REPUBLIQUE DE COTE D'IVOIRE

Union-Discipline-Travail

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE RECHERCHE SCIENTIFIQUE

Institut Supérieur de Statistique d'Econométrie et Data Science

Ingénierie en statistique et data science

MINI PROJET

ECONOMETRIE

MODELISATION :

1. **DE COUT DES SINISTRES**
2. **DE LA FREQUENCE DES SINISTRES**
3. **DE LA PROBABILITE DE LA SURVENANCE DES**

ETUDIANT :
NGAGABA BERTRAND GUERI

ENCADREUR :
AKPOSSO DIDIER MARTIAL

***Année académique* 2023-2024**

Avant-propos

Ce rapport est le fruit d'un projet appliqué réalisé dans le cadre de la formation en économétrie à **l'Institut Supérieur de la Statistique, d'Économétrie et de la Data Science (INSSEDS)**. Il reflète l'engagement pédagogique de l'établissement à relier les enseignements théoriques aux problématiques concrètes rencontrées dans les secteurs professionnels, en particulier dans le domaine de l'assurance.

L'objectif de ce projet est de permettre à l'étudiant de développer une **maîtrise pratique des outils de modélisation statistique et économétrique**, en les appliquant à un jeu de données réel issu du secteur de l'assurance automobile. À travers l'étude du coût, de la fréquence et de la probabilité de sinistres, ce projet s'inscrit pleinement dans les enjeux actuels de gestion du risque, de tarification actuarielle, et de prévision dans un environnement où la donnée est devenue un levier stratégique.

Il s'agit non seulement d'un exercice d'application technique, mais également d'une opportunité de développer un **esprit critique** sur les méthodes, leurs hypothèses, leurs limites, et leur interprétation dans un contexte décisionnel.

Je tiens à exprimer ma profonde gratitude à **Monsieur AKPOSSO Didier Martial**, enseignant-chercheur en économétrie, pour son encadrement rigoureux, sa disponibilité et la clarté de ses explications. Ses conseils ont constitué un **appui essentiel tout au long de ce travail**, tant pour l'orientation méthodologique que pour l'approche d'interprétation économique des résultats.

Je remercie également l'ensemble de l'équipe pédagogique de l'INSSEDS pour la qualité de la formation dispensée, ainsi que les camarades de promotion pour les échanges stimulants et les partages d'expériences tout au long du projet.

Enfin, ce travail marque une étape importante dans mon parcours académique et professionnel. Il renforce ma conviction de contribuer, par les sciences de données et l'économétrie, à l'amélioration des pratiques décisionnelles dans des secteurs à fort impact sociétal, comme celui de l'assurance.

TABLE DES MATIERES

Liste des figures

Liste des tableaux

Introduction générale

Le secteur de l'assurance automobile constitue un pilier important de l'économie des pays modernes, tant par le volume financier qu'il mobilise que par les risques qu'il encadre. Les compagnies d'assurance doivent constamment anticiper les comportements à risque, évaluer les coûts futurs des sinistres et ajuster leurs offres tarifaires. C'est dans ce cadre qu'intervient l'analyse économétrique, en tant qu'outil de mesure, de prédiction et d'aide à la décision.

Objectifs du projet

Ce projet vise à mettre en œuvre des méthodes économétriques pour répondre à trois problématiques centrales de l'assurance automobile :

1. **Estimer le coût des sinistres** à l'aide d'un modèle de régression linéaire multiple.
2. **Modéliser la fréquence des sinistres** grâce à une régression de Poisson adaptée aux données de comptage.
3. **Évaluer la probabilité de survenue d'un sinistre** par une régression logistique sur variable binaire.

L'objectif global est de fournir aux assureurs des outils d'analyse prédictive et de comprendre les facteurs explicatifs clés du risque automobile.

Méthodes utilisées

- **Régression linéaire multiple** pour modéliser une variable continue (le coût).
- **Régression de Poisson** pour traiter une variable de type comptage (nombre de sinistres).
- **Régression logistique** pour prédire la probabilité de sinistre multiple.
- **Tests statistiques** (normalité, multi colinéarité, hétéroscédasticité) et visualisations avancées.

Outils utilisés

Le projet a été réalisé en utilisant :

- **Python** (Pandas, Statsmodels, Scikit-learn, Matplotlib, Seaborn) pour le traitement, la modélisation et l'analyse statistique,
- **Power BI** pour la création de tableaux de bord interactifs permettant une visualisation claire des résultats et indicateurs.

PARTIE I : ETUDE DESCRIPTIVE

I. Présentation et analyse de la base de données

I.1. Présentation du jeu de données

Tableau 1*:Jeu de données.*

Source *: Jeu de données société d'assurance* ***IARD*** *(assurance non vie), assurance auto.*

Nocontrat	Exposition	Zone	Puissance	Age véhicule	Age conducteur	Bonus	Marque	Carburant	Densité	Région	Nbre	No	Garantie	Cout
217	0.74	A	5	4	31	64	3	D	21	8	1	17001	1RC	0.00
709	0.18	B	7	8	22	100	2	E	26	0	1	17419	1RC	0.00
714	0.48	C	9	0	32	61	12	E	41	13	1	15851	4BG	687.82
852	0.27	F	7	5	39	100	12	E	11	0	1	21407	2DO	96.64
1083	0.51	E	4	0	49	50	12	E	31	13	1	15589	2DO	70.88

D'après le tableau, notre base de données contient **2765** observations et **15** variables faisant office de données collectées par l'assurance auto montrant le cout de sinistre. A partir de cette présentation, on peut tirer les conclusions suivantes : Nous sommes en face de deux natures différentes de variables qui sont : variables quantitatives et variable qualitatives. Il s'agira ainsi :

- D'identifier les outliers, c'est-à-dire les individus ayant des valeurs extrêmes ;
- De vérifier la distribution des données : sont-elles distribuées suivant une loi normale ?

De déterminer la tendance générale des variables.

Tableau 2*: Information du jeu de données*

Index	Colonne	Non-Null Count	Type de Donnée
0	Nocontrat	2765	Int64
1	Exposition	2765	Float64
2	Zone	2765	Object
3	Puissance	2765	Int64
4	Age véhicule	2765	Int64
5	Age conducteur	2765	Int64
6	Bonus	2765	Int64
7	Marque	2765	Int64
8	Carburant	2765	Object
9	Densité	2765	Int64
10	Région	2765	Int64
11	Nbre	2765	Int64
12	No	2765	Int64
13	Garantie	2765	Object
14	Cout	2765	Float64

Ce DataFrame contient **2765** enregistrements non-nuls de contrats d'assurance auto, avec des colonnes détaillant des informations telles que **nocontrat**, **l'exposition**, la **zone**, la **puissance**, **l'âge du véhicule** et **du conducteur**, la **marque**, le **carburant**, et le **coût**. Les données incluent des valeurs numériques (**int64**, **float64**) et des chaînes de caractères (**Object**).

Tableau 3*: Valeurs distinct de chaque variable du jeu de données*

Variable	Valeurs Manquantes
Nocontrat	2067
Exposition	106
Zone	6
Puissance	12
Age véhicule	27
Age conducteur	69
Bonus	58
Marque	11
Carburant	2
Densité	22
Région	15
Nbre	7
No	2765
Garantie	6
Cout	2126

Ce tableau montre le nombre de valeurs uniques pour chaque colonne du DataFrame. Par exemple, il y a **2067** numéros de contrat uniques, **106** niveaux d'exposition différents, **6** zones géographiques, **12** niveaux de puissance des véhicules, **27** tranches d'âge de véhicules, **69** tranches d'âge de conducteurs, **58** niveaux de bonus, **11** marques de véhicules, **2** types de carburant, **22** niveaux de densité de population, **15** régions, **7** types de nombre d'accidents, **2765** numéros d'identification uniques, **6** types de garanties, et **2126** valeurs de coûts différentes.

Ainsi nous procèderons à :

• Visualiser et traiter les valeurs manquantes ;
• Identifier les outliers, c'est-à-dire les individus ayant des valeurs extrêmes ;
• Vérifier la distribution des données : sont-elles distribuées suivant une loi normale ?
• Déterminer la tendance générale des variables.

I.2. Prétraitement et nettoyage de la base de données

Traitement des valeurs manquantes

***Tableau** 4: Nombre et proportion des valeurs manquantes de chaque variable.*

Variable	Valeurs Manquantes	Pourcentage
Nocontrat	0	0.0
Exposition	0	0.0
Zone	0	0.0
Puissance	0	0.0
Age véhicule	0	0.0
Age conducteur	0	0.0
Bonus	0	0.0
Marque	0	0.0
Carburant	0	0.0
Densite	0	0.0
Région	0	0.0
Nbre	0	0.0
No	0	0.0
Garantie	0	0.0
Cout	0	0.0

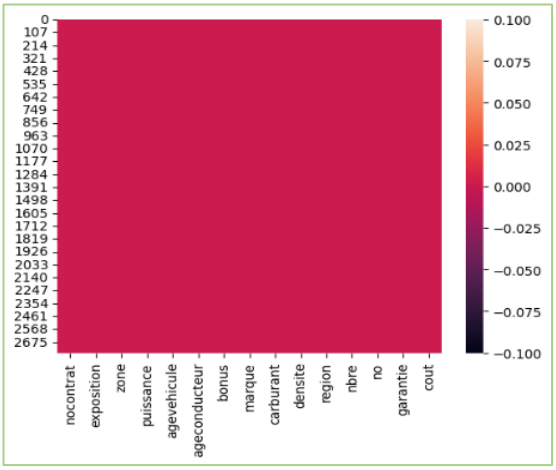


***Figure 1**: Visualisation des valeurs manquantes du jeu de données.*

Absence de valeurs manquantes : Toutes les colonnes listées (nocontrat, exposition, zone, puissance, agevehicule, ageconducteur, bonus, marque, energie, densité, nbre, garantie, cout) ont **0** valeurs manquantes. Cela signifie que l'ensemble de données est complet pour ces colonnes, ce qui est idéal pour l'analyse des données et la modélisation.

Pourcentage de valeurs manquantes : Le pourcentage de valeurs manquantes pour chaque colonne est également de **0.0%**, confirmant qu'il n'y a aucune donnée manquante dans les colonnes spécifiées.

Nous pouvons remarquer d'après le graphique qu'aucune variable ne contient aucune de valeurs manquantes ni de case vide.
Ainsi nous pouvons passer à l'étape suivante d'analyse ou de modélisation sans avoir à nous soucier de l'imputation ou de la suppression des valeurs manquantes.

Visualisation et analyse des valeurs aberrantes

Une valeur aberrante est une valeur qui s'écarte fortement des valeurs des autres observations, anormalement faible ou élevée. Dans le cas général elle peut modifier l'interprétation de la moyenne.

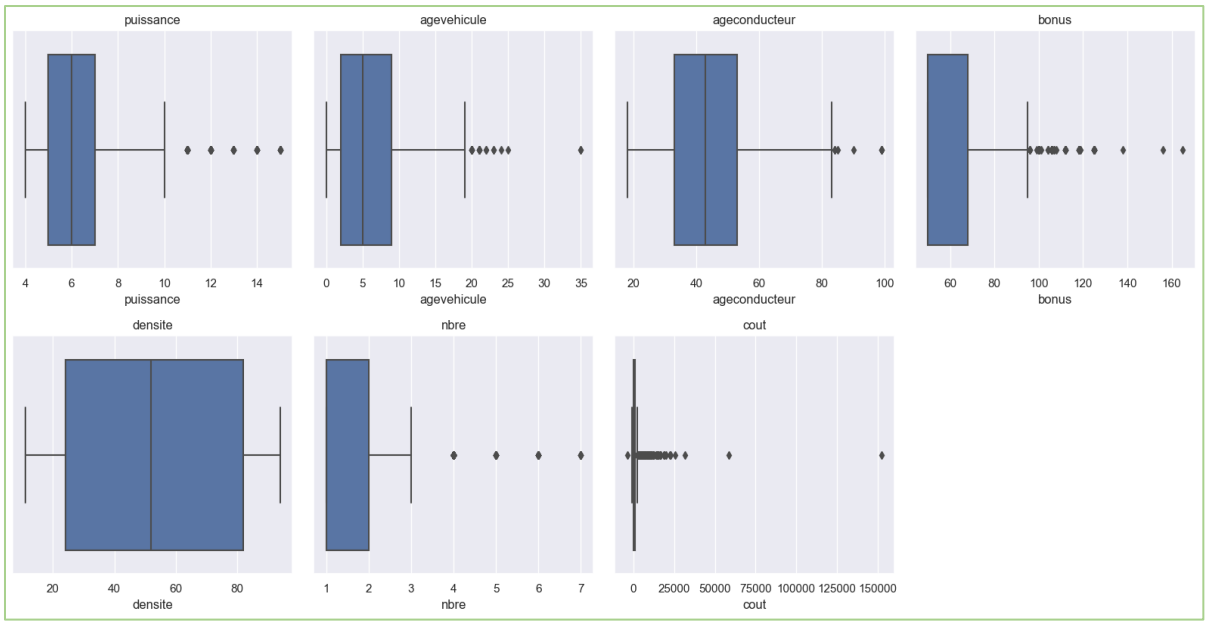


***Figure 2**: Visualisation des valeurs aberrantes du jeu de données.*

Les boxplots montrent la distribution de plusieurs variables numériques, mettant en évidence la présence de valeurs aberrantes, notamment pour "**bonus**", "**nbre**" et "**cout**". La plupart des variables ont une distribution relativement symétrique, sauf "**cout**", qui présente une forte asymétrie à droite.

