MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE RECHERCHE SCIENTIFIQUE

Institut Supérieur de Statistique D'Econométrie

REPUBLIQUE DE COTE D'IVOIRE

Union-Discipline-Travail

MASTER 1

STATISTIQUES – ECONOMETRIE – DATA SCIENCE

MINI PROJET

ANALYSE STATISTIQUES ECONOMETRIQUES

MODÉLISATION DES ACCIDENTS AUTOMOBILES

Nom: YOBO

Prénom(s): BAYE GUY ANGE HENOC

Enseignant – Encadreur

AKPOSSO DIDIER MARTIAL


Accident

Avant-propos

Le présent projet a pour objectif d'analyser un jeu de données issu du secteur de l'assurance IARD (assurance de dommages), plus précisément dans le domaine de l'assurance automobile. Le jeu de données étudié, intitulé assurance_auto_makani.csv, contient les informations de 374 393 clients d'une société d'assurance. Ces données permettent d'examiner différents profils d'assurés ainsi que les sinistres déclarés, dans le but de mieux comprendre les facteurs associés aux accidents automobiles.

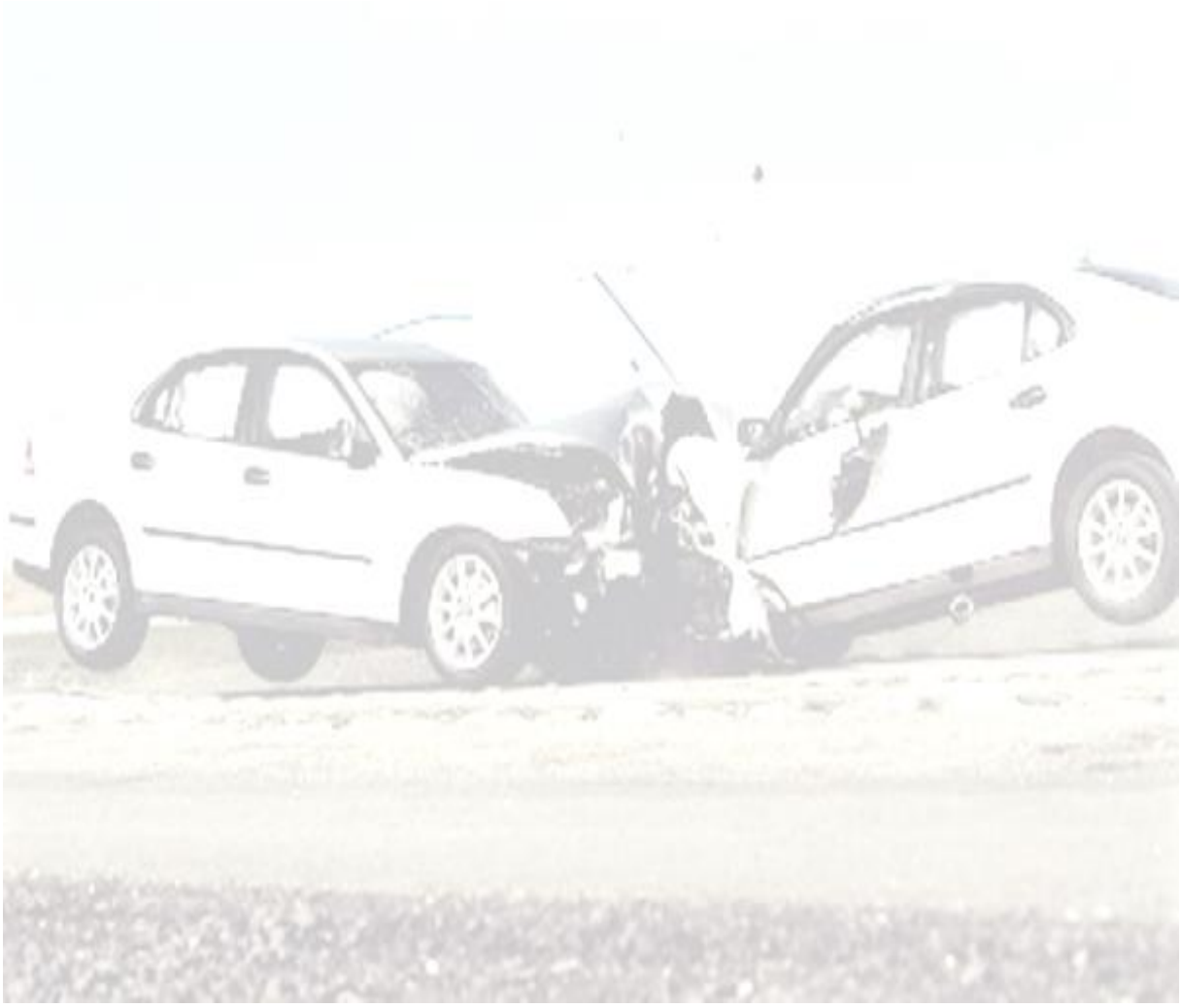
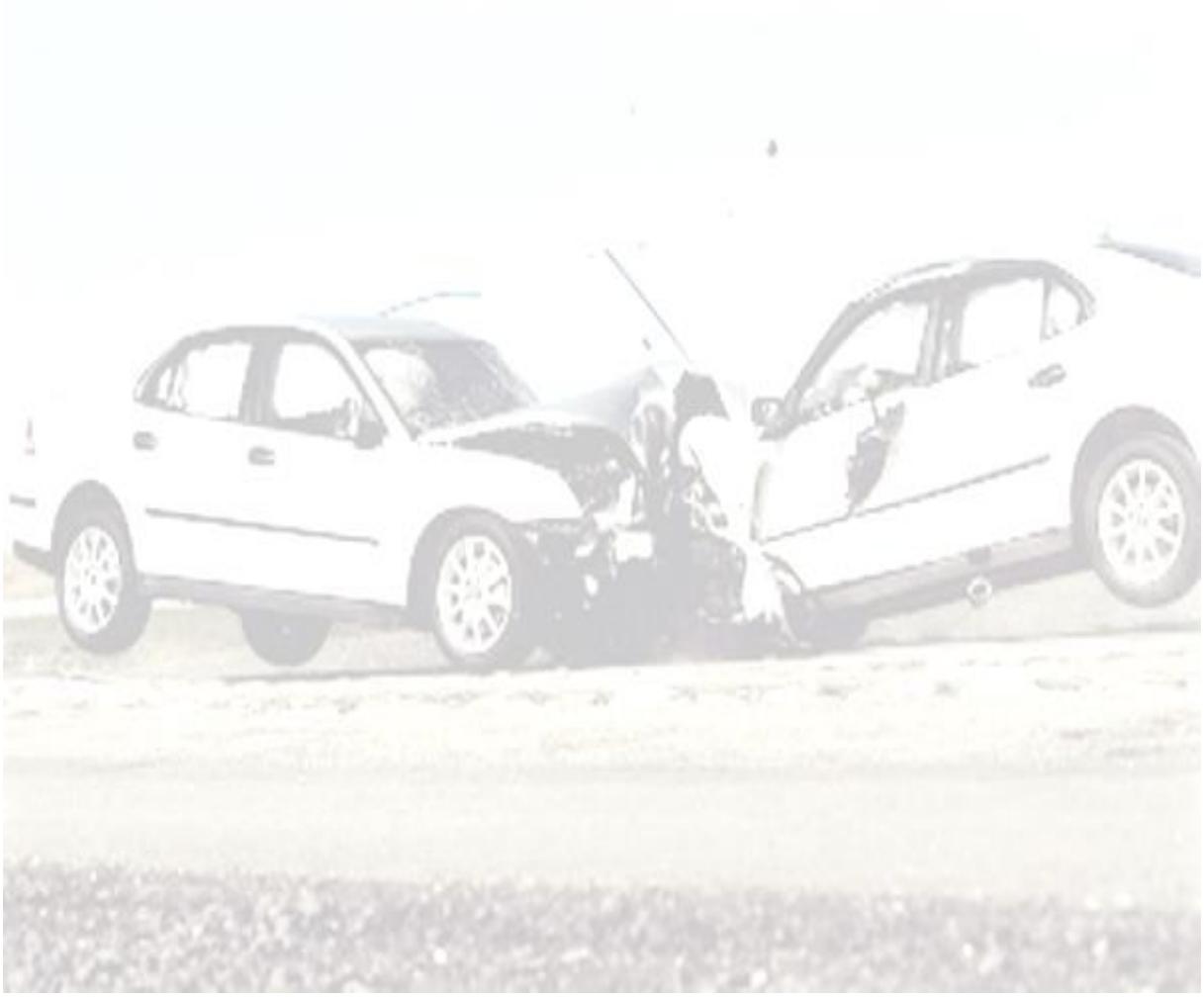


Table des matières

LISTE DES FIGURES



INTRODUCTION

Dans le secteur de l'assurance automobile, la compréhension fine du risque associé à chaque assuré est essentielle pour adapter les tarifs, prévenir la sinistralité et optimiser la gestion du portefeuille client. Le jeu de données **assurance_auto_makani.csv**, utilisé dans cette étude, regroupe les informations de **374 393 assurés** d'une compagnie d'assurance IARD, incluant les caractéristiques personnelles des clients, celles de leurs véhicules, ainsi que les circonstances des éventuels accidents. L'objectif principal de ce projet est de modéliser le risque d'accident automobile à partir des variables disponibles, à travers deux axes : (1) estimer la **probabilité de survenance** d'un accident selon le profil de l'assuré, et (2) prédire le **nombre d'accidents** pour chaque assuré. Cette double approche vise à améliorer la segmentation des assurés selon leur niveau de risque, en vue d'optimiser les stratégies tarifaires et préventives.

Les résultats attendus sont : une modélisation fiable de la variable accident (oui/non) pour détecter les facteurs explicatifs majeurs, une estimation du nombre d'accidents via une régression adaptée aux données de comptage, une meilleure compréhension des variables influençant le risque automobile (liées au véhicule, au comportement de l'assuré et aux conditions de circulation), ainsi que des recommandations concrètes sur l'ajustement des primes d'assurance en fonction des profils de risque.

La méthodologie suivie se divise en deux grandes étapes. D'abord, un **prétraitement des données** comprenant le nettoyage, la gestion des valeurs manquantes, la transformation des variables (notamment la **binarisation** de la variable « fréquence » en une variable cible « accident »), l'encodage des variables catégorielles et, si besoin, la normalisation. Ensuite, deux **analyses statistiques** sont menées : une **régression logistique** pour modéliser la probabilité de survenance d'un accident, et une **régression de Poisson** pour modéliser le nombre d'accidents. La qualité de ces modèles sera évaluée à l'aide d'indicateurs appropriés (AUC pour la logistique, déviance ou RMSE pour la Poisson), afin d'assurer la robustesse et l'interprétabilité des résultats.

PARTIE I : ANALYSE DESCRIPTIVE ET EXPLORATOIRE DES DONNEES.

