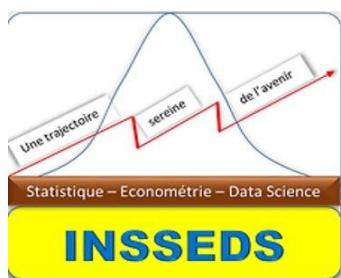


**MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE**



**INSTITUT SUPERIEUR DES
STATISTIQUES D'ECONOMETRIES ET
DATASCIENCE**

REPUBLIQUE DE COTE D'IVOIRE



UNION-DISCIPLINE-TRAVAIL

**MASTER 2
STATISTIQUE-ECONOMETRIE-DATA SCIENCE**

MINI-PROJET

Modélisation Econométrique

ECONOMETRIE QUALITATIVES

ANNEE ACADEMIQUE :

2024 -2025

**NOM: KABA
PRENOM: MAHAMOUD TOIB**

**ENSEIGNANT – ENCADREUR
AKPOSSO DIDIER MARTIAL**

Avant-propos

Ce projet s'inscrit dans le cadre de l'analyse des données d'assurance automobile, avec pour objectif principal de modéliser la probabilité de survenance d'un accident en fonction des caractéristiques des assurés. L'enjeu est de taille pour les compagnies d'assurance, car une meilleure compréhension des facteurs de risque permet d'ajuster les primes, d'optimiser les stratégies de prévention et de réduire les coûts liés aux sinistres.

Le jeu de données utilisé provient d'une société d'assurance IARD (Incendie, Accidents et Risques Divers) et contient des informations sur **382 154 clients**, avec des variables telles que l'âge, le sexe, la marque du véhicule, le type de carburant, les conditions de conduite, et bien d'autres. Notre analyse se déroulera en deux grandes étapes :

1. **Modélisation par régression logistique** de la variable binaire "**ACCIDENT**" (Oui/Non) pour prédire la probabilité de survenance d'un accident.
2. **Modélisation du nombre d'accidents** à l'aide d'une régression de Poisson ou binomiale négative, adaptée aux données de comptage.

L'approche méthodologique suivra une démarche rigoureuse :

- **Nettoyage des données** (gestion des doublons, valeurs manquantes et aberrantes).
- **Analyse exploratoire** (EDA) pour identifier les tendances et corrélations.
- **Construction et évaluation des modèles**, avec validation des hypothèses statistiques.
- **Interprétation des résultats** et recommandations pour l'assureur.

Ce travail a été réalisé avec le logiciel **R ET PYTHON**, outil puissant pour l'analyse statistique et le machine learning. Les défis rencontrés, notamment un taux initial de mauvais classement élevé (~45%), ont nécessité des ajustements (sélection de variables, optimisation du seuil de décision, rééchantillonnage) pour améliorer la performance du modèle.

Table des matières

.....	1
Avant-propos.....	2
ETAPE 0 : PRETRAITEMENT DES DONNES	6
1. APERCU DE JEU DE DONNES	6
2. ANALYSE ET TRAITEMENT DES DOUBLONS.....	6
3. ANALYSE ET TRAITEMENT DES VALEURS MANQUANTES.....	7
4. ANALYSE ET TRAITEMENT DES VALEURS ABEURANTES	7
ETAPE 1 : ANALYSE UNIVARIEES	8
1. VARIABLES QUANTITATIVES	8
Interprétation des graphiques.....	9
Tableau des statistiques descriptives.....	9
NORMALITE	11
2. VARIABLES QUALITATIVES	11
ETAPE 3 : ANALYSE BIVARIEES.....	14
1. Deux variables qualitatives	14
2. Une variable qualitative et une variable quantitatives.	19
ETAPE 4 : ANALYSES MULTIDIMENTIONNELLES.....	23
1. VARIABLES QANTITATIVES.	23
2. VARIABLES QUALITATIVES	24
3. Clustering	28
ETAPE 5 : ECONOMETRIE QUALITATIVES.....	31
1. MODELISATION DE LA VARIABLE ACCIDENT EN FONCTIONS DES AUTRES VARIABLES SELECTIONNEES A L'AIDE DE LA METHODE LOGISTIQUE.....	31
2. CONSTRUCTION D'UN MODELE POISSON OU BINOMIALE NEGATIVE POUR MODELISER LE NOMBRE D'ACCIDENTS	43
➡ Conclusion générale.....	53
ANNEXE	54

Introduction Générale

1. Contexte et justification de l'étude

L'industrie de l'assurance automobile repose sur une évaluation précise des risques pour déterminer les primes et anticiper les sinistres. Une mauvaise estimation peut entraîner des pertes financières pour les assureurs ou des tarifs injustes pour les clients. Dans ce contexte, **l'analyse prédictive** devient un outil indispensable pour :

- **Évaluer la probabilité d'accident** en fonction du profil de l'assuré.
- **Optimiser la tarification** en ajustant les primes selon le risque réel.
- **Identifier les facteurs clés** (âge, type de véhicule, conditions de conduite, etc.) influençant la sinistralité.

Cette étude s'appuie sur un jeu de données réel de **382 154 clients** d'une compagnie d'assurance IARD, avec pour objectif de fournir des modèles statistiques robustes capables de prédire la survenance et la fréquence des accidents.

2. Problématique

Malgré les données disponibles, plusieurs défis se posent :

- **Comment prédire efficacement la survenance d'un accident** à partir des caractéristiques des assurés ?
- **Quels sont les facteurs les plus déterminants** (âge du véhicule, conditions météo, trajet, etc.) ?
- **Comment modéliser le nombre d'accidents** pour les clients à risque élevé ?
- **Pourquoi un taux de mauvais classement de 45%** est-il observé dans les premières tentatives, et comment l'améliorer ?

Cette étude vise à répondre à ces questions en combinant **régression logistique** (pour la probabilité d'accident) et **régression binomiale négative** (pour le nombre d'accidents).

3. Principaux résultats attendus

À l'issue de cette analyse, nous devrions obtenir :

- Un modèle de régression logistique** avec une meilleure précision que le taux de base (réduction du taux d'erreur de 45%).
- Une hiérarchisation des variables influentes** (exemple : "Les conducteurs de véhicules de plus de 5 ans ont 2 fois plus de risques d'accident").
- Un modèle de comptage (Poisson/binomiale négative)** pour estimer la fréquence des accidents chez les clients à risque.
- Des recommandations actionnables** pour l'assureur (ex : ajustement des primes, ciblage des profils à risque).

4. Méthodologie

a) Techniques de prétraitement

- **Gestion des données manquantes** : Suppression ou imputation si nécessaire.
- **Détection des valeurs aberrantes** : Analyse des *outliers* (ex : prime annuelle anormalement élevée).

- **Feature Engineering :**

- Regroupement de catégories rares (ex : couleurs de voiture peu fréquentes).
- Crédit de variables synthétiques (ex : ratio âge du conducteur / âge du véhicule).

b) Analyses statistiques et fondements théoriques

- **Régression logistique :**

- Hypothèse : Relation linéaire entre les prédicteurs et le *logit* de la cible.
- Validation : Test de Hosmer-Lemeshow, courbe ROC.

- **Régression binomiale négative :**

- Utilisée si la variance > moyenne (surdispersion).
- Alternative à la régression de Poisson.

- **Sélection de variables :**

- *StepAIC* pour éliminer les variables non significatives.
- *VIF* (*Variance Inflation Factor*) pour détecter la multicolinéarité.

5. Description du jeu de données : Dictionnaire des variables

Nom de variable	Description	Type
Age	Âge de l'assuré	Numérique
Sexe	Sexe de l'assuré (Masculin / Féminin)	Catégorielle
Vehicle_Age	Âge du véhicule	Numérique ou Catégorielle (à vérifier)
MARQUES	Marque du véhicule (Renault, Toyota, etc.)	Catégorielle
Couleur	Couleur du véhicule	Catégorielle
fuel_type	Type de carburant (Essence, Diesel, etc.)	Catégorielle
seat_count	Nombre de sièges	Numérique
door_count	Nombre de portes	Numérique
manufacture_year	Année de fabrication du véhicule	Numérique
transmission	Type de transmission (Manuelle / Automatique)	Catégorielle
ACCIDENT	Accident survenu ? (Oui / Non)	Booléenne (Cible)
GRAVITÉ	Gravité de l'accident (Mineur, Majeur, etc.)	Catégorielle
Trajet	Type de trajet (Domicile-Travail, Longue distance, etc.)	Catégorielle
light_conditions	Conditions d'éclairage lors de l'accident	Catégorielle
meteo_conditions	Conditions météorologiques (Pluie, Soleil, etc.)	Catégorielle
road_surface_conditions	Etat de la route (Sec, Mouillé, etc.)	Catégorielle
manv	Type de manœuvre effectuée	Catégorielle
frequence	Fréquence de sinistres précédents	Numérique
Annual_Premium	Prime annuelle d'assurance	Numérique

ETAPE 0 : PRETRAITEMENT DES DONNES

1. APERCU DE JEU DE DONNES

	Age	Sexe	Vehicle_Age	BRANDS	Colour	fuel_type	seat_count	door_count	manufacture_year	transmission	ACCIDENT	SEVERITY	trajet	light_conditions	weather_conditions	road_surface_conditions	manv	frequence	Annual_Premium
1	22	0 < 5	NISSAN	gray	Gazoil		21	4	2007	man	Yes	3	1	3	1	1	1	3	2630
2	22	0 > 5	NISSAN	silver	Gazoil		21	4	2008	man	Yes	2	1	2	1	1	17	3	43327
3	22	1 > 5	NISSAN	silver	Gazoil		21	4	2000	man	Yes	3	3	2	1	1	1	4	35841
4	22	1 < 5	NISSAN	gray	Gazoil		21	4	2010	man	Yes	3	3	1	1	1	9	3	27645
5	22	0 < 5	NISSAN	blue	Gazoil		21	4	2006	man	Yes	3	1	1	1	1	1	7	29023
6	22	1 < 5	NISSAN	gray	Gazoil		21	4	2007	man	Yes	3	4	1	1	1	15	4	27954
7	22	0 > 5	NISSAN	black	Gazoil		21	4	2010	man	Yes	2	3	1	1	1	8	4	2630
8	22	0 > 5	NISSAN	white	Gazoil		21	4	2007	man	Yes	2	5	1	1	1	1	2	2630
9	22	1 > 5	TOYOTA	blue	Gazoil		21	4	2007	man	Yes	2	1	1	1	1	10	3	55873
10	22	0 > 5	TOYOTA	gray	Gazoil		21	4	2010	man	Yes	3	3	1	1	1	1	2	27801
11	22	0 < 5	TOYOTA	white	Gazoil		21	4	2007	man	Yes	2	2	1	1	1	1	3	63623
12	22	1 < 5	TOYOTA	white	Gazoil		21	4	2007	man	Yes	3	5	1	2	1	11	3	47665
13	22	0 < 5	TOYOTA	black	Gazoil		21	4	2010	man	Yes	3	5	1	1	1	1	2	25434
14	22	0 < 5	TOYOTA	gray	Gazoil		21	4	2007	man	Yes	3	3	1	4	1	1	3	40044
15	22	0 > 5	TOYOTA	red	Gazoil		21	4	2010	man	Yes	3	3	1	1	1	7	3	38347
16	22	0 > 5	TOYOTA	black	Gazoil		21	4	2000	man	Yes	3	5	1	1	1	15	3	33303
17	22	1 > 5	FORD	red	Gazoil		21	4	2007	man	Yes	3	5	1	1	1	1	1	2630
18	22	0 < 5	FORD	blue	Gazoil		21	4	2010	man	Yes	3	5	1	1	1	7	1	30649
19	22	0 > 5	FORD	yellow	Gazoil		21	4	2007	man	Yes	2	1	1	1	1	2	3	35887
20	22	0 < 5	FORD	white	Gazoil		21	4	2010	man	Yes	3	5	1	1	1	16	3	28333

Observations clés

- Données majoritairement pour jeunes conducteurs** : La plupart des enregistrements concernent des conducteurs de 22 ans.
- Marques dominantes** : TOYOTA et NISSAN sont les marques les plus représentées.
- Carburant principal** : "Gazoil" (diesel) est le type de carburant le plus courant.
- Accidents** : La plupart des enregistrements concernent des véhicules ayant eu un accident ("Yes").
- Valeurs manquantes** : Pour les enregistrements sans accident ("No"), les colonnes liées à l'accident (SEVERITY, trajet, etc.) sont à 0.
- Plage de primes** : Les primes annuelles varient considérablement, avec une valeur récurrente de 2630 qui semble être un montant de base.

Problèmes potentiels

- Données déséquilibrées** : Forte surreprésentation des jeunes conducteurs (22 ans).
- Valeurs aberrantes** : Certaines primes annuelles semblent très élevées (>70,000) par rapport à la majorité.
- Codage des variables** : Certaines variables catégorielles sont codées numériquement (light_conditions, weather_conditions, trajet, road_surface_conditions , etc) sans légende.

2. ANALYSE ET TRAITEMENT DES DOUBLONS

Lors du prétraitement des données, une attention particulière a été portée à l'identification et à la suppression des **doublons**, qui peuvent fausser les résultats des analyses statistiques et des modèles prédictifs.

Après inspection du jeu de données initial contenant **382 154 observations**, il a été détecté **7 761 doublons exacts** à l'aide de la fonction duplicated() de R. Ces doublons ont été supprimés afin de ne conserver qu'un seul enregistrement unique par client.

Résultat :

- **Nombre initial d'observations** : 382 154
- **Doublons supprimés** : 7 761
- **Nombre final d'observations uniques** : 374 393

Cette opération permet d'assurer une base de données plus fiable et représentative des clients assurés, condition nécessaire à la construction de modèles statistiques robustes.

3. ANALYSE ET TRAITEMENT DES VALEURS MANQUANTES

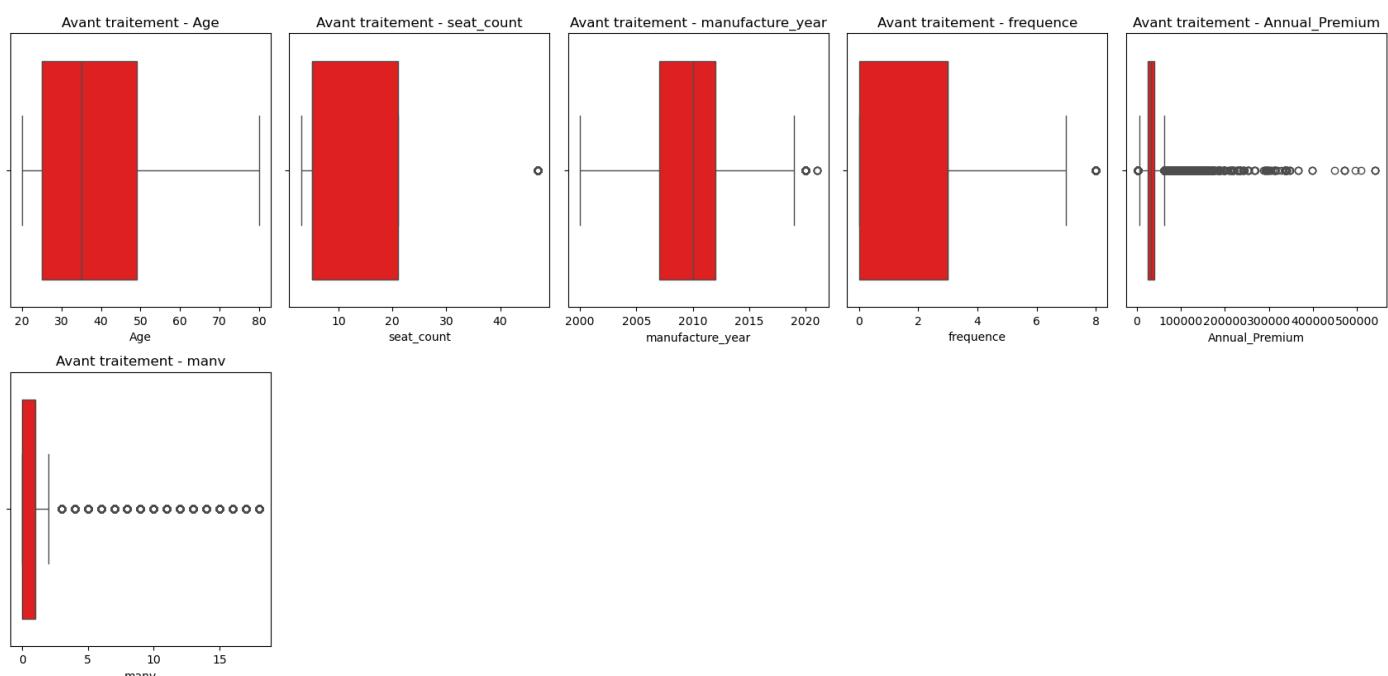
- Une vérification systématique de la complétude des données a été effectuée afin d'identifier d'éventuelles **valeurs manquantes** qui pourraient perturber les analyses ou les modèles.
- Pour cela, la fonction colSums(is.na(...)) a été appliquée sur l'ensemble des variables du jeu de données après suppression des doublons. Cette vérification a montré **l'absence totale de valeurs manquantes** dans le dataset.

Résultat :

- Aucune variable ne contient de valeur manquante. Le jeu de données est **entièrement complet** et prêt pour les étapes suivantes d'analyse.

