DataScientest • com

Décembre 2023 – Juillet 2024

Accidents routiers en France

Projet mené par :
Matthieu Claudel
Vanessa Ibert
Camille Pelat
Nadège Reboul

Table des matières

Table des illustrations 2

Introduction au projet 4

Contexte 4

Objectifs 5

Compréhension et manipulation des données 8

Cadre 8

Pertinence 10

Pre-processing et feature engineering 11

Exemples détaillés avec visualisation 17

Analyse des variables continues 29

Age_usager 29

Heure et mois 29

Latitude et longitude 29

Visualisations et Statistiques 31

Evolution Temporelle 31

Disparités spatiales 36

Influence de l'âge et du sexe dans l'accidentalité routière 37

Conclusion 40

Références bibliographiques 42

Table des illustrations

Figure 1 : Liste des variables présentes dans les bases de données mises à disposition du grand public 9
Figure 2 (a). Nombre d'accidents par an, (b). Nombre de tués par an 10
Figure 3 : Pourcentage d'accidents selon la modalité pour les années 2019 à 2022 11
Figure 4 : Pourcentage de valeurs manquantes pour chaque variable 12
Figure 5 : Proportion de chaque modalité de la variable grav avant et après traitement 13
Figure 6 : Liste des variables présentes après traitement du jeu de données 15
Figure 7 : Valeurs de V_Cramer pour chaque variable par ordre d'importance 16
Figure 8 : Regroupement des catégories pour la variable catv 17
Figure 9 : Répartition de la gravité en fonction de la catégorie de véhicule 18
Figure 10 : Valeurs du V_Cramer pour les modalités de la catégorie de véhicule initiales et recodées 18
Figure 11 : Numérotation des places dans la variable place 19
Figure 12 : (a). Comparaison du χ^2 et du V de Cramer pour les variables prox_pt_choc et choc, (b). Tableau de contingence de la variable prox_pt_choc avec la variable cible 20
Figure 13 : Répartition des modalités de la variable place 21
Figure 14 : χ^2 et V de Cramer pour les différentes modalités (a) de la variable place, (b). de la variable place_rec. 21
Figure 15 : Corrélation entre les différentes modalités de place_rec et de la variable cible. Les niveaux de gravité 1, 2, 3 et 4 correspondent respectivement à indemne, tué, blessé hospitalisé et blessé léger. 22
Figure 16 : Dichotomisation des équipements de sécurité 23
Figure 17 : χ2 et V de Cramer pour les différentes variables équipement et catégories de véhicule recodées 23
Figure 18 : Corrélation entre les différentes variables équipements, les différentes modalités de véhicule recodées et de gravité 24
Figure 19 : (a). Proportion de la gravité selon les conditions atmosphériques, (b). χ2 et V de Cramer pour les différentes modalités de la variable atm 25
Figure 20 : (a). Corrélation entre la variable atm et les différentes modalités de la variable cible, (b). χ2 et V de Cramer pour la variable atm recodée en binaire. 26
Figure 21 : Regroupement des catégories pour la variable manv 27
Figure 22 : Proportion de la gravité selon la manoeuvre principale avant l'accident 28
Figure 23 : (a). Corrélation entre les différentes modalités de la variable manv et de la variable cible, (b). χ2 et V de Cramer pour les différentes modalités de la variable manv 28
Figure 24 : Distribution de age_usager. (a). Histogramme des valeurs, (b). Graphique des quantiles, (c). Boîte à moustaches 29
Figure 25 : Distributions des variables lat et long 30
Figure 26 : Boîtes à moustaches des variables lat et long 30
Figure 27 : Distributions des variables lat et long, pour la métropole uniquement 30
Figure 28 : Boîtes à moustaches des variables lat et long pour la métropole uniquement 30
Figure 29 : Proportion mensuelle de chaque classe de gravité 31
Figure 30 : Courbes du nombre d'usagers pour chaque classe de gravité selon le mois pour les années 2019 à 2022 32
Figure 31 : Proportion des différents modalités de gravité selon s'il s'agit d'un jour de vacances et jours fériés ou non 32
Figure 32 : χ^2 et V de Cramer pour la variable jour_chome 32

Figure 33 : Courbes du nombre d'usagers pour chaque classe de gravité selon le jour de la semaine pour les années 2019 à 2022 33
Figure 34 : Tableaux de contingence pour la variable weekend (a) excluant le vendredi, (b) incluant le vendredi 33
Figure 35 : pour les variables (a). week-end sans le vendredi, (b). weekend avec le vendredi, (c) jour_semaine 34
Figure 36 : Répartition de la gravité pour la variable weekend 34
Figure 37 : Courbes du nombre d'usagers pour chaque classe de gravité selon l'heure pour les années 2019 à 2022 35
Figure 38 : Carte de la localisation des accidents selon la gravité en France métropolitaine 36
Figure 39 : Cartes (a) de la répartition des états de gravité des accidents par région, (b) des proportions de victimes décédées selon le département de France métropolitaine. La taille des camemberts est proportionnelle au nombre d'usagers impliqués dans les accidents de chaque région. 37
Figure 40 : Graphique de la répartition des modalités de la gravité en fonction de l'âge 37
Figure 41 : Graphique de la mortalité des accidents de la route selon le sexe. Le calcul des proportions se base sur 9556 hommes tués, contre 2645 femmes (soit une mortalité masculine 3 à 4 fois supérieure). 38
Figure 42 : Proportions, par classes d'âge, des catégories de véhicules associées aux usagers décédés 39
Figure 43 : Comparaison de la position de la personne accidentée selon le sexe 39

Contexte

La France dispose d'infrastructures routières particulièrement importantes avec environ 1,7 millions de kilomètres de routes en 2021 déclinées comme ceci (Routes de France 2023):

Longueur du réseau routier français métropolitain (2021)

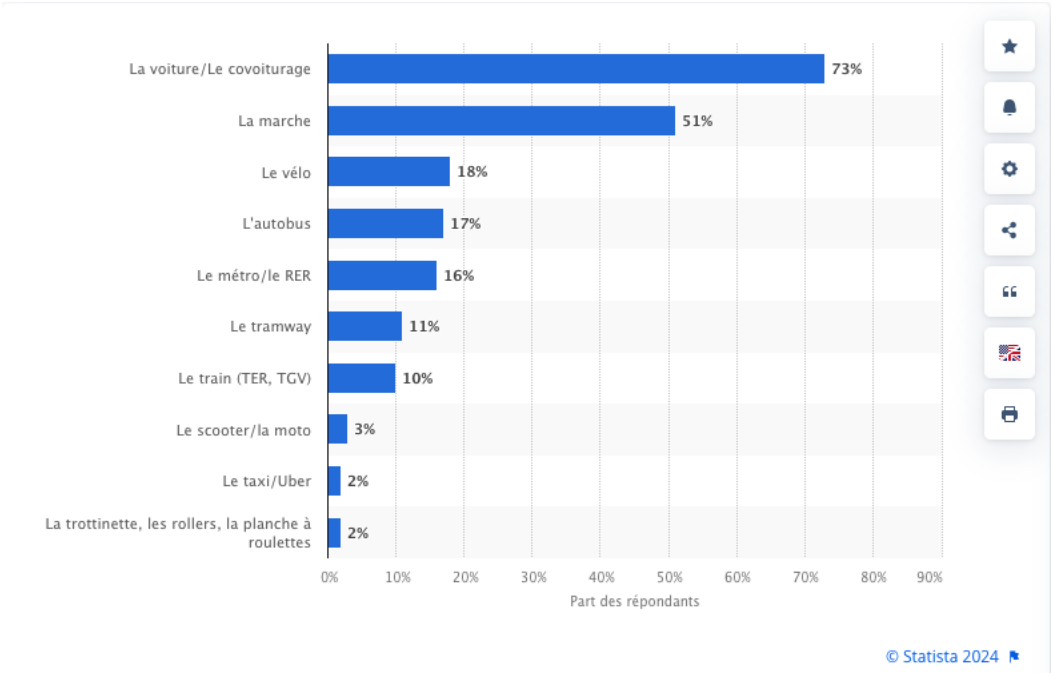
	Km	% du trafic	Observations
Autoroutes concédées	9 221	16 %	dont 2 372 Km à 2 x 3 voies
Autoroutes non concédées	3 309	16 %	
Routes nationales	8 380	4 %	dont environ 2 836 km à chaussées séparées
Routes départementales	378 834	} 64 %	dont environ 1 500 km à chaussées séparées
Routes communales et rues	705 000		
Total	**1 104 744**		
Chemins ruraux	env. 600 000 km		

Sources : Cerema , ASFA , SDES.

Au 1er janvier 2023, sur ces routes circulait le parc automobile français suivant (Ministère de la transition écologique et de la cohésion des territoires 2024) :

- 38,9 millions de voitures particulières,
- 6,4 millions de véhicules utilitaires légers (VUL),
- 620 000 poids lourds,
- 94 000 autobus et autocars.

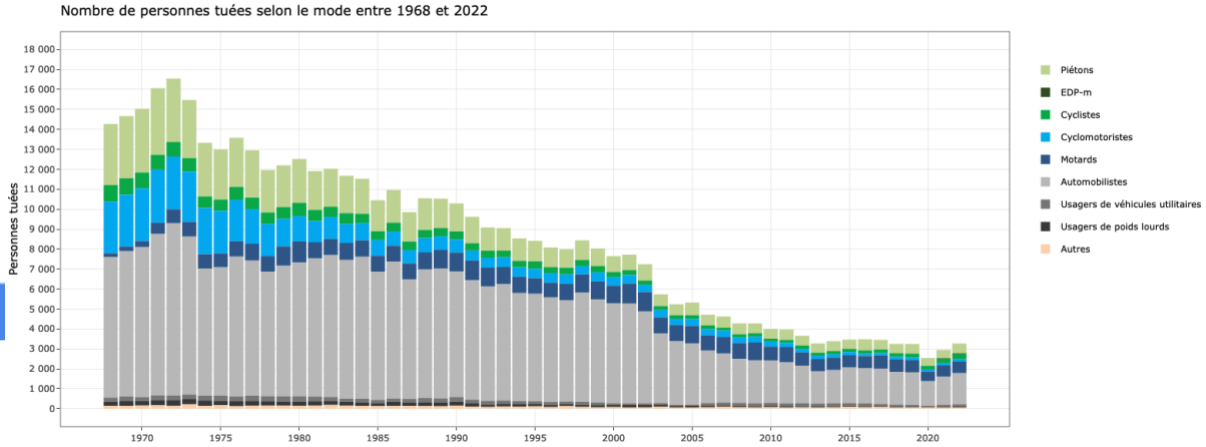
Les Français n'utilisent pas uniquement ces types de véhicules pour leurs déplacements. En effet, les moyens de transport utilisés par les Français pour les déplacements quotidiens en 2023 sont (Statista Research Department 2024) :



Malheureusement, chaque année, de nombreux accidents de la route se produisent. D'ailleurs, le bilan 2023 provisoire de la sécurité routière est le suivant (Observatoire national interministériel de la sécurité routière 2024) :

- **3 402 personnes sont décédées** en 2023 sur les routes de France métropolitaine ou d'outre-mer (estimation ONISR au 22/01/2024),
- **232 000 personnes ont été blessée**s en 2023 sur les routes de France métropolitaine, dont **16 000 gravement**, d'après la méthode d'estimation ONISR-Université Gustave Eiffel (Registre du Rhône).

En regardant plus précisément le nombre de personnes tuées selon le mode de transport utilisé, on voit une évolution du nombre de morts qui a diminué depuis 1968 pour stagner depuis 2013 (hors année 2020 impactée par le covid). De plus, on se rend compte que la mortalité est différente selon le mode de transport (CEREMA 2024):



Objectifs

Le projet a pour but de **prédire la gravité des accidents routiers en France** d'après des données historiques.

Plus spécifiquement, l'objectif ici sera de **prédire la catégorie de gravité de l'accident** (indemne, blessé léger, blessé hospitalisé, décès) pour chaque usager entré dans la base de données, en fonction de caractéristiques individuelles (âge, sexe, utilisation d'équipements de sécurité, place dans le véhicule…), des caractéristiques de son mode de transport (voiture, 2-roues motorisé, vélo, transport en commun…), des caractéristiques du lieu de l'accident (type de voie, intersection, double-sens…), et de caractéristiques contextuelles (date, heure, luminosité, météo…).

Dans un deuxième temps, **le poids de chaque facteur dans la classification sera étudié**, afin de classer les facteurs par ordre d'importance : qu'est-ce qui fait qu'une personne va être plus sévèrement atteinte lors d'un accident routier ? Cette étape *d'interprétabilité* du modèle permettra de dégager les principaux axes d'amélioration de la sécurité routière afin de réduire l'accidentalité à différents niveaux :

- Matériel : équipement de sécurité, type de véhicule…
- Humain : genre, âge, mode de déplacement…
- Localisation : type de route, intersection, urbain ou non…
- Conditions atmosphériques : météo, luminosité…

Une troisième partie, si le temps le permet, pourrait être de **simuler, à partir du modèle, la gravité des accidents si l'on améliorait certains des facteurs identifiés** : par exemple, s'il ressort que les équipements de sécurité jouent un rôle important, de combien aurions-nous pu réduire le nombre de morts en 2022 si toutes les voitures étaient équipées d'airbag ?

L'étude de ce projet est réalisée par l'équipe suivante :

- **Camille Pelat :** statisticienne chez Santé publique France, j'ai plus souvent affaire à des outcomes binaires (malade / non malade), parfois multinomiaux (allaitement exclusif / mixte / non), et plus souvent dans une visée explicative que prédictive. Le rééquilibrage du jeu de données, la sélection des variables par cross-validation et le test de performance sur un jeu de test sont des étapes que je trouve intéressantes à aborder dans ce projet. L'étape d'interprétabilité me paraît aussi importante pour faire le lien entre prédiction et explication.
- **Matthieu Claudel :** ingénieur d'études en maintenance chez SNCF Matériel, j'ai participé dans mon précédent poste à un projet de prédiction basé sur le computer vision mais jamais sur des données numériques ou catégorielles. Ce projet représente pour moi une opportunité de compléter de nouvelles compétences.
- **Nadège Reboul :** enseignant-chercheur dans le domaine du génie civil, actuellement en disponibilité pour reconversion, j'ai travaillé par le passé sur des données quantitatives, issues de mesures expérimentales, et utilisé des méthodes de clustering sous R et/ou matlab. Le travail avec Python, sur des données fortement catégorielles, avec une finalité de classification supervisée est donc une première expérience pour moi.
- **Vanessa Ibert :** docteur en chimie organique et professeur des écoles en reconversion professionnelle. En dehors de la formation, je n'ai pas encore eu l'occasion de traiter de telles problématiques.

Des recherches dans la littérature montrent des projets similaires. Nous avons sélectionné trois d'entre eux :

- La première étude est un article paru dans la revue de l'IA (Talbi 2020). Dans cet article, le nettoyage du jeu de données est très rapide (suppression de colonnes) sans faire de regroupement de modalités (hormis les variables 'lat' et 'long'). Il est à noter le travail de regroupement des latitudes et des longitudes en 15 modalités en utilisant la méthode des K-Means. Pour la modélisation, il compare les algorithmes de Random Forest et de XGBoost avec de meilleurs résultats avec XGBoost. En perspectives d'améliorations, il propose l'optimisation des hyperparamètres et le rééquilibrage des catégories de la variable cible.

- La seconde proposition est un repo GitHub (Maxime 2019). Ce repos traite plus en profondeur le nettoyage des données : suppression de colonnes et remplacement des valeurs manquantes (généralement par la modalité la plus fréquente). En revanche, la variable cible est remaniée pour obtenir une variable cible binaire (regroupement des modalités 1 et 4 pour devenir 'Light Injury', regroupement des modalités 2 et 3 pour devenir 'Serious Injury and Death'). Pour la modélisation, l'algorithme Random Forest est utilisé dans un premier temps. Puis les variables les plus importantes sont sélectionnées pour refaire un entraînement avec uniquement ces variables. Enfin, une optimisation des hyperparamètres est effectuée.

- Le troisième est un article néo-zélandais, (Ahmed, et al. 2023), dont les données et l'objectif sont très proches des nôtres. Leur objectif est de prédire la gravité d'un accident (et pas d'un usager accidenté, à la différence de notre projet) en 4 classes : accident sans blessé, avec blessé léger, avec blessé sévère, avec tué. Leur analyse contient une étape de rééquilibrage du jeu de données, des méthodes d'ensemble (XGBoost notamment), une étape d'interprétabilité et de sélection des variables les plus contributives, et un ré-entraînement du modèle sur ce sous-ensemble de variables.

