

TRAVAIL PRATIQUE

APPRENTISSAGE STATISTIQUE EN ACTUARIAT
ACT-4114

ÉQUIPE 07

Premier rapport
Analyse exploratoire

Par

Justin Poulin	<i>Numéro d'identification</i>
Cédric Provencher	536 778 899
Philippe Gélinas	536 770 011
Randy Lefebvre	536 782 326
Alex Maggioni	536 776 846
	536 783 575

Travail présenté à
Monsieur
OLIVIER CÔTÉ

13 MARS 2023



Table des Matières

Introduction	2
Analyse Exploratoire	3
Nettoyage des données	3
Somme des réclamations	4
Exposition	5
Genre	6
Âge du conducteur	7
Année du véhicule	8
Modèle du véhicule	9
Région	10
État	11
Prime Totale	12
Montant d'assurance moyen	14
Traitement des valeurs manquantes	15
Visualisation	15
Création d'une nouvelle variable explicative	17
Imputation des données manquantes	18
Conclusion	19
Bibliographie	20
Annexes	21
Description du jeu de données	21
Tests du khi-carré de Pearson	21
Patron de non-réponse de la variable «VehModel»	21
Patron de non-réponse de la variable «VehGroup»	22
Patron de non-réponse de la variable «VehManuf»	23
Patron de non-réponse de la variable «DrivAge»	24
Patron de non-réponse de la variable «Gender»	25
Code de l'imputation des valeurs manquantes	26

Introduction

Pour ce travail, l'objectif est de construire un modèle de prime pure pour la somme des réclamations en assurance automobile (tous types confondus). La base de données utilisée s'appelle *brvehins1a* et est basée sur une partie de l'*AUTOSEG*, un système de statistiques sur les automobiles assurées au Brésil [SUSEP - Superintendência de Seguros Privados, 2013]. Sa structure est très intuitive. En bref, chaque profil véhicule-assuré est associé à une ligne. Chacune des lignes renferme l'ensemble des informations connues sur l'assuré et sur les montants de réclamations dont il est responsable en date de la production du jeu de données. Puisque les mêmes informations sont demandées pour chaque client, les lignes sont divisées en colonnes homogènes, ce qui nous permet de faire une analyse efficace du comportement des données.

En tout, chaque ligne comporte 23 variables, dont 5 variables décrivant les montants de réclamations, chacune étant associée à un type de dommage spécifique (par exemple une collision ou un vol de la voiture). Nous avons choisi de nous intéresser au montant total des réclamations pour un client et c'est donc la somme des 5 types de réclamations possibles qui sera notre variable réponse. Notre modèle utilisera les informations collectés par le système pour établir une prédition de l'ensemble des coûts de réclamations que le client engendrera pour une exposition unitaire. À cette fin, il y a 9 variables explicatives sur lesquelles concentrer nos efforts. En effet, les autres variables sont dépendantes ou encore collige l'information rétrospectivement par rapport aux montants de réclamations.

Pour ces données, la documentation d'*AUTOSEG* indique que l'exposition se mesure en année-véhicule. Une exposition unitaire correspond donc à un véhicule assuré pendant une année. Nous pourrons donc l'utiliser pour calibrer nos modèles. Ce type d'analyse pourrait avoir plusieurs utilités, notamment pour planter une tarification avantageuse ou encore pour faire des études de marchés.

[SUSEP - Superintendência de Seguros Privados, 2013] SUSEP - Superintendência de Seguros Privados, (2013). Autoseg - Sistema de estatísticas de automóveis da susep. <https://www2.susep.gov.br/menues/tatistica/Autoseg/principal.aspx>.

Analyse Exploratoire

Nettoyage des données

Avant de lancer le développement de notre modèle, quelques ajustements sont requis. Notamment, les variables «State» et «StateAb» représentent la même information, l'une étant simplement l'abréviation de l'autre. Nous allons seulement conserver la variable «State».

Également, une des observations a une année de véhicule de 0, ce qui est clairement impossible, donc on remplace le 0 par un **NA**.

Nous créons aussi les variables représentant le nombre total de sinistres ainsi que le montant total de réclamations pour chaque assuré. Puisque l'information initiale est trop granulaire pour les besoins de cette analyse, nous consolidons les 5 variables de réclamations en une seule, nouvelle, qui constitue la somme des 5 variables préexistante.

Les montants de sinistre présentés dans les graphiques sont normalisés pour une exposition unitaire.

Somme des réclamations

La somme des réclamations associées à chaque assuré est la variable réponse pour ce projet. Cette variable prend fréquemment la valeur de 0, soit 319910 fois sur 382621. Cela revient à dire qu'un assuré ne fera pas de réclamation 83.6% du temps. La Table 1 présente d'autres informations notables sur cette variable.

Table 1 : Statistiques sommaires des réclamations

Moyenne	Médiane	Écart-type	Minimum	Maximum
1983.2	0	74641.4	0	23051939.4

Puisqu'un montant de réclamation peut prendre des valeurs assez élevées, les graphiques peuvent devenir plutôt difficiles à lire. Pour remédier à ce problème, le logarithme naturel du montant total des réclamations est utilisé dans les graphiques. Cela a cependant un coût : toutes les réclamations de 0 doivent être retirées du jeu de données. Des informations supplémentaires par rapport à ces réclamations seront ajoutées au long de l'analyse exploratoire pour pallier ce défaut.

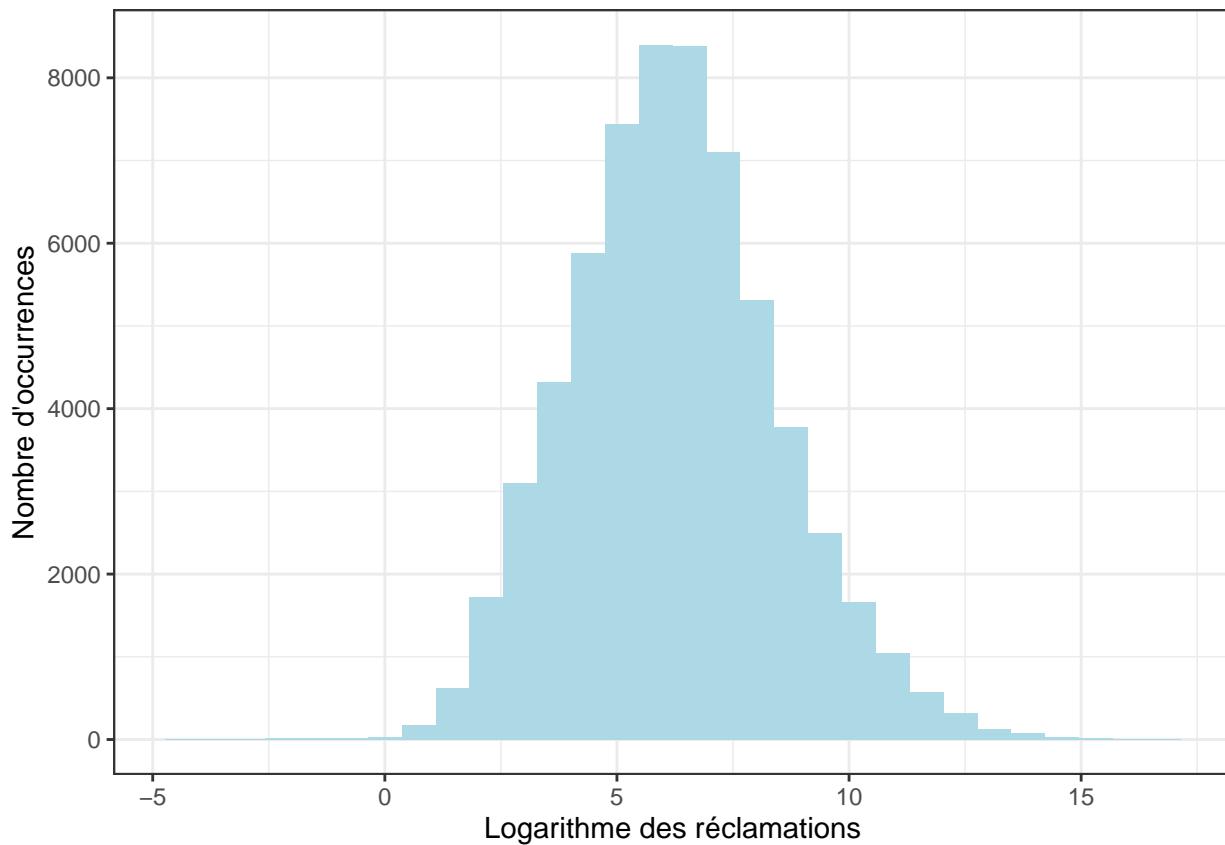


Figure 1: Histogramme du logarithme des montants de réclamation

Comme on peut voir dans la Figure 1, la distribution est quelque peu asymétrique et la queue est très lourde. En effet, on observe une fréquence toute élevée de réclamations très importantes. Il n'est pas rare que le logarithme naturel du montant de réclamations atteigne 10 et même plus, alors que la médiane est plutôt autour de 6.

