Premier rapport

Apprentissage statistique en actuariat
ACT-4114

Équipe 09

---

**Rapport**

# Inondations en Californie

---


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

Le jeu de données utilisé est une base de données de réclamations d'assurance faites, par contrat, à la suite d'inondations aux États-Unis par FEMA. Étant donné que le jeu de données est trop volumineux, l'analyse se fera seulement sur l'état de la Californie. La variable réponse est le montant total payé par réclamation en dollar USD (totalAmount). Cette variable est obtenue en additionnant le montant payé sur la réclamation du bâtiment (amountPaidOnBuildingClaim), le montant payé sur la réclamation des biens (amountPaidOnContentsCaim) et le montant payé sur l'augmentation des coûts de la conformité (amountPaidOnIncreasedCostOfComplianceClaim).

# Sélection des variables

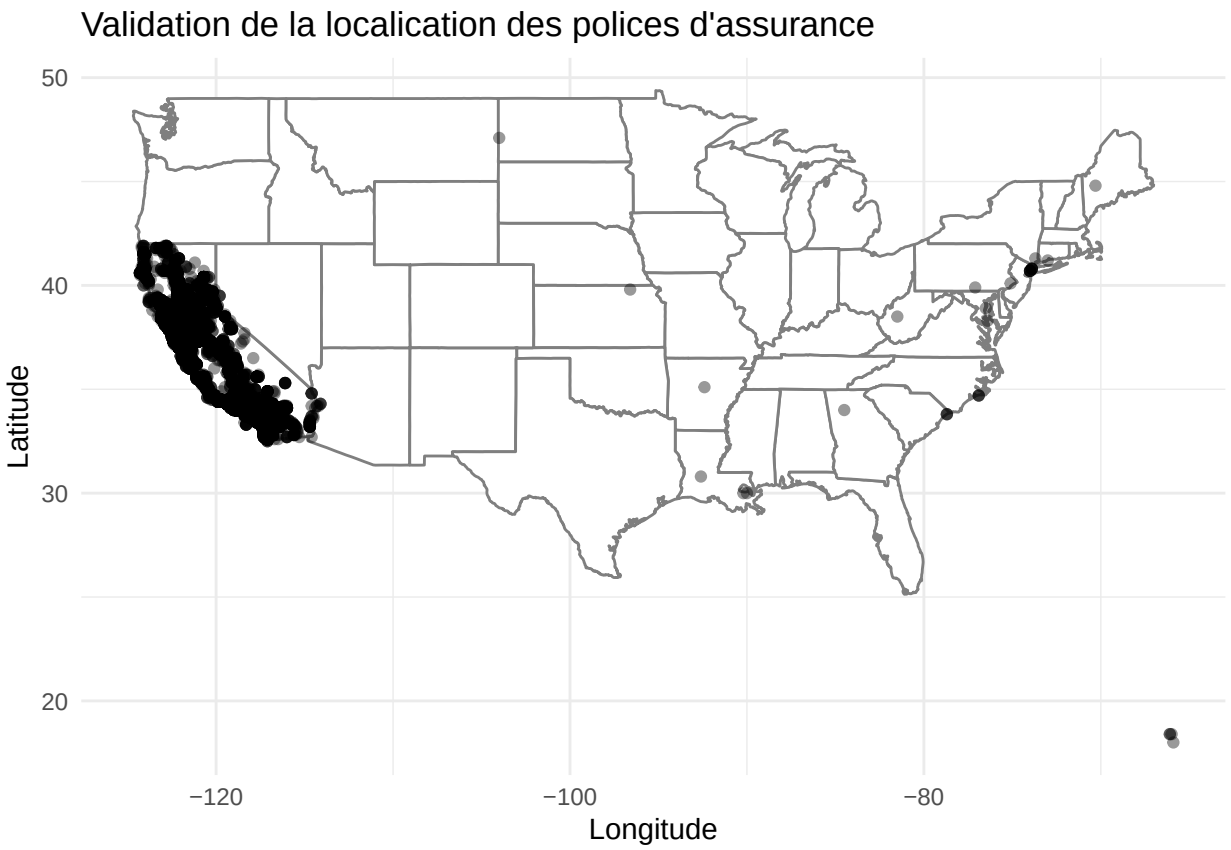
La première étape du travail a consisté à réduire la dimension du jeu de données. En effet, celui-ci est constitué de 41 variables, dont une bonne partie n'étant pas utiles dans le contecte de l'analyse des montants de réclamation.

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

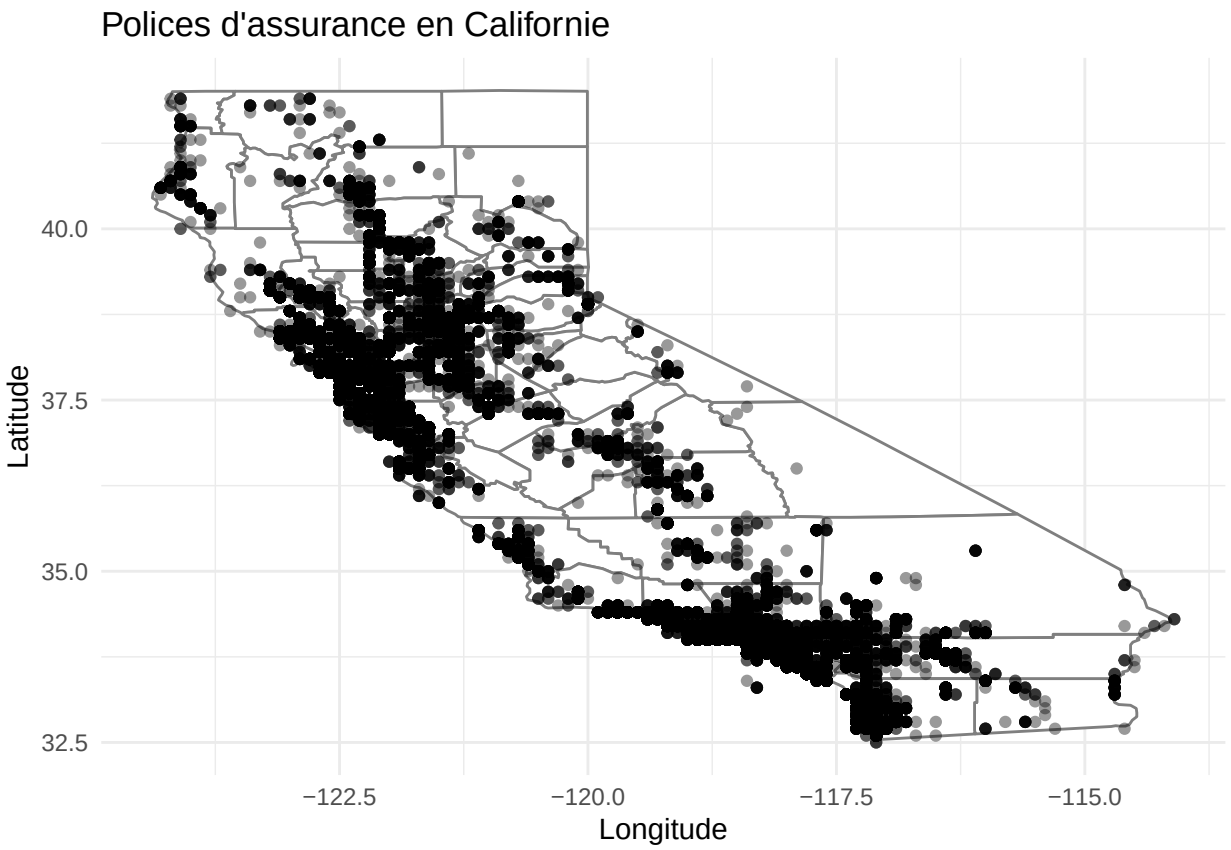
Les sujets d'intérêts sont, pour se remémoriser, les polices d'assurances couvrant le risque d'inondation pour des bâtiments de l'État de la Californie. Il est donc important de s'assurer que les données proviennent uniquement de la Californie. On peut effectuer cette sélection grâce aux données de coordonnées (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.

