Premier rapport

Apprentissage statistique en actuariat
ACT-4114

Équipe 09

Rapport
Nom de votre TP

Par	*Numéro d'identification*
Danny Larochelle	111 174 586
Maryjane Bastille	111 268 504
Isabelle Legendre	536 768 666
Henri Lebel	111 286 185
Félix-Antoine Paris	536 776 223

Travail présenté à
Monsieur
Olivier Côté

13 mars 2023

UNIVERSITÉ LAVAL

Faculté des sciences et de génie
École d'actuariat

Table des Matières

Introduction

Premier taitement des variables

Sélection des variables

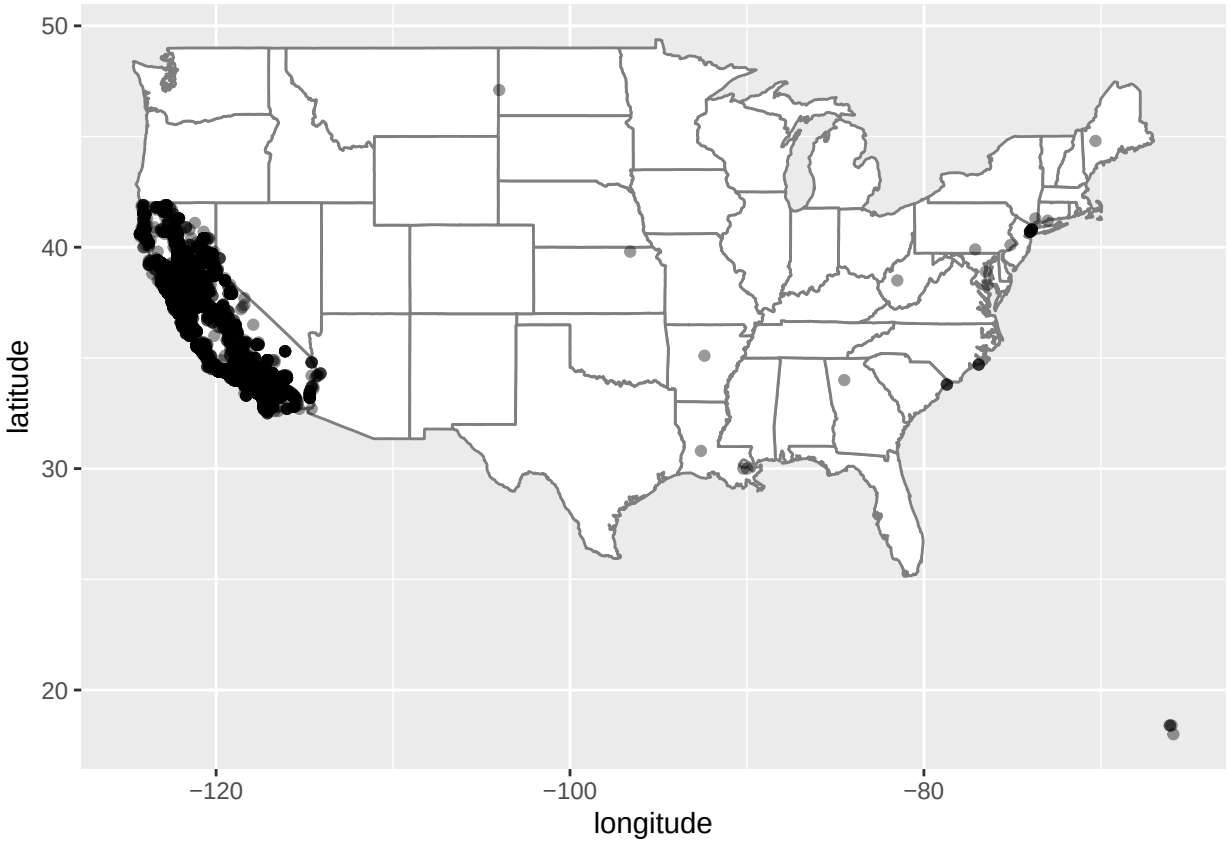
La première étape du travail a consisté à réduire la dimension du jeu de données. En effet, celui-ci est constitué de 41 variables, dont une bonne partie n'étant pas utiles dans le contecte de l'analyse des montants de réclamation.

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.

Sélection des observations

Les sujets d'intérêts sont, pour se remémoriser, les polices d'assurances couvrant le risque d'inondation pour des bâtiments de l'État de la Californie. Il est donc important de s'assurer que les données proviennent uniquement de la Californie. On peut effectuer cette sélection grâce aux données de coordonnées (latitude et logitude) ainsi qu'avec les code de comtés disponibles dans le jeu de données.