Exposition

Sans surprise, la Table 2, traitant de l'exposition, présente un minimum de 1/365, soit une journée. Certaines combinaisons d'assuré-véhicule sont beaucoup plus communes que d'autres, de sorte que l'écart-type est particulièrement élevé, de même que le maximum.

Table 2 : Statistiques sommaires de l'exposition

Moyenne	Médiane	Écart-type	Minimum	Maximum
3.2946203	0.59	15.1437203	0.0027397	1829.91

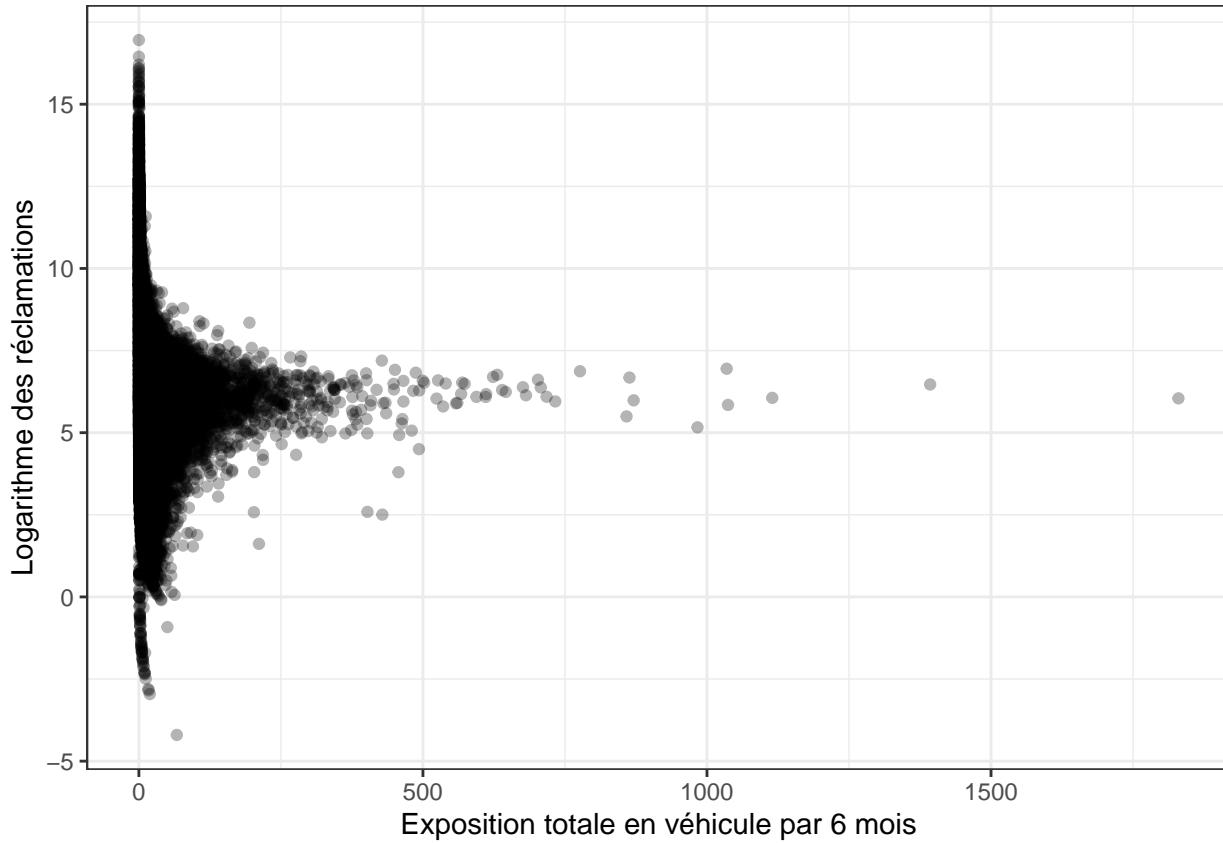


Figure 2: Logarithme des réclamations en fonction de l'exposition

On voit aisément dans la Figure 2 que le montant de réclamations normalisé tend vers le centre plus la mesure d'exposition est élevée. C'est ce à quoi on peut s'attendre par le théorème centrale limite. La mesure d'exposition semble donc satisfaisante pour l'analyse.

Genre

La variable **Gender** a quatre niveaux : *Male*, *Female*, *Corporate* et *NA*, dont les fréquences sont présentées à la Table 3.

Table 3 : Fréquences associées à la variable "Genre"

	Male	Female	Corporate	NA
Fréquence	168825	128259	75067	10470

En analysant le lien avec les réclamations à la Table 4, on remarque qu'il n'y a absolument aucune réclamation pour le genre *Corporate*. Étant donné la très grande exposition qui y est associée, c'est évidemment absurde et on en déduit que l'assureur demande l'information sur le conducteur lorsqu'un accident survient et qu'une réclamation est nécessaire. Ainsi, il ajoute l'information rétrospectivement. Pour faire des prédictions, il sera possiblement nécessaire de regrouper le genre *Corporate* avec les clients pour lesquels l'information sur le genre est inconnue.

Table 4 : Fréquences de non-réclamation pour la variable "Genre"

	Male	Female	Corporate	NA
Aucune réclamation	135245	101846	75067	7752

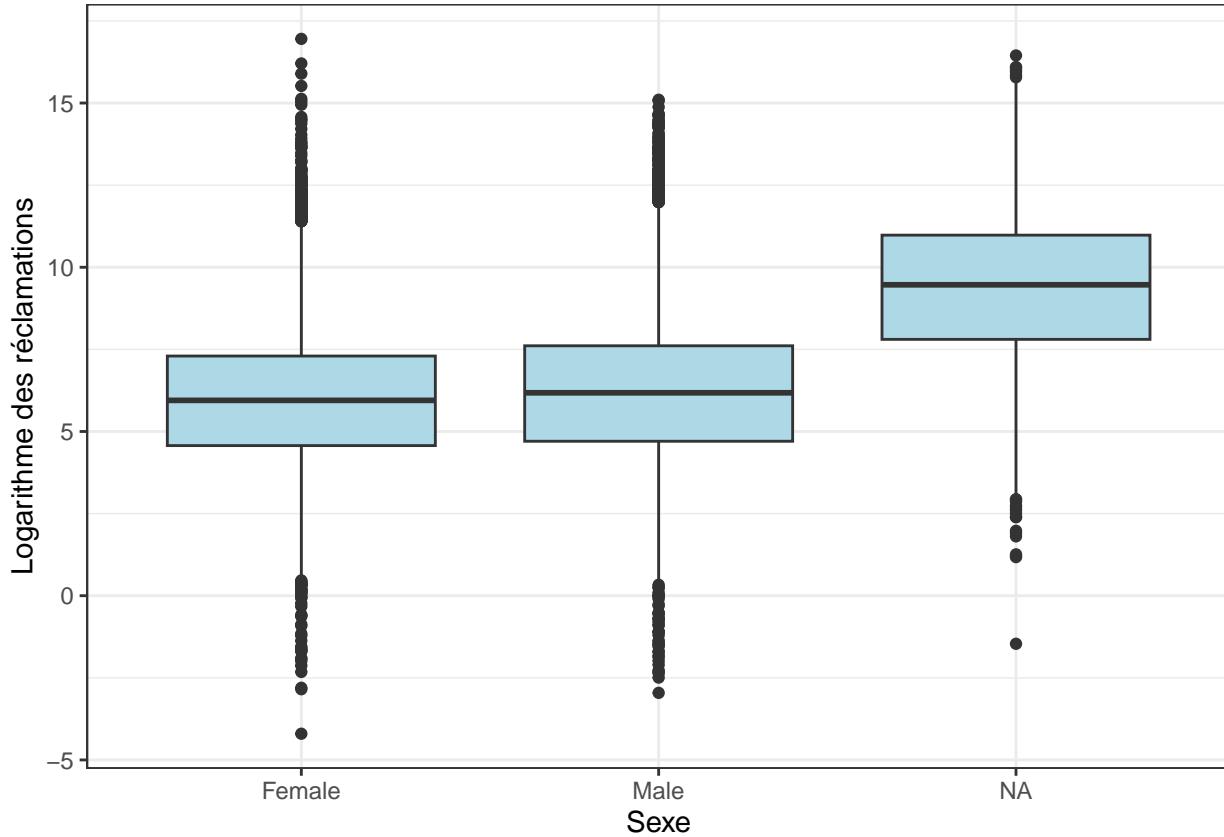


Figure 3: Logarithme des réclamations par genre

Les moyennes des réclamations ne semblent pas changer d'un genre à un autre, mais il est possible d'observer à la Figure 3 des réclamations plus élevées en moyenne lorsque le genre est inconnu. Il faut admettre que c'est plutôt surprenant de prime abord. Il convient de noter que c'est l'échantillon le plus faible des trois catégories, mais avec plus de 10 000 réclamations, il est difficile de blâmer le hasard. Une analyse du patron de non-réponse incluant les autres variables s'impose.

