KUNASEELAN Sangeerthan
KOFFI Eunice
MO'MANTE SOH Micelle-Ange
TENAILLE Léa

Langage R en Actuariat :

Tarification en Assurance Dommage

Mai 2024

TABLE DES MATIERES

INTRODUCTION 3
Partie 1 : EXPLORATION DES DONNÉES 3
I. Statistiques descriptives 3
1. Le paquet CASdatasets 3
2. Les données freMTPLfreq 3
3. Les données freMTPLsev 4
II. Analyse univariée 4
1. Les variables numériques 4
2. Les variables catégorielles 7
III. Analyse multivariée 8
1. Corrélation entre variables numériques 9
2. Corrélation entre variables catégorielles 9
3. Corrélation entre variables numériques et catégorielles 9
4. Sinistralité par âge 10
Partie 2 : MODÉLISATION 11
I. Modification de la base de données 11
II. Modélisation 11
1. Modélisation du nombre de sinistres 11
2. Modélisation du coût de sinistres 12
Partie 3 : TARIFICATION 13
Partie 4 : RÉGRESSION LASSO 13
CONCLUSION 14

INTRODUCTION

Le but de ce projet est de tarifer un contrat d'assurance de responsabilité civile automobile grâce au langage R, à partir de données d'un portefeuille d'assurés.

Pour parvenir à notre objectif, nous utiliserons comme outil principal les Modèles Linéaires Généralisés qui permettent d'estimer le coût et le nombre de sinistres moyen d'un assuré.

Premièrement, nous présenterons les données utilisées. Dans un second temps, nous utiliserons les Modèles Linéaires Généralisés, aussi bien sur le coût des sinistres que sur le nombre de sinistres, afin d'obtenir un tarif pour notre portefeuille d'assurés. Nous vérifierons également que notre estimation est fiable grâce à la validation croisée.

Partie 1 : EXPLORATION DES DONNÉES

Dans cette partie, nous allons apprendre à connaître les données que nous allons traiter dans ce projet. Nous commencerons par faire une description générale puis nous nous attarderons sur de potentiels liens entre variables.

I. Statistiques descriptives

1. Le paquet CASdatasets

Les données que nous exploitons pour ce projet sont des données provenant du paquet CASdatasets. Elles concernent le nombre de sinistres et le coût des sinistres dans un portefeuille de contrats d'assurance de responsabilité civile automobile, en France. Au total, il y a 413 169 contrats et il y a eu 16 181 sinistres.

Nous utiliserons plus précisément deux fichiers : *freMTPLfreq* et *freMTPLsev*.

2. Les données freMTPLfreq

Ces données sont celles liées au nombre de sinistres dans le portefeuille des assurés. Cette table contient 413 169 lignes et 10 colonnes. Il y a donc 10 variables :

- PolicyID : Numéro de police du contrat *(variable catégorielle)*;
- ClaimNB : Nombre de sinistres survenus dans l'année pour chaque contrat *(variable numérique)* ;
- Exposure : Temps dans l'année pendant lequel le contrat était effectif *(variable numérique)* ;
- CarAge : Age de la voiture *(variable numérique)* ;
- DriverAge : Age du conducteur *(variable numérique)* ;
- Density : Nombre d'habitants au km² dans la ville française de résidence de l'assuré *(variable numérique)* ;
- Power : Puissance du moteur *(variable catégorielle)* ;
- Brand : Marque *(variable catégorielle)* ;

- Gas : Type de carburant *(variable catégorielle)* ;
- Region : Région française de résidence de l'assuré *(variable catégorielle)*.

Pour chaque contrat, aucune donnée parmi ces 10 variables n'est manquante.

```
── Data Summary ────────────────────────
                           Values
Name                       freMTPLfreq
Number of rows             413169
Number of columns          10
_________________________
Column type frequency:
  factor                   5
  numeric                  5
```

3. *Les données freMTPLsev*

Ces données sont celles liées au coût des sinistres dans le portefeuille des assurés. Cette table contient 16 181 lignes et 2 colonnes. Il y a donc 2 variables, pour lesquelles aucune donnée n'est manquante :

- PolicyID : Numéro de police du contrat *(variable catégorielle)* ;
- ClaimAmount : Coût du sinistre *(variable catégorielle)*.

