MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE RECHERCHE SCIENTIFIQUE

REPUBLIQUE DE COTE D'IVOIRE

Institut Supérieur de Statistique
d'Econométrie et de Data Science

MASTER 2

STATISTIQUE – ECONOMETRIE – DATA SCIENCE

Mini-projet

Econométrie

MODELISATION DE COUT DE PROBABILITE DE LA SURVENANCE FREQUENCE DU SINISTRE

2023 – 2024

Etudiant

WAWA LAURALIE MARIE MICHELLE

Enseignant – Encadreur

AKPOSSO DIDIER MARTIAL

AVANT PROPOS

Cher lecteur,

C'est avec un grand enthousiasme que je vous présente ce mini-projet, fruit de Plusieurs semaines de réflexion, de recherche et de travail acharné. Ce mini-Projet est le résultat de mon engagement envers l'apprentissage, la créativité et la mise en pratique des connaissances acquises.

Ce mini-projet est de démontrer ma capacité à appliquer les compétences et les concepts que j'ai acquis dans le cadre de l'obtention du diplôme de Master Professionnel en StatistiqueEconométrie-Machine Learning à l'institut Supérieur de Statistique d'Econométrie et de Data Science (INSSEDS). Le sujet abordé dans le cadre de ce mini-projet me semble à la fois stimulant et pertinent.

Dans le contexte économique et social actuel, l'industrie de l'assurance joue un rôle crucial en offrant protection et sécurité financière à ses clients. Spécifiquement, le secteur de l'assurance automobile est d'une importance capitale car il touche directement à la mobilité individuelle, un aspect fondamental de la vie quotidienne moderne. Ce rapport s'intéresse à une analyse détaillée des données relatives à 2765 clients d'une société d'assurance IARD (Incendie, Accidents et Risques Divers) spécialisée dans l'assurance auto.

Ces données, qui couvrent divers aspects tels que la puissance du véhicule, l'âge du véhicule et du conducteur, la zone géographique, la marque du véhicule, et bien d'autres paramètres, offrent une opportunité unique d'explorer les facteurs qui influencent les coûts et les décisions en matière de polices d'assurance. L'analyse de ces informations permettra non seulement d'améliorer la compréhension des risques associés mais aussi de mieux adapter les produits d'assurance aux besoins spécifiques des assurés.

Ce travail a pour but de mettre en lumière les relations entre les différentes variables et leur impact sur les coûts des sinistres. En identifiant les tendances et les anomalies au sein de ces données, nous espérons fournir des insights précieux pour la prise de décision stratégique au sein de la société d'assurance, contribuant ainsi à une gestion plus efficace et à une tarification plus équitable des polices d'assurance auto.

Je tiens à remercier M. AKPOSSO Didier Martial, Directeur des études et Encadreur de ladit formation, les enseignants de l'INSSEDS, ma famille et mes amis qui m'ont soutenu tout au long de la rédaction de ce mini-projet en m'apportant leur expertise, leurs conseils et leur encouragement.

Enfin, je vous invite à parcourir ce document avec attention, en espérant qu'il vous apportera une vision claire et détaillé de mon travail, tout en suscitant votre intérêt et votre réflexion.

Merci de prendre le temps de découvrir ce mini-projet. Je vous souhaite une agréable lecture.

Cordialement

INTRODUCTION GENERALE 4
CONTEXTE ET JUSTIFICATION DE L'ETUDE 5
PROBLEMATIQUE 5
PRINCIPAUX RESULTATS ATTENDUS 5
PREMIERE PARTIE : PRETAITEMENT DES DONNEES 5
I-PRETAITEMENT DES DONNEES 5
1-DESCRIPTION DU JEU DE DONNEE : DICTIONNAIRE DES DONNEES 5
2-Struture des données 6
3-Visualisation du jeu de donnée 6
II-TRAITEMENT DES VALEURS MANQUANTE 7
a-VISUALISATION DES DONNEES MANQUANTES 7
III-TRAITEMENT DES VALEURS ABERRRANTES ET EXTREMES 8
DEUXIEME PARTIE: ANALYSE UNIVARIEE ET BIVARIEE 9
I-ANALYSE UNI VARIEE 9
A-ANALYSES DES VARIABLES QUANTITATIVES 10
1-Analyse de la variable puissance 10
2-Analyse de la variable Age véhicule 11
3-Etude de la variable Age conducteur 12
4-Etude de la variable nombre de sinistre 13
5-Etude de la variable cout 14
B-ANALYSES DES VARIABLES QUALITATIVES 15
1-Analyse de la variable zone 15
2-Analyse de la variable carburant 16
3-Analyse de la variable garantie 17
II-ANALYSE BI VARIEE 17
A-ANALYSES DES VARIABLES QUANTITATIVES 17
1-Matrice de corrélation 18
B-ANALYSES DES VARIABLES QUANTITATIVE ET QUALITATIVE 19
1-Analyse de la variable cout et zone 19
2-Analyse de la variable cout et garantie 19
3-Analyse de la variable cout et marque 19
TROISIEME PARTIE : ANALYSE EN COMPOSANTE PRINCIPALE 20
I-DISTRIBUTION DE L'INERTIE 20
QUATRIEME PARTIE : MODELISATION DU COUT DES SINISTRES 22

I-CONSTRUCTION DU MODELE 22
1-Modèle et sélection automatiques des explicatives 22
2-Anova du model 23
3-Regression sur Les variables retenus 23
a-Analyse du résidu 24
4-Test de la validité du modèle 24
a-Test de linéarité du modèle 24
b-Test d'homoscédasticité 25
CINQUIEME PARTIE : MODELISATION DE LA FREQUENCE DES SINISTRES 26
I-CONTRUCTION DU MODELE 26
1-Anova du model 26
2-Anova du model final 26
3-Resultat final du model 27
4-Les odd ratios et effets marginaux 27
a-Affichage des paramètres estimés 27
b- Les intervalles de confiance 28
c- Les odd ratio 28
d-Les effets marginaux 28
e- Analyse des résidus 29
5-Taux de mauvais classement et Matrice de Confusion 29
SIXIEME PARTIE : MODELISATION DE LA PROBABILITE DE LA SURVENANCE DES SINISTRE 30
I-TEST D'ADEQUATION A LA LOI DE POISSON 30
1-Test d'adéquation 30
II-CONSTRUCTION DU MODELE COMPLET 31
1-Anova du modèle 31
2-Construction du modèle avec les variables significatives 31
III- TABLEAU DE BORD DE LA COMPAGNIE D'ASSURANCE 33
Perspectives 33
CONCLUSION 34
SOURCE DU CODE R 35

INTRODUCTION GENERALE

L'assurance automobile constitue une part significative de l'activité des sociétés d'assurance IARD (Incendie, Accidents et Risques Divers). Ce secteur, tout en étant très compétitif, exige une compréhension approfondie des facteurs de risque associés à chaque police pour optimiser la tarification et la gestion des sinistres. L'évaluation précise de ces risques est d'autant plus pertinente dans un contexte où les consommateurs sont de plus en plus sensibles aux variations de prix et aux offres personnalisées.

Cette étude se penche sur un échantillon de 2765 clients d'une compagnie d'assurance, où divers aspects tels que l'exposition au risque, les caractéristiques des véhicules et des conducteurs, ainsi que les informations géographiques et les détails des sinistres sont enregistrés. Ces données offrent une opportunité inestimable de déceler des patterns et des corrélations qui peuvent aider à prédire les coûts futurs et à ajuster les stratégies de tarification et de couverture.

L'objectif principal de cette analyse est de comprendre comment les différentes variables, telles que la zone géographique, la puissance du véhicule, l'âge du véhicule et du conducteur, influencent la fréquence et le coût des sinistres. Cela aidera la compagnie à optimiser ses pratiques en termes de gestion des risques, de politique de prix, et ultimement, d'amélioration de la satisfaction client.

À travers des techniques statistiques avancées et des analyses exploratoires, ce rapport cherche à fournir des recommandations stratégiques fondées sur les données pour affiner les modèles de souscription et de tarification de l'assurance auto. Ainsi, nous espérons contribuer à l'efficacité opérationnelle et à la compétitivité de la compagnie dans un marché exigeant.

Le présent document s'articulera autour de six grandes parties :

- **PREMIERE PARTIE: PRETRAITEMENT DES DONNEES**
- **DEUXIEME PARTIE: ANALYSE UNIVARIEE ET BIVARIEE**
- **TROISIEME PARTIE: ANALYSE EN COMPOSANTE PRINCIPALE**
- **QUATRIEME PARTIE: MODELISATION DU COUT DES SINISTRES**
- **CINQUIEME PARTIE: MODELISATION DE LA FREQUENCE DES SINISTRES**
- **SIXIEME PARTIE: MODELISATION DE LA PROBABILITE DE LA SURVENANCE DES SINISTRES**

CONTEXTE ET JUSTIFICATION DE L'ETUDE

Dans l'industrie de l'assurance, la capacité à évaluer avec précision le risque et à adapter en conséquence les stratégies de tarification est fondamentale pour garantir la viabilité et la compétitivité des entreprises. L'assurance automobile, en particulier, représente un segment dynamique mais également complexe en raison de la variabilité des facteurs de risque impliqués, allant de l'âge et de l'expérience du conducteur aux spécificités techniques du véhicule assuré.

PROBLEMATIQUE

Quels sont les principaux facteurs de risque associés aux sinistres dans le portefeuille actuel de clients de la société d'assurance ?

Comment les caractéristiques des véhicules et des conducteurs influencent-elles les coûts et la fréquence des sinistres ?

PRINCIPAUX RESULTATS ATTENDUS

L'analyse détaillée du jeu de données des clients de notre compagnie d'assurance vise à produire des résultats significatifs qui auront un impact direct sur la prise de décisions stratégiques et opérationnelles au sein de l'entreprise.

PREMIERE PARTIE : PRETAITEMENT DES DONNEES

I-PRETAITEMENT DES DONNEES

Le prétraitement des données est une étape cruciale dans le processus d'analyse des données, visant à préparer les données brutes en vue de les rendre appropriées pour une analyse ultérieure. Cette partie consiste, dans un premier temps, à importer le jeu de données dans un logiciel statistique. Ensuite, on procèdera à la visualisation du jeu de données afin d'en dégager la structure. Enfin, la dernière étape consistera à apurer le jeu de données. L'apurement de données, souvent appelé "data cleansing" en anglais, est le processus de nettoyage et de correction des données stockées dans une base de données ou un ensemble de données. L'objectif de l'apurement de données est d'identifier et de corriger les incohérences, les erreurs, les doublons et les données obsolètes ou incorrectes afin d'assurer la qualité et la précision des données.

1-DESCRIPTION DU JEU DE DONNEE : DICTIONNAIRE DES DONNEES

Le dictionnaire de données est une documentation détaillée qui répertorie et décrit chacune des variables du jeu de données. Il sert de référence pour comprendre la signification, la

structure et les propriétés des données stockées, ce qui facilite la gestion, la maintenance, l'analyse et l'utilisation des données. Le jeu de données de notre étude se décrit comme suit :

Tableau 1: Dictionnaire des données

VARIABLE	NATURE	DESCRIPTION	MODALITES
EXPOSITION	QUANTITATIVE	Durée pendant laquelle le véhicule a été assuré au cours de l'année	NUMERIQUE
ZONE	QUALITATIVE	Catégorisation géographique de l'adresse de stationnement du véhicule	CATEGORIELLE
PUISSANCE	QUANTITATIVE	Puissance du moteur du véhicule	NUMERIQUE
AGEVEHICULE	QUANTITATIVE	Âge du véhicule en années	NUMERIQUE
AGECONDUCTEUR	QUANTITATIVE	Âge du conducteur principal	NUMERIQUE
BONUS	QUANTITATIVE	Coefficient de réduction-majoration	NUMERIQUE
MARQUE	QUALITATIVE	Marque du véhicule, par exemple, Marque3, Marque12, etc	CATEGORIELLE
CARBURANT	QUALITATIVE	Type de carburant utilisé par le véhicule	CATEGORIELLE
DENSITE	QUANTITATIVE	Densité de population dans la région de stationnement du véhicule	NUMERIQUE
REGION	QUANTITATIVE	Code numérique indiquant la région administrative	CATEGORIELLE
NBRE	QUANTITATIVE	Nombre de sinistres déclarés pendant la période de couverture	NUMERIQUE
GARANTIE	QUALITATIVE	Type de garantie sous laquelle le sinistre a été déclaré	CATEGORIELLE
COUT	QUANTITATIVE	Coût des sinistres en euros	NUMERIQUE

2-Struture des données

Notre jeu de données est constitué de 2765 observations et de 13 variables.

Tableau 2: Structure du jeu de donnée

```
data.frame':   2765 obs. of  13 variables:
 $ exposition   : num  0.74 0.18 0.48 0.27 0.51 0.64 0.64 0.11 0.11 0.1 ...
 $ zone         : Factor w/ 6 levels "A","B","C","D",..: 1 2 3 6 5 4 4 3 3 5 ...
 $ puissance    : int  5 7 9 7 4 10 10 5 5 9 ...
 $ agevehicule  : int  4 8 0 5 0 0 0 0 0 0 ...
 $ ageconducteur: int  31 22 32 39 49 58 58 52 52 78 ...
 $ bonus        : int  64 100 61 100 50 50 50 50 50 50 ...
 $ marque       : Factor w/ 10 levels "Marque1","Marque10",..: 7 6 3 3 3 3 3 3 3 3 ...
 $ carburant    : Factor w/ 2 levels "D","E": 1 2 2 2 2 1 1 2 2 2 ...
 $ densite      : int  21 26 41 11 31 72 72 73 73 72 ...
 $ region       : int  8 0 13 0 13 13 13 13 13 13 ...
 $ nbre         : int  1 1 1 1 1 2 2 2 2 1 ...
 $ garantie     : Factor w/ 6 levels "1RC","2DO","3VI",..: 1 1 4 2 2 1 2 1 2 1 ...
 $ cout         : num  0 0 687.8 96.6 70.9 ...
```

3-Visualisation du jeu de donnée

Le tableau ci-dessous présente les cinq (5) premières et dernières observations de notre jeu de données.