Compréhension et manipulation des données

Cadre

Afin de réaliser le projet, nous avions la possibilité d'utiliser des jeux de données publiques des deux sites suivants :

- Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2021 - data.gouv.fr (Ministère de l'Intérieur et des Outre-Mer 2013, màj 2023)
- https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016 (Lahlou Mimi 2018)

Après étude des données du site Kaggle, nous nous sommes rendu compte qu'elles étaient issues du site du gouvernement et ne concernaient que les années 2005 à 2016. Nous avons donc opté pour **l'utilisation des données gouvernementales**. Cette base de données recense **l'ensemble des accidents corporels survenus en France entre 2005 et 2022**.

Est considéré comme accident corporel, « **tout accident survenu sur une voie ouverte à la circulation publique, impliquant au moins un véhicule et ayant fait au moins une victime ayant nécessité des soins** ». Lorsqu'un tel accident survient, les forces de l'ordre interviennent sur place et remplissent des Bulletins d'Analyse des Accidents Corporels (BAAC) administrés par l'Observatoire National Interministériel de la Sécurité Routière (ONISR). Ce sont les éléments recensés dans ces BAAC, et donc **potentiellement sujets à des erreurs de saisie ou des défauts de saisie** (les forces de l'ordre ne sont pas toujours informées lorsque l'accident n'est pas mortel)**,** qui constituent notre base de données.

Certaines informations qui pourraient nuire à la vie privée des usagers concernés (conduite sous l'emprise d'alcool ou de drogue, défaut de permis de conduire…) **ont été éliminées** avant la diffusion de cette base de données au grand public. **Il est probable que ces différents facteurs exercent une influence non négligeable sur l'accidentalité routière et cela devra être conservé à l'esprit à la lecture des conclusions du présent rapport**.

Les bases de données sont annuelles et se composent chaque année de 4 fichiers au format .csv : «Caractéristiques - Lieux - Véhicules - Usagers ». La Figure 1 présente l'ensemble des variables contenues dans ces fichiers. Il est intéressant de noter que les fichiers Caractéristiques, Lieux et Véhicules pourront être fusionnés grâce à l'identifiant de l'accident (Num_Acc), tandis que le fichier Usagers pourra être relié aux autres par l'intermédiaire du fichier Véhicules puisqu'ils ont en commun les identifiants du véhicule (id_vehicule et num_veh). **Pour chaque usager, il est donc possible d'avoir accès à l'ensemble de ces variables.**

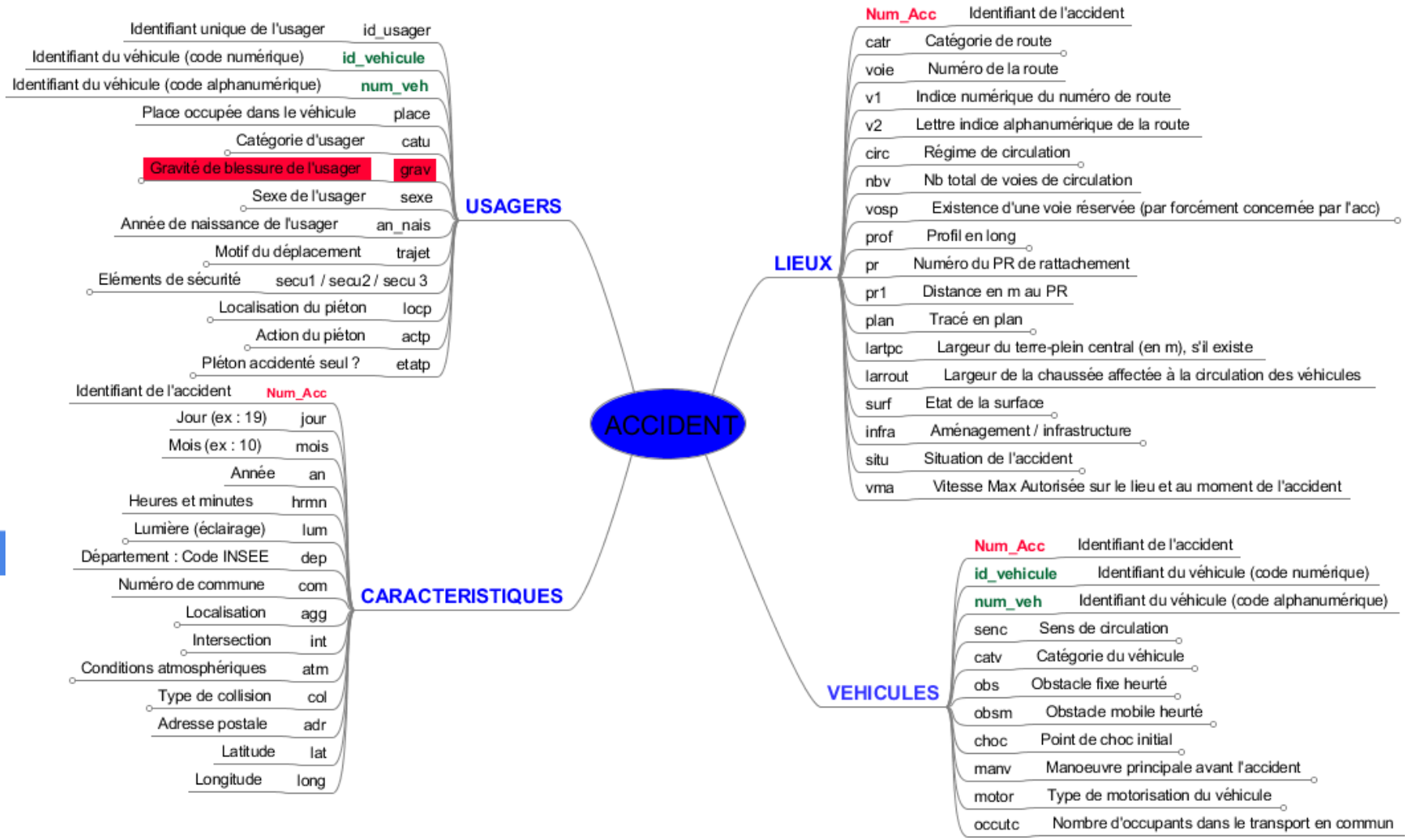
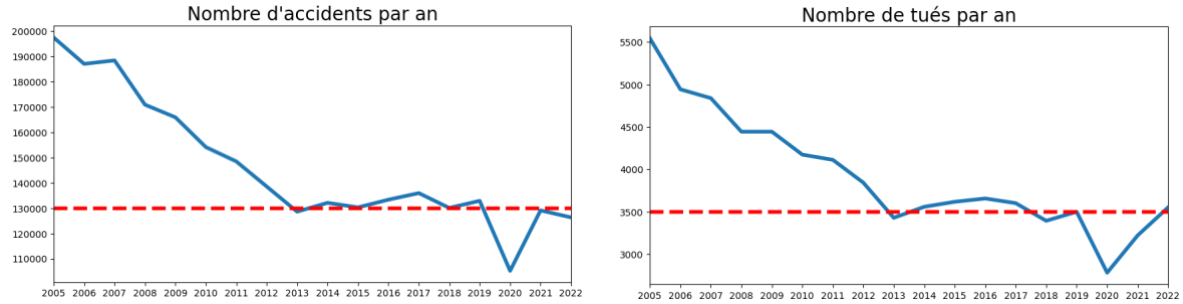


Figure 1 : Liste des variables présentes dans les bases de données mises à disposition du grand public

La fusion de l'ensemble des fichiers crée un jeu de données de plus de 2 millions de lignes. Une première étude du nombre d'accidents et de tués par an (Figure 2) montre une diminution du nombre de 2005 à 2013, puis une stabilisation du nombre d'accidents à environ 130000 par an et du nombre de tués à environ 3500 par an (en dehors de l'année 2020 correspondant à l'année du confinement).



(a) (b)

Figure 2 (a). Nombre d'accidents par an, (b). Nombre de tués par an

A la suite de modifications du processus de saisie des forces de l'ordre, les données depuis l'année 2018 ne peuvent être comparées à celles des années précédentes. Et, depuis 2019, des territoires d'Outre-mer ont été ajoutés, de nouvelles variables ont été créées (vma, motor, secu1, secu2, secu3 et id_vehicules), alors que d'autres ont été supprimées (gps et secu). En conséquence, nous avons fait le choix de **réduire notre analyse sur la période 2019 à 2022**. La fusion de l'ensemble des fichiers sur cette période permet d'obtenir **un jeu de données de 494 182 lignes avec 55 variables.**

Pertinence

Parmi ces variables, compte-tenu de la problématique de notre projet, la **gravité de blessure de l'usager (grav)** retient particulièrement notre attention. Chaque usager peut être considéré comme :

- tué s'il décède du fait de l'accident, sur le coup, ou dans les 30 jours qui suivent l'accident,
- blessé hospitalisé, s'il est hospitalisé plus de 24 heures,
- blessé léger, s'il a reçu des soins médicaux mais n'a pas été admis à l'hôpital plus de 24 heures,
- ou indemne.

Cette variable, grav, sera la variable cible de notre problème de prévision. Les 54 autres variables seront des variables explicatives.

Afin de vérifier si le jeu de données n'est pas biaisé (surtout sur l'année 2020 à cause du confinement), nous vérifions si le pourcentage d'accidents selon la gravité est homogène chaque année (Figure 3).

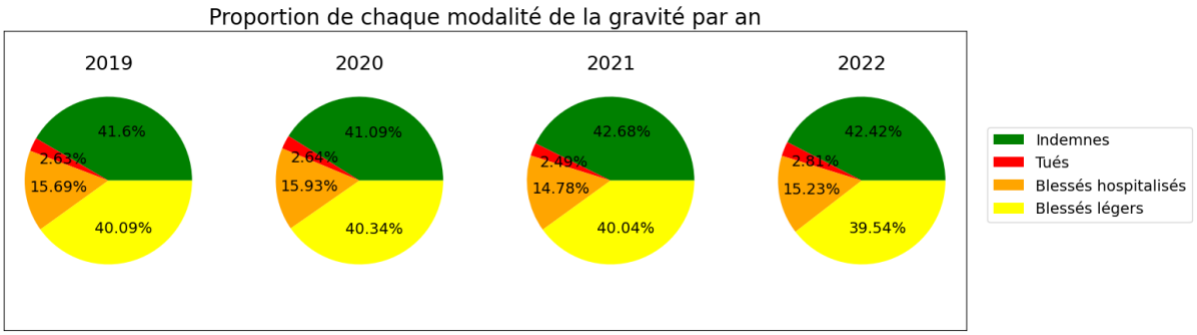


Figure 3 : Pourcentage d'accidents selon la modalité pour les années 2019 à 2022

Nous pouvons voir qu'il y a très peu de différence de proportion pour chaque modalité quelle que soit l'année. Donc nous pouvons garder le jeu de données sur les années 2019 à 2022 pour essayer de prédire la gravité des accidents routiers en France.

Pre-processing et feature engineering

Démarche globale

Le jeu de données résultant de la fusion des fichiers .csv usagers, véhicules, lieux et caractéristiques 2019 à 2022 contenait 55 variables dont une variable cible, 4 variables "identifiants" ('Num_Acc', 'id_vehicule', 'num_veh','id_usager') et 50 variables potentiellement explicatives.

Nous avons d'abord supprimé **164 doublons** (lignes exactement identiques).

Puis nous avons regardé le pourcentage de valeurs manquantes dans chaque variable. Parmi nos 55 variables, **14 avaient plus de 8% de valeurs manquantes** (Figure 4). Il s'agissait majoritairement de :

- variables **"administratives"** (ex. "v1 : Indice numérique du numéro de route, "id_usager : Identifiant unique de l'usager", "pr1 : Distance en mètres au PR (par rapport à la borne amont)."),
- variables renseignées uniquement dans un **sous-ensemble des accidents** (ex : "lartpc : Largeur du terre-plein central (TPC) s'il existe", "occutc : Nombre d'occupants dans le transport en commun",'locp : Localisation du piéton :'), donc à faible potentiel explicatif.

Nous les avons donc toutes supprimées, à l'exception des variables secu3 (98% de valeurs manquantes) et secu2 (39% de valeurs manquantes).

Les variables secu2 et secu3 sont en effet deux variables indiquant l'utilisation d'équipements de sécurité, complémentaires dans leur construction à secu1, et qui feront l'objet d'un recodage spécial, qui sera abordé dans la section "exemples détaillés avec visualisation".

Nous supprimons aussi les 4 variables "identifiants" ayant servi à faire la jonction entre les différents fichiers csv.

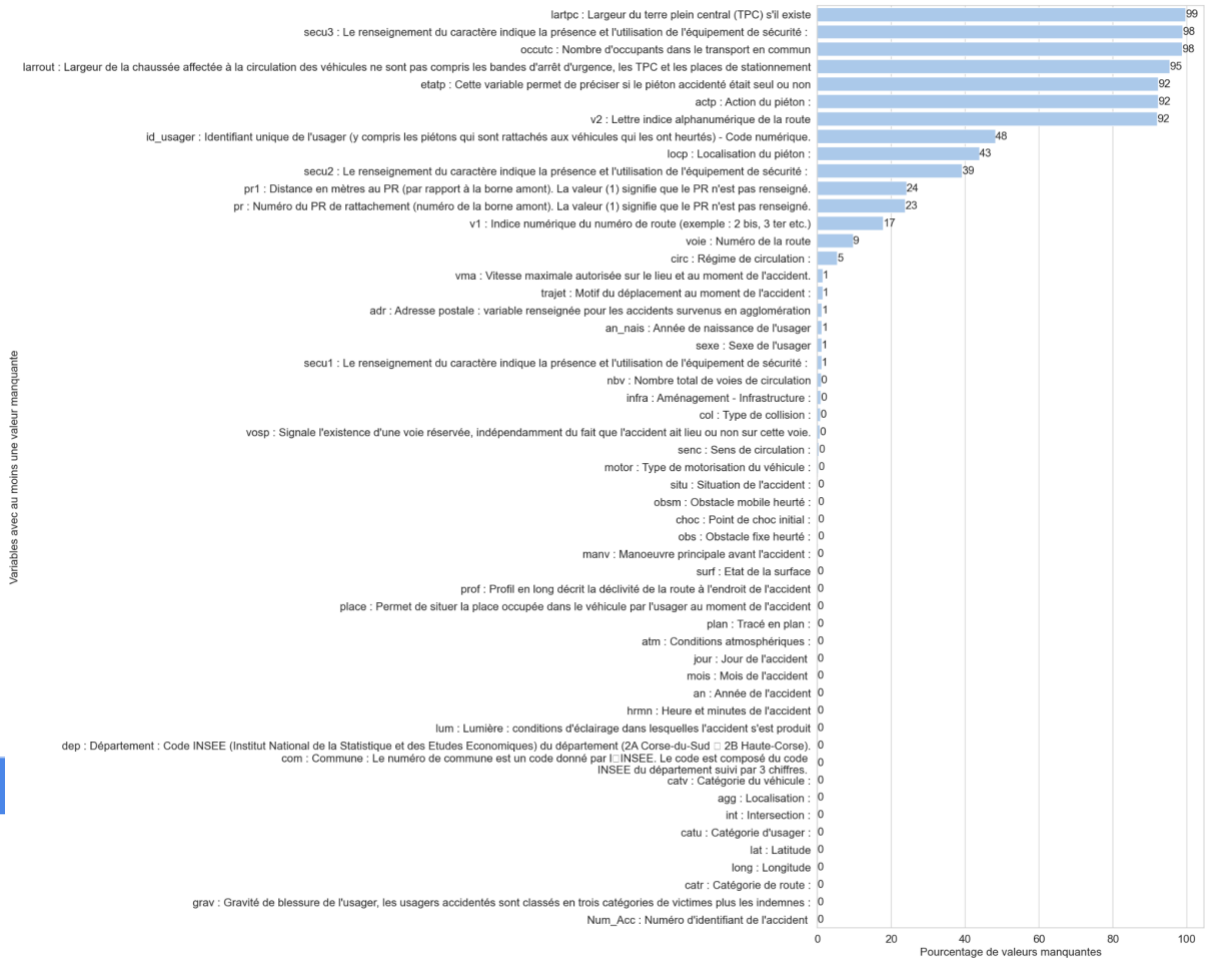


Figure 4 : Pourcentage de valeurs manquantes pour chaque variable

Nous avons choisi de supprimer aussi :

- d'autres variables **''administratives''** (''adr : adresse postale'', ''senc : sens de circulation'')
- d'autres variables renseignées uniquement dans un **sous-ensemble des accidents** (''vosp : présence d'une voie réservée, indépendamment du fait que l'accident ait eu lieu ou non sur cette voie''),
- des variables **trop corrélées** entre elles (''com : numéro de commune'' corrélée avec ''dep : numéro de département'', ''nbv : nombre total de voie de circulation'' et ''vma : vitesse maximale autorisée'' corrélées avec ''catr : catégorie de route'', ''catu : catégorie d'usagers'' corrélée avec ''place'' que l'on va modifier),
- des variables possédant trop de valeurs manquantes et ne pouvant pas être renseignée sans risquer de biaiser le jeu de données (''trajet : motif de déplacement au moment de l'accident'').