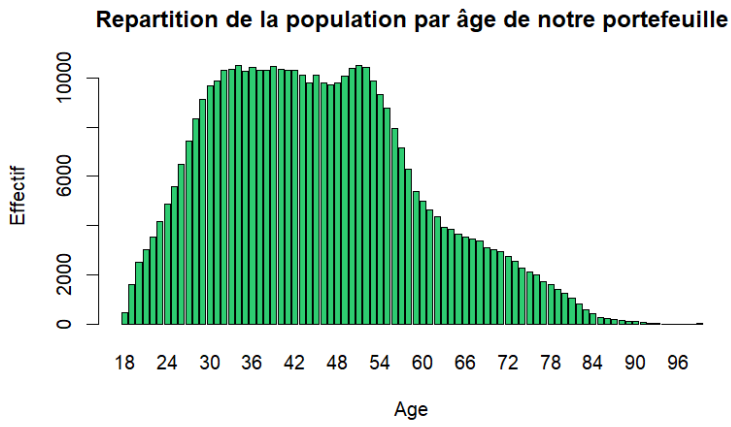
```
── Data Summary ────────────────────────
                           Values
Name                       freMTPLsev
Number of rows             16181
Number of columns          2
_________________________
Column type frequency:
  numeric                  2
```

II. Analyse univariée

Nous allons dorénavant analyser plusieurs graphiques, construits à partir des différentes variables numériques et catégorielles des 2 tables.

1. *Les variables numériques*

Les données freMTPLfreq



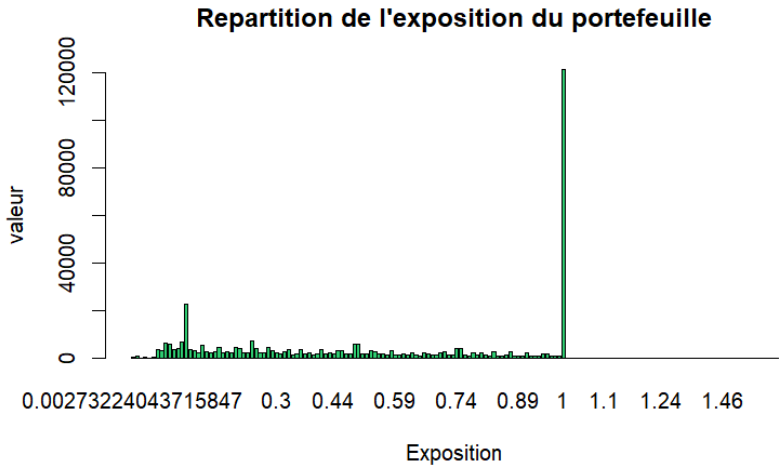
La répartition de l'âge a la forme d'une courbe de Gauss : il y a peu de jeunes assurés et peu (voire très peu, dans les très grands âges) de personnes âgées assurées.

Les assurés ont au moins 18 ans, car pour être assurés, ils doivent avoir le permis, qui est délivrable à partir de 18 ans en France. Mais les jeunes n'ont pas tous le permis à 18 ans et s'ils l'ont, ils n'ont pas spécialement de voiture directement, donc ils n'ont pas besoin d'être assurés. Ainsi, il y a peu de jeunes assurés.

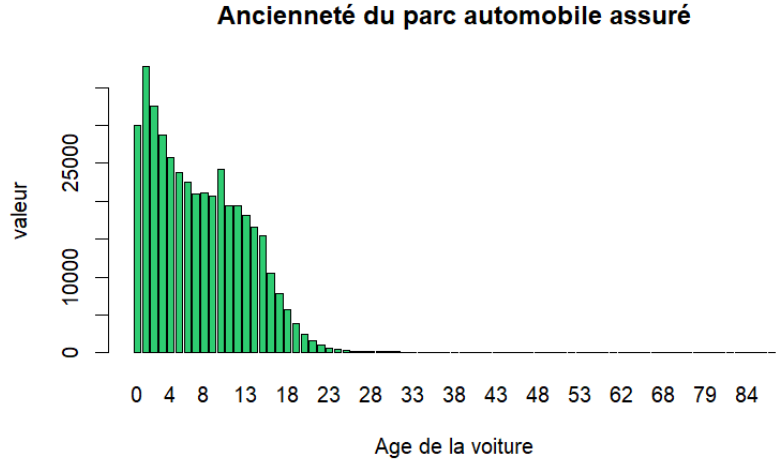
Mais au fur et à mesure de l'âge, il y a une augmentation du nombre de personnes assurés, jusqu'à environ 35 ans : de plus en plus de personnes ont le permis et de plus en plus de personnes possèdent une voiture qu'ils assurent.

Le nombre d'assurés entre 35 ans et 55 ans environ reste plutôt stable.

Ensuite, il y a une baisse du nombre de personnes assurées. Plus les personnes vieillissent, moins elles conduisent, donc moins elles ont de voiture et donc moins elles sont assurées. L'effectif de personnes assurées de plus de 90 ans est très faible.