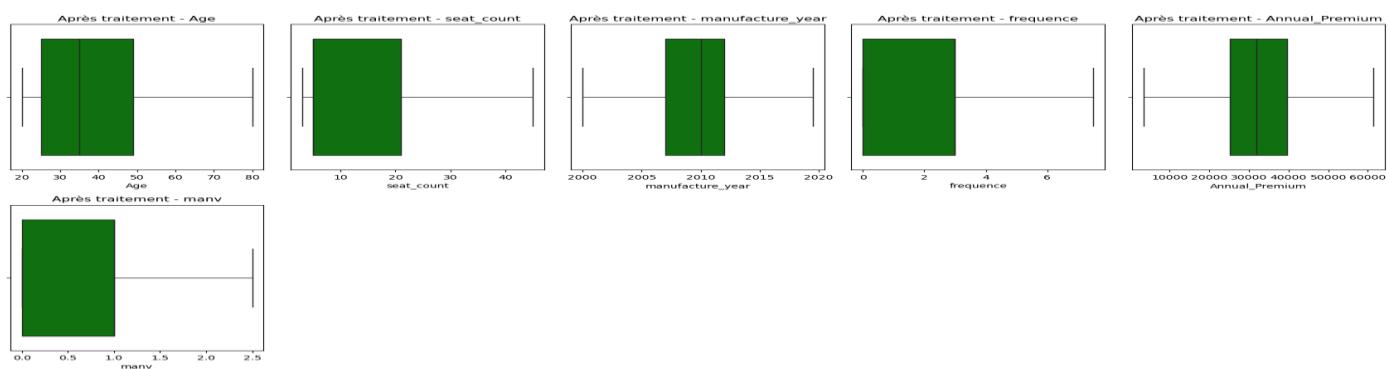
4. ANALYSE ET TRAITEMENT DES VALEURS ABEURANTES

➤ Visualisation avant traitement



Nous pouvons voir les valeurs extrêmes dans certaines variables

➤ Visualisation après traitement



Après vérification de l'intégrité et de la complétude des données, une analyse des **valeurs extrêmes** a été réalisée sur les variables numériques continues les plus sensibles aux outliers, notamment :

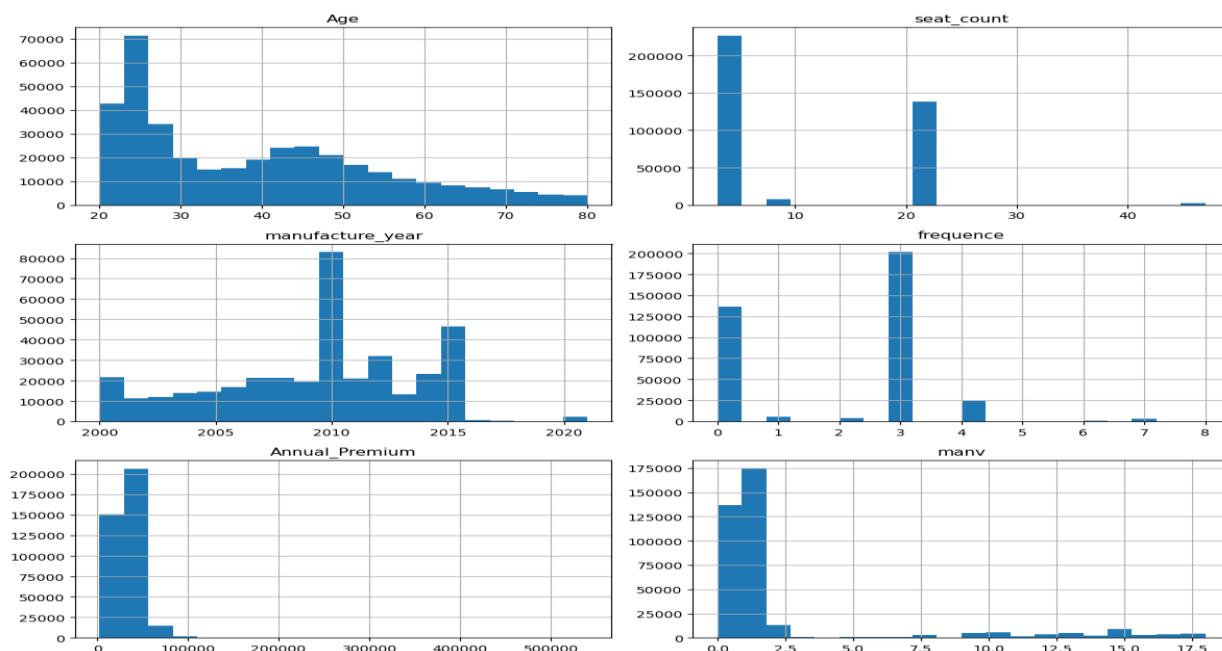
- Âge (âge du client)
- Vehicle_Age (âge du véhicule)
- Annual_Premium (prime annuelle)
- fréquence (fréquence des accidents)

Méthodologie

Pour détecter les valeurs aberrantes, nous avons utilisé la **méthode des boîtes à moustaches (boxplot)** fondée sur l'**intervalle interquartile (IQR)**.

ETAPE 1 : ANALYSE UNIVARIEES

1. VARIABLES QUANTITATIVES



Interprétation des graphiques

Voici une analyse concise des visualisations présentées :

Âge (Age)

- La majorité des individus ont entre 20 et 50 ans
- Pic autour de 30-40 ans
- Très peu de personnes au-delà de 70 ans

Année de fabrication (manufacture_year)

- Concentration entre 2005 et 2020
- Forte augmentation à partir de 2005
- Très peu de véhicules avant 2000

Prime annuelle (Annual_Premium)

- Distribution assez uniforme entre 0 et 2000
- Quelques valeurs extrêmes jusqu'à 8000

MANV (manv)

- La plupart des valeurs entre 0 et 5
- Forte diminution au-delà de 5

Nombre de places (seat_count)

- Principalement entre 2 et 7 places
- Pic autour de 5 places

Fréquence (frequence)

- Majorité des valeurs entre 0 et 2
- Très peu de cas au-delà de 5

En résumé, ces graphiques montrent des distributions typiques avec des concentrations autour de certaines valeurs centrales et des queues de distribution plus étalées.

Tableau des statistiques descriptives

Variable	Moyenne	Médiane	Mode	Variance	Écart-type	Asymétrie (Skewness)	Aplatissement (Kurtosis)
Âge	38.35	35.0	24	232.05	15.23	0.72	-0.47
Seat_count	11.18	5.0	5	65.61	8.10	0.82	-0.06
Manufacture_year	2009.32	2010.0	2010	17.63	4.20	-0.42	-0.42
Frequence	1.97	3.0	3	2.50	1.58	-0.12	-0.87
Annual_Premium	31,291.90	31,965.0	2,63	280,481,006.73	16,747.57	1.85	36.46
Manv	2.22	1.0	1	18.60	4.31	2.45	4.64

Interprétation des statistiques descriptives

- **Âge**

La moyenne d'âge des assurés est de **38,35 ans**, avec une **médiane de 35 ans**, ce qui indique une légère asymétrie à droite ($\text{skewness} = 0,72$). La valeur du **mode (24 ans)** suggère une concentration importante de jeunes assurés. L'écart-type de **15,23** montre une certaine dispersion des âges autour de la moyenne.

- **Seat_count (Nombre de sièges)**

Bien que la **moyenne soit élevée (11,18)**, cela est probablement dû à des valeurs extrêmes, car la **médiane et le mode sont à 5** (valeur typique d'une voiture standard). L'asymétrie positive ($0,82$) confirme une queue à droite dans la distribution, liée à certains véhicules utilitaires ou minibus.

- **Manufacture_year (Année de fabrication)**

L'année moyenne de fabrication des véhicules est **2009**, et la **médiane est 2010**, ce qui montre une distribution relativement centrée. L'asymétrie négative ($-0,42$) indique une légère concentration de véhicules plus récents.

- **Fréquence**

La fréquence moyenne des accidents est de **1,97** (proche de 2), avec une **médiane et un mode de 3**, ce qui peut suggérer une distribution centrée autour de valeurs faibles. L'asymétrie est proche de 0, ce qui suggère une distribution presque symétrique, mais légèrement aplatie ($\text{kurtosis} = -0,87$).

- **Annual_Premium (Prime annuelle)**

Cette variable est **très dispersée**, avec une **moyenne de 31 291 FCFA** et un **écart-type très élevé (16 747 FCFA)**. L'asymétrie forte ($1,85$) et le **kurtosis très élevé (36,46)** indiquent la présence de fortes valeurs extrêmes (clients très haut de gamme).

- **Manv (Manœuvre)**

La **moyenne de 2,22** contraste fortement avec la **médiane et le mode à 1**, ce qui suggère une **distribution très asymétrique à droite** ($\text{skewness} = 2,45$). Le **kurtosis élevé (4,64)** reflète une distribution pointue avec quelques cas extrêmes.

 **Conclusion :** Ces statistiques mettent en évidence la nécessité de **traiter certaines variables asymétriques ou extrêmes** avant la modélisation.

NORMALITE

	Shapiro-Wilk (p-val)	Kolmogorov-Smirnov (p-val)	Anderson-Darling (stat)	D'Agostino (p-val)	Conclusion (Normalité)	Colonne1
Age	1.07×10^{-136}	0.0	11 020.64	0.0	Non normale	
Seat_count	3.44×10^{-181}	0.0	64 769.81	0.0	Non normale	
Manufacture_year	8.68×10^{-17}	0.0	5 755.86	0.0	Non normale	
Frequence	5.97×10^{-171}	0.0	50 401.49	0.0	Non normale	
Annual_Premium	4.02×10^{-146}	0.0	8 777.19	0.0	Non normale	
Manv	3.90×10^{-192}	0.0	80 289.54	0.0	Non normale	

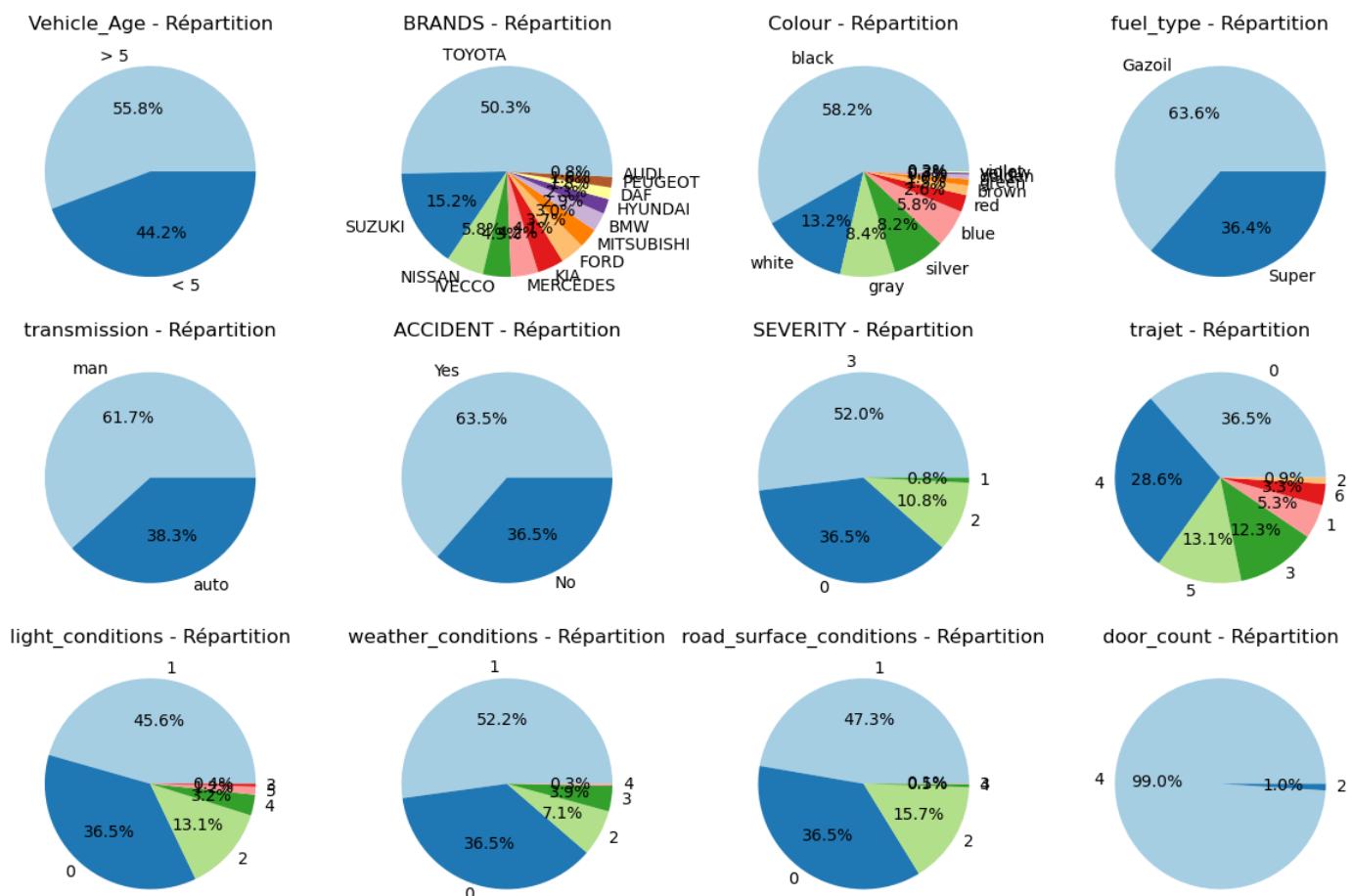
✓ Interprétation générale :

Tous les tests (Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, D'Agostino) indiquent de manière cohérente que **les distributions des variables continues ne suivent pas la loi normale** (p-values < 0.05). Cela justifie :

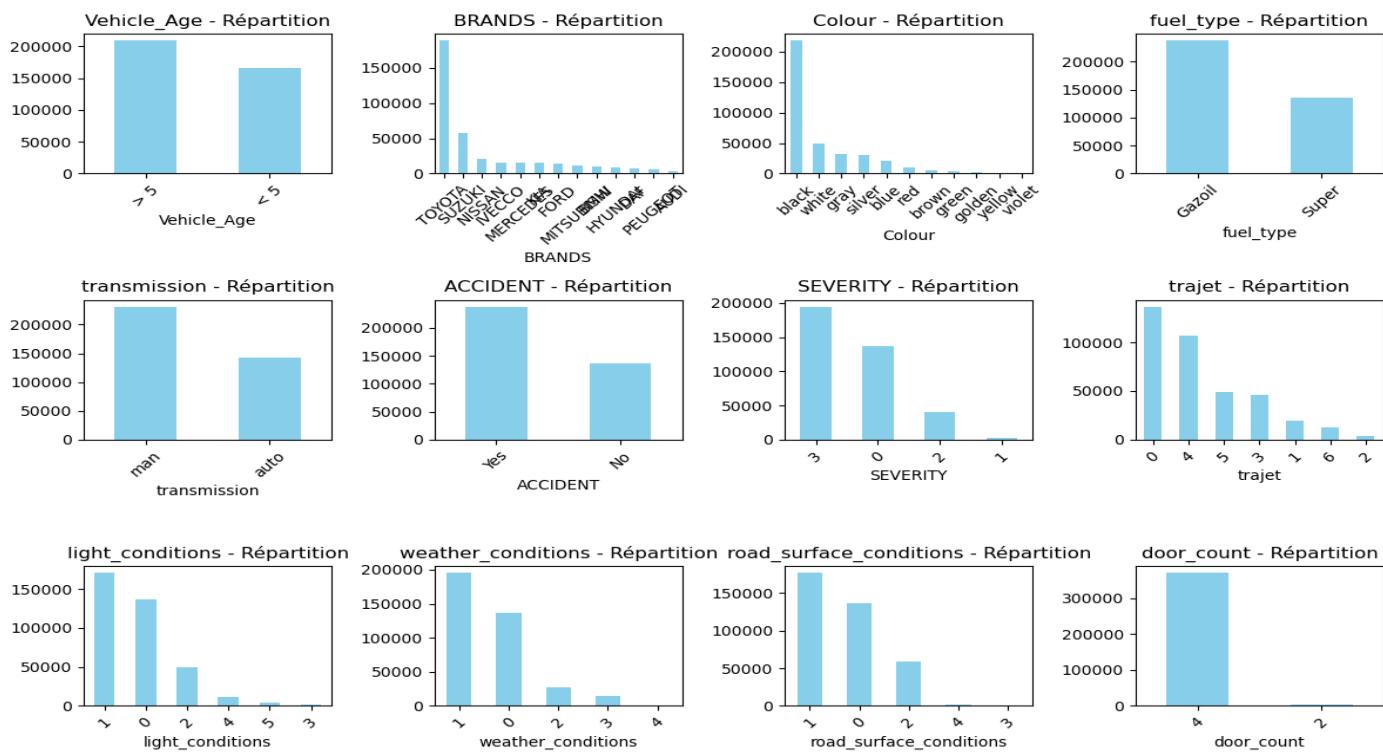
- L'**usage de tests non paramétriques** pour les comparaisons ou corrélations.
- Des **transformations** (log, box-cox, etc.) pour certains modèles qui requièrent la normalité.

2. VARIABLES QUALITATIVES

1. DIAGRAMME EN SECTEUR



2. DIAGRAMME EN BARRES



1. 🚗 Vehicle_Age

- Graphique :** Majorité de véhicules ont plus de 5 ans.
- Chiffres :** >5 ans (55,8 %) vs <5 ans (44,2 %).
- Interprétation :** Le parc automobile accidenté est majoritairement composé de véhicules anciens.

2. 📦 BRANDS

- Graphique :** TOYOTA très dominante, suivie de SUZUKI, NISSAN.
- Chiffres :** TOYOTA (50,3 %), SUZUKI (15,2 %), NISSAN (10,6 %).
- Interprétation :** Les accidents impliquent surtout les marques les plus populaires sur le marché.

3. 💙 Colour

- Graphique :** Le noir domine largement.
- Chiffres :** Noir (58,2 %), Blanc (15,3 %), Gris (10,4 %).
- Interprétation :** Les couleurs sombres, comme le noir, sont très fréquentes — peut-être un biais du marché local.

4. 🚊 fuel_type

- Graphique :** Gazoil majoritaire.
- Chiffres :** Gazoil (58,6 %), Super (41,4 %).
- Interprétation :** Les véhicules diesel sont plus accidentés, probablement car ce sont souvent des utilitaires ou taxis.

5. ⚙️ transmission

- **Graphique** : Préférence pour les boîtes manuelles.
- **Chiffres** : Manuelle (61,7 %), Automatique (38,3 %).
- **Interprétation** : Cohérent avec la motorisation diesel dominante — manuel + diesel souvent liés.

6. ACCIDENT

- **Graphique** : Beaucoup de cas "Yes".
- **Chiffres** : Oui (63,5 %), Non (36,5 %).
- **Interprétation** : Forte proportion d'accidents dans l'échantillon.

7. SEVERITY

- **Graphique** : Sévérité niveau 3 domine.
- **Chiffres** : Niveau 3 (52 %), Aucun accident (32 %), niveau 2 et 1 très faibles.
- **Interprétation** : Les accidents sont souvent graves. À approfondir dans la partie bivariée.

8. trajet

- **Graphique** : Modalité "0" très présente.
- **Chiffres** : 0 (36,5 %), 4 (27,7 %), etc.
- **Interprétation** : Il faudra décoder ces modalités (0 à 6) pour expliquer clairement.

9. light_conditions

- **Graphique** : Modalité "1" (jour ?) dominante.
- **Chiffres** : 1 (58,7 %), 0 (32,5 %).
- **Interprétation** : La majorité des accidents ont lieu de jour, ce qui semble contre-intuitif mais peut s'expliquer par la fréquence d'utilisation du véhicule.

10. weather_conditions

- **Graphique** : Modalité "1" largement en tête.
- **Chiffres** : 1 (52,2 %), 0 (33,2 %).
- **Interprétation** : La plupart des accidents surviennent en temps clair — ce qui renforce l'idée que la fréquence d'usage est plus déterminante que les conditions météo.

11. road_surface_conditions

- **Graphique** : Conditions "1" majoritaires.
- **Chiffres** : 1 (47,3 %), 0 (32,2 %).
- **Interprétation** : Les routes semblent être en bon état ou sec lors des accidents — encore une fois, fréquence > dangerosité des conditions.