Étant donné la taille de notre jeu de données et compte-tenu de la faible proportion de valeurs manquantes pour les autres variables, nous avons décidé de supprimer les lignes avec des valeurs manquantes pour les autres variables. In fine, le retrait de ces valeurs manquantes a majoritairement impacté la catégorie des personnes indemnes (11.0 %), puis celle des blessés légers (8.7 %), puis celle des blessés hospitalisés (7.9 %), et enfin celle des tués (6.5 %). Ce choix n'a donc pas eu de répercussion sur l'une des modalités de la variable cible en particulier et n'a pas aggravé le déséquilibre du jeu de données initial.

Nous avons aussi modifié les modalités de certaines variables. (cf. tableau résumé des traitements de chaque variable).

Enfin, nous avons créé des variables qui nous paraissent plus pertinentes.

Après ces traitements, nous obtenons un jeu de données avec **447 136 lignes et 41 variables**. Certaines de ces variables, comme 'dep' et 'an' ne sont conservées que pour la data visualisation et seront supprimées pour la modélisation.

Enfin, nous réalisons un traitement spécifique pour la modélisation :

- création de dummies pour les variables catégorielles non binaires (avec conservation de tous les dummies dans le but de faire un Random Forest et /ou un XGBoost). Dans le cas où nous utiliserions d'autres algorithmes sensibles à la multi-colinéarité, nous enlèverions une colonne de chaque variable dummisée.
- réflexion sur les modes de normalisation/standardisation des variables continues (age_usager, latitude, longitude, mois et heure)

Finalement, le jeu de données final est composé de **447136 lignes et 98 variables**. Le traitement des données n'a pas changé les proportions des modalités de la variable cible (Figure 5). Cependant, le jeu de données reste largement **déséquilibré**. Des procédures de **ré-échantillonnage** devront donc être mises en œuvre au moment de la modélisation.

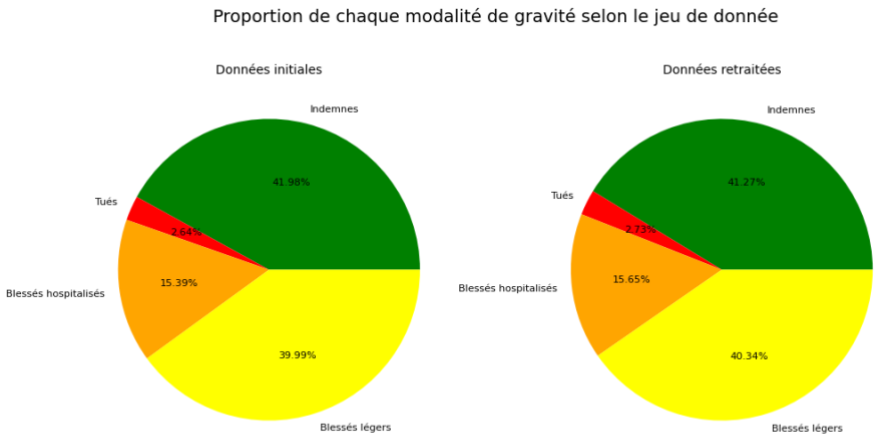


Figure 5 : Proportion de chaque modalité de la variable grav avant et après traitement

L'ensemble des modifications effectuées est regroupé dans le tableau de synthèse (Tableau 1), les variables de la base de données finalisée sont présentées sur la Figure 6 et la valeur du V_Cramer pour chaque variable est présentée par ordre d'importance dans la Figure 7.

Tableau 1 : Synthèse des traitements effectués sur les variables

Variables	Nombre NaN	Gestion NaN	Gestion valeurs abbérantes	Catégorisation	Traitement pour modélisation
Num_Acc	0	Suppression de la colonne	X	X	X
jour	0	X	X	Création variable 'weekend' : 0 - non 1 - oui Création variable 'jour_chome' : 0 - non 1 - oui	X
mois	0	X	X	X	X
an	0	X	X	X	Supprimé pour la modélisation
hrmn	0	X	X	Création colonne 'heure' et suppression 'hrmn'	X
lum	9	Suppresion des lignes avec NaN	X	Regrouper les catégories 3 et 4 pour devenir : 0 - Plein jour (1) 1 - Crépuscule ou aube (2) 2 - Nuit sans éclairage (3 et 4) 3 - Nuit avec éclairage public (5)	Dummies
dep	0	X	X	Remplacement des '1', ..., '9' par '01', ..., '09'	Supprimée pour la modélisation
com	0	Suppression de la colonne	X	X	X
agg	0	X	X	Remplacement de 1 en 0 et de 2 en 1 (pour être binaire)	X
int	21	Suppresion des lignes avec NaN	X	Regrouper les catégories en 2 catégories : 0 - hors intersection (1) 1 - hors intersection (2 à 9)	X
atm	33	Suppresion des lignes avec NaN	X	Regrouper les catégories en 2 catégories : 0 - Normal (1) 1 - Autres (2, 3, 4, 5, 6, 7, 8, 9)	X
col	3870	Suppresion des lignes avec NaN	X	X	Dummies
adr	6042	Suppression de la colonne	X	X	X
lat	0	X	Beaucoup de valeurs aberrantes retraitées au cas par cas	X	X
long	0	X	Beaucoup de valeurs aberrantes retraitées au cas par cas	X	X
catr	0	X	X	X	X
voie	47519	Suppression de la colonne	X	X	X
v1	87998	Suppression de la colonne	X	X	X
v2	454208	Suppression de la colonne	X	X	X
circ	26227	Suppresion des lignes avec NaN	X	Regrouper les catégories en 2 catégories : 0 - Unidirectionnel (1, 3, 4) 1 - Bidirectionnel (2)	X
nbv	4775	Suppression de la colonne	X	X	X
vosp	2830	Suppression de la colonne	X	X	X
prof	78	Suppresion des lignes avec NaN	X	Regrouper en 2 catégories : 0 - plat (1) 1 - pente (2, 3, 4)	X
pr	117212	Suppression de la colonne	X	X	X
pr1	118945	Suppression de la colonne	X	X	X
plan	60	Suppresion des lignes avec NaN	X	Regrouper les catégories en 2 catégories : 0 - Rectiligne 1 - Courbe	X
lartpc	492635	Suppression de la colonne	X	X	X
larrout	470903	Suppression de la colonne	X	X	X
surf	114	Suppresion des lignes avec NaN	X	X	Dummies
infra	4458	Suppresion des lignes avec NaN	X	X	Dummies
situ	274	Suppresion des lignes avec NaN	X	X	Dummies
vma	7141	Suppression de la colonne	X	X	X
id_vehicule	0	Suppression de la colonne	X	X	X
num_veh	0	Suppression de la colonne	X	X	X
place	3	Suppresion des lignes avec NaN	X	Création de la variable 'place_rec' regroupée en 4 catégories : 1 - Conducteur 2 - Passager avant 3 - Passager arrière 4 - Piéton	Dummies
catu	0	Suppression de la colonne	X	X	X
grav	0	Suppression des lignes avec NaN	X	X	X
sexe	5506	Suppression des lignes avec NaN	X	Remplacement de 1 par 0 et de 2 par 1 (pour être binaire)	X
an_nais	5641	Suppression des lignes avec NaN	Suppression des lignes où l'usager est une femme de 118 ans et + Suppression des lignes où l'usager est un homme de 112 ans et +	Création de la variable 'age_usager'	X
trajet	6600	Suppression de la colonne	X	X	X
secu1	5320	On remplacer les Nan par des 0.	X	Créer 7 variables avec les modalités 0 (non) ou 1 (oui) : 'Eq_ceinture' 'Eq_casque' 'Eq_dispositif_enfants' 'Eq_Gilet_reflechissant' 'Eq_Airbag' 'Eq_Gants' 'Eq_autre'	X
secu2	193102				
secu3	488124				
locp	216447	Suppression de la colonne	X	X	X
actp	455749	Suppression de la colonne	X	X	X
etatp	455834	Suppression de la colonne	X	X	X
id_usager	238108	Suppression de la colonne	X	X	X
senc	1642	Suppression de la colonne	X	X	X
catv	13	Suppression des lignes avec NaN	X	Regrouper les catégories en 5 catégories : 0 - Voiture (3, 7, 10) 1 - Moto (2, 30, 31, 32, 33, 34, 35, 36, 41, 42, 43) 2 - Poids lourds (13, 14, 15, 16, 17, 20, 21) 3 - Transport en commun (37, 38, 39, 40) 4 - Vélo/Trotinette (1, 50, 60, 80) 5 - Autre véhicule (0, 99)	Dummies
obs	163	Suppression des lignes avec NaN	X	Regrouper les catégories en 2 catégories : 0 - Sans obstacle 1 - Avec obstacle	X
obsm	231	Suppression des lignes avec NaN	X	Regrouper les catégories en 4 catégories : 0 - Sans obstacle 1 - Piéton 2 - Véhicule 3 - Animaux/Autres	Dummies
choc	208	Suppression des lignes avec NaN	X	Création de la variable 'proximite_choc' : 0 - Pas de proximité 1 - Proximité	X
manv	148	Suppression des lignes avec NaN	X	Regrouper les catégories en 4 catégories : 0 - Même sens (1, 2, 3, 7, 25) 1 - Contre sens (4, 5, 8) 2 - Immobile (22, 23, 24) 3 - Changement de direction (6, 9, 10 à 21, 26)	Dummies
motor	977	Suppression des lignes avec NaN	X	X	Dummies
occutc	487581	Suppression de la colonne	X	X	X

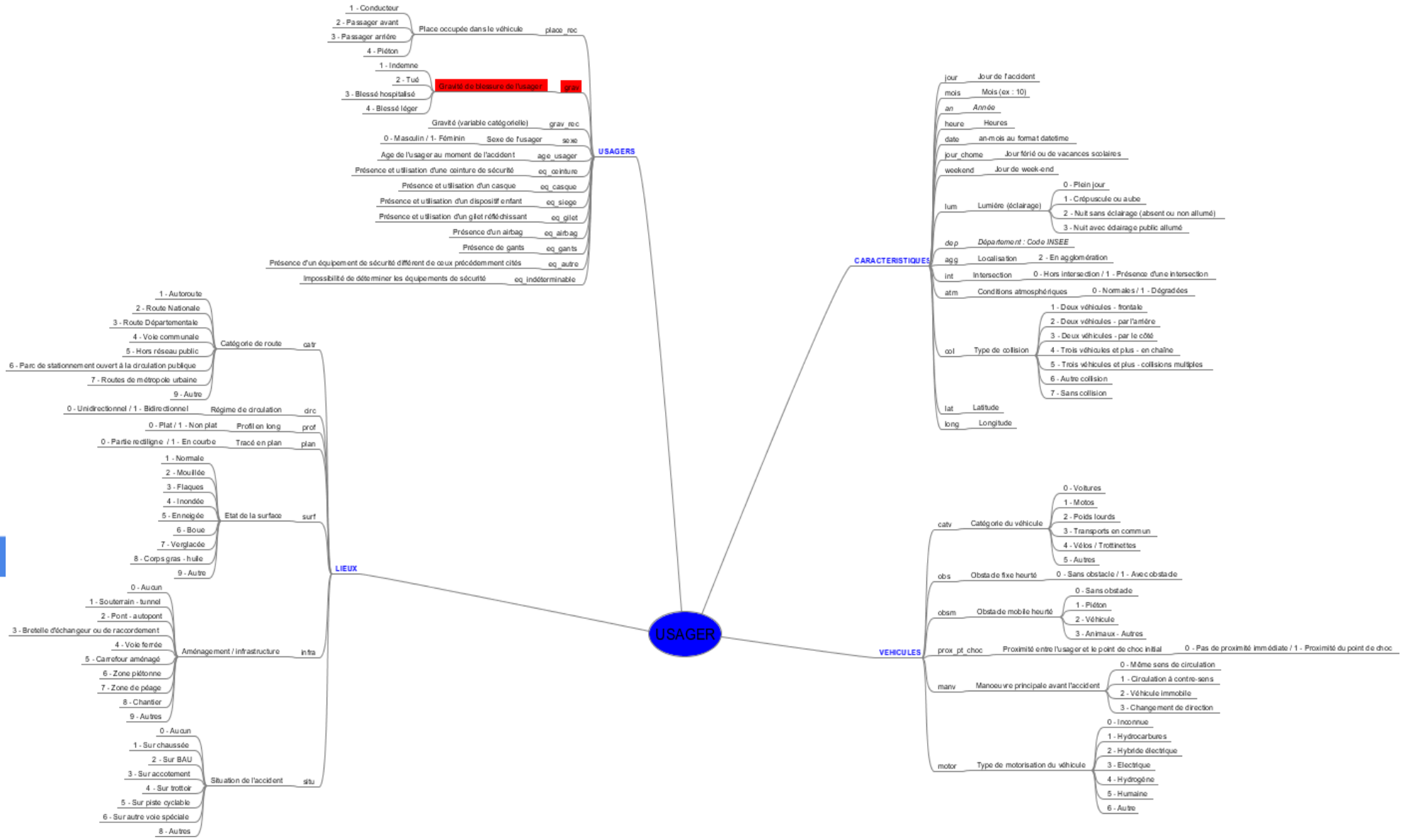


Figure 6 : Liste des variables présentes après traitement du jeu de données

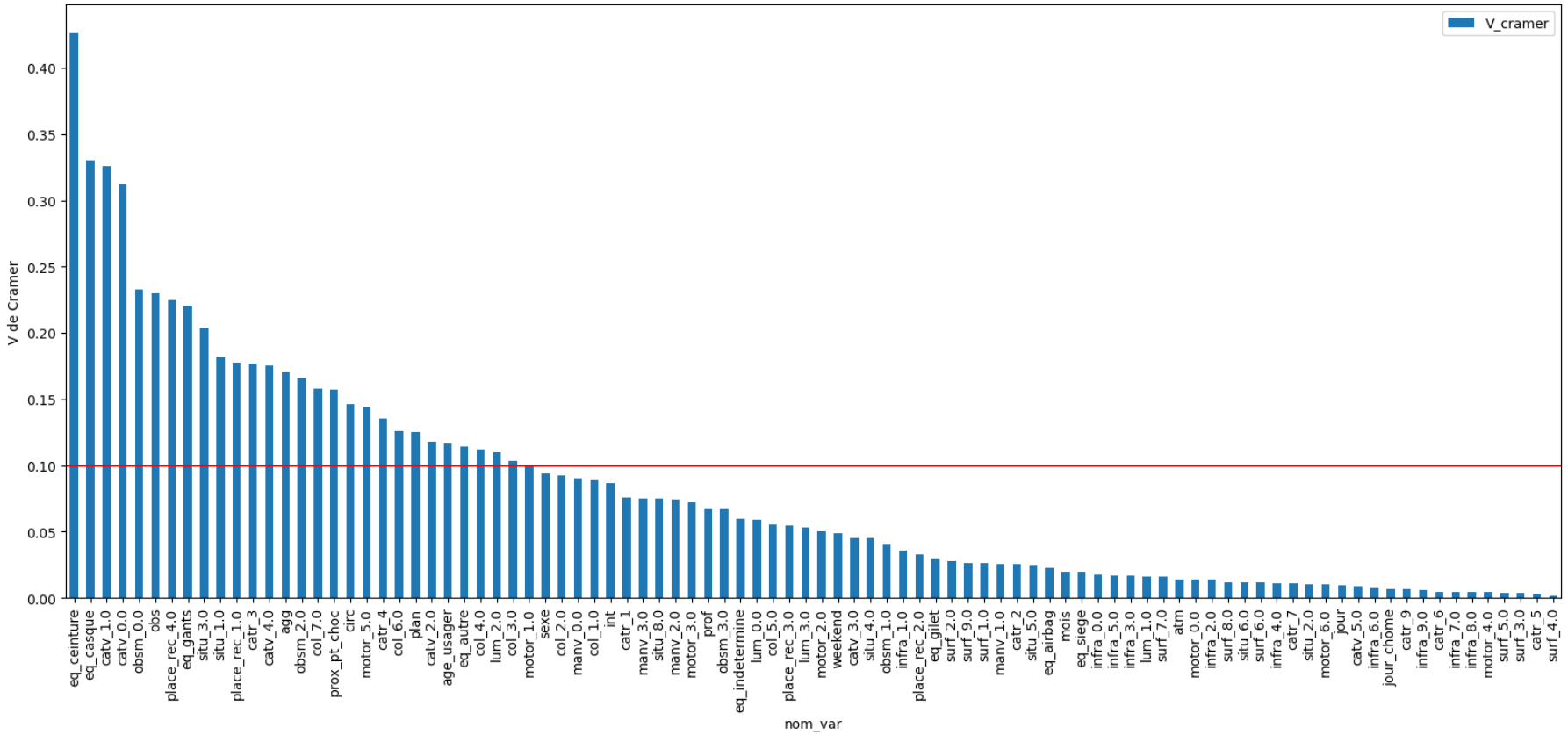


Figure 7 : Valeurs du V de Cramer pour chaque variable par ordre d'importance

Exemples détaillés avec visualisation

Nous présentons ci-après les argumentaires nous ayant conduit aux modifications de la base de données initiale pour 5 variables particulières.