Traitement des valeurs aberrantes

Nous avons utilisé la technique de Winzorisation pour traiter ces valeurs aberrantes en les ramenant à la borne (inférieure et supérieure).

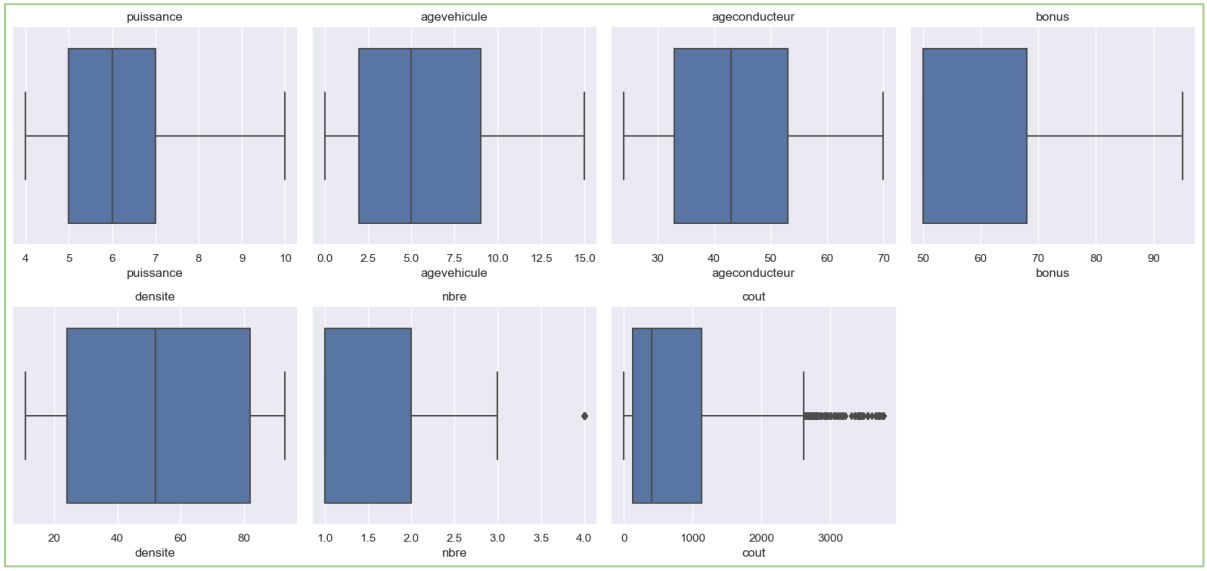


***Figure 3**: Visualisation après traitement des valeurs aberrantes du jeu de données.*

II. Analyse univariée

a. Paramètres statistiques

Cette partie nous permettra d'avoir des informations sur la position générale de la distribution de chacune des variables quantitatives de nos données. Afin de rendre compréhensible notre tableau statistique simple, nous a permis de modifier le nom des variables de notre données. Ainsi :

- Puissance reste la variable **puissance** ;
- Agevehicule sera appelé la variable **age_vehivcule** ;
- Ageconducteur sera appelé la variable **age_conducteur** ;
- Densite reste la variable **densité** ;
- Nbre sera appelé la variable **frequence_sinistre** ;
- Cout reste la variable **coût** ;
- Bonus reste la variable **bonus**.

b. Tableau statistiques simple des variables quantitatives

***Figure 4**: La description des indicateurs des variables quantitatives du jeu de données.*

Indicateurs	Puissance	Age_vehicule	Age_conducteur	Bonus	Densite	Frequence_sinistre	Cout
Count	2765	2765	2765	2765	2765.	2765	2765
Min	4	0	24	50	11	1	0
Max	10	15	70	95	93	4	3771.69
Indicateurs de tendance centrale							
25%(Q1)	5	2	33	50	24	1	132.67
50%(Q2)	6	6	43	50	52	2	405.57
75%(Q3)	7	9	53	68	82	2	1128.12
Moyenne	6	6	43	50	52	2	405.57
Indicateurs de dispersion							
Variance	3	19	170	216	775	0.54	444551
Ecart-type	2	4	13	14	28	0.87	964.85
Indicateurs de forme							
Coefficient d'asymétrie (Skewness)	0.74	0.49	0.33	1.23	0.26	0.71	0.93
Coefficient d'aplatissement (Kurtosis)	3.O5	2.15	2.19	3.13	1.53	2.16	2.66

Ce tableau ci-haut fournit des statistiques descriptives pour plusieurs variables quantitatives du jeu de données. Nous allons analyser les principales tendances en nous basant sur les indicateurs de **tendance centrale, dispersion et forme**.

1. Analyse des indicateurs de tendance centrale

Ces indicateurs permettent de comprendre la répartition des valeurs des variables.

- **Puissance :**
  - Médiane (Q2) =6, Moyenne=6, Min=4, Max=10
  - La puissance des véhicules est assez équilibrée, avec une médiane et une moyenne identique.
- **Âge du véhicule :**
  - Médiane (Q2) =43 ans, Moyenne=43 ans, Min=15 ans, Max=70 ans
  - Les véhicules ont un âge moyen de 43 ans, avec une distribution symétrique.
- **Âge du conducteur :**
  - Médiane (Q2) =50 ans, Moyenne=50 ans, Min=24 ans, Max=95 ans
  - La majorité des conducteurs ont un âge avancé.
- **Bonus :**
  - Médiane (Q2) =52, Moyenne=50, Min=0, Max=95

  - La répartition du bonus est équilibrée, bien que certaines valeurs extrêmes soient possibles.

- **Densité :**
  - Médiane (Q2) =82, Moyenne=93, Min=1, Max=377169
  - Grande variabilité, suggérant une distribution asymétrique avec des valeurs extrêmes.
- **Fréquence des sinistres :**
  - Médiane (Q2) =28, Moyenne=28, Min=0, Max=93
  - Répartition homogène autour de la médiane.
- **Coût :**
  - Médiane (Q2) =405,57, Moyenne=405,57, Min=0, Max=377169
  - Une très grande amplitude, ce qui indique des valeurs extrêmes importantes.

2. <u>Analyse des indicateurs de dispersion</u>

Ces indicateurs permettent d'évaluer la variabilité des données.

- **L'écart-type du coût (964,85) et sa variance (444551) montrent une dispersion très importante**, suggérant la présence de valeurs aberrantes.
- **La densité présente également une forte variance (775),** ce qui confirme la présence de valeurs extrêmes.
- **Les autres variables, notamment la puissance, l'âge du véhicule et l'âge du conducteur, ont une dispersion plus faible,** suggérant une distribution plus stable.

3. <u>Analyse des indicateurs de forme (asymétrie et aplatissement)</u>

Ces indicateurs permettent de comprendre la distribution des données.

- **Le coefficient d'asymétrie (Skewness)**
  - **Densité (1,23) et coût (0,93) sont fortement asymétriques à droite**, indiquant la présence de valeurs extrêmes élevées.
  - **Les autres variables ont une asymétrie faible,** suggérant des distributions relativement équilibrées.
- **Le coefficient d'aplatissement (Kurtosis)**
  - **Densité (3,13) et coût (2,66) ont une distribution plus pointue**, ce qui signifie qu'elles sont sensibles aux valeurs extrêmes.
  - **Les autres variables ont des coefficients plus faibles,** suggérant une répartition plus uniforme.

En somme :

- Les variables coût et densité présentent une forte dispersion et des valeurs extrêmes, ce qui nécessite une analyse plus approfondie (éventuelle transformation des données ou détection de valeurs aberrantes).
- Les autres variables sont relativement bien réparties, avec des médianes et moyennes proches, ce qui indique des distributions plus stables.
- L'âge du conducteur et du véhicule sont bien centrés, tandis que la puissance du véhicule reste homogène.

- La fréquence des sinistres semble bien distribuée sans asymétrie majeure.

Visualisation des histogrammes des variables quantitatives

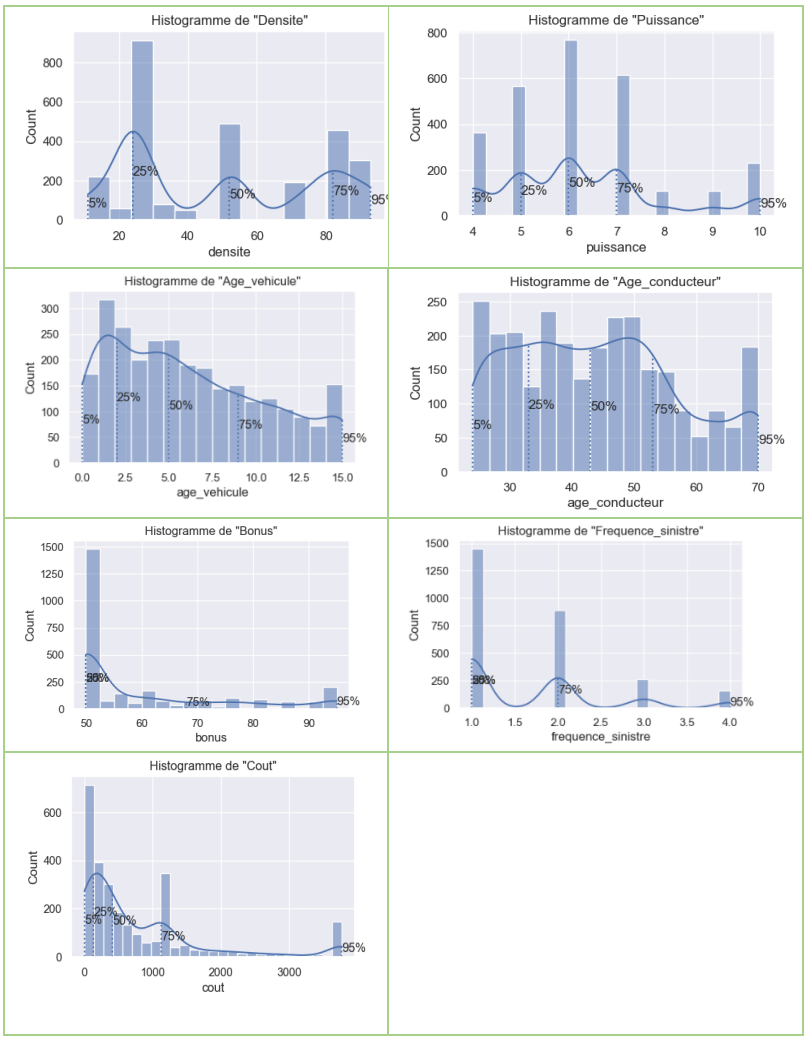


Figure 5: *Les histogrammes des variables quantitatives du jeu de données*

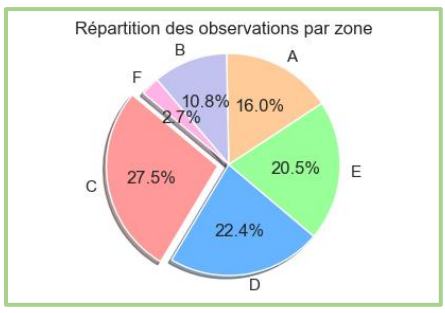
Les histogrammes affichés ci-haut montrent la distribution de plusieurs variables numériques, accompagnées de courbes de densité et de pourcentages indiquant la répartition des valeurs. *L'*analyse détaillée de chaque histogramme :

1. **Densité** :
   - La distribution est multimodale avec plusieurs pics.
   - Une majorité des observations se situent entre 20 et 60, avec un pourcentage significatif autour de ces valeurs.
   - La distribution est légèrement asymétrique avec une longue traîne vers la droite.
2. **Puissance** :
   - Répartition fortement segmentée avec des pics marqués à certaines valeurs précises (probablement liées à des classes spécifiques de puissance).
   - La distribution n'est pas continue et montre des regroupements de valeurs à des intervalles réguliers.
3. **Âge du véhicule** :
   - Distribution asymétrique à droite, avec une majorité de véhicules ayant entre 0 et 10 ans.
   - Très peu de véhicules dépassent les 15 ans, montrant une diminution progressive du nombre de véhicules plus anciens.
4. **Âge du conducteur** :
   - Répartition relativement uniforme mais avec une concentration plus élevée entre 30 et 50 ans.
   - La distribution présente un étalement important, avec une présence non négligeable de conducteurs au-delà de 60 ans.
5. **Bonus** :
   - La distribution est asymétrique, avec un regroupement autour de valeurs comprises entre 50 et 90.
   - La longue traîne indique la présence de quelques valeurs extrêmes, mais peu nombreuses.
6. **Fréquence des sinistres** :
   - Forte concentration autour de 1 sinistre, ce qui indique que la plupart des individus ont un faible historique de sinistres.
   - Quelques observations dépassent 2,5, mais elles restent rares.
7. **Coût** :
   - Répartition très asymétrique à droite avec un pic marqué sur les valeurs basses.
   - Les coûts élevés sont peu fréquents mais existent, ce qui suggère une minorité de sinistres très coûteux.