A) APPROCHE METHODOLOGIQUE DES DONNEES

L'approche méthodologique des données englobe l'organisation, la collecte, l'analyse et l'interprétation des données dans le cadre d'une étude ou d'une recherche. Elle repose sur un ensemble de principes, de techniques et de processus visant à traiter les données de manière systématique et rigoureuse, afin d'obtenir des résultats fiables et pertinents.

A.1) Information sur le jeu de donnée

Âge	Sexe	Âge Véhicule	Marques	Couleur	Carburant	Sièges	Portes	Année Fabric.	Transmission	Accident	Gravité	Trajet	Lumière	Météo	Route	Manœuvre	Fréquence	Prime annuelle
22		0<5	NISSAN	gris	Gazoil	21	4	2007 man		Oui	3	1	...	...	...	...	...	...
22		0>5	NISSAN	argent	Gazoil	21	4	2008 man		Oui	2	1	...	...	...	...	...	...
22		1>5	NISSAN	argent	Gazoil	21	4	2000 man		Oui	3	3	...	...	...	...	...	...
22		1<5	NISSAN	gris	Gazoil	21	4	2010 man		Oui	3	3	...	...	...	...	...	...
22		0<5	NISSAN	bleu	Gazoil	21	4	2006 man		Oui	3	1	...	...	...	...	...	...
22		1<5	NISSAN	gris	Gazoil	21	4	2007 man		Oui	3	4	...	...	...	...	...	...
47		0>5	TOYOTA	noir	Gazoil	21	4	2007 man		Non	0	0	...	...	...	...	...	...
24		0<5	TOYOTA	noir	Gazoil	21	4	2010 man		Oui	2	4	...	...	...	...	...	...
27		0<5	TOYOTA	noir	Gazoil	21	4	2010 man		Oui	3	4	...	...	...	...	...	...
45		0>5	MERCEDES	noir	Gazoil	21	4	2010 man		Non	0	0	...	...	...	...	...	...
28		1<5	SUZUKI	noir	Gazoil	21	4	2007 man		Oui	3	4	...	...	...	...	...	...
29		1<5	SUZUKI	noir	Gazoil	21	4	2007 man		Oui	3	4	...	...	...	...	...	...

Figure 1extrait du jeu de donnée

A.2) Détection et apurement des valeurs Manquantes et Aberrantes / extrême

Dans cette section, nous allons chercher à identifier visuellement les éventuelles valeurs manquantes dans notre jeu de données, puis à les traiter. Ces valeurs peuvent provenir d'erreurs de mesure, de saisie, de calcul ou bien de valeurs extrêmes réelles présentes dans les données. Les valeurs atypiques peuvent avoir un impact majeur sur les résultats des analyses statistiques, en faussant des indicateurs comme la moyenne ou l'écart-type, mais aussi en influençant les tests d'hypothèse. Il est donc crucial de détecter et de traiter ces valeurs extrêmes avant de procéder à toute analyse statistique.

A.2.1) Visualisation des valeurs manquantes manquantes

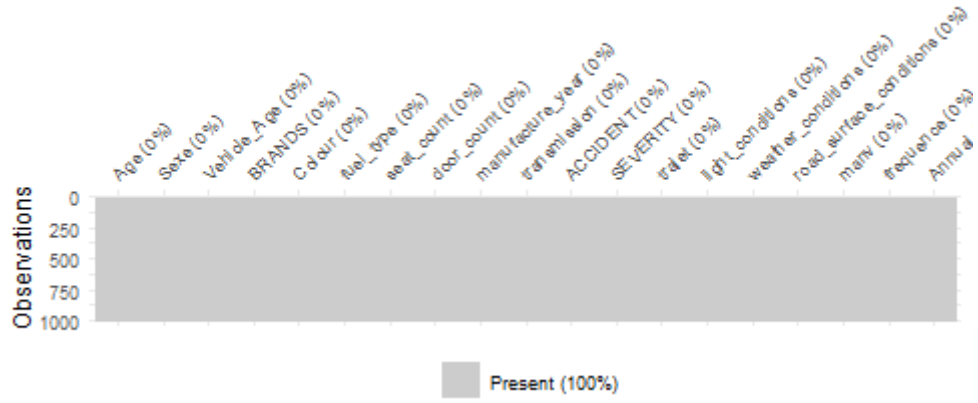


Figure 2 Valeur manquantes

Notre jeu de donnée ne présente aucunes valeurs manquantes

A.2.2) Visualisation des valeurs abberentes

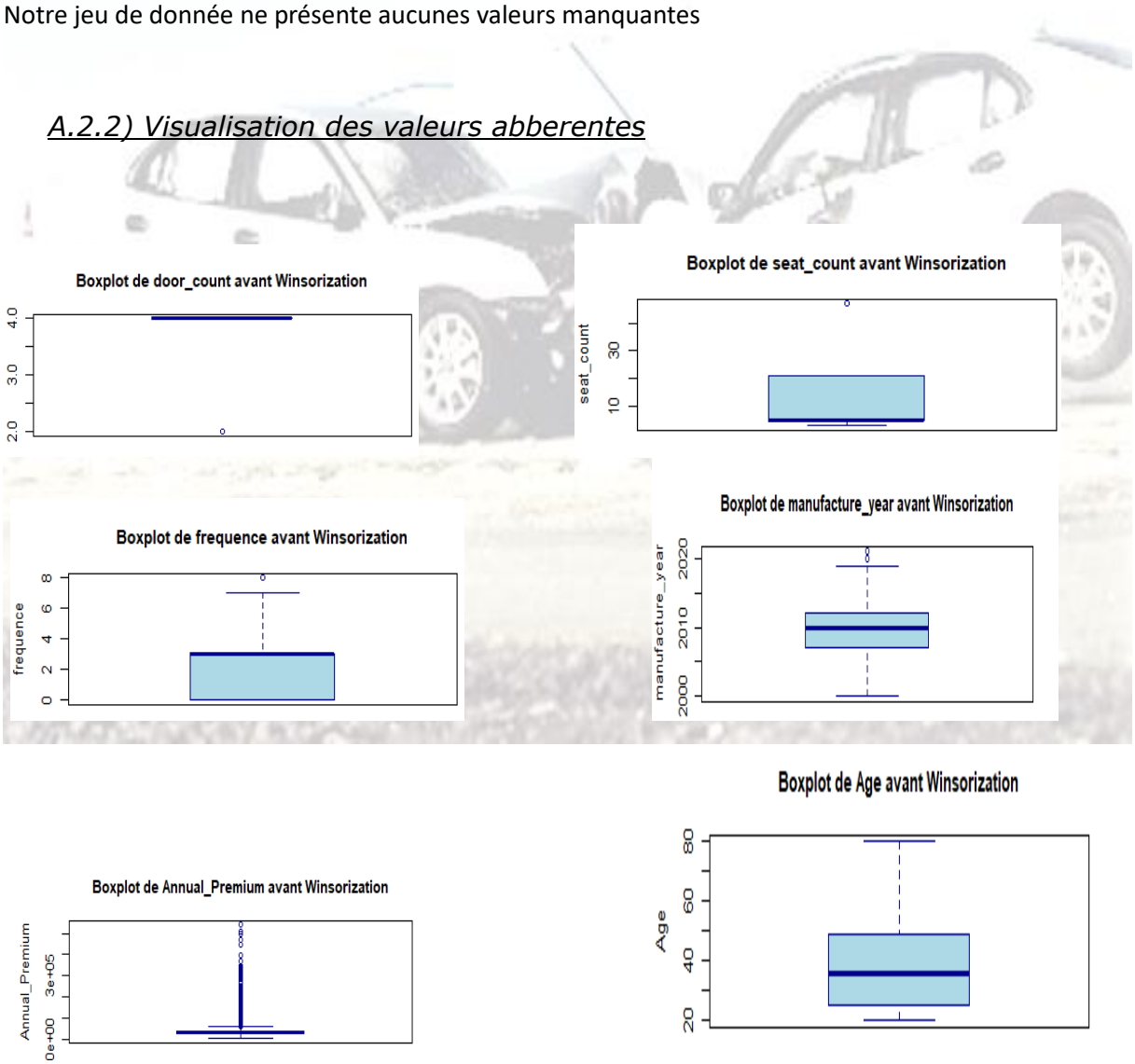


Figure 3 Valeurs abberentes

A.2.3) Visualisation des valeurs abberentes winzorisées

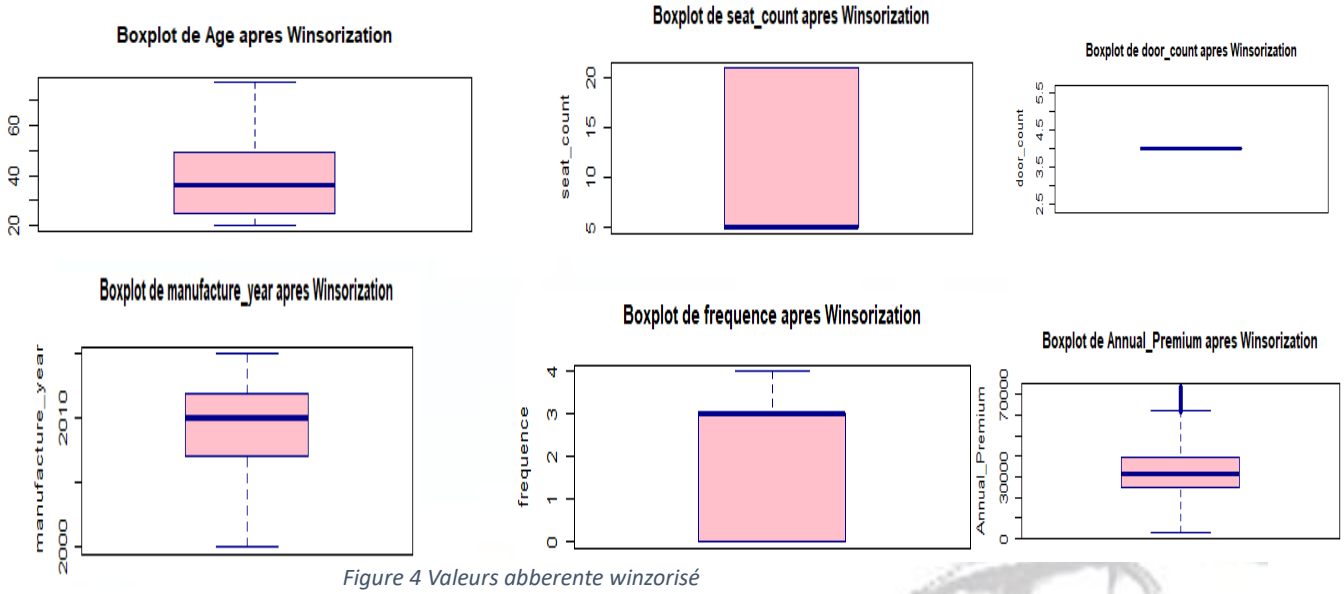
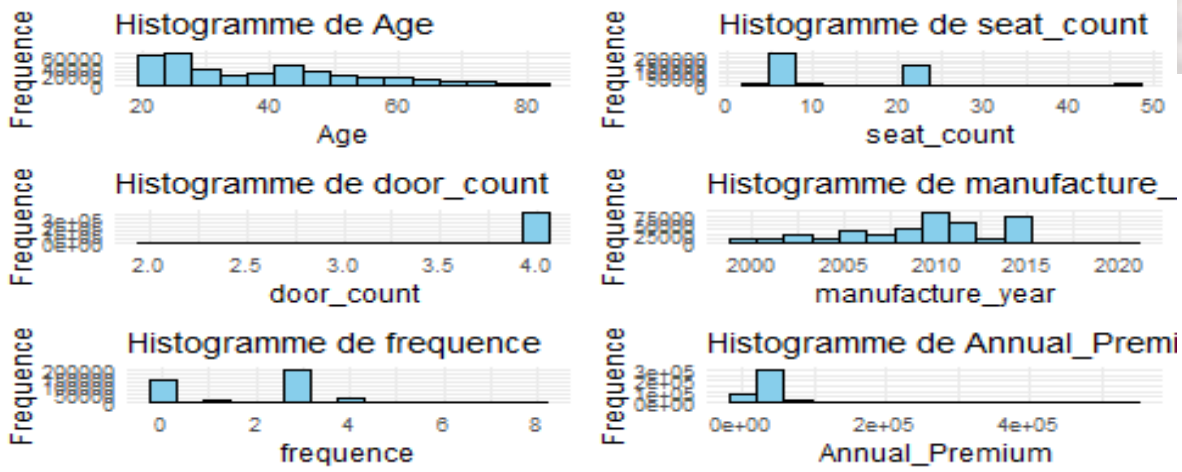


