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

# 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

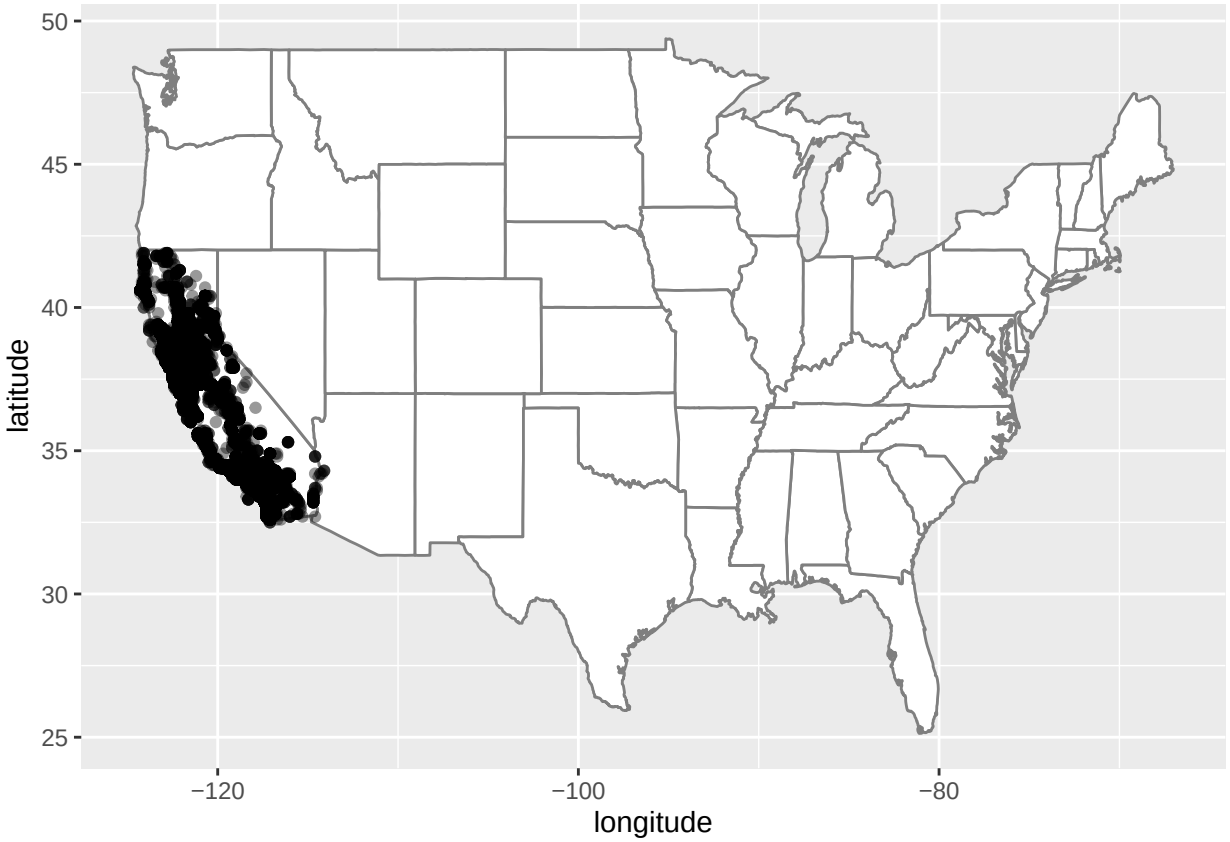
Dans le jeu de données se retrouvent trois colonnes contenant des informations sur les montants de prestations payés en lien avec le bâtiment (amountPaidOnBuildingClaim), les biens (amountPaidOnContentsClaim) et l'augmentation des coûts en lien avec la conformité (amountPaidOnIncreasedCostOfComplianceClaim).

