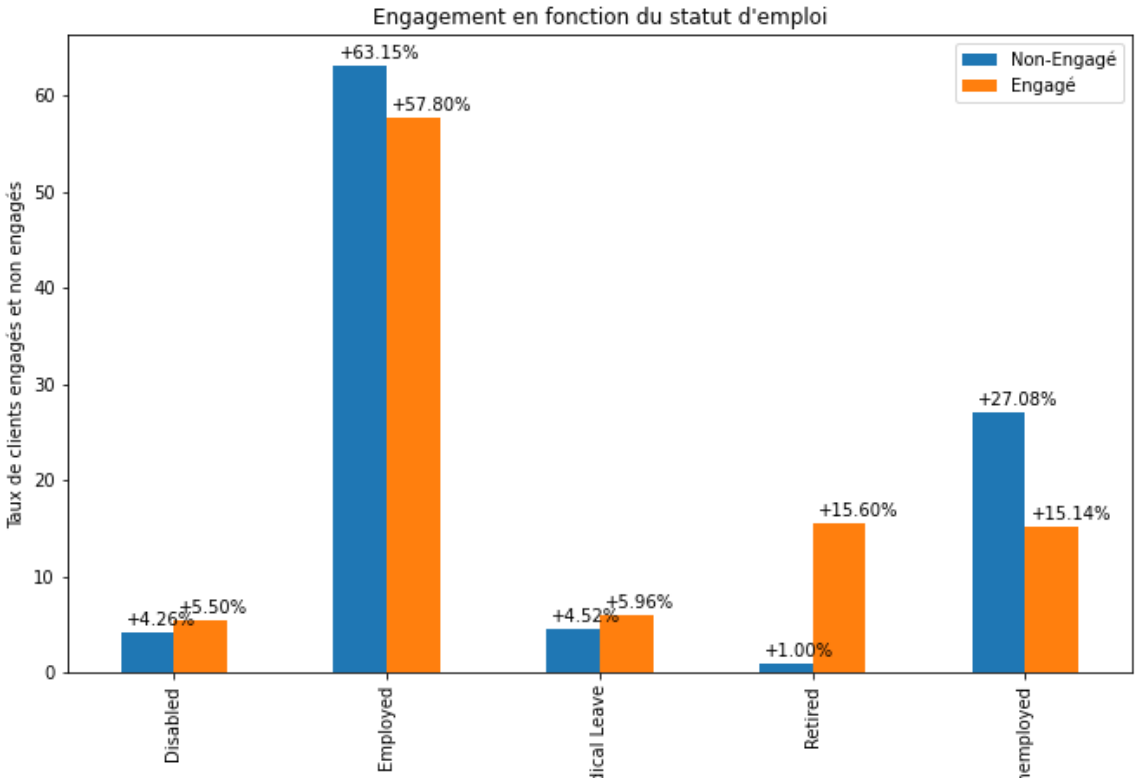
I.6.1 **Nombre d'employés engagés.**

	Nombre de Clients Non-Engagés	Nombre de Clients Engagés
EmploymentStatus		
Disabled	333	72

	Nombre de Clients Non-Engagés	Nombre de Clients Engagés
EmploymentStatus		
Employed	4942	756
Medical Leave	354	78
Retired	78	204
Unemployed	2119	198

Tableau : Nombre d'employés.

I.6.2 Taux d'engagement d'employés



Interprétation :

57.80% des clients engagés (contre 63.15% de clients non-engagés) ont un job. 15.14 % des clients engagés n'ont pas de travail. Remarquons aussi que presque tous les clients retraités (' «Retired ») se sont engagés.

I.6.3 Test d'indépendance.

57.8% d'employés et 15.14% non-employés, parmi les engagés, sont d'un écart énorme : est-ce dû au fait que le nombre des clients avec un job s'engagent plus que ceux sans job ?

La réponse serait « oui » s'il y a corrélation entre clients engagés avec un job et clients engagés sans job.

Résultats du test :

Ttest_indResult (statistic=5.930856425855, pvalue=3.1386268311641257e-09)

Interprétation:

T-test= 5.9 et P-value=3,13.e^{-9} : il y a une différence de fréquence d'engagement entre clients employés et non-employés. On peut donc bel et bien dire que les employés répondent favorablement à l'appel marketing que ceux sans emploi. Il faut noter que cette conclusion n'est pas suffisante : il faut vérifier, grâce à une analyse explicative, si le **fait d'être employé serait une raison suffisante d'engagement** de ces clients. Nous ferons cela dans l'analyse explicative de l'engagement. Effectuons toutefois une analyse sur l'engagement du statut d'emploi selon le sexe.

I.6.4 Engagement du statut d'emploi selon le sexe

Problème :

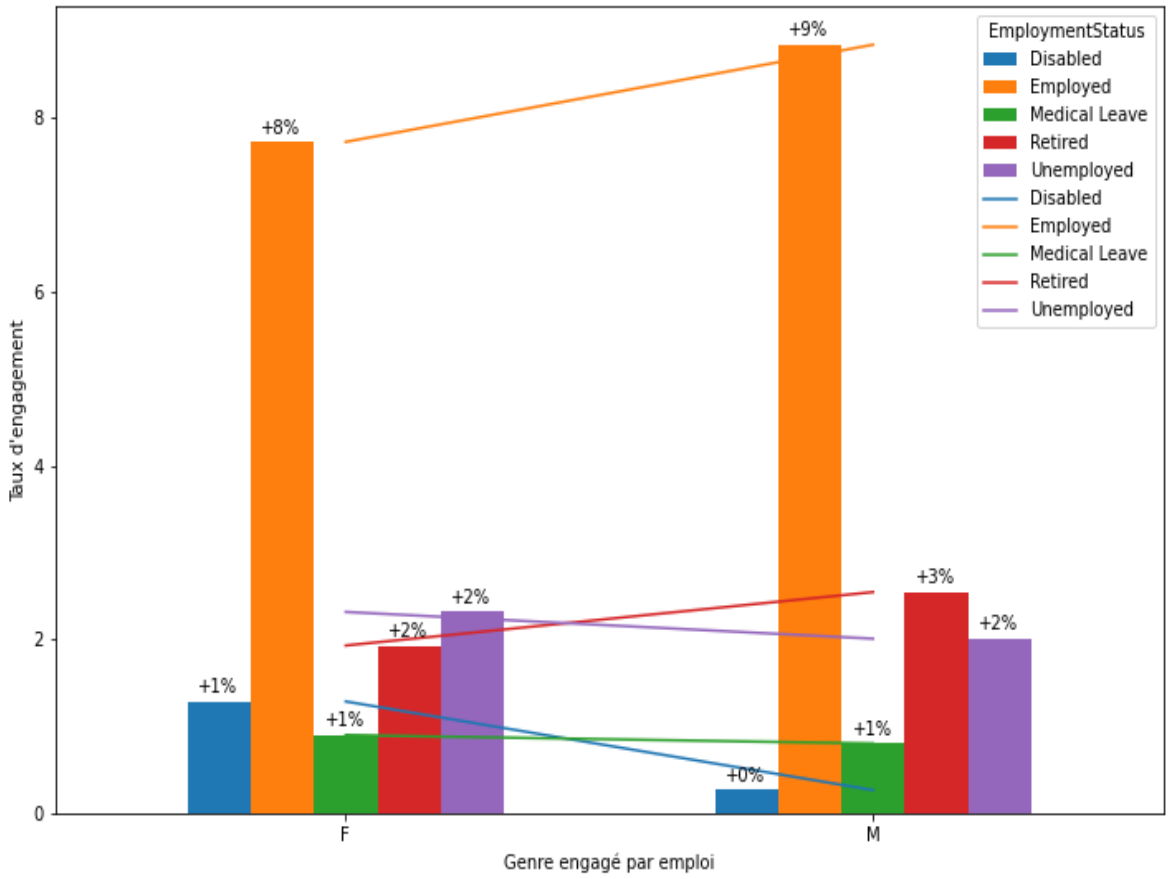
Il n'y a pas de différence d'engagement des clients selon le sexe d'après l'analyse d'engagement du genre. Mais l'est-elle si l'on ajoute à la fois le sexe et le statut d'emploi ? Voyons cela tout de suite :

I.6.4 .1 : Le genre qui s'engage le plus suivant les différents statuts d'emplois

```
Gender  EmploymentStatus
F       Disabled           0.012881
        Employed           0.077286
        Medical Leave      0.009017
        Retired            0.019322
        Unemployed         0.023186
M       Disabled           0.002681
        Employed           0.088472
        Medical Leave      0.008043
        Retired            0.025469
        Unemployed         0.020107
Name: Customer, dtype: float64
```

I.6.4.2 Forme croisée de la table ci-haute.

EmploymentStatus	Disabled	Employed	Medical Leave	Retired	Unemployed
Gender					
F	0.012881	0.077286	0.009017	0.019322	0.023186
M	0.002681	0.088472	0.008043	0.025469	0.020107



<u>Interprétation :</u>

Les clients avec un job s'engagent plus que le reste parmi toutes les différentes modalités d'emplois. Cependant, les employés de sexe masculin semblent plus s'engager que les employés de sexe féminin. De même, les clients retraités engagés sont en deuxième position par rapport aux clients employés engagés. Parmi les retraités, les hommes semblent s'engager plus que les femmes. Vérifions ces différences par des tests A/B.

I.6.4.2.1 <u>Test t d'indépendance entre hommes et femmes employés.</u>

```
tf = df[(df['Gender']=='F')&(df['EmploymentStatus']=='Employed')]['Engaged']
tm = df[(df['Gender']=='M')&(df['EmploymentStatus']=='Employed')]['Engaged']
print(stat.ttest_ind(tf,tm, equal_var=True))
```

```
Ttest_indResult(statistic=-2.3196135294597044, pvalue=0.020396975304434954)
```

<u>Conclusion</u> 1

La corrélation est réelle : Les hommes employés s'engagement plus que les femmes employées.

I.6.4.2.2 Test t d'indépendance entre hommes et femmes démissionnés

```
fr = df[(df['Gender']=='F')&(df['EmploymentStatus']=='Retired')]['Engaged']
mr = df[(df['Gender']=='M')&(df['EmploymentStatus']=='Retired')]['Engaged']
print(stat.ttest_ind(fr,mr, equal_var=True))
```

```
Ttest_indResult(statistic=-0.6922032570455581, pvalue=0.48938332014420927)
```

Conclusion 2 :

La différence d'engagement observée dans le visuel entre hommes et femmes **retraités** est fausse :
Les hommes retraités s'engagent autant que les femmes retraités

I.6.3.4.2.3 Test t d'indépendance entre femmes employées et retraitées

```
import scipy.stats as stat
tf = df[(df['Gender']=='F')&(df['EmploymentStatus']=='Employed')]['Engaged']
tm = df[(df['Gender']=='F')&(df['EmploymentStatus']=='Retired')]['Engaged']
print(stat.ttest_ind(tf,tm, equal_var=True))
```

```
Ttest_indResult(statistic=-19.225023858942038, pvalue=7.387683286482735e-78)
```

Conclusion 3 :

Les femmes employées s'engagent plus que les femmes retraitées.

I.6.4.2.4 Test t d'indépendance entre hommes employées et retraités

```
import scipy.stats as stat
tf = df[(df['Gender']=='M')&(df['EmploymentStatus']=='Employed')]['Engaged']
tm = df[(df['Gender']=='M')&(df['EmploymentStatus']=='Retired')]['Engaged']
print(stat.ttest_ind(tf,tm, equal_var=True))
```

```
Ttest_indResult(statistic=-20.25789093743959, pvalue=1.7910675263860418e-85)
```

Conclusion 4 :
Les hommes employés s'engagent plus que les hommes retraités.

I.6.5 Engagement du statut d'emploi selon le statut marital :

Problème :

Les clients avec un job et de sexe masculin s'engagent plus que les hommes ou femmes sans emploi. Les femmes avec un emplois s'engagent plus que celles sans emploi ou **retraitées** etc. **Qu'en est-il alors des employés, des sans emploi ou des retraités qui sont mariés, divorcés ou célibataires** ?

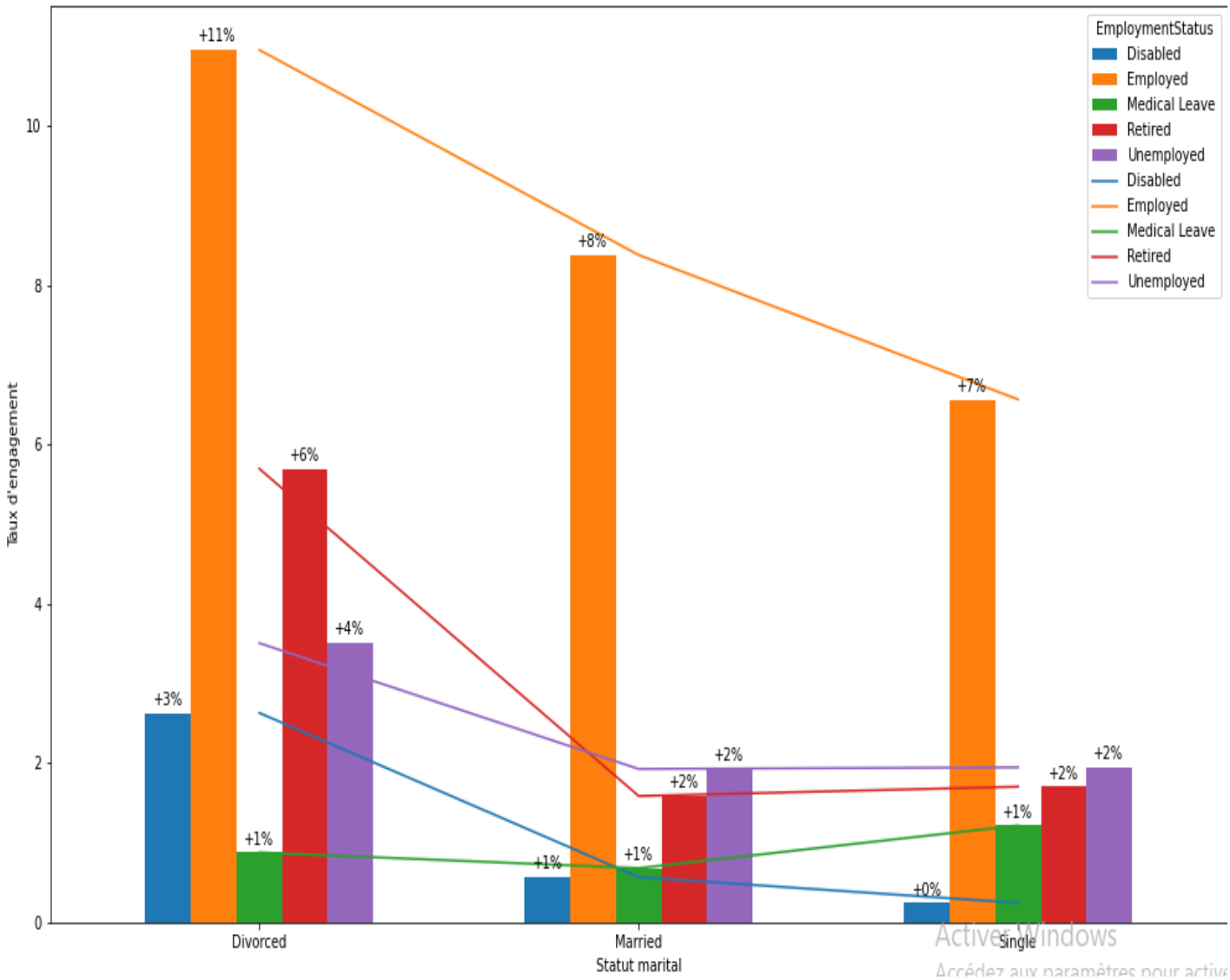
I.6.5.1 : Taux d'engagement des modalités d'emplois selon le statut marital.

```
Marital Status  EmploymentStatus
Divorced        Disabled           0.026297
                Employed           0.109569
                Medical Leave      0.008766
                Retired            0.056976
                Unemployed         0.035062
Married         Disabled           0.005663
                Employed           0.083805
                Medical Leave      0.006795
                Retired            0.015855
                Unemployed         0.019253
Single          Disabled           0.002432
                Employed           0.065667
                Medical Leave      0.012161
                Retired            0.017025
                Unemployed         0.019457
Name: Customer, dtype: float64
```

I.6.5.2 Table Croisée :

EmploymentStatus	Disabled	Employed	Medical Leave	Retired	Unemployed
Marital Status					
Divorced	0.026297	0.109569	0.008766	0.056976	0.035062
Married	0.005663	0.083805	0.006795	0.015855	0.019253
Single	0.002432	0.065667	0.012161	0.017025	0.019457

I.6.5.3 Visualisation :



Interprétation :

Les clients avec un job ont le taux d'engagent le plus élevé que tout le reste lorsqu'ils sont divorcés que lorsqu'ils sont mariés (les employés mariés se placent en deuxième position) ou célibataires. Les clients retraités et divorcés sont aussi un peu plus engagés que les clients retraités et mariés ou retraité et célibataires. Notons que les clients retraités et ceux sans emploi ont le même taux d'engagement lorsqu'ils sont divorcés ou célibataires etc.

I.8 Engagement des clients par lieu de résidence et Positionnement du marché.

Problème :

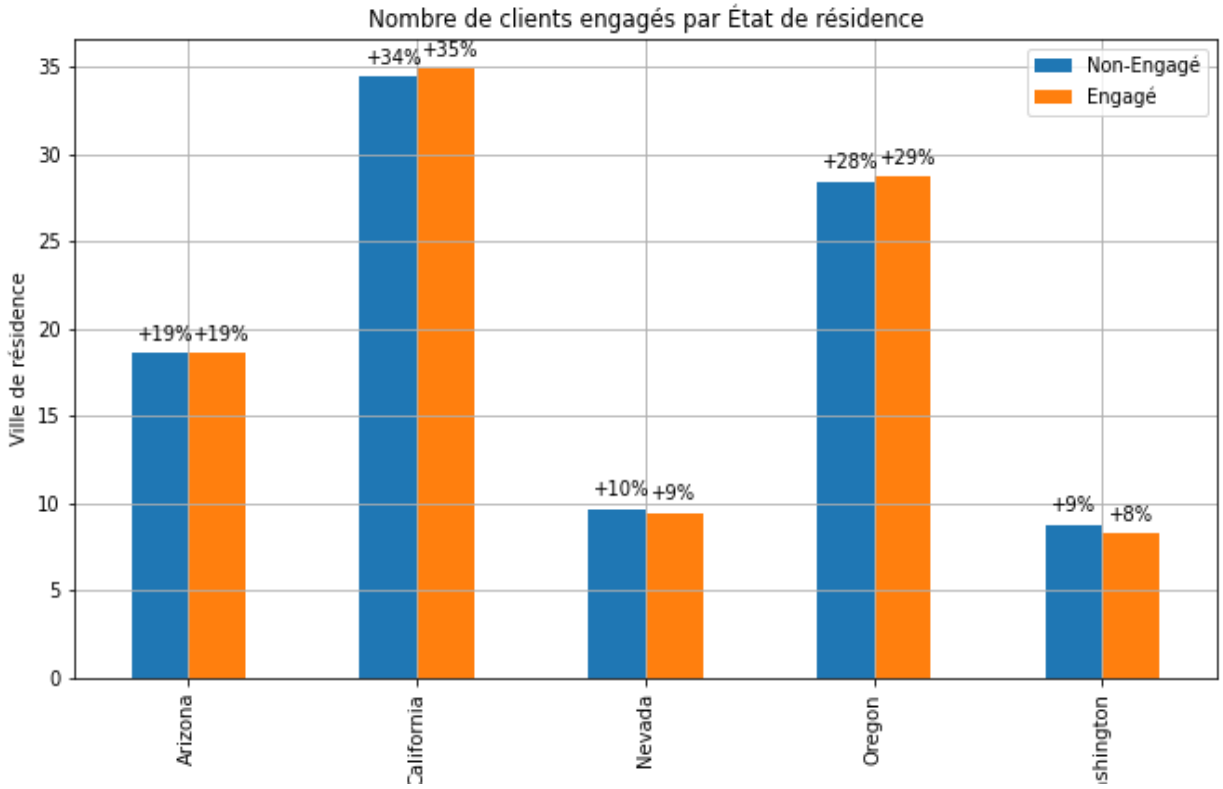
D'où peuvent provenir le plus de nos clients qui ont répondu favorablement à nos offres ? Telle est la question pour développer une stratégie marketing de positionnement du marché. Ici, l'étude se concentre dans un seul pays : Dans quelle ville des usa se trouve le plus de clients engagés à l'appel marketing?

I.8.1 Le nombre de clients engagés par villes des Etats-Unis.

	Nombre de Clients Non-Engagés	Nombre de Clients Engagés
State		
Arizona	1460	243
California	2694	456
Nevada	758	124
Oregon	2225	376
Washington	689	10

Tableau 6 : Nombre de clients par État

I.8.2 Taux d'engagement des clients par État.



Interprétation :

La majorité des clients répondants à l'appel markéting provient de la Californie et d'Oregon. Mais c'est aussi de ces deux États que proviennent le plus de clients non-engagés : Plus de 2000 clients (environ 28 à 34 %) de ces deux États ne se sont pas engagés, contre 400 clients engagés (29 à 35% de clients engagés). Clairement, il y a un marché potentiel à exploiter dans ces deux États

I.9 Engagement par niveau intellectuel

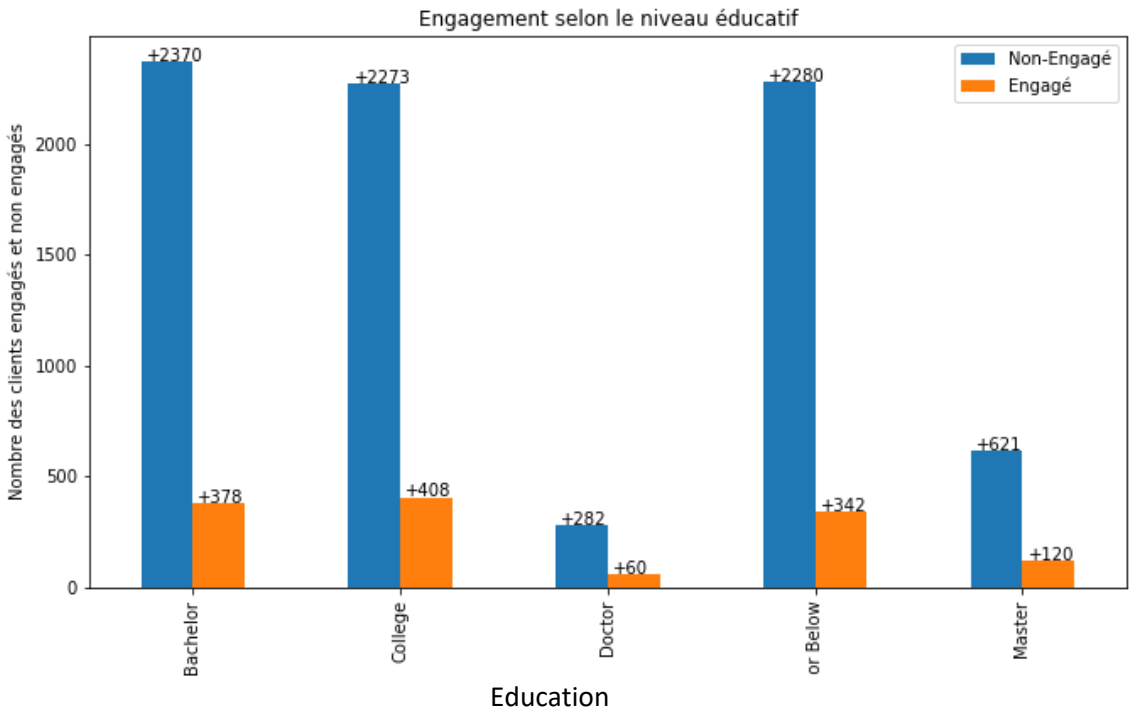
Problème :

Une manière de parvenir à faire comprendre le contenu du massage marketing par le marché cible consiste à identifier le niveau intellectuel des clients susceptibles à l'engagement. On a du mal à répondre favorable sur quelque chose que l'on ne comprend pas ou qui nous parait banale. Cette analyse permet de mesurer le niveau d'obstacle cognitif ou psychologique à l'engagement : **quelle catégorie de clients s'engageait sans trop réfléchir ou après avoir bien réfléchis… ?**

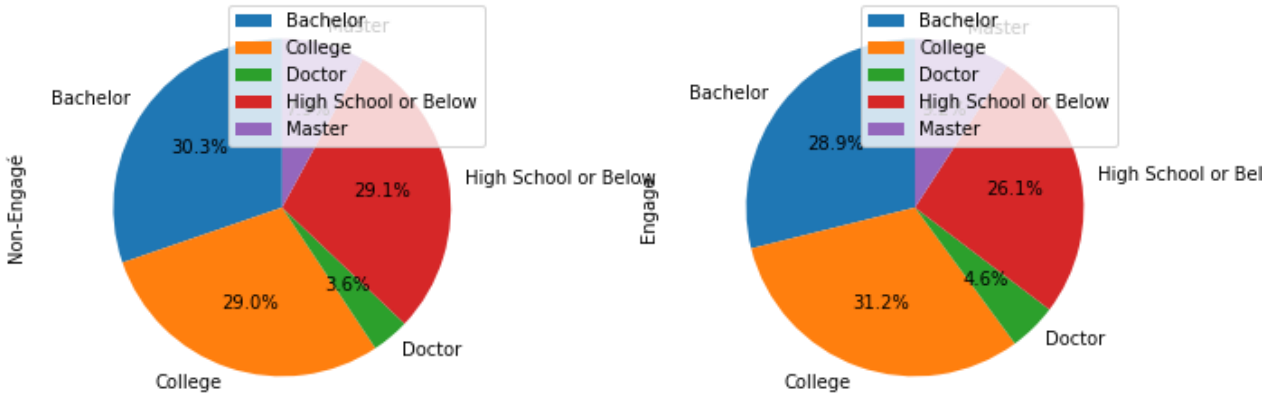
I.9.1 Nombre de clients engagés par niveau intellectuel (éducatif) :

Education	Non-Engagé	Engagé
Bachelor	2370	378
College	2273	408
Doctor	282	60
High School or Below	2280	342
Master	621	120

Tableau 8 : Nombre de clients engagés par éducation



I.9.2 **Taux d'engagement selon l'éducation.**



Interprétation :
Trois catégories intellectuelles des clients engagés semblent bien se démarquer : ceux du niveau « **College** » en premier, suivit par ceux du niveau « **Bachelor** » en 2ème position, et enfin ceux du niveau « **High School or Below** » en dernier position.

