TP3

zhang

4/22/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Exo1:

1.1)

```
#setwd("~/Users/shurongzhang/Downloads/TPbuzz")
twitter=read_csv2("twitter_small.csv")
```

```
## i Using "’,’" as decimal and "’.’" as grouping mark. Use ‘read_delim()‘ for more control.
```

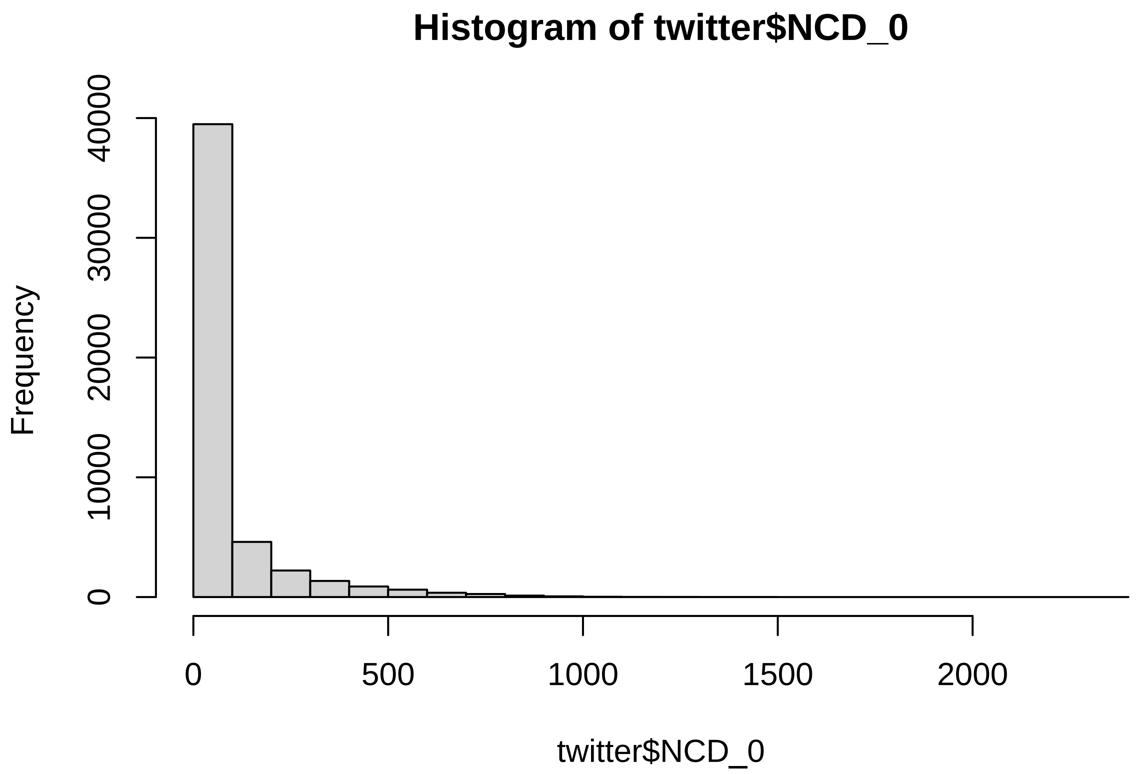
```
## Rows: 50000 Columns: 67
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## dbl (67): NCD_0, NCD_1, NCD_2, NCD_3, NCD_4, NCD_5, AI_0, AI_1, AI_2, AI_3, ...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
#summary(twitter)
nb_var=length(colnames(twitter))
nb_indivi=length(twitter[,1])
```

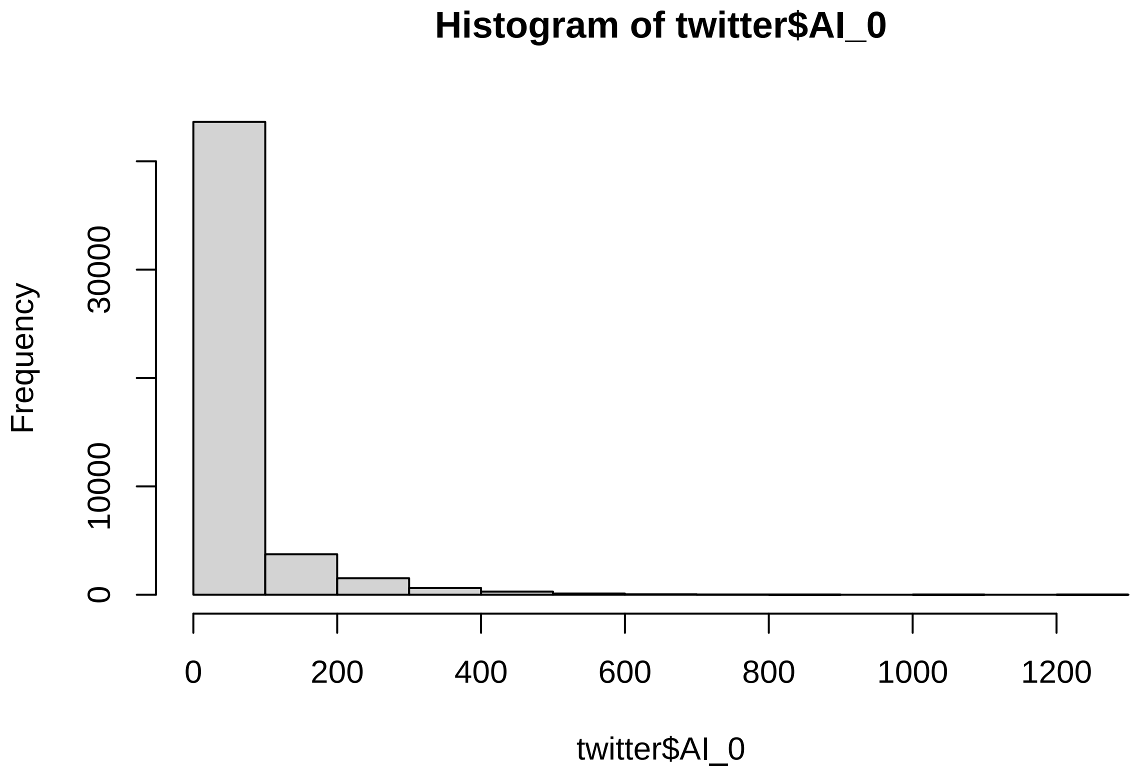
1.2)

Histogramme pour les variables explicatives au temps 0 et pour le label:

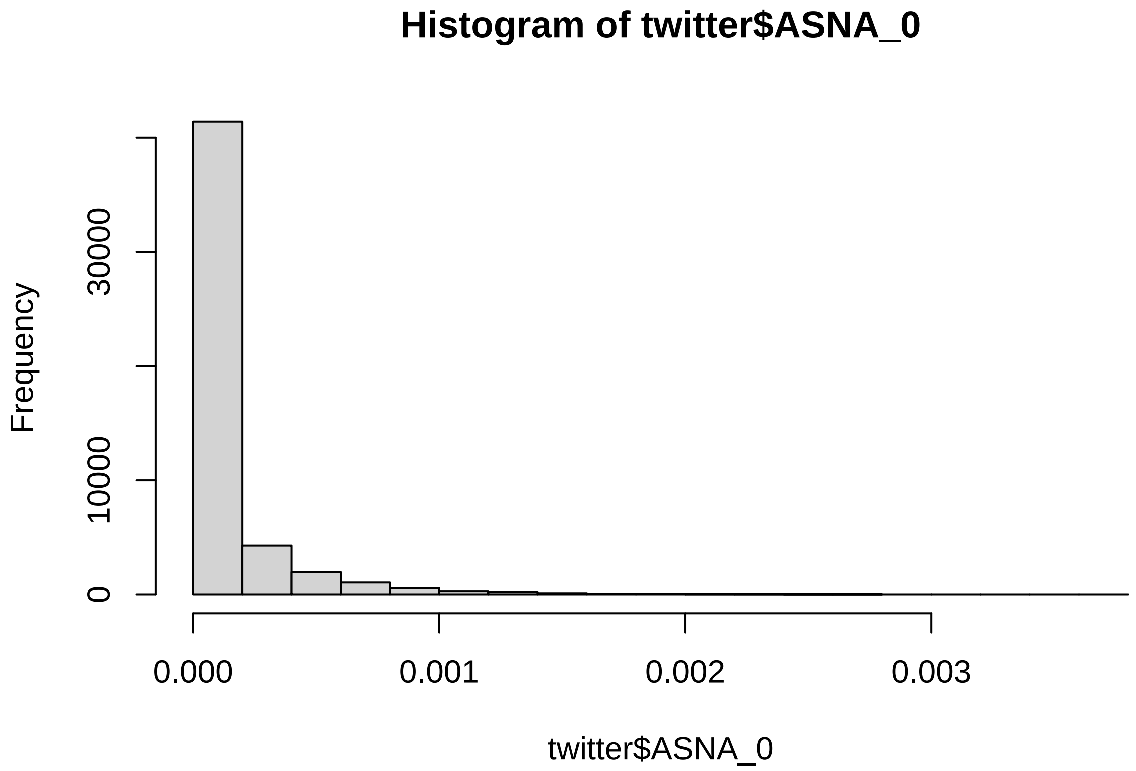
```
hist(twitter$NCD_0)
```



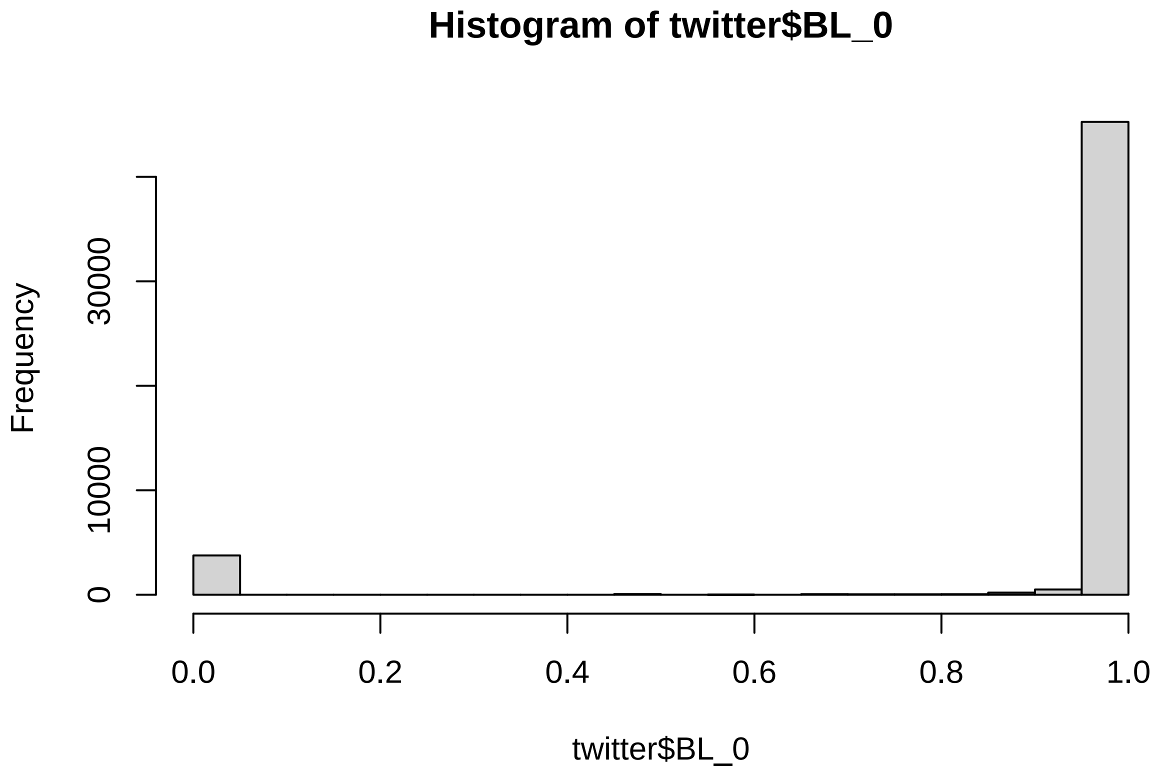
```
hist(twitter$AI_0)
```



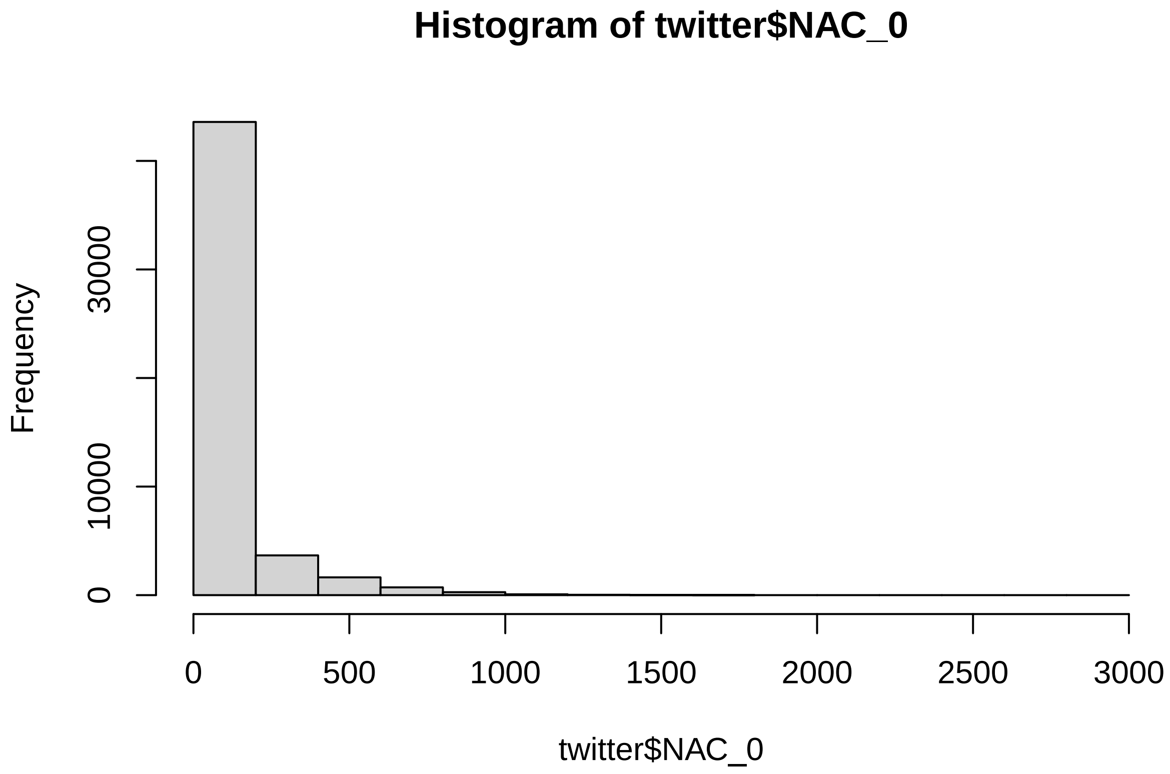
```
hist(twitter$ASNA_0)
```



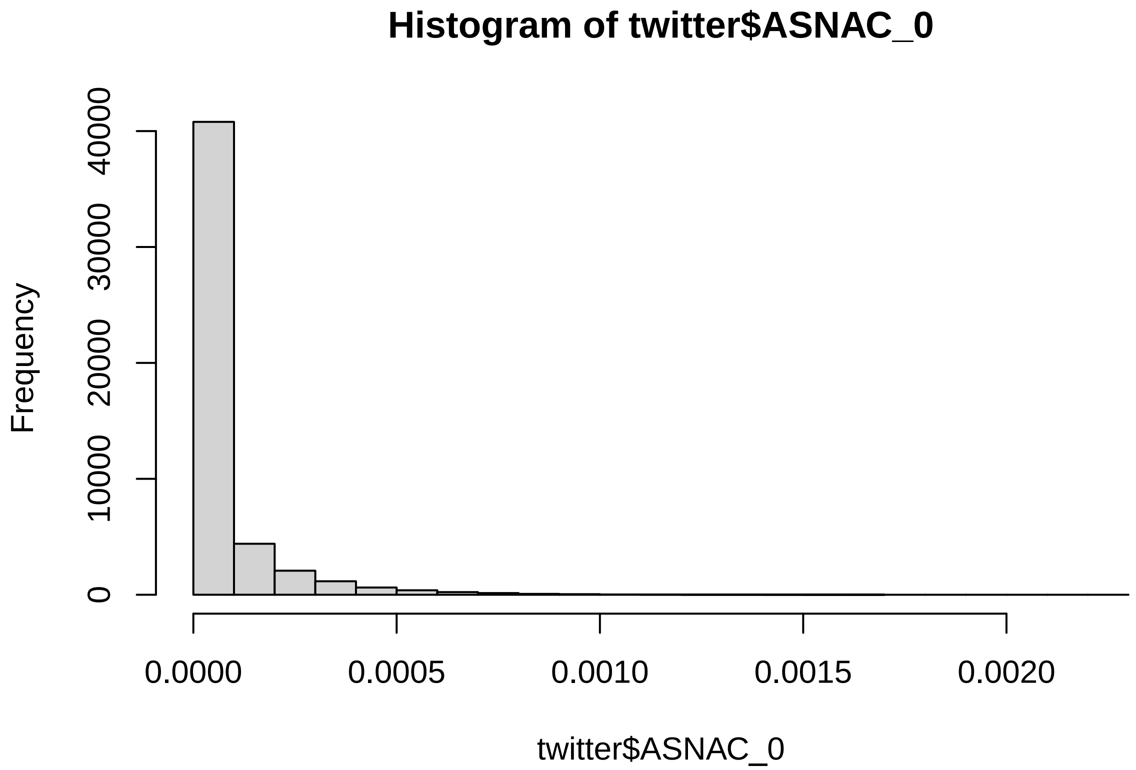
```
hist(twitter$BL_0)
```



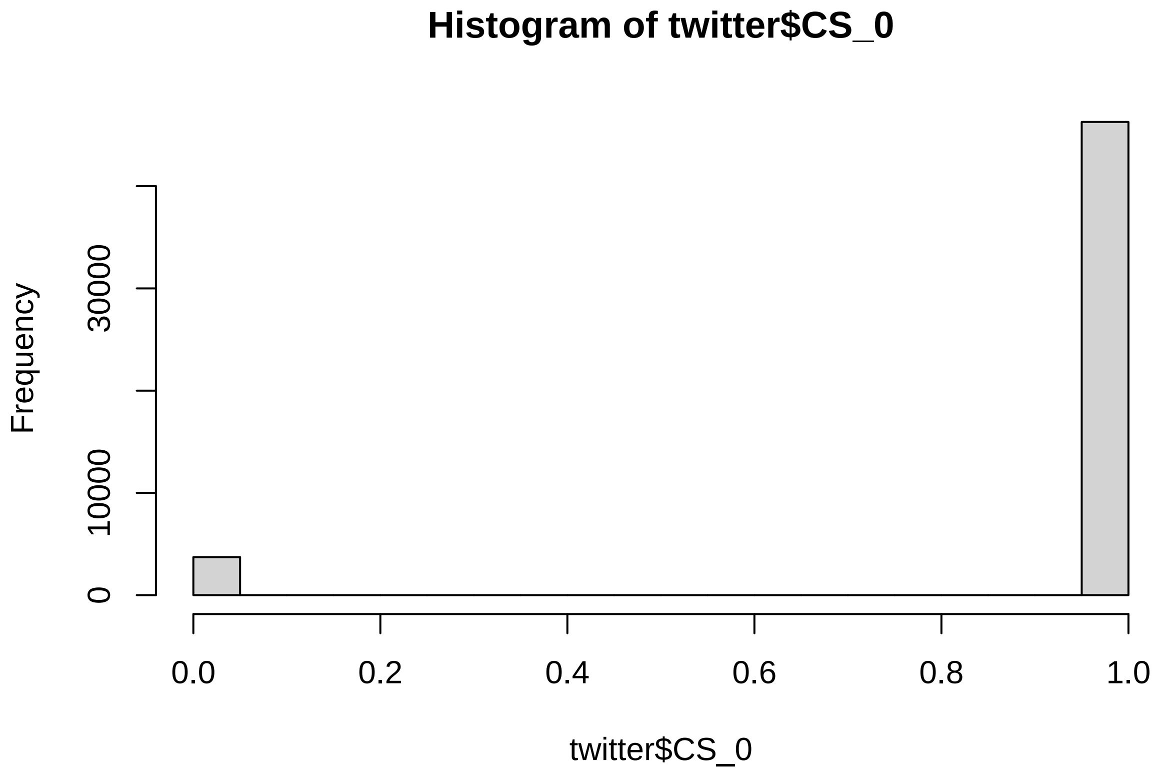
```
hist(twitter$NAC_0)
```



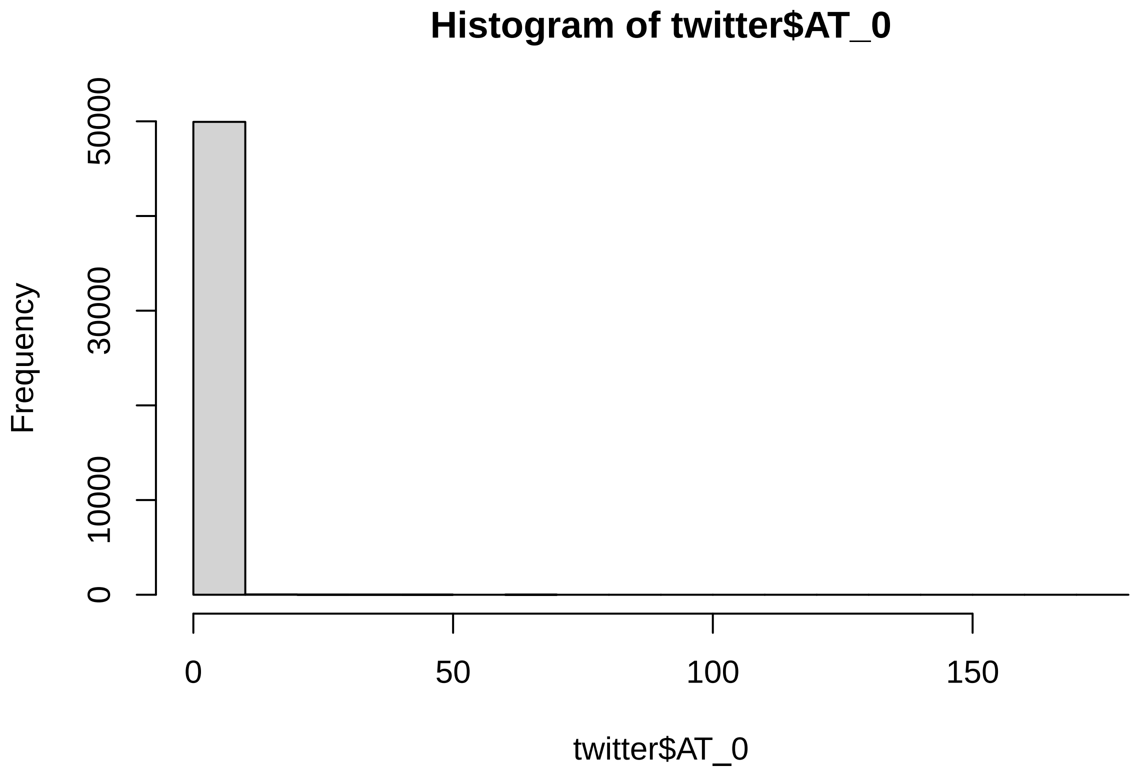
```
hist(twitter$ASNAC_0)
```



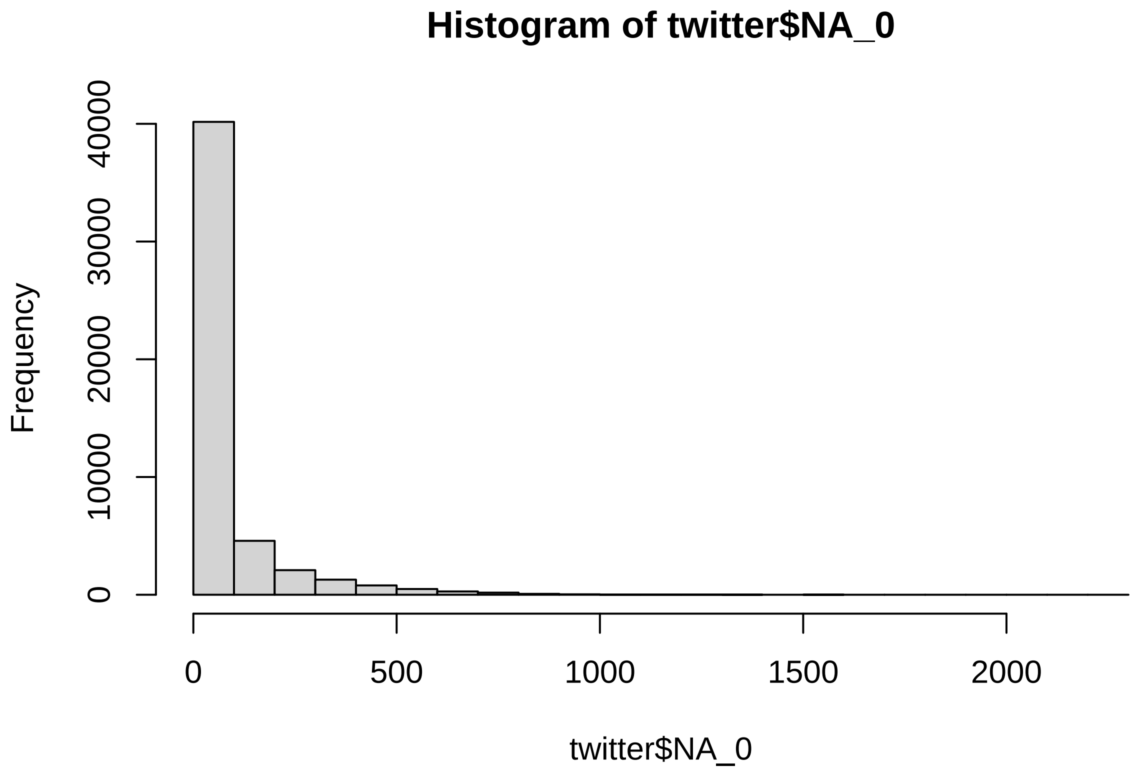
```
hist(twitter$CS_0)
```



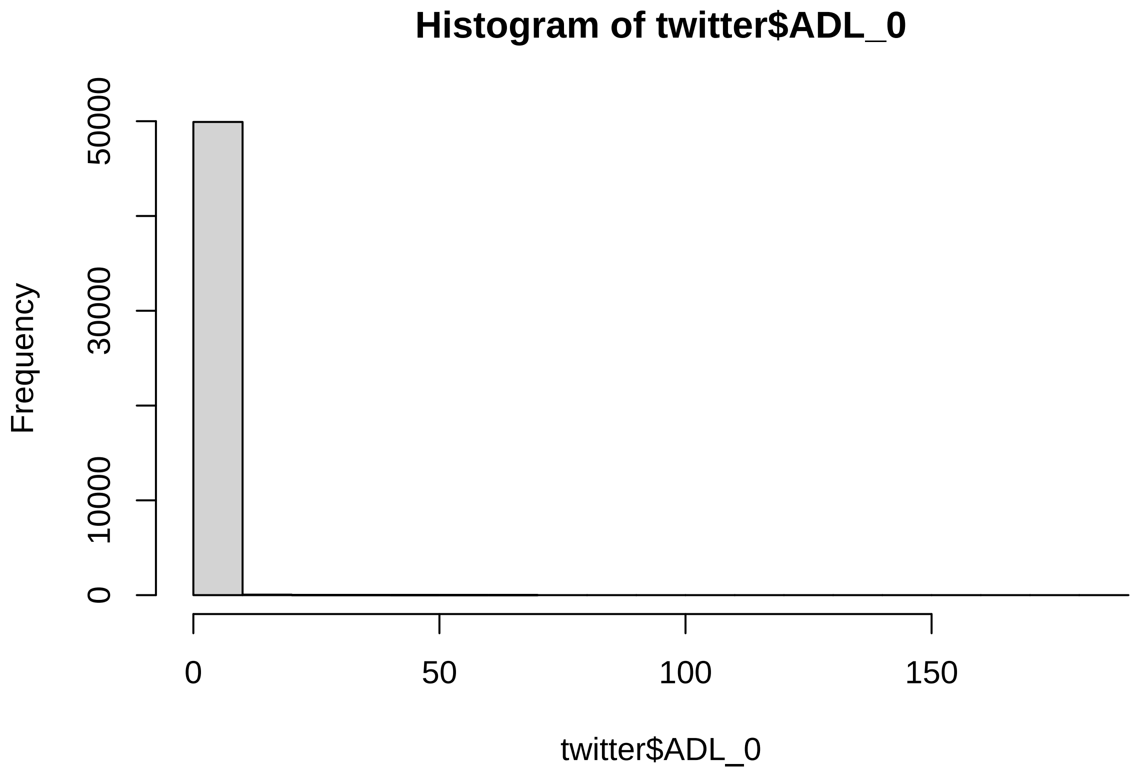
```
hist(twitter$AT_0)
```



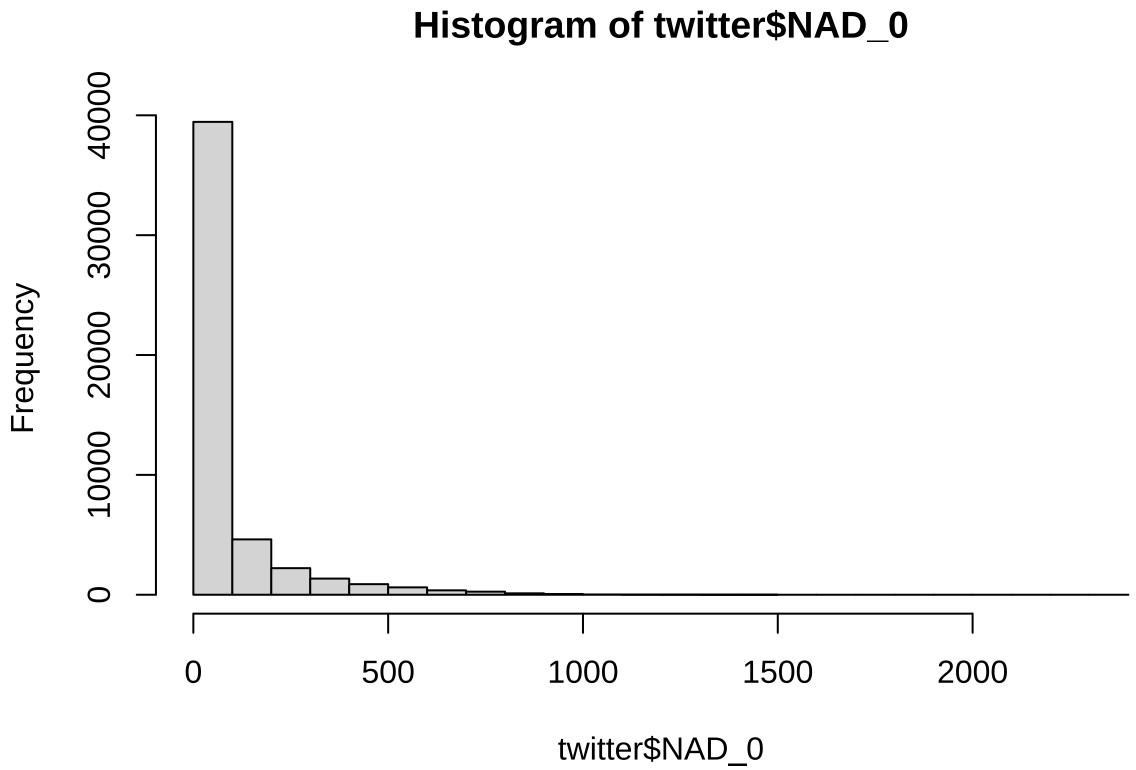
```
hist(twitter$NA_0)
```



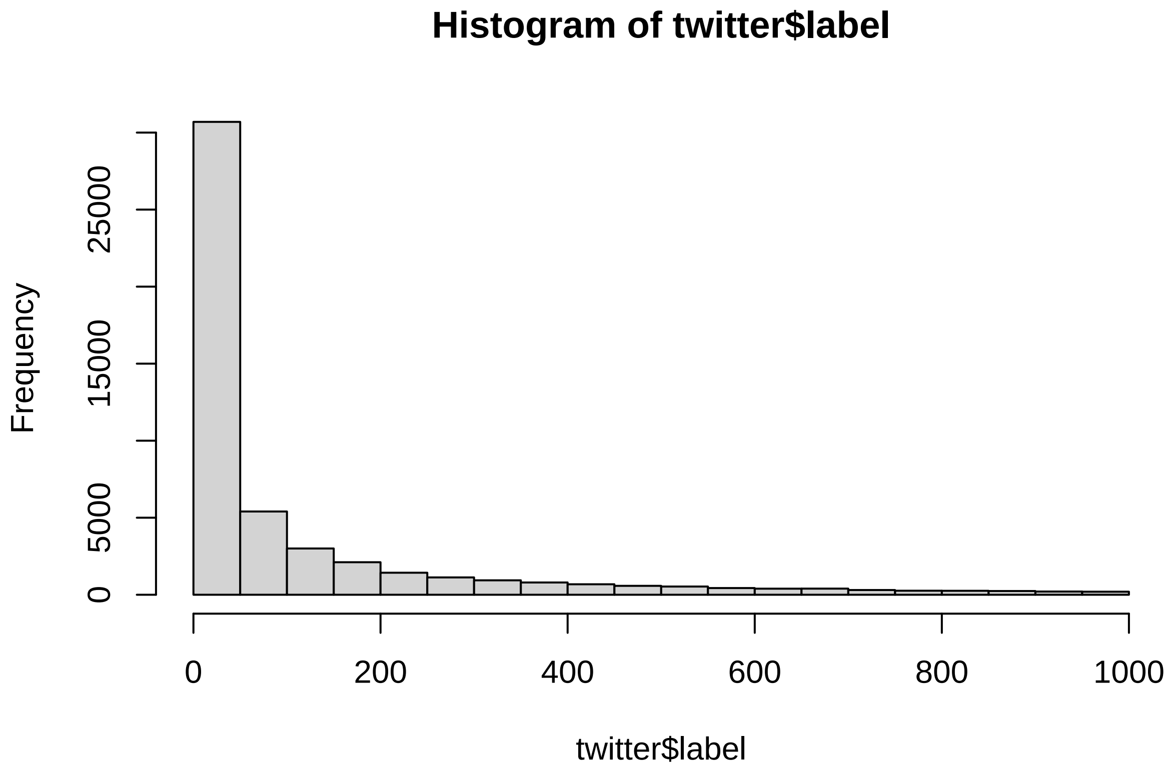
```
hist(twitter$ADL_0)
```



```
hist(twitter$NAD_0)
```



```
hist(twitter$label)
```



A partir de ces histogrammes, on peut modéliser les différentes variables par les lois de poisson avec des paramètres différents.

1.3) Transformation log (appliqué à la variable +1) pour toutes les features dont le maximum dépasse 5:

```
for (i in 1:(nb_var-1)){
  if(max(twitter[,i])>5){
    twitter[,i]=log(twitter[,i]+1)
  }
}
```

1.4) 80% Training-set et 20% Test-set

```
library(splitTools)
library(ranger)

dt = sort(sample(nrow(twitter), nrow(twitter)*.8))
train<-twitter[dt,]
test<-twitter[-dt,]
```

1.5) la moyenne et l'écart-type sur le train et standardisé les features du train et du test avec ces valeurs pour chaque feature:

```
# Charger les packages R requis
library(dplyr)
```

```r
# Préparation des données
df <- as_tibble(train)
head(df)
```

```
## # A tibble: 6 x 67
##   NCD_0 NCD_1 NCD_2 NCD_3 NCD_4 NCD_5  AI_0  AI_1  AI_2  AI_3  AI_4  AI_5
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  1.39 1.61  2.08  1.39   1.79 2.71  1.39   1.39 1.79  1.10  0.693 2.40
