# *NOTE RAPPORT PROJET R*

*Rstudio et Rshiny*

*DAI Elena HOGENMULLER Edmée PAUSE Léa ZORKA Adriana*
*ESILV Master 1 Actuariat*

*Table des matières*

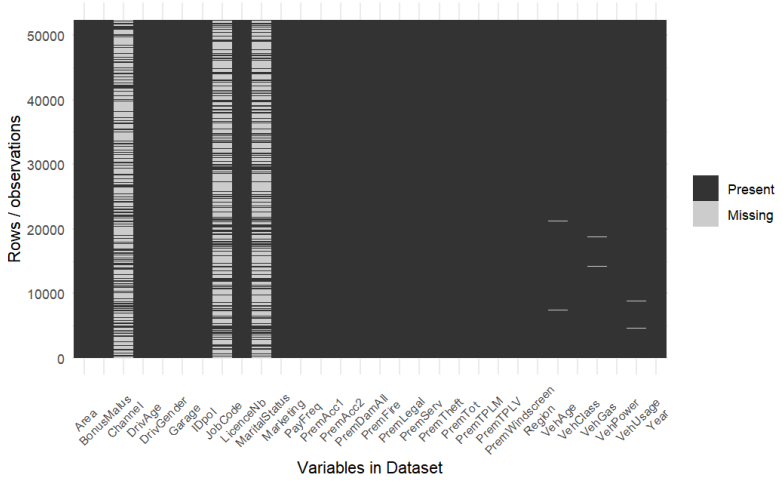
# I) Nettoyage de la base de données

Nous avons recueilli 3 bases de données pour cette étude : la base *Prem*, la base *Sev* et la base *Freq*.

La base *Prem* contient 30 variables et 52 372 lignes, correspondant à des assurés. Ce sont leurs informations des années 2003 et 2004.
La base *sev* contient 8931 observations, correspondant à des sinistres et 6 variables. Ils correspondent aux informations des sinistres des assurés.
La base *freq* contient des informations que nous n'avons ni dans *prem* ni dans *sev*. Nous ne l'avons pas utilisé dans notre étude.

## *Étude des valeurs manquantes*

En étudiant nos bases de données, nous nous sommes aperçus qu'il y avait un grand nombre de données manquantes. Afin d'avoir une vision globale, nous les avons représentées sur un graphe :




Nous observons ainsi que les variables Channel, MartialStatue et JobCode ont énormément de valeurs manquantes. Nous avons décidé de supprimer la colonne channel, et de laisser dans l'état les deux autres colonnes. Ensuite, pour remplacer les données manquantes des variables quantitatives nous avons utilisé la fonction MICE (Multivariate Imputations by Chained Equations).

Pour les caractéristiques du véhicule (power, gaz, usage) nous avons remplacé en fonction de la classe de la voiture la valeur majoritaire de la variable recherchée.

## *Étude des outliers*

Pour cette partie, nous avons enlevé toutes les valeurs aberrantes.

- Âge (DrivAge) : Nous avons admis que l'âge d'un conducteur se trouvait entre 18 et 99 ans. Nous avons remplacé l'âge par la médiane des âges pour les assurés qui avait des âges inférieurs à 18 ou supérieur à 99.
- Genre (DrivGender) : Nous avons rassemblé les genres des assurés en 2 catégories : « F » ou « M ». En effet, on pouvait retrouver des « F » et « M » mais aussi des « Male » et des « Female ».
- Coût des sinistres (Payment) : Nous avons remarqué que nous avions des paiements catégorisé « Waiting », « ? » et négatifs. Nous avons considéré qu'il s'agissait d'une erreur de comptabilité et avons remplacé ces paiements par 0.

Nous avons également remarqué qu'il y avait un sinistre de 632 893 € en 2003, qui était largement supérieur aux autres. Nous avons considéré qu'il s'agissait d'une exception et avons décidé de l'écarter de notre étude, pour ne pas fausser nos résultats.

### *Simplification*

Tout d'abord, nous voulions simplifier les catégories de voiture en 3 types : « Cheap », « Medium » et « Expensive ». Nous avons également enlevé les « p » de VehPow pour avoir des données quantitatives. Pour finir, nous avons remplacé le gaz gazole par diesel (car il s'agit de la même chose). Cette simplification facilitera nos interprétations et analyses dans les parties suivantes.

### *Étude des doublons*

Nous avons d'abord supprimé les lignes doublons de la base *prem*. Nous avons ensuite supprimé les doublons de IDPol au sein d'une même année.

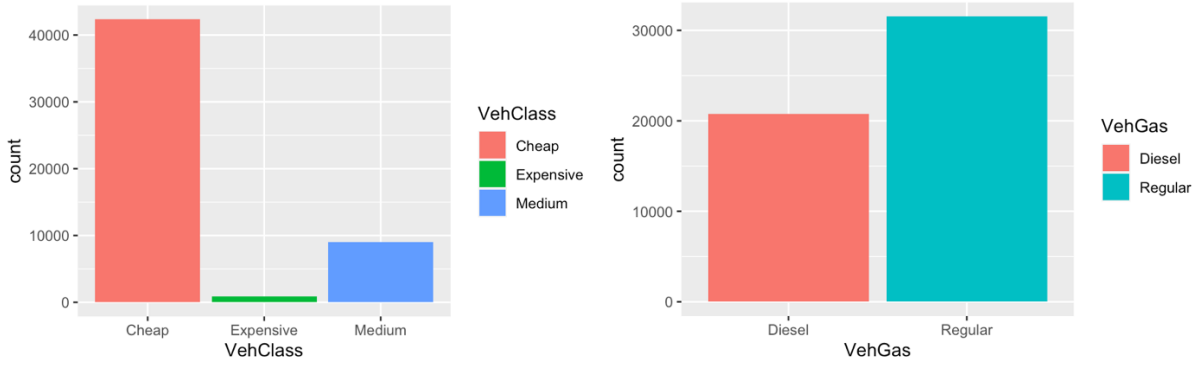
### *Jointure*

Nous avons décidé de rajouter une colonne avec le total de coût de sinistre, la moyenne de coût de sinistres et le nombre de sinistre par personne par année à la table *prem*. Avec la fonction aggregate nous avons créé une nouvelle table avec l'année, l'IDpol et ces informations, que nous avons joint avec la table *prem*. Pour ceux qui n'avaient pas eu de sinistres, nous avons remplacé les valeurs manquantes par 0.

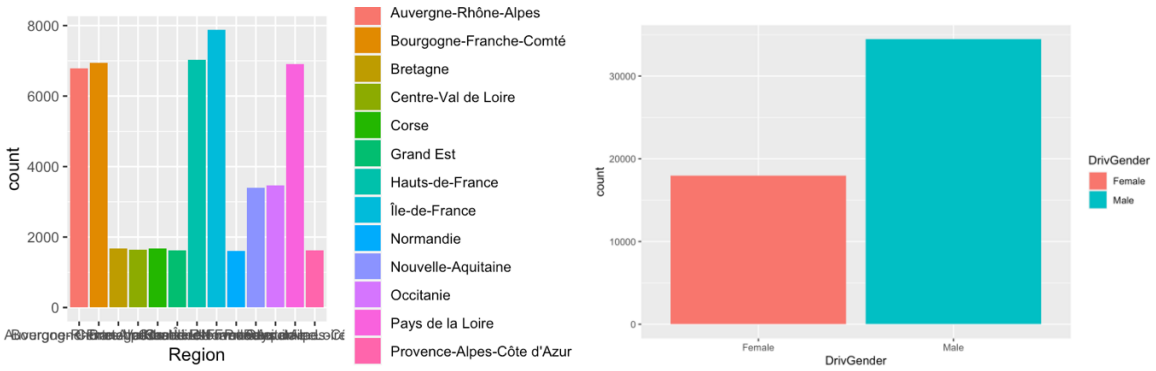
## II) Étude univariée

### *Variables qualitatives*

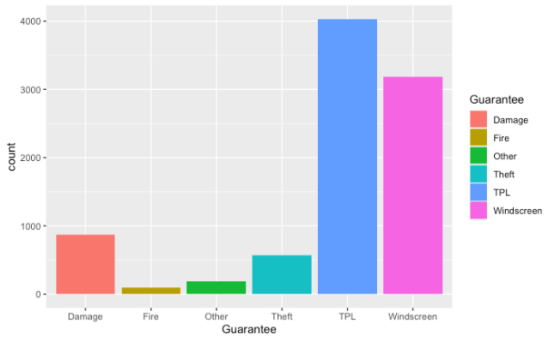
Pour étudier deux variables qualitatives, nous pouvons faire un tableau de fréquences (absolues et relatives) ainsi que les diagramme bâton/secteurs.




Les voitures de notre base sont en général des voitures peu chères, avec le statut « cheap », on retrouve ensuite des voitures de catégorie moyenne et très peu de voitures chères. La majorité de ces voitures consomment de l'essence « Regular » (3/5 des voitures) et on note également la présence de voiture « Diesel » ou Gazole (2/5) des voitures.

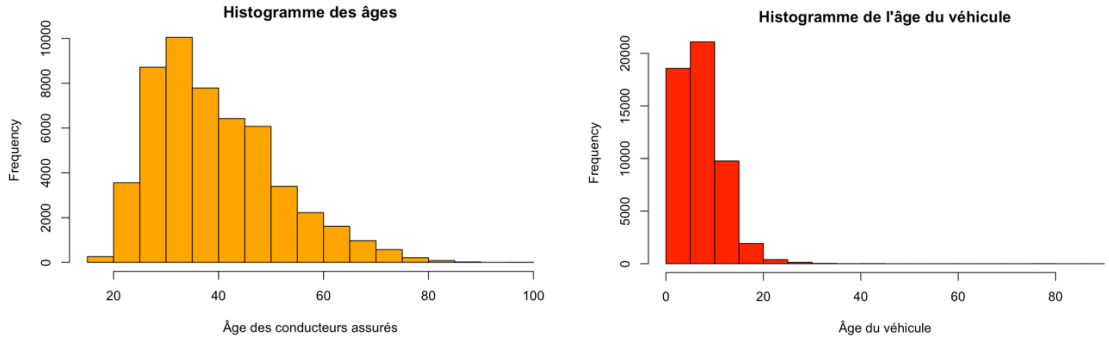



Concernant les régions, les assurés sont présents dans toute la France métropolitaine, on note tout de même une forte concentration dans les régions Auvergne, Bourgogne, Hauts-de-France, Île de France et Pays de la Loire. Nous avons une majorité du genre masculin en termes d'assurés.

