Facteurs influençants la prise des transports en commun pour l'agglomération grenobloise

RACHIDI Mustapha & SAUNIER Florent & SAADALLAH Malek

Janvier 2023

```
#J'ai tenté l'image
#knitr::include_graphics("Image_BUS_TRAM.jpg")
```

Introduction

Ce projet se base sur des données récoltées en 2010 dans la région Grenobloise. L'étude a pour but de déterminer les facteurs influençants la prise des transports en commun . Pour cela nous nous sommes pris comme limites : le réseau Mtag qui comprend les bus qualifiés de "ville" (Nous n'avons pas pris en compte les bus régionaux comme par exemple le bus Grenoble - Chamrousse) et le réseau du trammway dont les lignes depuis 2010 ont été augmentées.

Articles de la littérature

Familiarisation avec la base de données

La base de données contient 30 702 lignes et 116 colonnes ce qui correspond à nos variables , on peut la qualifier de base de données "moyenne" mais qui saura nous occuper. Concernant le nombre de valeurs manquantes, toutes variables confondues nous avons 971 658 valeurs manquantes soit 27.3% de notre base de données. De plus, 0% des lignes ont toutes leurs valeurs et c'est 21% des colonnes qui n'ont pas de valeurs manquantes. Il peut être intéressant de voir où sont les valeurs manquantes.
L'échantillon comporte 5189 personnes

Visualisation valeurs manquantes titre à changer peut ête

En annexe, quelques graphiques permettant de visualiser quelles variables ont le plus de valeurs manquantes. Ces graphiques nous permettrons d'adopter un regard critique sur les variables que nous choisirons par la suite. Cependant, on peut établir quelques critères avec r : ration de valeurs manquantes dans la colonne.

Bon : r<=5% Moyen : 5%<r<=20% Mauvais 20%<r<=45% Très mauvais : r>45%

Plusieurs variables ont entre 80% 99% de valeurs manquantes JAI TROUVÉ PQ c'est jusque que beaucoup de gens n'ont tout simplement pas plus de 1 véhicule, ce qui fait que les variables correspondantes sont vides. À CHANGER

Variables du projet

Frecqtcu : Variable d'intérêt (Y) catégorielle qui indique la fréquence d'utilisation des transports en communs chez une personne.
Elle prend les valeurs :

1 : Utilisation des transports en commun tous les jours
2 : Utilisation des transports en commun au moins deux fois par semaine
3 : Utilisation des transports en commun au moins deux fois par mois
4: Utilisation des transports en commun très rare
5 : Utilisation des transports en commun inexistante

Nous avons décidié de construire frecqtcu de manière à ce qu'elle prenne la valeur 0 ou 1

```
 DB_projet_full<-DB_projet_full%>%mutate(freqtcu=ifelse(freqtcu<=3,1,0))
DB_projet_full$freqtcu<-factor(DB_projet_full$freqtcu)
```

Pour toutes les personnes qui prennent les transports de manière : régulière/tous les jours , au moins deux fois par semaine et au moins deux fois par mois se sont vues attribuées la valeur 1 car le "au moins" présage une prise des transports en communs plus élevée.

Tailmng : Variable qui indique le nombre de personnes composant le ménage.

```
DB_projet_full<-rename(DB_projet_full,"tailmng"="NO_PERS")
```

On change simplement le nom de la variable “NO_PERS” qui indique le nombre de personne dans le ménage
Permis :Variable indiquant si la personne effectuant le trajet possède le permis ou pas.

```
DB_projet_full<-DB_projet_full%>%mutate(permis=ifelse(any(permis==1 | permis==3),"YES","NO"))
DB_projet_full$permis<-factor(DB_projet_full$permis)
```

Car_ownership : Variable indiquant si la personne effectuant le trajet possède une voiture

```
DB_projet_full<-DB_projet_full%>%mutate(car_ownership=ifelse(DB_projet_full$VP_DISPO>0 & (DB_projet_full

DB_projet_full$car_ownership<-factor(DB_projet_full$car_ownership)
```

Cette variable dépent de trois variables qui sont VP_dispo qui doit être strictement supérieur à 0, puis GENRE (type de véhicule utilisé) , nous avons exclu les campings cars car notre sujet se prête au milieu urbain et de POSSE (Est ce que la voiture appartient à la personne).Nous nous sommes contentés de prendre exclusivement les véhicules possedés par la personne.