I.9.3 Test de la variance des moyennes de ces catégories (ANOVA) :

Problème :

Peut-on conclure que les clients s'engagent-ils plus à mesure qu'ils soient moins éduqués ? On ne peut pas vraiment donner une réponse réaliste avant d'avoir effectué un test d'indépendance et une analyse explicative. Le test d'indépendance nous permettra de confirmer ou infirmer la différence observée entre clients de niveau College par rapport aux autres niveaux éducatifs, et l'analyse explicative nous permettra de mesurer le degré et le sens de la corrélation entre éducation et engagement des clients.

Résultats du test :

```
# Anova
M = df[df['Education']=='Master']['Engaged']
H = df[df['Education']=='High School or Below']['Engaged']
D = df[df['Education']=='Doctor']['Engaged']
C = df[df['Education']=='College']['Engaged']
B = df[df['Education']=='Bachelor']['Engaged']

print(stat.f_oneway(C,B,D,M,B))
```

```
F_onewayResult(statistic=1.9598301610516418, pvalue=0.09774079604755687)
```

Interprétation :

Pris globalement, les niveaux éducatifs testés donnent les résultats suivants :
F = 1.96 et P-value = 0.097 < 5% : rejet de l'hypothèse nulle et prise en compte de l'hypothèse alternative : la différence des proportions observées dans les 5 groupes (les 5 modalités éducatives) est statistiquement significative : les clients avec un niveau éducatif élevé se sont moins engagés que ceux avec un niveau éducatif « bas ». L'analyse explicative dans la partie-II va nous dire si le fait d'avoir un niveau scolaire moins élevé constitue une raison suffisante à l'engagement.

Cependant nous remarquons que les clients de niveau « Bachelor » s'emblent autant s'engager avec les clients de niveau « College » et ceux du niveau « High School or Below » tandis que les clients de niveau Doctor s'emblent autant s'engager avec les clients de niveau « Master» . Effectuons des tests de ces attributs pris deux-à-deux pour avoir les idées claires.

1- Test-t d'indépendance entre Bachelor et College :

```
import scipy.stats as stat
C = df[df['Education']=='College']['Engaged']
B = df[df['Education']=='Bachelor']['Engaged']
print(stat.ttest_ind(C,B, equal_var=True))
```

```
Ttest_indResult(statistic=1.5313981002245889, pvalue=0.12572931930564488)
```

Conclusion 1 :

Les clients de niveau College et Bachelor répondent autant à l'engagement.

2- Test-t d'indépendance entre Bachelor et « High School or Below » :

```
import scipy.stats as stat
H = df[df['Education']=='High School or Below']['Engaged']
B = df[df['Education']=='Bachelor']['Engaged']
print(stat.ttest_ind(H,B, equal_var=True))
```

```
Ttest_indResult(statistic=-0.7652967638379059, pvalue=0.44412849543045874)
```

Conclusion 2 :
Les clients de niveau Bachelor et High School or Below répondent autant à l'engagement: donc par déduction les 3 niveau High School or Below, College et Bachelor ont le même taux de réponse à l'engagement.

3- Test-t d'indépendance entre Doctor et « High School or Below » :

```
import scipy.stats as stat
H = df[df['Education']=='High School or Below']['Engaged']
B = df[df['Education']=='Doctor']['Engaged']
print(stat.ttest_ind(H,B, equal_var=True))
```

```
Ttest_indResult(statistic=-2.2874515872636128, pvalue=0.022239672231116406)
```

Conclusion 3:
Les clients de niveau Doctor s'engagent largement moins que les 3 types de clients de niveau Bachelor, College et High School or Below.

4- Test-t d'indépendance entre Doctor et Master :

```
import scipy.stats as stat
H = df[df['Education']=='Master']['Engaged']
B = df[df['Education']=='Doctor']['Engaged']
print(stat.ttest_ind(H,B, equal_var=True))
```

```
Ttest_indResult(statistic=-0.5541129969555515, pvalue=0.5796161208187984)
```

Conclusion 4 :

Les clients de niveau Doctor et Master ont le même taux de réponse.

Conclusion générale :

Les clients de niveau College, Bachelor et «High School or Below» ont le même taux d'engagement entre-eux mais largement supérieur au taux d'engagement des clients de niveau éducatif élevé (Doctor et Master). Notons cependant que ces derniers (Doctor et Master) ont le même taux d'engagement.

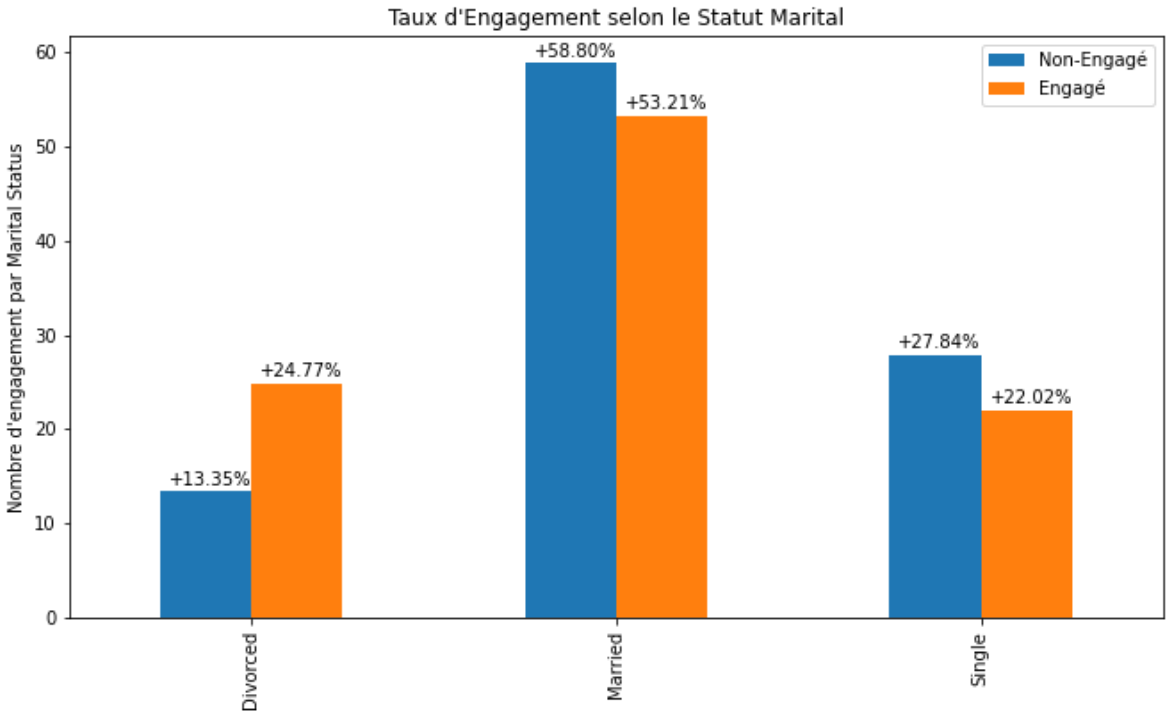
I.10 Engagement selon le statut marital des clients

Problème :
Marié ou célibataire, lesquels se sont le plus engagés à un appel marketing. Le statut marital, est-ce une raison d'engagement marketing?

I.10.1 Nombre de clients mariés et non-mariés.

	Fréquence de Clients Non-Engagés	Fréquence de clients Engagés
Marital Status		
Divorced	1045	324
Married	4602	696
Single	2179	288

I.10.2 Taux d'engagements selon le statut marital.



Interprétation :

Les clients mariés ont, à 53 %, répondu à l'engagement. Les divorcés (25 %) engagés sont en deuxième position et les célibataires en dernière position (22 %). Un test statistique permettra de confirmer ou infirmer le caractère réelle ou fausse de cette distribution .

Anova :

```
F_onewayResult(statistic=59.546718755896556,
      pvalue=2.0248113788379816e-26):
```

Rejet de L'hypothèse nulle et considération de l'hypothèse alternative.

Conclusion :

Les mariés se sont bel et bien plus engagés à l'appel marketing que les divorcés et les célibataires. Cela n'est pas vraie si l'on ajoute d'autres attributs (voire ci-haut).On peut aussi effectuer des tests des modalités de statut marital prises deux-à-deux pour affiner encore mieux ces résultats si nécessaire. L'analyse explicative prochaine nous fera savoir si cette distribution décrite constitue une raison suffisante d'engagement de ces clients.

II. ANALYSE EXPLICATIVE DE L'ENGAGEMENT MARKETING.

Introduction :

L'analyse descriptive que nous venons de réaliser cherchait à expliciter **en quoi l'engagement de nos clients consistait-il en fonctions de leurs caractéristiques relatives**. Elle nous a permis de révéler des différences significatives et d'autres non significatives dans l'engagement de nos clients, nous permettant de formuler des décisions stratégiques réalistes selon l'objectif marketing.

Ces différences observées « à l'œil nu » pouvaient s'avérer statistiquement fausses, qui amèneraient à la mise en place de stratégies marketing erronées avec comme conséquence la perte d'énormes sommes d'argents dépensés en terme de coût sur investissement. C'est là qu'apparait l'importance des test A/B que nous avons réalisé durant cette analyse mais aussi l'analyse explicative à l'engagement que nous allons effectuer tout de suite.

En effet, l'analyse descriptive nous permet seulement de découvrir le « quoi » et le « Comment » d'une chose, mais elle ne nous permet pas de comprendre le « pourquoi » de cette chose. C'est l'analyse explicative qui permet d'expliquer la cause de ce qui est découvert par l'analyse descriptive.

II.1 Impact des attributs de clients dans leurs décisions à refuser ou à répondre à un engagement marketing?

<u>Problème :</u>

Quelles attributs de nos clients influencent-elles ou sont susceptibles de pousser à un engagement ou à éviter un engagement marketing de nos clients ? En effet, certains caractéristiques, une fois présentes chez certaines clients, peuvent constituer des raisons fondamentales d'un engagement ou d'un refus à l'engagement.

Connaitre ces caractéristiques permet de s'en servir comme base d'une stratégie de ciblage sur des clients ou prospects dont nous ne savons pas forcément qu'ils peuvent s'engager mais qu'on peut

qualifier par ces attributs étudiés. Nous allons réaliser cette analyse explicative grâce à une régression logistique multivariée.

Résultats :

Logit Regression Results

Dep. Variable:	Engaged	**No. Observations:**	9134
Model:	Logit	**Df Residuals:**	9113
Method:	MLE	**Df Model:**	20
Date:	Tue, 11 Oct 2022	**Pseudo R-squ.:**	0.06951
Time:	14:56:25	**Log-Likelihood:**	-3490.8
converged:	True	**LL-Null:**	-3751.6
Covariance Type:	Nonrobust	**LLR p-value:**	8.793e-98

	Coef	**Std err**	**Z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-2.7461	0.185	-14.849	0.000	-3.109	-2.384
CLTV	-5.042e-06	5.07e-06	-0.994	0.320	-1.5e-05	4.9e-06
Income	1.368e-05	1.29e-06	10.570	0.000	1.11e-05	1.62e-05
MonthlyPremiumAuto	-0.0008	0.002	-0.450	0.652	-0.005	0.003
MonthsSinceLastClaim	-0.0034	0.003	-1.082	0.279	-0.009	0.003
MonthsSincePolicyInception	0.0001	0.001	0.098	0.922	-0.002	0.002
NumberOfOpenComplaints	-0.0260	0.035	-0.751	0.453	-0.094	0.042
NumberOfPolices	-0.0153	0.013	-1.143	0.253	-0.041	0.011
TotalClaimAmount	0.0002	0.000	1.317	0.188	-0.000	0.001
GenreFactorized	0.0183	0.062	0.293	0.769	-0.104	0.141
EducationFacctorized	0.0740	0.022	3.393	0.001	0.031	0.117
CoverageFactorized	0.0250	0.061	0.412	0.681	-0.094	0.144
MaritalStatus	0.2480	0.040	6.173	0.000	0.169	0.327
RenewOfferType	0.0106	0.031	0.347	0.728	-0.049	0.070
SalesChannel	-0.2133	0.026	-8.217	0.000	-0.264	-0.162
EmploymentStatus	0.5806	0.030	19.515	0.000	0.522	0.639
VehicleSize	0.0327	0.046	0.708	0.479	-0.058	0.123
VehicleClass	0.0602	0.046	1.323	0.186	-0.029	0.150
PolicyType	-0.0467	0.070	-0.666	0.505	-0.184	0.091
Policy	0.0153	0.014	1.059	0.290	-0.013	0.044
State	0.0105	0.024	0.442	0.659		

La variable à expliquer c'est l'engagement ; elle est présente dans ce tableau au nom de « **Engaged** »
Ici, les variables qui vont expliquer l'engagement sont à gauche, en dessous de la constate **« intercept».**

Trois choses fondamentales nous intéressent dans ce tableau :

1. La colonne **« P>|z| »** : elle porte les P-values de chaque variable explicative
2. La colonne « **Z** » : elle porte les statistique t de chaque variable explicative.
3. La colonne « **Coef** » : elle porte les coefficients linéaires de la relation entre variables explicatives et la variable expliqués (Engaged).

II.2 Interprétation :

1. **Étude explicative** de la variable « **Income** (revenu) » sur **l'engagement** (« Engaged »).

Résultats :

- Les statistiques de cette variable explicative sont : **state = 10.570, et P-value = 0 <5%** : elles montrent qu'il y a un rejet de l'hypothèse nulle et une considération de l'hypothèse alternative : existence d'impact des revenus des clients sur leurs susceptibilité à l'engagement.
- Le coefficient de corrélation (coef=1.368e-05) de cette variable sur l'engagement est positif mais très proche de zéro.

Conclusion :

- **Point 1 :** Le **revenu** des clients constitue une cause susceptible à leur engagement marketing.

- **Point 2** : cet engagement sera d'autant possible que le revenu de chaque client soit grand, cependant la corrélation est faible.

2. **Étude explicative** de l'attribut « **Renew Offer Type**» sur l'engagement des clients.

Résultats :

- Cette variable explicative correspond à **la state = 0.347 et à une P-value = 0.728 >5% :** on garde l'hypothèse nulle : existence d'indépendance…
- Le coefficient de corrélation est positif

Conclusion :

- **Point1** : bien que l'analyse descriptive précédente nous ait révélé que les clients engagés s'intéressaient plutôt à loffre2 qu'à l'offre1, il se trouve, selon l'analyse explicative, qu'un tel intérêt n'a pas constitué une raison assez suffisante d'engagement de ces clients.
- **Point 2** : le coefficient d'impact de cet attribut sur l'engagement est quasi nul (coef = .0106) et cela justifie l'indépendance entre type d'offre et engagement des clients.

3. **Étude explicative** de l'attribut **« Education »** sur **l'engagement** des clients**.**

Résultats :

- **State = 3.393, P-value = 0.001 < 5%** : rejet de l'hypothèse nulle et considération de l'hypothèse alternative.
- Le coefficient de corrélation est positif

Conclusion :

- **Point 1** : dans l'analyse descriptive, nous avons découvert que les clients dont l'éducation est de niveau « College » et « Bachelor » sont les plus engagés à l'appel marketing et cela a été confirmé par l'Anova.

Ici, avec l'analyse explicative , nous constations aussi que la P-value est très inférieur à 5% : **la distribution des modalités des niveaux éducatifs décrites par l'analyse descriptive précédente constituait une raison d'engagement des clients.**

- **Point 2** : Plus les différences de niveaux éducatifs augmentent suivant les mêmes proportions observées, plus les clients s'engagent. C'est-à-dire : plus il y aura un nombre élevé de clients de moindre niveau éducatif , plus il y aura d'engagements.

4. **Analyse explicative** de la variable « **Marital Status** » sur **l'engagement**.

Résultats :

- **State = 6.173 et P-value = 0 < 5% :** pvalue inférieur à 5% : rejet de l'hypothèse nulle et considération de l'hypothèse alternative.
- Coefficient : **coef = 0.2480>0.**

Conclusion :

- **Point 1** : Nous avons découvert dans l'analyse descriptive que les clients mariés (pas les employés mariés car ceux-ci sont plutôt en deuxième position dans l'engagement) sont les plus engagés. Mais nous ne savions pas si le fait d'être dans un « couple marié» constituait une raison suffisante d'engagement. L'analyse explicative que nous venons de réaliser ici montre que c'est effectivement le cas : être partenaire d'un couple marié constitue une raison suffisante à l'engagement.
- **Point 2** : cet engagement est d'autant plus possible que le nombre de clients mariés augmente.

5. **Analyse explicative** de l'attribut **« SalesChannel** » à **l'engagement** des clients

Résultats :

- **State = -8.217, P-value =0.000 < 5%** : : rejet de l'hypothèse nulle et considération de l'hypothèse alternative.
- Coefficient de corrélation : **coef = - 0.2133 < 0**

Conclusion :

- **Point 1** : l'analyse descriptive nous a montré que les clients qui se sont le plus engagés provenaient des agences de marketing (quasiment la moitié : 50.9%). Et le test d'indépendance l'approuve : la proportion d'engagement était différente selon le canal de vente. Il se trouve, ici, que l'analyse explicative approuve l'hypothèse posée dans l'analyse descriptive **: les agences de marketing de cette entreprise constituent une raison d'engagement des clients.**

- **Point 2** :Cependant, la probabilité à l'engagement via ces agences est d'autant moindre à mesure que la fréquence d'engagement est grande (coef = < 0).

6. **Analyse explicative** de la variable **« Employment Status »** sur **l'engagement** des clients.

Résultats :

- **State =** 19.515, P-value = 0 < 5% : rejet de l'hypothèse nulle et considération de l'hypothèse alternative : le statut d'emploi influence la décision des clients à l'engagement, et cela avec une intensité élevée (state =19.515 >> 0)
- **Coefficient : coef = +** 0.5806 > 0

Conclusion :

- **Point 1 :** L'analyse descriptive nous a montré, avec une différence statistiquement significative, que la moitié des clients avec un job sont engagés. Ici, l'analyse explicative que nous venons d'effectuer nous montre que cette information (le fait d'avoir un job) constitue une raison suffisante des clients à s'engager.
- **Point 2** : Et plus le nombre des clients ayant un emploi est grand, plus la tendance à l'engagement est grande (le coefficient de linéarité est positif)

Conclusion générale :

Le revenu des clients est une raison suffisante d'engagement, ces clients préfèrent l'offre 2 que l'offre 1 surtout pour les clients ayants des voitures à 4 et 2 portes (sauf les clients ayants des voiture de classes SUV : ils apprécient l'offre 1 que l'offre 2) **mais cet intérêt** (le fait d'éprouver un intérêt pour l'offre 2 ou l'offre 1) **n'est pas la raison de leur engagement.**

Ils s'engagent toujours moins à mesure que la majorité d'entre eux soit plus éduquée tandis qu'ils s'engagent toujours plus à mesure que la majorité d'entre eux soit moins éduquée.

Le fait d'être un mari ou une femme marié constitue aussi une raison suffisante d'engagement, cependant le statut « femme » ou « homme » tout court, ne constitue pas une raison d'engagement.

Ces clients aiment les agences marketing que tout autre canal de vente et cette préférence constitue aussi pour eux une raison d'engagement.

Plus le nombre de ces clients est employé, plus ils s'engagent (surtout pour les hommes avec un job que pour les femmes avec un job, ou les clients divorcés avec un job que les clients mariés ou célibataires avec un job). Et ce statut d'emploi constitue aussi une cause d'engagement.

Toutes ces informations extraites de cette analyse doivent être prises en compte lors d'une future mise en place d'une ou plusieurs stratégies marketings…

III. Analyse prédictive de l'engagement des clients

Introduction :

C'est bien une analyse descriptive ou explicative des données d'une entreprise pour comprendre les clients. Vous arrivez à identifier les raisons qui les ont poussés à effectuer des achats, à s'engager, à se désabonner ou à rester fidèles à vous. En gros, vous cherchez à comprendre l'historique des clients pour essayer de formuler des stratégies marketings sûres. Cependant il y a un problème : **l'analyse descriptive et explicative cherchent à comprendre ce qui s'est déjà passé mais pas ce qui va arriver**. Elles ne vous disent absolument rien de ce qui va arriver du future. Elles dévoilent par exemple quoi et comment des clients ont effectué des achats ou se sont engagés dans le passé mais ne vous disent jamais si ces clients vont encore une fois s'engager ou effectuer des achats durant les 2 mois prochains par exemple.

C'est là qu'intervient l'analyse prédictive : vous arrivez, avec un score de certitude de 99 %, à découvrir qu'un client va s'engager ou va te rapporter plus d'argent ou te faire perdre de l'argent durant ces 2 ou 3 mois prochains par exemple.

Dans cette partie, nous allons développer un modèle prédictif qui va permettre à notre entreprise d'identifier le client qui va s'engager à ses offres et le client qui ne va pas le faire : pour celui qui va s'engager, on peut se permettre de dépenser des sommes pour le pousser à l'engagement et pour celui qui ne va pas s'engager, on évitera de dépenser bêtement son argent pour l'engager. Ceci augmente les marges et réduit les coûts. Si vous n'êtes pas familier au code informatique, lisez seulement le texte pour bien comprendre.

III.1 Développement du modèle prédictive de l'engagement

1) Définition des variables explicative et de la variable cible (expliquée).

Nous séparons nos variables explicatives à la variable cible d'abord :

```
[10]: # Séparartion des variables en features et target
      X = df.drop('Engaged',axis=1)
      y = df['Engaged']
```

2) Encodage des variables catégorielles :

Toutes nos données ne sont pas quantitatives (numériques). Or les algorithmes d'intelligence artificielles utilisent des nombres et des formules mathématiques pour effectuer des calculs. Il est donc important de transformer toutes les attributs qualitatives des clients (variables catégorielles) en variables numériques. On appelle ça de **l'encodage**.

2.1) listons les variables numériques et catégorielles

[11]:
```
# classer les variable catégorielles et numériques dans deux listes
list_categ = []
list_numeric = []
for i,j in enumerate(list(X.columns)):
    if X[X.columns[i]].dtypes == object:
        list_categ.append(j)
    else:
        list_numeric.append(j)
print(list_categ,'\n')
print(list_numeric)
```

2.2) Encodage et fusion

[13]:
```
# Encodage des variable catégorielles et
dfx = X[list_categ]
dfx = pd.get_dummies(dfx)

#fusion des variables catégorielles encodées avec les variables numériqued
X_final = pd.concat([X[list_numeric], dfx], axis=1)
X_final.head(3)
```

[13]:

	Customer Lifetime Value	Income	Monthly Premium Auto	Months Since Last Claim	Months Since Policy Inception	Number of Open Complaints	Number of Policies	Total Claim Amount	Customer_AA10041	Customer_AA11235	...	Sales Channel_Web
0	2763.519279	56274	69	32	5	0	1	384.811147	0	0	...	0
1	6979.535903	0	94	13	42	0	8	1131.464935	0	0	...	0
2	12887.431650	48767	108	18	38	0	2	566.472247	0	0	...	0

3 rows × 9258 columns

3) Sélection des variables utiles

- Vous pouvez remarquer que la méthode d'encodage que nous venons d'utiliser augmente le nombre des variable explicatives (9258 colonnes contre 24 au début): un grand nombre de variable peut créer de l'« overfiting » (surapprentissage). Et cela créer un modèle qui parait bien entrainé (près de 100%) mais est incapable à effectuer de bonnes prédictions (sa performance est moins de 60 %)
- Comme nous l'avons vu dans l'analyse explicative, certaines de nos variables explicatives sont totalement indépendantes de la variable cible (l'engagement) et quasi constantes(variance proches de zéros): ce sont des variables sans importance prédictives ; elles ne sont pas des variables à grande variance.

Ces deux point nous obligent à identifier les variables explicatives à grande variance pour les utiliser, et laisser les variables inutiles.

[58]:
```
# Variables de grandes variances
X1 = X_final
X11 = pd.DataFrame(X1.var())
```

[59]:
```
X11 = X11.reset_index()
X11.columns = ['Nom_variable','Variance']
X11 = X11.sort_values(by='Variance', ascending=False)
X11.shape
X111 = X11.iloc[:20]
X111
```

[59]:

	Nom_variable	Variance
1	Income	9.229386e+08
0	Customer Lifetime Value	4.721020e+07
7	Total Claim Amount	8.439030e+04
2	Monthly Premium Auto	1.183908e+03
4	Months Since Policy Inception	7.787443e+02
3	Months Since Last Claim	1.014705e+02
6	Number of Policies	5.712969e+00
5	Number of Open Complaints	8.287982e-01
9249	Vehicle Class_Four-Door Car	2.499924e-01

Les variables sont classées suivant leurs variances de manières décroissante. Puis nous avons sélectionné les 20 premières variables à grande variance.

