

Accidents routiers en France

Rendu 1: rapport d'exploration, de data visualisation et de pre-processing des données

Table des matières

0. Contexte.....	3
1. Objectifs.....	3
2. Présentation du jeu de données.....	3
3. Visualisations et Statistique	5
3.1 Data-frame : Véhicules.....	5
3.1.1 Exploration univariée et bivariée (avec variable cible).....	6
3.1.2 Analyses statistiques.....	12
3.1.3 Conclusions	13
3.2 Data-frame : Caractéristiques	13
3.2.1 Exploration univariée et bivariée (avec variable cible).....	14
3.2.2 Analyses statistiques.....	21
3.2.3 Conclusions	22
3.3 Data-frame : Usager.....	22
3.3. 1 Exploration univariée et bivariée (avec variable cible).....	23
3.3.2 Analyses statistique	27
3.3.3 Conclusions	28
3.4 Data-frame : Lieux	29
3.4.1 Exploration univariée et bivariée (avec variable cible).....	30
3.4.2 Analyses statistique	39
4.1.3 Conclusions	41
4. Pre-processing et feature engineering.....	41
4.1 Nettoyage du data-frame initial : Suppression des erreurs.....	41
4.2 Traitement des valeurs manquantes (Nan).....	42
4.3 Création de variables supplémentaires, et regroupement de catégories	43
4.3.1 Data-frame : Véhicules	43
4.3.2 Data-frame : Caractéristiques	46
4.3.3 Data-frame : Usager.....	52
4.3.4 Data-frame : Lieux	53
4.4 Gestion des doublons dans le data frame	53
4.5 Encodage des variables catégorielles.....	53
4.6 Data-frame final.....	54

Introduction au projet

0. Contexte

Du point de vue scientifique, ce projet offre une opportunité d'utiliser différentes techniques de Data Science, de tester divers algorithmes de Machine Learning et d'évaluer leur performance. Il permettra de mieux comprendre les facteurs influençant la gravité des accidents routiers. De plus, ce projet permettra d'approfondir les connaissances dans les problèmes de prédiction et de classification en général, et spécifiquement dans le domaine de la sécurité routière.

Du point de vue économique, les résultats de ce projet peuvent aider à réduire les coûts associés aux accidents routiers en France. En développant un modèle prédictif de la gravité des accidents, le projet permettra de prendre des mesures préventives plus efficaces. Cela permettra de réduire les dépenses liées aux soins médicaux, aux réparations de véhicules et optimiser les dépenses du budget dans son ensemble.

1. Objectifs

L'objectif principal de ce projet est de développer un modèle prédictif précis qui peut estimer la gravité des accidents routiers en France se basant sur les données historiques.

En utilisant les données historiques sur la période de 2018 - 2021, nous allons nettoyer et préparer les données, extraire les caractéristiques les plus pertinentes, puis créer un modèle prédictif. Nous allons tester différents modèles et méthodes d'évaluation pour trouver l'approche la plus performante. Le modèle final sera entraîné sur les données historiques.

L'objectif ultime est de fournir des informations pour pouvoir prendre des mesures de prévention et d'intervention ciblées, contribuant ainsi à la réduction du nombre d'accidents graves sur les routes françaises.

Compréhension et manipulation des données

2. Présentation du jeu de données

Le jeu de données dans le cadre de ce projet est une base de données d'accidents corporels de la circulation routière disponible sur site data.gouv.fr ([web](#)). Cette base est constituée annuellement de plusieurs fichiers. Les données utilisées dans le projet portent sur la période 2018-2021 uniquement. Ces données sont produites par le ministère de l'intérieur, et sont disponibles librement (licence ouverte).

Le jeu de données est réparti en quatres rubriques sous forme de fichiers au format csv :

- CARACTERISTIQUES qui décrit les circonstances générales de l'accident
- LIEUX qui décrit le lieu principal de l'accident
- VEHICULES impliqués
- USAGERS impliqués

Voici l'information sur la volumétrie du dataset (l'information est présentée par fichier, par année et pour la période étudiée de 2018 à 2021) :

Lieux	Caractéristiques	véhicules	usagers
2018f (57783, 18)	2018f (57783, 16)	2018f (98876, 9)	2018f (130169, 12)
2019 (58840, 18)	2019 (58840, 15)	2019 (100710, 11)	2019 (132977, 15)
2020 (47744, 18)	2020 (47744, 15)	2020 (81066, 11)	2020 (105295, 15)
2021 (56518, 18)	2021 (56518, 15)	2021 (97315, 11)	2021 (129153, 15)
Total 2018-2021 : 220885	Total 2018-2021 : 220885	Total 2018-2021 : 377967	Total 2018-2021 : 497594

On peut constater que chaque rubrique du dataset a un nombre de lignes différent. Cette variation est due au fait que chaque rubrique contient des informations spécifiques sur les accidents. Pour les rubriques "Lieux" et "Caractéristiques", chaque ligne représente un accident unique, tandis que les colonnes contiennent des détails sur les circonstances générales, et le lieu de l'accident. Par conséquent, le nombre de lignes est le même pour ces deux rubriques.

Cependant, la taille des rubriques "Véhicules" et "Usagers" varie. Dans la rubrique "Véhicules", les informations sont présentées pour chaque voiture impliquée dans un accident. Chaque ligne correspond à chaque accident, et chaque voiture impliquée. En revanche, dans la rubrique "Usagers", les informations sont fournies pour chaque accident, chaque voiture impliquée et chaque usager associé à cet accident. Ainsi, plusieurs lignes peuvent être associées à un même accident, chacune représentant un usager différent ou une voiture différente (fig. 2.1).

The diagram illustrates a dataset row with 5 lines. A green bracket on the left groups the first 5 lines under '1 accident (5 lignes)'. A green bracket on the right groups the first 5 lines under '5 usagers (2 conducteurs, 3 passagers) pour deux voitures'. An orange bracket at the bottom groups the last two lines under '2 voitures'.

Num_Acc	place	catu	grav	sexe	trajet	secu	locp	actp	etatp	an_nais	num_veh	id_vehicule	secu1	secu2	secu3
73	201800000037	1.0	1	1	1	5.0	3.0	0.0	0.0	2000.0	A01		NaN	NaN	NaN
74	201800000037	2.0	2	1	1	5.0	3.0	0.0	0.0	2002.0	A01		NaN	NaN	NaN
75	201800000037	1.0	1	4	1	5.0	3.0	0.0	0.0	1968.0	B01		NaN	NaN	NaN
76	201800000037	2.0	2	3	2	5.0	11.0	0.0	0.0	1971.0	B01		NaN	NaN	NaN
77	201800000037	4.0	2	3	2	5.0	11.0	0.0	0.0	1993.0	B01		NaN	NaN	NaN

Figure 2.1 – Exemple de la présentation d'un accident dans le dataset

Pour regrouper toutes les rubriques en un seul jeu de données, nous avons utilisé les variables suivantes : Num_acc - identifiant unique de l'accident, num_veh - identifiant du véhicule et id_veh - identifiant unique du véhicule (ajouté à partir de 2019).

Voici la volumétrie du dataset final (l'information est présentée par année et pour la période utilisée dans le projet 2018 – 2021) :

2018f (130169, 51)
2019 (132977, 54)
2020 (105295, 54)
2021 (129153, 54)
Total 2018-2021 : (497594, 57)

La variable cible : grav qui est représenté la gravité de blessure de l'usager. les usagers accidentés sont classés en trois catégories 1 – Indemné , 2 – Tué, 3 – Blessé, 4 – Blessé léger.

Dans ce rapport, nous avons choisi d'étudier les quatres parties du jeu de données final de manière séparée. Cela nous a permis de répartir le travail au sein de l'équipe et d'analyser les variables par groupes plus similaires. Les variables de chaque partie correspondent aux rubriques initiales (Véhicules, Caractéristiques, Usagers et Lieux) et ont été analysées en détail. Pour étudier l'impact des variables explicatives sur la variable cible (grav), la variable cible (grav) a été ajoutée à chaque partie de l'analyse. Une fois les analyses individuelles terminées, nous avons regroupé et synthétisé les résultats pour l'ensemble du jeu de données.

3. Visualisations et Statistique

La chapitre 3 se concentre à l'exploration et à l'analyse du data-frame. Comme expliqué dans le paragraphe précédent les résultats sont présentés en 4 parties. Nous aborderons les data-frames "Véhicules", "Caractéristiques", "Usager" et "Lieux" de manière individuelle afin de mieux comprendre les informations qu'ils contiennent.

Dans la première partie de cette exploration, nous réaliserons une analyse univariée et bivariée (avec la variable cible) afin de mettre en évidence les caractéristiques et les relations entre les variables. Cette analyse nous permettra d'obtenir une vision détaillée de chaque data-set, en identifiant les tendances, les distributions et les corrélations potentielles.

Dans la deuxième partie l'analyse statistique est effectuée. Deux tests statistiques sont utilisés : le test du Chi2 et le test V de Cramer. Le test du Chi2 est utilisé pour conclure sur la dépendance statistique entre les variables catégorielles, tandis que le test V de Cramer est utilisé pour mesurer la force de cette dépendance. En utilisant ces tests, nous pouvons évaluer la relation statistique entre les variables catégorielles et la variable cible. Cela renforce la fiabilité de nos conclusions formulées dans la première partie.

Dans la troisième partie, les conclusions sur les variables sont présentées. Nous déterminerons quelles variables sont pertinentes pour la modélisation, et quels paramètres doivent être pris en compte.

Cette approche nous permettra de sélectionner les variables les plus significatives et de définir les paramètres clés pour notre modèle prédictif.

3.1 Data-frame : Véhicules

Dans cette partie nous allons étudier le data-frame « Véhicules ». L'objectif est de comprendre et d'analyser les caractéristiques et les relations entre les variables afin de sélectionner les variables les plus pertinentes pour l'étape de la modélisation. Cette sélection se fonde sur trois aspects : la pertinence métier, l'exploration bivariée et la dépendance statistique avec la variable cible. En combinant ces aspects, on peut choisir les paramètres les plus significatifs pour prédire la variable cible avec précision.

```
Total dataset pour la période 2018–2021 : (497594, 12)

Type de variable : qualitatives: 10, quantitatives : 1

- Num_Acc:----- variable qualitative
- senc:----- variable catégorielle (valeurs discrètes et non ordonnées)_(NB=4)
- id_vehicule:----- variable qualitative
- Num_Veh:----- variable qualitative (identifiant unique et non numérique)
- catv:----- variable catégorielle (valeurs discrètes et non ordonnées)
- obs:----- variable catégorielle (valeurs discrètes et non ordonnées)
- obsm:----- variable catégorielle (valeurs discrètes et non ordonnées)
- choc:----- variable catégorielle (valeurs discrètes et non ordonnées)
- manv:----- variable catégorielle (valeurs discrètes et non ordonnées)
- motor:----- variable catégorielle (valeurs discrètes et non ordonnées)
- occutc:----- variable quantitative (valeurs entières positives)
- grav:----- variable cible/cat

• Num_Acc - Identifiant de l'accident identique\ 
• senc - Sens de circulation
  • ○ ○ -1 – Non renseigné 0 – Inconnu 1 – PK ou PR ou numéro d'adresse postale croissant 2 – PK ou PR ou numéro d'adresse postale décroissant 3 – Absence de repère
• id_vehicule - Identifiant unique du véhicule\ 
• Num_Veh - Identifiant du véhicule\ 
• catv - Catégorie du véhicule :
  • ○ ○ 00 – Indéterminable 01 – Bicyclette etc ... 50 – EDP à moteur 60 – EDP sans moteur 80 – VAE 99 – Autre véhicule
• obs - Obstacle fixe heurté :
  • ○ ○ -1 – Non renseigné 0 – Sans objet 1 – Véhicule en stationnement 2 – Arbre 3 – Glissière métallique 4 – Glissière béton 5 – Autre glissière 6 – Bâtiment, mur, pile de pont 7 – Support de signalisation verticale ou poste d'appel d'urgence 8 – Poteau 9 – Mobilier urbain 10 – Parapet 11 – Ilot, refuge, borne haute 12 – Bordure de trottoir 13 – Fossé, talus, paroi rocheuse 14 – Autre obstacle fixe sur chaussée 15 – Autre obstacle fixe sur trottoir ou accotement 16 – Sortie de chaussée sans obstacle 17 – Buse – tête d'aqueduc
• obsm - Obstacle mobile heurté :
  • ○ ○ -1 – Non renseigné 0 – Aucun 1 – Piéton 2 – Véhicule 4 – Véhicule sur rail 5 – Animal domestique 6 – Animal sauvage 9 – Autre
• choc - Point de choc initial
  • ○ ○ 1 - Avant 2 – Avant droit 3 – Avant gauche 4 – Arrière 5 – Arrière droit 6 – Arrière gauche 7 – Côté droit 8 – Côté gauche 9 – Chocs multiples (tonneaux)
```

- **manv** - Manœuvre principale avant l'accident :
 - ○ -1 – Non renseigné 0 – Inconnue 1 – Sans changement de direction 2 – Même sens, même file 3 – Entre 2 files 4 – En marche arrière 5 – A contresens 6 – En franchissant le terre-plein central 7 – Dans le couloir bus, dans le même sens 8 – Dans le couloir bus, dans le sens inverse 9 – En s'insérant 10 – En faisant demi-tour sur la chaussée **Changeant de file** 11 – A gauche 12 – A droite **Déporté** 13 – A gauche 14 – A droite **Tournant** 15 – A gauche 16 – A droite **Dépassant** 17 – A gauche 18 – A droite **Divers** 19 – Traversant la chaussée 20 – Manœuvre de stationnement 21 – Manœuvre d'évitement 22 – Ouverture de porte 23 – Arrêté (hors stationnement) 24 – En stationnement (avec occupants) 25 – **Circulant sur trottoir** 26 – **Autres manœuvres** 25,26 depuis 2019
- **motor** - Type de motorisation du véhicule :
 - ○ -1 – Non renseigné 0 – Inconnue 1 – Hydrocarbures 2 – Hybride électrique 3 – Electrique 4 – Hydrogène 5 – Humaine 6 – Autre
- **occutc** - Nombre d'occupants dans le transport en commun.

Les variables **Num_acc**, **id_veh** et **Num_veh** seront utilisées dans la partie Pre-processing pour créer un nouveau paramètre : la quantité de voitures pour chaque accident.

En revanche, les variables **senc**, **motor** et **occutc** sont considérées comme non pertinentes d'un point de vue métier et ne seront pas prises en compte.

Les variables **obs**, **obsm**, **choc**, **manv** et **catv** seront étudiées dans les prochains paragraphes.

3.1.1 Exploration univariée et bivariée (avec variable cible)

Dans ce paragraphe, nous allons examiner de près les caractéristiques de nos variables, à la fois individuellement (exploration univariée) et en relation avec la variable cible (exploration bivariée). Cette analyse approfondie nous permettra de mieux comprendre les distributions, les tendances et les interactions entre les variables, et d'identifier les variables les plus significatives pour prédire la gravité.

Dans le cadre de l'analyse bivariée, nous souhaitons étudier l'impact de chaque variable sur la variable cible, qui représente le niveau de gravité des accidents routiers. Afin de réaliser cette analyse, nous avons effectué une modification de la variable cible en regroupant les catégories de la manière suivante :

- La catégorie "Indemne" (1) et "Blessé léger" (4) a été regroupée en une seule catégorie (0) correspondant aux accidents considérés comme non graves.
- La catégorie "Tué" (2) et "Blessé hospitalisé" (3) a été regroupée en une seule catégorie (1) correspondant aux accidents considérés comme graves.

Cette modification nous permet de simplifier l'analyse en créant une variable cible binaire, où 0 représente les accidents non graves et 1 représente les accidents graves.

a. Variable **obs** – obstacle fixe heurte

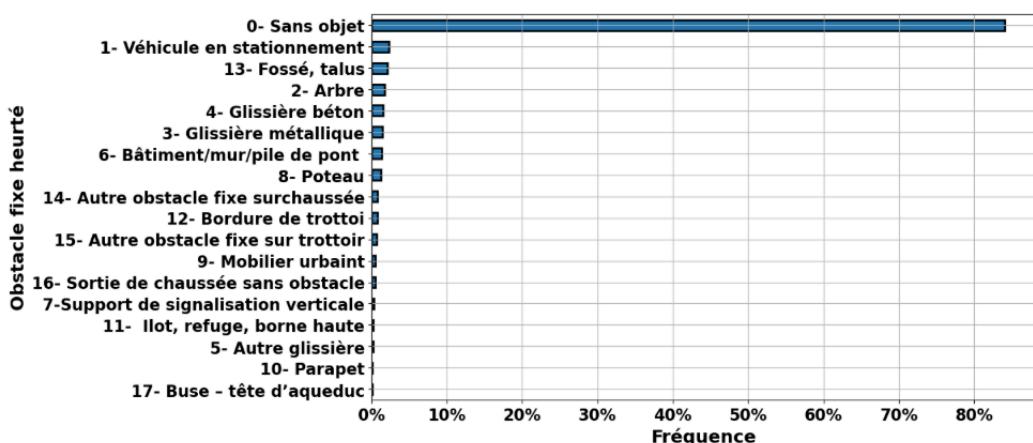
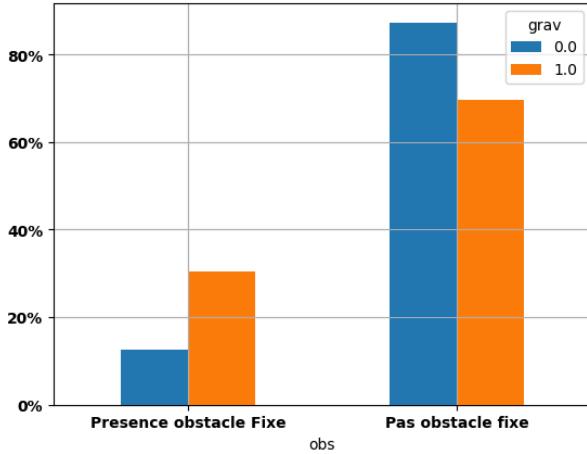


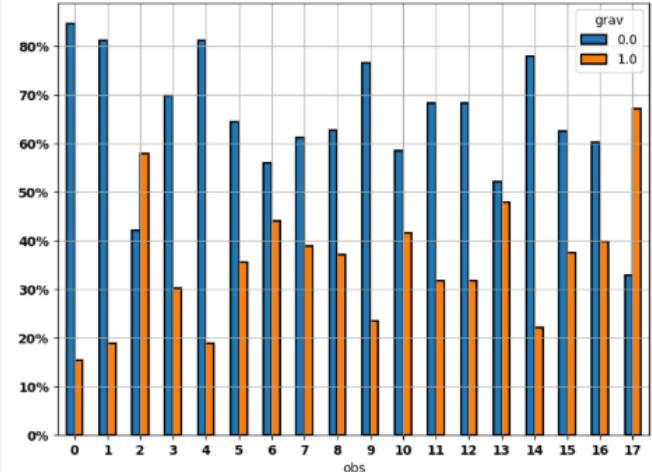
Figure 3.1.1 Fréquence des différents types d'obstacles fixes heurtés lors des accidents

- La figure 3.1.1 présente la distribution des différents types d'obstacles fixes heurtés lors des accidents de la route. On peut constater que les obstacles les plus courants sont les véhicules en stationnement, les fossés ou les talus, les arbres et les glissières en béton.
- La majorité des accidents (plus de 85 %) ne déclarent pas la présence d'obstacles fixes, ce qui suggère qu'ils sont causés par d'autres facteurs.
- Seulement 15 % des accidents déclarent la présence d'obstacles fixes.

Malgré la faible présence d'obstacles (15%) dans l'ensemble des accidents, ces accidents (avec obstacle fixe) représentent 30% de tous les cas graves, comme illustré dans la figure 3.1.2 (a). Le graphique 3.1.2 (a) représente la distribution des accidents graves et non-graves en fonction de la présence ou non d'un obstacle fixe. On peut remarquer que parmi les accidents graves (représentés en orange), environ 30% ont impliqué la présence d'un obstacle fixe, ce qui souligne l'importance de prendre en compte ce paramètre dans l'analyse des accidents de la route.



a - Distribution des accidents graves et non-graves en fonction de la présence obstacle fixe



b - Distribution des accidents graves et non-graves pour chaque type obstacle fixe

Figure 3.1.2

La figure 3.1.2 (b) présente la distribution des accidents graves et non-graves pour chaque type obstacle fixe. On peut observer que les obstacles "2 - Arbre" et "17 - Buse - tête d'aqueduc" représentent presque 60% et 70% des cas graves, respectivement, par rapport à 42% et 30% des cas non graves pour ces types d'obstacles. Cela suggère que ces deux types d'obstacles ont plus de chances de conduire à des accidents graves.

b. Variable obsm – obstacle mobile heurté

La figure 3.1.3 présente la distribution des différents types d'obstacles mobiles heurtés lors des accidents de la route. On peut constater que les obstacles les plus courants sont les véhicules (70%) et piétons (15%). Dans 18% des accidents, aucun obstacle mobile n'est présent.

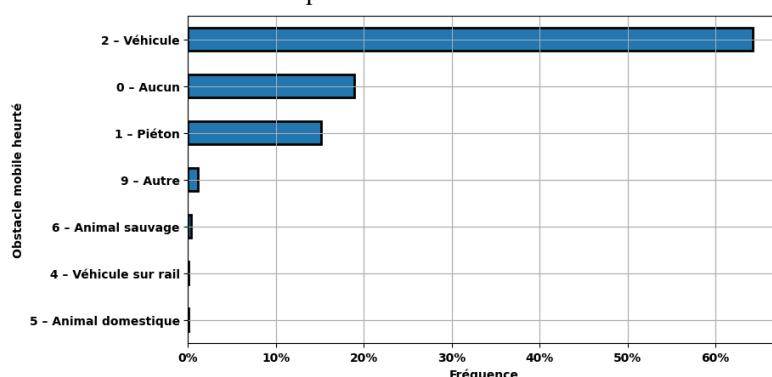


Figure 3.1.3 – Distribution obstacle mobile

La figure 3.1.4 illustre la répartition des accidents graves et non-graves en fonction de la présence ou de l'absence d'un obstacle mobile. L'analyse de cette figure révèle que parmi les accidents graves (représentés en orange), environ 65% ont impliqué la présence d'un obstacle mobile. Cette observation souligne l'importance de prendre en compte ce paramètre dans l'analyse des accidents de la route, car il semble jouer un rôle significatif dans la gravité des accidents.

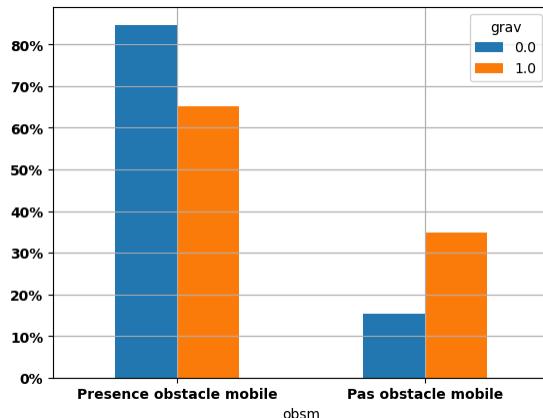


Figure 3.1.4 - Répartition des accidents graves et non-graves en présence d'un obstacle

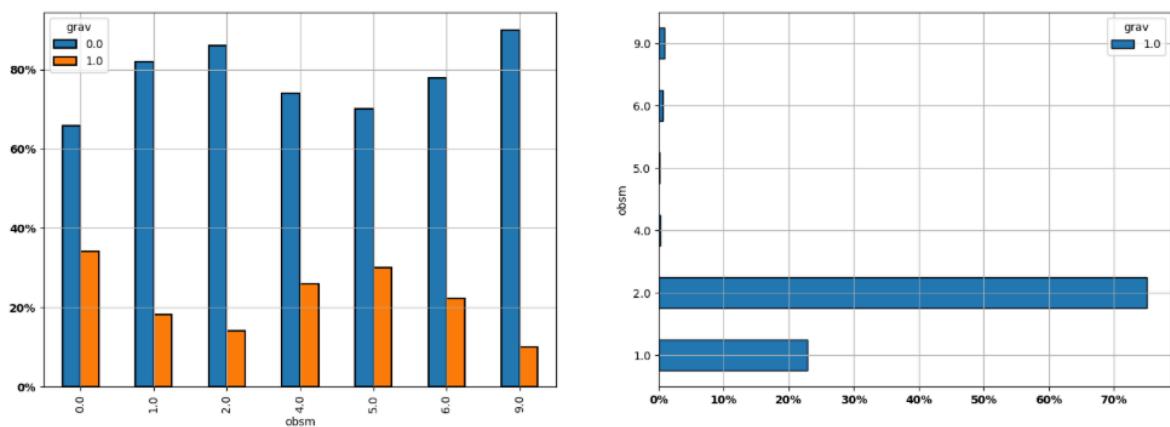


Figure 3.1.5

La figure 3.1.5 (à gauche) présente la distribution des accidents graves et non-graves pour chaque type d'obstacle mobile. On constate que les catégories "5 - animal domestique" (30% de cas graves) et "4 - véhicule sur rail" (27% de cas graves) sont associées à un plus grand nombre de cas graves. La figure 3.1.5 (à droite) présente la répartition des types d'obstacles mobiles parmi tous les cas graves. On observe que parmi tous les accidents graves impliquant un obstacle mobile, 75% sont des collisions avec des voitures et 23% sont des collisions avec des piétons.

c. Variable choc – point de choc initial

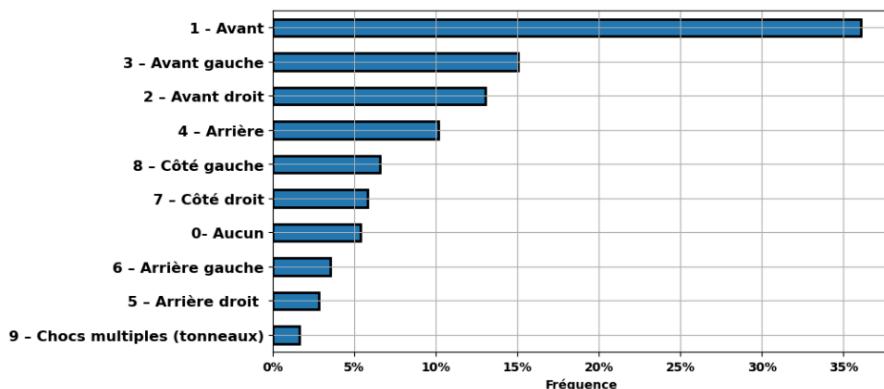


Figure 3.1.6 - Répartition des différents types de points de choc initial

La figure 3.1.6 illustre la répartition des différents types de points de choc initial lors des accidents de la route. On observe que dans 62% des cas, le point de choc initial est situé à l'avant du véhicule, ce qui inclut les chocs

avant frontal, avant gauche et avant droit. Cela indique que la majorité des accidents de la route ont lieu lors d'une collision frontale ou d'une collision frontale oblique.