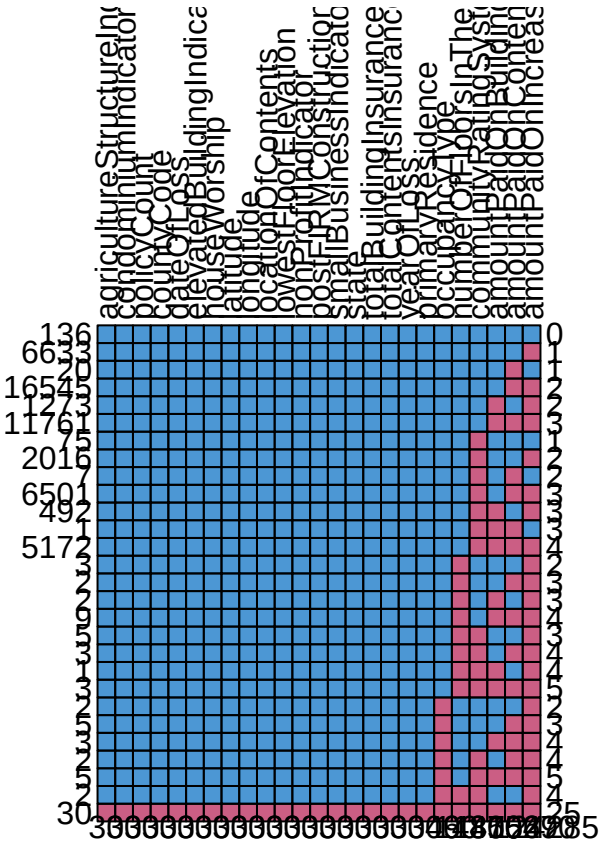
Ensuite, en observant les codes de comtés des observations restantes, on peut réaliser que trois observations arborent le code “32031”, qui appartient au comté de Washoe au Nevada. Ces trois données sont donc retirées du jeu de données.

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




## 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 celle 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.

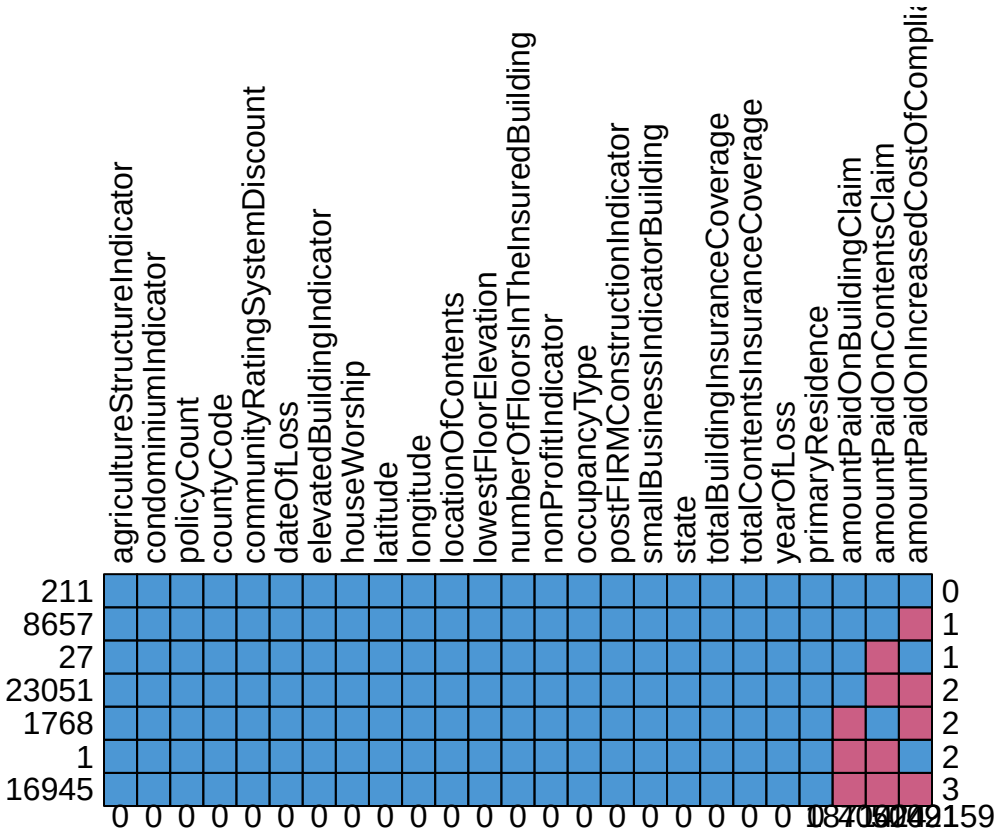
- Niveau 1: Résidences familiales
- Niveau 2: Copropriétés résidentielles
- Niveau 3: Non-résidentiel

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 1, on lui impute une valeur de "1" et une valeur de "0" dans le cas contraire.

Observons 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

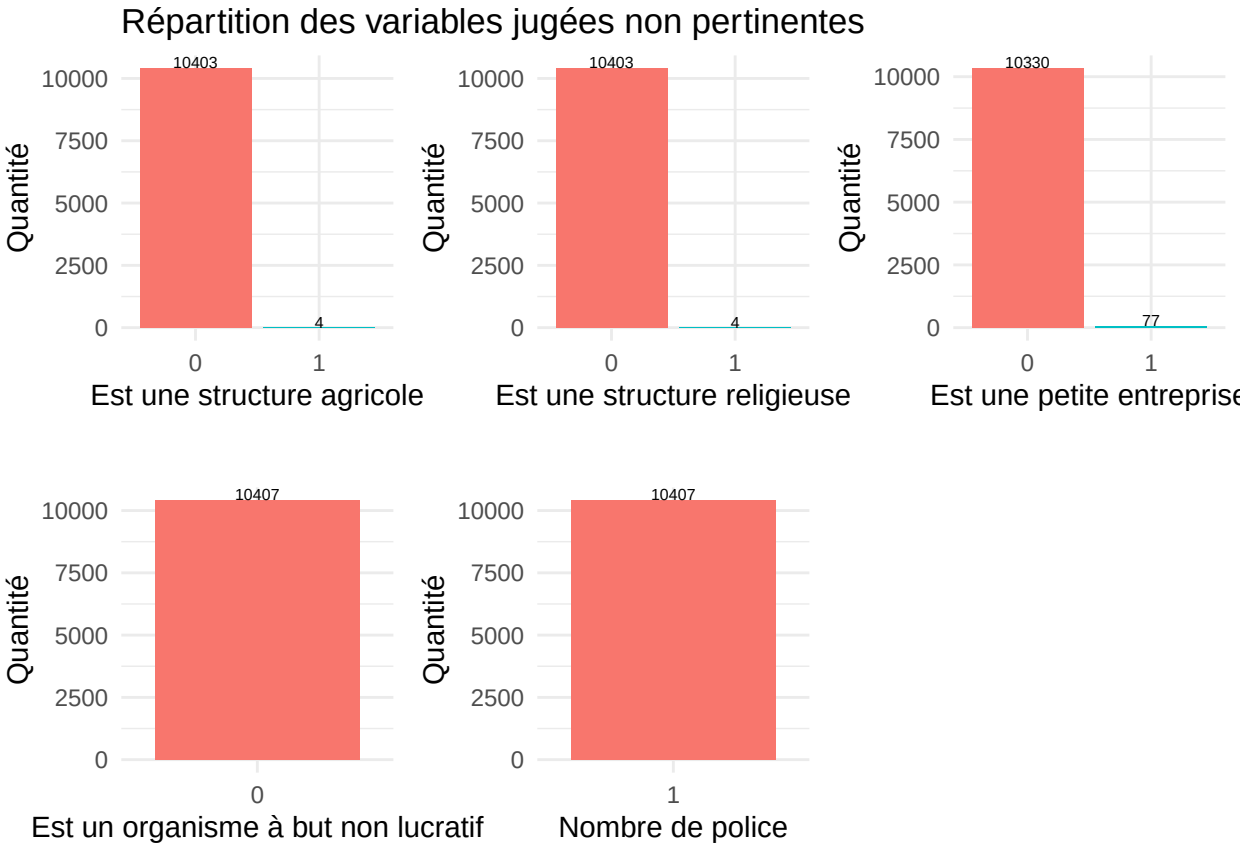
Dans le jeu de données se retrouvent trois colonnes contenant des informations sur les montants de prestations payés en lien avec le bâtiment (amountPaidOnBuildingClaim), les biens (amountPaidOnContentsClaim) et l'augmentation des coûts en lien avec la conformité (amountPaidOnIncreasedCostOfComplianceClaim).

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

Puisque nous voulons étudier les réclamations causés par un sinistre, nous allons retenir les montants supérieurs à 10$. Ce seuil a été choisi car, logiquement les réclamations en dessous de ce montant devraient être quasi inexistantes.

## Variables non pertinentes




Tel qu'on peut le voir dans les graphiques précédants, les cinq variables indicatices (smallBusinessIndicator-Building, agricultureStructureIndicator, houseWorship, policyCount et nonProfitIndicator) sont trop peu fréquentes pour être significatives, ce qui indique que les données conservées sont quasi homogènes par rapport à ces variables. Nous pouvons donc les retirer.

