

RAPPORT DU PROJET DE MACHINE LEARNING

Accidents routiers en France
- de 2005 à 2022 -

Emmanuel GAUTIER
Erika MÉRONVILLE

Table des matières

PARTIE I	2
1 Exploration des données	2
1.1 Préambule	2
1.2 Chargement des données	2
1.3 Découverte des données brutes	3
1.3.1 Analyse des données originales :	3
1.3.2 Analyse des données brutes concaténées :	4
1.3.3 Analyse des données brutes fusionnées :	5
1.4 Constats préalables avant le preprocessing des données :	7
2 Premier prétraitement des données	8
2.1 Etape 1 – Restructuration des dataframes :	8
2.1.1 Création de fonctions spécifiques	8
2.1.2 Considérations préalables	8
2.1.3 Lancement du processus	9
2.1.4 Contrôles avant fusion	9
2.1.5 Fusion des dataframes harmonisés	9
2.2 Etape 2 - Prétraitement après fusion des dataframes :	10
2.2.1 Contrôles après fusion	10
2.2.2 Création de fonctions spécifiques	10
2.2.3 Lancement du processus	10
3 Deuxième prétraitement des données	10
3.1 Base de structuration des dataframes :	11
3.1.1 Structure pour 'caracteristiques' :	11
3.1.2 Structure pour 'lieux' :	12
3.1.3 Structure pour 'usagers' :	13
3.1.4 Structure pour 'vehicules' :	13
3.2 Etapes de transformation des dataframes	14
3.2.1 Initialisation de l'environnement	14
3.2.2 Enregistrement des 4 rubriques de 2019 à 2022	15
3.2.3 Jointure des 4 rubriques	15
3.2.4 Suppression d'observations	15
3.2.5 Crédit de variables	16
3.2.6 Dichotomisation / catégorisation	16
3.2.7 Suppression de variables	17
3.2.8 Suppression de doublons	18
3.2.9 Équilibrage de la dimension	18
3.2.10 Synthèse des actions de préprocessing	20
3.2.11 Sauvegarde des données	21
PARTIE II	22
1 Modélisation 1	22
1.1 Implémentation d'un premier modèle	22
1.2 Évaluation des performances du modèle	24
1.3 Analyse des résultats	24
2 Modélisation 2	25
2.1 Implémentation de multiples modèles de classification	25
2.2 Évaluation des performances des modèles	28
2.3 Analyse des résultats	29

3 Modélisation	33
3.1 Implémentation d'un modèle de deep learning	33
3.2 Evaluation des performances	34
3.3 Analyse des résultats	35
4 Conclusion	39
CONCLUSION	41

Ce rapport présente le processus d'exploration et de prétraitement d'un jeu de données complexe sur les accidents de la route en France, couvrant la période de 2005 à 2022. L'objectif principal de cette analyse est de préparer les données pour un projet de machine learning visant à modéliser et prédire la gravité des accidents routiers.

De manière générale, les accidents routiers peuvent être déterminés par plusieurs facteurs. En ce sens, les comportements des conducteurs jouent parfois un rôle central : vitesse excessive, alcool, téléphones portables, fatigue sont des acteurs récurrents dans cette tragédie. Les conditions environnementales, telles que la météo et l'état des routes, ajoutent une couche d'incertitude et de danger. Les véhicules eux-mêmes, de par leur âge et leur entretien, influencent le risque d'accident. Les infrastructures routières, avec leurs forces et leurs faiblesses, façonnent également le paysage de cette problématique.

Bien que les statistiques montrent une tendance constante à la baisse des accidents de la route, leur complexité réside dans la multitude de facteurs et de circonstances uniques à chaque incident. Aussi, l'analyse des données est le phare qui éclaire la voie à suivre. En examinant les statistiques des accidents, les profils des conducteurs, et les conditions spécifiques de chaque incident, nous pouvons discerner des patterns, anticiper les risques, et élaborer des stratégies préventives.

Dans ce contexte, notre projet de machine learning a pour objet de recourir aux algorithmes sophistiqués qui permettront de transformer des événements isolés en précieuses données capables de prédire la gravité de futurs accidents. Notre investigation se porte ainsi sur un dataset constitué de plusieurs fichiers téléchargés librement à partir de l'adresse suivante : <https://www.data.gouv.fr/en/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2022/>.

Le jeu de données, fourni par le gouvernement français, est composé de quatre fichiers principaux :

- caractéristiques : informations générales sur les circonstances de l'accident ;
- lieux : détails sur l'emplacement de l'accident ;
- véhicules : informations sur les véhicules impliqués ;
- usagers : données sur les personnes impliquées dans l'accident.

Ces données sont d'une importance cruciale pour la sécurité routière, car elles peuvent aider à identifier les facteurs de risque et à élaborer des stratégies de prévention plus efficaces.

L'analyse présentée dans ce rapport suit plusieurs étapes clés :

- exploration initiale des données : examen de la structure, des types de variables, et des valeurs pour chaque fichier ;
- analyse détaillée de chaque variable : étude de la distribution, des valeurs manquantes, des outliers, et de l'évolution temporelle ;
- identification des problèmes potentiels : repérage des incohérences, des changements de codification, et des valeurs aberrantes ;
- propositions de prétraitement : suggestions pour le nettoyage, la transformation et la création de nouvelles variables.
- modélisation : application de modèles de machine learning et évaluation de leurs performances à l'aide de métriques.

Une attention particulière a été portée aux changements survenus dans la collecte et le codage des données au fil des années, notamment les modifications importantes intervenues à partir de 2019.

Ce travail d'exploration et de préparation des données représente une étape décisive dans un processus plus large de data science, allant de la compréhension initiale du problème à la mise en œuvre de solutions basées sur les données. Il est crucial pour assurer la qualité et la fiabilité du modèle de machine learning qui sera développé par la suite, l'objectif final étant de créer un outil capable de prédire la gravité des accidents.

Cette démarche illustre parfaitement comment un Data Scientist contribue à transformer des données brutes en connaissances actionnables et en modèles prédictifs performants, jetant ainsi les bases solides nécessaires pour des analyses avancées et des prises de décision éclairées dans le domaine de la sécurité routière.

PARTIE I

1 Exploration des données

Les données concernent 72 dataframes au total, soit 1 dataframe par année et par rubrique. Voici les rubriques concernées :

- caractéristiques : qui prend en compte les circonstances générales de l'accident.
- lieux : qui décrit l'endroit de l'accident.
- véhicules : qui énonce les véhicules impliqués dans l'accident.
- usagers : qui relate les usagers impliqués dans l'accident.

Chaque nom de fichier contient le nom de la rubrique, un tiret normal ou bas, l'année avec 4 chiffres et l'extension '.csv'. Les séparateurs sont la virgule, le point-virgule ou la tabulation. Les noms des colonnes sont très courts et principalement écrits en lettres minuscules : ils ont quelques chiffres et majuscules, et suivent les conventions de nommage des variables de python. Avant d'effectuer l'assemblage des dataframes, il est procédé à leur analyse distincte (année par année pour chaque rubrique) en vue d'une meilleure appréhension des données. Plusieurs étapes se succèdent alors :

1.1 Préambule

Toutes les bibliothèques nécessaires à la manipulation des données, l'analyse statistique, la visualisation des données, et la préparation des données sont préalablement chargées.

Par ailleurs, un système de journalisation est configuré dès le début, pour nous assurer que toutes les actions, messages d'information, erreurs, et avertissements soient correctement capturés et enregistrés, pour faciliter la maintenance, le débogage et la documentation du processus.

1.2 Chargement des données

Les fichiers CSV sur lesquels nous avons travaillé ont comme particularités des variations de format, d'encodage, de nommage ou autres :

- encodage : « utf-8 », « latin1 », « ISO-8859-1 »
- séparateur : « , », « ; », « \t »
- connecteur : « - », « _ »
- nommage : « caractéristiques », « carcteristiques »
- année : de « 2005 » à « 2022 »

Il nous a donc fallu prendre en compte ces différenciations pour parvenir à une uniformisation de ces derniers. Le code que nous avons créé pour automatiser le processus de chargement des fichiers prend notamment en considération les différents points suivants :

- Détection automatique du délimiteur : La fonction 'get_delimiter' analyse un échantillon du fichier CSV pour identifier automatiquement le caractère utilisé comme séparateur (virgule, point-virgule, etc.).
- Lecture flexible des fichiers CSV : La fonction 'read_csv_file' tente de lire le fichier CSV en utilisant différents encodages (UTF-8, Latin-1, ISO-8859-1) pour gérer les variations possibles dans l'encodage des caractères. Elle utilise également le délimiteur détecté précédemment.
- Chargement systématique des datasets : La fonction 'load_datasets' parcourt une série de fichiers CSV basés sur des préfixes spécifiques (caractéristiques, lieux, usagers, véhicules) et des années (de 2005 à 2022). Elle gère les différences de nommage des fichiers selon l'année (utilisation de '_' ou '-' comme connecteur).

- Gestion des erreurs et logging : Le code intègre un système de gestion des erreurs et de logging pour suivre le processus de chargement, signaler les problèmes rencontrés (fichiers manquants, erreurs de lecture) et fournir un résumé des datasets chargés avec succès.
- Flexibilité et réutilisabilité : Le code est conçu pour être facilement adaptable à différents ensembles de données en permettant la spécification des préfixes de fichiers et des années à traiter.

Nota bene : La prise en compte des caractéristiques propres à chaque dataframe permet la lecture et le chargement de tous les dataframes téléchargés.