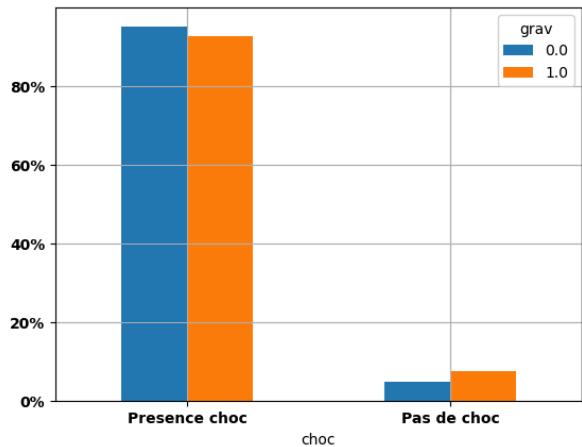


Figure 3.1.7

La figure 3.1.7 représente la répartition des accidents graves et non-graves en fonction de la présence ou de l'absence d'un choc initial. En analysant cette figure, nous pouvons constater qu'environ 92% des accidents graves (représentés en orange) ont impliqué la présence d'un choc initial. Cette observation met en évidence l'importance de prendre en compte ce paramètre dans l'analyse des accidents de la route, car il semble jouer un rôle significatif dans la gravité des accidents.

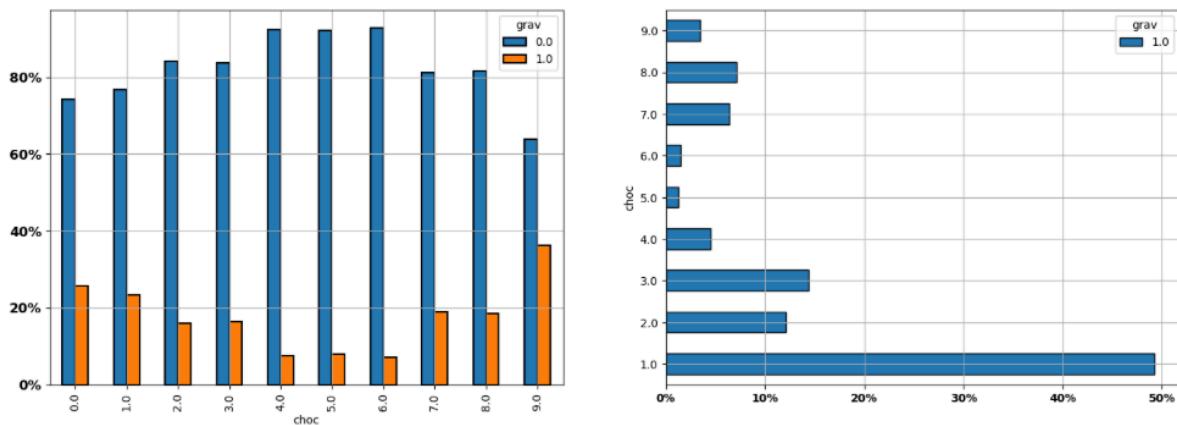


Figure 3.1.8

La figure 3.1.8 (à gauche) montre la distribution des accidents graves et non-graves selon le type de choc initial. On remarque que la catégorie 9 (chocs multiples) présente un pourcentage plus élevé de cas graves, environ 40%.

La figure 3.1.8 (à droite) présente la répartition des types de chocs parmi tous les cas graves. On constate que parmi les accidents graves avec un choc initial, environ 75% d'entre eux sont caractérisés par un point de choc initial situé à l'avant du véhicule, ce qui inclut les chocs avant frontal, avant gauche et avant droit.

d. Variable manv – Manœuvre principale avant l'accident

La figure 3.1.9 présente la répartition des différents types de manœuvres principales avant les accidents de la route. On observe que les manœuvres principales les plus courantes sont : 1 - sans changement de direction (45%), 2 - même sens, même file (11%) et 15 - tournant à gauche (8%). De plus, dans 6% de tous les accidents, le type de manœuvre principale est inconnu.

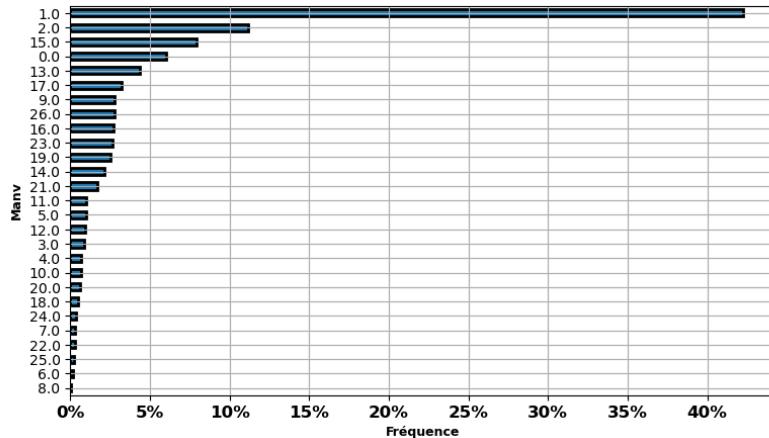


Figure 3.1.9

- **manv - Manœuvre principale avant l'accident :**
 - □ ○ -1 – Non renseigné 0 – Inconnue 1 – Sans changement de direction 2 – Même sens, même file 3 – Entre 2 files 4 – En marche arrière 5 – A contresens 6 – En franchissant le terre-plein central 7 – Dans le couloir bus, dans le même sens 8 – Dans le couloir bus, dans le sens inverse 9 – En s'insérant 10 – En faisant demi-tour sur la chaussée **Changeant de file** 11 – A gauche 12 – A droite **Déporté** 13 – A gauche 14 – A droite **Tournant** 15 – A gauche 16 – A droite **Dépassant** 17 – A gauche 18 – A droite **Divers** 19 – Traversant la chaussée 20 – Manœuvre de stationnement 21 – Manœuvre d'évitement 22 – Ouverture de porte 23 – Arrêté (hors stationnement) 24 – En stationnement (avec occupants) 25 – **Circulant sur trottoir** 26 – **Autres manœuvres** 25,26 depuis 2019

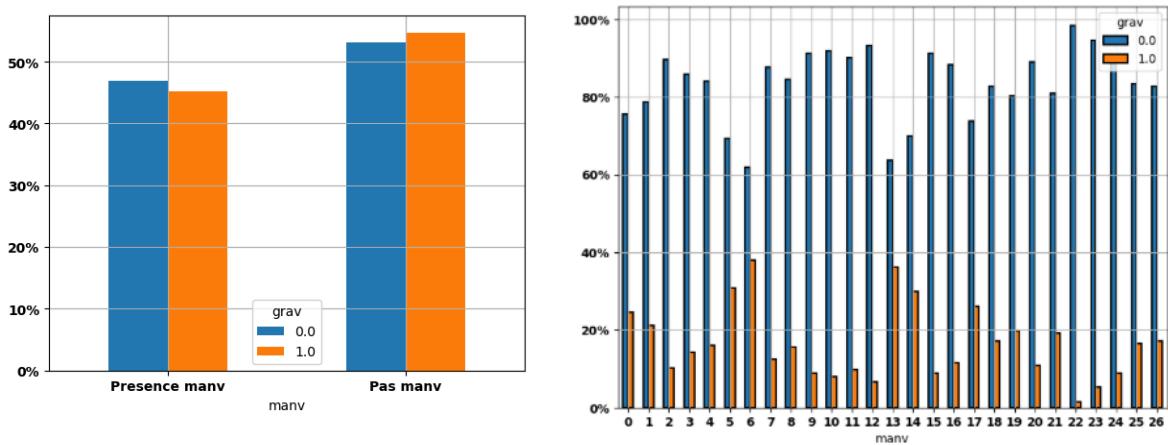


Figure 3.1.10

La figure 3.1.10 (à gauche) représente la répartition des accidents graves et non-graves en fonction de la présence ou de l'absence d'un manœuvres principales. Les catégories 1- (sans changement de direction) et 2 (même sens, même file) ont été regroupées pour définir le terme « pas de manv ». En analysant cette figure, nous pouvons constater qu'environ 45% des accidents graves (représentés en orange) ont impliqué la présence d'un manœuvre principale.

La figure 3.1.10 (à droite) présente la répartition des accidents graves et non-graves en fonction du type de manœuvres principales. Une observation importante est que les manœuvres : 5 - à contresens, 6 - en franchissant le terre-plein central, 13 - déporté à gauche et 14 - déporté à droite sont plus dangereuses et peuvent entraîner un plus grand nombre de cas graves par rapport aux autres manœuvres.

e. Variable catv – Catégorie du véhicules

La figure 3.1.11 présente la répartition des différentes catégories de véhicules impliqués dans les accidents. On constate que les catégories de véhicules les plus fréquemment rencontrées sont les suivantes : 7 - VL (véhicule léger) avec 55% des cas, 33 - motocyclette avec 7% des cas, 7 - VU (véhicule utilitaire) avec 7% des cas, et la catégorie de bicyclette avec 5% des cas.

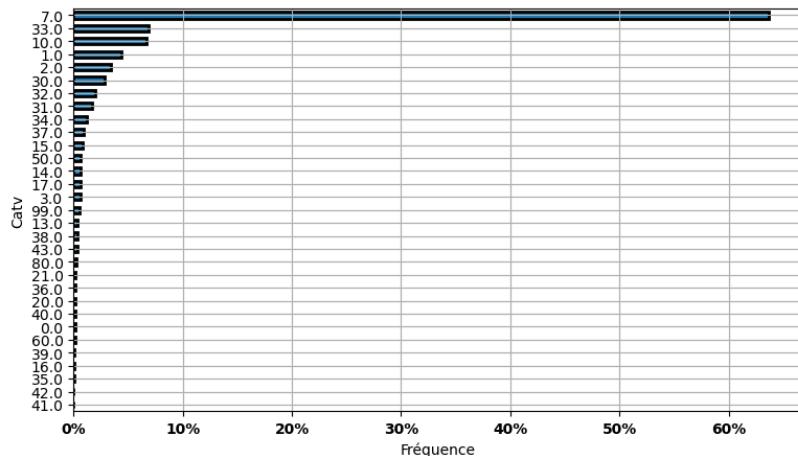


Figure 3.1.11

Catégorie du véhicule :

00 – Indéterminable 01 – Bicyclette 02 – Cyclomoteur <50cm3 03 – Voiturette (Quadricycle à moteur carrossé) (anciennement "voiturette ou tricycle à moteur") 04 – Référence inutilisée depuis 2006 (scooter immatriculé) 05 – Référence inutilisée depuis 2006 (motocyclette) 06 – Référence inutilisée depuis 2006 (side-car) 07 – VL seul 08 – Référence inutilisée depuis 2006 (VL + caravane) 09 – Référence inutilisée depuis 2006 (VL + remorque) 10 – VU seul 1,5T <= PTAC <= 3,5T avec ou sans remorque (anciennement VU seul 1,5T <= PTAC <= 3,5T) 11 – Référence inutilisée depuis 2006 (VU (10) + caravane) 12 – Référence inutilisée depuis 2006 (VU (10) + remorque) 13 – PL seul 3,5T <PTCA <= 7,5T 14 – PL seul > 7,5T 15 – PL > 3,5T + remorque 16 – Tracteur routier seul 17 – Tracteur routier + semi-remorque 18 – Référence inutilisée depuis 2006 (transport en commun) 19 – Référence inutilisée depuis 2006 (tramway) 20 – Engin spécial 21 – Tracteur agricole 30 – Scooter < 50 cm3 31 – Motocyclette > 50 cm3 et <= 125 cm3 32 – Scooter > 50 cm3 et <= 125 cm3 33 – Motocyclette > 125 cm3 34 – Scooter > 125 cm3 35 – Quad léger <= 50 cm3 (Quadricycle à moteur non carrossé) 36 – Quad lourd > 50 cm3 (Quadricycle à moteur non carrossé) 37 – Autobus 38 – Autocar 39 – Train 40 – Tramway 41 – 3RM <= 50 cm3 42 – 3RM > 50 cm3 <= 125 cm3 43 – 3RM > 125 cm3 50 – EDP à moteur 60 – EDP sans moteur 80 – VAE99 – Autre véhicule.

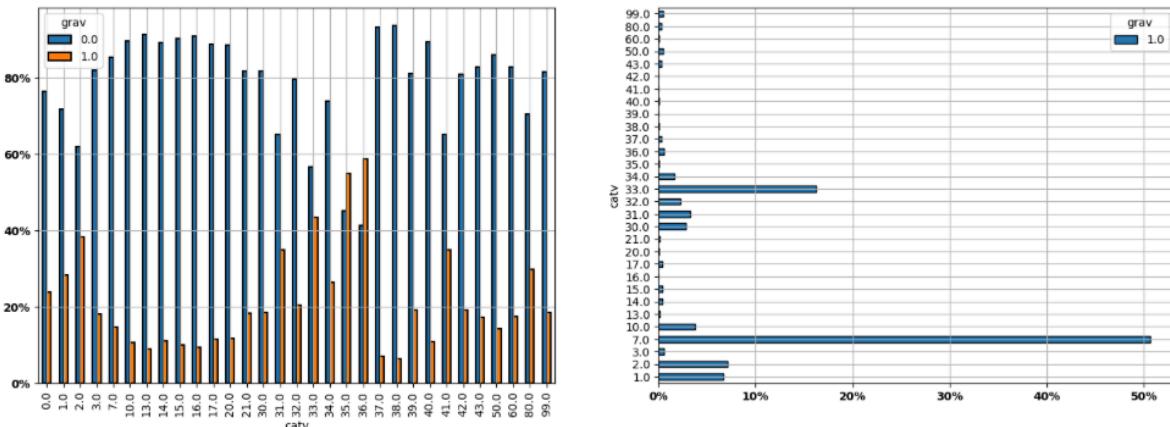


Figure 3.1.12

La figure 3.1.12 (à gauche) présente la répartition des accidents graves et non-graves en fonction des catégories de véhicules impliqués dans les accidents. On observe que les catégories 35 - "Quad léger <= 50 cm3" et 36 - "Quad lourd > 50 cm3" représentent respectivement près de 57% et 60% des cas graves, par rapport à 43% et 40% des cas non graves pour ces types de véhicules. Ces résultats montrent que ces deux catégories de véhicules sont plus susceptibles d'être associées à des accidents graves.

La figure 3.1.8 (à droite) présente la répartition de chaque type de véhicule parmi tous les cas graves. On observe que les deux types de véhicules les plus fréquemment impliqués dans tous les accidents graves : 7 - VL (véhicule léger) et 33 - Motocyclette > 125 cm3, représentant respectivement 50% et 15% de tous les cas graves.

3.1.2 Analyses statistiques

Dans cette section, nous approfondissons notre analyse en utilisant des tests statistiques pour confirmer les observations lors de l'exploration univariée et bivariée avec la variable cible.

a. Test chi2

```
senc   : p-value=1.1214882108888122e-73
catv   : p-value=0.0
obs    : p-value=0.0
obsm   : p-value=0.0
choc   : p-value=0.0
manv   : p-value=0.0
num_veh: p-value=0.0
motor   : p-value=0.0
grav    : p-value=0.0
                                              , correlation=correlated , chi2 = 367.6
                                              , correlation=correlated , chi2 = 88103.7
                                              , correlation=correlated , chi2 = 36988.0
                                              , correlation=correlated , chi2 = 29322.7
                                              , correlation=correlated , chi2 = 22814.2
                                              , correlation=correlated , chi2 = 40028.6
                                              , correlation=correlated , chi2 = 13715.2
                                              , correlation=correlated , chi2 = 9987.5
                                              , correlation=correlated , chi2 = 1492599.0
```

Les résultats du test du Chi2 révèlent des valeurs de p très faibles (p-value=0.0) pour toutes les variables examinées, indiquant une dépendance statistiquement significative entre ces variables et la variable cible (gravité des accidents). Ces résultats rejettent l'hypothèse nulle selon laquelle il n'y a aucune association entre ces variables.

b. Test V Cramer

La figure 3.1.2 présente une heatmap illustrant les résultats du test V de Cramer, qui mesure la force de l'association entre les variables catégorielles et la variable cible (gravité des accidents). Les couleurs dans la heatmap indiquent le degré de corrélation, avec des valeurs significatives observées pour les variables catv, obs, obsm, choc, manv et grav. Cela confirme une dépendance statistique entre ces variables et la gravité des accidents.

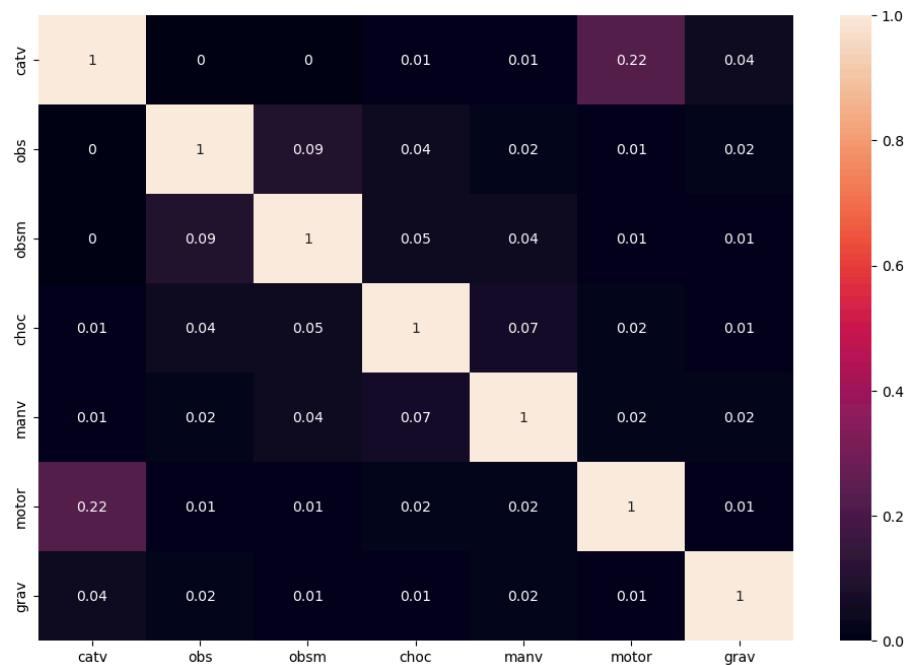


Figure 3.1.2 - Heatmap V-Cramer

3.1.3 Conclusions

Le tableau 3.1.3 synthétise l'analyse des variables du dataframe "Véhicules". Les critères de pertinence métier, l'exploration bivariée et la dépendance statistique avec la variable cible ont été utilisés pour sélectionner les variables les plus pertinentes. Seules les variables qui ont été jugées pertinentes selon ces trois critères ont été conservées pour la prochaine étape de modélisation.

		Pertinence selon métier	Pertinence de la variable concernant la variable cible (observation graphique)	Test statistiques dépendances variable explicative et la cible	Décision
[manv]	Manœuvre principale avant l'accident	OK	OK	OK	à garder
[obs]	Obstacle fixe heurté	OK	OK	OK	à garder
[obsm]	Obstacle mobile heurté	OK	OK	OK	à garder
[choc]	Point de choc initial	OK	OK	OK	à garder
[catv]	Catégorie du véhicule	OK	OK	OK	à garder
[Num_acc]	Identifiant de l'accident identique	-	-	-	potentiellement utiles pour créer de nouvelles variables. -> recommandons de les supprimer après
[id_vehicule]	Identifiant unique du véhicule (depuis 2019)	-	-	-	
[num_veh]	Identifiant du véhicule	-	-	-	
[sens]	Sens de circulation	NOK	NOK	OK	à supprimer
[occutc]	Nombre d'occupants dans le transport en commun	NOK	NOK	OK	à supprimer
[motor]	Type de motorisation du véhicule	NOK	NOK	OK	à supprimer

NOK	pas pertinente
OK	pertinente
-	sans signification

Tableau 3.1.3

3.2 Data-frame : Caractéristiques

Dans cette section, notre focus sera sur l'analyse du data-frame "Caractéristiques". Notre but est d'appréhender et d'examiner les caractéristiques et les interactions entre les variables, afin de déterminer celles qui sont les plus pertinentes pour le processus de modélisation. Le choix de ces variables se base sur trois critères : leur pertinence dans le contexte métier, l'analyse bivariée, et leur corrélation statistique avec la variable cible. En intégrant ces facteurs, nous serons en mesure de sélectionner les paramètres les plus importants pour prédire avec précision la variable d'intérêt.

```
Total dataset pour la periode 2018-2021 : (497594, 17)
```

```
Type de variable : qualitatives: 17, quantitatives : 0
```

```
- Num_Acc:----- variable nominale (catégorielle)
- an:----- variable ordinaire (traitée comme une variable catégorielle)
- mois:----- variable ordinaire (traitée comme une variable catégorielle)
- jour:----- variable ordinaire (traitée comme une variable catégorielle)
- hrmn:----- variable ordinaire (traitée comme une variable catégorielle)
- lum:----- variable nominale (catégorielle)
- agg:----- variable nominale (catégorielle)
- int:----- variable nominale (catégorielle)
- atm:----- variable nominale (catégorielle)
- col:----- variable nominale (catégorielle)
- com:----- variable nominale (catégorielle)
- adr:----- variable nominale (catégorielle)
- grav:----- variable cible/nominale
```

Num_Acc : Numéro d'identifiant de l'accident.

jour : Jour de l'accident.

mois : Mois de l'accident.

An : Année de l'accident.

Hrmn : Heure et minutes de l'accident.

Lum : Lumière : conditions d'éclairage dans lesquelles l'accident s'est produit : 1 – Plein jour, 2 – Crépuscule ou aube, 3 – Nuit sans éclairage public, 4 – Nuit avec éclairage public non allumé

5 – Nuit avec éclairage public allumé

Dep : Département : Code INSEE (Institut National de la Statistique et des Etudes Economiques) du département (2A Corse-du-Sud – 2B Haute-Corse).

Com : Commune : Le numéro de commune est un code donné par l'INSEE. Le code est composé du code INSEE du département suivi par 3 chiffres.

Agg : Localisation : 1 – Hors agglomération, 2 – En agglomération

Int : Intersection : 1 – Hors intersection, 2 – Intersection en X, 3 – Intersection en T, 4 – Intersection en Y, 5 – Intersection à plus de 4 branches, 6 – Giratoire, 7 – Place, 8 – Passage à niveau, 9 – Autre intersection

Atm : Conditions atmosphériques : -1 – Non renseigné, 1 – Normale, 2 – Pluie légère, 3 – Pluie forte

4 – Neige – grêle, 5 – Brouillard – fumée, 6 – Vent fort – tempête, 7 – Temps éblouissant, 8 – Temps couvert, 9 – Autre

Col Type de collision : -1 – Non renseigné, 1 – Deux véhicules – frontale, 2 – Deux véhicules – par l'arrière, 3 – Deux véhicules – par le côté, 4 – Trois véhicules et plus – en chaîne, 5 – Trois véhicules et plus - collisions multiples, 6 – Autre collision, 7 – Sans collision

Adr : Adresse postale : variable renseignée pour les accidents survenus en agglomération.

Lat, Long : Latitude et Longitude

- Les variables "jours", "mois" et "an" seront exploitées dans la section de prétraitement des données afin de générer plusieurs nouvelles variables :
 - La variable "date" : cette variable permet de connaître la date de l'accident. Elle sera utilisée exclusivement pour la réalisation de diagrammes et ne sera pas intégrée dans notre modèle.
 - La variable "jour de la semaine" : cette variable permet de déterminer le jour de la semaine auquel l'accident s'est produit. Elle sera remplacée par la variable suivante :
 - "week-end" : cette variable indique si l'accident s'est produit pendant le week-end ou non.
- La variable "Hrmn" sera utilisée pour créer la variable "période", qui permet de connaître la période de la journée de l'accident.
- En revanche, les variables "An", "jour", "hrmn", "com", "adr", "gps", "dep" sont considérées comme non pertinentes du point de vue métier et ne seront donc pas prises en compte.
- Les variables "mois", "lum", "agg", "int", "atm" et "col" seront examinées dans les paragraphes suivants.