Figure 4 Valeurs abberente winzorisé

B) STATISTIQUES DESCRIPTIVES (PARAMETRES)

	mean	sd	min	max	median	IQR
Age	38.49	15.17	20	77	36	24.00
seat_count	11.05	7.71	5	21	5	16.00
door_count	4.00	0.00	4	4	4	0.00
manufacture_year	2009.31	4.13	2000	2015	2010	5.00
frequence	1.91	1.52	0	4	3	3.00
Annual_Premium	30486.02	15786.23	2630	72852	31692	14901.75

B.1) Graphiques des quantitatives



◆ Age

- **Distribution fortement asymétrique à droite (positive)** : la majorité des assurés sont jeunes (entre 20 et 40 ans).
- Très peu d'assurés ont plus de 60 ans.

◆ seat_count (Nombre de sièges)

- La distribution est **fortement concentrée autour de 21 sièges**, ce qui est surprenant et pourrait indiquer une erreur ou une codification particulière (ex. : véhicule collectif ou erreur d'unité).
- Ce point mérite une **vérification dans les données sources**.

◆ door_count (Nombre de portes)

- La **majorité des véhicules ont 4 portes**.
- Quelques rares cas de véhicules à 2 ou 3 portes.

◆ manufacture_year (Année de fabrication)

- La majorité des véhicules sont fabriqués **entre 2005 et 2015**, ce qui indique une flotte assez récente.
- Très peu de véhicules datent d'avant 2000 ou après 2018.

◆ frequence (Fréquence des sinistres)

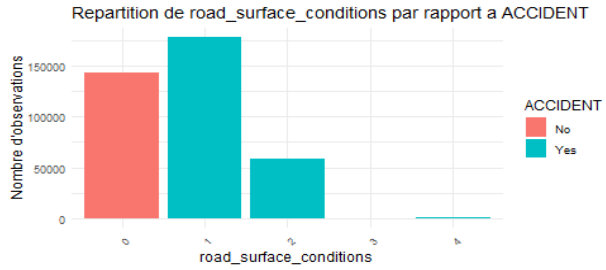
- Distribution **fortement asymétrique à droite** : la majorité des clients n'ont **aucun ou un seul accident**.
- Quelques cas rares avec une fréquence plus élevée (jusqu'à 8).

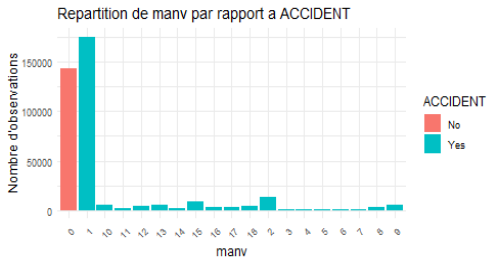
◆ Annual_Premium (Prime annuelle)

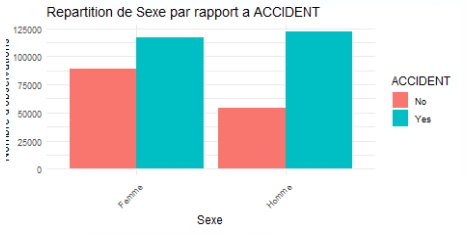
- Distribution également **asymétrique à droite** : la majorité des primes sont **basses à modérées**.
- Quelques clients ont une **prime très élevée** (> 200 000), ce qui pourrait indiquer des cas particuliers ou des profils à risque.

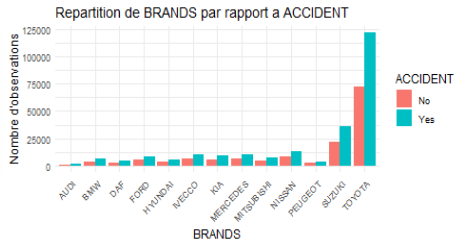
C) Relation entre les variables et la Variable cible « ACCIDENT »

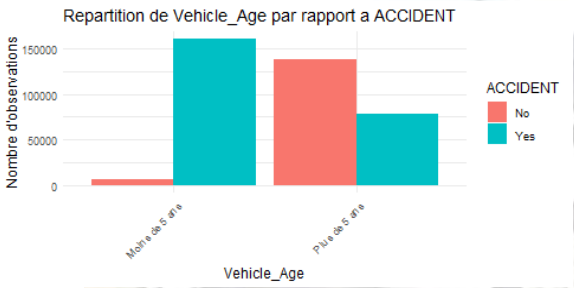
C.1) distribution des variables qualitatives

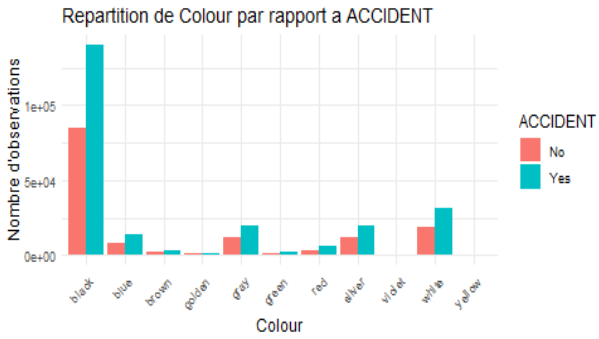


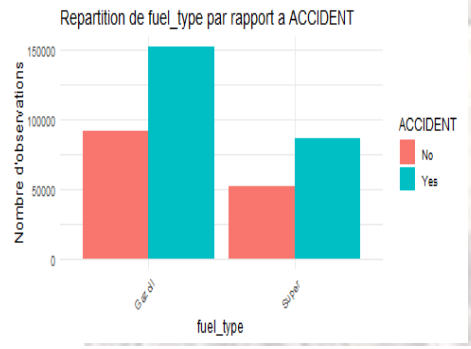


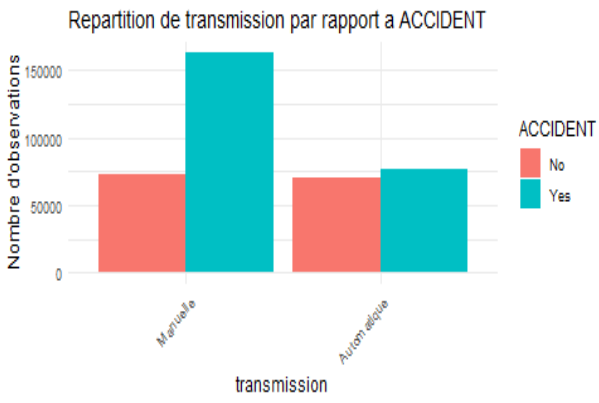


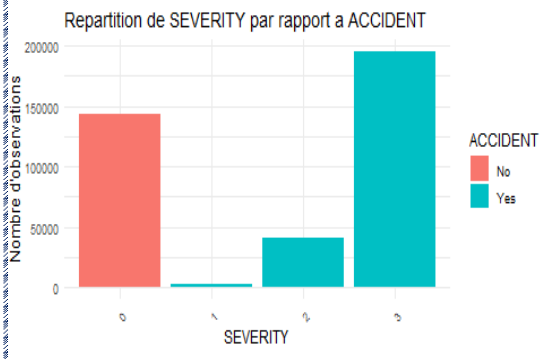












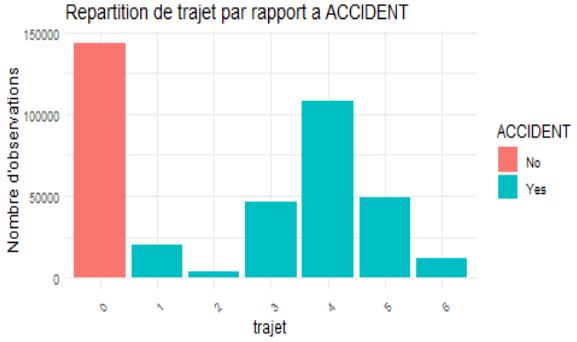
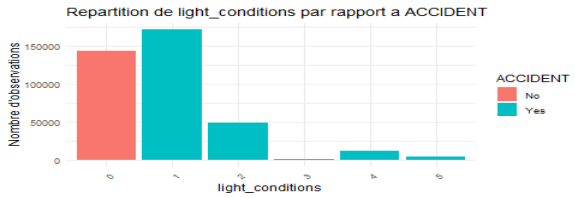
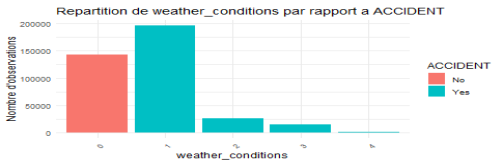
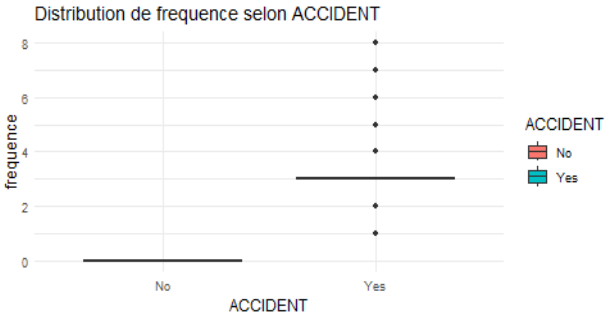