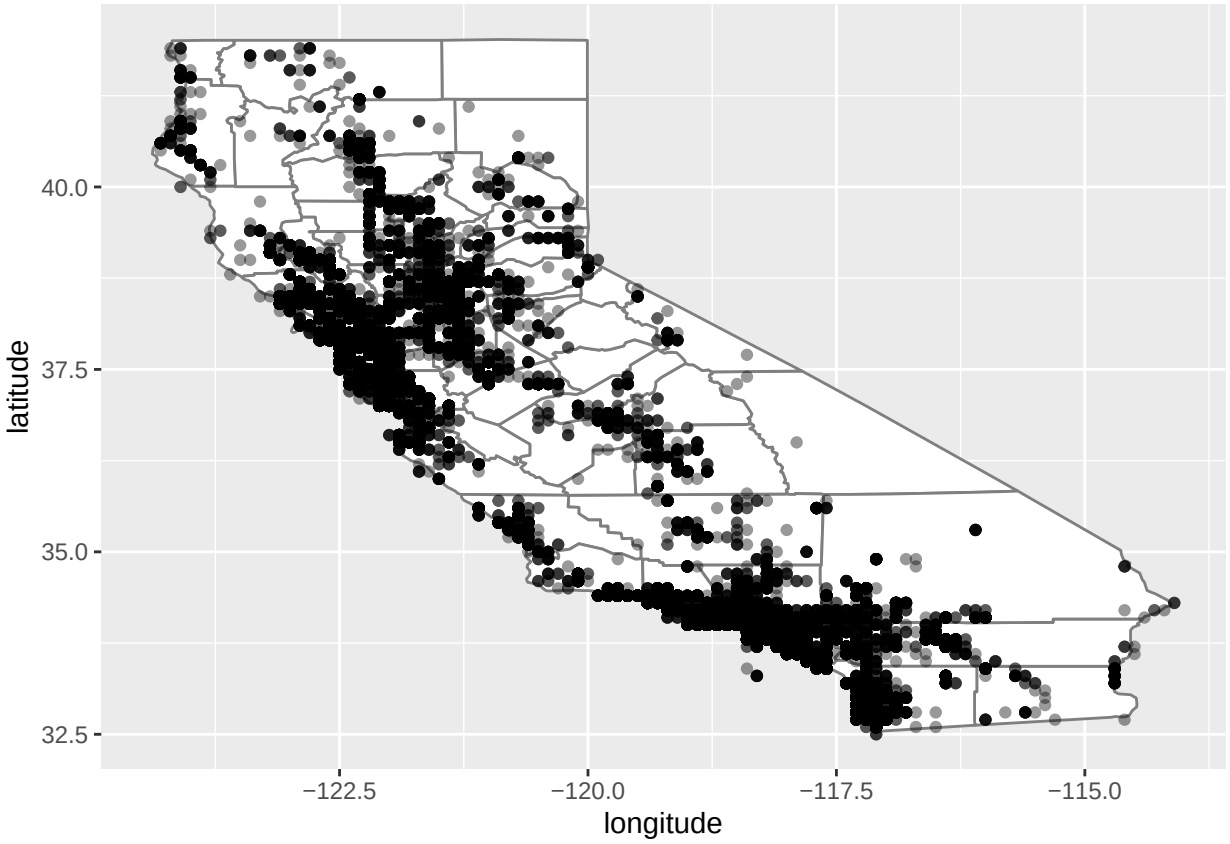


Grâce à cette carte, on peut facilement voir que certaines données ne sont visiblement pas situées en Californie. Ces observations sont retirées du jeu de données.

Ensuite, en observant les codes de comtés des observations restantes, on peut réaliser que trois observations arborent le code “32031”, qui appartient au comté de Washoe au Nevada. Ces trois données sont donc retirées du jeu de données.

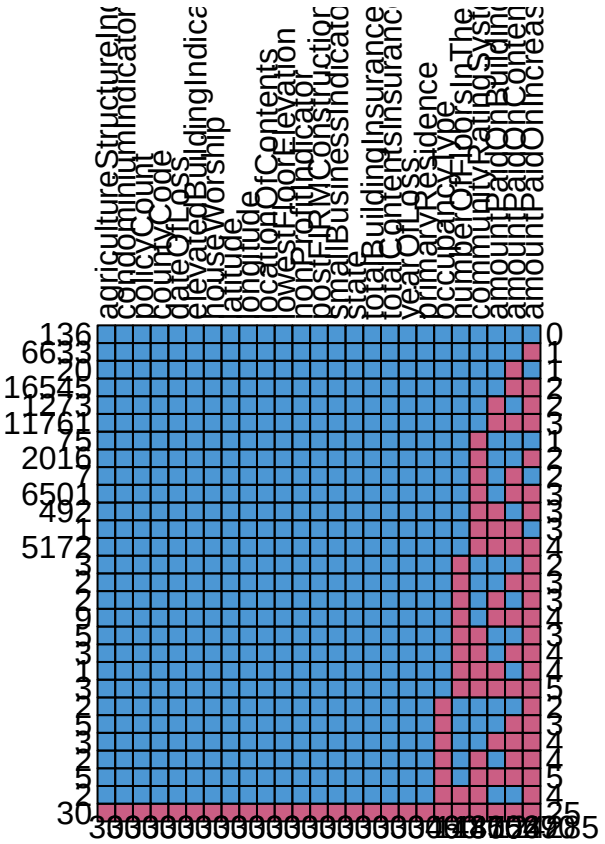
Voici à quoi ressemble la distribution des données restantes sur la carte de la Californie.

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



Imputation des données manquantes

Le jeu de données comporte plusieurs données manquantes réparties dans multiples variables explicatives. Explorons le patron de non réponse.



Attaquons la variable communityRatingDiscount en premier. En effet, celle-ci indique le niveau auquel la police d'assurance a droit à un rabais sur sa prime en fonction de de la zone d'inondation dans laquelle le bâtiment se retouve. La cote est sur une échelle de 1 à 10, du plus gros rabais pour la classe 1 à l'absence de rabais. La façon la plus intuitive que nous avons pu trouver de gérer les données manquante est de leur attribuer arbitrairement la classe 10, puisque selon l'organisme qui publie ces données, les données manquantes ne participent tout simplement pas au programme de primes.