### Variable *condominiumIndicator*

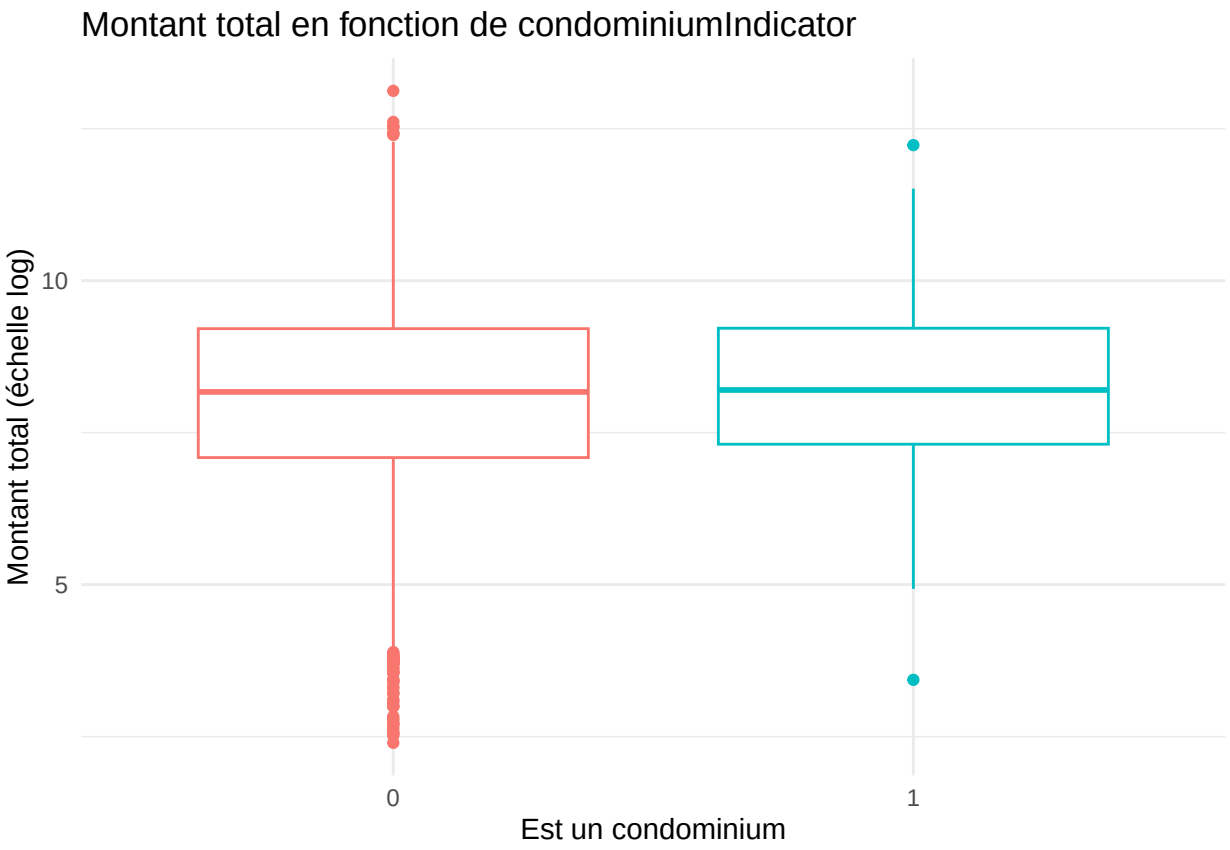



Table 1: Statistiques sur totalAmount selon condominiumIndicator

| | Min | Max | Médiane | Moyenne | Écart-type |
|---|---|---|---|---|---|
| Condominium | 31 | 204800 | 3647.65 | 9993.642 | 21356.01 |
| Autre | 11 | 500000 | 3532.50 | 9029.393 | 18282.19 |

Cette variable est indicatrice, elle prendre une valeur de 1 si l'habitation est un condominium, 0 sinon. Pour faciliter la visualisation, une transformation logarithmique est effectuée sur le montant total, cette échelle sera aussi utilisée pour de futurs graphiques. Comme on peut le voir avec le tableau et le graphique, lorsqu'il y a une réclamation dans un condominium elles tend à être en moyenne plus élevée.

## Variable *communityRatingSystemDiscount*

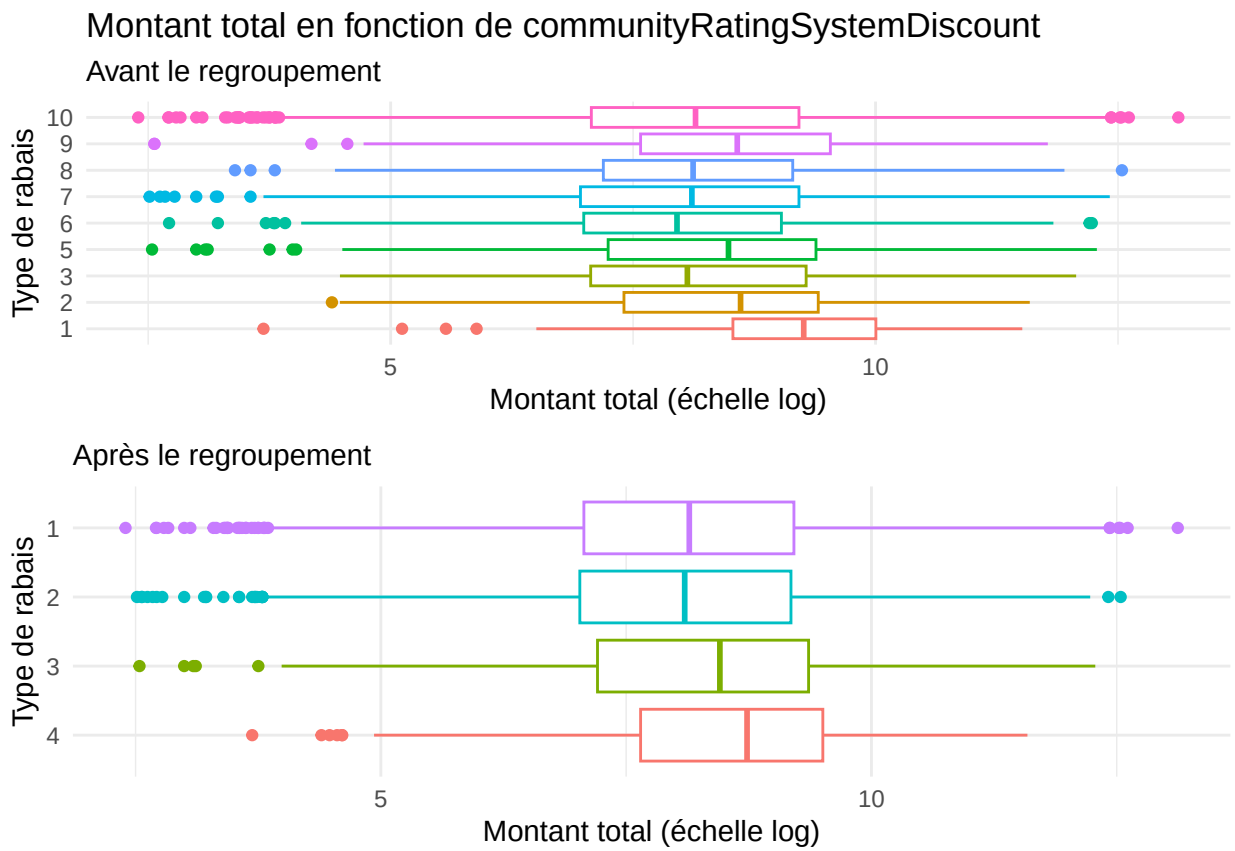



Table 2: Statistiques sur totalAmount selon communityRatingSystemDiscount

| | Min | Max | Médiane | Moyenne | Écart-type |
|---|---|---|---|---|---|
| Aucun | 11.00 | 500000.0 | 3441.470 | 8827.199 | 19055.38 |
| 5% - 20% | 12.36 | 279376.8 | 3284.715 | 8940.088 | 17691.94 |
| 25% - 35% | 12.68 | 215732.9 | 4700.230 | 9910.646 | 18843.42 |
| 40% et plus | 40.00 | 108006.3 | 6201.000 | 10817.843 | 13708.86 |

