Premier rapport

Apprentissage statistique en actuariat
ACT-4114

Équipe 09

Rapport

Nom de votre TP

Par	*Numéro d'identification*
Danny Larochelle	111 174 586
Étudiant 2	XYZ XYZ XYZ
Étudiant 3	YZX YZX YZX
Étudiant 4	ZYX ZYX ZYX
Étudiant 5	XYD XYD XYD

Travail présenté à
Monsieur
Olivier Côté]

13 mars 2023

UNIVERSITÉ LAVAL

Faculté des sciences et de génie
École d'actuariat

Table des Matières

Introduction **2**

Analyse exploratoire des données **2**
Sélection des variables 2
Création de la nouvelle variable réponse 2
Explication des variables 6
Transformation des variables 8

Conclusion **15**

Bibliographie **16**

Introduction

Analyse exploratoire des données

Sélection des variables

La première étape du travail a consisté à réduire la dimension du jeu de données. En effet, celui-ci est constitué de 41 variables, dont une bonne partie n'étant pas utiles dans le contecte de l'analyse des montants de réclamation.

Sans effectuer aucune analyse statistique, nous avons jugé adéquat de retirer plusieurs variables du modèle, notamment, toutes les variables contenant beaucoup de valeurs manquantes, comme baseFloodElevation, basementEnclosureCrawlspace, elevationCertificateIndicator, elevationDifference, rateMethod et lowestAdjacentGrade. Ces variables sont aussi toutes issues de l'évaluation de quelques uns des bâtiments assurés, alors que plusieurs autres variables telles que numberOfFloorsInTheInsuredBuilding, originalConstructionDate ou encore lowestFloorElevation auront un impact probablement plus marqué sur le modèle sans devoir nécessiter un travail ardu et approximatif d'estimation d'une grande quantité de données manquantes.

Nous avons aussi pris la décision d'enlever les variables temporelles à l'exception de la date de construction du bâtiment (originalConstructionDate) et la date du sinistre (dateOfLoss), puisqu'elles sont les seules variables temporelles pertinentes à notre analyse selon nous.

Création de la nouvelle variable réponse

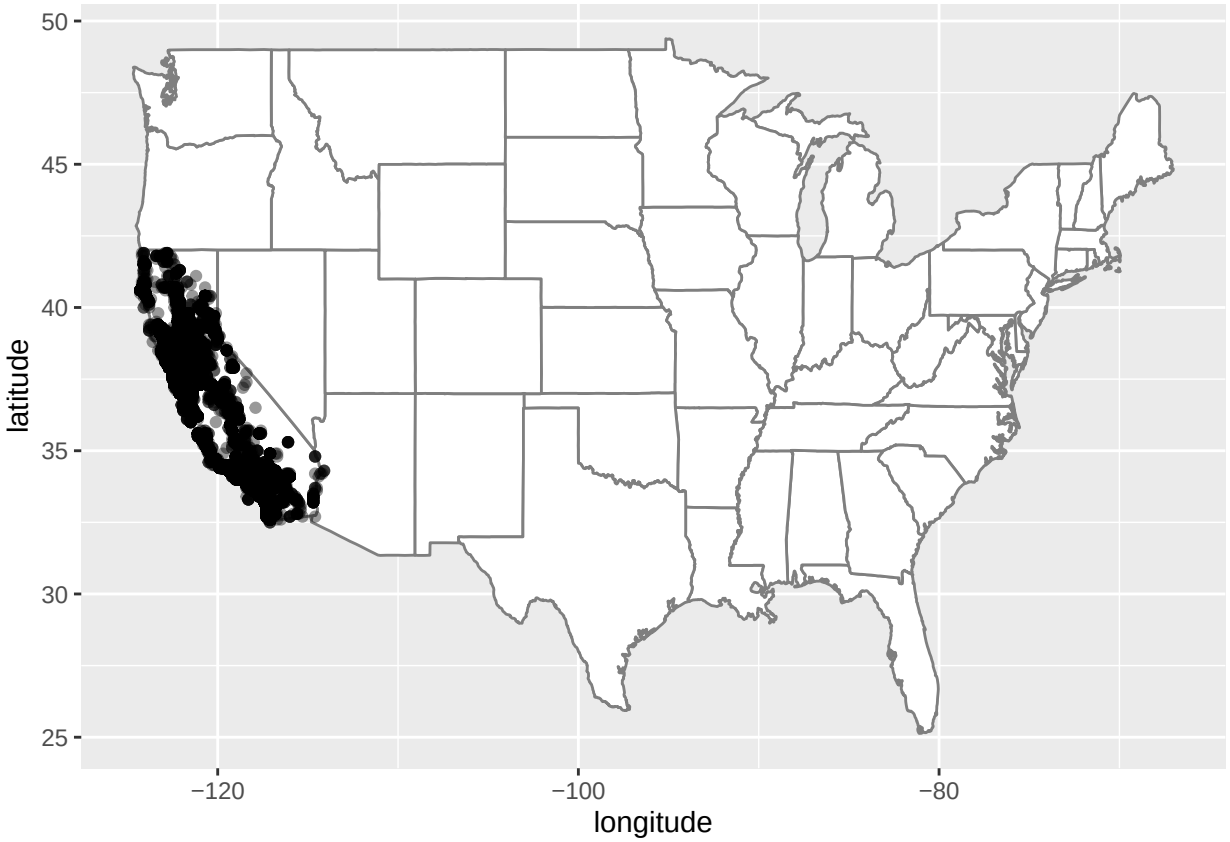
Dans le jeu de données se retrouvent trois colonnes contenant des informations sur les montants de prestations payés en lien avec le bâtiment (amountPaidOnBuildingClaim), les biens (amountPaidOnContentsClaim) et l'augmentation des coûts en lien avec la conformité (amountPaidOnIncreasedCostOfComplianceClaim).

```
data.raw <- read.csv("Flood_California.csv")
```