Ensuite, la variable du nombre d'étages du bâtiment (variable numberOfFloorsInTheInsuredBuilding) arborait un certain nombre de données manquantes pouvant être imputées. Nous avons commencé par reclassifier cette la du type d'occupartion du bâtiment (occupancyType) en trois classes plus intuitives selon les descriptions des 14 classes offertes par le publicateur des données. Une de ces 14 classes n'avait aucune description, nous l'avons attribué à la classe ayant la proportion du nombre d'étages la plus semblable, qui est la classe 2.

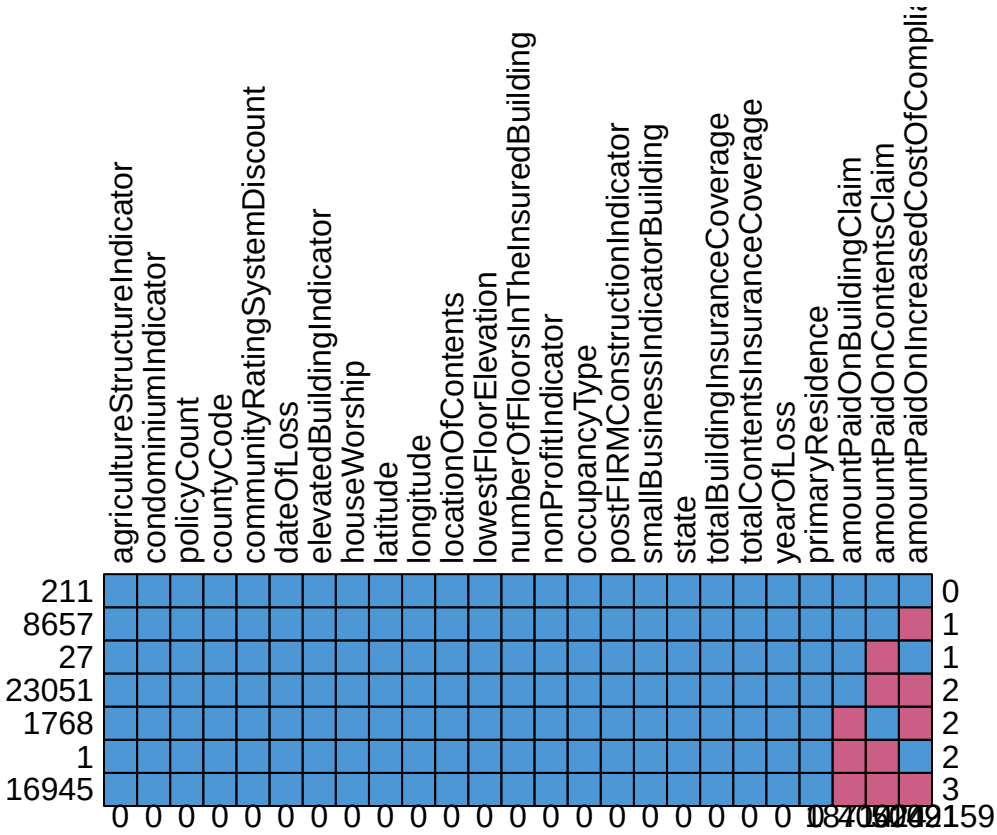
Niveau 1: Résidences familiales Niveau 2: Copropriétés résidentielles Niveau 3: Non-résidentiel

En effet, nous avons identifié la variable du type d'occupation intuitivement comme une variable intimement corrélée avec le nombre d'étages du bâtiment, car elle donnc des indices sur la nature du bâtiment. Elle servira donc à l'imputation des données du nombre d'étages.

Nous avons donc imputé les données à l'aide d'un modèle accordant le nombre d'étages aléatoirement en fonction de la distribution du nombre d'étages à l'intérieur d'une même classe d'occupation du bâtiment.

La dernière variable à imputer est l'indicateur du bâtiment étant un condo ou non. On utilise tout simplement notre variable du type d'occupation du bâtiment; si l'observation est dans la catégorie 2, on lui impute une valeur de "1" et une valeur de "0" dans le cas échéant.

Observons maintenant à nouveau notre patron de non réponse.



On remarque que les données manquantes restantes se retrouvent dans les trois variables contenant les montants de réclamation. Nous gèrerons ceci dans la prochaine section.

