STA212 - Méthodes de rééchantillonnage

Enseignant: Mohammed Sedki

Devoir : aspects pratiques

Romin DURAND
Loukman Eltarr

May 9, 2020

Arbre de décision unique

```
setwd('~/Cours/STA212/STA212DM')
rm(list = objects())
graphics.off()
OJ=read.csv("oj.csv", header = TRUE)
#View(OJ)
```

On regarde la nature de nos données. On a 1070 observations pour 18 variables différentes. Les variables categorielles sont `Purchase` qui admet deux niveaux, et `Store 7` qui admet aussi deux niveaux. Les autres sont numériques.

```
str(OJ)
```

```
## 'data.frame':    1070 obs. of  18 variables:
##  $ Purchase      : Factor w/ 2 levels "CH","MM": 1 1 1 2 1 1 1 1 1 1 ...
##  $ WeekofPurchase: int  237 239 245 227 228 230 232 234 235 238 ...
##  $ StoreID       : int  1 1 1 1 7 7 7 7 7 7 ...
##  $ PriceCH       : num  1.75 1.75 1.86 1.69 1.69 1.69 1.69 1.75 1.75 1.75 ...
##  $ PriceMM       : num  1.99 1.99 2.09 1.69 1.69 1.99 1.99 1.99 1.99 1.99 ...
##  $ DiscCH        : num  0 0 0.17 0 0 0 0 0 0 0 ...
##  $ DiscMM        : num  0 0.3 0 0 0 0 0.4 0.4 0.4 0.4 ...
##  $ SpecialCH     : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ SpecialMM     : int  0 1 0 0 0 1 1 0 0 0 ...
##  $ LoyalCH       : num  0.5 0.6 0.68 0.4 0.957 ...
##  $ SalePriceMM   : num  1.99 1.69 2.09 1.69 1.69 1.99 1.59 1.59 1.59 1.59 ...
##  $ SalePriceCH   : num  1.75 1.75 1.69 1.69 1.69 1.69 1.69 1.75 1.75 1.75 ...
##  $ PriceDiff     : num  0.24 -0.06 0.4 0 0 0.3 -0.1 -0.16 -0.16 -0.16 ...
##  $ Store7        : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 2 2 2 2 ...
##  $ PctDiscMM     : num  0 0.151 0 0 0 ...
##  $ PctDiscCH     : num  0 0 0.0914 0 0 ...
##  $ ListPriceDiff : num  0.24 0.24 0.23 0 0 0.3 0.3 0.24 0.24 0.24 ...
##  $ STORE         : int  1 1 1 1 0 0 0 0 0 0 ...
```