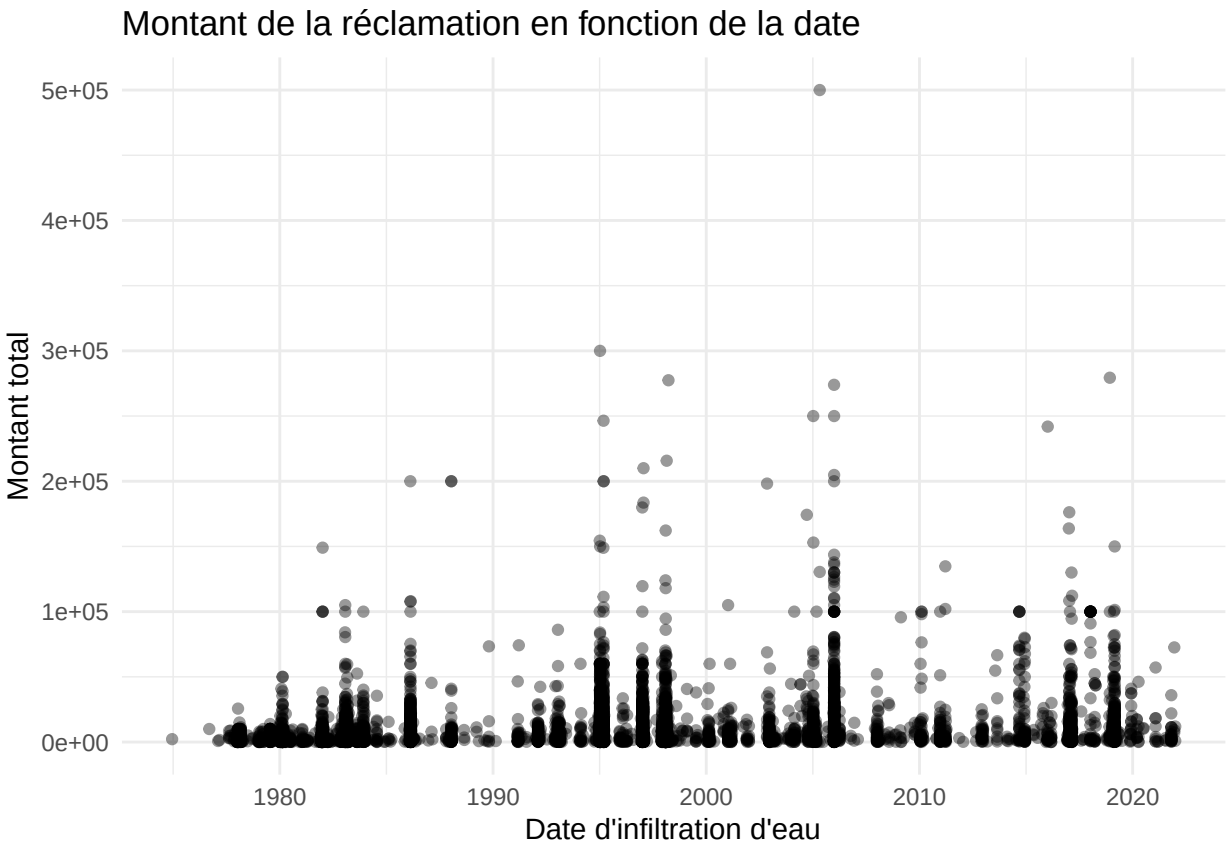
Cette variable représente le remboursement par le gouvernement pour les individus éligibles, donc à faible revenu et situés dans une zone à risque. Par conséquent un grand remboursement indique un plus grand risque d'inondation.

On remarque qu'il y a un trop grand nombre de catégories. Pour cette raison, elles sont réunies en quatres niveaux:

- Niveau 1 : Aucun
- Niveau 2 : 5% - 20%
- Niveau 3 : 25% - 35%
- Niveau 4 : 40% et plus

Comme le montre le graphique et le tableau plus le remboursement est élevé, plus les réclamations augmentent.

### Variable *dateOfLoss*




Cette variable temporelle, au jours près, indique la date où s'est produit l'infiltration d'eau dans le bâtiment. La période commence en 1968 et se termine en 2021.