Tableau 3: Visualisation des 5 premières observations

	exposition	zone	puissance	agevehicule	ageconducteur	bonus	marque	carburant	densite	region	nbre	garantie	cout
1	0.74	A	5	4	31	64	Marque3	D	21	8	1	1RC	0.00
2	0.18	B	7	8	22	100	Marque2	E	26	0	1	1RC	0.00
3	0.48	C	9	0	32	61	Marque12	E	41	13	1	4BG	687.82
4	0.27	F	7	5	39	100	Marque12	E	11	0	1	2DO	96.64
5	0.51	E	4	0	49	50	Marque12	E	31	13	1	2DO	70.88

Tableau 4: Visualisation des 5 dernières observations

	exposition	zone	puissance	agevehicule	ageconducteur	bonus	marque	carburant	densite	region	nbre	garantie	cout
2761	0.24	B	7	8	31	54	Marque1	E	52	11	1	4BG	156.23
2762	0.91	C	5	4	32	57	Marque1	D	24	10	1	1RC	74.15
2763	0.16	A	5	17	44	50	Marque2	E	24	13	1	4BG	325.57
2764	0.07	C	5	7	48	50	Marque4	D	24	13	1	4BG	387.04
2765	1.00	A	6	4	45	50	Marque1	D	24	13	1	4BG	231.20

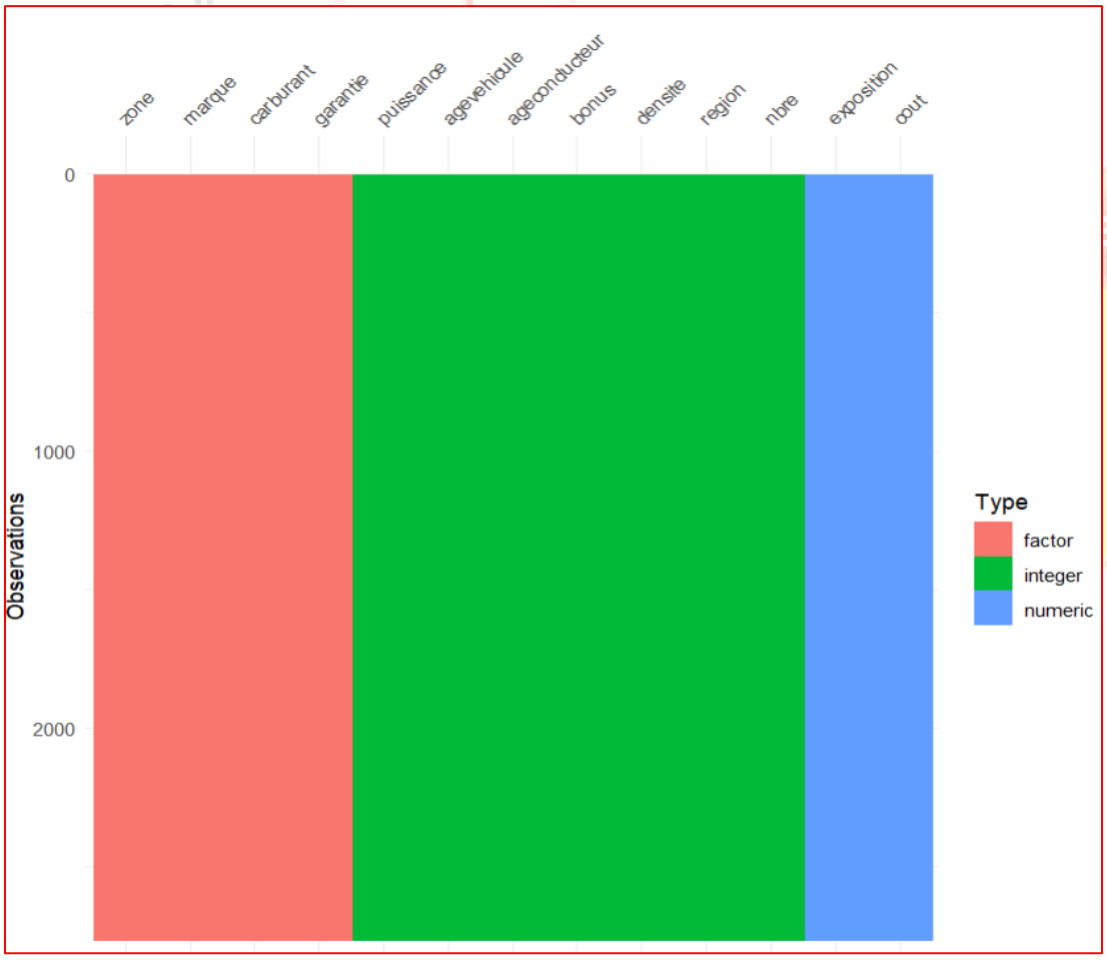
II-TRAITEMENT DES VALEURS MANQUANTE

En statistique, une valeur manquante, également appelée donnée manquante ou observation manquante, fait référence à l'absence d'une valeur pour une variable particulière dans un ensemble de données ou un échantillon.

a-VISUALISATION DES DONNEES MANQUANTES

La visualisation du graphique 1 montre bien que le jeu de données ne contient pas des valeurs manquantes.

Graphique 1: Visualisation des valeurs manquantes



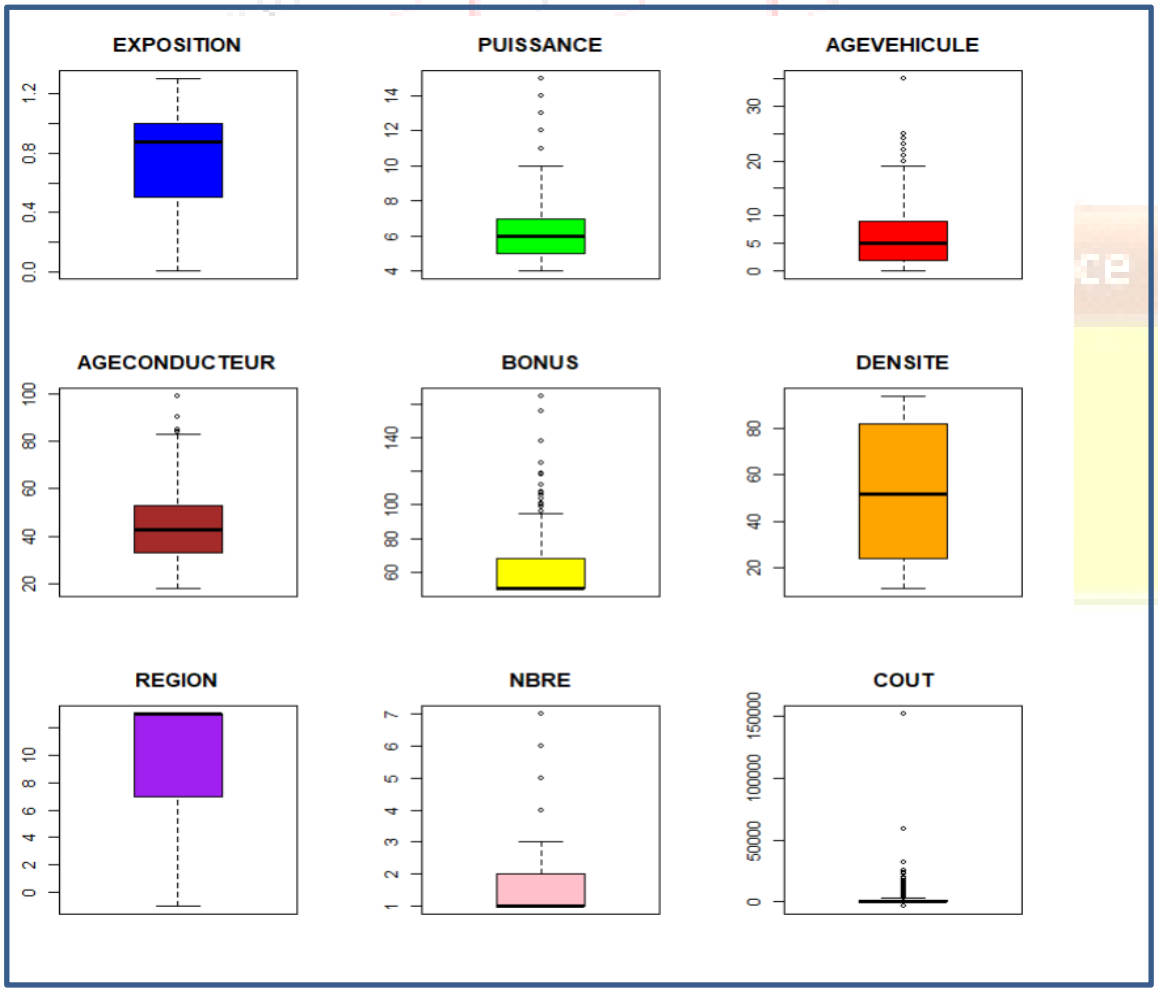
III-TRAITEMENT DES VALEURS ABERRRANTES ET EXTREMES

Les valeurs aberrantes (ou outliers) Ce sont des observations qui sont très différentes de la majorité des autres observations dans un ensemble de données. Elles peuvent résulter d'erreurs de mesure, de saisie de données incorrecte, ou représenter des événements rares ou inhabituels.

Les valeurs aberrantes peuvent fausser les analyses statistiques et doivent souvent être identifiées et traitées de manière appropriée pour éviter des conclusions erronées. Les valeurs extrêmes, Ces valeurs se trouvent à l'extrémité de la distribution des données et représentent les valeurs les plus élevées ou les plus basses dans un ensemble de données. Elles peuvent être importantes pour comprendre la variabilité des données et les tendances générales, mais elles peuvent également être sources de biais si elles ne sont pas correctement gérées dans l'analyse.

Elles se font uniquement avec des variables quantitatives. La visualisation de ces valeurs se fait à l'aide des boîtes à moustache. Une valeur est considérée comme aberrante ou extrême lorsqu'elle est située au-delà des moustaches.

Graphique 2: Visualisation des valeurs aberrantes et extrêmes

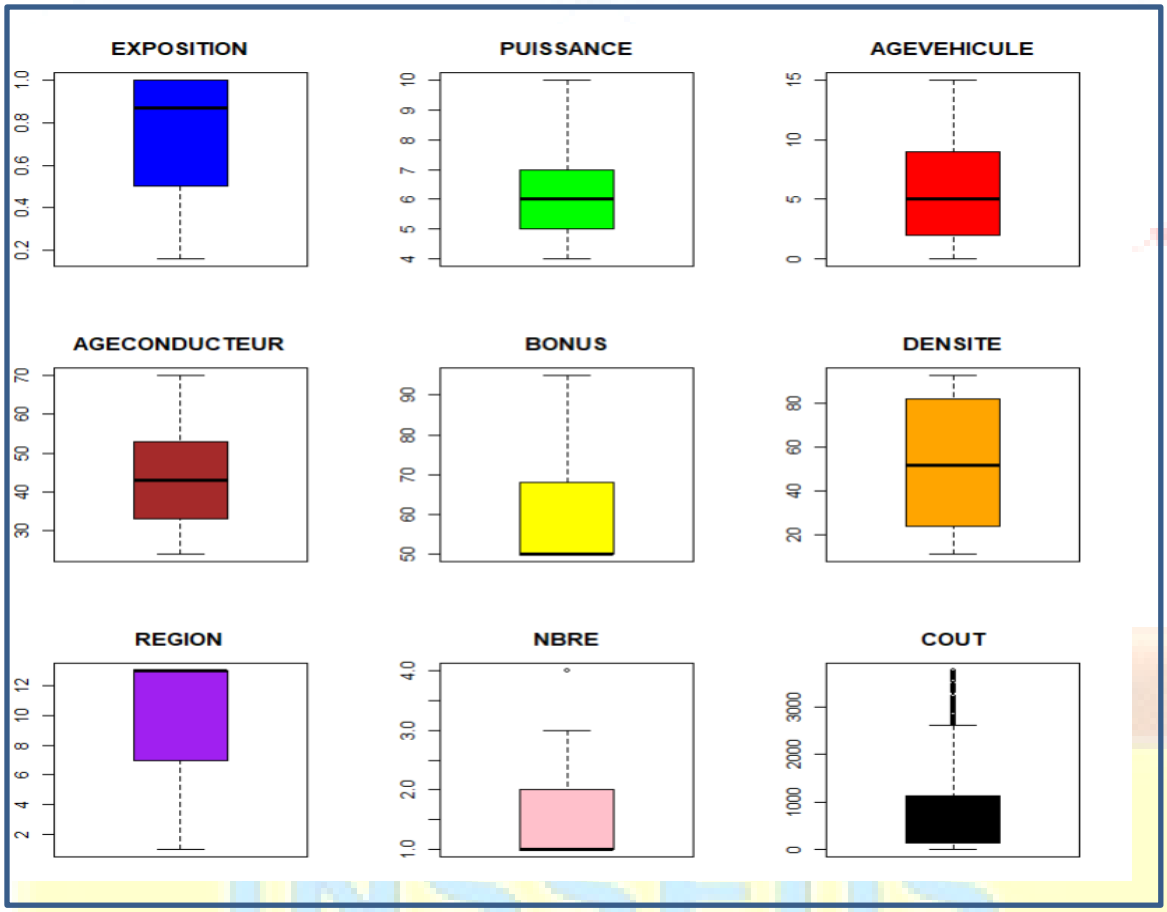


On voit bien que sur le graphique 2, les variables telles que Puissance, Age véhicule, Age conducteur, Bonus, Nombre et cout contiennent des valeurs aberrantes et extrêmes.