[60]:
```
# Top 40 des variables de grande variance (donc de grande importance predictive)
top20 = X111['Nom_variable'].tolist()
top20
```

[60]:
```
['Income',
 'Customer Lifetime Value',
 'Total Claim Amount',
 'Monthly Premium Auto',
 'Months Since Policy Inception',
 'Months Since Last Claim',
 'Number of Policies',
```

4) <u>Séparons nos des données en données d'entrainement et d'évaluation.</u>

Une manière de s'assurer que le modèle qu'on développe se fasse bien est de séparer les données : celles avec lesquelles nous créons le modèle et celles qui vérifient sa performance prédictive.

```
# Séparation des données en données d'entrainement et de teste
from sklearn.model_selection import train_test_split
X_train, X_test, y_train,y_test = train_test_split(X1[top20].values,y, test_size= 0.3, random_state=49, stratify= y)
```

```
print("Mombre de lignes et de colonnes des données d'entrainement:",X_train.shape)
print("Nombre de lignes et de colonnes des données d'évaluation:",X_test.shape)
```

```
Mombre de lignes et de colonnes des données d'entrainement: (6393, 20)
Nombre de lignes et de colonnes des données d'évaluation: (2741, 20)
```

III.2 **Modélisation**

Avant de développer le modèle, il faut s'assurer de deux choses :

- Utiliser le bon algorithme de machine learning.
- Utiliser les bons hyperparamètres du dit algorithme choisi.

Ce que nous allons faire en premier est de trouver le bon algorithme. Et pour ça, nous développons plusieurs algorithmes de classification en même temps puis les évaluons tous simultanément pour choisir le plus performant :

Recherche de bon algorithme

[120]:
```
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
```

[121]:
```
seed =7
models = []
models.append(('LogisticRegression', LogisticRegression()))
models.append(('LinearDiscriminantAnalysis', LinearDiscriminantAnalysis()))
models.append(('KNeighborsClassifier', KNeighborsClassifier()))
models.append(('DecisionTreeClassifier', DecisionTreeClassifier()))
models.append(('GaussianNB', GaussianNB()))
models.append(('SystemVectorMachineClassifier', SVC()))
models.append(('RandomForestClassifier', RandomForestClassifier()))
```

[122]:
```
results_mean =[]
results_std =[]
names = []
scoring = 'accuracy'
import warnings
warnings.filterwarnings("ignore")

for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed, shuffle=True)
    cv_results = model_selection.cross_val_score(model, X_train, y_train, cv=kfold, scoring=scoring)
    results_mean.append(cv_results.mean())
    results_std.append(cv_results.std())
    names.append(name)
    msg = "%s, %f, (%f)"% (name,cv_results.mean(), cv_results.std())
    #print(msg)
```

[123]:
```
results={'Model':names,'Accuracy':list(results_mean),'Écart_type':list(results_std)}
frame=pd.DataFrame(results)
frame.sort_values(by ='Accuracy', ascending = False)
```

[123]:

	Model	Accuracy	Écart_type
6	RandomForestClassifier	0.996403	0.002324
3	DecisionTreeClassifier	0.953387	0.007721
2	KNeighborsClassifier	0.877991	0.007314
0	LogisticRegression	0.856878	0.014995
5	SystemVectorMachineClassifier	0.856878	0.014995
4	GaussianNB	0.856409	0.014938
1	LinearDiscriminantAnalysis	0.856408	0.015519

Le **RandomForestClassifier** (algorithme de forêt aléatoires) et le **DecisionTreeClassifier** (algorithme d'arbre de décisions) semblent élus (scoring : 0.9965 pour le forêt aléatoires, et 0.953 pour l'arbre de décision).

Nous allons donc utiliser l'algorithme des forêt aléatoire pour développer le modèle final. Mais, comme vous allez le voir, utiliser le meilleur algorithme ne suffit pas. Il faut aussi savoir bien le paramétrer pour pouvoir le rendre performant :

Nous allons imbriquer notre algorithme dans une Pipeline contenant d'autre algorithmes de préparation et traitement des données(mise à l'échelle des données, croisement des données, du feature sélection…)

On va donc définitivement utiliser le rendom Forest Classifier

```
132]: from sklearn.pipeline import Pipeline
model = Pipeline([
    ('SelFirst',VarianceThreshold()),
    ('SelSecond', SelectKBest(score_func=f_classif, k=6)),
    ('Scaler', StandardScaler()),
    ('RFC',RandomForestClassifier())])

model.fit(X_train, y_train)

#print(model.score(X_train, y_train))
#print(model.score(X_test,y_test),'\n')

x_pred = model.predict(X_train)
y_pred = model.predict(X_test)
print('Training set:%0.4f'% accuracy_score(y_train,x_pred),'\n')
print('Testing set:%0.4f'% accuracy_score(y_test,y_pred))
```

```
Training set:0.8569

Testing set:0.8566
```

Et voilà : malgré le fait que nous avons utilisé l'algorithme qui nous paraissait mieux, nous n'avons pas obtenu sa performance optimale : le modèle a appris 85.69 % des formations historiques des clients mais n'est capable d'effectuer une prédiction d'engagement correcte qu'avec 85.66% de certitude (14,53 % de ses prédictions seront fausse.).

Nous allons effectuer un ajustement des hyperparamètres nous permettant de réduire encore ces 14.53% de fausses prédictions.
Remarquez que nous avons choisi au hasard une valeur de k égale à 6. Nous allons créer une listes de valeurs de k, puis laisser notre algorithme GridSearchCV nous choisir la bonne valeur correspondant aux bon hyperparamètres de notre algorithme.

- **<u>Les valeurs de k.</u>**

```
parametres = [{'SelSecond__k':[2,3,10,17,19,20]}]
```

- **<u>Recherche des bons hyperparamètres :</u>**

```
[127]: from sklearn.model_selection import StratifiedKFold
cvStrat = StratifiedKFold(n_splits=6,shuffle=True,random_state=49)

from sklearn.model_selection import GridSearchCV
grid = GridSearchCV(model, param_grid=parametres, cv=cvStrat, refit=True)

grid.fit(X_train, y_train)
```

- **<u>Résultats :</u>**

[128]:
```
grid.cv_results_
```

[128]:
```
{'mean_fit_time': array([0.80432121, 0.8176225 , 1.42132652, 2.21905677, 2.06700432,
        2.05949589]),
 'std_fit_time': array([0.07321676, 0.01667699, 0.16612269, 0.22292655, 0.12649748,
        0.22509682]),
 'mean_score_time': array([0.07887173, 0.07886994, 0.11520807, 0.0962172 , 0.11288929,
        0.11492896]),
 'std_score_time': array([0.01634693, 0.00544565, 0.05154606, 0.01027182, 0.0395823 ,
        0.04535394]),
 'param_SelSecond__k': masked_array(data=[2, 3, 10, 17, 19, 20],
              mask=[False, False, False, False, False, False],
        fill_value='?',
             dtype=object),
 'params': [{'SelSecond__k': 2},
  {'SelSecond__k': 3},
  {'SelSecond__k': 10},
  {'SelSecond__k': 17},
  {'SelSecond__k': 19},
  {'SelSecond__k': 20}],
 'split0_test_score': array([0.8564728 , 0.8564728 , 0.97748593, 0.99812383, 0.99906191,
        0.99812383]),
 'split1_test_score': array([0.8564728 , 0.8564728 , 0.84521576, 0.99061914, 0.99530957,
        0.99155722]),
 'split2_test_score': array([0.8564728 , 0.8564728 , 0.97748593, 0.99155722, 1.        ,
        0.99812383]),
 'split3_test_score': array([0.857277  , 0.857277  , 0.98122066, 0.99906103, 1.        ,
        1.        ]),
 'split4_test_score': array([0.857277  , 0.857277  , 0.9258216 , 0.99342723, 0.9943662 ,
        0.9943662 ]),
 'split5_test_score': array([0.857277  , 0.857277  , 0.98873239, 0.99812207, 0.99624413,
        0.9971831 ]),
 'mean_test_score': array([0.8568749 , 0.8568749 , 0.94932704, 0.99515175, 0.99749697,
        0.99655903]),
 'std_test_score': array([0.0004021 , 0.0004021 , 0.05090431, 0.00340043, 0.00227799,
        0.00279742]),
 'rank_test_score': array([5, 5, 4, 3, 1, 2])}
```

[130]:
```
LearnBest = grid.best_estimator_.predict(X_train)
predBest = grid.best_estimator_.predict(X_test)
```

[131]:
```
print('LearnBest:',accuracy_score(y_train, LearnBest))
print('PredBest:',accuracy_score(y_test, predBest))
```

```
LearnBest: 1.0
PredBest: 0.9985406785844583
```

Maintenant, les bonnes hyperparamètres ont été trouvés : notre modèle a appris les 100% des infos historiques des clients avec une capacité prédictive correcte de 99.85% (le model commettra la modique erreur de 0.15 % au lieu de 14.53%).

- **<u>Mais quelle est donc la bonne valeurs de k en accord avec les bonnes hyperparamètre de notre modèle ?</u>**

Découvrons-la :

```
[216]: pd.DataFrame(grid.cv_results_)[['params','mean_test_score']]
```

[216]:

	params	mean_test_score
0	{'SelSecond__k': 2}	0.856875
1	{'SelSecond__k': 3}	0.856875
2	{'SelSecond__k': 10}	0.950109
3	{'SelSecond__k': 17}	0.995464
4	{'SelSecond__k': 19}	0.996403
5	{'SelSecond__k': 20}	0.996402

On découvre que 4 valeurs de k peuvent convenir pour augmenter la performance prédictive de notre modèle final : 10, 17 , 19 et 20. Clairement, 19 et 20 se placent en première position.
En définitive, le bon algorithme c'est le **RandomForestClassifier** et la bonne valeur **d'hyperparamétrage** de **k** c'est **19.**

Nous pouvons donc utiliser ces deux éléments dans notre modèle déjà développé pour atteindre sa performance optimale comme suit (Nous avons utilisé k=20 et le résultat est parfait):

```
from sklearn.pipeline import Pipeline
model = Pipeline([
    ('SelFirst',VarianceThreshold()),
    ('SelSecond', SelectKBest(score_func=f_classif, k=20)),
    ('Scaler', StandardScaler()),
    ('RFC',RandomForestClassifier())])

model.fit(X_train, y_train)

#print(model.score(X_train, y_train))
#print(model.score(X_test,y_test),'\n')

x_pred = model.predict(X_train)
y_pred = model.predict(X_test)
print('Training set:%0.4f'% accuracy_score(y_train,x_pred),'\n')
print('Testing set:%0.4f'% accuracy_score(y_test,y_pred))
```

```
Training set:1.0000

Testing set:1.0000
```

Le modèle a appris toutes les informations historiques des clients qui se sont précédemment engagés (Training set = 1.0000).

Chaque fois que l'entreprise voudra savoir si un client donnée va s'engager à un appel marketing pour une offre, notre modèle révèlera à 100% de certitude qu'il va s'engager ou pas (Testing set = 1.0000) : il y aura 0.0000 (0.0 %) fausse prédiction au lieu de 14.53 % de fausse prédiction. Il ne reste maintenant qu'à intégrer ce modèle dans une page web pour que n'importe quel novice puisse l'utiliser pour détecter les possibilités d'engagement des clients de cette entreprise.

A titre d'exemple nous pouvons créer une petite fonction dans laquelle nous pouvons intégrer notre modèle pour effectuer une prédiction. Faisons cela tout de suite :

- **Prédiction**

  Soient i un numéro représentant le client C_i donnée, i étant un nombre de la liste [0,56,1,569, 600]. Vous pouvez choisir les nombres que vous voulez... Nous avons alors les clients C_0, C_56, C_1, C_569, C_600.

1- **Création de notre fonction prédictive de l'engagement** :

```
77]: # Prédiction
def Engagement(i):

    if y_pred[i]==1:
        return print(f"Le client C_{i}, va s'engager; vous pouvez le cibler")
    else:
        y_pred[i]==0
        return print(f"Il y a de forte chance que le client C_{i} ne s'engage plus")
```

2- <u>Vérifions si les clients ci-dessus vont s'engager ou pas à un appel marketing.</u>

a. <u>Est-ce que le client C_0 va-t-il s'engager ou pas ?</u>

```
Engagement(i=0)
```

```
Il y a de forte chance que le client C_0 ne s'engage plus
```

Marketing : on ne va pas cibler ce client.

b. <u>Est-ce que le client C_56 va-t-il s'engager ou pas ?</u>

```
[263]: Engagement(i=56)
```

```
Ce client va s'engager; vous pouvez le cibler
```

c. <u>Est-ce que le client C_569 va-t-il s'engager ou pas ?</u>

```
[279]: Engagement(i=569)
```

```
Le client C_569, va s'engager; vous pouvez le cibler
```

<u>Conclusion :</u>

Vous avez maintenant compris l'idée : vous prenez n'importe quel client que vous estimez potentiellement prometteur et vous le testez pour bien prendre une décision marketing. L'objectif de l'analyse prédictive est de rendre possible les analyses prescriptives. Connaissant ce qui a de grandes chances d'arriver, vous pouvez prescrire des recommandations, des stratégies etc.

Notez que nous avons juste associé des numéros au 3 clients que nous venons de tester mais dans la réalité on se sert des valeurs numériques de leurs attributs (les caractéristiques explicatives analysées ci-hautes) pour effectuer nos prédictions au travers d'une application web intégrant notre modèle.

Voici que se termine notre projet d'analyse de l'engagement. Il ne reste que le spécialiste en marketing pour se servir en toute conscience de toutes les informations découvertes dans cette analyse pour mettre en place des puissantes stratégies marketings prescriptives susceptibles d'engager les clients.

Notez cependant qu'il n'existe nulle part ailleurs une méthode définitive capable de cerner ce qui se passe dans la tête des humains. Il ne faut donc pas prendre l'analyse prédictive ou toute autre analyse comme une solution miracle pour cerner les décisions des humaine. Ces méthodes servent surtout d'aide aux prises de décisions sous la base de calcul que via l'intuition…

PROJET 2 :

Prédiction du chiffre d'affaire qu'un client de votre entreprise serai susceptible de vous rapporter durant les 3 prochains mois !

Problématique :

Si vous connaissez d'avance, avec 99% de certitude, la valeur monétaire qu'un client donné est susceptible de vous rapporter, que feriez-vous ? Et que feriez-vous si vous découvrez, avec une certitude de 99%, qu'un client donné ne va jamais effectuer des commandes dans les 3 prochains mois qui suivront sa dernière commande ?

Voici ce que je pense que vous ferez :

Cas 1 : cas où vous découvrez que votre client va certainement effectuer des achats durant les 3 prochains mois et la sommes exacte qu'il va vous apporter.

Vous allez chercher à faire en sorte :

1. Qu'à la fin de ces trois mois ce client vous rapporte l'argent prédit
2. Qu'il ne se désabonnes pas de vous. Qu'un concurrent ne te le pique.
3. Quand arrive le moment de proposer votre offre, tu sais d'avance ce que tu vas gagner et tu formuleras alors ce qu'il te faudra dépenser en terme de coût de tel sorte qu'il soit toujours inférieur au retour sur investissement.
4. Vous allez chercher à maintenir votre potentiel acheteur à chaud, grâce à un marketing taillé sur mesure, bien spécialisé.

Cas 2 : cas où votre client ne va certainement pas effectuer des commandes durant ces 3 prochains mois :

1. Vous allez éviter de le harceler par vos propositions de ventes, puis continuer plutôt à lui offrir de la valeur ajouté : le faire avancer…
2. Ou décider qu'il vaut mieux éviter de trop dépenser pour lui, si vous trouver qu'il est un vrai mauvais client.

Connaitre le client porteur de valeur et celui qui va vous faire dépenser des sommes sans retour dans une période donnée c'est tout ce qu'une entreprise veut savoir pour augmenter le chiffre d'affaire. **Nous allons développer un modèle qui va aider l'entreprise à réaliser ce travail.**

I. Analyse exploratoire des transactions passées des clients.

- Chargement des données

[3]:
```
df= pd.read_csv('C:/Users/user/Downloads/customer_segmentation.csv',encoding='latin1')
df['CustomerID'] = df['CustomerID'].astype(object)
df.head()
```

[3]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

I.1 Regroupement des commandes réalisées pour chaque client.

Apres avoir effectuer quelques nettoyages des données. Nous allons d'abord calculer les montants des commandes réalisées, puis regrouper toutes ces commandes pour chaque client. L'objectif est de regrouper, pour chaque client, le nombre des commandes, et l'historique des commandes. Cela nous aide à comprendre le comportement chorologique d'achats des clients.

```
commandes_df.head(10)
```

```
<
```

]:

CustomerID	InvoiceNo	Ventes	InvoiceDate
12346.0	541431	77183.60	2011-01-18 10:01:00
12347.0	537626	711.79	2010-12-07 14:57:00
	542237	475.39	2011-01-26 14:30:00
	549222	636.25	2011-04-07 10:43:00
	556201	382.52	2011-06-09 13:01:00
	562032	584.91	2011-08-02 08:48:00
	573511	1294.32	2011-10-31 12:25:00
12348.0	539318	892.80	2010-12-16 19:09:00
	541998	227.44	2011-01-25 10:42:00
	548955	367.00	2011-04-05 10:47:00

Interprétation :

1- Le client 12346.0 a effectué une seule commande le 18 janvier 2011 à 10h, d'une valeur de 77183.60 €.
2- Contrairement au premier client, le client 12347.0 a effectué 6 commandes entre le 7 décembre 2010 et le 31 octobre 2011, dont les valeurs monétaires sont successivement : 711.79€, 475.39€, 636.25€, 382.52€, 584.91€, et 1294.32€.
3- D'autres clients pas affichés sont bien évidemment présents dans cette table.

Nous allons réarranger cette tables de commande pour mieux fournir plus d'informations précieuses.

[12]:

	Ventes					InvoiceDate			
	Revenu_minimal	Revenu_maximal	Montant_total	Moyenne	Nombre	Date_ancienne_achat	Date_recente_achat	period_achat	nb_jours
CustomerID									
12346.0	77183.60	77183.60	77183.60	77183.600000	1	2011-01-18 10:01:00	2011-01-18 10:01:00	0	0.000000
12347.0	382.52	1294.32	4085.18	680.863333	6	2010-12-07 14:57:00	2011-10-31 12:25:00	327	54.500000
12348.0	227.44	892.80	1797.24	449.310000	4	2010-12-16 19:09:00	2011-09-25 13:13:00	282	70.500000
12349.0	1757.55	1757.55	1757.55	1757.550000	1	2011-11-21 09:51:00	2011-11-21 09:51:00	0	0.000000
12350.0	334.40	334.40	334.40	334.400000	1	2011-02-02 16:01:00	2011-02-02 16:01:00	0	0.000000
12352.0	120.33	840.30	2506.04	313.255000	8	2011-02-16 12:33:00	2011-11-03 14:37:00	260	32.500000
12353.0	89.00	89.00	89.00	89.000000	1	2011-05-19 17:47:00	2011-05-19 17:47:00	0	0.000000
12354.0	1079.40	1079.40	1079.40	1079.400000	1	2011-04-21 13:11:00	2011-04-21 13:11:00	0	0.000000
12355.0	459.40	459.40	459.40	459.400000	1	2011-05-09 13:49:00	2011-05-09 13:49:00	0	0.000000
12356.0	58.35	2271.62	2811.43	937.143333	3	2011-01-18 09:50:00	2011-11-17 08:40:00	302	100.666667

Interprétation

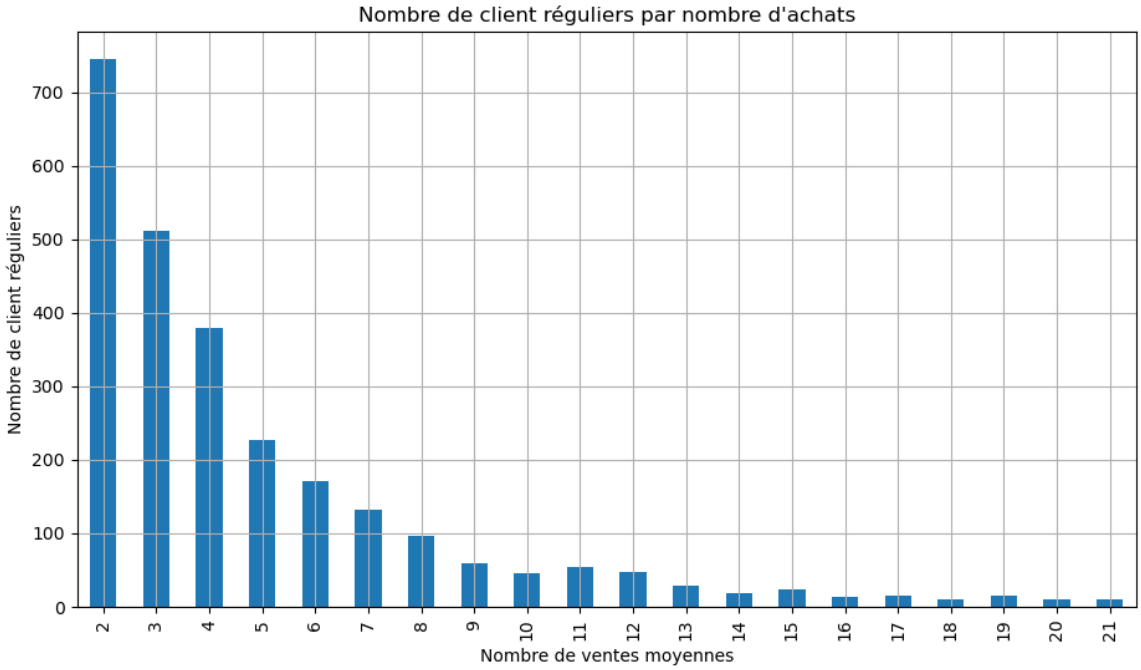
Cette table donne une vue d'ensemble des comportements chronologiques des commandes des clients, comme la fréquence des commandes, les périodes actives de ces commandes, le temps écoulé entre deux commandes successives, les commandes régulières etc. Ainsi on a par exemple :

1. Le client 12346.0 (la 1ère ligne) n'a effectué qu'une seule commande le 18 janvier 2011(Sa fréquente d'achat est de zéros . Rappel : 0 fréquence de commande = pas de répétition de la première commande). Dans cette Commande, ce client a opéré plusieurs transactions qui s'élèvent à une valeur totale de commande unique de 77183.6 en moyenne (Valeur moyenne monétaire). Son Montant total monétaire est de 77183.60: le même que la moyenne car il n'a effectué qu'une seule commande. La durée écoulée entre sa première et sa dernière commande est zéros(car il n'en effectué qu'une seule, comme c'est déjà dit)

2. Par contre, le client 12348.0 (ligne 3) a effectué 4 commandes entre Le 16 décembre 2010 et le 25 septembre 2011, et cela durant une période de 282 jours entre sa première et sa dernière commande. Le montant moyen dépensé par ce client pour chaque commande s'élève à une valeur de 449.31, et ce client, en moyenne, effectuait chaque commande tous les 70.5 jours etc... Présentons la régularité des commandes sous un visuelle pour mieux comprendre ces informations.

I.2 Visualisation de la régularité des commandes

:



Interprétation

Chaque bar de ce visuel représente un groupe de clients dont le nombre est représenté par l'axe vertical tandis que le nombre de commandes (la régularité des commandes) est représenté par l'axe horizontal. Ainsi, le groupe de clients qui effectuent au plus 2 ventes est majoritaire.

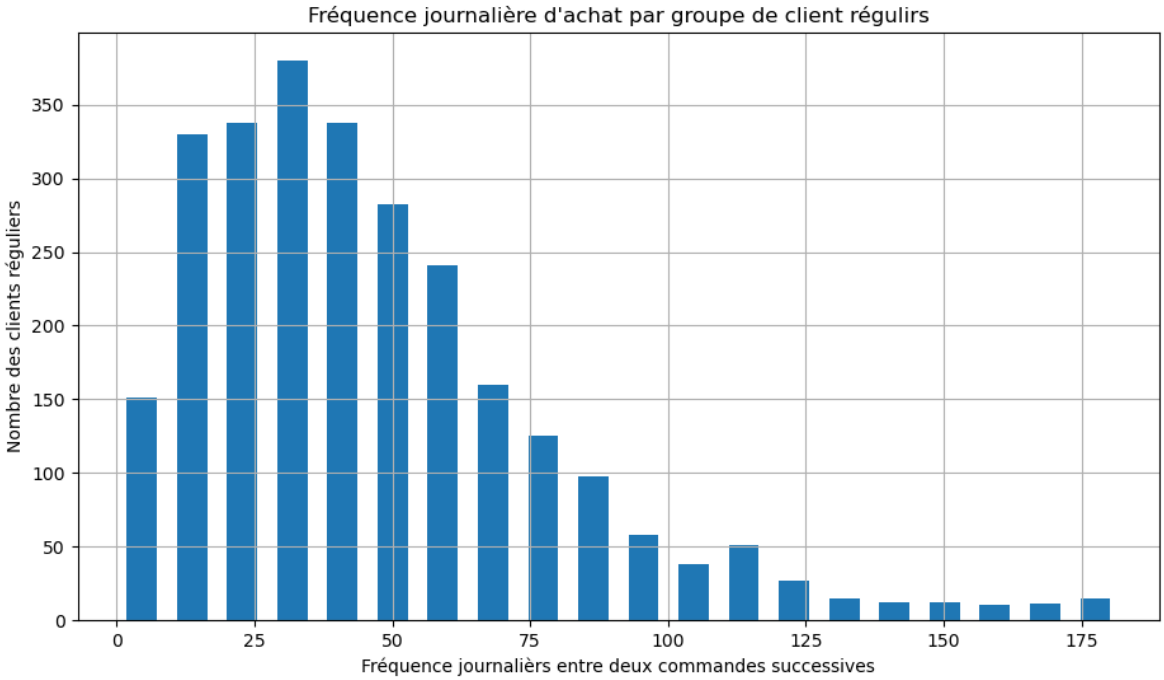
Si on observe bien : plus les clients sont réguliers (ont effectué plus d'achats), plus ils se raréfient en nombre.

