Projet de clustering : segmentation de clients

DIARRASSOUBA SAKARIA, DOUBA JAFUNO, OUMAR BALDE, SIDY MAMADOU DIALLO

23/01/2020

Contexte

Dans cette analyse, nous utiliserons une base de données contenant des données de 200 clients d'un centre commercial et cette base de données contient l'identité du client, son âge, son sexe, son revenu annuel et son score des dépenses. L'objectif de cette analyse est d'identifier le segment de clientèle à l'aide du CAH (Classifiaction hiérarchique ascendante) ou du K-Means Clustering, afin de comprendre quel est le segment de clientèle qui devient la cible de l'équipe marketing pour planifier les stratégies de marketing.

1.Etudes des données

a. Description des données:

Tableau 1: Résumé des données

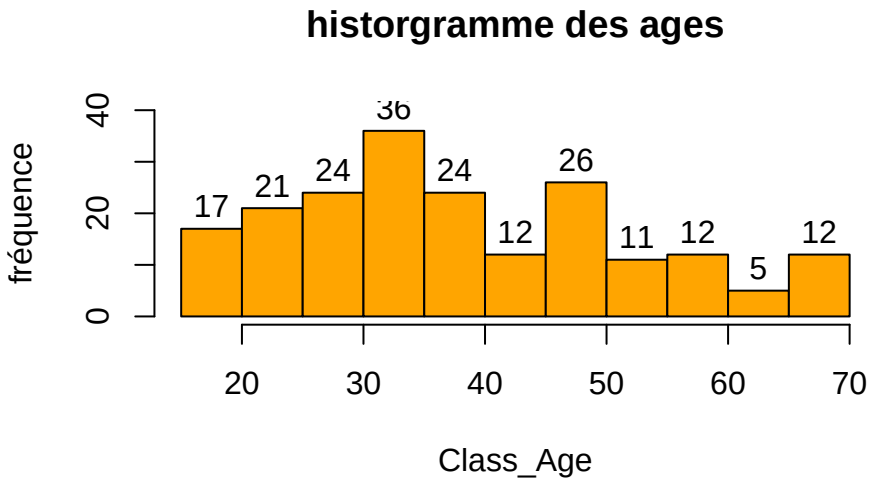
	Age	Annual_Income	Spending_Score
Min	18.00	15.00	1.00
1st Qu	28.75	41.50	34.75
Median	36.00	61.50	50.00
Mean	38.85	60.56	50.20
3rd Qu	49.00	78.00	73.00
Max	70.00	137.00	99.00

	Female	Male	Total
Nb_clients	112	88	200
poucentage	56%	44%	100%

Fait intéressant on voit que les scores de dépenses varie de 1 K US Dollar et 99 K US Dollar, ce qui montre que le centre commercial répond aux besoins et exigences variés des clients et les revenus annuels sont compris entre 15 K US Dollar et 137 K US Dollar.

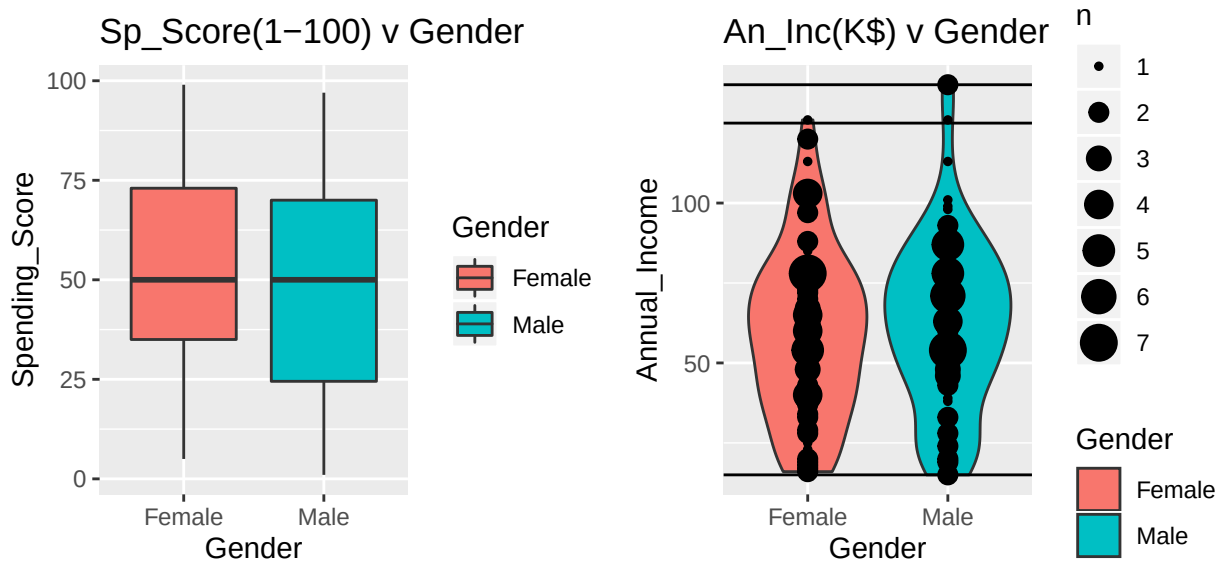
Notons qu'on n'a pas de **valeurs manquantes** dans notre dataset.