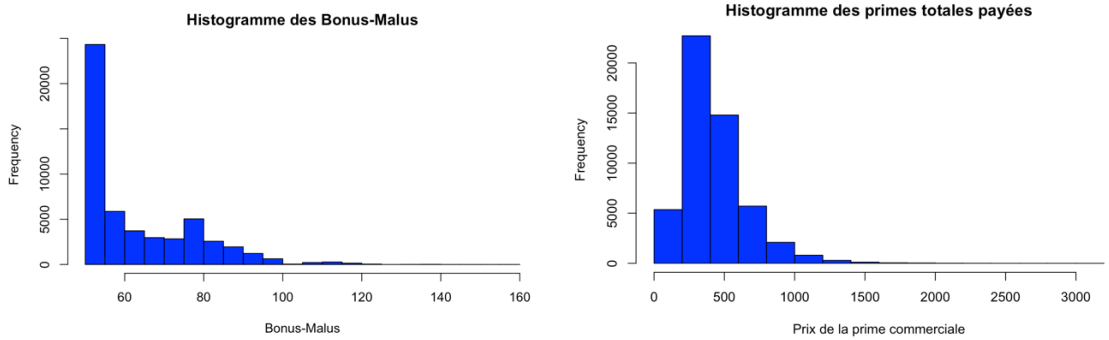



Les garanties qui rapportent le plus d'argent dans notre échantillon sont les garanties « TPL » et « Windscreen » elles correspondent respectivement aux garanties liées à l'assurance au tiers et pare-brise.

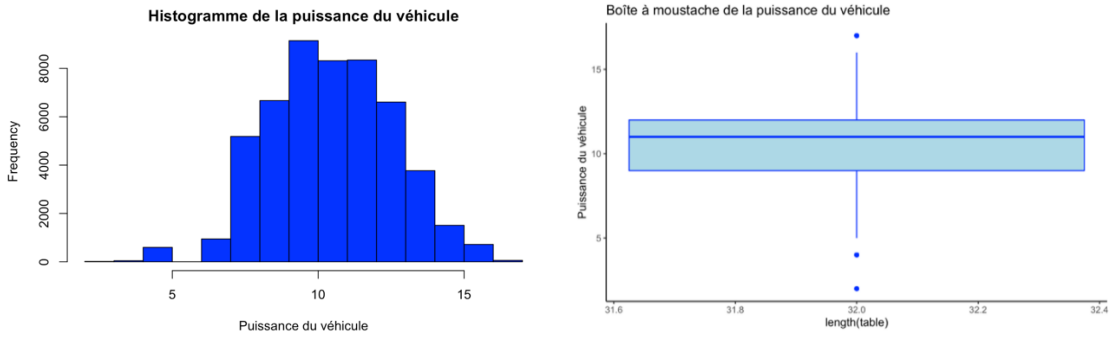
## *Variables quantitatives*



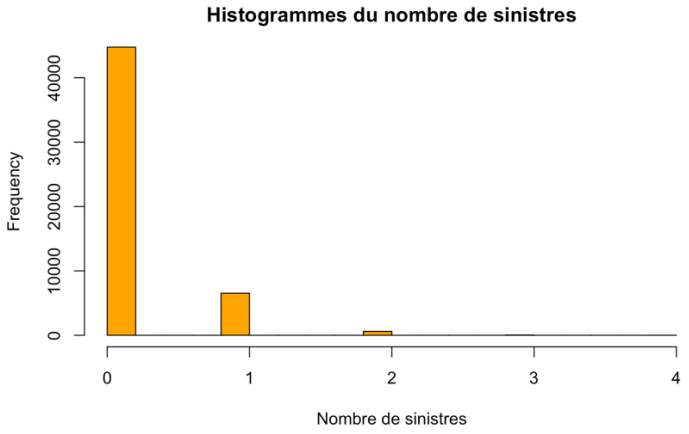

Les âges sont compris entre 18 et 97 ans, avec une moyenne à 40 ans, 50% des assurés de notre population ont entre 31 et 47 ans. L'âge du véhicule est en moyenne à 8 ans, la majorité des véhicules se situe entre 4 et 10 ans.

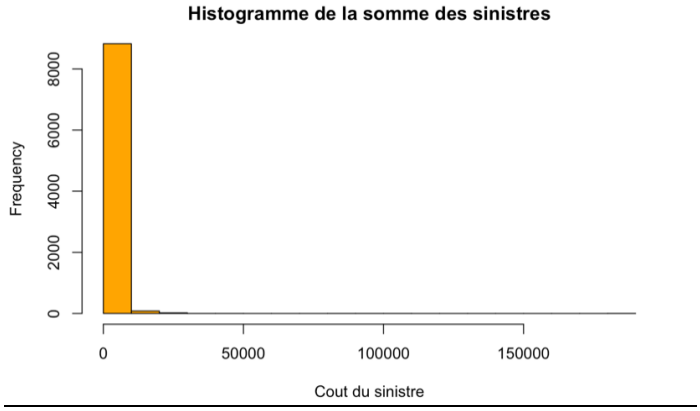



Une valeur de bonus-malus en dessous de 100 est un Bonus et une supérieur à 100 un malus. On peut voir que la majorité de nos assurés n'ont pas forcément de malus, avec une moyenne étant à 63 < 100. Il n'y a que des très peu d'individus qui possède un malus. La prime totale est en moyenne de 428 par assuré, ces derniers paient en général entre 269 et 530€.




On peut voir que la puissance du véhicule à l'allure d'une loi normale centrée en 10. 50% de notre échantillon a un véhicule possédant une puissance entre 9 et 12, à savoir que l'ensemble des voitures ont des puissances comprises entre 2 à 17.




| | Minimum | 1er quartile | Médiane | Moyenne | 3ème quartile | Maximum |
|---|---|---|---|---|---|---|
| Nombre sinistres | 0 | 0 | 0 | 0,1 | 0 | 4 |
| Coût des sinistres | 2 | 225 | 453 | 1173 | 997 | 185750 |

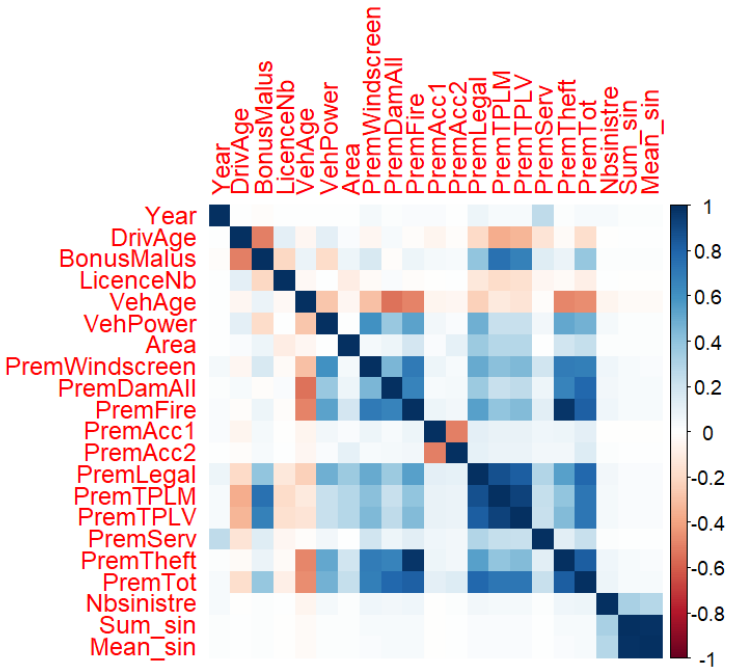
La variable Nombre de sinistre nous montre que nos assurés n'ont en grande majorité pas de sinistre, il y a moins de 10 000 assurés qui en ont 1 (sur plus de 40 000 qui en ont 0). Très peu d'individus ont entre 2 et 4 sinistres.

Les ¾ des sinistres coûtent moins de 1 000 €. Il y a cependant quelques sinistres très couteux, qui font augmenter la moyenne, qui est bien au-dessus de la médiane.

## III) Étude bivariée

L'étude bivariée consiste à étudier les relations entre différentes variables. On parle ici de 2 variables qualitatives, 2 quantitatives ou une qualitative et une quantitative. Notre étude n'a pas forcément montré de corrélations immédiates nous permettant de conclure quant à une liaison évidente entre 2 variables, pouvant expliquer les sinistres ou les primes payé par les assurés. Nous avons voulu nous concentrer sur le nombre de sinistres et le coût moyen des sinistres.

### *Matrice de corrélation*

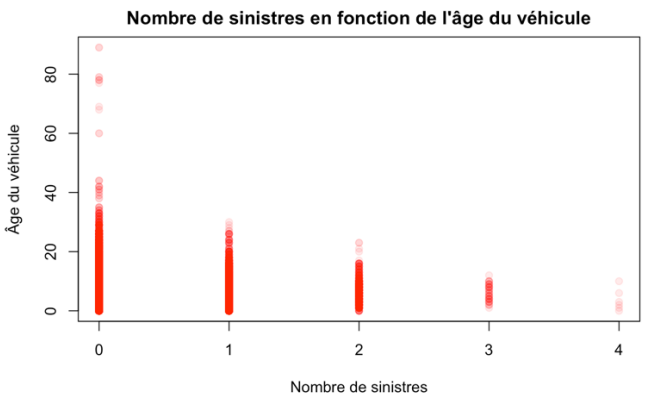



Nous observons que à première vue, nous n'avons pas de corrélation linéaire entre les variables, hormis entre les primes, et entre le total des coûts et la moyenne des coûts de sinistres par personne.

## *En fonction du nombre de sinistres*

Cette variable nous permet d'étudier le nombre de sinistres par individus, nous allons l'étudier en fonction de plusieurs variables pour chercher un lien entre celles-ci.

- Nombre de sinistres en fonction de l'âge du véhicule et âge du conducteur





| P-value | |
|---|---|
| Pearson | 2.2e-16 |
| Kendall | 2.2e-16 |
| Spearman | 2.2e-16 |

Pour le nombre de sinistre en fonction de l'âge du véhicule, nous obtenons la même p-value = 2.2e-16 pour les 3 tests (Pearson, Kendall et Spearman), on ne rejette pas H0 ce qui signifie que les variables nombre du sinistre et Age du véhicule semblent être corrélés à première vue.

Le graphique de gauche nous montre les voitures ayant eu des sinistres (entre 1 et 2 principalement) sont des voitures qui ont au maximum 20 ans. Les voitures n'ayant eu aucuns sinistres sont également les voitures très anciennes (plus de 40 ans). Les voitures ayant eu le plus sinistres (maximum, c'est-à-dire 4) sont aussi des voitures assez jeunes (moins de 10 ans).

| P-value | |
|---|---|
| Pearson | 0,05527 |
| Kendall | 0,2223 |
| Spearman | 0,2225 |

Pour le nombre de sinistre en fonction de l'âge du conducteur, on obtient une p-value qui nous permet de rejeter l'hypothèse H0, à savoir qu'il existe un lien entre les 2 variables concernées.

Le graphique de droite nous montre que les personnes ayant des sinistres ont un âge étalé par rapport à l'ensemble des âges présents dans notre échantillon.

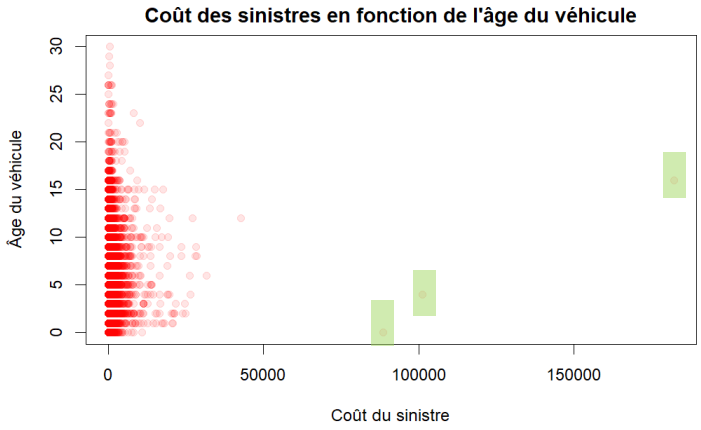
⇒ Nombre de sinistre en fonction des autres variables qualitatives : voir en annexe.