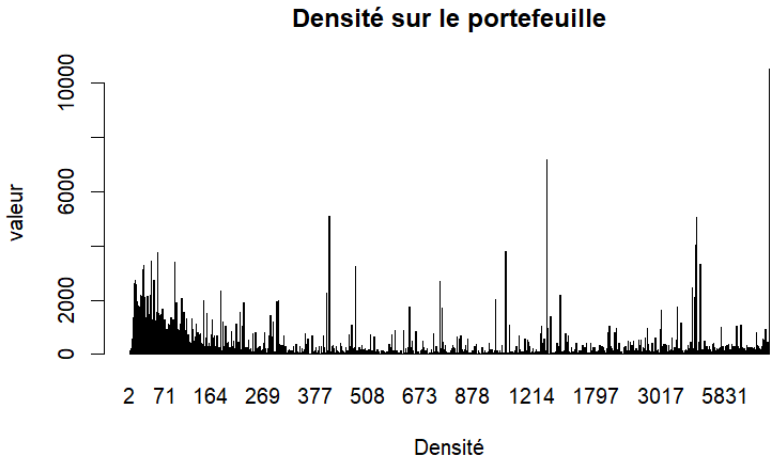


La durée de la plupart des contrats dans le portefeuille est une année, qui est la durée maximale d'un contrat. Cependant, certains contrats ont une durée d'exposition plus courte. Cela peut être dû à des départs d'assurés de l'assurance.

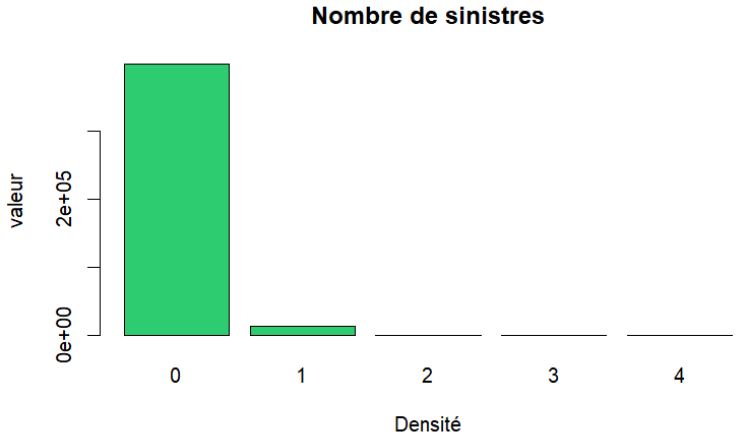


La plupart des voitures ont moins de 15 ans. Le pic du nombre de voitures par âge se situe autour des 5 ans.

Il y a peu de voitures âgées (plus de 20 ans) et très peu de voitures très âgées (au-delà de 35 ans). Ces dernières sont sûrement pour la plupart des voitures de collection.

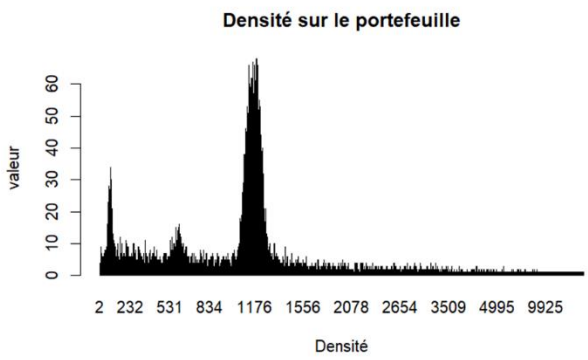


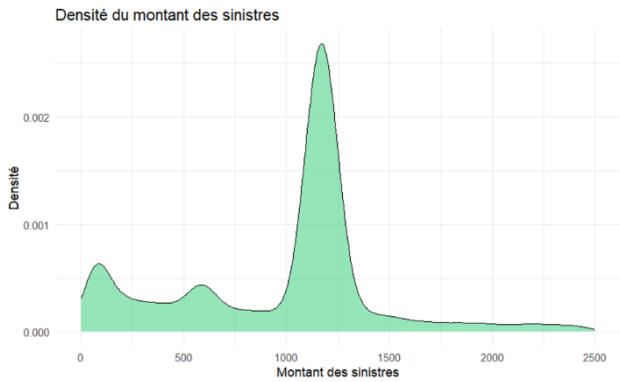
Il n'y a pas de schéma spécifique qui ressort du graphique de la densité. Certaines villes sont peu denses, d'autres sont très denses.