b. Répartions des ages



Là, on constate que la majeure partie des clients ont des ages compris entre 25 et 40 ans donc on peut dire qu'on a une clientèle jeune.

c. Répartitions des scores de dépenses et revenus anneuls en fonction du genre



Avec ce graphe des revenus annuels ci-dessus, on constate que deux choses, premiérement que les femmes, leur revenu annuel varie entre **15 K** et **125 K US Dollar** et pour les hommes, leur revenu annuel varie entre **15 K** et **137 K US Dollar**, clairement on peut affirmer que le genre a un impact sur le revenu annuel. Pour le graphe des scores de dépenses on voit clairement que la plupart des hommes ont un score de dépenses d'environ 25 à 70 alors que les femmes ont un score de dépenses d'environ 35 à 75. Ce qui confirme encore une foisque les femmes sont les leaders en matière de shopping. Par conséquent, on voit que le genre du client à une influence sur le revenu annuel et le score de dépense pour un client donné.

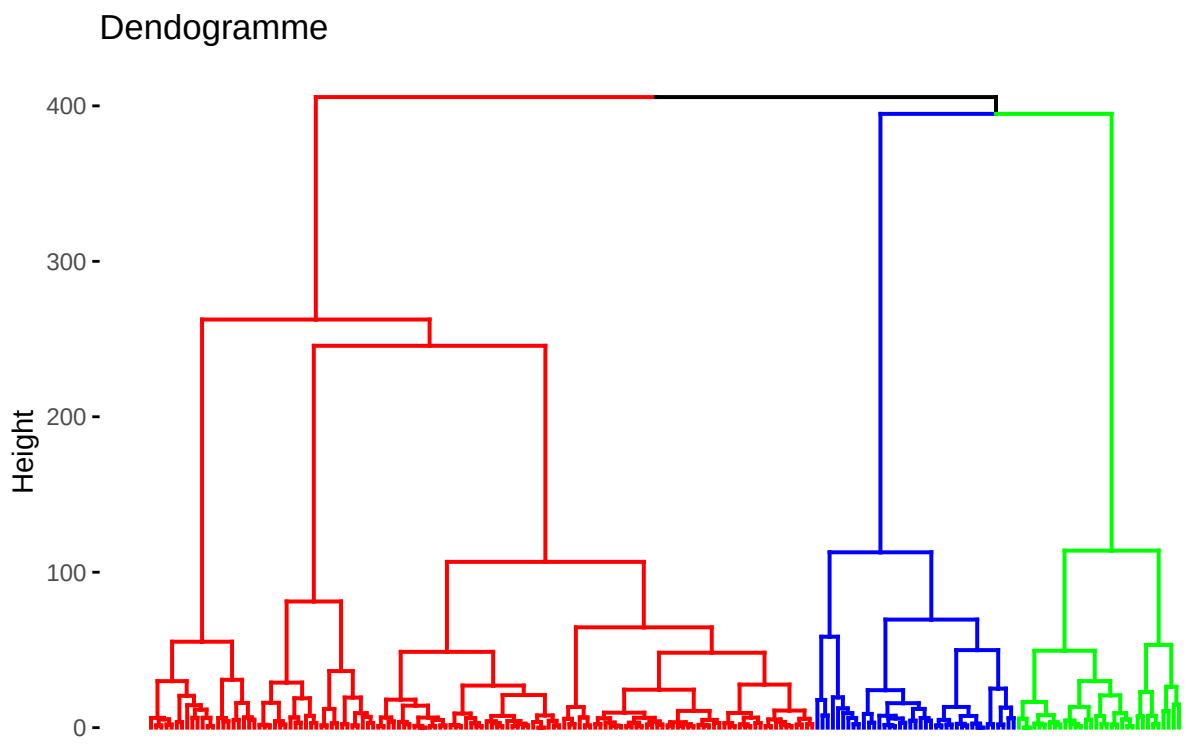
2.Clustering

a. CAH

Principe du CAH

Le principe de la CAH est de rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existant entre chaque individu pris deux à deux. Deux observations identiques auront une distance nulle. Plus les deux observations