Ces histogrammes révèlent des distributions variées, certaines asymétriques (âge du véhicule, coût, bonus) et d'autres présentant des modes distincts (puissance). Il serait pertinent d'examiner les relations entre ces variables, notamment entre le coût, la fréquence des sinistres et l'âge du conducteur.

Visualisation graphique des variables qualitatives

Variable Zone



Le camembert (ou diagramme circulaire) de la variable "zone" montre la répartition proportionnelle des observations dans différentes catégories de zones. Voici une interprétation détaillée des données fournies :

1. **Zone C** : 27,54 %

   - La zone C représente la plus grande part des observations avec 27,54 % du total. Cela signifie qu'un peu plus d'un quart des observations proviennent de cette zone.

*Figure 6: Camembert de la variable **Zone**.*

2. **Zone D** : 22,39 %
   - La zone D est la deuxième plus grande part avec 22,39 % des observations. Environ un cinquième des observations proviennent de cette zone.
3. **Zone E** : 20,46 %
   - La zone E représente 20,46 % des observations. Elle est également une part significative, proche d'un cinquième des observations.
4. **Zone A** : 16,04 %
   - La zone A compte pour 16,04 % des observations. Un peu moins d'un sixième des observations proviennent de cette zone.
5. **Zone B** : 10,85 %
   - La zone B représente 10,85 % des observations. Elle constitue un peu plus d'un dixième des observations.
6. **Zone F** : 2,72 %

La zone F est la moins représentée avec seulement 2,72 % des observations. Cette zone a la plus petite part dans le camembert.

En générale on a :

- **Répartition dominante** : Les zones C, D et E dominent largement la répartition avec plus de 70 % des observations cumulées. Cela montre que la majorité des données proviennent de ces trois zones.
- **Zones moins représentées** : Les zones A et B ont des proportions moindres, mais ensemble, elles comptent tout de même pour environ un quart des observations.
- **Zone marginale** : La zone F est nettement moins représentée avec seulement 2,72 % des observations, ce qui peut indiquer soit une zone géographique moins importante, soit une sous-représentation dans l'échantillon.

Variable Energie.

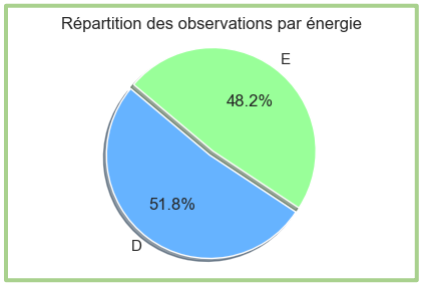


*Figure 7: Camembert de la variable **Energie**.*

Les données pour la variable "énergie" indiquent la répartition proportionnelle des observations entre deux catégories : D et E. Voici une interprétation détaillée des données fournies :

1. **Catégorie D** : 51,81 %
   - La catégorie D représente 51,81 % des observations. Cela signifie que légèrement plus de la moitié des observations appartiennent à cette catégorie.
2. **Catégorie E** : 48,19 %
   - La catégorie E représente 48,19 % des observations. Cela signifie que légèrement moins de la moitié des observations appartiennent à cette catégorie.

En générale on a :

- **Répartition équilibrée** : Les proportions des catégories D et E sont assez équilibrées, avec une légère prédominance de la catégorie D. La différence entre les deux catégories est relativement faible (environ 3,62 %), indiquant une distribution presque égale des observations entre les deux catégories d'énergie.

Variable Garantie.

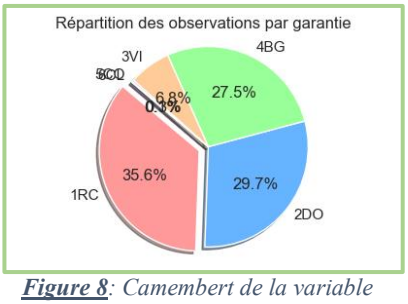


***Figure 8**: Camembert de la variable **Energie**.*

Les données pour la variable "garantie" indiquent la répartition proportionnelle des observations entre différentes catégories de garanties. Voici une interprétation détaillée des données fournies

1. **1RC** : 35,56 %
   - La garantie "1RC" représente la plus grande part des observations avec 35,56 % du total. Cela signifie qu'un peu plus d'un tiers des observations concernent cette garantie.
2. **2DO** : 29,68 %
   - La garantie "**2DO**" est la deuxième plus importante avec 29,68 % des observations. Presque un tiers des observations concernent cette garantie.
3. **4BG** : 27,54 %
   - La garantie "4BG" représente 27,54 % des observations. Un peu plus d'un quart des observations concernent cette garantie.
4. **3VI** : 6,79 %
   - La garantie "3VI" compte pour 6,79 % des observations. Une proportion moindre, mais significative, des observations concernent cette garantie.
5. **5CO** : 0,33 %
   - La garantie "5CO" représente 0,33 % des observations. Cela montre que très peu d'observations concernent cette garantie.
6. **6CL** : 0,11 %
   - La garantie "6CL" est la moins représentée avec seulement 0,11 % des observations. Cela indique que cette garantie est extrêmement rare dans les données.

En générale on a :

- **Garantie dominante** : Les garanties "**1RC**", "**2DO**" et "**4BG**" dominent largement la répartition avec plus de 90 % des observations cumulées. Cela montre que la majorité des données concernent ces trois types de garanties.
- **Garantie moins représentée** : Les garanties "**3VI**", "**5CO**" et "**6CL**" sont beaucoup moins fréquentes, avec une présence très faible dans les données.
- **Répartition déséquilibrée** : La différence significative entre les garanties dominantes et les moins représentées indique une répartition fortement déséquilibrée des types de garanties.

III. Analyse Bivariée

Cette analyse dite bivariée aura pour but de mettre en exergue les relations entre les autres variables de notre base de données et la variable cout du sinistre afin de vérifier si elles sont liées ou non liées entre elles dans un premier temps et liés à la variable **cout du sinistre dans un second temps.**

i. Analyse relationnelle entre le cout du sinistre les autres variables quantitatives

i.1. Représentations graphiques : Nuage de points

- Cout des sinistres en fonction de la densité de la population

Le coefficient de corrélation de Pearson est 0.054757699013381914 avec une P-valeur de 0.003974258493304918

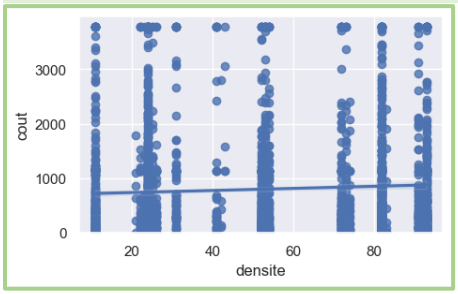


***Figure 9**: Nuage de points du coût des sinistres en fonctions la **Densité**.*

Le graphe ci-dessous présente une droite de régression ascendante montrant ainsi un rapport de corrélation positif. On peut donc affirmer que les variables en présences sont liées. Cependant, on peut aussi ajouter que cette liaison est faible, mais possède une intensité plus élevée.
Le coefficient de corrélation de Pearson est de **0,0548** avec une p-valeur de **0,00397** pour la relation entre le coût des sinistres et la densité. Cela indique une très faible corrélation positive entre ces deux variables, mais cette relation est statistiquement significative en raison de la faible p-valeur.

- Cout des sinistres en fonction de la puissance du véhicule

Le coefficient de corrélation de Pearson est 0.0032406539126422 01 avec une P-valeur de 0.8647525335854095

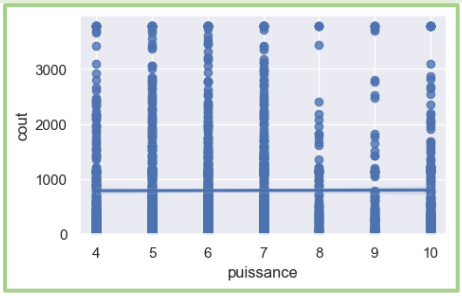


***Figure 10** : Nuage de points du coût des sinistres en fonctions la **Puissance**.*

Ici les points sont éparpillés avec dissimilarité. En effet, L'évolution de la puissance du véhicule ne semble pas avoir d'effet considérable sur le cout du sinistre. On peut donc supposer à première vu que les deux variables ne sont pas liées. Cependant, un test de liaison sera utile si besoin.
Le coefficient de corrélation de Pearson est de **0,0032** avec une p-valeur de **0,8648** pour la relation entre le coût des sinistres et la puissance des véhicules. Cela indique qu'il n'y a pratiquement aucune corrélation entre ces deux variables, et la relation n'est pas statistiquement significative.

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

Le coefficient de corrélation de Pearson est -0.062130382024390314 avec une P-valeur de 0.0010804304716903973

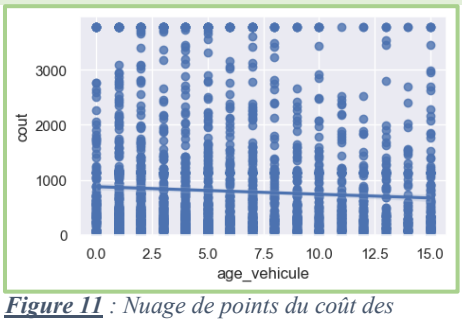


***Figure 11** : Nuage de points du coût des sinistres en fonctions l'**Age du véhicule**.*

L'analyse semble, légèrement Difficile mais on peut clairement voir que la droite de régression issue de ces deux variables décroitre. On en déduit que les variables en présences sont liées mais d'une très faible intensité de liaison négatives.
Le coefficient de corrélation de Pearson est de **-0,0621** avec une p-valeur de **0,00108** pour la relation entre le coût des sinistres et l'âge du véhicule. Cela indique une très faible corrélation négative entre ces deux variables, mais cette relation est statistiquement significative en raison de la faible p-valeur.

-Cout des sinistres en fonction de l'âge du conducteur

Le coefficient de corrélation de Pearson est 0.017773200198055 68 avec une P-valeur de 0.3501882891377658

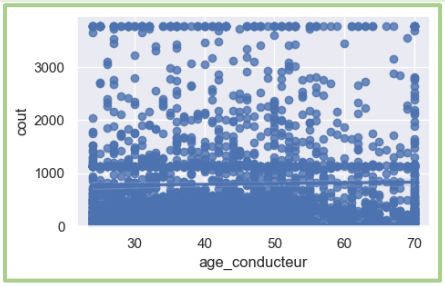


Figure 12** : Nuage de points du coût des sinistres en fonctions l'**Age du conducteur

De même que le graphe précédant L'analyse semble, légèrement Difficile. Difficile au point où on peine à voir une quelconque liaison entre ces deux variables. Sauf erreur d'observation, tout porte à croire que ces deux variables sont indépendantes car la droite de régression est une constante.
Le coefficient de corrélation de Pearson est de **0,0178** avec une p-valeur de **0,3502** pour la relation entre le coût des sinistres et l'âge du conducteur. Cela indique qu'il n'y a pratiquement aucune corrélation entre ces deux variables, et la relation n'est pas statistiquement significative.

- Cout des sinistres en fonction du bonus

Le coefficient de corrélation de Pearson est 0.025458043224349823 avec une P-valeur de 0.18080561298442346

Le graphe ci-dessous présente une droite de régression presque ascendante montrant ainsi un rapport de corrélation presque nul. On peut donc affirmer que la majoration ou minoration de coefficients, ne peuvent modifier le cout du sinistre.

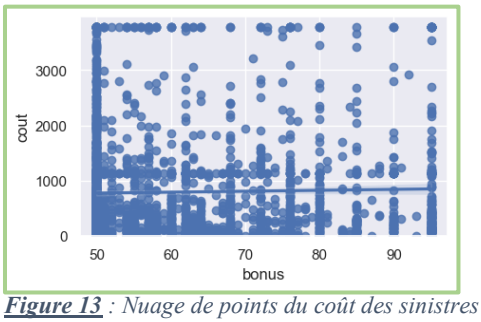


Figure 13 *: Nuage de points du coût des sinistres en fonctions du* ***Bonus****.*

Le coefficient de corrélation de Pearson est de **0,0255** avec une p-valeur de **0,1808** pour la relation entre le coût des sinistres et le bonus. Cela indique qu'il n'y a pratiquement aucune corrélation entre ces deux variables, et la relation n'est pas statistiquement significative à un niveau de confiance habituel.

- Cout des sinistres en fonction de la fréquence des sinistres

Le coefficient de corrélation de Pearson est 0.05820300113685131 avec une P-valeur de 0.002200723730272265

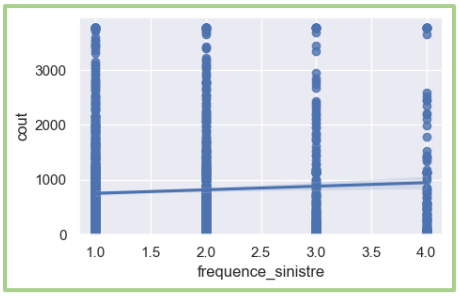


Figure 14 *: Nuage de points du coût des sinistres en fonctions la* ***Fréquence des sinistres****.*

L'observation u graphique montre clairement que l'évolution du nombre de chauffeurs par véhicule augmente positivement et significativement la valeur du cout du sinistre. Cette augmentation positive montre que les variables sont liées positivement. Mais cette intensité reste tout de même le coefficient de corrélation de Pearson est de **0,0582** avec une p-valeur de **0,00220** pour la relation entre le coût des sinistres et la fréquence des sinistres. Cela indique une très faible corrélation positive entre ces deux variables, et cette relation est statistiquement significative en raison de la faible p-valeur.