La plupart des assurés n'ont pas eu de sinistre durant l'année. En effet, il y a eu seulement 16 181 sinistres, alors qu'il y a 413 169 contrats d'assurance. Pour les assurés qui ont reporté un sinistre, la plupart en ont un seul dans l'année. Le maximum de sinistres par police est de 4. C'est un chiffre encore raisonnable pour un nombre de sinistres pour un seul assuré en une année.

<u>Les données freMTPLsev</u>

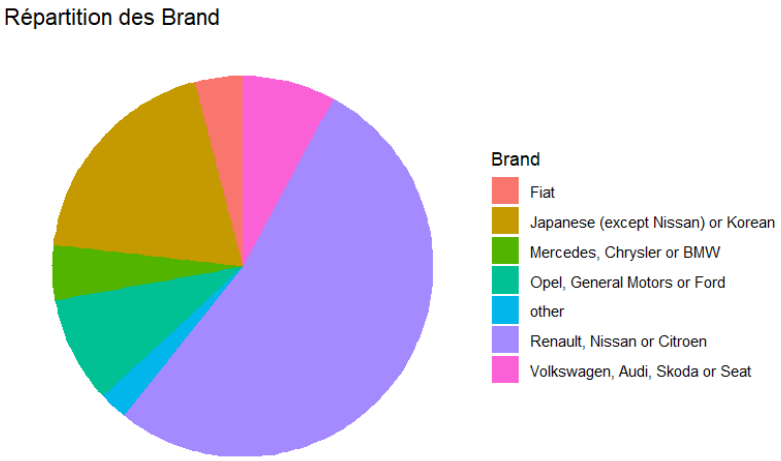




Ce diagramme en bâtons représente le coût des sinistres.

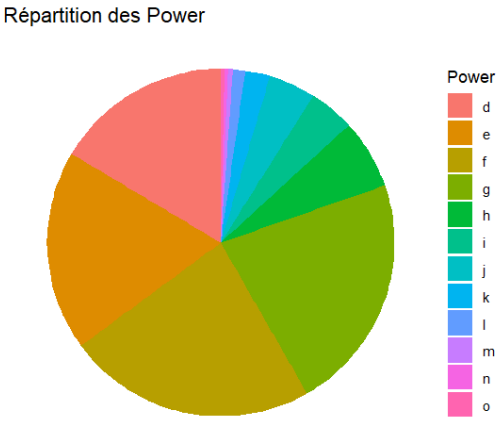
Une majorité des sinistres a un coût aux environs de 1200€. Il y a plus de sinistres dont le coût est inférieur à 1000€ que de sinistres dont le coût est supérieur à 1400€, mais il y a des sinistres pouvant aller jusqu'à quasiment 10000€.

2. *Les variables catégorielles*



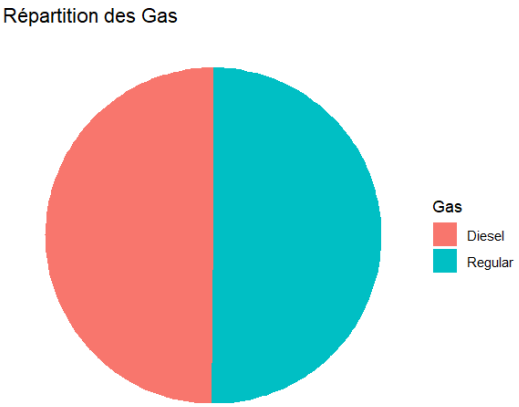
Les marques des voitures sont regroupées par paquets de plusieurs marques.

Les marques les plus représentées sont Renault, Nissan et Citroën. Cela fait sens car la première et la dernière sont des marques françaises.

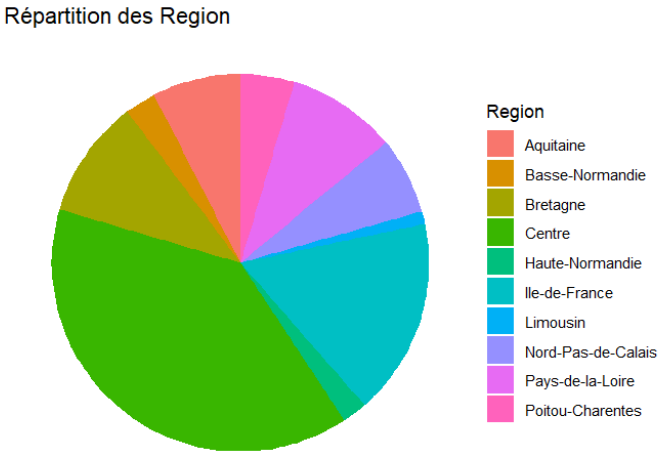


Les puissances des voitures sont représentées par des lettres.

4 types de puissance représentent plus de 75% de la totalité des puissances : les « d », « e », « f » et « g ».