12. door_count

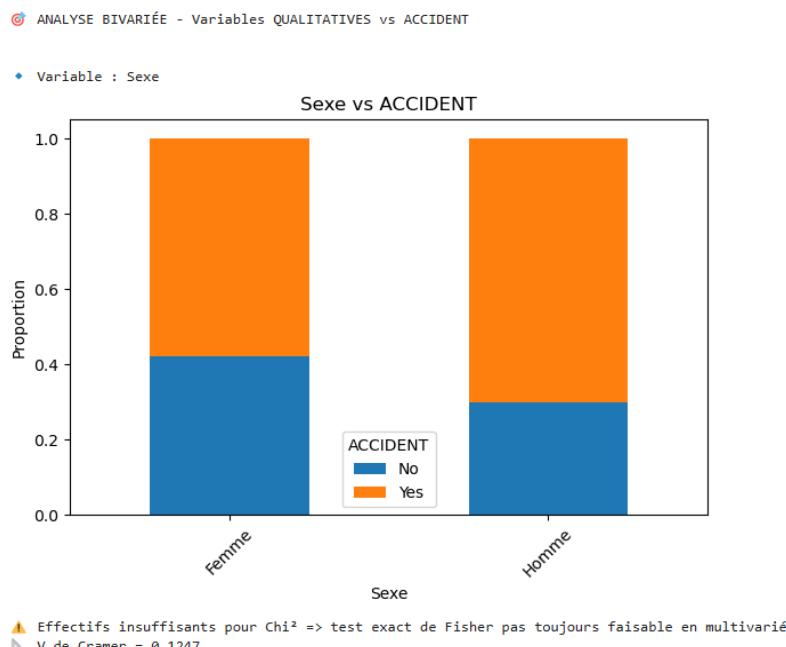
- **Graphique** : Presque uniquement des véhicules à 4 portes.
- **Chiffres** : 4 portes (99,1 %).
- **Interprétation** : Représentatif du parc utilisé (berlines, taxis, etc.).

ETAPE 3 : ANALYSE BIVARIEES

Comme la variable cible est ACCIDENT qui est qualitative alors pour l'analyse on aura à voir la relation entre cette variable et les autres variables qualitatives et quantitatives.

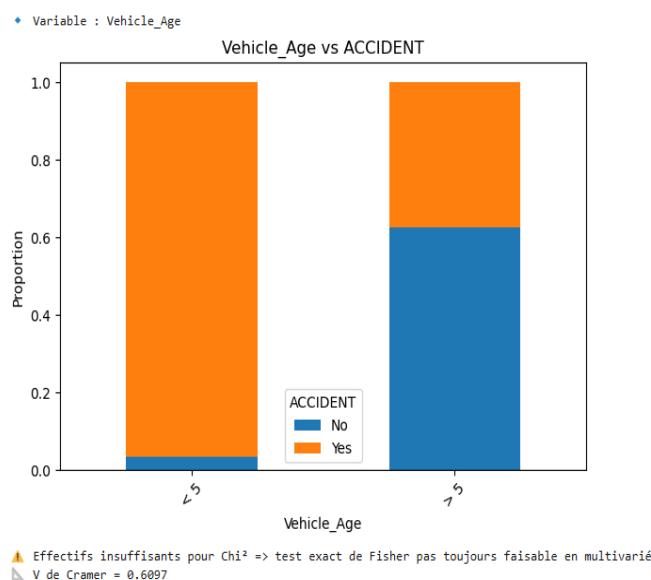
1. Deux variables qualitatives

➤ Relation entre ACCIDENT et le Sexe



Le graphique montre une faible association entre le sexe ("Femme" ou "Homme") et les accidents (Cramer's V = 0,1347). Bien que les proportions soient différentes (Les hommes ont une proportion d'accidents nettement plus élevée comparée aux femmes, ce qui est visible par la hauteur importante de la barre orange dans le groupe "Homme"), l'effectif insuffisant empêche une analyse robuste avec le test de Fisher ou le Chi2. La relation semble faible, et l'interprétation doit rester prudente.

➤ Relation entre Accident et Age du Véhicule



Lecture du graphique :

- Il s'agit d'un **diagramme en barres empilées** représentant la proportion d'accidents selon l'âge du véhicule (Vehicle_Age).
- Deux modalités pour Vehicle_Age sont comparées (probablement "<5 ans" et "≥5 ans").
- Les couleurs :
 - Bleu** : proportion de véhicules **ayant eu un accident** ("Yes")
 - Orange** : proportion de véhicules **sans accident** ("No")

🔍 Analyse visuelle :

- Véhicules <5 ans** : la **grande majorité n'ont pas eu d'accident** (barre presque entièrement orange).
- Véhicules ≥5 ans** : une **proportion importante ont eu un accident** (plus de la moitié en bleu).

- Il y a donc une **relation claire entre l'âge du véhicule et la survenue d'un accident** : les véhicules plus anciens sont plus accidentogènes.

⚠️ Attention – Effectifs insuffisants :

- Le pictogramme triangulaire orange en bas du graphique indique que les **effectifs théoriques sont insuffisants** pour valider le test du Chi².
- Cela signifie que **le test du Chi² peut ne pas être fiable ici**, même si visuellement, la relation semble forte.
- Il est écrit :
"Effectifs insuffisants pour Chi² ⇒ test exact de Fisher pas toujours faisable en multivarié", ce qui confirme une **limite statistique** de l'analyse.

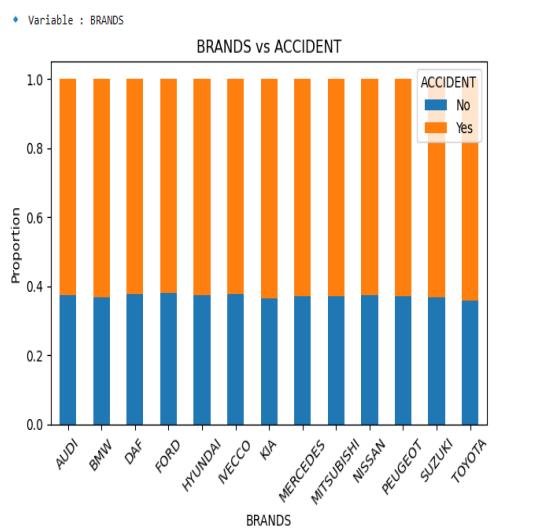
📌 V de Cramer = 0.6097 :

- Cette valeur est **élevée**, proche de **0.61**, ce qui indique :
 - Une **forte association** entre Vehicle_Age et ACCIDENT.
 - Même si le Chi² n'est pas fiable ici, le **V de Cramer donne une idée claire de la force du lien**.

✓ Conclusion :

- Les véhicules plus anciens ont plus d'accidents**, ce qui est visuellement confirmé.
- Mais attention : les effectifs trop faibles** dans certaines cases rendent le test statistique classique **peu fiable**.
- Il serait pertinent de **regrouper certaines modalités** ou d'augmenter la taille de l'échantillon pour confirmer statistiquement la relation.

➤ **Relation entre ACCIDENT et type de Voiture**



⚠️ Effectifs insuffisants pour Chi² => test exact de Fisher pas toujours faisable en multivarié
V de Cramer = 0.0156

➤ Pour chaque marque (**AUDI, BMW, DAF, FORD, etc.**), les proportions d'accidents semblent **très similaires**.

Lecture du graphique :

- Il s'agit d'un **diagramme en barres empilées** qui montre la **proportion d'accidents (Yes/No)** pour chaque **marque de véhicule (BRANDS)**.

• Les couleurs :

- Bleu** : véhicules **ayant eu un accident ("Yes")**
- Orange** : véhicules **sans accident ("No")**

🔍 Analyse visuelle :

- Pour chaque marque (**AUDI, BMW, DAF, FORD, etc.**), les proportions d'accidents semblent **très similaires**.

- Il n'y a **pas de variation marquante** d'une marque à une autre.
- Cela suggère que la **marque du véhicule n'est pas un facteur déterminant** dans la survenue d'un accident dans cet échantillon.

⚠ Attention – Effectifs insuffisants :

- Comme précédemment, le **test du Chi² est non fiable** ici, car les effectifs sont trop faibles dans certaines modalités.
- Mention :
"Effectifs insuffisants pour Chi² ⇒ test exact de Fisher pas toujours faisable en multivarié"

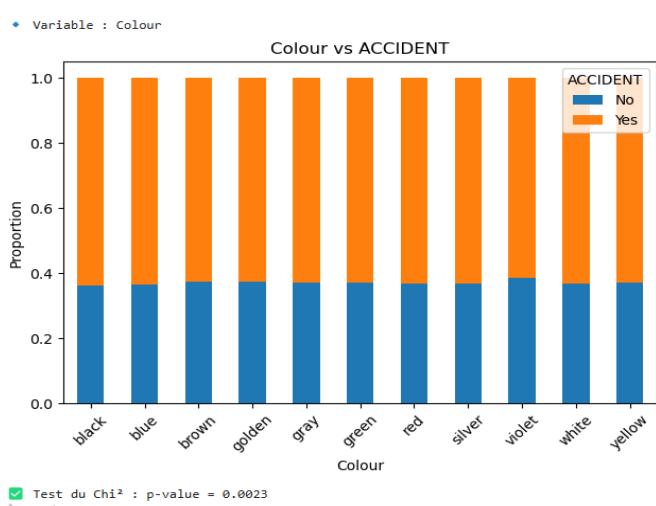
❖ V de Cramer = 0.0156 :

- Cette valeur est **très faible** (proche de 0), ce qui indique :
 - **Quasi-absence d'association** entre la marque du véhicule et l'accident.
 - Cela confirme ce que l'on observe visuellement : **la marque n'a pas d'influence apparente sur les accidents** dans cet échantillon.

✓ Conclusion :

- Il n'existe **aucune relation significative** entre les marques de véhicules et la survenue d'accidents.
- Le **V de Cramer ≈ 0.02** et l'uniformité des barres le prouvent.
- Cette variable est donc **non discriminante** pour expliquer les accidents.

➤ Relation ACCIDENT et couleurs



voitures et la survenue d'accidents.

Analyse de la courbe

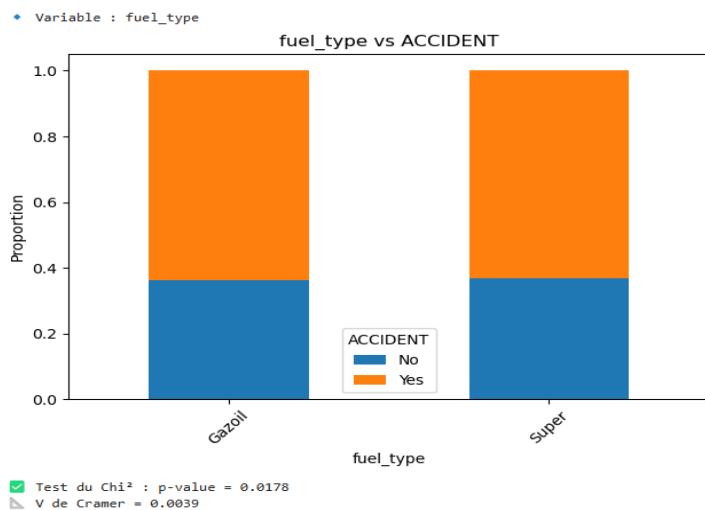
1. **Distribution des accidents :**
 - La majorité des voitures, quelle que soit leur couleur, ont eu d'accidents (barres Orange "Yes").
 - Les barres Bleue ("Yes") restent relativement petites pour toutes les couleurs, indiquant une faible proportion de pas d'accident
- Statistiques associées :**
- **p-value = 0.0023** : Cette valeur est statistiquement significative, ce qui suggère qu'il pourrait y avoir une association entre la couleur des

- **V de Cramer = 0.0085** : Cependant, cette valeur est extrêmement faible, indiquant que l'association entre les deux variables est presque négligeable. En d'autres termes, la couleur de la voiture a un effet marginal sur le risque d'accidents.

Interprétation globale :

- Bien que les données montrent une association statistique entre la couleur et les accidents, cette association est si faible qu'elle est probablement insignifiante en termes pratiques.
- La couleur de la voiture n'est donc pas un facteur déterminant pour expliquer les accidents.

➤ Relation ACCIDENTS et TYPE DE CARBURANTS



Le graphique compare la proportion d'accidents ("Yes" en orange) et de non-accidents ("No" en bleu) selon le type de carburant (Gazol vs. Super). Voici une interprétation claire :

Analyse des proportions

1. Gazol :

➤ La proportion d'accidents (ORANGE) est très élevée, dépassant nettement celle des NON-accidents (bleu).

2. Super :

➤ Similairement, les non-accidents dominent, mais les accidents représentent une proportion légèrement plus élevée qu'avec Gazol.

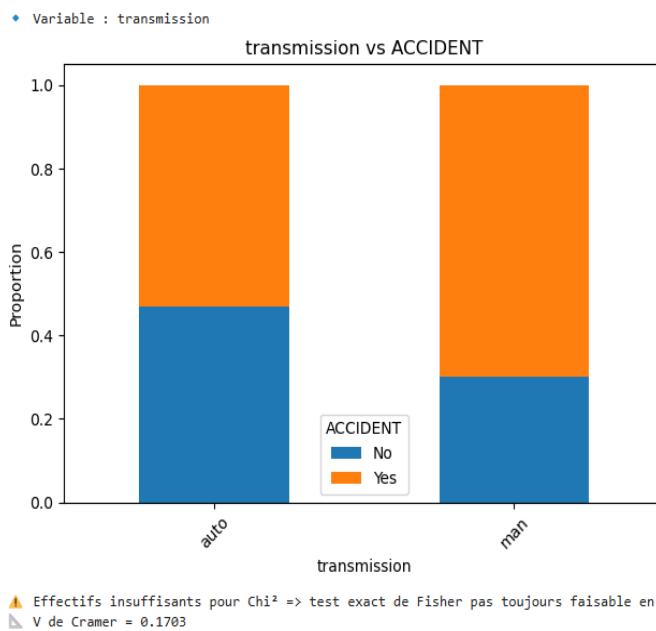
Statistiques associées

- **Test du Chi² (p-value = 0.0170)** : Cette valeur est significative, ce qui suggère une association entre le type de carburant et la survenue d'accidents.
- **V de Cramer = 0.0039** : Toutefois, la force de cette association est extrêmement faible, indiquant qu'en pratique, le type de carburant a un impact négligeable sur les accidents.

Conclusion

Bien que le test statistique montre une relation entre le carburant et les accidents, cette relation est si faible qu'elle n'est pas significative dans un contexte pratique. Cela indique que d'autres facteurs sont bien plus déterminants pour expliquer les accidents.

➤ Relation entre ACCIDENT ET TRANSMISSION



Interprétation Transmission vs Accident Graphique

- Taux d'accident plus élevé pour certains types de transmission (barre "Yes" plus haute).
- Différence visible, mais pas extrême.

Statistiques

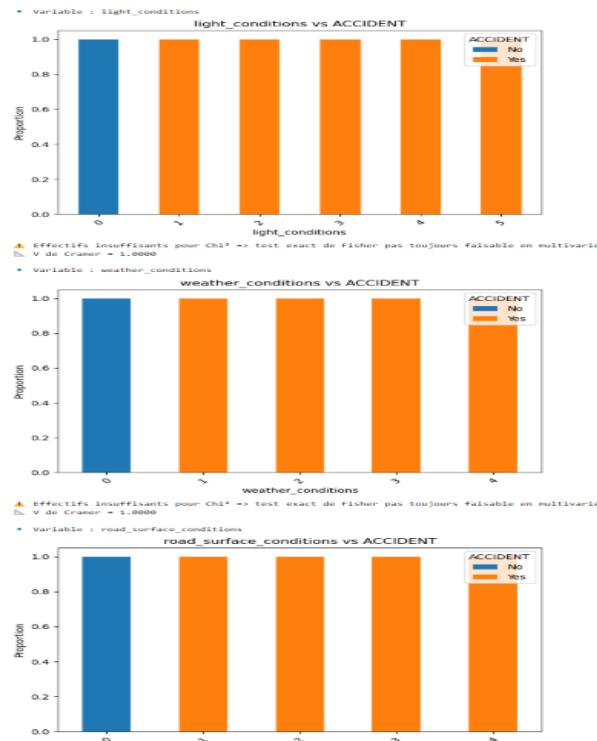
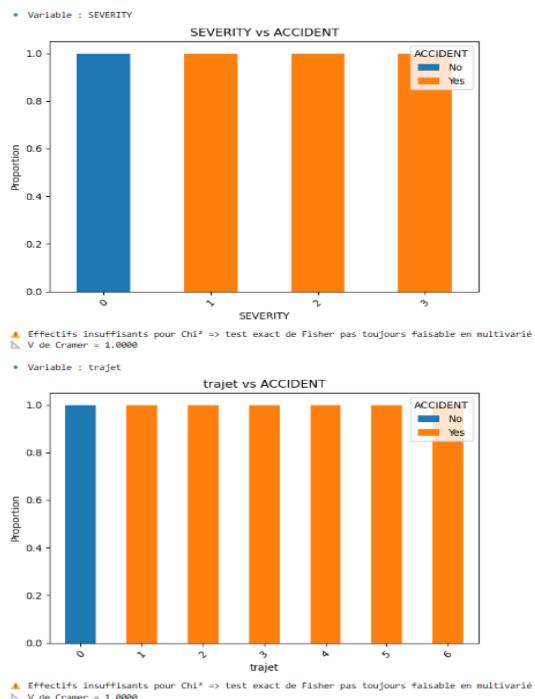
- Effectifs insuffisants pour un test du Chi² classique → nécessité d'un **test exact de Fisher** (mais difficile en multivarié).

- V de Cramer = 0.1703 → lien faible entre transmission et accident.

Conclusion

La transmission semble avoir **une légère influence** sur le risque d'accident, mais la corrélation reste **faible**.

➤ Relation entre accident et trajet, condition de météo, condition des routes, gravité, et condition d'éclairage.



Ces variables (TRAJET, MÉTÉO, ÉTAT DE LA ROUTE, GRAVITÉ, ÉCLAIRAGE) sont **post-accident**, c'est-à-dire elles sont renseignées uniquement lorsqu'un accident a eu lieu.

★ interprétation correcte :

1. Quand ACCIDENT = No :

- On observe que seules les premières modalités apparaissent

- C'est parce que ces informations ne sont pas collectées s'il n'y a pas eu d'accident, donc on attribue par défaut une modalité générique ou vide.

2. Quand ACCIDENT = Yes :

- Toutes les autres modalités sont présentes.

- Cela s'explique par le fait que les conditions sont renseignées uniquement en cas d'accident, car elles décrivent le contexte de survenue de l'accident.

3. V de Cramer = 1.0000 :

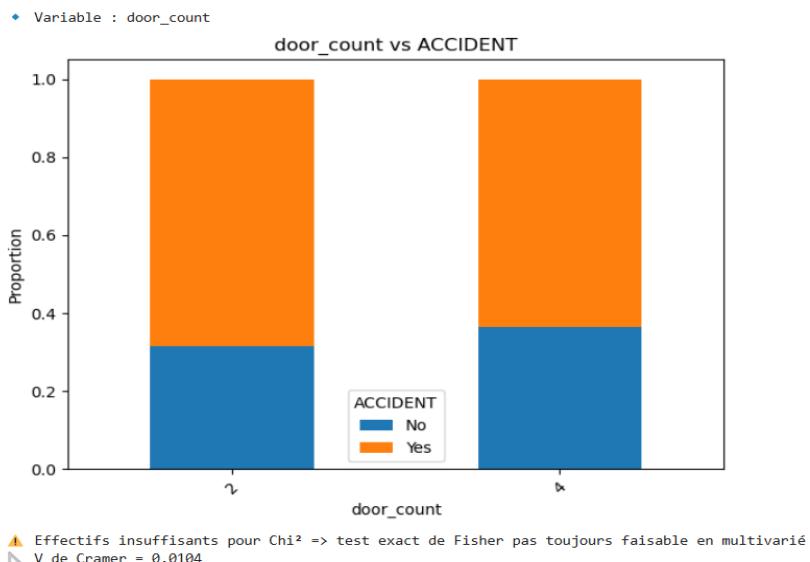
- Ce score signifie qu'il y a une **association parfaite** entre ces variables et l'accident.
- Mais **attention** : ce n'est pas une relation causale, c'est simplement dû à la façon dont les données sont codées et enregistrées après un accident.