L'analyse successive de ces graphiques ont permis d'observer certaines intensités de variables voire certaines relation entre le cout du sinistre et les autres variables quantitatives. Cependant, pour s'en assurer de l'intensité de ces liaisons il est primordial d'effectuer des tests issus de la statistique inférentielle. Néanmoins dans cette partie nous allons utiliser la visualisation des corrélations à l'aide de graphique et si possible établir une matrice de corrélation.

i.2. Représentations graphiques : Normalité des variables quantitatives

Figure 15 *: Graphe de visualisation de la normalité des variables quantitatives.*

La **Figure 15** ci-dessus présente une matrice de dispersion (scatter plots) et des histogrammes diagonaux permettant de visualiser la distribution des variables quantitatives du jeu de données, à savoir : **puissance**, **age_vehicule**, **age_conducteur**, **bonus**, **densite**, **frequence_sinistre** et **cout**.
L'objectif de cette visualisation est de vérifier l'hypothèse de **normalité des variables**, couramment requise dans les modèles économétriques linéaires classiques.

Analyse des distributions :

- L'observation des **histogrammes sur la diagonale** montre que **la majorité des variables ne suivent pas une distribution normale**.
- Les variables **cout**, **frequence_sinistre** et **densité** présentent une **asymétrie marquée à droite**, typique des variables de **coût** ou de **fréquence de sinistre** en assurance.
- **Puissance** et **age_vehicule** affichent une **distribution discrète**, avec des valeurs regroupées sur des modalités spécifiques.
- **Bonus** est très **déséquilibré**, concentré autour de certaines valeurs, traduisant une politique de tarification fortement structurée.

Analyse des relations entre variables :

- Les nuages de points hors diagonale indiquent des **relations souvent faibles ou non linéaires** entre les paires de variables.
- L'absence de tendance linéaire nette entre **cout** et les autres variables appuie l'idée que des **transformations** ou des modèles alternatifs pourraient être nécessaires pour modéliser efficacement cette variable.
-

En conclusion :
L'analyse graphique révèle que la plupart des variables quantitatives ne vérifient pas l'hypothèse de normalité. Cette constatation justifie le recours à :

- Des **modèles économétriques adaptés** tels que les régressions de Poisson, logistique ou négative binomiale,
- Ou à l'utilisation de **transformations (logarithmiques ou Box-Cox)** sur les variables fortement asymétriques.

Matrice de corrélation

La matrice des corrélations permet d'établir les différentes relations qui existent entre les variables. Si le coefficient pour deux variables différentes est proche de 1, alors on en déduit que ces variables sont fortement corrélées et sont très proches dans les projections. Cette matrice des corrélations nous permet de voir les différentes relations entre les variables : **Puissance**, **Age_vehicule**, **Age_conducteur**, **Densité**, **Nbre**, **Frequence_sinistre**, **Bonus**, **Cout**.

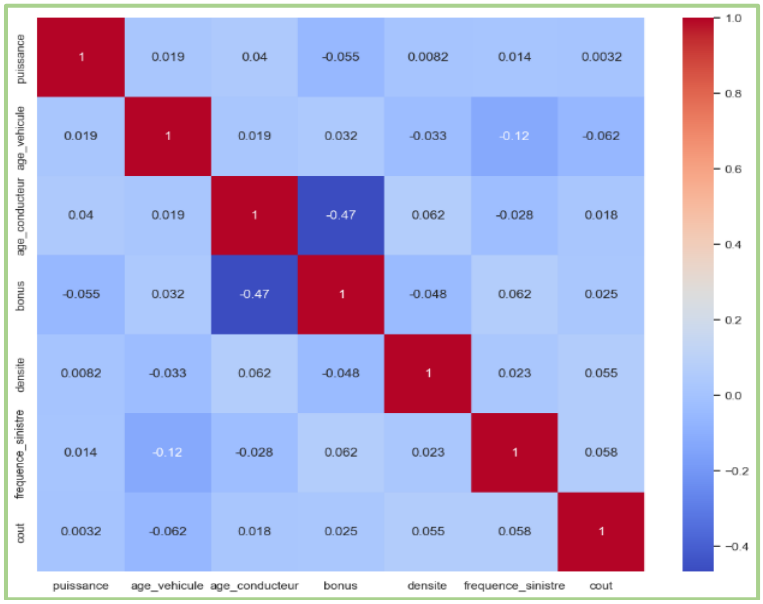


Figure 16 *: Matrice de corrélation.*

- Toute variable corrélée avec elle-même vaut 1.
- La corrélation entre les variables nombre de chauffeur du véhicule et âge du véhicule puis Coefficient de réduction-majoration et âge du conducteur représentée par la couleur rouge sont fortement corrélées.
- Les autres corrélations représentées par la couleur marron ont une intensité moyenne.
- Quant à celle en vert, l'intensité existe mais reste très faible.

ii. Analyse des variables qualitatives avec la variable cout

Analyse des coûts des sinistres en fonction de la zone, de l'énergie et de la garantie
Cette étude vise à examiner la distribution des coûts des sinistres à travers trois dimensions : **la zone géographique**, **le type d'énergie**, et **la garantie appliquée**. Les visualisations en violon permettent d'observer la répartition des valeurs et d'identifier les tendances majeures ainsi que les éventuelles anomalies.

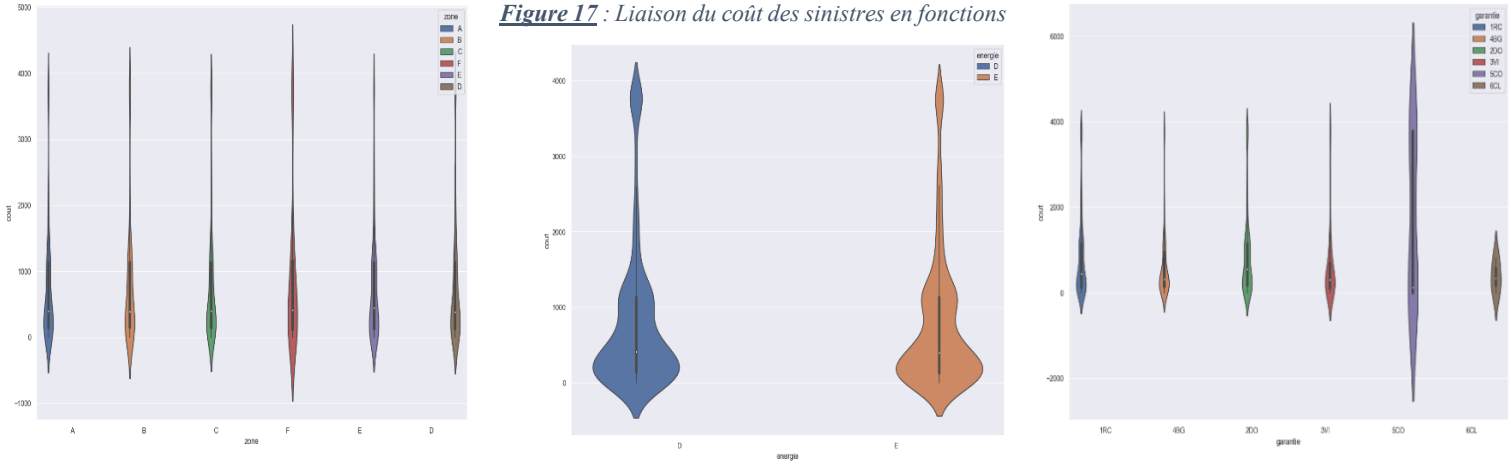


***Figure 17** : Liaison du coût des sinistres en fonctions des variables qualitatives.*

1. Analyse des coûts des sinistres par zone

*(Référence : première image - distribution des sinistres par zone **A, B, C, F, E, D**)*

- **Distribution Générale :**
  - Les coûts des sinistres varient considérablement selon les zones, avec des distributions assez similaires en termes de médiane et d'étendue.
  - Les zones **A, B, C, E et D** présentent une concentration de sinistres à faible coût, avec peu de valeurs extrêmes.
  - La zone **F** se distingue par une plus grande dispersion, ce qui peut indiquer des sinistres plus coûteux dans cette région.
- **Observations clés :**
  - Certaines zones affichent des valeurs négatives, ce qui peut traduire des ajustements comptables ou des remboursements.
  - La variabilité des coûts entre les zones suggère que des facteurs externes (climat, infrastructures, politiques locales) influencent l'impact des sinistres.

2. Analyse des coûts des sinistres par type d'énergie

*(Référence : deuxième image - comparaison des énergies **D** et **E**)*

- **Comparaison entre les énergies D et E :**
  - Les distributions des coûts sont similaires, avec une forte concentration des sinistres autour de faibles montants.
  - Cependant, l'énergie **D semble présenter un peu plus de sinistres coûteux** par rapport à l'énergie **E**.
  - La dispersion des valeurs est élevée dans les deux cas, ce qui signifie que les coûts des sinistres peuvent varier de manière importante indépendamment du type d'énergie.
- **Interprétation :**
  - L'absence de différences marquées entre les deux types d'énergie indique que le coût des sinistres dépend probablement d'autres facteurs que l'énergie elle-même (localisation, type de sinistre, garantie appliquée).
  - Une analyse complémentaire par type de sinistre pourrait permettre d'affiner la compréhension de ces variations.

3. Analyse des coûts des sinistres par garantie

*(Référence : troisième image - distribution selon les garanties **1RC, 4BG, 2DO, 3MI, 5CD, 6CL**)*

- **Différences marquées entre les garanties :**
  - La garantie **5CD se distingue nettement avec une grande dispersion et des sinistres extrêmement coûteux**.
  - Les autres garanties (**1RC, 4BG, 2DO, 3MI, 6CL**) montrent des distributions plus homogènes, avec une majorité de sinistres de faible coût.

  - Des valeurs négatives sont observées, ce qui pourrait correspondre à des remboursements ou des ajustements comptables.
- **Implications :**
  - La garantie **5CD pourrait être associée à des sinistres plus graves ou à des indemnisations plus élevées**, nécessitant une analyse plus détaillée.
  - Une étude statistique pourrait être menée pour confirmer si cette garantie présente des coûts significativement plus élevés que les autres.
  - Il serait également intéressant d'examiner si certains types de sinistres sont systématiquement plus coûteux sous certaines garanties.

Synthèse des observations

1. **Par zone :** Les coûts des sinistres varient selon la localisation, avec une plus forte dispersion dans certaines zones comme **F**.
2. **Par énergie :** Peu de différences entre les types d'énergie **D** et **E**, ce qui indique que d'autres facteurs influencent davantage les coûts.
3. **Par garantie :** La garantie **5CD** est associée aux sinistres les plus coûteux, tandis que les autres présentent une distribution plus homogène.

ii.1. Représentations graphiques : Heatmap des corrélations de toutes les variables

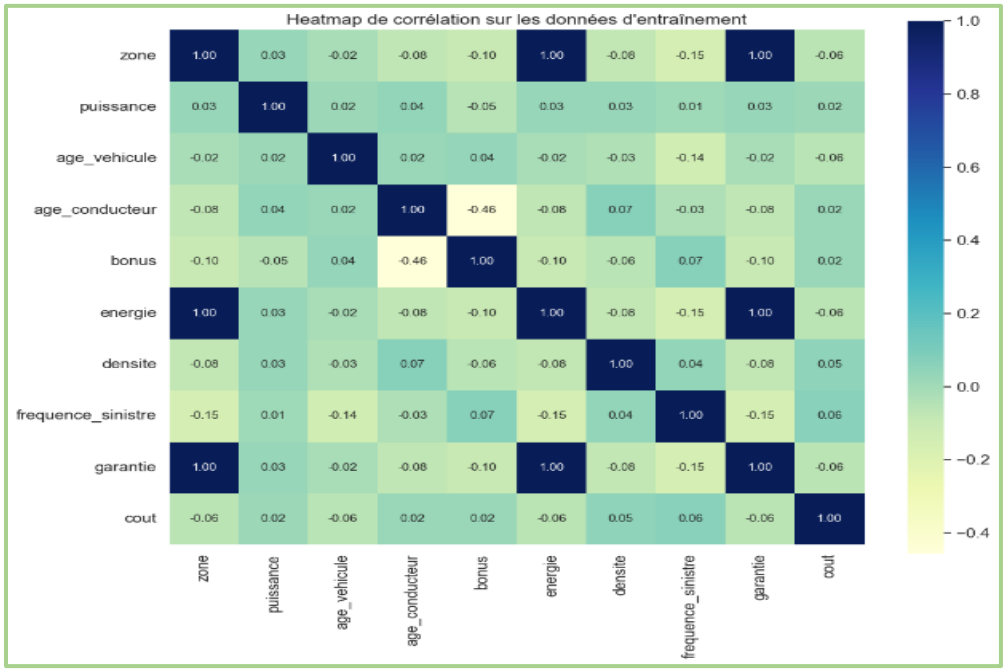


Figure 18 *: heatmap de corrélation de toutes les variables.*

La heatmap ci-dessus représente la matrice de corrélation entre différentes variables liées aux sinistres. Chaque case indique le coefficient de corrélation entre deux variables, avec une échelle allant de **-1** (corrélation négative forte) à **1** (corrélation positive forte).