### Variable *elevatedBuildingIndicator*

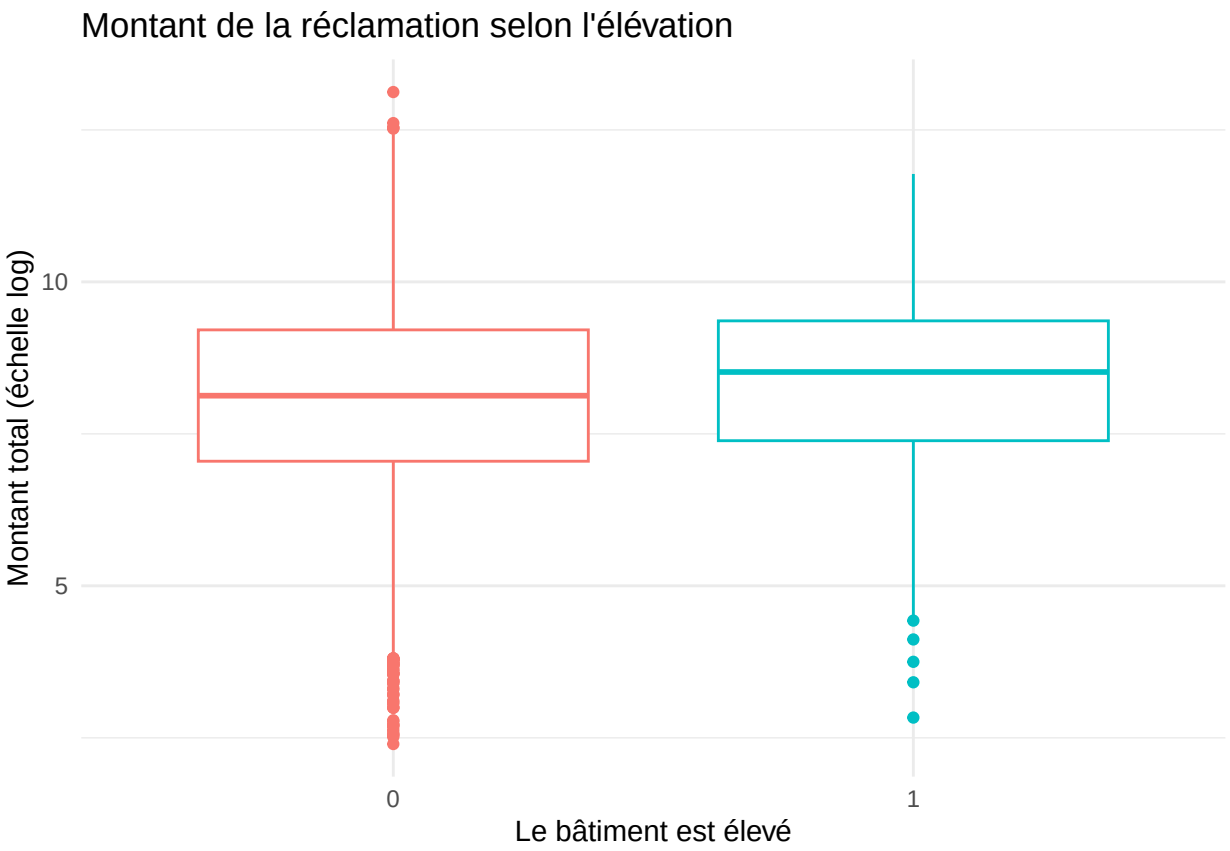


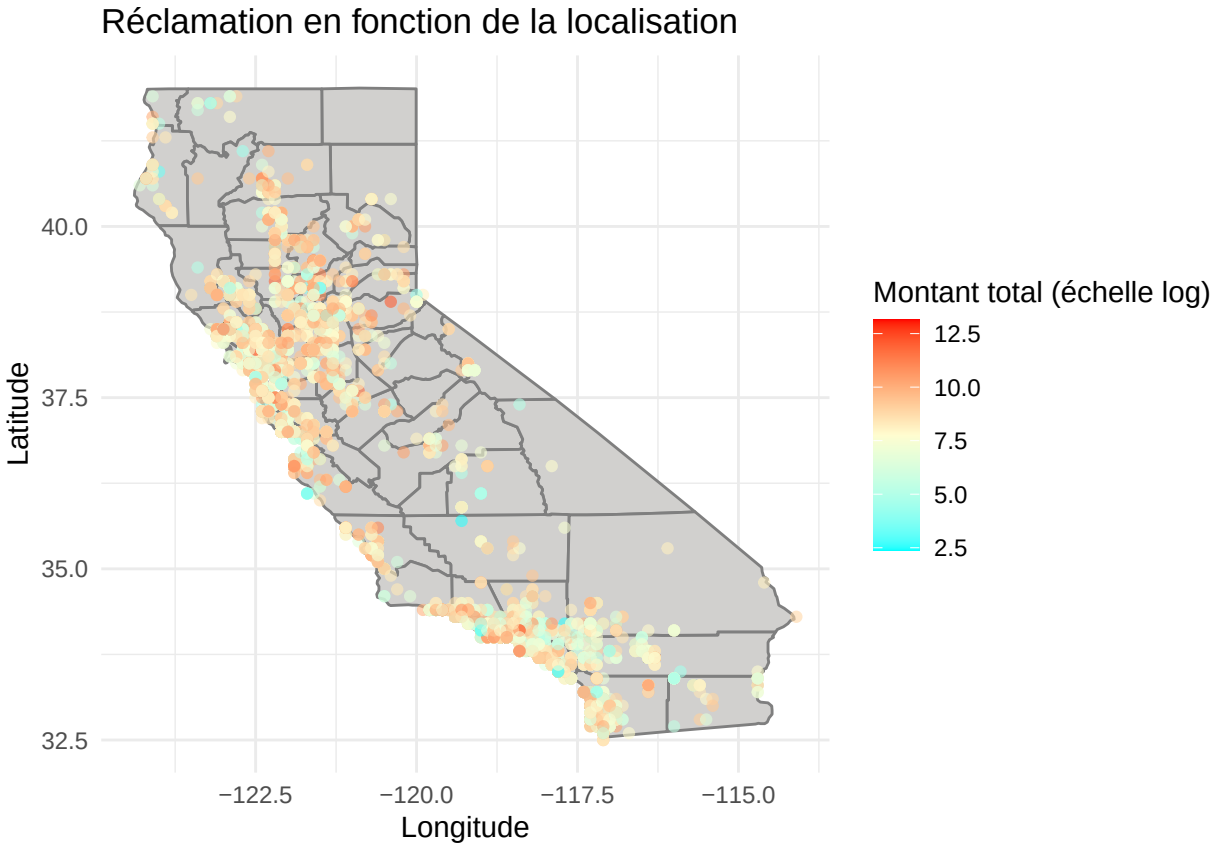

Table 3: Statistiques sur totalAmount selon elevatedBuildingIndicator

| | Min | Max | Médiane | Moyenne | Écart-type |
|---|---|---|---|---|---|
| Le bâtiment est élevé | 17 | 130001 | 5000.000 | 9978.578 | 15519.48 |
| Ne l'est pas | 11 | 500000 | 3387.575 | 8919.537 | 18660.28 |

Cette variable est indicatrice, elle prend une valeur de 1 si le bâtiment est élevé, 0 sinon. Un bâtiment est considéré élevé si le plancher le plus bas est au dessus du niveau du sol.

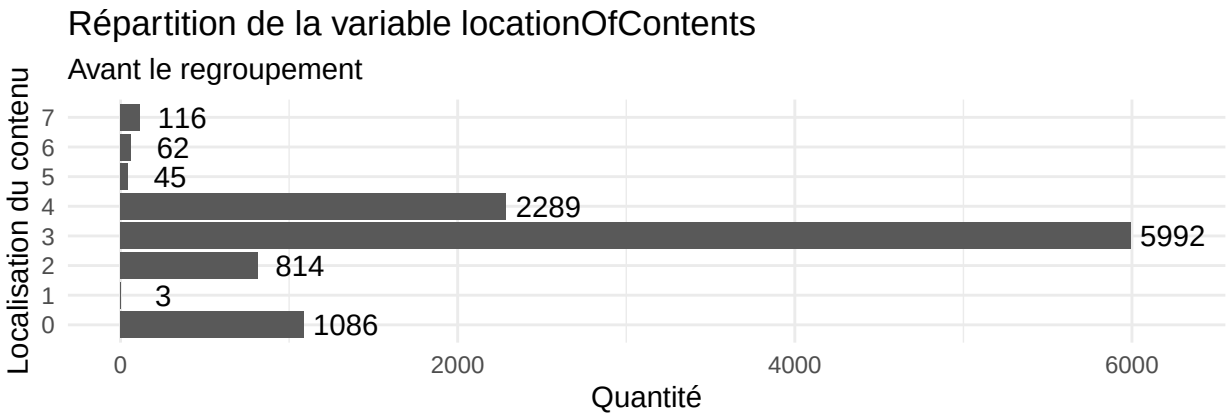
Selon le tableau et le graphique, nous observons des montants plus grands de réclamation pour les bâtiments élevés. En effet, si le le bâtiment est élevé et qu'il y a une réclamation, nous en déduisons que l'inondation était plus catastrophique, donc des dégâts plus importants.

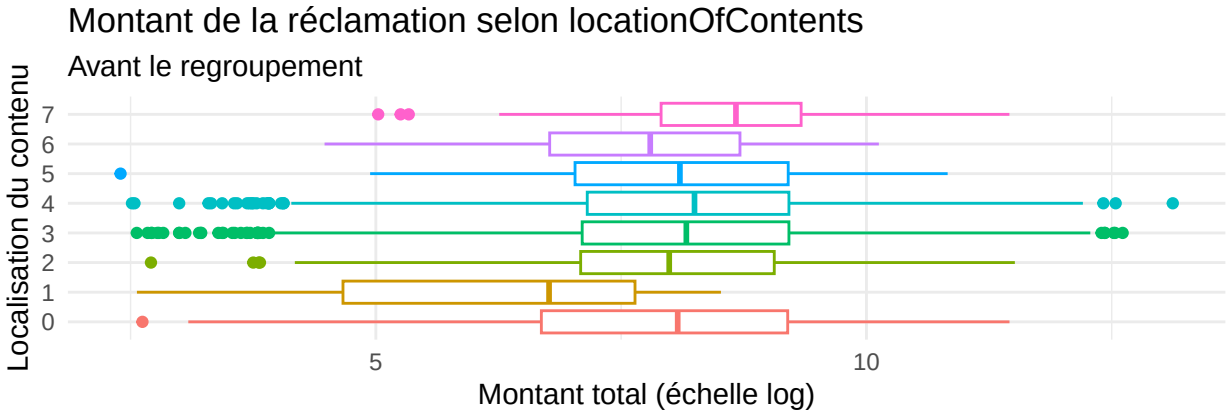
## Variables *latitude* et *longitude*

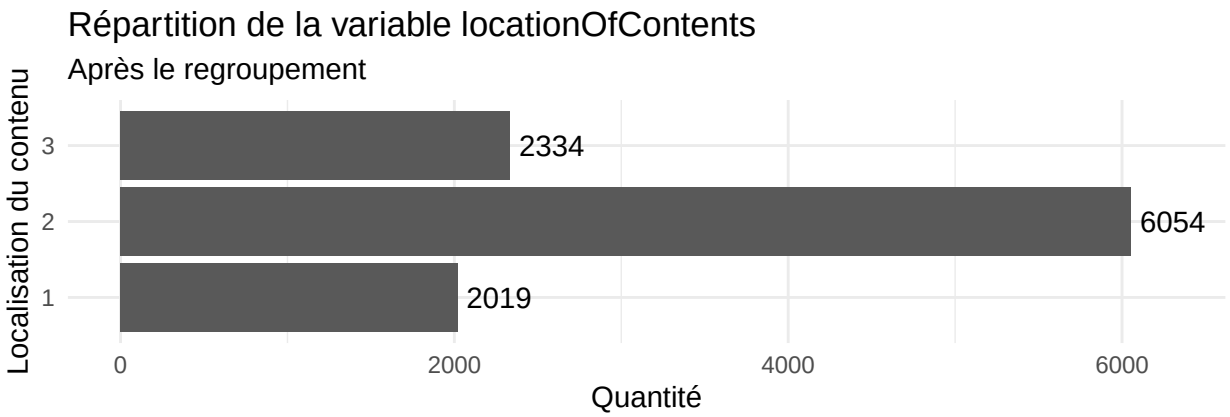



Grâce au graphique précédent, nous observons que la majorité des réclamations sont situés aux alentours des grands centres urbains. Pour en nommer quelques uns: Sacramento, Los Angeles, San Fransisco.

### Variable *locationOfContents*





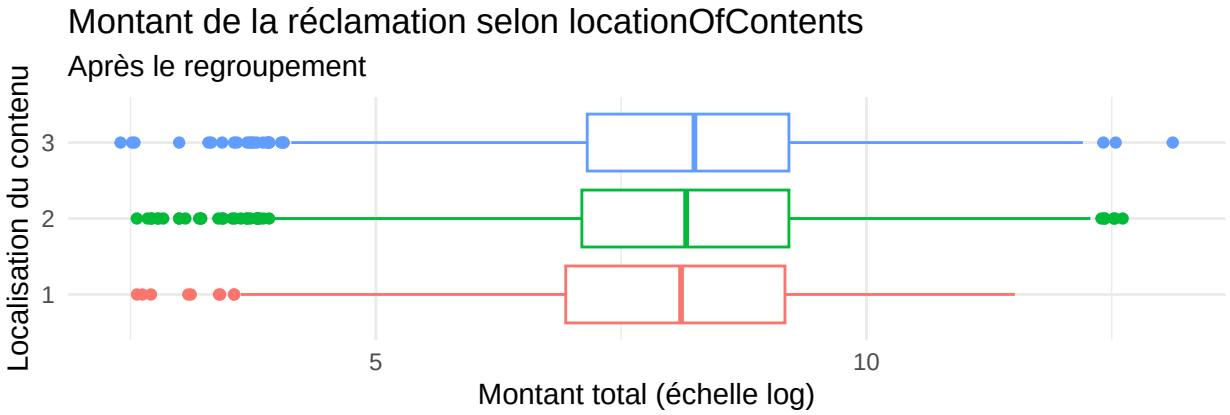




Table 4: Statistiques sur totalAmount selon locationOfContents

| | Min | Max | Médiane | Moyenne | Écart-type |
|---|---|---|---|---|---|
| Sous-sol | 13.00 | 100001 | 3330.22 | 7687.823 | 12424.05 |
| Premier étage seulement | 12.98 | 300000 | 3506.81 | 9458.976 | 19274.74 |
| Premier étage et plus | 11.00 | 500000 | 3816.48 | 9133.476 | 19963.14 |

Cette variable catégoritielle indique où se trouve le contenu endomagé dans le bâtiment assuré.