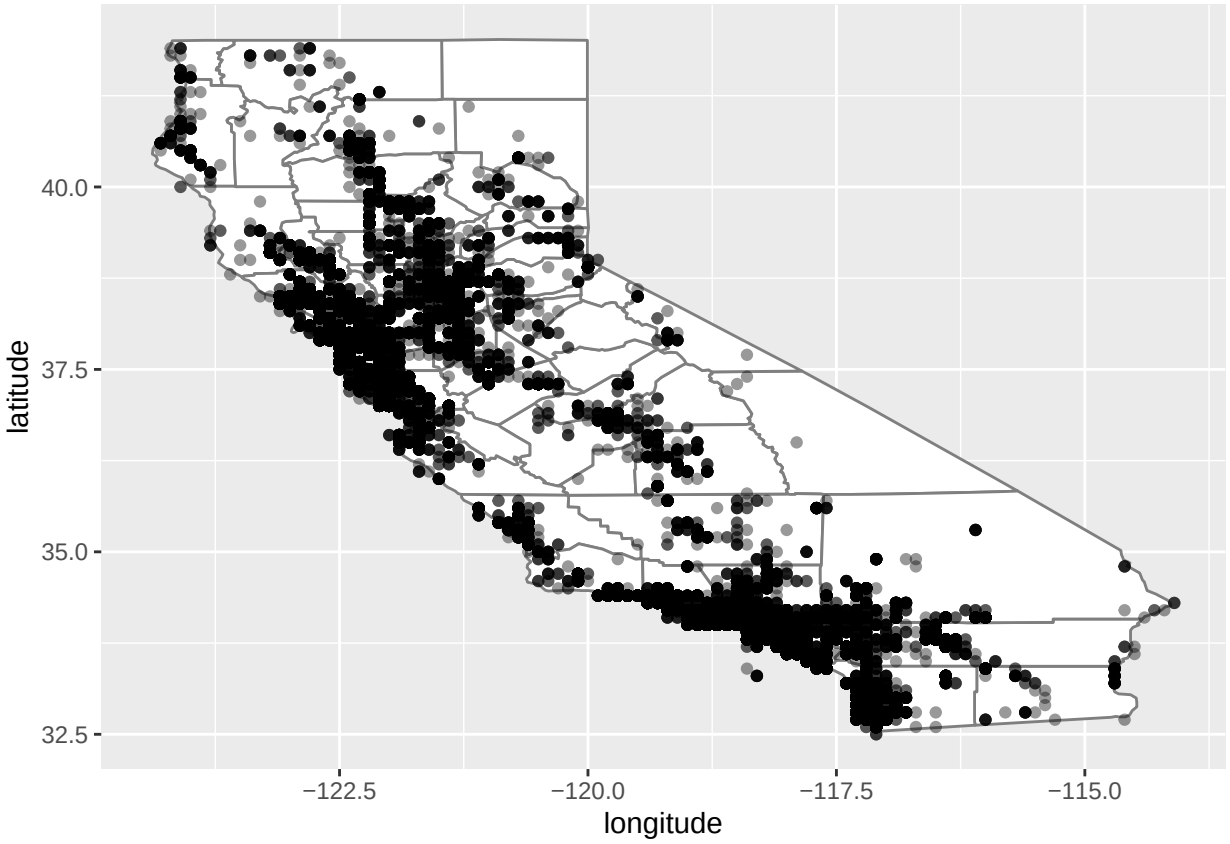
```
## Retirer les variables inutiles
data.rm <- data.raw[, c(1, 3, 4, 5, 6, 13, 14, 15, 16, 21, 25, 28, 33, 39, 41)]
data <- data.raw[, -c(1, 3, 4, 5, 6, 13, 14, 15, 16, 21, 25, 28, 33, 39, 41)]

# Combiner les variables réponses (totalAmount)
data$amountPaidOnBuildingClaim[is.na(data$amountPaidOnBuildingClaim)] <- 0
data$amountPaidOnBuildingClaim <-
  abs(data$amountPaidOnBuildingClaim)
data$amountPaidOnContentsClaim[is.na(data$amountPaidOnContentsClaim)] <-0
data$amountPaidOnContentsClaim <-
  abs(data$amountPaidOnContentsClaim)
data$amountPaidOnIncreasedCostOfComplianceClaim[is.na(data$amountPaidOnIncreasedCostOfComplianceClaim)]
data$amountPaidOnIncreasedCostOfComplianceClaim <-
  abs(data$amountPaidOnIncreasedCostOfComplianceClaim)
data$totalAmount <- apply(data[, 17:19], 1, sum)
data <- data[, -c(17, 18, 19)]

# Retirer les lignes n'étant pas localisées en Californie
data <- data[!is.na(data$longitude),]
data <- data[data$longitude <= -110,]

xdf <- which(is.na(data$countyCode))

# Imputation par régression linéaire des codes de régions (countyCode)
```

```
mod.county <- lm(countyCode ~ latitude + longitude, data = data)
pred.county <- predict(mod.county, newdata = data[is.na(data$countyCode),], type = "response")
data$countyCode[is.na(data$countyCode)] <- pred.county
data <- data[data$countyCode != 32031,]

# Imputation par régression linéaire des codes de régions (countyCode)
# data$communityRatingSystemDiscount <- as.factor(data$communityRatingSystemDiscount)
# levels(data$communityRatingSystemDiscount) <- c("A", "B", "C", "D", "E", "F", "G", "H", "I", "H")
# mod.communityRating <- lm(communityRatingSystemDiscount ~ ., data = data)
# pred.CR <- predict(mod.communityRating, newdata = data[is.na(data$communityRatingSystemDiscount),], t
# data$communityRatingSystemDiscount[is.na(data$communityRatingSystemDiscount)] <- pred.CR
```

