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.

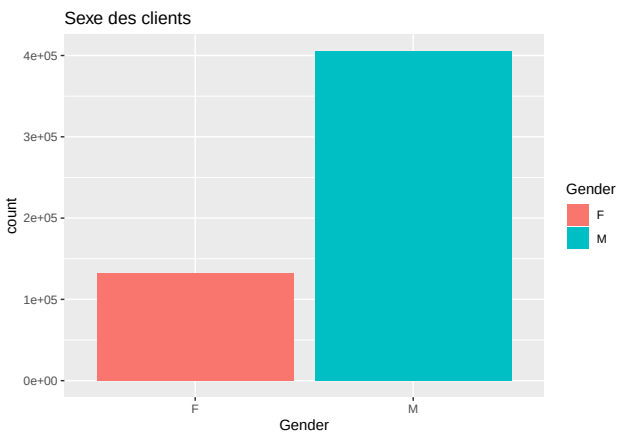
Types et granularité

Remarque :

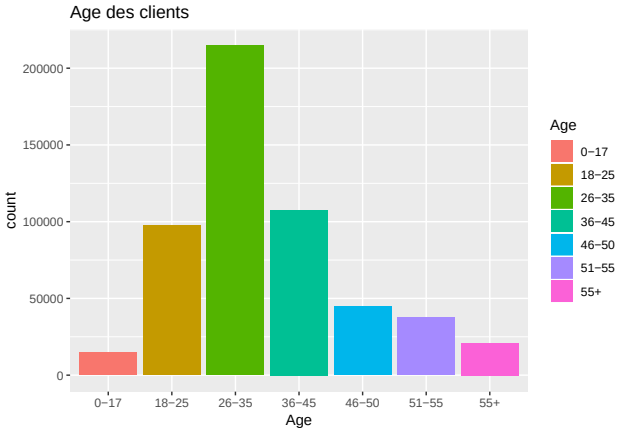
- Un produit a obligatoirement une catégorie (Product_Category_1 ne contient pas de NA). Il n'a pas obligatoirement de 2ème et 3ème catégorie.

Distribution

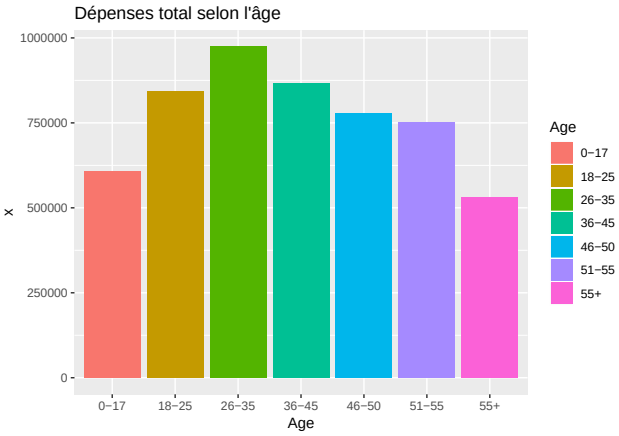
Sexe



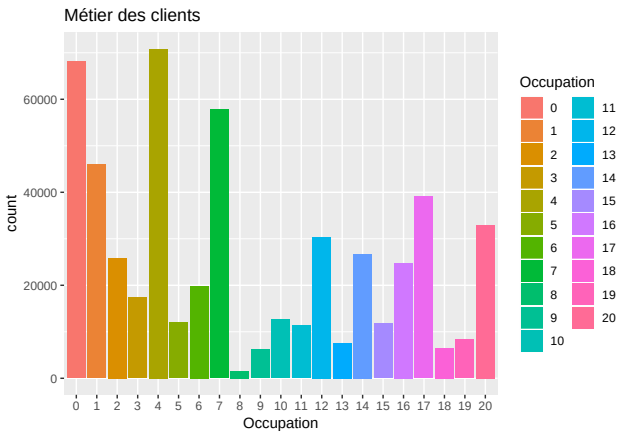
Age



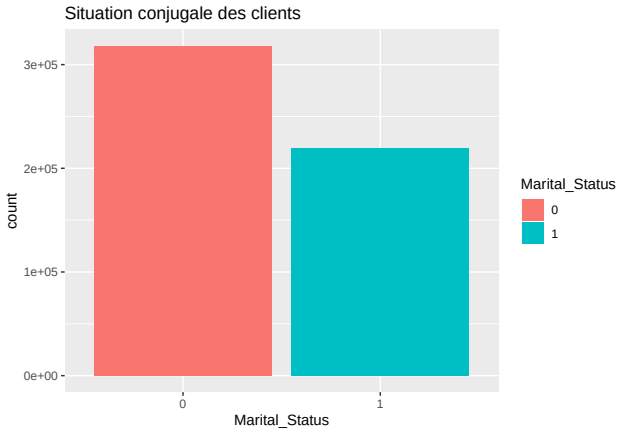
La distribution de l'age dans une population suit habituellement une loi normale. On observe ici que ce n'est pas le cas. Il apparait que la tranche d'âge des 26-35 effectue le plus d'achats lors du Black Friday.