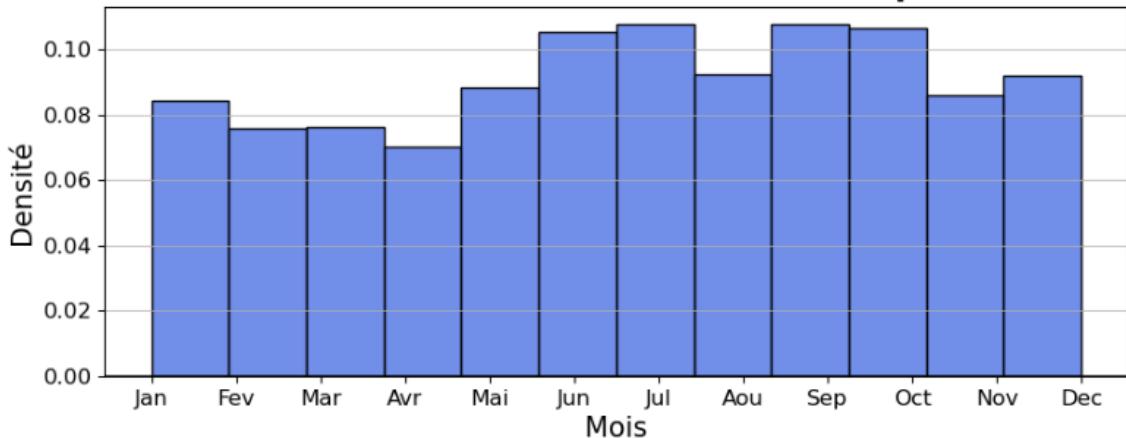
3.2.1 Exploration univariée et bivariée (avec variable cible)

Dans le présent paragraphe, nous nous pencherons minutieusement sur les caractéristiques de nos variables, en les examinant tant individuellement (à travers une exploration univariée) qu'en relation avec la variable cible (via une exploration bivariée). Cette analyse en profondeur nous permettra de mieux saisir les distributions, les tendances et les interactions entre les variables, et d'identifier celles qui sont les plus déterminantes pour prédire la gravité des accidents.

Au cours de l'analyse bivariée, notre objectif est d'étudier l'impact de chaque variable sur la variable cible, cette dernière représentant le degré de gravité des accidents de la route.

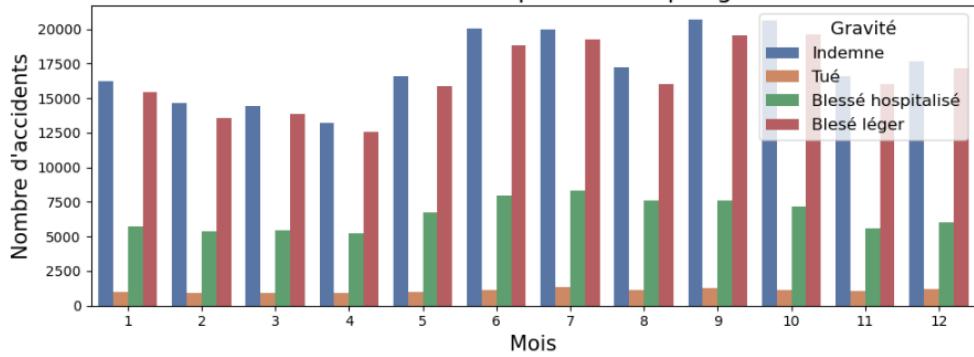
a. Variable mois

Distribution normalisée des accidents par mois

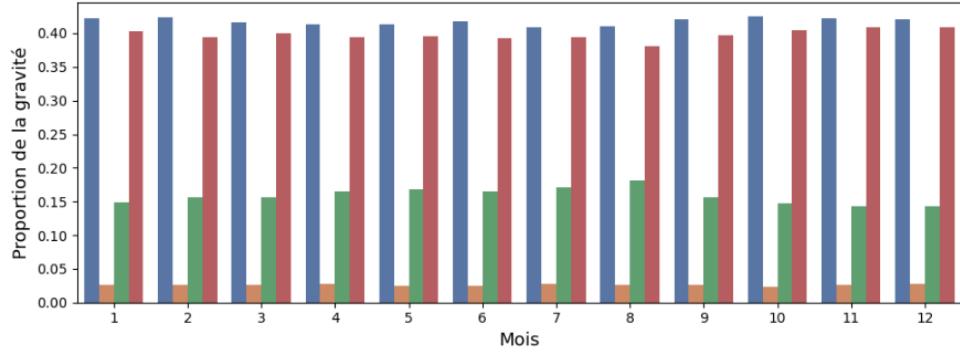


- La figure illustre la distribution normalisée du nombre total d'accidents par mois.
- On observe que le mois d'avril enregistre le pourcentage le plus faible d'accidents, avec 6,41% du total.
- Par contre, septembre se démarque comme le mois le plus accidentogène, avec 9,86% du total d'accidents.
- Il est important de noter que les mois de mars et d'avril ont été spécifiquement impactés en 2020 en raison de la pandémie de COVID-19.

Nombre d'accidents par mois et par gravité

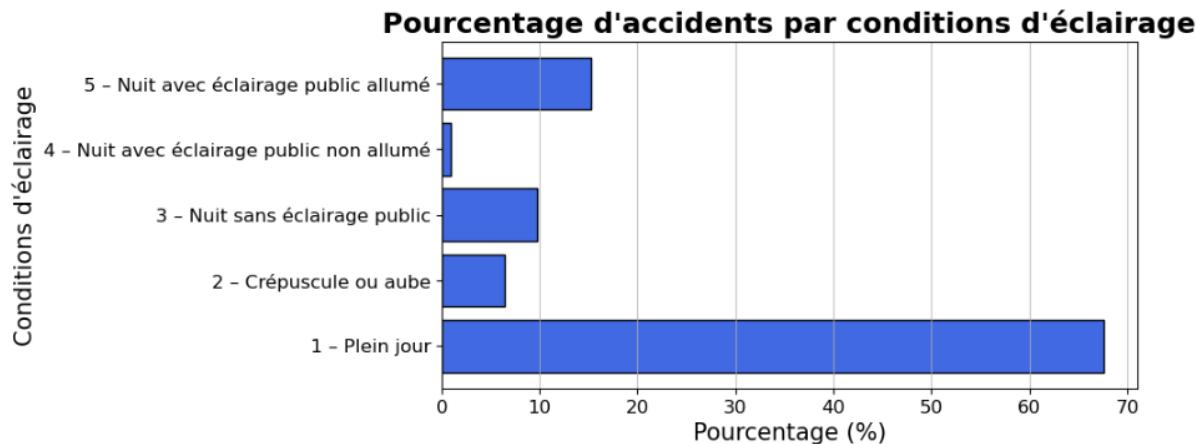


Proportion d'accidents par mois et gravité

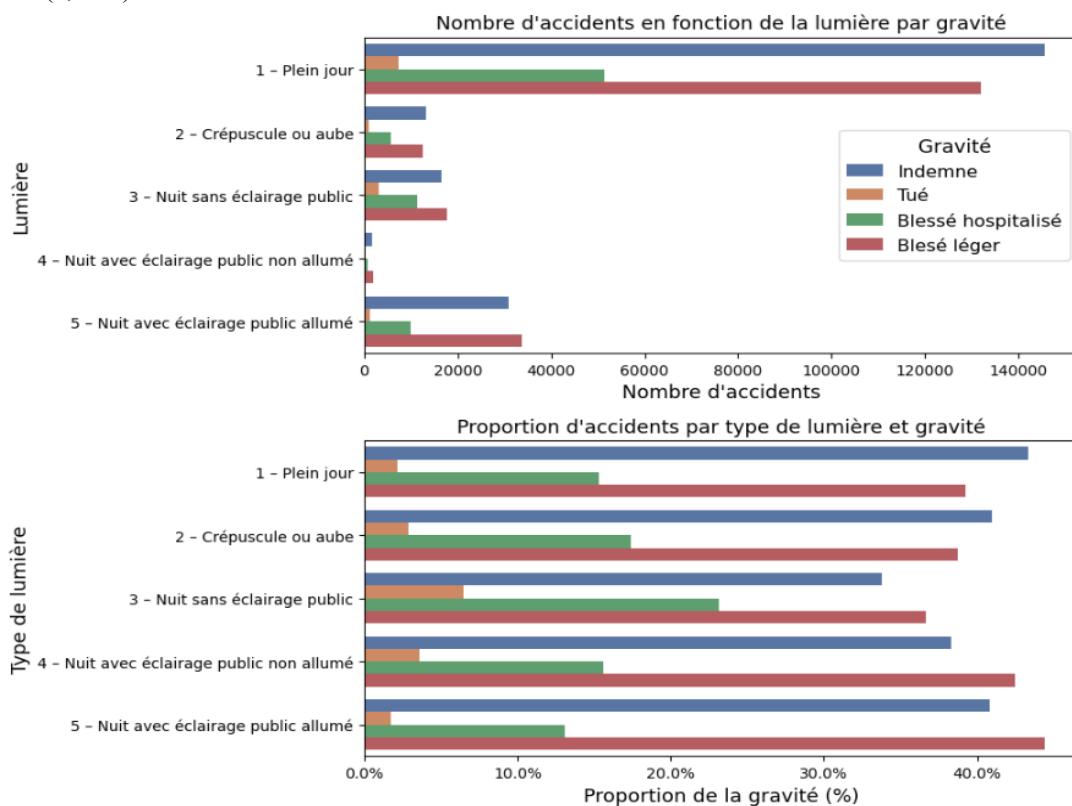


- Le diagramme du dessus illustre le nombre d'accidents par mois en fonction de leur gravité, tandis que le graphique inférieur démontre la proportion de la gravité des accidents au sein de chaque mois.
- La proportion d'accidents mortels fluctue mensuellement, allant de 0.0229 en septembre à 0.0279 en décembre.

b. Variable Lum - Lumière



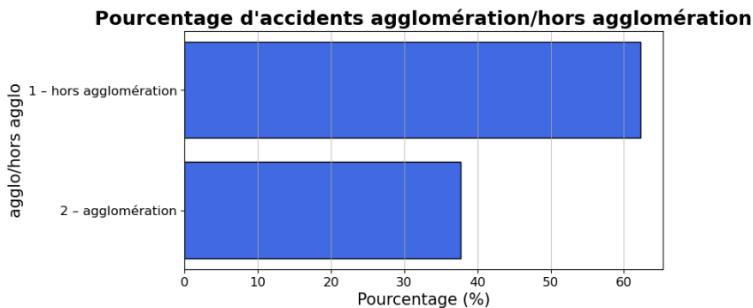
- L'illustration représente la répartition mensuelle des accidents en fonction des diverses conditions d'éclairage.
- Près des deux tiers (67,61%) des accidents se produisent en plein jour, démontrant que la majorité des incidents surviennent dans des conditions de lumière optimale.
- 15,28% des accidents ont lieu la nuit lorsque l'éclairage public est allumé, suggérant que même avec une visibilité améliorée par l'éclairage artificiel, des accidents significatifs se produisent.
- Les accidents se produisant la nuit avec l'éclairage public non allumé ne représentent qu'une petite fraction (0,88%) du total.



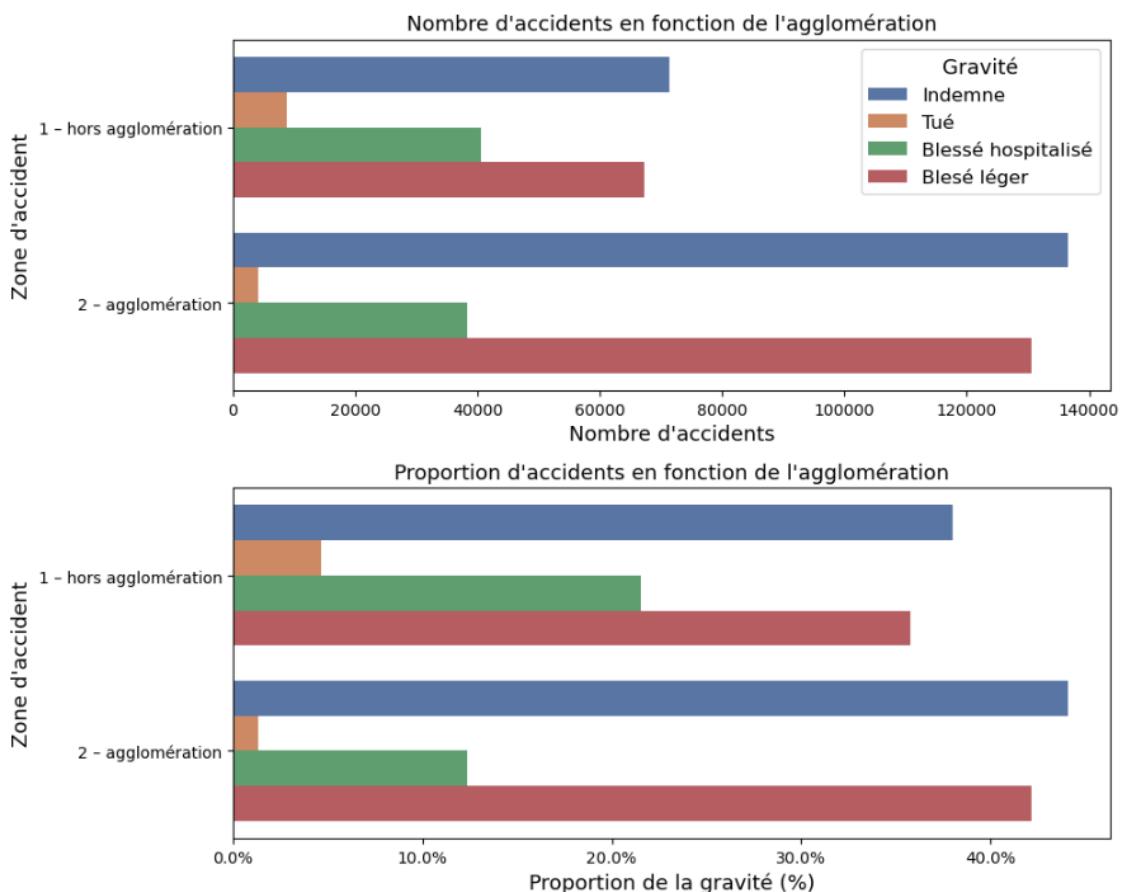
- Le diagramme du dessus illustre le nombre d'accidents par mois en fonction de la lumière, tandis que le graphique inférieur montre la proportion de la gravité des accidents au sein de chaque catégorie de lumière.

- La proportion d'accidents mortels fluctue de 1.7% la nuit avec éclairage public jusqu'à 6.45% (nuit sans éclairage public). Plein jour qui représente la catégorie majoritaire est à 2.2%

c. Variable agg - agglomération

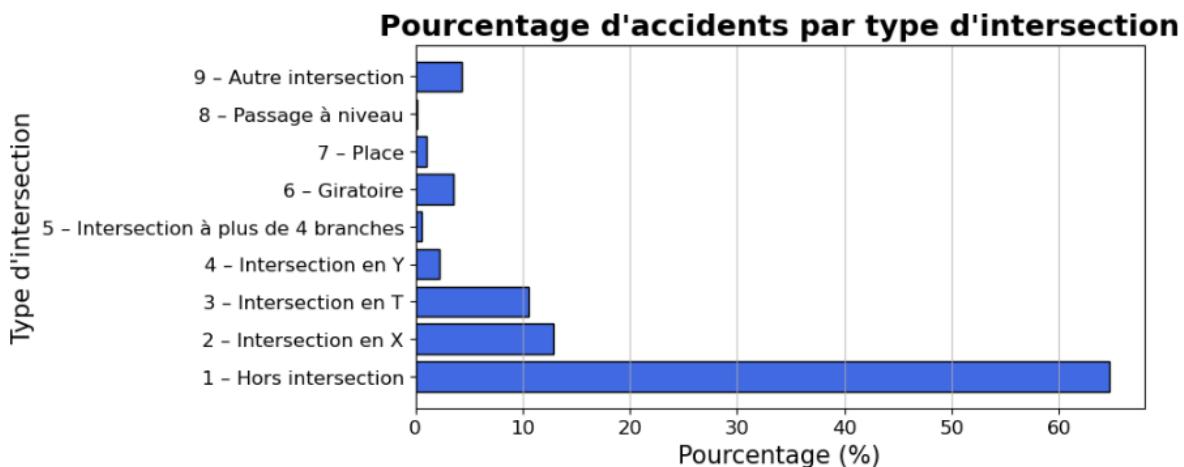


- L'illustration représente la répartition mensuelle des accidents selon qu'ils ont lieu en agglomération ou hors agglomération.
- Près des deux tiers (62,23%) des accidents se produisent en agglomération.
- 37.77% des accidents ont lieu hors-agglomération.

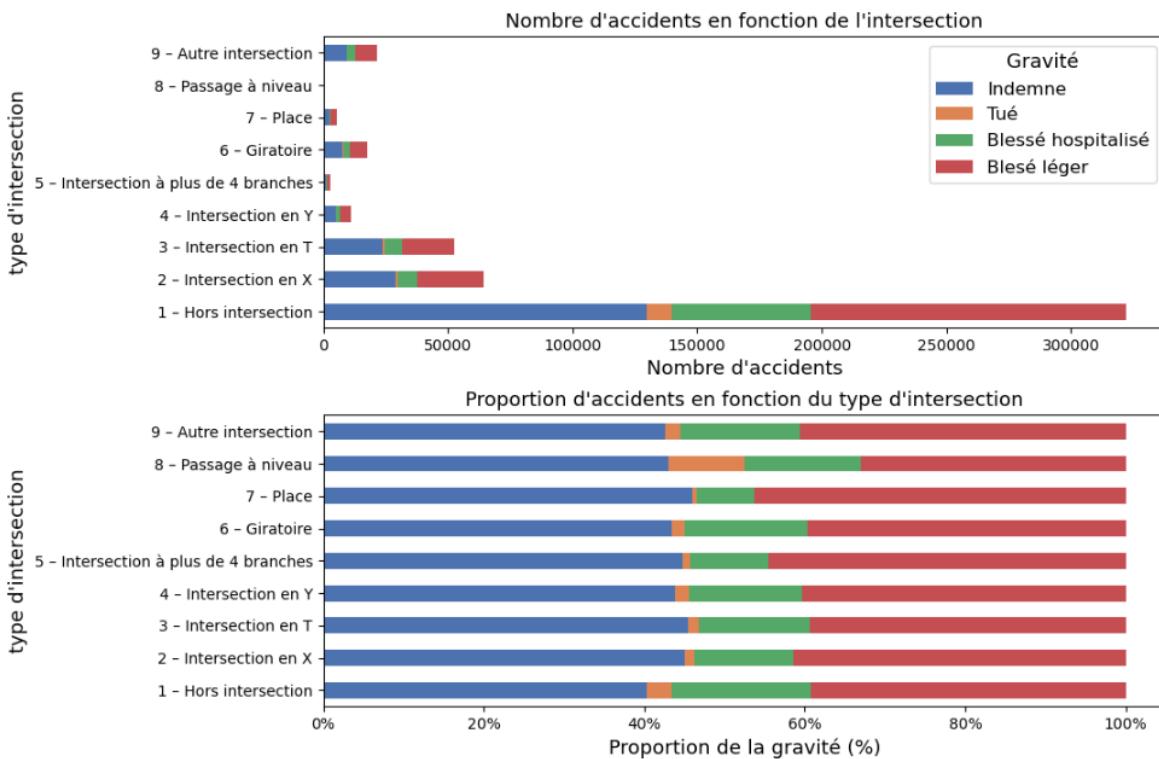


- Le diagramme du dessus illustre le nombre d'accidents par mois en fonction de la zone d'accident (agglo/hors-agglo), tandis que le graphique inférieur montre la proportion de la gravité des accidents au sein de chaque catégorie.
- La proportion d'accidents mortels fluctue de 4.65% hors agglomération et 1.34% en agglomération

d. Variable int - intersection

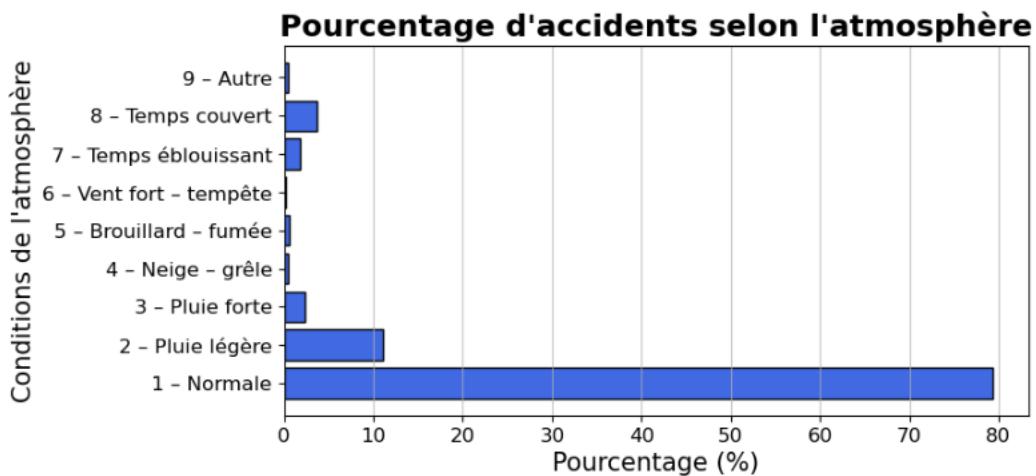


- L'illustration représente la répartition mensuelle des accidents selon le type d'intersection.
- Près des deux tiers (64,7%) des accidents se produisent hors intersection.
- A contrario seulement 0,16% des accidents ont lieu sur un passage à niveau.

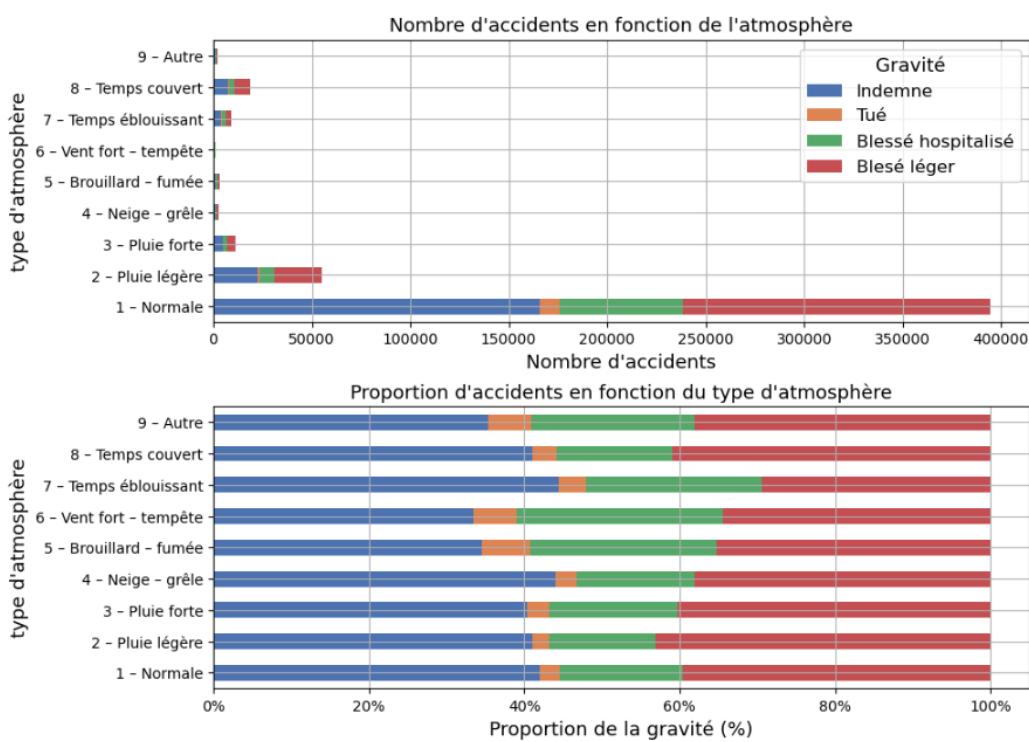


- L'illustration représente le nombre d'accidents en fonction du type d'intersection.
- Les accidents sur passage à niveau ont un taux de mortalité nettement supérieur aux autres à 9.5% contre 3.2% pour la seconde catégorie la plus touché, soit les accidents hors intersection.

e. Variable atm - atmosphère

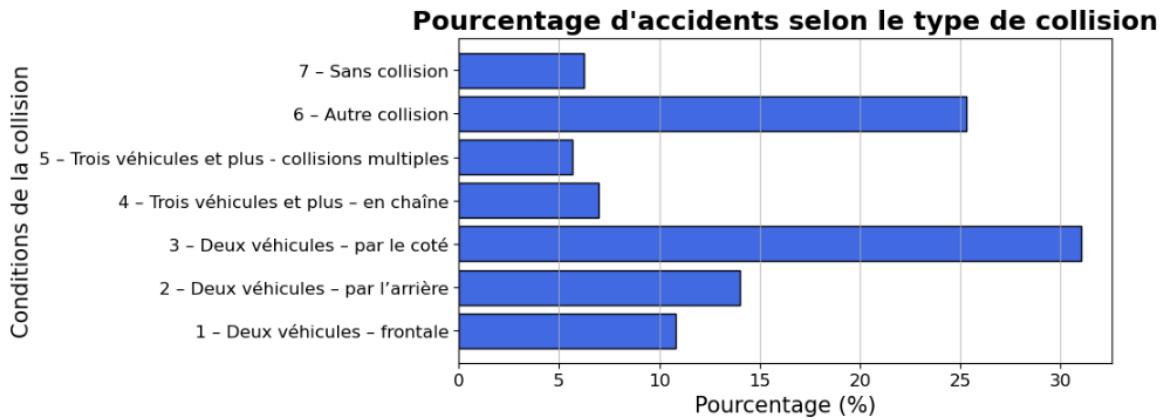


- L'illustration représente la répartition mensuelle des accidents selon l'atmosphère.
- Près des 4/5 (79.33%) des accidents se produisent dans des conditions d'atmosphère « normale ».
- A contrario seulement 0.25% des accidents ont lieu en cas de « vent fort-tempête ».