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

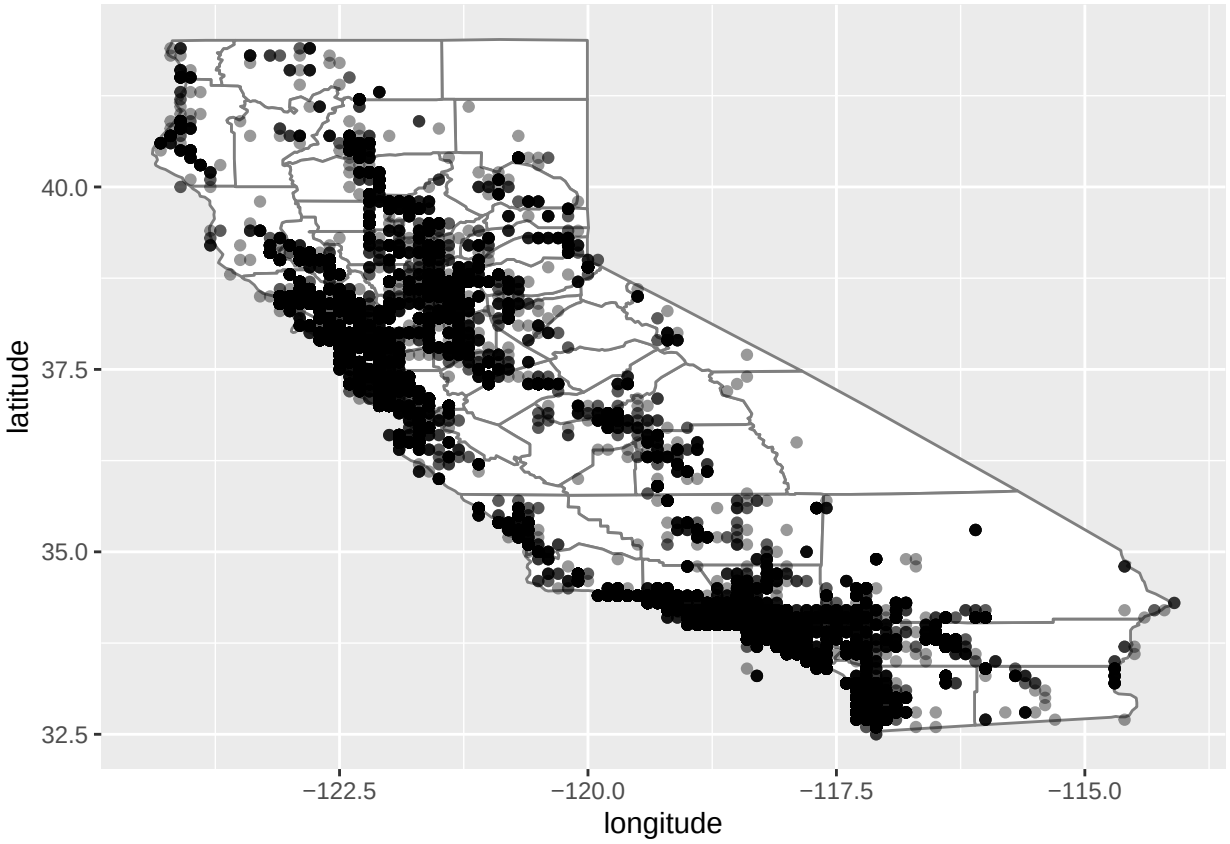
```
## Retirer les variables inutiles
data.rm <- data.raw[, c(1, 3, 4, 5, 6, 13, 14, 15, 16, 21, 25, 28, 33, 39, 41)]
data <- data.raw[, -c(1, 3, 4, 5, 6, 13, 14, 15, 16, 21, 25, 28, 33, 39, 41)]

# 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
```