1.3 Découverte des données brutes

1.3.1 Analyse des données originales :

- **Exploration primaire :**

Chacun des dataframes fait l'objet d'une première analyse isolément, afin de repérer d'éventuelles anomalies. La méthode consiste alors à :

- créer une fonction d'extraction du nom du fichier CSV pour obtenir l'année concernée en utilisant une expression régulière (fonction extract_year).
- créer une fonction visant à analyser chaque colonne dans tous les datasets (analyze_column), en collectant les informations suivantes :
 - l'année du dataset,
 - le type de données de la colonne,
 - la valeur la plus fréquente (mode),
 - la proportion de valeurs nulles dans le fichier,
 - la proportion de valeurs nulles par rapport à l'ensemble des données.
- analyser globalement toutes les colonnes de chacun des datasets (fonction analyze_all_columns) en :
 - identifiant toutes les colonnes uniques,
 - calculant le nombre total de lignes dans tous les datasets,
 - générant un logging permettant de suivre le processus de chargement de chaque fichier,
 - appellant la fonction analyze_column pour chaque colonne unique.
- présenter les résultats de l'analyse sous forme de tableau pour chaque colonne de tous les datasets, pour faciliter la comparaison entre les différentes années.

L'analyse ainsi effectuée a permis de relever les points suivants :

- La variable ‘Num_Acc’ (ou son équivalent ‘Accident_Id’) est présente dans chaque dataframe, elle permet de lier les dataframes des différentes rubriques entre eux.
- D'une année à l'autre, le type des variables varie parfois, ce qui doit être pris en compte par la suite pour le prétraitement des données.
- Il existe quelques variables ayant des valeurs nulles, et parfois ces valeurs sont encore plus présentes pour les années antérieures à 2019.
- Certaines variables (comme ‘vma’) ne sont existantes que depuis 2019, alors que d'autres (comme ‘secu’) ont disparu à partir de 2019 lorsqu'elles ne sont pas remplacées par une autre (exemple : ‘Num_Acc’ qui devient ‘Accident_Id’ en 2022).

- **Visualisation graphique :**

Pour une analyse plus approfondie de la structure et de l'évolution des données, nous avons fait appel aux outils de visualisation. L'objectif était de :

- comprendre la distribution des valeurs numériques et noter la présence d'éventuelles valeurs aberrantes (boxplots) ;

- visualiser la répartition des modalités pour les variables catégorielles (barplots empilés).
 - suivre l'évolution de la distribution des modalités par année (lineplots).
- Dans ce cadre, notre démarche a consisté à :
- créer des boxplots ('create_boxplot_grid') qui :
 - génèrent une grille de boxplots pour visualiser la distribution des valeurs numériques d'une colonne spécifique à travers les années,
 - ignorent les colonnes non numériques,
 - permettent une comparaison visuelle rapide ;
 - créer des barplots empilés ('create_total_stacked_barplot') qui :
 - visualisent la distribution des modalités (valeurs uniques) d'une colonne à travers les années,
 - limitent le nombre de modalités affichées pour éviter la surcharge visuelle,
 - présentent les données sous forme de barres empilées, en montrant l'évolution de la proportion de chaque modalité au fil du temps ;
 - créer des lineplots ('create_lineplot_evolution') qui :
 - montrent l'évolution de la distribution des modalités d'une colonne au fil des années,
 - affichent chaque modalité comme une ligne distincte, permettant de suivre son évolution en pourcentage,
 - limitent également le nombre de modalités pour maintenir la lisibilité ;
 - orchestrer les visualisations ('one_column_generate_plots') :
 - en combinant les trois types de visualisation (boxplot, barplot, lineplot) pour une colonne donnée,
 - analyser globalement ('all_columns_generate_plots') :
 - en générant toutes les visualisations graphiques de chaque colonne unique trouvée dans l'ensemble des datasets.

Nota bene : Tous les résultats obtenus sont présentés en détails dans la partie annexe du présent rapport lorsqu'ils s'avèrent pertinents. Combinés aux éléments statistiques qui suivront, ils nous seront utiles notamment à l'occasion du nettoyage des données pour juger de l'intérêt de les conserver, regrouper, supprimer, ...

1.3.2 Analyse des données brutes concaténées :

Pour une exploration plus poussée des données, il est procédé à une concaténation des fichiers CSV en un seul DataFrame pour chaque type de données. Il s'agit donc d'aller un peu plus loin dans la recherche d'éventuelles anomalies.

- **Concaténation des dataframes par rubrique :**

Notre première étape de prétraitement des données consiste à transformer les ensembles de données fragmentés en une ressource cohérente et plus facilement exploitable. Elle passe par plusieurs étapes :

- Le processus de transformation commence par l'initialisation d'un dictionnaire qui servira à stocker les données consolidées.
- Ensuite, le code parcourt systématiquement chaque préfixe (type de données) et ses fichiers associés. Pour chaque préfixe, il collecte tous les DataFrames correspondants, quelle que soit leur année d'origine.
- Une fois tous les DataFrames d'un type particulier rassemblés, le code les fusionne en un seul grand DataFrame. Cette opération est répétée pour chaque type de données, résultant en un ensemble de DataFrames consolidés, chacun représentant un aspect spécifique des données sur toute la période étudiée.

L'intérêt principal de cette approche est de simplifier considérablement la structure des données. Au lieu d'avoir de nombreux fichiers séparés par année, on obtient un ensemble réduit de DataFrames, chacun contenant toutes les données d'un type particulier sur l'ensemble de la période. Cette consolidation facilite grandement les analyses ultérieures, notamment des analyses statistiques plus complexes.

Au final, on obtient un aperçu de la taille de chaque DataFrame consolidé, qui offre une vision plus claire de la quantité de données disponibles pour chaque aspect étudié. Cette information est précieuse pour comprendre la portée des données à disposition.

Les dimensions des 4 DataFrames ainsi obtenus (soit 1 par rubrique) suivent :

- 'caracteristiques' : 1 176 873 lignes x 17 colonnes
- 'lieux' : 1 176 873 lignes x 19 colonnes
- 'usagers' : 2 636 377 lignes x 17 colonnes
- 'vehicules' : 2 009 395 lignes x 11 colonnes

- **Description des variables :**

Après avoir fourni leur brève description, chaque variable fait l'objet de statistiques, pour obtenir une vue d'ensemble de la structure et du contenu de chaque ensemble de données. La liste suivante reprend les éléments ayant servi de base à une meilleure compréhension de ces données, et les détails correspondants sont rendus disponibles en annexe du présent rapport :

- **Définition des variables :**

Une documentation en ligne décrivant toutes les variables existe.

- **Etendue des valeurs :**

Le code correspondant utilise la fonction 'stats_descriptives' pour calculer des statistiques descriptives sur plusieurs DataFrames stockés dans le dictionnaire 'concatenated_dfs'. Pour chaque colonne de chaque DataFrame, il détermine les éléments suivants : le nombre de valeurs non nulles, le nombre de valeurs uniques, la valeur la plus fréquente (mode), et la fréquence de cette valeur. Les résultats sont ensuite stockés dans le dictionnaire 'resultats_par_dataframe' et affichés de manière détaillée pour chaque DataFrame, ainsi qu'un résumé synthétique, facilitant ainsi l'exploration et l'analyse des données.

- **Valeurs nulles :**

Le code correspondant définit la fonction 'analyser_valeurs_nulles', qui calcule et affiche les valeurs nulles et non nulles ainsi que le taux de valeurs nulles pour chaque colonne d'un DataFrame. Pour chaque colonne, il utilise les méthodes notnull() et isnull() pour compter les valeurs non nulles ('val_notnull') et nulles ('val_null'), et calcule le pourcentage de valeurs nulles ('tx_null'). Les résultats sont stockés dans un dictionnaire 'resultats' et affichés de manière détaillée, suivis d'un résumé synthétique pour chaque DataFrame.

- **Outliers :**

Le code correspondant est conçu pour détecter les valeurs aberrantes dans chaque colonne d'un DataFrame, qu'elle soit numérique ou catégorielle. La fonction 'detect_outliers_numeriques' utilise l'écart interquartile (IQR) pour identifier les outliers dans les colonnes numériques, tandis que la fonction 'detect_outliers_catégoriels' détecte les outliers dans les colonnes catégorielles en se basant sur la fréquence des valeurs. La fonction 'analyser_outliers' compile ces informations dans un dictionnaire nommé 'resultats', qui contient le type de données, le nombre d'outliers, le nombre d'outliers uniques et une liste des premiers outliers détectés pour chaque colonne. Les résultats sont ensuite stockés dans le dictionnaire 'resultats_par_dataframe' pour chaque DataFrame concaténé, et affichés de manière détaillée et résumée.

- **Répartition :**

Le code créé permet d'analyser le nombre de modalités et leur taux pour chaque variable dans plusieurs DataFrames. La fonction 'safe_percentage' calcule le pourcentage de chaque modalité en fonction du nombre total de valeurs. La fonction 'analyser_modalites' crée un dictionnaire nommé 'resultats', qui contient pour chaque colonne un DataFrame avec les modalités, leur nombre (Count), et le pourcentage de valeurs (% valeurs) calculé à l'aide de 'safe_percentage'. Les résultats sont ensuite stockés dans le dictionnaire 'resultats_par_dataframe' pour chaque DataFrame concaténé, et affichés de manière détaillée pour chaque colonne.

1.3.3 Analyse des données brutes fusionnées :

Pour continuer notre analyse, il s'avère nécessaire de fusionner les DataFrames et de réaliser d'autres tests statistiques.