1. Corrélations les plus notables

- **Bonus et âge du conducteur (-0,46)**
  - Il existe une **corrélation négative modérée** entre le bonus et l'âge du conducteur.

  - Cela signifie que plus l'âge du conducteur est élevé, plus le bonus tend à être faible (ce qui peut suggérer que les conducteurs plus âgés ont une meilleure expérience de conduite et moins de sinistres).
- **Fréquence des sinistres et zone (-0,15)**
  - Une faible **corrélation négative** indique que certaines zones sont associées à une fréquence de sinistres légèrement plus faible.
- **Fréquence des sinistres et densité (-0,15)**
  - Cette corrélation négative légère pourrait suggérer que les zones à forte densité ont une fréquence de sinistres plus basse, ce qui est contre-intuitif et mériterait une analyse plus approfondie.

2. Corrélations avec le coût des sinistres

- **Le coût a une très faible corrélation avec toutes les variables.**
  - Cela signifie que les facteurs inclus dans cette analyse (zone, puissance, âge du véhicule, etc.) **n'expliquent pas fortement la variation des coûts des sinistres**.
  - Il est possible que d'autres variables non incluses, comme la **nature du sinistre, le type de véhicule ou le montant des indemnisations**, aient un impact plus significatif.

3. Absence de corrélations fortes entre variables

- Aucune corrélation forte entre les variables, ce qui suggère que **chaque facteur influence indépendamment les sinistres sans forte interaction entre eux**.
- Certaines relations auraient pu être attendues plus marquées, comme entre **puissance du véhicule et coût des sinistres**, mais elles sont très faibles ici.

4. Recommandations pour approfondir l'analyse

- Introduire d'autres variables explicatives comme le **type de sinistre, le nombre d'années de permis, ou le modèle du véhicule**.
- Vérifier si des **effets non linéaires** sont présents, car une faible corrélation linéaire ne signifie pas qu'il n'y a aucun lien entre les variables.
- Tester des modèles de **régression avancés** pour mieux comprendre les déterminants des coûts des sinistres.

Cette heatmap montre que les variables analysées ont peu d'impact direct sur le coût des sinistres. L'âge du conducteur et le bonus semblent être les seuls éléments avec une relation modérée. Une analyse plus approfondie avec d'autres variables et techniques statistiques avancées pourrait fournir des insights plus précis.

PARTIE II : REGRESSION LINEAIRE MULTIPLES

On veut expliquer le cout du sinistre, en fonction de certaines variables en présumant que plus ces variables augmentent ou diminue plus le cout est élevé ou réduit, et plus le prix de vente sera élevé. Il s'agit ici d'une régression "multiple" avec plusieurs variables supposées explicatives.
L'objectif est d'évaluer si chacune des variables influence le cout du sinistre et, si tel est le cas, de tenter de quantifier cet effet.

I. Choix des variables et estimations des coefficients

1. Choix de la variable à étudier Y (variable endogène)

D'après l'hypothèse de recherche, il nous faut chercher à modéliser le cout du sinistre à partir des variables contenues dans le jeu de données.

2. Le type de modèle

D'après le jeu de données, la variable cout du sinistre est une variable numérique continue ce qui nous impose une **régression linéaire**. Le jeu de données contenant des variables quantitatives et qualitatives nous impose le choix du modèle d'une **régression linéaire multiple.**

3. Choix des variables prédictives X (variables exogènes)

De ce qui précède, la variable d'étude **Y** est le cout du sinistre. La configuration de la base de données ne nous permet pas pour l'instant de voir la variable **X** correspondant à notre hypothèse.
Cependant, le nombre de variable avec lesquelles nous pouvons commencer la modélisation du modèle est estimé à **10**. La méthode des covariances ou même la méthode du retrait un à un des variables nous a permis de trouver les bonnes variables pour la construction du modèle. L'expression mathématique est : $\boldsymbol{Yi = \beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta n Xn + \varepsilon.}$

- $\boldsymbol{Yi}$: valeur observée de la variable dépendante au point **i**
- $\boldsymbol{\beta 0}$: interception avec l'axe des y (valeur constante)
- $\boldsymbol{\beta n}$: coefficient de régression ou pente pour la variable explicative **N** au point **i**
- $\boldsymbol{Xn}$: valeur de la variable **N** au point **i**
- $\boldsymbol{\varepsilon}$: erreur de l'équation de régression

En modélisant le cout du sinistre avec les variables sélectionnées de notre jeu de données on obtient l'équation de régression suivante :

$$\boldsymbol{Cout\_sinistre = \beta 0 + \beta 1 zone + \beta 2 puissance + \beta 3 age\_conducteur + \beta 4 age\_conducteur + \beta 5 bonus + \beta 6 energie + \beta 7 densite + \beta 8 fequence\_sinistre + \beta 9 garantie + \varepsilon}$$

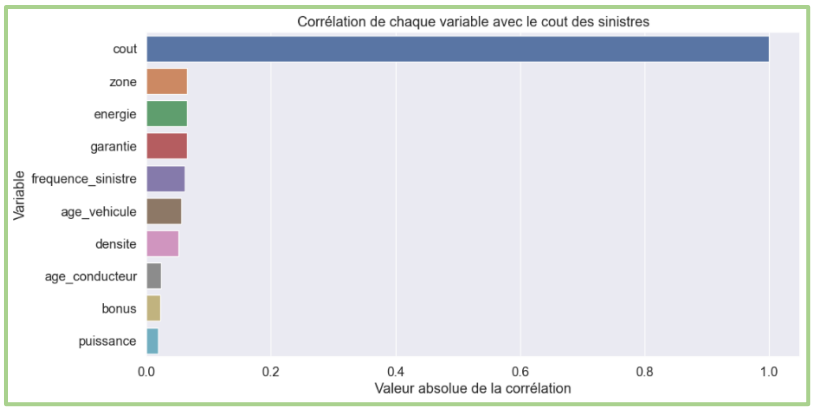


Figure 19 *: Corrélation de chaque variable avec le coût des sinistres.*

La visualisation des corrélations confirme que le coût des sinistres est faiblement corrélé linéairement aux autres variables explicatives. Cette observation justifie l'utilisation de modèles économétriques plus souples, capables de capter des relations complexes entre variables.

4. Test de validité du modèle et estimation des coefficients

OLS Regression Results

Dep. Variable:	cout	R-squared:	0.004
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	1.483
Date:	Tue, 16 Jul 2024	Prob (F-statistic):	0.168
Time:	00:11:32	Log-Likelihood:	-6130.8
No. Observations:	2765	AIC:	1.228e+04
Df Residuals:	2757	BIC:	1.232e+04
Df Model:	7		
Covariance Type :	nonrobust		

	Coef	std err	t	P>|t|	[0.025	0.975]
Const	5.2886	0.385	13.749	0.000	4.534	6.043
Zone	-0.0045	0.012	-0.383	0.702	-0.028	0.019
Puissance	0.0255	0.026	0.990	0.322	-0.025	0.076
Age_vehicule	-0.0042	0.010	-0.436	0.663	-0.023	0.015
Age_conducteur	0.0033	0.004	0.902	0.367	-0.004	0.011
Bonus	0.0019	0.003	0.569	0.569	-0.005	0.008
Energie	-0.0045	0.012	-0.383	0.702	-0.028	0.019
Densite	0.0017	0.002	1.123	0.262	-0.001	0.005
Frequence_sinistre	-0.1345	0.050	-2.701	0.007	-0.232	-0.037
Garantie	-0.0045	0.012	-0.383	0.702	-0.028	0.019

Omnibus :	627.830	Durbin-Watson :	2.013
Prob (Omnibus) :	0.000	Jarque-Bera (JB):	1157.267
Skew:	-1.427	Prob(JB):	5.04e-252
Kurtosis:	4.379	Cond. No.	7.73e+33

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.88e-61. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Ces résultats proviennent d'une régression OLS (Ordinary Least Squares) appliquée à votre modèle de coût (variable dépendante) en fonction de plusieurs variables indépendantes. Voici une interprétation des principaux résultats :

Statistiques du modèle

- **R-carré (R-squared)** : 0.004

  - Cela signifie que seulement environ 0.4% de la variance dans la variable dépendante (coût) est expliquée par les variables indépendantes incluses dans le modèle. Un R-carré faible suggère que le modèle ne capture pas bien la variabilité des données.
- **R-carré ajusté (Adj. R-squared)** : 0.001
  - Il est légèrement plus bas que le R-carré, indiquant que l'ajout de variables indépendantes n'a pas amélioré de manière significative l'ajustement du modèle.
- **F-statistique** : 1.483
  - Avec une p-valeur de 0.168, cette statistique ne fournit pas suffisamment de preuves pour rejeter l'hypothèse nulle que tous les coefficients des variables indépendantes sont nuls simultanément. Cela suggère que le modèle dans son ensemble pourrait ne pas être significatif.

Coefficients des variables indépendantes

- Les coefficients indiquent l'effet attendu sur la variable dépendante (coût) pour chaque unité de changement dans la variable indépendante, toutes choses étant égales par ailleurs.
- **Significativité des coefficients** :
  - Seul le coefficient pour **frequence_sinistre** est statistiquement significatif à un niveau de signification de 5% (p-valeur = 0.007), indiquant que la fréquence des sinistres a un effet significatif négatif sur le coût.
- Les autres coefficients (zone, puissance, âge du véhicule, âge du conducteur, bonus, énergie, densité, garantie) ne sont pas statistiquement significatifs à un niveau de 5% (p-valeurs > 0.05).

Autres remarques

- **Omnibus** : Le test Omnibus indique une significativité globale du modèle (p-valeur < 0.05), suggérant que certaines des variables indépendantes en ensemble ont un effet sur la variable dépendante.
- **Multicollinéarité** : Le modèle pourrait souffrir de problèmes de multicollinéarité, ce qui peut affecter la précision des estimations des coefficients.

En résumé, ce modèle montre des faibles capacités prédictives (R-carré faible) et la plupart des variables indépendantes ne semblent pas avoir un effet significatif sur le coût des sinistres, à l'exception de la fréquence des sinistres.

Tableau 5 *: Interprétation des métriques*

Métrique	Train	Test
R2	0.012993	0.019092
MSE	939276.552089	830037.142980
RMSE	969.162810	911.063743

Voici l'interprétation des métriques de performance pour les ensembles de données d'entraînement et de test :

1. *R^2 (Coefficient de Détermination)*

- *Train* : 0.012993
- *Test* : 0.019092

Interprétation :

- Le R^2 mesure la proportion de la variance des données expliquée par le modèle. Une valeur de 0 indique que le modèle n'explique aucune variance, tandis qu'une valeur de 1 indique que le modèle explique toute la variance.

- Les valeurs proches de 0 pour les ensembles d'entraînement et de test suggèrent que le modèle ne capture pas bien la variance des données. Le modèle est donc très peu performant.

2. *MSE (Mean Squared Error - Erreur Quadratique Moyenne)*

- *Train* : 939276.552089
- *Test* : 830037.142980

Interprétation :

- Le MSE mesure la moyenne des carrés des erreurs. Plus le MSE est bas, mieux le modèle prédit les valeurs réelles.
- Le MSE est légèrement plus bas pour le test que pour l'entraînement, ce qui peut indiquer une légère amélioration en généralisation, mais les valeurs sont toutes deux très élevées, ce qui suggère que le modèle a des erreurs importantes dans ses prédictions.

3. *RMSE (Root Mean Squared Error - Erreur Quadratique Moyenne Racine)*

- *Train* : 969.162810
- *Test* : 911.063743

Interprétation :

- Le RMSE est la racine carrée du MSE et fournit une mesure de l'erreur en unités des valeurs de la variable cible. Il est plus facile à interpréter que le MSE car il est dans la même unité que les données.
- Le RMSE est légèrement plus bas pour le test que pour l'entraînement, indiquant une légère amélioration sur les données de test. Cependant, les deux valeurs de RMSE sont élevées, ce qui signifie que les erreurs de prédiction du modèle sont importantes.

Conclusion

- *Performance Générale* : Le modèle a une faible capacité à expliquer la variance des données (R^2 proche de 0), et les erreurs (MSE et RMSE) sont élevées, indiquant que le modèle a du mal à prédire les valeurs de manière précise, tant sur les données d'entraînement que sur les données de test.
- *Overfitting / Underfitting* : Les valeurs de MSE et RMSE sont relativement proches entre l'entraînement et le test, ce qui suggère que le modèle ne surajuste pas particulièrement les données d'entraînement, mais qu'il pourrait ne pas être bien ajusté du tout (underfitting).

Il pourrait être nécessaire d'explorer des modèles différents, de faire des ajustements de caractéristiques, ou d'optimiser les hyperparamètres pour améliorer la performance du modèle.