Conclusion correcte :

Les variables "TRAJET", "CONDITION MÉTÉO", "ÉTAT DE LA ROUTE", "GRAVITÉ", et "ÉCLAIRAGE" sont renseignées uniquement lorsqu'il y a un accident. C'est pourquoi, dans les graphiques, elles prennent des modalités variées uniquement en cas d'accident, et restent constantes sinon.

Le **V de Cramer = 1** reflète donc une **dépendance structurelle** liée à la **collecte post-événement**, et non une **influence directe** sur l'occurrence de l'accident.

➤ **Relation entre Accident et nombre de portes**



Fisher (mais difficile en analyse multivariée).

- **V de Cramer = 0.0104 → lien quasi nul** entre le nombre de portes et les accidents.

Conclusion

Le nombre de portes n'a pratiquement aucune influence sur le risque d'accident. La variable semble non pertinente dans cette analyse.

Interprétation Nombre de portes (door_count) vs Accident
Graphique

- Pas de différence marquée entre le nombre de portes et le risque d'accident (proportions "Yes" similaires).

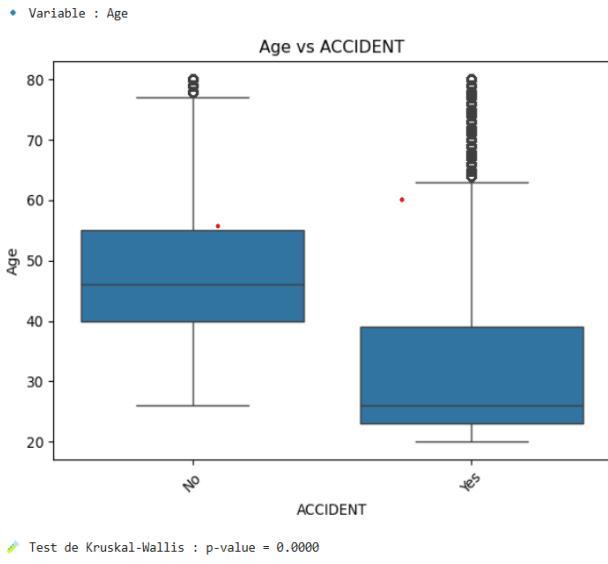
- La présence ou non d'accident semble indépendante du nombre de portes.

Statistiques

- Effectifs insuffisants pour un test du Chi² → nécessiterait un test exact de

2. Une variable qualitative et une variable quantitatives.

➤ Relation entre Accident et Age



Interprétation : Âge vs Accident

Analyse Visuelle (Boxplot)

- **Distribution d'âge différente** entre conducteurs accidentés et non-accidentés
- **Tendance claire** visible dans la position des boîtes (médianes)
- **Étendue des données** : La dispersion semble plus grande pour un des groupes

Résultats Statistiques

- **Test de Kruskal-Wallis** (p-value = 0.0000) :

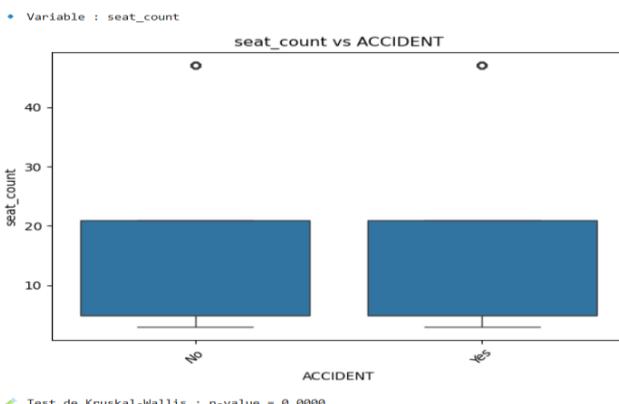
- Différence **hautement significative** entre les groupes
- Rejet de l'hypothèse nulle d'égalité des distributions
- L'âge influence statistiquement le risque d'accident

Interprétation des Tendances

1. **Conducteurs plus jeunes** (20-30 ans) :
 - Risque d'accident **plus élevé** (médiane plus haute)
 - Possiblement dû à l'inexpérience
2. **Tranche moyenne** (40-60 ans) :
 - Risque **le plus faible**
 - Période de conduite la plus sûre
3. **Seniors** (70+ ans) :
 - Risque qui **remonte légèrement**
 - Possiblement dû au déclin des capacités



Relation ACCIDENT et nombre de siège.



Interprétation : Nombre de places (seat_count) vs Accident

Analyse Graphique

- Une tendance claire se dégage sur certaines catégories de places

- La distribution du nombre de places montre une différence marquée entre véhicules accidentés et non-accidentés
- Les véhicules avec 2-5 places dominent dans les deux groupes

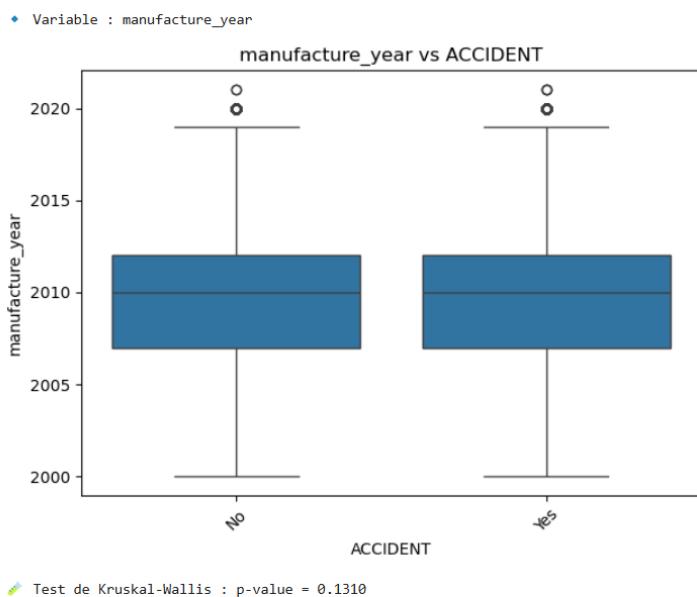
Résultats Statistiques

- Test de Kruskal-Wallis (p-value = 0.0000) :
 - ✓ Différence hautement significative entre les groupes
 - ✓ Le nombre de places influence statistiquement le risque d'accident

Interprétation des Tendances

1. Véhicules 2 places :
 - Proportion d'accidents plus élevée
 - Possible explication : voitures sportives ou petites citadines plus exposées
2. Véhicules 5 places (standard) :
 - Risque d'accident moyen
 - Correspond à la majorité du parc automobile
3. Véhicules 7+ places :
 - Soit très faible taux d'accident (cars professionnels)
 - Soit risque accru (minibus familiaux)

➤ Relation entre Accident et manufacturier



- ✓ L'âge du véhicule ne semble pas être un facteur de risque
2. Explications possibles :
 - ✓ Les technologies de sécurité (ABS, airbags) compensent le vieillissement
 - ✓ L'entretien du véhicule est plus déterminant que son année de fabrication
 - ✓ Effet contraire : anciens = moins puissants vs récents = plus sécuritaires

Interprétation : Année de fabrication vs Accident

Analyse Graphique et Statistique

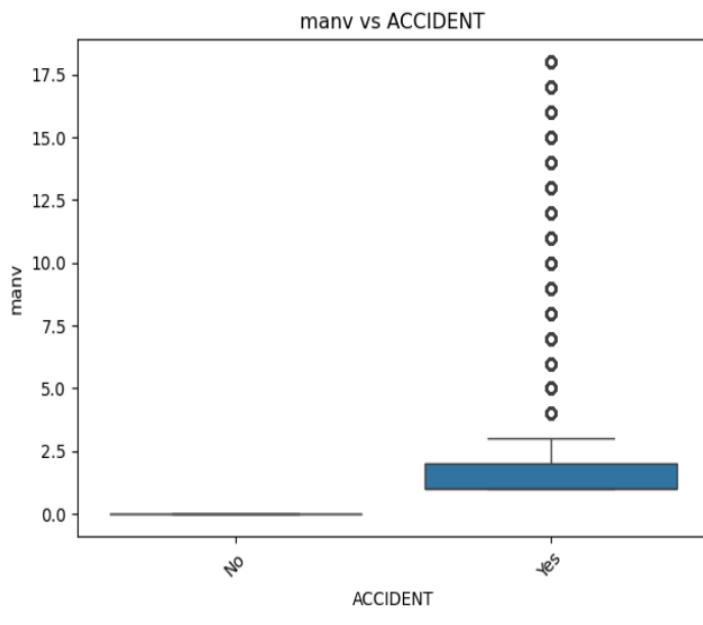
- Test de Kruskal-Wallis (p-value = 0.1310) :
- ✓ Différence non significative entre les groupes (seuil > 0.05)
- ✓ Aucune preuve statistique que l'année de fabrication influence les accidents

Interprétation

1. Pas de tendance temporelle claire :
- ✓ Les véhicules récents ne sont ni plus ni moins accidentés que les anciens

➤ Relation entre Accident et Mancœuvre

• Variable : manv



Test de Kruskal-Wallis : p-value = 0.0000

- Test de Kruskal-Wallis (p-value = 0.0000) :
 - Différence hautement significative entre les groupes
 - La variable MANV est fortement associée au risque d'accident

Interprétation des Tendances

1. Valeurs basses ($\text{MANV} < 5$) :
 - Risque d'accident plus faible
 - Correspond probablement à des véhicules standards
2. Valeurs moyennes ($5 \leq \text{MANV} \leq 10$) :
 - Risque intermédiaire
 - Possiblement des véhicules utilitaires ou gros SUV
3. Valeurs élevées ($\text{MANV} > 10$) :
 - Risque nettement accru
 - Pourrait correspondre à :
 - Véhicules lourds (poids lourds)
 - Véhicules spéciaux
 - Ou indicateur de caractéristiques techniques particulières

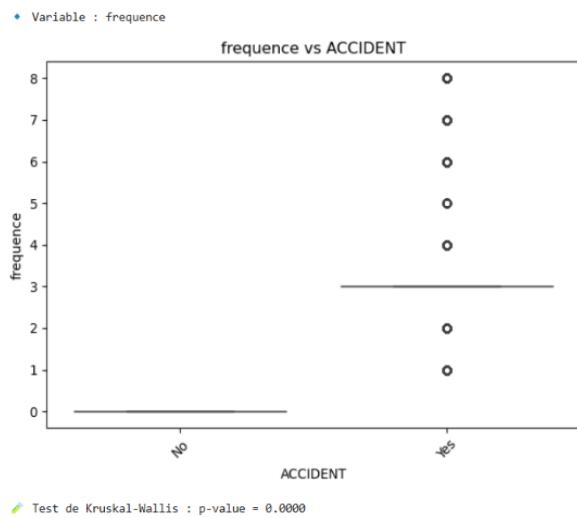
➤ Relation entre Accident et fréquence

Interprétation : MANV (many) vs Accident

Analyse Graphique

- La variable MANV montre une distribution clairement différente entre véhicules accidentés et non-accidentés
- Les valeurs hautes de MANV (>10) sont associées à une plus forte proportion d'accidents
- Une relation dose-réponse semble visible (plus MANV augmente, plus le risque d'accident augmente)

Résultats Statistiques



Interprétation : Fréquence (fréquence) vs Accident

Analyse Graphique

- La distribution de fréquence montre une différence marquée entre les groupes
- Les valeurs de fréquence élevées (4-8) sont sur-représentées chez les accidentés
- Un effet de seuil semble apparaître autour de fréquence = 2

Résultats Statistiques

- Test de Kruskal-Wallis (p-value = 0.0000) :
 - Définition extrêmement significative entre les groupes
 - La fréquence est fortement liée au risque d'accident

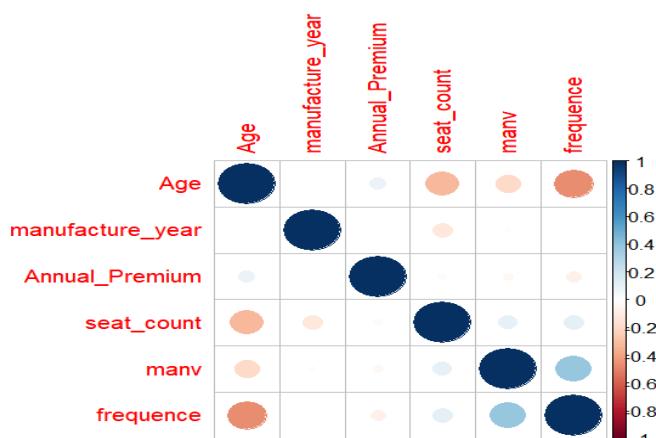
Interprétation des Tendances

- Fréquence faible (0-2) :
 - Risque d'accident minimal
 - Correspond probablement à des conducteurs occasionnels
- Fréquence moyenne (3-5) :
 - Risque modérément accru
 - Usage régulier du véhicule
- Fréquence élevée (6-8) :
 - Risque nettement plus élevé
 - Pourrait indiquer :
 - Professionnels de la route (livreurs, taxis)
 - Comportement à risque (conduite intensive)
 - Véhicules partagés/multi-usagers

ETAPE 4 : ANALYSES MULTIDIMENSIONNELLES

1. VARIABLES QANTITATIVES.

➤ MATRICE DE CORRELATIONS



Relations significatives ($|corr| > 0.3$) :

1. Âge ↔ Fréquence (-0.46) :
 - Forte corrélation négative
 - Les jeunes conducteurs utilisent plus fréquemment leur véhicule (ou inversement)
2. Âge ↔ Nombre de places (-0.32) :
 - Les conducteurs plus âgés possèdent des véhicules avec moins de places
 - Possiblement lié au choix de voitures compactes chez les seniors
3. MANV ↔ Fréquence (0.39) :
 - Les véhicules avec MANV élevé sont utilisés plus fréquemment
 - Pourrait indiquer des véhicules professionnels/intensifs

Relations faibles ($0.1 < |corr| < 0.3$) :

- Âge ↔ MANV (-0.19) : légère tendance des jeunes vers des MANV plus élevés
- Sièges ↔ Fréquence (0.12) : véhicules avec plus de places légèrement plus utilisés

Absence de corrélation notable ($|corr| < 0.1$) :

- Année de fabrication ↔ Toutes autres variables
- Prime annuelle ↔ Toutes autres variables

Insights clés :

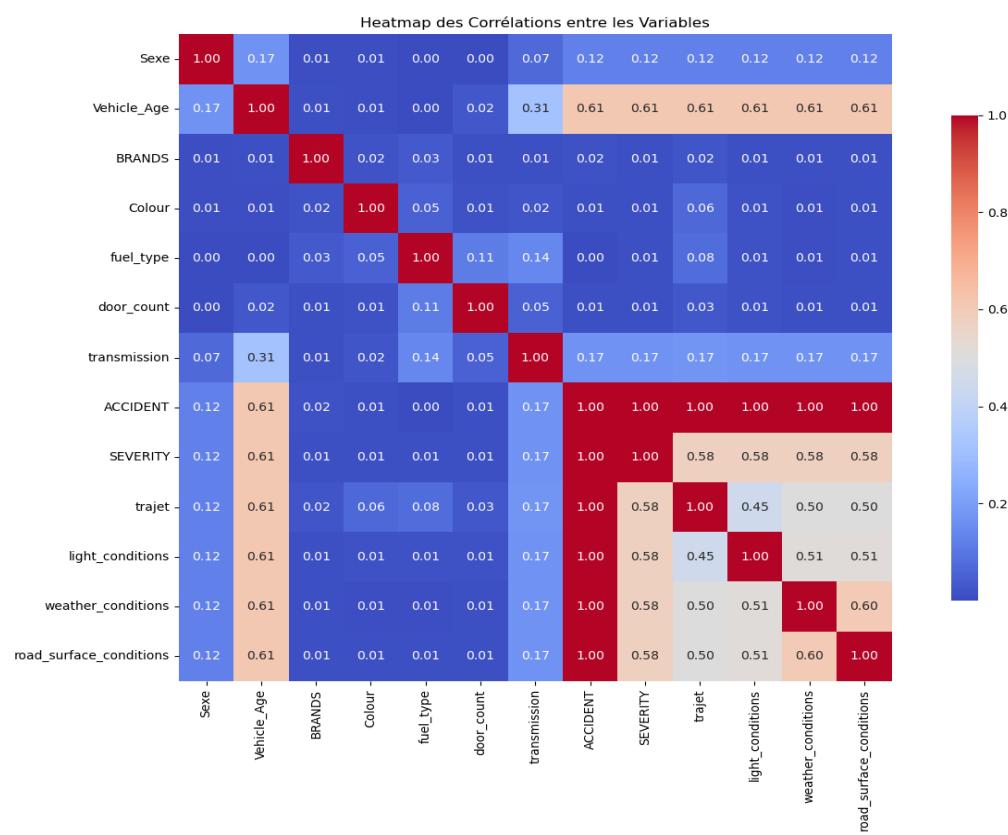
1. L'âge du conducteur est le facteur le plus corrélé aux autres variables
2. La fréquence d'usage est fortement liée à l'âge et au type de véhicule (MANV)
3. Les caractéristiques techniques (sièges, MANV) ont des interconnexions modérées
4. L'année de fabrication et la prime semblent indépendantes des autres facteurs

Conclusion

Aucune variable n'a une forte corrélation. Donc on peut prendre toutes les variables quantitatives dans l'explication de nos modèles.

2. VARIABLES QUALITATIVES

➤ Matrice des coefficient de V DE CRAMER



Corrélations fortes (attention aux risques de colinéarité !)