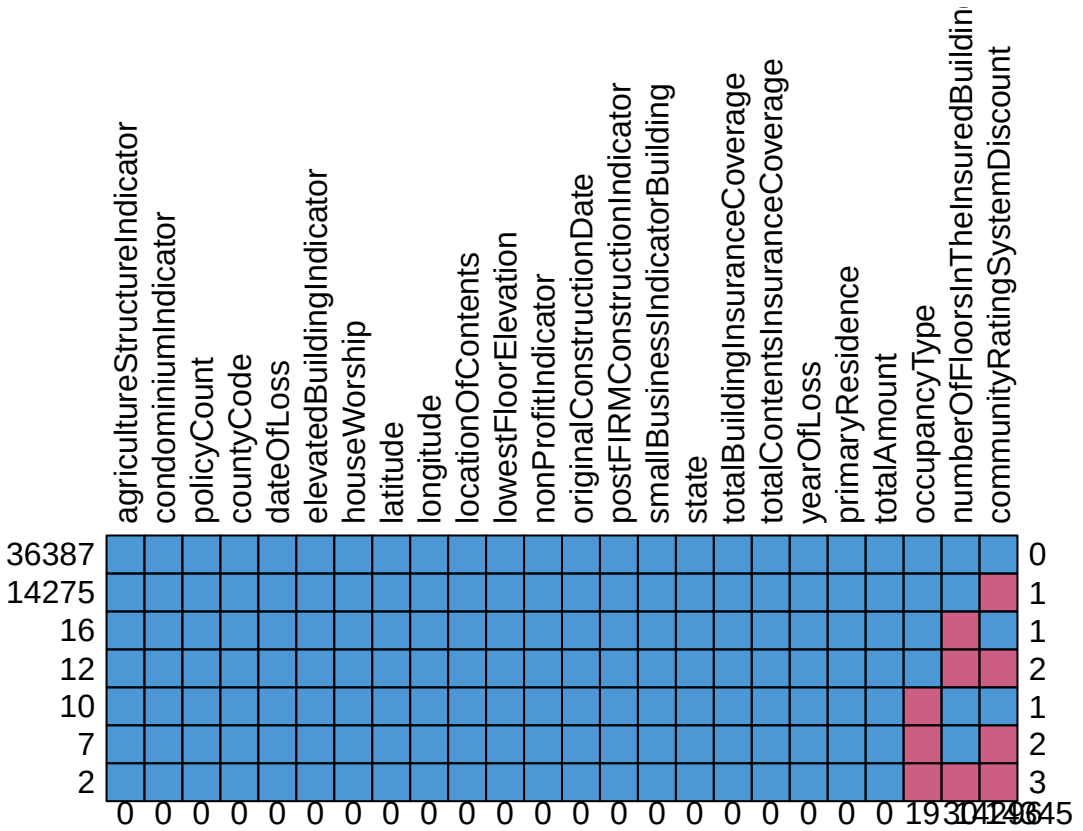
```
mapUSA <- borders(database = "state",
                  colour="gray50", fill="white")
ggplot(data = data, aes(x = longitude, y = latitude)) +
    mapUSA + geom_point(alpha = .4)
```




```
mapCalifornia <- borders(database = "county", region = "california",
                  colour="gray50", fill="white")
ggplot(data = data, aes(x = longitude, y = latitude, col= )) +
    mapCalifornia + geom_point(alpha = .4)
```




```
md.pattern(data, rotate.names = T)
```




```
##       agricultureStructureIndicator condominiumIndicator policyCount countyCode
## 36387                             1                    1           1          1
## 14275                             1                    1           1          1
## 16                                1                    1           1          1
## 12                                1                    1           1          1
## 10                                1                    1           1          1
## 7                                 1                    1           1          1
## 2                                 1                    1           1          1
##                                   0                    0           0          0
##       dateOfLoss elevatedBuildingIndicator houseWorship latitude longitude
## 36387          1                         1           1        1         1
## 14275          1                         1           1        1         1
## 16             1                         1           1        1         1
## 12             1                         1           1        1         1
## 10             1                         1           1        1         1
## 7              1                         1           1        1         1
## 2              1                         1           1        1         1
##                0                         0           0        0         0
##       locationOfContents lowestFloorElevation nonProfitIndicator
## 36387                  1                    1                  1
## 14275                  1                    1                  1
## 16                     1                    1                  1
## 12                     1                    1                  1
## 10                     1                    1                  1
## 7                      1                    1                  1
## 2                      1                    1                  1
```

```
##                        0                   0                  0
##       originalConstructionDate postFIRMConstructionIndicator
## 36387                        1                             1
## 14275                        1                             1
## 16                           1                             1
## 12                           1                             1
## 10                           1                             1
## 7                            1                             1
## 2                            1                             1
##                              0                             0
##       smallBusinessIndicatorBuilding state totalBuildingInsuranceCoverage
## 36387                              1     1                              1
## 14275                              1     1                              1
## 16                                 1     1                              1
## 12                                 1     1                              1
## 10                                 1     1                              1
## 7                                  1     1                              1
## 2                                  1     1                              1
##                                    0     0                              0
##       totalContentsInsuranceCoverage yearOfLoss primaryResidence totalAmount
## 36387                              1          1                 1           1
## 14275                              1          1                 1           1
## 16                                 1          1                 1           1
## 12                                 1          1                 1           1
## 10                                 1          1                 1           1
## 7                                  1          1                 1           1
## 2                                  1          1                 1           1
##                                    0          0                 0           0
##       occupancyType numberOfFloorsInTheInsuredBuilding
## 36387             1                                  1
## 14275             1                                  1
## 16                1                                  0
## 12                1                                  0
## 10                0                                  1
## 7                 0                                  1
## 2                 0                                  0
##                  19                                 30
##       communityRatingSystemDiscount
## 36387                             1     0
## 14275                             0     1
## 16                                1     1
## 12                                0     2
## 10                                1     1
## 7                                 0     2
## 2                                 0     3
##                                14296 14345
```