Pour traiter les valeurs aberrantes observées sur le graphique, nous allons utiliser la winsorisation. C'est une technique de traitement des valeurs aberrantes dans un jeu de données qui consiste à ramener ces valeurs dans la limite des bornes (inférieure et supérieure). Cette méthode doit son nom à son concepteur, Charles P. Winsor, un statisticien américain.

Après avoir traité les valeurs aberrantes, la visualisation nous donne ci-dessous (graphique 3).

Graphique 3: Visualisation des valeurs aberrantes et extrêmes après traitement



DEUXIEME PARTIE: ANALYSE UNIVARIEE ET BIVARIEE

I-ANALYSE UNI VARIEE

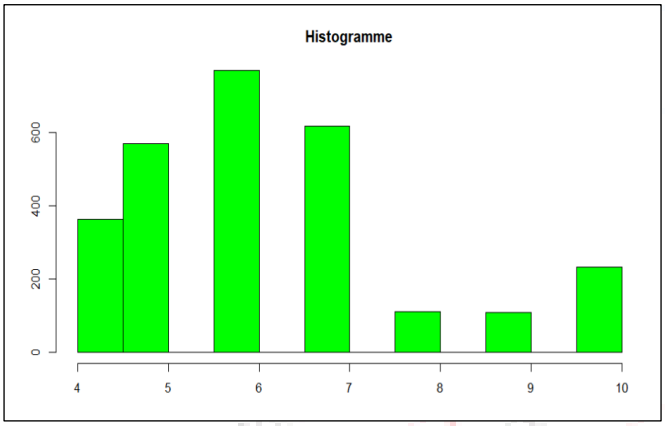
Dans cette partie du document, il s'agira pour nous de définir et interpréter les indicateurs de tendance centrale, les indicateurs de dispersion et les indicateurs de forme.

A-ANALYSES DES VARIABLES QUANTITATIVES

1-Analyse de la variable puissance

La variable "puissance" dans un jeu de données d'assurance automobile fait généralement référence à la puissance du moteur du véhicule, exprimée en chevaux-vapeur (CV) ou en kilowatts (kW).

Graphique 4: Histogramme de la variable puissance



Le diagramme ci-dessus présente visiblement une asymétrie potée vers la droite, donc la distribution ne suit pas une loi normale.

Tableau 5: Tableau de la variable puissance

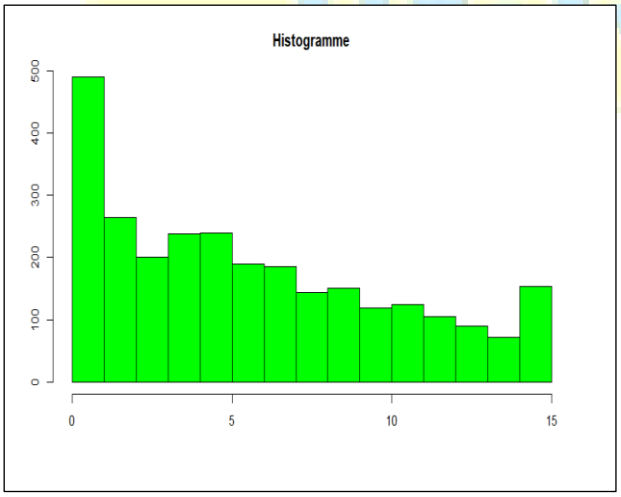
	Effectifs	Eff_Cum_crois	Eff_Cum_decrois	Frequence	Freq_Cum_crois	Freq_Cum_decrois
4	363	363	2765	0.1313	0.1313	1.0000
5	568	931	2534	0.2054	0.3367	0.9165
6	768	1699	2426	0.2778	0.6145	0.8774
7	617	2316	2316	0.2231	0.8376	0.8376
8	110	2426	1699	0.0398	0.8774	0.6145

Tableau 6: Résumé numérique de la variable puissance

INDICAEURS	VALEURS	INTERPRETATION
INDICAEURS DE TENDANCE CENTRALE ET DE POSITION		
MINIUM	4	La valeur minimale de la puissance est de 4
MAXIMUM	10	La valeur maximale de la puissance est de 10
MODE	6	La majorité de la société d'assurance a connu une puissance d'une valeur de 6
MOYENNE	6,286076	La puissance moyenne est de 6,286076
1er QUARTILE Q1	5	25% des clients de la société d'assurance ont connu une puissance maximum de 5
2ème QUARTILE Q2	6	La moitié des clients de la société d'assurance ont connu une puissance maximum de 6
3ème QUARTILE Q3	7	75% des clients de la société d'assurance ont connu une puissance maximum de 7
INDICATEURS DE DISPERSION		
VARIANCE	2,720229	On interprètera l'écart type qui est la racine carrée de la variance
ECART TYPE	1,649312	La puissance moyenne est de 6,286076 , Un écart type de 1,649312 indique que la puissance est dispersée en moyenne
COEFFICIENT DE VARIATION	26,23754	Le coefficient de variation est de 26,23754 , ce qui indique que la distribution de la puissance es hétérogène
INDICATEURS DE FORME		
SKEWNESS	0,7426321	Le skewness étant positif, ce qui indique que la distribution de la puissance est étalée à droite
KURTOSIS	3,051584	Le Kurtosis est élevé, ce qui indique que la distribution de la puissance est plus concentrée avec des queues plus épaisses (leptokurtique)

2-Analyse de la variable Age véhicule

Graphique 5: Histogramme de la variable Age véhicule



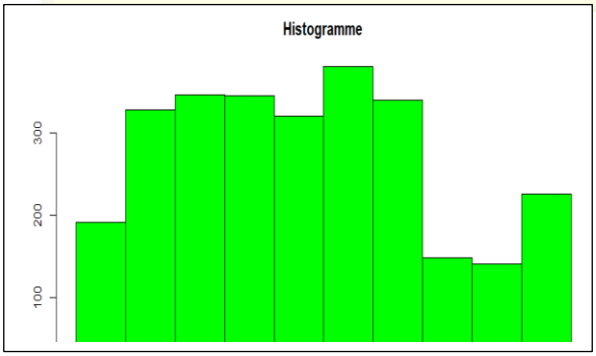
Le diagramme ci-dessus présente visiblement une asymétrie potée vers la droite, donc la distribution ne suit pas une loi normale.

Tableau 7: Résumé numérique de la variable Age véhicule

INDICAEURS	VALEURS	INTERPRETATION
INDICAEURS DE TENDANCE CENTRALE ET DE POSITION		
MINIUM	0	La valeur minimale de l'Age véhicule est de 0
MAXIMUM	15	La valeur maximale de l'Age véhicule est de 15
MODE	1	La majorité de la société a connu une Age véhicule d'une valeur de 1
MOYENNE	6,088969	L'Age véhicule moyenne est de 6,088969
1er QUARTILE Q1	2	25% des clients de la société d'assurance ont connu un Age véhicule maximum de 2
2ème QUARTILE Q2	5	La moitié des clients de la société d'assurance ont connu un Age véhicule maximum de 5
3ème QUARTILE Q3	9	75% des clients de la société d'assurance ont connu un Age véhicule maximum de 9
INDICATEURS DE DISPERSION		
VARIANCCE	19,42334	On interpretera l'écart type qui est la racine carrée de la variance
ECART TYPE	4,407192	L'Age véhicule moyenne est de 6,088969, Un ecart type de 4,407192 indique que la puissance est dipersée en moyenne
COEFFICIENT DE VARIATION	72,37993	Le coefficient de variation est de 26,23754 , ce qui indique que la distribution de l'Age véhicule est héterogène
INDICATEURS DE FORME		
SKEWNESS	0,4854656	Le skewness étant positif, ce qui indique que la distribution de l'Age véhicule est éalée à droite
KURTOSIS	2,149844	Le Kurtosis est faible, ce qui indique que la distribution de l'Age véhicule est plus aplaties avevc des queues plus minces (platikurtique)

3-Etude de la variable Age conducteur

Graphique 6: Histogramme de la variable Age conducteur



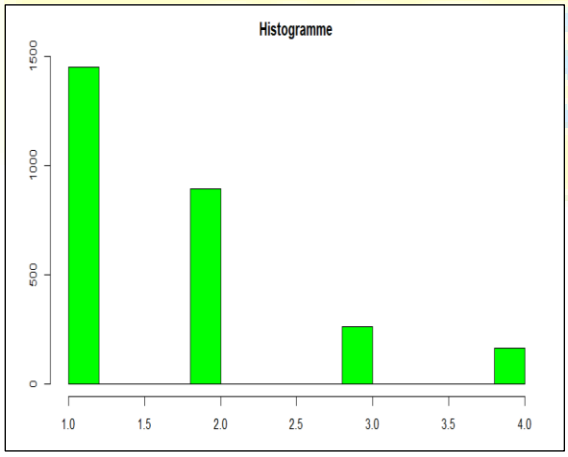
Le diagramme ci-dessus présente visiblement une asymétrie potée vers la droite, donc la distribution ne suit pas une loi normale.

Tableau 8: Tableau de la variable Age conducteur

INDICAEURS	VALEURS	INTERPRETATION
INDICAEURS DE TENDANCE CENTRALE ET DE POSITION		
MINIUM	24	La valeur minimale de l'Age conducteur est de 24
MAXIMUM	70	La valeur maximale de l'Age conduceur est de 70
MODE	24	La majorité de la société a connu un Age conducteur d'une valeur de 24
MOYENNE	43,86112	L'Age du conducteur moyen est de 43,86112
1er QUARTILE Q1	33	25% des clients de la société d'assurance ont connu un Age conducteur maximum de 33
2ème QUARTILE Q2	43	La moitié des clients de la société d'assurance ont connu un Age conducteur maximum de 43
3ème QUARTILE Q3	53	75% des clients de la société d'assurance ont connu un Age conducteur maximum de 53
INDICATEURS DE DISPERSION		
VARIANCE	170,8903	On interpretera l'écart type qui est la racine carrée de la variance
ECART TYPE	13,0752	L'Age du conducteur moyen est de 43,86112, Un ecart type de 13,0752 indique que l'Age du conducteur est dipersée en moyenne
COEFFICIENT DE VARIATION	29,8043	Le coefficient de variation est de 29,8043 , ce qui indique que la distribution de l'Age du conducteur est hétérogène
INDICATEURS DE FORME		
SKEWNESS	0,3293893	Le skewness étant positif, ce qui indique que la distribution de l'Age du conducteur est éalé à droite
KURTOSIS	2,193014	Le Kurtosis est faible, ce qui indique que la distribution de l'Age du conducteur est plus aplaties avevc des queues plus minces (platikurtique)

4-Etude de la variable nombre de sinistre

Graphique 7: Histogramme de la variable nombre de sinistre



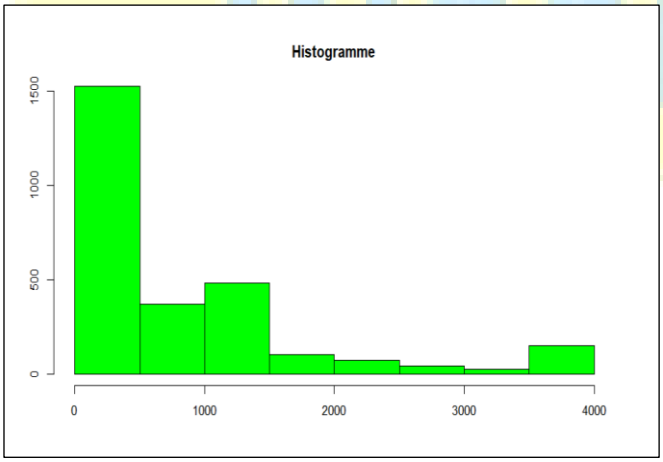
Le diagramme ci-dessus présente visiblement une asymétrie potée vers la droite, donc la distribution ne suit pas une loi normale.