Âge du conducteur

La variable de l'âge du conducteur n'est pas continue, mais plutôt divisée en 5 groupes. La fréquence de chaque niveau est illustrée ci-dessous à la Table 5.

Table 5 : Fréquences associées à la variable "Âge du conducteur"

	18-25	26-35	36-45	46-55	>55
Fréquence	76396	25184	67554	85547	79032

À la Figure 4, il est possible d'observer que la tranche d'âge 26-35 ans a, en moyenne, les réclamations les plus élevées. Cela est plutôt intuitif car les jeunes conducteurs sont reconnus pour être plus téméraires et ainsi causer plus d'accidents. Une hypothèse vraisemblable pour expliquer que le groupe de conducteurs de 18 à 25 ans n'a pas des réclamations plus élevées est qu'ils n'ont tout simplement pas les moyens de se procurer des véhicules de grande valeur.

Encore une fois, lorsque l'information est manquante, les réclamations sont bien plus élevées en moyenne.

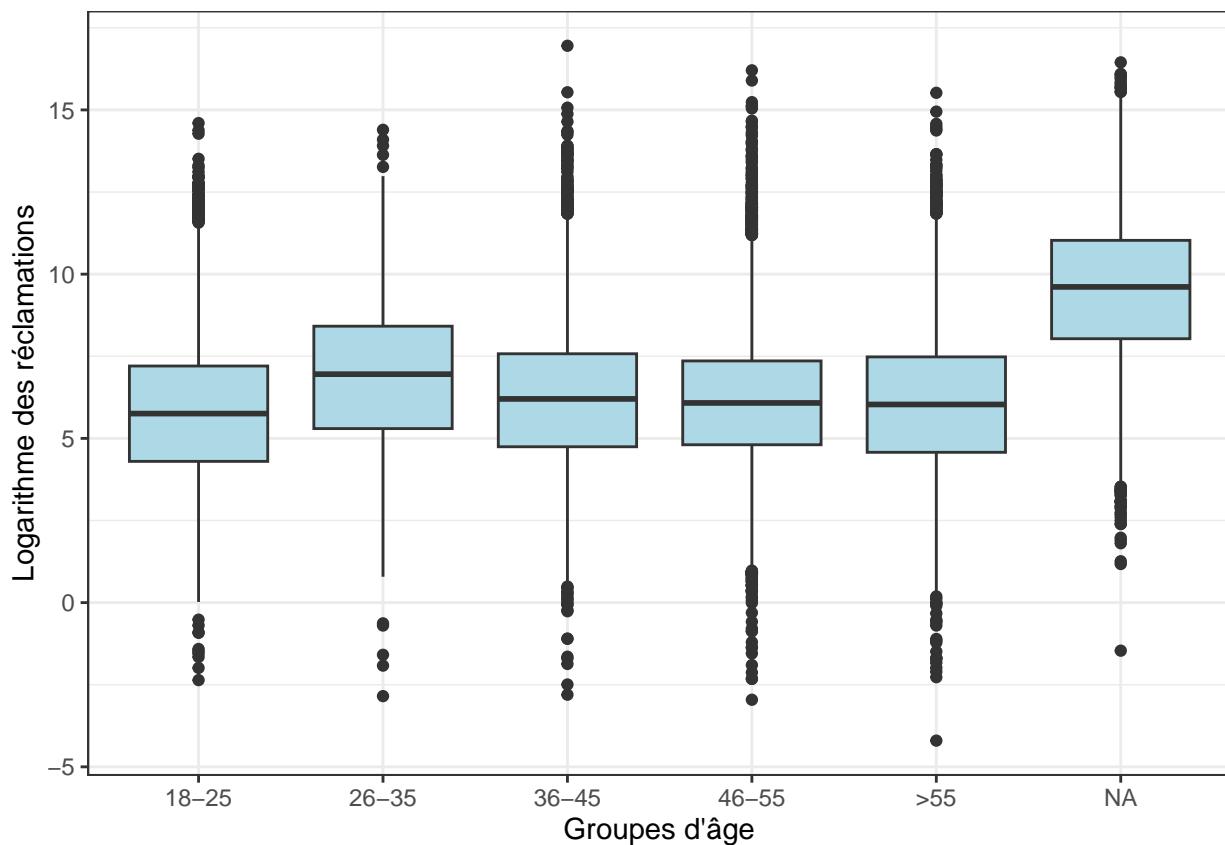


Figure 4: Logarithme des réclamations par Âge de conducteur

Année du véhicule

Une autre variable explicative disponible est l'année du véhicule. Tel que attendu, on constate que la grande majorité du parc automobile est récent. En effet, la Table 6 démontre un faible écart-type de même qu'une médiane très récente.

Table 6 : Statistiques sommaires de l'année du véhicule

Moyenne	Médiane	Écart-type	Minimum	Maximum
2005.2	2007	5.4	1944	2012

Selon ce qu'on peut observer à la Figure 5 à l'aide de la régression, l'année du véhicule semble avoir une relation croissante avec le montant des réclamations. Le volume avant les années 90 semble trop faible pour se prononcer sur la tendance de cette époque.

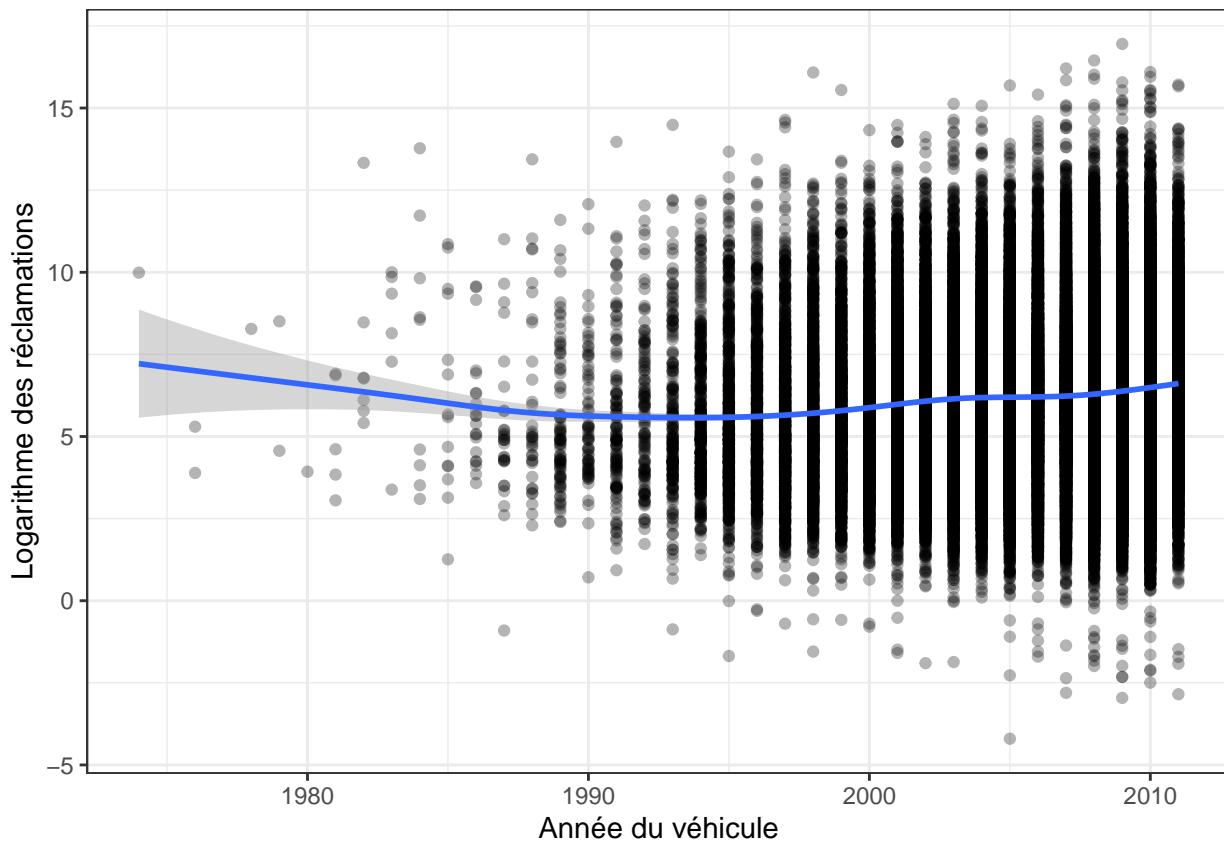


Figure 5: Logarithme des réclamations en fonction de l'année du véhicule

Modèle du véhicule

Pour l'analyse par marque de véhicule, nous avons décidé de regrouper les marques qui ont moins de 500 observations dans la catégorie *Other*. Nous pouvons donc nous concentrer sur les marques les plus importantes.