- **Fonction pour fusionner les DataFrames :**

Les étapes du code créé sont les suivantes :

- D'abord, il concatène les données de même type provenant de différentes sources ou périodes. Pour chaque catégorie (usagers, caractéristiques, lieux, véhicules), il combine tous les DataFrames disponibles en un seul. Cette étape consolide les informations par type, regroupant par exemple toutes les données sur les usagers en un seul DataFrame.
 - Ensuite, le code prépare la fusion finale de ces DataFrames consolidés. Il définit l'ordre dans lequel les différentes catégories de données seront combinées et spécifie les clés de jonction pour chaque fusion.
- Le processus de fusion commence alors avec le DataFrame des usagers comme base. Le code y ajoute successivement les informations des autres catégories : d'abord les caractéristiques des accidents, puis les données sur les lieux, et enfin les informations sur les véhicules impliqués.
- Chaque fusion est réalisée en utilisant des identifiants spécifiques, principalement le numéro d'accident ('Num_Acc'), pour s'assurer que les informations sont correctement associées. Pour les véhicules, la fusion utilise plusieurs clés pour gérer la complexité supplémentaire de cette catégorie. Le type de fusion utilisé ('left') garantit que toutes les données des usagers soient conservées, même si certaines informations correspondantes manquent dans les autres catégories. Cela permet de préserver l'intégrité des données sur les personnes impliquées dans les accidents.
- Tout au long du processus, le code enregistre des informations sur le nombre de lignes dans chaque DataFrame fusionné et dans le DataFrame final. Ces logs permettent de suivre l'évolution de la taille des données et de vérifier que la fusion se déroule comme prévu.
 - Le résultat final est un unique DataFrame nommé 'accidents', qui contient toutes les informations sur les accidents, les usagers impliqués, les lieux où ils se sont produits et les véhicules concernés. Ce DataFrame unifié facilite grandement les analyses à venir en fournissant une vue complète et intégrée de chaque accident dans un seul ensemble de données cohérent.

- **Tests statistiques :**

Afin d'appréhender l'orientation de nos recherches, quelques tests statistiques sont réalisés en vue d'identifier les facteurs pouvant avoir une relation significative avec la gravité des accidents de la route.

Voici la description linéaire de notre approche effectuée en plusieurs étapes :

- Nous avons créé un code qui prépare les données en nettoyant le DataFrame 'accidents'. Celui-ci remplace les valeurs problématiques par NaN et convertit certaines colonnes en format numérique. Il calcule également l'âge des personnes impliquées dans les accidents.
- Ensuite, le code définit une fonction d'analyse statistique nommée 'run_statistical_tests'. Cette fonction est conçue pour appliquer différents tests statistiques selon le type de variables. Pour les variables numériques, elle utilise des tests de corrélation de Spearman et de Pearson. Pour les variables catégorielles, elle applique le test du Chi-carré. De plus, elle effectue une analyse de variance (ANOVA) pour toutes les variables.
- Le code établit ensuite une liste exhaustive de caractéristiques à analyser. Ces caractéristiques couvrent divers aspects des accidents, incluant les informations sur les usagers, les véhicules, les conditions de l'accident, la route, le temps et la localisation.
- Après cette préparation, le code exécute les tests statistiques. Il applique la fonction 'run_statistical_tests' à toutes les caractéristiques sélectionnées, en utilisant 'grav' (la gravité de l'accident) comme variable cible.
- Enfin, le code organise et présente les résultats. Il compile tous les résultats des tests dans un DataFrame, les trie par valeur-p pour mettre en évidence les relations les plus significatives, et les affiche de manière complète et lisible.

Types de Tests Utilisés :

- Spearman (Corrélation de Spearman) : Mesure la corrélation monotone entre deux variables. Utilisé pour des données ordinaires ou lorsque les relations ne sont pas linéaires.
- Pearson (Corrélation de Pearson) : Mesure la corrélation linéaire entre deux variables continues.
- ANOVA (Analyse de la Variance) : Teste les différences entre les moyennes de plusieurs groupes pour une variable continue.

- Chi2 (Test du Chi-carré d'indépendance) : Utilisé pour déterminer s'il existe une association significative entre deux variables catégorielles.

P-Values et Significativité :

Une p-value proche de 0 (par exemple, 0.000000e+00) indique que le test statistique a trouvé une relation significative (corrélation, différence, ou association) entre les variables testées.

Une p-value supérieure au seuil de 0,05 signifie qu'il n'y a pas de relation significative.

Interprétation des Résultats des Tests :

- Corr_significative : Indique une corrélation significative entre la variable d'intérêt et une autre variable.
- Diff_significative : Indique une différence significative entre les groupes de la variable testée.
- Ass_significative : Indique une association significative entre deux variables catégorielles.
- Corr_non_significative / Diff_non_significative : Indique l'absence de corrélation ou de différence significative.

Voici un extrait des résultats des tests statistiques :

Index	Feature	Test	P-value	Résultat
0	age	Spearman	0.000000e+00	Corr_significative
59	obsm	Spearman	0.000000e+00	Corr_significative
58	obs	ANOVA	0.000000e+00	Diff_significative
57	obs	Pearson	0.000000e+00	Corr_significative
56	obs	Spearman	0.000000e+00	Corr_significative
55	catv	ANOVA	0.000000e+00	Diff_significative
54	catv	Pearson	0.000000e+00	Corr_significative
52	trajet	ANOVA	0.000000e+00	Diff_significative
51	trajet	Pearson	0.000000e+00	Corr_significative
49	choc	ANOVA	0.000000e+00	Diff_significative

Résultats Significatifs

- CORRÉLATIONS ET DIFFÉRENCES SIGNIFICATIVES :

Les variables telles que **age**, **obs**, **catv**, **trajet**, **choc**, **secu1**, **secu2**, **sexe**, **manv**, et **etatp** ont des résultats de tests Spearman, Pearson, ou ANOVA avec des p-values très proches de zéro (0.000000e+00). Cela signifie qu'il existe des corrélations significatives ou des différences significatives entre ces variables et la variable cible ou entre elles. Par exemple, pour la variable age, les tests de corrélation de Spearman et de Pearson ainsi que le test ANOVA ont tous une p-value de 0.000000e+00, ce qui indique une forte corrélation et des différences significatives liées à l'âge.

- ASSOCIATIONS SIGNIFICATIVES (CHI2) :

Les variables **voie**, **pr**, **com**, **dep**, **gps**, **v2**, **actp**, et **adr** montrent des associations significatives avec des p-values de 0.000000e+00 ou proches de zéro dans les tests du Chi2. Cela suggère que ces variables catégorielles sont statistiquement associées entre elles ou avec d'autres variables d'intérêt.

- RÉSULTATS NON SIGNIFICATIFS :

Certaines variables, comme **vosp**, **jour**, **agg**, **voie**, **adr**, **lat**, et **long**, ont des résultats de tests avec des p-values élevées (par exemple, 2.143277e-01, 3.827203e-01, jusqu'à 1.000000e+00). Cela signifie que ces tests n'ont pas trouvé de corrélation, de différence, ou d'association statistiquement significative.

1.4 Constats préalables avant le preprocessing des données :

L'exploration approfondie des données nous a permis d'avoir une meilleure compréhension de notre jeu de données. Nous avons examiné la structure des données, identifié les principales variables, visualisé les distributions, et effectué des tests statistiques préliminaires.

Ces analyses exploratoires ont révélé plusieurs points importants :

- La complexité des informations disponibles dans les différentes rubriques (caractéristiques, lieux, usagers, véhicules).

- La présence de valeurs manquantes, aberrantes ou codées de manière inconsistante dans certaines variables.
- Des relations statistiquement significatives entre plusieurs variables et la gravité des accidents.
- La nécessité de traiter les problèmes de qualité des données et de préparer celles-ci pour une modélisation efficace.

Ces constats mettent en évidence la nécessité d'un prétraitement des données avant toute modélisation fiable. Le prochain chapitre sera donc consacré au nettoyage des données en vue de notre première itération. Cette étape nous permettra d'évaluer l'efficacité du prétraitement et d'identifier les améliorations potentielles pour optimiser notre approche de modélisation.

2 Premier prétraitement des données

2.1 Etape 1 – Restructuration des dataframes :

2.1.1 Création de fonctions spécifiques

La fonction 'convert_dtypes' permet de corriger les variations de type constatées dans certaines colonnes des DataFrames 'df'. Elle parcourt les colonnes du DataFrame, tente de les convertir aux types spécifiés dans le dictionnaire de référence 'reference_dtypes', et gère les erreurs de conversion. La fonction retourne le DataFrame avec les types de données ajustés, tout en enregistrant les éventuels échecs de conversion.

La fonction 'extract_reference_structure' crée un dictionnaire 'reference_structures' à partir d'une liste de 'dataframes'. Pour chaque paire prefix-dataframe, elle extrait les 'dtypes' du dernier DataFrame de la liste, les associe au 'prefix' correspondant dans 'reference_structures'. Le résultat est un dictionnaire où chaque clé 'prefix' est associée aux types de données de son dernier DataFrame.

La fonction 'preprocess' prend en entrée un DataFrame 'df' et son 'prefix'. Elle applique une série de transformations au DataFrame selon la valeur de son 'prefix' et le retourne nettoyé et standardisé.

La fonction 'preprocess_datasets' permet d'appliquer la fonction 'preprocess' à chaque DataFrame 'df' contenu dans chaque 'df_dict' de chaque 'df_list', en utilisant le 'prefix' associé. Les DataFrames traités sont ensuite replacés dans 'df_dict' sous leur 'file_name' d'origine. La fonction retourne les 'dataframes' prétraités, en conservant leur structure initiale.

La fonction 'harmonize_dataframes' vise à harmoniser les colonnes et le type de données pour chaque dataframe, selon une structure de référence 'reference_structures' définie par la fonction 'convert_dtypes'.

La fonction 'compare_structures' compare les structures des dataframes harmonisés (résultant de la fonction 'harmonize_dataframes') par rapport à la structure de référence (émanant de la fonction 'convert_dtypes'), afin de repérer les différences pouvant exister en termes de colonnes et de types de données.

La fonction 'concat_harmonized_dataframes' crée une liste de dataframes concaténés 'concatenated_dataframes', en vue d'une utilisation ultérieure.