Création de la nouvelle base de données

Variables complémentaires
Grâce aux variables précédentes et aux articles que l’on a trouvé dans la littérature, nous allons construire notre base de données pour notre modèle.
Nous exploiterons un ensemble de caractéristiques socio-économiques puis certaines variables liées au “confort” du trajet.

Restriction géographique Définissons ce que l’on entend par “transports en communs”.
Pour notre étude nous nous concentrons sur les transports en communs de la société MTag,c’est à dire les tram et bus du réseau.
Notre délimitation géographique sera simplement les terminaux des différentes lignes de tram/bus confondues.
Par la suite, quand on parlera de transports en communs, on se refère à la définition au dessus.

Toutes les zones répertoriées dans le vecteur “Vec_zone” ont au moins un arrêt du réseau Mtag.

vérifier notre choix avec un clustering

```
test <- New_DB %>% mutate(is_in_zones = ifelse(tir %in% Vec_zone, 100, 0))
library(dbscan)
```

```
## Warning: package 'dbscan' was built under R version 4.3.2
```

```
##
## Attaching package: 'dbscan'
```

```
## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

```
a <- nrow(filter(test, is_in_zones == 100))
b <- nrow(filter(test, is_in_zones == 0))
print(b)
```

```
## [1] 17444
```

```
print(a)
```

```
## [1] 13258
```

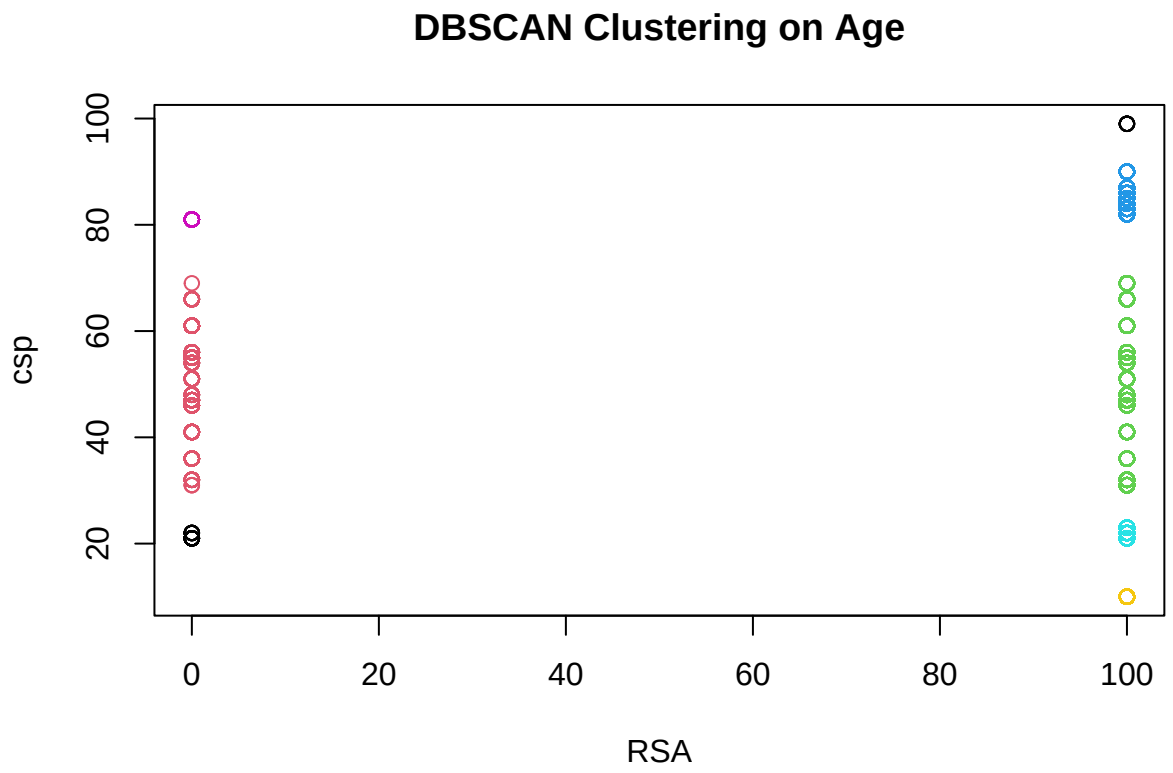
```
test_age_db <- test %>% mutate(RSA = ifelse(OCCU1 == "RSA", 0, 100))
test_age_db <- select(test_age_db, RSA, csp)
```