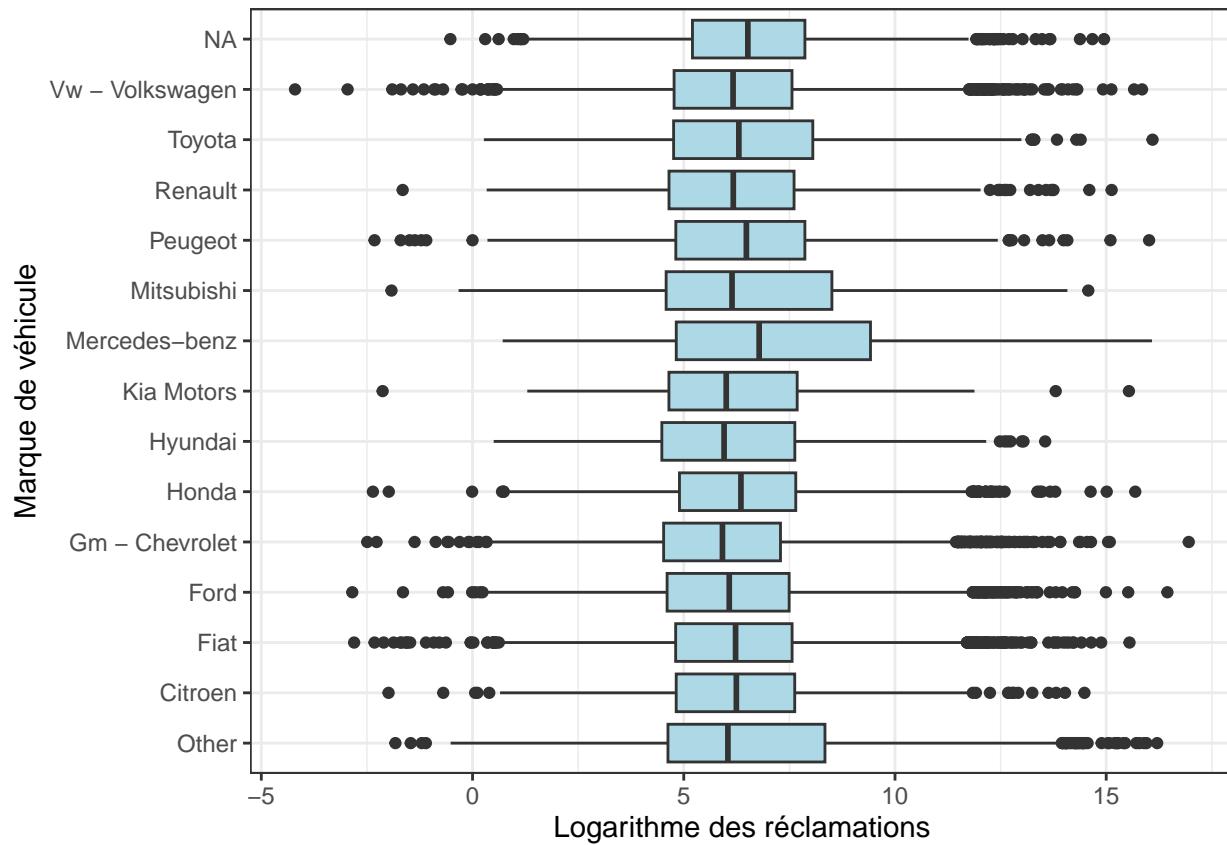


Figure 6: Logarithme des réclamations par région

La Figure 7 ci-haut présente la distribution du logarithme des réclamations selon les principales marques de véhicules présentes dans le jeu de données. Sans surprise, les véhicules les plus luxueux comme *Mercedes-Benz* sont titulaires des réclamations les plus importantes en moyenne. Cependant, comme cette base de données n'origine pas du Québec, il est important de rappeler qu'une partie non négligeable des réclamations sert à payer les dommages corporels de même que les dommages matériels à autrui. Donc même si l'assuré possède un véhicule plus abordable comme une *Kia*, on observe quand même quelques réclamations très substantielles. Enfin, conformément aux variables précédentes, les réclamations sont significativement plus couteuses en moyenne lorsque la marque du véhicule est inconnue.

Région

Pour l'analyse par région, nous avons regroupé les régions qui présentent moins de 2000 observations dans la catégorie *Other*.

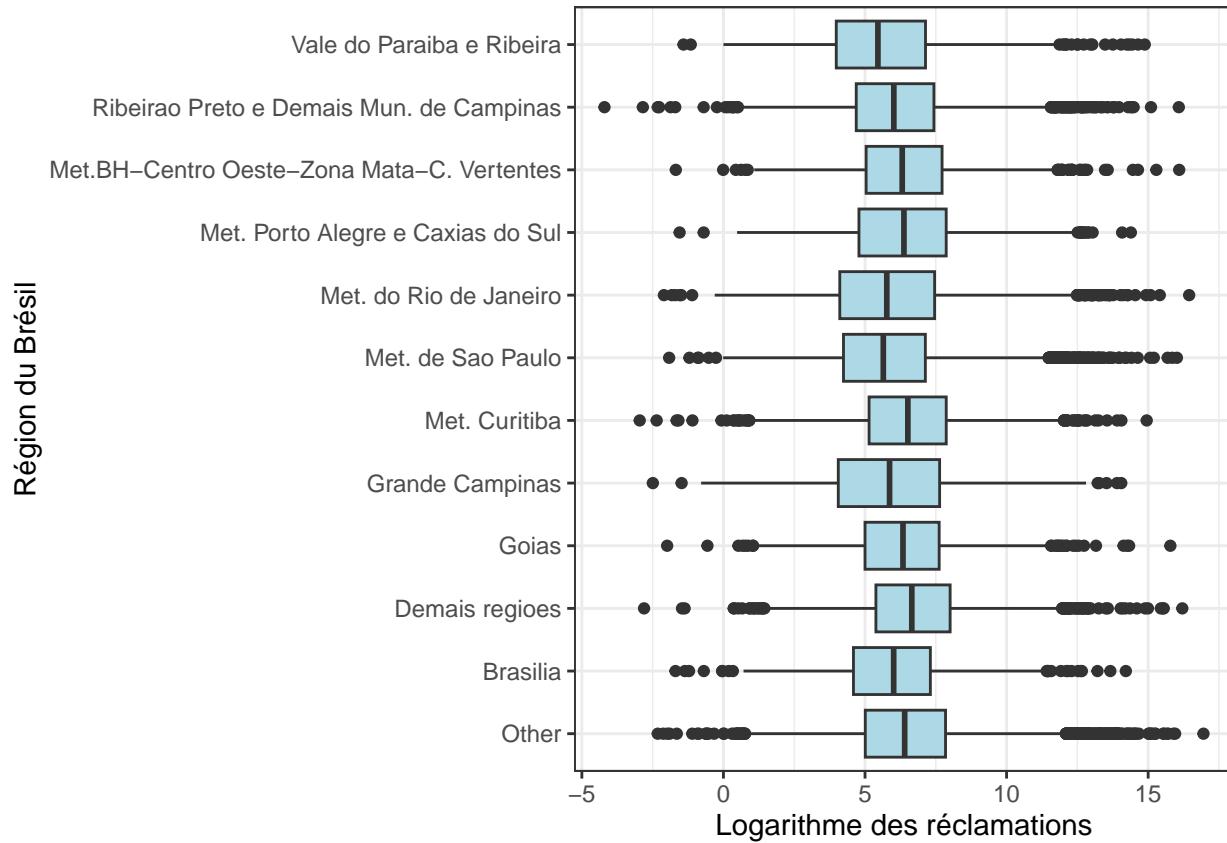


Figure 7: Logarithme des réclamations par région

Il est possible d'observer à la Figure 7 les réclamations dans les régions les plus récurrentes de la base de données. Sans surprise, les réclamations varient beaucoup même au sein d'une même région et il est difficile d'établir des constats en lesquels l'analyste avisé pourrait réellement avoir confiance.