Ces informations permettent de comprendre à quel moment il faut arrêter , selon le groupe, de pousser toujours vos clients à se convertir: vos clients ne doivent pas se sentir comme des vaches à lait. Ainsi, dans le groupe majoritaire, il ne faut pas par exemple pousser les clients qui le compose à effectuer plus de 2 achats. Mais si vous le faites, il y aura à la fois une augmentation de commandes et une diminution du nombre de clients dans le groupe : le groupe va se rapprocher de celui du 2ème groupe où les clients effectuent maximum 3 ventes (plus de vente mais moins de nombre de clients dans le groupe par rapport au group 1).

Ainsi, selon l'objectif marketing, vous savez maintenant quel groupe de clients il faut cibler…

I.3 Durée écoulée entre deux commandes successives

Là, nous avons représenté la régularité des commande, mais il serait encore intéressant que l'on visualise la durée écoulée entre deux commandes successives régulières de chaque groupe de clients. Faisons cela tout de suite :



Interprétation :

La majorité des clients achètent à environs tous les 25 jours ou les 50 jour: les clients qui passent plus de 75 jours avant de revenir à l'entreprise effectuer des commandes se raréfient. Avec ces informations, nous savons maintenant les clients les plus actifs de ceux qui sont moins actifs : les premiers sont souvent ceux qui sont nouveaux (leurs durées entre deux commandes successives sont très courtes) et les derniers sont souvent des anciens clients. Ceux-ci passent plus de temps avant de revenir à l'entreprise effectuer des commandes.

Qu'ils soient nouveaux ou anciens, les clients dont l'activité d'achats diminue semblent entrain de se désabonner : **ce sont des clients qui nécessitent une stratégie marketing de rétention…**

I.4 Les transactions trimestrielles des commandes.

Il nous reste une dernière chose à faire avant le développement de notre modèle. Notre objectif est justement de prédire des valeurs monétaires trimestrielles par client. Nous devons donc d'abord transformer nos données à correspondre à des transactions trimestrielles. Faisons cela tout de suite :

[91]:

	CustomerID	InvoiceDate	Ventes_Montant_total	Ventes_Moyenne	Ventes_Nombre	Mois_trimestriel
1	12347.0	2010-12-31	711.79	711.790	1	Décembre_précédent
6	12348.0	2010-12-31	892.80	892.800	1	Décembre_précédent
0	12346.0	2011-03-31	77183.60	77183.600	1	Mars
2	12347.0	2011-03-31	475.39	475.390	1	Mars
7	12348.0	2011-03-31	227.44	227.440	1	Mars
3	12347.0	2011-06-30	1018.77	509.385	2	Juin
8	12348.0	2011-06-30	367.00	367.000	1	Juin
4	12347.0	2011-09-30	584.91	584.910	1	Septembre
9	12348.0	2011-09-30	310.00	310.000	1	Septembre
5	12347.0	2011-12-31	1294.32	1294.320	1	Décembre

Là , nous voyons que les dates de facturation des commandes passées se différentient d'un écart de 3 mois : chaque commande successive de chaque client se fait 3 mois après la précédante commande. Seulement cela se fait selon les lignes. Ce qui n'est pas très pratique . Nous allons donc croiser cette table et faire en sorte que les commandes trimestrielles, pour chaque client, se fasse suivant la direction des colonnes : chaque mois trimestriel présent dans la colonne « **Mois_trimestriel »** va être associer aux 3 colonnes « **Ventes_Montant_total**», « **Ventes_Moyenne** » et « **Ventes_Nombre »** . Ce qui va alors nous donner le comportement trimestriel des commandes de chaque client. Faisons cela tout de suite :

CustomerID	ventes_montant_total_décembre_précédent	ventes_montant_total_juin	ventes_montant_total_mars	ventes_montant_total_septembre
12346.0	0.00	0.00	77183.60	0.00
12347.0	711.79	1018.77	475.39	584.91
12348.0	892.80	367.00	227.44	310.00
12350.0	0.00	0.00	334.40	0.00
12352.0	0.00	0.00	1561.81	632.50
12353.0	0.00	89.00	0.00	0.00
12354.0	0.00	1079.40	0.00	0.00
12355.0	0.00	459.40	0.00	0.00

Nous n'avons pas pris toute la table car elle est très grande. Là maintenant, nous pouvons s'en servir pour développer notre modèle prédictif.

I.5 Développement et déploiement du modèle prédictif de la valeur trimestriel des commandes de clients.

La manière générale de développer un modèle de machine learning est déjà étudiée dans notre projet 1. Nous n'allons pas le répéter ici. De même, le déploiement d'un modèle c'est tout un chapitre. On ne va pas le faire ici aussi. Bien en tendu, nous avons déjà fait cela mais nous avons choisi de ne pas les publier ici car c'est trop long. Mais pour comprendre le rôle, voici le lien pour le tester si tu en a envie : https://monclv-1.herekuapp.com
Notez que l'application a été développée grâce à Flask, le html et css. C'est une application simple dont le but est de simplement permettre à l'entreprise d'effectuer ses prédictions. Voilà la fin de ce projet.

Projet 3 :

1- Analyse chronologique des produits et des interactions clients-entreprises et

2- Étude de système de recommandation des bons produits aux bons clients.

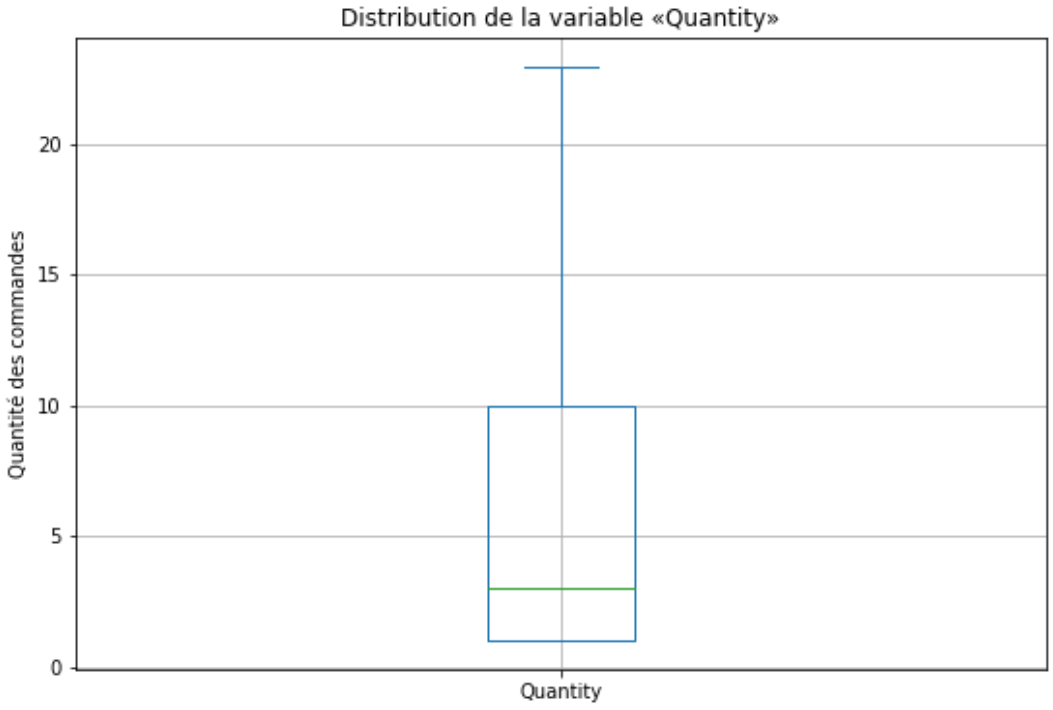
Problématique :

Quand vous vendez des produits, vous avez envie de comprendre l'interaction qui existe entre les clients et vos produits : **vous voudrez connaitre les produits qui paraissent populaires aux yeux des gens d'une part, et le type de gens qui semblent attirés par ces produits populaires d'autre part.**

La connaissance de ces deux informations vous pernet de **raccommoder** correctement les **produits qui semblent tendances ou populaires** aux gens qui semblent plus intéressés. Ce qui augmente les ventes.

I.1 Analyse chronologique des produits.

I.1.1 Distribution de la quantité des commandes des produits.



La quantité médiane des commandes des produits n'est pas au milieu de la boite : elle est proche de la première quartil et très loin du 3ème quartil. De plus, le 3ème quartil est confondu au nombre moyen des commandes. Cette médiane est d'environ 3 commandes :

```
[365]: count    531285.000000
       mean         10.655262
       std         156.830323
       min           1.000000
       25%           1.000000
       50%           3.000000
       75%          10.000000
       max       80995.000000
       Name: Quantity, dtype: float64
```

Là, on voit clairement que les commandes médianes sont au nombre de 3, les commandes moyennes sont au nombre de 10.66 (quasi la même valeur que dans le cas des 75% des commandes c'est-à-dire du 3ème quartil).

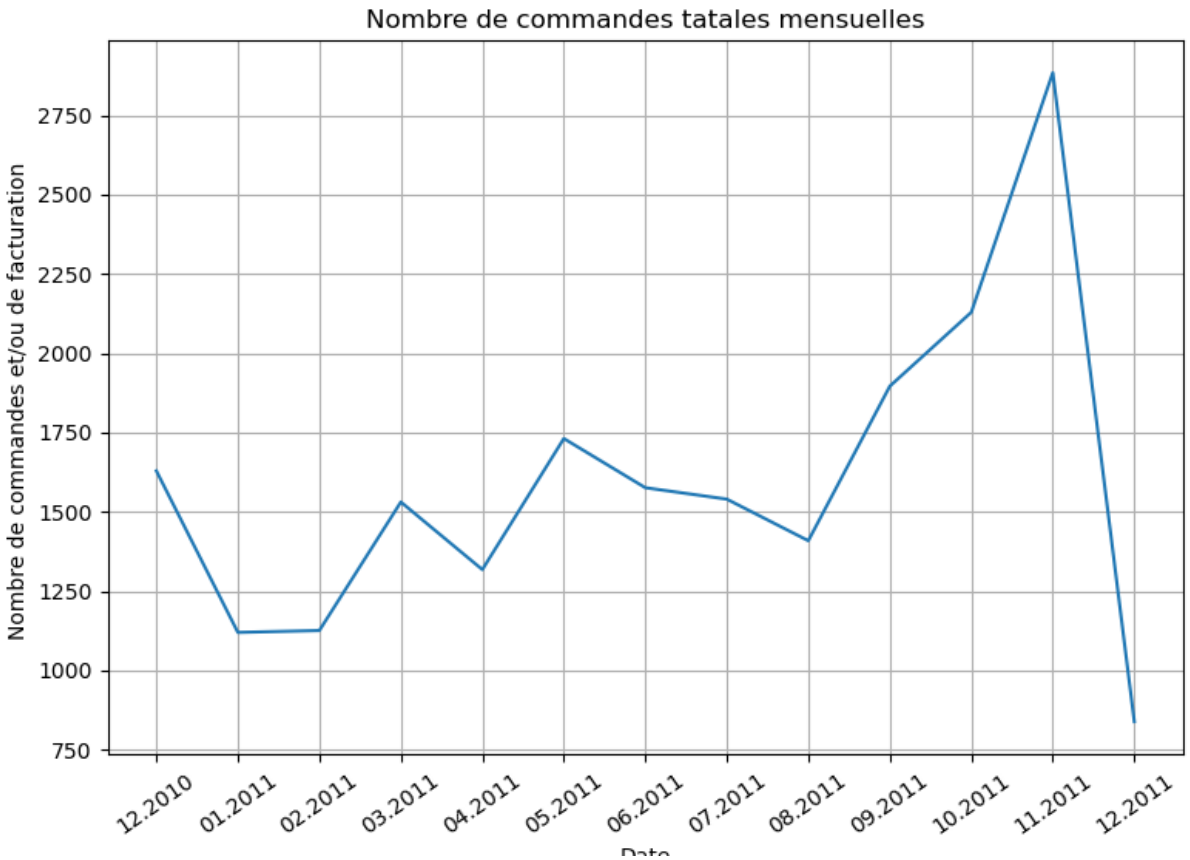
I.1.2 Série chronologique des commandes :

I1.2.1 Chronologie mensuelle des commandes globales:

InvoiceDate	InvoiceNo
2010-12-31	1629
2011-01-31	1120
2011-02-28	1126
2011-03-31	1531
2011-04-30	1318
2011-05-31	1731
2011-06-30	1576
2011-07-31	1540
2011-08-31	1409
2011-09-30	1896
2011-10-31	2129
2011-11-30	2884
2011-12-31	839

Nous avons la relation en série entre les dates de factures mensuelles et le nombre des numéros de ces factures. Ces deux informations donnent le nombres de commandes par mois.

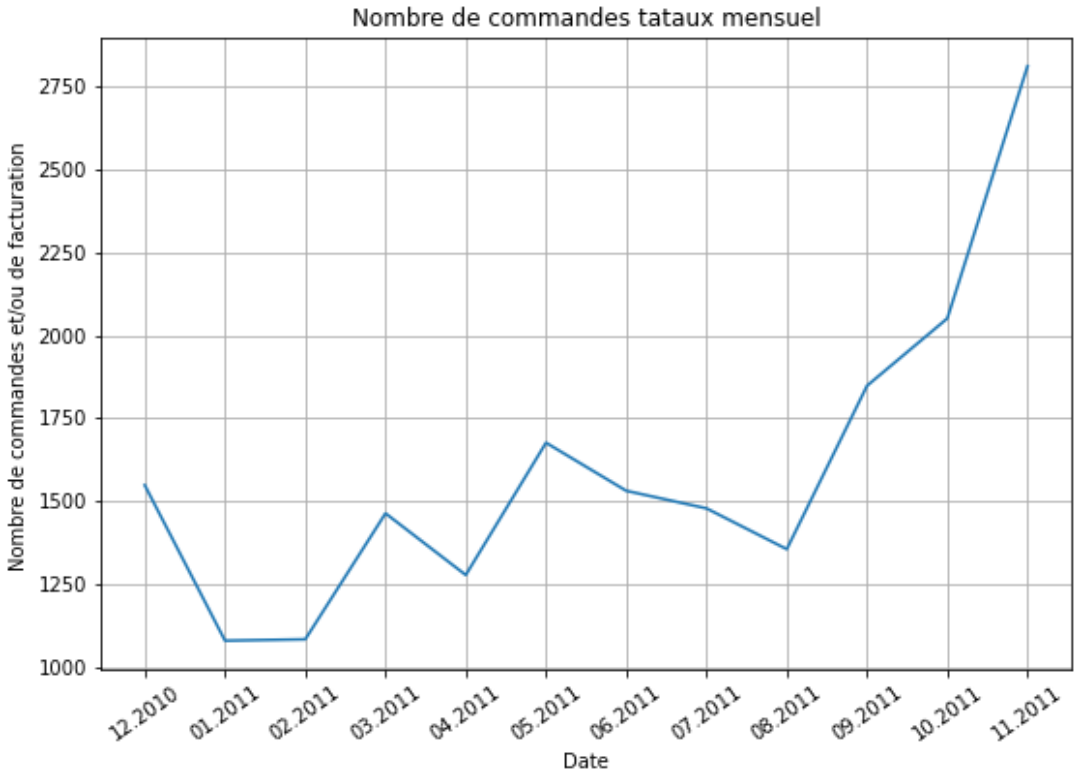
Remarquons qu'en **décembre**, **le nombre de commande est très inférieurs** que les autres nombres de commandes de la série. Visualisons cela pour voir clairement cette baisse.



La tendance des commandes décrite par cette courbe montre qu'elle est croissante jusqu'à décembre puis commence à chuter. En effet, la courbe montre qu'en novembre il y a eu beaucoup d'achats contrairement en décembre où le nombre d'achats diminue. Cette diminution est dû au fait qu'en ce mois de décembre les données sont incomplètes. Nous allons donc supprimer ce mois de décembre dans la liste des dates.

[46]:
```
InvoiceDate
2010-12-31    1629
2011-01-31    1120
2011-02-28    1126
2011-03-31    1531
2011-04-30    1318
2011-05-31    1731
2011-06-30    1576
2011-07-31    1540
2011-08-31    1409
2011-09-30    1896
2011-10-31    2129
2011-11-30    2884
Freq: M, Name: InvoiceNo, dtype: int64
```

Nous venons de supprimer les données incomplètes des commandes en éliminant carrément le mois de décembre dans la série. Effectuons une visualisation pour bien observer la tendance chronologique des commandes en dehors de décembre 2011:



Là, tout est bon. La manière dont la croissance des commandes s'est faite en novembre est très différente. Dans ce mois de novembre, il y a eu un pic de commandes qui nous pousse à se poser cette question : **qu'est-ce qui est à l'origine de cette croissance exponentielle soudaine de commandes en novembre seul, au lieux de tous les autres mois précédents novembre ?**

Cette question cherche à rassurer l'entreprise par rapport aux actions marketing qu'elle a entreprise avant le lancement des ventes**: la croissance en novembre 2011 est-ce due à un phénomène saisonnier ou bien à la régularité des clients** ? En effet, si la tendance des commandes change soudainement dans un mois donnée, c'est qu'il doit y avoir une raison de plus par rapport aux actions marketing menées : soit c'est donc dû à la saisonnalité de l'année, soit cela est dû à la régularité des clients acheteurs.

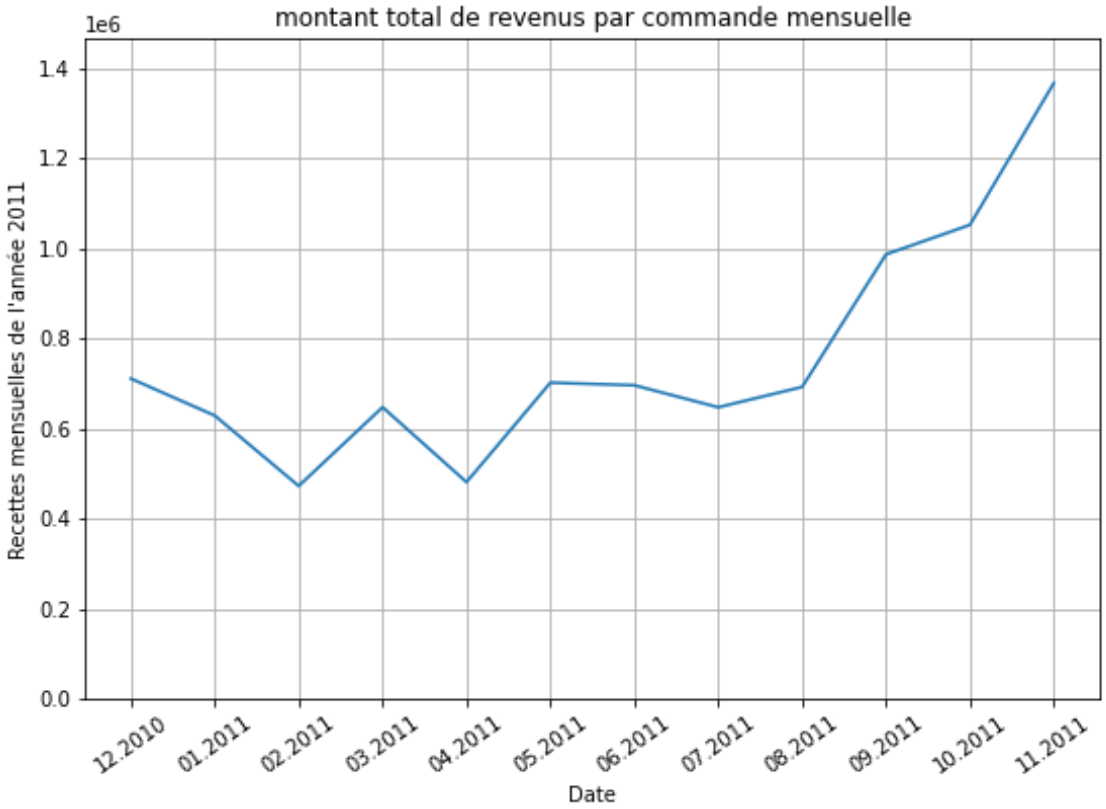
Pour que l'entreprise puisse décider quelle stratégie marketing adopter, il faut que l'une de ces deux raisons soit choisie. Dans la partie suivante, nous allons chercher à savoir si la croissance est due à la saisonnalité ou bien à la régularité des clients.

<u>Etape 1 : comparaison des courbes de tendance de nombre et de revenu de commandes</u>.

Ce que nous venons de faire concernait le nombre de commandes. Ici, nous allons d'abord étudier la tendance des revenus et voir si cette tendance suit la même logique de croissance.

InvoiceDate	Ventes
2010-12-31	823746.140
2011-01-31	691364.560
2011-02-28	523631.890
2011-03-31	717639.360
2011-04-30	537808.621
2011-05-31	770536.020
2011-06-30	761739.900
2011-07-31	719221.191
2011-08-31	737014.260
2011-09-30	1058590.172
2011-10-31	1154979.300
2011-11-30	1509496.330

Avec cette séries de revenu issus des commandes de clients, nous pouvons maintenant tracer la courbe de tendance chronologiques de ces revenus afin de la comparer à la courbe précédante du nombre de commandes.



On voit clairement que la tendance chronologique des revenus (des vents) semble la même que la tendance chronologique du nombre de commandes faite en haut. **Là, nous somme certain que les ventes générées et le nombre de commandes réalisées proviennent des mêmes clients**

<u>Etape 2 : recherche des clients réguliers.</u>

Notons que parmi ces clients, certains ont juste effectué une seule commande tandis que d'autres en ont effectué plusieurs : nous allons exclure ceux qui ont effectué une seule commande et garder ceux qui ont effectué plusieurs commandes (les clients réguliers).

1. **Série chronologique du nombre de tous les clients acheteurs (à une commande et à plusieurs commandes)**

InvoiceDate	CustomerID
2010-12-31	885
2011-01-31	741
2011-02-28	758
2011-03-31	974
2011-04-30	856
2011-05-31	1056
2011-06-30	991
2011-07-31	949
2011-08-31	935
2011-09-30	1266
2011-10-31	1364
2011-11-30	1665

2. **Série chronologique du nombre de clients régulier (à plusieurs commandes seulement))**

InvoiceDate	CustomerID
2010-12-31	263
2011-01-31	153
2011-02-28	153
2011-03-31	203
2011-04-30	170
2011-05-31	281
2011-06-30	220
2011-07-31	227
2011-08-31	198
2011-09-30	272
2011-10-31	324
2011-11-30	541

3. Pourcentage chronologique des clients réguliers

```
InvoiceDate
2010-12-31    29.717514
2011-01-31    20.647773
2011-02-28    20.184697
2011-03-31    20.841889
2011-04-30    19.859813
2011-05-31    26.609848
2011-06-30    22.199798
2011-07-31    23.919916
2011-08-31    21.176471
2011-09-30    21.484992
2011-10-31    23.753666
2011-11-30    32.492492
Freq: M, Name: CustomerID, dtype: float64
```

4. Résumé : clients totaux, réguliers et à une commande par mois.

InvoiceDate	nb_clients_regulier	nb_total_clients	client_à_1_achat
2010-12-31	263	885	622
2011-01-31	153	741	588
2011-02-28	153	758	605
2011-03-31	203	974	771
2011-04-30	170	856	686
2011-05-31	281	1056	775
2011-06-30	220	991	771
2011-07-31	227	949	722
2011-08-31	198	935	737
2011-09-30	272	1266	994
2011-10-31	324	1364	1040
2011-11-30	541	1665	1124

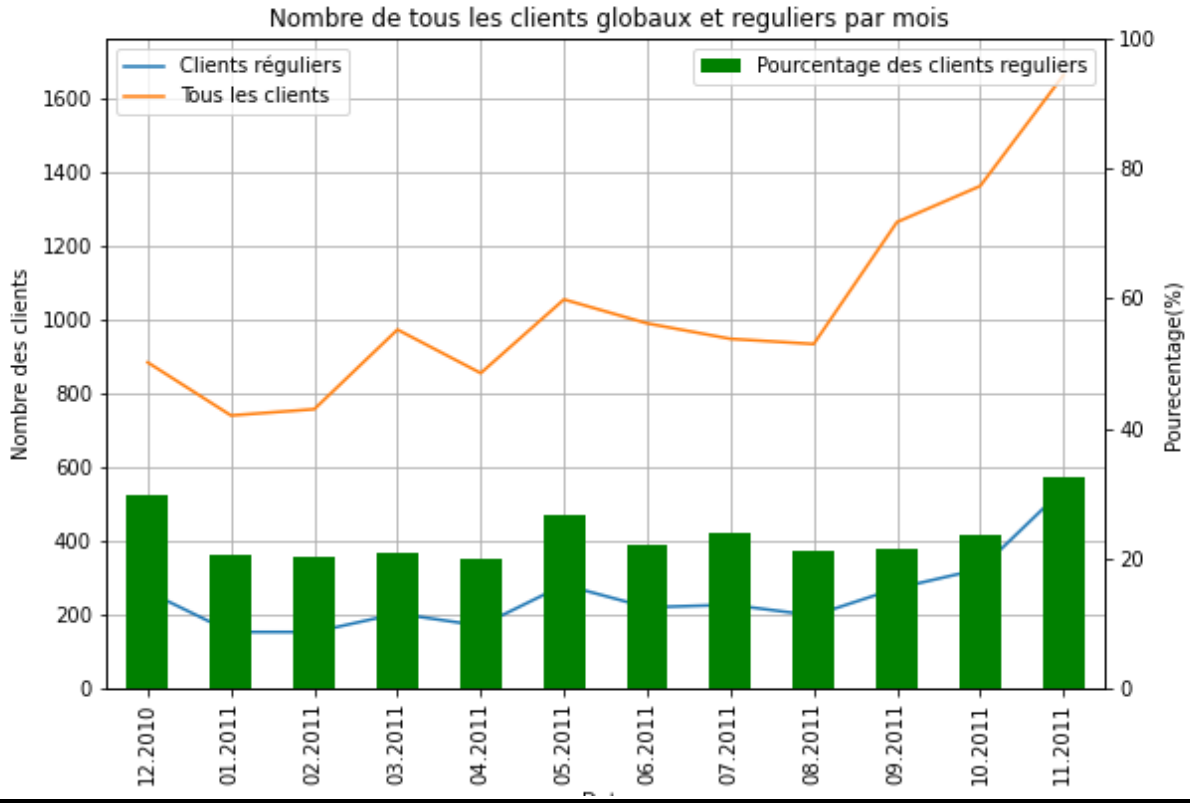
Nous avons maintenant le nombre de clients réguliers le nombre de clients totaux et le nombre de clients ayant effectués une seule commande.