Tableau 9: Tableau de la variable nombre de sinistre

INDICAEURS	VALEURS	INTERPRETATION
INDICAEURS DE TENDANCE CENTRALE ET DE POSITION		
MINIUM	1	La valeur minimale du nombre d'accident est de 1
MAXIMUM	4	La valeur maximale du nombre d'accident est de 4
MODE	1	La majorité de la société a connu un nombre d'accident d'une valeur de 1
MOYENNE	1,686076	Le nombred'accident moyen est de 1,686076
1er QUARTILE Q1	1	25% des clients de la société d'assurance ont connu un nombre d'accident de 1
2ème QUARTILE Q2	1	La moitié des clients de la société d'assurance ont connu un nombre d'accident maximum de 43
3ème QUARTILE Q3	2	75% des clients de la société d'assurance ont connu un nombre d'accident maximum de 2
INDICATEURS DE DISPERSION		
VARIANCE	0,7538039	On interpretera l'écart type qui est la racine carrée de la variance
ECART TYPE	0,8682188	Le nombre d'accident moyen est de 0,8682188 , Un ecart type de 0,8682188 indique que le nombre d'accident est dipersée en moyenne
COEFFICIENT DE VARIATION	51,49346	Le coefficient de variation est de 51,49346 , ce qui indique que la distribution du nombre d'accident est héterogène
INDICATEURS DE FORME		
SKEWNESS	1,186394	Le skewness étant positif, ce qui indique que la distribution du nombre d'accident est étalé à droite
KURTOSIS	3,645608	Le Kurtosis est élevé, ce qui indique que la distribution du nombre d'accident est plus concentré avec des queues plus épaise (leptokurtique)

5-Etude de la variable cout

Graphique 8: Histogramme de la variable cout



Le diagramme ci-dessus présente visiblement une asymétrie potée vers la droite, donc la distribution ne suit pas une loi normale

Tableau 10: Tableau de la variable cout

INDICAEURS	VALEURS	INTERPRETATION
INDICAEURS DE TENDANCE CENTRALE ET DE POSITION		
MINIUM	0	La valeur minimale du cout est de 0
MAXIMUM	3771,638	La valeur maximale du cout est de 3771,638
MODE	146,2795	La majorité de la société a connu un cout d'une valeur de 146,2795
MOYENNE	795,2499	Le cout moyen est de 795,2499
1er QUARTILE Q1	132,67	25% des clients de la société d'assurance ont connu un cout maximum de 132,67
2ème QUARTILE Q2	405,57	La moitié des clients de la société d'assurance ont connu un cout maximum de 405,57
3ème QUARTILE Q3	1128,12	75% des clients de la société d'assurance ont connu un cout maximum de 1128,12
INDICATEURS DE DISPERSION		
VARIANCE	930933	On interpretera l'écart type qui est la racine carrée de la variance
ECART TYPE	964,8487	Le cout moyen est de 795,2499 , Un ecart type de 964,8487 indique que le cout est dipersée en moyenne
COEFFICIENT DE VARIATION	121,3265	Le coefficient de variation est de 121,3265 , ce qui indique que la distribution du cout est héterogène
INDICATEURS DE FORME		
SKEWNESS	1,839278	Le skewness étant positif, ce qui indique que la distribution du nombre cout est étalé à droite
KURTOSIS	5,84179	Le Kurtosis est élevé, ce qui indique que la distribution du cout est plus concentré avec des queues plus épaise (leptokurtique)

B-ANALYSES DES VARIABLES QUALITATIVES

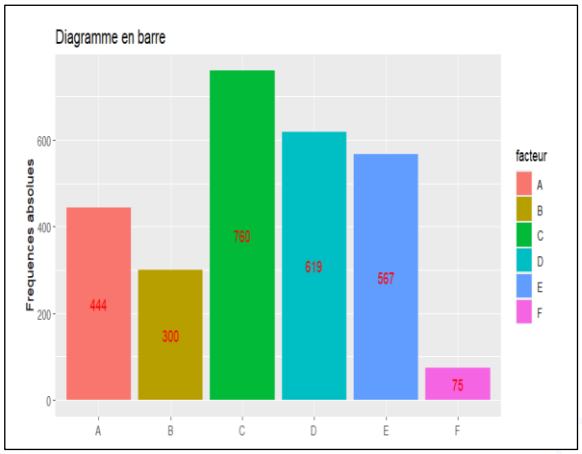
1-Analyse de la variable zone

Tableau 11: Tableau de la variable zone

	Effectif	Fréquence
A	444	0.16057866
B	300	0.10849910
C	760	0.27486438
D	619	0.22386980
E	567	0.20506329
F	75	0.02712477

Zone A : Cette zone est représentée par 444 clients, soit environ 16,06% de l'échantillon total. C'est l'une des zones les moins peuplées parmi celles répertoriées.

Graphique 9: Diagramme en barre de la variable zone



Cette répartition des clients selon les zones géographiques peut être utile pour la compagnie d'assurance afin de mieux comprendre la répartition de sa clientèle et de personnaliser ses offres en fonction des besoins et des caractéristiques spécifiques de chaque zone. Elle peut également être utilisée pour ajuster les stratégies de tarification et de distribution en fonction des différences régionales.

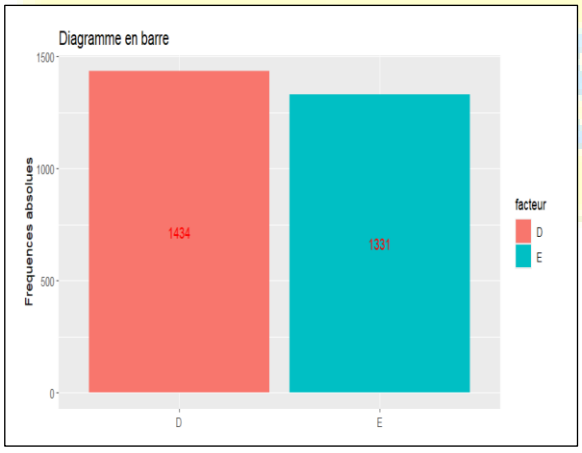
2-Analyse de la variable carburant

Tableau 12: Tableau de la variable carburant

Effectif Fréquence		
D	1434	0.5186257
E	1331	0.4813743

Type de carburant D : Ce type de carburant est le plus courant parmi les clients, représentant environ 51,86% de l'échantillon total. Cela indique que la majorité des clients utilisent ce type de carburant pour leurs véhicules assurés.

Graphique 10: Diagramme en barre de la variable carburant



Cette répartition peut être utile pour la compagnie d'assurance pour comprendre les tendances de préférence en matière de type de carburant parmi sa clientèle. Cela pourrait également être utilisé pour ajuster les offres et les politiques d'assurance en fonction des préférences des clients ou des tendances du marché.

3-Analyse de la variable garantie

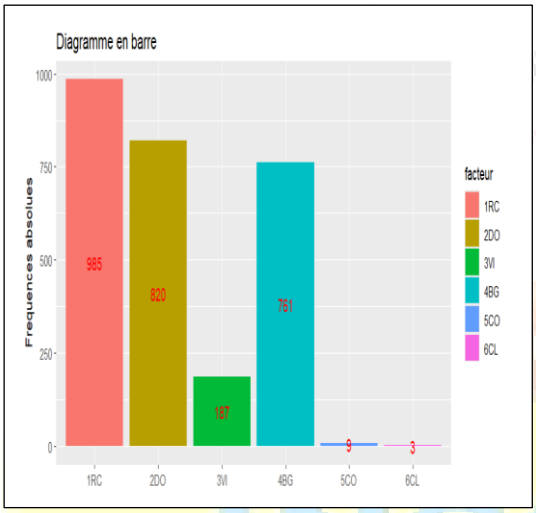
Tableau 13: Tableau de la variable garantie

	Effectif	Fréquence
1RC	985	0.356238698
2DO	820	0.296564195
3VI	187	0.067631103
4BG	761	0.275226040
5CO	9	0.003254973
6CL	3	0.001084991

1RC : 985 est la garantie la plus fréquente parmi les clients, représentant environ 35,62% de l'échantillon total. Cela indique que la majorité des clients ont souscrit à cette garantie spécifique.

2DO : La deuxième garantie la plus courante est 2DO, représentant environ 29,66% de l'échantillon. Bien qu'elle soit moins fréquente que 1RC, elle reste tout de même significative en termes de souscriptions.

Graphique 11: Diagramme en barre de la variable garantie



Les garanties 1RC, 2DO et 4BG sont les plus populaires parmi les clients de la société d'assurance, tandis que les garanties 3VI, 5CO et 6CL sont moins fréquentes. Cette analyse peut aider la compagnie d'assurance à mieux comprendre les préférences de ses clients et à adapter ses offres en conséquence.

II-ANALYSE BI VARIEE

Une analyse bi variée est une méthode d'analyse statistique qui examine la relation entre deux variables à la fois. Contrairement à une analyse uni variée qui se concentre sur une seule variable à la fois, l'analyse bi variée explore comment deux variables différentes sont liées ou interagissent entre elles.

A-ANALYSES DES VARIABLES QUANTITATIVES

1-Matrice de corrélation

Une matrice de corrélation est un outil puissant dans l'analyse exploratoire des données car elle permet de visualiser rapidement les relations entre les différentes variables d'un ensemble de données. Elle peut aider à identifier les variables qui sont fortement corrélées entre elles, ce qui peut être utile pour réduire la dimensionnalité des données ou pour sélectionner des variables pertinentes pour des analyses ultérieures.

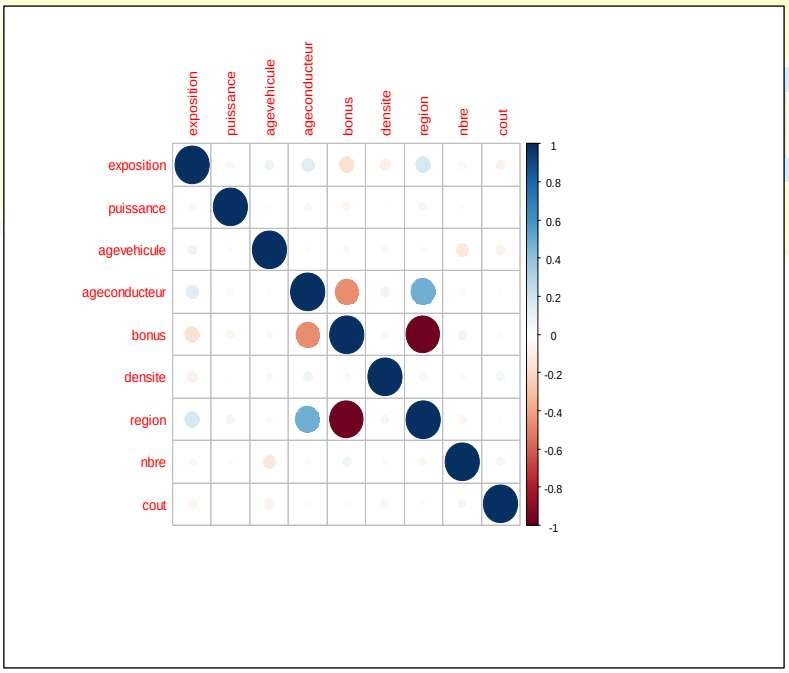
Tableau 14: tableau des variables

	exposition	puissance	agevehicule	ageconducteur	bonus	densite	region	nbre	cout
exposition	1.00000000	-0.044525926	0.07018302	0.13051587	-0.16874366	-0.085957259	0.17130527	0.04618602	-0.066122711
puissance	-0.04452593	1.000000000	0.01899479	0.03966624	-0.05542070	0.008154578	0.05420627	0.01448072	0.003240666
agevehicule	0.07018302	0.018994786	1.00000000	0.01895327	0.03243877	-0.032517377	-0.02628355	-0.12422005	-0.062130028
ageconducteur	0.13051587	0.039666241	0.01895327	1.00000000	-0.46656189	0.062016984	0.48444480	-0.02845150	0.017773495
bonus	-0.16874366	-0.055420697	0.03243877	-0.46656189	1.00000000	-0.047864514	-0.97274493	0.06180113	0.025457847
densite	-0.08595726	0.008154578	-0.03251738	0.06201698	-0.04786451	1.000000000	0.05325764	0.02348625	0.054757916
region	0.17130527	0.054206267	-0.02628355	0.48444480	-0.97274493	0.053257638	1.00000000	-0.05362756	-0.017788906
nbre	0.04618602	0.014480720	-0.12422005	-0.02845150	0.06180113	0.023486251	-0.05362756	1.00000000	0.058202800
cout	-0.06612271	0.003240666	-0.06213003	0.01777349	0.02545785	0.054757916	-0.01778891	0.05820280	1.000000000

Exposition et Puissance : Il y a une corrélation négative très faible entre l'exposition et la puissance du véhicule, ce qui suggère qu'il n'y a pas de relation linéaire claire entre ces deux variables.

Exposition et Âge du Véhicule : Il y a une corrélation positive très faible entre l'exposition et l'âge du véhicule, indiquant qu'il peut y avoir une légère tendance à ce que les véhicules plus anciens soient assurés pendant de plus longues périodes.

Graphique 12: Visualisation des variables



B-ANALYSES DES VARIABLES QUANTITATIVE ET QUALITATIVE

1-Analyse de la variable cout et zone

Tableau 15: Tableau de la variable cout et zone

Response: vecteur				
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
facteur	5	1771114	354223	0.3801	0.8627
Residuals	2759	2571327577	931978		

Dans notre cas, la valeur p associée au facteur est élevée (0.8627), ce qui suggère qu'il n'y a pas de différence significative entre les moyennes des groupes définis par le facteur. Cela signifie que le facteur n'a probablement pas d'effet significatif sur la variable de réponse.

2-Analyse de la variable cout et garantie

Tableau 16 : Tableau de la variable cout et garantie

Response: vecteur					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
facteur	5	37816534	7563307	8.2307	9.959e-08 ***
Residuals	2759	2535282157	918913		

Le facteur a un effet significatif sur la variable de réponse "vecteur" dans votre analyse. Les moyennes des groupes définis par le facteur sont différentes les unes des autres.