La fonction 'merge_dataframes' fusionne l'ensemble des dataframes concaténés ('concatenated_dataframes') en considération des spécificités mentionnées au niveau de chaque jointure :

Une fois la définition des fonctions achevées, celles-ci sont exécutées pour mettre en œuvre les étapes d'harmonisation des dataframes.

2.1.2 Considérations préalables

La modification de la codification de la variable cible 'grav' en 2019 représente une cause majeure de l'hétérogénéité structurelle des dataframes annuels, marquant une rupture significative dans la continuité des données. Pour assurer une cohérence optimale de notre étude, nous avons décidé de ne garder que les données de la période 2019-2022. Ce choix est conforté par l'introduction en 2019 de nouvelles variables, telles que 'vma' (vitesse maximale autorisée), et par une restructuration générale des données (exemple : 'secu1', 'secu2', etc.). Cette approche permet de travailler avec des données plus récentes et

plus homogènes, facilitant l'uniformisation nécessaire pour les prochaines étapes de prétraitement et de modélisation.

2.1.3 Lancement du processus

Un code prépare les données des accidents routiers pour analyse. Il commence par définir la plage des années considérées (2019-2022) et les catégories de données (caractéristiques, lieux, usagers, véhicules). Ensuite, il charge ces données au moyen des fonctions préalablement définies.

2.1.4 Contrôles avant fusion

Un code fournit un aperçu rapide du nombre de datasets chargés pour chaque catégorie de données (caractéristiques, lieux, usagers, véhicules) et indique le nombre total de datasets chargés pour toutes les catégories confondues.

Un autre code parcourt un dictionnaire pour examiner les structures de référence définies pour chaque catégorie de données, et log chacune d'elles.

La fonction 'compare_structures' examine chaque dataframe harmonisé et le compare à la structure de référence correspondante. Elle permet de détecter les anomalies potentielles résultant du processus d'harmonisation, telles que des colonnes manquantes, des colonnes inattendues, ou des incohérences dans les types de données.

Voici un extrait des résultats observés :

2024-09-10 12 :04 :02,235 - INFO - Comparaison pour le fichier data/raw/usagers-2019.csv :	
Type de différence	Colonnes
Colonnes manquantes	id_usager
Colonnes supplémentaires	
Différences de types	an_nais (réf : float64, fichier : int64)
2024-09-10 12 :04 :02,246 - INFO - Comparaison pour le fichier data/raw/usagers-2020.csv :	
Type de différence	Colonnes
Colonnes manquantes	id_usager
Colonnes supplémentaires	
Différences de types	an_nais (réf : float64, fichier : int64)

Grâce à l'affichage des premières lignes de chaque dataframe harmonisé, un code est créé pour permettre de vérifier visuellement si toutes les colonnes attendues sont présentes dans chaque dataframe et de repérer si les types de données semblent corrects et cohérents.

Un nouveau code met en mémoire la fonction 'concat_harmonized_dataframes' destinée à fusionner tous les dataframes harmonisés de chaque catégorie (préfixe) en un seul dataframe par catégorie.

Un autre code appelle le précédent en créant une boucle qui parcourt les dataframes concaténés pour associer chaque dataframe à son préfixe. Le code affiche la dimension de chaque ensemble de données consolidé (nombre de lignes et de colonnes) pour chaque dataframe concaténé.

Observations :

L'exécution progressive de cet ensemble de boucles et de fonctions a permis de vérifier le bon déroulement du processus de chargement et de structuration des données avant d'aboutir à la fusion des dataframes.

2.1.5 Fusion des dataframes harmonisés

Il est procédé à une fusion des DataFrames harmonisés et les résultats obtenus appellent de nouvelles vérifications.

2.2 Etape 2 - Prétraitement après fusion des dataframes :

2.2.1 Contrôles après fusion

La fonction 'check_nan_presence' effectue un contrôle sur le dataframe final fusionné et calcule le pourcentage des valeurs manquantes (NaN) pour chaque colonne concernée. Les résultats sont affichés pour fournir un aperçu de la qualité des données après fusion.

La fonction 'check_nunique' fournit le nombre de valeurs uniques dans chaque colonne et affiche ces informations pour aider à comprendre la structure et la complexité du dataframe final fusionné.

Observations :

Ces contrôles successifs ont permis de relever les valeurs sur lesquelles il sera utile de revenir :

- au niveau des valeurs nulles pour les colonnes 'an_naiss' et 'id_usager' et,
- au niveau des valeurs uniques qui restent trop élevées pour certaines variables.

Les éléments concernés ont été traduits dans la fonction 'preprocessing_final_dataframe' qui suit.

2.2.2 Crédation de fonctions spécifiques

La fonction 'preprocessing_final_dataframe(df)' permet de réaliser un prétraitement complémentaire des données sur certains aspects spécifiques.

La fonction 'encode_dataframe(df)' transforme les données catégorielles du dataframe final en variables numériques au moyen d'un codage one-hot.

2.2.3 Lancement du processus

Un code applique la fonction 'preprocessing_final_dataframe' (définie en amont) au dataframe pour réaliser un nettoyage résiduel de points sépcifiques, avant de vérifier notamment le résultat obtenu sur l'amplitude horaire des données.

Un autre code finalise la préparation des données, assurant une conversion des colonnes de type 'object' en type 'int64', pour que toutes les variables du dataframe aient un format approprié pour la modélisation. La sauvegarde du fichier CSV permet de conserver une version propre et structurée des données pour une utilisation ultérieure.

3 Deuxième prétraitement des données

Le premier prétraitement des données nous a permis d'effectuer une restructuration initiale des dataframes, de créer des variables de base et d'éliminer certaines variables non pertinentes. Cependant, la modélisation qui en a résulté (cf. infra : modélisation n°1 - régression logistique) nous a révélé des insuffisances au niveau des métriques obtenues. C'est pourquoi nous en avons conclu que certaines variables nécessitaient une catégorisation plus fine, pour que les relations complexes entre elles soient mieux exploitées et que le déséquilibre des classes puisse être moins persistant.

Dans cet objectif, nous appréhendons les choses sous un tout nouvel angle qui prend nécessairement en considération les résultats observés jusqu'ici pour parvenir à leur amélioration. Egalement, nous nous appuierons sur les observations suivantes pour définir la structuration ultérieure de nos dataframes.

3.1 Base de structuration des dataframes :

3.1.1 Structure pour 'caracteristiques' :

Colonne	Problématique	Modification/Suppression
'Num_Acc' : dtype('int64')	Nom de variable qui change en 2022 + Trop de valeurs uniques	Supprimer colonne après avoir remplacé 'Accident_Id' par 'Num_Acc'
'jour' : dtype('int64')	Faible variation des valeurs	Colonne à supprimer
'mois' : dtype('int64')	-	-
'an' : dtype('int64')	-	Supprimer colonne après avoir calculé l'âge des usagers ('an' - 'an_nais')
'hrmn' : dtype('O')	2 formats horaires existant : HHMM et HH :MM	Convertir HHMM en HH :MM
'lum' : dtype('int64')	Pas de définition pour -1	Dichotomiser toutes les lignes, en excluant celles contenant -1
'dep' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'com' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'agg' : dtype('int64')	-	-
'int' : dtype('int64')	-	-
'atm' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1 + Regrouper des valeurs de météo extrême
'col' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1
'adr' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'lat' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'long' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer

3.1.2 Structure pour 'lieux' :

Colonne	Problématique	Modification/Suppression
'Num_Acc' : dtype('int64')	Trop de valeurs uniques	Colonne à supprimer
'catr' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant 9
'voie' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'v1' : dtype('int64')	Informations non pertinentes + Trop de valeurs nulles	Colonne à supprimer
'v2' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'circ' : dtype('int64')	Pas de définition pour 0 + Valeurs manquantes existantes	Dichotomiser toutes les lignes, en excluant celles contenant -1
'nbv' : dtype('O')	-	Dichotomiser avec regroupement des valeurs supérieures à 5
'vosp' : dtype('int64')	Valeurs manquantes existantes	Dichotomiser toutes les lignes, en excluant celles contenant -1 et NaN
'prof' : dtype('int64')	Valeurs manquantes existantes	Dichotomiser toutes les lignes, en excluant celles contenant -1 et NaN
'pr' : dtype('O')	Trop de valeurs uniques + Valeurs manquantes existantes	Colonne à supprimer
'pr1' : dtype('O')	Trop de valeurs uniques + Valeurs manquantes existantes	Colonne à supprimer
'plan' : dtype('int64')	Valeurs manquantes existantes	Dichotomiser toutes les lignes, en excluant celles contenant -1 et NaN
'lartpc' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'larrout' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'surf' : dtype('int64')	Valeurs manquantes existantes	Dichotomiser toutes les lignes en excluant celles content -1 et NaN
'infra' : dtype('int64')	-	Dichotomiser avec regroupement des valeurs supérieures à 2 et exclusion des lignes content -1 et 0
'situ' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1
'vma' : dtype('int64')	Valeurs parfois aberrantes de : -1 à 901 kmh	Dichotomiser toutes les lignes avec une vitesse comprise entre 0 et 130 km/h

3.1.3 Structure pour 'usagers' :

Colonne	Problématique	Modification/Suppression
'Num_Acc' : dtype('int64')	Trop de valeurs uniques	Colonne à supprimer
'id_usager' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'id_véhicule' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'num_veh' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'place' : dtype('int64')	Valeurs manquantes existantes	Dichotomiser toutes les lignes, e, excluant celles contenant -1
'catu' : dtype('int64')	Valeurs manquantes existantes	Dichotomiser toutes les lignes, en excluant celles contenant -1
'grav' : dtype('int64')	Pas de définition pour -1	Dichotomiser toutes les lignes, en excluant celles contenant -1
'sexe' : dtype('int64')	-	Dichotomiser toutes les lignes
'an_nais' : dtype('float64')	-	Supprimer colonne après avoir calculé l'âge des usagers ('an' - 'an_nais')
'trajet' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1
'secu1' : dtype('int64')	-	-
'secu2' : dtype('int64')	-	-
'secu3' : dtype('int64')	-	-
'locp' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1
'actp' : dtype('O')	Valeurs manquantes existantes	Dichotomiser toutes les lignes, en excluant celles contenant -1, 0 et B
'etatp' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1

3.1.4 Structure pour 'véhicules' :

Colonne	Problématique	Modification/Suppression
'Num_Acc' : dtype('int64')	Trop de valeurs uniques	Colonne à supprimer
'id_véhicule' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'num_veh' : dtype('O')	Trop de valeurs uniques	Colonne à supprimer
'senc' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1
'catv' : dtype('int64')	Pas de définition de -1	Dichotomiser toutes les lignes, en excluant celles contenant -1 et 0
'obs' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1 et 0
'obsm' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1 et 0
'choc' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1 et 0
'manv' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1 et 0
'motor' : dtype('int64')	-	Dichotomiser toutes les lignes, en excluant celles contenant -1 et 0
'occute' : dtype('float64')	Trop de valeurs uniques	Colonne à supprimer

3.2 Etapes de transformation des dataframes

3.2.1 Initialisation de l'environnement

Dans cette section, nous mettons en place un environnement de travail pour le projet. Il a notamment pour but de structurer notre flux de travail, documenter nos actions effectuées, et garantir la reproductibilité des résultats.