Nous allons faire deux choses : réunir les courbes entre nombre de client totaux, nombre de clients réguliers et pourcentage du nombre de clients réguliers dans un même visuel d'une part, puis les courbes des revenus générés par le nombre de clients taux, le nombre de clients réguliers et le pourcentage de revenu des clients réguliers d'autres part.

Dans chacun de ces deux visuels, nous voulons comparer les comportements d'achats de tous les clients par rapport aux comportements d'achats des clients réguliers :

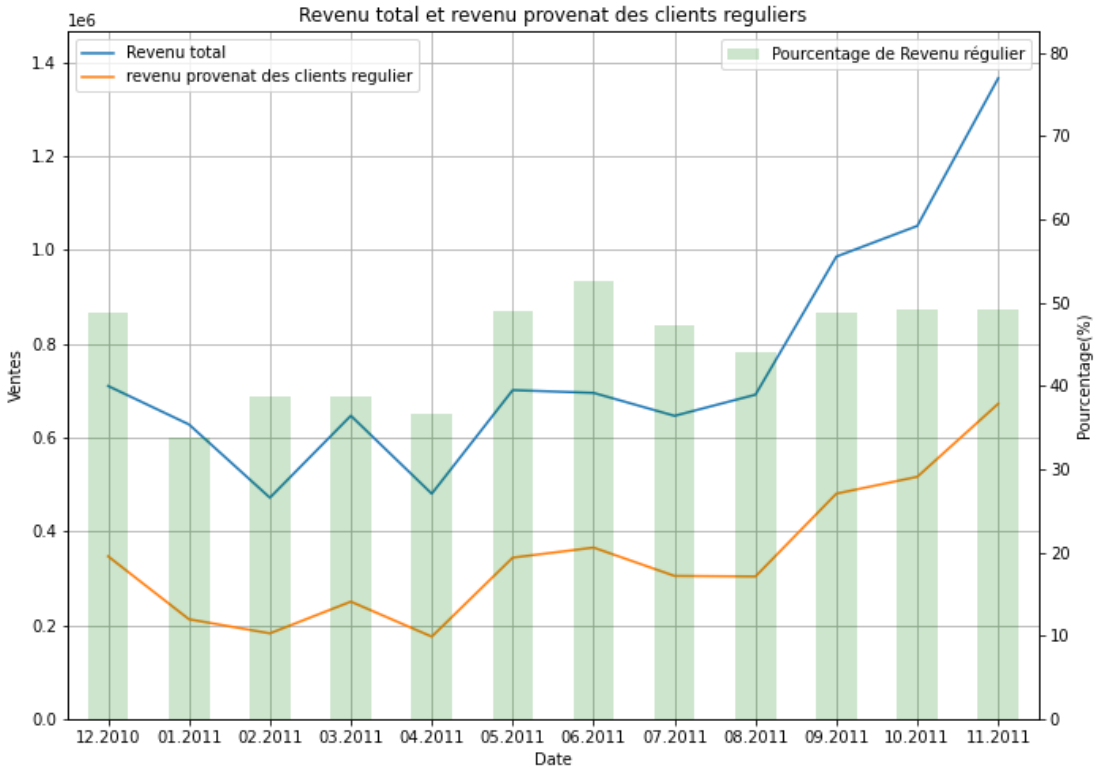
- Si le comportement d'achats de tous les clients se diffère du comportement d'achats des clients réguliers, alors la croissance des ventes en novembre n'est pas due à la régularité des clients : elle doit alors provenir d'une autre raison, comme la saisonnalité par exemple.
- Mais si le comportement d'achats de tous les clients suit la même tendance chronologique que celui de la tendance des clients réguliers, alors la croissance observée en novembre est due à la régularité des clients. Effectuons ce travail.

Cas, selon le nombre de commandes.



Clairement, la distribution du nombre des clients réguliers (courbe bleue et barres vertes) se comporte de la même manière que la distribution du nombre de tous les clients (courbe orangée) : l'hypothèse de la régularité des commandes semble crédible que l'hypothèse de la saisonnalité. Mais pour nous rassurer, faisons la même chose dans le cas des revenus :

Cas selon les revenus des commandes.



Interprétation :

- La distribution relative aux revenus réguliers (courbe orangée et barres vertes) suit la même tendance chronologique que la distribution chronologique des revenus totaux (provenant de tous les clients : courbe bleue)
- Le pourcentage de revenu oscille entre 35 et 50 %

Conclusion :

- **La croissance observée en novembre** est due non pas à un phénomène saisonnier mais bien grâce à la **régularité** de commandes de clients réguliers.
- Cette régularité génère un revenu important.

La stratégie marketing à mettre en place :

Étant donné que la croissance n'est pas saisonnière, cela veut dire que ces clients n'achètent pas en fonction des mois de l'année : **chaque mois pouvait donc être un mois de croissance, car ces clients sont susceptibles d'effectuer des achats à tout moment.**

1- **Ces clients doivent être fidélisés et veiller quand ils sont actifs et inactifs, et puis contrôler leurs susceptibilité aux désabonnements futures.**
2- Il faut les identifier et identifier les produits qu'ils ont commandés
3- Il faut ensuite chercher les mois passés où ils ont effectué des achats et quels types de produits ils ont acheté.

5. Recherche du top 5 des produits tendances.

5.1 Nombre de fois de ventes de chaque produit par mois :

```
InvoiceDate  StockCode
2010-12-31   10002        251
             10120         16
             10123C         1
             10124A         4
             10124G         5
             10125        154
             10133        130
             10135        411
             11001         74
             15034         45
Name: Quantity, dtype: int64
```

5.2 Quantité de ventes de chaque produit en novembre.

	StockCode	Quantity
0	23084	14954
1	84826	12551
2	22197	12460
3	22086	7908
4	85099B	5909
5	22578	5366
6	84879	5254
7	22577	5003
8	85123A	4910
9	84077	4559

5.3 Le top 5 des produits populaires.

Dans la table 5.2 nous avons arrangé, de manières décroissante, les quantité de ventes de chaque produit **: les 5 premiers sont les plus appréciés par les clients :**

```
list(tri_produits['StockCode'])[:5]
```

```
['23084', '84826', '22197', '22086', '85099B']
```

Voici donc **les codes de stock des 5 produits tendances de novembre 2011**. Trouvons, parmi ces 5 produits, ceux qui ont été commandées durant les autres mois de l'année.

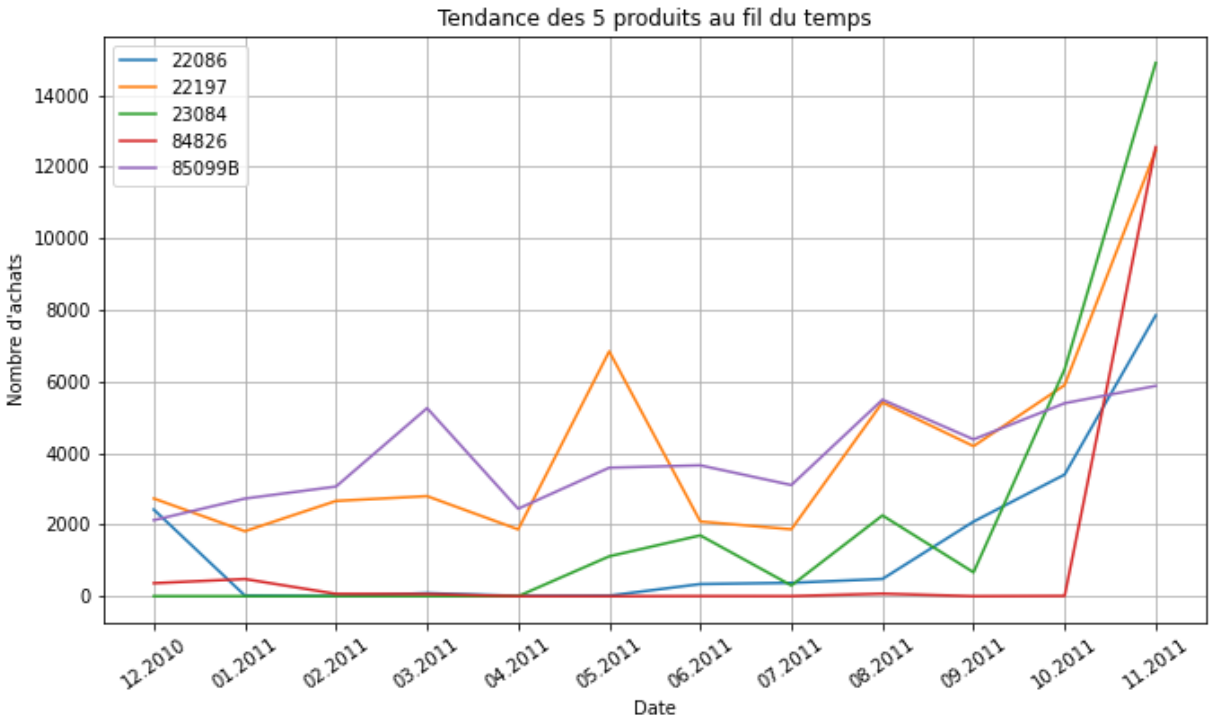
```
: InvoiceDate  StockCode
  2010-12-31   22086        2460
               22197        2738
               84826         366
               85099B       2152
  2011-01-31   22086          24
               22197        1824
               84826         480
               85099B       2747
  2011-02-28   22086           5
               22197        2666
               84826           66
               85099B       3080
  2011-03-31   22086          87
               22197        2803
               84826           60
               85099B       5282
  2011-04-30   22086          13
               22197        1869
               84826           1
               85099B       2456
  Name: Quantity, dtype: int64
```

Nous n'avons pas affiché toute la série ici. On remarque que parmi les 5 produits tendances en novembre 2011, 4 seulement ont été commandés en décembre 2010 tandis que le produit le plus acheté en novembre 2011 (le produit 23084) n'a fait l'objet d'aucune commande en décembre 2010 : c'est le produit 22 197 qui est le plus commandé en décembre 2010.
Ainsi, suivant chaque date, nous pouvons décrire les produits tendances commandés dans ce mois.

Pour comprendre et mieux interpréter la chronologie tendance des 5 produits populaires, visualisons cela. Effectuons d'abord un réarrangement de nos données.

StockCode	22086	22197	23084	84826	85099B
InvoiceDate					
2010-12-31	2460.0	2738.0	0.0	366.0	2152.0
2011-01-31	24.0	1824.0	0.0	480.0	2747.0
2011-02-28	5.0	2666.0	0.0	66.0	3080.0
2011-03-31	87.0	2803.0	0.0	60.0	5282.0
2011-04-30	13.0	1869.0	0.0	1.0	2456.0
2011-05-31	17.0	6849.0	1131.0	0.0	3621.0
2011-06-30	344.0	2095.0	1713.0	4.0	3682.0
2011-07-31	383.0	1876.0	318.0	2.0	3129.0
2011-08-31	490.0	5421.0	2267.0	72.0	5502.0
2011-09-30	2106.0	4196.0	680.0	0.0	4401.0
2011-10-31	3429.0	5907.0	6348.0	11.0	5412.0
2011-11-30	7908.0	12460.0	14954.0	12551.0	5909.0

Nous pouvons maintenant réaliser notre visualisation :



<u>Interprétation</u> :

1. Les ventes de ces 5 produits ont en commun un pic en novembre 2011 notamment pour le produit 23084.

2. La plupart de ces produits ne se vendaient pas, surtout durant les mois du début de l'année 2011. Par exemple , le produit 84826 avait zéro vente depuis février jusqu'en octobre. Et c'est à partir d'octobre que le nombre de ventes (pour ce produit) s'est soudainement monté en flèche (environ 12200 ventes).

Conclusion :

L'analyse des tendances et des changement dans la popularité des produits aide à la fois à **comprendre ce que vos clients aiment et achètent le plus**, mais aussi elle permet **d'adapter** et **de personnaliser** vos stratégies marketings**.**

I.2 Système de recommandation.

Introduction :

Dans une étude menée par **Salesforce**, il est scientifiquement prouvé que les clients qui reçoivent des recommandations de produits personnalisés apportent à l'entreprise plus de 24% du chiffre provenant des commandes, et 26% provenant des revenus. Ce qui montre **l'importance du système de recommandations de produits sur le volume des commandes et du chiffre d'affaire global.**

Cette étude montre aussi que les recommandations de produits conduisent à des achats répétitifs, de valeur moyenne toujours élevée, et que les clients achètent plus les produits qui leurs sont recommandés que les produits qui ne leurs sont pas recommandés.

Un système de recommandation de produits est un système dont le but est de **prédire et de compiler une liste de produits qu'un client est susceptible d'acheter.** Il existe deux façons de produire une liste de recommandation:

1. Le filtrage collaboratif
2. Le filtrage basé sur le contenu.

Le filtrage collaboratif :

Un client A qui a acheté le produit **p1** et le produit **p2** est **similaire** au **client B** si ce **client B** a acheté au moins le produit **p1** : on peut alors recommander le produit **p2 au client B.** Notez qu'il y a deux types de recommandation derrière le filtrage collaboratif :

1. Une recommandation des produits par filtrage collaboratif basé sur les clients ;
2. Une recommandation des produits par filtrage collaboratif basé sur les produits

Le filtrage basé sur le contenu

Un produit p1 et un produit p2 sont différents mais ont des caractéristiques similaires. Il se trouve qu'un client C a acheté le produit p1 dans le passé : il y a donc de fortes chances que ce client C achète l'autre produit p2 si nous le lui recommandons. Commençons l'étude sur les recommandations des produits par filtrage collaboratif.

Le filtrage collaboratif basé sur les clients

1. **La matrice client-client :**

La recherche de cette matrice client-client nous permettra de trouver les clients similaires. Pour y arriver, nous devons d'abord construire une autre matrice appelée matrice **client-produit que voici :**

1.1 matrice Clients-produits :

StockCode	10002	10080	10120	10123C	10124A	10124G	10125	10133
CustomerID								
12346.0	-0.041239	-0.043852	-0.049611	-0.022999	-0.032059	-0.03036	-0.038379	-0.104173
12347.0	-0.041239	-0.043852	-0.049611	-0.022999	-0.032059	-0.03036	-0.038379	-0.104173
12348.0	-0.041239	-0.043852	-0.049611	-0.022999	-0.032059	-0.03036	-0.038379	-0.104173
12349.0	-0.041239	-0.043852	-0.049611	-0.022999	-0.032059	-0.03036	-0.038379	-0.104173
12350.0	-0.041239	-0.043852	-0.049611	-0.022999	-0.032059	-0.03036	-0.038379	-0.104173
12352.0	-0.041239	-0.043852	-0.049611	-0.022999	-0.032059	-0.03036	-0.038379	-0.104173
12353.0	-0.041239	-0.043852	-0.049611	-0.022999	-0.032059	-0.03036	-0.038379	-0.104173
12354.0	-0.041239	-0.043852	-0.049611	-0.022999	-0.032059	-0.03036	-0.038379	-0.104173

Les valeurs à l'intérieur de cette matrice clients-produit sont négatives car elles sont normalisées. Comme vous pouvez le voir, les lignes portent **les identifiants des clients** tandis que les colonnes portent les **code de stock des produits.** D'où le nom matrice clients-produits. Là maintenant, grâce à l'algorithme de cosinus par similarité, nous pouvons construire la matrice clients-clients :

1.**2 Matrice Clients-clients :**

CustomerID	12346.0	12347.0	12348.0	12349.0	12350.0	12352.0	12353.0	12354.0
CustomerID								
12346.0	1.000000	-0.003806	-0.000439	0.001089	0.004674	0.001825	0.030228	0.005026
12347.0	-0.003806	1.000000	0.007184	0.000229	0.004157	-0.014313	-0.038047	-0.014385
12348.0	-0.000439	0.007184	1.000000	-0.007937	-0.002753	-0.005959	-0.003692	0.000149
12349.0	0.001089	0.000229	-0.007937	1.000000	-0.001940	0.061443	0.013004	-0.006710
12350.0	0.004674	0.004157	-0.002753	-0.001940	1.000000	-0.001261	0.048835	0.004990
12352.0	0.001825	-0.014313	-0.005959	0.061443	-0.001261	1.000000	0.020616	0.000275
12353.0	0.030228	-0.038047	-0.003692	0.013004	0.048835	0.020616	1.000000	0.053087
12354.0	0.005026	-0.014385	0.000149	-0.006710	0.004990	0.000275	0.053087	1.000000

Le voici donc notre matrice clients-clients (clients en lignes et en colonnes). Nous pouvons maintenant choisir n'importe client de référence que nous voulons dans cette matrice et nous servir à trouver les clients qui lui sont similaire, c-à-dire les clients qui ont acheté certains produits communs à ceux achetés par le client de référence. Choisissons donc comme **client de référence le client dont l'identifiant est 12350.0.** Cherchons donc 10 clients similaires à ce clients de référence :

1.1 Clients similaires au client 12350.0

```
CustomerID
12350.0    1.000000
17041.0    0.614465
12538.0    0.561272
12560.0    0.550065
18209.0    0.518250
12758.0    0.292034
14974.0    0.262869
13189.0    0.249993
14107.0    0.227253
13028.0    0.181192
Name: 12350.0, dtype: float64
```

Plus un client de cette série est proche du client 12350.0, plus il est similaire à lui. Ainsi le client 17041.0 est le plus similaire du client de référence 12350.0 (ce sont les clients de la série qui ont acheté le plus de produits communs)

1.1.2 Les produits à recommander à notre client

- **Les produits achetés par le clients 17410.0**

```
produits_achtee_par_17041
```

```
{'10002',
 '10080',
 '10120',
 '10123C',
 '10124A',
 '10124G',
 '10133',
 '15056P',
 '15058A'}
```

- **Les produits achetés par le client de référence 12350.0 :**

```
produits_achtee_par_12350
```

```
{'10002', '10080', '10123C', '10124A', '10125', '10135', '15036', '15056P'}
```

Nous remarquons que tous les produits achetés par le client de référence (12350.0) sont aussi achetés par le client 17410.0 : **ils sont les produits communs**. Par exemple, le produits 10002 est acheté à la fois par le client de référence 12350.0 et par le client 17410.0.

Par contre, certains produits achetés par le clients 17410.0 ne sont pas acheté par le client de référence : ce sont les produits susceptibles d'être appréciés par le client de référence (12350.0) et qu'il faut les lui recommander. Voici donc les produits à recommander au client de référence 12350.0 :

```
produit_recommander_a_12350
```

```
{'10120', '10124G', '10133', '15058A'}
```

Bien entendu, nous ne connaissons pas exactement ces produits ; nous connaissons uniquement leurs codes de stock. Cherchons alors les noms de ces produits que le client de référence n'a pas acheté mais peut acheter :

```
noms_produits_recommandes_a_12350
```

StockCode	Description
10133	COLOURING PENCILS BROWN TUBE
10124G	ARMY CAMO BOOKCOVER TAPE
10120	DOGGY RUBBER
15058A	BLUE POLKADOT GARDEN PARASOL

Et voilà les noms des produits à recommander au client 12350.

Stratégie marketing à mettre en place : un marketing personnalisé.

1. Vous connaissez que le client 12350.0 n'a jamais acheté ces produits,
2. Mais vous savez aussi qu'il y a une grande probabilité qu'il achète ces produit
3. Ces deux informations te permettront de cibler uniquement ce client et de continuer à lui proposer un en un ces produit pour le convertir . Et voilà !

<u>Filtrage collaboratif basé sur les produits (articles commerciaux)</u>

Dans le cas du filtrage collaboratif basé sur les utilisateurs (ou clients), nous avons d'abord construit une matrice clients-produits pour obtenir la matrice clients-clients. Ce que nous allons faire ici, c'est de construire une matrice produits-clients (la transposée de la matrice clients-produits) pour obtenir une matrice produits-produits. C'est cette matrice produits-produits qui va nous permettre de découvrir les produits tendances à recommander.

1. **Matrice produits-clients .**

Cette matrice n'est rien d'autre que la transposée de la patrice clients-produits déjà construite ci-haut.

```
[49]: matrice_produits_clients = matrice_client_produit.T
      matrice_produits_clients.iloc[0:8,0:8]
```

[49]:

CustomerID	12346.0	12347.0	12348.0	12349.0	12350.0	12352.0	12353.0	12354.0
StockCode								
10002	-0.041239	-0.041239	-0.041239	-0.041239	-0.041239	-0.041239	-0.041239	-0.041239
10080	-0.043852	-0.043852	-0.043852	-0.043852	-0.043852	-0.043852	-0.043852	-0.043852
10120	-0.049611	-0.049611	-0.049611	-0.049611	-0.049611	-0.049611	-0.049611	-0.049611
10123C	-0.022999	-0.022999	-0.022999	-0.022999	-0.022999	-0.022999	-0.022999	-0.022999
10124A	-0.032059	-0.032059	-0.032059	-0.032059	-0.032059	-0.032059	-0.032059	-0.032059
10124G	-0.030360	-0.030360	-0.030360	-0.030360	-0.030360	-0.030360	-0.030360	-0.030360
10125	-0.038379	-0.038379	-0.038379	-0.038379	-0.038379	-0.038379	-0.038379	-0.038379
10133	-0.104173	-0.104173	-0.104173	-0.104173	-0.104173	-0.104173	-0.104173	-0.104173

Là, nous pouvons maintenant construire notre matrice produits-produits avec laquelle nous pouvons identifier les produits similaires.

2. **Matrice produits-produits :**

```
matrice_produit_produit.iloc[0:8,0:8]
```

StockCode	10002	10080	10120	10123C	10124A	10124G	10125	10133
StockCode								
10002	1.000000	-0.001809	-0.000389	0.000042	-0.001322	-0.001252	0.854750	0.048114
10080	-0.001809	1.000000	-0.002176	-0.001009	-0.001406	-0.001332	0.003289	0.016209
10120	-0.000389	-0.002176	1.000000	0.004105	-0.001591	-0.001507	-0.000190	0.040581
10123C	0.000042	-0.001009	0.004105	1.000000	-0.000737	-0.000698	0.003544	-0.002396
10124A	-0.001322	-0.001406	-0.001591	-0.000737	1.000000	0.491290	-0.000129	0.011715
10124G	-0.001252	-0.001332	-0.001507	-0.000698	0.491290	1.000000	0.002763	0.010256
10125	0.854750	0.003289	-0.000190	0.003544	-0.000129	0.002763	1.000000	0.007717
10133	0.048114	0.016209	0.040581	-0.002396	0.011715	0.010256	0.007717	1.000000

Nous pouvons maintenant choisir n'importe quel code de stock de référence au hasard dans cette Matrice et nous s'en servir pour identifier les produits similaires à ce produit de référence.

3. **Top10 des produits similaires au produits de référence 23166 (nous ne voyons pas ce produit dans cette matrice mais il est bien là).**

```
to_10_produit_si

['23166',
 '23165',
 '22985',
 '22652',
 '21984',
 '23564',
 '23566',
 '21531',
 '21980',
 '21844']
```

Et voilà, la liste de codes de stock des produits similaires au produit de référence « 3166» dont voici leurs noms :

StockCode	Description
23166	MEDIUM CERAMIC TOP STORAGE JAR
23165	LARGE CERAMIC TOP STORAGE JAR
22985	WRAP, BILLBOARD FONTS DESIGN
22985	WRAP BILLBOARD FONTS DESIGN
22652	TRAVEL SEWING KIT
21984	PACK OF 12 PINK PAISLEY TISSUES
23564	EGG CUP MILKMAID INGRID
23566	EGG CUP MILKMAID HEIDI
21531	RED RETROSPOT SUGAR JAM BOWL
21980	PACK OF 12 RED RETROSPOT TISSUES
21844	RED RETROSPOT MUG

Nous pouvons recommander tous ces produits à un client donné qui aura effectué un achat d'au moins un de ces produits.

Conclusion :

Avec le filtrage basé sur les utilisateurs, on passe par l'identification de certains utilisateurs similaires à partir desquels on identifie les produits qu'ils peuvent acheter. Mais avec le filtrage basé sur les articles(produits) on passe par l'identification de certains produits similaires à recommander à tout client qui aura acheté au moins un d'eux.

Projet 4 : Analyse des cohorte.

Objectifs :

a. Nombre de clients chronologiquement actifs par cohorte,
b. Taux de rétention chronologique par cohorte des clients actifs et
c. Taux de désabonnement chronologique par cohorte

Introduction :

L'analyse de cohorte permet, comme son nom l'indique, de comprendre et de suivre par cohorte de clients les périodes d'activité (les achats ici) et les périodes d'inactivité (comme la tendance aux désabonnements, l'arrêt des achats etc.). Elle permet aussi de déduire les taux de rétention et de désabonnement des clients dans le temps…
Nous allons diviser les données en différentes cohortes temporelles de clients pour mieux identifier leurs niveaux d'activité commerciales (leurs comportements d'achats).

4.1 Les mois d'achats de chaque client.

1-Au niveau de chaque date de commande d'un produit donné par un client, il y a l'année, le mois et le nombre de jours comptés à partir du 1er jour de commande du produit par ce client.

2- De cette date, on va exclure les jours et ne garder qu'un seul. On crée ainsi une date ne contenant uniquement que l'année et les mois des commandes.