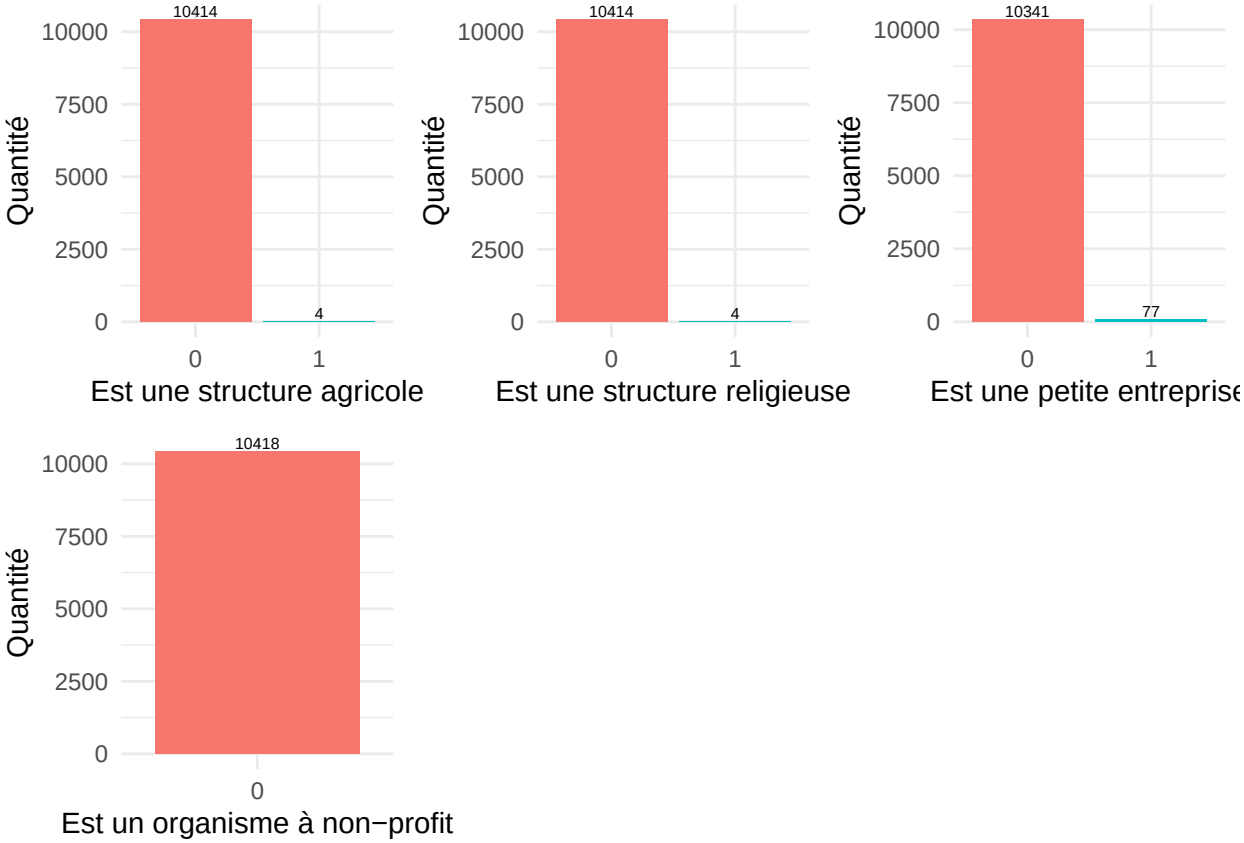
Création de la nouvelle variable réponse

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).

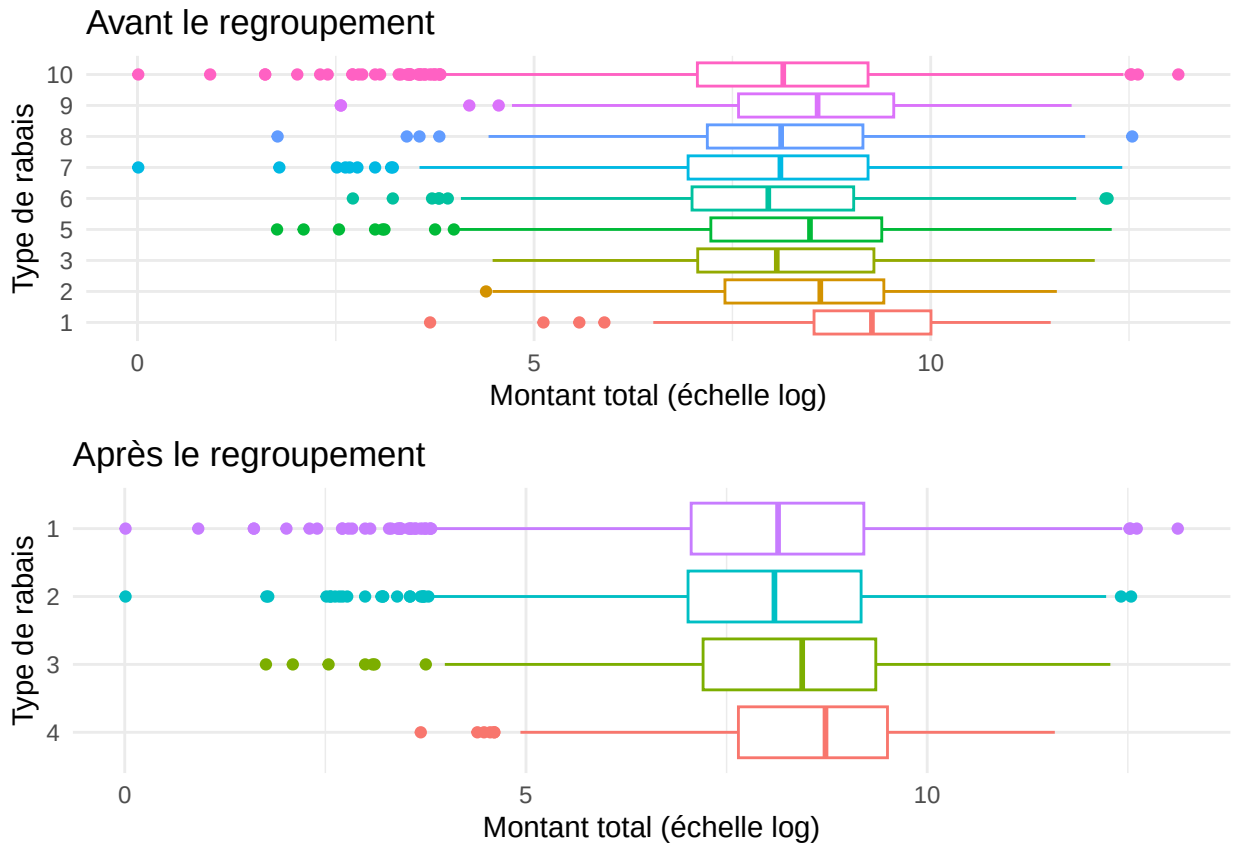
On suppose dans ce cas que les données manquantes peuvent tout simplement se faire attribuer la valeur de 0, indiquant l'absence de paiement dans cette catégorie. Ensuite, nous combinons ces trois variables en créant une nouvelle variable du paiement de prestation total versé au détenteur de police. Celle-ci sera la variable réponse du modèle.