Voici plus précisément la description des étapes concrétisées dans notre code :

- Il permet d'importer les bibliothèques nécessaires, définir les chemins de répertoires où seront stockés les différents types de fichiers, mettre en place un système de journalisation des actions effectuées au cours du traitement. Ce système utilise la fonction 'log_action' qui affiche les actions et alimente en même temps la liste 'actions' permettant de stocker les opérations réalisées. La liste d'actions est affichée en fin de notebook pour servir à la rédaction du rapport.
- Une liste 'var_ecartees' est simplement initialisée pour servir à l'enregistrement des variables écartées. Celle-ci sera complétée ultérieurement pour permettre la prise en compte des noms et motifs d'écartement de ces variables.
- Dans ce même code, deux fichiers de configuration au format JSON sont chargés :
 - le premier (desc_fic_raw.json) contient des informations sur les fichiers de données brutes, comme leurs noms, les séparateurs utilisés, et les périodes concernées. La description des fichiers a été motivée par les disparités des fichiers et elle a permis de charger facilement toutes les données.
 - le second (desc_vars.json) décrit les variables du jeu de données, incluant leurs libellés et les correspondances pour les modalités, ce qui permettra des affichages avec les libellés correspondants aux noms de variables et codes de modalités bien plus explicites.
- Un dictionnaire 'dfrub' est initialisé en vue d'y stocker des dataframes dans les prochaines étapes du traitement.

A la base, les données sont réparties en 4 rubriques qui ont été volontairement réduites sur 4 années successives (2019-2022).

Pour mettre en oeuvre leur prétraitement, nous utiliserons le fichier 'desc_fic_raw.json' qui contient les informations couvrant divers aspects des accidents de la route, des véhicules impliqués et des personnes concernées. Ci-après quelques éléments correspondants :

- Le nom du fichier ;
- Le séparateur ;
- La phase : jusqu'à 2018 ou à partir de 2019 ;
- Les conversions de types dans dtypes à réaliser lors du chargement.

Un code effectue ensuite le traitement et la consolidation de données à partir de plusieurs fichiers CSV, organisés par rubriques. Voici comment il fonctionne :

- Il commence à parcourir chaque rubrique définie dans un dictionnaire appelé 'des_fic_raw'.
- Pour chaque rubrique, le code initialise un compteur pour suivre le nombre total d'observations. Il prépare également une liste vide pour stocker les DataFrames pandas qui seront créés pour chaque fichier annuel.
- Ensuite, le script traite chaque fichier annuel associé à la rubrique en cours. Il ne traite que les fichiers marqués comme étant en "phase 2". Pour chaque fichier éligible, il effectue les opérations suivantes :
 - Lecture du fichier CSV en utilisant pandas, en spécifiant les paramètres appropriés tels que le séparateur, l'encodage et les types de données.
 - Comptage du nombre d'observations dans le fichier.
 - Si nécessaire, renommage des colonnes du DataFrame selon les spécifications fournies.
 - Ajout du DataFrame à la liste préparée précédemment.
- Une fois tous les fichiers d'une rubrique traités, le code concatène tous les DataFrames de la liste en un seul grand DataFrame. Ce DataFrame consolidé subit ensuite un nettoyage : les valeurs "-1" (avec un espace) sont remplacées par "-1" pour corriger un problème de codage des valeurs manquantes.

- Le script affiche ensuite des statistiques sur le traitement effectué, notamment le nombre de DataFrames traités, le nombre total d'observations et le nombre de colonnes dans le DataFrame final.
- Enfin, le DataFrame consolidé et nettoyé est sauvegardé dans un nouveau fichier CSV, avec le nom de la rubrique.
- Le processus se répète pour chaque rubrique, créant ainsi un ensemble de fichiers CSV consolidés, chacun correspondant à une rubrique spécifique.
- Pour terminer, le code crée des DataFrames distincts pour chaque rubrique principale : usagers, véhicules, lieux et caractéristiques.

Le script correspondant a donc permis d'automatiser le processus de consolidation et de nettoyage initial des données provenant des multiples fichiers annuels, en les regroupant par catégorie et en effectuant quelques corrections basiques.

3.2.2 Enregistrement des 4 rubriques de 2019 à 2022

Après avoir nettoyé et uniformisé les données brutes, nous enregistrons les jeux de données de chaque rubrique — caractéristiques, lieux, usagers et véhicules — couvrant la période de 2019 à 2022. Ces enregistrements servent de fichiers intermédiaires qui pourront être utilisés pour des explorations et des analyses plus approfondies, tout en permettant de valider les étapes de prétraitement effectuées.

En sauvegardant les données à ce stade, nous nous assurons que les ajustements et transformations appliqués sont bien capturés avant de procéder à des étapes plus avancées, telles que la jointure des jeux de données. Cela facilite également la détection et la correction d'éventuelles erreurs ou incohérences rencontrées lors du prétraitement, en permettant un retour en arrière sans devoir répéter les étapes initiales de nettoyage.

3.2.3 Jointure des 4 rubriques

Nous cherchons à prédire la gravité des accidents pour les personnes en fonction des circonstances des accidents, la gravité est dans la variable `grav` de la rubrique usagers, c'est notre cible, la rubrique usagers contient déjà quelques informations, nous relions alors les informations des autres rubriques à la rubrique usagers. Cette étape implique la jointure des données sur la base de l'identifiant unique d'accident (`Num_Acc`). Les jointures ne doivent pas perdre des données ou introduire des erreurs dues à des correspondances incorrectes. Nous les vérifions à chaque fois en affichant les nombres de colonnes (variables) et de lignes (observations).

Notre code réalise ainsi plusieurs opérations menant à la jointure des différents DataFrames créés. Quelques remarques préalables à ce sujet :

- Les jointures sont toutes faites avec le champ `Num_Acc` ;
- Le type de `Num_Acc` est forcé à "int" lors de la lecture par `read_csv()` ;
- La dernière jointure sur les véhicules est faite avec, en plus, les champs `num_veh` et `id_vehicule` ;
- Les jointures sont "à gauche" ("left") pour conserver le nombre d'usagers ;
- Les nombres d'observations et de variables affichés avant et après chaque jointure permettent de vérifier les jointures ;
- Il y a 494 182 usagers avant les jointures et le DataFrame final 'df' résultant a ce même nombre d'observations.

Après les jointures, le code utilise la méthode `info()` pour afficher des informations détaillées sur le DataFrame, y compris les types de données et le nombre de valeurs non nulles pour chaque colonne. Enfin, il libère de la mémoire en mettant à `None` les DataFrames qui ne seront plus utilisés (`dfc`, `dfl`, `dfu`, `dfv`, `dfrub`).

Tout au long du processus, le code utilise la fonction '`log_action()`' pour enregistrer les étapes principales du processus de jointure.

3.2.4 Suppression d'observations

Lors de la préparation des données pour l'analyse et la modélisation, il est essentiel de s'assurer que les observations incluses soient complètes et cohérentes. La variable '`grav`' représente la gravité de l'accident, qui est une information clé pour toute analyse visant à comprendre ou prédire les facteurs influençant la gravité des accidents de la route.

Cependant, certaines observations contiennent une valeur de 'grav' égale à -1, indiquant que la gravité est inconnue ou non renseignée. Ces observations ne donnent pas d'information et risquent de biaiser nos modèles.

Pour cette raison, nous décidons de supprimer toutes les observations pour lesquelles la gravité est inconnue ('grav' = -1). Cette étape garantit que le jeu de données final est composé uniquement de cas où la gravité de l'accident est clairement définie, ce qui améliore la qualité des données pour les analyses futures.

3.2.5 Crédation de variables

La création de nouvelles variables est une étape clé du prétraitement des données, car elle permet de capturer des informations supplémentaires qui ne sont pas directement présentes dans les variables brutes. En générant de nouvelles variables à partir des données existantes, nous pouvons identifier des motifs et des relations qui pourraient être déterminants pour la prédiction de la gravité des accidents de la route. Intuitivement, nous pensons que le jour de la semaine, les jours fériés et l'âge des usagers ont une influence sur les conséquences des accidents.