Explication des variables

```
colnames(data) <- c("EstAgricole",
                    "EstCondo",
                    "NbPolice",
                    "CountyCode",
```

```
                    "TypeRabais",
                    "DateSiniste",
                    "EstElevé",
                    "EstReligieux",
                    "Latitude",
                    "Longitude",
                    "LocContenu",
                    "ÉlévationPlancher",
                    "NbÉtage",
                    "EstNonProfit",
                    "TypeHabitation",
                    "DateConstruction",
                    "ApresFIRM",
                    "EstPME",
                    "État",
                    "CouvertureBatiment",
                    "CouvertureContenu",
                    "AnneePerte",
                    "EstResPrimaire",
                    "PerteTotal")


data$EstAgricole <- factor(data$EstAgricole)

data$EstCondo[data$EstCondo == ""] <- "NA"
data$EstCondo[data$EstCondo == "N"] <- "0"
data$EstCondo[data$EstCondo == "A" | data$EstCondo == "H" | data$EstCondo == "L" | data$EstCondo == "U"]
data$EstCondo <- factor(data$EstCondo)

data$TypeRabais <- factor(data$TypeRabais)

#data$DateSinistre

data$EstElevé <- factor(data$EstElevé)

data$EstReligieux <- factor(data$EstReligieux)

data$LocContenu <- factor(data$LocContenu)

data$NbÉtage <- factor(data$NbÉtage)

data$EstNonProfit <- factor(data$EstNonProfit)

data$TypeHabitation <- factor(data$TypeHabitation)
# On rassemble des categories?

# DateConstruction

data$ApresFIRM <- factor(data$ApresFIRM)

data$EstPME <- factor(data$EstPME)

# Enlever etat?
```

```
data$EstResPrimaire <- factor(data$EstResPrimaire)

data$PerteTotal <- abs(data$PerteTotal)
```

Transformation des variables

```
str(data)
```

```
## 'data.frame':    50709 obs. of  24 variables:
##  $ EstAgricole       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EstCondo          : Factor w/ 3 levels "0","1","NA": 1 1 1 1 1 1 1 1 1 1 ...
##  $ NbPolice          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ CountyCode        : num  6013 6059 6085 6059 6059 ...
##  $ TypeRabais        : Factor w/ 9 levels "1","2","3","5",..: NA 7 5 NA NA NA 1 NA NA 5 ...
##  $ DateSiniste       : chr  "1995-01-09T00:00:00.000Z" "2017-01-22T00:00:00.000Z" "1998-02-03T00:00:0
##  $ EstElevé          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
##  $ EstReligieux      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Latitude          : num  37.9 33.9 37.5 33.8 33.4 39.3 38.7 37.6 32.8 38.1 ...
##  $ Longitude         : num  -122 -118 -122 -118 -118 ...
##  $ LocContenu        : Factor w/ 8 levels "0","1","2","3",..: 1 1 1 1 5 1 1 1 1 1 ...
##  $ ÉlévationPlancher : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NbÉtage           : Factor w/ 6 levels "1","2","3","4",..: 2 1 1 1 2 2 1 2 1 3 ...
##  $ EstNonProfit      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ TypeHabitation    : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ DateConstruction  : chr  "1970-01-01T00:00:00.000Z" "1969-01-01T00:00:00.000Z" "1954-01-01T00:00:0
##  $ ApresFIRM         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EstPME            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ État              : chr  "CA" "CA" "CA" "CA" ...
##  $ CouvertureBatiment: int  185000 250000 203500 120000 250000 20000 173700 185000 86400 20000 ...
##  $ CouvertureContenu : int  60000 100000 63000 0 100000 8000 30000 10000 20800 5000 ...
##  $ AnneePerte        : int  1995 2017 1998 1999 2017 2017 1995 1997 1993 1995 ...
##  $ EstResPrimaire    : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 1 1 1 ...
##  $ PerteTotal        : num  2261 12183 75745 0 9766 ...
```

```
summary(data)
```

```
##  EstAgricole EstCondo      NbPolice         CountyCode     TypeRabais
##  0:50682     0 :48664   Min.   :  1.000   Min.   :6001   7      :13637
##  1:   27     1 :  952   1st Qu.:  1.000   1st Qu.:6037   10     : 7081
##              NA: 1093   Median :  1.000   Median :6065   6      : 4102
##                         Mean   :  1.056   Mean   :6063   8      : 3158
##                         3rd Qu.:  1.000   3rd Qu.:6087   5      : 3092
##                         Max.   :103.000   Max.   :6115   (Other): 5343
##                                                          NA's   :14296
##  DateSiniste        EstElevé  EstReligieux    Latitude       Longitude
##  Length:50709       0:45059   0:50681      Min.   :32.50   Min.   :-124.3
##  Class :character   1: 5650   1:   28      1st Qu.:34.10   1st Qu.:-122.5
##  Mode  :character                          Median :37.40   Median :-121.5
##                                            Mean   :36.49   Mean   :-120.6
##                                            3rd Qu.:38.50   3rd Qu.:-118.4
##                                            Max.   :41.90   Max.   :-114.1
##
##    LocContenu    ÉlévationPlancher NbÉtage      EstNonProfit TypeHabitation
##  0      :21164   Min.   : -10.90   1   :30191   0:50698      1      :39871
```