Analyse exploratoire des données

Transformation des variables



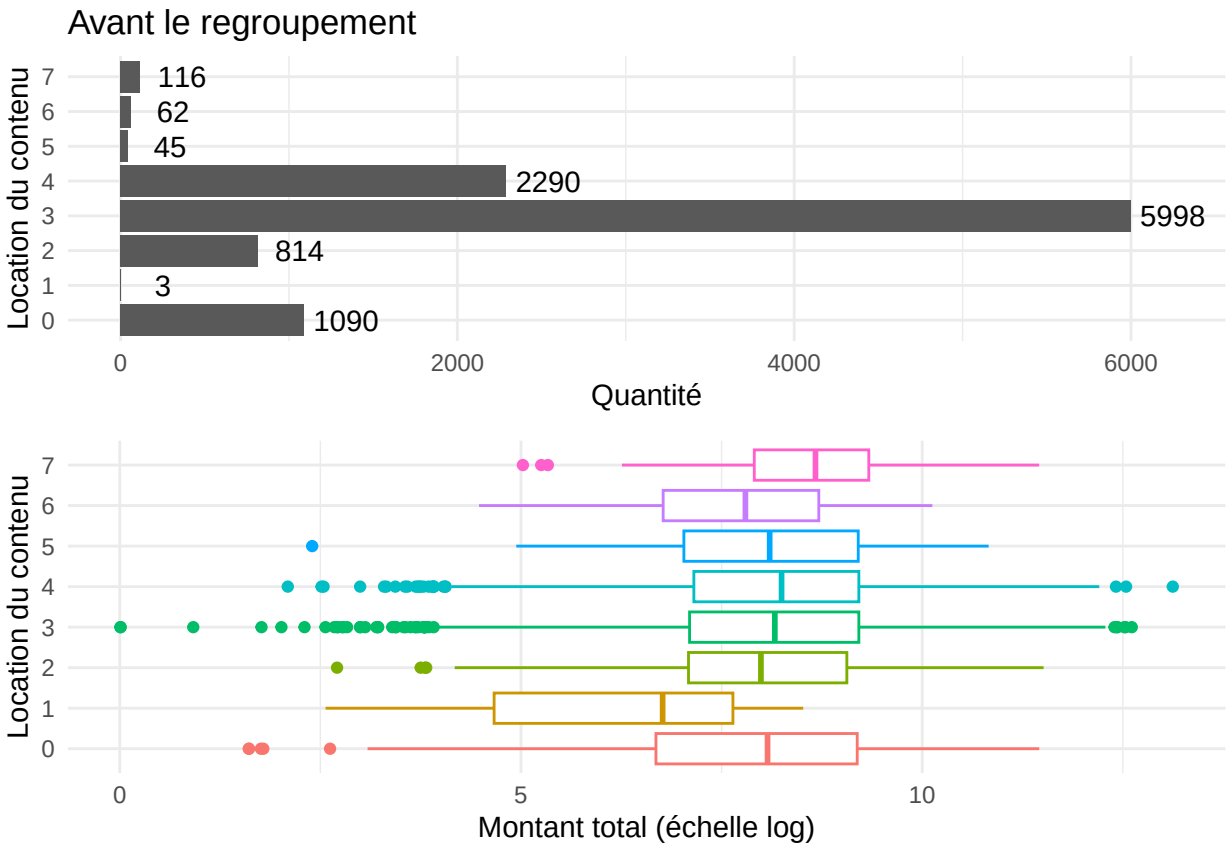
Tel qu'on peut le voir dans les graphiques précédants, les quatres variables indicatices (smallBusinessIndicatorBuilding, agricultureStructureIndicator, houseWorship et nonProfitIndicator) sont trop peu fréquentes pour être significatives, elles seront donc élminées.