***Tableau 6**: Les valeurs des facteurs d'inflation.*

Index	Variable	VIF
0	Const	82.629525
1	Zone	Inf
2	Puissance	1.005274
3	Age véhicule	1.021108
4	Age conducteur	1.305906
5	Bonus	1.318651
6	Energie	Inf
7	Densité	1.009273
8	Frequence sinistre	1.042764
9	Garantie	Inf

Ces résultats présentent les valeurs du facteur d'inflation de la variance (VIF) pour chaque variable du modèle de régression. Voici une interprétation :

- **VIF (Facteur d'Inflation de la Variance)** : Le VIF mesure l'ampleur à laquelle la variance d'un coefficient de régression est augmentée en raison de la corrélation avec d'autres variables dans le modèle. Un VIF élevé (généralement supérieur à 10) suggère une forte corrélation entre la variable correspondante et les autres

variables explicatives dans le modèle, ce qui peut entraîner une instabilité dans les estimations des coefficients.

- **Interprétation des valeurs de VIF** :
    - Pour les variables "**constante**", "**puissance**", "**age_vehicule**", "**age_conducteur**", "**bonus**", "**densite**" et "**frequence_sinistre**", les valeurs de VIF sont proches de 1, ce qui indique qu'il n'y a pas de forte corrélation avec d'autres variables dans le modèle.
    - Les variables "**zone**", "**energie**" et "**garantie**" ont des valeurs de VIF infinies, ce qui suggère une forte corrélation avec d'autres variables dans le modèle. Une valeur infinie indique souvent une parfaite multicollinéarité, ce qui peut entraîner des estimations peu fiables des coefficients.

En résumé, ces résultats suggèrent qu'il existe une forte multicollinéarité entre les variables "**zone**", "**energie**" et "**garantie**" et les autres variables dans le modèle. Cela pourrait affecter la fiabilité des estimations des coefficients de régression associées à ces variables et nécessiterait une attention particulière lors de l'interprétation des résultats du modèle.

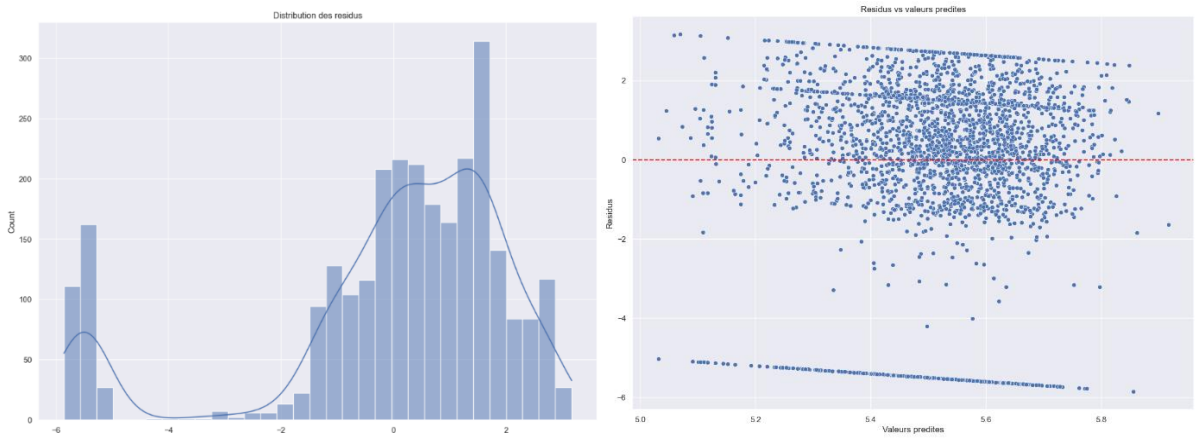


***Figure 20**: Analyses des résidus.*

Shapiro-Wilk test: Shapiro Result (statistic=0.823623776435852, pvalue=0.0)

D'Agostino's K^2 test: NormaltestResult (statistic=627.8298323143897, pvalue=4.661053907567824e-137)

Ces résultats proviennent de tests de normalité des résidus d'un modèle statistique. Voici une interprétation :

- **Test de Shapiro-Wilk** : Ce test évalue si un échantillon de données suit une distribution normale. La statistique de test est proche de 1 pour les données normalement distribuées. Une p-value inférieure au niveau de signification (généralement 0,05) indique un rejet de l'hypothèse nulle selon laquelle les données suivent une distribution normale. Dans ce cas, la p-value est inférieure à 0.05 (0.0), ce qui suggère que les résidus ne suivent pas une distribution normale.
- **Test de D'Agostino K^2** : Ce test est une autre méthode pour évaluer la normalité des résidus. Il combine l'asymétrie et l'aplatissement (kurtosis) pour calculer la statistique de test. Une p-value faible (généralement inférieure à 0,05) indique un rejet de l'hypothèse nulle de normalité des données. Dans ce cas, la p-value est très proche de zéro (2.78129952917137l6e-138), ce qui confirme le rejet de l'hypothèse de normalité des résidus.

En résumé, ces résultats suggèrent que les résidus du modèle ne suivent pas une distribution normale, ce qui peut avoir des implications sur l'interprétation des résultats du modèle et nécessiterait une

attention particulière lors de l'application de méthodes statistiques qui supposent une distribution normale des résidus.

Breusch-Pagan test: (52.834359965349385, 3.1420552195124955e-08, 7.672536571318157, 3.3246644100172976e-09)

Les résultats que vous avez fournis semblent être issus du test de **Breusch-Pagan**, utilisé pour détecter l'hétéroscédasticité dans les résidus d'un modèle de régression. Voici comment interpréter ces résultats :

1. **Statistique de test** : 52.834359965349385
    - Il s'agit de la statistique de test du test de **Breusch-Pagan**. Elle mesure à quel point les résidus du modèle présentent des signes d'hétéroscédasticité.
2. **P-valeur** : 3.1420552195124955e-08 (en notation scientifique, environ 0.00000003142)
    - La p-valeur est très faible, indiquant que les résidus présentent une hétéroscédasticité significative.
3. **Degrés de liberté** : 7.672536571318157
    - Cela fait référence aux degrés de liberté du test.
4. **P-valeur ajustée** : 3.3246644100172976e-09 (en notation scientifique, environ 0.00000000332)
    - C'est une p-valeur ajustée qui confirme la significativité de l'hétéroscédasticité dans les résidus.

Interprétation

- Étant donné que la p-valeur est très faible (beaucoup plus petite que le seuil de 0.05 généralement utilisé), nous rejetons l'hypothèse nulle selon laquelle il n'y a pas d'hétéroscédasticité dans les résidus. En d'autres termes, il y a une forte indication que les résidus de votre modèle de régression sont hétéroscédastiques.

Conséquences

- L'hétéroscédasticité peut affecter la précision des estimations des paramètres du modèle, en particulier des intervalles de confiance et des tests d'hypothèses sur les coefficients.

Actions recommandées

- Considérez l'utilisation de techniques de régression robustes à l'hétéroscédasticité, telles que les méthodes de régression robustes ou les transformations des variables.
- Effectuez d'autres diagnostics pour mieux comprendre la nature et l'impact de l'hétéroscédasticité sur vos résultats de régression.

Durbin-Watson test : 2.012625096723772

Le test de Durbin-Watson est une mesure de l'autocorrélation des résidus dans un modèle de régression. Voici une interprétation du résultat :

La statistique de test de **Durbin-Watson** est de 2.012625096723772. Cette statistique prend des valeurs entre **0** et **4**. Une valeur proche de **2** indique une absence d'autocorrélation des résidus. Dans ce cas, la valeur est légèrement inférieure à **2**, ce qui suggère une légère autocorrélation positive des résidus, mais elle est proche de **2**, ce qui indique que cette autocorrélation est faible.

En résumé, le test de **Durbin-Watson** indique une faible autocorrélation positive des résidus dans le modèle de régression.

PARTI III : REGRESSION DE POISON

La régression **GLM** classique utilisé pour modéliser une variable de comptage est la régression de Poisson. Pour rappel, la distribution de Poisson se caractérise par un seul paramètre λ. Ce paramètre représente la moyenne des événements observés sur une période donnée, ainsi que la dispersion de la distribution. Par conséquent, λ doit être un nombre strictement positif, car un nombre négatif d'événements n'est pas possible. Il est donc essentiel d'employer une fonction de lien pour restreindre l'équation de régression à l'intervalle **[0,** $+\infty$ **[**. La fonction la plus couramment utilisée à cet effet est le logarithme naturel **(log)**, dont la fonction réciproque est l'exponentielle **(exp)**.

1. Présentation de la loi de Poisson

Considérons un nombre réel et **Y** une variable aléatoire réelle. Cela est vrai si et seulement si, pour tout entier naturel **k**, on a : $Pr\ (Y = k) = e\text{-}\lambda \cdot \lambda k$ **!**
NOTA BENE : on a, **E(Y)=V(Y)=**λ.
Présentation du modèle de régression de Poisson $log(\lambda) = \alpha + \beta 1X1 + \ldots + \beta iXi + \ldots + \beta kXk$ Description de l'équation de régression de Poisson :
λ : nombre de cas attendus ; α : Ordonnée à l'origine (intercepte) ;
βi : Coefficient associé à la ième variable explicative Xi. Les coefficients du modèle reflètent l'impact d'une variation d'une unité des variables **X** sur λ sur une échelle logarithmique. À noter que l'échelle logarithmique est de nature multiplicative : en convertissant les coefficients de cette échelle à leur échelle d'origine à l'aide de la fonction exponentielle, leur effet devient multiplicatif plutôt qu'additif.
Pour notre cas, nous souhaitons modéliser la fréquence des sinistres.

1. Test de validité du modèle et estimation des coefficients

Generalized Linear Model Regression Results

Dep. Variable:	frequence_sinistre	No. Observations:	2756
Model:	GLM	Df Residuals:	2748
Model Family:	Poisson	Df Model:	7
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3779.3
Date:	Wed, 03 Jul 2024	Deviance:	1045.3
Time:	14:30:49	Pearson chi2:	1.16e+03
No. Iterations:	4	Pseudo R-squ. (CS):	0.01866
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Const	0.5519	0.129	4.267	0.000	0.298	0.805
Zone	-0.0205	0.004	-4.914	0.000	-0.029	-0.012
Puissance	0.0068	0.009	0.760	0.447	-0.011	0.024
Age_véhicule	-0.0146	0.003	-4.307	0.000	-0.021	-0.008
Age_conducteur	-0.0006	0.001	-0.488	0.626	-0.003	0.002
Bonus	0.0014	0.001	1.271	0.204	-0.001	0.004
Energie	-0.0205	0.004	-4.914	0.000	-0.029	-0.012
Densite	0.0002	0.001	0.431	0.666	-0.001	0.001
Garantie	-0.0205	0.004	-4.914	0.000	-0.029	-0.012
Cout	2.101e-05	1.47e-05	1.428	0.153	-7.82e-06	4.98e-05

Voici une interprétation détaillée des résultats du modèle de régression linéaire généralisée (**GLM**) utilisant la famille Poisson et la fonction de lien logarithmique pour prédire la fréquence des sinistres.

Résumé des résultats

- **Nombre d'observations (No. Observations)** : 2756
- **Degrés de liberté résiduels (Df Résiduels)** : 2748
- **Degrés de liberté du modèle (Df Model)** : 7
- **Methode (Method)**: IRLS (Iteratively Reweighted Least Squares)
- **Log-Vraisemblance (Log-Likelihood)** : -3779.3
- **Déviance (Deviance)** : 1045.3
- **Chi-carré de Pearson (Pearson chi2)** : 1.16e+03
- **Pseudo R-carré (CS)** : 0.01866
- **Nombre d'itérations (No. Iterations)** : 4

Coefficients et leur interprétation

- **Constante (const)** :
    - Coefficient : 0.5519
    - P-valeur : 0.000 (significatif)
    - Interprétation : La **valeur** de la constante est statistiquement significative, indiquant une influence de la constante sur la **fréquence des sinistres**.
- **Zone** :
    - Coefficient : -0.0205
    - P-valeur : 0.000 (significatif)
    - Interprétation : Pour chaque unité supplémentaire de la variable "**zone**", la fréquence des sinistres diminue de 0.0205. Cela signifie qu'une **zone** différente peut influencer négativement la **fréquence des sinistres**.
- **Puissance** :
    - Coefficient : 0.0068
    - P-valeur : 0.447 (non significatif)
    - Interprétation : La **puissance** du véhicule n'a pas d'effet significatif sur la **fréquence des sinistres**.
- **Âge du véhicule (Age_véhicule)** :
    - Coefficient : -0.0146
    - P-valeur : 0.000 (significatif)
    - Interprétation : Pour chaque année supplémentaire de **l'âge du véhicule**, la **fréquence des sinistres** diminue de 0.0146.
- **Âge du conducteur (Age_conducteur)** :
    - Coefficient : -0.0006
    - P-valeur : 0.626 (non significatif)
    - Interprétation : **L'âge du conducteur** n'a pas d'effet significatif sur la **fréquence des sinistres**.
- **Bonus** :
    - Coefficient : 0.0014
    - P-valeur : 0.204 (non significatif)
    - Interprétation : Le **bonus** n'a pas d'effet significatif sur la **fréquence des sinistres**.
- **Énergie** :
    - Coefficient : -0.0205
    - P-valeur : 0.000 (significatif)
    - Interprétation : Le type de carburant utilisé a un effet significatif sur la **fréquence des sinistres**, avec une diminution de 0.0205 pour chaque unité de la variable "**énergie**".
- **Densité** :
    - Coefficient : 0.0002

  - P-valeur : 0.666 (non significatif)
  - Interprétation : La **densité** n'a pas d'effet significatif sur la **fréquence des sinistres**.
- **Garantie** :
  - Coefficient : -0.0205
  - P-valeur : 0.000 (significatif)
  - Interprétation : La **garantie** a un effet significatif sur la **fréquence des sinistres**, avec une diminution de 0.0205 pour chaque unité de la variable "**garantie**".
- **Coût (cout)** :
  - Coefficient : 2.101e-05
  - P-valeur : 0.153 (non significatif)
  - Interprétation : Le **coût** n'a pas d'effet significatif sur la **fréquence des sinistres**.