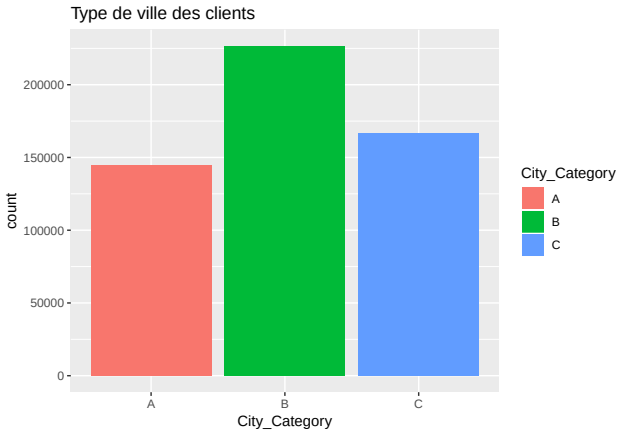
Métier



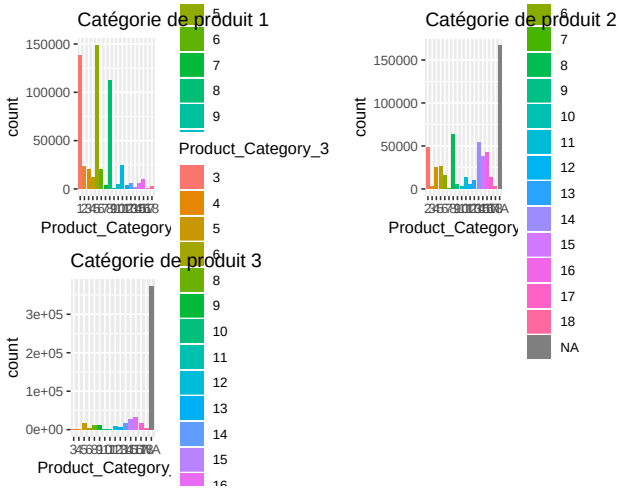
Situation conjugale



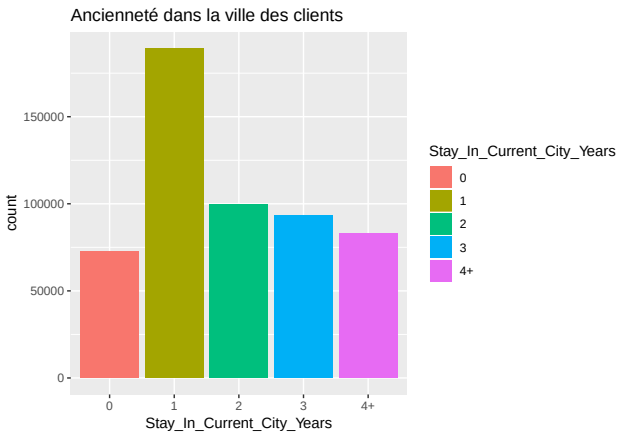
Type de ville



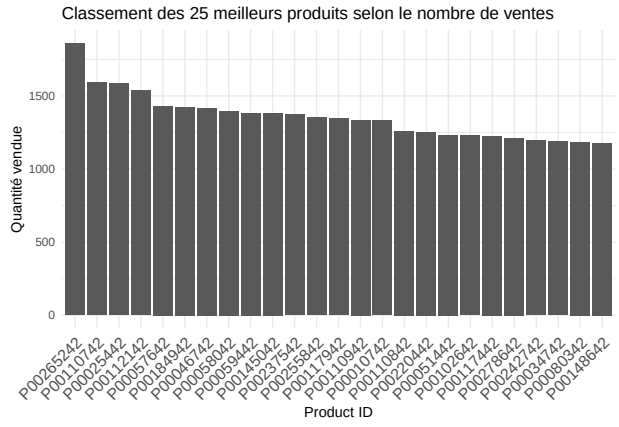
Catégories des produits



Ancienneté dans la ville



Performance des produits



Le premier produit se détache du reste.