Pour la variable communityRatingSystemDiscount, il y a un trop grand nombre de catégories. Pour cette raison, elles sont réeunies en quatres catégories de crédit:

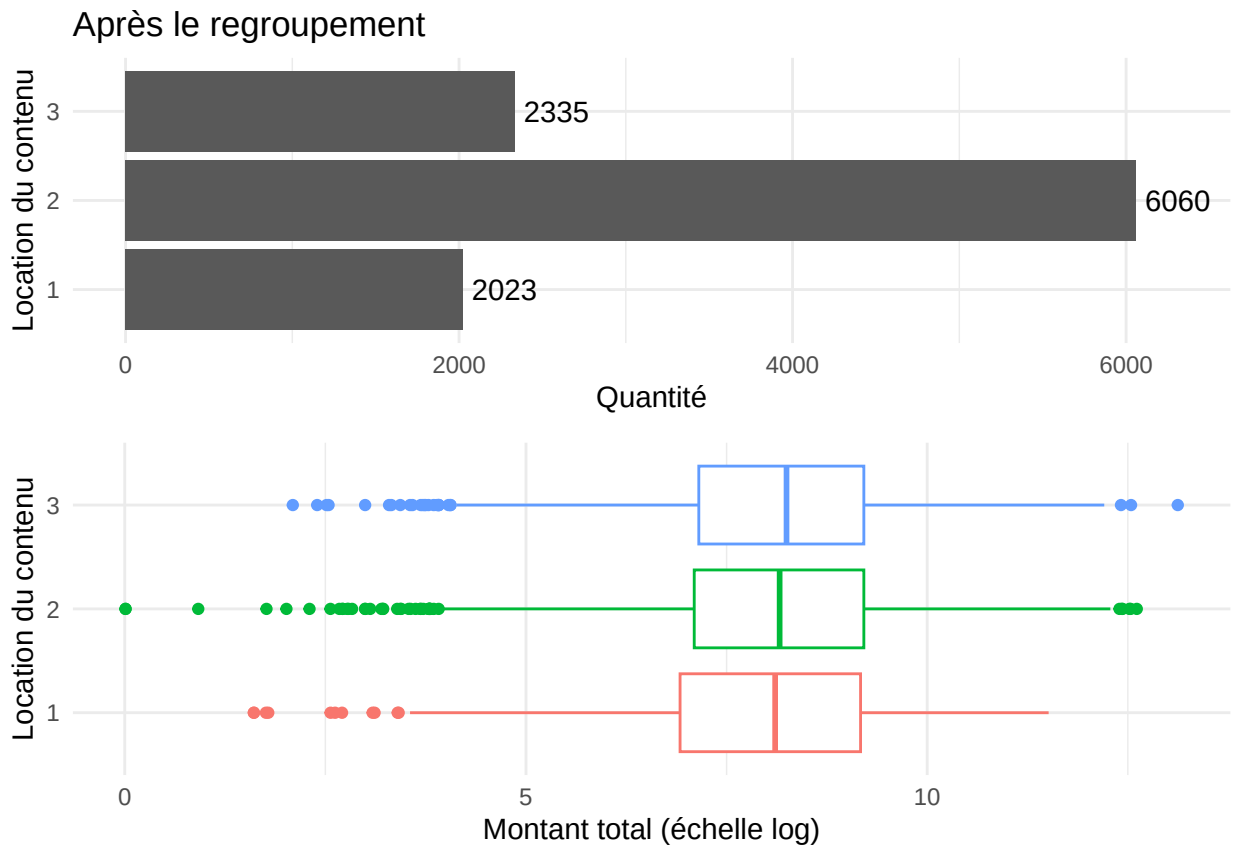
- Niveau 1 : Aucun
- Niveau 2 : 20% - 5%
- Niveau 3 : 35% - 25%
- Niveau 4 : Plus de 40%

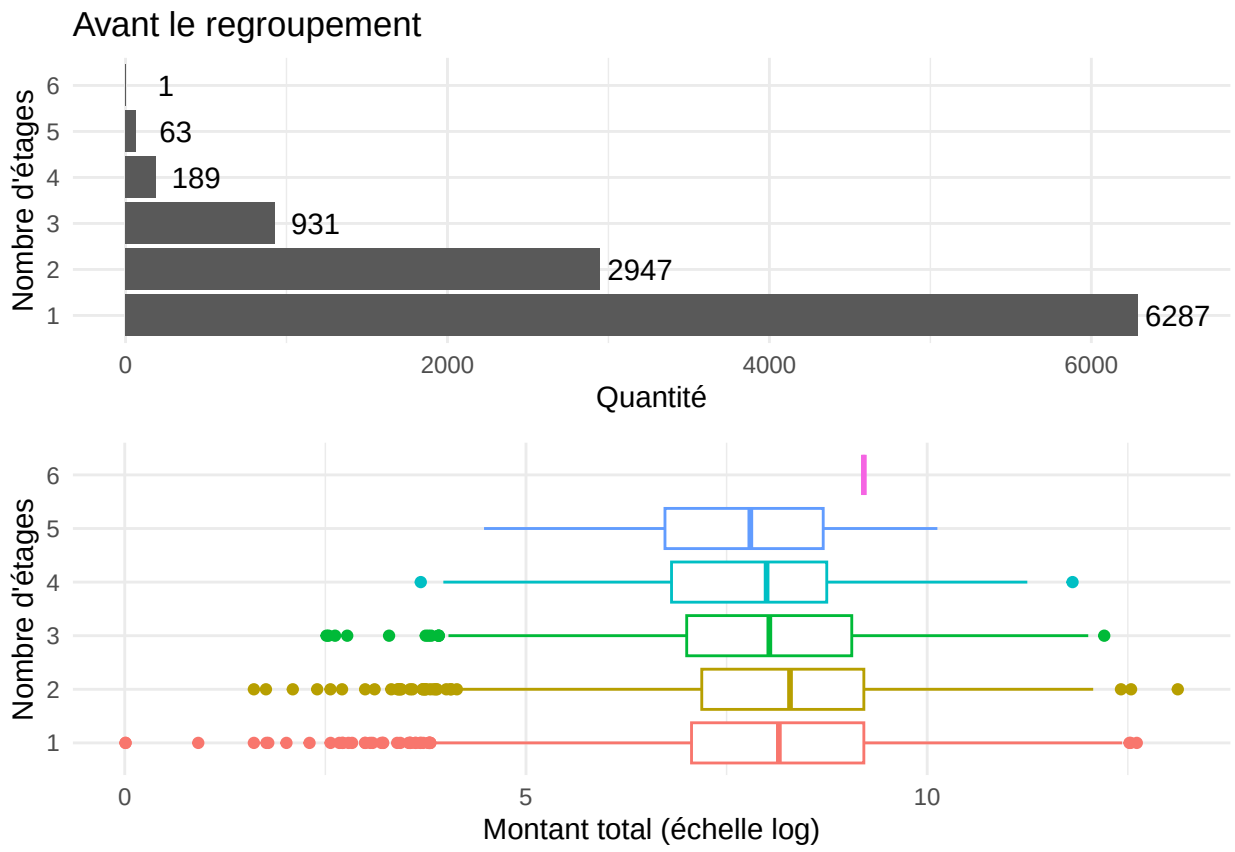
Pour faciliter la visualisation une transformation log est effectuée sur le montant total, cette échelle sera utilisée pour de futurs graphiques.