Dans cette section, nous créons alors trois nouvelles variables qui pourraient apporter des explications supplémentaires de la gravité des accidents :

- **Jour de la semaine (jsem)** : Cette variable est créée pour identifier le jour de la semaine (lundi à dimanche) où l'accident s'est produit. Cette information peut être utile pour comprendre les tendances hebdomadaires, telles que l'augmentation des accidents le week-end ou les jours de semaine chargés.
- **Jour férié** : Les jours fériés peuvent avoir un impact significatif sur le trafic et le comportement des conducteurs, ce qui pourrait influencer la gravité des accidents. Cette variable binaire indique si l'accident s'est produit un jour férié ou non (incluant les dimanches et les jours de fête).
- **Âge des usagers (age)** : Calculée comme la différence entre l'année de l'accident et l'année de naissance de l'usager, cette variable permet d'analyser l'impact de l'âge sur la gravité des accidents. L'âge des conducteurs, par exemple, peut être un facteur très important, les jeunes conducteurs et les conducteurs plus âgés pouvant avoir des comportements et des risques différents. Lorsque l'année de naissance est inconnue, l'âge est codé -1. Cette variable sera dichotomisée ultérieurement.

Ces variables créées visent à enrichir le jeu de données et à fournir des attributs supplémentaires qui peuvent améliorer la performance des modèles prédictifs en capturant des dimensions supplémentaires de la dynamique des accidents de la route.

3.2.6 Dichotomisation / catégorisation

Dans le cadre de la préparation des données pour la modélisation prédictive, certaines variables sont transformées par dichotomisation (transformation en variables binaires) ou par catégorisation pour mieux capturer des informations pertinentes et améliorer la performance des modèles d'apprentissage.

Dichotomisation des variables :

La dichotomisation consiste à convertir une variable en une variable binaire (0 ou 1). Cette méthode est utilisée lorsque nous souhaitons simplifier une variable en la réduisant à deux catégories distinctes. Par exemple, la variable 'lum' (luminosité) pourrait être transformée en une variable binaire pour distinguer les conditions de jour (1) des conditions de nuit (0). Les valeurs non définies (comme -1) sont généralement exclues pour éviter les biais.

Catégorisation des variables :

La catégorisation consiste à diviser une variable en plusieurs catégories distinctes. Cette méthode est utilisée pour regrouper des valeurs en classes significatives, gérer les valeurs aberrantes ou simplifier des variables avec de nombreuses modalités. Par exemple, pour la variable 'vma' (vitesse maximale autorisée), qui peut contenir des valeurs allant de -1 à 901 km/h, il est possible de créer plusieurs catégories comme "0-50 km/h", "51-90 km/h", et "91-130 km/h". Cette transformation permet de structurer les données de manière plus interprétable et d'améliorer la robustesse des modèles.

Regroupement de modalités et gestion des valeurs manquantes :

Les variables telles que circ (circulation) ou prof (profil de la route) contiennent des valeurs non définies ou manquantes. Celles-ci sont regroupées ou exclues pour s'assurer que seules des données valides et pertinentes sont utilisées dans les analyses et la modélisation. Par exemple, la variable circ est

dichotomisée en excluant les valeurs -1 et autres modalités non pertinentes, ce qui permet de rendre les données plus cohérentes.

En appliquant ces techniques, nous simplifions la structure des données, ce qui peut améliorer les performances des modèles prédictifs et faciliter l'interprétation des résultats.

Voici un résumé des principales transformations effectuées :

1. Variables temporelles :

- 'jsem' (jour de la semaine) : dichotomisée en 7 variables binaires
- 'hrmn' (heure) : catégorisée en 5 périodes (matin, midi, après-midi, soir, nuit)
- 'mois' : transformée en 12 variables binaires

2. Variables démographiques :

- 'age' : catégorisée en 4 groupes (enfant, jeune, adulte, 3ème âge)
- 'sexe' : transformée en variables binaires (homme/femme)

3. Variables liées à la sécurité :

- 'secu1', 'secu2', 'secu3' : combinées en 6 variables binaires représentant différents équipements de sécurité

4. Variables liées à l'environnement :

- 'surf' (état de la surface) : catégorisée en 4 types (normale, mouillée, glissante, autre)
- 'atm' (conditions atmosphériques) : dichotomisée en excluant les valeurs non pertinentes
- 'lum' (luminosité) : dichotomisée en excluant les valeurs non définies

5. Variables liées au véhicule et à la route :

- 'catv' (catégorie de véhicule) : transformée en plusieurs variables binaires en excluant les valeurs non pertinentes
- 'nbv' (nombre de voies) : regroupée en 5 catégories (1 à 4 voies, et 5+)
- 'vma' (vitesse maximale autorisée) : catégorisée en groupes de vitesse pertinents (30km/h et moins, 40-50km/h, 60-70km/h, 80-90km/h, 100km/h et plus)

6. Variable cible :

- 'grav' (gravité) : dichotomisée en 'grave' et 'non grave'

Autres variables dichotomisées : 'choc', 'manv', 'obs', 'obsm', 'catu', 'trajet', 'motor', entre autres.

Note : Pour toutes ces transformations, nous avons veillé à exclure les valeurs non pertinentes ou inconnues (souvent codées comme -1) afin de ne pas biaiser notre analyse.

3.2.7 Suppression de variables

Certaines variables peuvent contenir des informations redondantes, des valeurs manquantes importantes, des valeurs uniques ou des données non pertinentes qui peuvent nuire à la performance et à la robustesse des modèles prédictifs. En éliminant ces variables, nous réduisons la dimensionnalité des données, ce qui simplifie le modèle, améliore sa capacité à généraliser et réduit les risques de surajustement.

Les raisons principales de la suppression de variables incluent :

Faible Variabilité ou Valeurs Uniques Trop Nombreuses :

Certaines variables, comme les identifiants (Num_Acc, id_vehicule), contiennent des valeurs uniques pour chaque observation, ce qui n'apporte aucune information utile pour l'analyse prédictive. Ces variables sont souvent éliminées car elles ne contribuent pas à la compréhension des relations entre les données.

Valeurs Manquantes ou Erronées :

Certaines variables contiennent un nombre excessif de valeurs manquantes ou de valeurs aberrantes, rendant leur utilisation impraticable. Par exemple, une variable qui a plus de 50% de valeurs manquantes peut être supprimée car sa reconstruction ou son imputation pourrait introduire des biais.

Redondance avec d'autres Variables :

Les variables qui sont fortement corrélées avec d'autres variables peuvent être redondantes (exemple, les variables dichotomisées). Dans ce cas, une seule variable représentative est conservée pour éviter des calculs inutiles et minimiser la multicolinéarité.

Pertinence pour l'Analyse :

Certaines variables peuvent être considérées comme non pertinentes pour les objectifs de l'analyse. Par exemple, des variables géographiques très détaillées comme adr (adresse), pr (point de repère), ou voie (nom de la rue) peuvent être supprimées si elles n'apportent pas de valeur ajoutée significative à l'analyse des facteurs influençant la gravité des accidents.

En appliquant cette stratégie de suppression de variables, nous nous assurons que seules les données les plus pertinentes et les plus fiables sont utilisées pour la modélisation, augmentant ainsi la qualité et l'efficacité des résultats finaux.

Voici un résumé des principales raisons pour lesquelles certaines variables ont été supprimées :

1. Dispersion trop importante ou valeurs douteuses :

- Variables géographiques détaillées : 'adr', 'com', 'dep', 'voie', 'v1', 'v2'
- Variables de localisation précise : 'lat', 'long', 'pr', 'pr1'
- Mesures de route imprécises : 'larrout', 'lartpc'

2. Redondance après transformation :

- Variables temporelles : 'an', 'jour' (utilisées pour calculer d'autres variables)
- Variables démographiques : 'an_nais' (utilisée pour calculer l'âge)

3. Variables d'index ou d'identification :

- 'Num_Acc', 'id_usager', 'id_vehicule', 'num_veh'

4. Variables avec trop de valeurs nulles :

- 'occutc' (taux d'occupation du véhicule)

5. Variables déjà dichotomisées :

- toutes les variables qui ont été dichotomisées dans la section précédente.

3.2.8 Suppression de doublons

La suppression de doublons est une étape essentielle pour s'assurer que chaque observation dans le jeu de données est unique et représentative. Les doublons peuvent se produire en raison de diverses erreurs de collecte ou d'intégration de données, telles que des entrées multiples pour le même accident ou des erreurs lors de la fusion des bases de données. Conserver ces doublons dans le jeu de données pourrait fausser les résultats analytiques et les prédictions des modèles en donnant un poids excessif à certaines observations.

La suppression des doublons est effectuée en identifiant les enregistrements qui ont les mêmes valeurs dans toutes les colonnes pertinentes. Après cette étape, seules les observations uniques et significatives sont conservées pour garantir la qualité et l'intégrité des données.

3.2.9 Équilibrage de la dimension

L'équilibrage de la dimension est essentiel lorsque la variable cible est déséquilibrée, c'est-à-dire qu'une modalité est significativement plus représentée qu'une autre. Notre objectif étant de prédire les conditions des accidents ayant entraîné une hospitalisation de plus de 24h ou la mort, le déséquilibre entre les deux classes de la variable 'grav_grave' indiquant si un accident est "grave" (1) ou "non grave" (0) doit être corrigé pour éviter que le modèle d'apprentissage soit biaisé en faveur de la classe majoritaire (les accidents "non graves"), un équilibrage de ces modalités est nécessaire.

Nous avons sous-échantilloné la classe dominante pour rééquilibrer les données avec RandomUnderSampler en lui précisant random_state=8421 pour assurer la reproductibilité du processus. Le sous-échantillonnage consiste à réduire le nombre d'exemples de la classe majoritaire (ici les conséquences "non graves") afin de correspondre au nombre d'exemples de la classe minoritaire (conséquences "graves"). Cette technique permet de créer un jeu de données équilibré où les deux classes sont représentées par le même nombre d'observations.