Corrélation entre les variables

Mesure de l'entropie et du gain d'entropie

TODO: A expliquer et justifier et synthétiser.

```
##                        Age
##                0.069374015
## Stay_In_Current_City_Years
##                0.013115811
##             Marital_Status
##                0.007119339
```

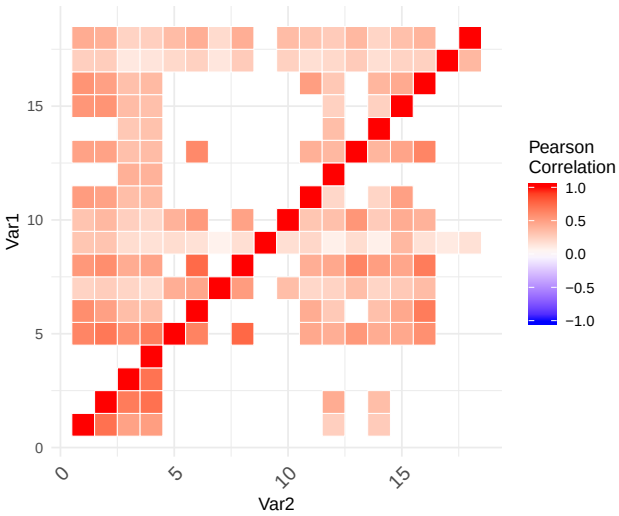
```
##                 Occupation
##                0.091392940
##              City_Category
##                0.019069753
```

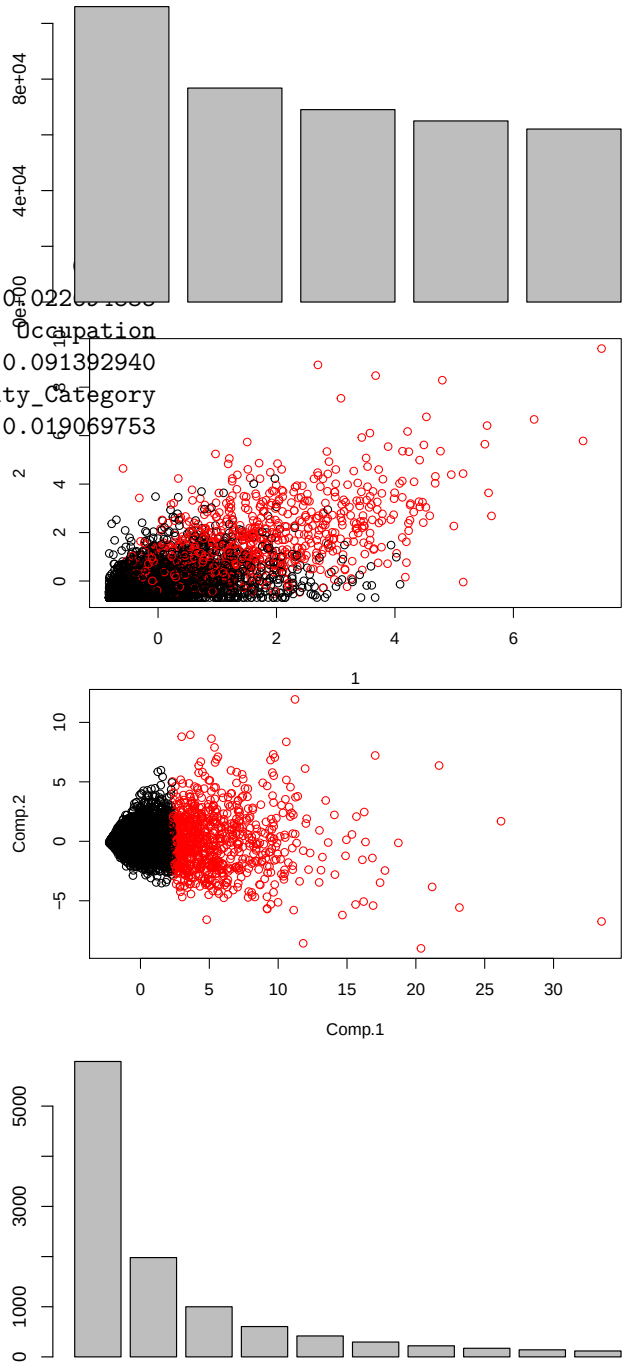
Classification

Notre jeu de données concerne les achats effectués par des consommateurs lors de la période du Black Friday, on peut imaginer que l'on souhaite tirer de ces données, des groupes de consommateur.

- On dispose déjà de certains groupes, par exemple la variable sexe en forme un, le métier également. On peut donc essayer de déterminer si ces groupes sociaux ont des comportements de consommation similaires. C'est un problème de classification supervisée.
- On peut également souhaiter déterminer des groupes indépendamment des paramètres sociaux. C'est un problème de classification non supervisée.

Classificatin non supervisée





Classificatin supervisée

```
## [1] 1.186441
```

Régression logistique

```
## Loading required package: lattice
##
```

```
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```