Analyse Univariée et Bivariée

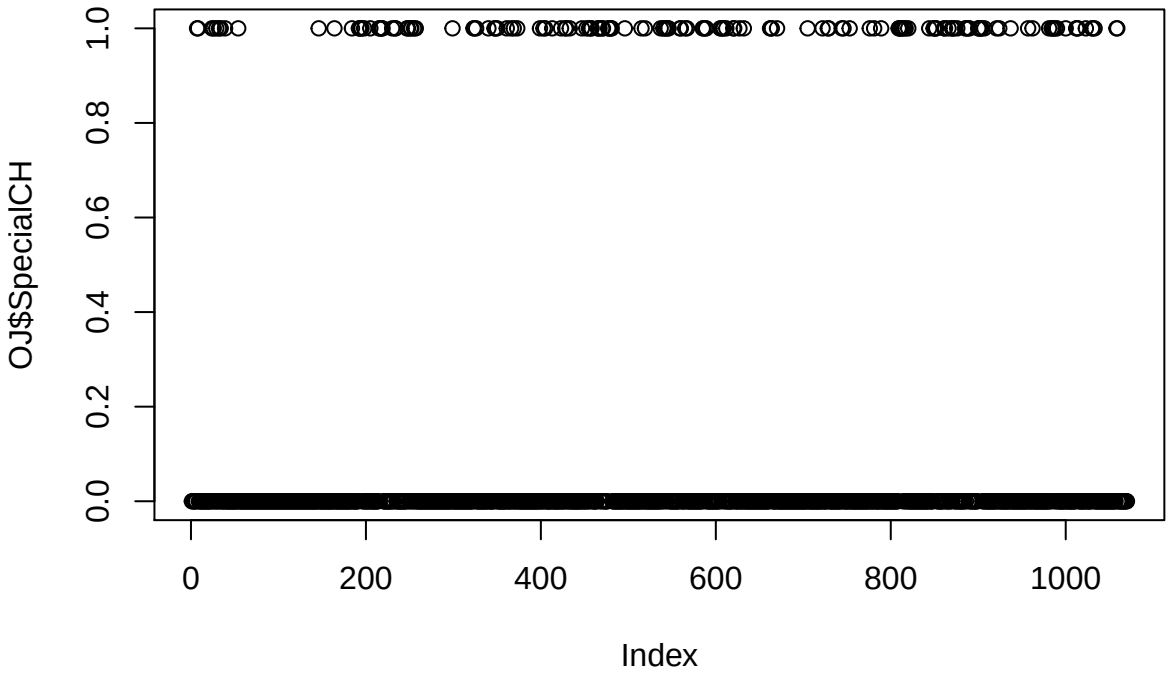
On procéde à une analyse univariée des variables. On se sert de la description des variables ainsi que des commandes `summary`, `plot` et `table`.

Par exemple, on peut voir que les variables `SpecialCH` et `SpecialMM` prennent seulement les valeurs 0 et 1.

```
table(OJ$SpecialCH)
```

```
##
##   0   1
## 912 158
```

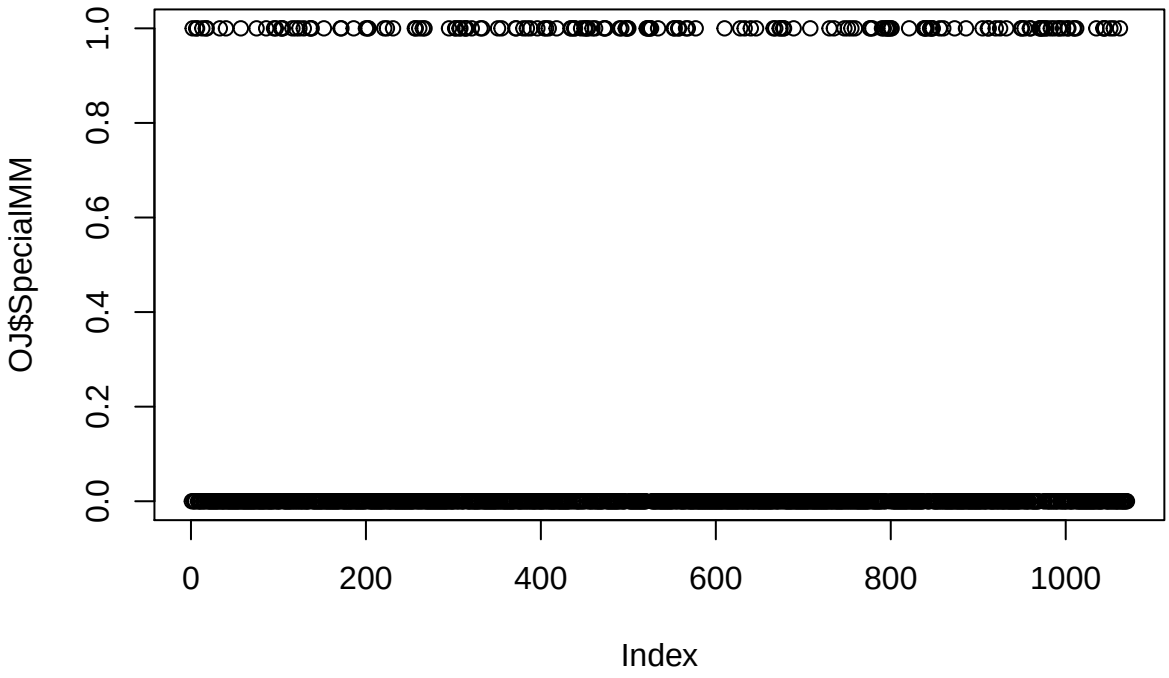
```
plot(OJ$SpecialCH)
```




```
table(OJ$SpecialMM)
```

```
##
##   0   1
## 897 173
```

```
plot(OJ$SpecialMM)
```