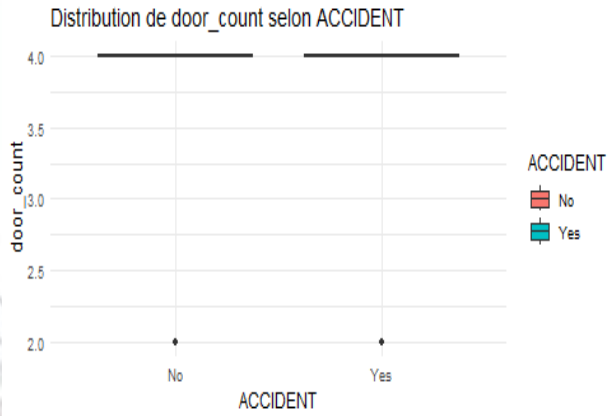
Figure 5 variable quali

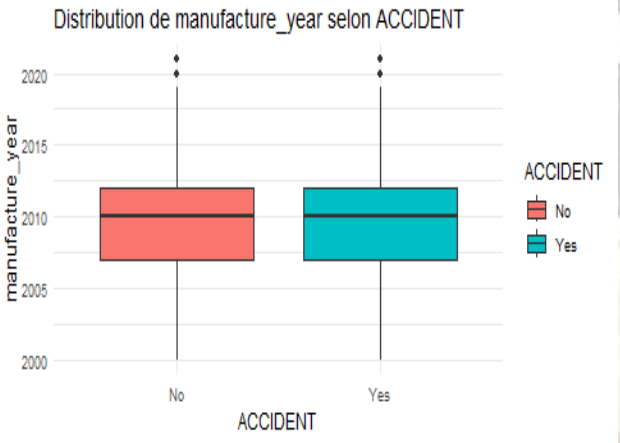


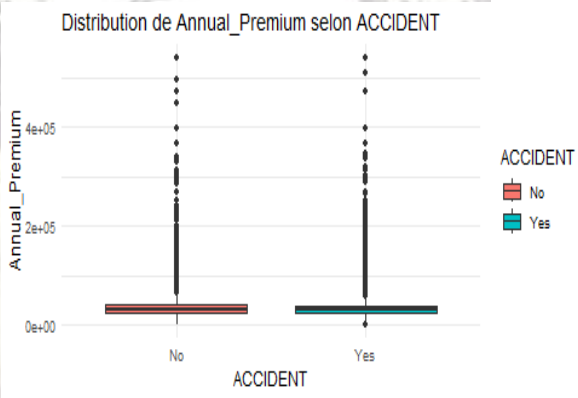


C.2) Distribution des variables quantitatives









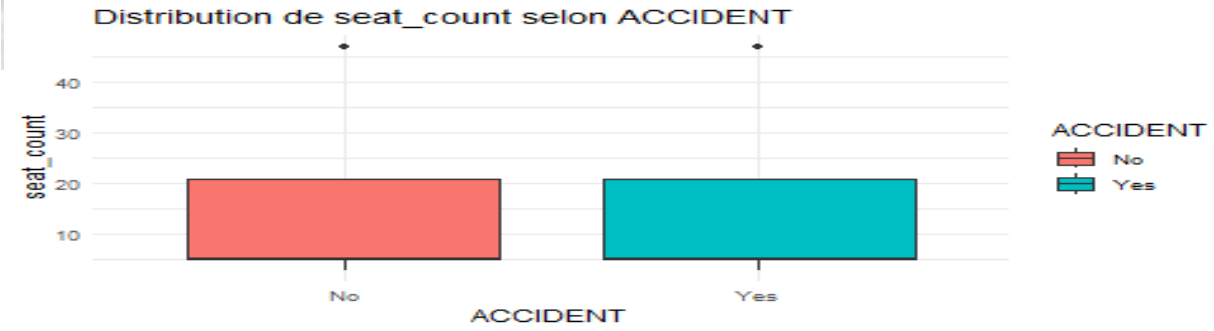


Figure 6 variable quanti

PARTIE II : MODELISATION DE L'ACCIDENT PAR RÉGRESSION LOGISTIQUE

1) RÉGRESSION LOGISTIQUE

La **régression logistique** est une méthode statistique utilisée pour modéliser une **variable dépendante binaire** (c'est-à-dire qui prend deux valeurs possibles, comme 0 ou 1, oui ou non, succès ou échec) en fonction d'une ou plusieurs **variables indépendantes**.

🔍 *Principe de base :*

Contrairement à la **régression linéaire** qui prédit des valeurs continues, la **régression logistique** prédit la **probabilité** qu'un événement se produise.

Par exemple :

Est-ce qu'un email est un spam ou non ?
Est-ce qu'un patient est malade ou pas ?
Est-ce qu'un client fera un accident ou non ?

⚠ *Formule mathématique*

La régression logistique utilise la **fonction sigmoïde (logistique)** pour transformer une valeur réelle en une probabilité entre 0 et 1 :

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

- $P(Y = 1 \mid X)$: probabilité que la sortie soit 1
- β_0 : l'ordonnée à l'origine (intercept)
- β_i : coefficients associés aux variables explicatives X_i

🔧 *En pratique :*

- **Entrée** : données avec une colonne "cible" (0 ou 1) et des variables explicatives
- **Sortie** : probabilité d'appartenir à la classe 1

2) CRÉATION DU MODÈLE DE RÉGRESSION LOGISTIQUE

```
Coefficients:
                          Estimate Std. Error  z value Pr(>|z|)
(Intercept)               2.7925215  0.0147438  189.403   <2e-16 ***
Age                      -0.0033663  0.0003701   -9.096   <2e-16 ***
SexeHomme                 0.1557593  0.0084317   18.473   <2e-16 ***
Vehicle_AgePlus de 5 ans -3.2880390  0.0144896 -226.924   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

🧠 Interprétation intelligente et synthétique

Ce modèle nous dit : quels profils sont plus ou moins susceptibles d’avoir un accident, toutes choses égales par ailleurs.

🔎 Intercept : 2.79 - Quand toutes les variables explicatives sont à zéro (ce qui veut dire "âge = 0", sexe = "Femme", véhicule de moins de 5 ans), la log-odds d'accident est **2.79**. - En pratique, l’intercept seul n’est pas très interprétable ici, mais il sert de point de départ au modèle.

📉 Age : -0.00337 (très significatif, p \< 2e-16)** - Chaque **année supplémentaire** réduit **l’odds (la cote)** d’avoir un accident d’environ **0.34%** :

e\^{-0.00337} \approx 0.9966

- Autrement dit, **les conducteurs plus âgés sont légèrement moins susceptibles d’avoir un accident**, selon le modèle.

- Effet modéré mais statistiquement très fort.

👦 3. **SexeHomme : +0.156** - Les hommes ont une **odds d'accident environ 17% plus élevée** que les femmes :

e\^{0.156} \approx 1.17

- Autrement dit, **à âge et ancienneté de véhicule constants**, les hommes sont **statistiquement plus à risque** que les femmes.

🚗 4. **Vehicle_AgePlus de 5 ans : -3.29** les véhicules de plus de 5 ans ont une **odds d'accident presque 96% plus faible** que ceux de moins de 5 ans :

$$ e\^{-3.29} \approx 0.037 $$

NB : Ce résultat est **contre-intuitif** ! Il mérite probablement une investigation

- Est-ce un biais dans les données ?
- Les véhicules plus anciens sont-ils conduits plus prudemment (propriétaires plus âgés, trajets courts) ?
- Variable mal codée ?

📊 Qualité du modèle

- **Null deviance** : 510,987 → modèle sans prédicteurs
- **Residual deviance** : 353,611 → modèle avec prédicteurs
- → Donc, **le modèle améliore bien l'ajustement**, et l'AIC (353,619) est relativement bas.

💡 Résumé Le modèle logistique révèle que les hommes ont un risque d'accident significativement plus élevé que les femmes, et que ce risque diminue légèrement avec l'âge du conducteur. Contre toute attente, les véhicules de plus de 5 ans sont associés à un risque beaucoup plus faible d'accident, un résultat qui mérite une attention particulière quant à l'interprétation (biais ou effet indirect possible). Le modèle est statistiquement très significatif et explique une part substantielle de la variabilité.

3) CALCUL DES OOD RATIO ET DES EFFETS MARGINAUX

3.1) Ordre ratio (ODD RATIO)

Estimatiion des parametres

```
PARAMETRES
     (Intercept)              Age        SexeHomme  Vehicle_AgePlus de 5 ans
     2.792521454      -0.003366282      0.155759306             -3.288039006
```

Variable	Coefficient	Effet sur la probabilité d'accident	Interprétation
(Intercept)	2.7925	Base (quand toutes les variables = 0)	Log-cote initiale relativement élevée de survenue d'un accident
Âge	-0.0034	**Diminue** légèrement	Plus l'assuré est âgé, moins il a de probabilité d'avoir un accident
Sexe : Homme	0.1558	**Augmente** légèrement	Les hommes ont une probabilité légèrement plus élevée d'avoir un accident
Âge du véhicule > 5 ans	-3.2880	**Diminue fortement**	Les véhicules plus anciens sont associés à une probabilité beaucoup plus faible

Intervalles de confiance

```
confint(modele)
                              2.5 %       97.5 %
(Intercept)               2.763661275  2.821456842
Age                      -0.004091805 -0.002641122
SexeHomme                 0.139231665  0.172283426
Vehicle_AgePlus de 5 ans -3.316475540 -3.259676443
```

Interpretation :

Variable	Coefficient	Intervalle de confiance (95 %)	Interprétation
(Intercept)	2.7925	[2.7637 ; 2.8215]	L'ordonnée à l'origine est bien estimée ; l'intervalle est étroit, donc bonne précision
Âge	-0.0034	[-0.0041 ; -0.0026]	Effet **négatif significatif** : l'âge diminue la proba d'accident, et l'intervalle ne contient pas 0
Sexe : Homme	0.1558	[0.1392 ; 0.1723]	Effet **positif significatif** : être un homme augmente la proba d'accident, intervalle exclut 0
Âge du véhicule > 5 ans	-3.2880	[-3.3165 ; -3.2597]	Effet **fortement négatif** : les véhicules anciens sont **beaucoup moins accidentogènes**

Calcul des ordres ratio

```
ODD_RATIO
   (Intercept)           Age      SexeHomme Vehicle_AgePlus de 5 ans
  16.32212344    0.99663938    1.16854491            0.03732698
```

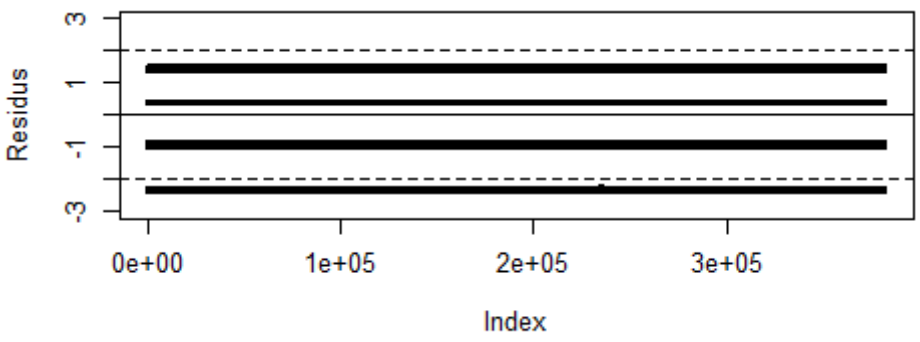
Variable	Odds Ratio (OR)	Interprétation
(Intercept)	16.32	Pour les modalités de référence (âge = 0, femme, véhicule < 5 ans), les chances d'accident sont élevées.
Âge	0.9966	Chaque année de plus **diminue légèrement** les chances d'accident (0.34 % en moins environ par an).
Sexe : Homme	1.17	Les hommes ont **16.8 % de chances en plus** d'avoir un accident comparé aux femmes.
Âge du véhicule > 5 ans	0.037	Les véhicules de plus de 5 ans ont **96.3 % de chances en moins** d'avoir un accident.

Effets marginaux