En étudiant la fréquence des sinistres en fonction des autres variables, nous n'avons pas vu de tendance particulière.

## *En fonction de la somme des sinistres*

- Somme des sinistres en fonction de l'âge du véhicule

| P-value | |
|---|---|
| Pearson | 3.45e-8 |
| Kendall | 2.2e-16 |
| Spearman | 2.2e-16 |

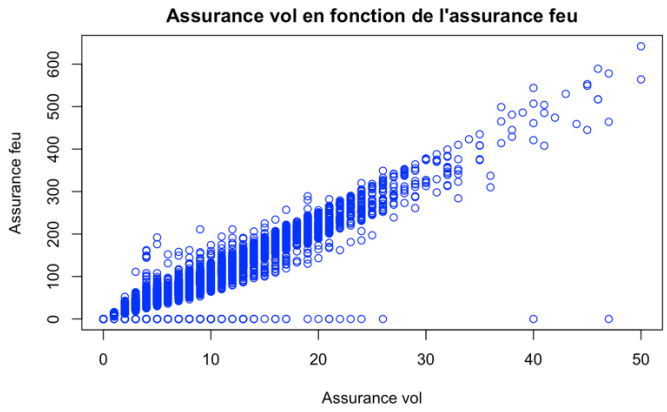


Même interprétation que pour le coût des sinistres en fonction de l'âge du véhicule, on trouve des coefficients de corrélation qui nous permettent de ne pas rejeter l'hypothèse nulle. Cela nous permet de penser qu'il existe une relation entre le coût des sinistres et l'âge du véhicule.

Le graphique à droite ne nous permet pas de conclure quant à une potentielle relation entre les 2 variables. On observe le fait qu'il y a beaucoup de sinistres peu chers, qui représente la grande majorité des indemnisations. Néanmoins, on note la présence de sinistres très couteux ce qui « allonge » notre graphique, comme on peut le voir en vert.

## *En fonction des primes*

On note une forte corrélation entre les variables « PremTheft » et « PremFire », ayant un coefficient de corrélation de 0,98 (très proche de 1).




Nous avons une relation pratiquement linéaire. Cela signifie que les variables liées à l'assurance du vol et du feu sont corrélées. Nous pouvons interpréter cela en pensant que ces deux garanties sont souvent vendues ensemble.

## *Étude année 2003 et 2004*

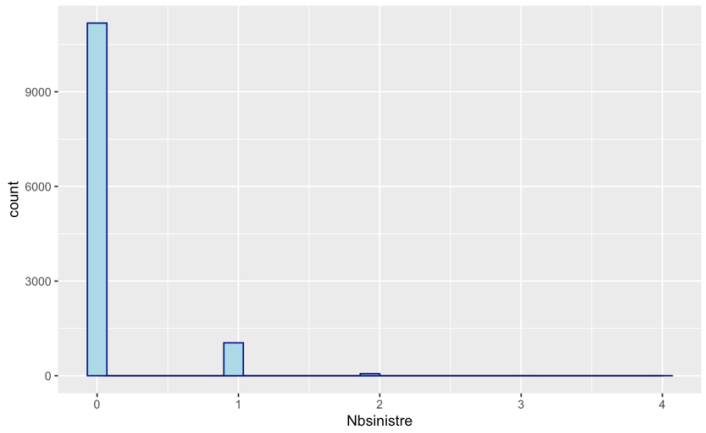
Dans le cadre de cette étude, on sépare les deux années de notre base de données. On s'intéresse aux raisons pour lesquelles les assurés ont résiliés ou sont arrivés. On s'interroge sur les différences entre ces deux années.

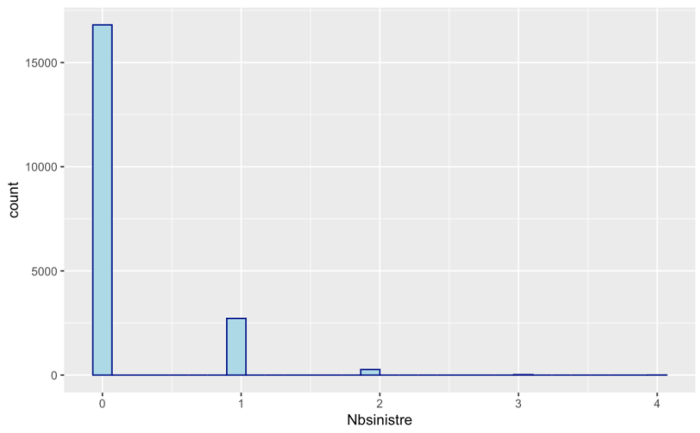
| Années | Nombres d'assurés |
|---|---|
| 2003 | 32 114 |
| 2004 | 19 829 |

On note ici une différence de 12 285 assurés : l'assureur a perdu des assurés entre les 2 années.

Nous avons tenté de trouver un profil type des résiliés, en comparant les caractéristiques de ces clients avec ceux non résilié. Nous n'avons noté aucune différence notable hormis la fréquence des sinistres.

Titres des 2 graphiques : Nombre de sinistre de la table 2004, résiliés (à gauche) ; et non résiliés (à droite).



Nous pouvons cependant noter que les assurés résiliés ont en moyenne moins de sinistre que les non résiliés (0,09 contre 0,16 en fréquence), même si la répartition est similaire (histogrammes ci-dessus, avec les résiliés à gauche et les non résilié à droite).

| Années | Coût total des sinistres | Cout moyen des sinistres | Fréquence des sinistres | Chiffre d'affaires annuel |
|---|---|---|---|---|
| 2003 | 4 874 254 | 1216 | 0.14 | 8 777 449 |
| 2004 | 3 891 219 | 1296 | 0.17 | 4 889 268 |

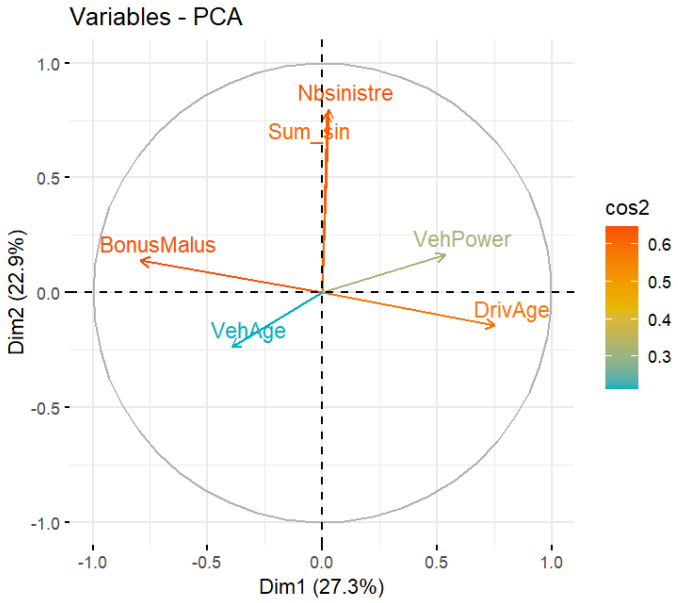
Premièrement, le coût total des sinistres en 2003 est plus important qu'en 2004. On note la présence davantage d'assurés en 2003 qu'en 2004, donc nous nous intéresserons plutôt à la moyenne des coûts de sinistres et de la fréquence des sinistres par personne. Nous pouvons constater qu'ils sont plus élevés en 2004 qu'en 2003.

En calculant le chiffre d'affaires de l'assurance en 2003 et 2004 (prime-sinistre), nous observons que l'assurance à gagné moins d'argent entre 2003 et 2004 : 8 777 449 VS 4 889 268 €.

L'assureur a perdu 38% de ses assurés et 44% de son chiffre d'affaires.

## IV) Étude multivariée

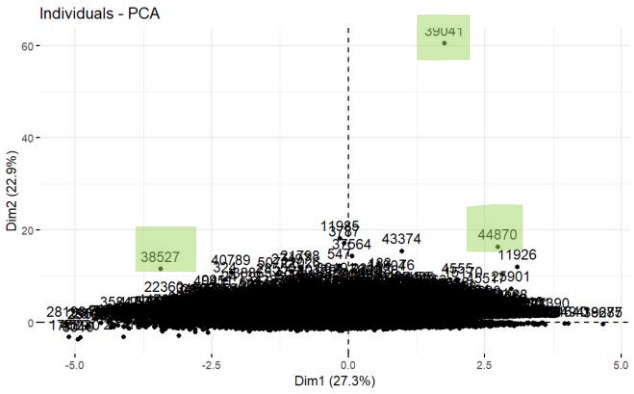
Nous avons réalisé une Analyse en Composante Principale (ACP) avec les variables : âge du conducteur, bonus-malus, âge/puissance du véhicule, coût et fréquence de sinistre.




Hormis l'âge du véhicule, toutes les variables sont bien représentées. En observant le cercle de corrélation, nous observons que l'axe 1 oppose bonus-malus et âge du véhicule avec respectivement âge du conducteur et puissance du véhicule.

Ainsi nous pouvons en déduire que les conducteurs les plus jeunes ont un véhicule peu puissant et plus ancien et beaucoup de bonus-malus (qui correspond à des malus), tandis que les plus âgés ont des véhicules plus puissant et plus récent et plus de bonus.

Les variables de fréquence et coût de sinistres contribuent beaucoup à l'axe 2 et sont du même côté. Ce qui est cohérent car plus il y a de sinistres, plus le coût augmente.

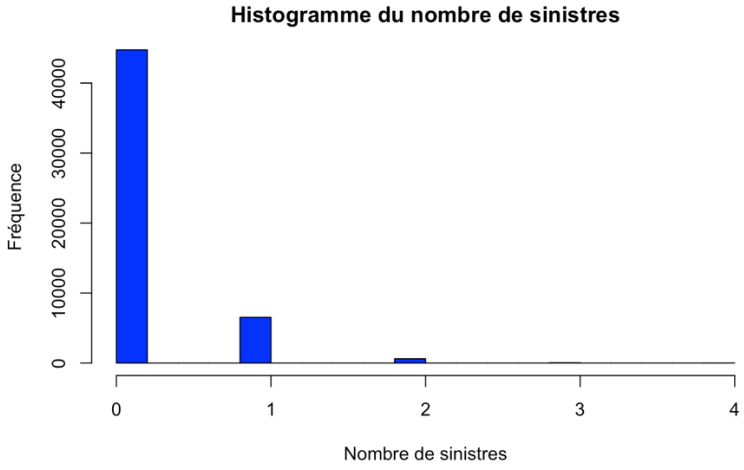



Ayant énormément d'individus, le graphique individu n'est pas lisible. Nous avons tout de même étudié 3 individus qui ressortaient, ce qui a confirmé nos analyses du graphique précédent.