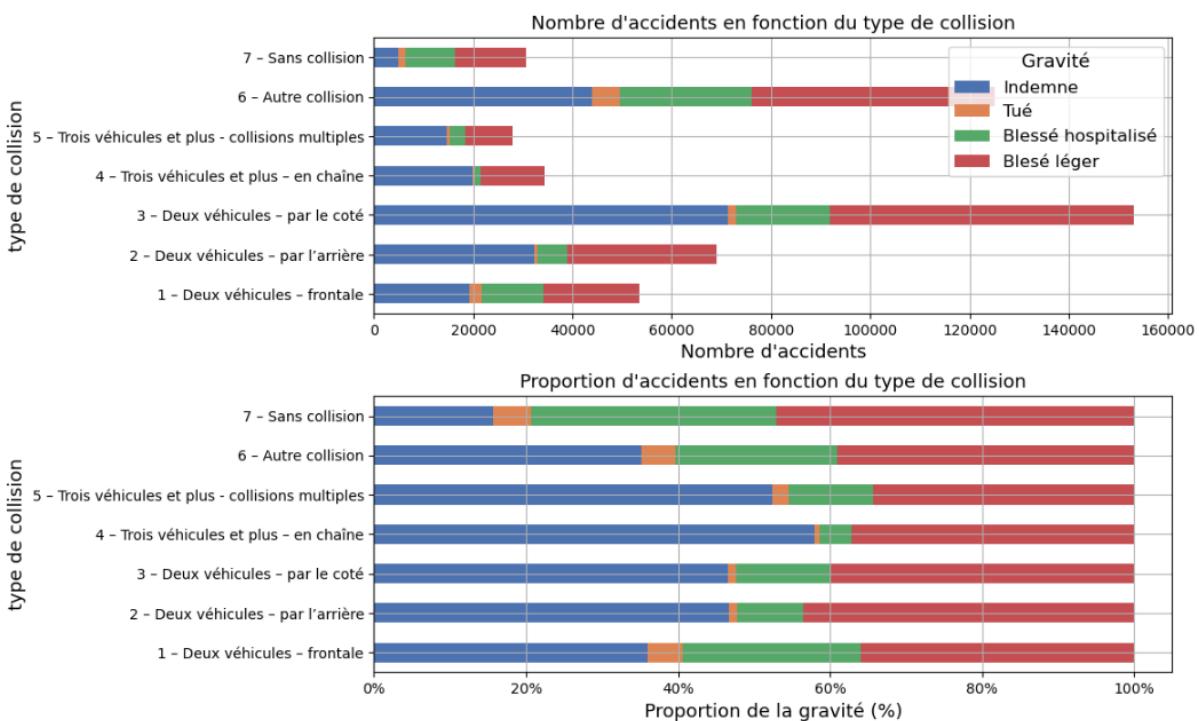


- Ces deux graphiques servent à visualiser et à comprendre la répartition des accidents et de leur gravité en fonction du type d'atmosphère.
- Le premier affiche le nombre total d'accident en fonction du type d'atmosphère avec pour chaque barre la répartition suivant le type de gravité
- Le deuxième graphique montre la proportion en pourcentage des différents niveaux de gravités par type d'atmosphère.
- La classe 5 « brouillard-fumée » a le plus haut taux de mort avec 6.21% tandis que pluie légère a le plus faible taux à 2.2%.

f. Variable col - collision



- L'illustration représente la répartition mensuelle des accidents selon le type de collision.
- La distribution des valeurs est un peu plus équilibrée que pour les autres variables
- La classe la plus représentée, 31.02%, implique deux véhicules par côtés, et celle la moins la représentée, 5.66%, implique trois véhicules et plus ».



- Ces deux graphiques servent à visualiser et à comprendre la répartition des accidents, et de leur gravité, en fonction du type de collision.
- Le premier affiche le nombre total d'accidents en fonction du type de collision avec pour chaque barre la répartition suivant le type de gravité
- Le deuxième graphique montre la proportion en pourcentage des différents niveaux de gravités par type de collision.
- La classe 7 « sans collision » a le plus haut taux de mort avec 5 % tandis que « trois véhicules ou plus » ont un taux de mortalité de 0.65%

3.2.2 Analyses statistiques

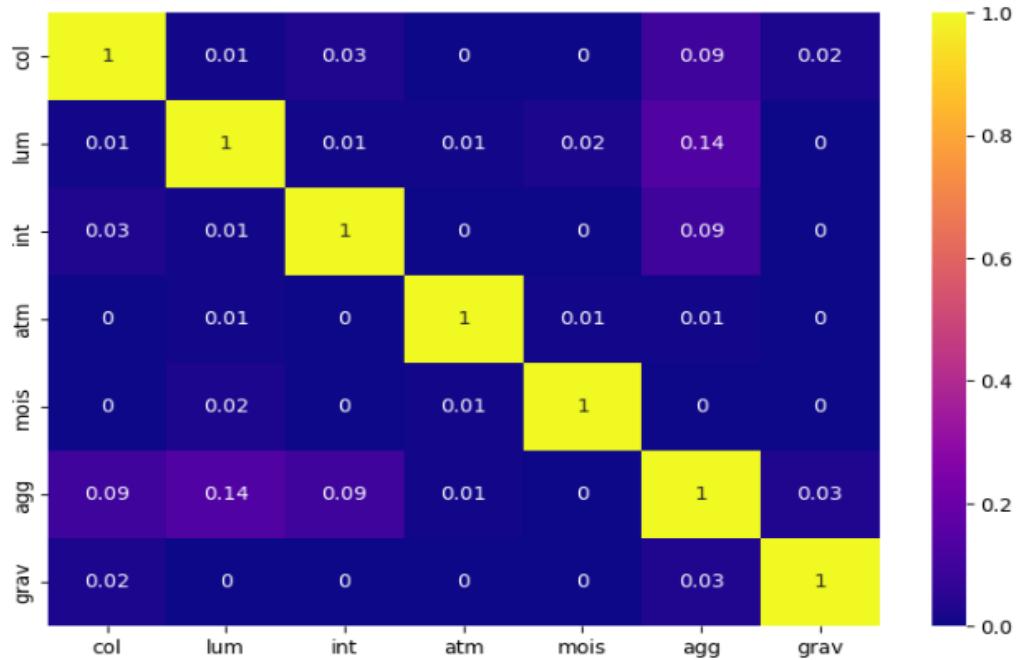
Dans cette section, nous approfondissons notre analyse en utilisant des tests statistiques pour confirmer les observations lors de l'exploration univariée et bivariée avec la variable cible.

a. Test chi2

	var	chi2	p-value	Degrees of freedom
0	com	171340.646402	0.000000e+00	59676
0	col	33933.919776	0.000000e+00	18
0	lum	6897.873620	0.000000e+00	12
0	int	3754.523173	0.000000e+00	24
0	atm	1632.423885	0.000000e+00	24
0	mois	559.798165	4.779413e-97	33
0	agg	0.000000	1.000000e+00	0

Les conclusions du test Chi2 démontrent que les valeurs de p sont extrêmement basses (p-value=0.0) pour toutes les variables étudiées. Cela signifie qu'il existe une corrélation statistiquement notable entre ces variables et la variable cible, qui est la gravité des accidents. Ces résultats réfutent donc l'hypothèse de départ, qui postulait qu'il n'y avait aucune liaison entre ces variables.

b. Test V Cramer



- Les valeurs du V de Cramer en relation avec la variable cible sont égales à 0 ou s'en approchent, ce qui indique que l'intensité de la corrélation statistique est faible, même si les variables sont dépendantes.

3.2.3 Conclusions

Le tableau suivant résume notre analyse des variables présentes dans le dataframe intitulé "Caractéristiques". Nous avons utilisé trois critères clés pour sélectionner les variables les plus pertinentes : leur pertinence selon notre connaissance du domaine, l'exploration bivariée et leur dépendance statistique avec la variable cible. Seules les variables qui ont répondu avec succès à ces trois critères ont été retenues pour la phase de modélisation suivante

Variable	description	Pertinence selon métier	pertinence de la variable cible (observation graphique)	Test statistiques dépendances variables explicative et la cible	Décision
jour	Jour du mois de l'accident.	NOK	OK	OK	à supprimer après la création des variables week end et date
an	Année de l'accident.	NOK	OK	OK	
mois	Mois de l'accident.	OK	OK	OK	à garder
hrmn	Heure et minutes de l'accident.	NOK	OK	OK	à supprimer après la création de la variable heure
lum	Lumière : conditions d'éclairage	OK	OK	OK	à garder
dep	Département	NOK	OK	OK	à supprimer
com	Commune	NOK	NOK	OK	à supprimer
agg	Localisation	OK	OK	OK	à garder
int	Intersection	OK	OK	OK	à garder
atm	Conditions atmosphériques	OK	OK	OK	à garder
col	Type de collision	OK	OK	OK	à garder
adr	Adresse postale	NOK	NOK	OK	à supprimer
lat	Latitude	NOK	NOK	OK	Utilise pour créer un diagramme de géolocalisation mais à supprimer du modèle
long	Longitude	NOK	NOK	OK	

NOK	pas pertinente
OK	pertinente
-	sans signification

3.3 Data-frame : Usager

Dans cette section, notre focus sera sur l'analyse du data-frame "Usagers". Notre but est d'appréhender et d'examiner les caractéristiques et les interactions entre les variables, afin de déterminer celles qui sont les plus pertinentes pour le processus de modélisation. Le choix de ces variables se base sur trois critères : leur pertinence dans le contexte métier, l'analyse bivariée, et leur corrélation statistique avec la variable cible. En intégrant ces facteurs, nous serons en mesure de sélectionner les paramètres les plus importants pour prédire avec précision la variable d'intérêt.

Les variables présentes dans le dataset sont toutes pertinentes.

En 2018, une nouvelle représentation des équipements de sécurité (ceinture, casque...) est apparue. Jusqu'en 2018 seule une variable « secu » sur 2 caractères représentait si un équipement de sécurité existait et s'il était utilisé lors de l'accident. Depuis 2019, trois variables « secu1 », « secu2 », « secu3 » représentent la présence d'au maximum trois équipements de sécurité. Il n'y plus de notion d'équipements utilisés. La variable « secu » a été intégrée dans la variable « secu1 ».

En 2018, une nouvelle identification des véhicules est survenue. Jusqu'en 2018, les véhicules étaient identifiés par « num_véhicule » avec les mêmes valeurs par accident. Pour retrouver un véhicule il faut donc faire le lien avec « num_acc » + « num_véhicule ». En 2019, la variable « id_véhicule » a été rajoutée pour identifier directement de manière unique un véhicule. Toutefois, la variable « num_véhicule » a continué d'être alimentée, ce qui a généré de la redondance.

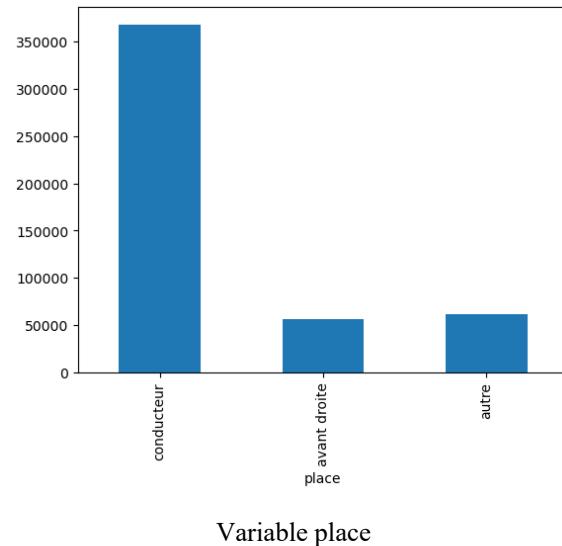
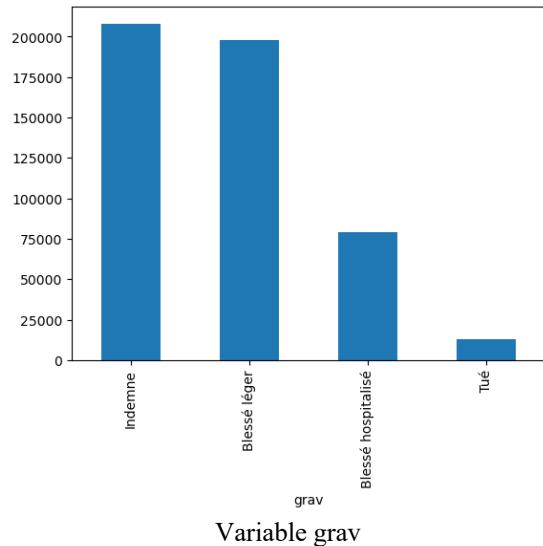
Particularités et limites du jeu de données :

Le jeu de données est déséquilibré.

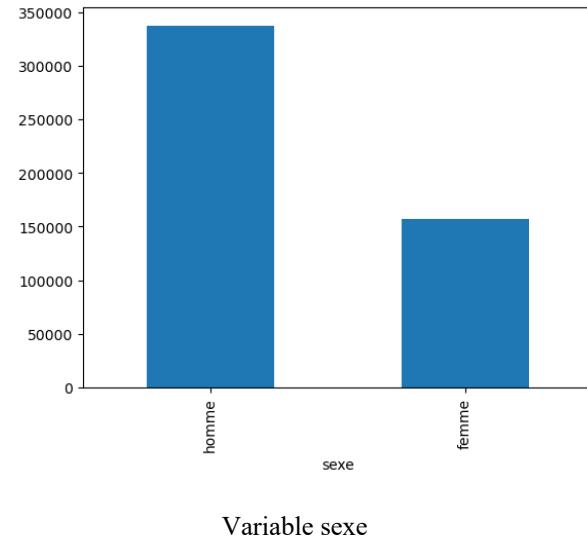
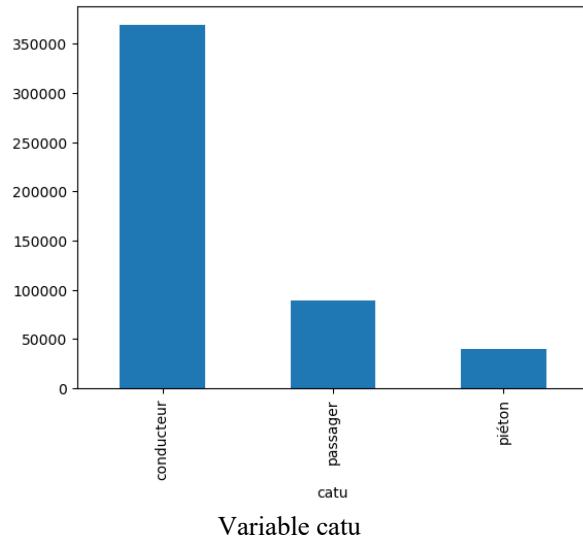
Le jeu de données comporte beaucoup de NaN.

3.3. 1 Exploration univariée et bivariée (avec variable cible)

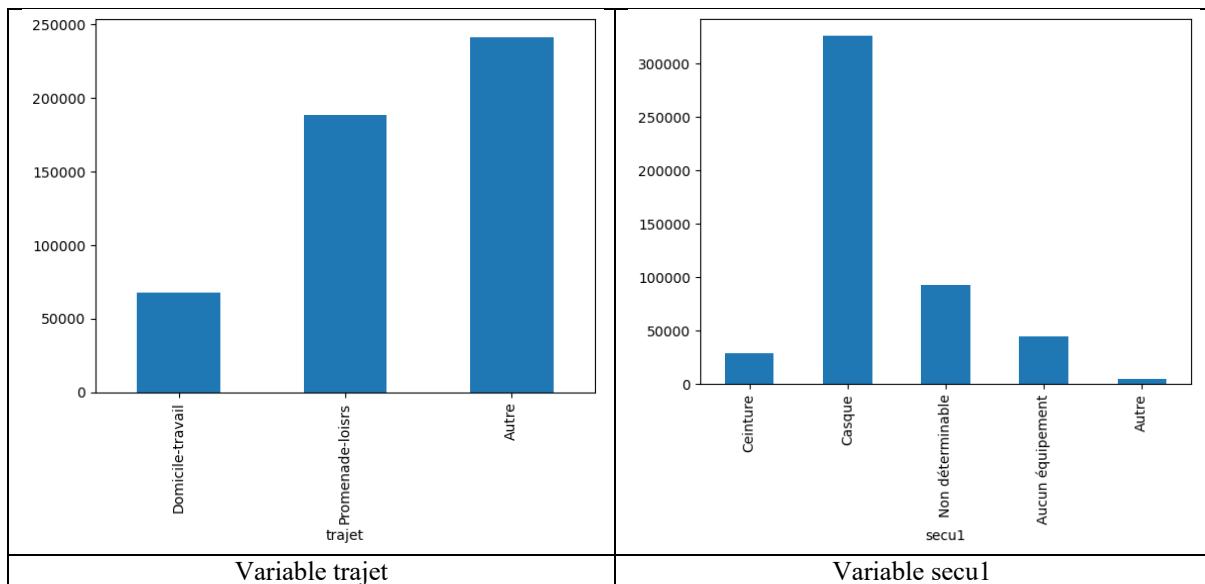
Analyse univariée des variables



- Analyse de la variable cible : grav => Variable déséquilibrée avec une grande majorité d'accidents sans conséquence grave (indemne ou blessé léger)
- Variable place : variable déséquilibrée avec une très grande majorité d'usagers physiquement à la place des conducteurs lors de l'accident.



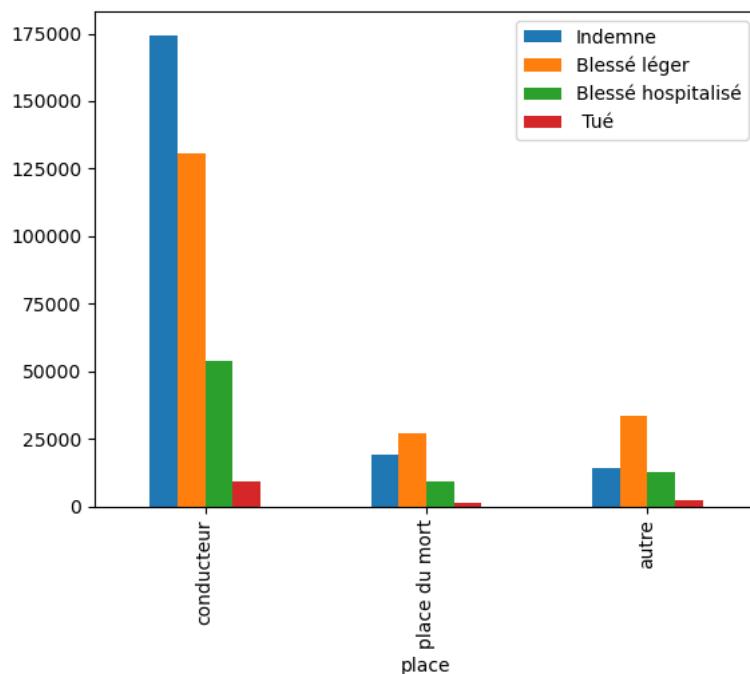
- Variable catu : variable déséquilibrée avec une très grande majorité d'usagers qui sont effectivement des conducteurs.
- Sexe : variable très déséquibrée qui montre que la majorité des usagers sont des hommes.



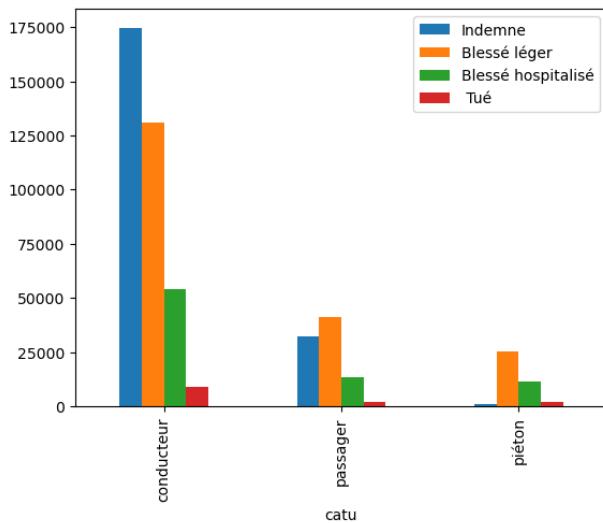
- Variable trajet : variable relativement équilibrée
- Secu1 : l'équipement de sécurité le plus utilisé est le casque

Analyse bi-variée avec la variable cible

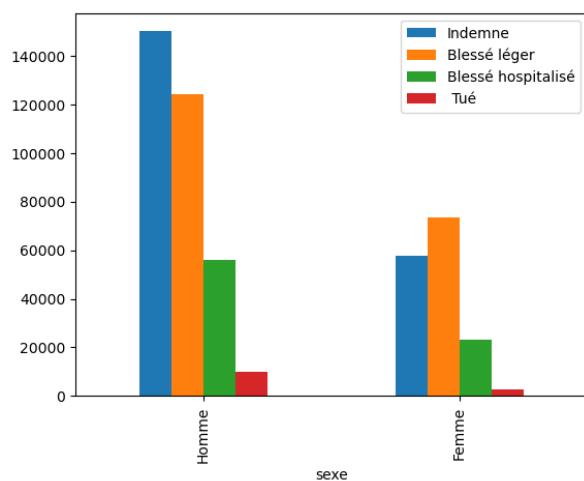
Place et variable cible. Les usagers assis physiquement à la place du conducteur sont les usagers les plus impactés lors des accidents, peu importe la gravité de l'accident.



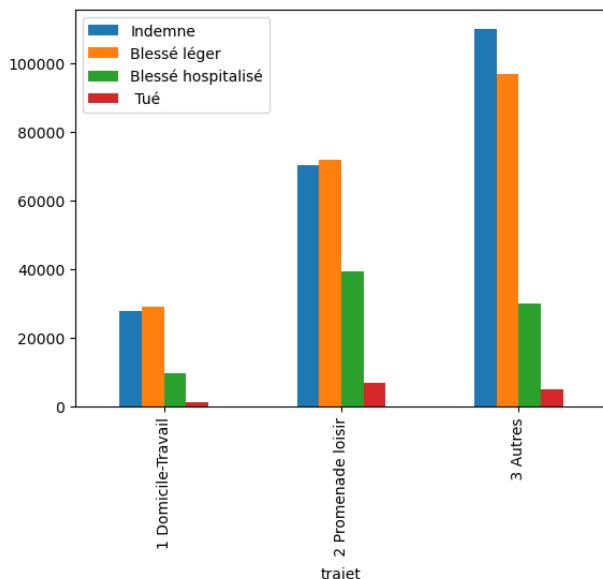
Catégorie variable cible : Les conducteurs sont les usagers les plus impactés lors des accidents, peu importe la gravité de l'accident.



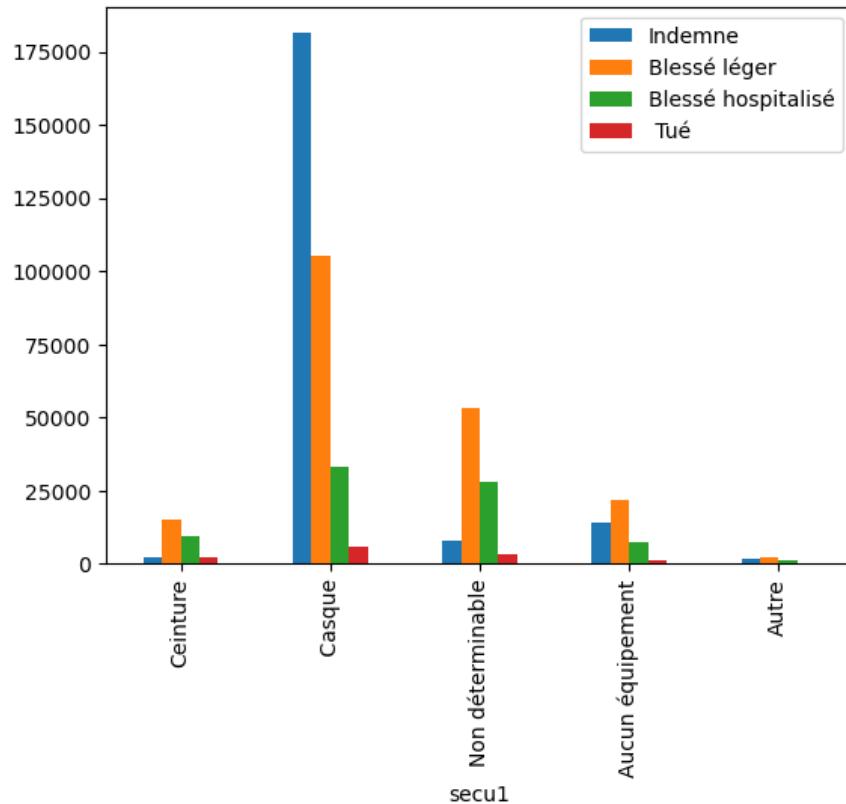
Sexe et variable cible. Les hommes sont les usagers les plus impactés lors des accidents et le plus gravement.