Il y a une répartition quasiment égalitaire entre les voitures fonctionnant au diesel et celles fonctionnant à l'essence.



Tout d'abord, on peut remarquer que les régions où habitent les assurés sont encore nommées selon les anciennes régions, c'est-à-dire celles d'avant 2016, et qu'elles sont toutes en France métropolitaine.

Les assurés résident pour la majorité dans la région Centre, puis en Ile-de-France.

III. Analyse multivariée

A présent, nous allons nous intéresser aux corrélations entre les différentes variables, aussi bien du côté des variables numériques que catégorielles.

Nous avons vérifié en amont, que le nombre de polices n'excédait pas strictement le nombre de déclarations de sinistres indiquées dans la base de données. Dans le cas où il n'y a pas de sinistre, il peut exister une police d'assurance, mais le contraire n'est pas possible : il n'existe pas un sinistre qui ne fait pas partie d'une police d'assurance.

Après avoir fait une jointure entre les deux tables avec pour clef le numéro de police, nous avons étudié les corrélations de différentes variables.

1. *Corrélation entre variables numériques*

	ClaimNb	Exposure	CarAge	DriverAge	Density
ClaimNb	1.000000000	0.07603922	0.004505673	-0.006127537	0.005163136
Exposure	0.076039220	1.00000000	0.140103931	0.194260409	-0.112348816
CarAge	0.004505673	0.14010393	1.000000000	-0.046413527	-0.142318327
DriverAge	-0.006127537	0.19426041	-0.046413527	1.000000000	-0.001692481
Density	0.005163136	-0.11234882	-0.142318327	-0.001692481	1.000000000

Grâce à cette matrice de corrélation, on peut constater si certaines variables sont corrélées entre elles : celles dont le coefficient est proche de 1.

Il n'y a pas de coefficient étant proche de 1. Ils sont même plutôt proches de 0. Donc les variables ne sont pas vraiment corrélées entre elles.

On peut cependant constater que les plus gros coefficients de corrélations (ceux supérieurs à 0,10) sont entre les variables :

- Density et Exposure ;
- DriverAge et Exposure ;
- CarAge et Exposure ;
- CarAge et Density.

2. *Corrélation entre variables catégorielles*

La corrélation entre les variables catégorielles se fait grâce au test du Chi2 et au V de Cramer. Ce dernier permet de donner les densités de la corrélation entre variables catégorielles.

La première étape est de faire un test du Chi2, qui va nous renseigner sur s'il y a une dépendance entre les variables catégorielles ou pas.

La deuxième étape est de faire le V de Cramer, qui va permettre de quantifier l'intensité de la dépendance entre les variables catégorielles.

	Power <dbl>	Brand <dbl>	Gas <dbl>	Region <dbl>
Power	1.00000000	0.19225800	0.33224440	0.04766947
Brand	0.19225800	1.00000000	0.09996278	0.18978339
Gas	0.33224440	0.09996278	1.00000000	0.10228913
Region	0.04766947	0.18978339	0.10228913	1.00000000

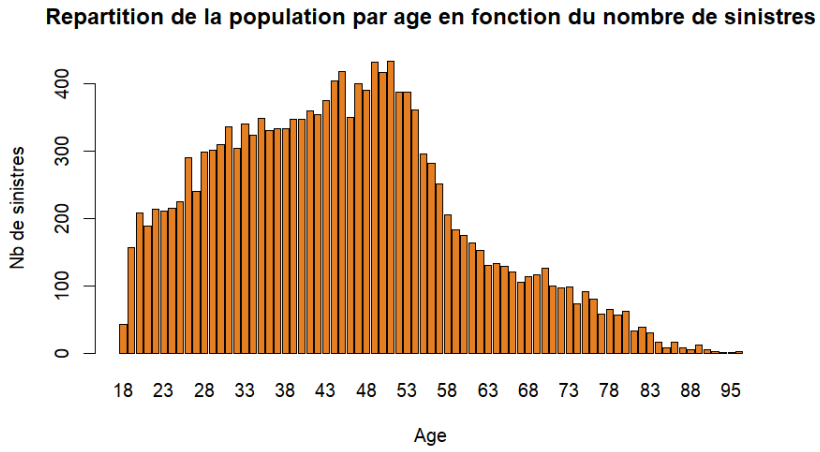
Pour les variables catégorielles, les coefficients de corrélation sont un peu plus élevés que pour celles numériques, mais sont quand même plus proches de 0 que de 1.

Les variables Puissance et Gas sont les variables les plus corrélées. Les autres variables entre elles ont des coefficients qui sont à peu près de l'ordre de la corrélation des variables les plus corrélées des variables numériques ci-dessus.