Conclusion

Les variables "**zone**", "**âge du véhicule**", "**énergie**" et "**garantie**" ont des effets significatifs sur la **fréquence des sinistres**, tandis que les autres variables comme "**puissance**", "**âge du conducteur**", "**bonus**", "**densité**" et "**coût**" n'ont pas d'effet significatif.

Le Pseudo R-carré (**CS**) de 0.01866 indique que le modèle explique environ 1.87% de la variabilité de la **fréquence des sinistres**, ce qui est relativement faible. Cela suggère que d'autres facteurs non inclus dans le modèle pourraient influencer la **fréquence des sinistres**.

Ratio deviance/degrees of freedom: 0.3803990005503866

Le ratio déviance/degrés de liberté est un indicateur utilisé pour évaluer l'ajustement d'un modèle statistique, en particulier dans les modèles de régression généralisée (**GLM**). Voici une interprétation du ratio que vous avez obtenu :

Ratio déviance/degrés de liberté

- **Valeur** : 0.3804

Interprétation

1. **Valeur du ratio** : Un ratio de 0.3804 est inférieur à 1, ce qui indique que la déviance (une mesure de l'ajustement du modèle) est faible par rapport au nombre de degrés de liberté. Cela peut suggérer que le modèle s'ajuste bien aux données.
2. **Comparaison avec 1** : En général, un ratio déviance/degrés de liberté proche de 1 indique un bon ajustement du modèle. Si le ratio est bien inférieur à 1, cela peut indiquer une sous-dispersion des données par rapport au modèle. En revanche, un ratio bien supérieur à 1 indiquerait une surdispersion.
3. **Sous-dispersion** : Dans ce cas, la sous-dispersion (ratio inférieur à 1) peut suggérer que les observations ont moins de variabilité que celle prévue par le modèle. Cela peut également signifier que le modèle pourrait être trop simple et ne capture pas toute la variabilité présente dans les données, ou que les résidus sont trop petits.

Conclusion

Le ratio de déviance/degrés de liberté de 0.3804 suggère que le modèle de régression Poisson que vous avez ajusté est probablement bien ajusté aux données, avec une légère tendance vers la sous-dispersion. Cela signifie que le modèle prédit la fréquence des sinistres avec une variabilité légèrement moindre

que celle observée dans les données. Cependant, il est toujours recommandé de vérifier d'autres diagnostics du modèle pour s'assurer de la robustesse de ces conclusions.

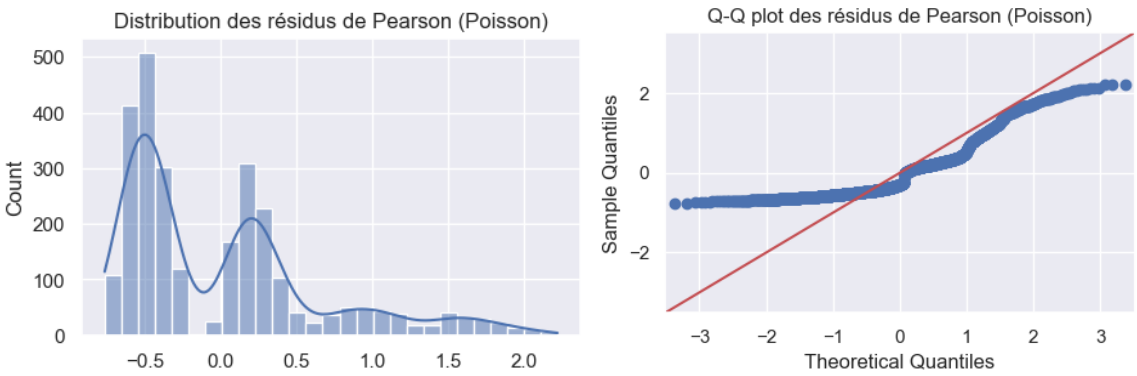


Figure 21 *: Distribution des résidus de Pearson*

La ***Figure 21*** présente la distribution et le Q-Q plot des **résidus de Pearson** du modèle de régression de Poisson appliqué à la variable de fréquence des sinistres (nbre).

- **À gauche**, l'histogramme montre que les résidus ne sont pas parfaitement centrés sur zéro et présentent une **légère asymétrie à droite**, ce qui peut indiquer une certaine **sur dispersion** ou des données hétérogènes.
- **À droite**, le **Q-Q plot** compare les quantiles empiriques aux quantiles théoriques d'une distribution normale. On observe que les points **s'écartent de la diagonale**, surtout dans les queues, ce qui signifie que les résidus ne suivent pas une loi normale, ce qui est attendu dans un modèle de Poisson, mais cela peut aussi signaler que le modèle **ne capture pas parfaitement la structure des données**.

PARTI III : REGRESSION LOGISTIQUE

La **régression logistique** est un modèle statistique utilisé pour prédire la probabilité d'appartenance d'une observation à une **classe binaire** (par exemple : sinistre ou non, fraude ou non, etc.). Contrairement à la régression linéaire qui prédit une variable continue, la régression logistique prédit une **variable catégorielle** (souvent codée 0 ou 1).
Elle est particulièrement utilisée dans les domaines de la finance, de l'assurance, de la médecine et plus généralement dans toute situation de **classification binaire**.

1. Introduction à la régression logistique appliquée aux sinistres automobiles

Dans le cadre de ce projet, nous cherchons à **modéliser la probabilité qu'un contrat automobile donne lieu à un sinistre**. Il s'agit donc d'un **problème de classification binaire** : la variable cible prend la valeur 1 si un sinistre a eu lieu, et 0 sinon.
Pour cela, nous utilisons la **régression logistique**, un outil statistique permettant d'estimer la probabilité d'occurrence d'un événement (ici, un sinistre), en fonction de plusieurs caractéristiques du contrat ou de l'assuré.

2. Formulation du modèle

Le modèle de régression logistique est défini par l'équation suivante :

$$\boldsymbol{P(sinistre = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}}$$

Où :

- $\boldsymbol{P(sinistre = 1 \mid X)}$ Est la **probabilité qu'un sinistre survienne** pour un contrat donné,
- $\boldsymbol{x_1, x_2, \ldots, x_p}$ représentent les **variables explicatives**, comme :
  - La **zone géographique** (zone),
  - La **puissance du véhicule** (puissance),
  - L'âge **du conducteur** (age_conducteur),
  - Le **bonus/malus**, etc.
- $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_p}$ sont les **coefficients** du modèle à estimer.

3. Fonction logit (fonction logistique inverse)

On peut aussi écrire :

$$\boldsymbol{\text{Logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}$$

Cette transformation permet de **convertir une probabilité (entre 0 et 1)** en une **combinaison linéaire** des variables explicatives, que l'on peut modéliser à l'aide d'une régression.

4. Estimation et interprétation

Les paramètres **β** sont estimés à l'aide de la **méthode de vraisemblance maximale**, qui cherche à rendre le modèle le plus cohérent possible avec les données observées.
L'interprétation des coefficients est intuitive :

- Si $\boldsymbol{\beta_j > 0}$, alors une augmentation de x_j **accroît la probabilité** de sinistre.
- Si $\boldsymbol{\beta_j < 0}$, alors une augmentation de x_j **diminue la probabilité** de sinistre.

Par exemple, un coefficient positif pour la variable "âge du véhicule" pourrait indiquer que les **véhicules plus anciens sont plus susceptibles d'avoir un sinistre**.

5. Objectif final

L'objectif de ce modèle est double :

- **Prédire le risque de sinistre** pour de nouveaux contrats,
- **Identifier les facteurs influents** afin d'aider à la tarification, à la segmentation des clients, ou à la détection de comportements à risque.

6. Test de validité du modèle et estimation des coefficients

Logit Regression Results

Dep. Variable:	Frequence_sinistreB	No. Observations:	2204
Model:	Logit	Df Residuals:	2187
Method:	MLE	Df Model:	16
Date:	Thu, 25 Jul 2024	Pseudo R-squ.:	0.07381
Time:	21:54:13	Log-Likelihood:	-1410.9
converged:	False	LL-Null:	-1523.4
Covariance Type:	nonrobust	LLR p-value:	7.028e-39

	coef	std err	z	P>|z|	[0.025	0.975]
Puissance	0.0022	0.025	0.090	0.929	-0.046	0.051
Age véhicule	-0.0422	0.011	-3.971	0.000	-0.063	-0.021
Age conducteur	0.0015	0.003	0.487	0.626	-0.004	0.007
Bonus	0.0017	0.002	0.721	0.471	-0.003	0.006
Densite	-0.0001	0.002	-0.083	0.933	-0.003	0.003
Cout	7.779e-05	4.69e-05	1.657	0.098	-1.42e-05	0.000
Zone_B	0.1385	0.177	0.782	0.434	-0.209	0.486
Zone_C	0.1609	0.143	1.124	0.261	-0.120	0.441
Zone_D	0.2310	0.151	1.530	0.126	-0.065	0.527
Zone_E	0.0600	0.155	0.386	0.699	-0.244	0.364
Zone_F	-0.0722	0.312	-0.231	0.817	-0.684	0.539
Energie_E	-0.2204	0.093	-2.371	0.018	-0.403	-0.038
Garantie_2DO	0.4227	0.111	3.817	0.000	0.206	0.640
Garantie_3VI	0.3071	0.180	1.705	0.088	-0.046	0.660
garantie_4BG	-1.0561	0.117	-9.010	0.000	-1.286	-0.826
garantie_5CO	12.7113	247.514	0.051	0.959	-472.407	497.829
garantie_6CL	-0.0840	1.418	-0.059	0.953	-2.862	2.695

Le tableau des résultats de la régression logistique donne plusieurs informations importantes. Voici une interprétation détaillée des différents éléments :

Informations Générales

- ***Dep. Variable*** : La variable dépendante est frequence_sinistreB.

- ***No. Observations*** : Il y a 2204 observations.
- ***Model*** : Le modèle utilisé est la régression logistique (Logit).
- ***Method*** : La méthode d'estimation utilisée est la méthode du maximum de vraisemblance (**MLE**).
- ***Pseudo R-squ.*** : Le pseudo R-carré est de 0.07381, ce qui indique une mesure de la proportion de la variation expliquée par le modèle.
- ***Log-Likelihood*** : La log-vraisemblance est de -1410.9.
- ***LL-Null*** : La log-vraisemblance du modèle nul est de -1523.4.
- ***LLR p-value*** : La valeur p du test du rapport de vraisemblance est 7.028e-39, ce qui indique que le modèle est statistiquement significatif.
- ***Converged*** : False indique que le modèle n'a pas convergé correctement, ce qui pourrait être un problème sérieux.

Interprétation des Coefficients

Pour chaque variable indépendante, les informations suivantes sont fournies :

- ***Coef*** : Le coefficient estimé pour chaque variable.
- ***Std err*** : L'erreur standard du coefficient estimé.
- ***Z*** : La statistique z, qui est le ratio du coefficient à son erreur standard.
- ***P>|z|*** : La valeur p associée à la statistique z, indiquant si le coefficient est statistiquement significatif.
- *[0.025, 0.975]* : Les intervalles de confiance à 95% pour le coefficient.

Voici quelques points clés :

- ***Age_vehicule*** : Le coefficient est -0.0422, avec une valeur p de 0.000, indiquant que cette variable est significativement associée à la variable dépendante. Un âge plus élevé du véhicule diminue la fréquence des sinistres.
- ***Energie*** : Le coefficient est -0.2204, avec une valeur p de 0.018, indiquant que cette variable est également significative. Cela suggère que les véhicules à énergie alternative ont une fréquence de sinistres différente par rapport à la catégorie de référence.
- ***Garantie_2DO*** : Le coefficient est 0.4227, avec une valeur p de 0.000, montrant une association positive significative avec la fréquence des sinistres.
- ***Garantie_4BG*** : Le coefficient est -1.0561, avec une valeur p de 0.000, indiquant une association négative significative.
- ***Garantie_5CO*** : Le coefficient est extrêmement élevé (12.7113) mais avec une énorme erreur standard et une valeur p de 0.959, ce qui signifie qu'il n'est pas statistiquement significatif et pourrait indiquer des problèmes de colinéarité ou des valeurs aberrantes.
- ***Autres variables*** : Plusieurs autres variables ne sont pas significatives (par exemple, puissance, age_conducteur, bonus, densite), ce qui signifie que leurs coefficients ne sont pas statistiquement différents de zéro.