```r
test_age_db <- subset(test_age_db, !is.na(RSA), !is.na(csp))
summary(test_age_db)
```

```
##       RSA              csp
##  Min.   :  0.00   Min.   :10.00
##  1st Qu.:100.00   1st Qu.:41.00
##  Median :100.00   Median :54.00
##  Mean   : 96.51   Mean   :56.05
##  3rd Qu.:100.00   3rd Qu.:82.00
##  Max.   :100.00   Max.   :99.00
```

```r
dbscan_age <- dbscan(test_age_db, eps = 5, minPts = 100 )

plot(test_age_db, col = dbscan_age$cluster + 1L, main = "DBSCAN Clustering on Age")
```




```r
New_DB<-dplyr::filter(New_DB,tir %in% Vec_zone) #on garde que les zones où il y a des transports en com
head(New_DB)
```

```
##   tir NO_MEN        TYPE_HAB        TYPE_OCU LIEU_STAT1 TYPE_STAT1  sexe age
## 1 101     12 GRAND_COLLECTIF AUTRE_LOCATAIRE          0          0 Homme  24
## 2 101     12 GRAND_COLLECTIF AUTRE_LOCATAIRE          0          0 Homme  24
## 3 101     12 GRAND_COLLECTIF AUTRE_LOCATAIRE          0          0 Homme  24
## 4 101     12 GRAND_COLLECTIF AUTRE_LOCATAIRE          0          0 Homme  24
## 5 101     12 GRAND_COLLECTIF AUTRE_LOCATAIRE          0          0 Homme  24
## 6 101     12 GRAND_COLLECTIF AUTRE_LOCATAIRE          0          0 Homme  24
##   OCCU1   OCCU2 csp ABO_TC travdom PBM_STAT_GENERAL STAT_TRAV duree nbmodemec
## 1   RSA NON_Con  55    NON    <NA>             <NA>      <NA>     5         0
```