- **39041** (situé en haut) : Il s'agit en effet d'un individu ayant eu plusieurs sinistres à des coûts importants.
- **38527** (situé à gauche) : Il s'agit d'un jeune assuré (20 ans), il a beaucoup de malus (120), un véhicule ancien (6 ans) et une puissance moyenne (8).
- **44870** (à droite) : Il s'agit d'un assuré de 60 ans, il a un véhicule qui âgé de 3 ans avec une puissance de 10. Il a 64 de bonus.

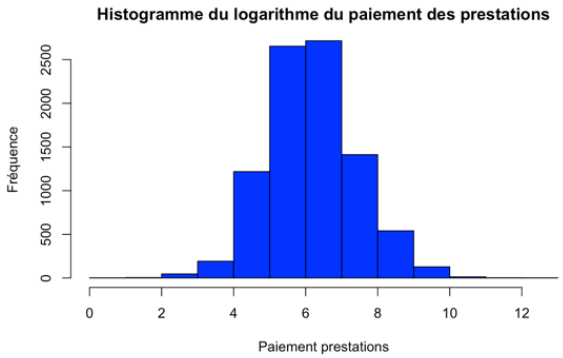
# V) Tarification et proposition prime

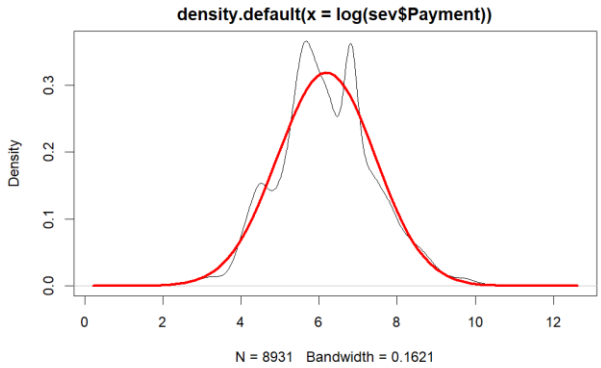
## *Estimation de la fréquence*

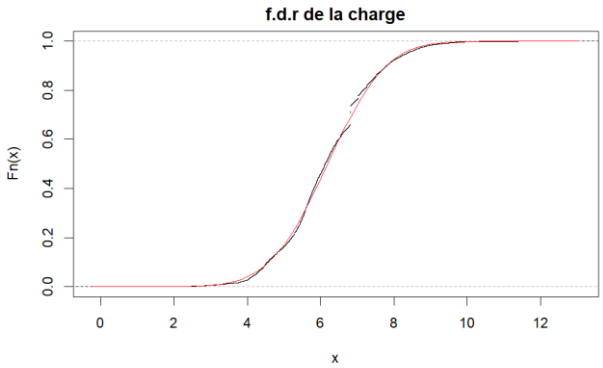



La moyenne et la variance de la fréquence des sinistres sont similaires (0,1521996 contre 0,1591304). En effectuant le Chi-test, nous ne rejetons pas l'hypothèse que la fréquence suit une loi de poisson. Nous estimons ainsi une fréquence de 0,16 avec la fonction densité de la loi poisson.

## *Estimation du coût des sinistres*






Le coût des sinistres semble suivre une loi log normal, d'après l'histogramme. En superposant les fonctions de répartition et de densité de notre variable avec celles de la loi log normal, nous observons qu'elles se superposent plutôt bien. Nous pouvons ainsi supposer que les coûts de sinistres suivent une loi log normal.

Nous estimons ainsi le coût moyen à 1065 €.

### *Prime pure*

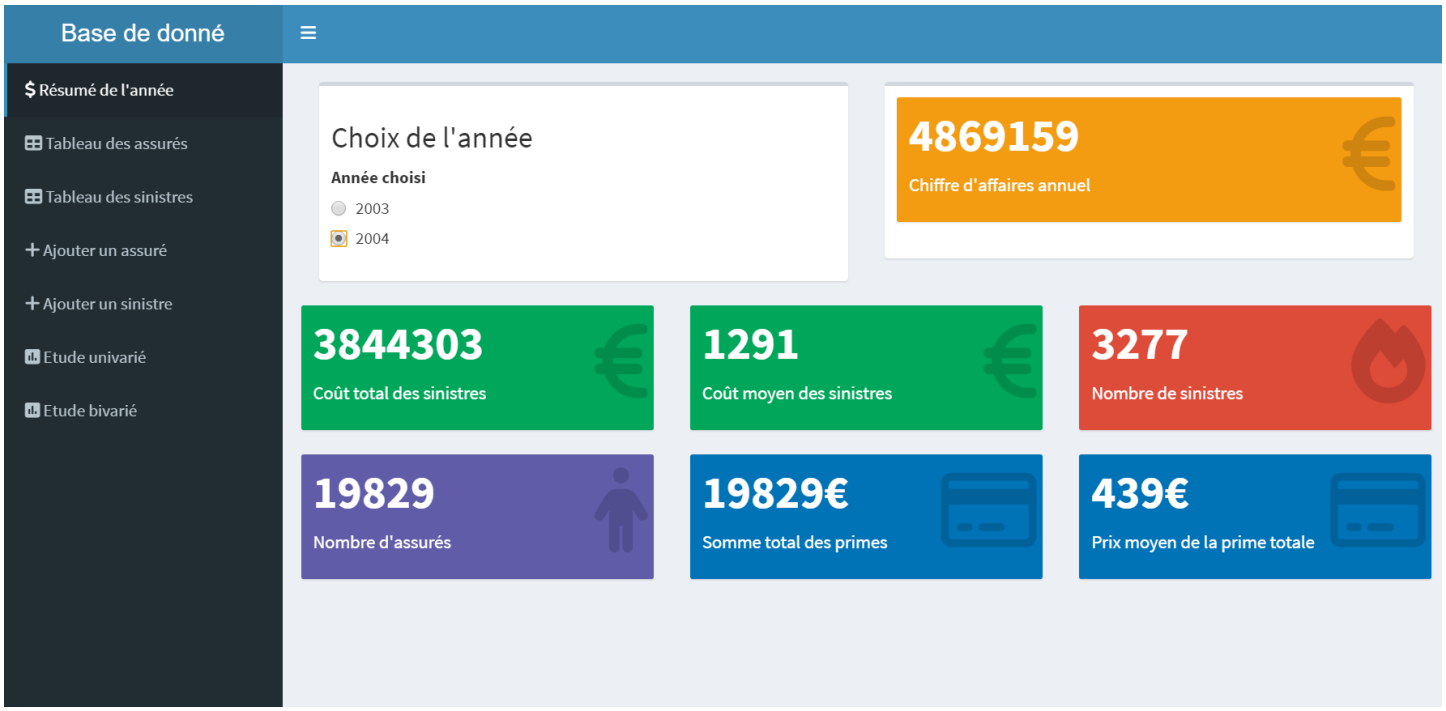
En multipliant la fréquence et le coût moyen estimé, nous obtenons une prime pure de 175 €, qui est légèrement inférieur à la prime pure empirique (179 €).

### *Chargement de sécurité*

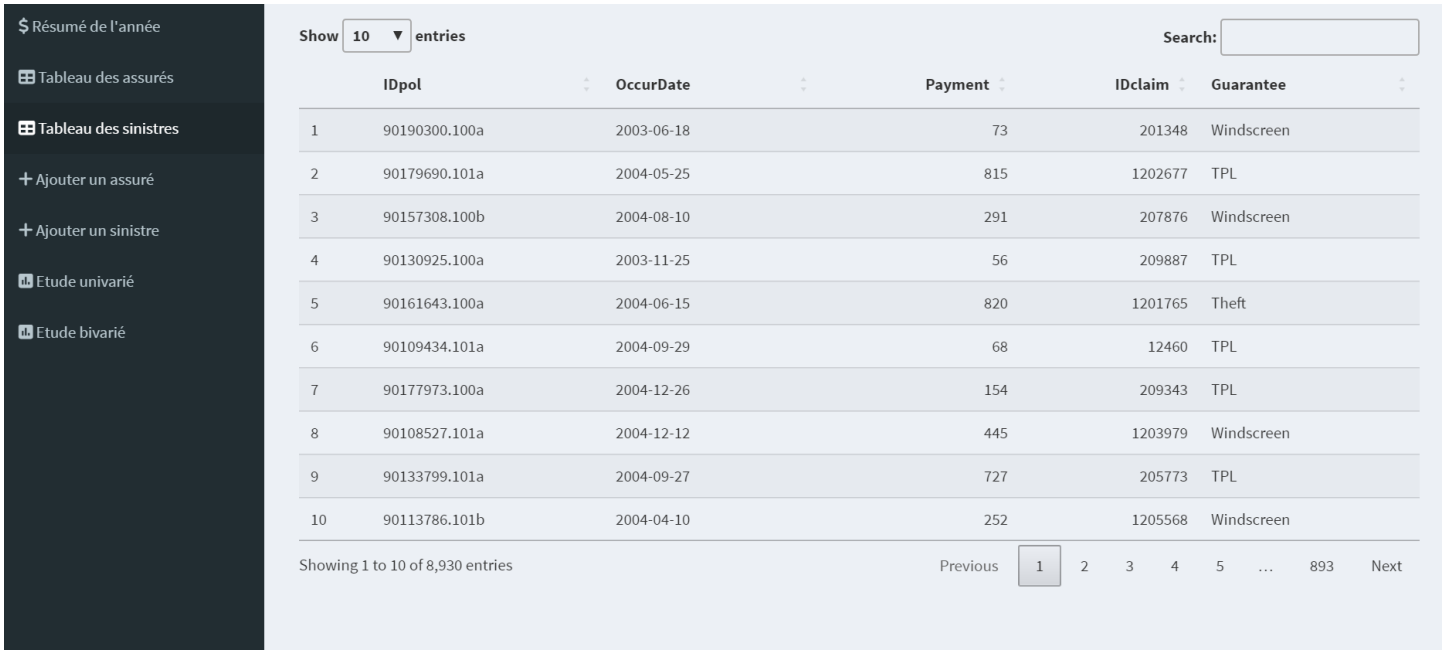
Afin de permettre à l'assurance de pouvoir payer tous les sinistres des assurés sans craindre de se ruiner, nous rajoutons une charge de sécurité. Cette charge correspond à 30% de la prime pure.