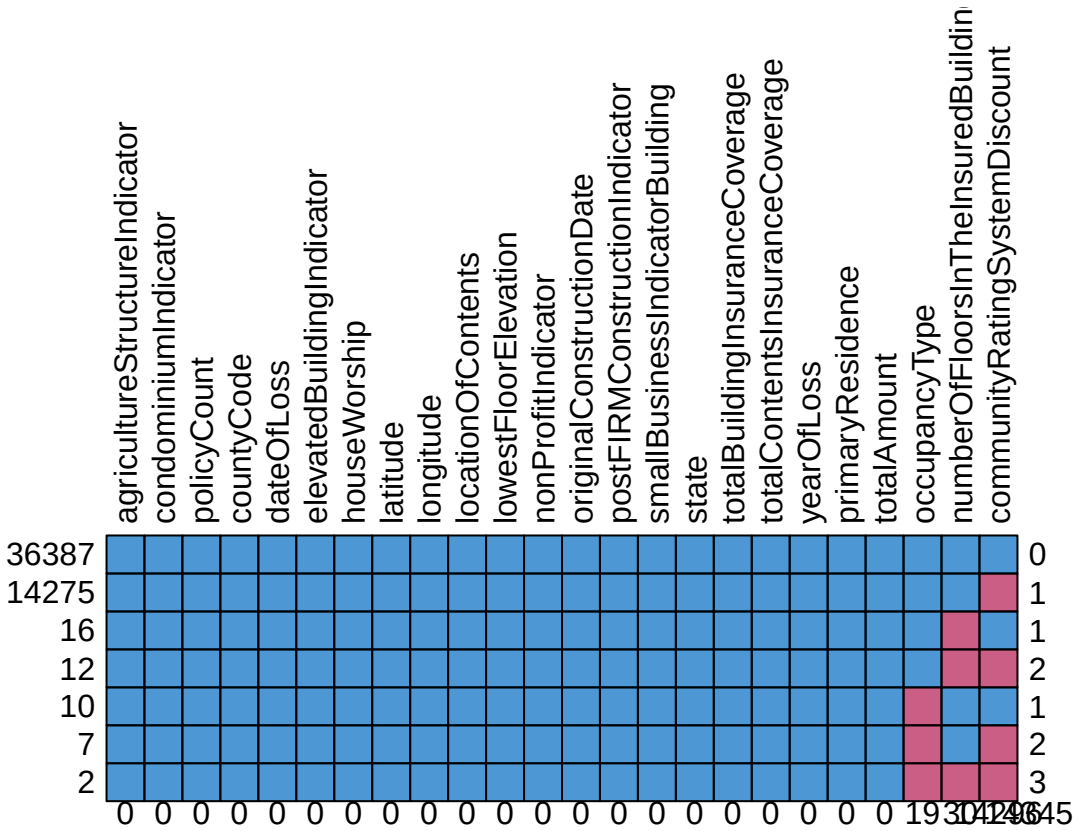
```
mapUSA <- borders(database = "state",
                  colour="gray50", fill="white")
ggplot(data = data, aes(x = longitude, y = latitude)) +
    mapUSA + geom_point(alpha = .4)
```




```
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

```
formatted.data <- data
colnames(formatted.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")
```

## Transformation des variables