État

Pour l'analyse par ville, tout comme celle pour la région, nous avons regroupé les états qui présentent moins de 2000 observations dans la catégorie *Other*.

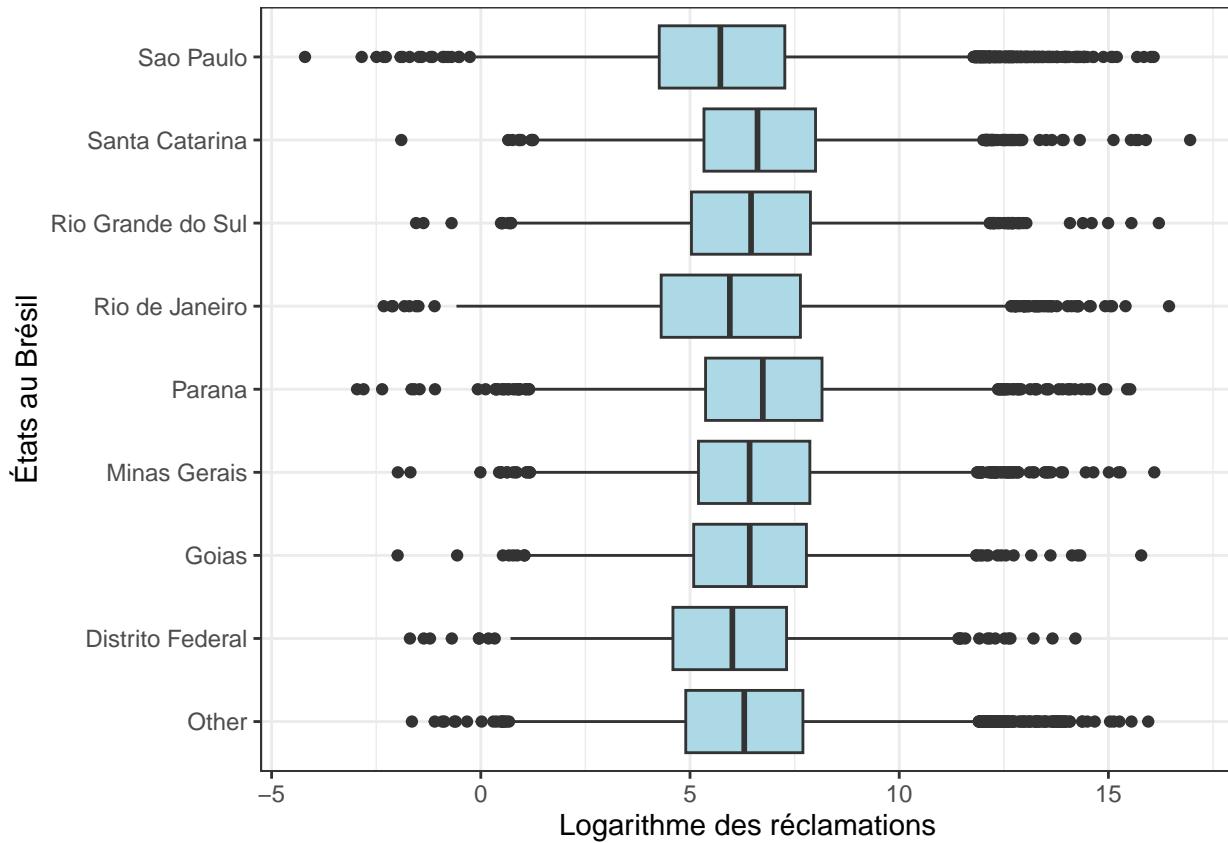


Figure 8: Logarithme des réclamations par état

Un constat similaire est de mise quant au lien entre la ville (ou région administrative selon le cas) et les réclamations, comme le présente la Figure 8. S'il est vrai qu'en moyenne les réclamations sont plus faibles dans la ville Rio de Janeiro qui a assurément la réputation d'accueillir une population moins aisée que Santa Catarina par exemple, il n'en demeure pas moins que cela varie énormément. Des extrêmes sont observés dans toutes les villes du graphique et il est difficile, de prime abord, de ne pas attribuer les écarts au simple hasard.

Prime Totale

La prime totale est une variable qui reflète la somme totale de primes pour une combinaison véhicule-assuré donnée. Afin de mieux comprendre la distribution de cette variable, plusieurs analyses peuvent être effectuées:

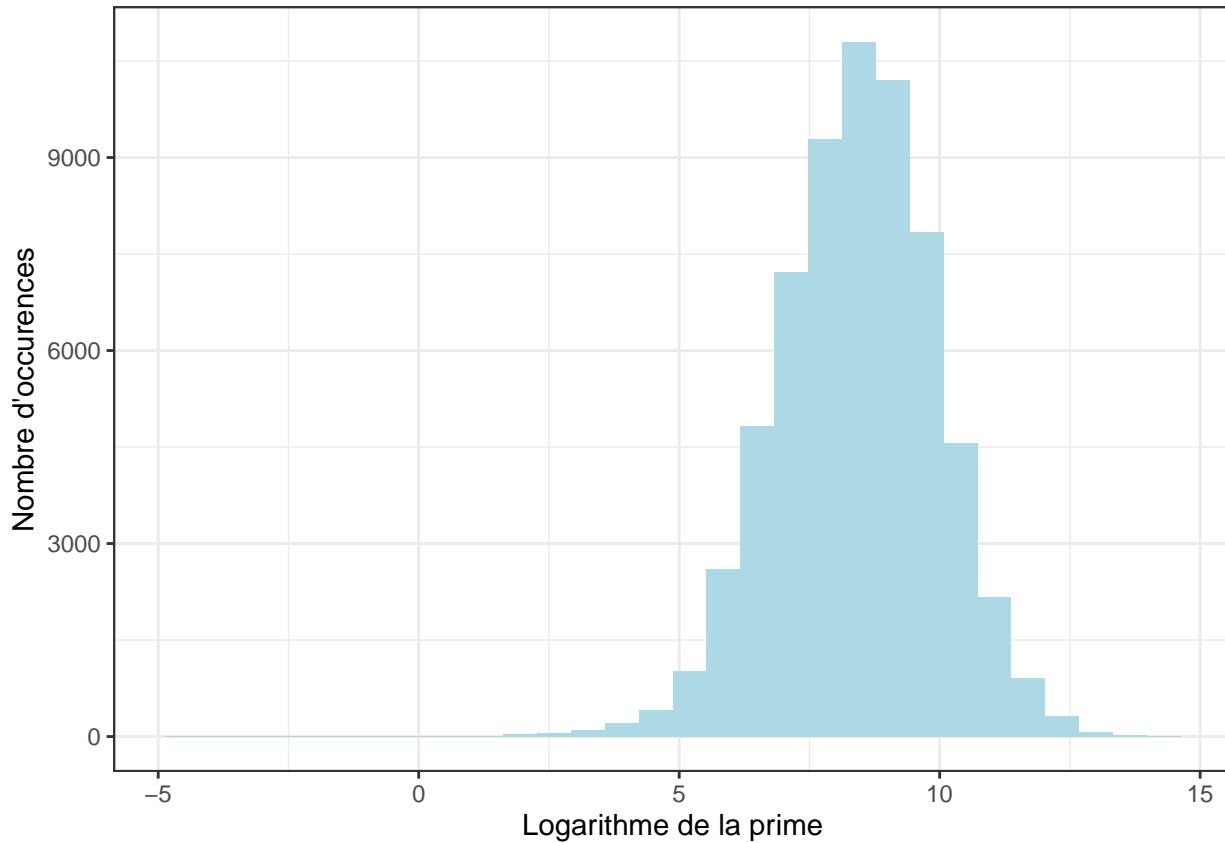


Figure 9: Logarithme des réclamations en fonction de la prime totale

Tout d'abord, l'histogramme de la Figure 9 révèle que la distribution des primes totales présente une asymétrie négative, avec un nombre plus important de primes de faible montant par rapport à celles de montant élevé.