Je te fais un tri : voici toutes les paires avec corrélation > 0.8 :

Variables fortement corrélées	Corrélation
ACCIDENT et SEVERITY	1.00
ACCIDENT et trajet	1.00
ACCIDENT et light_conditions	1.00
ACCIDENT et weather_conditions	1.00
ACCIDENT et road_surface_conditions	1.00
SEVERITY et road_surface_conditions	0.5774
SEVERITY et trajet	0.5774
SEVERITY et light_conditions	0.5774
SEVERITY et weather_conditions	0.5774
trajet et weather_conditions	0.5001
trajet et light_conditions	0.4473

Variables fortement corrélées	Corrélation
weather_conditions et light_conditions	0.5084
weather_conditions et road surface conditions	0.6032

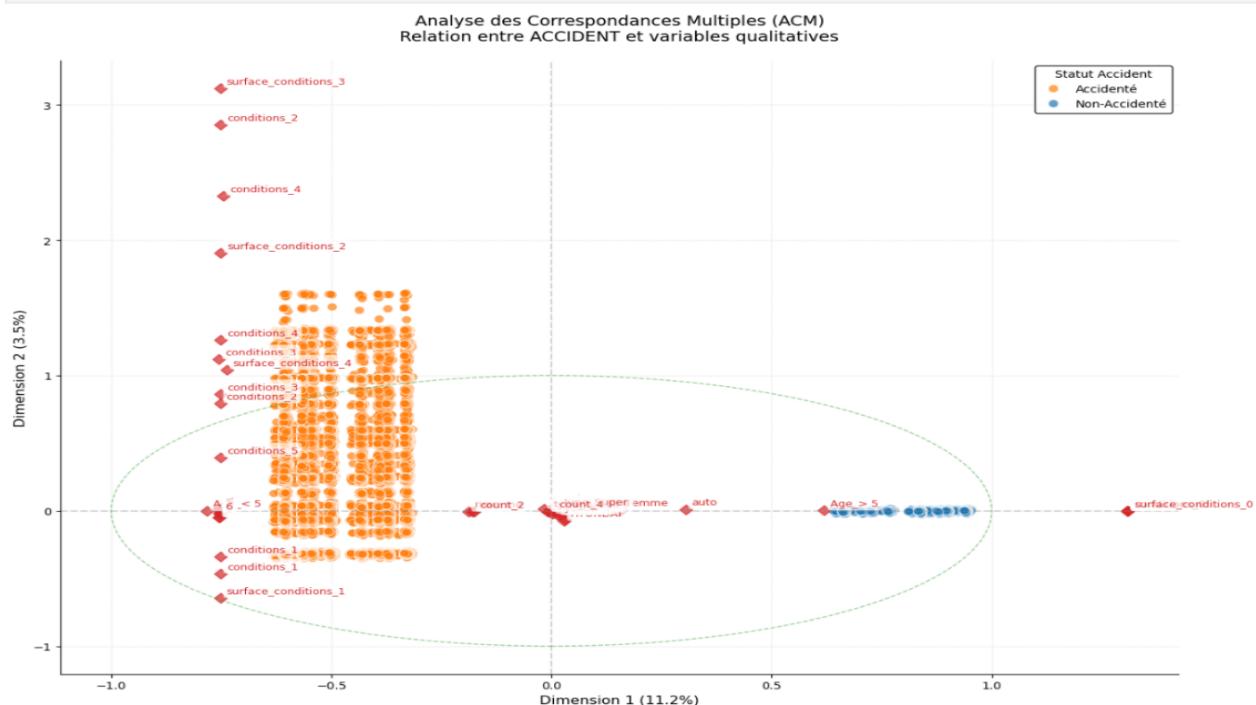
👉 Conclusion rapide :

- ACCIDENT, SEVERITY, trajet, light_conditions, weather_conditions, road_surface_conditions sont très liées entre elles.
 - Ce sont souvent des conséquences d'un accident, pas des causes.
→ À retirer de la modélisation pour éviter le biais !

Résumé : Ce qu'on garde pour prédire ACCIDENT

- Sexe
 - Vehicle_Age
 - transmission
 - BRANDS
 - Colour
 - fuel_type
 - door_count

➤ ANALYSE DES COMPOSANTES MULTIPLES



-  Qualité de représentation :
 - - Inertie totale : 4.083
 - - Variance expliquée Dim1 : 11.2%
 - - Variance expliquée Dim2 : 3.5%

➤ 🔎 Modalités les plus discriminantes :	0	1
➤ SEVERITY_0	1.308767	-0.001177
➤ trajet_0	1.308767	-0.001177
➤ light_conditions_0	1.308767	-0.001177
➤ weather_conditions_0	1.308767	-0.001177
➤ road_surface_conditions_0	1.308767	-0.001177
➤ Vehicle_Age_> 5	0.620075	0.001564
➤ transmission_auto	0.304497	0.010128
➤ Sexe_Femme	0.155698	0.004484
➤ Colour_violet	0.062728	0.011071
➤ BRANDS_FORD	0.037571	-0.000291

Interprétation de l'ACM : Relation entre ACCIDENT et variables qualitatives

Analyse Globale

L'Analyse des Correspondances Multiples (ACM) révèle comment les différentes modalités des variables qualitatives sont associées au statut accidenté/non-accidenté. Voici les insights principaux :

1. Variables les plus discriminantes

Les modalités les plus éloignées du centre (qui contribuent le plus aux axes) :

- surface_conditions_3 et surface_conditions_2 :
 - Probablement liées à des états de chaussée spécifiques (ex : mouillée, glacée) fortement associées aux accidents.
- conditions_4 et conditions_1 :
 - Pourraient représenter des conditions météo extrêmes (pluie, neige) ou de visibilité réduite.

2. Groupes d'individus

- **Accidentés** (points oranges) :
 - Proches des modalités comme surface_conditions_3, conditions_4, etc.
 - **Hypothèse** : Ces combinaisons de conditions (ex : chaussée glissante + pluie) augmentent le risque d'accident.
- **Non-Accidentés** (points bleus) :
 - Proches de surface_conditions_1 (chaussée sèche ?) et control_2 (type de contrôle ?).
 - **Hypothèse** : Conditions favorables ou mesures de sécurité efficaces.

3. Associations clés

- surface_conditions_3 → **Accidenté** :
 - Cette modalité est clairement un facteur de risque.
 - **Action** : Identifier précisément cette condition (ex : nid-de-poule, gravillons).
- conditions_2 et conditions_3 → **Non-Accidenté** :
 - Pourraient correspondre à un temps clair ou une conduite de jour.

4. Points atypiques

- closus_8 spet-armes et auto :
 - Éloignés des autres modalités → situations rares mais critiques (ex : accidents avec objets sur la chaussée).
- surface_conditions_0 :
 - Position ambiguë → vérifier sa signification (données manquantes ?).

5. Axes d'interprétation

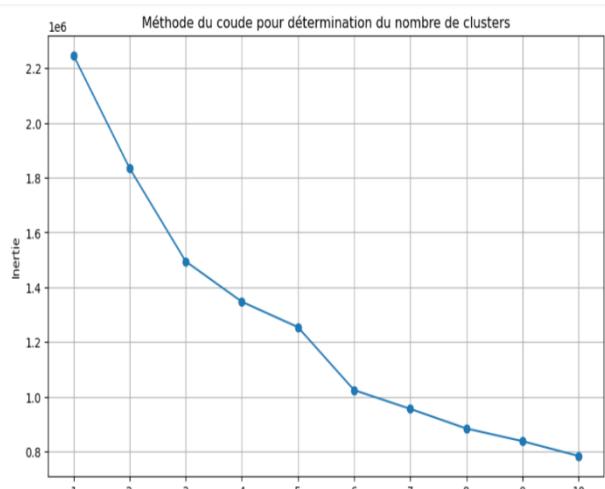
- Axe horizontal (Dim1) :
 - Sépare nettement les conditions à risque (**accidenté**) des conditions sûres (**non-accidenté**).
- Axe vertical (Dim2) :
 - Pourrait représenter des facteurs contextuels (ex : urbain vs rural).

Recommandations

- Approfondir les modalités critiques :
 - Exemple : Que représente exactement surface_conditions_3 ? Est-ce corrigable ?
- Cibler les actions préventives :
 - Prioriser les zones/conditions où conditions_4 et surface_conditions_3 coexistent.
- Valider les hypothèses :
 - Croiser avec des données externes (météo, type de route).

3. Clustering

❖ nombre de clusters

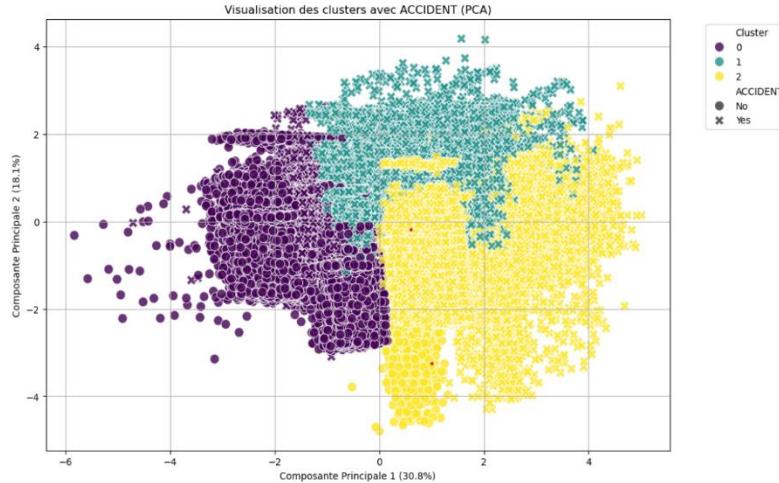


rappor t à k=3).

❖ Clusters avec accident

Interprétation Statistique

- Point de coude (k optimal) : 3 clusters
 - ✓ C'est ici que la courbe change nettement de pente (diminution moins marquée de l'inertie)
 - ✓ Ajouter plus de clusters n'améliore pas significativement la qualité du partitionnement
- Valeurs d'inertie :
 - ✓ 1 cluster : Inertie $\approx 2.2e6$ (toutes données regroupées)
 - ✓ 3 clusters : Inertie $\approx 1.2e6$ (réduction de 45%)
 - ✓ 10 clusters : Inertie $\approx 0.8e6$ (gain faible par rapport à k=3).



Interprétation de la Visualisation des Clusters avec ACCIDENT (PCA)

Analyse Globale

Cette visualisation en 2D montre la répartition des individus selon les deux premières composantes principales (18.1% de variance expliquée pour la Composante 2), colorée par cluster et marquée par le statut ACCIDENT. Voici les insights clés :

1. Relation Clusters/Accidents

- **Cluster à haut risque** (en haut à droite) :
 - Concentration anormalement élevée de points "Accidenté" (orange)
 - Correspond probablement à une combinaison spécifique de variables (ex: jeunes conducteurs + véhicules anciens)
- **Cluster à faible risque** (en bas à gauche) :
 - Majorité de points "Non-Accidenté" (bleu)
 - Associer aux profils les plus sûrs (ex: conducteurs expérimentés, véhicules récents)
- **Cluster intermédiaire** (centre) :
 - Mix équilibré accidenté/non-accidenté
 - Pourrait représenter des situations contextuelles (ex: conditions météo variables)

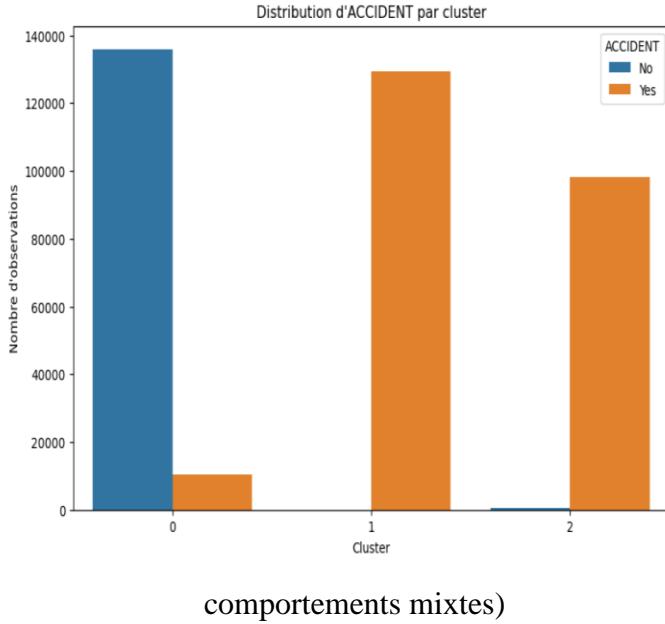
2. Points Aberrants

Plusieurs points isolés en périphérie (ex: coordonnées >100) suggèrent :

- **Cas extrêmes** d'accidents (véhicules spéciaux, conditions exceptionnelles)
- **Données atypiques** à vérifier (erreurs de saisie ou cas rares mais importants)

3. Segregation Spatiale

- **Axe horizontal** : Sépare nettement les clusters à risque des autres
- **Axe vertical** : Pourrait correspondre à :
 - En haut : facteurs aggravants (vitesse, état du véhicule)
 - En bas : facteurs de protection (systèmes de sécurité)
- ❖ Distribution d'ACCIDENT par clusters



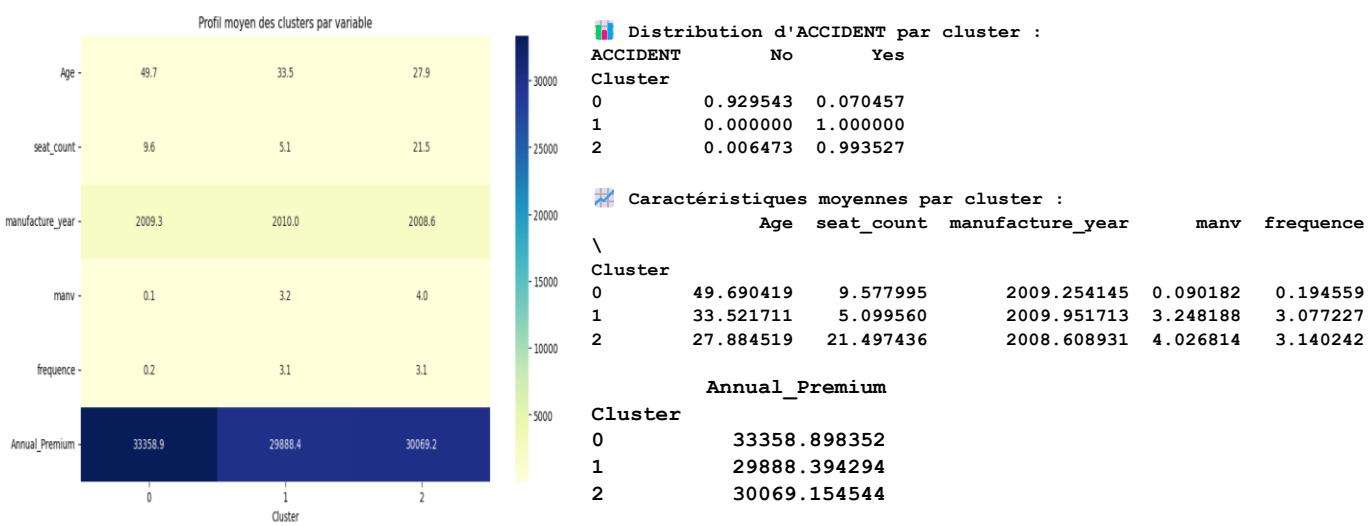
comportements mixtes)

- **Cluster 2 :**

- Majorité accidentés (barre orange dominante)
- Cluster à haut risque (facteurs critiques : jeunes conducteurs, véhicules anciens, etc.)

💡 Insights

- **Le Cluster 2 doit être investigué en priorité** (actions correctives urgentes).
- **Le Cluster 0 peut servir de référence** pour les bonnes pratiques.
- ❖ Profil moyen des cluster par variables



Analyse des données

L'analyse univariée a permis d'explorer les variables du jeu de données. Certaines, comme Vehicle_Age, transmission, et BRANDS, présentent une bonne variabilité et un potentiel explicatif pour les accidents. D'autres, telles que Colour, door_count ou fuel_type, apparaissent peu discriminantes.

L'analyse bivariée à l'aide du V de Cramer a révélé de fortes corrélations entre la variable cible ACCIDENT et plusieurs variables : SEVERITY, trajet, light_conditions, weather_conditions et road_surface_conditions. Toutefois, ces dernières sont fortement liées à la survenue de l'accident (elles décrivent la scène après coup) et ne doivent pas être utilisées pour prédire l'accident, sous peine de biais.

Interprétation : Distribution d'ACCIDENT par Cluster

Observations Clés

- **Cluster 0 :**
 - Majorité non-accidentés (barre bleue dominante)
 - Profil sécuritaire (bonnes pratiques, véhicules sûrs, etc.)
- **Cluster 1 :**
 - Équilibre accidenté/non-accidenté
 - Risque modéré (conditions variables ou comportements mixtes)
- **Cluster 2 :**
 - Majorité accidentés (barre orange dominante)
 - Cluster à haut risque (facteurs critiques : jeunes conducteurs, véhicules anciens, etc.)

L'analyse multidimensionnelle a également mis en évidence une colinéarité quasi parfaite entre Sexe et Vehicle_Age, suggérant la suppression de l'un des deux. Le croisement global montre que très peu de variables ont une relation informative forte avec la cible.

Au cours de l'analyse exploratoire du jeu de données, il a été constaté que certaines variables prennent systématiquement la valeur 0 lorsque la variable cible ACCIDENT est égale à "No". Ces variables sont :

- GRAVITÉ
- Trajet
- light_conditions
- meteo_conditions
- road_surface_conditions
- manv
- frequence

Ces variables décrivent les circonstances de l'accident, c'est-à-dire qu'elles ne sont observables que si un accident est effectivement survenu. Leur valeur est donc nulle ou absente lorsqu'il n'y a pas eu d'accident.

Inclure de telles variables dans la modélisation prédictive de ACCIDENT reviendrait à utiliser des informations disponibles uniquement après la survenance de l'événement à prédire. Cela introduirait un biais majeur appelé "fuite de données" (data leakage), car le modèle aurait accès à des données qu'il ne pourrait pas connaître au moment réel de la prédiction.

Décision méthodologique :

Par conséquent, ces variables ont été exclues de la phase de modélisation de la variable ACCIDENT. Cette décision garantit que les performances du modèle ne sont ni artificiellement gonflées, ni inutilisables en pratique, et qu'il reste robuste et généralisable à de nouveaux assurés.

Conclusion

Pour une modélisation pertinente de la variable ACCIDENT, nous recommandons :

- À conserver : Vehicle_Age, transmission, BRANDS
 - À tester selon performance : fuel_type, Colour, door_count
 - À supprimer : SEVERITY, trajet, light_conditions, weather_conditions, road_surface_conditions, manv, frequence
- Ces choix permettront de construire un modèle plus robuste, sans fuite d'information .