```
str(formatted.data)
```

```
## 'data.frame':    50709 obs. of  24 variables:
##  $ EstAgricole       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ EstCondo          : chr  "N" "N" "N" "N" ...
##  $ NbPolice          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ CountyCode        : num  6013 6059 6085 6059 6059 ...
##  $ TypeRabais        : int  NA 8 6 NA NA NA 1 NA NA 6 ...
##  $ DateSiniste       : chr  "1995-01-09T00:00:00.000Z" "2017-01-22T00:00:00.000Z" "1998-02-03T00:00:0
##  $ EstElevé          : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ EstReligieux      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 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        : int  0 0 0 0 4 0 0 0 0 0 ...
##  $ ÉlévationPlancher : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NbÉtage           : int  2 1 1 1 2 2 1 2 1 3 ...
##  $ EstNonProfit      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ TypeHabitation    : int  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         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ EstPME            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 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    : int  0 1 1 1 1 1 1 0 0 0 ...
##  $ PerteTotal        : num  2261 12183 75745 0 9766 ...
```

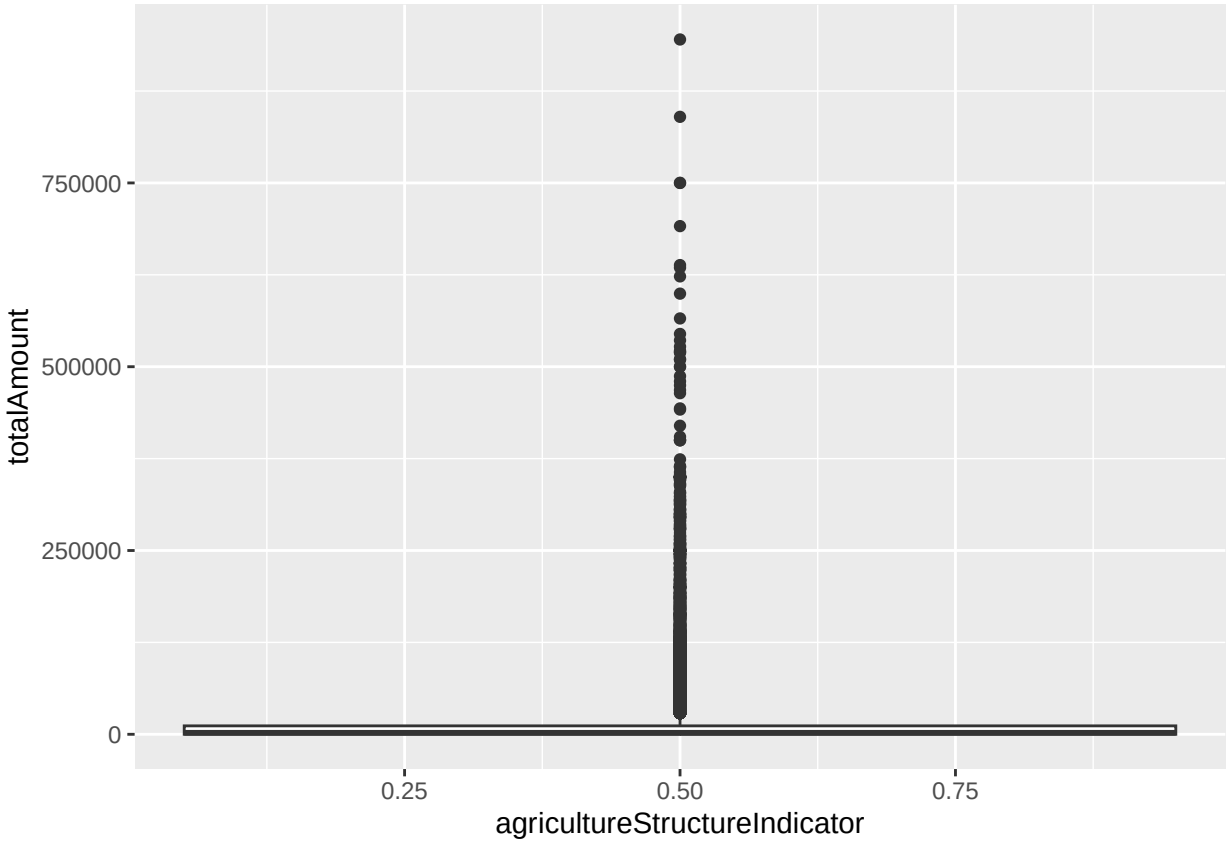
```
summary(formatted.data)
```

```
##   EstAgricole          EstCondo            NbPolice         CountyCode
```