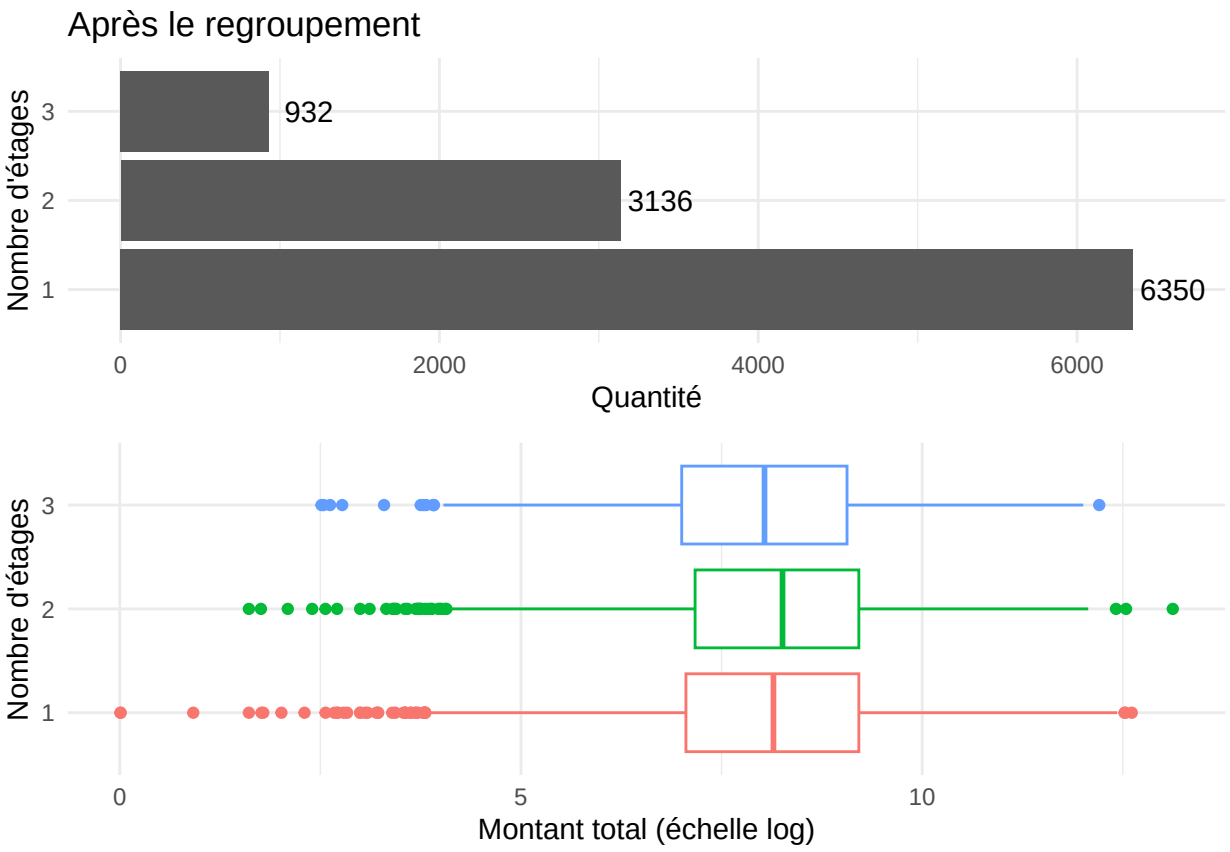


On remarque pour la variable locationOfContents, qu'il y a plusieurs catégories ne comportant qu'un faible nombre d'observations. Certaines seront donc combinées pour former les nouvelles catégories indiquant la location des objets endomagés dans le bâtiment assuré:

