Exploration des données

Arnaud Vanholderbeke, Benjamin Loriot et Natan Danous

11 Mai 2019

Contents

Description

Ces données sont les ventes provenant d'un seul magasin lors du Black Friday. Le magasin veut mieux comprendre le comportement des consommateur sur différents produits. Selon Kaggle, le problème principal est un problème de regression. On essaie de prédire la variable quantité d'achat à l'aide des autres variables. Ce problème peut également être vu comme un problème de régression.

Types et granularité

Table 1: Dictionnaire des données

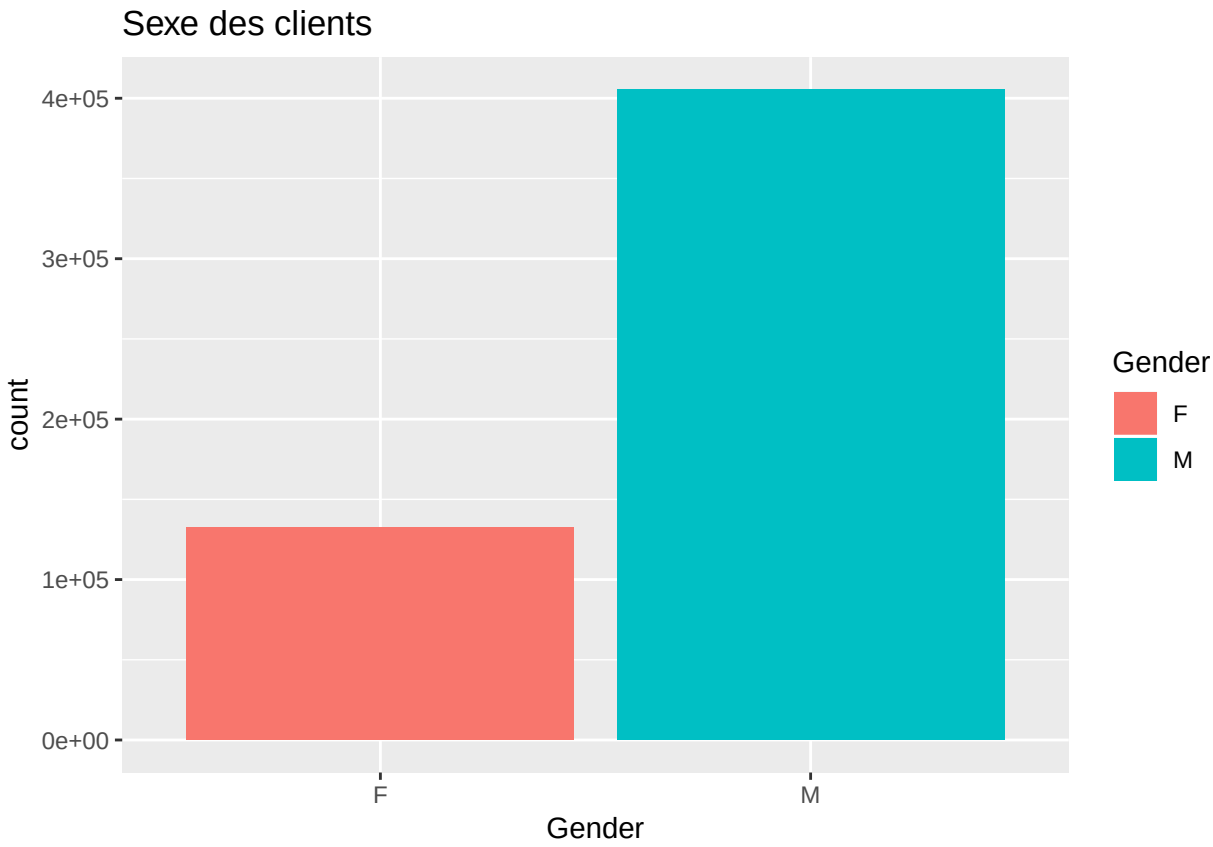
Champ	Type	Modalités
Purchase	Quantitative	
User_ID	Ordinal	
Product_ID	Nominal	
Gender	Nominal	F, M
Age	Ordinal	0-17 -> ... -> 51-55 -> 55+
Occupation	Nominal	0 -> 20
City_Category	Nominal	A, B, C
Stay_In_Current_City_Years	Ordinal	0, 1, 2, 3, 4+
Marital_Status	Nominal	0, 1
Product_Category_1	Nominal	1 -> 18
Product_Category_2	Nominal	2 -> 18 (avec NA)
Product_Category_3	Nominal	3 -> 18 (avec NA)

Remarque :