CATÉGORIES DE VÉHICULES

Traitement de la variable catv

La variable **catv**, renseignant sur la catégorie de véhicules, a initialement 30 modalités, ce qui la rend difficilement exploitable en l'état.

Il a donc été décidé de procéder à des regroupements, tels que présentés sur la Figure 8. La variable catv réencodée présente donc 6 modalités : *0 – Voiture, 1 – Moto, 2 – Poids lourds, 3 – Transport en commun, 4 – Vélo/Trottinette, 5 – Autre véhicule*.

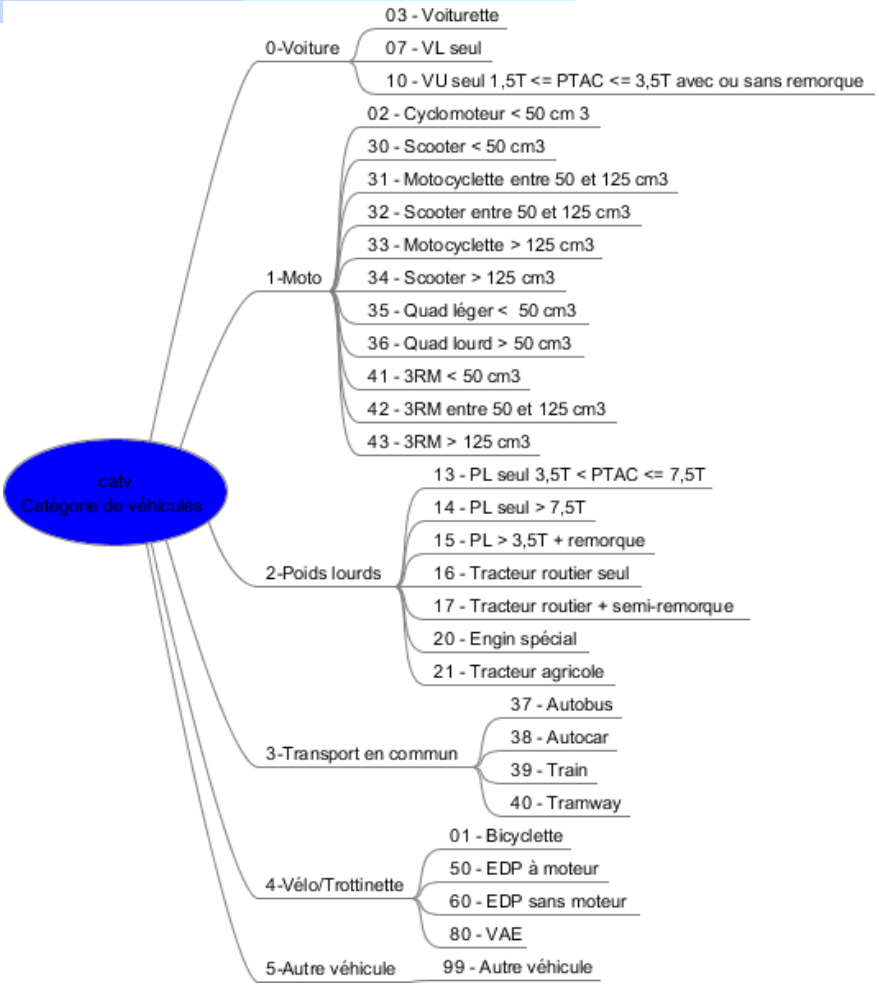


Figure 8 : Regroupement des catégories pour la variable catv

Le recodage en 6 modalités permet de conserver la différence de gravité selon le type de véhicule (Figure 9). On voit très nettement que la proportion de gravité des accidents est plus importante lorsque l'usager est en moto ou en vélo/trottinette. A l'inverse, les poids lourds sont ceux qui ont la proportion de personnes indemnes à l'issue d'un accident la plus élevée.

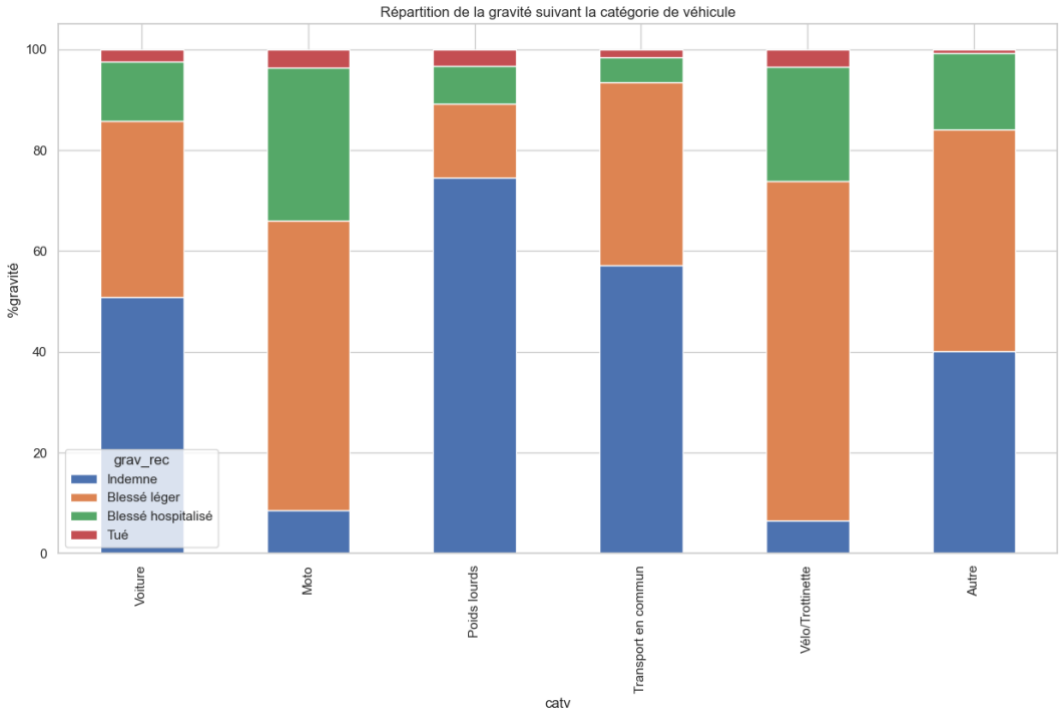


Figure 9 : Répartition de la gravité en fonction de la catégorie de véhicule

La Figure 10 présente les valeurs du V de Cramer pour les modalités de la nouvelle variable en rouge et de l'ancienne variable en bleu. Le nouvel encodage a tendance à réduire le nombre de modalités faiblement influentes et conserve les variables les plus influentes. Le regroupement a tendance à accentuer l'intensité de la relation des variables encodées avec la variable cible.

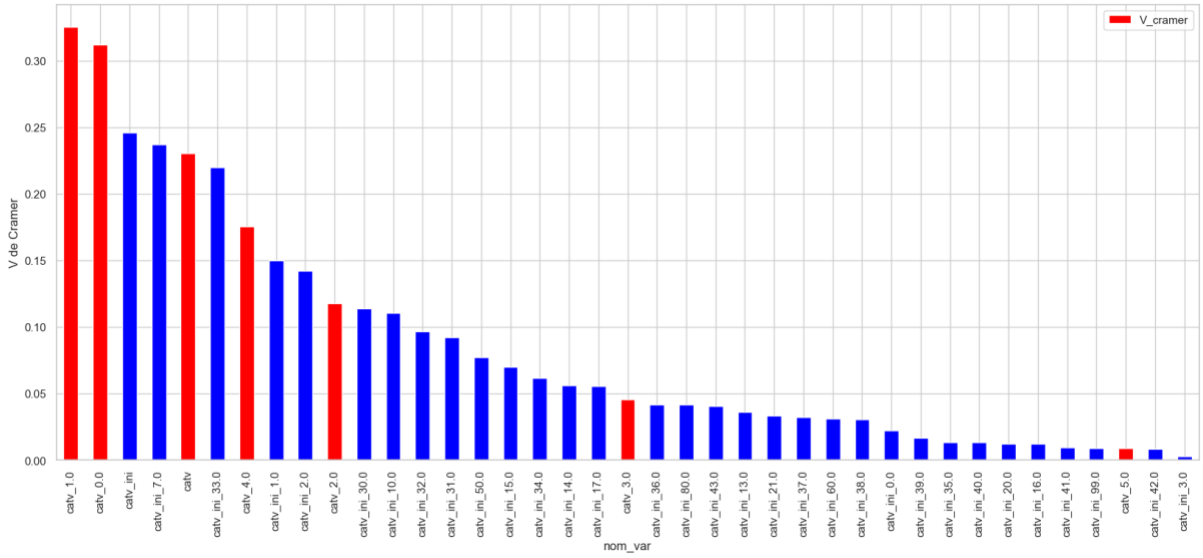


Figure 10 : Valeurs du V_Cramer pour les modalités de la catégorie de véhicule initiales et recodées

PLACE OCCUPÉE PAR L'USAGER et POINT DE CHOC INITIAL

Traitement des variables catu, place, catv et choc

Dans sa forme initiale, la base de données contient :

- Une variable **catu**, à 3 modalités, correspondant à la catégorie d'usager :
  1 - Conducteur, 2 - Passager, 3 – Piéton

- Une variable **place**, à 10 modalités, permettant de situer la place occupée dans le véhicule par l'usager au moment de l'accident. La Figure 11 explique la catégorisation adoptée, la valeur 10, absente sur cette figure, étant associée aux piétons.

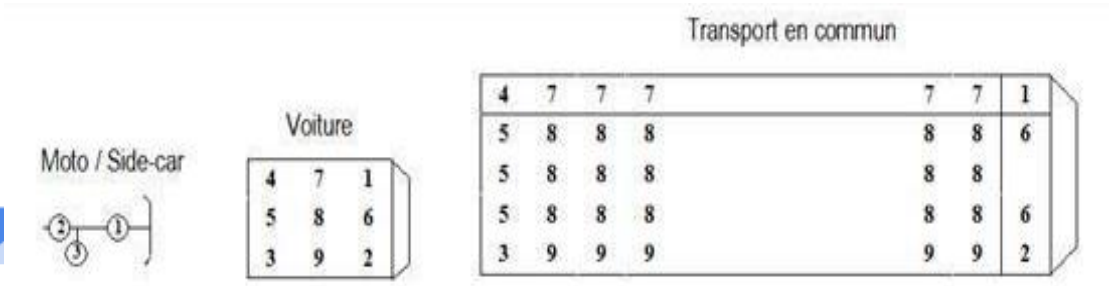


Figure 11 : Numérotation des places dans la variable place

- La catégorie de véhicules est connue grâce à la variable **catv**, à 5 modalités : *0 - Voitures, 1 - Motos, 2 - Poids lourds, 3 – Transports en commun, 4 – Vélos et trottinettes, 5 – Autres véhicules*

- Une variable **choc**, à 10 modalités, informant sur le 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

Ces variables ont été **recombinées** de façon à **limiter les modalités** (pour réduire la dimension du modèle), tout en conservant un maximum d'informations sur **la position de l'usager dans le véhicule** ainsi que sur **sa proximité avec le point de choc**. Ces deux aspects nous semblent en effet importants pour estimer la gravité d'un accident pour un usager donné. C'est pourquoi, nous avons réalisé les modifications suivantes :

- création d'une nouvelle variable proximité du point de choc,
- recodage de la variable place en 4 modalités.

Création d'une nouvelle variable binaire : prox_pt_choc

Une variable **prox_pt_choc**, binaire, est ainsi créée. Elle prend la valeur 1 lorsque l'usager est considéré à proximité du point de choc, 0 sinon (Tableau 2).

Tableau 2 : Configurations pour lesquelles la variable prox_pt_choc est prise égale à 1

Catégorie de véhicules	Point de choc initial	Place de l'usager
Voitures, Poids lourds, Transports en commun	1- Avant	1 – 6 – 2
	2- Avant droit	2
	3- Avant gauche	1
	4- Arrière	3 – 4 – 5
	5- Arrière droit	3
	6- Arrière gauche	6
	7- Côté droit	2 – 3 – 9
	8- Côté gauche	1 – 7 – 4
	9- Chocs multiples	Toutes les places
Motos, vélos et trottinettes	Toutes les places, quel que soit le point de choc	

nom_var	stat_chi2	p_value	V_cramer
prox_pt_choc	11034.800313	0.0	0.157095
choc	21979.802725	0.0	0.128006

(a)

prox_pt_choc	grav_rec	Blessé hospitalisé	Blessé léger	Indemne	Tué
0		0.268604	0.355711	0.47269	0.292025
1		0.731396	0.644289	0.52731	0.707975

(b)

Figure 12 : (a). Comparaison du χ^2 et du V de Cramer pour les variables prox_pt_choc et choc, (b). Tableau de contingence de la variable prox_pt_choc avec la variable cible

Le test d'indépendance du χ^2 et le calcul du V de Cramer tendent à montrer une **relation plus importante de la variable cible avec la variable prox_pt_choc, qu'avec choc** (Figure 12 (a)). Le tableau de contingence (Figure 12 (b)) souligne que les proportions de blessés ou tués sont significativement plus importantes lorsque les usagers sont à proximité du choc.

Recodage de la variable place en 4 modalités

La Figure 13 montre que la variable "place" se répartit majoritairement sur les modalités

conducteur, passager avant des véhicules ou arrière des motos, et piéton. Il a donc été choisi de rassembler certaines modalités pour en diminuer le nombre.

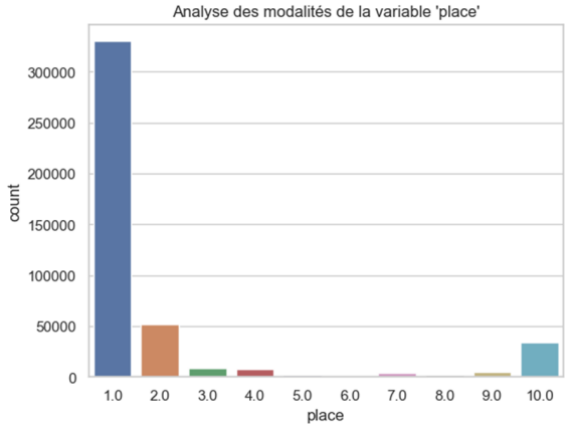


Figure 13 : Répartition des modalités de la variable place

La variable place est ainsi recodée pour apporter une précision supplémentaire à la variable catu, et la remplacer. La nouvelle variable place_rec contient 4 modalités : *1 – Conducteur, 2 – Passager avant, 3 – Passager arrière, 4 – Piéton*. Le Tableau 3 donne les correspondances entre les places de la Figure 11 et les modalités de notre nouvelle variable.

Tableau 3 : Correspondances entre les modalités de la variable place

Modalités de la variable place	**Voiture, Poids lourds, Transport en commun**	1	2 - 6	3 à 9	10
	Moto	1		2 - 3	
Modalités de la nouvelle variable, place_rec		**1 - Conducteur**	**2 – Passager avant**	**3 – Passager arrière**	**4 – Piéton**

Les tableaux des Figure 14 (a) et (b) soulignent que **le recodage permet de conserver les modalités les plus en relation avec la variable cible**, ce qui est prometteur quant à la réorganisation effectuée.

nom_var	stat_chi2	p_value	V_cramer
place_10.0	22631.424444	0.000000e+00	0.224976
place_1.0	14046.536073	0.000000e+00	0.177241
place	27087.228931	0.000000e+00	0.142103
place_2.0	1433.533133	1.558411e-310	0.056622
place_3.0	505.871751	2.547776e-109	0.033636
place_4.0	479.035627	1.666337e-103	0.032731
place_5.0	92.385153	6.732664e-20	0.014374
place_7.0	56.668810	3.024175e-12	0.011258
place_8.0	26.269312	8.375938e-06	0.007665
place_6.0	7.833142	4.958902e-02	0.004186

(a)

nom_var	stat_chi2	p_value	V_cramer
place_rec_4.0	22631.424444	0.000000e+00	0.224976
place_rec_1.0	14046.536073	0.000000e+00	0.177241
place_rec	26247.681571	0.000000e+00	0.139883
place_rec_3.0	1353.501966	3.620533e-293	0.055019
place_rec_2.0	477.481651	3.618279e-103	0.032678

(b)

Figure 14 : χ^2 et V de Cramer pour les différentes modalités (a) de la variable place, (b). de la variable place_rec.