seront dissemblables, plus la distance sera importante. La CAH va ensuite rassembler les individus de manière itérative afin de produire un dendrogramme ou arbre de classification. La classification est ascendante car elle part des observations individuelles ; elle est hiérarchique car elle produit des classes ou groupes de plus en plus vastes, incluant des sous-groupes en leur sein. En découpant cet arbre à une certaine hauteur choisie, on produira la partition désirée.

Mise en oeuvre



On voit qu'on a 3 groupes distincts. Découpons maintenant ce dendogramme en 3 groupes à la hauteur h=300 en utilisant la fonction **cutree** .

Tableau 2 : les 3 différents clusters

	cluster1	cluster2	cluster3	Total
Nb_clients	129	39	32	200
Pourcentage	64.5%	19.5%	16%	100%

Résumé cluster 1 :

	Age	Annual_Income	Spending_Score
Min	18.00	15.00	3.00
1st Qu	25.00	33	41.00
Median	38.00	48.00	49.00
Mean	40.18	45.55	49.13
3rd Qu	51.00	60.00	58.00
Max	70.00	79.00	99.00

	Female	Male	Total
Nb_clients	77	52	129
Pourcentage	60%	40%	100%

Résumé cluster 2

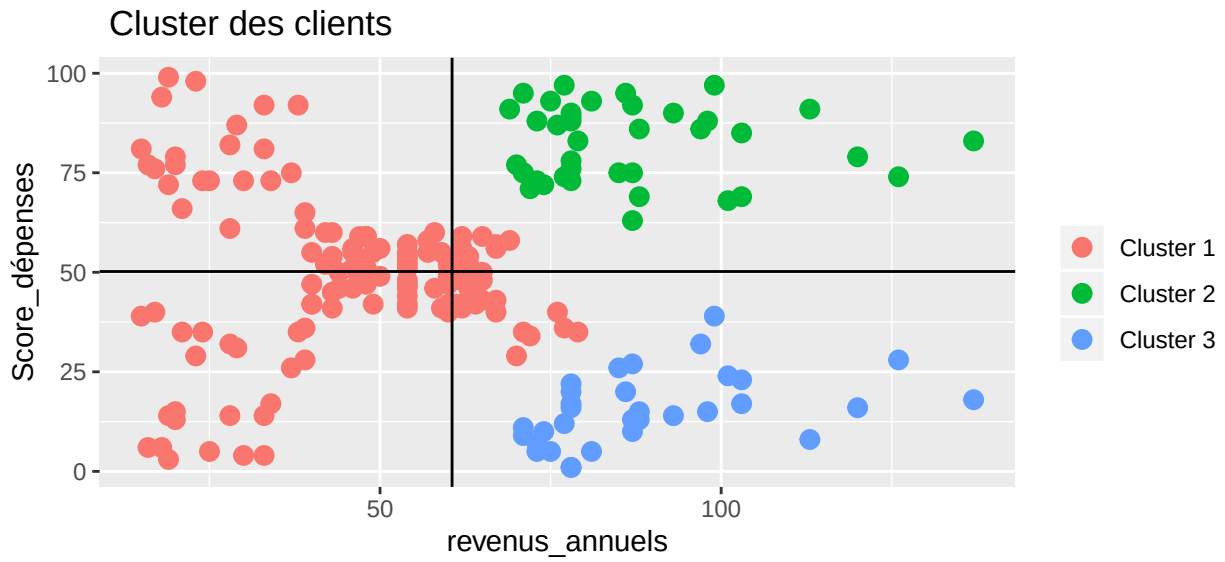
	Age	Annual_Income	Spending_Score
Min	27.00	69.00	63.00
1st Qu	30.00	75.50	74.50
Median	32.00	79.00	83.00
Mean	32.69	86.54	82.13
3rd Qu	35.50	95.00	90.00
Max	40.00	137.00	97.00