```
## 2   RSA NON_Con  55    NON    <NA>             <NA>      <NA>     5          0
## 3   RSA NON_Con  55    NON    <NA>             <NA>      <NA>    20          0
## 4   RSA NON_Con  55    NON    <NA>             <NA>      <NA>    20          0
## 5   RSA NON_Con  55    NON    <NA>             <NA>      <NA>    25          0
## 6   RSA NON_Con  55    NON    <NA>             <NA>      <NA>    25          0
##   prisecharge   id_pers    id_depl ntraj btt freqtcu permis tailmng
## 1        <NA> 101012001 1010120011    NA 100       1    YES       1
## 2        <NA> 101012001 1010120012    NA 100       1    YES       1
## 3        <NA> 101012001 1010120013    NA 100       1    YES       1
## 4        <NA> 101012001 1010120014    NA 100       1    YES       1
## 5        <NA> 101012001 1010120015    NA 100       1    YES       1
## 6        <NA> 101012001 1010120016    NA 100       1    YES       1
##   car_ownership
## 1           NON
## 2           NON
## 3           NON
## 4           NON
## 5           NON
## 6           NON
```

Restriction sur l'âge

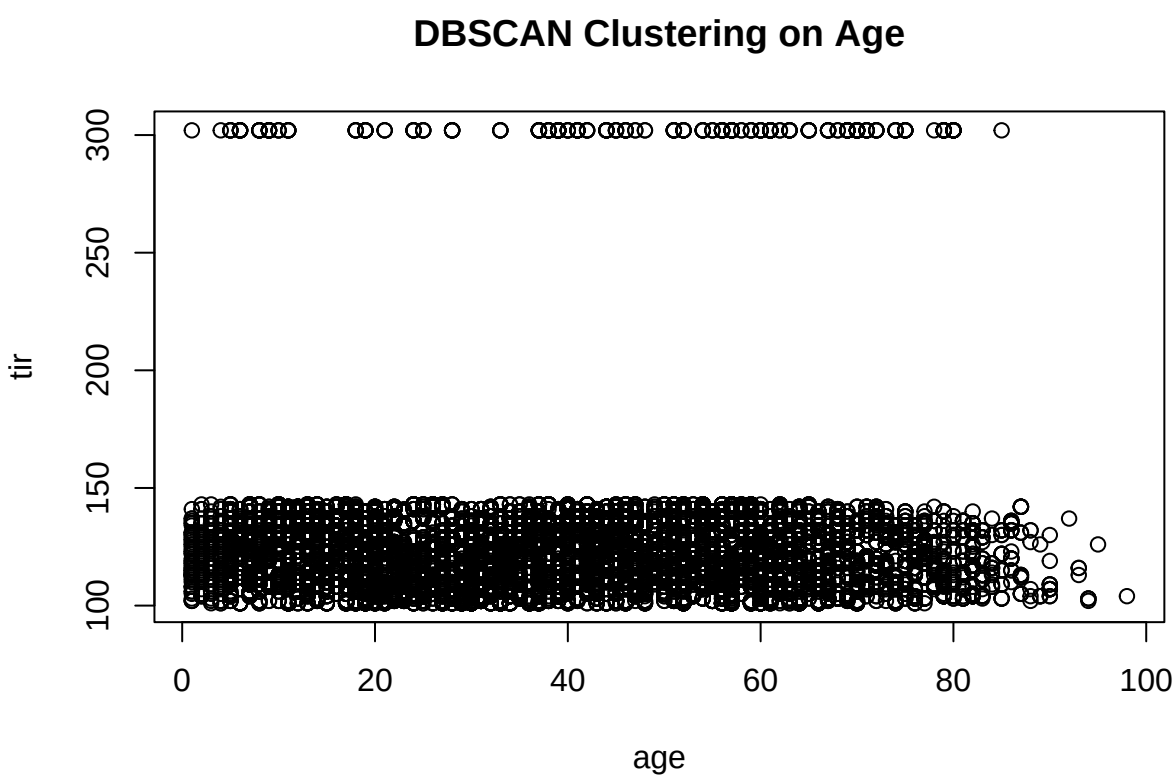
Faire un dbscan pour choisir les tranches d'âge

```
library(dbscan)
test_age_db <- select(New_DB, age, tir)
summary(test_age_db)
```

```
##       age             tir
##  Min.   : 1.00   Min.   :101.0
##  1st Qu.:22.00   1st Qu.:111.0
##  Median :38.00   Median :122.0
##  Mean   :39.32   Mean   :125.5
##  3rd Qu.:55.00   3rd Qu.:132.0
##  Max.   :98.00   Max.   :302.0
```

```
dbscan_age <- dbscan(test_age_db, eps = 1, minPts = 500)

plot(test_age_db, col = dbscan_age$cluster + 1L, main = "DBSCAN Clustering on Age")
```



Il est nécessaire de préciser que les mineurs se déplacent majoritairement via les transports en communs car ils n'ont tout simplement pas le choix...
Pour ne pas être biaisé, il est judicieux de filtrer les mineurs de notre base de données ainsi que les personnes âgées de plus de 80ans.

Notre nouvelle base de données comprend maintenant 10 879 observations et 22 variables ## Analyse Univariée

Analyse Univariée : freqtcu

Dans notre base de données, il y a 46% des gens qui prennent les transports en communs de manière plus ou moins régulière.

Analyse Univariée :permis

Toutes les personnes de notre échantillonnage possède le permis de conduire.

Analyse Univariée : tailmng

Pour ce qu'il en est de tailmng, la moyenne étant plus élevée que la médiane nous avons une asymétrie du côté droit , c'est à dire qu'il y a une concentration plus importante de valeurs à gauche de la moyenne.

Analyse Univariée : car_ownership

84.4% des gens qui ont le permis sont propriétaires d'un véhicule dans notre étude.
Pas de conclusions hâtives, cela sera explicité dans l'analyse bivariée.

Analyse Bivariée

Pour cette partie, nous allons faire appel à plusieurs tests statistiques pour tenter de comprendre les relations qu'il peut y avoir entre nos varialbes.

Le test statistique du Chi^2 est utile pour établir ou non une relation entre deux variables qualitatives. Tandis que le test statistique d