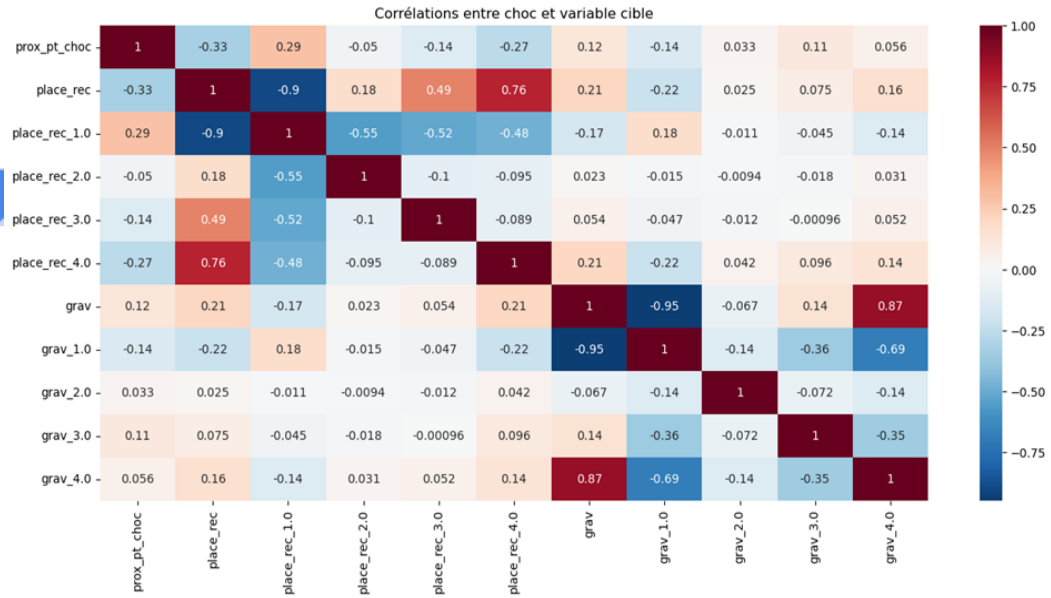


Figure 15 : Corrélation entre les différentes modalités de place_rec et de la variable cible. Les niveaux de gravité 1, 2, 3 et 4 correspondent respectivement à indemne, tué, blessé hospitalisé et blessé léger.

L'analyse des corrélations (Figure 15) souligne notamment que les piétons (à la place_rec 4) sortent rarement indemnes des accidents répertoriés dans les BAAC. Du côté des véhicules, la place de conducteur est négativement corrélée au fait d'être blessé léger et positivement corrélée au fait d'être indemne (protection de l'habitacle ?). Pour le reste des corrélations, les valeurs sont relativement faibles.

ÉLÉMENTS DE SÉCURITÉ

Traitement des variables secu1, secu2, secu3

La présence et l'utilisation d'éléments de sécurité par les usagers sont prises en compte dans la base de données par 3 variables distinctes : secu1, secu2, secu3. Secu1 renseigne sur le type d'un premier élément de sécurité, secu2 et secu3 informent sur un deuxième et troisième équipement le cas échéant. Chacune de ces variables a 10 modalités, recensées dans le tableau de la Figure 16.

En l'état, ces variables sont peu intéressantes car elles contiennent des informations identiques et ne peuvent être analysées indépendamment les unes des autres. Il a donc été choisi **de dichotomiser ces variables et de générer une nouvelle variable pour chaque équipement de sécurité, indiquant son utilisation, ou non, par l'usager impliqué dans l'accident.**

Base de données initiale : 3 variables secu1, secu2, secu3 à 10 modalités chacune

0	Aucun équipement
1	Ceinture
2	Casque
3	Dispositif enfants
4	Gilet réfléchissant
5	Airbag (2RM/3RM)
6	Gants (2RM/3RM)
7	Gants + Airbag (2RM/3RM)
8	Non déterminable
9	Autre

Après transformation : 8 variables binaires

eq_ceinture
eq_casque
eq_siege
eq_gilet
eq_airbag
eq_gants
eq_indetermine
eq_autre

Figure 16 : Dichotomisation des équipements de sécurité

Cette variable recodée permet de mieux appréhender la gravité de l'accident en fonction du port ou non de l'équipement de sécurité. Le test d'indépendance du χ^2 sur les équipements de sécurité et sur la catégorie de véhicules (Figure 17) fait ressortir la significativité de ces variables et l'intensité relativement forte de leur relation avec la variable cible.

nom_var	stat_chi2	p_value	V_cramer
eq_ceinture	82249.401604	0.000000e+00	0.426129
eq_casque	49392.580652	0.000000e+00	0.330221
catv_1.0	48049.346525	0.000000e+00	0.325700
catv_0.0	44067.709326	0.000000e+00	0.311914
eq_gants	22104.529370	0.000000e+00	0.220910
catv_4.0	13916.418985	0.000000e+00	0.175282
catv_2.0	6286.570219	0.000000e+00	0.117810
eq_autre	5944.895508	0.000000e+00	0.114564
eq_indetermine	1660.844001	0.000000e+00	0.060553
catv_3.0	949.501312	1.620146e-205	0.045785
eq_gilet	379.056727	7.608711e-82	0.028929
eq_airbag	240.205309	8.592272e-52	0.023029
eq_siege	186.454065	3.560829e-40	0.020289
catv_5.0	36.655608	5.441811e-08	0.008996

Figure 17 : χ2 et V de Cramer pour les différentes variables équipement et catégorie de véhicule recodées

L'analyse des corrélations (Figure 18) montre que le port de la ceinture a une influence importante sur la variable cible, en étant positivement corrélée au fait d'être indemne, et négativement corrélée au fait d'être blessé (léger ou hospitalisé). Les résultats sont inversés sur le port du casque, ce qui est relativement contre-intuitif (plus de conduite à risque ?).

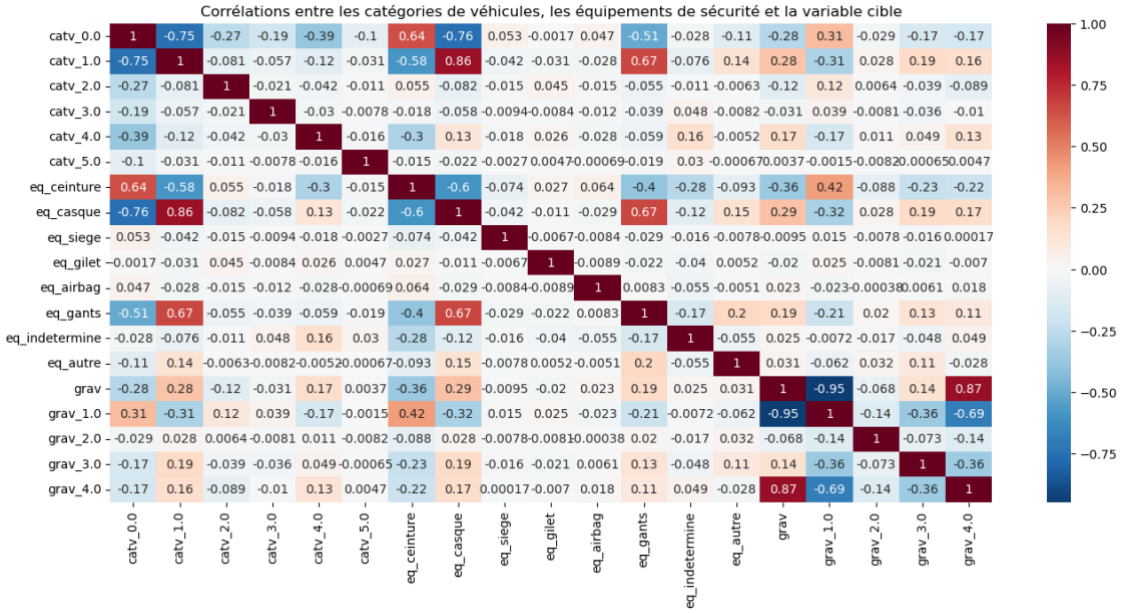


Figure 18 : Corrélation entre les différentes variables équipement, les différentes modalités de véhicule recodées et de gravité

Les variables sur les équipements sont assez logiquement fortement corrélées à certaines modalités des catégories de véhicules. Ainsi on observe :

- pour les gants et le casque : une corrélation positive avec la moto et une corrélation négative avec la voiture,
- pour la ceinture : une corrélation positive avec la voiture et une corrélation négative avec la moto

Cela permet aussi de connaître les équipements qui sont régulièrement utilisés ensemble :

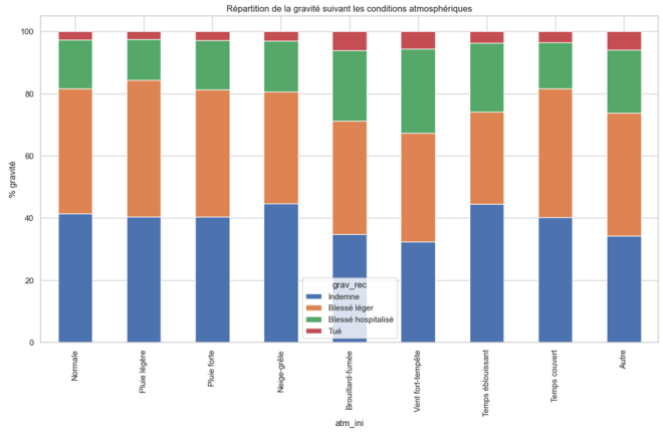
- le casque a une corrélation positive avec les gants, et dans une moindre mesure, avec les équipements répertoriés 'autre' par les forces de l'ordre.

CONDITIONS ATMOSPHÉRIQUES

Traitement de la variable atm

La variable atm de la base de données initiale recense 9 modalités : *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*.

Cette variable traite à la fois de l'impact des conditions atmosphériques sur la tenue de route et sur la visibilité. La Figure 19 (a) montrant les proportions des états de gravité en fonction des modalités de cette variable, souligne l'impact des conditions atmosphériques sur la gravité, confirmé par des p-valeurs au test d'indépendance du χ^2 inférieures à 5%. En revanche, les intensités des relations de ces variables avec la variable cible, mesurées avec le V de Cramer (Figure 19 (b)), s'avèrent très faibles. Les modalités "temps éblouissant" et "pluie légère" apparaissent comme les plus influentes. Mais ces modalités nous semblent largement soumises à l'interprétation des forces de l'ordre : un temps éblouissant pour un officier pouvant apparaître comme un temps normal pour un autre. Il en est de même pour les modalités pluies, temps couvert, etc.



(a)

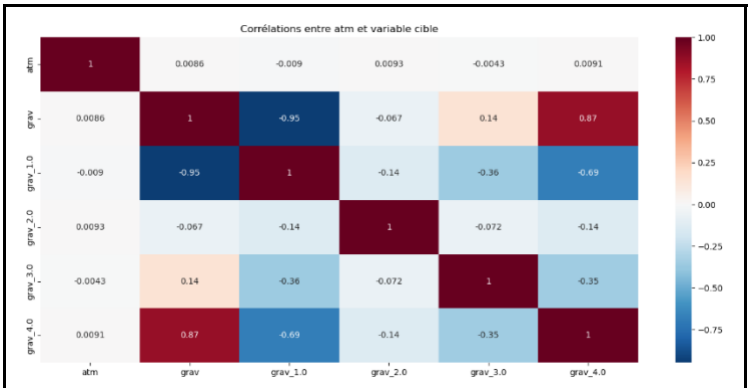
nom_var	stat_chi2	p_value	V_cramer
atm_7.0	527.871568	4.346804e-114	0.034359
atm_ini	1544.063595	7.584133e-312	0.033928
atm_2.0	423.255626	2.029890e-91	0.030767
atm_5.0	273.962336	4.287428e-59	0.024753
atm_6.0	165.899967	9.766472e-36	0.019262
atm_9.0	125.943611	4.046957e-27	0.016783
atm	87.500705	7.538590e-19	0.013989
atm_1.0	87.500705	7.538590e-19	0.013989
atm_8.0	50.473381	6.333998e-11	0.010625
atm_4.0	15.581712	1.381347e-03	0.005903
atm_3.0	3.230597	3.574196e-01	0.002688

(b)

Figure 19 : (a). Proportion de la gravité selon les conditions atmosphériques, (b). χ2 et V de Cramer pour les différentes modalités de la variable atm

Ce biais potentiel dans le remplissage des bulletins nous a conduits à ne pas souhaiter multiplier les modalités sur cette variable et nous avons opté pour un ré-encodage binaire de la variable atm, prenant désormais *0 en conditions normales*, et *1 en conditions dégradées* (pluie légère, pluie forte, neige-grêle, brouillard-fumée, vent fort-tempête, temps éblouissant, temps couvert et autre).

La corrélation entre la variable ‘atm’ recodée en binaire et la variable cible indique une influence des conditions atmosphériques sur la survenue d’accidents avec des blessés hospitalisés ou tués (Figure 20 (a)). Cependant, le V de Cramer indique une faible influence de cette variable sur notre variable cible (Figure 20 (b)).



(a)

	stat_chi2	p_value	V_cramer
nom_var			
atm	87.500705	7.538590e-19	0.013989

(b)

Figure 20 : (a). Corrélation entre la variable atm et les différentes modalités de la variable cible, (b). χ2 et V de Cramer pour la variable atm recodée en binaire.

MANŒUVRE PRINCIPALE AVANT L'ACCIDENT

Traitement de la variable manv

La variable 'manv' de la base de données initiale recense 27 modalités ce qui ne permet pas de l'utiliser ainsi.

Nous avons donc effectué des regroupements, tels que présentés sur la Figure 21. La variable manv réencodée présente 4 modalités : *0 – Même sens, 1 – Contresens, 2 – Immobile, 3 – Changement de direction*.

Figure 21 : Regroupement des catégories pour la variable manv

Le recodage de la variable en 4 modalités permet d'avoir une meilleure vision de l'impact du mouvement du véhicule au moment de l'accident. Ainsi, on peut constater que les accidents les plus graves surviennent lorsque le véhicule roule à contresens (Figure 22). Réciproquement, les accidents les moins graves interviennent lorsque le véhicule est à l'arrêt.