```
summary(margins_model)
                 factor     AME     SE         z      p   lower   upper
                    Age -0.0005 0.0001   -9.1007 0.0000 -0.0006 -0.0004
              SexeHomme  0.0239 0.0013   18.3920 0.0000  0.0214  0.0265
 Vehicle_AgePlus de 5 ans -0.5734 0.0017 -333.6835 0.0000 -0.5767 -0.5700
```

Variable	AME	Interprétation
Âge	-0.0005	Chaque année supplémentaire **diminue la probabilité d'accident de 0,05 point de pourcentage**.
Sexe : Homme	0.0239	Être un homme **augmente la probabilité d'accident de 2,39 points de pourcentage** par rapport aux femmes.
Âge du véhicule > 5 ans	-0.5734	Avoir un véhicule de plus de 5 ans **réduit la probabilité d'accident de 57,34 points de pourcentage**.

Analyse des residus



4) CALCULER LE TMC : taux de mauvais classement et Matrice de Confusion

Caclculer le taux de mauvais classement à partir de la matrice de confusion

```
              accident     no accident
0                9600          139281
1              156753           76520
```

5) INDICATEURS DE PERFORMANCES

```
AUC: 0.8150445
```

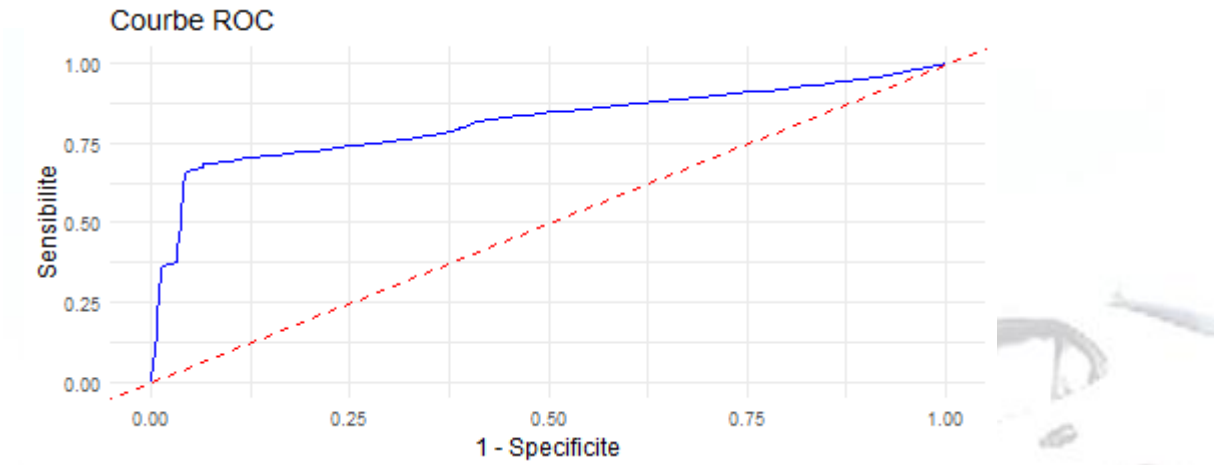
◆ Qu'est-ce que l'AUC ?

L'**AUC** (Area Under the Curve) est l'aire sous la courbe ROC. Elle mesure la **capacité du modèle àdistinguer entre deux classes** : ici, les assurés ayant eu un **accident** (classe 1) et ceux n'en ayant **pas eu** (classe 0).

🔍 Interprétation de ton AUC = 0.815

- Cela signifie que **dans 81,5 % des cas**, ton modèle **prédit correctement** qu'un individu ayant eu un accident a un score de risque supérieur à celui d'un individu n'en ayant pas eu.
- Plus précisément, si tu prends **au hasard un assuré accidenté** et **un non-accidenté**, ton modèle a **81,5 % de chances** d'assigner un **score de probabilité plus élevé** à l'accidenté.

Calculer la courbe ROC



Interpetation :

. **Compréhension des Axes** - **Axe X (1 - Spécificité)** : Représente le taux de faux positifs. Une valeur proche de 0 indique une faible probabilité de faux positifs, tandis qu'une valeur proche de 1 indique une forte probabilité. - **Axe Y (Sensibilité)** : Représente le taux de vrais positifs, c'est-à-dire la capacité du modèle à identifier correctement les positifs. Une valeur proche de 1 indique une bonne performance.

. **Interprétation de la Courbe** - **Forme de la Courbe** : La courbe commence à (0,0) et finit à (1,1). Plus la courbe est proche du coin supérieur gauche du graphique, meilleure est la performance du modèle. - **Zone sous la courbe (AUC)** : Bien que non visible sur l'image, l'AUC (Area Under the Curve) est une mesure clé. Elle varie de 0 à 1. Une AUC proche de 1 indique un excellent modèle, tandis qu'une AUC de 0.5 indique un modèle aléatoire.

. **Comparaison avec la Diagonale Rouge** - La ligne rouge diagonale représente un modèle aléatoire. Si la courbe ROC est au-dessus de cette ligne, cela signifie que le modèle a une meilleure capacité de discrimination qu'un modèle aléatoire.

. **Analyse des Points Clés** - **Points de Coupure** : Chaque point sur la courbe correspond à un seuil de classification. Il est important d'examiner ces seuils pour trouver un équilibre entre sensibilité et spécificité selon le contexte d'application. - **Performance à Différents Seuils** : Évalue comment la sensibilité et la spécificité changent avec différents seuils. Cela peut influencer le choix de seuil basé sur les coûts des faux positifs et des faux négatifs.

6) PREDICTION DE LA PROBABILITE DE FAIRE UN ACCIDENT

Affichage des parametres estimés

```
PARAMETRES
 (Intercept)           Age            SexeHomme          Vehicle_AgePlus de 5ans
   2.792521454    -0.003366282     0.155759306               -3.288039006
```

Calcul des probabilités

```
PROBABILITE_PREDITE MODALITE_PREDITE
                <dbl> <chr>
 1              0.938 accident
 2              0.361 no accident
 3              0.398 no accident
 4              0.947 accident
 5              0.938 accident
 6              0.947 accident
 7              0.361 no accident
 8              0.361 no accident
 9              0.398 no accident
10              0.361 no accident
```

Après avoir exploré la régression logistique, qui est particulièrement adaptée à la **modélisation de variables binaires**, il est naturel de s'intéresser à un autre type de modèle linéaire généralisé : la **régression de Poisson**.

Alors que la régression logistique permet d'estimer la probabilité d'occurrence d'un événement (succès/échec), la **régression de Poisson** s'intéresse à un autre type de variable dépendante : **les variables de comptage**.

En d'autres termes, lorsque la variable cible ne se limite plus à deux états (0 ou 1), mais représente un **nombre entier de fois** qu'un événement se produit (par exemple, nombre d'appels à un service client, nombre de visites sur un site, nombre de cas dans une épidémie), la régression de Poisson devient un outil statistique adapté et puissant.

PARTIE III : MODELISATION POISSONNIENE DE "frequence"

Representation graphique de la frequence

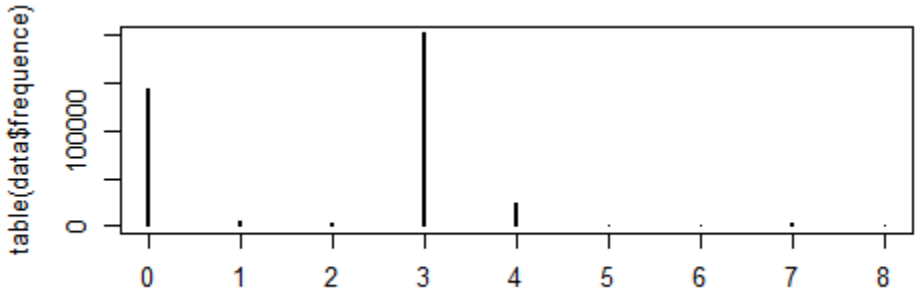


Figure 7 Frequence

la distribution des fréquences

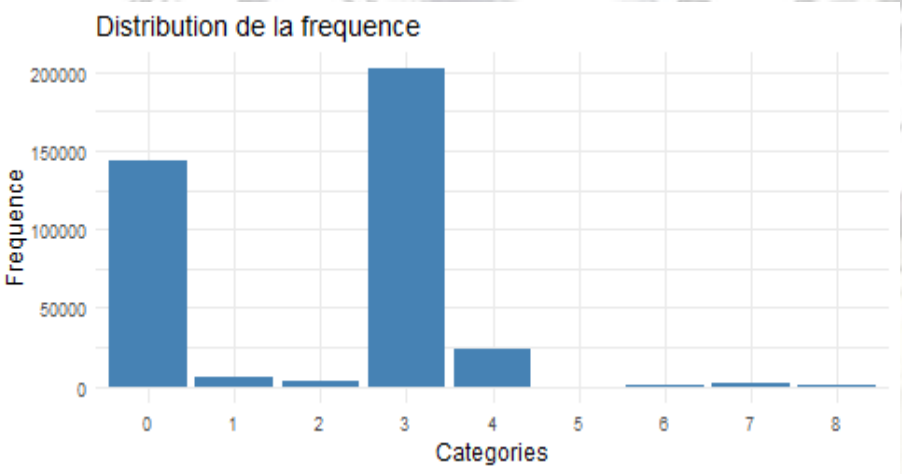
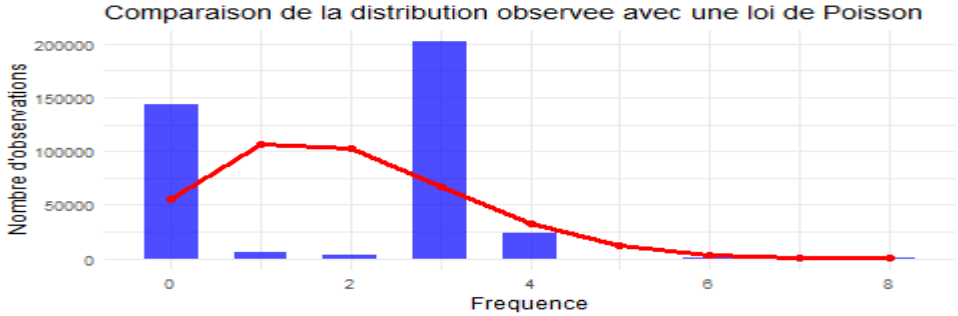


Figure 8 distribution des frequences

Calculer la moyenne empirique (λ estimé) :

```
Lambda estime : 1.934712
```



Calcul des fréquences théoriques (même n et même lambda)