On remarque pour la variable locationOfContents, qu'il y a plusieurs catégories ne comportant qu'un faible nombre d'observations. Certaines seront donc combinées selon leur nombre d'étage pour former les nouvelles catégories suivantes:

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

Comme nous pouvons l'observer dans le graphique à boîtes à moustache après le regroupement, plus le niveau d'eau monte dans la maison plus les réclamations sont importantes, puisque plus d'objets sont affectés.

### Variable *lowestFloorElevation*

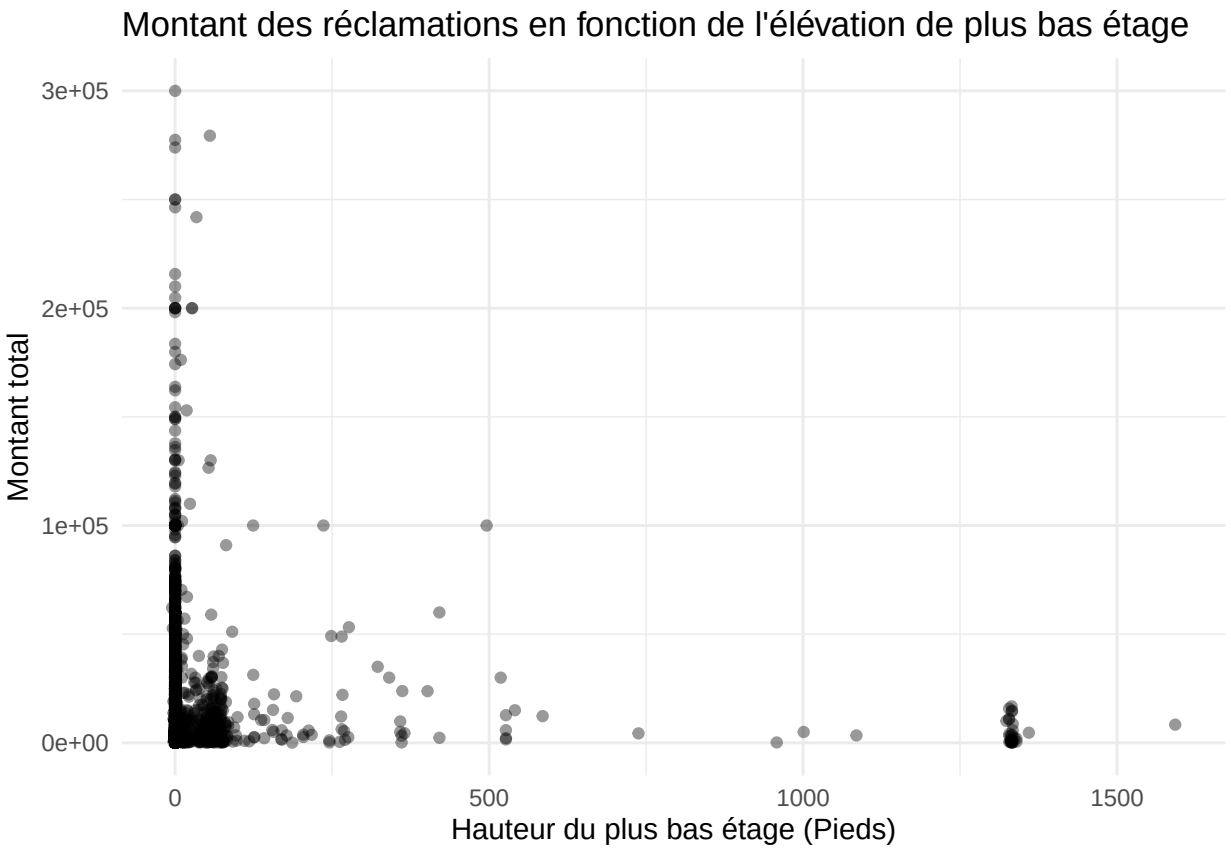
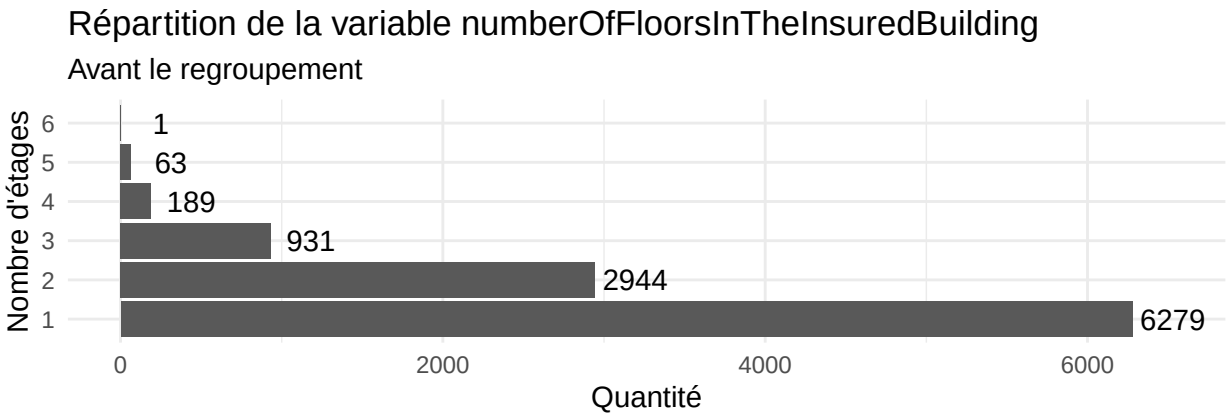


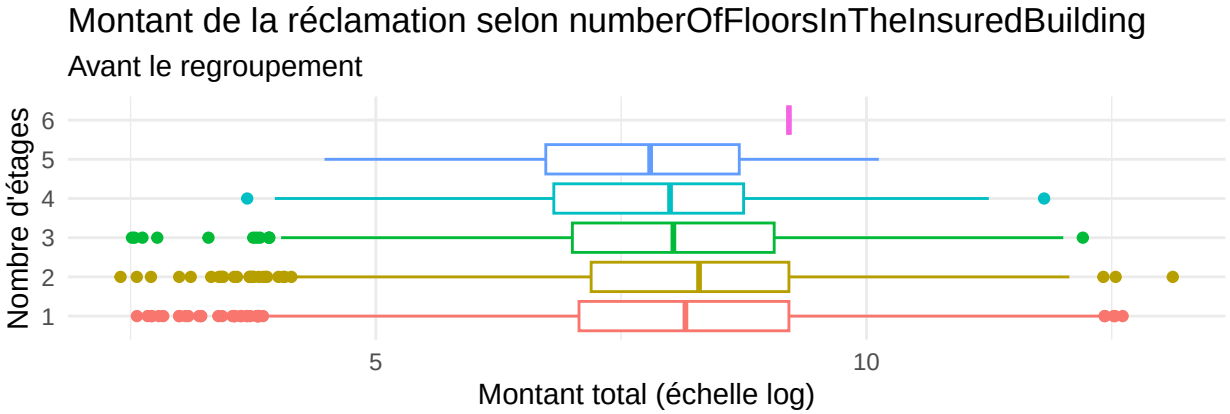

Table 5: Statistiques de lowestFloorElevation

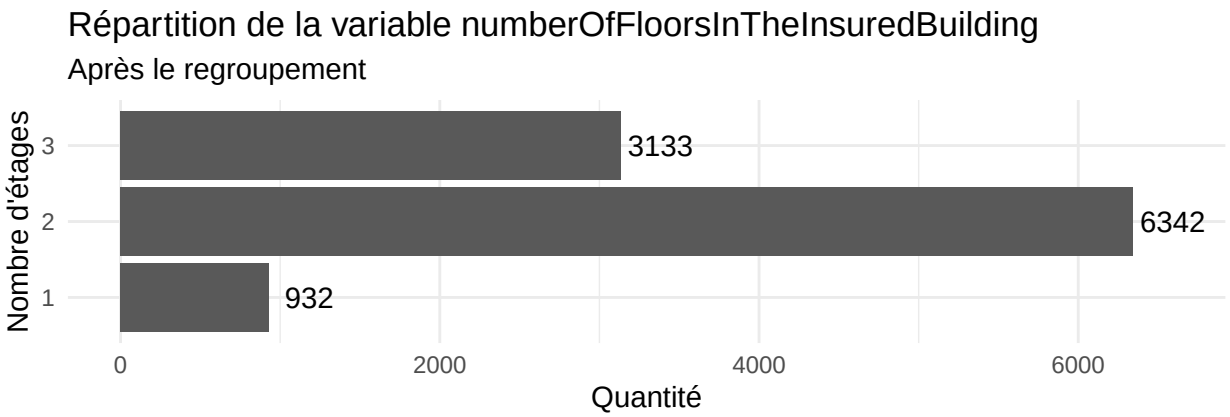
| | Min | Max | Médiane | Moyenne | Écart-type |
|---|---|---|---|---|---|
| Hauteur (en pieds) | -4.8 | 5940.4 | 0 | 8.297175 | 100.9575 |

Cette variable représente la hauteur du plus bas étage de l'habitation, en pieds, où il y a eu réclamation. Tel qu'observer sur le graphique les réclamation sont généralements plus élevés lorsque la hauteur du plus bas étage est petite. De plus, dans le tableau, on remarque que la moyenne est de 8.29717 pieds. Alors, les habitations sont habituellement près du niveau du sol.