3-Analyse de la variable cout et marque

Tableau 17: Tableau de la variable cout et marque

Response: vecteur					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
facteur	9	14359800	1595533	1.7179	0.0796 .
Residuals	2755	2558738892	928762		

Dans ce cas, il y a une tendance à l'effet du facteur sur la variable de réponse "vecteur", mais cette tendance n'est pas statistiquement significative au seuil de 5%.

TROISIEME PARTIE : ANALYSE EN COMPOSANTE PRINCIPALE

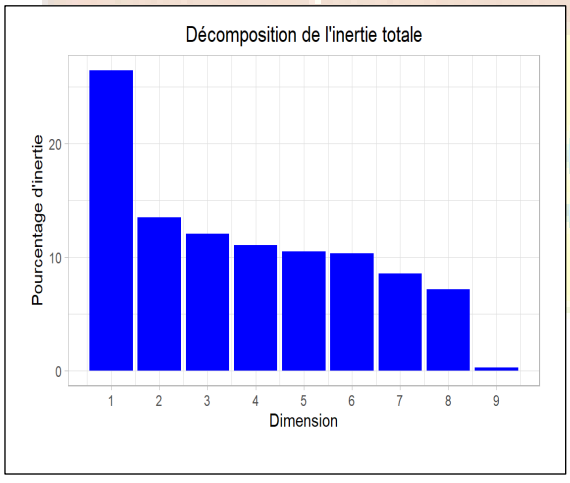
L'Analyse en Composantes Principales (ACP) est une technique statistique puissante utilisée pour réduire la dimensionnalité des données tout en préservant au mieux leur variance. Elle permet de transformer un ensemble de variables corrélées en un ensemble de variables non 14 corrélées, appelées composantes principales, qui capturent l'essentiel de l'information contenue dans les données d'origine.

L'ACP est souvent utilisé pour explorer la structure sous-jacente des données, identifier les tendances et les relations entre les variables, et faciliter la visualisation des données dans un espace de dimensions réduites. Elle est particulièrement utile dans les cas où les données sont complexes et comportent de nombreuses variables, ce qui rend difficile l'interprétation et l'analyse directe.

I-DISTRIBUTION DE L'INERTIE

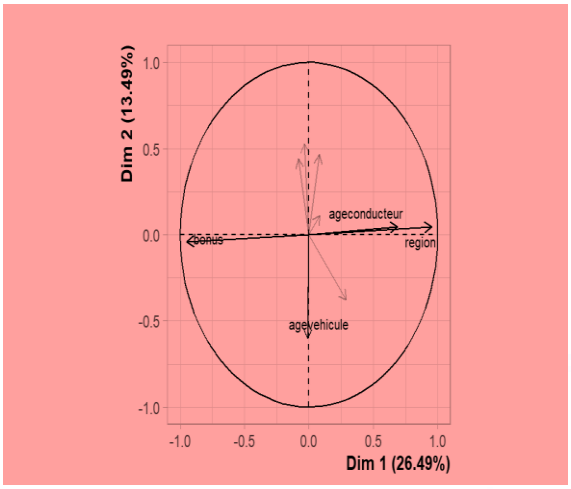
Une distribution de l'inertie, également connue sous le nom de "scree plot" en anglais, est un graphique qui montre la quantité d'inertie expliquée par chaque composante principale dans une analyse en composantes principales (ACP) ou une autre technique de réduction de dimensionnalité. L'inertie, dans ce contexte, fait référence à la quantité de variance expliquée par chaque composante principale. Plus une composante principale explique de variance, plus elle est importante pour décrire la structure des données.

Graphique 13: Visualisation de la distribution de l'inertie

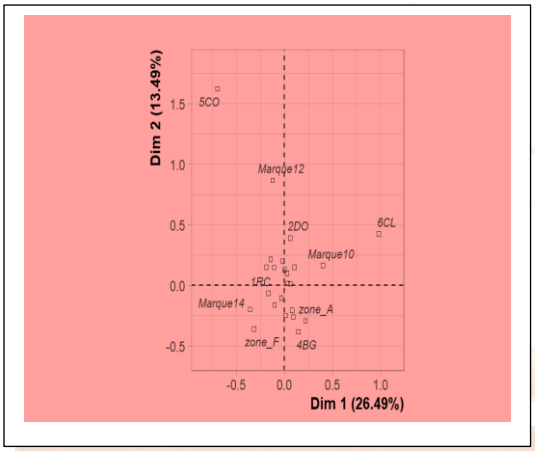


Les 2 premiers axes de l'analyse expriment 39.97% de l'inertie totale du jeu de données ; cela signifie que 39.97% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage relativement moyen, et le premier plan représente donc seulement une part de la variabilité contenue dans l'ensemble du jeu de données actif. Cette valeur est supérieure à la valeur référence de 24.22%, la variabilité expliquée par ce plan est donc significative.

Graphique 14: Graphe des individus



Graphique 15: Graphe des variables



Dimension 1 :

Le groupe 1 (coordonnée positives) se caractérise par de fortes valeurs pour les variables densité, coût, région, nombre, âge du conducteur et puissance, et par des faibles valeurs pour les variables âge du véhicule, bonus et exposition.

Le groupe 2 (coordonnée positives) présente des fortes valeurs pour les variables région, âge du conducteur, exposition et âge du véhicule, et des faibles valeurs pour les variables bonus, nombre, coût, densité et puissance.

Le groupe 3 (coordonnée négatives) montre des fortes valeurs pour les variables bonus et nombre, et des faibles valeurs pour les variables région, âge du conducteur, exposition et densité. Notamment, la variable région est fortement corrélée à cette dimension.

Dimension 2 :

Le groupe 1 (coordonnée positives) partage des caractéristiques similaires au groupe 1 de la dimension 1.

Le groupe 2 (coordonnée négatives) partage des caractéristiques similaires au groupe 2 de la dimension 1.

QUATRIEME PARTIE : MODELISATION DU COUT DES SINISTRES

La modélisation des coûts des sinistres en assurance consiste à prédire les coûts futurs que l'assureur devra payer pour les sinistres basés sur les caractéristiques des clients et des véhicules. Cette analyse est cruciale pour déterminer les primes d'assurance adéquates et pour la gestion des risques. En termes simples, elle vise à établir une relation entre les variables explicatives (par exemple, l'âge du véhicule, l'âge du conducteur, la puissance du véhicule, etc.) et le coût des sinistres (la variable cible).

I-CONSTRUCTION DU MODELE

(Intercept)	actuarNV1$exposition	actuarNV1$puissance	actuarNV1$agevehicule	actuarNV1$ageconducteur
141.763456	-165.852477	-1.567316	-8.637077	1.693870
actuarNV1$bonus	actuarNV1$densite	actuarNV1$nbre	actuarNV1$marqueMarque10	actuarNV1$marqueMarque12
6.860259	1.787615	37.593039	136.641317	51.369220
ctuarNV1$marqueMarque13	actuarNV1$marqueMarque14	actuarNV1$marqueMarque2	actuarNV1$marqueMarque3	actuarNV1$marqueMarque4
233.510151	-257.206011	67.834520	-24.340963	48.898745
actuarNV1$marqueMarque5	actuarNV1$marqueMarque6	actuarNV1$region	actuarNV1$garantie2DO	actuarNV1$garantie3VI
24.321794	30.896250	18.536336	58.614890	-169.414232
actuarNV1$garantie4BG	actuarNV1$garantie5CO	actuarNV1$garantie6CL	actuarNV1$zoneB	actuarNV1$zoneC
-122.721482	738.616893	-531.763712	33.836174	22.961174
actuarNV1$zoneD	actuarNV1$zoneE	actuarNV1$zoneF	actuarNV1$carburantE	
19.807613	-4.649643	147.435961	-32.350581	

1-Modèle et sélection automatiques des explicatives

En économétrie, la sélection automatique des variables explicatives consiste à utiliser des méthodes systématiques pour choisir les variables les plus pertinentes à inclure dans un modèle de régression. L'objectif est de construire un modèle qui explique au mieux la variation de la variable dépendante sans inclure de variables redondantes ou non significatives, ce qui permet d'obtenir des estimations plus précises et interprétables.

Tableau 18: Tableau du model

```
Call:
lm(formula = cout ~ exposition + agevehicule + densite + nbre +
    garantie, data = actuarNV1)

Residuals:
    Min      1Q  Median      3Q     Max
-1806.0  -616.0  -340.5   297.5  3256.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  867.8469    80.8975  10.728  < 2e-16 ***
exposition  -168.7217    63.5137  -2.656  0.00794 **
agevehicule   -9.2267     4.3128  -2.139  0.03249 *
densite        1.5958     0.6582   2.424  0.01540 *
nbre          37.1298    21.8144   1.702  0.08885 .
garantie2DO   56.7313    46.8389   1.211  0.22592
garantie3VI -194.5461    76.6475  -2.538  0.01120 *
garantie4BG -137.9993    47.0912  -2.930  0.00341 **
garantie5CO  756.3943   322.2407   2.347  0.01898 *
garantie6CL -525.1173   552.9096  -0.950  0.34233
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 955.3 on 2755 degrees of freedom
Multiple R-squared:  0.0228,    Adjusted R-squared:  0.01961
F-statistic: 7.142 on 9 and 2755 DF,  p-value: 2.67e-10
```

2-Anova du model

L'analyse de la variance (ANOVA) pour un modèle de régression est une technique statistique utilisée pour décomposer la variabilité totale des données en composantes associées aux différentes sources de variation dans le modèle. En d'autres termes, ANOVA permet de déterminer l'importance relative des variables explicatives dans le modèle de régression.

Tableau 19: Tableau de l'ANOVA du model

```
Analysis of Variance Table

Response: cout
              Df     Sum Sq  Mean Sq F value    Pr(>F)
exposition     1   11250135 11250135 12.3265 0.0004537 ***
agevehicule    1    8546248  8546248  9.3639 0.0022343 **
densite        1    5863790  5863790  6.4248 0.0113086 *
nbre           1    7295863  7295863  7.9939 0.0047274 **
garantie       5   25712585  5142517  5.6345 3.567e-05 ***
Residuals   2755 2514430070   912679
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L'ANOVA montre que les variables expositions, âge du véhicule, densité, nombre et garantie ont toutes des effets significatifs sur le coût des sinistres. Les p-values associées à ces variables sont toutes inférieures au seuil de significativité conventionnel (0.05), indiquant que les effets de ces variables ne sont probablement pas dus au hasard.

3-Regression sur Les variables retenus

Tableau 20: Tableau des variables retenues

(Intercept)	actuarNV1$exposition	actuarNV1$agevehicule	actuarNV1$densite	actuarNV1$nbre
867.846936	-168.721696	-9.226694	1.595766	37.129816
actuarNV1$garantie2DO	**actuarNV1$garantie3VI**	**actuarNV1$garantie4BG**	**actuarNV1$garantie5CO**	**actuarNV1$garantie6CL**
56.731285	-194.546146	-137.999290	756.394338	-525.117284

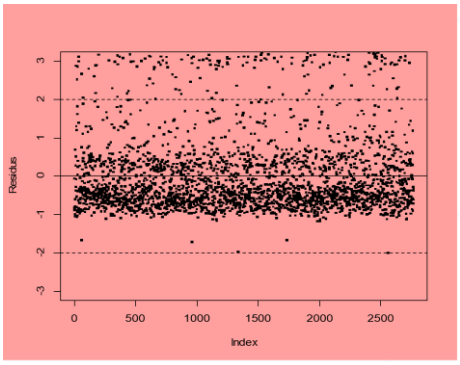
Tableau 21: Résultat final du model

```
Call:
lm(formula = actuarNV1$cout ~ actuarNV1$exposition + actuarNV1$agevehicule +
    actuarNV1$densite + actuarNV1$nbre + actuarNV1$garantie)

Residuals:
    Min      1Q  Median      3Q     Max
-1806.0  -616.0  -340.5   297.5  3256.6

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              867.8469     80.8975  10.728  < 2e-16 ***
actuarNV1$exposition    -168.7217     63.5137  -2.656  0.00794 **
actuarNV1$agevehicule     -9.2267      4.3128  -2.139  0.03249 *
actuarNV1$densite          1.5958      0.6582   2.424  0.01540 *
actuarNV1$nbre            37.1298     21.8144   1.702  0.08885 .
actuarNV1$garantie2DO     56.7313     46.8389   1.211  0.22592
actuarNV1$garantie3VI   -194.5461     76.6475  -2.538  0.01120 *
actuarNV1$garantie4BG   -137.9993     47.0912  -2.930  0.00341 **
actuarNV1$garantie5CO    756.3943    322.2407   2.347  0.01898 *
actuarNV1$garantie6CL   -525.1173    552.9096  -0.950  0.34233
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 955.3 on 2755 degrees of freedom
Multiple R-squared:  0.0228,    Adjusted R-squared:  0.01961
F-statistic: 7.142 on 9 and 2755 DF,  p-value: 2.67e-10
```

Le modèle montre que certaines variables explicatives, telles que exposition, âge du véhicule, densité, et certaines catégories de garantie, ont un impact significatif sur le coût des sinistres. Cependant, la faible valeur du R^2 suggère que le modèle explique une petite partie de la variabilité totale du coût des sinistres, ce qui indique que d'autres variables importantes peuvent être absentes du modèle.

a-Analyse du résidu

Graphique 16: Visualisation du résidu



En théorie 95% des résidus studentisés se trouvent dans l'intervalle [-2;2]. Ici on a visuellement beaucoup de résidus qui se trouvent dans cet intervalle. Ce qui est acceptable. La moyenne des résidus est de 92,94756.

4-Test de la validité du modèle

a-Test de linéarité du modèle

Le test de **Rainbow**, également connu sous le nom de test de linéarité de la régression, est un test utilisé pour évaluer la linéarité des relations dans un modèle de régression linéaire. Contrairement à certains tests de linéarité qui se concentrent sur des aspects spécifiques de la linéarité, le test de **Rainbow** évalue la linéarité globale du modèle de régression dans son ensemble.

H0 : le modèle est linéaire

H1 : le modèle n ' est pas linéaire

Tableau 22: Tableau du modèle