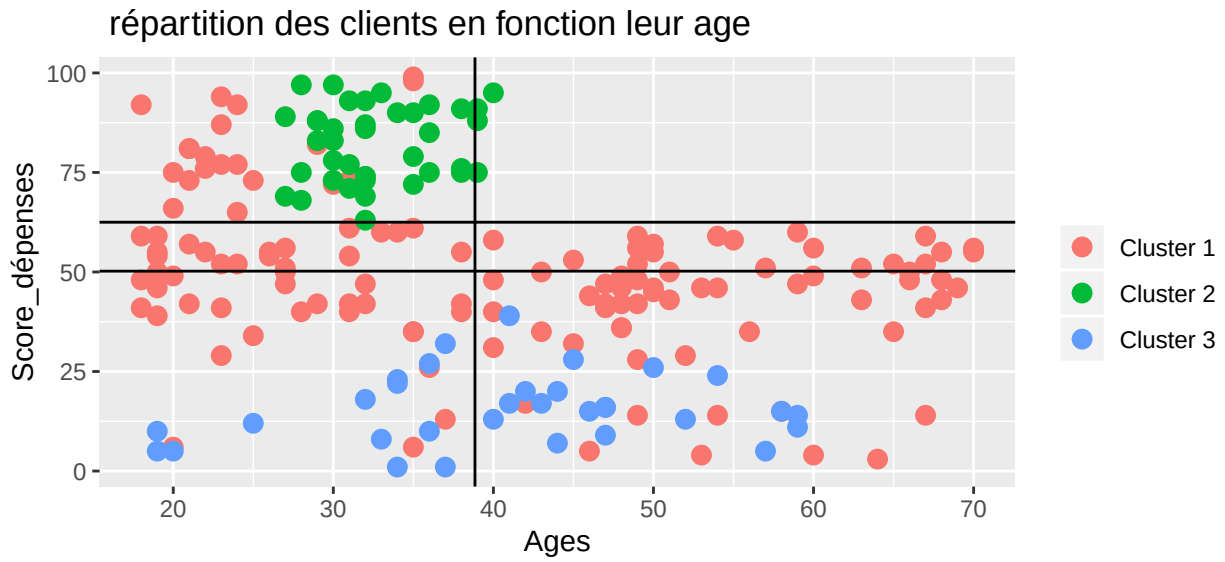
	Female	Male	Total
Nb_clients	21	18	39
Pourcentage	54%	46%	100%

Résumé cluster 3:

	Age	Annual_Income	Spending_Score
Min	19.00	71.00	1.00
1st Qu	34.00	78.00	9.75
Median	41.50	86.50	15.00
Mean	41.00	89.41	15.59
3rd Qu	47.00	98.25	20.50
Max	59.00	137.00	39.00

	Female	Male	Total
Nb_clients	14	18	32
Pourcentage	43.75%	56.25%	100%





Analyse et interprétation des 3 clusters

- Pour le graphe revenus_annuels et Score_dépenses : Le cluster 1 est caractérisé par pour la quasi-totalité des clients de ce groupe par un faible revenu annuel qui varie entre **15 K** et **76 K US Dollar** et par un score de dépense variant entre **3 K** et **99 K US Dollar**.Ce cluster, on peut le voir comme étant le groupe de clients standard. Les personnes du cluster 2 font plus d'achats(de **63 K** à **97 K US Dollar**) que les autres personnes des deux autres groupes,donc on peut dire qu'elles sont plutôt ravis par les services qu'offre le centre commercial. Notons aussi qu'elles ont des revenus annuels élevés variant entre **69 K** et **137 K US Dollar** et que la gente féminine est représentée à **54%**. Pour les clients du cluster 3, on constate qu'ils ont un revenu annuel élevé qui est entre **71 K** et **137 K US Dollar** mais ils ont un faible score de dépense variant entre **1 K** et **39 K US Dollar**. De plus, la gente masculine est représentée à 56.25%, on peut dire que les clients de ce groupe ne sont pas satisfaits des services du centre commercial surtout la gente masculine par conséquent le centre commercial a une carte à jouer pour inverser cette tendance. En ce qui concerne le **graphe Score_dépenes vs Age**, on observe que les clients ayant plus de 40 ans indépendement de leur sexe ont des scores de dépenses faible en moyenne et ne dépassent pas **62.5 US Dollar** donc le centre a intérêt de changer ses offres afin de répondre aux attentes de cette tranche d'ages.