```
DB_var_zg<-New_DB_filtered[!duplicated(New_DB_filtered$id_pers),]
```

Création des variables pour l'analyse bivariée

```
#DB_var_zg : zone grenoble
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Plus_jeune=sum(min(age))) #variable plus jeune de la régi
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_actif=sum(OCCU1=="TravailPleinT" | OCCU1=="TravailPart
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_inactif=sum(OCCU1=="Chomeur" | OCCU1=="Reste_auFoyer"
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_retraites=sum(OCCU1=="Retraite")) #Nb_retraités
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_etu=sum(OCCU1=="Scolaire" | OCCU1=="Etudiant" | OCCU1=

DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_log_collec=sum(TYPE_HAB=="GRAND_COLLECTIF" | TYPE_HAB=
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_log_indiv=sum(TYPE_HAB=="INDIVIDUEL_ISO" | TYPE_HAB=="

DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_proprietaire=sum(TYPE_OCU=="PROPRIETAIRE" | TYPE_OCU=
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_locataire=sum(TYPE_OCU=="LOCATAIRE" | TYPE_OCU=="AUTRE_

DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_stat_parking=sum(LIEU_STAT1=="PARKING PUBLIC" | LIEU_S
#DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_stat_garage=sum(LIEU_STAT1=="GARAGE/BOX"))
#DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_stat_rue=sum(LIEU_STAT1=="RUE"))
```

```
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_stat_parking=sum(LIEU_STAT1=="PARKING PUBLIC" | LIEU_S
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_stat_garage=sum(LIEU_STAT1=="GARAGE/BOX"))
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_stat_rue=sum(LIEU_STAT1=="RUE"))

DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_stat_interdit=sum(TYPE_STAT1=="INTERDIT"))
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_stat_gratuit=sum(TYPE_STAT1=="GRATUIT" | TYPE_STAT1=="
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Nb_stat_payant=sum(TYPE_STAT1=="PAYANT"))

DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Haut_stat_social=sum(csp ==21 | csp==22 | csp==23 |csp==3
DB_var_zg<-DB_var_zg%>%group_by(tir)%>%mutate(Bas_stat_social=sum(csp==10 | csp==47 | csp==48 |csp==56
```

```
##Permis
table(New_DB_filtered$freqtcu,New_DB_filtered$permis)
```

```
##
##      YES
##   0 5823
##   1 5028
```

```
chisq.test(table(New_DB_filtered$freqtcu,New_DB_filtered$permis))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(New_DB_filtered$freqtcu, New_DB_filtered$permis)
## X-squared = 58.246, df = 1, p-value = 2.313e-14
```