Résultats de l'équilibrage :

Les nombres de modalités de notre variable cible et les nombres d'observations avant et après réduction par échantillonage sont reportées dans le tableau suivant :

	"Graves"	"Non graves"	Total
Avant	88 821	401 827	490 648
Après	88 821	88 821	177 642

Ces valeurs issues d'un affichage dans un notebook nous assurent que nos modalités sont également réparties.

Une fois cet ensemble d'étapes effectué, nous réalisons une analyse préliminaire pour explorer le contenu de chaque colonne, notamment pour des variables binaires ou catégorielles converties en indicateurs numériques.

Synthèse des variables clés et des statistiques principales obtenues :

Variable	Description	Pourcentage
grav_grave	Accident grave	50.000%
sexe_m	Conducteur masculin	69.271%
sexe_f	Conducteur féminin	29.985%
age_adulte	Conducteur adulte	55.641%
age_jeune	Jeune conducteur	24.473%
secu_ceinture	Port de la ceinture	49.354%
secu_casque	Port du casque	24.793%
surf_norm	Surface normale	81.101%
surf_mouil	Surface mouillée	17.164%
vma_50	Vitesse max autorisée 50 km/h	46.801%
vma_80	Vitesse max autorisée 80 km/h	20.589%
atm_1	Condition atmosphérique normale	79.979%
lum_1	Plein jour	65.967%
int_1	Hors intersection	67.038%
catu_1	Autoroute	72.507%
obsm_2	Absence d'obstacle mobile	58.188%
senc_1	Sens de circulation unique	44.430%
trajet_5	Trajet loisirs	41.870%

Interprétation :

— Gravité des accidents :

La variable cible 'grav_grave' est parfaitement équilibrée à 50% grâce à l'opération d'équilibrage.

— Caractéristiques des conducteurs :

Les hommes sont surreprésentés (69.27%) par rapport aux femmes (29.99%). Les adultes constituent la majorité des conducteurs (55.64%), suivis par les jeunes (24.47%). Le port de la ceinture est observé dans environ la moitié des cas (49.35%). Le port du casque concerne environ un quart des cas (24.79%), probablement lié aux accidents de deux-roues.

— Conditions de l'accident :

La majorité des accidents se produit sur une surface normale (81.10%), avec une part non négligeable sur surface mouillée (17.16%). Les conditions atmosphériques sont généralement normales (79.98%). La plupart des accidents ont lieu en plein jour (65.97%).

— Localisation des accidents :

Les accidents hors intersection sont majoritaires (67.04%). Les autoroutes représentent une part importante des lieux d'accidents (72.51%).

— Limites de vitesse :

Les zones à 50 km/h (46.80%) et 80 km/h (20.59%) sont les plus représentées, suggérant une prédominance des accidents en ville et sur routes secondaires.

— **Autres observations notables :**

L'absence d'obstacle mobile est notée dans 58.19% des cas. Les trajets de loisirs sont les plus représentés (41.87%).

3.2.10 Synthèse des actions de préprocessing

La liste des 'actions' réalisées au cours du traitement des données permet de garder une trace claire et structurée de toutes les manipulations et transformations effectuées sur le jeu de données. Dans cette section, nous effectuons un récapitulatif des étapes importantes du pipeline de prétraitement, qui peut inclure le nettoyage des données, la création de variables, l'équilibrage des classes, la suppression des doublons, et bien d'autres actions.

Nous obtenons le résultat suivant :

Actions réalisées :

- Jointure usagers <— caractéristiques
- Jointure (usagers et caractéristiques) <— lieux
- Jointure (usagers, caractéristiques et lieux) <— véhicules
- Suppression de 0 observations dont la gravité est inconnue (codée -1)
- Création de la variable jsem : jour de la semaine
- Création de la variable ferie : jour férié, dimanches et autres fêtes
- Création de la variable age : différence entre l'année de l'accident et l'année de naissance
- Dichotomisation des champs secu1, secu2 et secu3
- Dichotomisation de l'âge
- Dichotomisation de l'heure
- Dichotomisation du sexe
- Dichotomisation de la gravité
- Dichotomisation du nombre de voies avec regroupement 1 à 4 puis 5 et plus
- Dichotomisation du l'état de la surface : sèche, mouillée, glissante (3 à 9)
- Dichotomisation de la vitesse maximale autorisée avec regroupement
- Dichotomisation en ou hors agglomération (agg), 1 agglomération, 0 hors agglomoration
- Dichotomisation de l'action du piéton (actp), modalité -1 0 et B exclues
- Dichotomisation des cond. atmosphériques (atm), modalité -1 et 9 exclues
- Dichotomisation de la catégorie de route (catr)), modalité -1 et 9 exclues
- Dichotomisation de la catégorie d'usager (catu), modalité -1 exclue
- Dichotomisation de la catégorie de véhicule (catv), modalité -1 et 0 exclues
- Dichotomisation du point de choc initial (choc), modalité -1 et 0 exclues
- Dichotomisation du régime de circulation (circ), modalité -1 exclue
- Dichotomisation du type de collision (col), modalité -1 exclue
- Dichotomisation de (etatp), modalité -1 exclue
- Dichotomisation de Aménagement - infrastructure (infra), modalités -1 et 0 exclues
- Dichotomisation du type d'intersection (int), modalité -1 exclues
- Dichotomisation du jour de la semaine (jsem), toutes modalité
- Dichotomisation de la localisation du piéton (locp), modalités -1, 0(non renseigné) exclues
- Dichotomisation des conditions lumineuses (lum), modalité -1 exclue
- Dichotomisation de la manœuvre (manv), modalité -1 et 0 exclues
- Dichotomisation du mois (mois), toutes modalités
- Dichotomisation de la motorisation (motor), modalité -1 et 0 exclues
- Dichotomisation de l'obstacle fixe heurté (obs), modalité -1 et 0 exclues
- Dichotomisation de l'obstacle mobile heurté (obsm), modalité -1 et 0 exclues

- Dichotomisation de la place dans le véhicule (place), modalité -1 exclue
- Dichotomisation du tracé en plan (plan), modalité -1 exclue
- Dichotomisation de la déclivité (prof), modalité -1 exclue
- Dichotomisation du sens de circulation (senc), modalité -1, 0 et 3 exclues
- Dichotomisation de la situation de l'accident (situ), modalité -1 exclue
- Dichotomisation du motif du trajet (trajet), modalité -1 exclue
- Dichotomisation de la présence d'une voie réservée (vosp), modalités -1 exclue
- Équilibrage 2 parts égales dans df2

3.2.11 Sauvegarde des données

Une fois que toutes les étapes de nettoyage, transformation, équilibrage, et ingénierie des variables ont été réalisées, il est nécessaire de sauvegarder le jeu de données final ainsi que la documentation associée. Ces enregistrements permettent de garantir que les données et les métadonnées soient prêtes et disponibles pour les analyses ou modélisations futures.

PARTIE II

Dans le cadre de notre projet d'étude, nous cherchons à prédire si un accident sera "grave" (entraînant une hospitalisation de plus de 24 heures ou un décès) ou "non grave" (blessures légères ne nécessitant pas d'hospitalisation prolongée), en nous basant sur les circonstances et caractéristiques de l'accident.

Cette prédiction binaire constitue un problème de classification où :

- Classe positive (1) = accident "grave",
- Classe négative (0) = accident "non grave".

Notre approche consiste alors à tester différents types de modèles :

- 1. La régression logistique qui servira de base de référence.
- 2. Les multiples modèles classiques (SVM, Random Forest, Gradient Boosting, etc.) avec optimisation des hyperparamètres.
- 3. Le Deep learning avec un réseau de neurones.

Pour chaque modèle, nous examinerons plusieurs métriques de classification :

- Accuracy (Taux de Précision globale) : proportion totale de prédictions correctes (graves et non graves).
- Precision (Taux d'Exactitude) : proportion des accidents qui sont réellement graves.
- Recall (Taux de Rappel) : proportion d'accidents graves correctement identifiés parmi tous les accidents réellement graves.
- F1-Score : moyenne harmonique entre la 'precision' et le 'recall'.
- AUC-ROC : capacité du modèle à distinguer les différentes classes entre elles.

Etant donné que nous souhaitons accorder une importance particulière au recall, il convient de clarifier préalablement les différents types de prédictions possibles dans notre contexte :

- Vrai Positif (VP) : Le modèle prédit correctement un accident "grave".
- Vrai Négatif (VN) : Le modèle prédit correctement un accident "non grave".
- Faux Positif (FP) : Le modèle prédit un accident "grave", alors qu'il ne l'est pas.
- Faux Négatif (FN) : Le modèle ne prédit pas un accident "grave", alors qu'il l'est.

A partir des différents résultats attendus, nous espérons identifier le meilleur modèle qui nous permettra de prédire les accidents réellement graves, c'est-à-dire ayant le meilleur taux de vrais positifs (cf. recall ou taux de rappel pour la classe 1). Ce choix est motivé par le fait que la non-détection d'un accident grave (faux négatif) est plus problématique qu'une fausse alerte (faux positif) dans le contexte de la sécurité routière.

1 Modélisation 1

1.1 Implémentation d'un premier modèle

Sur la base du fichier obtenu lors du premier préprocessing, il est procédé à une 1ère modélisation de type classique, à savoir une régression logistique. Le choix de ce modèle se justifie entre autres par les raisons suivantes :

- il est adapté à la nature binaire de la variable cible (accident grave ou non grave),
- il est capable de gérer les nombreuses variables catégorielles présentes dans le jeu de données,
- il offre une bonne base pour servir de référence à des modèles plus complexes, etc.

Le code créé met en place un pipeline complet pour l'entraînement et l'optimisation d'un modèle de régression logistique.

• Importation des librairies

Le code commence par importer les bibliothèques essentielles pour la manipulation des données (Numpy et Pandas), la visualisation (Matplotlib), le prétraitement (StandardScaler), la division des données (train_test_split), la modélisation (LogisticRegression), l'optimisation des hyperparamètres (GridSearchCV), et l'évaluation des performances (diverses métriques et validation croisée).