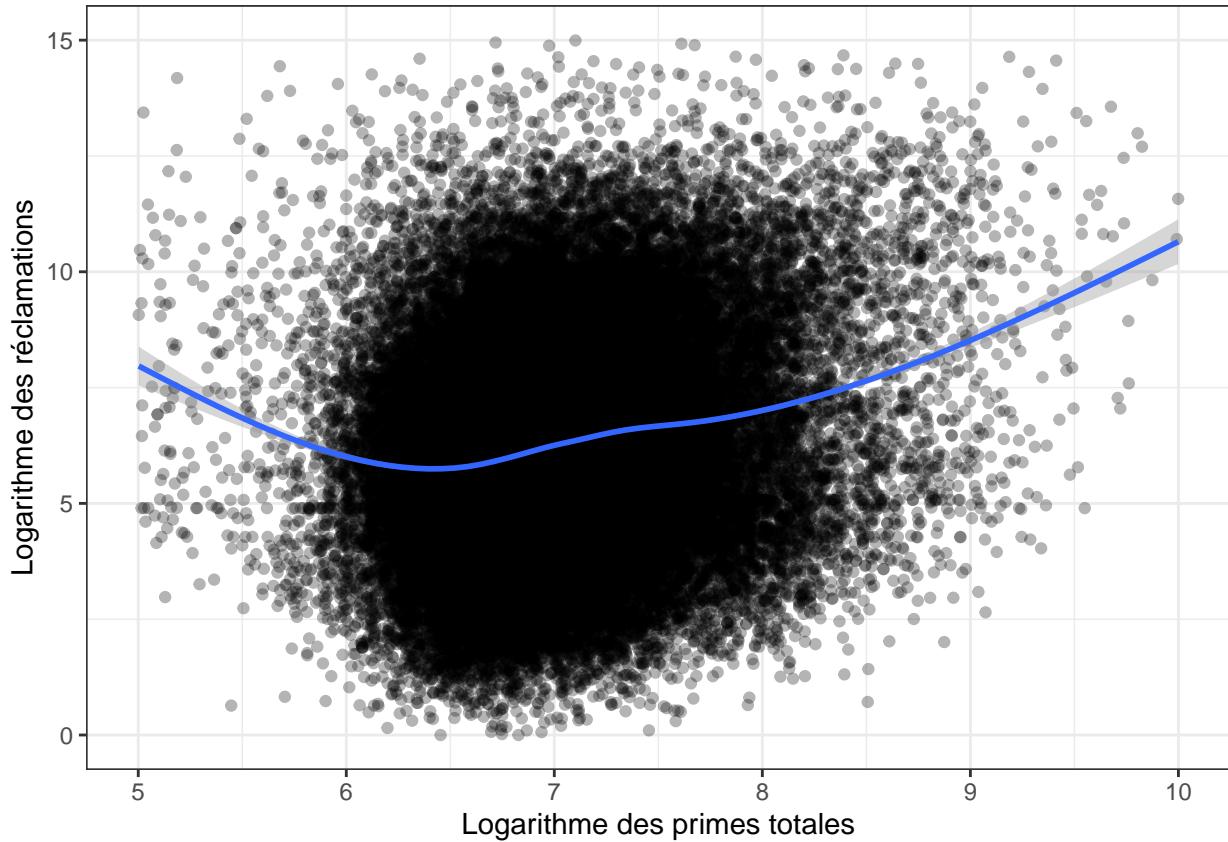


Figure 10: Logarithme des réclamations en fonction de la prime totale

On voit dans la Figure 10 qu'il y a une énorme corrélation entre le montant de prime par exposition et le montant des réclamations. C'est un constat auquel on peut s'attendre considérant que le but de la prime est de couvrir les réclamations.

La prime totale présente clairement la meilleure puissance prédictive, mais elle ne peut être utilisée pour le modèle. Effectivement, il va sans dire que le but de notre projet est de développer notre propre modèle de calcul de la prime. Il est important de garder à l'esprit que la prime est assurément calculée à partir des variables présentées plus haut, en soi, elle n'apporte donc aucune nouvelle information puisqu'elle est elle-même une combinaison de toutes les autres variables. Ainsi, il serait contreproductif de tenter de réutiliser la prime dans notre modèle.

Montant d'assurance moyen

Selon la documentation d'*AUTOSEG*, cette variable correspond à la moyenne pondérée par exposition du montant d'assurance. On s'attend donc à ce que les réclamations moyennes soient inférieures à cette limite d'assurance moyenne.

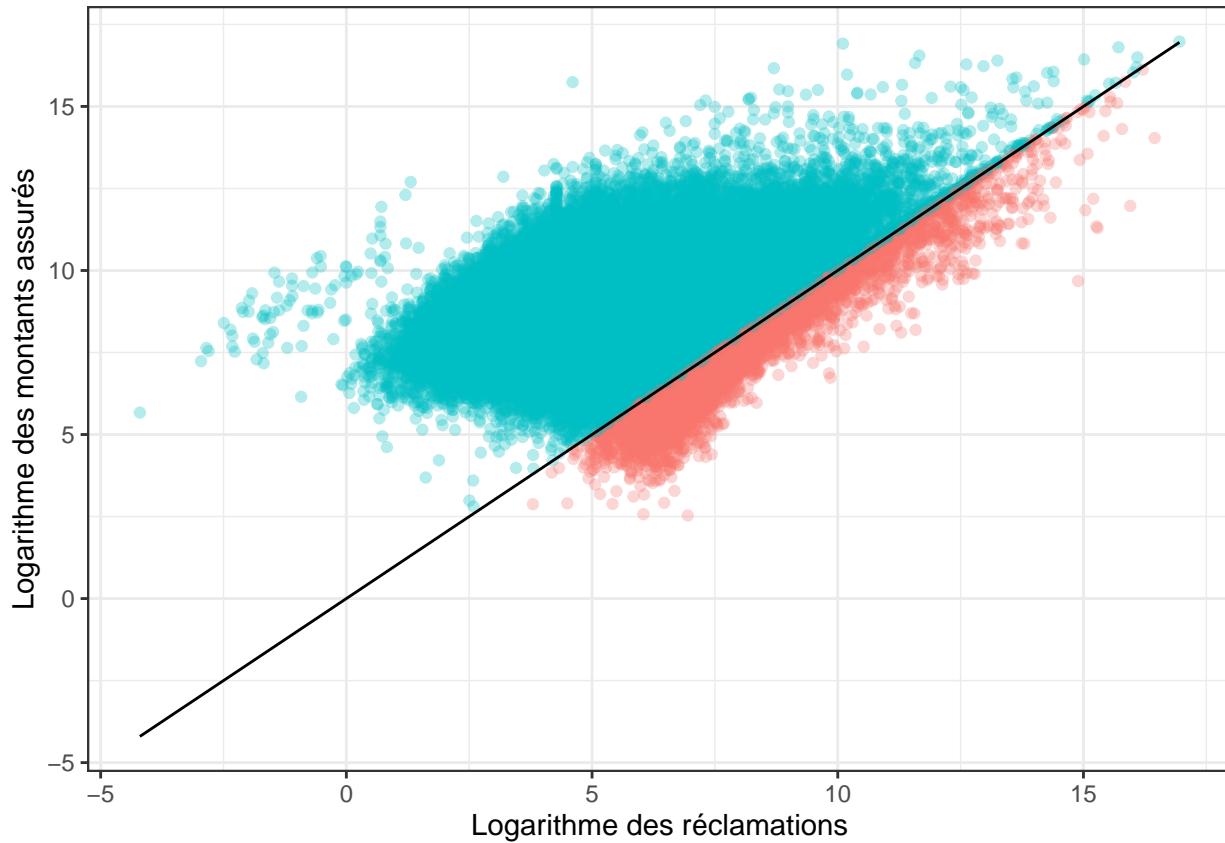


Figure 11: Logarithme des réclamations en fonction de SumInsAvg

Effectivement, on voit dans la Figure 11 que dans la très grande majorité des cas, le montant assuré moyen est supérieur au montant réclamé moyen. Les points sous la courbe sont principalement attribuable au hasard. En effet, si la proportion de réclamation est plus importante pour ceux qui ont une limite plus élevée, il est tout à fait concevable que la réclamation moyenne dépasse la limite d'assurance moyenne.

Traitement des valeurs manquantes

Visualisation

Visualisons les données manquantes.



Figure 12: Visualisation des données manquantes

Dans la Figure 12, les variables n’ayant pas d’observations manquantes n’ont pas été affichées afin de simplifier le diagramme. Au total, 19.39% des rangées ont au moins une variable manquante.