3-Comme l'étude des commandes est souvent choisie suivant le début et la fin d'une année (cela n'est pas obligatoire), l'année ne va pas changer: seule la distributions suivant les mois de l'année va se faire. C'est pour cela qu'on parle « dates mensuelles des commandes» ou «Mois d'achats». Voici un exemple des mois d'achats :

```
MoisAchat
2011-11-01
2011-10-01
2011-09-01
2011-05-01
2011-06-01
2011-03-01
2011-08-01
2011-07-01
2010-12-01
2011-04-01
2011-01-01
2011-02-01
2011-12-01
dtype: int64
```

L'année 2010 est là car l'étude a débuté en décembre 2010. Ainsi, depuis janvier 2011, seuls les mois varient.

Notez que chacune de ces date est répétée un certain nombre de fois car chacune correspond à plusieurs transactions :

```
MoisAchat
2011-11-01    65598
2011-10-01    50695
2011-09-01    40822
2011-05-01    28908
2011-06-01    27836
2011-03-01    27822
2011-08-01    27662
2011-07-01    27502
2010-12-01    26850
2011-04-01    23198
2011-01-01    21912
2011-02-01    20363
2011-12-01    17661
dtype: int64
```

4.2 **Cohortes mensuelles des achats.**

Il peut se trouver qu'un client donnée ait effectué plusieurs commandes dans certains mois de cette année: il a alors son ancien mois de commande et son dernier mois de commande. La date de cohorte d'un client donnée c'est celle correspondant au mois de sa premier commande de cette année d'étude: Si on considère tous les clients, on a alors tous les anciens mois de commande des clients pour cette année-là. On alors la «série» suivante appelée Mois de cohorte :

```
MoisCohort
2010-12-01    177272
2011-01-01     49047
2011-03-01     33646
2011-02-01     30136
2011-04-01     19547
2011-10-01     16428
2011-05-01     15607
2011-09-01     14419
2011-06-01     13686
2011-08-01     12987
2011-11-01     12401
2011-07-01     10657
2011-12-01       996
dtype: int64
```

Oui vous avez raison ; ce sont les mêmes dates que celles des mois d'achats que nous venons de calculer juste en haut. Et je paris que vous êtes entrains de vous demander : mais où est la différence alors, entre «mois de cohorte» et «mois d'achats» ?

Je vais vous expliquer : comme nous l'avons déjà dit, les mois d'achats sont tous les mois de l'année d'étude qu'un client donnée a effectué des achats (commandes) . Le premier mois d'entre eux pour ce client est un mois de cohorte : ce mois apparait donc à la fois dans les mois d'achats et dans les mois des cohortes. La différence ne se situe pas au niveau des mois mais bien au niveau du nombre de répétition des dates.

Exemple :

Mois d'achat :

En décembre 2010 il a eu plus de 65598 commandes (on ne cherche pas à savoir qui a effectué ces commandes; on cherche plutôt le nombre de commandes effectués en décembre de cette date)

Mois de cohorte :

Plus de 177272 clients ont effectué leurs premiers achats en décembre 2010 : la cohorte de décembre 2010.
Seuls 996 clients ont effectué leurs premières commandes en décembre 2011 (la cohorte de décembre 2011).

4.3 Les différentes périodes mensuelles des cohortes d'achat

On extrait ensuite:

1-L'année d'achat, puis le mois d'achat au niveau des dates d'achats d'une part,

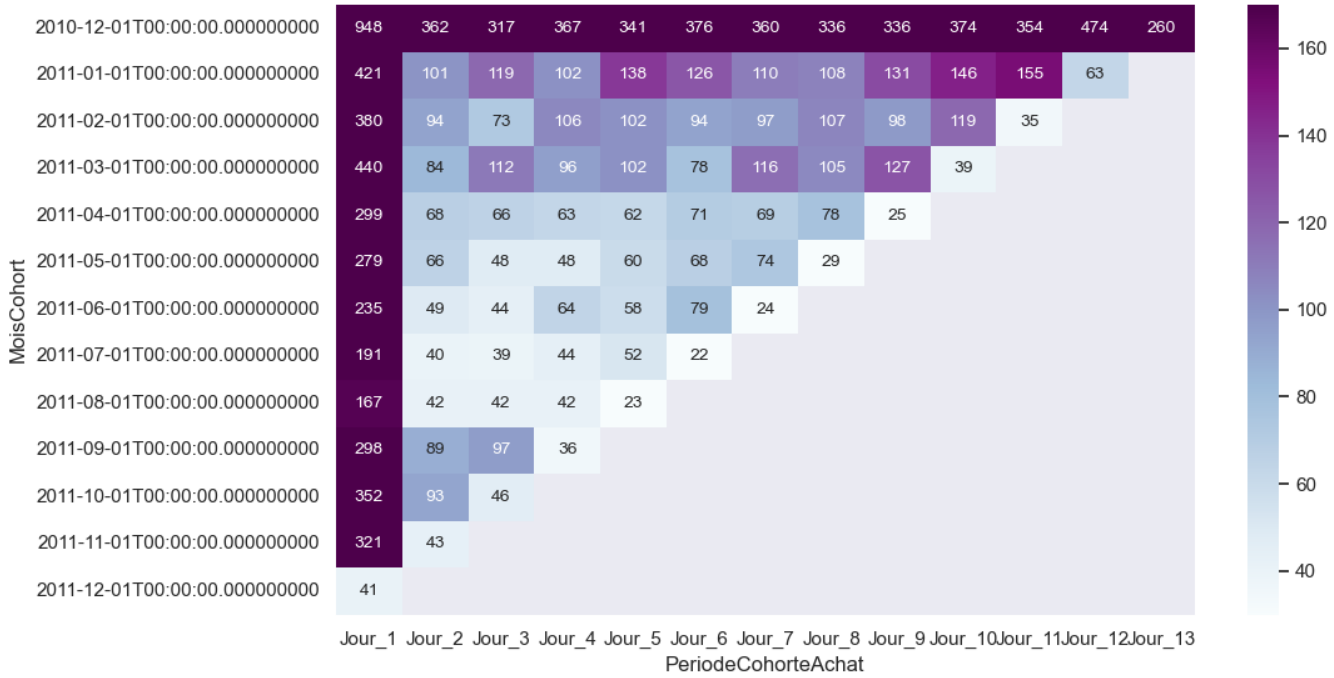
2-L'année de cohorte, puis le mois de cohorte au niveau des date de cohortes d'autre part.

3- On crée à partir de ces années et mois les indices des différentes périodes d'activités (commandes) de chaque cohorte de clients :

```
[193]: Jour_1     119191
       Jour_2      29147
       Jour_4      28075
       Jour_6      27576
       Jour_3      27493
       Jour_5      25926
       Jour_12     24520
       Jour_7      24214
       Jour_8      24200
       Jour_10     24075
       Jour_9      23672
       Jour_11     21331
       Jour_13      7409
       Name: PeriodeCohorteAchat, dtype: int64
```

4.4 Cohortes des clients actifs.

Maintenant, avec la série des mois d'achats(MoisAchats), la série des mois de cohortes (MoisCohorte) et la série des indices de chaque période d'activité de chaque cohorte (PeriodChorteAchat), nous pouvons créer le visuel représentant les cohortes. Le voici :



Interprétation :

Chaque ligne de ce visuel représente une cohorte de clients. Chacune de ces cohortes est repéré par une date (à gauche). Chaque colonne correspond à un jour de cette date de cohorte.

1- En décembre 2010 par exemple(1ère ligne), plus de 948 clients se sont inscrits dès le premier jour (colonne 1). Au jour suivant de la même cohorte, le nombres de commandes a baissé à 362 achats, puis à 317 commandes le 3ème jour, puis il a monté à 367 commandes au 4ème jour. À la fin du mois de cette cohorte, il ne restait que 260 clients acheteurs (260 achats).

2-Ainsi, suivant chaque date de cohorte est associé la cohortes décrivant les comportements d'achats journaliers durant cette date. Selon l'objectif marketing, vous choisissez la cohorte que vous voulez, puis vous identifier le jour où il a eu le plus de ventes pour l'exploiter. Dans la partie suivante, nous allons calculer les taux de rétentions de ces clients actifs dans chaque cohorte.

4.6 Taux de rétention des clients actifs :

Le taux de rétention c'est le rapport entre le nombre de clients actifs de chaque période d'activité d'une cohorte et le nombre de clients inscrits au premier jour de cette cohorte, multiplié par 100

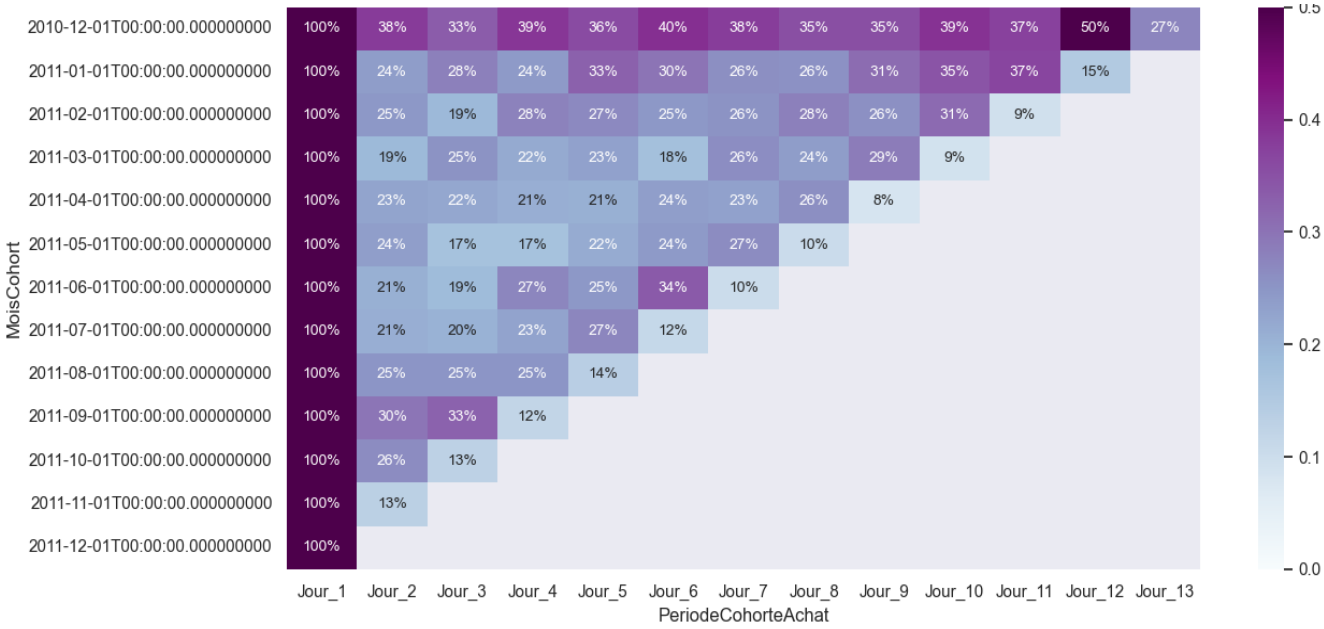
Exemple :
Taux de rétention de la période « jour_2 » de la cohorte de décembre 2010 :

- Nombre de clients inscrits le premier jour (jour_1) de cette cohorte = 968 clients
- Nombre de clients actifs (ayant effectué des achats) à la période suivante (jour_2) de cette cohorte = 362.

T = 362 / 948 * 100

T= 38.1 % : taux de rétention des clients dans la période «jour_2» de la première cohorte.

Voici le visuel des taux de rétention des clients actifs :



Interprétation :

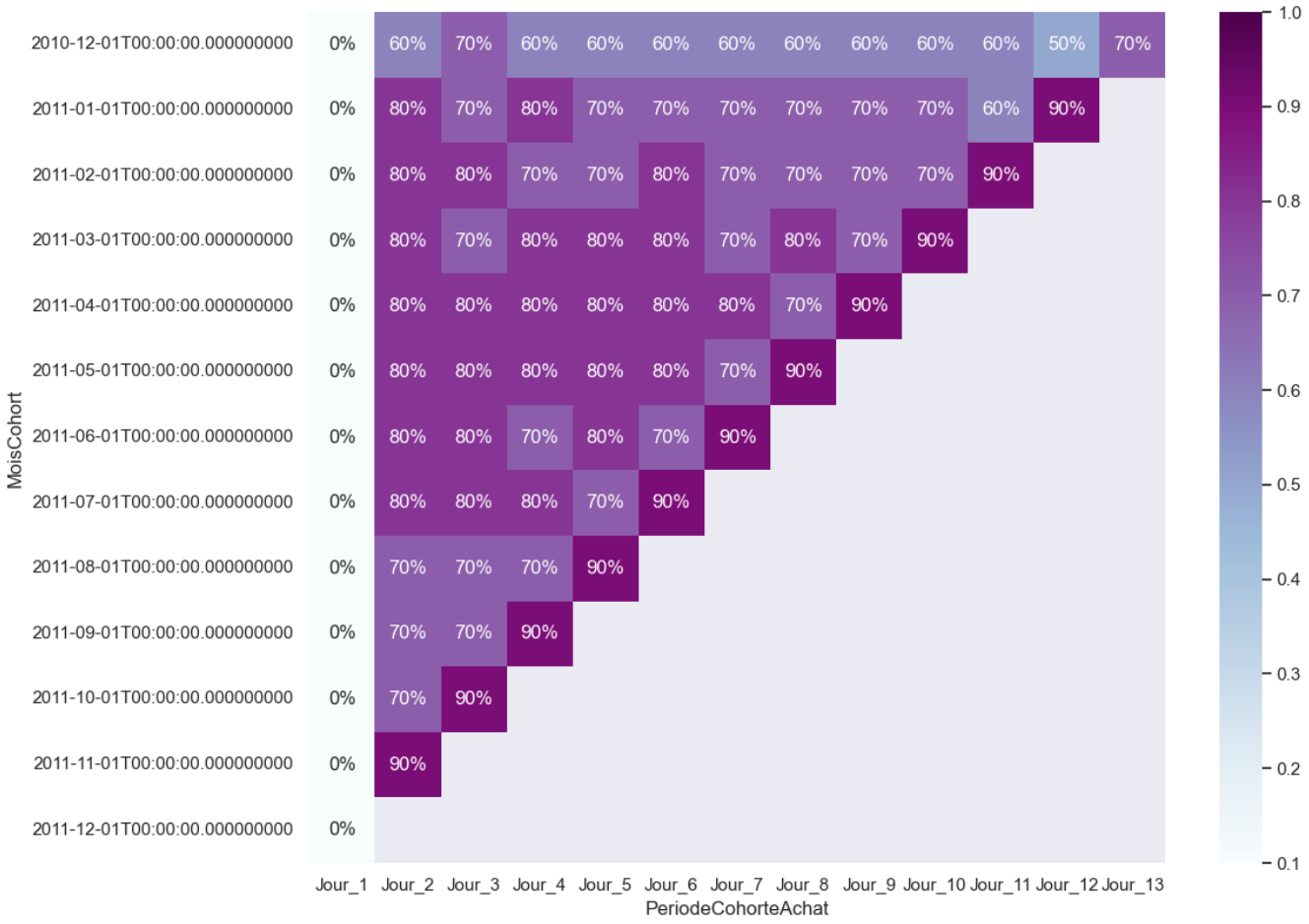
Ce visuel montre le degré d'activité(la force de rétention) des clients actifs: plus le pourcentage est élevé, plus les clients relatifs à ce taux d'activité sont actifs. Par exemple, au niveau de la cohorte de décembre 2010, les clients qui composent cette cohorte sont très actifs au 12ème jour de cette date(taux de rétention = 50.0 %.)

Et plus ce taux est moins élevé, plus les client tendent à se désabonner. Par exemple, vous remarquerez que la majorité des taux de rétentions les plus bas se trouve au dernier jour d'activité (dans la diagonal qui sépare la zone d'activité et la zone d'inactivité de chaque cohorte): le deuxième triangle où il n'y a rien correspond la zone d'inactivité ou de désabonnement. Nous allons calculer les taux de désabonnement de cette zone:

4.7. Taux de désabonnement(ou d'inactivité) des clients inactifs dans chaque mois de l'année de chaque cohorte

Introduction :

Le taux de désabonnement (ou d'inactivité) est la différence entre 100% des clients actifs (les clients inscrits au premier jour) et le taux de rétention des clients actifs.



Interprétation :

Les jours de chaque cohorte où le taux de désabonnement est très élevé ce sont des jours où les clients sont très inactifs ou carrément désabonnés.

Conclusion :

Voilà que se termine notre projet d'analyse de cohorte chronologique des clients de notre entreprise. Avec cette analyse nous pouvons découvrir les comportements journaliers passés de nos clients par mois de l'année d'activités commerciales de notre entreprise. Comprendre les tendances temporelles actives et inactives de ces clients. Ceci est important lorsque l'on veut avoir un œil sur la santé de l'entreprise, les tendances déficitaires et la force concurrentielle…

Projet 5 :

Segmentation des clients suivant leurs récences, leurs fréquence d'achat et leurs valeurs monétaires dépensées.

Introduction :

La segmentation RFM permet de grouper les clients selon leurs activités récente(récence), leur fréquence d'activité(Fréquence d'achats ici) et le montant mensuel, journalier ou annuel qu'ils dépensent. Cela donne:

- Ceux qui ont effectué des achats le plus récemment possible et donc actifs(Récence), qui répètent les achats après le premier(Fréquence) et qui achètent les produits chers
- Ceux qui ont effectué des achats le plus récemment possible, qui répètent les achats mais qui achètent les produit moins chers
- Ceux qui ont effectué des achats le plus récemment possible, mais qui n'achètent qu'une, 2 ou 3 fois et qui achètent le produit cher
- Ceux qui ont effectué des achats le plus récemment possible, mais qui n'achètent qu'une, 2 ou 3 fois et qui achètent le produit moins cher
- Ceux dont le dernier achat est très ancien (clients inactifs), mais qui ont effectué plusieurs commandes, sur les produits moins cher (ou plus chers) etc.

On peut encore continuer à reformuler d'autres types de groupes suivant la même logique de récence-fréquence-monétaire en fonction des objectifs du magasin ou entreprise, comme vous pouvez le deviner.

5.1 Calcule de la récence

5.1.1. La date d'analyse.

La date d'analyse est la date de la dernière commande réalisé durant toute l'année d'activité commercial de l'entreprise. De cette date. Nous ajourons 1 jour de plus pour un but jute technique :

```
dernier_date_achat = rfm['InvoiceDate'].max() + dt.timedelta(days=1)
print(dernier_date_achat)
```

```
2011-12-10 12:50:00
```

Notre date d'analyse c'est donc le 10 décembre 2011.

5.1.2. Date de la dernier commande de chaque client.

Après avoir calculer la date de la dernière commande (la date d'analyse), ce qu'il faut faire c'est de calculer la dernier date de commande de chaque client. Leur différence : la date d'analyse ne s'exprime pas en fonction de chaque client mais bien en fonction du dernier client acheteur seulement.

```
pd.DataFrame(rfm.groupby(['CustomerID'])['InvoiceDate'].max())
```

CustomerID	InvoiceDate
12346.0	2011-01-18 10:01:00
12347.0	2011-12-07 15:52:00
12348.0	2011-09-25 13:13:00
12349.0	2011-11-21 09:51:00
12350.0	2011-02-02 16:01:00
...	...
18280.0	2011-03-07 09:52:00

5.1.3 La récence :

Avec ces deux date nous pouvons calculer la récence : c'est le nombre de jour écoulés entre la date de la dernière commande (date d'analyse) et la date de la dernier commande de chaque client. En d'autres termes, c'est le temps(nombre des mois ici) écoulé depuis la dernière commande d'un client donné. Soit donc:

CustomerID	Recency
12346.0	326
12347.0	2
12348.0	75
12349.0	19
12350.0	310
...	...
18280.0	278
18281.0	181
18282.0	8
18283.0	4
18287.0	43

4338 rows × 1 columns

Interprétation :

La dernier commande du client 1236.0 remonte à 326 jours tandis que la dernier commande du client 12347.0 ne remonte qu'à seulement 2 jour.

Ainsi, le client 1236.0 dont sa commande remonte à 326 jour est donc moins actif que le client 12347.0 dont sa dernière commande ne remonte qu'à 2 jour.

Le moins actif semble donc moins intéressé et risque de basculer vers un autre concurrent (se désabonner) alors que le client actif reste intéressé à votre offre.

5.2 La fréquence de commande de chaque client.

La fréquence d'achat d'un client donnée c'est le nombre de répétition des commande : quand un client a effectué une commande, il n'a pas encore réalisé une répétition. Dans ce cas sa fréquence d'achat est égale à zéro.
Mais quand il achète à nouveau une deuxième fois, ce «deuxième fois» constitue sa première répétition : sa fréquence d'achat est alors égale à 1. La fréquence d'achat d'un client c'est donc son nombre des commandes moins 1 (dans les partie qui vont suivre nous négligerons cette différence entre fréquence et nombre d'achats).

Cette variable est importante car elle permet de comprendre le lien entre les montants dépensés et la fréquence des dépenses. D'après vous , qui est le meilleur clients dans ces deux cas : celui qui

achète plusieurs fois des produits moins cher ou bien celui qui achète une ou 2 fois mais qui dépense des montants colossaux ? vous avez compris l'idée.

```
[599]: pd.DataFrame(rfm.groupby([
```

[599]:

CustomerID	Frequency
12346.0	1
12347.0	182
12348.0	31
12349.0	73
12350.0	17
...	...
18280.0	10
18281.0	7
18282.0	12
18283.0	756
18287.0	70

4338 rows × 1 columns

Interprétation :

Dans l'analyse de récence faite juste en haut, nous avons vu que le client 12346.0 n'était pas un bon client car sa dernière commande remontait à 326 jour. Eh bien d'après sa fréquence de commande nous voyons qu'il n'a effectué qu'une commande (fréquence de commandes =1 fois).

Par contre, le client 12347.0 dont sa dernière commande ne remonte qu'à seulement 2 jours a déjà réalisé plus de 182 commandes avant (fréquence de commande = 182 fois). C'est clairement un client à ne jamais laisser partir.

5.4 Montant total dépensé de chaque client :

C'est l'argent total dépensé dans toutes les commandes réalisées par un client.

CustomerID	Montant
12346.0	77183.60
12347.0	4310.00
12348.0	1797.24
12349.0	1757.55
12350.0	334.40
...	...
18280.0	180.60
18281.0	80.82
18282.0	178.05
18283.0	2094.88
18287.0	1837.28

4338 rows × 1 columns

Interprétation :

Revenons encore à nos deux clients étudiés précédemment :
Le clients 12346.0 dont sa dernière commande remonte è 328 jours a dépensé plus de 77183.6 € dans sa commandes réalisé. Ainsi, bien que sa récence et sa fréquence d'achat lui aient collé le statut de mauvais client, il a quand même dépensé un montant d'argent colossal

Par contre, notre meilleur client (d'après sa récence et sa fréquence) dont sa dernière commande remonte à seulement 2 jour n'a dépensé que 4310 €. Le bon client n'est pas forcément celui qui vous a donné le plus d'argent mais bien celui qui continue d'acheter vos produits et restant fidèle malgré son ancienneté. Rassemblons ces trois séries en une table unique :

5.5 <u>Résumé</u>.

CustomerID	Recency	Frequency	Monetary
12346.0	326	1	77183.60
12347.0	2	182	4310.00
12348.0	75	31	1797.24
12349.0	19	73	1757.55
12350.0	310	17	334.40
...	...	...	...
18280.0	278	10	180.60
18281.0	181	7	80.82
18282.0	8	12	178.05
18283.0	4	756	2094.88
18287.0	43	70	1837.28

4338 rows × 3 columns

5.6 Discrétisation et scoring des Récences, Fréquences et Monétaires

5.6.1 Discrétisation :

Les caractéristiques récence, fréquence et monétaire de nos clients sont des variables continues. Nous pouvons donc imaginer qu'elles ont des valeurs maximales, minimales médianes, moyennes, leurs 3èmes, 2èmes et premiers quartiles. On peut donc les attribuer des numéros pour les classer : 4 pour le segment des valeurs de nos variables comprises entre leurs maximums et leurs 3ème quartiles, 3 pour le segment des valeurs comprises entre leurs 3ème et leurs 2ème quartiles, 2 pour le segment des valeurs comprises entre leurs 2èmes et leurs 1ere quartiles et 1 pour le segments des valeurs comprises entre leurs 1ers quartiles et 1 leurs valeurs minimales. On crée alors des segments de nos variables :

Cette table statistique va vous aider à comprendre :

:

	Recency	Frequency	Monetary
count	4.338000e+03	4.338000e+03	4.338000e+03
mean	-1.027980e-16	3.672591e-18	-8.022628e-16
std	1.000115e+00	1.000115e+00	1.000115e+00
min	-2.630445e+00	-2.775160e+00	-4.179280e+00
25%	-6.124235e-01	-6.384332e-01	-6.841832e-01
50%	1.147066e-01	2.550746e-02	-6.094235e-02
75%	8.296516e-01	6.979246e-01	6.542440e-01
max	1.505796e+00	3.988157e+00	4.721395e+00

Entre le maximum et le 3ème quartile (les 75%) se trouve le 1er segment de chacune de nos 3 variables récence, Fréquence et monétaire : nous lui donnons le numéro 4 pour la fréquence et la valeur monétaire , et 1 pour la récence . Nous avons alors le segment 144

Entre le 3ème quartile (les 75%) et le 2ème quartile (50% c'est-à-dire la valeur médiane de chaque variable) se situe le 2ème segment : nous lui attribuons 3 pour la fréquence et la valeur monétaire, et 2 pour la récence. Nous avons alors le segment 233

Entre le 2ème quartile (la médiane=50%) et le 1er quartile (les 25%) se situe le 3ème segment : nous lui attribuons le numéro 2 pour la fréquence et la valeur monétaire et 3 pour la récence. Nous avons alors le segment 322

Et entre le 1er quartile (les 25%) et la valeur minimale se situe le 4ème segment. Nous lui attribuons la valeur 1 pour la fréquence et la valeur monétaire, et 4 pour la récence. Nous avons alors le segment 411.