## VI) Rshiny

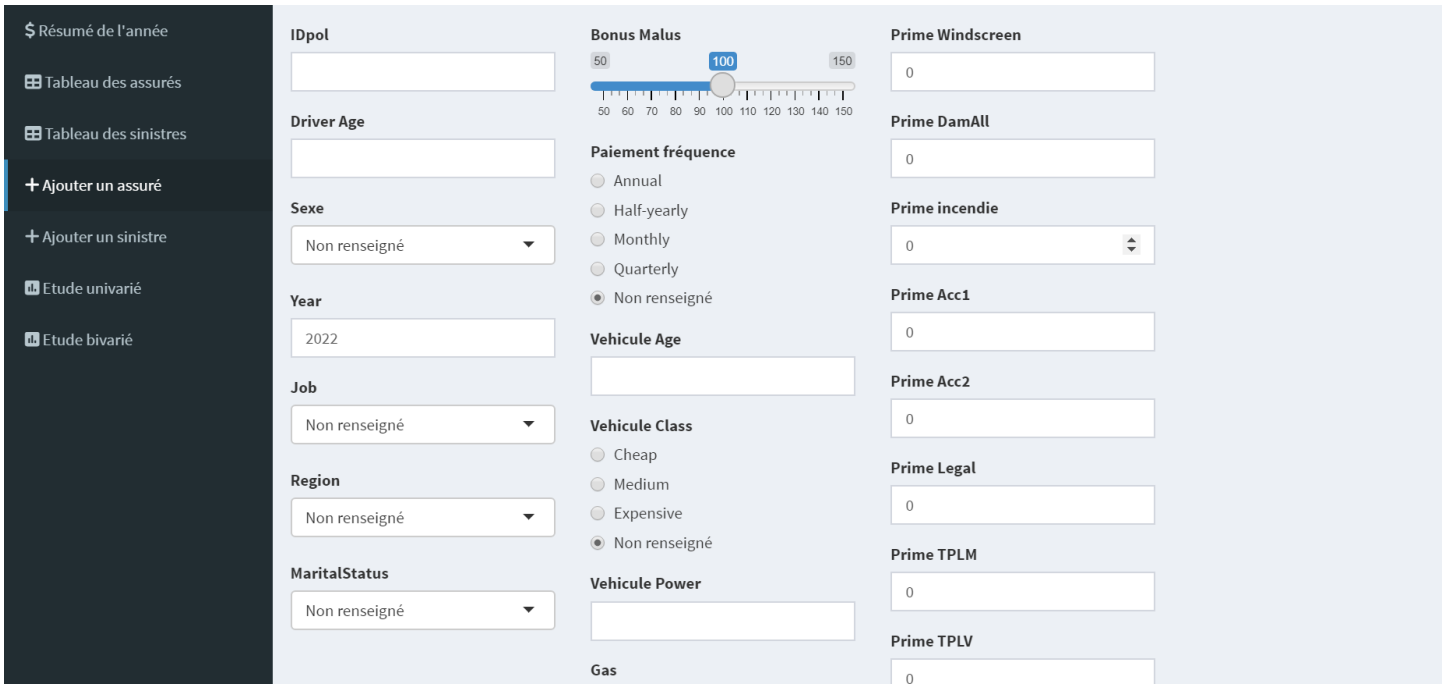
Pour permettre à l'utilisateur d'avoir une interface facile à utiliser et plus agréable que Rstudio, nous avons utilisé Rshiny. Nous avons créé un dashboard composé de plusieurs onglets.



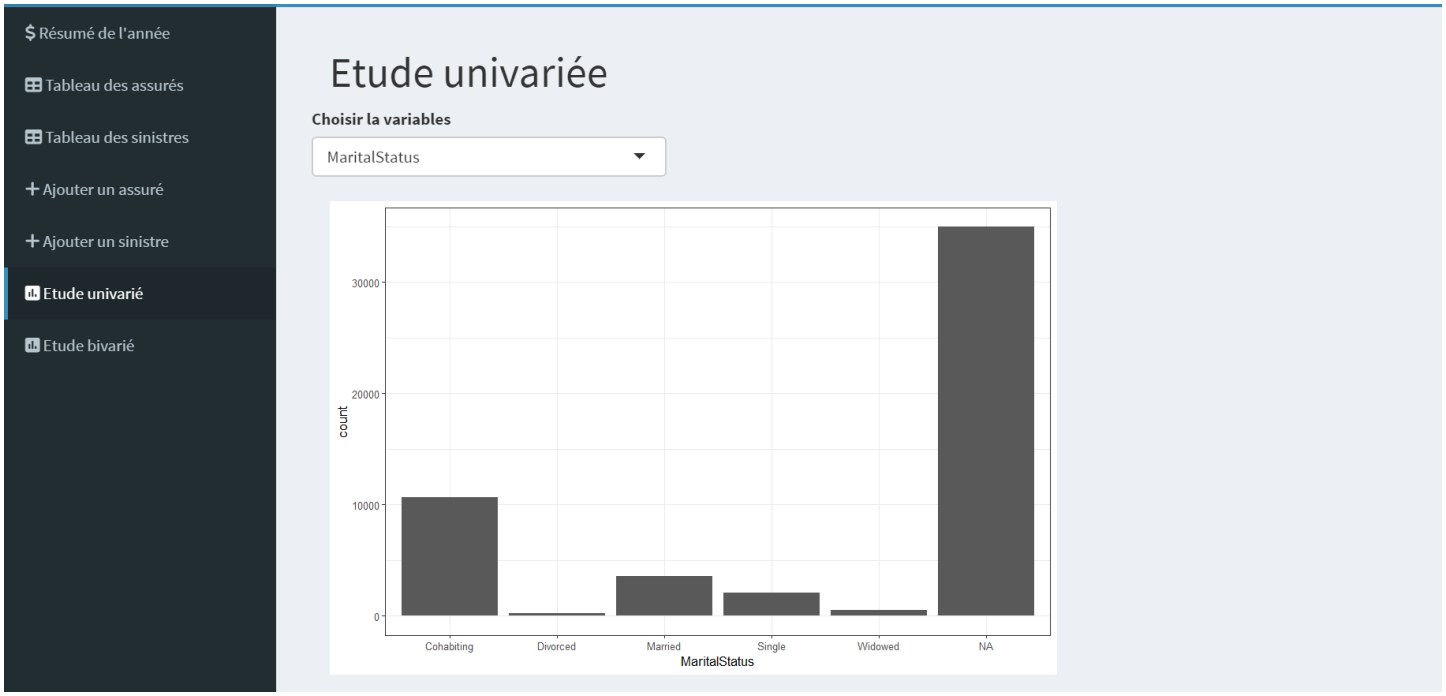
L'utilisateur peut tout d'abord avoir une vision globale de ses dépenses et entrées d'argent, avec pour chaque année le chiffre d'affaires, le nombre d'assuré, le nombre/coût des sinistres et d'autres informations.




L'utilisateur peut également avoir accès au tableau avec tous les assurés et les sinistres, ainsi que leurs caractéristiques. Il peut les chercher facilement grâce à une barre de recherche.




L'utilisateur peut également ajouter un nouvel assuré ou un sinistre. Ce nouvel assuré ou sinistre apparaîtra directement dans la table.




Il peut également obtenir les histogrammes de toutes les variables, ainsi que les nuages de points entre les variables qu'il souhaite.

*Remarque sur R shiny* : N'ayant jamais utilisé ce package, nous avons dû d'abord apprendre à l'utiliser. Grâce aux nombreux tutos internet, nous avons réussi à utiliser plusieurs de ses fonctionnalités.

Nous sommes cependant conscients que le résultat n'est pas parfait, par exemple l'utilisateur peut ajouter un assuré qui existe déjà, ou rentrer un sinistre d'une personne non assurée. De plus, nous ne pouvons ni modifier ni supprimer un assuré, ce qui aurait pu être intéressant.

Nous sommes tout de même satisfaits du résultat, qui est agréable à utiliser, et qui permet de faire de nombreuses choses.

## *Conclusion*

Cette étude nous a permis d'étudier un échantillon d'assurés du point de vue de l'assureur. Nous voulions dans un premier temps comprendre comment la prime commerciale était calculée en cherchant une certaine segmentation dans notre base de données en fonction du profil de nos assurés.

Nous n'avons pas forcément trouvé de variables corrélées qui permettrait de faire cette segmentation et de proposer des tarifs préférentiels en fonction des caractéristiques de notre client.

C'est la raison pour laquelle nous nous sommes intéressés aux différences entre les années 2003 et 2004. C'est en observant les fluctuations entre ces deux années que nous avons noté la présence d'un certain nombre de résiliations. La fréquence des sinistres et les coûts des sinistres ont été plus nombreux en 2004. Ainsi l'assureur a perdu 38% de ses assurés et 44% de son chiffre d'affaires.

Nous avons également réalisé une ACP qui nous a appris quelques généralités sur les assurés (comme le lien entre l'âge et l'âge du véhicule).

Pour finir, nous voulions proposer une interface facile et pratique à utiliser avec plusieurs fonctionnalités. C'est Rshiny qui nous a permis de construire cette interface et d'automatiser notre code sur R studio. Nous avons ainsi une interface qui permet à l'utilisateur de visualiser les chiffres clés de son business, de chercher un assuré, d'inscrire de nouveaux assurés et sinistres dans la base de données.

# Annexe

## Présentation de la base de données

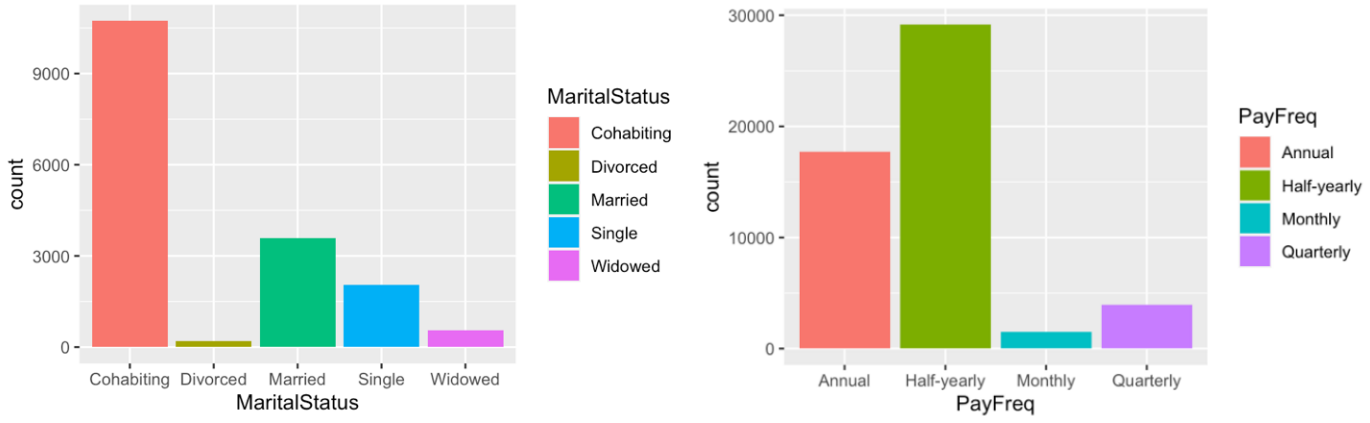
En termes de variables nous avons :

| Base de données | Variables quantitatives | Variables qualitatives |
|---|---|---|
| *Prem* | | IDpol |
| *Prem* | Year | |
| *Prem* | DrivAge | |
| *Prem* | | DrivGender |
| *Prem* | | MaritalStatus |
| *Prem* | BonusMalus | |
| *Prem* | LicenceNb | |
| *Prem* | | PayFreq |
| *Prem* | | JobCode |
| *Prem* | VehAge | |
| *Prem* | | VehClass |
| *Prem* | | VehPower |
| *Prem* | | VehGas |
| *Prem* | | VehUsage |
| *Prem* | | Garage |
| *Prem* | | Area |
| *Prem* | | Region |
| *Prem* | | Channel |
| *Prem* | | Marketing |
| *Prem* | PremWindscreen, PremDamAll, PremFire, PremAcc1, PremAcc2, PremLegal, PremTPLM, PremTPLV, PremServ, PremTheft, PremTot | |
| *Sev* | | IDpol |
| *Sev* | OccurDate | |
| *Sev* | | Payment |
| *Sev* | IDclaim | |
| *Sev* | | Guarantee |
| *Freq* | | IDpol |
| *Freq* | Year | |
| *Freq* | Damage, Fire, Other, Theft, TPL, Windscreen | |