```
        Rainbow test

data:  reg.fin
Rain = 1.0077, df1 = 1383, df2 = 1372, p-value = 0.4431
```

La p-value > 0.05, Nous ne pouvons pas rejeter l'hypothèse nulle de linéarité, Le modèle est donc linéaire.

b-Test d'homoscédasticité

Le test d'homoscédasticité, également appelé test d'homogénéité des variances, est une procédure statistique utilisée pour évaluer si la variance des résidus dans un modèle de régression est constante à travers toutes les valeurs des variables indépendantes.

TEST DE BREUSCH-PAGAN

H0 : il y homoscédasticité

H1 : il y a hétéroscédasticité

Tableau 23: Tableau du modèle

```
        studentized Breusch-Pagan test

data:  model_MCG
BP = 0.89242, df = 1, p-value = 0.3448
```

La p-value > 0.05, Nous ne pouvons pas rejeter l'hypothèse nulle donc il y a homoscédasticité.

CINQUIEME PARTIE : MODELISATION DE LA FREQUENCE DES SINISTRES

La modélisation de la fréquence des sinistres est un processus en actuariat et en assurance qui consiste à estimer le nombre d'incidents ou de réclamations (sinistres) qu'une police d'assurance va subir sur une période donnée. Cette estimation permet aux compagnies d'assurance de prévoir les coûts futurs et de fixer des primes appropriées.

I-CONTRUCTION DU MODELE

1-Anova du model

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: nbre

Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         2764     3826.3
exposition    1    6.302      2763     3820.0  0.012060 *
zone          5    5.748      2758     3814.3  0.331496
puissance     1    0.103      2757     3814.2  0.748525
agevehicule   1   40.257      2756     3773.9 2.227e-10 ***
ageconducteur 1    0.456      2755     3773.4  0.499268
bonus         1    4.731      2754     3768.7  0.029618 *
marque        9   12.646      2745     3756.1  0.179272
carburant     1    2.807      2744     3753.3  0.093826 .
densite       1    1.232      2743     3752.0  0.267069
region       14   32.931      2729     3719.1  0.002948 **
garantie      5  250.695      2724     3468.4 < 2.2e-16 ***
cout          1    1.173      2723     3467.2  0.278743
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Les variables qui ont un effet significatif sur la fréquence des sinistres dans ce modèle sont l'exposition, l'âge du véhicule, le bonus, le type de carburant, la région et la garantie.

2-Anova du model final

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: nbre

Terms added sequentially (first to last)

            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         2764     3826.3
exposition   1    6.302      2763     3820.0  0.012060 *
agevehicule  1   40.212      2762     3779.8 2.279e-10 ***
bonus        1    3.040      2761     3776.8  0.081256 .
carburant    1    2.773      2760     3774.0  0.095844 .
region      14   34.417      2746     3739.6  0.001792 **
garantie     5  247.301      2741     3492.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Les variables suivantes ont un effet significatif sur la fréquence des sinistres dans ce modèle : Exposition (effet significatif, $p < 0.05$) ,Âge du véhicule (effet très significatif, $p < 0.001$) ,Bonus (effet potentiellement significatif, $p < 0.1$) , Carburant (effet potentiellement significatif, $p < 0.1$) , Région (effet significatif, $p < 0.01$) ,Garantie(effet très significatif, $p < 0.001$) Ces résultats suggèrent que ces facteurs devraient être pris en compte lors de l'évaluation des risques et de la tarification des assurances, car ils influencent significativement la fréquence des sinistres.

3-Resultat final du model

```
Call:
glm(formula = nbre ~ exposition + agevehicule + garantie + bonus +
    region, family = binomial, data = actuarNV1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.653874   1.469527  -5.208 1.90e-07 ***
exposition    0.652950   0.143991   4.535 5.77e-06 ***
agevehicule  -0.040879   0.009668  -4.228 2.36e-05 ***
garantie2DO   0.454359   0.101667   4.469 7.86e-06 ***
garantie3VI   0.275787   0.165302   1.668  0.09524 .
garantie4BG  -1.123312   0.108500 -10.353  < 2e-16 ***
garantie5CO  14.547621 290.848516   0.050  0.96011
garantie6CL  -0.472724   1.235173  -0.383  0.70193
bonus         0.076349   0.015361   4.970 6.69e-07 ***
region2       1.266514   0.303273   4.176 2.96e-05 ***
region3       1.260911   0.320996   3.928 8.56e-05 ***
region4       1.426734   0.348618   4.093 4.27e-05 ***
region5       1.581975   0.387924   4.078 4.54e-05 ***
region6       2.208572   0.449364   4.915 8.88e-07 ***
region7       2.443164   0.496516   4.921 8.63e-07 ***
region8       1.742057   0.586081   2.972  0.00295 **
region9       2.881879   0.605679   4.758 1.95e-06 ***
region10      3.118881   0.649783   4.800 1.59e-06 ***
region11      3.570669   0.700436   5.098 3.44e-07 ***
region12      2.974579   0.752927   3.951 7.79e-05 ***
region13      3.520016   0.686937   5.124 2.99e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3826.3  on 2764  degrees of freedom
Residual deviance: 3484.6  on 2744  degrees of freedom
AIC: 3526.6

Number of Fisher Scoring iterations: 13
```

4-Les odd ratios et effets marginaux

L'odd ratio est une mesure statistique utilisée principalement dans les études de cas-témoins pour évaluer l'association entre une exposition (ou un facteur de risque) et un événement (ou une maladie). Quant aux effets marginaux elles représentent l'effet d'une petite variation dans une variable indépendante (par exemple, une augmentation d'une unité) sur la probabilité d'un événement ou sur la valeur espérée de la variable dépendante dans un modèle économétrique.

a-Affichage des paramètres estimés

```
(Intercept)  exposition agevehicule garantie2DO garantie3VI garantie4BG garantie5CO
-7.65387384  0.65295050 -0.04087852  0.45435891  0.27578661 -1.12331194 14.54762120
garantie6CL       bonus     region2     region3     region4     region5     region6
-0.47272434  0.07634875  1.26651379  1.26091072  1.42673405  1.58197467  2.20857235
    region7     region8     region9    region10    region11    region12    region13
 2.44316429  1.74205742  2.88187872  3.11888145  3.57066879  2.97457894  3.52001640
```

Intercept (-7,65) : il s'agit de la cote logarithmique de référence du résultat lorsque tous les prédicteurs sont à zéro. Étant donné qu'elle est négative et relativement importante, elle suggère que le résultat est moins probable lorsque tous les prédicteurs sont à leur valeur de référence.

b- Les intervalles de confiance

```
                  2.5 %       97.5 %
(Intercept) -10.57405703 -4.80664161
exposition    0.37157151  0.93616285
agevehicule  -0.05988121 -0.02197151
garantie2DO   0.25543591  0.65406154
garantie3VI  -0.04695202  0.60179624
garantie4BG  -1.33748825 -0.91203265
garantie5CO  -8.88989779          NA
garantie6CL  -3.55385367  1.89174680
bonus         0.04659011  0.10688044
region2       0.67704055  1.86821979
region3       0.63324600  1.89293740
region4       0.74578829  2.11338883
region5       0.82537246  2.34715585
region6       1.33369182  3.09659570
region7       1.47744776  3.42541121
region8       0.59381065  2.89451960
region9       1.70426046  4.08070903
region10      1.85599445  4.40554959
region11      2.20929504  4.95768096
region12      1.50844681  4.46319522
region13      2.18930987  4.88544714
```

c- Les odd ratio

```
  (Intercept)    exposition   agevehicule   garantie2DO   garantie3VI   garantie4BG   garantie5CO
4.742036e-04  1.921201e+00  9.599457e-01  1.575163e+00  1.317567e+00  3.252010e-01  2.079465e+06
  garantie6CL         bonus       region2       region3       region4       region5       region6
6.233019e-01  1.079339e+00  3.548460e+00  3.528634e+00  4.165074e+00  4.864552e+00  9.102712e+00
      region7       region8       region9      region10      region11      region12      region13
1.150940e+01  5.709077e+00  1.784777e+01  2.262106e+01  3.554035e+01  1.958138e+01  3.378498e+01
```

Par exemple :

Region13 un rapport de cotes de 33,78 indique que l'événement est beaucoup plus probable dans cette région que dans la région de référence.

Garantie4BG un rapport de cotes de 0,33 indique que ce niveau réduit considérablement la probabilité de l'événement.

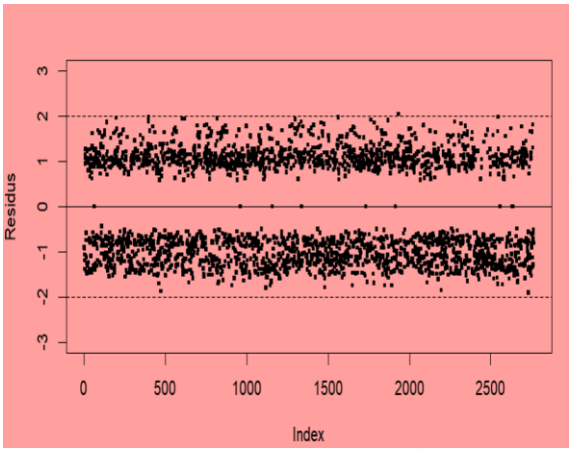
d-Les effets marginaux

Les effets marginaux montrent le changement de probabilité lorsque le prédicteur ou la variable indépendante augmente d'une unité.

```
Call:
logitmfx(formula = nbre ~ agevehicule + garantie + exposition +
    bonus + region, data = actuarNV1)

Marginal Effects:
                dF/dx  Std. Err.       z     P>|z|
agevehicule -0.0102023  0.0024451 -4.1725 3.013e-05 ***
garantie2DO  0.1131008  0.0250845  4.5088 6.520e-06 ***
garantie3VI  0.0688142  0.0410426  1.6767 0.0936102 .
garantie4BG -0.2666928  0.0453446 -5.8815 4.066e-09 ***
garantie5CO  0.5323683  0.0101543 52.4280 < 2.2e-16 ***
garantie6CL -0.1147462  0.2867426 -0.4002 0.6890303
exposition   0.1629615  0.0364905  4.4659 7.975e-06 ***
bonus        0.0190549  0.0039058  4.8786 1.069e-06 ***
region2      0.2885856  0.0842603  3.4249 0.0006150 ***
region3      0.2877911  0.0860568  3.3442 0.0008252 ***
region4      0.3181343  0.0964679  3.2978 0.0009744 ***
region5      0.3444086  0.1077787  3.1955 0.0013958 **
region6      0.4246252  0.1479491  2.8701 0.0041037 **
region7      0.4491367  0.1612831  2.7848 0.0053565 **
region8      0.3644920  0.1337710  2.7247 0.0064351 **
region9      0.4787077  0.1843588  2.5966 0.0094149 **
region10     0.4913054  0.1937205  2.5362 0.0112077 *
region11     0.5084665  0.2074632  2.4509 0.0142509 *
region12     0.4799594  0.1906781  2.5171 0.0118319 *
region13     0.6875805  0.1122126  6.1275 8.928e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

dF/dx is for discrete change for the following variables:

 [1] "garantie2DO" "garantie3VI" "garantie4BG" "garantie5CO" "garantie6CL" "region2"
 [7] "region3"     "region4"     "region5"     "region6"     "region7"     "region8"
[13] "region9"     "region10"    "region11"    "region12"    "region13"
```

e- Analyse des résidus

Graphique 17 : Visualisation des résidus



En théorie 95% des résidus studentisés se trouvent dans l'intervalle [-2;2]. Ici on a visuellement beaucoup de résidus qui se trouvent dans cet intervalle. Ce qui est acceptable.

5-Taux de mauvais classement et Matrice de Confusion

Tableau 24: Calcul des probabilités

	exposition	zone	puissance	agevehicule	ageconducteur	bonus	marque	carburant	densite	region	nbre	garantie	cout	PROBABILITE_PREDITE	MODALITE_PREDITE
1	0.74	A	5	4	31	64	3	D	21	8	0	1RC	0.00	0.3305215	pas accident
2	0.18	B	7	8	24	95	2	E	26	1	0	1RC	0.00	0.3520044	pas accident
3	0.48	C	9	0	32	61	12	E	41	13	0	4BG	687.82	0.4288659	pas accident
4	0.27	F	7	5	39	95	12	E	11	1	0	2DO	96.64	0.5063796	accident
5	0.51	E	4	0	49	50	12	E	31	13	0	2DO	70.88	0.6156070	accident
6	0.64	D	10	0	58	50	12	D	72	13	1	1RC	0.00	0.5253453	accident

Tableau 25: Transformation des probabilités en modalité prédites

	exposition	zone	puissance	agevehicule	ageconducteur	bonus	marque	carburant	densite	region	nbre	garantie	cout	PROBABILITE_PREDITE	MODALITE_PREDITE
1	0.74	A	5	4	31	64	3	D	21	8	0	1RC	0.00	0.3305215	pas accident
2	0.18	B	7	8	24	95	2	E	26	1	0	1RC	0.00	0.3520044	pas accident
3	0.48	C	9	0	32	61	12	E	41	13	0	4BG	687.82	0.4288659	pas accident
4	0.27	F	7	5	39	95	12	E	11	1	0	2DO	96.64	0.5063796	accident
5	0.51	E	4	0	49	50	12	E	31	13	0	2DO	70.88	0.6156070	accident
6	0.64	D	10	0	58	50	12	D	72	13	1	1RC	0.00	0.5253453	accident

Tableau 26: le taux de mauvais classement à partir de la matrice de confusion

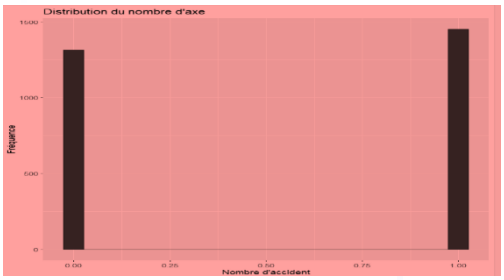
	accident	pas accident
accident	537	870
pas accident	914	444

35.5% signifie que seulement environ 35.5% des prédictions faites par le modèle sont correctes.

SIXIEME PARTIE : MODELISATION DE LA PROBABILITE DE LA SURVENANCE DES SINISTRE

La **modélisation de la probabilité de la survenance des sinistres** est une approche statistique utilisée pour estimer la probabilité qu'un sinistre (un événement dommageable ou une perte) se produise dans un certain contexte.

Graphique 18 : Visualisation de la variable cible



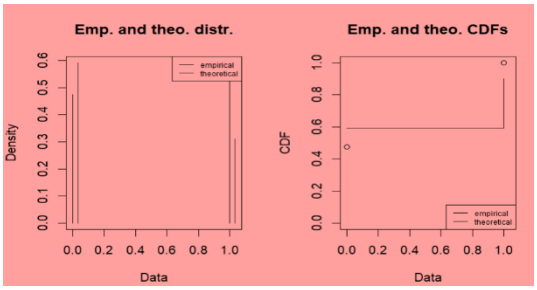
I-TEST D'ADEQUATION A LA LOI DE POISSON

1-Test d'adéquation

```
Chi-squared statistic:  742.6648
Degree of freedom of the Chi-squared distribution:  1
Chi-squared p-value:  1.57887e-163
Chi-squared table:
     obscounts theocounts
<= 0 1314.0000  1636.0203
<= 1 1451.0000   858.5409
> 1     0.0000   270.4388

Goodness-of-fit criteria
                                1-mle-pois
Akaike's Information Criterion   4775.174
Bayesian Information Criterion   4781.099
```

La p-value associée au test du Chi-carré est extrêmement petite (1.57887e-163), ce qui signifie qu'il est très peu probable que la différence entre les données observées et celles prévues par le modèle soit due au hasard négative **e**t les critères d'ajustement (AIC, BIC) sont élevés. Cela suggère qu'il y a probablement une surdispersion dans les données (la variance dépasse la moyenne), et on peut envisager un modèle de régression binomiale négative.

Graphique 19 : Visualisation du graphique



II-CONSTRUCTION DU MODELE COMPLET

1-Anova du modèle

```
Analysis of Deviance Table

Model: quasipoisson, link: log

Response: nbre

Terms added sequentially (first to last)


                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                 2764     1871.2
exposition           1    3.088      2763     1868.1  0.010877 *
zone                 5    2.741      2758     1865.3  0.330511
puissance            1    0.093      2757     1865.2  0.657873
agevehicule          1   19.569      2756     1845.7 1.447e-10 ***
ageconducteur        1    0.262      2755     1845.4  0.458142
bonus                1    3.461      2754     1842.0  0.007012 **
marque              10    7.163      2744     1834.8  0.130424
carburant            1    1.285      2743     1833.5  0.100374
densite              1    0.583      2742     1832.9  0.268315
region              12   24.762      2730     1808.2 6.184e-07 ***
garantie             5  107.974      2725     1700.2 < 2.2e-16 ***
cout                 1    1.115      2724     1699.1  0.125882
PROBABILITE_PREDITE  1    4.475      2723     1694.6  0.002172 **
MODALITE_PREDITE     1    0.009      2722     1694.6  0.890402
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Les variables significatives dans ce modèle sont : exposition, âge du véhicule, bonus, région, garantie, et probabilité prédite. Ces variables contribuent de manière importante à expliquer la fréquence des sinistres.
Les autres variables comme zone, puissance, âge du conducteur, marque, carburant, densité, et modalité prédite n'ont pas un effet significatif sur la fréquence des sinistres dans ce modèle.

2-Construction du modèle avec les variables significatives

```
Analysis of Deviance Table

Model: quasipoisson, link: log

Response: nbre

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          2764     1871.2
exposition    1    3.088      2763     1868.1  0.010615 *
agevehicule   1   19.501      2762     1848.6 1.353e-10 ***
ageconducteur 1    0.280      2761     1848.3  0.441610
marque       10    7.050      2751     1841.2  0.135532
bonus         1    3.896      2750     1837.4  0.004106 **
carburant     1    1.165      2749     1836.2  0.116484
garantie      5  110.438      2744     1725.8 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

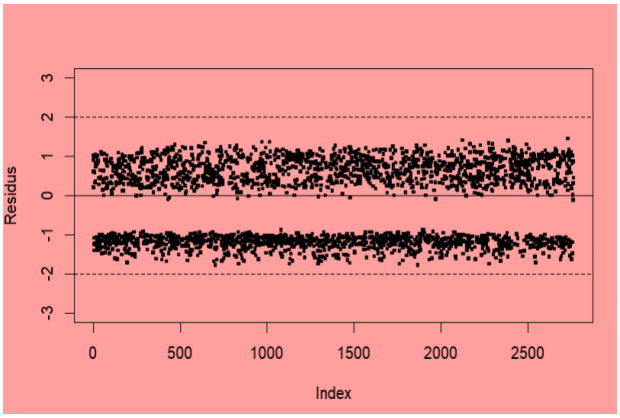
Les variables exposition, âge du véhicule, bonus, et garantie sont significatives et contribuent de manière importante à l'explication de la fréquence des sinistres.
En revanche, âge du conducteur, marque, et carburant n'ont pas un impact significatif dans ce modèle.

Tableau 27: Tableau des coefficients

(Intercept)	exposition	agevehicule	ageconducteur	marque2	marque3	marque4
6.516185e-01	7.589028e-01	1.015439e+00	9.991695e-01	9.664532e-01	8.367759e-01	8.554872e-01
marque5	marque6	marque10	marque11	marque12	marque13	marque14
9.406799e-01	8.708047e-01	9.164070e-01	7.708537e-01	1.048666e+00	1.125387e+00	8.405122e-01
bonus	carburantE	garantie2DO	garantie3VI	garantie4BG	garantie5CO	garantie6CL
9.977670e-01	1.101814e+00	7.570524e-01	8.786575e-01	1.515894e+00	4.223741e-07	1.188582e+00

Le modèle semble suggérer que les garanties et certains types de carburant ont un impact plus important sur la probabilité d'accidents, tandis que les marques de véhicules et l'âge du conducteur jouent un rôle moins prononcé.

Graphique 20 : Visualisation des résidus



Graphique 21 : Autre visualisation

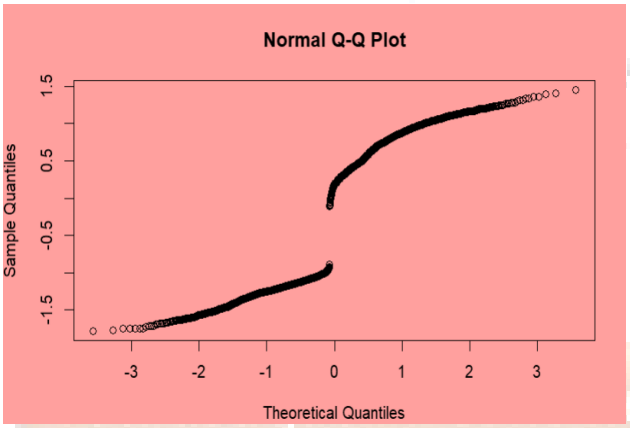
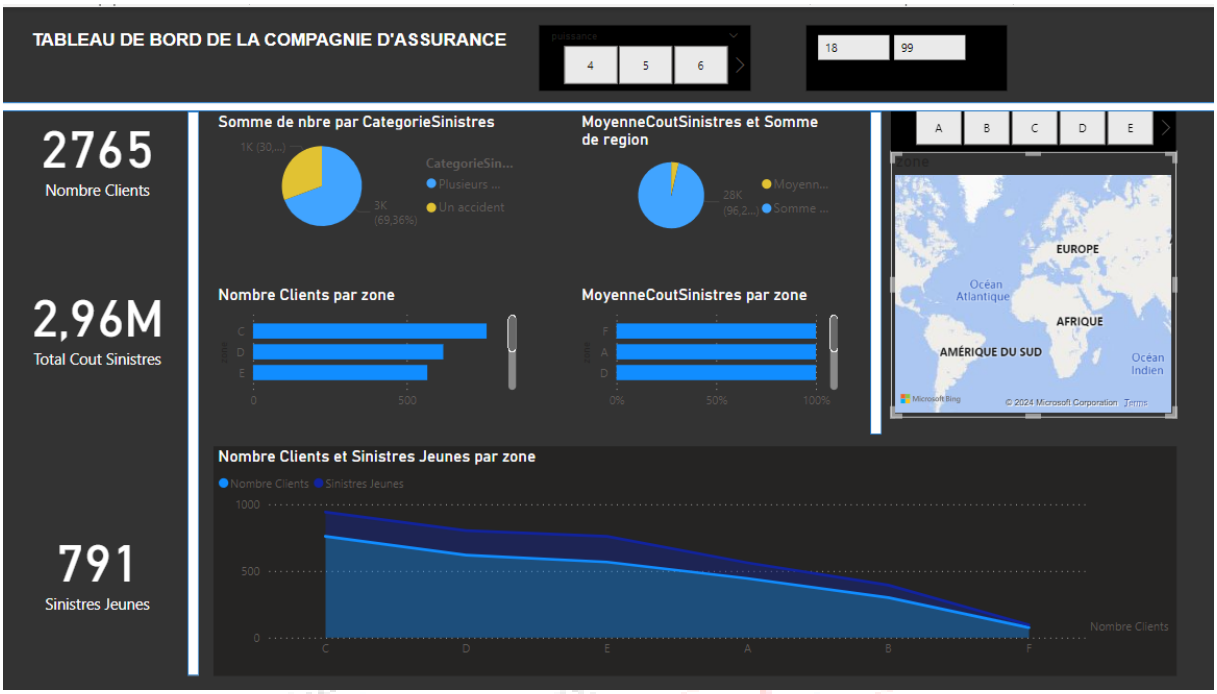


Tableau 28: Shaprio test

```
        Shapiro-Wilk normality test

data:  res.m
W = 0.84313, p-value < 2.2e-16
```

Avec une statistique W faible et une p-value extrêmement petite, le test de Shapiro-Wilk suggère que les résidus ne suivent pas une distribution normale.

III- TABLEAU DE BORD DE LA COMPAGNIE D'ASSURANCE



Perspectives

Ce projet montre comment l'exploitation des données à travers des modèles statistiques et des outils de visualisation peut transformer la gestion des risques dans l'assurance automobile. Les résultats peuvent guider la stratégie commerciale de l'entreprise, tant au niveau de la conception de produits que des initiatives de prévention des sinistres. Cela conforte l'idée que l'innovation dans l'utilisation des données est cruciale pour rester compétitif dans un marché de plus en plus complexe et concurrentiel.

CONCLUSION

L'analyse des données d'assurance auto a permis de mettre en lumière certains comportements et caractéristiques des clients en termes de sinistralité (probabilité d'accidents). Grâce à la modélisation en R et l'utilisation de Power BI pour visualiser les résultats clés, plusieurs observations ressortent :

Profil des clients à risque : Certains groupes de clients présentent un risque plus élevé d'accidents en fonction de facteurs tels que l'âge, l'expérience de conduite, l'historique des sinistres et d'autres variables démographiques ou comportementales.

Caractéristiques influençant la probabilité d'accidents : Les analyses ont mis en évidence les variables clés qui influencent fortement la probabilité d'accidents, permettant ainsi à la société d'assurance de mieux comprendre les profils des assurés les plus exposés au risque.

Optimisation de la tarification : En exploitant les résultats, l'entreprise peut ajuster sa politique de tarification de manière plus précise pour refléter les risques encourus par différents segments de clients, tout en maximisant la rentabilité et en minimisant les pertes liées aux sinistres.

Stratégie de prévention : La segmentation des clients en fonction de leur probabilité de sinistre offre des opportunités pour développer des programmes de prévention personnalisés, notamment pour les conducteurs à risque élevé, ce qui pourrait réduire le nombre d'accidents à long terme.

En conclusion, cette analyse fournit des leviers concrets pour améliorer la gestion des risques, optimiser les coûts et renforcer la satisfaction client. L'approche basée sur les données crée une dynamique d'efficacité opérationnelle et d'anticipation des risques, tout en garantissant la durabilité financière de l'entreprise.

SOURCE DU CODE R

```
#IMPORTATION DES DONNEES
actuarNV1 <- read.csv("C:/Users/hp/Desktop/INSSEDS/MINI PROJET/ECONOMETRIE/actuarNV1.csv",
sep=";", stringsAsFactors=TRUE)