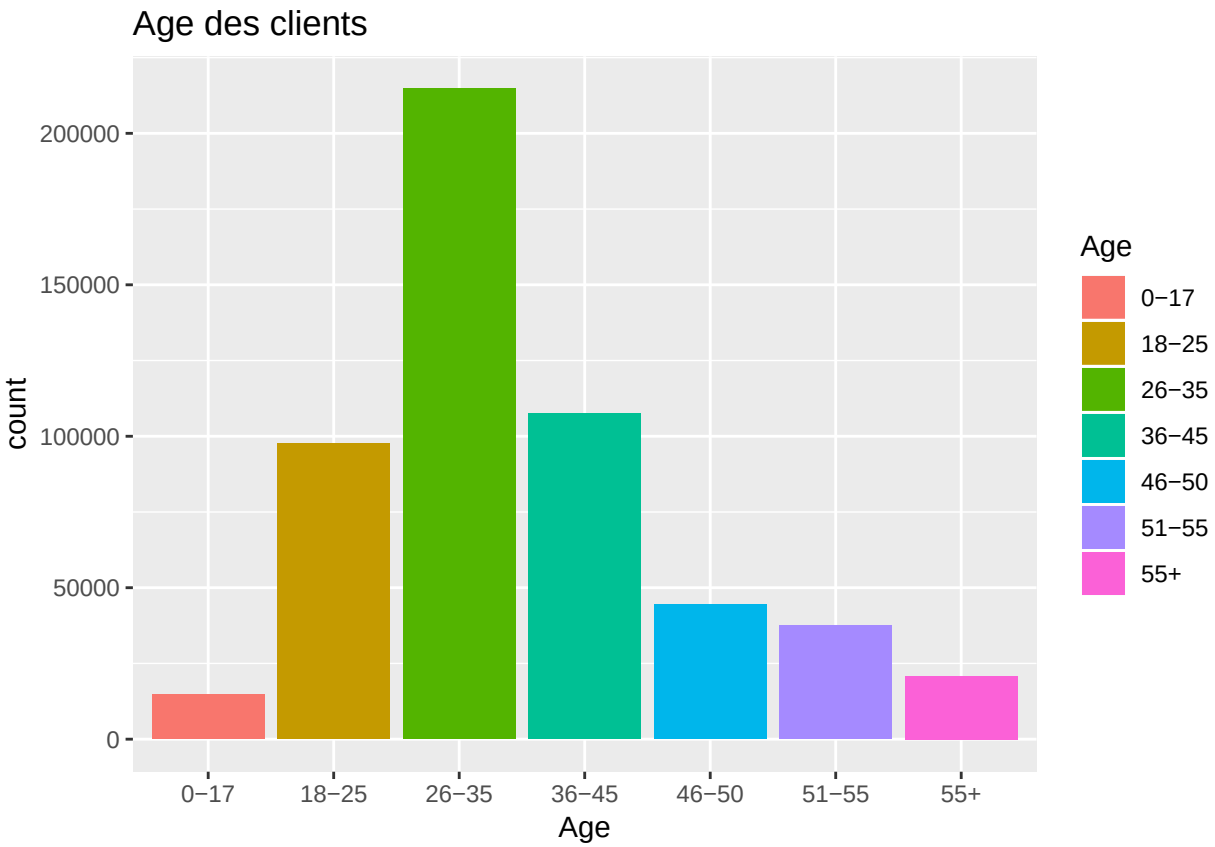
- Un produit a obligatoirement une catégorie (Product_Category_1 ne contient pas de NA). Il n'a pas obligatoirement de 2ème et 3ème catégorie.