## Variable *numberOfFloorsInTheInsuredBuilding*





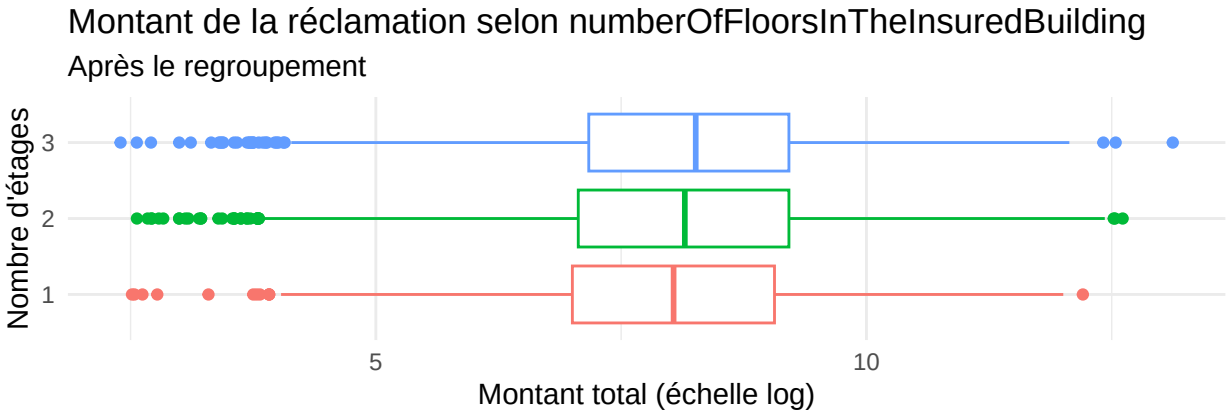




Table 6: Statistiques sur totalAmount selon numberOfFloorsInTheInsuredBuilding

| | Min | Max | Médiane | Moyenne | Écart-type |
|---|---|---|---|---|---|
| 1 étage | 12.36 | 2e+05 | 3085.940 | 8193.950 | 16902.48 |
| 2 étages | 13.00 | 3e+05 | 3460.725 | 9277.239 | 18846.50 |
| 3 étages et plus | 11.00 | 5e+05 | 3867.410 | 8819.302 | 17647.34 |

Cette variabe indique le nombre d'étage de l'habitation assuré. Elle comportait 6 catégories d'habitation, nous avons conservé les trois premiers niveaux, et nous avons combinés les trois dernières selon leur nombre d'étages. Nous avons ainsi obtenu les niveaux suivants :

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

Nous observons que la majorité des dommages sont concentés sur les maisons à 1 ou 2 étages.