```
Chi-squared test for given probabilities

data:  obs_df$obs
X-squared = 629971, df = 8, p-value < 2.2e-16
mean(data$frequence)
[1] 1.934712
> var(data$frequence)
[1] 2.516623
```

NB: "Dans ce cas, la moyenne est relativement proche de la variance, ce qui permet d'envisager l'utilisation d'un modèle de régression de Poisson. En revanche, si la variance est nettement supérieure à la moyenne, on parle alors de phénomène de surdispersion, ce qui peut rendre le modèle de Poisson inadapté."

A) 🔍 Test d'adéquation à la loi de Poisson

🧪 Hypothèses du test :

- H_0 (hypothèse nulle) : la variable suit une *loi de Poisson*
- H_1 (hypothèse alternative) : la variable **ne suit pas** une loi de Poisson.

✅ Règle de décision :

- Si la **p-value \< 0,05** : on **rejette H_0**, la variable ne suit **pas** une loi de Poisson.
- Si la **p-value ≥ 0,05** : on **ne rejette pas H_0**, on peut considérer que la variable **suit** une loi de Poisson.

A.1) Estimation des paramètres par la méthode du maximum de vraisemblance

```
> summary(fpois)
Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
       estimate  Std. Error
lambda 1.934712 0.002250033
Loglikelihood:  -721787.7   AIC:  1443577   BIC:  1443588
```

A.2) Tests d'adéquation

```
Chi-squared statistic:  623997.5
Degree of freedom of the Chi-squared distribution:  5
Chi-squared p-value:  0
Chi-squared table:
       obscounts  theocounts
<= 0 143473.0000  55208.1946
```

```
<= 1   5408.0000 106811.9717
<= 2   3530.0000 103325.2163
<= 3 202303.0000  66634.8547
<= 4  23983.0000  32229.8178
<= 7   3132.0000  17603.8560
> 7     325.0000    340.0889
Goodness-of-fit criteria
                               1-mle-pois
Akaike's Information Criterion    1443577
Bayesian Information Criterion    1443588
```

NB : p.value inférieur à 5%, les données suivent une loi de poisson

B) Construction du modèle BINOMOALE NEGATIVE

B.1) Spécification du modèle

```
summary(fnb)

Call:
glm.nb(formula = frequence ~ Age + Sexe + Vehicle_Age + fuel_type +
    manufacture_year, data = data, init.theta = 16352.57622,
    link = log)

Coefficients:
                          Estimate Std. Error  z value Pr(>|z|)
(Intercept)              4.0683591  0.5669911    7.175 7.21e-13 ***
Age                     -0.0020675  0.0001558  -13.266  < 2e-16 ***
SexeHomme                0.0373507  0.0023577   15.842  < 2e-16 ***
Vehicle_APlus de 5 ans -0.9201941  0.0044765 -205.563  < 2e-16 ***
fuel_typeSuper          -0.0243423  0.0024696   -9.857  < 2e-16 ***
manufacture_year        -0.0014609  0.0002823   -5.176 2.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for Negative Binomial(16352.58) family taken to be 1)

    Null deviance: 726760  on 382153  degrees of freedom
Residual deviance: 558676  on 382148  degrees of freedom
AIC: 1275492

Number of Fisher Scoring iterations: 1


              Theta:  16353
          Std. Err.:  3520
Warning while fitting theta: nombre limite d'iterations atteint

 2 x log-likelihood:  -1275478
```

📊 Interprétation du modèle fnb

Voici un résumé des points clés à commenter :

✅ Significativité des variables

Toutes les variables sont hautement significatives (**p < 0.001**), ce qui indique qu’elles ont un **effet statistique clair** sur la fréquence.

🔍 Interprétation des coefficients (en log)

- (Intercept) = 4.068 → C'est le log de la fréquence attendue pour une personne de référence (valeurs de base pour toutes les variables).
- Age = -0.00207 → À chaque année supplémentaire, la fréquence attendue diminue légèrement.
- SexeHomme = +0.0373 → Les hommes ont une fréquence légèrement plus élevée que les femmes.
- Vehicle_AgePlus de 5 ans = -0.92 → Les véhicules plus anciens ont une fréquence beaucoup plus faible.
- fuel_typeSuper = -0.0243 → L'utilisation de carburant "Super" est associée à une légère baisse de fréquence.
- manufacture_year = -0.00146 → Les véhicules plus récents ont une fréquence un peu plus faible.

💡 Tous les coefficients s'interprètent sur l'échelle logarithmique (log-linéaire).

⚠️ Theta élevé = surdispersion importante

- Theta ≈ 16 353 (écart-type ≈ 3520) indique une **forte dispersion**.
- Le message "nombre limite d'itérations atteint" est un **warning** : le modèle a eu du mal à estimer le paramètre de dispersion, ce qui suggère une complexité importante dans la structure des données. Ça mérite peut-être un ajustement ou une validation plus fine.

B.2) ◈AIC et ◈BIC

-AIC = Akaike Information Criterion