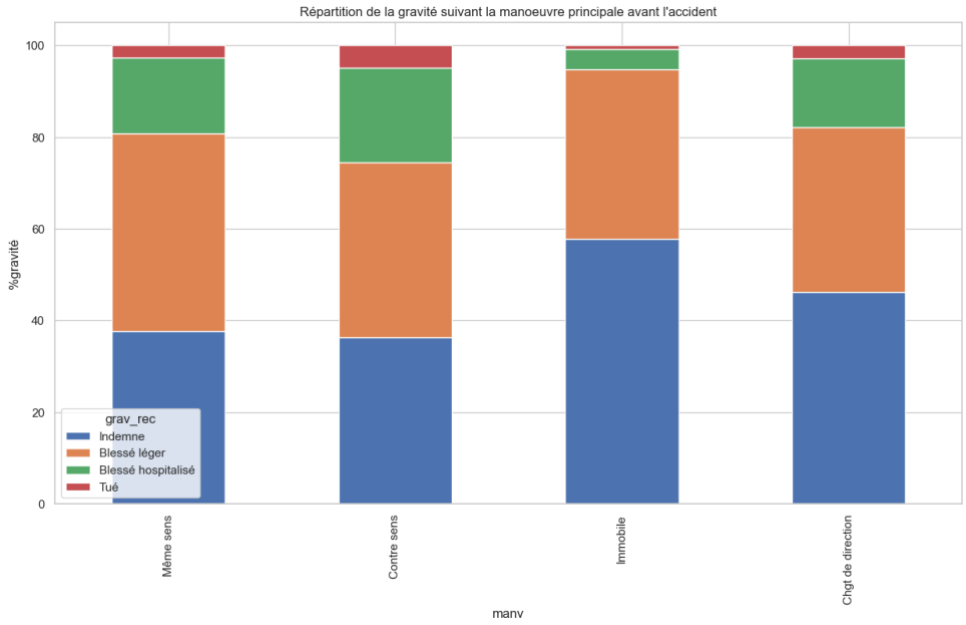
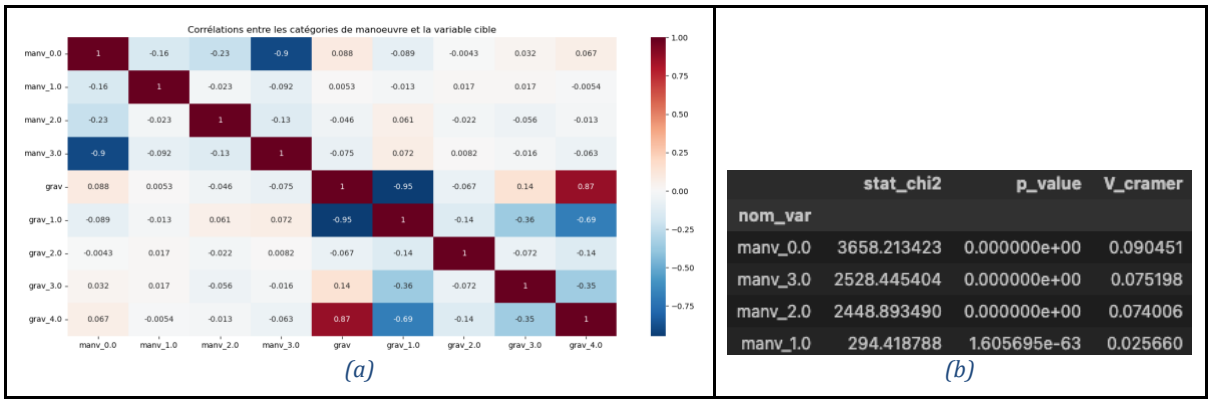


Figure 22 : Proportion de la gravité selon la manœuvre principale avant l'accident

L'étude de la corrélation entre la manœuvre au moment de l'accident et la gravité de l'accident (Figure 23 (a)) confirme les observations précédentes, mais montre aussi une faible corrélation en général de cette variable avec la variable cible.



(a)

nom_var	stat_chi2	p_value	V_cramer
manv_0.0	3658.213423	0.000000e+00	0.090451
manv_3.0	2528.445404	0.000000e+00	0.075198
manv_2.0	2448.893490	0.000000e+00	0.074006
manv_1.0	294.418788	1.605695e-63	0.025660

(b)

Figure 23 : (a). Corrélation entre les différentes modalités de la variable manv et de la variable cible, (b). χ2 et V de Cramer pour les différentes modalités de la variable manv

Le test d'indépendance du χ^2 (Figure 23 (b)) montre des p_valeurs inférieures à 5% confirmant l'impact de la manœuvre sur la gravité de l'accident. En revanche, les valeurs du V de Cramer, indiquant les intensités des relations de ces variables avec la variable cible, sont relativement faibles. Nous décidons de garder cette variable dans une première approche du modèle quitte à la supprimer par la suite.

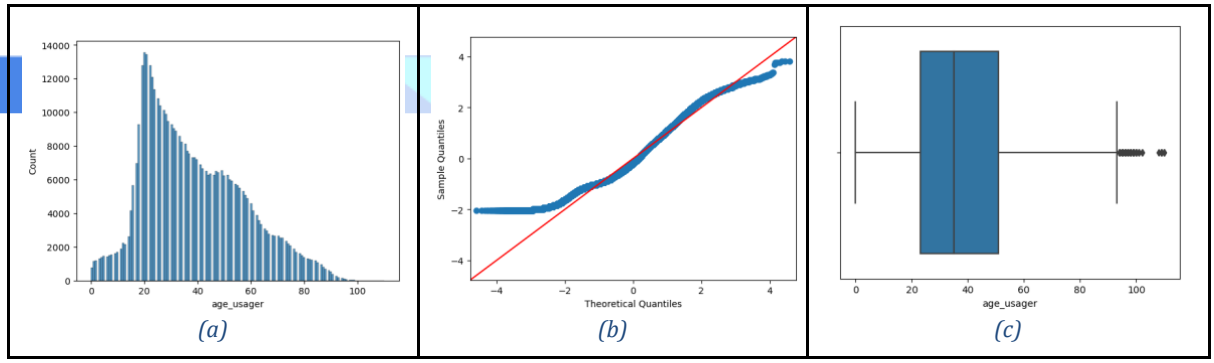
Analyse des variables continues

Les variables continues de la base de données sont au nombre de 5 : age_usager, latitude, longitude, mois et heure.

En fonction des modélisations envisagées, notamment si elles ont recours ou non à des calculs de distance, il peut être nécessaire de normaliser/standardiser les variables continues. Pour chacune des variables continues de notre base de données, il est précisé ci-après quel procédé pourra être utilisé en cas de besoin d'une normalisation, et les raisons de ce choix.

Age_usager

La variable age_usager ne suit pas une distribution normale (Figure 24 (a) et (b)) et présente quelques outliers (Figure 24 (c)) qui restent cependant dans des ordres de grandeur admissibles comparativement aux autres valeurs. **En cas de besoin, une normalisation min-max pourra être envisagée pour cette variable.**



(a) (b) (c)

Figure 24 : Distribution de age_usager. (a). Histogramme des valeurs, (b). Graphique des quantiles, (c). Boîte à moustaches

Heure et mois

Il est recommandé dans les forums dédiés à la data science (Kaleko 2017) de procéder à des transformations sinus/cosinus pour les variables temporelles de type heure et mois. En effet, ces modifications permettent de conserver le côté cyclique de ces variables temporelles.

$$\{= sin\left(2\pi \frac{heure}{24}\right) \quad = cos\left(2\pi \frac{heure}{24}\right) \quad \{= sin\left(2\pi \frac{mois-1}{12}\right) \quad = cos\left(2\pi \frac{mois-1}{12}\right)$$

Latitude et longitude

Les variables latitude et longitude ne suivent pas des distributions normales (Figure 25) et présentent de nombreux outliers (Figure 26) en raison de la présence des DOM/TOM dans la base de données. **En cas de besoin d'une normalisation, le recours à un procédé de Robust Scaling est envisagé pour ces variables.**

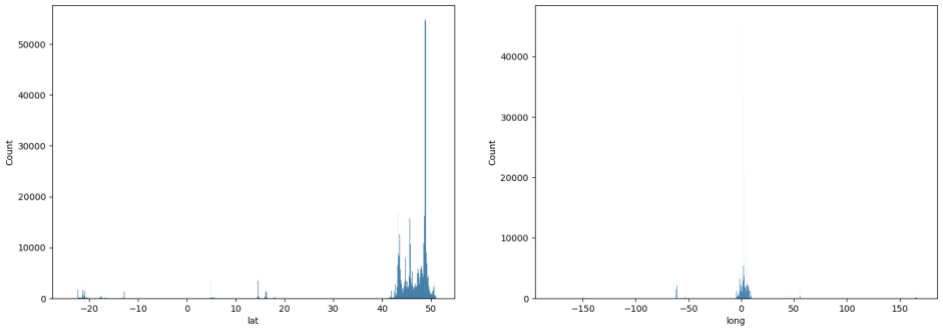


Figure 25 : Distributions des variables lat et long

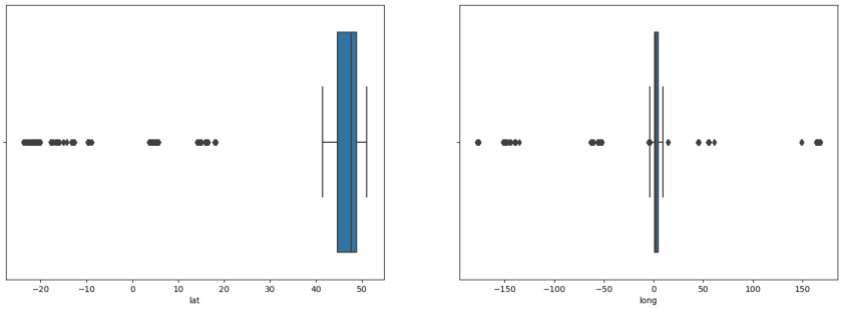


Figure 26 : Boîtes à moustaches des variables lat et long

En se focalisant uniquement sur la métropole, on observe toujours une distribution qui n'est pas normale (Figure 27) et la présence d'outliers (Figure 28). Il faudrait aussi avoir recours à **un procédé de Robust Scaling** dans le cas où une normalisation serait nécessaire.

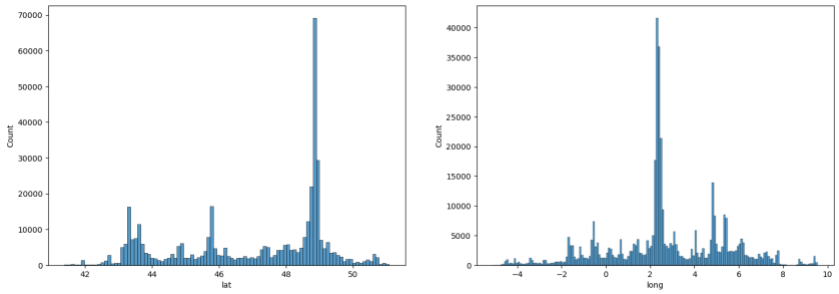


Figure 27 : Distributions des variables lat et long, pour la métropole uniquement

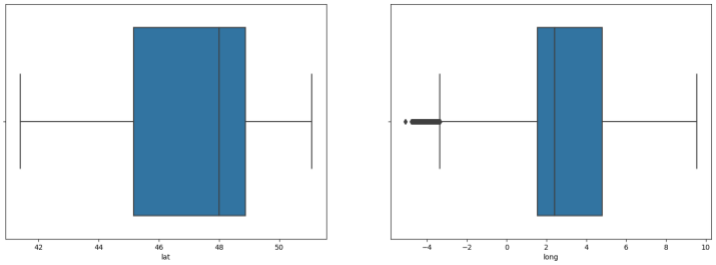


Figure 28 : Boîtes à moustaches des variables lat et long pour la métropole uniquement

Visualisations et Statistiques

Evolution Temporelle

Mois

On voit se dessiner une certaine saisonnalité dans la proportion mensuelle de tués et de blessés hospitalisés (Figure 29), avec plus de gravité chaque été, et un pic lors du premier confinement (avril 2020).

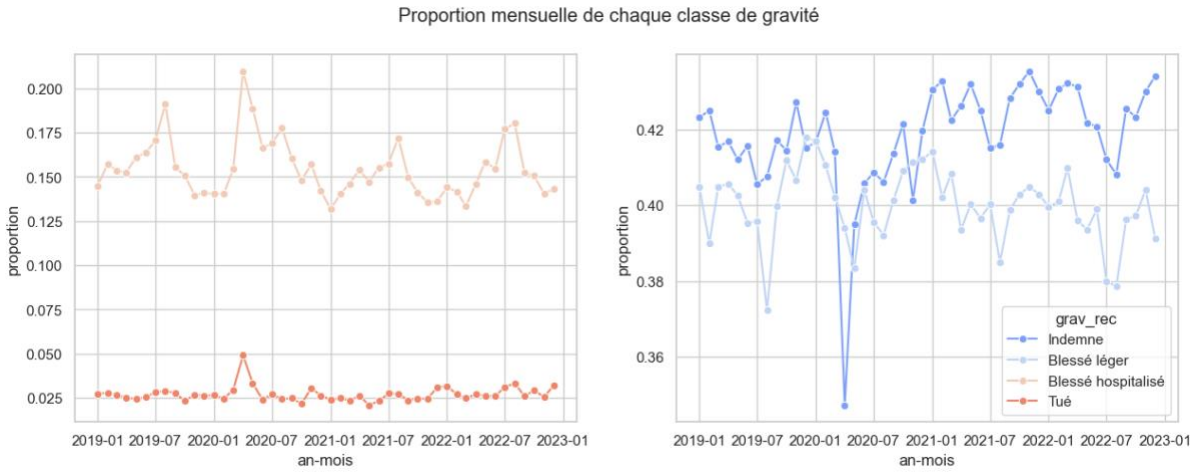


Figure 29 : Proportion mensuelle de chaque classe de gravité

Toutefois, dans un premier temps, nous allons axer la modélisation sur des méthodes de classification et par conséquent laisser de côté la variable “an-mois” au profit d’une variable “mois de l’année”.

En comparant la gravité des accidents mensuels sur les 4 années (Figure 30), on peut se rendre compte d’une saisonnalité :

- diminution du nombre de tués, blessés légers et usagers indemnes au mois d’août par rapport aux mois de juillet et septembre, peut-être due à la période estivale où la proportion de personnes en vacances est la plus importante,
- diminution de toutes les modalités de gravité aux mois d’avril et de novembre (correspondant aux vacances scolaires ?)
- augmentation de toutes les modalités de gravité aux mois de juin/juillet.
- augmentation du nombre de tués, blessés légers, usagers indemnes (et de blessés hospitalisés certaines années) aux mois de septembre/octobre.

L’année 2020 présente 2 pics bien en dessous des autres années correspondant aux 2 périodes de confinement. Le premier pic est plus important car il correspond au confinement strict pour la période du 17 mars au 11 mai 2020. Le second est moins accentué car la période de confinement du 30 octobre au 15 décembre 2020 était moins stricte.

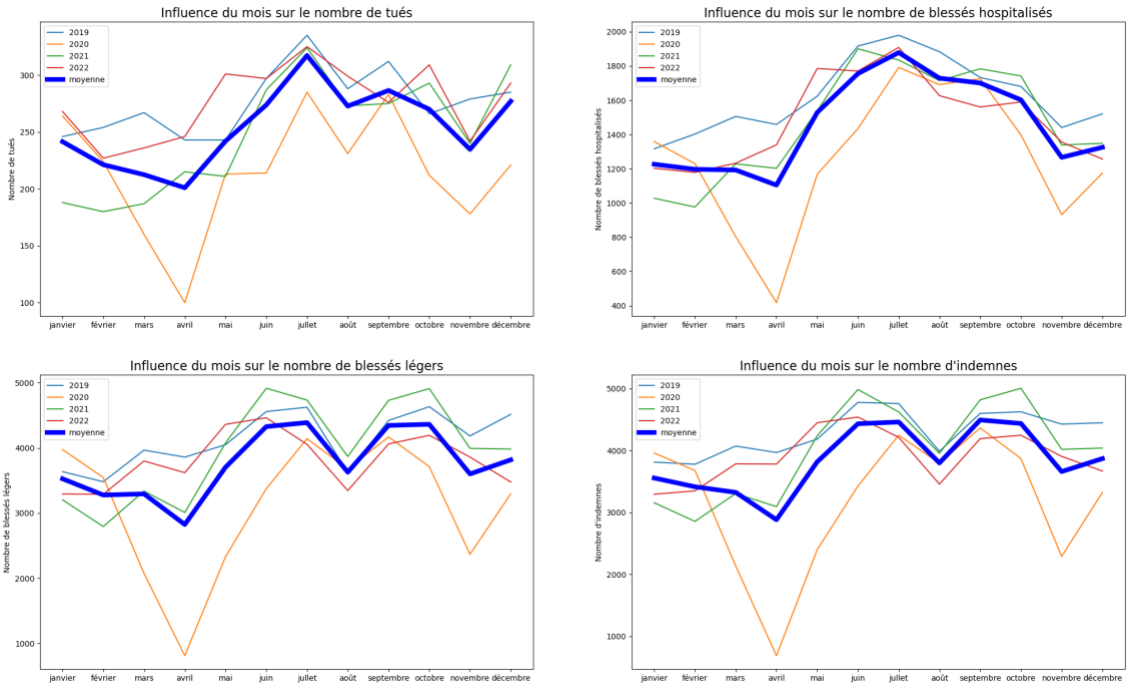


Figure 30 : Courbes du nombre d'usagers pour chaque classe de gravité selon le mois pour les années 2019 à 2022

La gravité subit donc de fortes variations selon les mois. Nous décidons donc de la conserver sans créer de catégories.

En revanche, nous observons des diminutions sur les courbes pour les mois d'avril, d'août et de novembre qui pourraient correspondre à des périodes de vacances scolaires. Nous décidons donc de créer une variable jour_chome qui regroupe les jours fériés et les vacances scolaires (Figure 31).