On remarque que lorsque la variable *Area* est inconnue, la variable *State* l’est aussi. Puisqu’il y a seulement 7 observations manquantes pour ces deux variables, nous les retirons du jeu de données.

Il y a seulement une observation avec une donnée manquante pour la variable *VehYear*. En raison de la faible importance de cette observation vis-à-vis la grande taille du jeu de donnée, nous la retirons également du jeu de données.

Lorsque *VehModel* est manquante, *VehGroup* et *VehManuf* le sont également. Cela est tout à fait attendu, car il n'est pas possible de classifier un véhicule si son modèle n'est pas connu et la variable *VehManuf* est créée à partir de la variable *VehModel*. On teste l'hypothèse que le patron de non-réponse de la variable *VehModel* est MCAR. Puisque la variable est catégorielle, le test de khi-carré de Pearson est utilisé. Les tests se retrouvent dans l'annexe. Les *p-values* de ces tests entre le patron de non-réponse de la variable *VehModel* et les variables *Gender*, *DrivAge*, *State* et *VehYear* se rapprochent de 0. En conséquence, on rejette l'hypothèse nulle. Le patron de non-réponse de la variable *VehModel* n'est selon toute vraisemblance pas MCAR. En reproduisant ces tests pour les patrons de non-réponse des variables *VehGroup* et *VehManuf*, la conclusion est malheureusement identique.

La variable *DrivAge* est la plus absente du jeu de données; elle est manquante dans 12.78% des observations. Faisons un diagramme boxplot pour voir s'il y a un lien entre la présence d'une observation manquante pour

cette variable et le montant des réclamations.

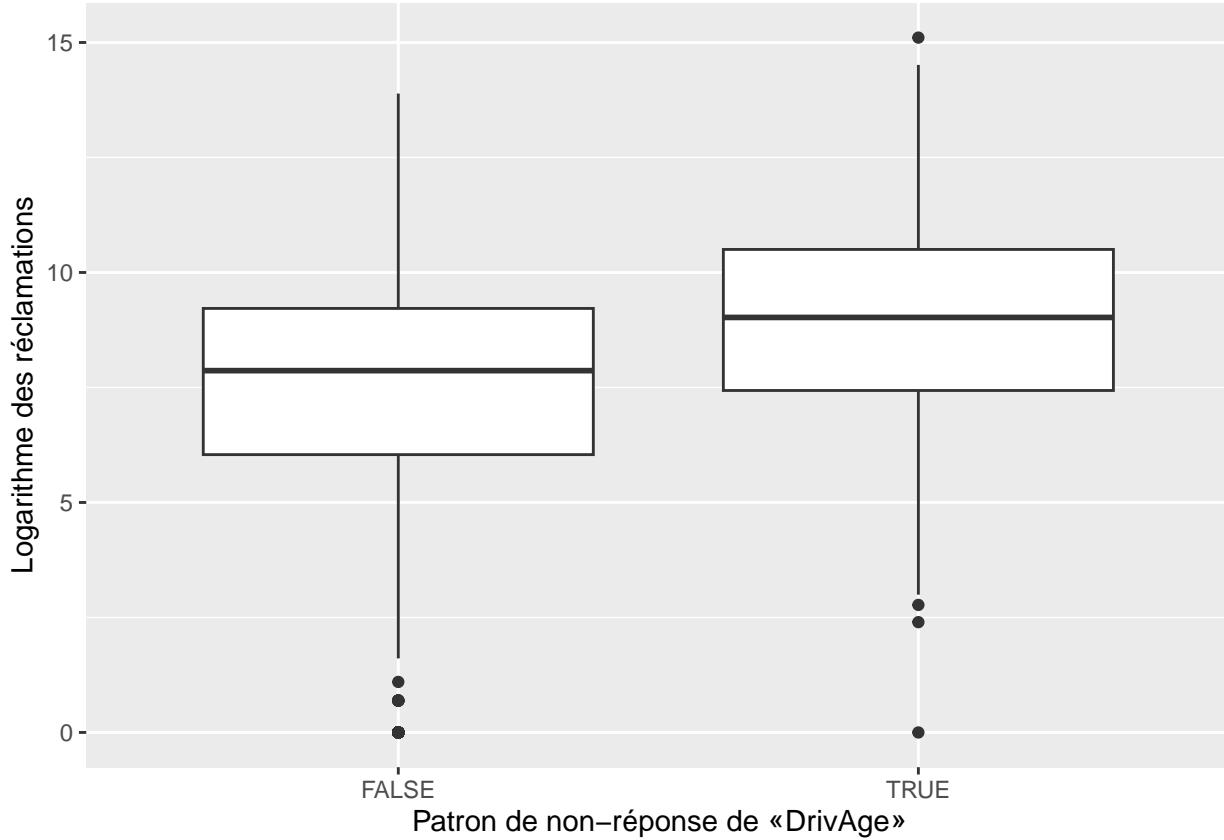


Figure 13: Lien entre le patron de non-réponse de la variable «DrivAge» et le logarithme des réclamations

Vu le grand écart entre le logarithme des réclamations quand la variable *DrivAge* est absente versus quand elle est présente dans la Figure 13, il semble peu probable que les données manquantes de «DrivAge» soient MCAR. Effectivement, les tests du khi-carré de Pearson rejettent l'hypothèse nulle que le Patron de non-réponse de cette variable est MCAR. Les *p-values* des tests entre le patron de non-réponse de *DrivAge* et les variables *Gender*, *VehManuf*, *State* et *Area* sont pratiquement nulles. L'approche des cas complets n'est pas appropriée quand le patron de non-réponse n'est pas MCAR, donc une méthode d'imputation stochastique sera privilégiée. Puisqu'il n'y a pas de test pour savoir si les données manquantes sont MAR ou NMAR, il est possible qu'un biais reste présent dans les données après l'imputation.

La variable *Gender* est rarement absente toute seule. Dans la majorité des observations où cette variable est absente, *DrivAge* l'est également. Cela devra être pris en considération s'il est nécessaire d'imputer les données manquantes puisque l'ordre dans lequel les variables sont imputées a une incidence. Testons l'hypothèse nulle que le patron de non-réponse de la variable *Gender* est MCAR avec des tests du khi-carré de Pearson. Ces tests se retrouvent dans l'annexe. Les *p-values* des tests comparant le patron de non-réponse de la variable *Gender* aux variables *DrivAge*, *VehManuf*, *State* et *Area* sont tous très près de 0. Une fois de plus, l'hypothèse nulle est rejetée : les observations manquantes de la variable *Gender* ne sont pas MCAR.

Création d'une nouvelle variable explicative

Nous croyons que la variable sur le modèle de véhicule contient trop de niveaux, ce qui nuit à leur puissance prédictive et à leur interprétabilité. Nous avons donc choisi d'utiliser les expressions régulières pour regrouper les véhicules par marque. Nous faisons implicitement l'hypothèse que la gamme de prix au sein d'une même marque reste relativement la même. Cette simplification permet une meilleure interprétabilité. De plus, un exemple simple de la création de la nouvelle variable (*VehManuf*) est présenté à la Table 7.

Table 7 : Exemple de la nouvelle variable

VehModel	VehManuf
Ford - Ecosport Xlt 2.0/ 2.0 Flex 16v 5p Mec.	Ford
Fiat - Stilo 1.8 Sporting Flex 8v 5p	Fiat
Toyota - Corolla Xei 1.8/1.8 Flex 16v Mec.	Toyota

Imputation des données manquantes

L'objectif de cette section est d'expliquer notre méthode d'imputation des données manquantes dans le jeu de données. En examinant le motif des données manquantes sur le jeu de données, on constate rapidement que lorsque le modèle de véhicule est manquant pour une observation, les valeurs correspondantes pour les variables *VehGroup* et *VehManuf* sont également manquantes. Pour déterminer si les variables *VehModel* et *VehGroup* fournissent des informations utiles pour notre modèle, nous en faisons une courte analyse:

Niveaux	
VehModel	4259
VehGroup	436

Le nombre de niveaux de ces variables est très élevé, ce qui les rend peu pratiques pour l'imputation des données manquantes. Afin de mieux comprendre les différences entre ces variables, nous avons décidé d'examiner plus en détail chacune d'entre elles:

VehManuf	VehGroup	VehModel
Gm - Chevrolet	Gm Chevrolet Kadett	Kadett GI 2.0 Mpfi / Efi
Harley-davidson	Harley-davidson Motos - Todas	Fat Boy
Volvo	Volvo Caminhoes - Todos	Fh 440 6x2 2p (diesel)

Nous avons remarqué que la variable *VehManuf* est plus sommative que les variables *VehModel* et *VehGroup*. La marque d'un véhicule peut avoir un impact significatif sur le risque associé à une observation, alors que la différence de risque entre différents modèles d'une même marque peut être relativement faible. Sur cette base, en utilisant uniquement la variable *VehManuf*, il est possible de réduire la complexité de l'analyse tout en conservant les informations les plus pertinentes pour l'objectif de notre analyse.