Distribution

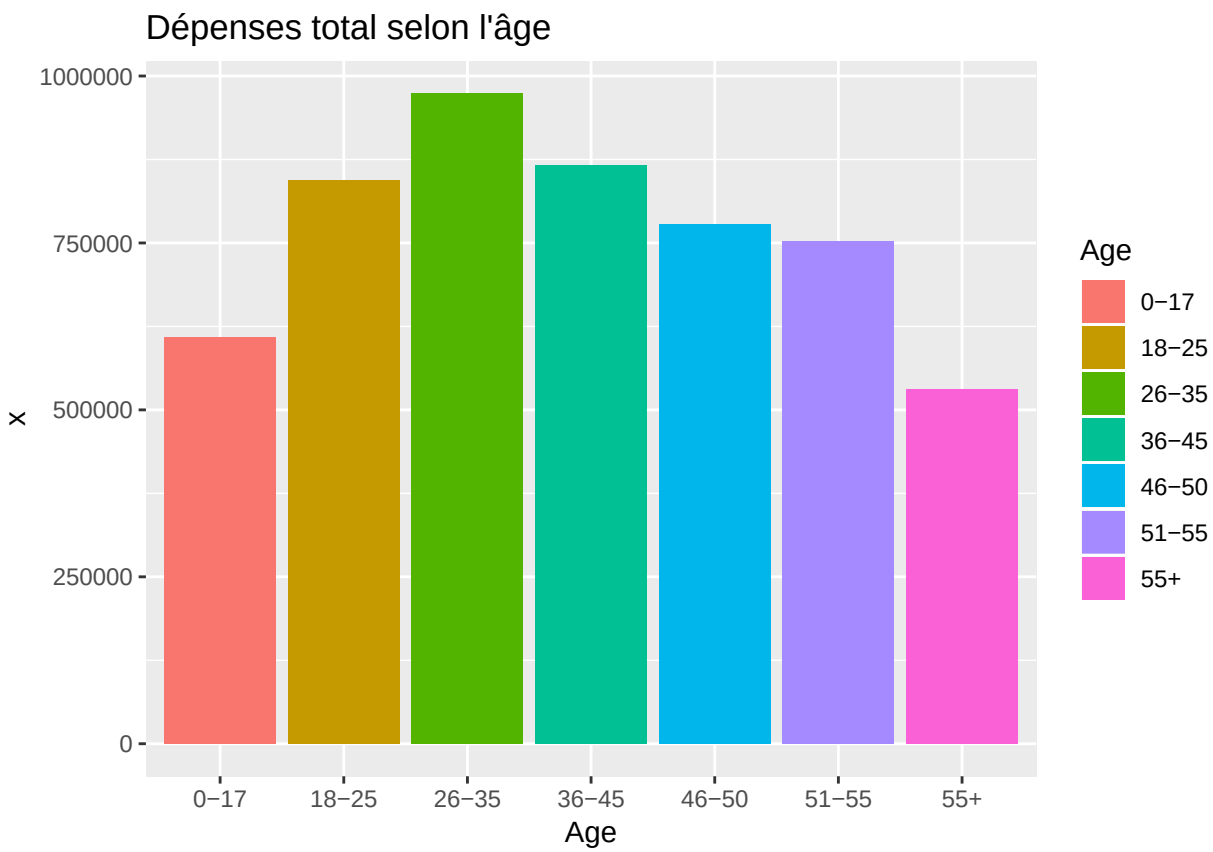
Sexe



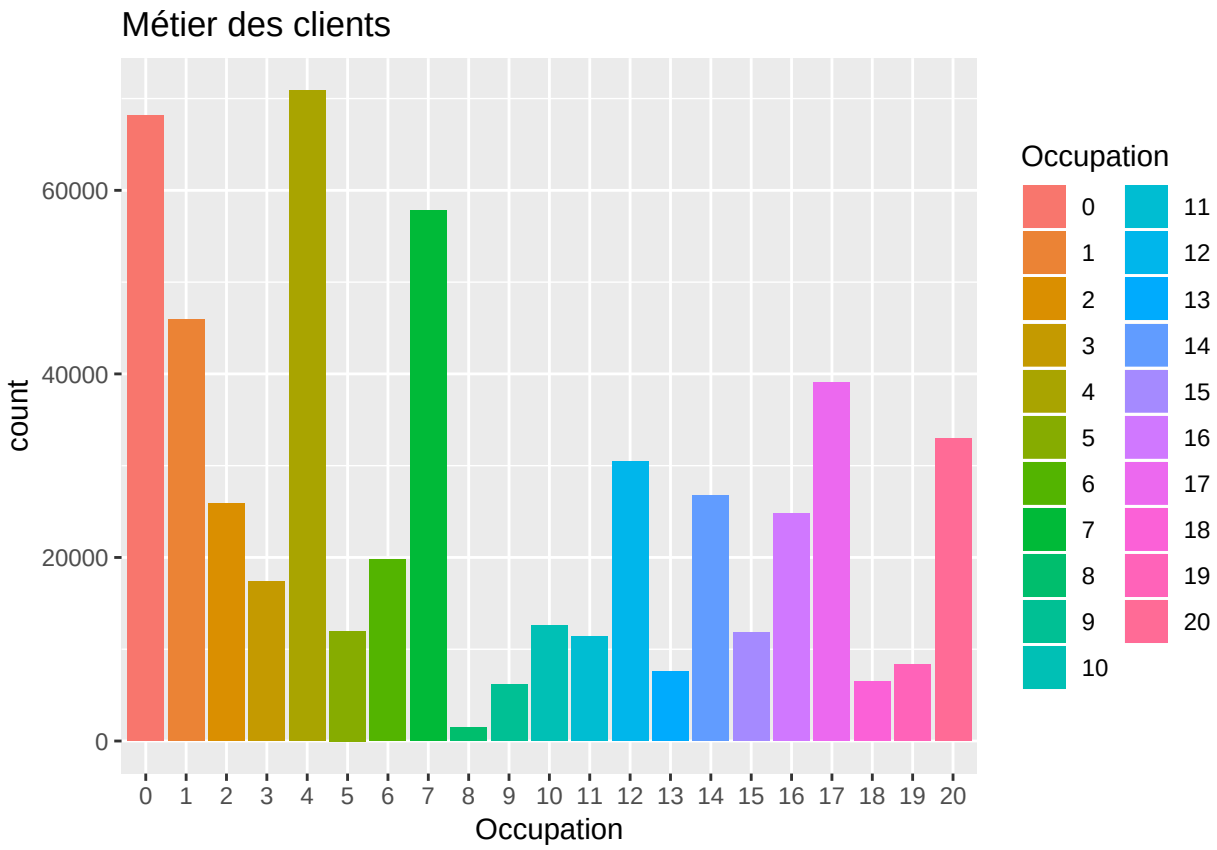
Age



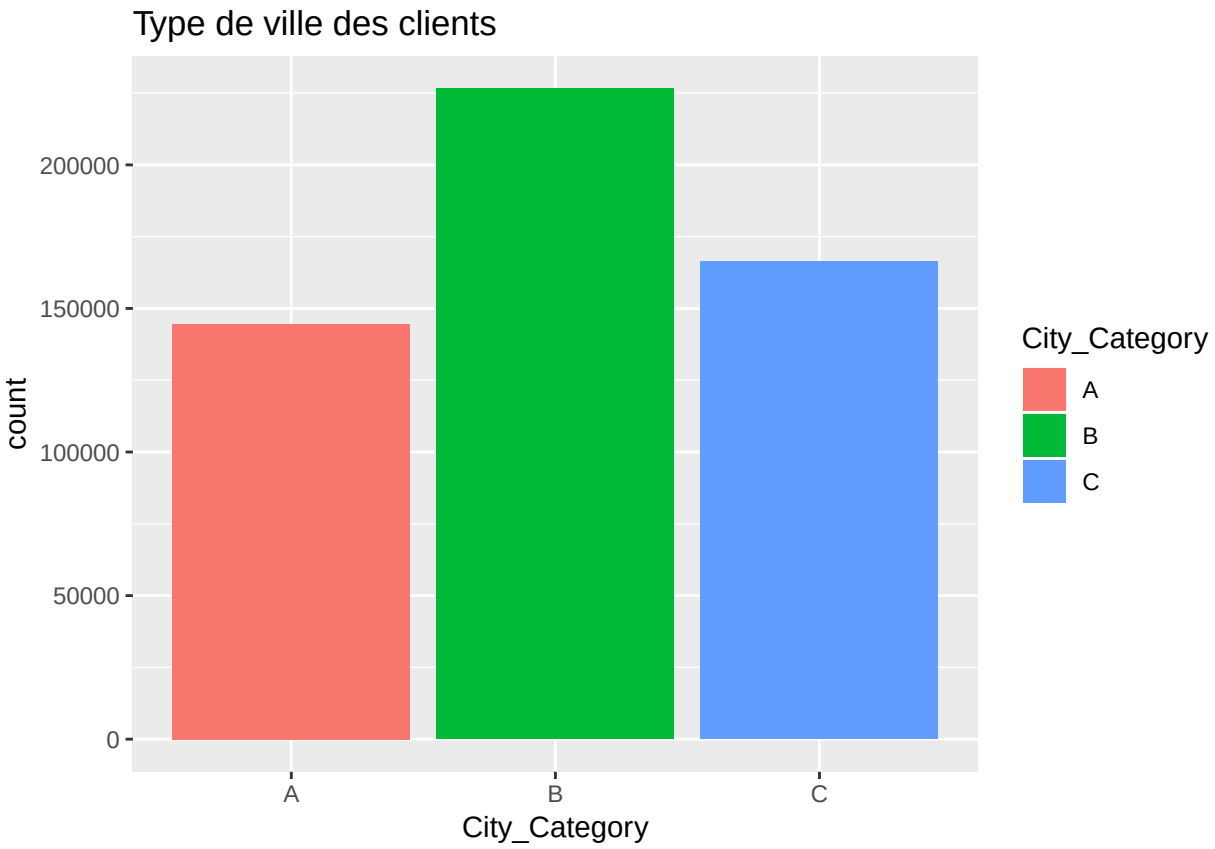
La distribution de l'age dans une population suit habituellement une loi normale. On observe ici que ce n'est pas le cas. Il apparait que la tranche d'âge des 26-35 effectue le plus d'achats lors du Black Friday.