Pour les personnes agées de moins 40 ans, on constate l'effet inverse et donc le centre commercial peut les accorder quelques opportunités afin qu'ils continuent à faire leurs achats.

Conclusion: Avec le CAH, on a pu classifier les clients en 3 groupes. Les informations qu'apportent cette classification permettra au centre commercial de planifier des plans marketing ciblés.

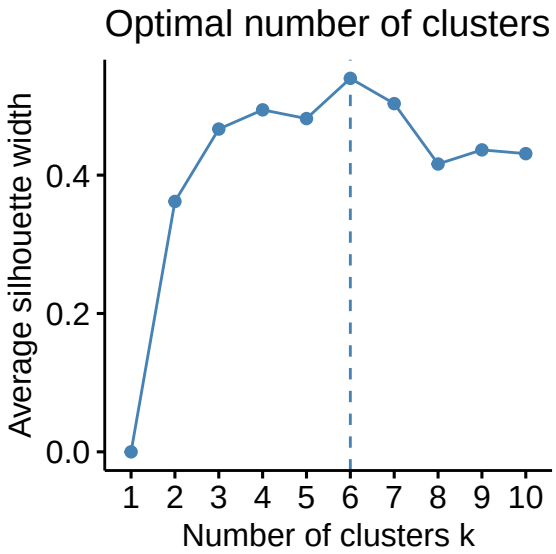
b. Kmeans

Principe du Kmeans

K-means est un algorithme non supervisé de clustering non hiérarchique.Il permet de regrouper en K clusters distincts les observations d'un dataset. Ainsi les données similaires se retrouveront dans un même cluster. Par ailleurs, une observation ne peut se retrouver que dans un cluster à la fois (exclusivité d'appartenance). Une même observation, ne pourra donc, appartenir à deux clusters différents.

Mise en oeuvre

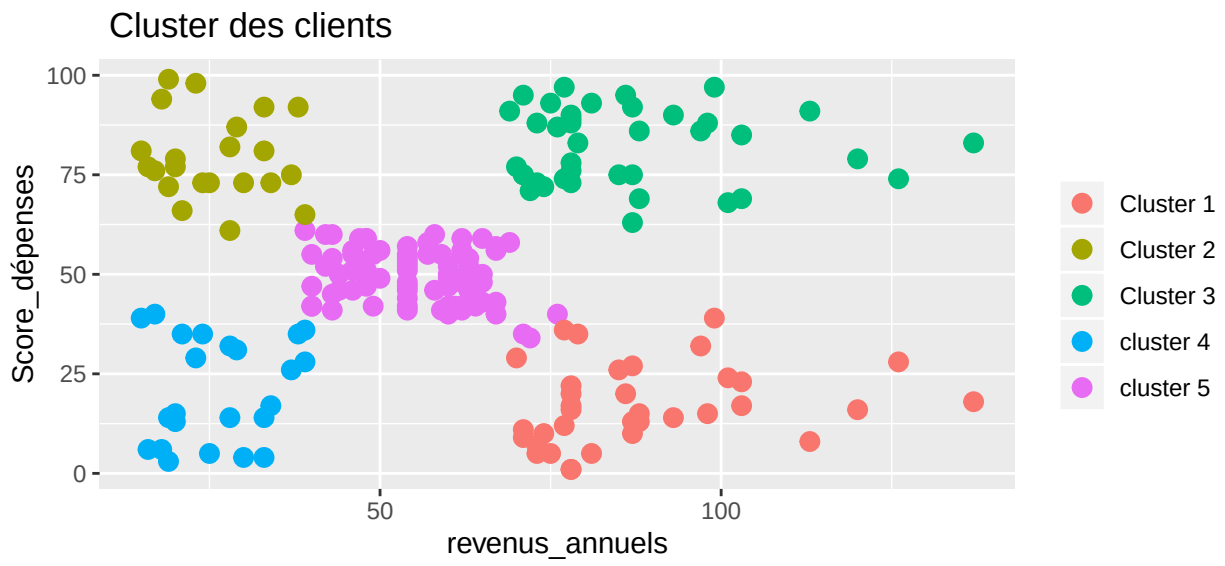
D'abord, on va chercher le **k optimal** en utilisant la méthode de la *silhouette*.