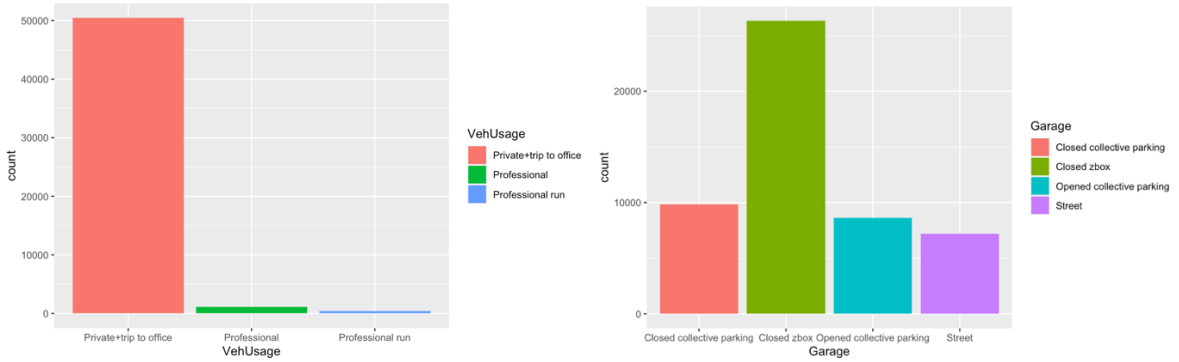
## Format des variables

Nous avons mis la variable OccurDate dans le bon format date et avons créé une colonne année dans *sev* avec l'année du sinistre. Nous avons également mis la plupart de nos variables qui était quantitatives au format *numeric.*

## Complément de l'analyse univariée



Pour ceux dont nous connaissons leur situation maritale, on note une grande majorité en situation de cohabitation. Notre échantillon d'assurés paie leurs primes 2 fois par an ou annuellement pour la plupart.

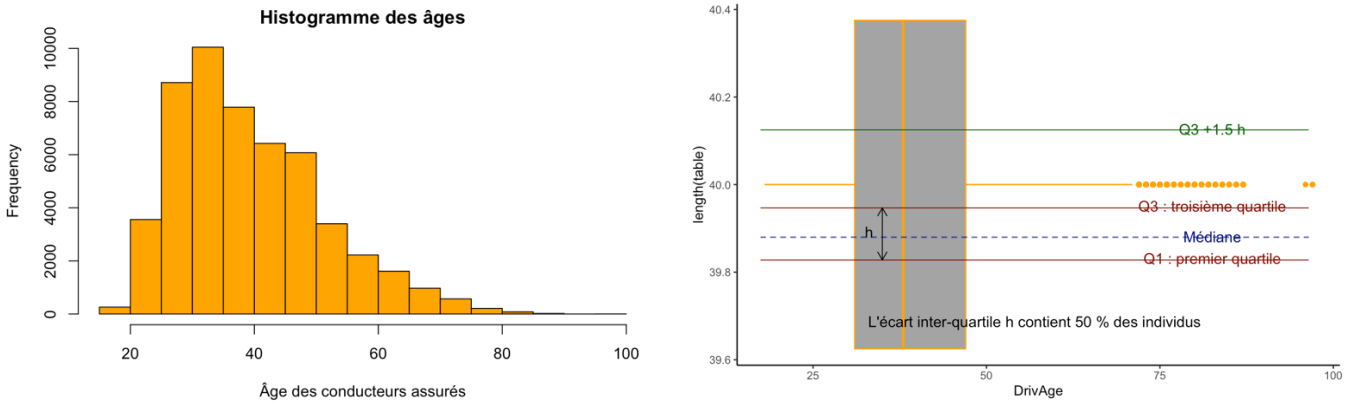



L'immense majorité des voitures sont utilisés dans le cadre privé ainsi que pour se rendre au travail. Les voitures sont généralement garées dans un parking clos (plus de 3/5 des voitures).

## *Variables quantitatives*

Pour cette partie, on peut étudier les moyennes, écart-type, quartiles, boîtes à moustaches etc.

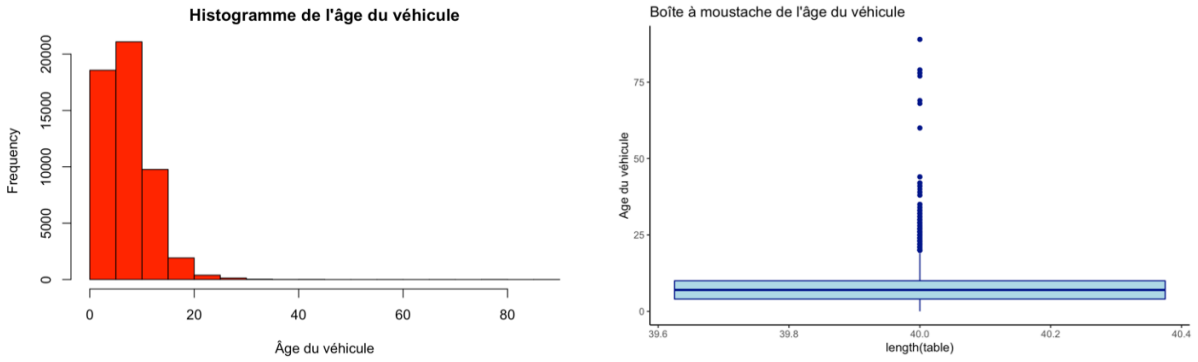
- **VARIABLE DRIVAGE**




| Minimum | 1er quartile | Médiane | Moyenne | 3ème quartile | Maximum |
|---|---|---|---|---|---|
| 18 | 31 | 38 | 40 | 47 | 97 |

Les âges sont compris entre 18 et 97 ans, avec une moyenne à 40 ans. Le graphique à gauche nous montre que 50% de notre population se trouve entre 31 et 47 ans.

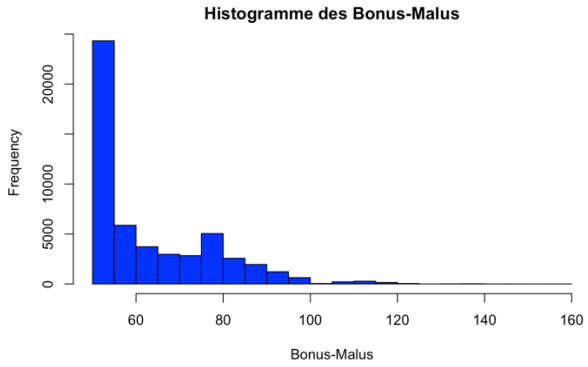
- **VARIABLE VEHAGE**

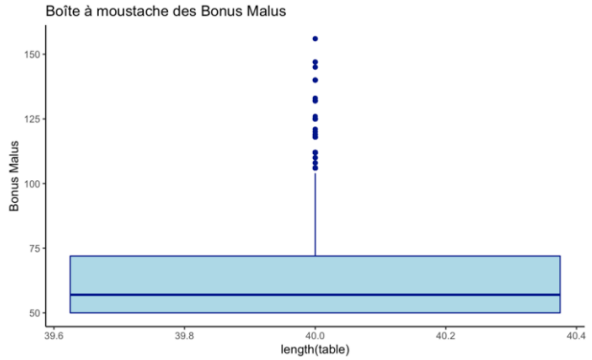



| Minimum | 1er quartile | Médiane | Moyenne | 3ème quartile | Maximum |
|---|---|---|---|---|---|
| 0 | 4 | 7 | 8 | 10 | 89 |

L'Âge du véhicule est en moyenne à 8 ans, 50% de notre échantillon du véhicule se situe entre 4 et 10 ans.

- **VARIABLE BONUS-MALUS**

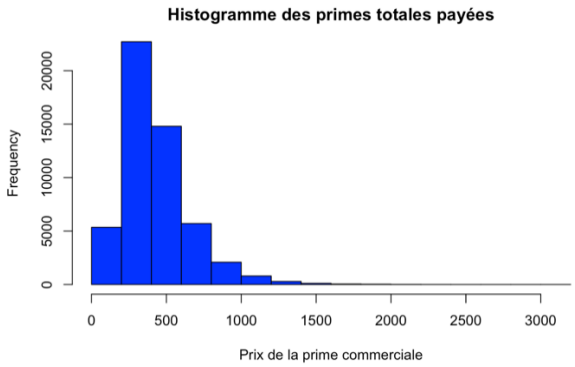


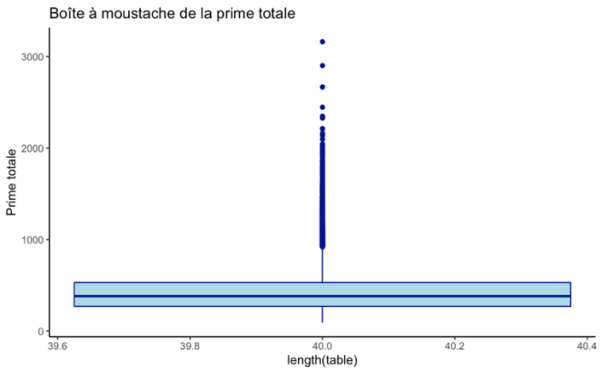




| Minimum | 1er quartile | Médiane | Moyenne | 3ème quartile | Maximum |
|---|---|---|---|---|---|
| 50 | 50 | 57 | 63 | 72 | 156 |

En dessous de 100 on considère que c'est un Bonus et supérieur à 100, un malus. On peut voir que la majorité de nos assurés n'ont pas forcément de malus, avec une moyenne étant à 63 < 100. Il n'y a que des très peu d'individus ayant un malus.