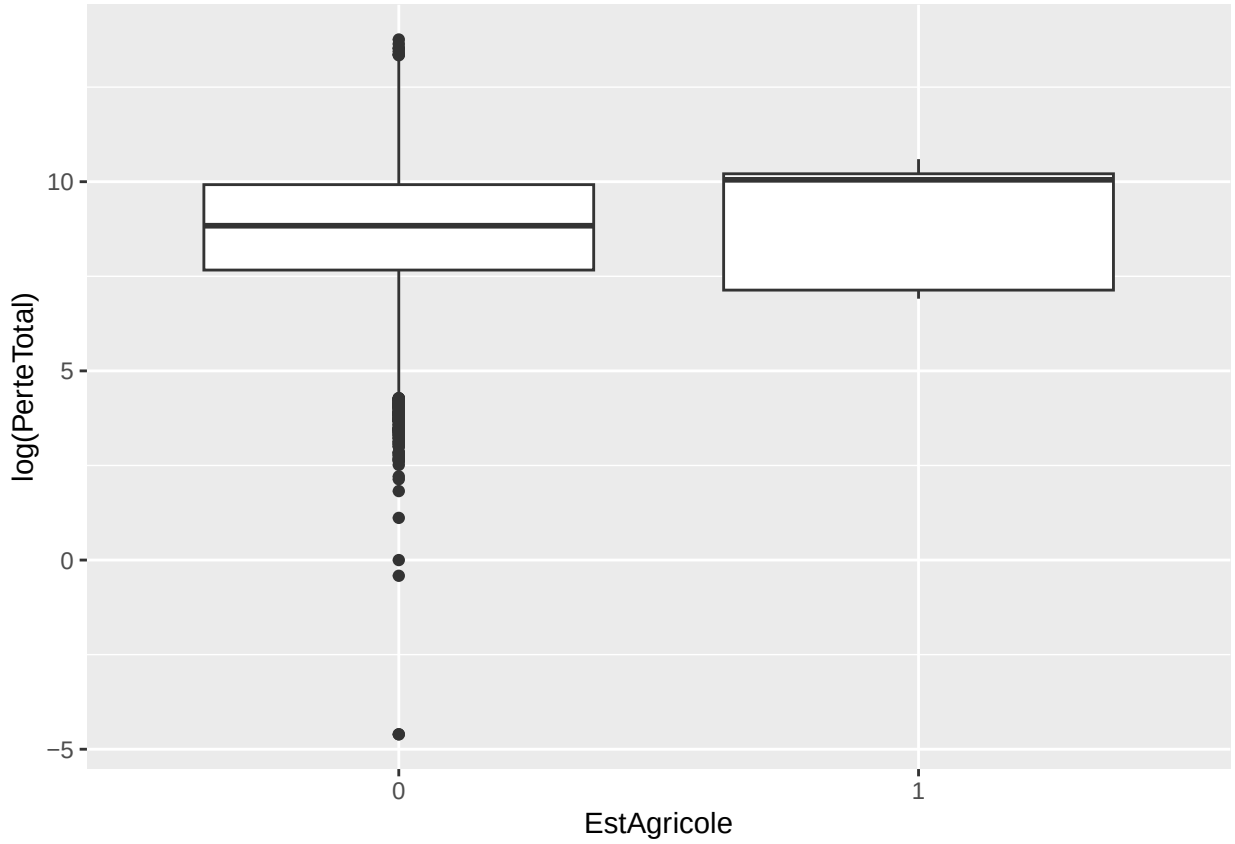
```
##  3      :18325   1st Qu.:   0.00   2   :14949   1:   11      4      : 5089
##  4      : 7865   Median :   0.00   3   : 4183                2      : 3318
##  2      : 2544   Mean   :  14.47   4   : 1013                3      : 1535
##  7      :  406   3rd Qu.:   0.00   5   :  310                6      :  720
##  6      :  246   Max.   :9992.00   6   :   33                (Other):  157
##  (Other):  159                     NA's:   30                NA's   :   19
##  DateConstruction   ApresFIRM EstPME        État           CouvertureBatiment
##  Length:50709       0:44359   0:50529   Length:50709       Min.   :       0
##  Class :character   1: 6350   1:  180   Class :character   1st Qu.:   35000
##  Mode  :character                       Mode  :character   Median :  100000
##                                                            Mean   :  125032
##                                                            3rd Qu.:  185000
##                                                            Max.   :22656800
##
##  CouvertureContenu   AnneePerte   EstResPrimaire   PerteTotal
##  Min.   :     0    Min.   :1974   0:36089        Min.   :     0
##  1st Qu.:     0    1st Qu.:1986   1:14620        1st Qu.:     0
##  Median :  5000    Median :1995                  Median :  2065
##  Mean   : 23027    Mean   :1996                  Mean   : 12480
##  3rd Qu.: 30000    3rd Qu.:2001                  3rd Qu.: 11468
##  Max.   :500000    Max.   :2022                  Max.   :945108
##
```

```
table(data$EstCondo)
```

```
##
##     0     1    NA
## 48664   952  1093
```