### Variable *occupancyType*

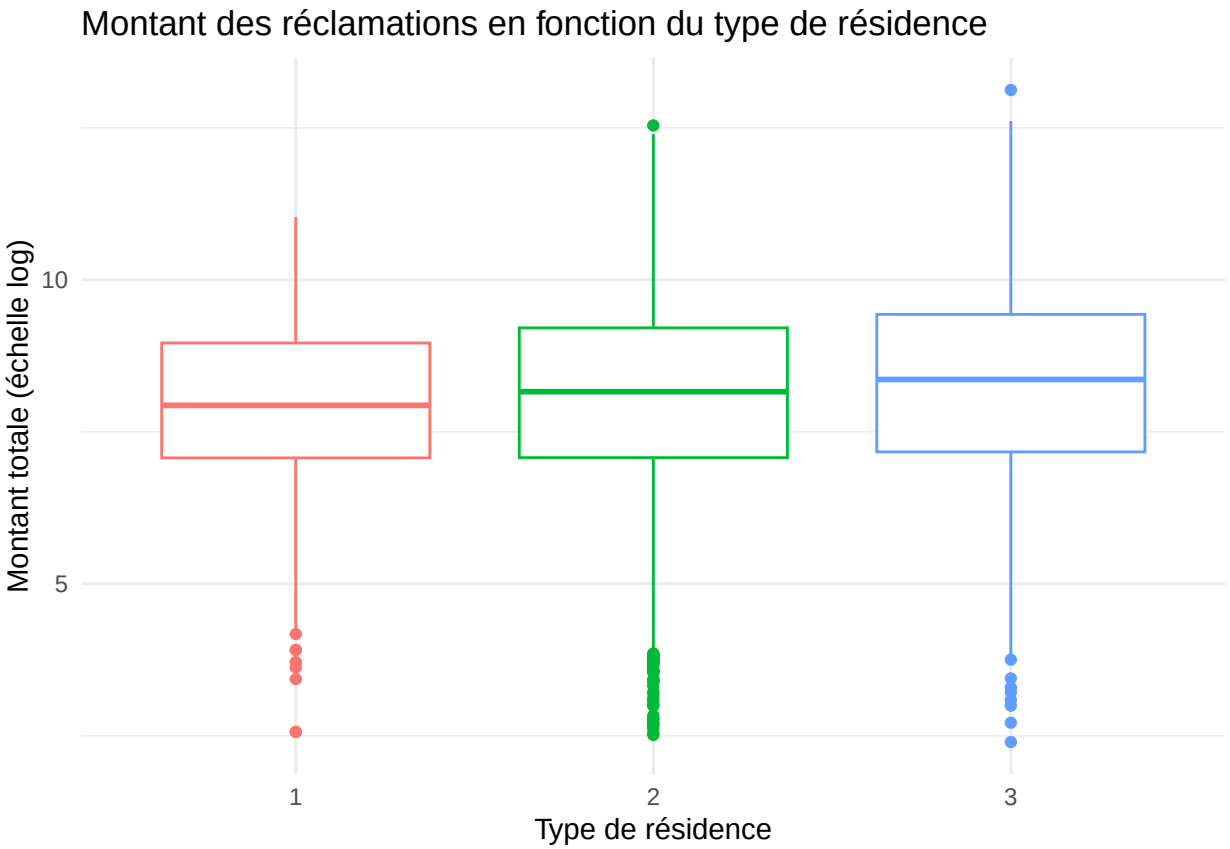



Table 7: Statistiques sur totalAmount selon occupancyType

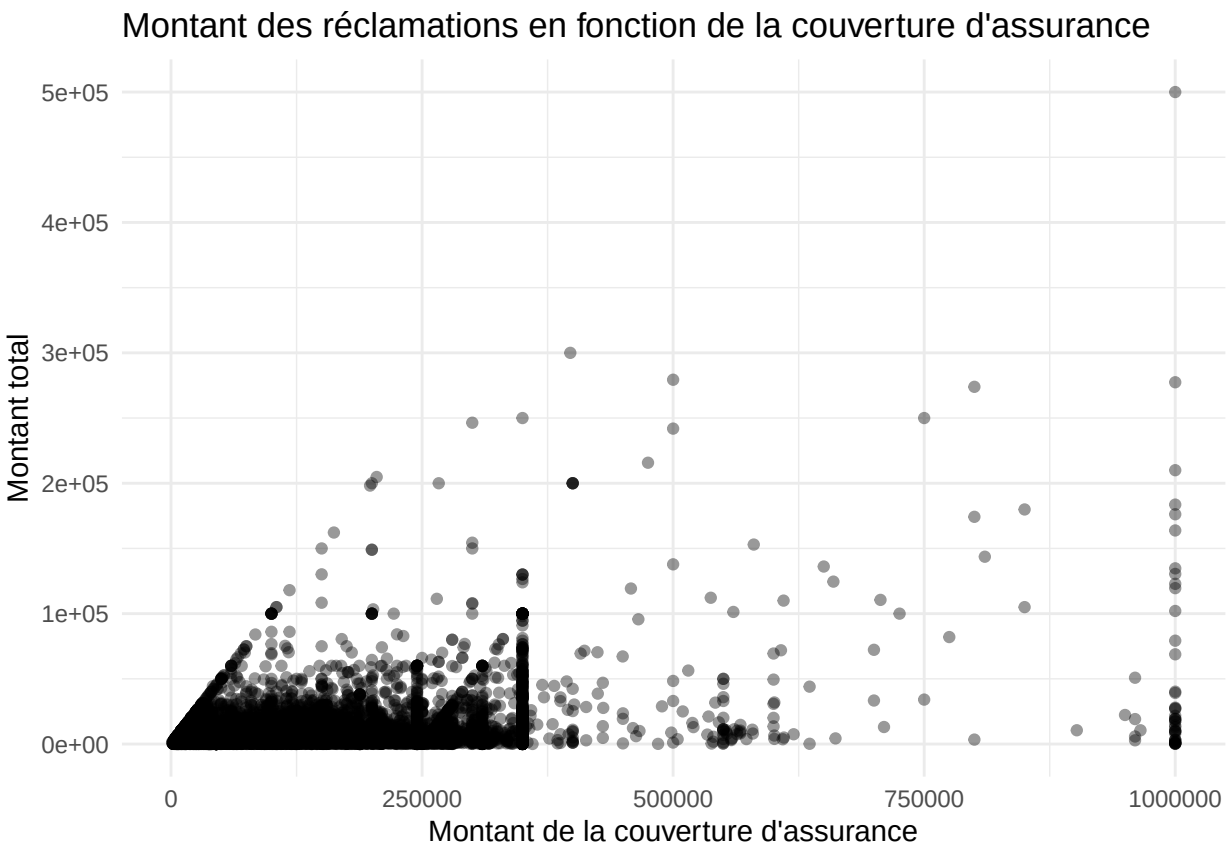
| | Min | Max | Médiane | Moyenne | Écart-type |
|---|---|---|---|---|---|
| Résidences familiales | 12.98 | 61931.6 | 2796.75 | 6316.045 | 9347.912 |
| Copropriétés résidentielles | 12.36 | 279376.8 | 3500.00 | 8026.237 | 13784.176 |
| Non-résidentiel | 11.00 | 500000.0 | 4278.93 | 13982.823 | 31163.459 |

Cette variable catégorielle indique le type de résidence du bâtiment.

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

Tel qu'observé dans le graphique et la table de statistiques, les résidences famillliales ont en général des dommages moins importants que les copropriétés et il évident que les édifices non résidentiels ont des réclamations plus élevées en cas de sinistres.

### Variable *totalCoverage*




Dans le jeu de données, les montants de couverture (totalBuildingInsuranceCoverage) pour le bâtiment et pour les biens personnels (totalContentsInsuranceCoverage) étaient séparés, nous avons jugés plus logique de combiner les deux variables en une seule (totalCoverage), puisque nous avions déjà réunis les montants des réclamations.

Nous remarquons que le graphique forme un triangle inférieur. Cette formation est logique puisque nous ne pouvons pas être dédommagé plus que notre couverture d'assurance.

### Variable *primaryResidence*

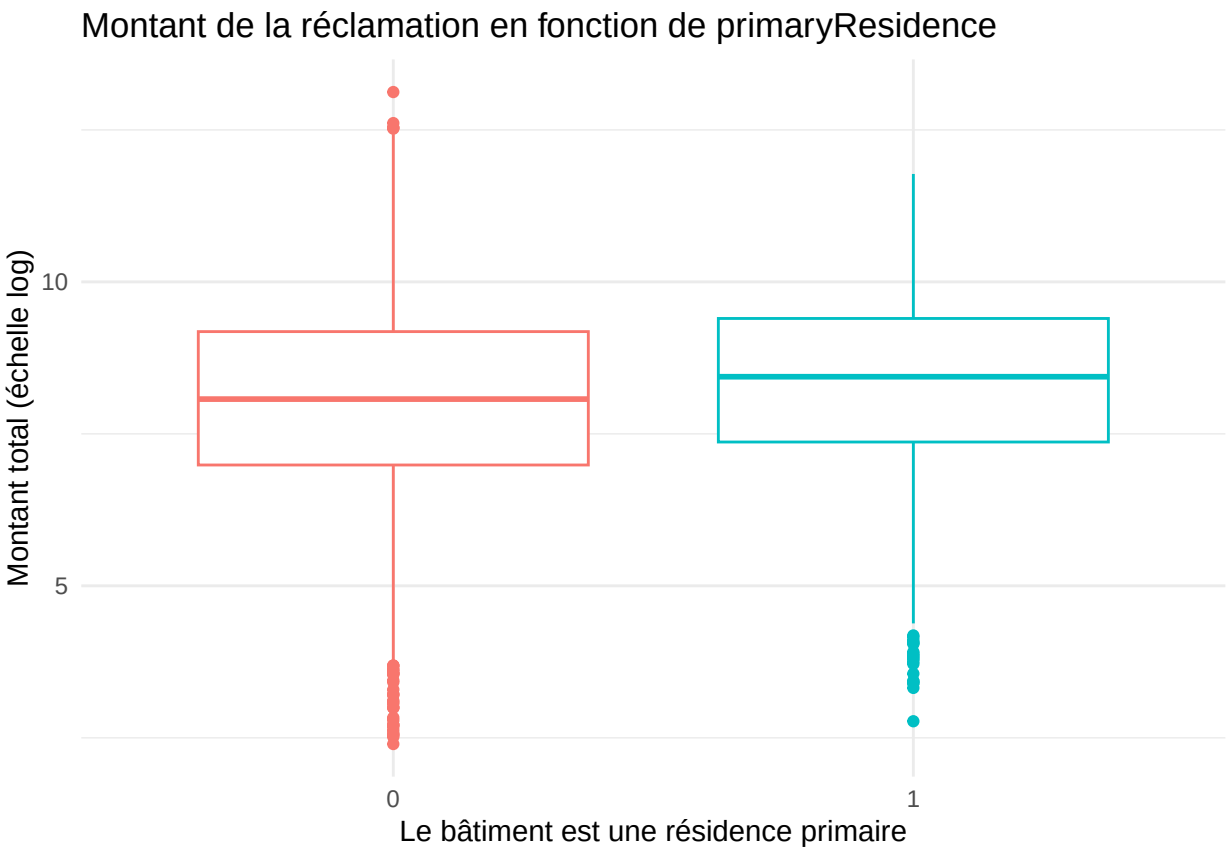



Table 8: Statistiques sur totalAmount selon primaryResidence

| | Min | Max | Médiane | Moyenne | Écart-type |
|---|---|---|---|---|---|
| Résidence primaire | 16 | 130001 | 4626.50 | 10506.283 | 16195.20 |
| Ne l'est pas | 11 | 500000 | 3202.18 | 8556.577 | 18956.43 |

Cette variable indique si le bâtiment assuré est la résidence primaire du client en prenant la valeur 1, dans le cas contraire 0.

En regardant le graphique et le table, il est évident que les réclamation sont plus importantes pour les résidences primaires.

# Conclusion

En conclusion, on s'intéresse au montant des réclamations à la suite d'inondations dans l'état de la Californie, aux États-Unis. Cette première partie du rapport a consisté à sélectionner et à traiter les variables significatives pour notre problème : en retirant les variables non significatives, en traitant les données manquantes et en regroupant des variables & obervations. Au départ, on avait 41 variables avec environ 50 000 observations, à la suite de l'analyse et du traitement des données on a maintenant 13 variables avec environ 10 000 observations qui vont nous permettre de résoudre le problème. Pour ce faire, un modèle linéaire généralisé Gamma ou une régression linéaire multiple pourraient être utilisés.

# Bibliographie

The Federal Emergency Management Agency (2023). FIMA NFIP Redacted Claims - v1.

Récupéré de https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v1

# Annexe

Source : FEMA

Lien : https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v1

Description : Base de données de réclamations d'assurance faites, par contrat, à la suite d'inondations aux États-Unis. Puisque le jeux de données est trop volumineux, on utilise un subset de l'état de Californie.

Taille du jeu de données : 50 000 observations, 30 variables

Variable réponse : Le montant total payé par réclamation en dollar USD. Pour ce faire, nous devons additionner amountPaidOnBuildingClaim, amountPaidOnContentsCaim et amountPaidOnIncreasedCostOfCompliance-Claim

Exposition: *policyCount* le nombre de polices au contrat

Variables explicatives : * 1. *primaryResidence* : Boolean, Y si résidence principale, N sinon * 2. *dateOfLoss* : Date, date à laquelle il y a eu l'infiltration d'eau * 3. *occupancyType* : Catégorielle, indique l'utilisation et le type du bâtiment * 4. *totalBuildingInsuranceCoverage* et *totalContentsInsuranceCoverage* : Numérique, le montant de la couverture au contrat * 5. *longitude* et *latitude* : Spatiale, longitude et latitude du bâtiment assuré