```
#cor.test(table(New_DB_filtered$freqtcu,New_DB_filtered$permis))
```

```
table(New_DB_filtered$freqtcu,New_DB_filtered$tailmng)
```

```
##
##        1    2    3    4    5    6
##   0 3345 2125  298   51    0    4
```

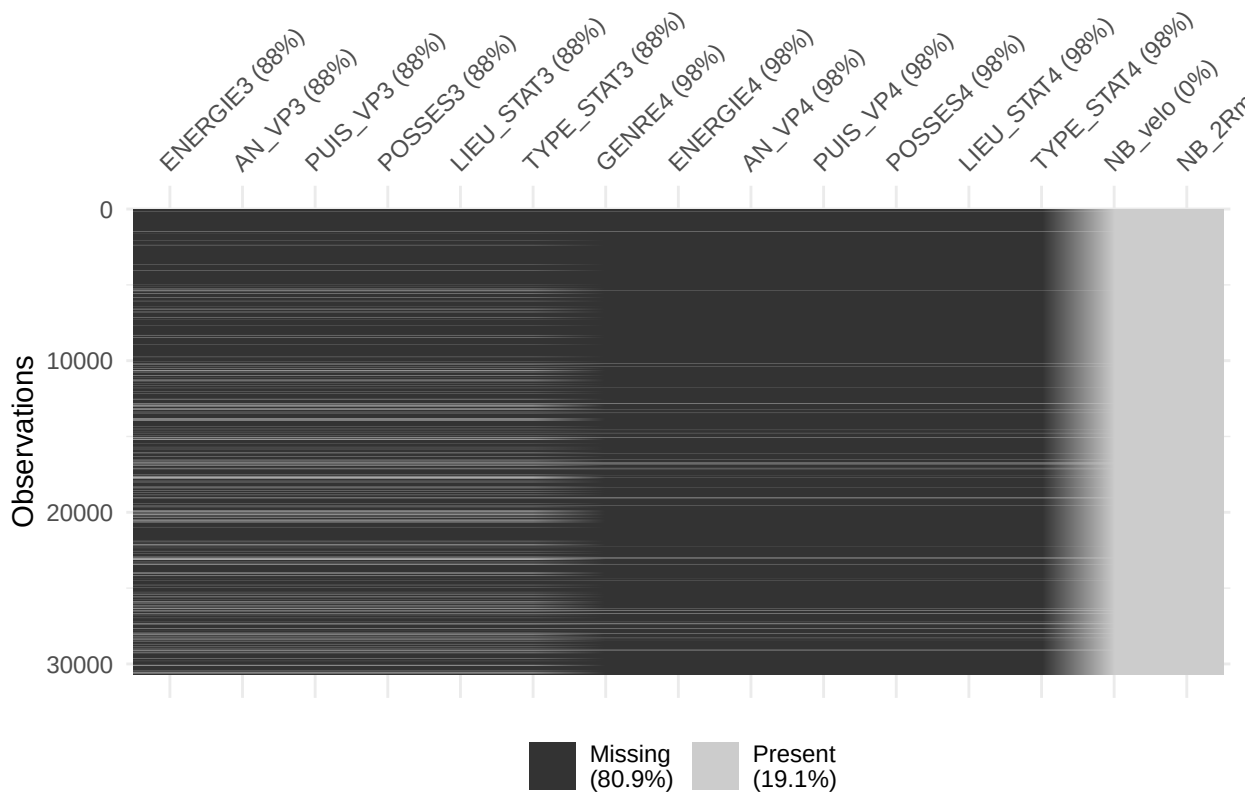
```
##   1 2821 1820  311   72    4    0
```

```
t.test(table(New_DB_filtered$freqtcu,New_DB_filtered$tailmng)
)
```

```
##
##  One Sample t-test
##
## data:  table(New_DB_filtered$freqtcu, New_DB_filtered$tailmng)
## t = 2.4937, df = 11, p-value = 0.02984
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   106.1447 1702.3553
## sample estimates:
## mean of x
##    904.25
```

Annexes

```
data_2<-DB_projet_full[,c(30:44)]
vis_miss(
  data_2,
  cluster = FALSE,
  sort_miss = FALSE,
  show_perc = TRUE,
  show_perc_col = TRUE,
  large_data_size = 9e+06,
  warn_large_data = TRUE
)
```



Listes variables à plus de 80% de valeurs manquantes

-motoracc -situveil -STAT_TRAV -TYPE_STAT4 -LIEU_STAT4 -POSSES4 -PUIS_VP4 -AN_VP4 -ENERGIE4 -GENRE4 -TYPE_STAT3 -LIEU_STAT3 -POSSES3 -PUIS_VP3 -AN_VP3 -ENERGIE3 -motdeacc -nbarret -abonpeage