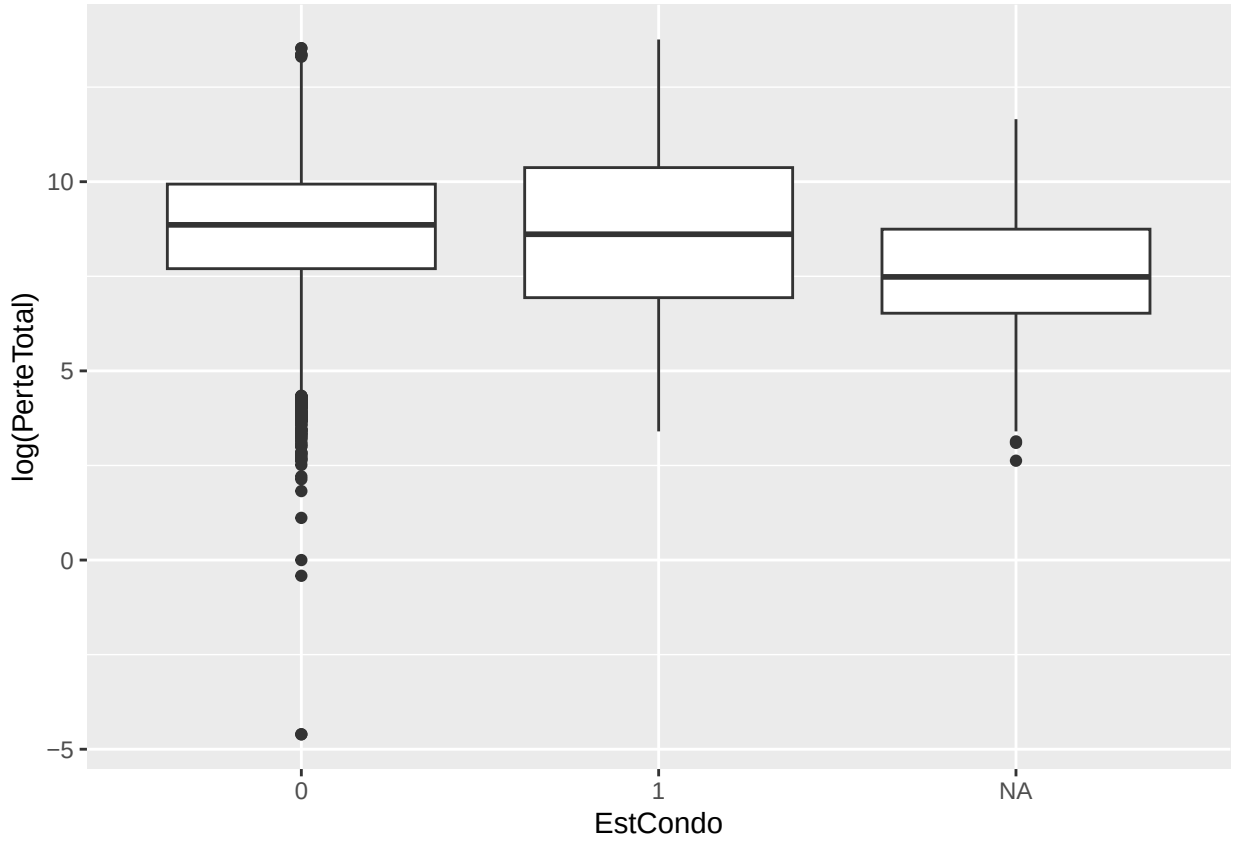
```
#Changement d'echelle, log
ggplot(data = data, aes(x = EstAgricole, y = log(PerteTotal) ))+
  geom_boxplot()
```

```
## Warning: Removed 17163 rows containing non-finite values (`stat_boxplot()`).
```