```
##  Min.   :0.0000000   Length:50709       Min.   :  1.000   Min.   :6001
##  1st Qu.:0.0000000   Class :character   1st Qu.:  1.000   1st Qu.:6037
##  Median :0.0000000   Mode  :character   Median :  1.000   Median :6065
##  Mean   :0.0005324                      Mean   :  1.056   Mean   :6063
##  3rd Qu.:0.0000000                      3rd Qu.:  1.000   3rd Qu.:6087
##  Max.   :1.0000000                      Max.   :103.000   Max.   :6115
##
##    TypeRabais     DateSiniste           EstElevé       EstReligieux
##  Min.   : 1.000   Length:50709       Min.   :0.0000   Min.   :0.0000000
##  1st Qu.: 6.000   Class :character   1st Qu.:0.0000   1st Qu.:0.0000000
##  Median : 7.000   Mode  :character   Median :0.0000   Median :0.0000000
##  Mean   : 6.966                      Mean   :0.1114   Mean   :0.0005522
##  3rd Qu.: 8.000                      3rd Qu.:0.0000   3rd Qu.:0.0000000
##  Max.   :10.000                      Max.   :1.0000   Max.   :1.0000000
##  NA's   :14296
##     Latitude       Longitude        LocContenu    ÉlévationPlancher
##  Min.   :32.50   Min.   :-124.3   Min.   :0.000   Min.   : -10.90
##  1st Qu.:34.10   1st Qu.:-122.5   1st Qu.:0.000   1st Qu.:   0.00
##  Median :37.40   Median :-121.5   Median :3.000   Median :   0.00
##  Mean   :36.49   Mean   :-120.6   Mean   :1.905   Mean   :  14.47
##  3rd Qu.:38.50   3rd Qu.:-118.4   3rd Qu.:3.000   3rd Qu.:   0.00
##  Max.   :41.90   Max.   :-114.1   Max.   :7.000   Max.   :9992.00
##
##     NbÉtage       EstNonProfit       TypeHabitation   DateConstruction
##  Min.   :1.000   Min.   :0.0000000   Min.   : 1.000   Length:50709
##  1st Qu.:1.000   1st Qu.:0.0000000   1st Qu.: 1.000   Class :character
##  Median :1.000   Median :0.0000000   Median : 1.000   Mode  :character
##  Mean   :1.548   Mean   :0.0002169   Mean   : 1.532
##  3rd Qu.:2.000   3rd Qu.:0.0000000   3rd Qu.: 1.000
##  Max.   :6.000   Max.   :1.0000000   Max.   :18.000
##  NA's   :30                          NA's   :19
##    ApresFIRM          EstPME            État           CouvertureBatiment
##  Min.   :0.0000   Min.   :0.00000   Length:50709       Min.   :       0
##  1st Qu.:0.0000   1st Qu.:0.00000   Class :character   1st Qu.:   35000
##  Median :0.0000   Median :0.00000   Mode  :character   Median :  100000
##  Mean   :0.1252   Mean   :0.00355                      Mean   :  125032
##  3rd Qu.:0.0000   3rd Qu.:0.00000                      3rd Qu.:  185000
##  Max.   :1.0000   Max.   :1.00000                      Max.   :22656800
##
##  CouvertureContenu   AnneePerte   EstResPrimaire     PerteTotal
##  Min.   :     0    Min.   :1974   Min.   :0.0000   Min.   :     0
##  1st Qu.:     0    1st Qu.:1986   1st Qu.:0.0000   1st Qu.:     0
##  Median :  5000    Median :1995   Median :0.0000   Median :  2065
##  Mean   : 23027    Mean   :1996   Mean   :0.2883   Mean   : 12480
##  3rd Qu.: 30000    3rd Qu.:2001   3rd Qu.:1.0000   3rd Qu.: 11468
##  Max.   :500000    Max.   :2022   Max.   :1.0000   Max.   :945108
##
```

```
ggplot(data = data, aes(x = agricultureStructureIndicator, y = totalAmount))+
  geom_boxplot()
```

```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?
```


750000
500000
250000
0
totalAmount
0.25
0.50
0.75
agricultureStructureIndicator

# Conclusion

## Bibliographie

]

# Annexe