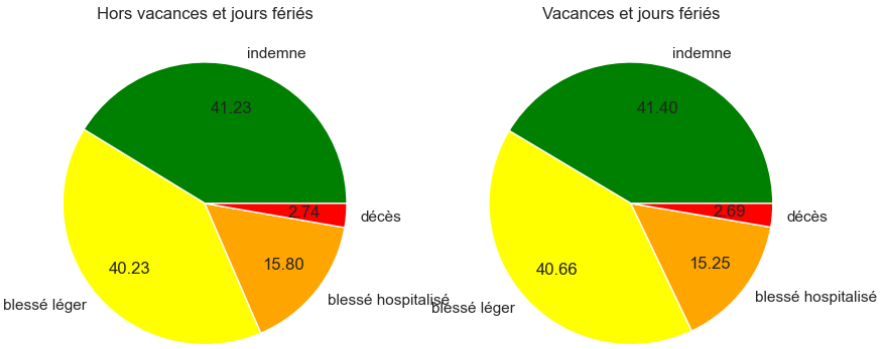


Figure 31 : Proportion des différentes modalités de gravité selon s'il s'agit d'un jour de vacances et jours fériés ou non

L'études des statistiques du χ^2 et du V de Cramer (Figure 32) montre que la gravité est bien dépendante de la variable jour_chome mais qu'elle influe très faiblement sur la gravité. Nous conservons cette variable dans un premier temps.

nom_var	stat_chi2	p_value	V_cramer
jour_chome	21.745782	0.000074	0.006974

Figure 32 : χ^2 et V de Cramer pour la variable jour_chome

Le jeu de données initial permet de connaître la date précise de l'accident grâce aux variables 'an', 'mois' et 'jour' et de créer la variable 'jour_semaine'. Nous avons pu ainsi retrouver le jour de la semaine où l'accident avait eu lieu. L'étude de la gravité en fonction du jour de la semaine (Figure 33) permet de visualiser une différence entre les jours de la semaine :

- un nombre relativement constant des modalités de gravité du lundi au jeudi,
- une augmentation de toutes les modalités de gravité le vendredi (due aux départs en week-end?),
- une augmentation du nombre de tués et blessés hospitalisés et, de manière concomitante, une diminution du nombre de blessés légers et usagers indemnes les samedis et dimanches (due à la pratique de loisirs ?).

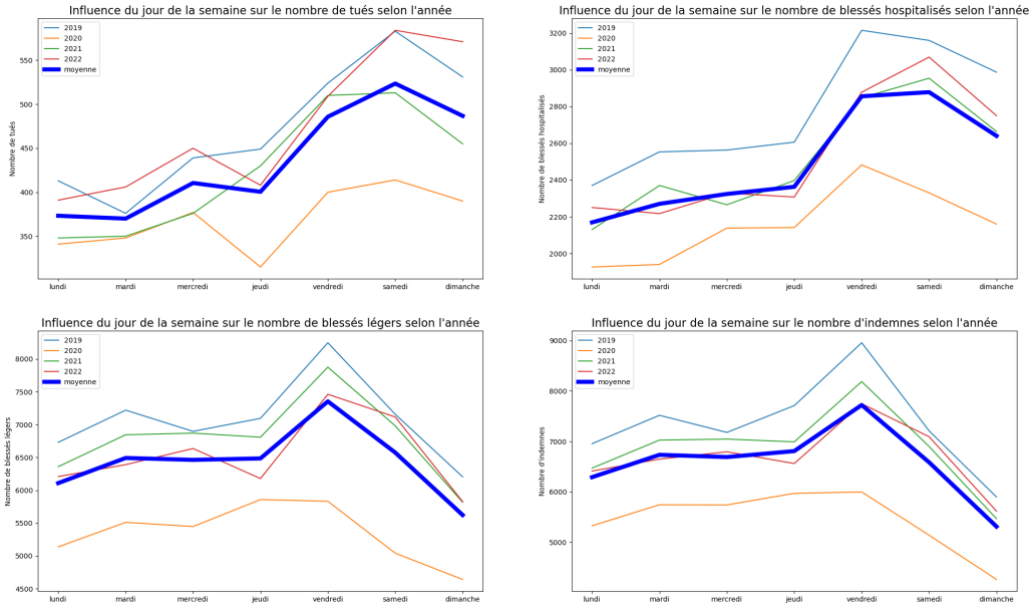


Figure 33 : Courbes du nombre d'usagers pour chaque classe de gravité selon le jour de la semaine pour les années 2019 à 2022

Suite à ces observations, un regroupement en 2 modalités nous semble intéressant : jours de semaine et week-end. En revanche, nous nous demandons s'il est intéressant d'inclure ou non le vendredi dans le weekend. Pour répondre à cela, nous avons comparé les proportions de chaque gravité dans le cas d'une variable binaire weekend (Figure 34 (a)) et d'une variable binaire vendredi et weekend (Figure 34 (b)). La proportion de tués passe de 2.43% à 3.20% le week-end, et celle de blessés hospitalisés de 14.49 à 17.78%, alors que en incluant vendredi dans le week-end, la proportion de tués passe de 2.40% à 2.95%, et celle de blessés hospitalisés de 14.28 à 16.81%. Comme la différence entre les modalités 0 et 1 étant plus importantes pour les blessés hospitalisés et les tués dans le cas de la variable weekend sans le vendredi, nous estimons que cette variable est le meilleur choix.

grav_rec	Indemne	Blessé léger	Blessé hospitalisé	Tué
weekend				
0	42.894608	40.182603	14.489761	2.433029
1	39.555243	39.465550	17.779919	3.199288

(a)

grav_rec	Indemne	Blessé léger	Blessé hospitalisé	Tué
vendredi_weekend				
0	42.953148	40.369549	14.278298	2.399005
1	40.737317	39.495822	16.812417	2.954445

(b)

Figure 34 : Tableaux de contingence pour la variable weekend (a) excluant le vendredi, (b) incluant le vendredi

Pour confirmer ces résultats, nous réalisons un test du χ^2 et calculons la valeur du V de Cramer pour ces deux variables (Figure 35 (a) et (b)). On remarque que dans le cas de la variable vendredi_weekend la p_valeur du test du χ^2 est supérieure à 5%, donc on ne peut pas rejeter que la variable vendredi_weekend soit indépendante de la gravité. Nous gardons donc la variable weekend qui a une p_valeur inférieure à 5%.

nom_var	stat_chi2	p_value	V_cramer
weekend	1057.571219	5.832204e-229	0.048633

(a)

nom_var	stat_chi2	p_value	V_cramer
vendredi_weekend	2.605329	0.456556	0.002414

(b)

nom_var	stat_chi2	p_value	V_cramer
jour_1	43.835374	1.635726e-09	0.009901
jour_3	14.384992	2.425307e-03	0.005672
jour_0	6.041107	1.096273e-01	0.003676
jour_6	5.787469	1.224208e-01	0.003598
jour_4	4.658266	1.986004e-01	0.003228
jour_5	2.636677	4.510958e-01	0.002428
jour_2	1.863872	6.011352e-01	0.002042

(c)

Figure 35 : χ^2 et V de Cramer pour les variables (a). week-end sans le vendredi, (b). weekend avec le vendredi, (c) jour_semaine

De plus, en comparant avec la variable jour_semaine initiale, on observe que la valeur du V de Cramer est nettement améliorée dans le cas de la variable weekend (V_Cramer de 0,048) par rapport à la variable jour_semaine dummisée pour chaque jour de la semaine (Figure 35 (b) et (c)). Le choix de regrouper les jours en weekend ou non nous permet donc d'obtenir une variable avec une relation à plus forte intensité avec la gravité qu'en prenant chaque jour de la semaine séparément.

Au final, nous obtenons une répartition de la gravité selon les jours du lundi au vendredi et le week-end (Figure 36).

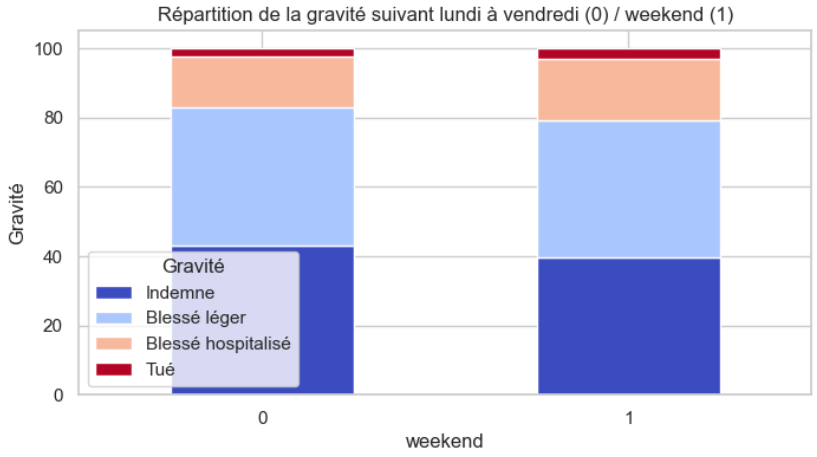


Figure 36 : Répartition de la gravité pour la variable weekend

Heure

Enfin, en récupérant l'heure de l'accident, en l'extrayant de la variable 'hrmn', nous étudions l'influence de l'heure de l'accident sur la gravité (Figure 37). Dans tous les cas de gravité, on observe un pic autour de 17h, s'étalant de 13h à 22h environ, ce qui correspond certainement à une augmentation du trafic routier l'après-midi. De même, le nombre diminue la nuit entre 22h et 4h correspondant aux heures où la circulation est la plus faible. En revanche, on note un léger pic des blessés légers et des usagers indemnes vers 8h, tandis que le nombre de tués augmente progressivement de 4h à 11h.

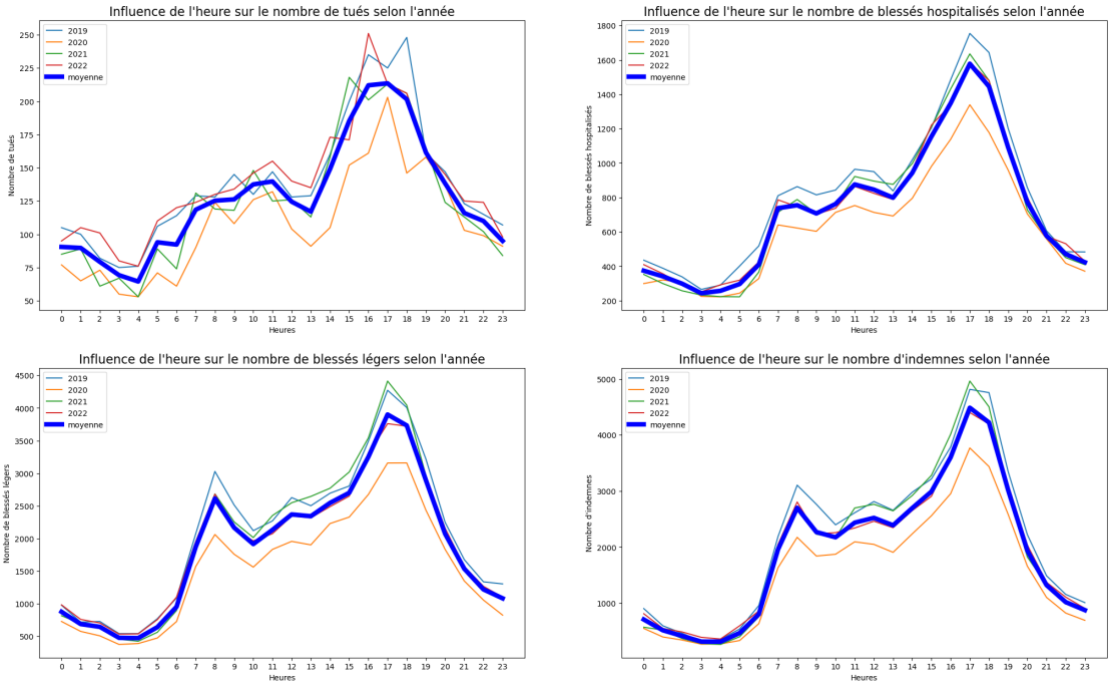


Figure 37 : Courbes du nombre d'usagers pour chaque classe de gravité selon l'heure pour les années 2019 à 2022

Ces variations différentes selon les gravités et l'évolution continue selon les heures pour une même gravité nous amène à conserver l'heure sans la catégoriser.

Disparités spatiales

Localisation des accidents en France métropolitaine

La précision du jeu de données avec les latitudes et longitudes permet de placer précisément la localisation des accidents avec leur gravité (Figure 38). Cette visualisation montre que les accidents ont plus fréquemment lieu au niveau des grandes agglomérations (Paris, Lyon, Lille, Marseille, Bordeaux…) mais aussi sur les principaux axes routiers (on reconnaît facilement le tracé de l'autoroute du Soleil par exemple). Enfin, les routes de la côte méditerranéenne semblent plus propices aux accidents. Ceci peut s'expliquer par la densité de population dans les agglomérations et la fréquentation plus importante des grands axes (notamment la zone méditerranéenne avec l'héliotropisme lors des vacances).

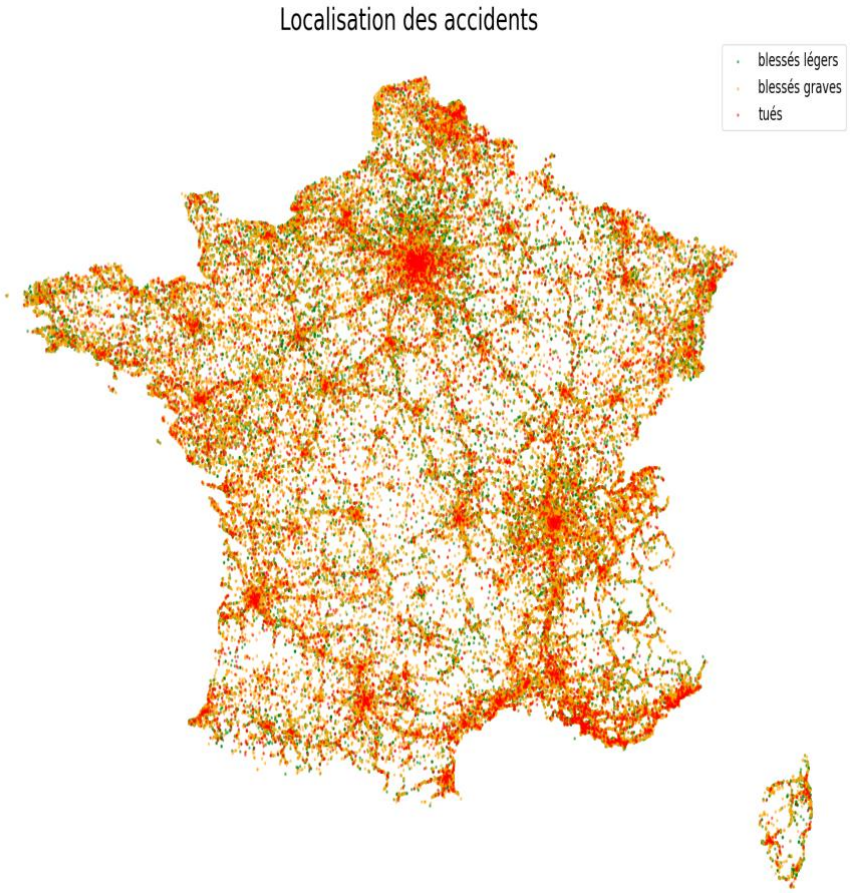


Figure 38 : Carte de la localisation des accidents selon la gravité en France métropolitaine

Si cette carte est révélatrice de la localisation des accidents, elle ne reflète cependant pas la proportion de la gravité des accidents en fonction du nombre d'accidents. C'est pourquoi nous nous sommes intéressés à la gravité des accidents selon la localisation.

Proportion de gravité des accidents selon la localisation

La Figure 39 (a) s'intéresse aux variations régionales dans la répartition des accidents. La taille des camemberts est directement indexée sur le nombre d'accidents par région, tandis que les portions de camembert renseignent sur les différents états de gravité des accidents.

Ainsi, il apparaît que si la majorité des accidents se produit en Ile de France, ce n'est pas dans cette région qu'ils sont les plus meurtriers. Les départements et territoires d'outre-mer, la Bourgogne-Franche-Comté, par exemple, se révèlent en proportion plus touchés par la mortalité routière. Si l'on regarde à l'échelle des départements (Figure 39 (b)), certains départements (Les Landes, la Haute Saône) enregistrent des proportions de décès plus importantes que les autres. Il serait intéressant de voir si les modèles permettront d'identifier quelles spécificités présentent ces départements pour comprendre ces disparités géographiques.