## 2  4.76 4.01  3.93  5.09   5.25 5.02  3.87   3.04 2.89  3.71  4.08  4.20
## 3  1.95 2.30  1.61  2.56   2.30 2.48  0.693  1.79 1.61  1.39  1.61  1.95
## 4  1.61 0     0.693 1.10   2.20 3.04  1.61   0    0.693 0     1.95  2.89
## 5  2.64 2.30  3.04  3.09   2.40 2.40  1.79   1.79 3.00  2.83  1.95  2.08
## 6  0    0.693 0.693 0.693  1.10 0.693 0      0    0.693 0.693 0     0.693
## # ... with 55 more variables: ASNA_0 <dbl>, ASNA_1 <dbl>, ASNA_2 <dbl>,
## #   ASNA_3 <dbl>, ASNA_4 <dbl>, ASNA_5 <dbl>, BL_0 <dbl>, BL_1 <dbl>,
## #   BL_2 <dbl>, BL_3 <dbl>, BL_4 <dbl>, BL_5 <dbl>, NAC_0 <dbl>, NAC_1 <dbl>,
## #   NAC_2 <dbl>, NAC_3 <dbl>, NAC_4 <dbl>, NAC_5 <dbl>, ASNAC_0 <dbl>,
## #   ASNAC_1 <dbl>, ASNAC_2 <dbl>, ASNAC_3 <dbl>, ASNAC_4 <dbl>, ASNAC_5 <dbl>,
## #   CS_0 <dbl>, CS_1 <dbl>, CS_2 <dbl>, CS_3 <dbl>, CS_4 <dbl>, CS_5 <dbl>,
## #   AT_0 <dbl>, AT_1 <dbl>, AT_2 <dbl>, AT_3 <dbl>, AT_4 <dbl>, AT_5 <dbl>, ...
```

```r
# Calculer la moyenne et le sd de toutes les colonnes numériques
df %>%
  summarise(across(
    .cols = is.numeric,
    .fns = list(Mean = mean, SD = sd), na.rm = TRUE,
    .names = "{col}_{fn}"
    ))
```

```
## Warning: Predicate functions must be wrapped in `where()`.
##
##   # Bad
##   data %>% select(is.numeric)
##
##   # Good
##   data %>% select(where(is.numeric))
##
## i Please update your code.
## This message is displayed once per session.
```

```
## # A tibble: 1 x 134
##   NCD_0_Mean NCD_0_SD NCD_1_Mean NCD_1_SD NCD_2_Mean NCD_2_SD NCD_3_Mean
##        <dbl>    <dbl>      <dbl>    <dbl>      <dbl>    <dbl>      <dbl>
## 1       2.88     1.82       2.84     1.83       2.98     1.87       3.09
## # ... with 127 more variables: NCD_3_SD <dbl>, NCD_4_Mean <dbl>,
## #   NCD_4_SD <dbl>, NCD_5_Mean <dbl>, NCD_5_SD <dbl>, AI_0_Mean <dbl>,
## #   AI_0_SD <dbl>, AI_1_Mean <dbl>, AI_1_SD <dbl>, AI_2_Mean <dbl>,
## #   AI_2_SD <dbl>, AI_3_Mean <dbl>, AI_3_SD <dbl>, AI_4_Mean <dbl>,
## #   AI_4_SD <dbl>, AI_5_Mean <dbl>, AI_5_SD <dbl>, ASNA_0_Mean <dbl>,
## #   ASNA_0_SD <dbl>, ASNA_1_Mean <dbl>, ASNA_1_SD <dbl>, ASNA_2_Mean <dbl>,
## #   ASNA_2_SD <dbl>, ASNA_3_Mean <dbl>, ASNA_3_SD <dbl>, ASNA_4_Mean <dbl>, ...
```