- **VARIABLE PRIME TOTALE**

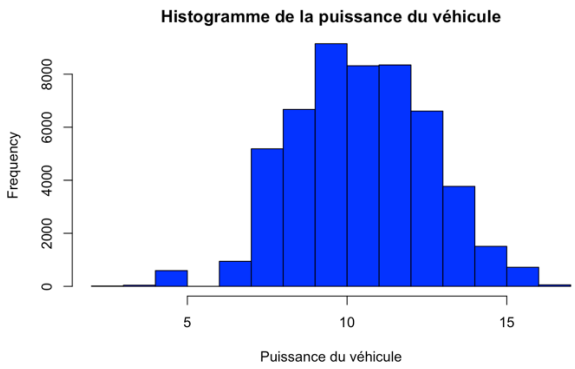


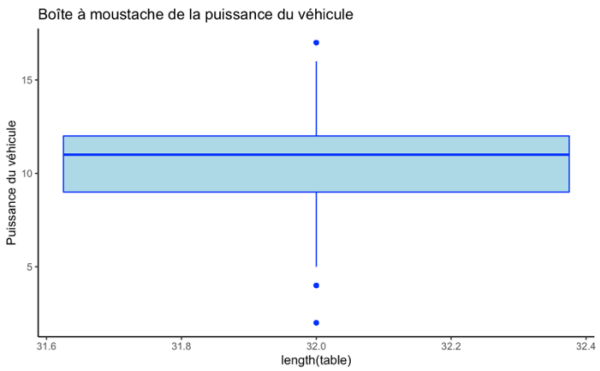




| Minimum | 1er quartile | Médiane | Moyenne | 3ème quartile | Maximum |
|---|---|---|---|---|---|
| 91 | 269 | 381 | 428 | 530 | 3163 |

La prime totale est en moyenne de 428 par assuré. Les assurés paient en général entre 269 et 530€.

- **VARIABLE VEHPOWER**



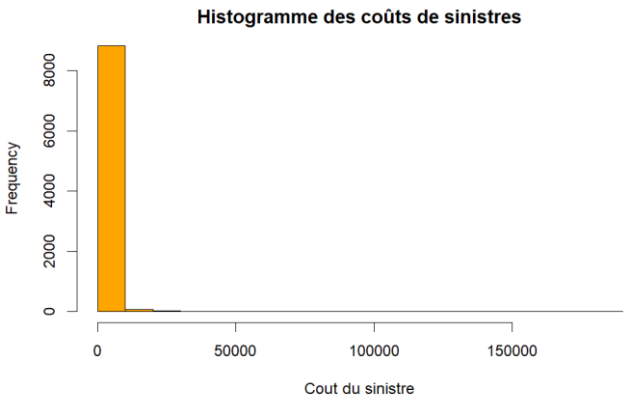




| Minimum | 1er quartile | Médiane | Moyenne | 3ème quartile | Maximum |
|---|---|---|---|---|---|
| 2 | 9 | 11 | 10,9 | 12 | 17 |

On peut voir que la puissance du véhicule à l'allure d'une loi normale centrée en 10. 50% de notre échantillon a un véhicule possédant une puissance entre 9 et 12, à savoir que l'ensemble des voitures ont des puissances comprises entre 2 à 17.

Nous avons créé 2 nouvelles variables : nombre de sinistres et somme du sinistre.

- **VARIABLE NOMBRE DE SINISTRES**

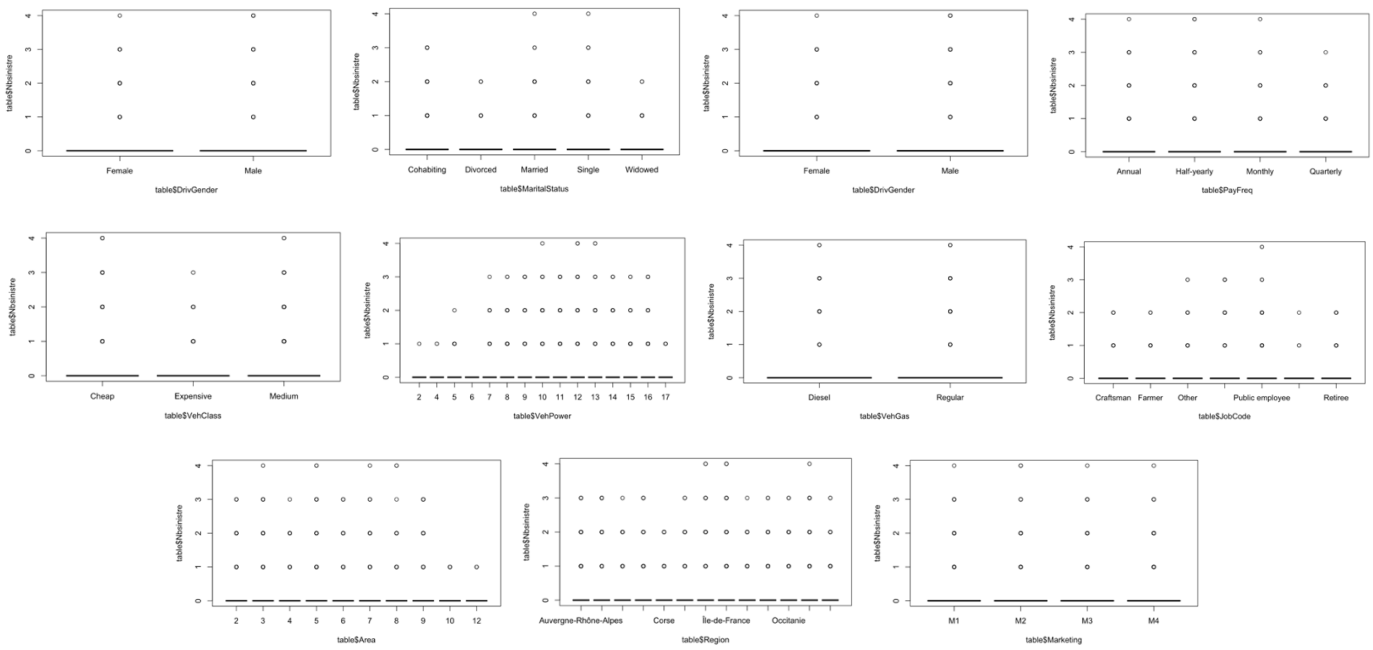

| | Minimum | 1er quartile | Médiane | Moyenne | 3ème quartile | Maximum |
|---|---|---|---|---|---|---|
| **Nombre sinistres** | 0 | 0 | 0 | 0,1 | 0 | 4 |
| **Somme du sinistre** | 0 | 0 | 0 | 151 | 0 | 182 108 |

Premièrement, la variable Nombre de sinistre nous montre que nos assurés ont en grande majorité pas de sinistre, il y a moins de 10 000 assurés qui en ont 1 (sur plus de 40 000 qui en ont 0). En effet, les quartiles nous montrent également que 50% de notre population se situe à 0 sinistre. La moyenne est moins d'un sinistre. Très peu d'individus ont entre 2 et 4 sinistres.

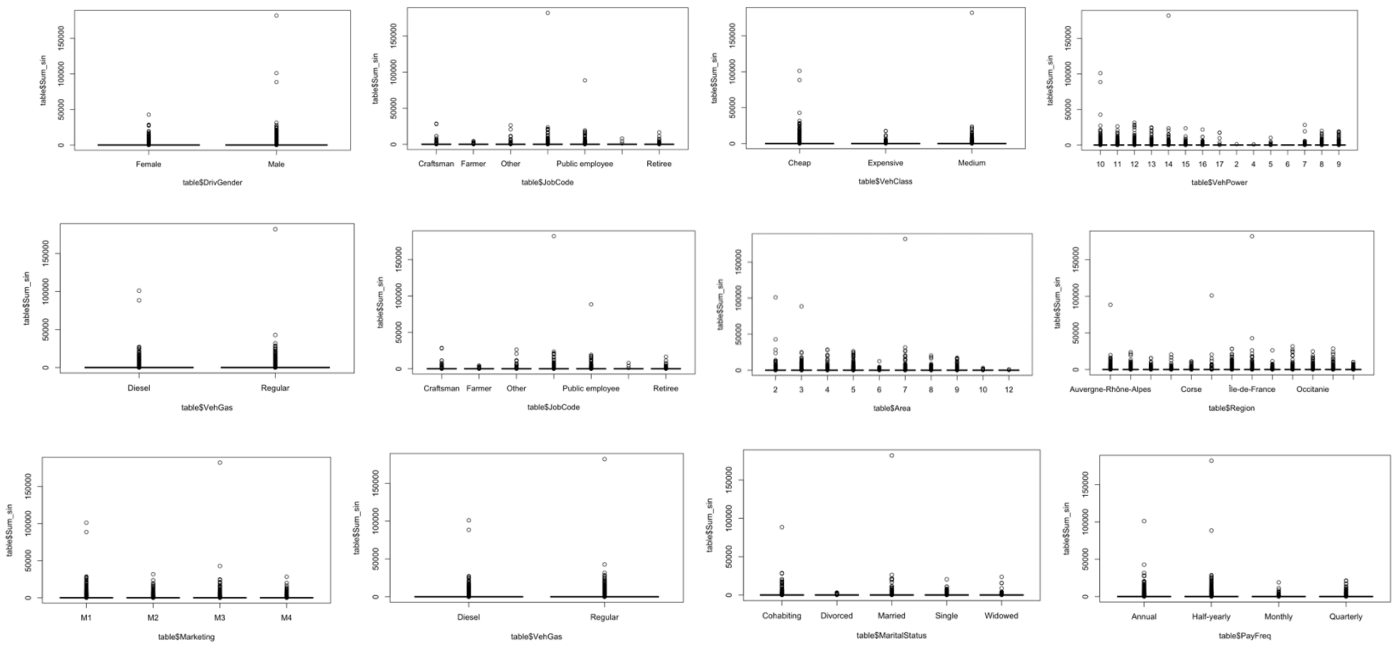
Deuxièmement, la variable somme des sinistres est en lien avec la variable nombre des sinistres dans le sens où l'immense majorité des évènements donne lieu à « 0 », ce qui signifie qu'aucun paiement est effectué et qu'il n'y a pas eu de sinistres.

⇒ Nombre de sinistre en fonction des autres variables qualitatives :

Nous avons effectué des boxplot du nombre de sinistre en fonction des différentes variables qualitatives pour chercher une potentielle corrélation entre la variable target (nombre de sinistre) et les autres variables, nous n'avons pas vu de tendance particulière.