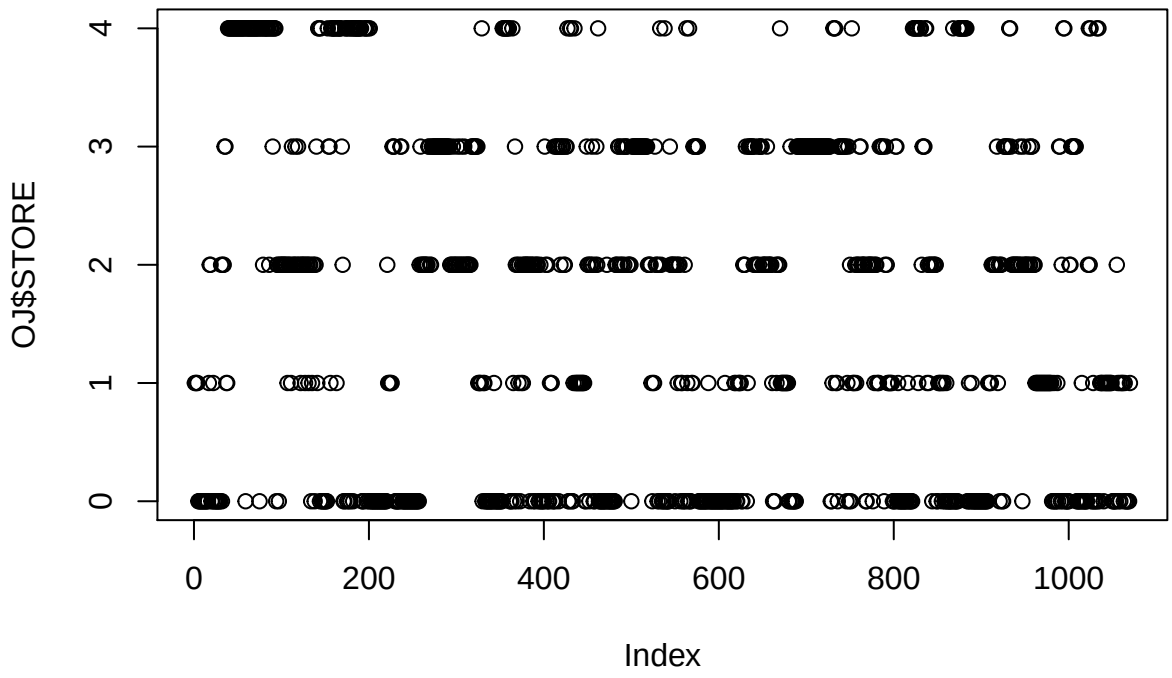


De la même manière `STORE` ne prend que les valeurs entre 0 et 4.

```
table(OJ$STORE)
```

```
##
##   0   1   2   3   4
## 356 157 222 196 139
```

```
plot(OJ$STORE)
```



On préfère alors les transformer en variables catégorielles:

```
OJ$SpecialMM <- as.factor(OJ$SpecialMM)
OJ$SpecialCH <- as.factor(OJ$SpecialCH)
OJ$STORE <- as.factor(OJ$STORE+1) ## On préfère avoir des valeurs entre 1 et 5.
```

Ensuite, on voit que la variable `PriceDiff` est une combinaison des deux variables `SalePriceMM` et `SalePriceCH`. On décide alors de la retrouver.

```
OJ <- subset(OJ, select=-PriceDiff)
```