- Niveau 1 : Sous-sol
- Niveau 2 : Premier étage seulement
- Niveau 3 : Premier étage et plus


Après le regroupement
Location du contenu
3
2
1
2335
6060
2023
0
2000
4000
6000
Quantité
Location du contenu
3
2
1
0
5
10
Montant total (échelle log)

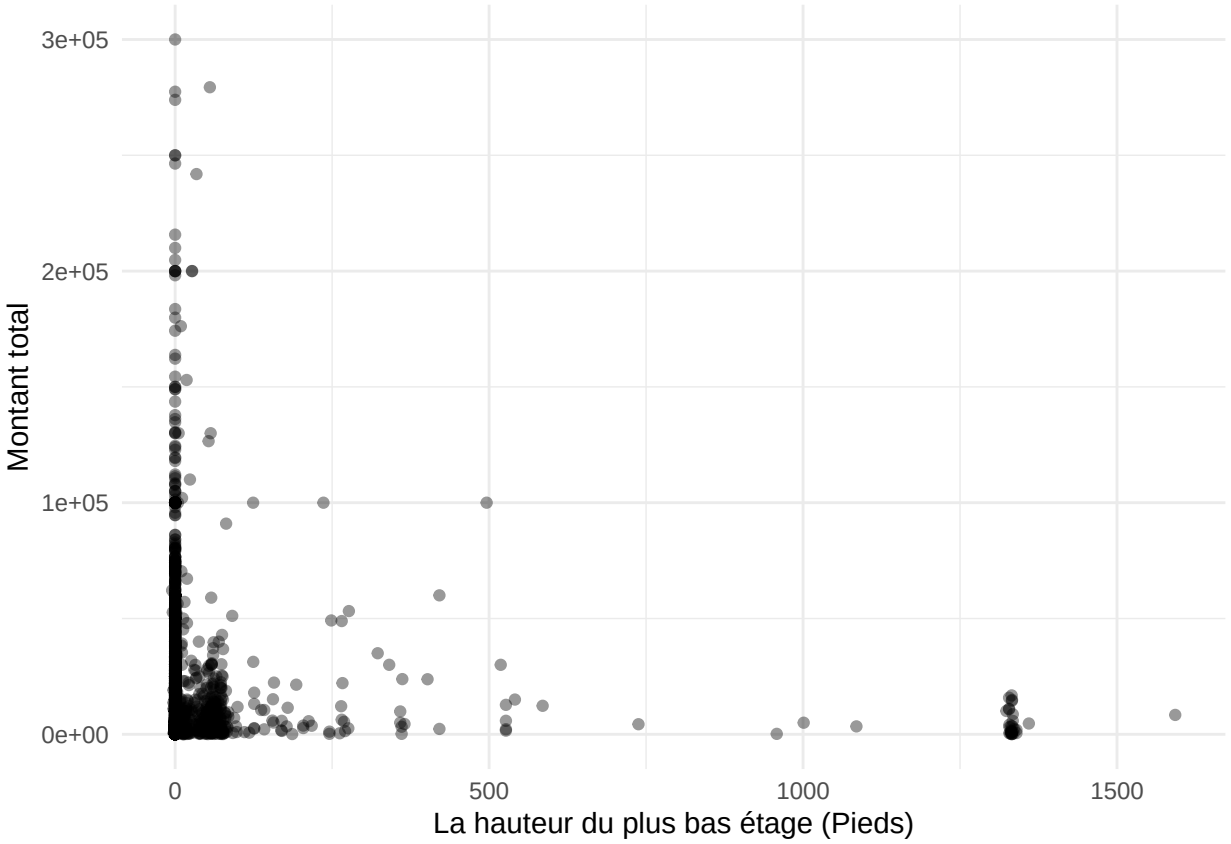

Avant le regroupement
Nombre d'étages
6
5
4
3
2
1
1
63
189
931
2947
6287
0
2000
4000
6000
Quantité
Nombre d'étages
6
5
4
3
2
1
0
5
10
Montant total (échelle log)



- Niveau 1 : 1 étage
- Niveau 2 : 2 étages
- Niveau 3 : 3 étages et plus

Explication des variables

- `condominiumIndicator` : Variable indicatrice, 1 si l'habitation est un condominium, 0 sinon.
- `policycount` : Le nombre de polices actives pour l'assuré.
- `county code` : Représente le code du comté.
- `communityRatingSystemDiscount` : Variable catégorielle indiquant le pourcentage de rabais accordé lors de la tarrification. (Plus de 40%, 35% - 25%, 20% - 5%, Aucun)
- `dateOfLoss` : La date où s'est produit l'infiltration d'eau dans le bâtiment.
- `elevatedBuildingIndicator` : Variable indicatrice, 1 si le bâtiment est élevé, c'est-à-dire, au dessus du nieveau ddu sol, 0 sinon.
- `latitude` et `longitude` : Position géoographique du bâtiment assuré, à une décimale près.
- `locationOfContents` : Variable catégoritielle indiquant où se trouve le contenu du bâtiment qui a été endomagé.
- `lowersFlorElevation` : La hauteur du plus bas étage de l'habitation, en pieds.



- `numberOfFloorsInTheInsuredBuilding` : Variable catécorielle de trois niveaux indiquant le nombre d'étages de la propriété. (1 étage, 2 étages, 3 étages ou plus)
- `occupancyType`

Conclusion

Bibliographie

]

Annexe