Trajet et cible. Les trajets pendant lesquels arrivent plus fréquemment les accidents sont les trajet de promenade-loisirs.

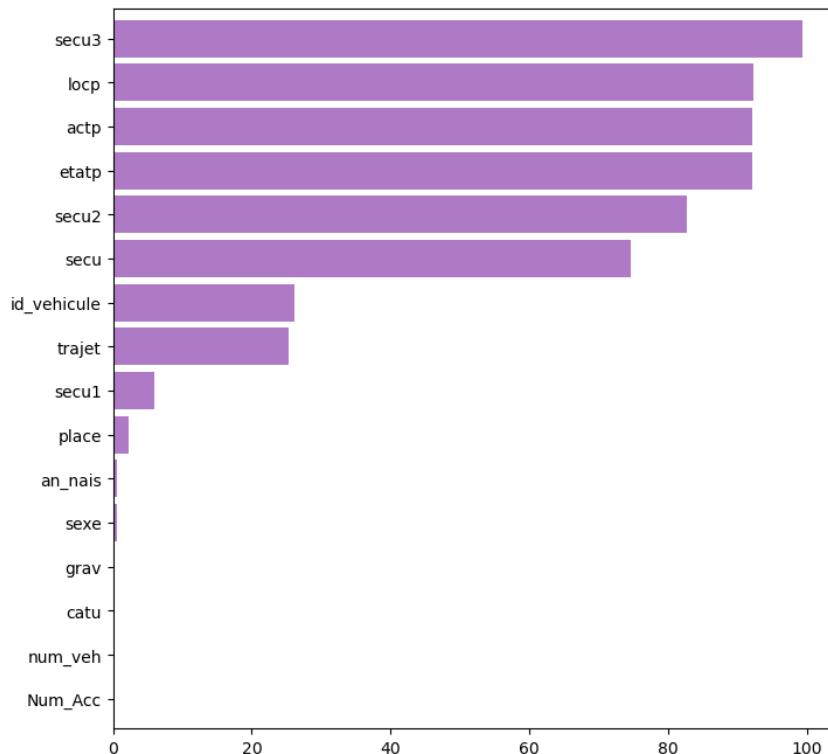


Secu1 et variable cible. L'équipement de sécurité le plus utilisé est le casque.



Les autres variables du dataframe contiennent beaucoup de NaN. Nous avons défini le seuil de 30% de NaN pour supprimer une variable.

Les variables secu2, secu3, locp, etat et actp ont des NaN à + de 90%. Elles seront supprimées dans l'étape de pre-processing. La variable secu possède également un nombre important de NaN qui justifie sa suppression.



3.3.2 Analyses statistique

Dans cette section, nous approfondissons notre analyse en utilisant des tests statistiques pour confirmer les observations lors de l'exploration univariée et bivariée avec la variable cible.

Nous présentons ici les analyses statistiques du test du chi2 et du V de cramer, uniquement sur les variables que l'on conserve : place, catu, sexe, trajet, secu1, âge et la variable cible grav.

a. Test chi2

Le calcul de chi2 pour les variables catégorielles par rapport à la variable cible donne les résultats suivants :

	var	chi2	p-value	Degrees of freedom
0	secu1	90696.890218	0.0	12
0	catu	31880.215334	0.0	6
0	place	14258.957120	0.0	6
0	trajet	8218.291144	0.0	6
0	sexe	4984.714248	0.0	3

Les résultats du test du Chi2 révèlent des valeurs de p très faibles (p-value=0.0) pour toutes les variables examinées, indiquant une dépendance statistiquement significative entre ces variables et la variable cible (gravité des accidents). Ces résultats rejettent l'hypothèse nulle selon laquelle il n'y a aucune association entre ces variables.

a. V de Cramer

La figure heatmap illustrant les résultats du test V de Cramer, mesure la force de l'association entre les variables catégorielles et la variable cible (gravité des accidents). Les couleurs dans la heatmap indiquent le degré de corrélation.



- Les valeurs du V de Cramer en relation avec la variable cible sont égales à 0 ou s'en approchent, ce qui indique que l'intensité de la corrélation statistique est faible, même si les variables sont dépendantes.
- Sauf pour les variables catu et place qui sont corrélées statiquement

3.3.3 Conclusions

Variable	Type variable	Description	Pertinence selon métier	Pertinence de la variable cible (observation graphique)	Test statistiques dépendances explicatives et la cible	Décision
Num_ACC	explicative	Identifiant de l'accident	Oui	Oui	Oui	A supprimer après fusion
place	explicative	Place occupée dans le véhicule par l'usager au moment de l'accident	Oui	Oui	Oui	A garder
catu	explicative	Catégorie d'usager	Oui	Oui	Oui	A garder
grav	cible	grav représente la gravité de blessure de l'usager	Oui	Oui	Oui	A garder
sexe	explicative	sexe de l'usager	Oui	Oui	Oui	A garder
trajet	explicative	Motif du déplacement au moment de l'accident	Oui	Oui	Oui	A garder
locp	explicative	Localisation du piéton	Oui	Non	Non	A supprimer
actp	explicative	Action du piéton	Oui	Non	Non	A supprimer
etatp	explicative	préciser si le piéton accidenté était seul ou non	Oui	Non	Non	A supprimer
an_naiss	explicative	Année de naissance de l'usager	Oui	Non	Non	A remplacer par age plus pertinent
id_véhicule	explicative	Identifiant unique du véhicule repris pour chacun des usagers occupant ce véhicule	Oui	Oui	Oui	A supprimer après fusion
num_vehicule	explicative	Identifiant unique du véhicule repris pour chacun des usagers occupant ce véhicule	Oui	Oui	Oui	A supprimer après fusion
secu	explicative	Existence et utilisation équipement sécurité	Oui	Oui	Oui	A fusionner avec secu1
secu1	explicative	Existence et utilisation équipement sécurité	Oui	Oui	Oui	A garder
secu2	explicative	Existence et utilisation équipement sécurité	Oui	Non	Non	A supprimer
secu3	explicative	Existence et utilisation équipement sécurité	Oui	Non	Non	A supprimer

3.4 Data-frame : Lieux

Le dataframe des Lieux représente un certain nombre de caractéristiques liées au lieu de l'accident. Ci-dessous se trouvent la description technique de chacune d'entre-elles et la description fonctionnelle.

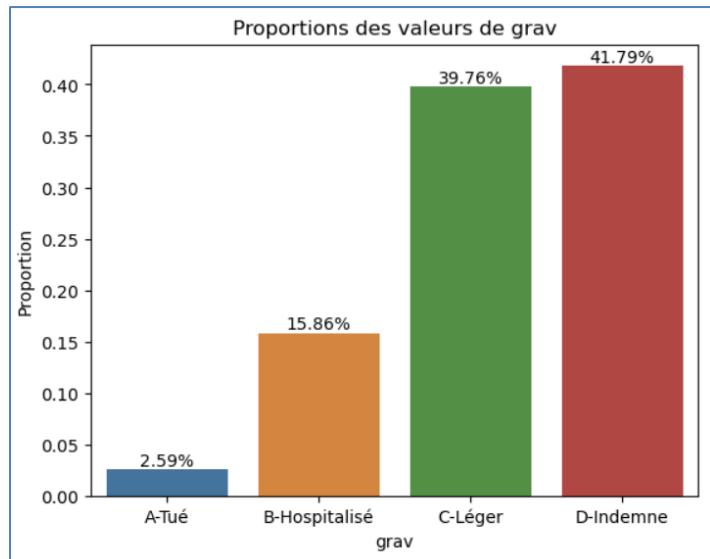
Column	Non-Null	Dtype	% de nan	Description
Num_Acc	497594	int64	0.00%	clef unique identifiant les accidents
an	497594	int64	0.00%	année
catr	497594	int64	0.00%	catégorie de route
voie	413492	object	16.90%	numéro et nom de voie
v1	343182	float64	31.04%	indice numérique de la route
v2	33594	object	93.25%	indice alphanumérique de la route
circ	496665	float64	0.19%	régime de circulation
nbv	496422	float64	0.24%	nombre total de voies de circulation
pr	464226	object	6.71%	points de rattachement
pr1	463620	object	6.83%	points de rattachement
vosp	496378	float64	0.24%	existence d'une voie réservée
prof	496587	float64	0.20%	Profil en long décrit la déclivité
plan	496621	float64	0.20%	tracé en plan, courbe ou rectiligne
lartpc	29731	object	94.02%	largeur du terre plein central s'il existe
larrout	264943	object	46.76%	largeur de la chaussée affectée à la circulation
surf	496568	float64	0.21%	état de la surface
infra	496519	float64	0.22%	aménagement de l'infrastructure
situ	496460	float64	0.23%	situation de l'accident
env1	129110	float64	74.05%	Variable non documentée
vma	367425	float64	26.16%	vitesse maximale autorisée
grav	497594	int64	0.00%	Gravité de l'accident (variable cible)
count	497594	int64	0.00%	Compteur ajouté temporairement

Parmi toutes ces variables beaucoup ne présentent aucun intérêt fonctionnel/métier en lien avec la détermination de la gravité de l'accident, ou présentent trop nan pour être exploitables. Ces variables ne seront donc pas étudiées, mais supprimées ultérieurement dans la phase de pré-processing.

Ces variables sont : voie, v1, v2, pr, pr1, lartpc, larrout, env1

Les autres variables sont étudiées une à une dans cette section, également sous forme univariée et bivariée avec la cible. La variable cible a été conservée dans le sens où il n'y a pas eu de regroupement effectué. En revanche, cette variable cible a été légèrement retravaillée pour être plus compréhensible, comme indiqué sur le graph ci-dessous.

Certaines valeurs de variables sont précisées dans la documentation du jeu de données, mais n'apparaissent pas dans les données, de même que d'autres données non-précisées apparaissent. Cela sera nettoyé lors du pré-processing. Ici, seules les informations essentielles sont revues.

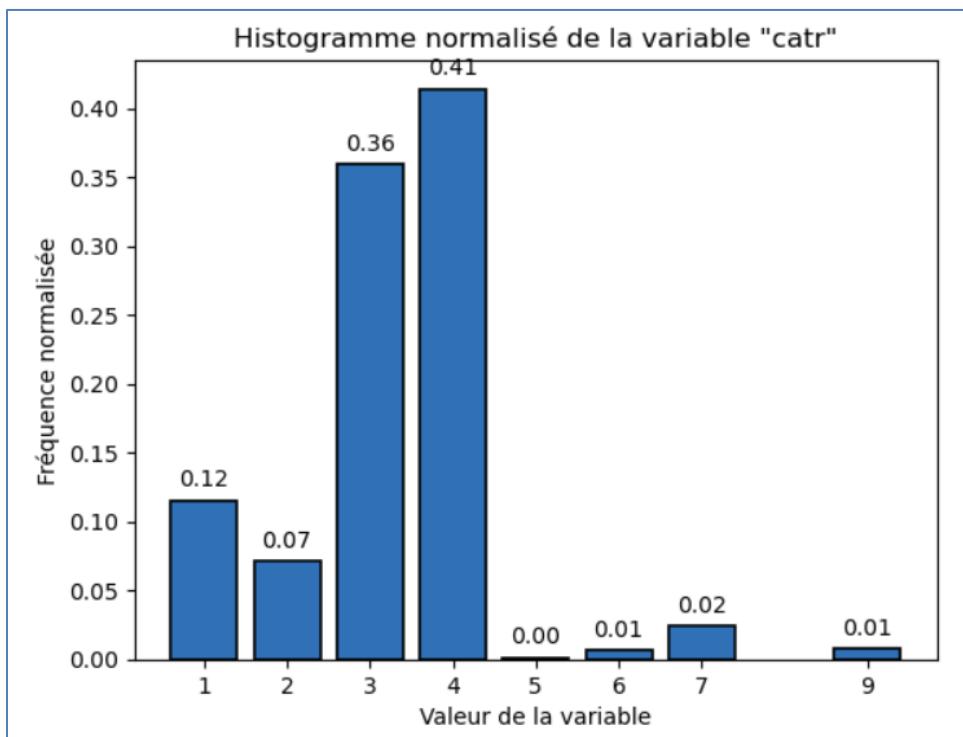


3.4.1 Exploration univariée et bivariée (avec variable cible)

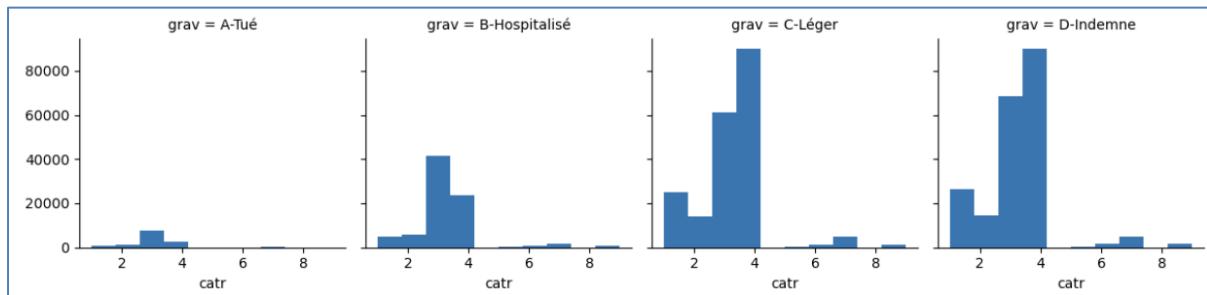
A. Variable catr - catégorie de route

Valeurs de cette variable:

- 1 – Autoroute
- 2 – Route nationale
- 3 – Route Départementale
- 4 – Voie Communales
- 5 – Hors réseau public
- 6 – Parc de stationnement ouvert à la circulation publique
- 7 – Routes de métropole urbaine
- 9 – autre



Les routes départementales (3) et voies communales (4) cumulent 77% des accidents. Le nombre d'accidents majoritaire est sur les voies communales.



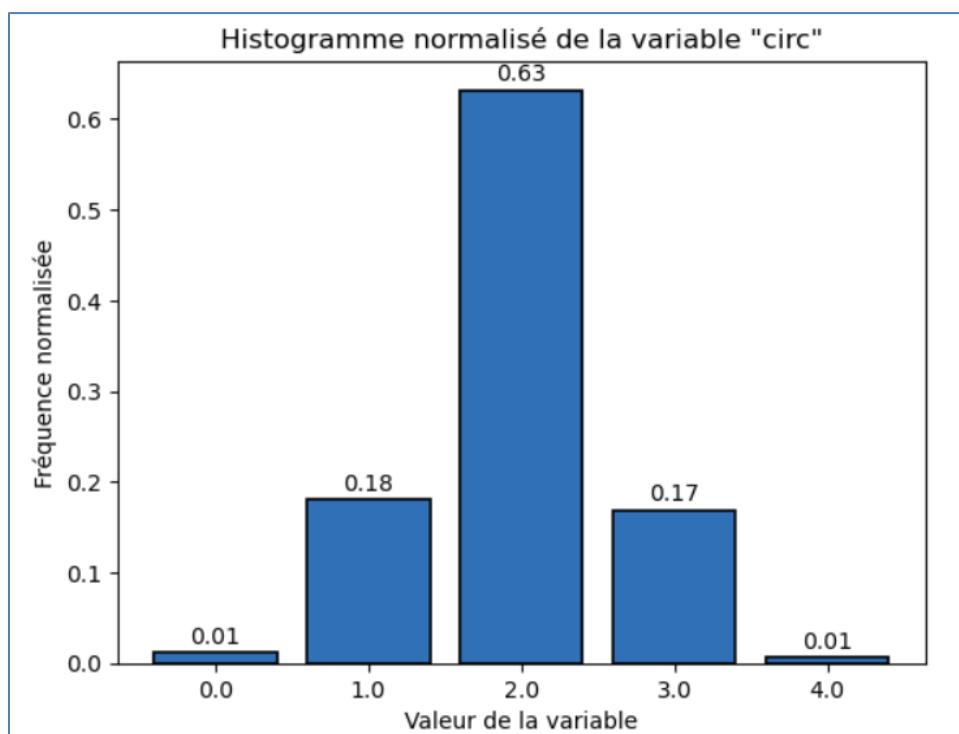
L'analyse bivariée avec la cible confirme la prépondérance des voies communales pour les blessés légers ou indemnes, en revanche pour les morts et hospitalisés, cela se produit essentiellement sur les routes départementales.

Nous regrouperons les catégories 5 à 9 en une seule catégorie 5 nommée 'autres'.

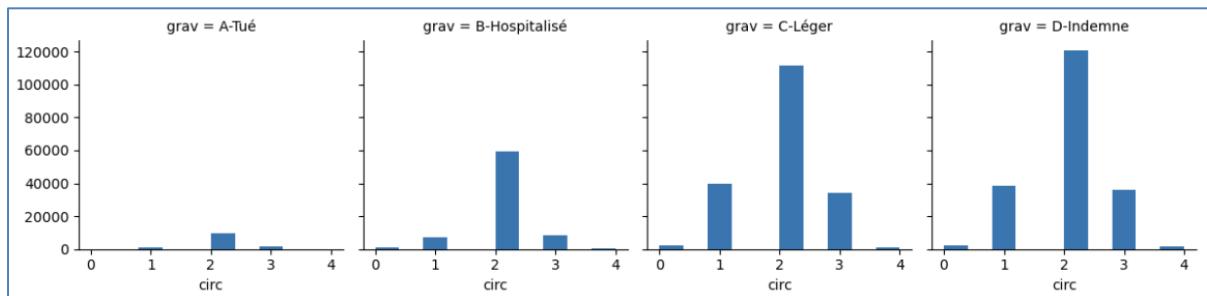
B. Variables circ – régime de circulation

Valeurs de cette variable:

- 1 – Non renseigné
- 1 – A sens unique
- 2 – Bidirectionnelle
- 3 – A chaussées séparées
- 4 – Avec voies d'affectation variable
- 7 – Routes de métropole urbaine
- 9 – autre



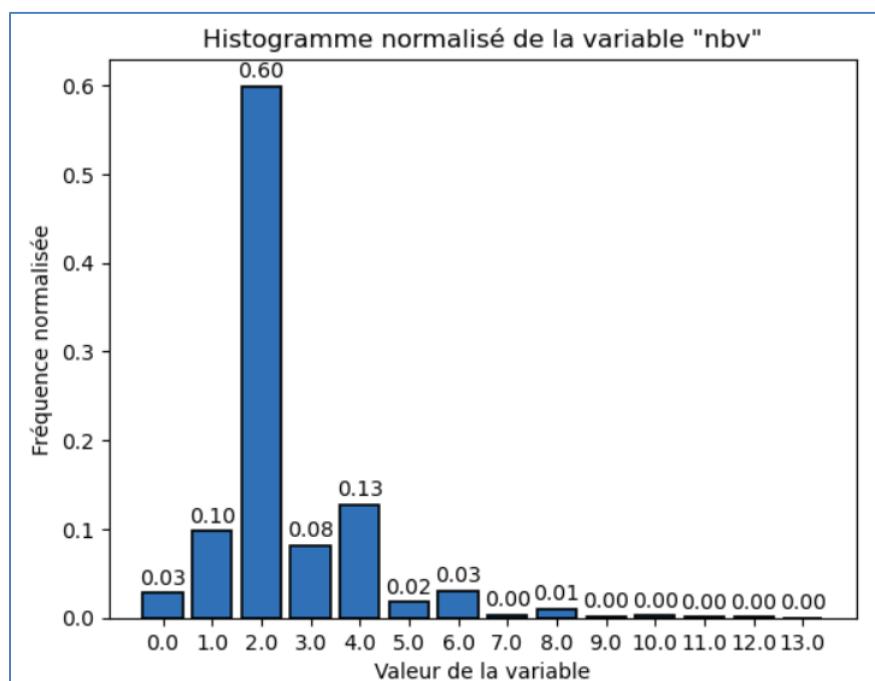
On note une nette prépondérance des accidents lors d'une circulation à deux sens, et dans une moindre mesure lorsque l'on a un sens unique ou des chaussées séparées.



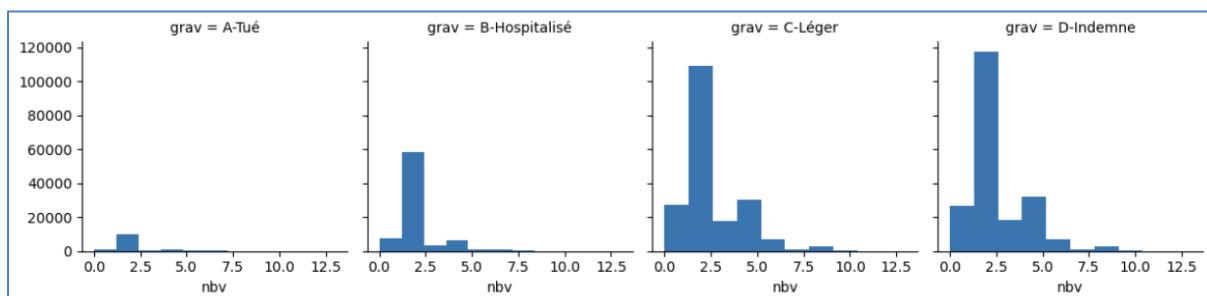
L'analyse multivariée montre que ces mêmes prédominances sont conservées quelque soit la catégorie de la gravité.

C. Variable nbv – nombre de voies

Cette variable ne contient pas de catégorie, c'est un entier qui se situe entre 0 et 13 selon les données recensées.



Il y a une forte prédominance d'accidents lorsqu'il y a deux voies, ce qui renforce l'analyse de la variable précédente qui indique que cela a lieu sur un sens bidirectionnel, donc sur deux voies en sens opposé.

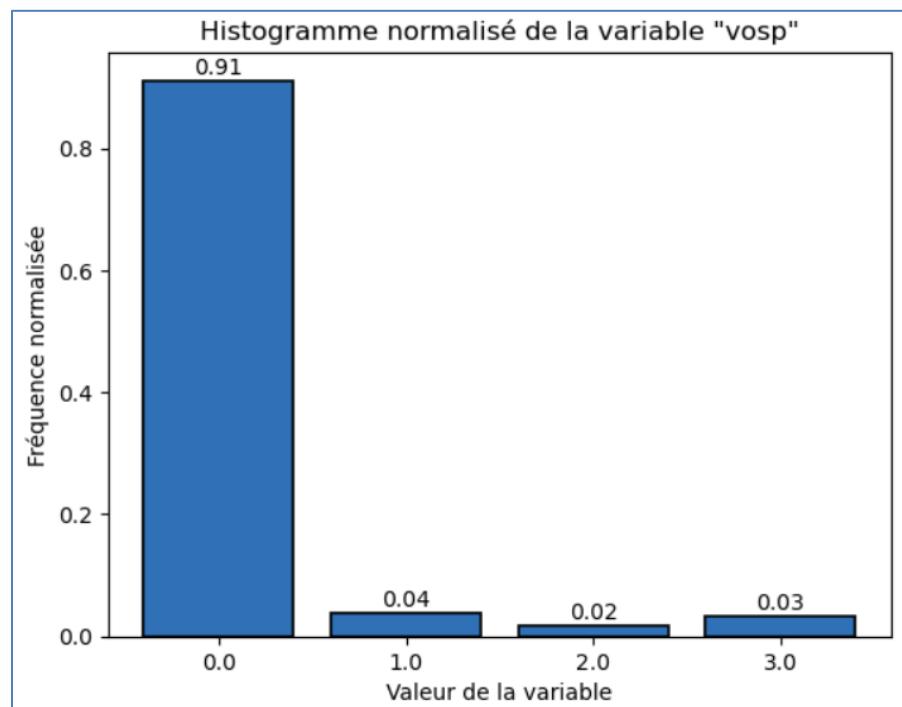


Quelle que soit la gravité, la prépondérance d'accident est sur une chaussée à deux voies.

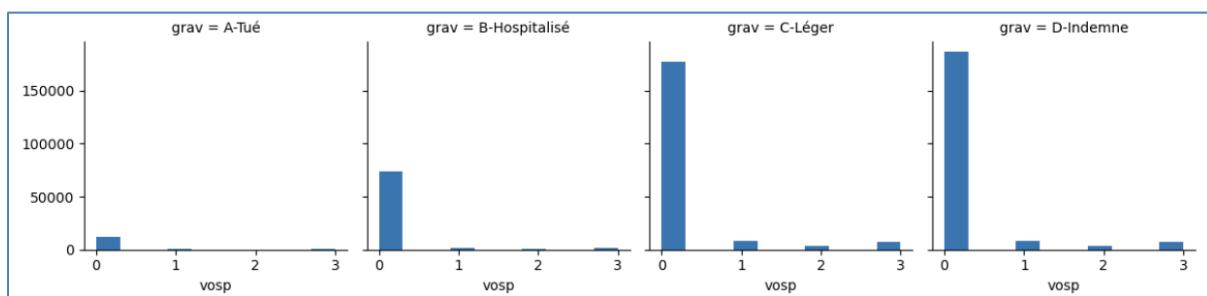
D. Variable vosp – voies réservées

Valeurs de cette variable

- 1 – Non renseigné
- 0 – Sans objet
- 1 – Piste cyclable
- 2 – Bande cyclable
- 3 – Voie réservée



91% des accidents se produisent en l'absence d'une voie réservée.