Rappelle : une valeur de récence élevée désigne un mauvais client alors qu'une fréquence et une valeur monétaire élevées désignent un bon client.

Voici une table qui résume ce que nous venons de faire :

CustomerID	Recency	Frequency	Monetary	R_quartile	F_quartile	M_quartile	RFM_Seg
12346.0	326	1	77183.60	1	1	4	114
12347.0	2	182	4310.00	4	4	4	444
12348.0	75	31	1797.24	2	2	4	224
12349.0	19	73	1757.55	3	3	4	334
12350.0	310	17	334.40	1	1	2	112

Interprétation :

Le client 12346.0 correspond à une pire récence (1), une pire fréquence (1) mais avec une parfaite valeur monétaire. Son segment RFM est le 114. C'est un bon client mais il n'est ni très bon ni parfait.

Le client 12347.0 correspond à une parfaite récence (4), une parfaite fréquence (4) et une parfaite valeur monétaire (4). Son segment est le 444 C'est un parfait client.

Le client 12350.0 correspond à une pire récence, à une pire fréquence mais avec une bonne valeur monétaire (juste une bonne mais pas une pire, ni une très bonne, ni même une parfaite valeur monétaire). Son segment est le 112. C'est un pire client.

Ainsi, on peut classer les clients en parfaits, très bons, bons et pires clients grâce à un principe de scoring. Faisons cela tout de suite.

5.6.2 Scoring RFM

CustomerID	Recency	Frequency	Monetary	R_quartile	F_quartile	M_quartile	RFM_Seg	RFM_Score
12346.0	326	1	77183.60	1	1	4	114	6
12347.0	2	182	4310.00	4	4	4	444	12
12348.0	75	31	1797.24	2	2	4	224	8
12349.0	19	73	1757.55	3	3	4	334	10
12350.0	310	17	334.40	1	1	2	112	4

Maintenant, nous pouvons associer à chaque intervalle de score le parfait au pire client :

```
rfm.head()
```

CustomerID	Recency	Frequency	Monetary	R_quartile	F_quartile	M_quartile	RFM_Seg	RFM_Score	Bon_Pire
12346.0	326	1	77183.60	1	1	4	114	6	3:Bon_Client
12347.0	2	182	4310.00	4	4	4	444	12	1:Parfait_Client
12348.0	75	31	1797.24	2	2	4	224	8	2:Très_Bon_Client
12349.0	19	73	1757.55	3	3	4	334	10	1:Parfait_Client
12350.0	310	17	334.40	1	1	2	112	4	4:Pire_Client

Cette table nous aide surtout à associer un client donné à son positionnement de parfait au pire client. Mais pour mieux effectuer une bonne interprétation générale, nous allons calculer les

valeurs moyennes des récences, fréquences et monétaires puis les attribuer en indexe les différentes modalités de la colonne «Bon_Pire». Soit donc:

```
rfm_analysis = rfm.groupby(['Bon_Pire']).agg({rfm1.columns[i]
rfm_analysis .head()
```

Bon_Pire	Recency	Frequency	Monetary
1:Parfait_Client	19.916996	228.290909	5248.510119
2:Très_Bon_Client	55.660161	67.895522	1380.604216
3:Bon_Client	111.732404	28.978397	638.844907
4:Pire_Client	218.268579	10.919166	199.209765

Et voilà ! On n'a même pas besoin d'effectuer des interprétations tellement elles sont évidentes :

Les parfaits clients : en moyenne, leurs dernières commandes remontent à 19.9 jours. Ils ont effectué environ 229 commandes. Et chaque client de cette catégorie dépense en moyenne 5248.5€

Les très bons clients : en moyenne, leurs dernières commandes remontent à 55.66 jours. Ils ont effectué environ 67.9 commandes. Et chaque client de cette catégorie dépense en moyenne 1380.60€

Les bons clients : en moyenne, leurs dernières commandes remontent à 111.73 jours. Ils ont effectué environ 29 commandes. Et chaque client de cette catégorie dépense en moyenne 638.84€

Les pires clients (ou clients basiques) : en moyenne, leurs dernières commandes remontent à 218.26 jours. Leur nombre de commandes moyennes est de 11 commandes. Et chaque client de cette catégorie dépense en moyenne 199.20€

5.7 Analyse des comportements d'achats par région

Connaitre la région la plus prometteuse aidera l'entreprise à stratégiquement bien se positionner ou à bien effectuer ses ciblages. Quelles sont donc ces régions ?

5.7.1 Clients totaux par pays

```
nb_pays = data.groupby(['Country']).size().to_
nb_pays.head()
```

	Country	Nombre_Clients_Totaux
0	Australia	1259
1	Austria	401
2	Bahrain	17
3	Belgium	2069
4	Brazil	32

5.7.2 Top 5 des pays performants pour les clients basiques ou pires

	Pire	Country	Nombre_Pires_Clients
85	4:Pire_Client	United Kingdom	8153
31	4:Pire_Client	France	185
35	4:Pire_Client	Germany	60
65	4:Pire_Client	Portugal	32
70	4:Pire_Client	Spain	31

5.7.3 Pourcentage de ce top5 des basique ou pires clients.

	Country	PourcentagePireClient
100	Saudi Arabia	100.00
99	Bahrain	100.00
87	Canada	10.60
61	Poland	4.99
7	Austria	4.49

Interprétation :

Les clients issus de l'Arabie Saoudite et du Bahrain sont 100% basique. 11% seulement des clients Canadiens est basique.

5.7.4 Top 7 des pays performants issus des parfaits clients.

5.7.4.1 Le nombre.

	ParfaitClient	Country	NombreParfaitClients
0	1:Parfait_Client	Australia	1064
4	1:Parfait_Client	Austria	158
8	1:Parfait_Client	Belgium	1536
12	1:Parfait_Client	Channel Islands	494
16	1:Parfait_Client	Cyprus	413
20	1:Parfait_Client	Denmark	184
23	1:Parfait_Client	EIRE	7323

5.7.4.2 Pourcentage.

	Country	PourcentageParfaitClient
66	Singapore	100.00
36	Iceland	100.00
23	EIRE	97.84
54	Norway	91.07
50	Netherlands	90.76
0	Australia	84.51
28	France	80.19

Les clients du Singapore et d'Island sont 100% des parfaits clients:

1- Leurs dernières commandes remontent à **seulement quelques jours** (récentes),
2- Ils achètent **plusieurs fois,**
3- Ils dépensent **beaucoup d'argents.**

Projet 6 :

Segmentation des clients en fonction de leurs RFM par un algorithme non-supervisé de clustering (Le K-Means).

Dans le projet 5, le regroupement des clients par récence-Fréquence-Monétaire a été fait grâce à un scoring réalisé par notre propre innervation. Ce qui représente un biais venant de nous-même.

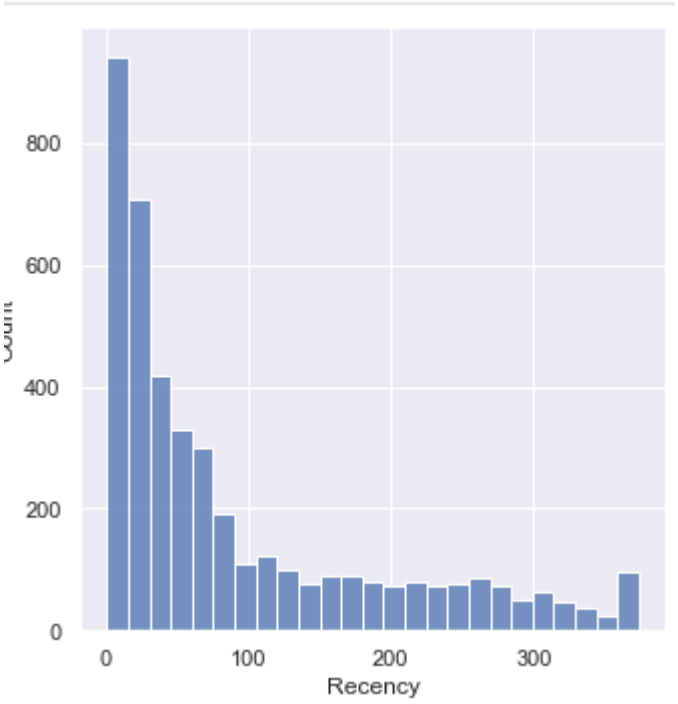
En effet, ce sont nous qui avons décidé le nombre de groupe (4 groupes allant du parait au pire client) à analyser et non pas une méthode quelconque à notre place. Ce qui n'est pas scientifiquement objectif.

Nous allons donc demander à un algorithme non-supervisé d'identifier le nombre de groupes réels de clients selon leurs récences, leurs fréquences et leurs valeurs monétaires.

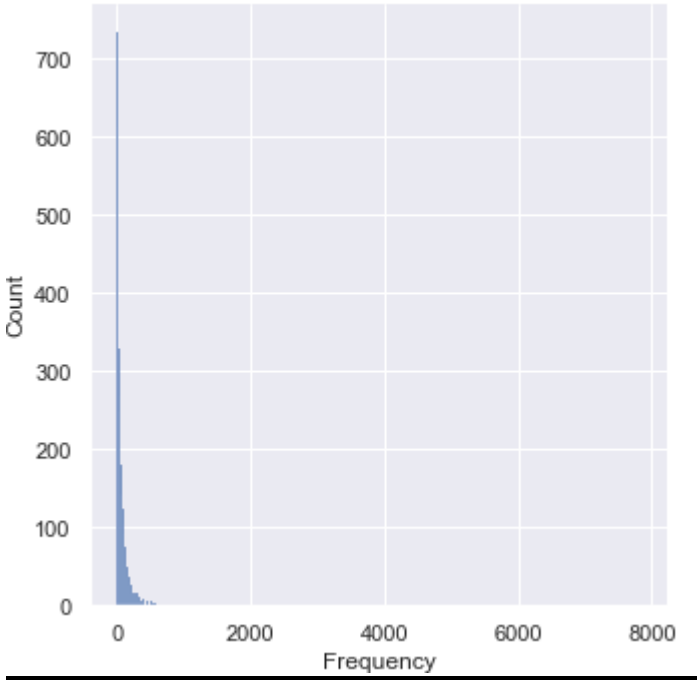
6.1 Suppression des asymétries et normalisation des données.

L'algorithme de k-means est très sensible à l'existence d'asymétrie des variables. Nous allons donc vérifier cela et tester l'égalité des moyennes: il a besoin que les moyennes soient égales et les variance aussi.
Les distributions suivantes des récences, fréquence et valeurs monétaires ne sont pas équilibrées : elles ne se comportent pas en « dos de chameau». Ce qui n'est pas bon pour notre algorithme.

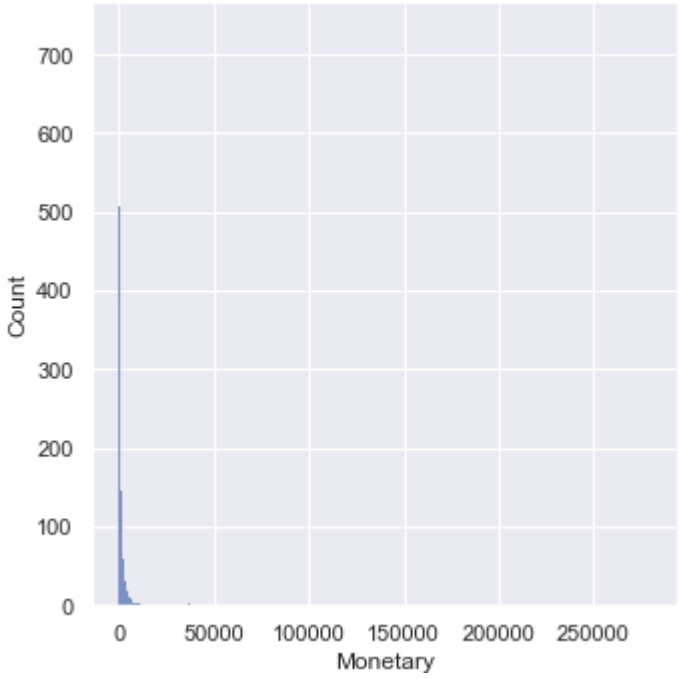
6.1.1. Récence :



6.1.2. **Fréquence.**



6.1.3 Valeurs monétaire.



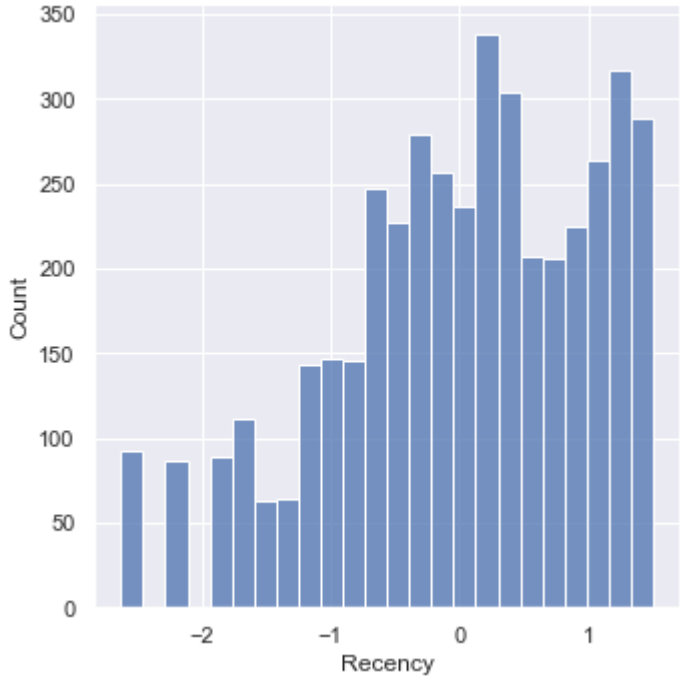
Nous allons donc utiliser fonction logarithmique pour supprimer l'asymétrie et une mise à l'échelle des données pour ramener les moyennes zéro et les variances à 1. Ce qui va normaliser la distribution des données (établissement de dos de chameau) de nos 3 variables :

```
rfm.describe()
```

	Recency	Frequency	Monetary
count	4.338000e+03	4.338000e+03	4.338000e+03
mean	-1.027980e-16	3.672591e-18	-8.022628e-16
std	1.000115e+00	1.000115e+00	1.000115e+00
min	-2.630445e+00	-2.775160e+00	-4.179280e+00
25%	-6.124235e-01	-6.384332e-01	-6.841832e-01
50%	1.147066e-01	2.550746e-02	-6.094235e-02
75%	8.296516e-01	6.979246e-01	6.542440e-01
max	1.505796e+00	3.988157e+00	4.721395e+00

Là, l'asymétrie est supprimée et nos données normalisées : les valeurs moyennes de nos 3 variables sont quasi nulles et leurs variance quasi égales à 1. Leurs distributions en bosse de chameau devraient apparaitre dans nos visuels :

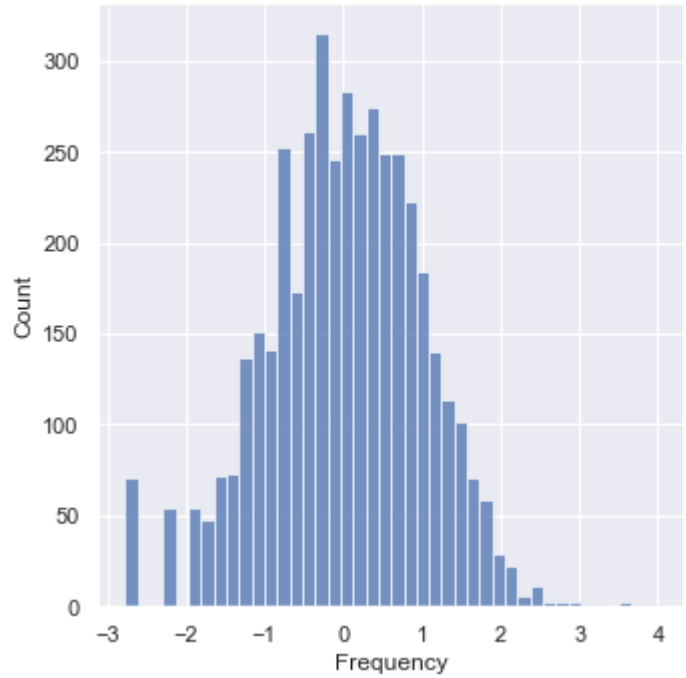
<u>Pour la Récence :</u>



Ça à l'air un peu bien amélioré mais la normalisation n'est pas top. Cependant cette amélioration reste suffisante

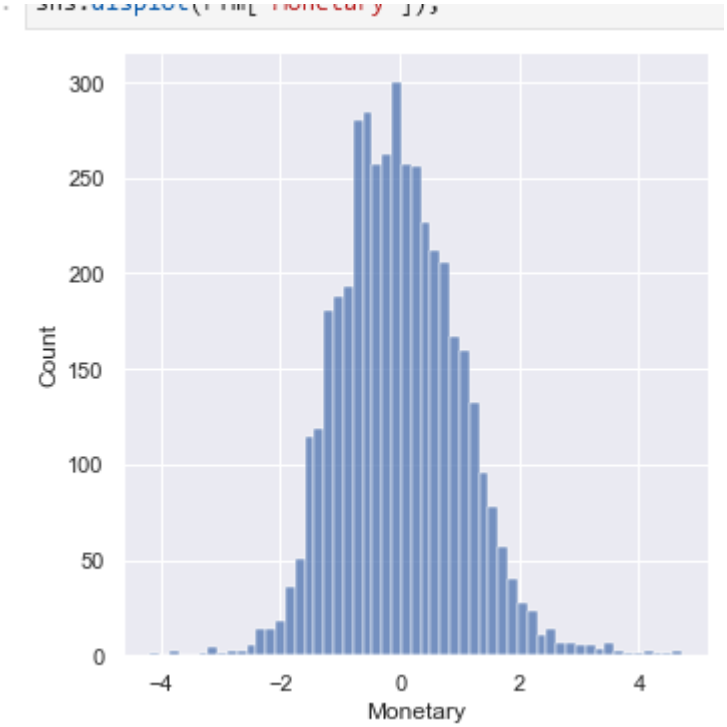
Pour la Fréquence .

```
sns.displot(rfm['Frequency']);
```



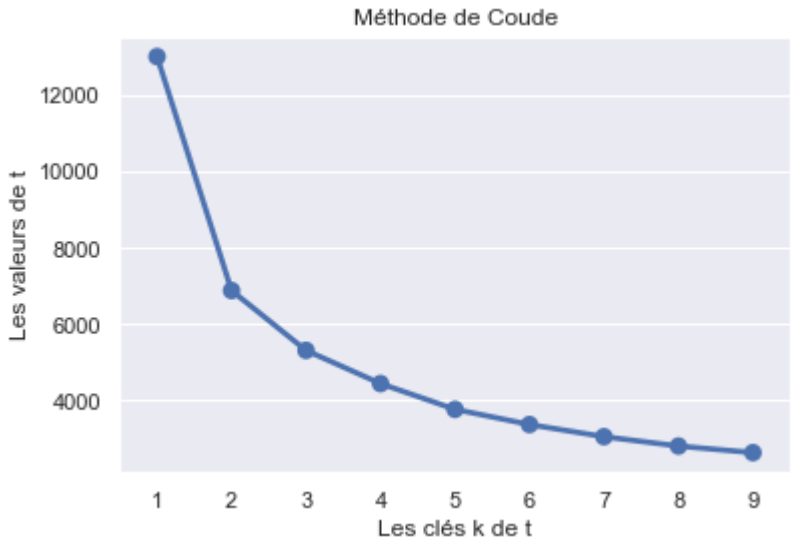
C'est super suffisant pour la normalisation de la fréquence d'achats,

Pour les valeurs monétaires :



Encore super pour les valeurs monétaires. Nous pouvons maintenant développer notre modèle de regroupement.

Etape 1 : recherche du nombre possible des groupes de clients grâce à la méthode dite de coude :



Quel nombre de directions possibles peut-on associer cette courbe :

Du point le plus haut au 2ème point, on peut tracer une première droite dessus : ce segment est donc un premier groupe.

Du 2ème au 4ème point de cette courbe on peut aussi tracer une 2ème droite non parallèle à la première droite : c'est un 2ème groupe de clients.

Du 4ème au dernier point de cette courbe on peut tracer une 3ème droite non parallèle aux deux premières : c'est un 3 groupe de clients.

Cependant depuis le 2ème point vers le dernier point (point 9) les deux segments qui sont là ne sont pas très distinctifs (ces deux groupes sont proches et peuvent en réalité ne constituer qu'un seul groupe). Une autre méthode automatique dite de silhouette va même nous le confirmer :

```
    print('silhouette Score for %i Clusters:%0.4f'%
```

```
silhouette Score for 2 Clusters:0.3951
silhouette Score for 3 Clusters:0.3032
silhouette Score for 4 Clusters:0.3026
silhouette Score for 5 Clusters:0.2784
silhouette Score for 6 Clusters:0.2761
```

Nous voyons que le score (0.3951) correspondant à 2 clusters (2 groupes de clients distincts) est le plus élevé. Ainsi, bien qu'il y ait normalement 3 groupes de clients, deux seulement sont clairement distincts. Pour des raison pratiques, nous allons considérer 3 groupes dans ce qui va suivre.

Etape 2 : Interprétation des 3 groupes :

```
]: # Remplaçons k par 3 pour voir
   kmeans = KMeans(n_clusters=3, random_state=1)
   kmeans.fit(rfm)

   kmeans.cluster_centers_
```

```
]: array([[-1.2112728 ,  1.13343002,  1.21507631],
         [ 0.72420279, -0.97798984, -0.9199018 ],
         [ 0.0216432 ,  0.22679183,  0.13739646]])
```

Cette tables array contient 3 lignes (les 3 types de groupes de clients):

1- Les clients du groupe de la première ligne (group 0) sont actifs (récence très basse), ils ont effectué plusieurs achats (Fréquence élevée) et ont dépensé beaucoup d'argent (Monétaire élevée). C'est le parfait groupe

2- Les clients du 2ème ligne (groupe 1) sont inactifs(récences élevées), ils n'ont pas effectué beaucoup de commandes (achats) ni même dépenser beaucoup d'argents: c'est le mauvais groupe de clients

3-Les clients du 3ème ligne (le groupe 2) sont actifs mais pas de façon extrême, ils font plusieurs commandes et dépensent beaucoup d'argents.

Etape 3 : Ajoutons une colonne des groupes identifiés dans notre tables rfm1 de départ

```
groupes.head()
```

CustomerID	Recency	Frequency	Monetary	Cluster
12346.0	326	1	77183.60	2
12347.0	2	182	4310.00	0
12348.0	75	31	1797.24	2
12349.0	19	73	1757.55	2
12350.0	310	17	334.40	1

6.2 Produits appréciés des clients pour chaque groupe

6.2.1 Le parfait groupe (groupe 0) :

CustomerID	Recency	Frequency	Monetary	Cluster
12347.0	2	182	4310.00	0
12357.0	33	131	6207.67	0
12358.0	2	19	1168.06	0
12359.0	58	248	6372.58	0
12362.0	3	266	5226.23	0
...	...	...	...	...
18237.0	3	61	987.10	0
18241.0	10	104	2073.09	0
18245.0	7	175	2567.06	0
18272.0	3	166	3078.58	0
18283.0	4	756	2094.88	0

951 rows × 4 columns

Interprétation :

On voit bien que dans ce groupe, les clients sont très actifs :ils dépensent beaucoup d'argent et ils ne tardent pas à revenir effectuer des achats car leurs dernières commandes ne remontent qu'à quelques jours et ils achètent plusieurs fois.

6.2.2. Top5 des produits achetés par les parfaits clients :

Description	StockCode
WHITE HANGING HEART T-LIGHT HOLDER	1190
JUMBO BAG RED RETROSPOT	1183
REGENCY CAKESTAND 3 TIER	1127
LUNCH BAG RED RETROSPOT	1014
PARTY BUNTING	835

6.3.1. Le basique ou pire groupe (groupe 1) :

CustomerID	Recency	Frequency	Monetary	Cluster
12350.0	310	17	334.40	1
12353.0	204	4	89.00	1
12355.0	214	13	459.40	1
12361.0	287	10	189.90	1
12363.0	110	23	552.00	1
...	...	...	...	...
18277.0	58	8	110.38	1
18278.0	74	9	173.90	1
18280.0	278	10	180.60	1
18281.0	181	7	80.82	1
18282.0	8	12	178.05	1

1529 rows × 4 columns

Interprétation :
Ce groupe est celui des clients basique ou pire : leurs récences moyennes (nombre de jours écoulés depuis leurs dernières commandes) sont tellement élevées. Leurs fréquences moyennes

d’achats sont les plus basses que dans les autres groupes. Et leurs dépenses moyennes sont aussi les plus basses.

6.3.2 Top7 des produits achetés par les clients basiques

Description	StockCode
WHITE HANGING HEART T-LIGHT HOLDER	177
REGENCY CAKESTAND 3 TIER	157
PARTY BUNTING	124
POSTAGE	112
ASSORTED COLOUR BIRD ORNAMENT	106
REX CASH+CARRY JUMBO SHOPPER	99
BAKING SET 9 PIECE RETROSPOT	94

6.4.1 Le bon groupe (groupe 2)

CustomerID	Recency	Frequency	Monetary	Cluster
12346.0	326	1	77183.60	2
12348.0	75	31	1797.24	2
12349.0	19	73	1757.55	2
12352.0	36	85	2506.04	2
12354.0	232	58	1079.40	2
...	...	...	...	...
18259.0	25	42	2338.60	2
18260.0	173	134	2643.20	2
18263.0	26	61	1213.16	2
18265.0	72	46	801.51	2
18287.0	43	70	1837.28	2

1858 rows × 4 columns

Interprétation :

Ce groupe se situe entre le groupe parfait et le groupe basique : les clients de ce groupe ont des récences moyennes basses que celles du groupe parfait mais élevées que celles du groupe basique.