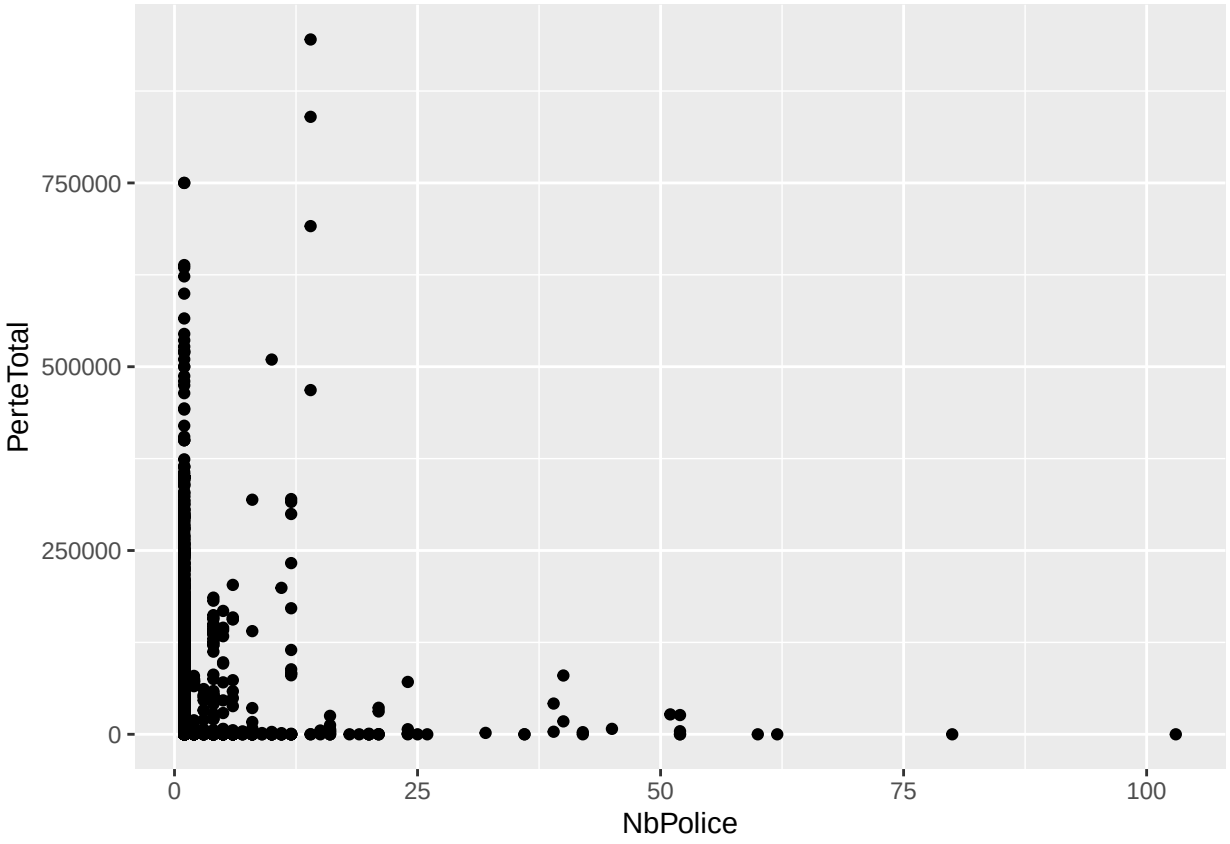



```
#Changement d'echelle, log
ggplot(data = data, aes(x = EstCondo, y = log(PerteTotal) ))+
  geom_boxplot()
```

```
## Warning: Removed 17163 rows containing non-finite values (`stat_boxplot()`).
```

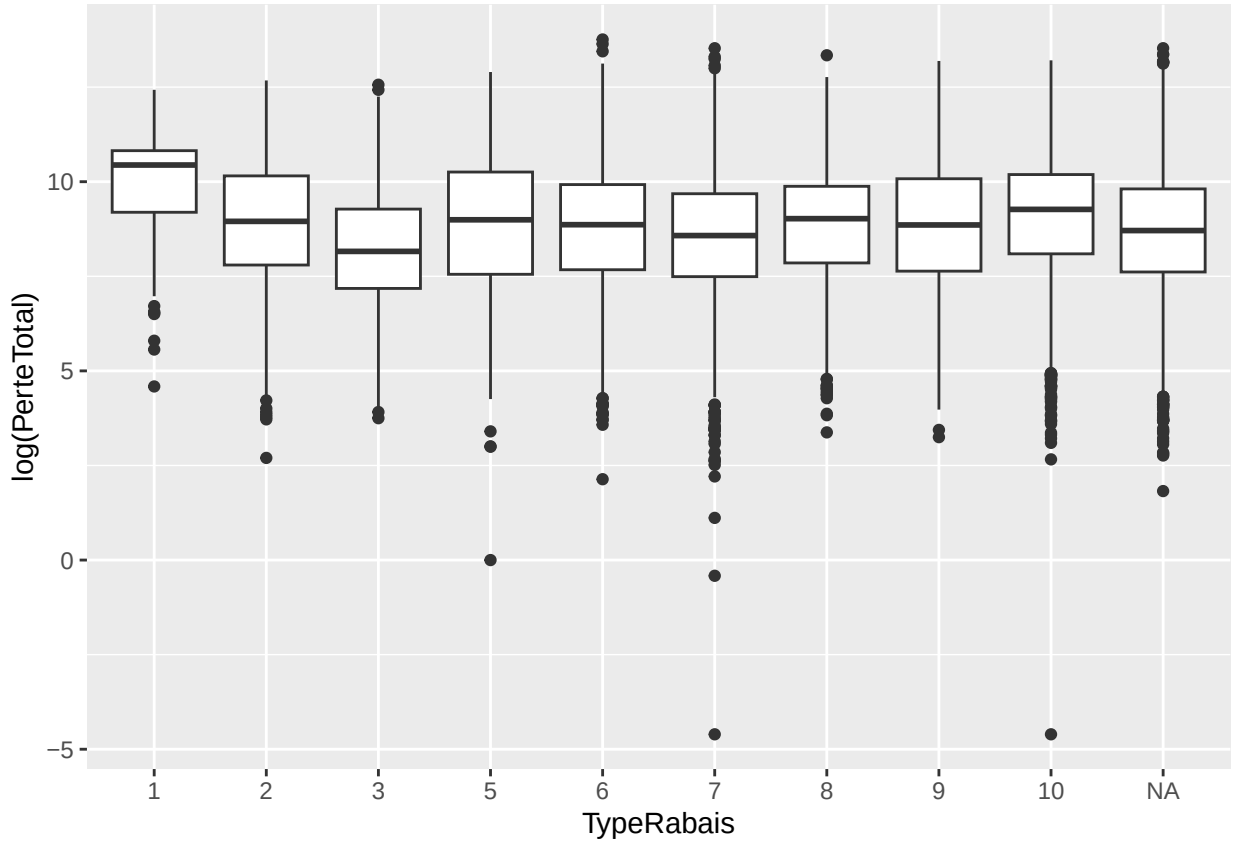



```
ggplot(data = data, aes(x = NbPolice, y = PerteTotal ))+
  geom_point()
```