Dépenses total selon l'âge
1000000
750000
500000
250000
0
x
0−17
18−25
26−35
36−45
46−50
51−55
55+
Age
Age
0−17
18−25
26−35
36−45
46−50
51−55
55+

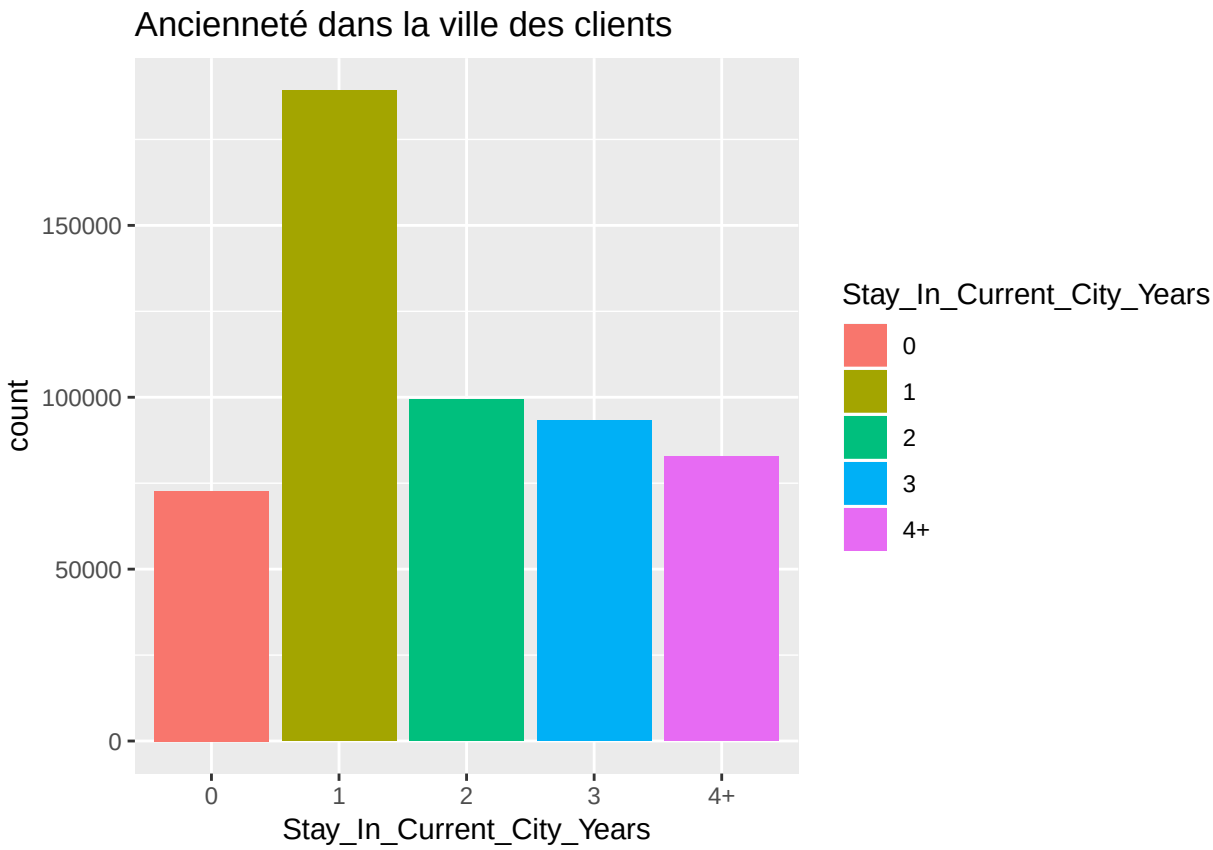
Métier


Métier des clients
count
60000
40000
20000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Occupation
Occupation
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