3. *Corrélation entre variables numériques et catégorielles*

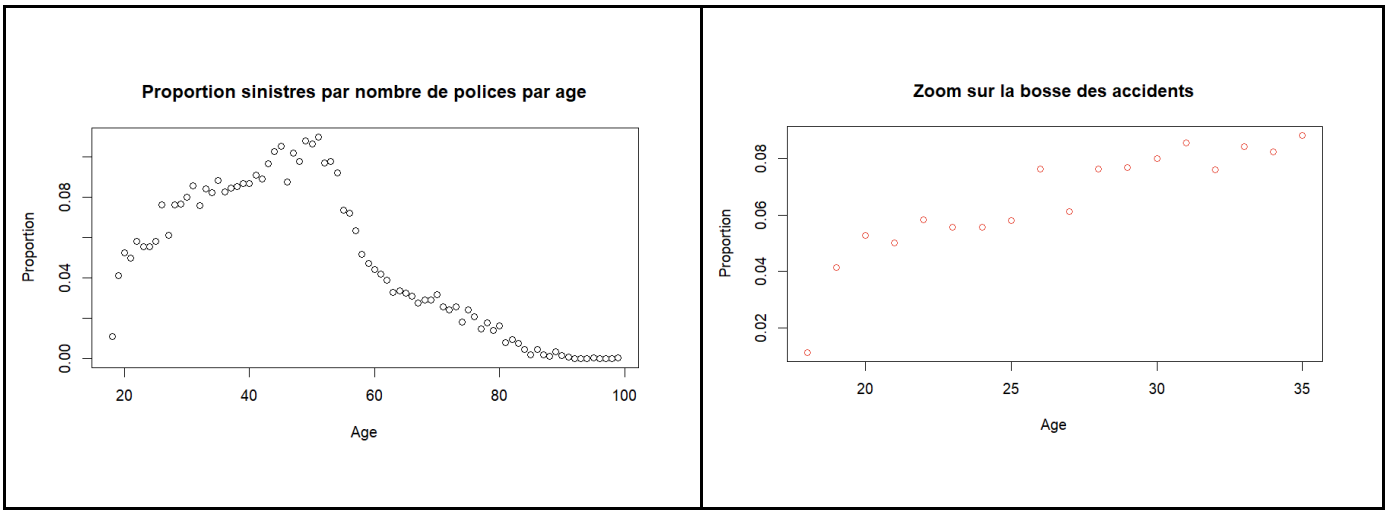
Nous avons pu déterminer qu'il avait une dépendance entre les variables numériques et catégorielles grâce à un test de Kruskall-Wallis.

4. *Sinistralité par âge*



Nous remarquons que la courbe est en forme de courbe de Gauss, tout comme pour le graphique « Répartition de la population par âge ».

Nous avons les nombre de sinistres les plus élevés autour des âges où il y a le plus de personnes assurées. Ceci est cohérent.



Il y a très peu d'assurés âgés, donc ce n'est pas vraiment représentatif.

En ce qui concerne les assurés très jeunes, ils ont beaucoup d'accidents par rapport à leur nombre. C'est ce qui est appelé la bosse des accidents.

Partie 2 : MODÉLISATION

Dans cette partie, nous allons modéliser le nombre et le coût des sinistres, après avoir fait quelques modifications de la base de données, grâce à des Modèles Linéaires Généralisés.

I. Modification de la base de données

Pour améliorer l'analyse lors de la modélisation, nous avons regroupé certaines modalités de variables qualitatives similaires afin de réduire leur nombre et faciliter l'interprétation des résultats. Les regroupements ont été effectués pour les régions de résidence des assurés, les marques des voitures, et la puissance des véhicules. Nous avons considéré le nombre de polices d'assurance, le montant moyen des indemnités, l'exposition, et l'âge des véhicules pour ces regroupements. Initialement, notre base de données contenait 10 régions, que nous avons regroupées en 4 grandes régions. En ce qui concerne les marques des voitures, nous avons réduit le nombre de catégories de 7 à 4, en fusionnant certaines marques similaires. Pour la puissance des véhicules, nous avons simplifié les catégories de 12 à 3. Ces regroupements permettent une meilleure gestion des données et une analyse plus cohérente des résultats, tout en préservant la pertinence et la signification des variables utilisées dans nos modèles.

II. Modélisation

Avant de procéder à la modélisation, nous avons divisé nos données en ensembles d'entraînement et de test, respectivement de 75% et 25%. Cette split stratifié garantit que les proportions des différentes catégories de la variable cible sont préservées dans les deux ensembles de données. Avant cela, nous avons converti nos variables catégorielles telles que Puissance, Région et Marque en variables dummy. Cette transformation nous permet de représenter chaque catégorie par une série de variables binaires distinctes, facilitant ainsi l'utilisation de ces variables dans notre modélisation.