#SUPPRIMER LES VARIABLES NOCONTRAT ET NO DANS LA BASE DE DONNEE
actuarNV1$nocontrat=NULL
View(actuarNV1)
actuarNV1$no=NULL
View(actuarNV1)


#EXPLORATION DES DONNEES
head(actuarNV1,5)
tail(actuarNV1,5)
str(actuarNV1)
summary(actuarNV1)

#VISUALISATION DES VALEURS MANQUANTES
library(visdat)
vis_dat(actuarNV1)
vis_miss(actuarNV1)

#PRETAITEMENT DES DONNEES
#TRAITEMENT DES DOUBLONS
actuarNV1_1 = unique(actuarNV1)
nrow(actuarNV1) - nrow(actuarNV1_1)

#IDENTIFICATION DU NOMBRE D'INDIVIDUS AYANT DES VALEURS MANQUANTES
actuarNV1[!complete.cases(actuarNV1),]
nrow(actuarNV1[!complete.cases(actuarNV1),])
```

#TRAITEMENT DES VALEURS ABERRANTES EXTREMES

```
par(mfrow=c(2,2), mar=c(3,3,3,3))
boxplot(actuarNV1$exposition, main = "EXPOSITION", col = "blue")
boxplot(actuarNV1$puissance, main = "PUISSANCE", col = "green")
boxplot(actuarNV1$agevehicule, main = "AGEVEHICULE", col ="red")
boxplot(actuarNV1$ageconducteur, main = "AGECONDUCTEUR", col = "brown")
boxplot(actuarNV1$bonus, main = "BONUS", col = "yellow")
boxplot(actuarNV1$densite, main= "DENSITE", col="orange")
boxplot(actuarNV1$region, main = "REGION", col="purple")
boxplot(actuarNV1$nbre, main = "NBRE", col="pink")
boxplot(actuarNV1$cout, main="COUT", col="black")
par(mfrow=c(2,2), mar=c(3,3,3,3))

library(DescTools)
actuarNV1$exposition <- Winsorize(actuarNV1$exposition)
actuarNV1$puissance <- Winsorize(actuarNV1$puissance)
actuarNV1$agevehicule <- Winsorize(actuarNV1$agevehicule)
actuarNV1$ageconducteur <- Winsorize(actuarNV1$ageconducteur)
actuarNV1$bonus <- Winsorize(actuarNV1$bonus)
actuarNV1$densite <- Winsorize(actuarNV1$densite)
actuarNV1$region <- Winsorize(actuarNV1$region)
actuarNV1$nbre <- Winsorize(actuarNV1$nbre)
actuarNV1$cout <- Winsorize(actuarNV1$cout)
boxplot(actuarNV1$exposition, main = "EXPOSITION", col = "blue")
boxplot(actuarNV1$puissance, main = "PUISSANCE", col = "green")
boxplot(actuarNV1$agevehicule, main = "AGEVEHICULE", col ="red")
boxplot(actuarNV1$ageconducteur, main = "AGECONDUCTEUR", col = "brown")
boxplot(actuarNV1$bonus, main = "BONUS", col = "yellow")
boxplot(actuarNV1$densite, main= "DENSITE", col="orange")
boxplot(actuarNV1$region, main = "REGION", col="purple")
boxplot(actuarNV1$nbre, main = "NBRE", col="pink")
boxplot(actuarNV1$cout, main="COUT", col="black")
```

```
#ANALYSE STATISTIQUE UNIVARIEE
#VARIABLE QUANTITATIVE
#ETUDE DE LA VARIABLE PUISSANCE
head(wawa.qt.tableau(actuarNV1$puissance),5)
wawa.qt.graph(actuarNV1$puissance)
#ETUDE DE LA VARIABLE AGEVEHICULE
wawa.qt.resume(actuarNV1$agevehicule)
head(wawa.qt.tableau(actuarNV1$agevehicule))
wawa.qt.graph(actuarNV1$agevehicule)
#ETUDE DE LA VARIABLE AGECONDUCTEUR
wawa.qt.resume(actuarNV1$ageconducteur)
head(wawa.qt.tableau(actuarNV1$ageconducteur))
wawa.qt.graph(actuarNV1$ageconducteur)
#ETUDE DE LA VARIABLE DU NOMBRE D'ACCIDENT
wawa.qt.resume(actuarNV1$nbre)
head(wawa.qt.tableau(actuarNV1$nbre))
wawa.qt.graph(actuarNV1$nbre)
#ETUDE DE LA VARIABLE COUT
wawa.qt.resume(actuarNV1$cout)
head(wawa.qt.tableau(actuarNV1$cout))
wawa.qt.graph(actuarNV1$cout)

#VARIABLE QUALITATIVE
#ETUE DE LA VARIABLE ZONE
head(wawa.ql.tableau(actuarNV1$zone))
wawa.ql.graph(actuarNV1$zone)

#ETUDE DE LA VARIABLE CARBURANT
head(wawa.ql.tableau(actuarNV1$carburant))
wawa.ql.graph(actuarNV1$carburant)

#ETUDE DE LA VARIABLE GARANTIE
head(wawa.ql.tableau(actuarNV1$garantie))
wawa.ql.graph(actuarNV1$garantie)
```

```
#ANALYSE STATISTIQUE BIVARIEE
#VARIABLE QUANTITATIVE
library(corrplot)
library(Factoshiny)
str(actuarNV1)
actuarNV1_numeric <- actuarNV1[1:9]
actuarNV1_numeric <- data.frame(lapply(actuarNV1_numeric, as.numeric))
cor_matrix <- cor(actuarNV1_numeric, use = "complete.obs")
corrplot(cor_matrix)
dev.off()

# REGRESSION LINEAIRE MULTIPLE
## estimation des parametres
#faire la regression
regM <- lm(actuarNV1$cout ~ actuarNV1$exposition + actuarNV1$puissance +
actuarNV1$agevehicule + actuarNV1$ageconducteur + actuarNV1$bonus + actuarNV1$densite +
actuarNV1$nbre + actuarNV1$marque + actuarNV1$region + actuarNV1$garantie + actuarNV1$zone
+ actuarNV1$carburant)
regM$coefficients
summary(regM)
# selection automatiques des variables explicatives
# modèle stepwise backward
modele_complet <- lm(cout ~ . , data = actuarNV1)
modele_2 <- step(modele_complet, direction="backward")
# Résumé du modèle
summary(modele_2)
# Anova du modèle
print(anova(modele_2, test="Chisq"))
# régression sur les variables retenues
reg.fin<-lm(actuarNV1$cout ~ actuarNV1$exposition  + actuarNV1$agevehicule + actuarNV1$densite
+ actuarNV1$nbre + actuarNV1$garantie)
reg.fin$coefficients
```

```
# ANALYSE DES RESIIDUS
# Graphe de l'analyse des résidus
res.m<-rstudent(reg.fin)
plot(res.m,pch=15,cex=.5,ylab="Residus",ylim=c(-3,3))
abline(h=c(-2,0,2),lty=c(2,1,2))
#Calcul de la Valeur des résidus
res.m <- rstudent(reg.fin)
pct_residus <- sum(abs(res.m) <= 2) / length(res.m) * 100
print(pct_residus)

# TEST DE VALIDITE DU MODELE
residus<-residuals(reg.fin)
res.normalise<-rstudent(reg.fin)
val.estimees<-fitted.values(reg.fin)
# Tester de linéarité du modèle
library(lmtest)
raintest(reg.fin)

# Générer des données avec hétéroscédasticité proportionnelle
set.seed(123)
n <- 100  # Nombre d'observations
x <- rnorm(n, mean = 5, sd = 2)  # Variable explicative
beta_0 <- 2  # Intercept
beta_1 <- 3  # Coefficient de la variable explicative
sigma <- 0.5 * exp(0.1 * x)  # Variance proportionnelle à exp(0.1*x)
errors <- rnorm(n, mean = 0, sd = sigma)  # Erreurs hétéroscédastiques
y <- beta_0 + beta_1 * x + errors  # Réponse

# Ajuster le modèle initial avec MCO
model_OLS <- lm(y ~ x)
summary(model_OLS)
```

```
# Estimer les résidus
residuals <- resid(model_OLS)



# Construire la matrice de poids pour les MCG
W <- diag(1 / sigma^2)

# Transformation des données
X <- cbind(1, x)  # Ajout de l'intercept
y_star <- sqrt(W) %*% y
X_star <- sqrt(W) %*% X

# Ajuster le modèle transformé avec MCO
model_MCG <- lm(y_star ~ X_star - 1)  # -1 pour éviter d'ajouter une intercept
summary(model_MCG)

# Extraire les coefficients estimés
coef_MCG <- coef(model_MCG)

# Afficher les coefficients estimés
cat("Les coefficients estimés avec les MCG sont :\n")
cat("Intercept :", coef_MCG[1], "\n")
cat("Pente :", coef_MCG[2], "\n")
library(lmtest)
bptest(model_MCG)

# REGRESSION LOGISTIQUE
## charger le jeu de donnees
actuarNV1 <- read.csv("C:/Users/hp/Desktop/INSSEDS/MINI
PROJET/ECONOMETRIE/actuarNV1.csv", sep=";", stringsAsFactors=TRUE)
unique_modalites <- unique(actuarNV1$nbre)
print(unique_modalites)
## Créer une copie du jeu de données sans les variables "nocontrat" et "no"
actuarNV1 <- actuarNV1[, !names(actuarNV1) %in% c("nocontrat", "no")]
actuarNV1$marque=factor(actuarNV1$marque)
actuarNV1$region=factor(actuarNV1$region)
str(actuarNV1)
## recodage de la variable nbre
actuarNV1$nbre <- ifelse(actuarNV1$nbre == 1, 0, 1)
actuarNV1
```

```
## construction du modèle complet
modele_complet <- glm(nbre ~ ., data = actuarNV1, family = binomial)

summary(modele_complet)
print(anova(modele_complet, test="Chisq"))
## determination du modele final
modele_complet <- glm(nbre ~ ., data = actuarNV1, family=binomial)
modele_final <- step(modele_complet, direction = "backward")
# faire l'anova du modèle
print(anova(modele_final, test="Chisq"))
# modele final
modele_final <- glm(nbre ~ exposition + agevehicule + garantie + bonus
                    + region, data = actuarNV1, family=binomial)
summary(modele_final)
##  calcul des odd ratio et des effets marginaux
# affichage des paramètres estimés
PARAMETRES = coefficients(modele_final)
PARAMETRES
# Calculons les intervalles de confiance
confint(modele_final)
# calcul des odd ratio
ODD_RATIO = exp(coefficients(modele_final))
ODD_RATIO
# calcul des effets marginaux
install.packages("mfx")
library(mfx)
logitmfx(nbre ~ agevehicule + garantie + exposition + bonus + region, data = actuarNV1)

## Analyse des résidus
res.m <- rstudent(modele_final)

plot(res.m,pch=15,cex=.5,ylab="Residus",ylim=c(-3,3))
abline(h=c(-2,0,2),lty=c(2,1,2))
res.m <- rstudent(modele_final)
sum(as.numeric(abs(res.m)<=3))/nrow(actuarNV1)*100

## calcul du taux de mauvais classement et Matrice de Confusion
# calcul des probabilité
actuarNV1$PROBABILITE_PREDITE <- predict(modele_final, actuarNV1, type="response")
```

```
head(actuarNV1)
# Transformer les probabilité en modalite predites
actuarNV1$MODALITE_PREDITE <- ifelse(actuarNV1$PROBABILITE_PREDITE < 0.5, "pas accident" , "accident")
head(actuarNV1)
# Caclculer le taux de mauvais classement à partir de la matrice de confusion
Matrice_confusion <- table(actuarNV1$MODALITE_PREDITE, actuarNV1$nbre)
rownames(Matrice_confusion) <- c("accident", "pas accident")
colnames(Matrice_confusion) <- c("accident", "pas accident")
print(Matrice_confusion)
#REGRESSION DE POISSON
## visualisation de la variable cible
ggplot(actuarNV1, aes(x = nbre)) + geom_histogram(bins = 20, fill = "red", color = "black") +labs(title = "Distribution du nombre d'axe", x = "Nombre d'accident", y = "Fréquence")
# test d'adequation a la loi de poisson
fpois <- fitdist(actuarNV1$nbre, "pois")
gofstat(fpois)
print(fpois)
plot(fpois)
## construction du modele complet
reg <- glm(nbre ~ ., data = actuarNV1, family = quasipoisson)
summary(reg)
anova(reg,test="Chisq") #Voir la significativité des variable à l'aide d'anova
## Construction du modèle avec les variables significatives
mod = glm(nbre ~exposition+agevehicule+ageconducteur+marque+bonus+carburant+garantie,data=actuarNV1,family = quasipoisson)
anova(mod,test="Chisq")
## Interprétation des coefficients
exp(mod$coefficients)
str(actuarNV1)
## Analyse des résidus
res.m<-rstudent(mod)
plot(res.m,pch=15,cex=.5,ylab="Residus",ylim=c(-3,3))
abline(h=c(-2,0,2),lty=c(2,1,2))
sum(as.numeric(abs(res.m)<=3))/nrow(actuarNV1)*100 #Calcul du pourcentage des résidus

## Test de normalité des résidus
qqnorm(res.m) #Graphique pour voir la normalité des résidus
shapiro.test(res.m) #Test de normalité des résidus
```