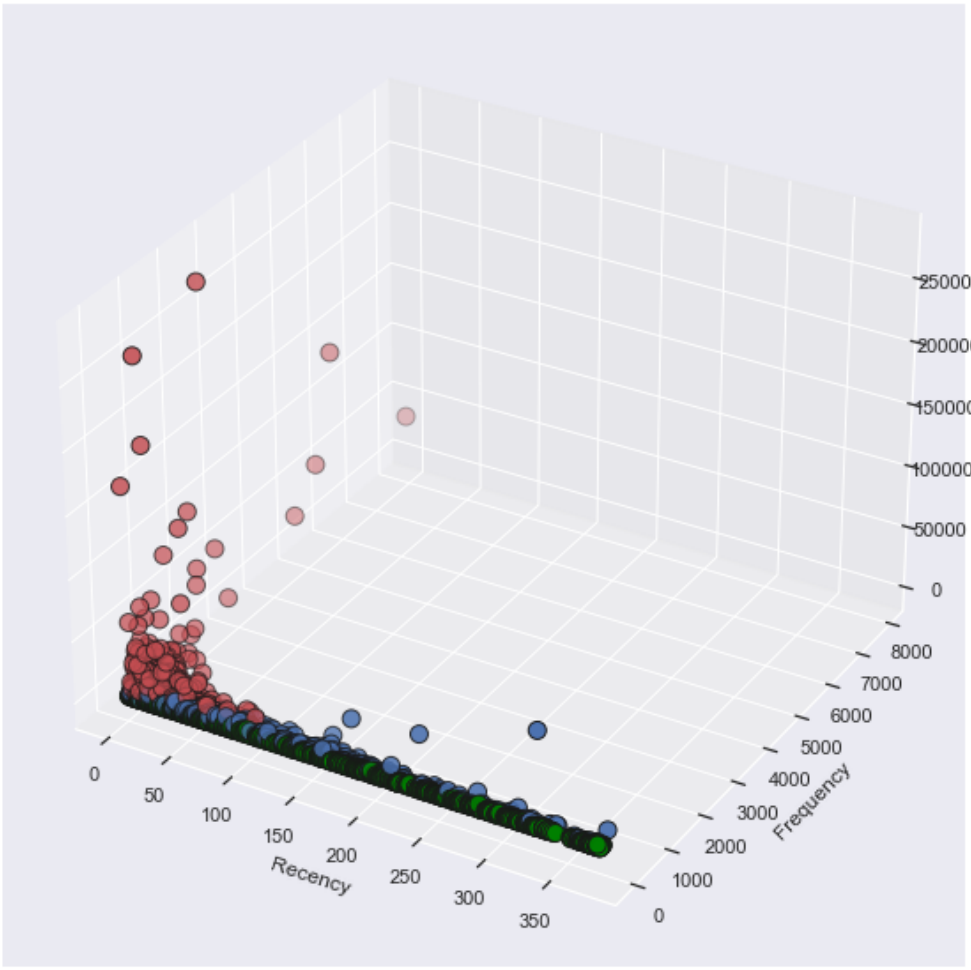
6.4.2 Le top7 des produits achetés par les bons clients

Description	StockCode
WHITE HANGING HEART T-LIGHT HOLDER	702
REGENCY CAKESTAND 3 TIER	620
ASSORTED COLOUR BIRD ORNAMENT	530
PARTY BUNTING	457
SET OF 3 CAKE TINS PANTRY DESIGN	430
JUMBO BAG RED RETROSPOT	394
POSTAGE	377

6.5 Valeurs moyennes de ces variables par groupe :

Cluster	Recency	Frequency	Monetary
0	12.93	263.99	6590.68
1	170.54	14.95	294.69
2	69.09	66.72	1180.35

Cette table résume parfaitement les 3 interprétations que nous venons de donner.



Les point rouges ont leurs valeurs de récence inférieurs à 100 jours, contrairement pour les autres points verts et bleu. Par contre, ils ont des valeurs monétaires et de fréquences élevées. Vous pouvez dévier quel groupe appartient ces points : le parfait groupe (groupe 0)

Les points bleus et verts ont quasiment les mêmes distribution de leurs récences, même si le nombre des points bleus semble beaucoup plus dominants entre 0 et 200 jours que le nombre des points verts. De plus, les points bleus ont des valeurs de fréquences et monétaires un peu plus élevées que celles des ponts verts : les points bleus appartiennent au groupe compris entre le parfait groupe et le groupe basique tandis que les points verts font partis du groupe basique (groupe 1 : pire groupe).

Projet 7 :
A/B Testing en stratégie de marketing promotionnel.

7.1 introduction

Étape 1 :

Supposons que vous aillez 2 versions d'une stratégie promotionnelle et vous voulez identifier laquelle rapporte un retour sur investissement performant en rapport à votre objectif marketing (l'augmentation du chiffre d'affaire par exemple).

Vous décidez donc de diviser vos clients en 2 groupes A et B distincts . Et durant 1, 2 ou 3 semaines vous leurs envoyez des emails de type promotionnel de vos produits. À la fin des 3 semaine d'étude. Vous classez le chiffre d'affaire moyen que chaque groupe vous ait rapporter

Étape 2 :

Normalement vous constaterez des différences marquées ou à peine distinctes au niveau des trois montants générés dans chaque groupe.

Supposons que vous avez remarquez que la ligne des emails du groupe A ait entrainé un taux de conversion un peu plus élevé que la ligne des emails du groupe B. Vous posez donc l'hypothèse que la promotion A rapporte plus que la promotion B. Vous voulez utiliser une deuxième fois la promotion A. Mais vous n'avez pas encore vérifié votre hypothèse. C'est là qu'intervient le test A/B sur ces deux groupes pour confirmer ou infirmer votre hypothèse. Dans ce projet, nous avons 3 versions de promotion et nous voulons identifier laquelle est performante pour l'adopter dans une future campagne promotionnelle.

Lors de l'analyse des résultats vous voudrons certifier s'il existe une différence statistiquement significative au niveau de ces 3 versions.

Data :

Cette étude se base sur des marchés de petite, moyenne et grande taille, dans lesquels siègent plus de 137 magasins uniques, d'âge allant de 1 à 14 ans d'existence. Et où les 3 promotions ont été réalisées durant 4 semaines, générant des montants de ventes.

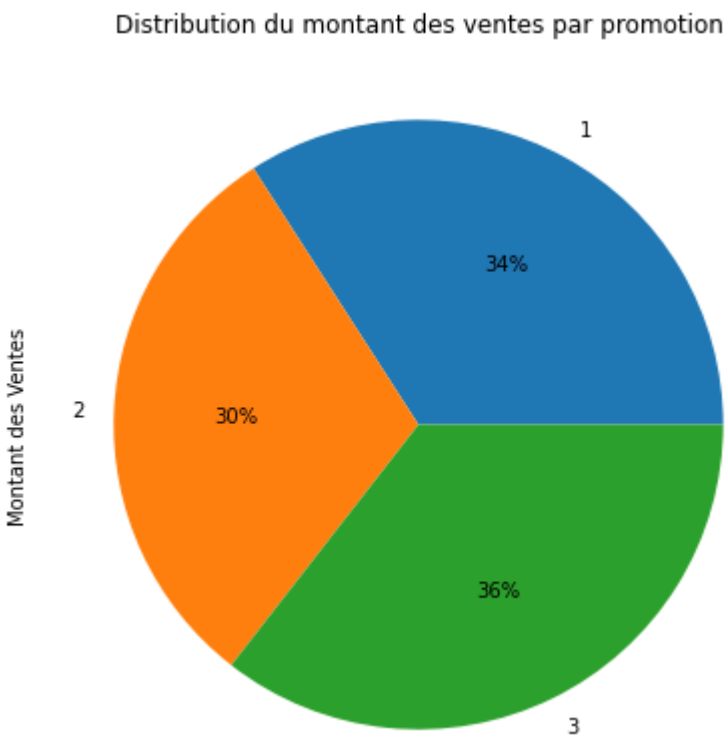
	MarketID	MarketSize	LocationID	AgeOfStore	Promotion	week	SalesInThousands
0	1	Medium	1	4	3	1	33.73
1	1	Medium	1	4	3	2	35.67
2	1	Medium	1	4	3	3	29.03
3	1	Medium	1	4	3	4	39.25
4	1	Medium	2	5	2	1	27.81
...	...	...	...	...	...	...	...
543	10	Large	919	2	1	4	64.34
544	10	Large	920	14	2	1	50.20
545	10	Large	920	14	2	2	45.75
546	10	Large	920	14	2	3	44.29
547	10	Large	920	14	2	4	49.41

548 rows × 7 columns

7.1 Problème.

La chose qu'il faut logiquement faire en premier , lorsque l'on veut savoir quelle version de stratégie promotionnelle fonctionne le mieux est de regarder la différence pouvant exister au niveau des montants moyens générés par promotion.

```
Promotion
1    58.099012
2    47.329415
3    55.364468
Name: SalesInThousands, dtype: float64
```



Nous remarquons que les montants moyens générés par les promo 1 et 3 semblent proches que ceux générés par les promo 1 et 2 ou 3 et 2. Et la distribution du visuel montre que le taux des montants générés par la promo 3 semble se démarquer un peu plus que les promos 1 et 2 respectivement. Cela nous mène à nous poser la question suivante : **la stratégie promotionnelle numéro 3 est-ce la meilleur?** La réponse serait affirmative si les différences que nous observons là sont bien réelles.

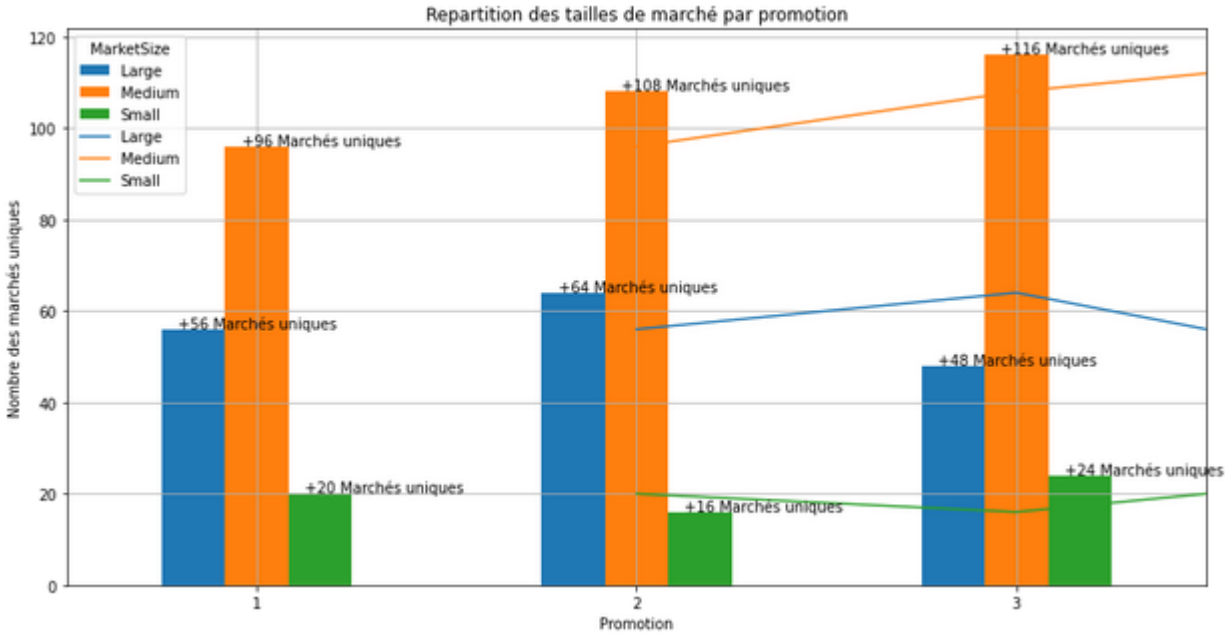
Le travails qui va suivre cherche à apporter une réponse à cette question. Cependant, il est important que nous fassions d'abord une analyse exploratoire afin que nous puissions comprendre un peu plus nos données.

7.2 La taille du marché par promotion.

7.2.1 table des tailles par promotion

MarketSize	Large	Medium	Small
Promotion			
1	56	96	20
2	64	108	16
3	48	116	24

7.2.2 Visualisation :



Interprétation :

Nous constations que les marchés moyens semblent plus facilement exploitables par la promotion 3 que par les promotions 2 et 1 respectivement

Le marché de large taille semble plus exploitable par la promotion 2 que par les promotions 1 et 3 respectivement.

Le marché de petite taille semble plus exploitable par la promotion 3 que par les promotions 1 et 2, respectivement.

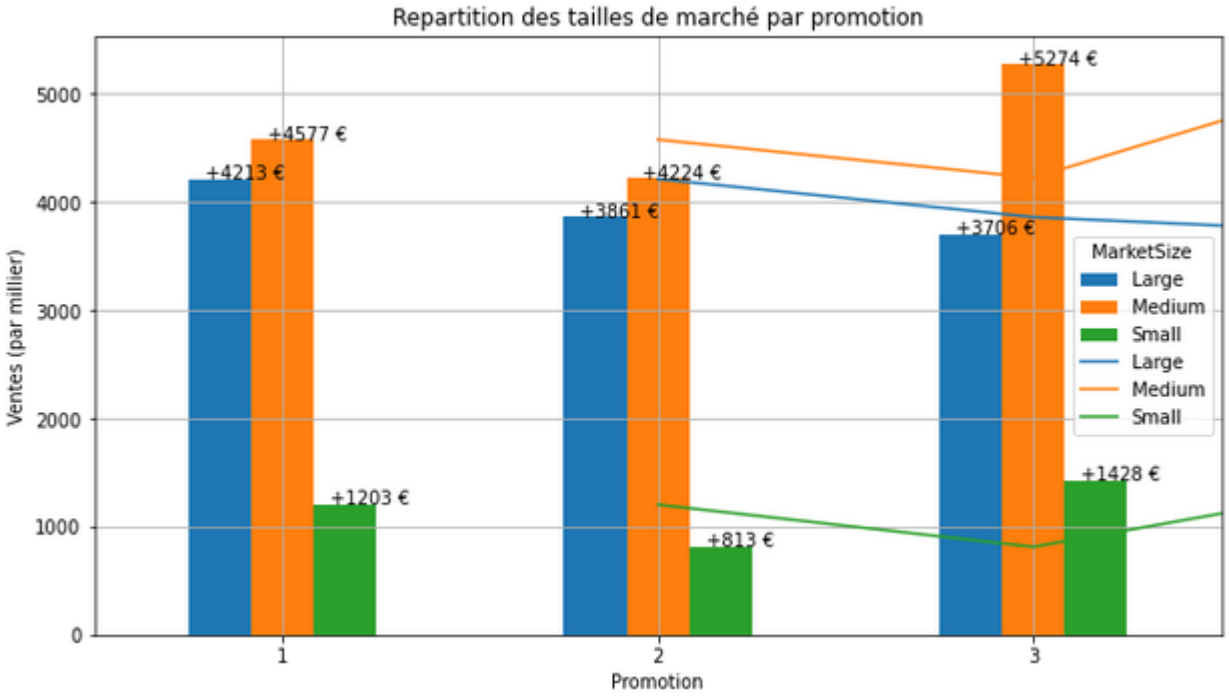
Bien évidemment, il faut valider ces différences par des tests A/B.

7.3 Distribution du montant des ventes générés par taille du marché via chaque promotion.

7.3.1 Table .

MarketSize	Large	Medium	Small
Promotion			
1	4213.21	4576.57	1203.25
2	3860.61	4224.35	812.97
3	3705.79	5274.39	1428.34

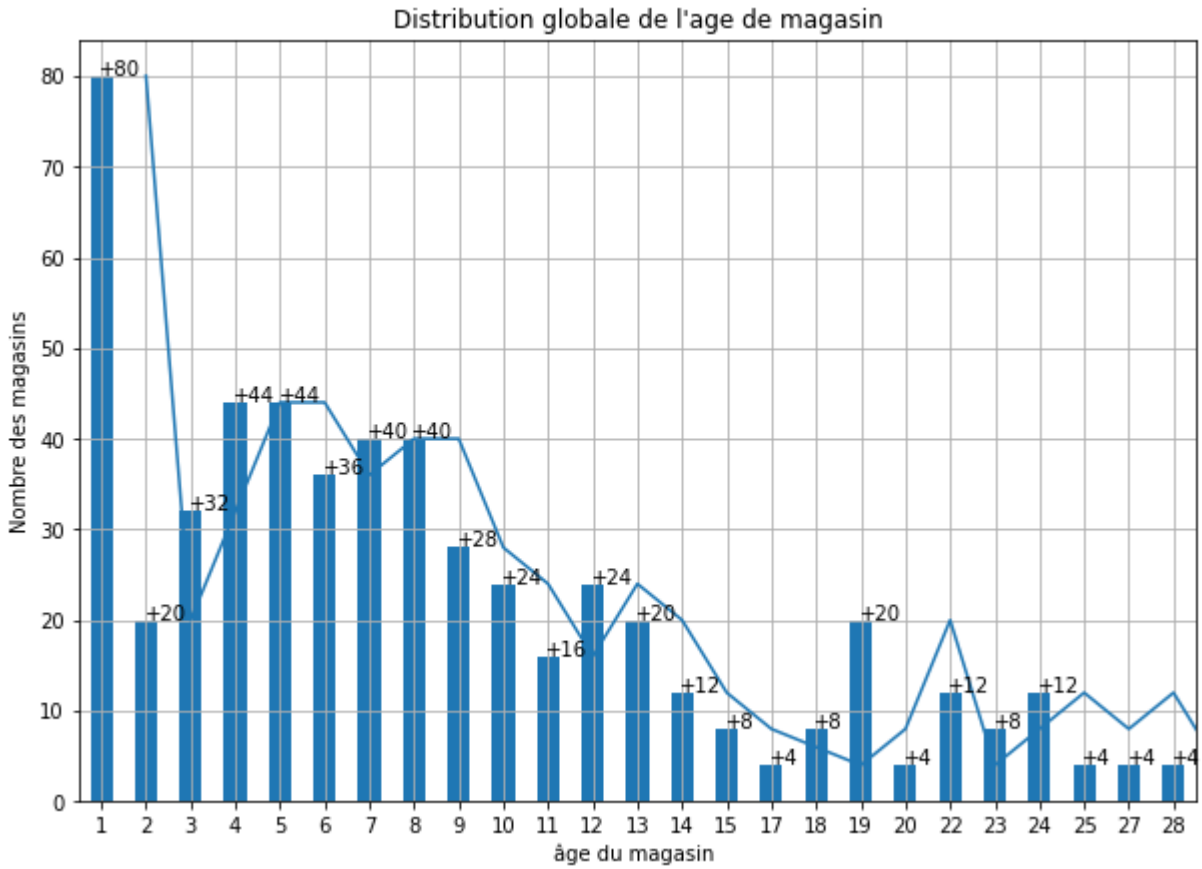
7.3.2 Visualisation de la distribution des montant générés par taille de marché suivant les promos.



Les marchés de taille moyenne et de petite taille semblent rapporter plus de chiffre d'affaire par la promotion 3 que par les promotions 1 et 2, tandis que le marché de grande taille rapporte plus de chiffre d'affaire par la promotion 1 que par la promo 2 ou 3.

Le marché de grande taille semble rapporter le même chiffre d'affaire par les promotion 2 et 3 (à vérifier par les tests).

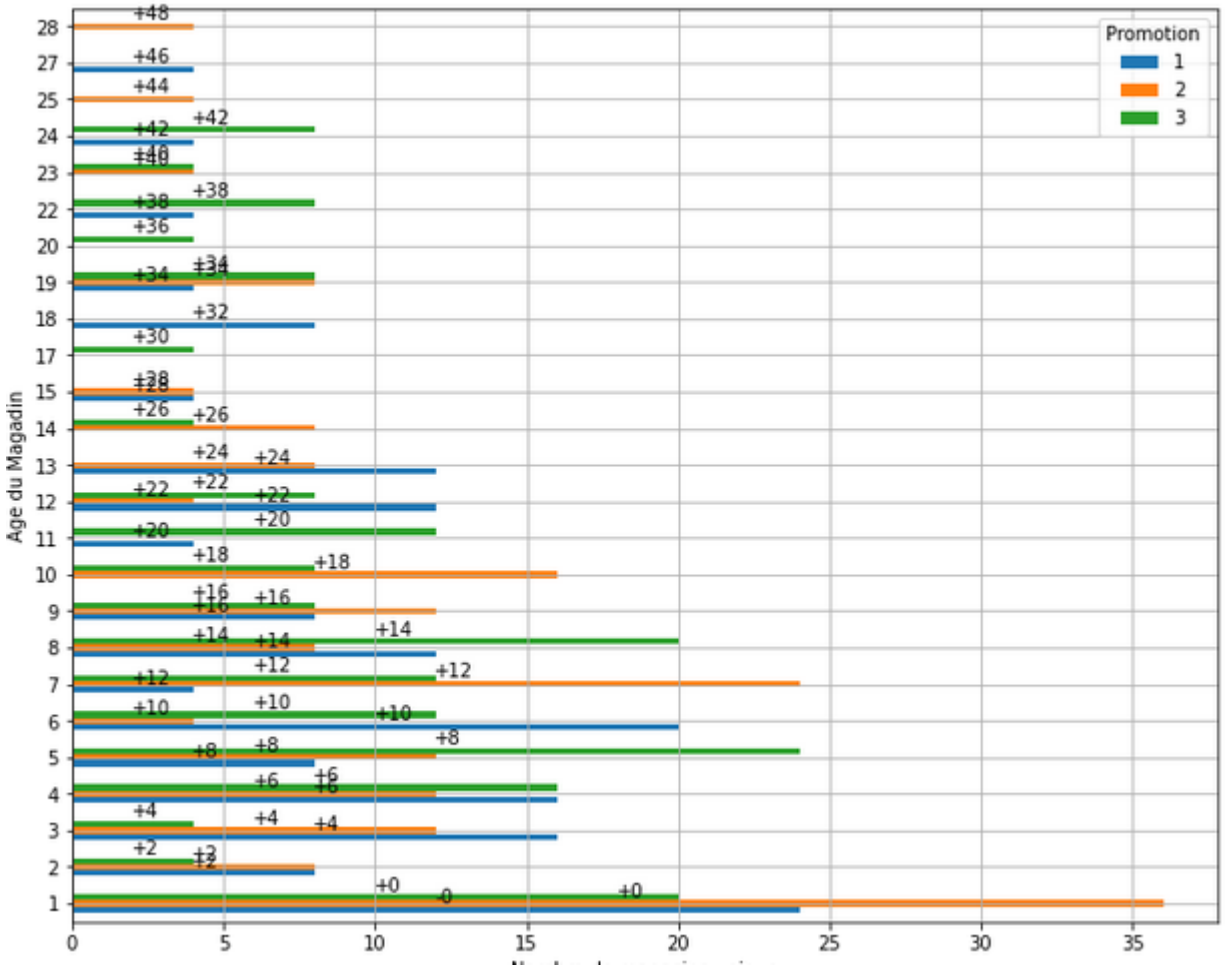
7.4 Le nombre des magasins par âge.



Interprétation :

La majorité des magasins (plus de 80) ont à peine une année d'existence. Mais à part cette catégorie, la majorité ont entre 3 et 9 ans d'existence.

7.5 Distribution de l'âge des magasins par type de promotion.



Interprétation :

Plus les magasins du marché prennent de l'âge, plus ils se raréfient. Et cela est d'autant beaucoup plus marquant lorsqu'ils pratiquent des stratégie promotionnelles de type 1 et 2 que de type 3, respectivement.

7.6 La promotion qui rapporte le plus d'argent :

7.6.1. Cas des promos 1 et 2.

Au début, nous avons remarqué que le montant moyen des ventes dans la promotion 1 est un peu plus élevé que le montant moyen des ventes dans la promotion 2 (58.9€ > 47.3€)): **est-ce vrai ou pas ?**

Résultat du test :

```
t1_2, p1_2 = stats.ttest_ind(df1.loc[df1['Promotion']==1]['SalesInThousands'],
                             df1.loc[df1['Promotion']==2]['SalesInThousands'],
                             equal_var=False)
t1_2, p1_2
```

```
(6.42752867090748, 4.2903687179871785e-10)
```

Interprétation :

T = 6.43 et Pvalue = 0,0000…< 0.05

Ces deux valeur t et p nous disent que **la différence observée entre les deux montants de ventes dans les promotions 1 et 2 est bel et bien statistiquement significative**.

Conclusion 2:

La promotion 1 génère bel et bien plus de ventes que **la promotion 2**

7.6.2 Cas des promos 1 et 3

Nous avons encore vu que le montant moyen des ventes dans la promotion 1 semble un peu plus élevé que le montant moyen des ventes dans la promotion 3 (58.9€ > 55.36€): **est-ce vrai ou pas ?**

Résultat du test :

```
t1_3, p1_3
```

```
(1.5560224307758634, 0.12059147742229478)
```

Interprétation

T = 1.55 et Pvalue = 0.121 >0.05

Ces deux valeur t et p nous disent que **la différence observée entre les deux montants de ventes dans les promotions 1 et 3 est statistiquement due au hasard** (non significative). En gros, la différence entre les montants de ventes dans les promotions 1 et 3 n'est pas vraie:

Conclusion 2 :
Les promotions 1 et 3 génèrent donc identiquement le plus de ventes et constituent les seules promotions qui apportent un résultats sûr: la mise en place d'une stratégie promotionnelle de type 3 ou 1 sera préférable que celle de type 2. Voilà la fin de ce projet.