Cette même prépondérance est valable quelle que soit la gravité de l'accident.

E. Variable prof – profil de la route

Valeurs de cette variable

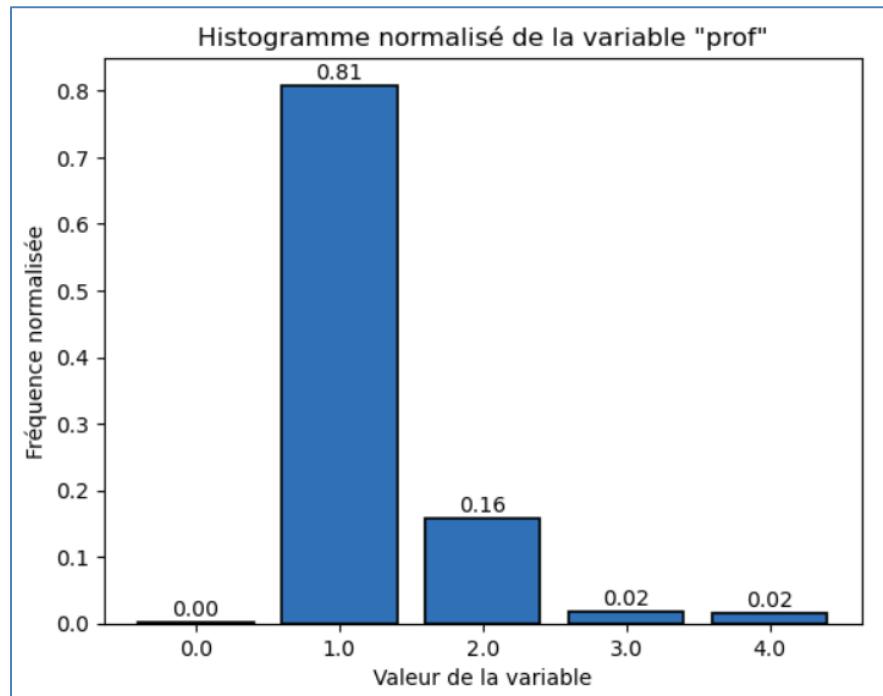
-1 – Non renseigné

1 – Plat

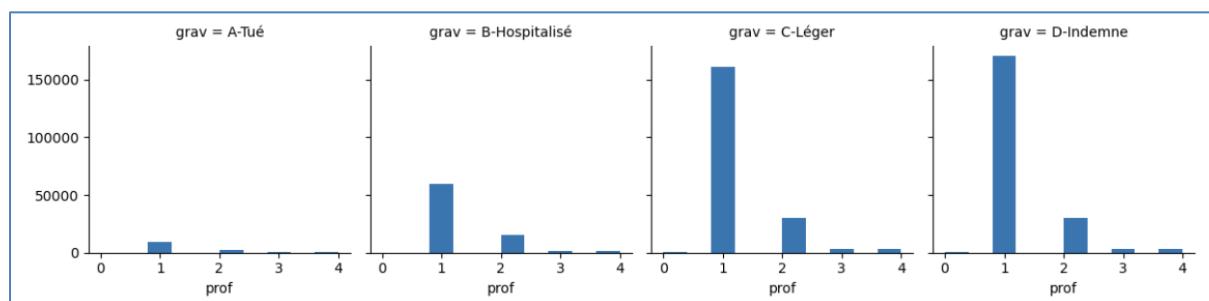
2 – Pente

3 – Sommet de côte

4 – Bas de côte



On note que 81% des accidents ont lieu sur une route plate, et 16% sur une route en déclivité.



Les accidents sur route plate restent majoritaires, quelle que soit la gravité.

F. Variable plan – tracé en plan

Valeurs de cette variable

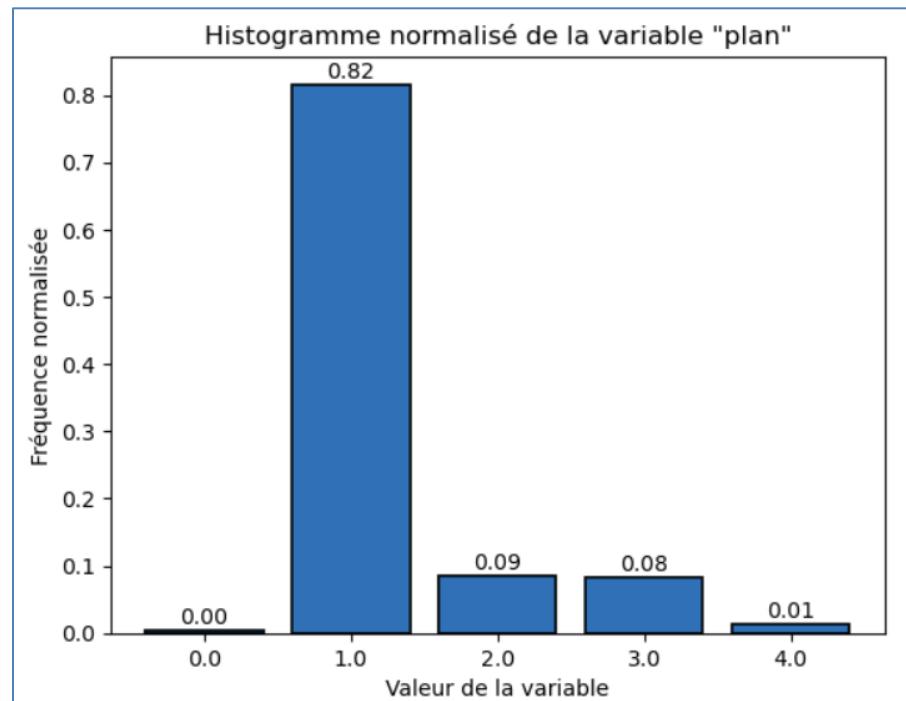
-1 – Non renseigné

1 – Partie rectiligne

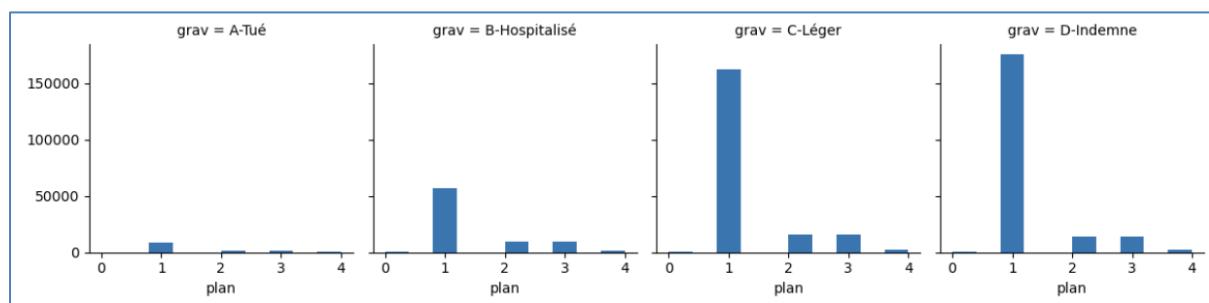
2 – En courbe à gauche

3 – En courbe à droite

4 – En « S »



On note que 82% des accidents ont lieux en ligne droite, alors que 17% se produisent en courbe/virage.

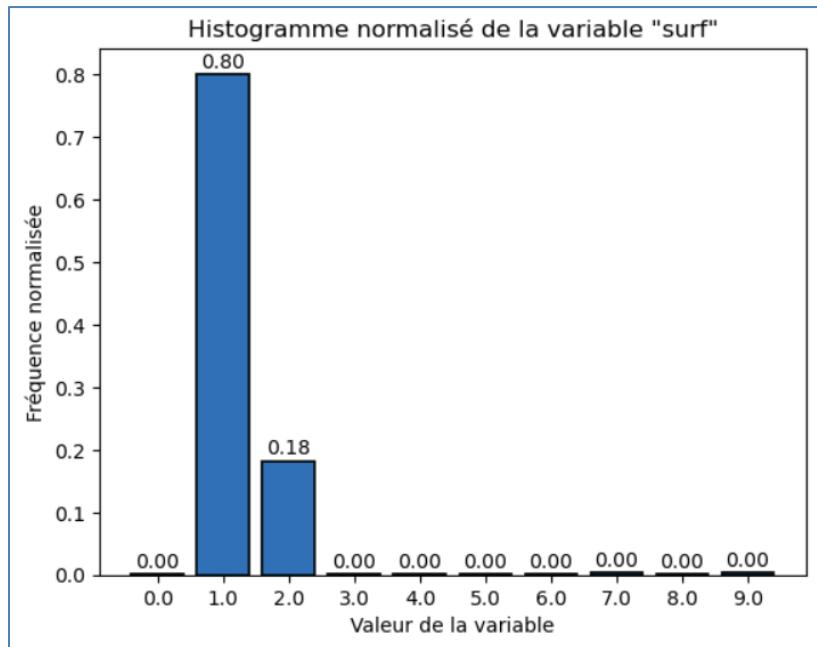


La prépondérance des accidents en ligne droite reste quelle que soit la gravité de l'accident.

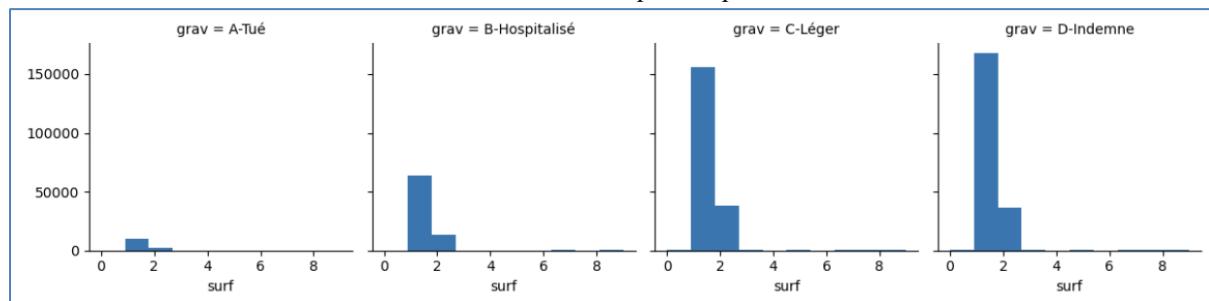
G. Variable surf – état de la surface

Valeurs de cette variable

- 1 – Non renseigné
- 1 – Normale
- 2 – Mouillée
- 3 – Flaque
- 4 – Inondée
- 5 – Enneigée
- 6 – Boue
- 7 – Verglacée
- 8 – Corps gras – huile
- 9 – Autre



On note que 80% des accidents se produisent sur une surface normale (sèche), alors que 18% se produisent sur une surface mouillée. Tous les autres états de surface ne dépassent pas 1%.

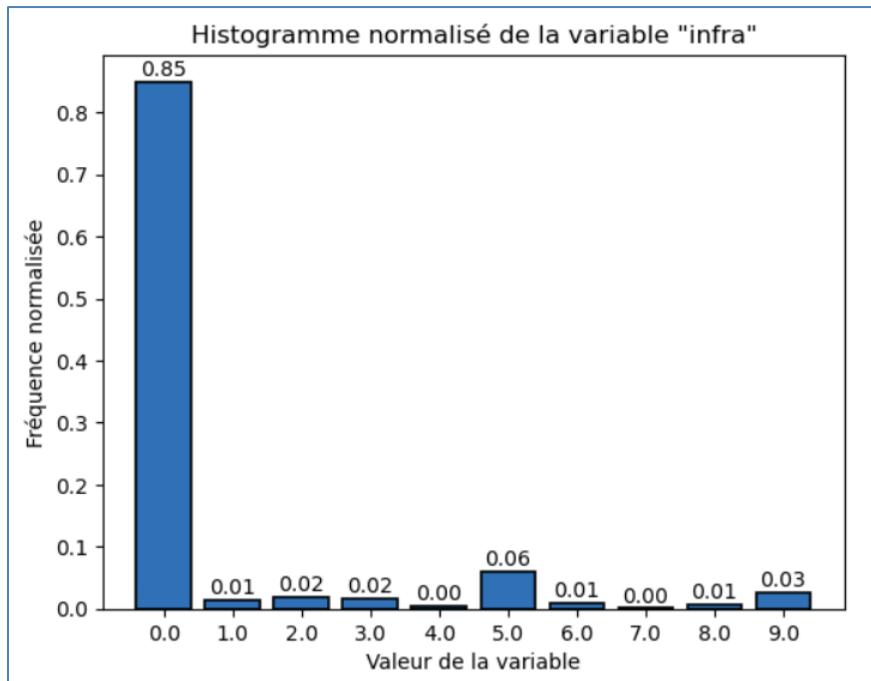


La prédominance des accidents sur route de surface normale reste quelle que soit la gravité de l'accident.

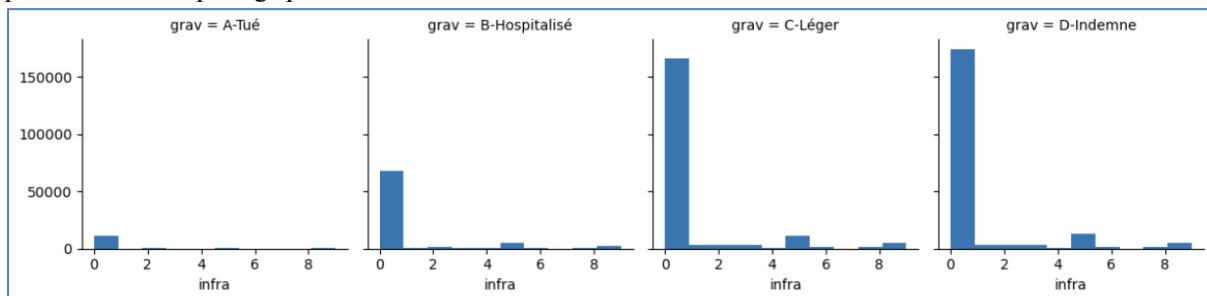
H. Variable infra – Aménagement - Infrastructure

Valeurs de cette variable

- 1 – Non renseigné
- 0 – Aucun
- 1 – Souterrain - tunnel
- 2 – Pont - autopont
- 3 – Bretelle d'échangeur ou de raccordement
- 4 – Voie ferrée
- 5 – Carrefour aménagé
- 6 – Zone piétonne
- 7 – Zone de péage
- 8 – Chantier
- 9 – Autres



La majorité, 85%, des accidents se produisent sur un lieu sans infrastructure particulière. 6% d'entre eux se produisent sur un passage piéton.

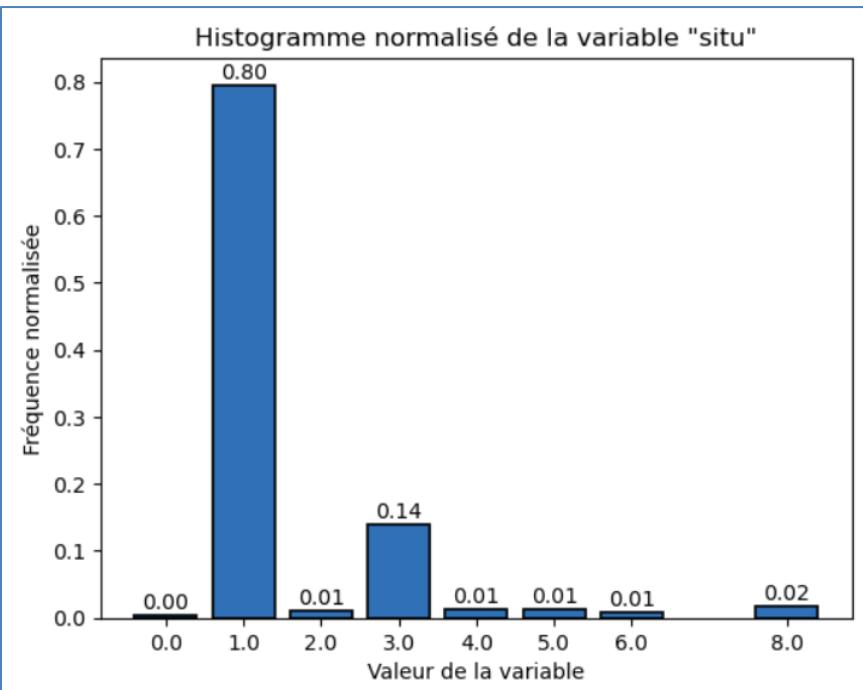


Il n'y a pas de changement de cette prépondérance lorsque l'on regarde les accidents par gravité.

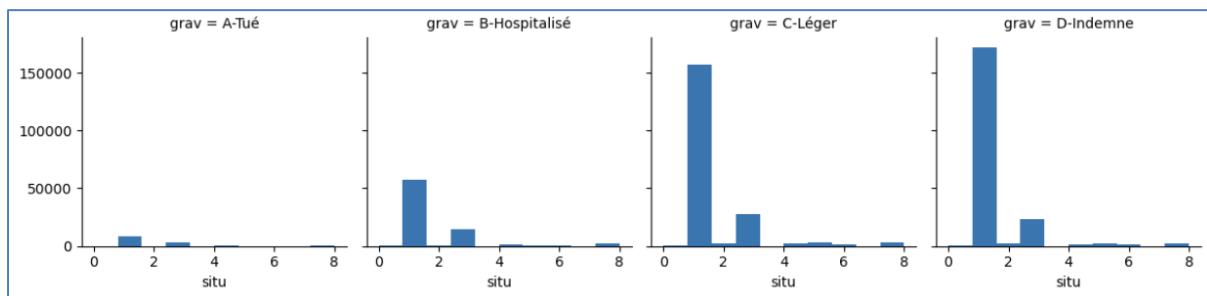
I. Variable situ – Situation de l'accident

Valeurs de cette variable

- 1 – Non renseigné
- 0 – Aucun
- 1 – Sur chaussée
- 2 – Sur bande d'arrêt d'urgence
- 3 – Sur accotement
- 4 – Sur trottoir
- 5 – Sur piste cyclable
- 6 – Sur autre voie spéciale
- 8 – Autres



Les accidents sur chaussée représentent 80% du total des accidents, et ceux sur accotement représentent 14%.

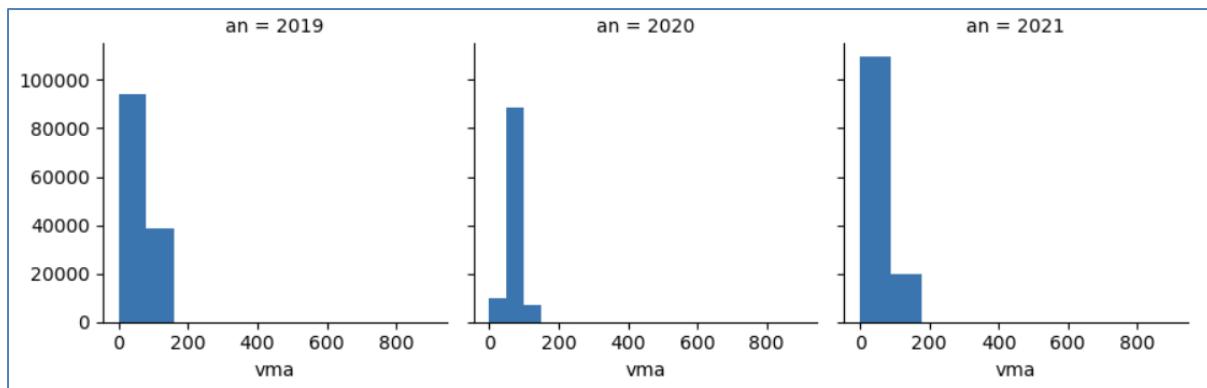


L'analyse par gravité ne change pas ces prédominances.

J. Variable vma – vitesse maximale autorisée

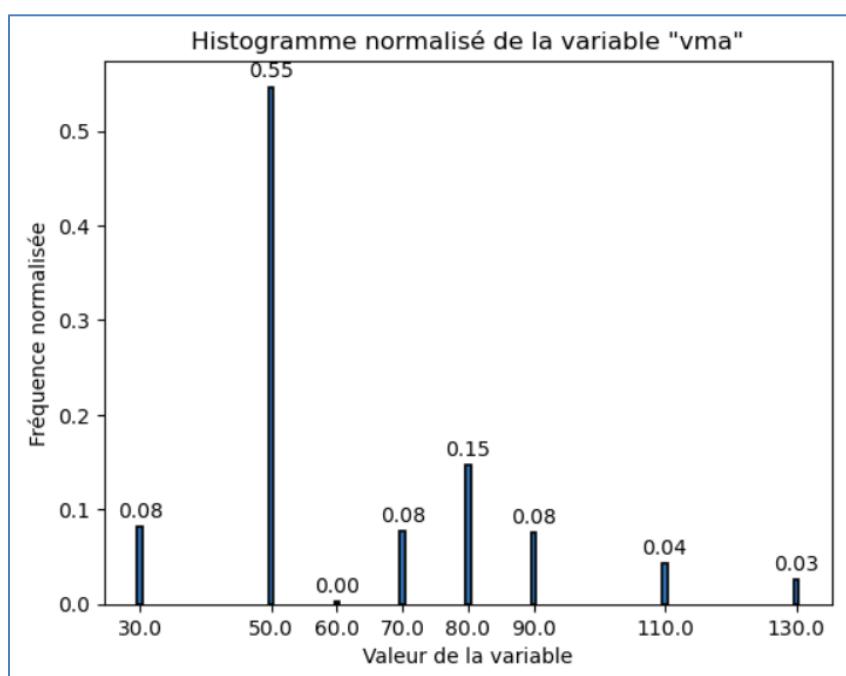
Cette variable est cardinale. En l'absence de précisions dans la documentation nous avons supposé qu'il s'agit de km/h. A défaut, d'avoir la vitesse des véhicules au moment de l'accident, nous sommes parti du principe que cette variable pourrait être intéressante, faute de mieux.

Néanmoins nous avons fini par abandonner cette variables pour 3 raisons : elle contient des saisies aberrantes, elle contient 26% de nan car absente sur 2018, et des hypothèses de regroupement sont faites pour exploiter cette variable. In fine cela revient à avoir les variables agg (agglomération) et catr (catégorie de route) qui ne présentent pas de problèmes de qualité ou de nan élevé.



Voici les histogrammes de vitesses par année. On note une échelle dépassant les 800 km/h...

Après nettoyage nous obtenons un graphique plus intéressant :



Néanmoins, l'hypothèse de départ de l'étude et de prendre les années de 2018 à 2021, et comme 2018 est vide, nous avons opté pour l'abandon de cette variable.

Il est néanmoins intéressant de remarquer que 63% des accidents se font à 50 km/h et moins, donc en agglomération essentiellement. Ce chiffre est confirmé par l'analyse de la variable agg (agglomération). Pour les autres valeurs de vitesse, la variable catr (catégorie de route) définit mieux si l'on est sur une départementale, nationale, etc...