EXO2 2.1 Premier modèle poissonnien en compte toutes les features:

```r
poisson_all = glm(label ~.,data=train,family ="poisson")
summary(poisson_all)
```

```
## 
## Call:
## glm(formula = label ~ ., family = "poisson", data = train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -41.842   -2.737   -1.285    0.442   65.506  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  1.558e+00  1.301e-02 119.687  < 2e-16 ***
## NCD_0       -2.560e+00  2.771e-01  -9.238  < 2e-16 ***
## NCD_1        2.451e+00  1.980e-01  12.376  < 2e-16 ***
## NCD_2       -1.156e+00  1.470e-01  -7.864 3.73e-15 ***
## NCD_3       -1.411e+00  3.294e-01  -4.283 1.84e-05 ***
## NCD_4        4.272e-01  3.290e-01   1.299 0.194086    
## NCD_5        3.010e+00  2.887e-01  10.428  < 2e-16 ***
## AI_0        -1.005e-01  2.966e-03 -33.881  < 2e-16 ***
## AI_1         3.997e-02  3.046e-03  13.122  < 2e-16 ***
## AI_2         1.276e-02  3.097e-03   4.120 3.80e-05 ***
## AI_3        -1.908e-03  3.090e-03  -0.618 0.536867    
## AI_4        -1.562e-02  3.141e-03  -4.971 6.65e-07 ***
## AI_5         3.795e-02  2.907e-03  13.053  < 2e-16 ***
## ASNA_0      -6.455e+01  1.085e+01  -5.947 2.74e-09 ***
## ASNA_1       1.948e+02  1.102e+01  17.671  < 2e-16 ***
## ASNA_2       3.499e+01  7.581e+00   4.615 3.93e-06 ***
## ASNA_3      -1.198e+02  8.764e+00 -13.670  < 2e-16 ***
## ASNA_4       1.402e+01  8.353e+00   1.678 0.093330 .  
## ASNA_5      -2.263e+02  3.183e+00 -71.081  < 2e-16 ***
## BL_0         1.866e+00  2.445e-01   7.633 2.30e-14 ***
## BL_1        -2.026e+00  1.557e-01 -13.008  < 2e-16 ***
## BL_2         5.502e-01  1.191e-01   4.620 3.84e-06 ***
## BL_3         1.158e+00  2.959e-01   3.913 9.12e-05 ***
## BL_4        -8.239e-02  2.928e-01  -0.281 0.778451    
## BL_5        -2.630e+00  2.477e-01 -10.616  < 2e-16 ***
## NAC_0       -1.803e-01  1.010e-02 -17.858  < 2e-16 ***
## NAC_1        5.766e-03  1.067e-02   0.540 0.589048    
## NAC_2        1.245e-01  1.048e-02  11.874  < 2e-16 ***
## NAC_3       -1.047e-01  1.100e-02  -9.523  < 2e-16 ***
## NAC_4        1.293e-02  1.052e-02   1.229 0.219187    
## NAC_5       -8.119e-02  1.026e-02  -7.913 2.51e-15 ***
## ASNAC_0      4.786e+02  1.827e+01  26.189  < 2e-16 ***
## ASNAC_1     -2.371e+02  1.747e+01 -13.570  < 2e-16 ***
## ASNAC_2     -5.493e+00  1.080e+01  -0.509 0.610960    
## ASNAC_3      1.490e+02  1.521e+01   9.791  < 2e-16 ***
## ASNAC_4     -1.926e+02  1.456e+01 -13.229  < 2e-16 ***
## ASNAC_5      8.155e+01  3.643e+00  22.387  < 2e-16 ***
## CS_0        -1.927e+00  2.446e-01  -7.882 3.23e-15 ***
## CS_1         1.722e+00  1.560e-01  11.036  < 2e-16 ***
## CS_2        -4.267e-01  1.196e-01  -3.566 0.000362 ***
## CS_3        -1.242e+00  2.960e-01  -4.196 2.72e-05 ***
## CS_4        -1.427e-03  2.929e-01  -0.005 0.996113
## CS_5         1.977e+00  2.479e-01   7.974 1.53e-15 ***
## AT_0        -9.643e-02  2.395e-02  -4.026 5.67e-05 ***
## AT_1         1.108e-01  2.413e-02   4.593 4.37e-06 ***
## AT_2         2.336e-01  2.390e-02   9.773  < 2e-16 ***
## AT_3        -7.501e-02  2.567e-02  -2.922 0.003476 **
## AT_4         1.249e-01  2.559e-02   4.881 1.06e-06 ***
## AT_5        -4.029e-01  2.529e-02 -15.930  < 2e-16 ***
## NA_0         2.755e-01  7.221e-03  38.145  < 2e-16 ***
## NA_1        -4.730e-02  7.311e-03  -6.470 9.79e-11 ***
## NA_2        -1.073e-01  6.925e-03 -15.495  < 2e-16 ***
## NA_3        -4.001e-03  7.096e-03  -0.564 0.572865
## NA_4        -2.011e-02  7.064e-03  -2.847 0.004415 **
## NA_5         9.660e-02  5.950e-03  16.235  < 2e-16 ***
## ADL_0        8.313e-03  2.421e-02   0.343 0.731326
## ADL_1       -5.018e-02  2.442e-02  -2.055 0.039903 *
## ADL_2       -2.785e-01  2.419e-02 -11.510  < 2e-16 ***
## ADL_3        1.922e-01  2.634e-02   7.300 2.87e-13 ***
## ADL_4       -1.612e-02  2.587e-02  -0.623 0.533194
## ADL_5        4.001e-01  2.607e-02  15.347  < 2e-16 ***
## NAD_0        2.810e+00  2.773e-01  10.134  < 2e-16 ***
## NAD_1       -2.340e+00  1.977e-01 -11.837  < 2e-16 ***
## NAD_2        1.190e+00  1.468e-01   8.105 5.28e-16 ***
## NAD_3        1.602e+00  3.295e-01   4.861 1.17e-06 ***
## NAD_4       -3.273e-01  3.290e-01  -0.995 0.319801
## NAD_5       -2.697e+00  2.887e-01  -9.345  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 8975018  on 39999  degrees of freedom
## Residual deviance:  919805  on 39933  degrees of freedom
## AIC: 1120340
##
## Number of Fisher Scoring iterations: 5
```

2.2

```
label_P=predict(poisson_all, test)
```

Prédictions obtenues avec le modèle sur les individus du train:

```
label_P_train=predict(poisson_all, train)
```

Prédictions obtenues avec le modèle sur les individus du test:

```
label_P_test=predict(poisson_all, test)
```

2.3) L'algorithme errors permet de trouver les logMSE et les MAPE de training-set et de test-set du modèle.

```
#fit,X_test,Y_test

errors = function( y_pred_train , y_pred_test ,
                   y_train, y_test){
  # y_pred_train : vector of predictions on train
  # y_pred_test :
  # y_train
  # y_test :
  if (length(y_pred_train)!=length(y_train)){
    print("y_pred_train and y_train do not have the same length")
  }
  logMSE_train = mean( ( log(y_train+1) - log(y_pred_train+1) )^2 ,
                     na.rm = T)
  logMSE_test = mean( (log(y_test+1) - log(y_pred_test+1) )^2,
                    na.rm = T)
  MAPE_train = mean( abs(y_train - y_pred_train)/(y_train+1),
                  na.rm = T)
  MAPE_test = mean( abs(y_test - y_pred_test)/(y_test+1),
                 na.rm = T)
  return( list(logMSE_train=logMSE_train,logMSE_test=logMSE_test,
               MAPE_train= MAPE_train,MAPE_test= MAPE_test))
}
```

2.4)

On calcule logMSE et MAPE de modèle poisionnien et les moyennes:

```
errors_full1 = data.frame(errors(label_P_train,label_P_test,train$label,test$label))
rownames(errors_full1)<-c("errors_full1")

Y_pred_train_null = rep(mean(train$label),nrow(train))
Y_pred_test_null = rep(mean(test$label),nrow(test))
errors_null = data.frame(errors(Y_pred_train_null,Y_pred_test_null,train$label,test$label))
rownames(errors_null)<-c("errors_null")

errors_all = rbind(errors_full1,errors_null)

errors_all
```

```
##              logMSE_train logMSE_test MAPE_train  MAPE_test
## errors_full1     5.770269    5.783533  0.7728315  0.7738743
## errors_null      5.270964    5.262991 14.5227421 14.5885468
```

On peut constater que le modèle poisionnien a une beaucoup plus petite valeur en MAPE par rapport aux moyennes.

EXO3 3.1) Une pénalisation elastic-net au modèle précédent (en prenant alpha = 0.8) et en considérant 3 folds pour la cross-validation:

```
X_train_interaction = model.matrix((poisson_all))
#X_train_interaction = data.frame(poisson_all)

library(glmnet)
```

```
## Le chargement a nécessité le package : Matrix

##
## Attachement du package : 'Matrix'

## Les objets suivants sont masqués depuis 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-4
```

```
library(Matrix)
Y_train = train$label

poisson_lasso=cv.glmnet(X_train_interaction, Y_train, family="poisson", alpha = 0.8,nfolds = 3)
```

```
summary(poisson_lasso)
```

```
##            Length Class   Mode
## lambda     76     -none-  numeric
## cvm        76     -none-  numeric
## cvsd       76     -none-  numeric
## cvup       76     -none-  numeric
## cvlo       76     -none-  numeric
## nzero      76     -none-  numeric
## call        6     -none-  call
## name        1     -none-  character
## glmnet.fit 12     fishnet list
## lambda.min  1     -none-  numeric
## lambda.1se  1     -none-  numeric
## index       2     -none-  numeric
```