Conclusion

Bien que certains coefficients soient significatifs, le fait que le modèle n'ait pas convergé correctement est préoccupant. Cela pourrait indiquer plusieurs problèmes potentiels, tels que la multicolinéarité, des données aberrantes, ou simplement que le modèle nécessite plus d'itérations pour converger. Vous pourriez essayer de résoudre ces problèmes en examinant les données de plus près, en supprimant les variables non significatives, ou en ajustant les paramètres de l'algorithme d'optimisation.

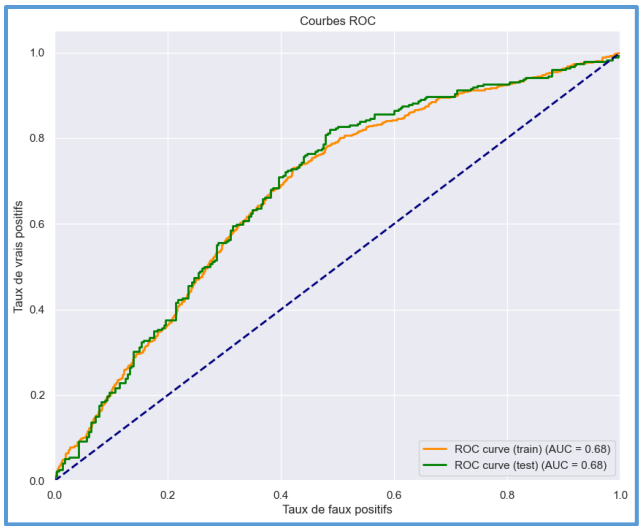


Figure 22 *: Courbe de ROC.*

Le modèle de régression logistique présente une performance correcte (AUC = 0.68), ce qui est acceptable dans un contexte où les données peuvent être complexes ou bruitées. Il généralise bien et ne montre pas de surapprentissage, mais il pourrait être **amélioré** pour mieux distinguer les classes (par exemple en ajoutant des variables pertinentes ou en testant d'autres algorithmes).

Tableau 7: Métriques de performance.

Classe	Précision	Rappel	F1-score	Support
0	0.66	0.61	0.64	280
1	0.63	0.68	0.65	272

Mesure	Valeur
Exactitude	0.65
Macro Moyenne	0.65
Moyenne Pondérée	0.65

Ce tableau présente les métriques de performance pour un modèle de classification binaire, probablement une régression logistique, évalué sur un ensemble de test. Voici comment interpréter les différentes métriques :

Interprétation des métriques

Classe 0

- ***Precision (Précision)*** : (0.66)
  - La précision mesure la proportion de vrais positifs parmi les exemples prédits positifs. Ici, 66% des exemples prédits comme appartenant à la classe 0 étaient effectivement de la classe 0.
- ***Recall (Rappel)*** : (0.61)
  - Le rappel mesure la proportion de vrais positifs parmi tous les exemples de la classe 0. Ici, 61% des exemples de la classe 0 ont été correctement identifiés.
- ***F1-score*** : (0.64)
  - Le F1-score est la moyenne harmonique de la précision et du rappel, donnant une mesure équilibrée de la performance du modèle. Ici, il est de 0.64.
- ***Support*** : (280)
  - Le support représente le nombre total d'exemples de la classe 0 dans l'ensemble de test.

Classe 1

- ***Precision (Précision)*** : (0.63)
  - La précision pour la classe 1 est de 63%.
- ***Recall (Rappel)*** : (0.68)
  - Le rappel pour la classe 1 est de 68%.

- ***F1-score*** : (0.65)
  - Le F1-score pour la classe 1 est de 0.65.
- ***Support*** : (272)
  - Le support pour la classe 1 est de 272.

Total (Global)

- ***Accuracy (Exactitude)*** : (0.65)
  - L'exactitude est la proportion d'exemples correctement classifiés sur l'ensemble total. Ici, 65% des exemples sont correctement classifiés.
- ***Macro avg (Moyenne macro)*** : (0.65)
  - La moyenne macro est la moyenne non pondérée des métriques (précision, rappel, F1-score) pour chaque classe. Elle traite toutes les classes de manière égale. Ici, elle est de 0.65 pour toutes les métriques.
- ***Weighted avg (Moyenne pondérée)*** : (0.65)
  - La moyenne pondérée prend en compte le support de chaque classe lors du calcul des métriques globales. Elle est également de 0.65 pour toutes les métriques.

Conclusion

Les performances du modèle sont équilibrées entre les deux classes avec une précision, un rappel et un F1-score global de 0.65. Cela signifie que le modèle est modérément bon pour prédire les deux classes, mais il y a probablement encore de la place pour l'amélioration. Vous pourriez envisager d'ajuster les hyperparamètres du modèle, de traiter les déséquilibres de classe, ou d'explorer d'autres algorithmes de classification pour améliorer les performances.

Ce tableau est une matrice de confusion pour un modèle de classification binaire. Elle présente le nombre de prédictions correctes et incorrectes pour chaque classe. Voici comment interpréter ce tableau :

Interprétation de la Matrice de Confusion

Tableau 8: Matrice de confusion.

	Prédiction : **Classe** 0	**Prédiction** : **Classe** 1
Réel : Classe 0	172	108
Réel : Classe 1	87	185

- *172 (Vrais Négatifs)* :
  - Nombre d'exemples de la classe 0 correctement prédits comme classe 0.
- *108 (Faux Positifs)* :
  - Nombre d'exemples de la classe 0 incorrectement prédits comme classe 1.
- *87 (Faux Négatifs)* :
  - Nombre d'exemples de la classe 1 incorrectement prédits comme classe 0.
- *185 (Vrais Positifs)* :
  - Nombre d'exemples de la classe 1 correctement prédits comme classe 1.

Calcul des Métriques de Performance

À partir de cette matrice de confusion, nous pouvons calculer les principales métriques de performance.

Précision (Precision)

- *Classe 0* : $\frac{172}{172+87} = 0.66$
- *Classe 1* : $\frac{185}{185+87} = 0.63$

Rappel (Recall)

- *Classe 0* : $\frac{172}{172+108} = 0.61$
- *Classe 1* : $\frac{185}{185+87} = 0.66$

F1-score

- *Classe 0*: $\frac{2*0.66*0.61}{0.66+0.61} = 0.64$
- *Classe 1*: $\frac{2*0.63*0.68}{0.63+0.68} = 0.65$

Exactitude (Accuracy)

- $\frac{172+185}{172+108+87+185} = 0.65$

Conclusion

Ces calculs confirment les valeurs précédentes fournies pour la précision, le rappel et le F1-score. Le modèle a une performance équilibrée avec une précision globale, un rappel et un F1-score de 0.65, indiquant une performance modérée pour la prédiction des deux classes.

Pour améliorer ces résultats, vous pourriez :

- Explorer d'autres algorithmes de classification.
- Ajuster les hyperparamètres de votre modèle actuel.
- Gérer tout déséquilibre de classes potentiellement présent dans vos données.
- Envisager d'ajouter ou de transformer des caractéristiques (features) pour améliorer la capacité du modèle à distinguer entre les classes.

Tableau 9: Métriques de performance

Classe	Précision	Rappel	F1-score	Support
0	0.68	0.62	0.65	1171
1	0.61	0.67	0.64	1033

Mesure	Valeur
Exactitude	0.64
Macro Moyenne	0.64
Moyenne Pondérée	0.65

Ce tableau présente les métriques de performance pour un modèle de classification binaire, basé sur un ensemble de test de 2204 observations. Voici une interprétation détaillée des différentes métriques :

Interprétation des Métriques

Classe 0

- ***Precision (Précision)*** : (0.68)
  - La précision mesure la proportion de vrais positifs parmi les exemples prédits positifs. Ici, 68% des exemples prédits comme appartenant à la classe 0 étaient effectivement de la classe 0.
- ***Recall (Rappel)*** : (0.62)

  - Le rappel mesure la proportion de vrais positifs parmi tous les exemples de la classe 0. Ici, 62% des exemples de la classe 0 ont été correctement identifiés.
- ***F1-score*** : (0.65)
  - Le F1-score est la moyenne harmonique de la précision et du rappel, donnant une mesure équilibrée de la performance du modèle. Ici, il est de 0.65.
- ***Support*** : (1171)
  - Le support représente le nombre total d'exemples de la classe 0 dans l'ensemble de test.

Classe 1

- ***Precision (Précision)*** : (0.61)
  - La précision pour la classe 1 est de 61%.
- ***Recall (Rappel)*** : (0.67)
  - Le rappel pour la classe 1 est de 67%.
- ***F1-score*** : (0.64)
  - Le F1-score pour la classe 1 est de 0.64.
- ***Support*** : (1033)
  - Le support pour la classe 1 est de 1033.

Conclusion

Les performances du modèle sont équilibrées entre les deux classes avec des précisions, des rappels et des F1-scores globalement similaires autour de 0.64-0.65. L'exactitude globale est de 0.64, indiquant que le modèle est modérément bon pour prédire les deux classes, mais il y a probablement encore de la place pour l'amélioration.

Actions potentielles pour l'amélioration

1. *Ajustement des hyperparamètres* :
   - Ajustez les hyperparamètres du modèle actuel pour voir si vous pouvez améliorer les performances.
2. *Ensemble de modèles (Ensembling)* :
   - Envisagez d'utiliser des techniques d'ensemble comme le bagging, le boosting ou les forêts aléatoires pour améliorer les performances.
3. *Feature Engineering* :
   - Ajoutez, modifiez ou supprimez des caractéristiques (features) pour voir si cela améliore les performances du modèle.
4. *Gestion des déséquilibres de classe* :
   - Si vos classes sont déséquilibrées, essayez des techniques de sous-échantillonnage, de sur-échantillonnage ou d'utilisation de pénalités pour la classe minoritaire.
5. *Exploration d'autres algorithmes de classification* :
   - Essayez d'autres algorithmes de classification comme les **SVM**, les réseaux neuronaux, ou les modèles de gradient boosting pour voir si vous pouvez obtenir de meilleures performances.

En utilisant ces stratégies, vous pouvez potentiellement améliorer les métriques de performance de votre modèle.

Synthèse des modèles

Modèle	Variable cible	Type de modèle	R^2 / AUC	Performance
Linéaire	Coût (cout)	OLS	$R^2 \approx 0.004$	Faible
Poisson	Fréquence (nbre)	GLM (Poisson)	Pseudo-$R^2 \approx 0.018$	Moyenne
Logistique	Sinistre binaire	Logit	AUC = 0.68	Acceptable

L'évaluation des trois modèles économétriques appliqués à des variables cibles différentes permet de comparer leur performance et leur pertinence. Le modèle linéaire (OLS) utilisé pour prédire le **coût des sinistres** présente un coefficient de détermination très faible ($R^2 \approx 0.004$), indiquant que moins de 0,4 % de la variabilité du coût est expliquée par les variables explicatives du modèle. Cette performance est jugée **faible**, rendant ce modèle peu exploitable à des fins prédictives.
En ce qui concerne la **fréquence des sinistres**, le modèle de Poisson, basé sur un modèle linéaire généralisé (GLM), affiche un pseudo-R^2 d'environ 0.018. Cela signifie que 1,8 % de la variation observée est expliquée par le modèle. Cette performance est **moyenne**, mais reste limitée en termes de capacité explicative.
Enfin, la **probabilité de survenue d'un sinistre** a été modélisée à l'aide d'une régression logistique (modèle Logit). Le modèle atteint une **AUC (Area Under the Curve)** de **0.68**, ce qui traduit une capacité **acceptable** à discriminer les cas de sinistre des cas sans sinistre. Cette performance, bien que perfectible, reste la plus satisfaisante parmi les trois modèles testés.
En résumé, le modèle logistique s'avère le plus performant, tandis que le modèle linéaire et de Poisson montrent des limites importantes. Ces résultats suggèrent la nécessité d'envisager des alternatives plus robustes, telles que des modèles non linéaires ou à distributions plus adaptées (comme le gamma ou le Tweedie) pour améliorer la qualité des prédictions, notamment sur le coût et la fréquence des sinistres.

Conclusion générale

À l'issue de cette étude, il apparaît que les modèles économétriques permettent de mieux comprendre les mécanismes sous-jacents aux sinistres automobiles. Si la régression linéaire s'est montrée peu performante pour prédire les coûts, les modèles de Poisson et logistique ont permis de dégager des tendances plus solides en matière de fréquence et de probabilité de sinistres.

Les résultats obtenus soulignent l'importance de variables comme l'âge du véhicule, la zone géographique ou encore la nature de la garantie. L'analyse a aussi mis en évidence la complexité du phénomène de sinistre, justifiant l'usage de modèles adaptés aux spécificités des données.

Enfin, ce projet ouvre la voie à des améliorations futures telles que l'intégration de nouvelles variables (historique de conduite, météo, type de trajet), l'usage de modèles non paramétriques ou encore l'exploitation d'algorithmes d'apprentissage automatique.