3.4.2 Analyses statistique

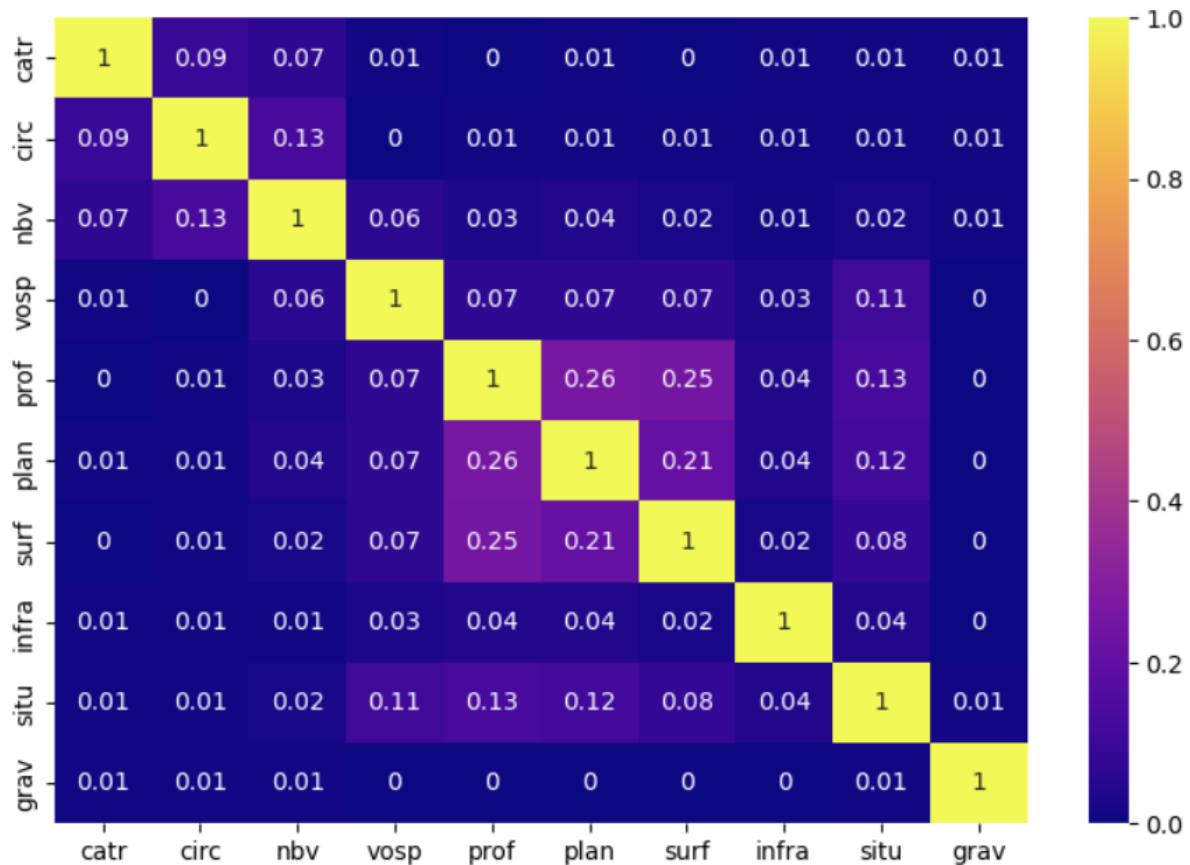
Comme décrit dans les précédentes parties, nous présentons ici les analyses statistiques du test du chi² et du V de cramer, uniquement sur les variables que l'on conserve.

a. Test chi2

	var	chi2	p-value	Degrees of freedom
0	catr	18334.399116	0.000000e+00	21
0	nbv	11755.218696	0.000000e+00	39
0	cicr	10741.465251	0.000000e+00	12
0	situ	8078.438987	0.000000e+00	21
0	plan	7221.903785	0.000000e+00	12
0	prof	2196.000432	0.000000e+00	12
0	infra	1127.630529	2.013434e-220	27
0	vosp	1118.803779	4.064027e-235	9
0	surf	1104.907689	1.342242e-215	27

L'ensemble des p-value sont nulles, le test de dépendance de chaque variable explicative avec la variable cible est donc vérifié.

b. V de Cramer



Le test de cramer fait apparaître qu'il y a bien des variables explicatives qui ont une « force » de relation entre elles, comme par exemple le profil de la route (prof) et la forme de la route (plan = tracé en plan), mais aucune de ces variables explicatives n'est fortement liée à la gravité de l'accident.

4.1.3 Conclusions

Sur la partie Lieux, nous regrettons que certaines variables ne soient apparues que tardivement et soient mal saisies, comme vma (vitesse maximale autorisée). Egalemente, il aurait été appréciable ici des données GPS en plus des adresses, qui sont incomplètes et mal saisies. Nous avions pensé pouvoir enrichir la vma à partir du d'une base de donnée des vitesses autorisées en fonction de leur coordonnées géographique. Mais, il n'existe actuellement pas une telle base. Elle est en cours de constitution auprès de chaque région en France, depuis plusieurs années déjà.

4. Pre-processing et feature engineering

Ce chapitre se concentre sur les étapes de prétraitement et de feature ingineering.

Dans la section 4.1, nous effectuerons le nettoyage du data-frame initial en supprimant les erreurs présentes. Cela nous permettra d'obtenir un jeu de données plus fiable et cohérent.

Dans la section 4.2, nous traiterons les valeurs manquantes (Nan) qui peuvent être présentes dans nos données. Nous utiliserons différentes techniques de gestion de ces valeurs afin de minimiser leur impact sur la variable cible.

La section 4.3 sera consacrée à la création de variables supplémentaires et au regroupement de catégories. Nous examinerons chaque data-frame (Véhicules, Caractéristiques, Usager, Lieux) individuellement et identifierons les possibilités d'enrichir nos données en créant de nouvelles variables ou en regroupant certaines catégories pour diminuer la taille de la dataset.

Enfin, dans la section 4.3.4, nous présenterons le data-frame final qui résultera de ces étapes de prétraitement.

4.1 Nettoyage du data-frame initial : Suppression des erreurs

	Nom de variables	Erreurs détectées
Véhicules	Id_vehicule	1) il faut remplacer '\xa' par ' ' : 138\x0306 -> 1380306 2)
	Num_veh	il faut supprimer les erreurs de frappe '\01' et '\01'
	Tout les variables	Remplacer -1 par NaN
Caractéristiques	an	Transformation de l'année 18 en 2018
	hrmn	Pour les datas 2018 mise au format 0000 et ajout du séparateur « : » entre heures et minutes
	col	Remplacer -1 par Nan
	int	Remplacer -1 par Nan
	lum	Remplacer -1 par Nan
	atm	Remplacer -1 par Nan
Usager	actp	Variable contenant valeurs int et str Remplacement des valeurs str par des valeurs int Changement du type de catégorie la variable en int
	secu	Copie de valeurs de secu dans la variable secu1 + mapping
	Etap, trajet, secu2, secu3, locp, actp	Remplacement valeur 0 "Non renseigné" par NaN
	catu	Remplacement valeur 4 non prévue dans la documentation par NaN
	Toutes les variables	Remplacement valeur -1 "Non renseigné" par NaN
Lieux		Pas d'erreur

4.2 Traitement des valeurs manquantes (NaN)

	Nom de variables	Taux de NA	Garder la colonne ?	Stratégies de gestion de NaN pour les colonnes gardées	Décision
Véhicules	obs	4%	Oui	Stratégie 1 : Suppression toutes les lignes avec Nan Variable cible : Répartition avant suppression des lignes : 81,6% cas non - graves (0) ; 18,4% cas grave (1) Variable cible : Répartition apres suppression des lignes : 82% cas non - graves (0) ; 18% cas grave (1) 497590 lignes -> 366396 lignes	Suppression des lignes
	obsm	3%	Oui		Suppression des lignes
	choc	3%	Oui		Suppression des lignes
	manv	3%	Oui		Suppression des lignes
Caractéristiques	Int	2 Nan	Oui	NS	Suppression des lignes
	Lum	6 Nan	Oui	NS	Suppression des lignes
	Atm	41 Nan	Oui	NS	Suppression des lignes
	Col	0.75% (3 753)	Oui	Variable cible : Répartition avant la suppression des lignes : 2,59% de décès. Répartition après la suppression des lignes : 2,60%	Suppression des lignes
Usager	secu	-	Non	Suppression de la colonne car devenue redondante avec secu1	Suppression de la colonne
	An_naiss	-	Non	Remplacement de la variable par une variable âge plus pertinente	Suppression de la colonne
	secu2, secu3	99%	Non	Suppression de la colonne si >30% de NaN	Suppression de la colonne
	locp, actp, etap	92%	Non	Suppression de la colonne si >=30% de NaN	Suppression de la colonne
	place	2%	Oui	Remplacement par valeur modale si <30% NaN	Remplacement NaN par valeur modale (1 conducteur)
	sexe	0.06%	Oui	Remplacement par valeur modale si <30% NaN	Remplacement NaN par valeur modale (1 homme)
	trajet	25%	Oui	Remplacement par valeur modale si <30% NaN	Remplacement NaN par valeur modale (2 Promenade loisirs)
	secu1	5%	Oui	Remplacement par valeur modale si <30% NaN	Remplacement NaN par valeur modale (1 ceinture)
	grav	60 lignes	Oui	Suppression des lignes dans la variable cible	Suppression des lignes
Lieux	catr	0.00%	oui	NS	Rien à faire
	voie	16.90%	non	NS	Suppression de colonne
	v1	31.04%	non	NS	Suppression de colonne
	v2	93.25%	non	NS	Suppression de colonne
	circ	0.19%	oui	NS	Suppression des lignes
	nbv	0.24%	oui	NS	Suppression des lignes
	pr	6.71%	non	NS	Suppression de colonne
	pr1	6.83%	non	NS	Suppression de colonne
	vosp	0.24%	oui	NS	Suppression des lignes
	prof	0.20%	oui	NS	Suppression des lignes
	plan	0.20%	oui	NS	Suppression des lignes
	lartpc	94.02%	non	NS	Suppression de colonne
	larrout	46.76%	non	NS	Suppression de colonne
	surf	0.21%	oui	NS	Suppression des lignes
	infra	0.22%	oui	NS	Suppression des lignes
	situ	0.23%	oui	NS	Suppression des lignes
	env1	74.05%	non	NS	Suppression de colonne
	vma	26.16%	non	NS	Suppression de colonne

4.3 Création de variables supplémentaires, et regroupement de catégories

Dans cette section, nous nous concentrerons sur la création de nouvelles variables et le regroupement des catégories dans nos jeux de données. Le regroupement des catégories des variables explicatives présente plusieurs avantages :

- Permet de réduire la taille du data set :

Lorsqu'une variable explicative contient un grand nombre de catégories, cela peut entraîner une augmentation de la taille du data set. Cela peut rendre l'analyse et la modélisation plus complexes, ainsi que nécessiter davantage de ressources. En regroupant les catégories similaires, nous pouvons réduire la taille du data set et simplifier la modélisation pour le ML.

- Peut améliorer la stabilité des modèles :

Lorsque certaines catégories de variables explicatives ont un nombre limité d'observations, elles peuvent avoir un impact disproportionné sur les modèles. En regroupant ces catégories avec des caractéristiques similaires, nous pouvons diminuer cet effet et obtenir des modèles plus stables.

Le regroupement des catégories dans les variables explicatives peut entraîner une perte d'information et avoir un impact sur la précision des prédictions. Les relations entre les variables explicatives et la variable cible peuvent également devenir moins évidentes. Il est important de trouver un équilibre entre la simplification des catégories et la préservation des informations pertinentes.

4.3.1 Data-frame : Véhicules

Comme expliqué dans le paragraphe 3.1.3, les critères de pertinence métier, l'exploration bivariée et la dépendance statistique avec la variable cible ont été utilisés pour sélectionner les variables les plus pertinentes. Seules les variables qui ont été jugées pertinentes selon ces trois critères ont été conservées pour la prochaine étape de modélisation. Ces variables sont présentées dans le tableau ci-dessous :

[manv]	Manœuvre principale avant l'accident	à garder	27 categories
[obs]	Obstacle fixe heurté	à garder	18 categories
[obsm]	Obstacle mobile heurté	à garder	7 categories
[choc]	Point de choc initial	à garder	10 categories
[catv]	Catégorie du véhicule	à garder	31 categories
[Num_acc]	Identifiant de l'accident identique	potentiellement utiles pour créer de nouvelles variables. -> recommandons de les supprimer après	
[id_vehicule]	Identifiant unique du véhicule (depuis 2019)		
[num_veh]	Identifiant du véhicule		

a. Crédit nouvelle variable : Nb_veh

En utilisant les variables Num_acc, id_veh et Num_veh, il est possible de créer une nouvelle variable, nb_veh, qui représente la quantité de véhicules dans chaque accident. Cette variable est quantitative.

En étudiant le graphique 4.3.1, nous constatons que 80% de tous les accidents impliquent 2, 3 ou 4 voitures, mais il y a également des accidents impliquant un grand nombre de voitures, jusqu'à un maximum de 51 voitures. En analysant la répartition de la gravité en fonction du nombre de voitures (fig. 4.3.2), nous remarquons que plus le nombre de voitures augmente, plus le pourcentage de cas graves diminue (45% - 1 voiture, 19% - 2 voitures, 10% - 8 voitures). Pour tous les cas graves (fig. 4.3.2 à droite), la répartition est la suivante : 2 voitures - 48%, 1 voiture - 22%, 3 voitures - 16%.

En conclusion, nous pouvons affirmer que la variable ajoutée (Nb_veh) a un impact sur la variable cible (gravité) et il est important de l'inclure dans la modélisation.

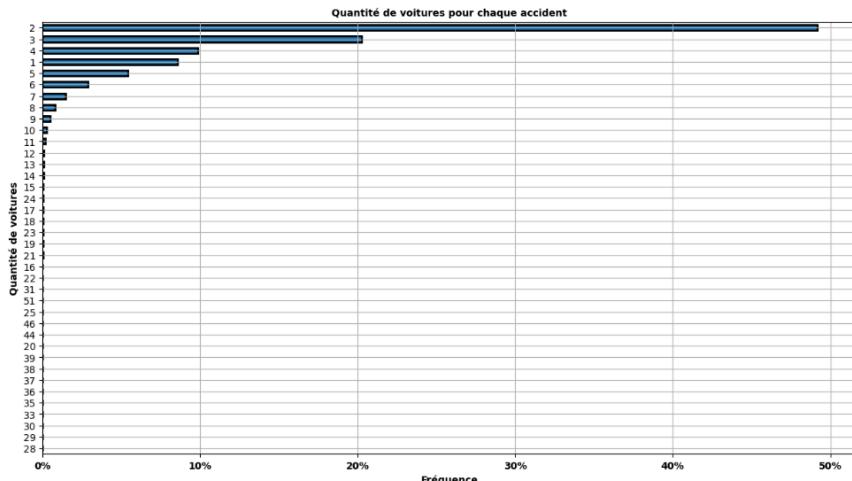


Figure 4.3.1 - Distribution du nombre de voitures dans chaque accident

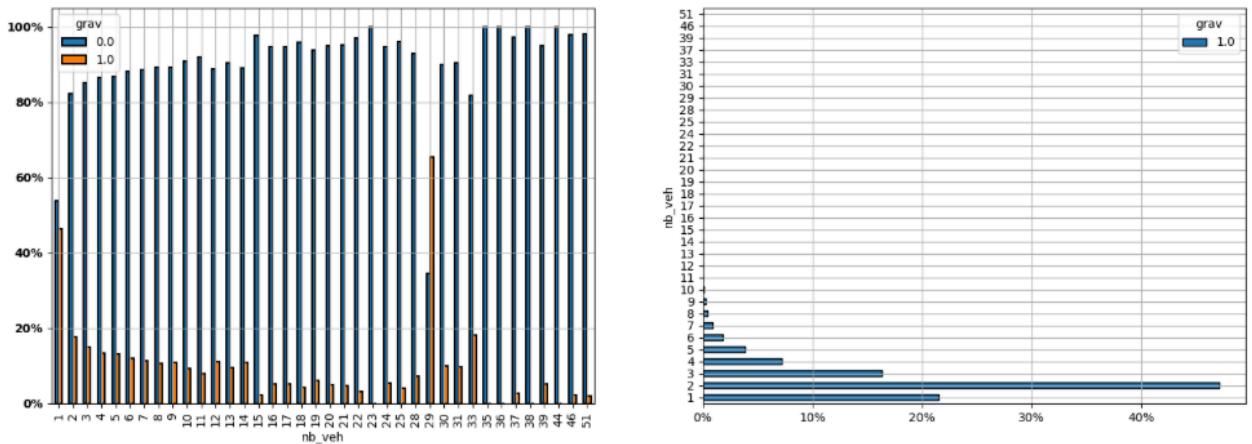


Figure 4.3.2 – Distribution du nombre de voitures pour les cas graves / non graves (à gauche) et uniquement pour les cas graves (à droite)

b. Regroupement des catégories

Dans le data set, nous envisageons de regrouper deux variables : [choc] et [catv].

Variable choc :

Variable : Choc

Point de choc initial :

- 0 – Aucun
- 1 – Avant
- 2 – Avant droit
- 3 – Avant gauche
- 4 – Arrière
- 5 – Arrière droit
- 6 – Arrière gauche
- 7 – Côté droit
- 8 – Côté gauche
- 9 – Chocs multiples (tonneaux)

Variable : Choc

Point de choc initial :

- 0 – Aucun
- 1 – Avant
- 2 – Arrière
- 3 – Côté droit
- 4 – Chocs multiples (tonneaux)

Cela nous permet de réduire le nombre de catégories de 10 à 5 pour la variable [choc].

Variable catv :

Catégorie du véhicule :

00 – Indéterminable 01 – Bicyclette 02 – Cyclomoteur <50cm3 03 – Voiturette (Quadricycle à moteur carrossé) (anciennement "voiturette ou tricycle à moteur") 04 – Référence inutilisée depuis 2006 (scooter immatriculé) 05 – Référence inutilisée depuis 2006 (motocyclette) 06 – Référence inutilisée depuis 2006 (side-car) 07 – VL seul 08 – Référence inutilisée depuis 2006 (VL + caravane) 09 – Référence inutilisée depuis 2006 (VL + remorque) 10 – VU seul 1,5T <= PTAC <= 3,5T avec ou sans remorque (anciennement VU seul 1,5T <= PTAC <= 3,5T) 11 – Référence inutilisée depuis 2006 (VU (10) + caravane) 12 – Référence inutilisée depuis 2006 (VU (10) + remorque) 13 – PL seul 3,5T <PTCA <= 7,5T 14 – PL seul > 7,5T 15 – PL > 3,5T + remorque 16 – Tracteur routier seul 17 – Tracteur routier + semi-remorque 18 – Référence inutilisée depuis 2006 (transport en commun) 19 – Référence inutilisée depuis 2006 (tramway) 20 – Engin spécial 21 – Tracteur agricole 30 – Scooter < 50 cm3 31 – Motocyclette > 50 cm3 et <= 125 cm3 32 – Scooter > 125 cm3 33 – Quad léger <= 50 cm3 (Quadricycle à moteur non carrossé) 36 – Quad lourd > 50 cm3 (Quadricycle à moteur non carrossé) 37 – Autobus 38 – Autocar 39 – Train 40 – Tramway 41 – 3RM <= 50 cm3 42 – 3RM > 50 cm3 <= 125 cm3 43 – 3RM > 125 cm3 44 – EDP à moteur 60 – EDP sans moteur 80 – VAE 31 – Motocyclette > 50 cm3 et <= 125 cm3 32 – Scooter > 50 cm3 et <= 125 cm3 33 – Motocyclette > 125 cm3 34 – Scooter > 125 cm3 35 – Quad lourd > 50 cm3 (Quadricycle à moteur non carrossé) 36 – Quad lourd > 50 cm3 (Quadricycle à moteur non carrossé) 42 – 3RM > 50 cm3 <= 125 cm3 43 – 3RM > 125 cm3 13 – PL seul 3,5T <PTCA <= 7,5T 14 – PL seul > 7,5T 15 – PL > 3,5T + remorque 16 – Tracteur routier seul 17 – Tracteur routier + semi-remorque 20 – Engin spécial 21 – Tracteur agricole 11 – Référence inutilisée depuis 2006 (VU (10) + caravane) 12 – Référence inutilisée depuis 2006 (VU (10) + remorque)

0 cat Indéterminable	1 cat Véhicules carrossés <3,5T	2 cat Véhicules non carrossés Pas puissantes (<=50cm3)	3 cat Véhicules non carrossés puissantes (>50cm3)	4 cat PL, tracteurs et etc	5 cat Transport en commun	6 cat Autre véhicule
00 – Indéterminable	07 – VL seul 10 – VU seul 1,5T <= PTAC <= 3,5T avec ou sans remorque (anciennement VU seul 1,5T <= PTAC <= 3,5T) 03 – Voiturette (Quadricycle à moteur carrossé) (anciennement "voiturette ou tricycle à moteur") 08 – Référence inutilisée depuis 2006 (VL + caravane) 09 – Référence inutilisée depuis 2006 (VL + remorque)	01 – Bicyclette 02 – Cyclomoteur <50cm3 50 – EDP à moteur 60 – EDP sans moteur 80 – VAE 30 – Scooter < 50 cm3 35 – Quad léger <= 50 cm3 (Quadricycle à moteur non carrossé) 04 – Référence inutilisée depuis 2006 (scooter immatriculé) 05 – Référence inutilisée depuis 2006 (motocyclette) 06 – Référence inutilisée depuis 2006 (side-car) 41 – 3RM <= 50 cm3	31 – Motocyclette > 50 cm3 et <= 125 cm3 32 – Scooter > 50 cm3 et <= 125 cm3 33 – Motocyclette > 125 cm3 34 – Scooter > 125 cm3 35 – Quad lourd > 50 cm3 (Quadricycle à moteur non carrossé) 42 – 3RM > 50 cm3 <= 125 cm3 43 – 3RM > 125 cm3	13 – PL seul 3,5T <PTCA <= 7,5T 14 – PL seul > 7,5T 15 – PL > 3,5T + remorque 16 – Tracteur routier seul 17 – Tracteur routier + semi-remorque 20 – Engin spécial 21 – Tracteur agricole 11 – Référence inutilisée depuis 2006 (VU (10) + caravane) 12 – Référence inutilisée depuis 2006 (VU (10) + remorque)	37 – Autobus 38 – Autocar 39 – Train 40 – Tramway 18 – Référence inutilisée depuis 2006 (transport en commun) 19 – Référence inutilisée depuis 2006 (tramway)	99 – Autre véhicule

Cela nous permet de réduire le nombre de catégories de 31 à 7 pour la variable catv. Les nouvelles catégories sont formées en fonction du type et de la protection du véhicule : 0 - Indéterminable, 1 - Véhicules carrossés <3,5T, 2 - Véhicules non carrossés pas puissants (<=50cm3), 3 - Véhicules non carrossés puissants (>50cm3), 4 - PL, tracteurs et autres, 5 - Transport en commun, 6 - Autre véhicule.

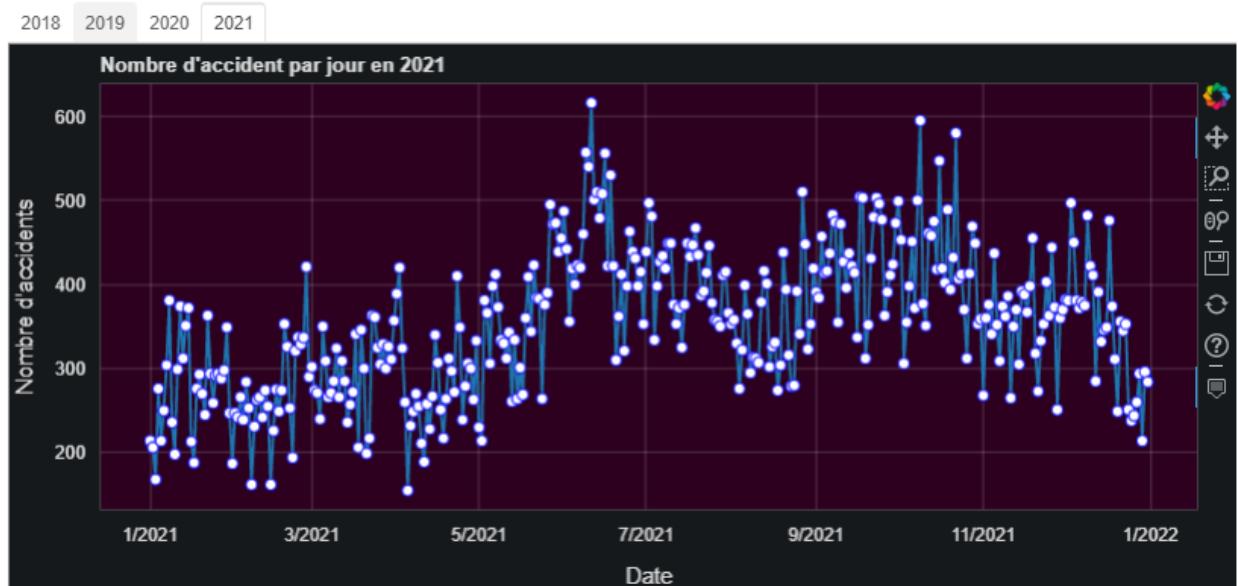
c. Conclusions

Nom variable	Description	Info sur variable
Num_Acc	Identifiant de l'accident identique	Passer à l'index
obs	Obstacle fixe heurté	Qualitative, 18 cat. Nan(0%), cardinal
obsm	Obstacle mobile heurté	Qualitative, 18 cat. Nan(0%), cardinal
choc	Point de choc initial	Qualitative, Nan(0%), cardinal 10 cat. Initial -> 5 categories
manv	Maneuvre principale avant l'accident	Qualitative, 27 cat. Nan, (0%), cardinal
catv	Catégorie du véhicule	Qualitative, Nan(0%), cardinal 31 cat. Initial -> 7 categories
nb_veh	Quantité de véhicule dans chaque accidents	Quantitative