3.2)

```
which(as.matrix(coef(poisson_lasso)!=0))# Trouver les indices de variables qui ont une coefficient non
```

```
##  [1]  1  3  4  5  6  7  8 16 19 20 27 29 31 32 33 51 55 56 63 64 65 66 67 68
```

```
X_train_interaction[1:10,1:10]
```

```
##   (Intercept)    NCD_0     NCD_1     NCD_2     NCD_3    NCD_4     NCD_5
## 1           1 1.386294 1.6094379 2.0794415 1.3862944 1.791759 2.7080502
## 2           1 4.762174 4.0073332 3.9318256 5.0875963 5.247024 5.0172798
## 3           1 1.945910 2.3025851 1.6094379 2.5649494 2.302585 2.4849066
## 4           1 1.609438 0.0000000 0.6931472 1.0986123 2.197225 3.0445224
## 5           1 2.639057 2.3025851 3.0445224 3.0910425 2.397895 2.3978953
## 6           1 0.000000 0.6931472 0.6931472 0.6931472 1.098612 0.6931472
## 7           1 2.302585 2.8332133 2.5649494 3.1780538 2.708050 3.1354942
## 8           1 1.098612 1.0986123 1.0986123 1.0986123 2.079442 1.0986123
## 9           1 1.386294 1.0986123 1.3862944 0.6931472 1.791759 1.3862944
## 10           1 3.784190 3.2580965 3.6888795 4.0430513 4.820282 5.7620514
##         AI_0     AI_1      AI_2
## 1  1.3862944 1.386294 1.7917595
## 2  3.8712010 3.044522 2.8903718
## 3  0.6931472 1.791759 1.6094379
## 4  1.6094379 0.000000 0.6931472
## 5  1.7917595 1.791759 2.9957323
## 6  0.0000000 0.000000 0.6931472
## 7  2.3025851 2.197225 2.3025851
## 8  0.6931472 0.000000 0.6931472
## 9  1.0986123 1.098612 1.3862944
## 10 3.4657359 2.639057 3.2188758
```

```
selected_features = c(1,which(as.matrix(coef(poisson_lasso))!=0)[-1]-1)# Features selection: trouver le

X_train_interaction[1:10,1:10]
```

```
##    (Intercept)    NCD_0     NCD_1     NCD_2     NCD_3    NCD_4     NCD_5
## 1            1 1.386294 1.6094379 2.0794415 1.3862944 1.791759 2.7080502
## 2            1 4.762174 4.0073332 3.9318256 5.0875963 5.247024 5.0172798
## 3            1 1.945910 2.3025851 1.6094379 2.5649494 2.302585 2.4849066
## 4            1 1.609438 0.0000000 0.6931472 1.0986123 2.197225 3.0445224
## 5            1 2.639057 2.3025851 3.0445224 3.0910425 2.397895 2.3978953
## 6            1 0.000000 0.6931472 0.6931472 0.6931472 1.098612 0.6931472
## 7            1 2.302585 2.8332133 2.5649494 3.1780538 2.708050 3.1354942
## 8            1 1.098612 1.0986123 1.0986123 1.0986123 2.079442 1.0986123
## 9            1 1.386294 1.0986123 1.3862944 0.6931472 1.791759 1.3862944
## 10           1 3.784190 3.2580965 3.6888795 4.0430513 4.820282 5.7620514
##         AI_0     AI_1      AI_2
## 1  1.3862944 1.386294 1.7917595
## 2  3.8712010 3.044522 2.8903718
## 3  0.6931472 1.791759 1.6094379
## 4  1.6094379 0.000000 0.6931472
## 5  1.7917595 1.791759 2.9957323
## 6  0.0000000 0.000000 0.6931472
## 7  2.3025851 2.197225 2.3025851
## 8  0.6931472 0.000000 0.6931472
## 9  1.0986123 1.098612 1.3862944
## 10 3.4657359 2.639057 3.2188758
```

```
coef(poisson_lasso)#Voir les coefficients de ces variables
```

```
## 68 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)  8.423810e-01
## (Intercept)  .
## NCD_0        1.131260e-01
## NCD_1        5.114454e-02
## NCD_2        3.947655e-02
## NCD_3        3.553544e-02
## NCD_4        3.840724e-02
## NCD_5        1.272490e-01
## AI_0         .
## AI_1          .
## AI_2          .
## AI_3          .
## AI_4          .
## AI_5          .
## ASNA_0        .
## ASNA_1        2.891933e+01
## ASNA_2        .
## ASNA_3        .
## ASNA_4       -3.998196e+01
## ASNA_5       -4.734712e+01
## BL_0          .
## BL_1          .
## BL_2          .
## BL_3          .
## BL_4          .
## BL_5          .
## NAC_0         3.412402e-03
## NAC_1         .
## NAC_2         2.683521e-04
## NAC_3         .
## NAC_4         4.397342e-05
## NAC_5         4.756862e-02
## ASNAC_0       1.564092e+02
## ASNAC_1       .
## ASNAC_2       .
## ASNAC_3       .
## ASNAC_4       .
## ASNAC_5       .
## CS_0          .
## CS_1          .
## CS_2          .
## CS_3          .
## CS_4          .
## CS_5          .
## AT_0          .
## AT_1          .
## AT_2          .
## AT_3          .
## AT_4          .
## AT_5          .
## NA_0          6.109997e-02
## NA_1          .
## NA_2          .
## NA_3          .
## NA_4          3.859684e-06
## NA_5          2.138467e-02
## ADL_0         .
## ADL_1         .
## ADL_2         .
## ADL_3         .
## ADL_4         .
## ADL_5         .
## NAD_0         1.046585e-01
## NAD_1        3.249884e-02
## NAD_2        4.575196e-02
## NAD_3        3.659285e-02
## NAD_4        3.052903e-02
## NAD_5        1.009974e-01
```

```
length(as.matrix(coef(poisson_lasso)!=0))
```

```
## [1] 68
```

```
# Refit le modèle glm avec ces variables
#X_train_interaction[,selected_features]

glm_selected = glm.fit(X_train_interaction[,selected_features], train$label, family = poisson())
```

```
summary(glm_selected)
```

```
##                   Length Class  Mode
## coefficients         24  -none- numeric
## residuals         40000  -none- numeric
## fitted.values     40000  -none- numeric
## effects           40000  -none- numeric
## R                   576  -none- numeric
## rank                  1  -none- numeric
## qr                    5  qr     list
## family               12  family list
## linear.predictors 40000  -none- numeric
## deviance              1  -none- numeric
## aic                   1  -none- numeric
## null.deviance         1  -none- numeric
## iter                  1  -none- numeric
## weights           40000  -none- numeric
## prior.weights     40000  -none- numeric
## df.residual           1  -none- numeric
## df.null               1  -none- numeric
## y                 40000  -none- numeric
## converged             1  -none- logical
## boundary              1  -none- logical
```

```
glm_selected = glm(train$label ~ X_train_interaction[,selected_features] -1, family = "poisson")
```

```
summary(glm_selected)
```

```
##
## Call:
## glm(formula = train$label ~ X_train_interaction[, selected_features] -
##     1, family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -43.143   -2.740   -1.216    0.465   65.246
```

```
##
## Coefficients:
##                                                        Estimate Std. Error
## X_train_interaction[, selected_features](Intercept)  6.647e-01  2.811e-03
## X_train_interaction[, selected_features]NCD_0       -2.841e-01  4.016e-02
## X_train_interaction[, selected_features]NCD_1        7.214e-02  3.494e-02
## X_train_interaction[, selected_features]NCD_2       -4.352e-01  2.965e-02
## X_train_interaction[, selected_features]NCD_3       -4.365e-01  4.531e-02
## X_train_interaction[, selected_features]NCD_4        1.565e-01  5.131e-02
## X_train_interaction[, selected_features]NCD_5       -1.134e-02  5.262e-02
## X_train_interaction[, selected_features]ASNA_1       6.377e+01  1.629e+00
## X_train_interaction[, selected_features]ASNA_4      -1.052e+02  1.876e+00
## X_train_interaction[, selected_features]ASNA_5      -1.563e+02  2.011e+00
## X_train_interaction[, selected_features]NAC_0       -1.633e-01  5.983e-03
## X_train_interaction[, selected_features]NAC_2       -2.726e-03  5.735e-03
## X_train_interaction[, selected_features]NAC_4        3.171e-02  6.174e-03
## X_train_interaction[, selected_features]NAC_5        3.536e-02  6.234e-03
## X_train_interaction[, selected_features]ASNAC_0      3.560e+02  4.638e+00
## X_train_interaction[, selected_features]NA_0         1.074e-01  4.131e-03
## X_train_interaction[, selected_features]NA_4        -1.433e-02  4.246e-03
## X_train_interaction[, selected_features]NA_5         5.864e-02  3.909e-03
## X_train_interaction[, selected_features]NAD_0        5.951e-01  4.121e-02
## X_train_interaction[, selected_features]NAD_1        7.382e-03  3.499e-02
## X_train_interaction[, selected_features]NAD_2        5.265e-01  3.096e-02
## X_train_interaction[, selected_features]NAD_3        5.084e-01  4.538e-02
## X_train_interaction[, selected_features]NAD_4       -8.890e-02  5.239e-02
## X_train_interaction[, selected_features]NAD_5        2.664e-01  5.364e-02
##                                                     z value Pr(>|z|)
## X_train_interaction[, selected_features](Intercept) 236.448  < 2e-16 ***
## X_train_interaction[, selected_features]NCD_0        -7.072 1.52e-12 ***
## X_train_interaction[, selected_features]NCD_1         2.065 0.038933 *
## X_train_interaction[, selected_features]NCD_2       -14.678  < 2e-16 ***
## X_train_interaction[, selected_features]NCD_3        -9.633  < 2e-16 ***
## X_train_interaction[, selected_features]NCD_4         3.049 0.002293 **
## X_train_interaction[, selected_features]NCD_5        -0.215 0.829409
## X_train_interaction[, selected_features]ASNA_1       39.138  < 2e-16 ***
## X_train_interaction[, selected_features]ASNA_4      -56.092  < 2e-16 ***
## X_train_interaction[, selected_features]ASNA_5      -77.754  < 2e-16 ***
## X_train_interaction[, selected_features]NAC_0       -27.292  < 2e-16 ***
## X_train_interaction[, selected_features]NAC_2        -0.475 0.634569
## X_train_interaction[, selected_features]NAC_4         5.136 2.80e-07 ***
## X_train_interaction[, selected_features]NAC_5         5.672 1.41e-08 ***
## X_train_interaction[, selected_features]ASNAC_0      76.747  < 2e-16 ***
## X_train_interaction[, selected_features]NA_0         25.991  < 2e-16 ***
## X_train_interaction[, selected_features]NA_4         -3.376 0.000736 ***
## X_train_interaction[, selected_features]NA_5         15.001  < 2e-16 ***
## X_train_interaction[, selected_features]NAD_0        14.442  < 2e-16 ***
## X_train_interaction[, selected_features]NAD_1         0.211 0.832903
## X_train_interaction[, selected_features]NAD_2        17.004  < 2e-16 ***
## X_train_interaction[, selected_features]NAD_3        11.203  < 2e-16 ***
## X_train_interaction[, selected_features]NAD_4        -1.697 0.089721 .
## X_train_interaction[, selected_features]NAD_5         4.966 6.84e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 42523213  on 40000  degrees of freedom
## Residual deviance:   931112  on 39976  degrees of freedom
## AIC: 1131561
##
## Number of Fisher Scoring iterations: 5
```

```
library(caret)
```

```
## Le chargement a nécessité le package : lattice
```

```
##
## Attachement du package : 'caret'
```

```
## L'objet suivant est masqué depuis 'package:purrr':
##
##     lift
```

```
X_train=train[,-67]
Y_train=train[,67]
X_test=test[,-67]
Y_test=test[,67]
cctrl1 <- trainControl(method = "cv", number = 3)
fit_complet <- train(X_train, train$label,
                             method = "glm",
                             family = poisson,
                             trControl = cctrl1)
```

```
## Warning: Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
## Setting row names on a tibble is deprecated.
```

```
fit_complet
```

```
## Generalized Linear Model
##
## 40000 samples
##    66 predictor
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 26667, 26667, 26666
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   78.65591  0.8334091  34.66126
```

```
Y_pred_train_complet = predict(fit_complet)
Y_pred_test_complet = predict(fit_complet,newdata = X_test)
```

```
selected_features = c(1,which(as.matrix(coef(poisson_lasso))!=0)[-1]-1)

poisson_lasso_refit = glm(train$label ~ X_train_interaction[,selected_features] -1, family = "poisson")


summary(poisson_lasso_refit)
```

```
## 
## Call:
## glm(formula = train$label ~ X_train_interaction[, selected_features] - 
##     1, family = "poisson")
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -43.143   -2.740   -1.216    0.465   65.246  
## 
## Coefficients:
##                                                        Estimate Std. Error
## X_train_interaction[, selected_features](Intercept)  6.647e-01  2.811e-03
## X_train_interaction[, selected_features]NCD_0       -2.841e-01  4.016e-02
## X_train_interaction[, selected_features]NCD_1        7.214e-02  3.494e-02
## X_train_interaction[, selected_features]NCD_2       -4.352e-01  2.965e-02
## X_train_interaction[, selected_features]NCD_3       -4.365e-01  4.531e-02
## X_train_interaction[, selected_features]NCD_4        1.565e-01  5.131e-02
## X_train_interaction[, selected_features]NCD_5       -1.134e-02  5.262e-02
## X_train_interaction[, selected_features]ASNA_1       6.377e+01  1.629e+00
## X_train_interaction[, selected_features]ASNA_4      -1.052e+02  1.876e+00
## X_train_interaction[, selected_features]ASNA_5      -1.563e+02  2.011e+00
## X_train_interaction[, selected_features]NAC_0       -1.633e-01  5.983e-03
## X_train_interaction[, selected_features]NAC_2       -2.726e-03  5.735e-03
## X_train_interaction[, selected_features]NAC_4        3.171e-02  6.174e-03
## X_train_interaction[, selected_features]NAC_5        3.536e-02  6.234e-03
## X_train_interaction[, selected_features]ASNAC_0      3.560e+02  4.638e+00
## X_train_interaction[, selected_features]NA_0         1.074e-01  4.131e-03
## X_train_interaction[, selected_features]NA_4        -1.433e-02  4.246e-03
## X_train_interaction[, selected_features]NA_5         5.864e-02  3.909e-03
## X_train_interaction[, selected_features]NAD_0        5.951e-01  4.121e-02
## X_train_interaction[, selected_features]NAD_1        7.382e-03  3.499e-02
## X_train_interaction[, selected_features]NAD_2        5.265e-01  3.096e-02
## X_train_interaction[, selected_features]NAD_3        5.084e-01  4.538e-02
## X_train_interaction[, selected_features]NAD_4       -8.890e-02  5.239e-02
## X_train_interaction[, selected_features]NAD_5        2.664e-01  5.364e-02
##                                                     z value Pr(>|z|)    
## X_train_interaction[, selected_features](Intercept) 236.448  < 2e-16 ***
## X_train_interaction[, selected_features]NCD_0        -7.072 1.52e-12 ***
## X_train_interaction[, selected_features]NCD_1         2.065 0.038933 *  
## X_train_interaction[, selected_features]NCD_2       -14.678  < 2e-16 ***
## X_train_interaction[, selected_features]NCD_3        -9.633  < 2e-16 ***
## X_train_interaction[, selected_features]NCD_4         3.049 0.002293 ** 
## X_train_interaction[, selected_features]NCD_5        -0.215 0.829409    
## X_train_interaction[, selected_features]ASNA_1       39.138  < 2e-16 ***
## X_train_interaction[, selected_features]ASNA_4      -56.092  < 2e-16 ***
## X_train_interaction[, selected_features]ASNA_5      -77.754  < 2e-16 ***
## X_train_interaction[, selected_features]NAC_0       -27.292  < 2e-16 ***
## X_train_interaction[, selected_features]NAC_2        -0.475 0.634569    
## X_train_interaction[, selected_features]NAC_4         5.136 2.80e-07 ***
## X_train_interaction[, selected_features]NAC_5         5.672 1.41e-08 ***
## X_train_interaction[, selected_features]ASNAC_0      76.747  < 2e-16 ***
## X_train_interaction[, selected_features]NA_0         25.991  < 2e-16 ***
## X_train_interaction[, selected_features]NA_4         -3.376 0.000736 ***
## X_train_interaction[, selected_features]NA_5         15.001  < 2e-16 ***
## X_train_interaction[, selected_features]NAD_0        14.442  < 2e-16 ***
## X_train_interaction[, selected_features]NAD_1         0.211 0.832903    
## X_train_interaction[, selected_features]NAD_2        17.004  < 2e-16 ***
## X_train_interaction[, selected_features]NAD_3        11.203  < 2e-16 ***
## X_train_interaction[, selected_features]NAD_4        -1.697 0.089721 .  
## X_train_interaction[, selected_features]NAD_5         4.966 6.84e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 42523213  on 40000  degrees of freedom
## Residual deviance:   931112  on 39976  degrees of freedom
## AIC: 1131561
##
## Number of Fisher Scoring iterations: 5
```

3.3

Calculer la MAPE et le log-MSE de ces deux modèles

```r
#fit,X_test,Y_test

errors = function( y_pred_train , y_pred_test ,
                   y_train, y_test){
  # y_pred_train : vector of predictions on train
  # y_pred_test :
  # y_train
  # y_test :
  if (length(y_pred_train)!=length(y_train)){
    print("y_pred_train and y_train do not have the same length")
  }
  logMSE_train = mean( ( log(y_train+1) - log(y_pred_train+1) )^2 ,
                       na.rm = T)
  logMSE_test = mean( (log(y_test+1) - log(y_pred_test+1) )^2,
                      na.rm = T)
  MAPE_train = mean( abs(y_train - y_pred_train)/(y_train+1),
                    na.rm = T)
  MAPE_test = mean( abs(y_test - y_pred_test)/(y_test+1),
                   na.rm = T)
  return( list(logMSE_train=logMSE_train,logMSE_test=logMSE_test,
               MAPE_train= MAPE_train,MAPE_test= MAPE_test))
}
```

```
errors_full = errors(Y_pred_train_complet,Y_pred_test_complet,train$label,test$label)
print(errors_full)
```

```
## $logMSE_train
## [1] 0.497836
##
## $logMSE_test
## [1] 0.5223486
##
## $MAPE_train
## [1] 0.9245675
##
## $MAPE_test
## [1] 1.102446
```

```
Y_pred_train_null = rep(mean(train$label),nrow(train))
Y_pred_test_null = rep(mean(test$label),nrow(test))
errors_null = errors(Y_pred_train_null,Y_pred_test_null,train$label,test$label)
print(errors_null)
```

```
## $logMSE_train
## [1] 5.270964
##
## $logMSE_test
## [1] 5.262991
##
## $MAPE_train
## [1] 14.52274
##
## $MAPE_test
## [1] 14.58855
```

```
errors_all = rbind(errors_full,errors_null)
```

```
errors_all
```

```
##             logMSE_train logMSE_test MAPE_train MAPE_test
## errors_full 0.497836     0.5223486   0.9245675  1.102446
## errors_null 5.270964     5.262991    14.52274   14.58855
```