Donc le **K optimal** vaut 5 et faisons maintenant le Kmeans avec K=5.

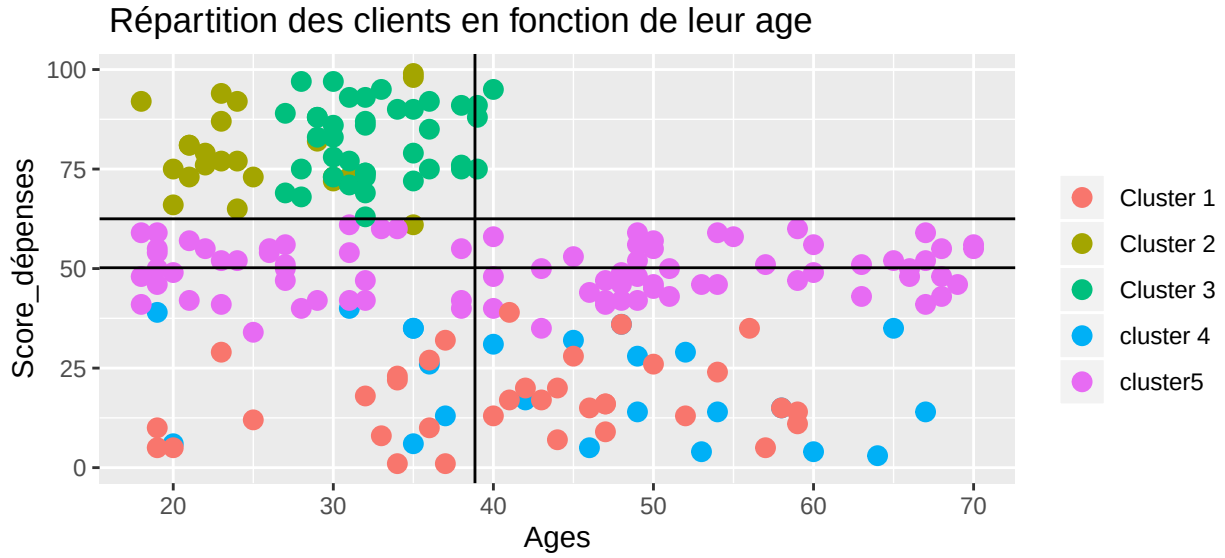
Tableau 3: les 5 clusters

	Cluser1	cluster2	cluster3	cluster4	cluster5	Total
Female	21	16	48	13	14	112
Male	18	19	33	9	9	88
Nb_clients	39	35	81	22	23	200
Pourcentage	19.5%	17.5%	40.5%	11%	11.5%	100%



Analyse et interprétation des 5 clusters

Avec ce graphique ci-dessus, on voit clairement les clusters 3, 1, 5, 2 et 4 sont caractérisés respectivement par des revenus annuels et score de dépenses trés élevès, par des revenus annuels élevés et des scores de dépenses faibles, par des revenus annuels et scores de dépenses moyens, par des revenus annuels faibles et des scores de dépenses trés élevés, par des revenus annuels et scores de dépenses trés faibles. Notons que le cluster 1 est dominé par la gente masculine.



On a la même configuration que celle obtenue avec le CAH (graphique Score_dépenses vs Age) donc on a la même interprétation.

c. CAH vs Kmeans lequel chosir !

On constate que le kmeans a éte plus sensible que le CAH dans la segmentation des clients de cette dataset donc il est préférable de montrer les résultats obtenus par le Kmeans aux responsables du centre commercial

et ces résultats permettront au centre commercial de planifier des plans marketing.

Conclusion:

Avec ces deux alogrithmes, on est arrivé à bien segmenter la clientèle du centre commercial. Cette segmentation sera d'aide précise lors de la prise de décision en ce qui concerne le plan marketing. Il devra consacrer une partie de son plan marketing pour la gente masculine.