```
> aic_val
[1] 1275492
```

BIC = Bayesian Information Criterion

```
> bic_val
[1] 1275568
```

📜 Interprétation de tes valeurs

- L'**AIC = 1 275 492** et le **BIC = 1 275 568** sont **relativement proches**, ce qui est **bon signe** : le modèle n'est **pas surparamétré**.
- Ces valeurs **ne sont pas interprétées seules**, mais **comparées à d'autres modèles**.
  👉 Un modèle alternatif avec **AIC ou BIC plus bas** serait **préférable**.

B.3) Vérification des hypothèses du modèle

Résidus simulés pour vérifier la distribution, l'hétéroscédasticité, l'indépendance

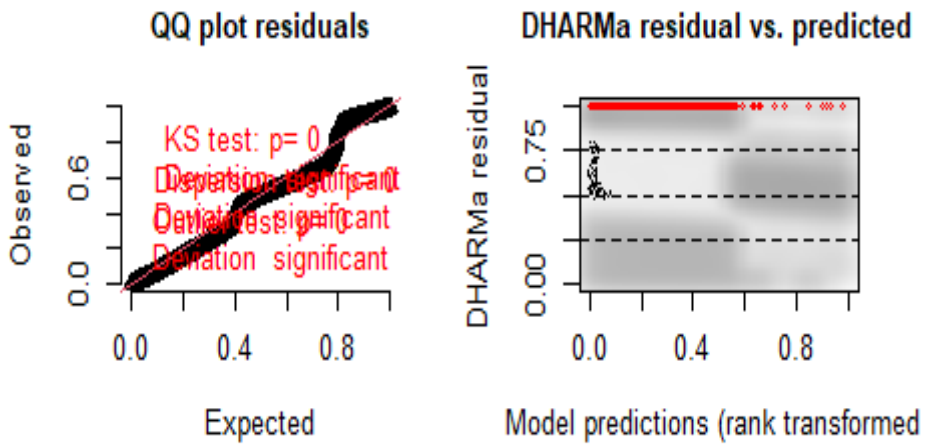


Figure 9 Residus

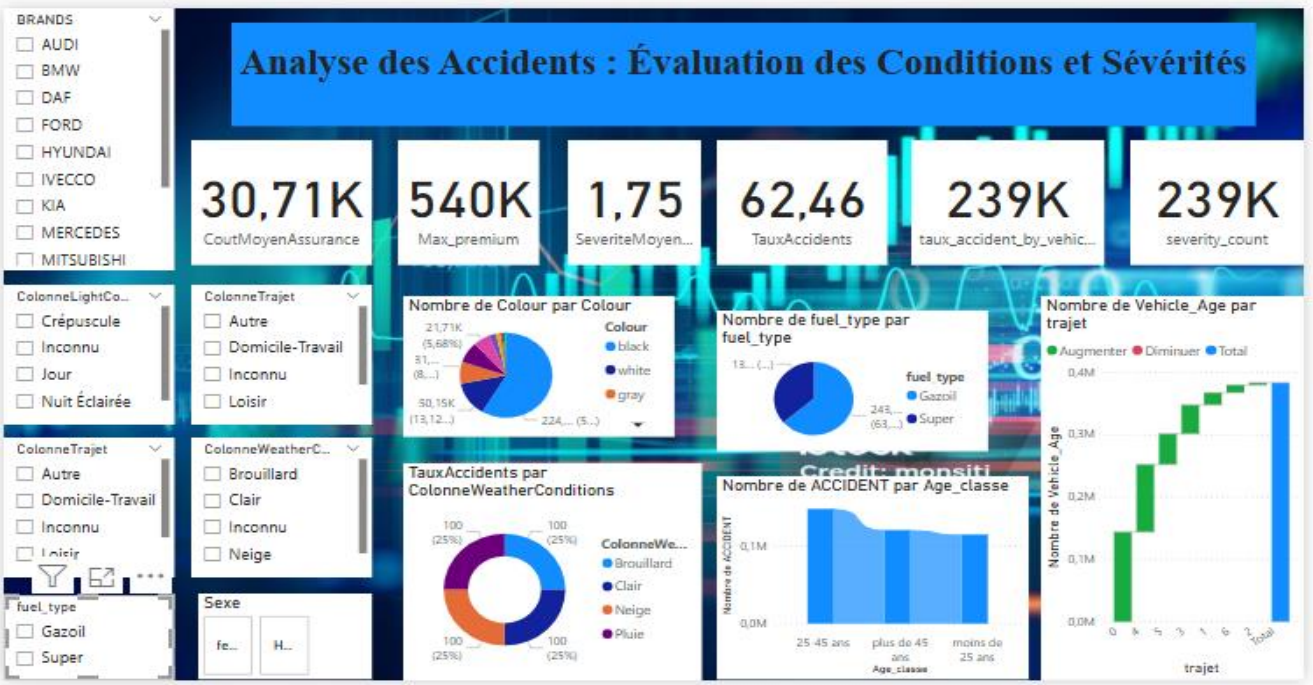
Multicolinéarité via VIF

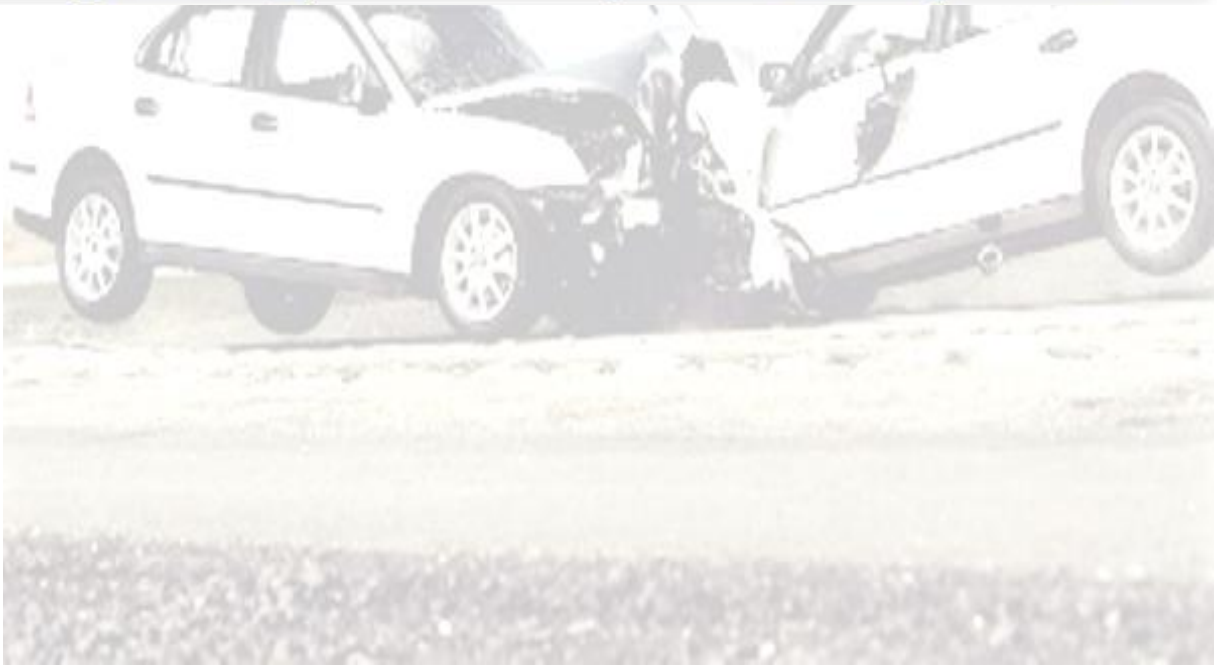
```
print(vif_vals)
     age            Sexe      Vehicle_Age        fuel_type manufacture_year
3.255026        1.026810        3.266908        1.035347        1.035339
```

🔍 Interprétation

- Toutes les valeurs sont **inférieures à 5**, ce qui signifie qu'il n'y a **pas de multicolinéarité préoccupante** dans ton modèle.
- Les variables Age et Vehicle_Age ont des VIF un peu plus élevés (~3.26), mais restent dans une zone **acceptable**. Cela peut simplement refléter un certain **recoupement logique** entre l'âge du conducteur et celui du véhicule, sans pour autant invalider le modèle.
- Les autres variables (Sexe, fuel_type, manufacture_year) ont des VIF très proches de 1 → **aucune corrélation notable** avec les autres variables.

DASHBORD


Analyse des Accidents : Évaluation des Conditions et Sévérités
30,71K
CoutMoyenAssurance
540K
Max_premium
1,75
SeveriteMoyen...
62,46
TauxAccidents
239K
taux_accident_by_vehic...
239K
severity_count
BRANDS
AUDI
BMW
DAF
FORD
HYUNDAI
IVECCO
KIA
MERCEDES
MITSUBISHI
ColonneLightCo...
Crépuscule
Inconnu
Jour
Nuit Éclairée
ColonneTrajet
Autre
Domicile-Travail
Inconnu
Loisir
ColonneWeatherC...
Brouillard
Clair
Inconnu
Neige
fuel_type
Gazoil
Super
Sexe
Nombre de Colour par Colour
Colour
black
white
gray
Nombre de fuel_type par fuel_type
fuel_type
Gazoil
Super
TauxAccidents par ColonneWeatherConditions
ColonneWe...
Brouillard
Clair
Neige
Pluie
Nombre de ACCIDENT par Age_classe
25-45 ans
plus de 45 ans
moins de 25 ans
Age_classe
Nombre de Vehicle_Age par trajet
Augmenter
Diminuer
Total
Nombre de Vehicle_Age
trajet



☑ CONCLUSION

Rappel de la problématique L'objectif principal de cette étude était de modéliser le risque d'accident automobile à partir d'un jeu de données fourni par une compagnie d'assurance IARD. L'analyse s'appuie sur un ensemble de 374 393 observations décrivant le profil des assurés et les caractéristiques de leurs véhicules. Deux axes ont été explorés : - La modélisation de la **probabilité de survenance d'un accident** à l'aide d'une **régression logistique**, après binarisation de la variable cible `fréquence`. - La modélisation du **nombre d'accidents** via une **régression de Poisson**, plus adaptée aux données de type "comptage".

Résumé des principaux résultats obtenus - **Régression logistique** :
Les variables telles que **l'âge**, le **sexe**, et l'**ancienneté du véhicule** se sont révélées significatives dans la prédiction du risque d'accident.
Par exemple : - Les **hommes** présentent une probabilité d'accident légèrement plus élevée que les femmes. - Les **véhicules plus anciens** sont associés à un **risque moindre** dans les données, ce qui peut refléter des comportements plus prudents ou d'autres facteurs non observés. - L'**âge de l'assuré** est **négativement corrélé** à la probabilité d'accident : plus l'assuré est âgé, moins il est susceptible d'avoir un accident.

- **Régression de Poisson** : L'écart modéré entre la moyenne et la variance du nombre d'accidents justifie l'usage du modèle de Poisson. Les résultats montrent que plusieurs variables explicatives influencent significativement la fréquence des sinistres.
- Les indicateurs AIC et BIC ont permis de comparer les modèles et de sélectionner les spécifications les plus efficaces tout en évitant le surajustement.
- L'**AUC** de la courbe ROC montre une bonne capacité discriminante du modèle logistique.

Recommandations - Utiliser le **modèle logistique** pour identifier les profils à risque élevé afin de cibler les campagnes de prévention ou ajuster les primes. - Utiliser le **modèle de Poisson** pour anticiper le **volume de sinistres**, ce qui peut guider les décisions actuarielles et budgétaires. - Intégrer ces modèles dans un outil de **scoring client** pour une tarification plus fine et un meilleur pilotage des risques.

Limites de l'analyse - **Binarisation** de la variable `fréquence` : une simplification qui peut entraîner une perte d'information. - Certaines variables potentiellement explicatives n'étaient pas présentes dans le jeu de données (ex : historique de conduite, bonus/malus, type de contrat). - **Surdispersion possible** dans les données de comptage qui pourrait justifier l'usage de modèles alternatifs (ex : régression négative binomiale).

Perspectives et approfondissements - Tester d'autres modèles comme : - **Régression binomiale négative** pour mieux traiter la surdispersion. - **Arbres de décision ou forêts aléatoires** pour capter d'éventuelles non-linéarités et interactions complexes. - Intégrer des **variables temporelles** (ex : saison, année) pour modéliser les tendances ou effets calendaires. - Croiser avec d'autres données externes (ex : géolocalisation, trafic routier, zones à risques) pour enrichir les modèles. - Mettre en place un **système de scoring automatisé** dans les processus de souscription ou de gestion des contrats.

Sources du code R

```
#--------------------------------#
#A-CHARGEMENT ET APERCU DES DONNEES#
#--------------------------------
# A.1. Charger le jeu de données
data <- read.csv("D:/INSSEDS/datasets/assurance_auto_makani.csv",
                 header = TRUE,
                 sep = ";") # Remplacez par le chemin correct
# A.2.Vérifier des lignes, du résumé et et de la structure du jeu de données
head(data)
summary(data)
str(data)
#--------------------------------
#B-APUREMENT DONNEES#
#--------------------------------
# B.1. Conversion de certaines variables en qualitatives
# Charger les bibliothèques nécessaires
library(dplyr)
# Recodage des variables
data <- data %>%
  mutate(
    Sexe = factor(Sexe, levels = c(0, 1), labels = c("Femme", "Homme")),
    Vehicle_Age = factor(Vehicle_Age, levels = c("< 5", "> 5"), labels = c("Moins de 5 ans", "Plus de
      5 ans")),
    transmission = factor(transmission, levels = c("man", "auto"), labels = c("Manuelle",
      "Automatique")),
    SEVERITY = as.character(SEVERITY),
    trajet = as.character(trajet),
    light_conditions = as.character(light_conditions),
    weather_conditions = as.character(weather_conditions),
    road_surface_conditions = as.character(road_surface_conditions),
    manv = as.character(manv)
```

```
 )
# Afficher la structure des données pour vérifier les modifications
str(data)
# B.2. Vérification des valeurs manquantes
sum(is.na(data))
# B.3. Visualiser le jeu de données entier avec les valeurs manquantes en pourcentage de NA
     pour chaque variable et global
# Charger les bibliothèques nécessaires
library(visdat)
library(dplyr)
# Échantillonner un sous-ensemble de données
data_sample <- data %>% slice_sample(n = 1000)  # Échantillonner 1000 lignes
# Visualiser les valeurs manquantes
vis_miss(data_sample)
# Séparer les variables quantitatives et qualitatives
library(dplyr)
quantitative_data <- data %>% select(where(is.numeric))
# B.4. Créer un boxplot pour chaque variable quantitative avant Winsorization
for (column in colnames(quantitative_data)) {
  boxplot(quantitative_data[[column]],
          main = paste("Boxplot de", column, "avant Winsorization"),
          ylab = column,
          col = "lightblue",
          border = "darkblue")
  # Ajouter une pause pour permettre de voir chaque graphique
  Sys.sleep(2) # Ajustez cette valeur pour contrôler la durée d'affichage de chaque graphique
}
# B.5. Fonction pour winsoriser une variable
winsorize <- function(x, lower_quantile = 0.01, upper_quantile = 0.99) {
  q_lower <- quantile(x, lower_quantile, na.rm = TRUE) # Ajout de na.rm = TRUE pour ignorer les
     NAs
  q_upper <- quantile(x, upper_quantile, na.rm = TRUE)
```

```
  x[x < q_lower] <- q_lower
  x[x > q_upper] <- q_upper
  return(x)
}
# B.6. Appliquer la fonction winsorize à toutes les colonnes quantitatives
quantitative_data[] <- lapply(quantitative_data, function(col) winsorize(col))
# B.7. Créer un boxplot pour chaque variable quantitative après Winsorization
for (column in colnames(quantitative_data)) {
  boxplot(quantitative_data[[column]],
          main = paste("Boxplot de", column, "après Winsorization"),
          ylab = column,
          col = "red",
          border = "darkblue")
  # Ajouter une pause pour permettre de voir chaque graphique
  Sys.sleep(2) # Ajustez cette valeur pour contrôler la durée d'affichage de chaque graphique
}
#--------------------------------
#C-STAT DESCRIPTIVE#
#--------------------------------
# C.1. Stat des quantis
# C.1.1 Calcul des statistiques descriptives pour chaque colonne
stats <- data.frame(
  mean = round(sapply(quantitative_data, mean, na.rm = TRUE), 2),
  sd = round(sapply(quantitative_data, sd, na.rm = TRUE), 2),
  min = sapply(quantitative_data, min, na.rm = TRUE),
  max = sapply(quantitative_data, max, na.rm = TRUE),
  median = sapply(quantitative_data, median, na.rm = TRUE),
  IQR = sapply(quantitative_data, function(x) IQR(x, na.rm = TRUE))
)
# Affichage du tableau formaté
library(knitr)
```