On voit que les nouveaux modèles ont une meilleures performances que le précédent,car ils ont une plus petites valeur en logMse et en MAPE.

```
X2=model.matrix(poisson_lasso_refit)
X2=X2[,-1]
X_train2=X2[1:31969,]
X_test2=X2[31970:40000,]

X2B=data.frame(X_train2)
X_train2_label=twitter[which(X2B$X_train_interaction...selected_features.NCD_0==twitter$NCD_0),]$label
```

```
## Warning in X2B$X_train_interaction...selected_features.NCD_0 == twitter$NCD_0:
## la taille d’un objet plus long n’est pas multiple de la taille d’un objet plus
## court
```

```
X2T=data.frame(X_test2)
X_test2_label=twitter[which(X2T$X_train_interaction...selected_features.NCD_0==twitter$NCD_0),]$label
```

```
## Warning in X2T$X_train_interaction...selected_features.NCD_0 == twitter$NCD_0:
## la taille d’un objet plus long n’est pas multiple de la taille d’un objet plus
## court
```

```
fit_aic <- train(X_train2,train[1:31969,]$label,
                             method = "glmStepAIC",
                             family = poisson,
                             trControl = cctrl1)
```

```
## Start:  AIC=610070.1
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_3‘ + ‘X_train_interaction[, selected_features]NCD_4
##     ‘X_train_interaction[, selected_features]NCD_5‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_4‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]NAC_0‘ + ‘X_train_interaction[, selected_features]NAC_2
##     ‘X_train_interaction[, selected_features]NAC_4‘ + ‘X_train_interaction[, selected_features]NAC_5
##     ‘X_train_interaction[, selected_features]ASNAC_0‘ + ‘X_train_interaction[, selected_features]NA_
##     ‘X_train_interaction[, selected_features]NA_4‘ + ‘X_train_interaction[, selected_features]NA_5‘ +
##     ‘X_train_interaction[, selected_features]NAD_0‘ + ‘X_train_interaction[, selected_features]NAD_1
##     ‘X_train_interaction[, selected_features]NAD_2‘ + ‘X_train_interaction[, selected_features]NAD_3
##     ‘X_train_interaction[, selected_features]NAD_4‘ + ‘X_train_interaction[, selected_features]NAD_5
##
##                                                     Df Deviance    AIC
## - ‘X_train_interaction[, selected_features]NCD_5‘    1   503161 610070
## <none>                                                   503159 610070
## - ‘X_train_interaction[, selected_features]NCD_4‘    1   503162 610071
## - ‘X_train_interaction[, selected_features]NAC_5‘    1   503163 610071
## - ‘X_train_interaction[, selected_features]NCD_0‘    1   503163 610072
## - ‘X_train_interaction[, selected_features]NAD_4‘    1   503164 610072
## - ‘X_train_interaction[, selected_features]NCD_3‘    1   503164 610073
## - ‘X_train_interaction[, selected_features]NA_4‘     1   503167 610076
## - ‘X_train_interaction[, selected_features]NAD_5‘    1   503168 610077
## - ‘X_train_interaction[, selected_features]NAD_3‘    1   503170 610079
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   503173 610082
## - ‘X_train_interaction[, selected_features]NAC_4‘    1   503183 610092
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   503191 610100
## - ‘X_train_interaction[, selected_features]NAC_2‘    1   503224 610133
## - ‘X_train_interaction[, selected_features]NAD_0‘    1   503229 610137
## - ‘X_train_interaction[, selected_features]NA_5‘     1   503273 610182
## - ‘X_train_interaction[, selected_features]NA_0‘     1   503348 610257
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   503427 610335
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   503432 610341
## - ‘X_train_interaction[, selected_features]NAC_0‘    1   503842 610750
## - ‘X_train_interaction[, selected_features]ASNA_1‘   1   504392 611300
## - ‘X_train_interaction[, selected_features]ASNA_4‘   1   505031 611939
## - ‘X_train_interaction[, selected_features]ASNAC_0‘  1   507015 613923
## - ‘X_train_interaction[, selected_features]ASNA_5‘   1   507061 613970
##
## Step:  AIC=610069.8
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_3‘ + ‘X_train_interaction[, selected_features]NCD_4
##     ‘X_train_interaction[, selected_features]ASNA_1‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_5‘ + ‘X_train_interaction[, selected_features]NAC_
##     ‘X_train_interaction[, selected_features]NAC_2‘ + ‘X_train_interaction[, selected_features]NAC_4
##     ‘X_train_interaction[, selected_features]NAC_5‘ + ‘X_train_interaction[, selected_features]ASNAC_
##     ‘X_train_interaction[, selected_features]NA_0‘ + ‘X_train_interaction[, selected_features]NA_4‘ +
##     ‘X_train_interaction[, selected_features]NA_5‘ + ‘X_train_interaction[, selected_features]NAD_0‘
##     ‘X_train_interaction[, selected_features]NAD_1‘ + ‘X_train_interaction[, selected_features]NAD_2
##     ‘X_train_interaction[, selected_features]NAD_3‘ + ‘X_train_interaction[, selected_features]NAD_4
##     ‘X_train_interaction[, selected_features]NAD_5‘
##
##                                                     Df Deviance    AIC
## - ‘X_train_interaction[, selected_features]NCD_4‘    1   503163 610070
## <none>                                                   503161 610070
## - ‘X_train_interaction[, selected_features]NAD_4‘    1   503165 610071
## - ‘X_train_interaction[, selected_features]NCD_0‘    1   503165 610071
## - ‘X_train_interaction[, selected_features]NAC_5‘    1   503165 610071
## - ‘X_train_interaction[, selected_features]NCD_3‘    1   503165 610072
## - ‘X_train_interaction[, selected_features]NA_4‘     1   503169 610075
## - ‘X_train_interaction[, selected_features]NAD_3‘    1   503171 610078
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   503174 610081
## - ‘X_train_interaction[, selected_features]NAC_4‘    1   503184 610091
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   503192 610099
## - ‘X_train_interaction[, selected_features]NAC_2‘    1   503225 610132
## - ‘X_train_interaction[, selected_features]NAD_0‘    1   503229 610136
## - ‘X_train_interaction[, selected_features]NA_5‘     1   503275 610182
## - ‘X_train_interaction[, selected_features]NA_0‘     1   503350 610256
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   503427 610334
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   503433 610339
## - ‘X_train_interaction[, selected_features]NAC_0‘    1   503843 610750
## - ‘X_train_interaction[, selected_features]ASNA_1‘   1   504393 611300
## - ‘X_train_interaction[, selected_features]NAD_5‘    1   504657 611563
## - ‘X_train_interaction[, selected_features]ASNA_4‘   1   505036 611943
## - ‘X_train_interaction[, selected_features]ASNAC_0‘  1   507016 613923
## - ‘X_train_interaction[, selected_features]ASNA_5‘   1   507061 613968
##
## Step:  AIC=610069.7
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_3‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_4‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]NAC_0‘ + ‘X_train_interaction[, selected_features]NAC_2
##     ‘X_train_interaction[, selected_features]NAC_4‘ + ‘X_train_interaction[, selected_features]NAC_5
##     ‘X_train_interaction[, selected_features]ASNAC_0‘ + ‘X_train_interaction[, selected_features]NA_
##     ‘X_train_interaction[, selected_features]NA_4‘ + ‘X_train_interaction[, selected_features]NA_5‘ +
##     ‘X_train_interaction[, selected_features]NAD_0‘ + ‘X_train_interaction[, selected_features]NAD_1
##     ‘X_train_interaction[, selected_features]NAD_2‘ + ‘X_train_interaction[, selected_features]NAD_3
##     ‘X_train_interaction[, selected_features]NAD_4‘ + ‘X_train_interaction[, selected_features]NAD_5
```

```
##
##                                                     Df Deviance    AIC
## <none>                                                   503163 610070
## - ‘X_train_interaction[, selected_features]NAC_5‘    1   503166 610071
## - ‘X_train_interaction[, selected_features]NCD_0‘    1   503167 610072
## - ‘X_train_interaction[, selected_features]NCD_3‘    1   503168 610073
## - ‘X_train_interaction[, selected_features]NA_4‘     1   503171 610075
## - ‘X_train_interaction[, selected_features]NAD_3‘    1   503175 610079
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   503177 610082
## - ‘X_train_interaction[, selected_features]NAD_4‘    1   503182 610087
## - ‘X_train_interaction[, selected_features]NAC_4‘    1   503187 610092
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   503196 610100
## - ‘X_train_interaction[, selected_features]NAC_2‘    1   503228 610132
## - ‘X_train_interaction[, selected_features]NAD_0‘    1   503234 610139
## - ‘X_train_interaction[, selected_features]NA_5‘     1   503277 610182
## - ‘X_train_interaction[, selected_features]NA_0‘     1   503351 610256
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   503432 610337
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   503438 610342
## - ‘X_train_interaction[, selected_features]NAC_0‘    1   503844 610748
## - ‘X_train_interaction[, selected_features]ASNA_1‘   1   504396 611300
## - ‘X_train_interaction[, selected_features]NAD_5‘    1   504657 611562
## - ‘X_train_interaction[, selected_features]ASNA_4‘   1   505041 611946
## - ‘X_train_interaction[, selected_features]ASNAC_0‘  1   507019 613923
## - ‘X_train_interaction[, selected_features]ASNA_5‘   1   507062 613967
## Start:  AIC=602866.5
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_3‘ + ‘X_train_interaction[, selected_features]NCD_4
##     ‘X_train_interaction[, selected_features]NCD_5‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_4‘ + ‘X_train_interaction[, selected_features]ASNA
##     ‘X_train_interaction[, selected_features]NAC_0‘ + ‘X_train_interaction[, selected_features]NAC_2
##     ‘X_train_interaction[, selected_features]NAC_4‘ + ‘X_train_interaction[, selected_features]NAC_5
##     ‘X_train_interaction[, selected_features]ASNAC_0‘ + ‘X_train_interaction[, selected_features]NA_
##     ‘X_train_interaction[, selected_features]NA_4‘ + ‘X_train_interaction[, selected_features]NA_5‘
##     ‘X_train_interaction[, selected_features]NAD_0‘ + ‘X_train_interaction[, selected_features]NAD_1
##     ‘X_train_interaction[, selected_features]NAD_2‘ + ‘X_train_interaction[, selected_features]NAD_3
##     ‘X_train_interaction[, selected_features]NAD_4‘ + ‘X_train_interaction[, selected_features]NAD_5
##
##                                                     Df Deviance    AIC
## - ‘X_train_interaction[, selected_features]NAC_5‘    1   495836 602865
## - ‘X_train_interaction[, selected_features]NCD_5‘    1   495836 602865
## <none>                                                   495836 602867
## - ‘X_train_interaction[, selected_features]NA_4‘     1   495839 602868
## - ‘X_train_interaction[, selected_features]NCD_3‘    1   495840 602869
## - ‘X_train_interaction[, selected_features]NAD_3‘    1   495845 602874
## - ‘X_train_interaction[, selected_features]NAD_5‘    1   495849 602878
## - ‘X_train_interaction[, selected_features]NA_5‘     1   495849 602878
## - ‘X_train_interaction[, selected_features]NCD_4‘    1   495851 602880
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   495851 602880
## - ‘X_train_interaction[, selected_features]NAD_4‘    1   495853 602882
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   495865 602894
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   495892 602921
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   495893 602922
## - ‘X_train_interaction[, selected_features]NCD_0‘    1   495901 602930
## - `X_train_interaction[, selected_features]NAC_2`    1   495917 602946
## - `X_train_interaction[, selected_features]NAC_4`    1   495930 602959
## - `X_train_interaction[, selected_features]NAD_0`    1   496042 603070
## - `X_train_interaction[, selected_features]ASNA_1`   1   496238 603267
## - `X_train_interaction[, selected_features]NA_0`     1   496378 603407
## - `X_train_interaction[, selected_features]NAC_0`    1   496790 603819
## - `X_train_interaction[, selected_features]ASNA_4`   1   498017 605046
## - `X_train_interaction[, selected_features]ASNAC_0`  1   498589 605618
## - `X_train_interaction[, selected_features]ASNA_5`   1   498996 606025
##
## Step:  AIC=602864.6
## .outcome ~ `X_train_interaction[, selected_features]NCD_0` +
##     `X_train_interaction[, selected_features]NCD_1` + `X_train_interaction[, selected_features]NCD_2
##     `X_train_interaction[, selected_features]NCD_3` + `X_train_interaction[, selected_features]NCD_4
##     `X_train_interaction[, selected_features]NCD_5` + `X_train_interaction[, selected_features]ASNA_
##     `X_train_interaction[, selected_features]ASNA_4` + `X_train_interaction[, selected_features]ASNA_
##     `X_train_interaction[, selected_features]NAC_0` + `X_train_interaction[, selected_features]NAC_2
##     `X_train_interaction[, selected_features]NAC_4` + `X_train_interaction[, selected_features]ASNAC_
##     `X_train_interaction[, selected_features]NA_0` + `X_train_interaction[, selected_features]NA_4` +
##     `X_train_interaction[, selected_features]NA_5` + `X_train_interaction[, selected_features]NAD_0`
##     `X_train_interaction[, selected_features]NAD_1` + `X_train_interaction[, selected_features]NAD_2
##     `X_train_interaction[, selected_features]NAD_3` + `X_train_interaction[, selected_features]NAD_4
##     `X_train_interaction[, selected_features]NAD_5`
##
##                                                      Df Deviance    AIC
## - `X_train_interaction[, selected_features]NCD_5`    1   495836 602863
## <none>                                                   495836 602865
## - `X_train_interaction[, selected_features]NA_4`     1   495839 602866
## - `X_train_interaction[, selected_features]NCD_3`    1   495840 602867
## - `X_train_interaction[, selected_features]NAD_3`    1   495845 602872
## - `X_train_interaction[, selected_features]NAD_5`    1   495850 602877
## - `X_train_interaction[, selected_features]NCD_4`    1   495851 602878
## - `X_train_interaction[, selected_features]NAD_1`    1   495851 602878
## - `X_train_interaction[, selected_features]NA_5`     1   495851 602878
## - `X_train_interaction[, selected_features]NAD_4`    1   495853 602880
## - `X_train_interaction[, selected_features]NCD_1`    1   495865 602892
## - `X_train_interaction[, selected_features]NCD_2`    1   495892 602919
## - `X_train_interaction[, selected_features]NAD_2`    1   495893 602920
## - `X_train_interaction[, selected_features]NCD_0`    1   495901 602928
## - `X_train_interaction[, selected_features]NAC_2`    1   495923 602950
## - `X_train_interaction[, selected_features]NAC_4`    1   495946 602973
## - `X_train_interaction[, selected_features]NAD_0`    1   496042 603069
## - `X_train_interaction[, selected_features]ASNA_1`   1   496239 603266
## - `X_train_interaction[, selected_features]NA_0`     1   496385 603412
## - `X_train_interaction[, selected_features]NAC_0`    1   496842 603869
## - `X_train_interaction[, selected_features]ASNA_4`   1   498021 605048
## - `X_train_interaction[, selected_features]ASNAC_0`  1   498590 605617
## - `X_train_interaction[, selected_features]ASNA_5`   1   499009 606036
##
## Step:  AIC=602862.8
## .outcome ~ `X_train_interaction[, selected_features]NCD_0` +
##     `X_train_interaction[, selected_features]NCD_1` + `X_train_interaction[, selected_features]NCD_2
##     `X_train_interaction[, selected_features]NCD_3` + `X_train_interaction[, selected_features]NCD_4
##     `X_train_interaction[, selected_features]ASNA_1` + `X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_5‘ + ‘X_train_interaction[, selected_features]NAC_
##     ‘X_train_interaction[, selected_features]NAC_2‘ + ‘X_train_interaction[, selected_features]NAC_4
##     ‘X_train_interaction[, selected_features]ASNAC_0‘ + ‘X_train_interaction[, selected_features]NA_
##     ‘X_train_interaction[, selected_features]NA_4‘ + ‘X_train_interaction[, selected_features]NA_5‘ +
##     ‘X_train_interaction[, selected_features]NAD_0‘ + ‘X_train_interaction[, selected_features]NAD_1
##     ‘X_train_interaction[, selected_features]NAD_2‘ + ‘X_train_interaction[, selected_features]NAD_3
##     ‘X_train_interaction[, selected_features]NAD_4‘ + ‘X_train_interaction[, selected_features]NAD_5
##
##                                                      Df Deviance    AIC
## <none>                                                    495836 602863
## - ‘X_train_interaction[, selected_features]NA_4‘      1   495840 602865
## - ‘X_train_interaction[, selected_features]NCD_3‘     1   495840 602865
## - ‘X_train_interaction[, selected_features]NAD_3‘     1   495846 602871
## - ‘X_train_interaction[, selected_features]NCD_4‘     1   495851 602876
## - ‘X_train_interaction[, selected_features]NA_5‘      1   495851 602876
## - ‘X_train_interaction[, selected_features]NAD_1‘     1   495852 602877
## - ‘X_train_interaction[, selected_features]NAD_4‘     1   495853 602878
## - ‘X_train_interaction[, selected_features]NCD_1‘     1   495866 602891
## - ‘X_train_interaction[, selected_features]NCD_2‘     1   495892 602917
## - ‘X_train_interaction[, selected_features]NAD_2‘     1   495893 602918
## - ‘X_train_interaction[, selected_features]NCD_0‘     1   495901 602926
## - ‘X_train_interaction[, selected_features]NAC_2‘     1   495923 602948
## - ‘X_train_interaction[, selected_features]NAC_4‘     1   495947 602972
## - ‘X_train_interaction[, selected_features]NAD_0‘     1   496043 603068
## - ‘X_train_interaction[, selected_features]ASNA_1‘    1   496239 603264
## - ‘X_train_interaction[, selected_features]NA_0‘      1   496386 603411
## - ‘X_train_interaction[, selected_features]NAC_0‘     1   496843 603868
## - ‘X_train_interaction[, selected_features]ASNA_4‘    1   498025 605050
## - ‘X_train_interaction[, selected_features]ASNAC_0‘   1   498590 605615
## - ‘X_train_interaction[, selected_features]ASNA_5‘    1   499009 606034
## - ‘X_train_interaction[, selected_features]NAD_5‘     1   499523 606548
## Start:  AIC=607191.6
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_3‘ + ‘X_train_interaction[, selected_features]NCD_4