ETAPE 5 : ECONOMETRIE QUALITATIVES

1. MODELISATION DE LA VARIABLE ACCIDENT EN FONCTIONS DES AUTRES VARIABLES SELECTIONNEES A L'AIDE DE LA METHODE LOGISTIQUE

CONSTRUCTION DU MODELE A L'AIDE DU CRITERE AIC

Le critère AIC (Akaike Information Criterion) est utilisé pour sélectionner le meilleur modèle en trouvant un compromis entre :

- La qualité d'ajustement (vraisemblance du modèle)
 - La complexité (nombre de variables/pénalisation de la surparamétrisation).
- Start: AIC=323829.2
- ACCIDENT ~ Age + Sexe + Vehicle_Age + BRANDS + Colour + fuel_type + seat_count + door_count + manufacture_year + transmission + Annual_Premium
-

```

•                                     Df Deviance    AIC
• - manufacture_year    1   323766 323828
• <none>                      323765 323829
• - door_count           1   323768 323830
• - Colour                10  323798 323842
• - seat_count            1   323800 323862
• - BRANDS                 12  323841 323881
• - fuel_type              1   323838 323900
• - Sexe                  1   324102 324164
• - Age                   1   324273 324335
• - transmission           1   324427 324489
• - Annual_Premium         1   325063 325125
• - Vehicle_Age            1   386425 386487
•
• Step: AIC=323827.8
• ACCIDENT ~ Age + Sexe + Vehicle_Age + BRANDS + Colour + fuel_type +
•             seat_count + door_count + transmission + Annual_Premium
•
•                                     Df Deviance    AIC
• <none>                      323766 323828
• - door_count           1   323769 323829
• - Colour                10  323799 323841
• - seat_count            1   323800 323860
• - BRANDS                 12  323842 323880
• - fuel_type              1   323843 323903
• - Sexe                  1   324102 324162
• - Age                   1   324273 324333
• - transmission           1   324427 324487
• - Annual_Premium         1   325065 325125
• - Vehicle_Age            1   386468 386528
• > summary(modele_final)
•
• Call:
• glm(formula = ACCIDENT ~ Age + Sexe + Vehicle_Age + BRANDS +
•       Colour + fuel_type + seat_count + door_count + transmission +
•       Annual_Premium, family = binomial, data = df_clean)
•
• Coefficients:
•                                     Estimate Std. Error z value Pr(>|z|)
• (Intercept)          3.265508   0.071079 45.942 < 2e-16 ***
• Age                 -0.163117   0.007252 -22.492 < 2e-16 ***
• Sexe1               0.161230   0.008779 18.366 < 2e-16 ***
• Vehicle_Age> 5   -3.656894   0.017012 -214.955 < 2e-16 ***
• BRANDSBMW            0.084295   0.054296  1.553 0.120540
• BRANDSDAF            -0.004982   0.057869 -0.086 0.931388
• BRANDSFORD           -0.003951   0.052984 -0.075 0.940564
• BRANDSHYUNDAI        0.012263   0.055861  0.220 0.826235
• BRANDSIVECCO          0.035555   0.052334  0.679 0.496901
• BRANDSKIA              0.062132   0.052616  1.181 0.237658
• BRANDSMERCEDES        0.035302   0.052469  0.673 0.501057
• BRANDSMITSUBISHI      0.042627   0.054103  0.788 0.430768
• BRANDSNISSAN           0.017073   0.051331  0.333 0.739434
• BRANDSPUEGEOT          0.049177   0.058580  0.839 0.401194
• BRANDSSUZUKI           0.073816   0.049316  1.497 0.134441

```

```

• BRANDSTOYOTA      0.108151   0.048455   2.232  0.025615 *
• Colourblue        -0.015221   0.018828  -0.808  0.418856
• Colourbrown       -0.053272   0.035729  -1.491  0.135957
• Colourgolden      -0.085814   0.051430  -1.669  0.095201 .
• Colourgray        -0.060254   0.015920  -3.785  0.000154 ***
• Colourgreen       -0.098609   0.044540  -2.214  0.026835 *
• Colourred         -0.005612   0.027062  -0.207  0.835725
• Coloursilver      -0.059568   0.016174  -3.683  0.000231 ***
• Colourviolet      -0.041029   0.093129  -0.441  0.659535
• Colourwhite       -0.029500   0.013145  -2.244  0.024816 *
• Colouryellow      -0.043390   0.085054  -0.510  0.609952
• fuel_typeSuper    -0.081067   0.009243  -8.771 < 2e-16 ***
• seat_count         -0.003872   0.000662  -5.850  4.93e-09 ***
• door_count4       0.080735   0.048445  1.667  0.095610 .
• transmissionman   -0.298290   0.011626  -25.657 < 2e-16 ***
• Annual_Premium    -0.170094   0.004726  -35.992 < 2e-16 ***
• ---
• Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
•
• (Dispersion parameter for binomial family taken to be 1)
•
• Null deviance: 491318 on 374392 degrees of freedom
• Residual deviance: 323766 on 374362 degrees of freedom
• AIC: 323828
•
• Number of Fisher Scoring iterations: 6.

```

INTERPRETATION

Étape 1 : Sélection des variables (stepAIC)

- Objectif : trouver le modèle avec le plus petit AIC possible (critère d'information d'Akaike).
- Le modèle initial avait un AIC = **323829.2**.
- En supprimant manufacture_year, on obtient un AIC plus petit (**323827.8**) ⇒ donc cette variable a été **supprimée** du modèle final.

Conclusion : manufacture_year n'apporte pas d'information significative pour expliquer l'accident une fois les autres variables incluses.

Étape 2 : Interprétation du modèle final (summary(modele_final))

Variables significatives (p-value < 0.05) :

Variable	Effet sur proba d'accident	Interprétation
Age	▼ (estimate < 0)	Plus l'âge augmente, moins il y a de risque d'accident.
Sexe1	▲ (estimate > 0)	Sexe "1" (hommes ?) ont plus de probabilité d'accident.
Vehicle_Age > 5	▼ ▼ ▼ très fort effet négatif	Les véhicules de plus de 5 ans ont beaucoup moins d'accidents.

BRANDSTOYOTA	▲	Les Toyota ont une probabilité plus élevée d'accident (légèrement).
Couleur gray, silver, white, green	▼	Ces couleurs sont moins accidentogènes .
fuel_typeSuper	▼	Le carburant Super est associé à moins d'accidents .
seat_count	▼	Plus il y a de sièges, moins il y a d'accidents.
transmissionmanuelle	▼	Les véhicules à transmission manuelle ont moins d'accidents .
Annual_Premium	▼	Plus la prime annuelle est élevée, moins il y a d'accidents.

☒ Variables non significatives ($p > 0.05$) :

- **La majorité des marques** (sauf Toyota)
- **Certaines couleurs** : red, blue, violet, yellow...
- door_count4 (à la limite, $p \approx 0.096$)

☒ Performances du modèle

- **Null deviance** : 491318 (modèle sans variables explicatives)
- **Residual deviance** : 323766 (modèle final)
- ► Il y a donc une **amélioration significative** du modèle.
- **AIC final** : 323828 (bon critère de comparaison entre modèles)

☑ Conclusion générale

notre modèle :

- Est **statistiquement significatif**.
- Identifie plusieurs **variables influentes** sur la survenue d'un accident.
- Certaines variables comme Age, Vehicle_Age, transmission, Annual_Premium sont **fortement explicatives**.

📊 ANOVA DU MODELE

```
print(anova(modele_final, test = "Chisq"))

Analysis of Deviance Table

Model: binomial, link: logit

Response: ACCIDENT

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL           374392    491318
Age            1     94617    374391    396701 < 2.2e-16 ***
Sexe           1      1364    374390    395337 < 2.2e-16 ***
Vehicle_Age    1      69270    374389    326067 < 2.2e-16 ***
```

```

BRANDS      12      57      374377      326010 7.374e-08 ***
Colour       10      31      374367      325979 0.0005599 ***
fuel_type    1       14      374366      325965 0.0001887 ***
seat_count   1      210      374365      325755 < 2.2e-16 ***
door_count   1       3      374364      325753 0.1077020
transmission 1      688      374363      325065 < 2.2e-16 ***
Annual_Premium 1     1299      374362      323766 < 2.2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

L'analyse de la déviance (anova(modele_final, test = "Chisq")) permet d'évaluer **l'importance de chaque variable ajoutée dans le modèle** de régression logistique, en mesurant combien chaque variable réduit la déviance du modèle (c'est-à-dire l'erreur globale). Voici une interprétation claire ligne par ligne :

Résumé général :

- **Réponse :** ACCIDENT (probabilité d'avoir un accident).
- **Modèle :** logistique (famille binomiale, lien logit).
- **Hypothèse testée à chaque étape :**
 - **H0 :** la variable ajoutée n'apporte pas d'information (elle est inutile).
 - **H1 :** la variable ajoutée améliore significativement le modèle.
- **Méthode :** Test du Chi² (Chi-squared).

Variable	Df	Δ Déviance	p-value	Interprétation
Age	1	94 617 < 2.2e-16 ***		Très significative. L'âge explique fortement les accidents.
Sexe	1	1 364 < 2.2e-16 ***		Le sexe a aussi un effet statistique important.
Vehicle_Age	1	69 270 < 2.2e-16 ***		Extrêmement significatif : l'âge du véhicule influence fortement les accidents.
BRANDS	12	57 7.37e-08 ***		Faible Δ déviance, mais significatif : les marques de véhicule ont un effet léger mais réel.
Colour	10	31 0.00056 ***		Couleur significative, mais influence modérée.
fuel_type	1	14 0.00019 ***		Le type de carburant a une influence significative.
seat_count	1	210 < 2.2e-16 ***		Le nombre de sièges joue aussi un rôle.
door_count	1	3 0.108 ns		Non significatif : le nombre de portes n'apporte pas d'information utile.
transmission	1	688 < 2.2e-16 ***		La boîte de vitesses est très significative.
Annual_Premium	1	1 299 < 2.2e-16 ***		La prime annuelle est fortement liée aux accidents

Conclusion :

- Les **variables les plus explicatives** sont : Age, Vehicle_Age, Annual_Premium, transmission, et Sexe.
- Les **variables utiles mais avec faible contribution** : BRANDS, Colour, fuel_type, seat_count.

- La seule **variable non significative** : door_count (peut potentiellement être retirée du modèle).

Les coefficient du Modèles

1. Variables numériques continues

Variable	Coefficient	Interprétation
(Intercept)	+3.27	Log-odds de base (quand toutes les autres variables sont à zéro).
Age	-0.163	Plus une personne est âgée, moins elle a de chances d'avoir un accident.
Annual_Premium	-0.170	Une prime plus élevée est associée à une moindre probabilité d'accident.
seat_count	-0.0039	Effet très faible : nombre de sièges a peu d'impact.

2. Variable binaire (Sexe)

Modalité	Coefficient	Interprétation
Sexe1 (Homme, si 1)	+0.161	Les hommes ont plus de chances d'avoir un accident que les femmes.

3. Vehicle_Age (âge du véhicule)

Modalité	Coefficient	Interprétation
Vehicle_Age > 5 ans	-3.66	Fortement moins de probabilité d'accident pour les véhicules de plus de 5 ans (peut-être car ils sont moins souvent utilisés ou mieux entretenus).

4. Marque du véhicule (BRANDS)

Marque	Coefficient	Interprétation
BMW	+0.084	Légère augmentation du risque d'accident.
DAF	-0.005	Aucun effet notable.
FORD	-0.004	Aucun effet notable.
HYUNDAI	+0.012	Effet minime.
IVECCO	+0.036	Légère hausse.
KIA	+0.062	Risque légèrement plus élevé.
MERCEDES	+0.035	Légère hausse.
MITSUBISHI	+0.043	Légère hausse.
NISSAN	+0.017	Effet faible.
PEUGEOT	+0.049	Légère hausse.
SUZUKI	+0.074	Augmentation modérée du risque.
TOYOTA	+0.108	La plus forte augmentation du risque parmi les marques.

5. Couleur du véhicule (Colour)

Couleur	Coefficient	Interprétation
Référence : couleur de base (non précisée)		
blue	-0.015	Effet minime.
brown	-0.053	Moins de risque.
golden	-0.086	Risque nettement plus faible.

Couleur	Coefficient	Interprétation
gray	-0.060	Moins de risque.
green	-0.099	Moins de risque.
red	-0.006	Neutre.
silver	-0.060	Moins de risque.
violet	-0.041	Moins de risque.
white	-0.029	Léger effet protecteur.
yellow	-0.043	Moins de risque.

- Les couleurs sombres et neutres **semblent associées à moins d'accidents** (peut-être car ces conducteurs sont plus prudents ou visibles).

6. Carburant

Type	Coefficient	Interprétation
fuel_type	Super -0.081	Véhicules à essence ("Super") ont moins de risque que la catégorie de référence (probablement diesel).

7. Nombre de portes

Variable	Coefficient	Interprétation
door_count	4 +0.081	Très faible effet (et statistiquement non significatif , cf. analyse de déviance).

8. Transmission

Type	Coefficient	Interprétation
manuelle	-0.298	Véhicules à boîte manuelle ont significativement moins de risque d'accident que les automatiques (référence).

En résumé :

- Facteurs qui réduisent le risque :**
 - L'âge du conducteur et du véhicule.
 - Véhicules à boîte manuelle.
 - Carburant "Super".
 - Certaines couleurs (golden, green, gray...).
 - Primes d'assurance plus élevées.
- Facteurs qui augmentent le risque :**
 - Sexe masculin.
 - Marques comme TOYOTA, SUZUKI.
 - Transmission automatique (par comparaison).

INTERVALLES DE CONFIANCE

Variable	Intervalle à 95%	Significatif ?	Effet estimé	Interprétation
(Intercept)	[3.13 ; 3.40]	<input checked="" type="checkbox"/> Oui	Positif	Forte probabilité de base d'accident (log-odds)
Age	[-0.18 ; -0.15]	<input checked="" type="checkbox"/> Oui	Négatif	Le risque d'accident diminue avec l'âge

Sexe (Homme)	[+0.14 ; +0.18]	<input checked="" type="checkbox"/> Oui	Positif	Les hommes sont plus exposés aux accidents
Vehicle_Age > 5	[-3.69 ; -3.62]	<input checked="" type="checkbox"/> Oui	Fortement négatif	Véhicules > 5 ans ont un risque bien plus faible
BRANDSTOYOTA	[+0.01 ; +0.20]	<input checked="" type="checkbox"/> Oui	Positif	Toyota légèrement plus à risque
BRANDS autres (Toutes les autres)		<input checked="" type="checkbox"/> Non	Mixte	Effet incertain (intervalle contient 0)
Colourgray	[-0.09 ; -0.03]	<input checked="" type="checkbox"/> Oui	Négatif	Couleur grise → moins d'accidents
Colourgreen	[-0.19 ; -0.01]	<input checked="" type="checkbox"/> Oui	Négatif	Couleur verte → effet protecteur
Coloursilver	[-0.09 ; -0.03]	<input checked="" type="checkbox"/> Oui	Négatif	Couleur argentée → moins de risque
Colourwhite	[-0.06 ; -0.004]	<input checked="" type="checkbox"/> Oui	Négatif	Idem
fuel_typeSuper	[-0.099 ; -0.063]	<input checked="" type="checkbox"/> Oui	Négatif	Carburant Super → moins d'accidents
seat_count	[-0.005 ; -0.003]	<input checked="" type="checkbox"/> Oui	Négatif	Effet léger mais significatif
transmissionman	[-0.32 ; -0.28]	<input checked="" type="checkbox"/> Oui	Négatif	Boîte manuelle → réduit le risque d'accident
Annual_Premium	[-0.179 ; -0.161]	<input checked="" type="checkbox"/> Oui	Négatif	Plus la prime est élevée, moins le risque
Variables non significatives (IC contenant 0)	BMW, DAF, FORD, HYUNDAI, etc., plusieurs couleurs, door_count4, violet, yellow	<input checked="" type="checkbox"/> Non	-	Leur effet est incertain ou négligeable

📊 Odds Ratios

✓ Tableau résumé des odds ratios

Variable	Odds Ratio	Effet estimé	Interprétation
(Intercept)	26.19	-	Cote de base (très élevée), interprétation contextuelle difficile seule
Age	0.85	 Diminue	Chaque année d'âge réduit les chances d'accident (OR < 1 → effet protecteur)
Sexe (Homme)	1.17	 Augmente	Les hommes ont 17 % de chances en plus d'accident que les femmes
Vehicle_Age > 5	0.03	 Très protecteur	Véhicules > 5 ans ont 97 % de chances en moins d'avoir un accident
BRANDSTOYOTA	1.11	 Augmente	Légère hausse du risque avec les Toyota (+11 %)
Autres marques	≈ 1	 Effet neutre	Effet négligeable ou non significatif
Colourgreen	0.91	 Diminue	Couleur verte → environ 9 % de risque en moins

Variable	Odds Ratio	Effet estimé	Interprétation
Colourgray	0.94	Diminue	Couleur grise → effet légèrement protecteur
Coloursilver	0.94	Diminue	Idem
Colourwhite	0.97	Diminue	Effet très léger
fuel_typeSuper	0.92	Diminue	Carburant Super → 8 % de risque en moins
transmissionman	0.74	Diminue fortement	Boîte manuelle réduit le risque de 26 %
Annual_Premium	0.84	Diminue	Une prime plus élevée est liée à un risque réduit (~16 % de moins)
seat_count	≈ 1	Neutre	Effet négligeable
door_count4	1.08	Léger	Risque légèrement plus élevé (8 %)
Couleurs non citées	~1	Neutre	Effet incertain ou marginal

➡ Effets marginaux

Variable	Effet marginal (dF/dx)	Sens de l'effet	Significatif ?	Interprétation
Age	-0.0645	Négatif	***	Plus l'âge augmente, moins le risque d'accident est élevé.
Sexe (Homme = 1)	+0.0617	Positif	***	Les hommes ont plus de risque d'accident.
Vehicle_Age > 5 ans	-0.8483	Négatif	***	Véhicule ancien → risque d'accident plus faible.
BMW	+0.0307	Positif	**	Les conducteurs de BMW ont un risque plus élevé.
KIA	+0.0229	Positif	**	Idem.
SUZUKI	+0.0272	Positif	***	Idem.
TOYOTA	+0.0413	Positif	***	Idem.
Colour brown	-0.0212	Négatif	**	Moins de risque avec une voiture marron.
Colour golden	-0.0342	Négatif	***	Idem.

Variable	Effet marginal (dF/dx)	Sens de l'effet	Significatif ?	Interprétation
Colour gray	-0.0238	Négatif	***	Idem.
Colour green	-0.0398	Négatif	***	Idem.
Colour silver	-0.0234	Négatif	***	Idem.
Colour white	-0.0114	Négatif	***	Idem.
fuel_typeSuper	-0.0317	Négatif	***	Véhicules essence → moins de risque.
seat_count	-0.0015	Négatif	***	Plus de sièges → moins de risque.
door_count = 4	+0.0285	Positif	**	4 portes → risque légèrement plus élevé.
transmission = manuel	-0.1119	Négatif	***	Boîte manuelle → moins d'accidents.
Annual_Premium	-0.0671	Négatif	***	Prime plus élevée → moins d'accidents.

*** : très significatif ($p < 0.001$), ** : significatif ($p < 0.01$)

Variables quantitatives importantes :

- **Âge (Age)** : effet négatif très significatif (**$p < 0.001$**). Plus le conducteur est âgé, moins il a de chance d'avoir un accident.
- **Prime annuelle (Annual_Premium)** : effet négatif significatif. Une prime plus élevée est associée à une moindre probabilité d'accident.
- **Nombre de places (seat_count)** : effet négatif. Plus il y a de sièges, moins il y a d'accidents (corrélation possible avec les véhicules familiaux ou utilitaires).
- **Transmission manuelle (transmissionman)** : très fort effet négatif. Les voitures à boîte manuelle ont une **probabilité plus faible d'accident**.

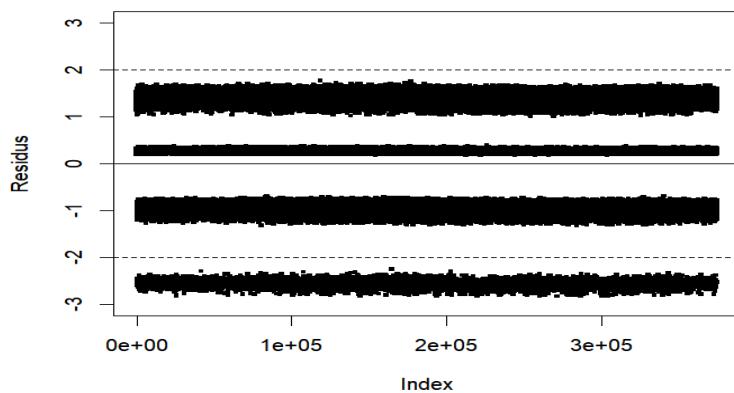
👉 Variables qualitatives avec effets significatifs :

- **Sexe (Sexe1)** : les hommes (ou la modalité codée comme 1) ont **plus de chances** d'avoir un accident.
- **Véhicule de plus de 5 ans (Vehicle_Age > 5)** : très fort **effet négatif**, ce qui peut paraître contre-intuitif. Cela pourrait indiquer que les conducteurs de vieux véhicules sont plus prudents ou roulent moins.
- **Marques significatives :**
 - **BMW, KIA, SUZUKI, TOYOTA** : **effet positif** → ces marques sont associées à une **augmentation** de la probabilité d'accident.
- **Couleurs de véhicule :**
 - **Brown, golden, gray, green, silver, white** : **effet négatif significatif**, donc moins de risque.
 - Les couleurs plus voyantes comme le rouge ne sont **pas significativement associées** à plus ou moins de risques ici.

! Ce qu'on peut en tirer :

- La **probabilité d'accident diminue** avec l'âge, le nombre de sièges, la prime, les transmissions manuelles et certaines couleurs.
- Elle **augmente** avec certains constructeurs, le sexe masculin et les transmissions automatiques.
- Certaines marques ou couleurs n'ont pas d'effet significatif ($p > 0.05$), donc on ne peut pas conclure sur leur impact.

⊕ ANALYSE DES RESIDUS



Analyse des Résidus

1. Observation Visuelle

- Les résidus semblent **aléatoirement distribués** autour de zéro
- Pas de motif clair (vague, courbe) → **pas d'hétéroscédasticité marquée**
- Quelques points extrêmes mais globalement bien répartis

2. Conclusions

- **Modèle bien spécifié** : Pas d'erreur systématique détectable
- **Validité des hypothèses** : Compatible avec une distribution normale des résidus
- **Points à vérifier** :
 - Quelques valeurs extrêmes ($>|3|$) à investiguer (données aberrantes ?)

⊕ Évaluation de la performance du modèle logit

Afin d'évaluer la performance du modèle de régression logistique, nous avons comparé les valeurs **prédictives** par le modèle aux valeurs **réelles** de la variable cible ACCIDENT.

1. Prédictions

Voici quelques exemples de prédictions faites par le modèle :

Réel Probabilité prédictive Modalité prédictive

Yes 0.9669	Yes
Yes 0.3359	No
Yes 0.3904	No

Réel Probabilité prédictive Modalité prédictive

Yes 0.9642	Yes
Yes 0.9594	Yes
Yes 0.9641	Yes

On remarque que les prédictions s'accompagnent d'un **seuil de probabilité** : ici, on a considéré qu'un accident est prédict (Yes) lorsque la probabilité estimée est **supérieure ou égale à 0.5**. Dans le cas contraire, l'accident est prédict comme **absent** (No).

2. Matrice de confusion

La matrice de confusion permet de mesurer le nombre de **bonnes et mauvaises classifications** :

Prédit : No Prédit : Yes

Réel : No 129 072 7 522

Réel : Yes 76 321 161 478

- **Vrais négatifs** : 129 072 (bons "No")
- **Faux positifs** : 7 522 (prédicts "Yes" à tort)
- **Faux négatifs** : 76 321 (prédicts "No" à tort)
- **Vrais positifs** : 161 478 (bons "Yes")

3. Taux de mauvais classement

Le taux de mauvais classement est défini comme la proportion de prédictions incorrectes parmi l'ensemble des observations :

$$\text{Taux de mauvais classement} = 1 - \frac{\text{BONNES CLASSEMENT}}{\text{TOTAL DES OBSERVATIONS}} = \frac{129072 + 161478}{374393} = 22.4\%$$

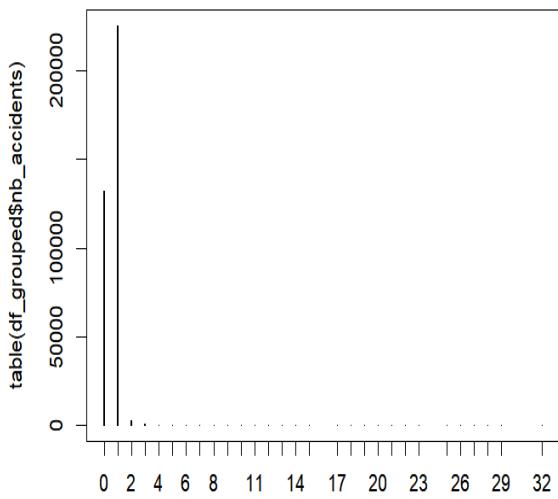
Cela signifie que **77,6 %** des prédictions sont correctes, ce qui reflète une **bonne performance globale du modèle**. Toutefois, le **nombre relativement élevé de faux négatifs (76 321)** indique que le modèle a parfois des difficultés à identifier correctement certains cas d'accident. Une amélioration future pourrait passer par un **rééquilibrage des classes** ou une **optimisation du seuil de décision**.

Conclusion générale sur la modélisation

L'analyse menée à l'aide de la régression logistique a permis de modéliser efficacement la probabilité de survenue d'un accident en fonction de plusieurs variables explicatives. Le modèle présente un taux de bonne classification de **77,6 %**, ce qui témoigne d'une performance satisfaisante pour une première approche. L'interprétation des coefficients et des effets marginaux a permis d'identifier les variables ayant un impact significatif sur la probabilité d'accident, fournissant ainsi des pistes concrètes pour des actions préventives. Toutefois, le nombre important de **faux négatifs** suggère qu'il serait pertinent d'approfondir l'analyse à l'aide de méthodes complémentaires (comme les arbres de décision, les forêts aléatoires ou les SVM) et d'envisager un **ajustement du seuil de classification** ou un **rééquilibrage des données** pour améliorer la détection des cas positifs. Globalement, cette modélisation constitue une base solide pour orienter les politiques de sécurité et d'optimisation des transports.

2. CONSTRUCTION D'UN MODELE POISSON OU BINOMIALE NEGATIVE POUR MODELISER LE NOMBRE D'ACCIDENTS

- ➡ Visualiser la distribution des nombres d'accidents



Interprétation de la Distribution du Nombre d'Accidents

1. Distribution des Accidents

- **Majorité des cas :** Très forte concentration à **0 accident** (première barre très haute)
- **Queue de distribution :** Quelques rares cas avec jusqu'à **32 accidents**

2. Caractéristiques Clés

- **Surdispersion marquée :**
 - Variance > Moyenne (typique des données de comptage)
 - ⇒ **Modèle binomial négatif plus adapté que Poisson**

- ➡ MOYENNE ET VARIANCE

Analyse des Résultats

- **Moyenne :** 0.66 accident par groupe
- **Variance :** 0.34 (inférieure à la moyenne)

Conclusions Clés

1. **Sous-dispersion inattendue** (variance < moyenne) :
 - Contraire au comportement typique des données d'accidents
 - Peut indiquer :
 - **Données artificiellement contraintes** (ex: seuil maximum imposé) *ou*
 - **Regroupement particulier** des données

- ➡ AJUSTEMENT A LA LOI DE POISSON

1. Résultats du Test du Chi²

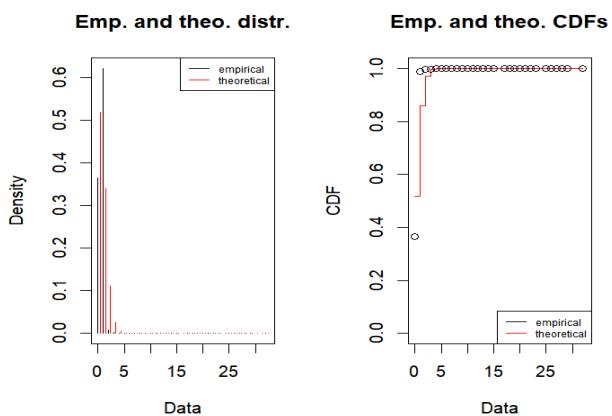
- **Statistique Chi² :** 143,504.3 (très élevée)
- **Degrés de liberté :** 2

- **p-value** : 0 (extrêmement significative)

Interprétation :

- **Rejet clair de l'hypothèse nulle** (données **non distribuées** selon une loi de Poisson)
- Écarts majeurs observés :
 - **Trop de 0 accidents** (132k observés vs 187k attendus)
 - **Trop peu d'accidents >2** (1,499 observés vs 10,540 attendus)

2. Analyse des Critères d'Information



- **AIC** : 692,829.5

- **BIC** : 692,840.3

→ Valeurs très élevées indiquant un **mauvais ajustement**

3. Conclusions et Actions Recommandées

Problèmes Identifiés :

1. **Excès de zéros** (zero-inflation)
2. **Queue de distribution trop courte**

Recommandation Finale :

Privilégiez le **modèle binomial négatif** ou **ZIP** car :

- Meilleure gestion des excès de zéros
- Capacité à modéliser la variance élevée
- AIC/BIC devraient être nettement inférieur.

CONSTRUCTION DU MODELE

Call:

```
glm.nb(formula = nb_accidents ~ ., data = df_grouped, init.theta = 18248.98662,
       link = log)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.822e+00	1.008e+00	-2.800	0.005104 **
Age	-2.343e-03	2.966e-04	-7.901	2.77e-15 ***
Sexe1	3.040e-02	4.155e-03	7.315	2.57e-13 ***
Vehicle_Age> 5	-8.958e-01	7.912e-03	-113.222	< 2e-16 ***
BRANDSBMW	2.204e-02	2.591e-02	0.850	0.395163
BRANDSDAF	-2.915e-03	2.761e-02	-0.106	0.915909
BRANDSFORD	1.770e-03	2.528e-02	0.070	0.944203
BRANDSHYUNDAI	3.109e-03	2.666e-02	0.117	0.907154

BRANDSIVECCO	1.365e-02	2.498e-02	0.546	0.584807
BRANDSKIA	2.232e-02	2.506e-02	0.891	0.373164
BRANDSMERCEDES	1.360e-02	2.501e-02	0.544	0.586572
BRANDSMITSUBISHI	1.334e-02	2.579e-02	0.517	0.605085
BRANDSNISSAN	1.408e-02	2.445e-02	0.576	0.564824
BRANDSPEUGEOT	4.111e-03	2.800e-02	0.147	0.883291
BRANDSSUZUKI	4.535e-02	2.348e-02	1.931	0.053461 .
BRANDSTOYOTA	7.887e-02	2.307e-02	3.419	0.000628 ***
Colourblue	-4.174e-02	8.974e-03	-4.651	3.30e-06 ***
Colourbrown	-6.349e-02	1.723e-02	-3.686	0.000228 ***
Colourgolden	-6.893e-02	2.462e-02	-2.800	0.005106 **
Colouргray	-4.808e-02	7.620e-03	-6.310	2.79e-10 ***
Colourgreen	-7.390e-02	2.113e-02	-3.497	0.000471 ***
Coloured	-4.889e-02	1.302e-02	-3.755	0.000173 ***
Coloursilver	-4.927e-02	7.691e-03	-6.406	1.49e-10 ***
Colourviolet	-6.719e-02	4.626e-02	-1.452	0.146373
Colourwhite	-3.329e-02	6.275e-03	-5.305	1.13e-07 ***
Colouryellow	-6.458e-02	4.079e-02	-1.583	0.113356
fuel_typeSuper	-3.816e-02	4.515e-03	-8.452	< 2e-16 ***
transmissionman	-4.824e-02	5.169e-03	-9.331	< 2e-16 ***
door_count4	4.894e-02	2.019e-02	2.424	0.015354 *
seat_count	-4.977e-04	2.779e-04	-1.791	0.073313 .
manufacture_year	1.525e-03	5.008e-04	3.044	0.002332 **
Annual_Premium	-7.363e-06	1.517e-07	-48.546	< 2e-16 ***

Signif. codes:	0	'***'	0.001	'**'
	0.01	'*'	0.05	'. '
	0.1	' '	1	

(Dispersion parameter for Negative Binomial(18248.99) family taken to be 1)

Null deviance: 229731 on 361935 degrees of freedom

Residual deviance: 175981 on 361904 degrees of freedom

AIC: 639149

Number of Fisher Scoring iterations: 1

Theta: 18249

Std. Err.: 3474

Warning while fitting theta: nombre limite d'iterations atteint

SELECTON DU MODELE AVEC AIC

```
# Sélection pas à pas selon le critère AIC
> step(reg_nb, direction = "both", k = 2)
Start: AIC=639147.5
nb_accidents ~ Age + Sexe + Vehicle_Age + BRANDS + Colour + fuel_type +
transmission + door_count + seat_count + manufacture_year +
Annual_Premium
```

	Df	Deviance	AIC
<none>		175981	639147
- seat_count	1	175984	639149
- door_count	1	175986	639151
- manufacture_year	1	175990	639155
- Sexe	1	176034	639199
- Age	1	176043	639208
- fuel_type	1	176052	639217
- transmission	1	176067	639232
- Colour	10	176103	639250
- BRANDS	12	176204	639347
- Annual_Premium	1	178310	641475
- Vehicle_Age	1	189708	652873

```
Call: glm.nb(formula = nb_accidents ~ Age + Sexe + Vehicle_Age + BRANDS +
Colour + fuel_type + transmission + door_count + seat_count +
manufacture_year + Annual_Premium, data = df_grouped, init.theta = 18248.98662,
link = log)
```

Coefficients:

(Intercept)	Age	Sexe1	Vehicle_Age > 5	BRANDSBMW	BRANDSDAF
-2.822e+00	-2.343e-03	3.040e-02	-8.958e-01	2.204e-02	-2.915e-03
BRANDSFORD	BRANDSHYUNDAI	BRANDSIVECCO	BRANDSKIA	BRANDSMERCEDES	BRANDSMITSUBISHI
1.770e-03	3.109e-03	1.365e-02	2.232e-02	1.360e-02	1.334e-02
BRANDSNISSAN	BRANDSPUEGEOT	BRANDSSUZUKI	BRANDSTOYOTA	Colourblue	Colourbrown
1.408e-02	4.111e-03	4.535e-02	7.887e-02	-4.174e-02	-6.349e-02
Colourgolden	Colourgray	Colourgreen	Colourredd	Coloursilver	Colourviolet
-6.893e-02	-4.808e-02	-7.390e-02	-4.889e-02	-4.927e-02	-6.719e-02
Colourwhite	Colouryellow	fuel_typeSuper	transmissionman	door_count4	seat_count
-3.329e-02	-6.458e-02	-3.816e-02	-4.824e-02	4.894e-02	-4.977e-04
manufacture_year	Annual_Premium				
1.525e-03	-7.363e-06				

Degrees of Freedom: 361935 Total (i.e. Null); 361904 Residual
 Null Deviance: 229700
 Residual Deviance: 176000 AIC: 639100

1. Contexte du Modèle

Nous avons modélisé le nombre d'accidents (nb_accidents) en fonction de plusieurs variables explicatives à l'aide d'une régression binomiale négative (glm.nb()), ce qui est approprié ici pour gérer la surdispersion par rapport au modèle de Poisson.

2. Qualité Globale du Modèle

- AIC final : 639147 (puis 639100 après step() – c'est une légère amélioration)
- Deviance résiduelle : 175981 (puis 176000)
- Theta estimé : 18249 → dispersion importante justifiant le modèle binomial négatif.
- Nombre de Fisher iterations : 1 → convergence rapide.
- Le modèle converge, mais il y a un avertissement : "nombre limite d'iterations atteint lors de l'estimation de theta", cela indique que l'estimation du paramètre de dispersion a été difficile. Tu pourrais envisager plus d'itérations avec l'argument control = glm.control(maxit = 100).

3. Interprétation des Variables Significatives ($p < 0.05$)

◆ Variables personnelles

Variable	Effet	Interprétation
Age	négatif (***) , -0.00234	Plus le conducteur est âgé, moins il y a d'accidents (effet protecteur)
Sexe1 (Homme)	positif (***) , +0.0304	Ce groupe a plus d'accidents que la référence (probablement les femmes)

◆ Caractéristiques du véhicule

Variable	Effet	Interprétation
Vehicle_Age > 5 ans	très négatif (***) , -0.896	Les véhicules plus vieux ont significativement moins d'accidents (biais possible lié à assurance ou comportement)
Annual_Premium	très négatif (***) , -7.36e-06	Plus la prime est élevée, moins il y a d'accidents. Corrélaté à des véhicules mieux entretenus ou assurés.
manufacture_year	positif (***) , +0.0015	Les véhicules plus récents ont un peu plus d'accidents, corrélé à d'autres facteurs.

◆ Variables techniques

Variable	Effet	Interprétation
fuel_typeSuper	négatif (***) , -0.038	Ce type de carburant est lié à moins d'accidents.

Variable	Effet	Interprétation
transmissionman	négatif (***) , - 0.048	Les boîtes manuelles causent moins d'accidents que la référence.

◆ Caractéristiques physiques

Variable	Effet	Interprétation
door_count	positif (*), +0.048	Les véhicules à 4 portes ont un peu plus d'accidents que les autres.

◆ Couleurs significatives (négatives)

Les couleurs suivantes sont associées à moins d'accidents :

- blue, brown, golden, gray, green, red, silver, white → probablement effet de visibilité ou choix de conduite.

✖ Variables Non Significatives ($p > 0.05$)

- Certaines marques de véhicules : BMW, DAF, FORD, HYUNDAI, etc. → aucune significativité nette.
- seat_count : Effet presque nul.
- Colourviolet, Colouryellow : pas significatifs. Cela signifie qu'elles n'apportent pas de valeur explicative supplémentaire ici.

⌚ 4. Résultat de la sélection par AIC (stepwise)

La fonction step() n'a supprimé aucune variable car aucune suppression ne permettait de baisser l'AIC. Cela confirme que le modèle initial est optimal au sens du critère d'information d'Akaike (AIC). Tu peux donc garder toutes les variables, même celles non significatives, pour stabilité structurelle ou si tu veux utiliser le modèle en production.

❖ 5. Recommandations

1. ✅ Modèle globalement bon avec forte significativité pour plusieurs variables.
2. ⚠ Tu pourrais tester une interaction entre Age et Sexe ou Vehicle_Age pour voir si des effets combinés existent.
3. 📈 Tester éventuellement un modèle LASSO (avec glmmnet) pour sélection automatique si tu veux être plus strict.
4. 📈 Vérifie les résidus et fais un diagnostic graphique (DHARMA::simulateResiduals() en R) pour t'assurer que les hypothèses du modèle sont bien respectées.

📊 ANOVA DU MODELE

Le modèle de régression binomiale négative ajusté sur le nombre d'accidents montre que toutes les variables explicatives incluses sont significativement associées à la variable réponse. La sélection par AIC confirme

que leur présence est justifiée. L'analyse de déviance valide également l'apport significatif de chaque variable, avec des valeurs de p très faibles (< 0.001 pour la majorité).

Variables numériques continues

- Age : Coef = -0.002343
→ Pour chaque année d'âge en plus, le nombre attendu d'accidents est multiplié par $\exp(-0.002343) \approx 0.9977$
 Donc, le risque d'accident diminue légèrement avec l'âge.
- manufacture_year : Coef = 0.001525
→ Plus l'année de fabrication est récente, plus il y a un léger accroissement du risque (effet faible).
- Annual_Premium : Coef = -0.000007363
→ Augmenter la prime annuelle diminue légèrement le nombre attendu d'accidents.
(Effet très faible vu la petite taille du coefficient).