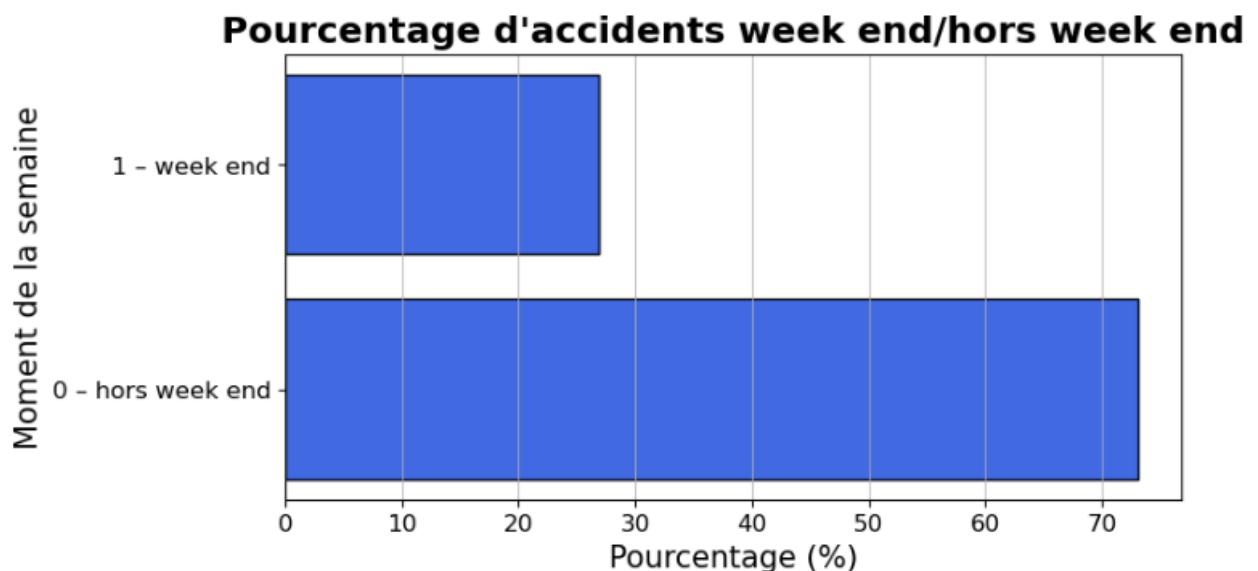
4.3.2 Data-frame : Caractéristiques

a. Variable date

L'intérêt de cette variable est double : d'une part elle doit nous permettre de créer la variable week-end et d'autre part elle nous permet de créer un diagramme très visuel sur bokeh

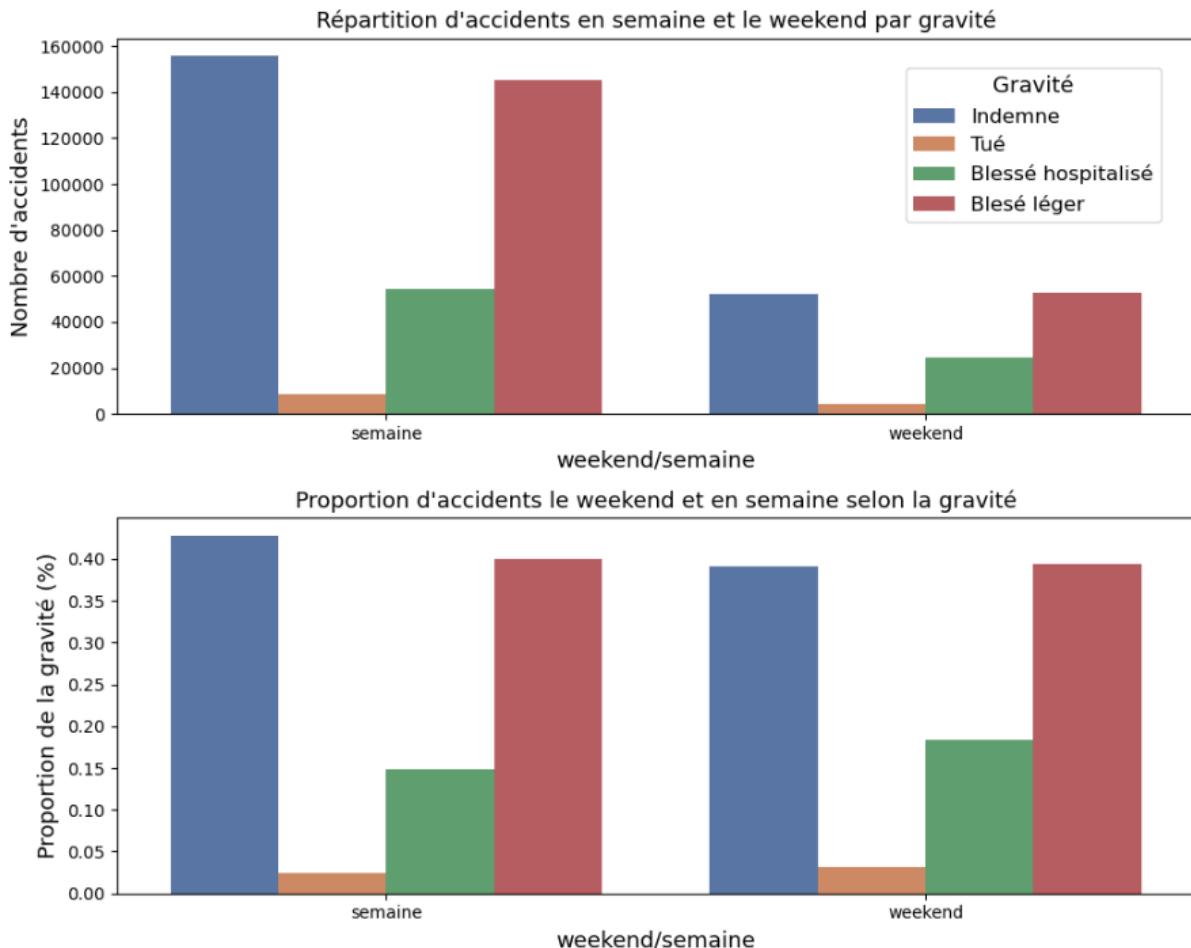


b. Variable Week-end



- Le graphique illustre la distribution des accidents en fonction du moment de la semaine - en semaine ou le week-end.
- Près des trois quarts (73,11%) des accidents se produisent en semaine, indiquant que la majorité se produisent pendant les jours de travail.

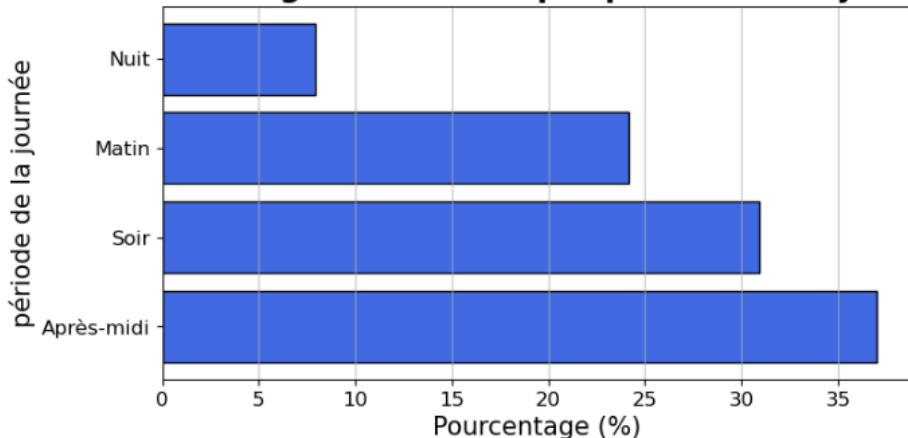
- Environ le quart (26,89%) des accidents ont lieu pendant le week-end. Ce chiffre suggère que, malgré une diminution générale du trafic par rapport aux jours de semaine, un nombre significatif d'accidents se produit encore pendant ces jours.



- Ces deux graphiques permettent de visualiser et de comprendre la répartition des accidents ainsi que leur gravité, en fonction du fait qu'ils se produisent en semaine ou pendant le week-end.
- Le premier graphique représente le nombre total d'accidents en fonction du moment de la semaine, avec pour chaque barre la répartition suivant le niveau de gravité.
- Le deuxième graphique montre la proportion en pourcentage des différents niveaux de gravité selon qu'ils se produisent en semaine ou pendant le week-end.
- La catégorie "week-end" présente un taux de mortalité légèrement plus élevé (3,14%) par rapport à la catégorie "en semaine" (2,39%). Les accidents qui se produisent pendant le week-end sont légèrement plus susceptibles d'être fatals.
- En revanche, la majorité des accidents, qu'ils se produisent en semaine ou le week-end, entraînent des blessures mineures ou aucune blessure. En effet, les taux de blessures mineures ou d'absence de blessures sont respectivement de 39,92% et 42,77% en semaine, et de 39,33% et 39,15% le week-end.

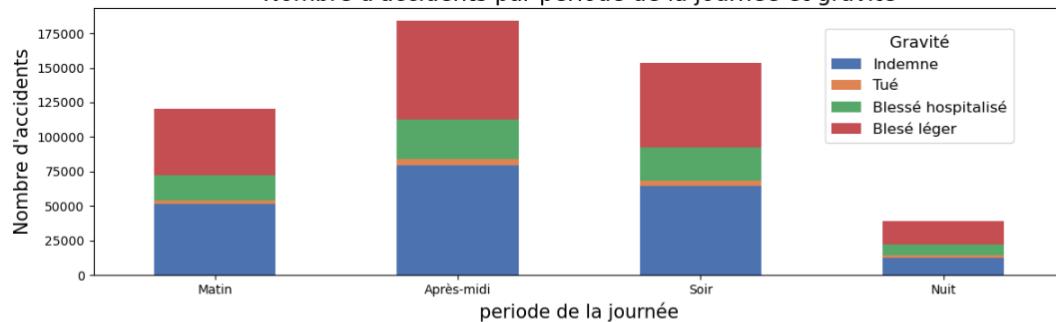
c. Variable période de la journée

Pourcentage d'accidents par période de la journée

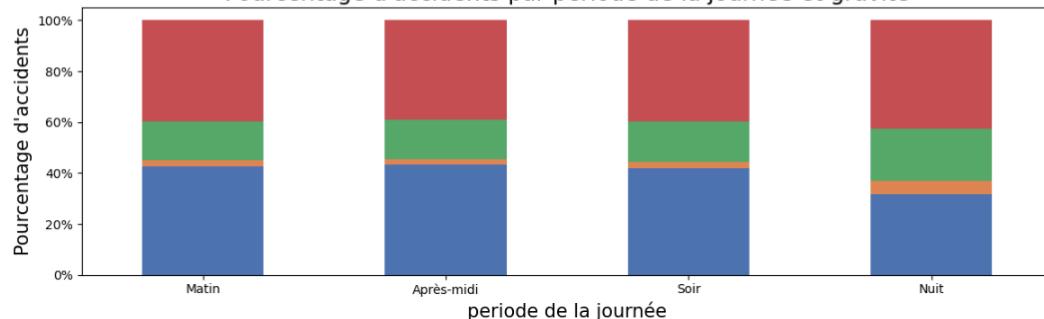


- L'illustration décrit la répartition des accidents selon les différentes tranches horaires de la journée.
- L'après-midi enregistre le plus grand nombre d'accidents, avec une représentation de 37,01%.
- Le soir suit avec 31,22% des accidents, suggérant que le changement de luminosité et la possible augmentation de la fatigue peuvent contribuer à une fréquence accrue des accidents.
- Les accidents survenant le matin constituent une part plus faible, avec 24,40% de l'ensemble des accidents.
- Malgré une visibilité potentiellement réduite, la période nocturne ne comptabilise qu'une petite part des accidents (7,96%). Ceci pourrait s'expliquer par une réduction significative du trafic routier pendant ces heures.

Nombre d'accidents par période de la journée et gravité



Pourcentage d'accidents par période de la journée et gravité



- Ces diagrammes ont pour objectif de visualiser et de comprendre la distribution des accidents et leur gravité en fonction des différentes périodes de la journée.
- Le premier graphique présente le nombre total d'accidents en fonction de la période de la journée, chaque barre décrivant la répartition selon le degré de gravité.

- Le second graphique expose la proportion en pourcentage des divers niveaux de gravité par période de la journée.
- Notamment, la période de la "nuit" affiche le taux de mortalité le plus élevé à près de 5%. Par contre, la période de "l'après-midi", a un taux de mortalité plus faible, seulement à 2.23%.

d. Regroupement de catégorie

Intersection

grav	1.0	2.0	3.0	4.0
int				
8.0	0.429963	0.095006	0.144945	0.330085
1.0	0.402498	0.032143	0.172563	0.392796
9.0	0.426099	0.018795	0.149663	0.405443
4.0	0.438734	0.017375	0.140999	0.402893
6.0	0.433910	0.015991	0.153445	0.396654
3.0	0.454205	0.013701	0.137891	0.394202
2.0	0.450158	0.012610	0.122707	0.414525
5.0	0.447622	0.009653	0.097962	0.444762
7.0	0.459300	0.006446	0.071639	0.462615

Répartition de la gravité des accidents en fonction des différentes catégories d'intersections

En se basant sur le tableau de la gravité des accidents, en particulier en ce qui concerne le nombre de décès, et en tenant compte de la sémantique des catégories, il est possible de réduire le nombre de catégories. Cette démarche simplifie l'analyse tout en conservant les informations essentielles sur la gravité des accidents.

Variable : int (intersection)
Avant modification

- 1 – Hors intersection
- 2 – Intersection en X
- 3 – Intersection en T
- 4 – Intersection en Y
- 5 - Intersection à plus de 4 branches
- 6 - Giratoire
- 7 - Place
- 8 – Passage à niveau
- 9 – Autre intersection

Variable : int (intersection)
Après modification

- 1 - Hors intersection
- 2 - Passage à niveau
- 3 - Intersection et autre



Lumière

grav	1.0	2.0	3.0	4.0
lum				
3.0	0.337940	0.064544	0.231135	0.366381
4.0	0.383061	0.035876	0.156222	0.424841
2.0	0.409703	0.028829	0.174250	0.387218
1.0	0.432990	0.021913	0.152874	0.392222
5.0	0.407926	0.017079	0.130844	0.444151

Répartition de la gravité des accidents en fonction des différentes catégories de lumières

En se basant sur le tableau de la gravité des accidents, en particulier en ce qui concerne le nombre de décès, et en tenant compte de la sémantique des catégories, il est possible de réduire le nombre de catégories. Cette démarche simplifie l'analyse tout en conservant les informations essentielles sur la gravité des accidents.

Variable : lum (lumière)

Avant modification

- 1 - plein jour
- 2 - crépuscule ou aube
- 3 - nuit sans éclairage public
- 4 - nuit avec éclairage public non allumé
- 5 - nuit avec éclairage public allumé

Variable : lum (lumière)

Après modification

- 1 - Conditions lumineuse optimale
- 2 - crépuscule ou aube
- 3 - nuit sans éclairage

Collision

grav	1.0	2.0	3.0	4.0
col				
7.0	0.156881	0.049980	0.321712	0.471426
6.0	0.351136	0.045181	0.212767	0.390916
1.0	0.360363	0.044910	0.234907	0.359821
5.0	0.524119	0.021956	0.109959	0.343966
3.0	0.465794	0.011222	0.123104	0.399881
2.0	0.467361	0.010424	0.086705	0.435510
4.0	0.579282	0.006567	0.041639	0.372512

Répartition de la gravité des accidents en fonction des différentes catégories de collision

En se basant sur le tableau de la gravité des accidents, en particulier en ce qui concerne le nombre de décès, et en tenant compte de la sémantique des catégories, il est possible de réduire le nombre de catégories. Cette démarche simplifie l'analyse tout en conservant les informations essentielles sur la gravité des accidents.

Variable : col (collision)

Avant modification

- 1 – Deux véhicules - frontale
- 2 – Deux véhicules - par l'arrière
- 3 – Deux véhicules - par le côté
- 4 – Trois véhicules et plus - en chaîne
- 5 – Trois véhicules et plus - collisions
- 6 – Autre collision
- 7 – Sans collision

Variable : col (collision)

Après modification

- 1 - Sans collision
- 2 - Autres collisions
- 3 - Deux véhicules frontales
- 4 - Autres collisions entre véhicules

Atmosphère

grav	1.0	2.0	3.0	4.0
atm				
5.0	0.345833	0.062179	0.239103	0.352885
9.0	0.354077	0.055365	0.209871	0.380687
6.0	0.334387	0.055336	0.266403	0.343874
7.0	0.444746	0.034576	0.227006	0.293672
8.0	0.410641	0.031405	0.148446	0.409508
3.0	0.403917	0.028456	0.165027	0.402600
4.0	0.440064	0.027353	0.151649	0.380933
1.0	0.420212	0.025336	0.159180	0.395272
2.0	0.410433	0.022091	0.136264	0.431212

Répartition de la gravité des accidents en fonction des différentes catégories d'atmosphère

Variable : atm (atmosphère)
Avant modification

- 1 – Normale (79.33%)
- 2 – Pluie légère (11.03%)
- 3 – Pluie forte (2.29%)
- 4 – Neige - grêle (0.50%)
- 5 – Brouillard - furnée (0.63%)
- 6 – Vent fort - tempête (0.25%)
- 7 – Temps éblouissant (0.78%)
- 8 – Temps couvert (3.72%)
- 9 – Autre (0.47%)

Variable : atm (atmosphère)
Après modification

- 1 - Normale
- 2 - Pluie
- 3 - Neige-grêle
- 4 - Brouillard - furnée
- 5 - vent fort - tempête

e. Récapitulatif

Nom variable	Description	Info sur variable	
mois	Mois de l'accident	Qualitative nominale, 12 cat. Nan : 0%	
lum	Lumière : conditions d'éclairage dans lesquelles l'accident se produit	Qualitative nominale, 3 cat. Nan : 0%	
agg	Localisation	Qualitative nominale, 2 cat. Nan : 0%	
int	Intersection	Qualitative nominale, 3 cat. Nan : 0%	
atm	Conditions atmosphériques	Qualitative nominale, 5 cat. Nan : 0%	
col	Type de collision	Qualitative nominale, 4 cat. Nan : 0%	
periode	Moment de la journée de l'accident	Qualitative nominale, 4 cat. Nan : 0%	
week end	Accident le week end ou en semaine	Qualitative nominale, 2 cat. Nan : 0%	
lat	Latitude	Quantitative continue (2018 fail)	Pour graph
long	Longitude	Quantitative continue (2018 fail)	Pour graph
Date	Date de l'accident	Quantitative continue, Nan : 0%	Pour graph

4.3.3 Data-frame : Usager

a. Crédation nouvelle variable : âge (âge de l'usager)

A partir de la variable année de naissance, calcul de l'âge de l'usager.

b. Encodage de la variable cible : mise sous forme ordinale

Mise sous forme ordinale des valeurs de la variable cible “grav”

Variable	Catégories initiales	Catégories ordinales
grav	1 Indenme 2 Tué 3 Blessé hospitalisé 4 Blessé léger	1 Indenme 2 Blessé léger 3 Blessé hospitalisé 4 Tué

c. Regroupement de catégories

Les catégories de certaines variables ont été regroupées pour donner des variables un peu moins déséquilibrées.

Variable	Catégories initiales	Regroupement catégories
place	1.0 conducteur 0.739840 2.0 avant droite 0.113382 10.0 piéton 0.058244 NaN 0.022060 3.0 arrière droit 0.019009 4.0 arrière gauche 0.016811 9.0 milieu droit 0.010448 7.0 milieu gauche 0.009488 5.0 arrière milieu 0.005036 8.0 milieu milieu 0.004401 6.0 devant milieu 0.001280	1.0 conducteur 0.739840 3.0 autre 0.1247182.0 2.0 avant droite 0.113382 NaN 0.022060
trajet	5.0 Promenade loisirs 0.37 NaN 0.2536201 Domicile travail 0.1356734 Utilisation prof 0.099.0 Autre 0.083.0 Courses – achats 0.032.0 Domicile – école 0.378634	2.0 Promenade loisirs 0.378634 NaN 0.253620 3.0 Autre 0.232073 1.0 Domicile travail 0.135673
secu1	1 Ceinture 0.594682 2 Casque 0.186433 8 Non déterminable 0.090411 NaN 0.059822 0 Aucun équipement 0.058258 3 Dispo enfants 0.006867 9 Autre 0.001803 4 Gilet réfléchis 0.000707 6 Gants 0.000607 5 Airbag 0.000390 7 Gants + airbags 0.000020	1 Ceinture 0.594682 2 Casque 0.186433 3 Non déterminable 0.090411 NaN 0.059822 0 Aucun équipement 0.0582584 Autre 0.010394

4.3.4 Data-frame : Lieux

Nom variable	Description	Info sur variable
catr	catégorie de route	fusion des catégories 5,6,7 et 9 en > catégorie 5
nbv	nombre total de voies de circulation	fusion du nombre de voies supérieur 9 en > catégorie 9 pour 9+
surf	état de la surface	fusion des catégories 0, 3, 4, 5, 6, 7, 8, 9 en > catégorie 3 pour autres

4.4 Gestion des doublons dans le data frame

Après avoir finalisé le jeu de données pour la modélisation, il est essentiel de détecter et éliminer les doublons afin de prévenir toute influence indue sur les résultats. Cette étape est cruciale pour assurer la qualité et la précision des données utilisées dans le processus de modélisation. En conséquence, tous les doublons ont été supprimés du jeu de données final.

4.5 Encodage des variables catégorielles

Dans notre ensemble de données, nous avons principalement des variables qualitatives qui nécessitent un encodage approprié. Initialement, ces variables étaient encodées en utilisant une numérotation où chaque catégorie de variable recevait une valeur numérique unique. Il est important de noter que les catégories ne présentent pas de relation d'ordre réelle entre elles. Ainsi, il est crucial d'identifier correctement les variables qui sont ordinaires (encodées par numérotation) et celles qui ne le sont pas.

La variable cible, qui représente la gravité des accidents, initialement ordonnée, a été transformée en une variable cardinale.

Pour les variables non ordinaires, nous avons utilisé la méthode de l'encodage en variables indicatrices en utilisant la technique du One Hot Encoding. Pour réaliser cet encodage, nous avons utilisé la fonction `pd.get_dummies()` de la librairie Pandas. Dans le but de réduire le nombre de colonnes indicatrices créées, nous avons spécifié l'argument `drop_first=True`.

Cela nous a permis de retirer pour chaque variable catégorielle la première modalité, de telle sorte que cette première modalité devient la modalité de base à laquelle les autres modalités de la variable catégorielle sont comparées.

4.6 Data-frame final

		Nom variable	Description	Type	Info sur variable	Nombre de colonnes dans data set final après encodage
Véhicules	1	Num_Acc	Identifiant de l'accident identique	Int64	Passer à l'index	-
	2	obs	Obstacle fixe heurté	Float64	Qualitative, Cardinal, 18 catégories	17 - colonnes – [1]
	3	obsm	Obstacle mobile heurté	Float64	Qualitative, Cardinal, 7 catégories	6 colonnes – [1]
	4	choc	Point de choc initial	Float64	Qualitative, Cardinal, 5 catégories	4 colonnes – [1]
	5	manv	Maneuvre principale avant l'accident	Float64	Qualitative, Cardinal, 27 catégories	26 colonnes – [1]
	6	catv	Catégorie du véhicule	Int64	Qualitative, Cardinal, Nan(0%) 7 catégories	6 colonnes – [1]
	7	nb_veh	Quantité de véhicule dans chaque accidents (Variable ajoutée)	Int64	Quantitative	1
Caractéristiques	8	mois	Mois de l'accident		Qualitative nominale, 12 cat.	11
	9	lum	Lumière : conditions d'éclairage dans lesquelles l'accident se produit		Qualitative nominale, 3 cat.	2
	10	agg	Localisation		Qualitative nominale, 2 cat.	1
	11	int	Intersection		Qualitative nominale, 3 cat.	2
	12	atm	Conditions atmosphériques		Qualitative nominale, 5 cat.	4
	13	col	Type de collision		Qualitative nominale, 4 cat.	3
	14	periode	Moment de la journée de l'accident		Qualitative nominale, 4 cat.	3
	15	week end	Accident le week end ou en semaine		Qualitative nominale, 2 cat.	1
	-	lat	Latitude		Quantitative continue (2018 fail)	Pour graphique
	-	long	Longitude		Quantitative continue (2018 fail)	Pour graphique
	-	Date	Date de l'accident		Quantitative continue	Pour graphique

Lieux	16	catr	catégorie de route	Int64	Qualitative, Cardinal, 5 catégories	5 colonnes – [1]
	17	circ	régime de circulation	Float64	Qualitative, Cardinal, 5 catégories	5 colonnes – [1]
	18	nbv	nombre total de voies de circulation	Float64	Qualitative, Cardinal, Nan(0%) 10 catégories	10 colonnes – [1]
	19	vosp	existence d'une voie réservée	Float64	Qualitative, Cardinal, 4 catégories	4 colonnes – [1]
	20	prof	Profil en long décrit la déclivité	Float64	Qualitative, Cardinal, 5 catégories	5 colonnes – [1]
	21	plan	tracé en plan, courbe ou rectiligne	Float64	Qualitative, Cardinal, 5 catégories	5 colonnes – [1]
	22	surf	état de la surface	Float64	Qualitative, Cardinal, 3 catégories	3 colonnes – [1]
	23	infra	aménagement de l'infrastructure	Float64	Qualitative, Cardinal, 10 catégories	10 colonnes – [1]
	24	situ	situation de l'accident	Float64	Qualitative, Cardinal, 8 catégories	8 colonnes – [1]
	25	place	Place de l'usager	float64	Qualitative, 3 catégories	2 colonnes – [1]
Usager	26	catu	Catégorie de l'usager	int64	Qualitative, 3 catégories	2 colonnes – [1]
	27	sexe	Sexe de l'usager	float64	Qualitative, 2 catégories	1 colonnes – [1]
	28	trajet	Type de trajet lors de l'accident	float64	Qualitative, 6 catégories	5 colonnes – [1]
	29	secu1	Equipement de sécurité 1 de l'usager	float64	Qualitative, 5 catégories	4 colonnes – [1]
	30	age	Age de l'usager	float64	Quantitative, variable continue	1
	31	grav	Gravité de l'accident pour l'usager	float64	Qualitative, 4 catégories	3 colonnes – [1]