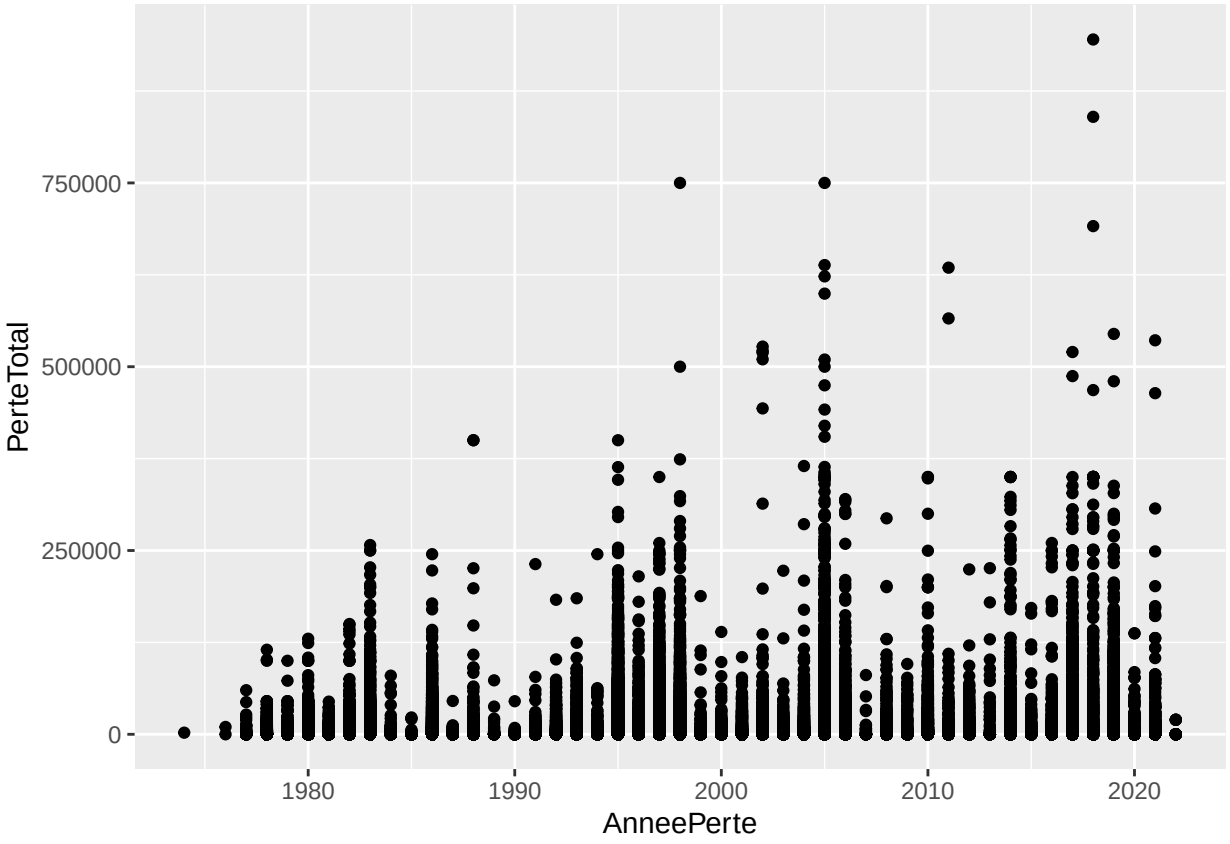



```
ggplot(data = data, aes(x = TypeRabais, y = log(PerteTotal) ))+
  geom_boxplot()
```

```
## Warning: Removed 17163 rows containing non-finite values (`stat_boxplot()`).
```




```
ggplot(data = data, aes(x = AnneePerte, y = PerteTotal ))+
  geom_point()
```


PerteTotal
750000
500000
250000
0
1980
1990
2000
2010
2020
AnneePerte

Conclusion

Bibliographie

]

Annexe