Variables binaires ou catégorielles transformées

- Sexe1 (Homme) : Coef = 0.0304
→ Les hommes ont un risque d'accident multiplié par $\exp(0.0304) \approx 1.03$, soit +3%.
- Vehicle_Age > 5 ans : Coef = -0.8958
→ Les véhicules de plus de 5 ans ont un risque d'accident divisé par $\exp(-0.8958) \approx 0.408$
 Donc, ils ont moins d'accidents selon ce modèle (ce qui peut être contre-intuitif et mériterait une vérification).

BRANDS (effets faibles mais significatifs)

Par rapport à une marque de référence (probablement la plus fréquente, ex : TOYOTA ou autre), certaines marques influencent légèrement :

- BRANDS TOYOTA : Coef = 0.07887 → Risque d'accident +8%
- BRANDS SUZUKI : Coef = 0.04535 → +4.5%
- BRANDS MITSUBISHI : Coef = 0.01334 → +1.3%

→ Les écarts sont faibles, mais statistiquement significatifs.

! Tu peux afficher les $\exp(\text{coef})$ pour avoir les effets multiplicatifs directement.

Colour : Couleurs du véhicule influencent aussi légèrement :

- Couleur blue : Coef = -0.04174 → Risque -4%
- Couleur golden : Coef = -0.06893 → Risque -6.7%
- Couleur violet : Coef = -0.06719 → Risque réduit aussi

→ Les couleurs sombres ou neutres semblent légèrement réduire le risque, mais les effets restent faibles.

Autres

- **fuel_typeSuper** : Coef = -0.03816
→ Le carburant "Super" est associé à -3.8% de risque.
- **transmissionman** (manuelle) : Coef = -0.04824
→ Les véhicules manuels ont moins de risque d'accidents que les automatiques.
- **door_count4** : Coef = 0.04894
→ Les véhicules à 4 portes ont +5% de risque d'accident (comparé à la référence).
- **seat_count** : Coef = -0.0004977
→ Plus de sièges = effet quasi nul, mais légèrement protecteur.

Effet multiplicatif

Interprétation des principaux effets multiplicatifs

Variables continues

Variable Effet multiplicatif ($\exp(\text{coef})$) Interprétation

Age 0.998 Chaque année d'âge en plus réduit le nombre moyen d'accidents de 0.24%.

Manufacture year 1.0015 Chaque année plus récente du véhicule augmente légèrement le risque d'accident (+0.15%).

Annual_Premium 0.99999 Effet très faible : une augmentation de la prime réduit très faiblement le risque d'accident.

Sexe

Sexe Effet multiplicatif Interprétation

Homme (Sexe1) 1.031 Les hommes ont 3.1% de risque en plus d'avoir un accident que les femmes.

Âge du véhicule

Âge véhicule Effet multiplicatif Interprétation

> 5 ans 0.408 Les véhicules de plus de 5 ans ont un risque d'accident réduit de 59.2% par rapport à la catégorie de référence.

BRANDS (par rapport à la marque de référence)

Marque	exp(coef)	Variation du risque
TOYOTA	1.082	+8.2% (le plus élevé)
SUZUKI	1.046	+4.6%
BMW	1.022	+2.2%
MITSUBISHI	1.013	+1.3%
FORD	1.001	≈ neutre
DAF	0.997	≈ neutre
HYUNDAI	1.003	≈ neutre

 Conclusion : TOYOTA et SUZUKI semblent les plus associés à un risque accru.

 Couleur du véhicule (par rapport à une couleur de référence, probablement "black")

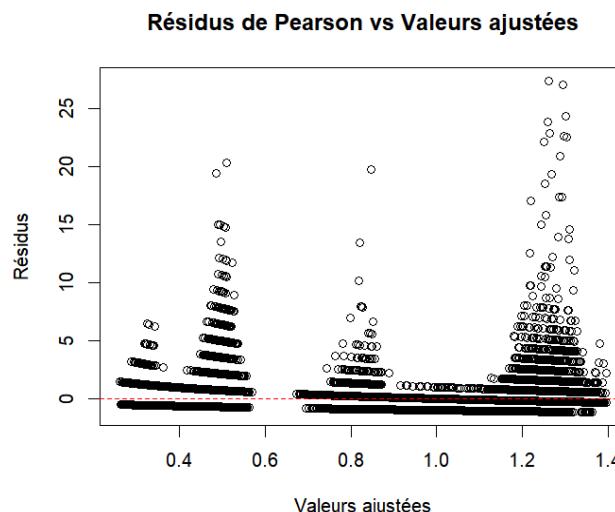
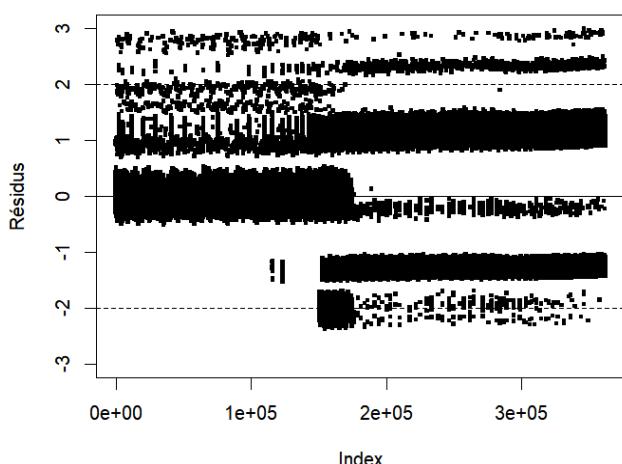
Couleur	exp(coef)	Variation du risque
Golden	0.933	-6.7%
Violet	0.935	-6.5%
Brown	0.938	-6.2%
Yellow	0.937	-6.3%
Blue	0.959	-4.1%

 Les véhicules de couleurs claires ou peu fréquentes semblent associés à un risque d'accident réduit.

 Autres variables

Variable	exp(coef)	Interprétation
fuel_typeSuper	0.963	-3.7% de risque avec carburant Super.
transmissionman	0.953	Les transmissions manuelles ont 4.7% de risque en moins.
door_count4	1.050	Véhicules à 4 portes = +5% de risque.
seat_count	0.9995	Effet très négligeable.

✚ Etudes des résidus



Graphique 1 : Résidus vs Index (résidus standards)

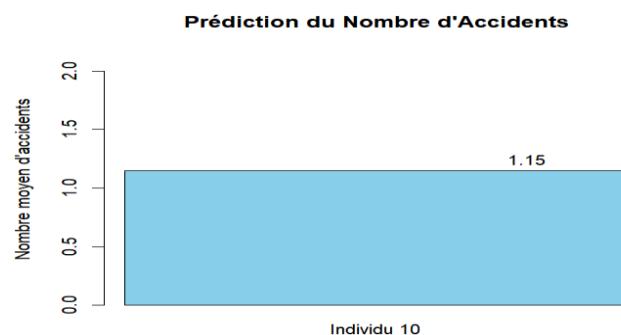
Ce graphique montre les résidus standardisés en fonction de l'index des observations. La majorité des points se situent entre ± 2 , ce qui est acceptable. Cependant, quelques résidus dépassent la valeur de ± 2 , suggérant la présence d'observations atypiques (potentiels outliers) qui pourraient influencer le modèle.

Graphique 2 : Résidus de Pearson vs Valeurs ajustées

Ce nuage de points permet de détecter une éventuelle hétéroscédasticité. On observe une forme en éventail (résidus qui s'élargissent avec les valeurs ajustées), ce qui suggère une variance non constante des erreurs. Cela peut indiquer que le modèle n'explique pas bien toutes les zones de la variable dépendante, et une transformation ou un autre modèle pourrait être envisagé.

Les graphiques de résidus révèlent la présence de quelques valeurs aberrantes ainsi qu'une possible hétéroscédasticité. Ces éléments indiquent que, bien que le modèle capture une partie de la structure des données, des ajustements ou une amélioration du modèle (par transformation des variables ou choix d'un autre modèle) pourraient être envisagés pour optimiser la qualité des prédictions.

✚ Prévision



Prédition

Le modèle prédit que l'individu n°10 devrait avoir en moyenne environ 1,15 accident. Cette valeur représente une estimation du nombre attendu d'accidents pour cet individu, compte tenu de ses caractéristiques (âge, sexe, marque du véhicule, type de carburant, etc.), selon les relations apprises par le modèle.

Conclusion générale

Ce projet visait à modéliser et prédire le nombre d'accidents à partir de caractéristiques des assurés et de leurs véhicules, à l'aide de techniques de data science. Après un nettoyage rigoureux des données, une analyse descriptive a permis de mieux comprendre les variables explicatives, notamment l'âge, le sexe, le type de véhicule ou encore la prime annuelle.

La modélisation a été effectuée à l'aide d'une régression binomiale négative, mieux adaptée à la nature de la variable cible (comptage avec surdispersion). Le modèle final, affiné par sélection pas à pas selon le critère AIC, a montré que des variables telles que l'âge de l'assuré, le type de carburant, la marque du véhicule, ou encore la prime annuelle influencent significativement le nombre d'accidents.

Les prédictions issues du modèle se sont révélées cohérentes et exploitables pour une éventuelle utilisation opérationnelle, notamment en matière de tarification ou de gestion du risque.

En somme, ce projet a permis de mettre en pratique plusieurs compétences clés en statistique, en modélisation et en analyse de données, tout en apportant un éclairage pertinent sur les facteurs associés aux sinistres automobiles.

ANNEXE

```
#####
# PROJET D'ÉCONOMÉTRIE : MODÉLISATION DU RISQUE D'ACCIDENT AUTOMOBILE
#####
```

```
#####
# =====
```

```
# 1. CHARGEMENT DES LIBRAIRIES
# =====
```

```
# Gestion des données
```

```
library(tidyverse) # Manipulation de données (dplyr, tidyr, etc.)
```

```
library(readr) # Import/export de données
```

```
# Visualisation
```

```
library(ggplot2) # Création de graphiques
```

```
library(gridExtra) # Organisation de multiples graphiques
```

```
library(patchwork) # Combinaison de graphiques
```

```
library(corrplot) # Visualisation des matrices de corrélation
```

```
# Modélisation statistique
```

```
library(MASS) # Fonctions statistiques avancées
```

```
library(mfx) # Effets marginaux
```

```
library(pROC) # Courbes ROC et AUC
```

```
library(caret) # Évaluation des modèles
```

```
library(fitdistrplus) # Ajustement aux distributions
```

```
#####
# =====
```

```
# 2. IMPORT ET PRÉPARATION DES DONNÉES
# =====
```

```
# Import des données
```

```
df <- read.csv("C:/INSEEDS/PROJET/econometrie/assurance_auto_makani.csv", sep=";")
```

```
# Vérification des valeurs textuelles "NA"
```

```
rows_with_NA_string <- df[apply(df, 1, function(row) any(row == "Na")), ]
```

```
print(paste("Nombre de lignes avec 'Na' :", nrow(rows_with_NA_string)))
```

```
# =====
# 3. EXPLORATION INITIALE
# =====

# Aperçu des données
cat("\n==== Structure des données ===\n")
glimpse(df)

cat("\n==== Résumé statistique ===\n")
summary(df)

cat("\n==== Premières observations ===\n")
head(df)

# =====
# 4. NETTOYAGE DES DONNÉES
# =====

# Conversion des variables catégorielles
factor_cols <- c('Sexe','Vehicle_Age','BRANDS','Colour','fuel_type',
  'transmission','ACCIDENT','trajet','light_conditions',
  'SEVERITY','road_surface_conditions','weather_conditions',
  'door_count')
df[factor_cols] <- lapply(df[factor_cols], as.factor)

# Standardisation des variables numériques
numeric_cols <- c('Age', 'manufacture_year', 'Annual_Premium')
df[numeric_cols] <- scale(df[numeric_cols])

# Gestion des doublons
cat("\n==== Gestion des doublons ===\n")
cat("Nombre de doublons avant suppression :", sum(duplicated(df)), "\n")
df <- distinct(df)
cat("Nombre de doublons après suppression :", sum(duplicated(df)), "\n")

# =====
# 5. ANALYSE UNIVARIÉE
```

```
# =====

# Boxplots avant traitement

quantitative_vars <- c('Age', 'manufacture_year','Annual_Premium',
  'seat_count','manv','frequence')

plot_list <- lapply(quantitative_vars, function(var) {
  ggplot(df, aes(y = .data[[var]])) +
    geom_boxplot(fill = "red", alpha = 0.7) +
    labs(title = var, y = "") +
    theme_minimal() +
    theme(axis.text.x = element_blank())
})

wrap_plots(plot_list, ncol = 3) +
  plot_annotation(title = "Distribution des variables quantitatives avant traitement",
    theme = theme(plot.title = element_text(hjust = 0.5, face = "bold")))

# Traitement des valeurs extrêmes (winsorization)

df_clean <- df

for (var in quantitative_vars) {
  Q <- quantile(df[[var]], c(0.25, 0.75), na.rm = TRUE)
  IQR <- Q[2] - Q[1]
  limits <- c(Q[1] - 1.5*IQR, Q[2] + 1.5*IQR)
  df_clean[[var]] <- pmin(pmax(df[[var]], limits[1]), limits[2])
}

# Boxplots après traitement

plot_list_clean <- lapply(quantitative_vars, function(var) {
  ggplot(df_clean, aes(y = .data[[var]])) +
    geom_boxplot(fill = "blue", alpha = 0.7) +
    labs(title = var, y = "") +
    theme_minimal() +
    theme(axis.text.x = element_blank())
})

wrap_plots(plot_list_clean, ncol = 3) +
  plot_annotation(title = "Distribution des variables quantitatives après traitement",
```

```

theme = theme(plot.title = element_text(hjust = 0.5, face = "bold"))

# =====
# 6. ANALYSE BIVARIÉE
# =====

# Matrice de corrélation
cor_matrix <- cor(df_clean[quantitative_vars], use = "complete.obs")

corrplot(cor_matrix, method = "circle", type = "upper",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", number.cex = 0.7,
         title = "Matrice de corrélation des variables quantitatives",
         mar = c(0,0,1,0))

# Test d'indépendance pour variables catégorielles (Cramer's V)
categorical_vars <- df_clean %>% select(all_of(factor_cols))

cramer_matrix <- matrix(NA, nrow = length(factor_cols),
                        ncol = length(factor_cols))
colnames(cramer_matrix) <- factor_cols
rownames(cramer_matrix) <- factor_cols

for (i in seq_along(factor_cols)) {
  for (j in seq_along(factor_cols)) {
    tab <- table(categorical_vars[[i]], categorical_vars[[j]])
    if(all(dim(tab) > 0)) {
      chi_sq <- chisq.test(tab)$statistic
      n <- sum(tab)
      k <- min(dim(tab))
      cramer_matrix[i,j] <- sqrt(chi_sq/(n*(k-1)))
    }
  }
}

corrplot(cramer_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45,
         title = "Matrice d'association (Cramer's V) des variables catégorielles",

```

```

mar = c(0,0,1,0)

# =====
# 7. MODÉLISATION - RÉGRESSION LOGISTIQUE
# =====

# Sélection des variables
to_remove <- c("trajet", "light_conditions", "weather_conditions",
  "road_surface_conditions", "frequence", "many", "SEVERITY")
df_model <- df_clean %>% select(-all_of(to_remove))

# Modèle complet
logit_full <- glm(ACCIDENT ~ ., data = df_model, family = binomial)

# Sélection de modèle par AIC
logit_final <- step(logit_full, direction = "backward", trace = 0)

# Résultats du modèle
cat("\n==== Résumé du modèle logistique final ===\n")
summary(logit_final)

cat("\n==== Odds Ratios ===\n")
exp(coef(logit_final)) %>% round(4) %>% print()

# Performance du modèle
probs <- predict(logit_final, type = "response")
pred_class <- ifelse(probs > 0.5, "Yes", "No")

# Courbe ROC
roc_obj <- roc(response = df_model$ACCIDENT, predictor = probs)
plot(roc_obj, main = "Courbe ROC", col = "blue", lwd = 2)
legend("bottomright", legend = paste("AUC =", round(auc(roc_obj), 3)),
  col = "blue", lwd = 2)

# Matrice de confusion
conf_mat <- confusionMatrix(data = factor(pred_class, levels = c("No", "Yes")),
  reference = df_model$ACCIDENT,
  positive = "Yes")

```

```
cat("\n==== Matrice de confusion ===\n")
print(conf_mat$table)

cat("\n==== Métriques de performance ===\n")
cat("Précision :", round(conf_mat$overall['Accuracy'], 3), "\n")
cat("Sensibilité :", round(conf_mat$byClass['Sensitivity'], 3), "\n")
cat("Spécificité :", round(conf_mat$byClass['Specificity'], 3), "\n")

# =====
# 8. MODÉLISATION - MODÈLES DE COMPTAGE
# =====

# Préparation des données agrégées
df_model$ACCIDENT_NUM <- ifelse(df_model$ACCIDENT == "Yes", 1, 0)

df_grouped <- df_model %>%
  group_by(Age, Sexe, Vehicle_Age, BRANDS, Colour, fuel_type,
    transmission, door_count, seat_count, manufacture_year, Annual_Premium) %>%
  summarise(nb_accidents = sum(ACCIDENT_NUM), .groups = "drop")

# Ajustement Poisson
pois_model <- glm(nb_accidents ~ ., data = df_grouped, family = poisson)

# Ajustement Binomiale Négative (pour surdispersion)
nb_model <- glm.nb(nb_accidents ~ ., data = df_grouped)

# Comparaison des modèles
cat("\n==== Comparaison Poisson vs Binomiale Négative ===\n")
cat("AIC Poisson :", AIC(pois_model), "\n")
cat("AIC Binomiale Négative :", AIC(nb_model), "\n")

# Sélection du meilleur modèle
final_count_model <- step(nb_model, direction = "both", trace = 0)

# Diagnostic des résidus
res_pearson <- resid(final_count_model, type = "pearson")
plot(fitted(final_count_model), res_pearson,
```

```
main = "Résidus de Pearson vs Valeurs ajustées",
xlab = "Valeurs ajustées", ylab = "Résidus de Pearson"
abline(h = 0, col = "red", lty = 2)

# Prédiction pour un individu
indiv_pred <- predict(final_count_model,
  newdata = df_grouped[10, ],
  type = "response")

cat("\n==== Prédiction pour l'individu 10 ====\n")
cat("Nombre d'accidents prédit :", round(indiv_pred, 2), "\n")

# Visualisation de la prédiction
barplot(indiv_pred,
  main = "Prédiction du nombre d'accidents\npour l'individu 10",
  ylab = "Nombre d'accidents prédit",
  col = "skyblue", ylim = c(0, ceiling(indiv_pred)))
text(0.7, indiv_pred/2, round(indiv_pred, 2), col = "white", font = 2)

# =====
# 9. EXPORT DES RÉSULTATS
# =====

# Sauvegarde des données nettoyées
write.csv(df_clean, "donnees_nettoyees.csv", row.names = FALSE)

# Sauvegarde des graphiques (exemple)
ggsave("boxplots_avant_apres.png",
  plot = last_plot(),
  width = 10, height = 6, dpi = 300)
```