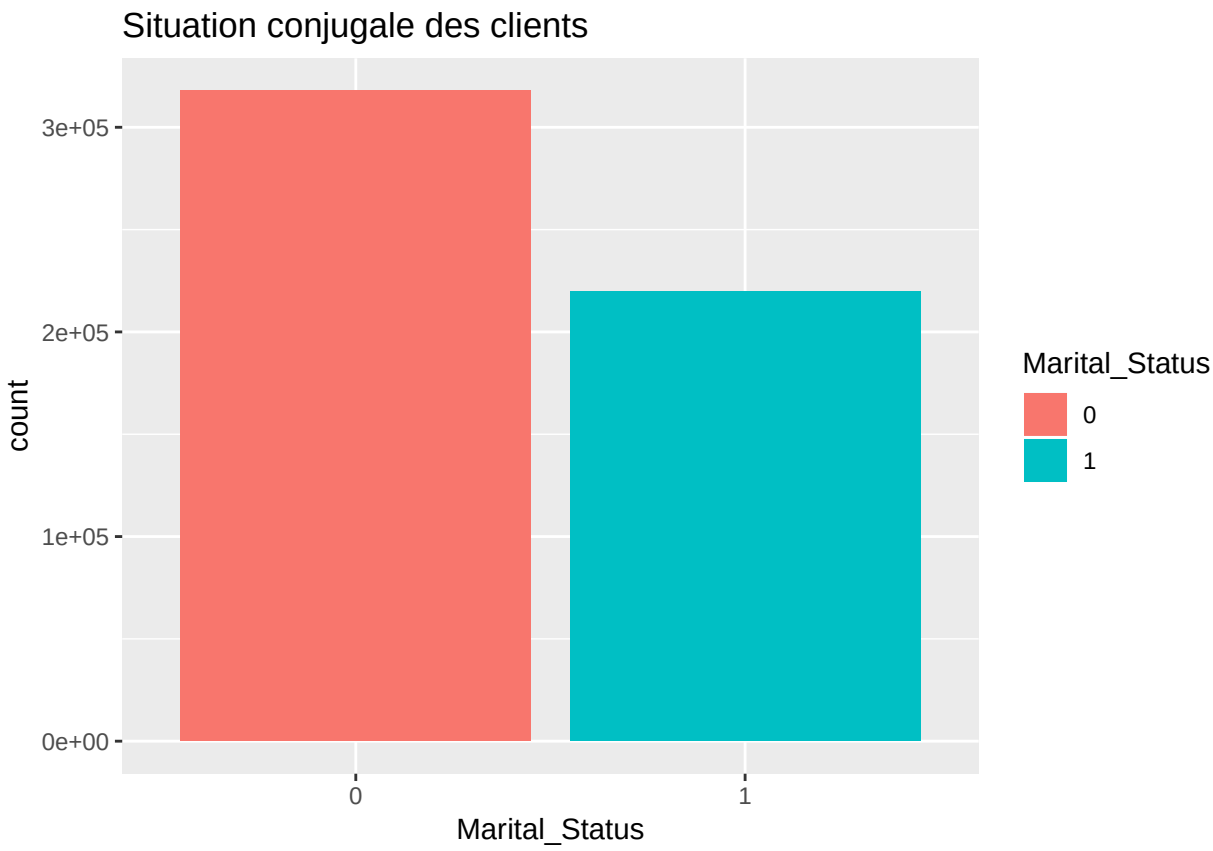
Type de ville



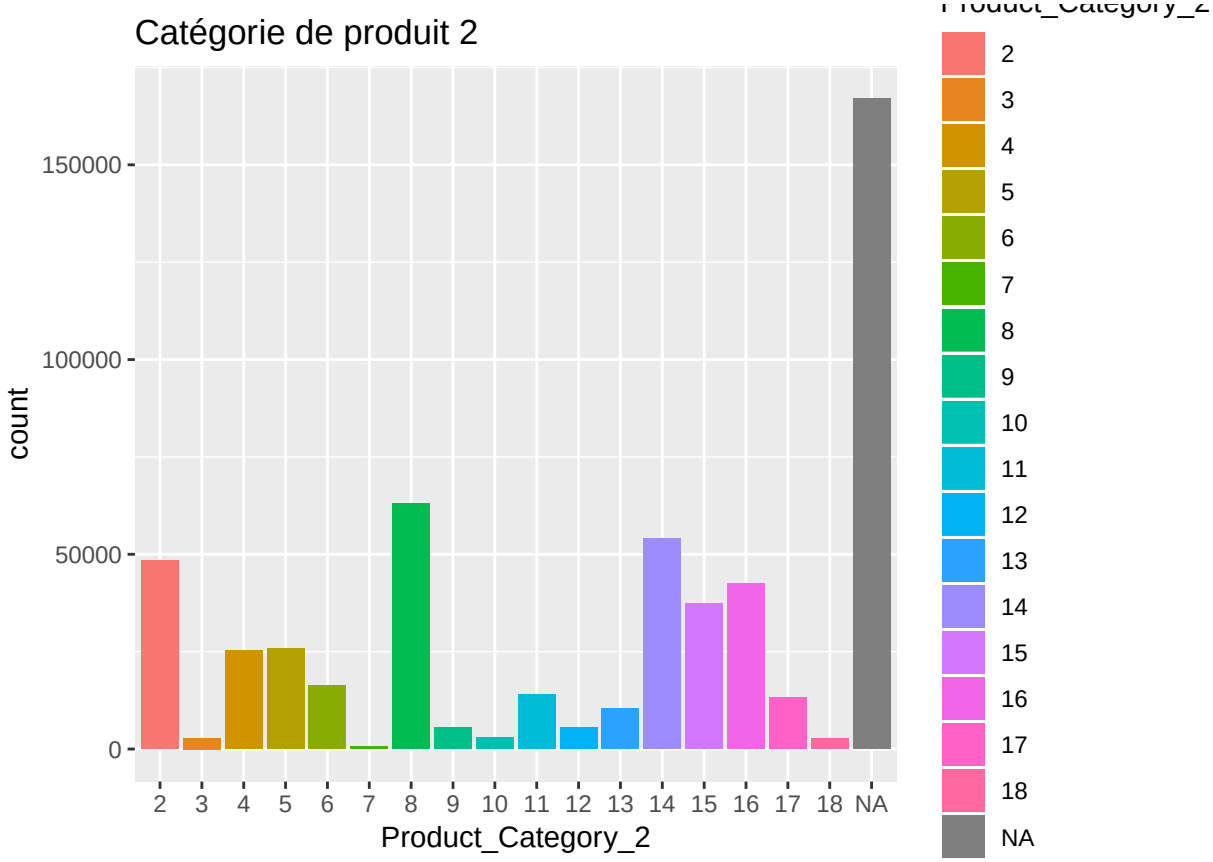
Ancienneté dans la ville

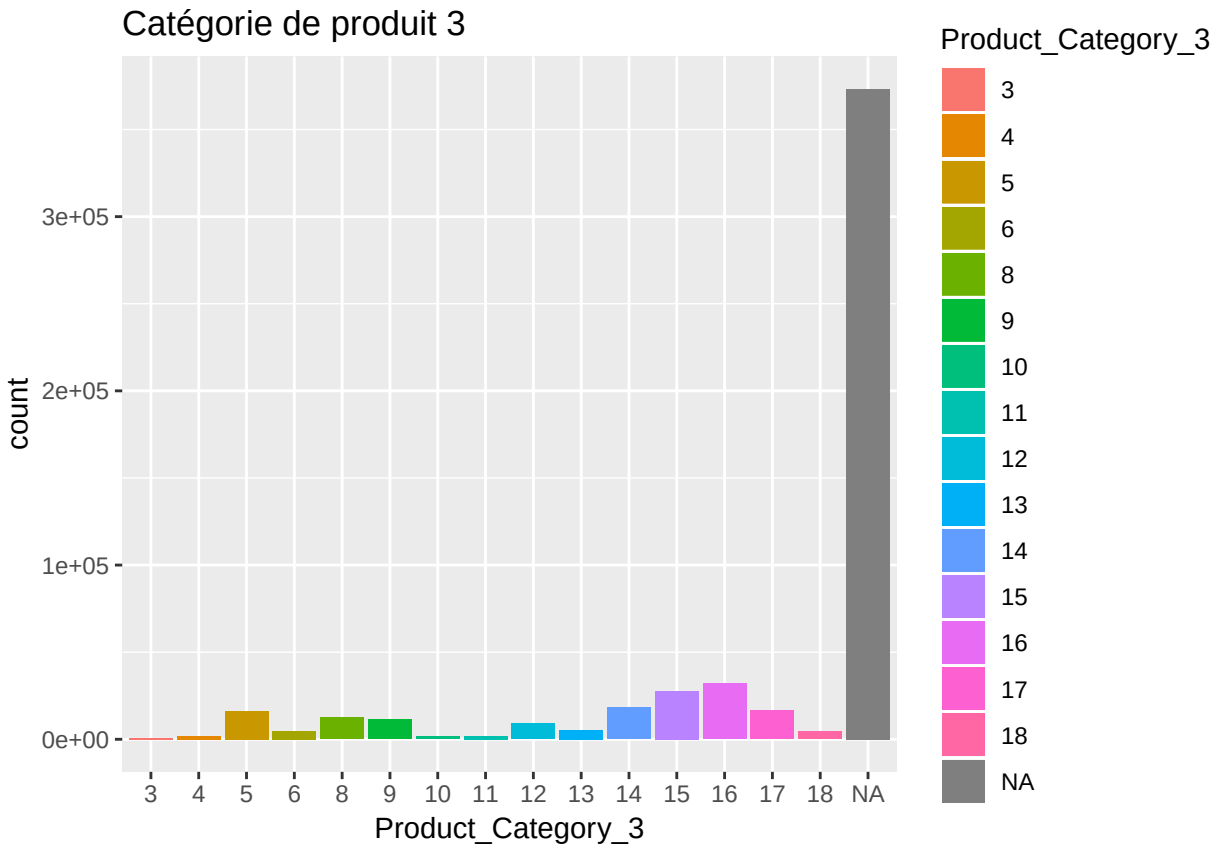


Situation conjugale

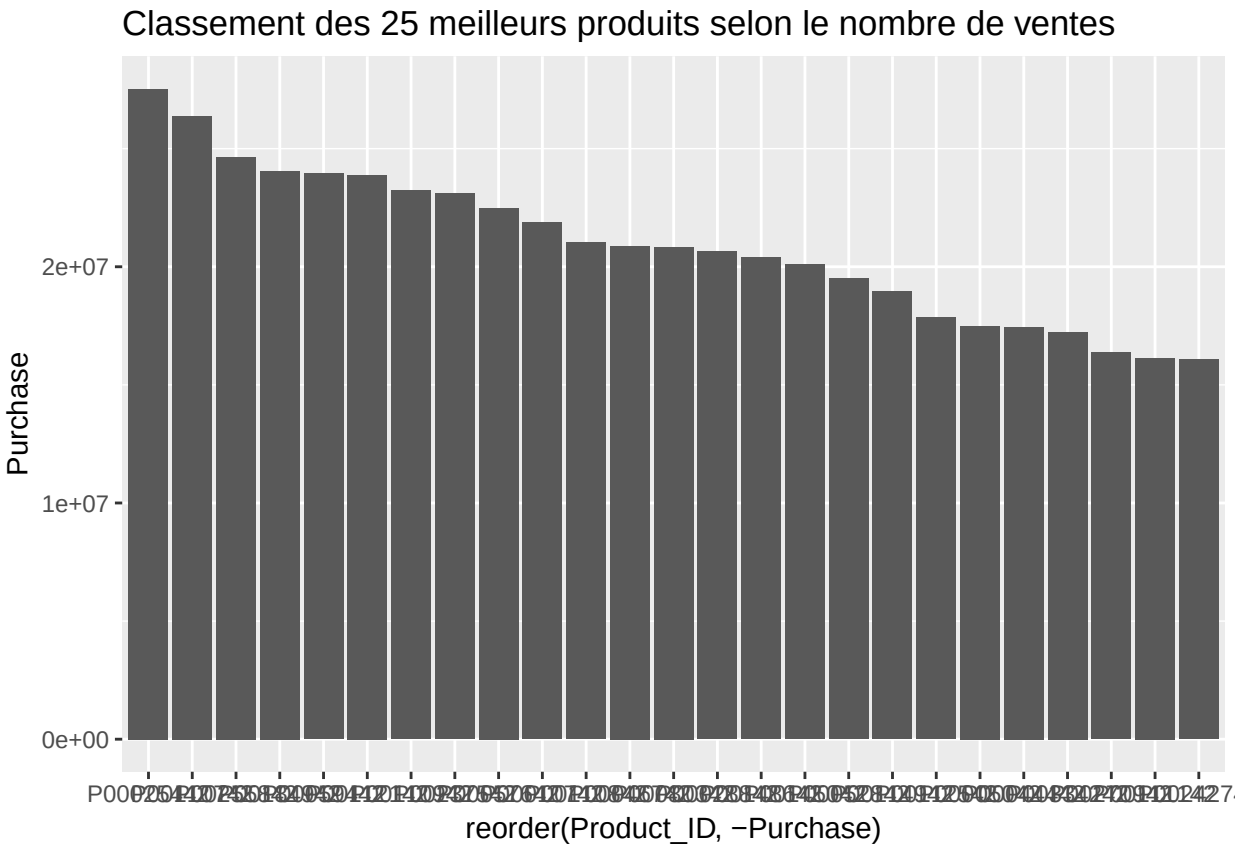


Catégories des produits


Catégorie de produit 2
count
150000
100000
50000
0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 NA
Product_Category_2
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
NA


Catégorie de produit 3
count
3e+05
2e+05
1e+05
0e+00
3 4 5 6 8 9 10 11 12 13 14 15 16 17 18 NA
Product_Category_3
Product_Category_3
3
4
5
6
8
9
10
11
12
13
14
15
16
17
18
NA

Performance des produits



Pas de produit que se détache du reste (du moins dans les 25 premiers, il faut garder à l'esprit qu'il y en a +3000).

Relations

Sociologie

- Age
- Métier
- Ancienneté dans la ville
- Marié, pas marié
- Sexe

=> Montant achat en fonction de l'age, métier, etc...

Commercial

- Performance des produit (via ID)
- Catégorie de produit qui se vendent le mieux

Mesure de l'entropie et du gain d'entropie

TODO: A expliquer et justifier et synthétiser.

```
##                    Age                 Gender
##          0.069374015           0.022694835
```

```
## Stay_In_Current_City_Years                 Occupation
##                 0.013115811                0.091392940
##             Marital_Status              City_Category
##                 0.007119339                0.019069753
```

Deux catégories importantes : Age et Occupation.