```
kable(stats)
# C.1.2. Graphiques des quantis
library(ggplot2)
library(gridExtra)
library(dplyr)
# Créer une liste de graphiques pour chaque variable quantitative
plots <- lapply(names(quantitative_data), function(var) {
  ggplot(data, aes(x = .data[[var]])) +  # Correction ici
    geom_histogram(bins = 15, fill = "skyblue", color = "black") +
    labs(title = paste("Histogramme de", var), x = var, y = "Fréquence") +
    theme_minimal()
}
# Afficher les graphiques (ajuster selon le nombre de variables)
do.call(grid.arrange, c(plots, ncol = 2))  # Organiser en 2 colonnes
# C.2. Stat des quali
library(ggplot2)
library(dplyr)
# Séparer les variables quantitatives et qualitatives (sauf ACCIDENT)
qual_vars <- data %>% select(where(~ is.factor(.) || is.character(.))) %>% select(-ACCIDENT)
# Variable cible
target_var <- "ACCIDENT"
# Pour les variables qualitatives
for (var in names(qual_vars)) {
  p <- ggplot(data, aes_string(x = var, fill = target_var)) +
    geom_bar(position = "dodge") +
    labs(title = paste("Répartition de", var, "par rapport à", target_var),
         x = var, y = "Nombre d'observations", fill = target_var) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
  print(p)
```

```
# Pour les variables quantitatives
for (var in names(quantitative_data)) {
  p <- ggplot(data, aes_string(x = target_var, y = var, fill = target_var)) +
    geom_boxplot() +
    labs(title = paste("Distribution de", var, "selon", target_var),
         x = target_var, y = var, fill = target_var) +
    theme_minimal()
  print(p)
}
# Tableau statistique pour la variable "ACCIDENT"
library(dplyr)
library(tidyr)
# Variables qualitatives
qual_vars <- data %>% select(where(~ is.factor(.) || is.character(.))) %>% select(-ACCIDENT)
cat_stat_tables <- lapply(names(qual_vars), function(var) {
  data %>%
    group_by(!!sym(var), ACCIDENT) %>%
    summarise(Effectif = n(), .groups = "drop") %>%
    group_by(!!sym(var)) %>%
    mutate(Fréquence = round(100 * Effectif / sum(Effectif), 1)) %>%
    mutate(Variable = var) %>%
    rename(Level = !!sym(var))
}) %>% bind_rows()
# Réorganisation du tableau
cat_stat_table <- cat_stat_tables %>%
  select(Variable, Level, ACCIDENT, Effectif, Fréquence)
# Variables quantitatives
quant_vars <- data %>% select(where(is.numeric))
quant_stat_table <- quantitative_data %>%
  mutate(ACCIDENT = data$ACCIDENT) %>%
  pivot_longer(cols = -ACCIDENT, names_to = "Variable", values_to = "Value") %>%
```

```
  group_by(Variable, ACCIDENT) %>%
  summarise(
    Moyenne = round(mean(Value, na.rm = TRUE), 2),
    Médiane = round(median(Value, na.rm = TRUE), 2),
    Écart_type = round(sd(Value, na.rm = TRUE), 2),
    Min = round(min(Value, na.rm = TRUE), 2),
    Max = round(max(Value, na.rm = TRUE), 2),
    .groups = "drop"
  )
library(knitr)
library(kableExtra)
# Variables qualitatives
kable(cat_stat_table, caption = "Tableau de contingence des variables qualitatives avec
    ACCIDENT") %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))
# Variables quantitatives
kable(quant_stat_table, caption = "Statistiques descriptives des variables quantitatives selon
    ACCIDENT") %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover")
# Transformer les données pour les rendre compatibles avec facet_wrap
data_long <- data %>%
  select(all_of(names(qual_vars)), all_of(target_var)) %>%
  pivot_longer(cols = -all_of(target_var), names_to = "variable", values_to = "value"
# Créer le plot avec facet_wrap
p <- ggplot(data_long, aes(x = value, fill = !!sym(target_var))) +
  geom_bar(position = "dodge") +
  labs(title = "Répartition des variables qualitatives par rapport à ACCIDENT",
      x = "Valeurs des variables", y = "Nombre d'observations", fill = "ACCIDENT") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_wrap(~ variable, scales = "free_x")
```

```
print(p)
# Transformer les données pour les rendre compatibles avec facet_wrap
data_long <- data %>%
  select(all_of(names(quantitative_data)), all_of(target_var)) %>%
  pivot_longer(cols = -all_of(target_var), names_to = "variable", values_to = "value"
# Créer le plot avec facet_wrap
p <- ggplot(data_long, aes(x = !!sym(target_var), y = value, fill = !!sym(target_var))) +
  geom_boxplot() +
  labs(title = "Distribution des variables quantitatives selon ACCIDENT",
       x = "ACCIDENT", y = "Valeurs des variables", fill = "ACCIDENT") +
  theme_minimal() +
  facet_wrap(~ variable, scales = "free_y")

print(p)
#-------------------------------
#D-MODELISATION DE "accident"
#-------------------------------
#  1. Binarisation de la variable cible fréquence
data$accident <- ifelse(data$frequence >= 2, 1, 0)
data$accident <- as.factor(data$accident)
str(data$accident)
# 2. Création du modèle de régression logistique
modele <- glm(accident ~ Age + Sexe + Vehicle_Age, data = data, family = binomial)
summary(modele)


# CALCUL DES OOD RATIO ET DES EFFETS MARGINAUX


# affichage des paramètres estimés
PARAMETRES = coefficients(modele)
PARAMETRES
```

```
# Calculons les intervalles de confiance
confint(modele)

# calcul des odd ratio
ODD_RATIO = exp(coefficients(modele))
ODD_RATIO

# Calcul des effets marginaux
library("margins")
margins_model <- margins(modele)
summary(margins_model)

# Analyse des résidus
res.m <- rstudent(modele)
plot(res.m,pch=15,cex=.5,ylab="Residus",ylim=c(-3,3))
abline(h=c(-2,0,2),lty=c(2,1,2))

# CALCULER LE TMC : taux de mauvais classement et Matrice de Confusion

# calcul des probabilités
data$PROBABILITE_PREDITE <- predict(modele, data,
                                    type="response")

# Transformer les probabilité en modalite predites
data$MODALITE_PREDITE <- ifelse(data$PROBABILITE_PREDITE < 0.5, "no accident" , "accident")
```

```
# Caclculer le taux de mauvais classement à partir de la matrice de confusion
Matrice_confusion = table(data$accident, data$MODALITE_PREDITE)
Matrice_confusion

#L'AUC

library(pROC)
library(ggplot2)

# Calculer la courbe ROC
roc_curve <- roc(data$accident, data$PROBABILITE_PREDITE)

# Calculer l'AUC
auc_value <- auc(roc_curve)
cat("AUC:", auc_value, "\n")

# Créer un dataframe pour ggplot
roc_df <- data.frame(
  specificity = 1 - roc_curve$specificities,
  sensitivity = roc_curve$sensitivities
)

# Visualiser la courbe ROC
ggplot(roc_df, aes(x = specificity, y = sensitivity)) +
  geom_line(color = "blue") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Courbe ROC", x = "1 - Spécificité", y = "Sensibilité") +
  theme_minimal()
```

```
# PREDICTION DE LA PROBABILITE DE FAIRE UN ACCIDENT


# affichage des paramètres estimés
PARAMETRES = coefficients(modele)
PARAMETRES


# calcul des probabilités sous R
data$PROBABILITE_PREDITE <- predict(modele, data,
                            type="response")
head(data)
#----------------------------------------
#E-MODELISATION POISSONNIENE DE "frequence"
#----------------------------------------
# Faisons la représentation graphique
plot(table(data$frequence))
library(ggplot2)
# Créer un tableau de fréquences
freq_table <- table(data$frequence)

# Convertir le tableau de fréquences en data.frame pour ggplot
freq_df <- as.data.frame(freq_table)

# Visualiser la distribution des fréquences avec ggplot
ggplot(freq_df, aes(x = Var1, y = Freq)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Distribution de la fréquence", x = "Catégories", y = "Fréquence") +
  theme_minimal()
```

```
#1.Calculer la moyenne empirique (λ estimé) :
lambda_hat <- mean(data$frequence)
cat("Lambda estimé :", lambda_hat)
#2. réer un tableau de fréquence observée :

observed_freq <- table(data$frequence)
#3. créer la fréquence théorique sous hypothèse de loi de Poisson :

# Transformer le tableau en data.frame pour faciliter la manipulation
obs_df <- as.data.frame(observed_freq)
colnames(obs_df) <- c("frequence", "obs")

# Convertir la variable frequence en entier
obs_df$frequence <- as.numeric(as.character(obs_df$frequence))

# Calcul des fréquences théoriques (même n et même lambda)
n_total <- sum(obs_df$obs)
obs_df$theo <- dpois(obs_df$frequence, lambda_hat) * n_total
#4. isualiser la comparaison graphique :
library(ggplot2)

ggplot(obs_df, aes(x = frequence)) +
  geom_bar(aes(y = obs), stat = "identity", fill = "blue", alpha = 0.7, width = 0.6) +
  geom_line(aes(y = theo), color = "red", size = 1.2) +
  geom_point(aes(y = theo), color = "red", size = 2) +
  labs(title = "Comparaison de la distribution observée avec une loi de Poisson",
       x = "Fréquence",
       y = "Nombre d'observations") +
  theme_minimal()

chisq.test(x = obs_df$obs, p = obs_df$theo / sum(obs_df$theo))
```

```
mean(data$frequence)
var(data$frequence)

#Ici la moyenne n'est pas très loin de la variance, on peut utiliser la régression
#de poisson. Si par contre la Variance est très supérieur à la moyenne :
#ce phénomène est appelé «sur-dispersion».

#Test d'adéquation à la loi de poisson
# H0 :la distribution ne suit pas la loi X
# H1 :la distribution suit la loi X
#Si p-value < 0.05, on rejette H0.
#Si p-value > 0.05, on ne peut rejeter H0.
# chargement du package
library(fitdistrplus)
# test d'adéquation avec une distribution de poisson
#1. Estimation des paramètres par la méthode du maximum de vraisemblance
fpois <- fitdist(data$frequence, "pois")
summary(fpois)
#2. Tests d'adéquation
fpois <- fitdist(data$frequence, "pois")
gofstat(fpois)

#NB : p.value inférieur à 5%, les données suivent une loi de poisson

# II.Construction du modèle BINOMOALE NEGATIVE

library(MASS)
```

```
# II.1.Spécification du modèle


fnb <- glm.nb(frequence ~ Age+Sexe+Vehicle_Age+
              fuel_type+ manufacture_year, data = data)
summary(fnb)  # Résumé du modèle ajusté


# II.2. AIC et BIC


aic_val <- AIC(fnb)
aic_val


bic_val <- BIC(fnb)
bic_val
# II.2.Vérification des hypothèses du modèle
# Chargement des packages nécessaires
library(MASS)      # pour glm.nb
library(car)       # pour vif
library(DHARMa)    # pour les résidus simulés
# 1. Résidus simulés pour vérifier la distribution, l'hétéroscédasticité, l'indépendance
sim_res <- simulateResiduals(fittedModel = fnb)
plot(sim_res)  # Génère 4 graphiques de vérification


# 2. Multicolinéarité via VIF
vif_vals <- vif(fnb)
print(vif_vals)
```