##     ‘X_train_interaction[, selected_features]NCD_5‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_4‘ + ‘X_train_interaction[, selected_features]ASNA
##     ‘X_train_interaction[, selected_features]NAC_0‘ + ‘X_train_interaction[, selected_features]NAC_2
##     ‘X_train_interaction[, selected_features]NAC_4‘ + ‘X_train_interaction[, selected_features]NAC_5
##     ‘X_train_interaction[, selected_features]ASNAC_0‘ + ‘X_train_interaction[, selected_features]NA_
##     ‘X_train_interaction[, selected_features]NA_4‘ + ‘X_train_interaction[, selected_features]NA_5‘ +
##     ‘X_train_interaction[, selected_features]NAD_0‘ + ‘X_train_interaction[, selected_features]NAD_1
##     ‘X_train_interaction[, selected_features]NAD_2‘ + ‘X_train_interaction[, selected_features]NAD_3
##     ‘X_train_interaction[, selected_features]NAD_4‘ + ‘X_train_interaction[, selected_features]NAD_5
##
##                                                      Df Deviance    AIC
## - ‘X_train_interaction[, selected_features]NAD_5‘     1   500266 607190
## <none>                                                    500265 607192
## - ‘X_train_interaction[, selected_features]NCD_5‘     1   500271 607195
## - ‘X_train_interaction[, selected_features]NAD_4‘     1   500272 607196
## - ‘X_train_interaction[, selected_features]NAD_3‘     1   500274 607199
## - ‘X_train_interaction[, selected_features]NCD_0‘     1   500274 607199
## - ‘X_train_interaction[, selected_features]NCD_4‘     1   500275 607199
## - ‘X_train_interaction[, selected_features]NAC_4‘     1   500279 607204
## - ‘X_train_interaction[, selected_features]NA_4‘     1   500280 607204
## - ‘X_train_interaction[, selected_features]NCD_3‘    1   500283 607208
## - ‘X_train_interaction[, selected_features]NAC_5‘    1   500285 607210
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   500306 607231
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   500314 607239
## - ‘X_train_interaction[, selected_features]NAD_0‘    1   500320 607244
## - ‘X_train_interaction[, selected_features]NAC_2‘    1   500329 607253
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   500334 607258
## - ‘X_train_interaction[, selected_features]NA_5‘     1   500354 607278
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   500367 607292
## - ‘X_train_interaction[, selected_features]NAC_0‘    1   500378 607302
## - ‘X_train_interaction[, selected_features]NA_0‘     1   500554 607478
## - ‘X_train_interaction[, selected_features]ASNA_1‘   1   500737 607661
## - ‘X_train_interaction[, selected_features]ASNAC_0‘  1   502754 609679
## - ‘X_train_interaction[, selected_features]ASNA_4‘   1   502802 609727
## - ‘X_train_interaction[, selected_features]ASNA_5‘   1   503074 609999
##
## Step:  AIC=607190.5
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_3‘ + ‘X_train_interaction[, selected_features]NCD_4
##     ‘X_train_interaction[, selected_features]NCD_5‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_4‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]NAC_0‘ + ‘X_train_interaction[, selected_features]NAC_2
##     ‘X_train_interaction[, selected_features]NAC_4‘ + ‘X_train_interaction[, selected_features]NAC_5
##     ‘X_train_interaction[, selected_features]ASNAC_0‘ + ‘X_train_interaction[, selected_features]NA_
##     ‘X_train_interaction[, selected_features]NA_4‘ + ‘X_train_interaction[, selected_features]NA_5‘ +
##     ‘X_train_interaction[, selected_features]NAD_0‘ + ‘X_train_interaction[, selected_features]NAD_1
##     ‘X_train_interaction[, selected_features]NAD_2‘ + ‘X_train_interaction[, selected_features]NAD_3
##     ‘X_train_interaction[, selected_features]NAD_4‘
##
##                                                     Df Deviance    AIC
## <none>                                                   500266 607190
## - ‘X_train_interaction[, selected_features]NAD_4‘    1   500272 607195
## - ‘X_train_interaction[, selected_features]NAD_3‘    1   500274 607197
## - ‘X_train_interaction[, selected_features]NCD_4‘    1   500275 607198
## - ‘X_train_interaction[, selected_features]NCD_0‘    1   500276 607198
## - ‘X_train_interaction[, selected_features]NAC_4‘    1   500280 607203
## - ‘X_train_interaction[, selected_features]NA_4‘     1   500281 607203
## - ‘X_train_interaction[, selected_features]NCD_3‘    1   500283 607206
## - ‘X_train_interaction[, selected_features]NAC_5‘    1   500289 607211
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   500308 607231
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   500316 607239
## - ‘X_train_interaction[, selected_features]NAD_0‘    1   500323 607245
## - ‘X_train_interaction[, selected_features]NAC_2‘    1   500330 607253
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   500336 607259
## - ‘X_train_interaction[, selected_features]NA_5‘     1   500355 607277
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   500370 607293
## - ‘X_train_interaction[, selected_features]NAC_0‘    1   500379 607301
## - ‘X_train_interaction[, selected_features]NA_0‘     1   500554 607477
## - ‘X_train_interaction[, selected_features]ASNA_1‘   1   500738 607660
## - ‘X_train_interaction[, selected_features]NCD_5‘    1   501226 608149
## - ‘X_train_interaction[, selected_features]ASNAC_0‘  1   502756 609679
## - ‘X_train_interaction[, selected_features]ASNA_4‘   1   502802 609725
## - ‘X_train_interaction[, selected_features]ASNA_5‘   1   503076 609999
## Start:  AIC=911306
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_3‘ + ‘X_train_interaction[, selected_features]NCD_4
##     ‘X_train_interaction[, selected_features]NCD_5‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_4‘ + ‘X_train_interaction[, selected_features]ASNA
##     ‘X_train_interaction[, selected_features]NAC_0‘ + ‘X_train_interaction[, selected_features]NAC_2
##     ‘X_train_interaction[, selected_features]NAC_4‘ + ‘X_train_interaction[, selected_features]NAC_5
##     ‘X_train_interaction[, selected_features]ASNAC_0‘ + ‘X_train_interaction[, selected_features]NA_
##     ‘X_train_interaction[, selected_features]NA_4‘ + ‘X_train_interaction[, selected_features]NA_5‘
##     ‘X_train_interaction[, selected_features]NAD_0‘ + ‘X_train_interaction[, selected_features]NAD_1
##     ‘X_train_interaction[, selected_features]NAD_2‘ + ‘X_train_interaction[, selected_features]NAD_3
##     ‘X_train_interaction[, selected_features]NAD_4‘ + ‘X_train_interaction[, selected_features]NAD_5
##
##                                                     Df Deviance    AIC
## - ‘X_train_interaction[, selected_features]NCD_3‘    1   750896 911304
## - ‘X_train_interaction[, selected_features]NCD_4‘    1   750897 911305
## - ‘X_train_interaction[, selected_features]NAC_5‘    1   750897 911305
## - ‘X_train_interaction[, selected_features]NAD_3‘    1   750898 911306
## <none>                                                   750896 911306
## - ‘X_train_interaction[, selected_features]NAD_4‘    1   750898 911306
## - ‘X_train_interaction[, selected_features]NCD_5‘    1   750899 911307
## - ‘X_train_interaction[, selected_features]NA_4‘     1   750899 911308
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   750903 911311
## - ‘X_train_interaction[, selected_features]NAD_5‘    1   750905 911313
## - ‘X_train_interaction[, selected_features]NAC_2‘    1   750910 911318
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   750920 911328
## - ‘X_train_interaction[, selected_features]NCD_0‘    1   750923 911331
## - ‘X_train_interaction[, selected_features]NAC_4‘    1   750952 911360
## - ‘X_train_interaction[, selected_features]NA_5‘     1   750990 911398
## - ‘X_train_interaction[, selected_features]NAD_0‘    1   751042 911450
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   751067 911475
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   751100 911508
## - ‘X_train_interaction[, selected_features]NA_0‘     1   751404 911812
## - ‘X_train_interaction[, selected_features]ASNA_1‘   1   751697 912105
## - ‘X_train_interaction[, selected_features]NAC_0‘    1   751702 912111
## - ‘X_train_interaction[, selected_features]ASNA_4‘   1   754275 914683
## - ‘X_train_interaction[, selected_features]ASNA_5‘   1   755771 916179
## - ‘X_train_interaction[, selected_features]ASNAC_0‘  1   755941 916349
##
## Step:  AIC=911304
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_4‘ + ‘X_train_interaction[, selected_features]NCD_5
##     ‘X_train_interaction[, selected_features]ASNA_1‘ + ‘X_train_interaction[, selected_features]ASNA
##     ‘X_train_interaction[, selected_features]ASNA_5‘ + ‘X_train_interaction[, selected_features]NAC_
##     ‘X_train_interaction[, selected_features]NAC_2‘ + ‘X_train_interaction[, selected_features]NAC_4
##     ‘X_train_interaction[, selected_features]NAC_5‘ + ‘X_train_interaction[, selected_features]ASNAC
##     ‘X_train_interaction[, selected_features]NA_0‘ + ‘X_train_interaction[, selected_features]NA_4‘
##     ‘X_train_interaction[, selected_features]NA_5‘ + ‘X_train_interaction[, selected_features]NAD_0‘
##     ‘X_train_interaction[, selected_features]NAD_1‘ + ‘X_train_interaction[, selected_features]NAD_2
##     ‘X_train_interaction[, selected_features]NAD_3‘ + ‘X_train_interaction[, selected_features]NAD_4
##     ‘X_train_interaction[, selected_features]NAD_5‘
```

```
##
##                                                   Df Deviance    AIC
## - ‘X_train_interaction[, selected_features]NCD_4‘    1   750897 911303
## - ‘X_train_interaction[, selected_features]NAC_5‘    1   750897 911303
## <none>                                                   750896 911304
## - ‘X_train_interaction[, selected_features]NAD_4‘    1   750898 911304
## - ‘X_train_interaction[, selected_features]NCD_5‘    1   750899 911305
## - ‘X_train_interaction[, selected_features]NA_4‘     1   750899 911306
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   750903 911309
## - ‘X_train_interaction[, selected_features]NAD_5‘    1   750905 911311
## - ‘X_train_interaction[, selected_features]NAC_2‘    1   750910 911316
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   750920 911327
## - ‘X_train_interaction[, selected_features]NCD_0‘    1   750923 911329
## - ‘X_train_interaction[, selected_features]NAC_4‘    1   750952 911358
## - ‘X_train_interaction[, selected_features]NA_5‘     1   750990 911396
## - ‘X_train_interaction[, selected_features]NAD_0‘    1   751042 911449
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   751068 911474
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   751100 911506
## - ‘X_train_interaction[, selected_features]NA_0‘     1   751404 911810
## - ‘X_train_interaction[, selected_features]ASNA_1‘   1   751697 912103
## - ‘X_train_interaction[, selected_features]NAC_0‘    1   751703 912109
## - ‘X_train_interaction[, selected_features]ASNA_4‘   1   754275 914681
## - ‘X_train_interaction[, selected_features]ASNA_5‘   1   755771 916177
## - ‘X_train_interaction[, selected_features]ASNAC_0‘  1   755941 916348
## - ‘X_train_interaction[, selected_features]NAD_3‘    1   757980 918387
##
## Step:  AIC=911303
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_5‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_4‘ + ‘X_train_interaction[, selected_features]ASNA
##     ‘X_train_interaction[, selected_features]NAC_0‘ + ‘X_train_interaction[, selected_features]NAC_2
##     ‘X_train_interaction[, selected_features]NAC_4‘ + ‘X_train_interaction[, selected_features]NAC_5
##     ‘X_train_interaction[, selected_features]ASNAC_0‘ + ‘X_train_interaction[, selected_features]NA_
##     ‘X_train_interaction[, selected_features]NA_4‘ + ‘X_train_interaction[, selected_features]NA_5‘
##     ‘X_train_interaction[, selected_features]NAD_0‘ + ‘X_train_interaction[, selected_features]NAD_1
##     ‘X_train_interaction[, selected_features]NAD_2‘ + ‘X_train_interaction[, selected_features]NAD_3
##     ‘X_train_interaction[, selected_features]NAD_4‘ + ‘X_train_interaction[, selected_features]NAD_5
##
##                                                   Df Deviance    AIC
## - ‘X_train_interaction[, selected_features]NAC_5‘    1   750898 911302
## <none>                                                   750897 911303
## - ‘X_train_interaction[, selected_features]NCD_5‘    1   750900 911304
## - ‘X_train_interaction[, selected_features]NA_4‘     1   750900 911305
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   750904 911308
## - ‘X_train_interaction[, selected_features]NAD_5‘    1   750907 911311
## - ‘X_train_interaction[, selected_features]NAC_2‘    1   750912 911316
## - ‘X_train_interaction[, selected_features]NAD_4‘    1   750920 911324
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   750922 911326
## - ‘X_train_interaction[, selected_features]NCD_0‘    1   750925 911329
## - ‘X_train_interaction[, selected_features]NAC_4‘    1   750955 911359
## - ‘X_train_interaction[, selected_features]NA_5‘     1   750991 911395
## - ‘X_train_interaction[, selected_features]NAD_0‘    1   751047 911451
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   751071 911475
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   751103 911507
## - ‘X_train_interaction[, selected_features]NA_0‘     1   751405 911809
## - ‘X_train_interaction[, selected_features]ASNA_1‘   1   751698 912103
## - ‘X_train_interaction[, selected_features]NAC_0‘    1   751703 912107
## - ‘X_train_interaction[, selected_features]ASNA_4‘   1   754282 914686
## - ‘X_train_interaction[, selected_features]ASNA_5‘   1   755771 916175
## - ‘X_train_interaction[, selected_features]ASNAC_0‘  1   755943 916348
## - ‘X_train_interaction[, selected_features]NAD_3‘    1   757998 918402
##
## Step:  AIC=911302
## .outcome ~ ‘X_train_interaction[, selected_features]NCD_0‘ +
##     ‘X_train_interaction[, selected_features]NCD_1‘ + ‘X_train_interaction[, selected_features]NCD_2
##     ‘X_train_interaction[, selected_features]NCD_5‘ + ‘X_train_interaction[, selected_features]ASNA_
##     ‘X_train_interaction[, selected_features]ASNA_4‘ + ‘X_train_interaction[, selected_features]ASNA
##     ‘X_train_interaction[, selected_features]NAC_0‘ + ‘X_train_interaction[, selected_features]NAC_2
##     ‘X_train_interaction[, selected_features]NAC_4‘ + ‘X_train_interaction[, selected_features]ASNAC
##     ‘X_train_interaction[, selected_features]NA_0‘ + ‘X_train_interaction[, selected_features]NA_4‘ +
##     ‘X_train_interaction[, selected_features]NA_5‘ + ‘X_train_interaction[, selected_features]NAD_0‘
##     ‘X_train_interaction[, selected_features]NAD_1‘ + ‘X_train_interaction[, selected_features]NAD_2
##     ‘X_train_interaction[, selected_features]NAD_3‘ + ‘X_train_interaction[, selected_features]NAD_4
##     ‘X_train_interaction[, selected_features]NAD_5‘
##
##                                                      Df Deviance    AIC
## <none>                                                   750898 911302
## - ‘X_train_interaction[, selected_features]NCD_5‘    1   750900 911302
## - ‘X_train_interaction[, selected_features]NA_4‘     1   750901 911303
## - ‘X_train_interaction[, selected_features]NCD_1‘    1   750905 911308
## - ‘X_train_interaction[, selected_features]NAD_5‘    1   750910 911312
## - ‘X_train_interaction[, selected_features]NAC_2‘    1   750916 911318
## - ‘X_train_interaction[, selected_features]NAD_4‘    1   750920 911322
## - ‘X_train_interaction[, selected_features]NAD_1‘    1   750924 911326
## - ‘X_train_interaction[, selected_features]NCD_0‘    1   750926 911329
## - ‘X_train_interaction[, selected_features]NAC_4‘    1   750973 911375
## - ‘X_train_interaction[, selected_features]NA_5‘     1   751011 911413
## - ‘X_train_interaction[, selected_features]NAD_0‘    1   751048 911450
## - ‘X_train_interaction[, selected_features]NCD_2‘    1   751072 911474
## - ‘X_train_interaction[, selected_features]NAD_2‘    1   751103 911505
## - ‘X_train_interaction[, selected_features]NA_0‘     1   751407 911809
## - ‘X_train_interaction[, selected_features]ASNA_1‘   1   751700 912102
## - ‘X_train_interaction[, selected_features]NAC_0‘    1   751736 912139
## - ‘X_train_interaction[, selected_features]ASNA_4‘   1   754285 914687
## - ‘X_train_interaction[, selected_features]ASNA_5‘   1   755796 916198
## - ‘X_train_interaction[, selected_features]ASNAC_0‘  1   755946 916349
## - ‘X_train_interaction[, selected_features]NAD_3‘    1   758001 918404
```

```r
#save(fit_aic,file = "./fit_aic")
#load("./fit_aic")
Y_pred_train_aic = predict(fit_aic)
Y_pred_test_aic = predict(fit_aic,newdata = X_test2)
errors_aic = errors(Y_pred_train_aic,Y_pred_test_aic,train[1:31969,]$label,train[31970:40000,]$label)

print(errors_aic)
```

```
## $logMSE_train
## [1] 0.4809899
##
## $logMSE_test
## [1] 0.4731883
##
## $MAPE_train
## [1] 0.9028123
##
## $MAPE_test
## [1] 0.8532535
```

```
errors_all=rbind(errors_full,errors_null,errors_aic)
```

```
errors_all
```

```
##             logMSE_train logMSE_test MAPE_train MAPE_test
## errors_full 0.497836     0.5223486   0.9245675  1.102446
## errors_null 5.270964     5.262991    14.52274   14.58855
## errors_aic  0.4809899    0.4731883   0.9028123  0.8532535
```