Par conséquent, nous avons décidé de retirer les variables *VehModel* et *VehGroup* du jeu de données pour simplifier l'imputation des données manquantes et l'analyse ultérieure. De plus, cette action permet de réduire la dimensionnalité du jeu de données, ce qui peut conduire à des modèles plus simples et plus facilement interprétables. Certains tests supplémentaires pourront être faits lors du développement du modèle afin de déterminer si les variables sont bel et bien, comme on le pense, trop spécifiques et granulaires pour être utilisées.

Il n'est pas possible d'utiliser les cas complets pour analyser les données manquantes dans notre jeu de données, car les données manquantes ne sont pas manquantes complètement au hasard (MCAR). En effet, il existe une certaine structure dans les données manquantes, ce qui peut introduire un biais important dans l'analyse si les observations manquantes sont simplement ignorées. Par conséquent, nous avons choisi d'utiliser le package **mice** pour effectuer une imputation par forêt aléatoire (voir annexe), qui permet de prendre en compte la structure des données manquantes et de minimiser les biais potentiels introduits par l'imputation.

Après avoir effectué l'imputation des données manquantes, toutes les valeurs manquantes ont été remplacées dans le jeu de données : ce dernier est désormais complet et prêt à être utilisé pour l'analyse.

Conclusion

En guise de conclusion, il importe de garder en tête que l'analyse suivante a pour but ultime de créer un modèle de prime pure pouvant être utilisé à des fins de tarification par des assureurs en Amérique du Sud. Malheureusement, toutes les variables explicatives n'ont pas été colligées exhaustivement dans le jeu de données et il a donc été nécessaire d'imputer stochastiquement les valeurs d'une proportion non négligeable de nos assurés pour les variables *Gender*, *DrivAge* et *VehManuf*. Il aurait clairement été préférable de travailler avec la bonne information directement, mais dans les circonstances, il s'agit de la meilleure avenue possible pour maintenir un niveau de variance acceptable et éviter d'introduire de biais dans nos résultats. Pour les prochaines étapes de cette analyse, puisque nous souhaitons modéliser la prime pure, une distribution tweedie pourrait s'avérer particulièrement appropriée, car beaucoup d'assurés n'ont pas effectué de réclamations. Effectivement, cette distribution est connue pour ses capacités à bien modéliser à la fois une proportion significative de réclamations inexistantes (donc lorsque la variable endogène est 0) et des seuils élevés. De plus, cela permet de ne pas développer deux modèles distincts pour la fréquence et la sévérité.

Bibliographie

[SUSEP - Superintendência de Seguros Privados, 2013] SUSEP - Superintendência de Seguros Privados, (2013). Autoseg - Sistema de estatísticas de automóveis da susep. <https://www2.susep.gov.br/menues/tatistica/Autoseg/principal.aspx>.

Annexes

Description du jeu de données

Nom du jeu de données: brvehins1a

Source: CASdatasets - <https://www2.susep.gov.br/menuestatistica/Autoseg/principal.aspx>

Brève description: Le jeu de données contient des informations sur tous les véhicules assurés par des compagnies d'assurances privées au Brésil. Il contient des informations sur le risque du véhicule et les transactions financières le concernant.

Variable réponse: La somme des montants réclamés, tout type confondus. La variable est quantitative.

Mesure d'exposition: Temps pour lequel le risque a été assuré.

Variables explicatives: Il y en a 9 si on ne considère pas les primes et l'exposition. En voici 5:

Sexe, qualitative. Âge du conducteur, ordinaire. Montant assuré, quantitative. Modèle du véhicule, qualitative. Endroit géographique, qualitative. Taille du jeu de données: 393071 observation et 23 variables.

Tests du khi-carré de Pearson

Patron de non-réponse de la variable «VehModel»

```
manq <- is.na(Veh$VehModel)
chisq.test(Veh$Gender, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$Gender and manq
## X-squared = 116.07, df = 2, p-value < 0.00000000000000022
chisq.test(Veh$DrivAge, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$DrivAge and manq
## X-squared = 109.58, df = 4, p-value < 0.00000000000000022
chisq.test(Veh$State, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$State and manq
## X-squared = 857.83, df = 26, p-value < 0.00000000000000022
chisq.test(Veh$VehYear, manq, correct = F)

## Warning in chisq.test(Veh$VehYear, manq, correct = F): Chi-squared approximation
## may be incorrect

##
## Pearson's Chi-squared test
##
## data: Veh$VehYear and manq
## X-squared = 79827, df = 65, p-value < 0.00000000000000022
```

Patron de non-réponse de la variable «VehGroup»

```
manq <- is.na(Veh$VehGroup)
chisq.test(Veh$Gender, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$Gender and manq
## X-squared = 116.07, df = 2, p-value < 0.00000000000000022
chisq.test(Veh$DrivAge, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$DrivAge and manq
## X-squared = 109.58, df = 4, p-value < 0.00000000000000022
chisq.test(Veh$State, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$State and manq
## X-squared = 857.83, df = 26, p-value < 0.00000000000000022
chisq.test(Veh$VehYear, manq, correct = F)

## Warning in chisq.test(Veh$VehYear, manq, correct = F): Chi-squared approximation
## may be incorrect

##
## Pearson's Chi-squared test
##
## data: Veh$VehYear and manq
## X-squared = 79827, df = 65, p-value < 0.00000000000000022
```

Patron de non-réponse de la variable «VehManuf»

```
manq <- is.na(Veh$VehManuf)
chisq.test(Veh$Gender, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$Gender and manq
## X-squared = 116.07, df = 2, p-value < 0.00000000000000022
chisq.test(Veh$DrivAge, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$DrivAge and manq
## X-squared = 109.58, df = 4, p-value < 0.00000000000000022
chisq.test(Veh$State, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$State and manq
## X-squared = 857.83, df = 26, p-value < 0.00000000000000022
chisq.test(Veh$VehYear, manq, correct = F)

## Warning in chisq.test(Veh$VehYear, manq, correct = F): Chi-squared approximation
## may be incorrect

##
## Pearson's Chi-squared test
##
## data: Veh$VehYear and manq
## X-squared = 79827, df = 65, p-value < 0.00000000000000022
```

Patron de non-réponse de la variable «DrivAge»

```
manq <- is.na(Veh$DrivAge)
chisq.test(Veh$Gender, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$Gender and manq
## X-squared = 185929, df = 2, p-value < 0.00000000000000022
chisq.test(Veh$VehManuf, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$VehManuf and manq
## X-squared = 22644, df = 46, p-value < 0.00000000000000022
chisq.test(Veh$State, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$State and manq
## X-squared = 817.48, df = 26, p-value < 0.00000000000000022
chisq.test(Veh$Area, manq, correct = F)

##
## Pearson's Chi-squared test
##
## data: Veh$Area and manq
## X-squared = 1225.2, df = 39, p-value < 0.00000000000000022
```

Patron de non-réponse de la variable «Gender»

```
manq <- is.na(Veh$Gender)
chisq.test(Veh$DrivAge, manq, correct = F)

##
##  Pearson's Chi-squared test
##
## data: Veh$DrivAge and manq
## X-squared = 179.48, df = 4, p-value < 0.00000000000000022
chisq.test(Veh$VehManuf, manq, correct = F)

## Warning in chisq.test(Veh$VehManuf, manq, correct = F): Chi-squared
## approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data: Veh$VehManuf and manq
## X-squared = 9478.3, df = 46, p-value < 0.00000000000000022
chisq.test(Veh$State, manq, correct = F)

##
##  Pearson's Chi-squared test
##
## data: Veh$State and manq
## X-squared = 2826.8, df = 26, p-value < 0.00000000000000022
chisq.test(Veh$Area, manq, correct = F)

##
##  Pearson's Chi-squared test
##
## data: Veh$Area and manq
## X-squared = 3501.9, df = 39, p-value < 0.00000000000000022
```

Code de l'imputation des valeurs manquantes

```
imp <- mice(temp_data, m = 1, method = "rf")
completedata <- complete(imp)
```