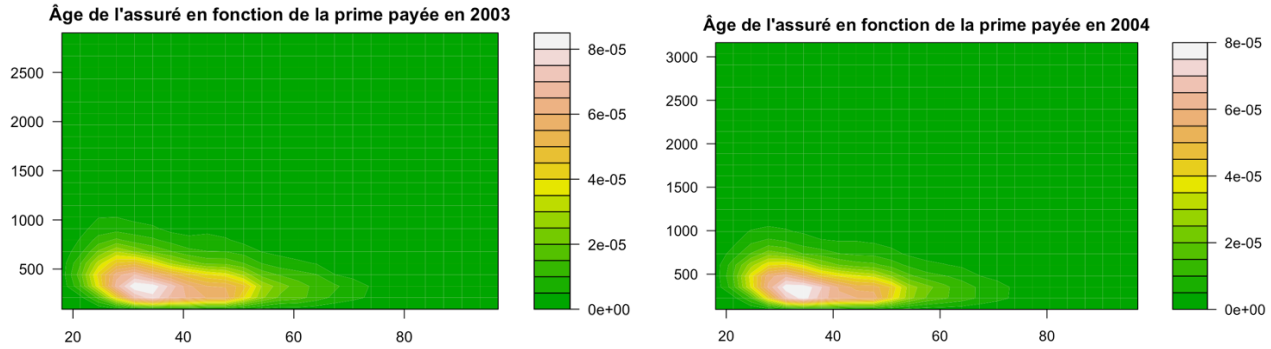


⇒ Somme des sinistres en fonction des autres variables qualitatives :

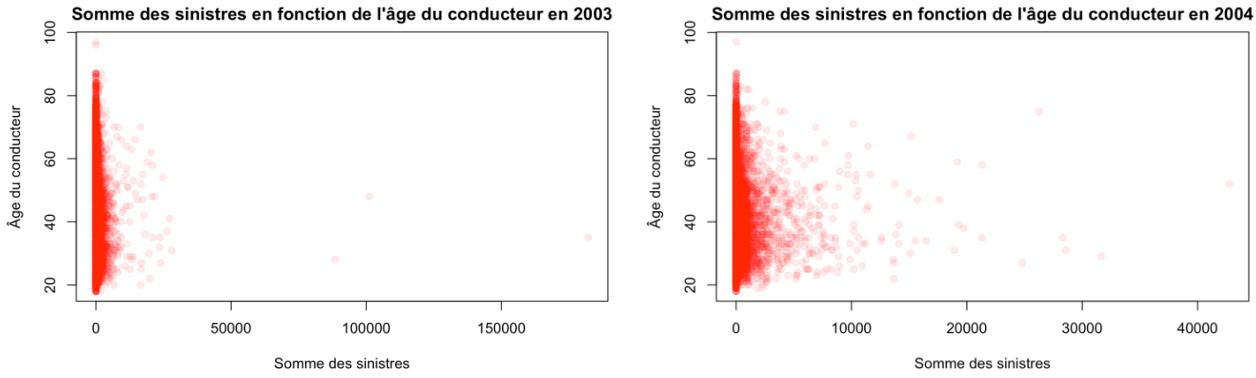



## *2003 VS 2004*

Nous cherchons à savoir pourquoi ces assurés ont résilié en se basant sur leurs profils.

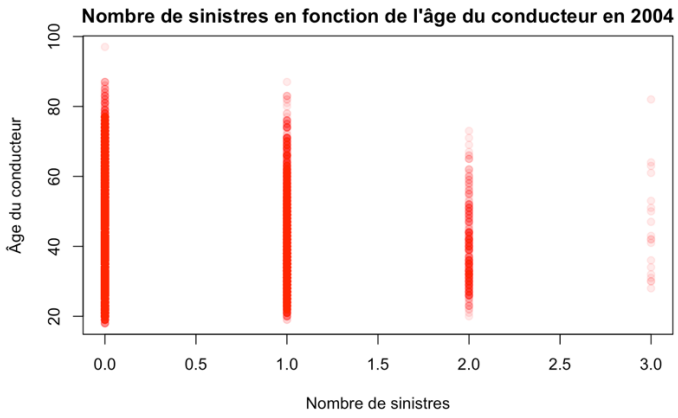



En fonction de la prime payée, nous ne notons pas de grosses différences entre les 2 années. En prenant en compte le départ de nombreux assurés, la prime reste relativement inchangée en fonction de l'âge.




Concernant les prestations de sinistres effectués, on voit tout d'abord à l'échelle que la somme des sinistres a diminuée, passant à une échelle avec des sinistres allant à plus de 150 000 € en 2003 à plus de 40 000 € en 2004.

Ainsi nous avons eu de moins « gros » sinistres en 2004 qu'en 2003.


Concernant le nombre de sinistres, on voit également que par rapport à 2003, 2004 enregistre un nombre de sinistre de 3 au maximum. Concernant la proportion du nombre 1 ou 2 (sinistres), les deux années semblent être plutôt proche.

Nous avons essayé de comparer les caractéristiques des assurés qui ont résilié avec ceux qui était assurés en 2003, en isolant les assurés présents en 2003 mais pas en 2004 :

Voici nos résultats avec les différentes variables :

| **Bonus-Malus** | **Minimum** | **1er quartile** | **Médiane** | **Moyenne** | **3ème quartile** | **Maximum** |
|---|---|---|---|---|---|---|
| Non résiliés | 50 | 50 | 57 | 63 | 72 | 156 |
| Résiliés | 50 | 50 | 60 | 64 | 76 | 140 |
| **Âge** | **Minimum** | **1er quartile** | **Médiane** | **Moyenne** | **3ème quartile** | **Maximum** |
| Non résiliés | 18 | 31 | 38 | 39 | 47 | 97 |
| Résiliés | 18 | 30 | 37 | 40 | 47 | 96 |
| **Prime totale** | **Minimum** | **1er quartile** | **Médiane** | **Moyenne** | **3ème quartile** | **Maximum** |
| Non résiliés | 95 | 264 | 375 | 420 | 520 | 2902 |
| Résiliés | 91 | 265 | 375 | 423 | 525 | 2208 |
| **Nombre sinistres** | **Minimum** | **1er quartile** | **Médiane** | **Moyenne** | **3ème quartile** | **Maximum** |
| Non résiliés | 0 | 0 | 0 | 0,16 | 0 | 4 |
| Résiliés | 0 | 0 | 0 | 0,09 | 0 | 4 |

Ainsi, les personnes ayant résiliés ont un profil très similaire à ceux qui n'ont pas résiliés. Nous aurions pu penser qu'il s'agirait de ceux payant le plus de prime, mais nous n'observons rien de semblable.

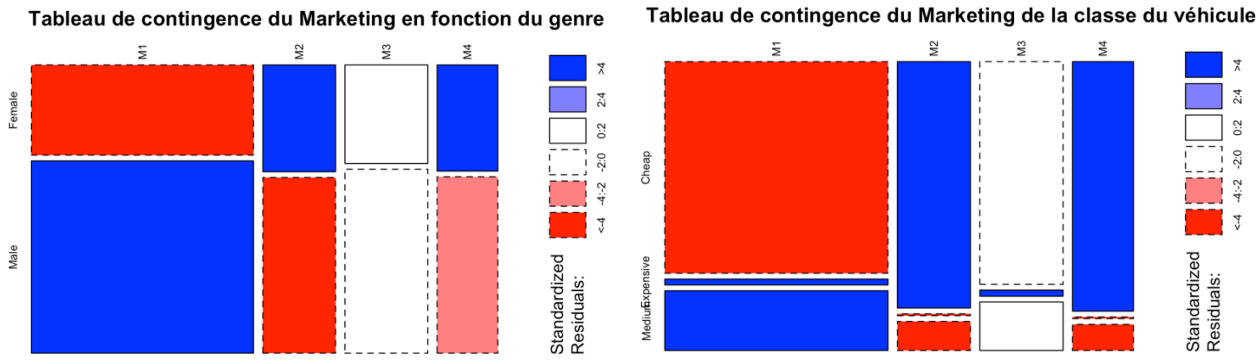
Nous pouvons cependant noter que ces derniers ont en moyenne moins de sinistre que les non résiliés (0,09 contre 0,16).

## Test $\chi^2$

Après avoir réalisé plusieurs tests de $\chi^2$, nous avons obtenu de bons résultats sur le Marketing utilisé. Nous avons pu faire un tableau de contingence avec 2 variables qualitatives et ainsi faire un diagramme en mosaïque. En effet, on trouve une p-value de 2.2e-16 (très petite), ce qui nous permet de rejeter l'hypothèse d'indépendance des lignes et des colonnes du tableau avec un seuil de 5%.

La surface des mosaïques est proportionnelle à l'effectif observé dans une cellule du tableau croisé. Ce type de représentation permet de voir la proportion d'une variable par rapport à l'effectif total mais également de croiser cette variable avec une autre et de voir les modalités qui compose cette dernière.

- Diagrammes en mosaïque des résidus standardisés sur un tableau qui croise la variable Marketing avec 2 variables qualitatives : le Genre et la Classe du véhicule

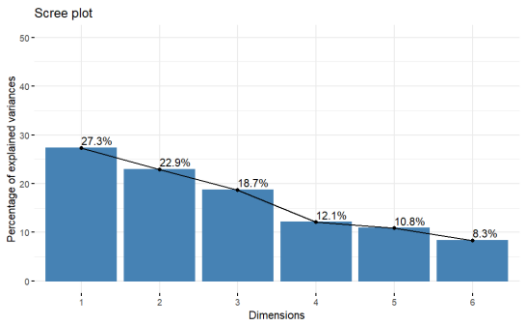



Pour les deux représentations graphiques ci-dessus, nous obtenons une p-value < 2.2e-16, ce qui nous permet de rejeter l'hypothèse d'indépendance des lignes et des colonnes des tableaux. C'est la raison qui nous a amené à faire une représentation graphique du tableau croisé entre Marketing et le genre ainsi que la classe du véhicule.

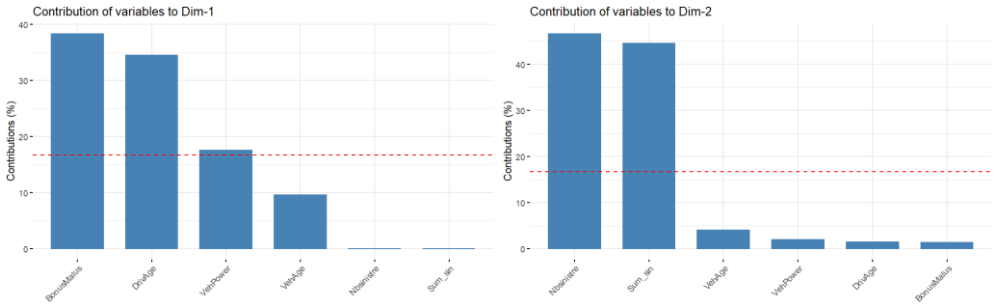
Pour le Marketing en fonction du genre, on note que le marketing « M1 » fonctionne très bien, d'autant plus chez le genre masculin que féminin (on se rappelle le fait que le marketing M1 est celui qui est majoritaire et qu'on a plus d'hommes que de femme dans notre échantillon). En effet, cette représentation nous montre que les hommes sont sur-représentés (colorés en bleu) et que les femmes sont sous-représentées (colorés en rouge). Pour le marketing « M2 » et « M3 » on voit que ce sont les femmes qui sont sur-représentées et les hommes sont sous-représentés. Le marketing « M3 » en blanc représente l'effectif attendu sous l'hypothèse d'indépendance.

Pour le Marketing en fonction de la classe du véhicule, on note que le marketing « M1 » touche la grande majorité des assurés possédant une voiture « Cheap », ce sont également ceux qui sont sous-représentés comme l'indique sa couleur. Le marketing « M2 » et « M4 » semblent identiques, avec les mêmes niveaux de sous et sur représentation, à savoir la classe « Cheap » étant sur-représentée et « Medium » sous représentée. Le marketing « M3 » est également représenté comme étant l'effectif attendu sous l'hypothèse d'indépendance, ce qui signifie que si les deux 2 variables étaient indépendantes, elle devrait obtenir ce type de résultat.

## *ACP*




Nous observons que les 2 premiers axes représentent quasi 50% de l'information, ce qui est satisfaisant.




De ces représentations, nous notons que les variables qui contribuent le plus à l'axe 1 sont liés au Bonus-Malus et l'âge du conducteur. La fréquence de sinistre et le coût du sinistre celles qui contribuent le plus à l'axe 2.