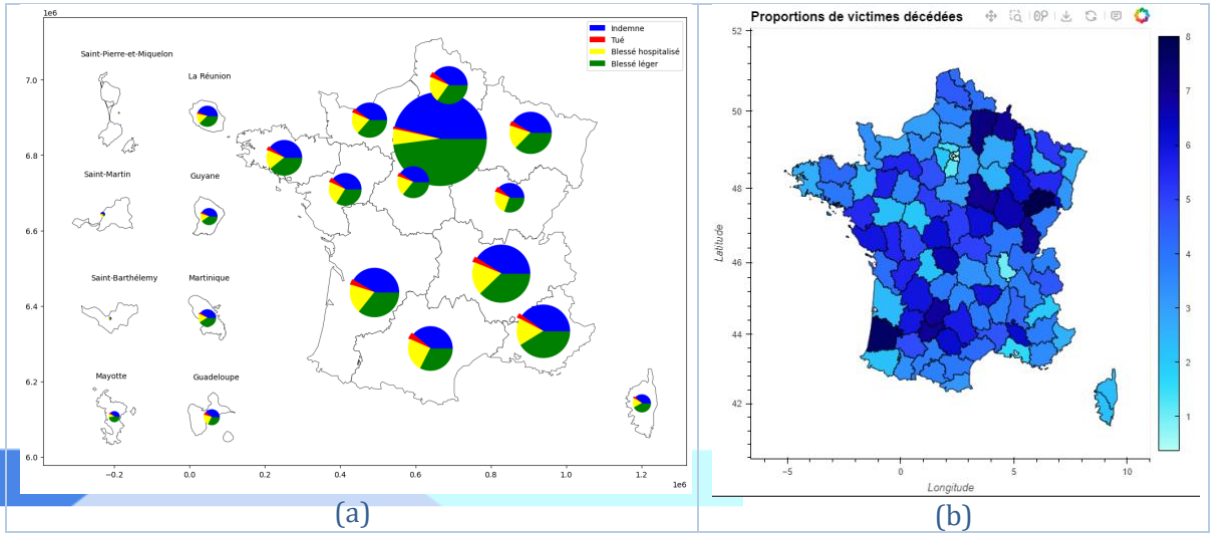


Figure 39 : Cartes (a) de la répartition des états de gravité des accidents par région (La taille des camemberts est proportionnelle au nombre d'usagers impliqués dans les accidents de chaque région), (b) des proportions de victimes décédées selon le département de France métropolitaine.

Influence de l'âge et du sexe dans l'accidentalité routière

Influence de l'âge sur la gravité

La Figure 40 présente la répartition des états de gravité en fonction de l'âge des usagers impliqués. Il est intéressant de noter sur cette figure que :

- **pour les moins de 18 ans**, la part de blessés (légers et hospitalisés) augmente dans les accidents avec l'âge,
- la gravité de l'accident tend à être **plus sévère lorsque l'usager est âgé**. La part de tués (en rouge sur la figure) augmente de manière importante à partir de 60 ans environ, aux dépens de la part des usagers indemnes.

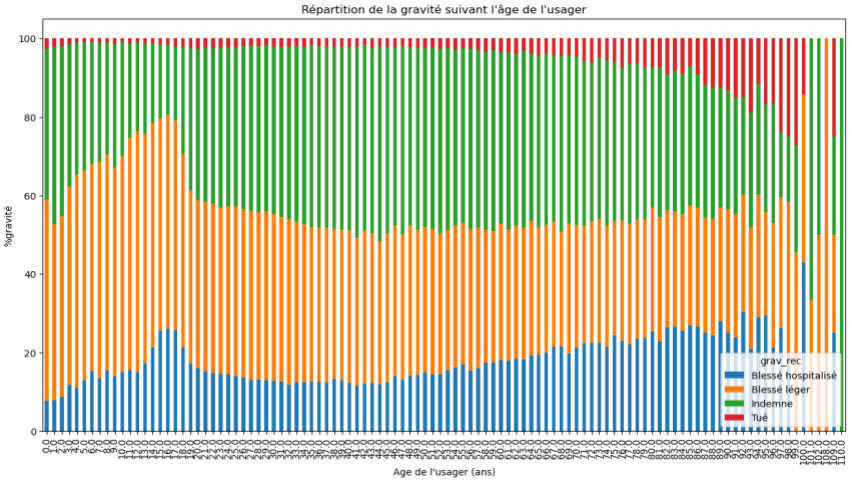


Figure 40 : Graphique de la répartition des modalités de la gravité en fonction de l'âge

Pour la modélisation, deux possibilités ont été envisagées : scinder cette variable en classes d'âges avec des bornes que nous aurions fixées, ou conserver cette variable en tant que variable continue. La seconde option a été choisie pour la modélisation de façon à ne pas perdre en informations en choisissant nous-mêmes des bornes, qui n'auraient donc pas été optimales. Des analyses par arbres de décision nous permettront peut-être de faire ressortir des bornes d'âge plus pertinentes que celles que nous aurions choisies initialement.

Pour l'analyse des dépendances entre variables en revanche, des classes d'âge ont été définies, par tranche de 5 ans.

Influence du sexe sur la mortalité lors d'un accident routier

La Figure 41 analyse la mortalité routière en fonction de l'âge et du sexe des usagers. La **classe d'âge des 20-24 ans est la plus impactée**, quelle que soit le sexe de l'usager. On recense au total 1433 tués dans cette tranche d'âge, contre 609 en moyenne, toutes tranches d'âges confondues. Du côté des hommes, les classes de 15 à 34 ans ont des taux de mortalité élevés. Ce taux tend à diminuer à mesure que l'âge augmente. En revanche, du côté des femmes, les 15 à 25 ans, mais également les plus âgées (au-delà de 70 ans) sont celles où l'on observe des proportions de décès plus importantes du côté des femmes que du côté des hommes.

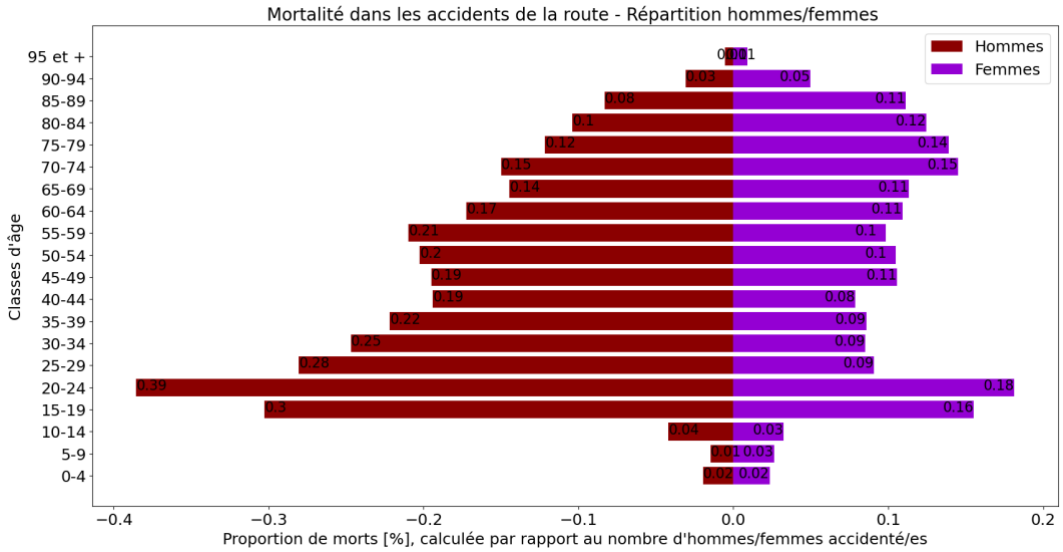


Figure 41 : Graphique de la mortalité des accidents de la route selon le sexe. Le calcul des proportions se base sur 9556 hommes tués, contre 2645 femmes (soit une mortalité masculine 3 à 4 fois supérieure).

Influence de la catégorie de véhicule selon l'âge sur la mortalité

La Figure 42 souligne que la voiture est le principal mode de déplacement impliqué dans les accidents, quelle que soit la catégorie d'âge. Viennent ensuite les motos pour les usagers de moins de 70 ans, et les vélos pour les plus de 70 ans. On note que les transports en commun et les poids lourds représentent une très faible partie des proportions de tués.

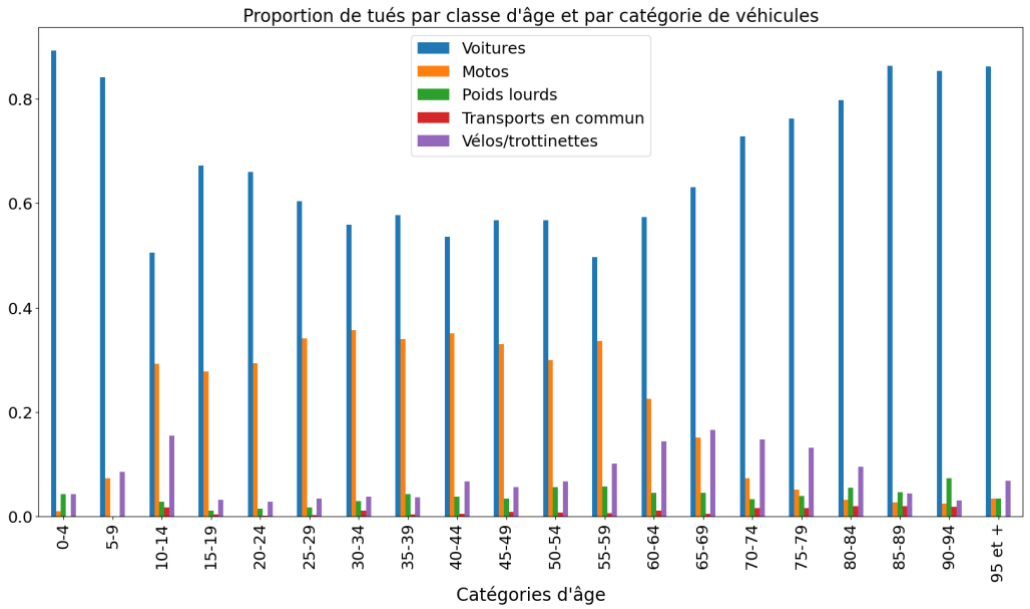


Figure 42 : Proportions, par classes d'âge, des catégories de véhicules associées aux usagers décédés

Comparaison des places occupées selon le sexe dans les accidents

Sur 100 hommes impliqués dans des accidents, 80 sont conducteurs, 7 sont passagers avant, 7 sont passagers arrière et 6 sont piétons. Sur 100 femmes impliquées dans des accidents, 59 sont conductrices, 16 sont passagères avant, 13 sont passagères arrière et 12 sont piétonnes. On observe donc une disparité importante entre les hommes et les femmes dans les places occupées dans les véhicules (Figure 43).

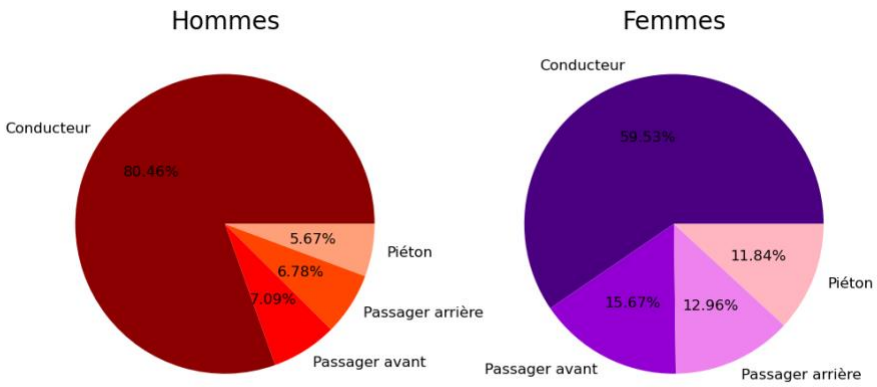


Figure 43 : Comparaison de la position de la personne accidentée selon le sexe

Conclusion

Le problème de prédiction de la gravité des accidents routiers est donc un problème de classification supervisée sur un jeu de données déséquilibré, issu de la base de données des accidents de la route administrée par l'ONISR.

Les grandes étapes de pre-processing et feature engineering ont consisté à :

- Restreindre notre champ d'étude aux années 2019 à 2022,
- Supprimer les doublons,
- Analyser les valeurs manquantes pour supprimer des observations, ou des variables selon l'ampleur du manque d'informations,
- Analyser les relations entre chaque variable et la variable cible pour, selon les cas :
    - Supprimer la variable (ex : numéro du PR de rattachement),
    - Regrouper certaines modalités de la variable, tout en conservant un maximum de relation avec la variable cible (ex : catégorie du véhicule),
    - Conserver la variable telle quelle (ex : type de collision)
- Créer de nouvelles variables, potentiellement intéressantes pour répondre à notre problématique (âge de l'usager, week-end, proximité du point de choc, jour chômé).

A l'issue de cette phase, nous disposons d'un jeu de données de 447136 observations et 98 variables, dont 5 sont des variables continues, les autres étant des variables catégorielles.

En prévision de modélisations impliquant des notions de distance, des procédés de normalisation ont été choisis pour les variables continues : un procédé de Robust Scaling pour la latitude et la longitude, des transformations sinus/cosinus pour l'heure et le mois, une normalisation min/max pour l'âge des usagers.

Une première analyse nous a conduit à identifier les éléments suivants comme facteurs potentiellement importants (à savoir avec une p-valeur à l'issue d'un test d'indépendance du χ^2 inférieure à 5% et un V de Cramer supérieur à 0,1):

- La présence ou non d'équipements de sécurité (ceinture, casque, gants)
- La catégorie d'usagers avec dans l'ordre d'importance identifiée, les motards, les automobilistes, les piétons, les cyclistes, l'implication d'un poids lourds
- La présence d'un obstacle fixe
- La circulation sur route départementale, puis sur voie communale
- Le fait d'être en agglomération
- La collision avec un autre véhicule, ou plusieurs autres véhicules
- La proximité de l'usager avec le point de choc
- Le caractère bidirectionnel de la voie empruntée
- La présence d'une courbe dans le tracé de la route
- L'âge de l'usager
- La circulation de nuit sans éclairage

Quelques facteurs ont été spécifiquement étudiés en fin de ce rapport (analyse temporelle, disparité spatiale, âge et sexe des usagers) pour avoir quelques constats généraux sur leurs liens avec la variable cible, et pour offrir des visualisations intéressantes des tendances observables avec ce jeu de données.

Le jeu de données que nous avons préparé doit dorénavant pouvoir nous servir de base pour l'analyse de notre problème de classification supervisé. Différentes perspectives sont envisagées : l'analyse à l'aide d'arbres de décision pour aller plus loin dans l'identification des facteurs influents, une approche multi-classes de la classification en conservant les 4 classes de gravité (tué, blessé hospitalisé, blessé léger, indemne), et une approche de classification à seulement 2 classes en considérant d'une part les tués et blessés hospitalisés, et d'autre part les blessés légers et usagers indemnes.

L'ensemble des éléments fournis dans ce rapport sont visibles sur le repo GitHub dont l'adresse est : https://github.com/DataScientest-Studio/sept23_cds_accidents2.

Références bibliographiques

Ahmed, Shakil, Md Akbar Hossain, Sayan Kumar Ray, et Md Mafijul Islam Bhuiyan. «A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance.» *Transportation Research Interdiciplinary Perspectives* 19, n° 100814 (2023).

CEREMA. *Evolution de la mortalité sur la période 1968-2022.* 11 02 2024. https://dataviz.cerema.fr/securite-routiere-series/.

Kaleko, David. *Feature Engineering - Handling Cyclical Features.* 30 10 2017. https://blog.davidkaleko.com/feature-engineering-cyclical-features.html (accès le 02 11, 2024).

Lahlou Mimi, Ahmed. *Accidents in France from 2005 to 2016, Kaggle.* 2018. https://www.kaggle.com/datasets/ahmedlahlou/accidents-in-france-from-2005-to-2016.

Maxime, maximeg1. *Predict Severity of Accidents, Kaggle.* 2019.

Ministère de la transition écologique et de la cohésion des territoires. *Données et études statistiques.* 11 02 2024. https://www.statistiques.developpement-durable.gouv.fr/.

Ministère de l'Intérieur et des Outre-Mer. *Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2022.* 2013, màj 2023. (accès le 02 2024).

Observatoire national interministériel de la sécurité routière. 11 02 2024. https://www.onisr.securite-routiere.gouv.fr/.

Routes de France. «Etat de la route - 2023.» 2023. https://www.routesdefrance.com/wp-content/uploads/2023/07/rdf-edlr-2023.pdf.

Statista Research Department. *Statista.* 11 02 2024. https://fr.statista.com/statistiques/943831/moyens-transport-utilises-deplacements-quotidiens-france/.

Talbi, Ilyes. *KGBoost vs Random Forest : prédire la gravité d'un accident de la route.* 6 09 2020. https://larevueia.fr/xgboost-vs-random-forest-predire-la-gravite-dun-accident-de-la-route/.