On peut constater que logMSE et MAPE de modèle “poisson_lasso_selected_AIC” sont très petites par rapport les autre modèles. Donc il est le modèle qui a les meilleurs performances

EXO4 4.1)

```
model_quasi = glm(label~X_train_interaction[,selected_features]-1 ,data=train ,family ="quasipoisson")
summary(model_quasi)
```

```
##
## Call:
## glm(formula = label ~ X_train_interaction[, selected_features] -
##     1, family = "quasipoisson", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -43.143   -2.740   -1.216    0.465   65.246
##
## Coefficients:
##                                                     Estimate Std. Error
## X_train_interaction[, selected_features](Intercept)  6.647e-01  1.645e-02
## X_train_interaction[, selected_features]NCD_0       -2.841e-01  2.350e-01
## X_train_interaction[, selected_features]NCD_1        7.214e-02  2.044e-01
## X_train_interaction[, selected_features]NCD_2       -4.352e-01  1.735e-01
## X_train_interaction[, selected_features]NCD_3       -4.365e-01  2.651e-01
## X_train_interaction[, selected_features]NCD_4        1.565e-01  3.002e-01
## X_train_interaction[, selected_features]NCD_5       -1.134e-02  3.078e-01
## X_train_interaction[, selected_features]ASNA_1       6.377e+01  9.533e+00
## X_train_interaction[, selected_features]ASNA_4      -1.052e+02  1.097e+01
## X_train_interaction[, selected_features]ASNA_5      -1.563e+02  1.176e+01
## X_train_interaction[, selected_features]NAC_0       -1.633e-01  3.500e-02
## X_train_interaction[, selected_features]NAC_2       -2.726e-03  3.355e-02
## X_train_interaction[, selected_features]NAC_4        3.171e-02  3.612e-02
## X_train_interaction[, selected_features]NAC_5        3.536e-02  3.647e-02
## X_train_interaction[, selected_features]ASNAC_0      3.560e+02  2.713e+01
## X_train_interaction[, selected_features]NA_0         1.074e-01  2.417e-02
## X_train_interaction[, selected_features]NA_4        -1.433e-02  2.484e-02
## X_train_interaction[, selected_features]NA_5         5.864e-02  2.287e-02
## X_train_interaction[, selected_features]NAD_0        5.951e-01  2.411e-01
## X_train_interaction[, selected_features]NAD_1        7.382e-03  2.047e-01
## X_train_interaction[, selected_features]NAD_2        5.265e-01  1.811e-01
## X_train_interaction[, selected_features]NAD_3        5.084e-01  2.655e-01
## X_train_interaction[, selected_features]NAD_4       -8.890e-02  3.065e-01
## X_train_interaction[, selected_features]NAD_5        2.664e-01  3.138e-01
##                                                     t value Pr(>|t|)
## X_train_interaction[, selected_features](Intercept)  40.417  < 2e-16 ***
## X_train_interaction[, selected_features]NCD_0        -1.209  0.22670
## X_train_interaction[, selected_features]NCD_1         0.353  0.72412
## X_train_interaction[, selected_features]NCD_2        -2.509  0.01211 *
## X_train_interaction[, selected_features]NCD_3        -1.647  0.09965 .
## X_train_interaction[, selected_features]NCD_4         0.521  0.60220
## X_train_interaction[, selected_features]NCD_5        -0.037  0.97062
## X_train_interaction[, selected_features]ASNA_1        6.690 2.26e-11 ***
## X_train_interaction[, selected_features]ASNA_4       -9.588  < 2e-16 ***
## X_train_interaction[, selected_features]ASNA_5      -13.291  < 2e-16 ***
## X_train_interaction[, selected_features]NAC_0        -4.665 3.09e-06 ***
## X_train_interaction[, selected_features]NAC_2        -0.081  0.93525
## X_train_interaction[, selected_features]NAC_4         0.878  0.37998
## X_train_interaction[, selected_features]NAC_5         0.970  0.33229
## X_train_interaction[, selected_features]ASNAC_0      13.119  < 2e-16 ***
## X_train_interaction[, selected_features]NA_0          4.443 8.90e-06 ***
## X_train_interaction[, selected_features]NA_4         -0.577  0.56392
## X_train_interaction[, selected_features]NA_5          2.564  0.01035 *
## X_train_interaction[, selected_features]NAD_0         2.469  0.01357 *
## X_train_interaction[, selected_features]NAD_1         0.036  0.97123
## X_train_interaction[, selected_features]NAD_2         2.907  0.00366 **
## X_train_interaction[, selected_features]NAD_3         1.915  0.05550 .
## X_train_interaction[, selected_features]NAD_4        -0.290  0.77178
## X_train_interaction[, selected_features]NAD_5         0.849  0.39598
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasipoisson family taken to be 34.22444)
##
##     Null deviance: 42523213  on 40000  degrees of freedom
## Residual deviance:   931112  on 39976  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
#Dispersion parameter for quasipoisson family taken to be 35.27498
```

On a trouvé la paramètre dispersion 34.12203.

4.2)

On voit que std.error pour chaque variable augmente par rapport les dernières modèles. Donc l’intervalle de confiance est plus large. Certaine variables deviennent non significatives. On sélectionne donc les variables qui restent encore significatives.

4.3)

les résidus de déviance:

```
X_beta=log(model_quasi$fitted.values)
X_beta2=model_quasi$fitted.values

residus=2*ifelse(train$label==0,0,(train$label*(log(train$label)-X_beta)-(train$label-X_beta2)))/34.122
```

4.4) Eliminer les indivdus pour lesquels le résidus dépasse 4 en valeur absolu et refaire une estimation:

```
residus2=2*abs(ifelse(train$label==0,0,(train$label*(log(train$label)-X_beta)-(train$label-X_beta2)))/34

train_filtre=train[-which(residus2==FALSE),]

model_quasi2 = glm(label~ NCD_0+NCD_1+NCD_2+ NCD_0+   NCD_1+   NCD_2+   NCD_3+   NCD_4+   NCD_5+   AI_0+
 NAC_2+   NAC_3+   NAC_4+   NAC_5+   ASNAC_0+ ASNAC_1+ ASNAC_2+ NA_1+    NA_5+    ADL_0+   NAD_1+
NAD_2+   NAD_3+   NAD_4+   NAD_5 ,data=train_filtre,family ="quasipoisson")
summary(model_quasi2)
```

```
##
## Call:
## glm(formula = label ~ NCD_0 + NCD_1 + NCD_2 + NCD_0 + NCD_1 +
##     NCD_2 + NCD_3 + NCD_4 + NCD_5 + AI_0 + ASNA_0 + ASNA_5 +
##     BL_0 + NAC_1 + NAC_2 + NAC_3 + NAC_4 + NAC_5 + ASNAC_0 +
##     ASNAC_1 + ASNAC_2 + NA_1 + NA_5 + ADL_0 + NAD_1 + NAD_2 +
##     NAD_3 + NAD_4 + NAD_5, family = "quasipoisson", data = train_filtre)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -44.090   -2.195   -0.733    0.925   16.725
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.023e-01  3.436e-02  17.531  < 2e-16 ***
## NCD_0        3.713e-01  7.053e-03  52.643  < 2e-16 ***
## NCD_1        2.585e-02  1.314e-01   0.197 0.843998
## NCD_2       -3.841e-01  1.104e-01  -3.479 0.000504 ***
## NCD_3       -5.479e-02  1.812e-01  -0.302 0.762389
## NCD_4        1.738e-01  2.005e-01   0.867 0.386033
## NCD_5        2.311e-01  2.041e-01   1.132 0.257558
## AI_0        -1.708e-02  5.264e-03  -3.245 0.001175 **
## ASNA_0       1.101e+02  2.831e+01   3.889 0.000101 ***
## ASNA_5      -2.124e+02  8.052e+00 -26.372  < 2e-16 ***
## BL_0        -1.641e-01  3.594e-02  -4.568 4.95e-06 ***
## NAC_1       -3.903e-02  2.325e-02  -1.678 0.093269 .
## NAC_2        1.387e-02  2.276e-02   0.609 0.542391
## NAC_3       -2.784e-02  2.328e-02  -1.196 0.231644
## NAC_4        2.476e-02  2.246e-02   1.102 0.270368
## NAC_5       -1.333e-02  2.380e-02  -0.560 0.575233
## ASNAC_0      6.488e+01  4.899e+01   1.325 0.185333
## ASNAC_1      1.803e+02  1.841e+01   9.797  < 2e-16 ***
## ASNAC_2     -3.008e+01  1.074e+01  -2.800 0.005107 **
## NA_1         9.638e-03  1.614e-02   0.597 0.550536
## NA_5         6.020e-02  1.273e-02   4.730 2.25e-06 ***
## ADL_0        4.254e-02  1.402e-02   3.034 0.002418 **
## NAD_1        7.455e-02  1.357e-01   0.549 0.582747
## NAD_2        4.651e-01  1.156e-01   4.025 5.72e-05 ***
## NAD_3        1.422e-01  1.848e-01   0.770 0.441370
## NAD_4       -1.555e-01  2.036e-01  -0.764 0.444967
## NAD_5        5.202e-02  2.083e-01   0.250 0.802753
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasipoisson family taken to be 12.2704)
##
##     Null deviance: 7935685  on 38735  degrees of freedom
## Residual deviance:  433433  on 38709  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
#Dispersion parameter for quasipoisson family taken to be 35.27498
```

On voit que la dispersion parameter est diminuée à 12.34157 et aussi le std.error des variables sont plus petit qu’avant. Donc ce nouveau modèle a des meilleures performances que les précédents.