1. Modélisation du nombre de sinistres

Nous avons décidé de tester plusieurs modèles pour modéliser le nombre de sinistres :

- Le modèle de la loi de Poisson : la famille de Poisson est utilisée pour modéliser des variables aléatoires discrètes qui prennent des valeurs entières positives. Elle est souvent utilisée pour modéliser des comptages d'évènements rares, tels que le nombre de sinistres.
- Le modèle de la loi de Poisson surdispersée ou la loi de QuasiPoisson : la famille de Poisson quasi est une extension de la famille de Poisson qui permet de prendre en compte la surdispersion des données. La surdispersion se produit lorsque la variance des données est supérieure à la moyenne, ce qui peut rendre le modèle de Poisson inadapté.
- Le modèle de la loi Binomiale Négative : la famille de négatif binomial est une autre extension de la famille de Poisson qui permet de prendre en compte la surdispersion des données. Elle est souvent utilisée pour modéliser des comptages d'événements qui présentent une forte variabilité.

Ces modèles conviennent pour ce qu'on cherche à modéliser.

Le modèle gaussien ne convient pas quant à lui car la distribution des sinistres ne semble pas symétrique autour de la moyenne et le nombre de sinistres n'est pas continu.

Le modèle de la loi Gamma ne peut pas être utilisé car il est possible d'avoir un nombre de sinistres nul.

Avec les trois modèles que nous avons retenus (Poisson, Poisson surdispersée, Binomiale négative) les variables Puissance et Région ne sont pas très significatives selon leur p-values, donc nous les avons retirées.

Pour le modèle de Poisson et celui de la Binomiale Négative, nous avons modélisé le modèle Zéro inflation de Poisson et le modèle Zéro inflation de Binomial Négatif. Ce sont des modèles qui sont basés sur une grande fréquence de la présence de 0 dans les données, ce qui est le cas pour le nombre de sinistres des assurés.

Afin de choisir parmi ces modèles, nous avons eu recours au critère d'information d'Akaike (AIC, en anglais Akaike Information Criterion). C'est un modèle permettant d'augmenter la vraisemblance du modèle en ajoutant un paramètre lors de l'estimation d'un modèle. Plus généralement, le critère AIC est utilisé pour comparer différents modèles et déterminer celui qui convient le mieux.

Nous calculons ce critère pour le modèle de Poisson, le modèle Binomial Négatif, le modèle Zéro inflation de Poisson et le modèle Zéro inflation de Binomial Négatif.

Le résultat nous indique que c'est le modèle Zéro inflation Binomial Négatif qui est le meilleur, car il a la plus petite valeur du critère AIC.

Nous avons tout de même souhaité regarder un autre critère pour le choix du modèle. C'est le critère RMSE (Root Mean Square Error). Il permet d'évaluer la précision d'un modèle de régression en comparant la racine carrée de la moyenne des carrés des écarts entre les valeurs prédites par le modèle et les valeurs observées dans l'ensemble des données de test. Il s'appuie sur les données de validation.

Avec ce critère RMSE (calculé sur les mêmes modèles que pour celui de l'AIC), nous avons trouvé que le meilleur modèle pour modéliser le nombre de sinistres était celui du Zéro inflation de la loi Poisson car il avait la valeur de RMSE la plus petite.

On note également que les précisions entre le modèle de Poisson et quasi-Poisson sont similaires. Ils ont quasiment la même valeur pour le RMSE. Cela signifie que les données ne présentent pas de surdispersion significative. Dans ce cas, la loi de quasi-Poisson se réduit à la loi de Poisson et les deux distributions sont équivalentes.

2. Modélisation du coût de sinistres

Nous avons adopté une approche similaire à celle utilisée pour modéliser le nombre de sinistres. Après avoir exploré plusieurs modèles, nous avons sélectionné le meilleur en fonction des critères AIC et RMSE. En ce qui concerne la modélisation du coût des sinistres, nous avons évalué trois modèles principaux : la loi Gamma, le modèle gaussien et la loi Binomiale Négative.

Le modèle de la loi Gamma peut être utilisé car nous nous sommes concentrés sur les coûts d'indemnisation de l'assurance, les coûts de sinistres positifs. Cependant, nous avons rejeté ce modèle ainsi que le modèle gaussien en raison d'un nombre insuffisant de variables significatives et d'une valeur d'AIC négative.

Ainsi, selon le critère AIC, il est apparu que le modèle de la loi Binomiale Négative était le plus approprié pour notre ensemble de données. Nous avons confirmé cette conclusion en examinant également le critère RMSE, qui a également indiqué que le modèle de la loi Binomiale Négative était le meilleur choix pour la modélisation du coût des sinistres.

Partie 3 : TARIFICATION

Comme vu en cours d'Assurance Non-Vie, nous appliquerons un modèle "Fréquence x Coût" pour établir une tarification adéquate pour notre portefeuille d'assurés. Pour ce faire, nous nous appuierons sur les résultats obtenus dans la partie précédente de notre étude, à savoir les modèles retenus en fonction des critères AIC et RMSE.

En utilisant le critère AIC, nous avons sélectionné le modèle Zéro Inflation de la loi Binomiale Négative pour modéliser le nombre de sinistres, tandis que pour le coût des sinistres, le choix s'est porté sur le modèle de la loi Binomiale Négative. Quant au critère RMSE, il a conduit à choisir le modèle Zéro Inflation de la loi de Poisson pour le nombre de sinistres, et le modèle de la loi Binomiale Négative pour le coût des sinistres.

Finalement, en tenant compte du tarif qui minimise l'écart avec les données observées, nous optons pour le critère RMSE. Cela se traduit par un écart de 6,47% par rapport aux données réelles. La somme des primes perçues par l'assurance est alors estimée à 34 210 609€.

Comme il y a 413 169 polices d’assurance, cela reviendrait à un tarif pur annuel moyen de 82,80€ par police d’assurance d’assurance de responsabilité civile du portefeuille. Cependant, la tarification est individuelle, donc les primes sont différentes d’un assuré à un autre. Il faut juste que la somme totale des primes que l’assurance reçoit soit au moins égale à 34 210 609€.

Partie 4 : RÉGRESSION LASSO

Les données de notre base de données ont été divisées de manière aléatoire en ensembles d'entraînement et de test, avec 75% des données utilisées pour l'entraînement et 25% pour les tests. Cette approche permet de s'assurer que le modèle est évalué sur des données non vues, fournissant ainsi une estimation plus fiable de sa performance sur des données futures.

La pénalisation de type Lasso (Least Absolute Shrinkage and Selection Operator) a été utilisée pour ajuster les modèles. Cette pénalisation permet de comprendre quelles variables ont le plus d'impact sur le nombre et le coût des sinistres. L'utilisation du Lasso est particulièrement utile dans les contextes où il y a beaucoup de variables explicatives, car elle peut améliorer la prédiction en réduisant le surajustement et en simplifiant le modèle. Avec votre jeu de données, nous avons obtenu les résultats suivants :

	Nombre de sinistres	Coût des sinistres
Gaussien	0.203554	7310.549
Poisson	0.2035885	7542.345

Pour la famille Gaussienne, les résultats montrent un RMSE (Root Mean Square Error) de 7310.549 pour le coût du sinistre, tandis que pour la famille Poisson, le RMSE est de 7542.345. En

termes de nombre de sinistres, les deux modèles montrent des résultats très proches avec une légère différence (0.203554 pour le modèle Gaussien et 0.2035885 pour le modèle Poisson).

Ainsi, le modèle Gaussien offre une performance légèrement meilleure en termes de prédiction du coût des sinistres par rapport au modèle Poisson, bien que la différence soit assez marginale. Cela pourrait indiquer que les données suivent plus étroitement une distribution normale pour le coût des sinistres plutôt qu'une distribution de Poisson, qui est souvent utilisée pour modéliser des données de comptage.

CONCLUSION

Tout d'abord, les variables que nous avons conservées nous semblent pertinentes pour la modélisation du nombre de sinistres et de leur coût. De plus, ce qui est ressorti de cette étude, est que le modèle de la loi Binomiale Négative semble le meilleur, en tout cas, celui qui avait le plus souvent les meilleurs critères.

Comme pistes d'amélioration, nous pourrions utiliser des modèles d'apprentissage, comme les Random Forests. Il serait également intéressant d'utiliser des modèles de durée pour la modélisation du nombre de sinistres. En effet, plus l'exposition d'une police d'assurance est grande, plus l'assuré est susceptible d'avoir un sinistre sur cette police.

Grâce à ce projet, nous avons pu mettre en pratique ce que l'on a étudié durant le cours d'Assurance Non-Vie. Nous avons remarqué qu'en pratique, certains problèmes se posent, ce qui n'est pas le cas quand on ne regarde que la théorie. Ce projet nous a donc permis de nous préparer pour notre avenir professionnel avec l'utilisation de données réelles et avec une problématique proche de ce que nous pourrions trouver dans nos futurs métiers.