• Chargement et prétraitement des données

Par la suite, le code charge et prépare les données pour une analyse de machine learning. Il commence par importer un fichier CSV prétraité ('data.csv'), utilisant des tabulations comme séparateurs. Ensuite, il convertit toutes les colonnes en nombres entiers pour assurer une cohérence des types de données. Finalement, il sépare les données en deux parties : X, qui contient toutes les variables explicatives, et y, qui isole la variable cible 'grav_grave'. Cette dernière sera l'objet de la prédiction dans le modèle à venir pour indiquer la gravité des accidents.

• Standardisation des variables numériques

Egalement, le code effectue la standardisation des variables numériques du jeu de données, en suivant cette démarche :

- Une liste 'dummy_columns' est définie, contenant les noms des colonnes catégorielles qui ne doivent pas être standardisées.
- Une nouvelle liste 'columns_to_scale' est créée pour y inclure toutes les colonnes de X qui ne sont pas dans 'dummy_columns', identifiant ainsi les variables numériques à standardiser.
- Un objet StandardScaler est instancié pour normaliser les variables numériques en les centrant et réduisant (moyenne de 0 et écart-type de 1).
- Une copie de X est créée et nommée 'X_scaled' pour préserver les données originales.
- La méthode fit_transform() du StandardScaler est appliquée uniquement aux colonnes identifiées dans 'columns_to_scale'. Cette opération ajuste le scaler aux données puis transforme ces colonnes.

Cette étape de prétraitement permet de mettre toutes les variables numériques à la même échelle, pour améliorer les performances de nos futurs modèles.

• Création d'un ensemble d'entraînement et d'un ensemble test

Au moyen de la fonction 'train_test_split' de scikit-learn, les données sont divisées en ensembles d'entraînement (80%) et de test (20%). Une graine aléatoire 'random_state' est fixée à 12 pour permettre d'assurer la reproductibilité de la division.

• Configuration du modèle et préparation des hyperparamètres

Le code commence par initialiser un classificateur de régression logistique avec des paramètres de base et configure une stratégie de validation croisée à 3 plis. Une grille de recherche d'hyperparamètres est définie pour explorer méticuleusement différentes combinaisons de force de régularisation, types de pénalité et, dans le cas de la pénalité elasticnet, divers ratios entre les normes L1 et L2. Cette approche exhaustive vise à identifier la configuration optimale du modèle.

• Création d'un GridSearchCV personnalisé pour sauvegarder les résultats partiels

Le code définit ensuite une classe personnalisée 'GridSearchWithProgress' qui hérite de GridSearchCV. Cette classe modifie la méthode fit() pour avoir plus de contrôle sur le processus de recherche par grille. Voici ses principales caractéristiques :

- Elle calcule le nombre total de combinaisons de paramètres à tester.
- Pour chaque combinaison de paramètres, il enregistre le temps de début, configure l'estimateur avec ces paramètres, effectue une validation croisée et enregistre le temps de fin.
- Pour chaque itération, il stocke les résultats (paramètres et score moyen de test), incrémente le compteur de combinaisons complétées, sauvegarde les résultats partiels dans un fichier joblib et affiche les résultats partiels, y compris les paramètres, le score moyen et l'écart-type, et le temps d'exécution.
- Après avoir testé toutes les combinaisons, il identifie les meilleurs paramètres et le meilleur score, configure le meilleur estimateur avec les meilleurs paramètres et entraîne le meilleur estimateur sur l'ensemble des données.

La recherche d'hyperparamètres est lancée avec cette classe personnalisée.

Voici les 3 premiers résultats obtenus :

Résultat partiel 1/24 :

Paramètres : 'C' : 0.1, 'max_iter' : 1000, 'penalty' : 'l1'

Score moyen : 0.5894 (+/- 0.0008)

Temps de fit : 10495.11 secondes

Résultat partiel 2/24 :

Paramètres : 'C' : 1, 'max_iter' : 1000, 'penalty' : 'l1'

Score moyen : 0.5895 (+/- 0.0008)

Temps de fit : 13500.07 secondes

Résultat partiel 3/24 :

Paramètres : 'C' : 10, 'max_iter' : 1000, 'penalty' : 'l1'

Score moyen : 0.5895 (+/- 0.0008)

Temps de fit : 14126.99 secondes

1.2 Évaluation des performances du modèle

A la suite des résultats obtenus, le code créé effectue l'évaluation finale et l'utilisation du meilleur modèle identifié par la recherche d'hyperparamètres :

- Il affiche les meilleurs hyperparamètres trouvés lors de la recherche par grille.
- Le meilleur score obtenu durant cette recherche est également affiché.
- Le meilleur modèle (celui avec les hyperparamètres optimaux) est extrait de l'objet GridSearchCV.
- Ce modèle optimal est ensuite utilisé pour faire des prédictions sur l'ensemble de test.
- Enfin, la précision (accuracy) du meilleur modèle sur l'ensemble de test est calculée et affichée.

Cette séquence permet de voir rapidement les résultats de l'optimisation des hyperparamètres et d'évaluer la performance du modèle optimisé sur des données non vues durant l'entraînement et l'optimisation.

Voici l'affichage correspondant :

Meilleurs paramètres : 'C' : 1, 'l1_ratio' : 0.4, 'max_iter' : 1000, 'penalty' : 'elasticnet'

Meilleur score : 0.5895009263286214

Accuracy du meilleur modèle : 0.5923140628339388

Le rapport de classification est édité pour permettre de noter les différentes métriques du meilleur modèle de prédiction retenu.

Voici l'affichage correspondant :

Rapport de classification du meilleur modèle :

	precision	recall	f1-score	support
-1	0.00	0.00	0.00	29
1	0.62	0.83	0.71	31181
2	0.00	0.00	0.00	2078
3	0.51	0.07	0.12	11552
4	0.56	0.59	0.58	30024
accuracy			0.59	74864
macro avg	0.34	0.30	0.28	74864
weighted avg	0.56	0.59	0.55	74864

1.3 Analyse des résultats

Du point de vue de la performance globale, l'accuracy de 59% indique que le modèle est moyennement satisfaisant, mais il y a encore une marge d'amélioration significative. Le modèle performe de manière cohérente entre la validation croisée (58.95%) et l'ensemble de test (59.23%), ce qui suggère une bonne généralisation.

Par contre, la performance par classe du modèle est clairement insatisfaisante : d'abord parce que les classes sont mal réparties entre elles (sur-représentation des classes 1 et 4), ensuite parce que la classe 2 relative aux accidents graves est clairement mal prédite (recall : 0.00 et F1-score : 0.00).

Au-delà du rééquilibrage nécessaire des classes, il est important de noter que la régression logistique est un modèle linéaire. De ce fait, elle peut ne pas capturer des relations complexes ou non linéaires de nos données.

Il apparaît intéressant d'essayer des algorithmes non linéaires comme les forêts aléatoires ou les gradient boosting machines pour aller plus loin. Aussi, essayer plusieurs modèles doit permettre d'améliorer les prédictions.

2 Modélisation 2

Après les résultats insuffisants du premier préprocessing, de nouvelles modélisations sont effectuées sur la base du deuxième préprocessing.

2.1 Implémentation de multiples modèles de classification

Le code correspondant a été conçu pour permettre d'essayer facilement plusieurs modèles en ajustant divers paramètres, en affichant des métriques de performances, temps d'entraînement et graphiques. Sa structure permet de réaliser facilement des modifications. La liste des modèles, pour laquelle quelques informations sont données, figure dans un dictionnaire très facile à compléter.

- **Importation des librairies**

D'abord, le code importe une vaste gamme de bibliothèques et modules Python essentiels pour l'analyse de données, l'apprentissage automatique et la visualisation. Il inclut des outils fondamentaux comme NumPy et Pandas pour la manipulation de données, des modules pour la gestion du temps et des fichiers, ainsi que des bibliothèques de visualisation comme Matplotlib et Seaborn, et Tabulate pour le formatage. Le code importe également de nombreux modules de scikit-learn, couvrant le prétraitement des données, la décomposition, la sélection de modèles, divers algorithmes de classification, des méthodes d'ensemble, et des métriques d'évaluation. Enfin, il intègre d'autres bibliothèques spécialisées comme LightGBM pour l'apprentissage automatique .

- **Fonction d'affichage des messages**

La fonction 'printlog()' est créée pour afficher à l'écran et enregistrer dans une chaîne de caractères 'logres' divers messages relatifs au processus d'entraînement et d'évaluation du modèle. Ces messages incluent les paramètres testés, les résultats de la recherche des meilleurs hyperparamètres, les scores de performance, la matrice de confusion, et d'autres informations pertinentes. Cette fonction permet ainsi de suivre en détail le déroulement de l'entraînement et de sauvegarder ces résultats pour une analyse ultérieure.

- **Fonction de formatage de la durée**

La fonction 'format_duree' sera utilisée spécifiquement pour formater le temps d'entraînement des modèles de manière plus lisible. Elle est appelée pour afficher la durée d'entraînement dans les résultats. Cette fonction permet de convertir le temps d'entraînement (initialement en secondes) en un format plus facile à lire, utilisant des heures, minutes et secondes selon la durée totale de l'entraînement.

- **Chargement et préparation des données**

Le code créé charge les données depuis un fichier CSV, puis sépare les variables explicatives (X) de la variable cible (y) nommée 'grav_grave'. Les noms des colonnes de (X) sont également enregistrés dans la variable 'feature_names' qui sera utile pour l'analyse de 'feature_importances'.

Cette étape prépare les données pour l'analyse et la modélisation en machine learning, en distinguant les informations utilisées pour faire des prédictions (X) de ce qui doit être prédit (y).

- **Réduction de dimension et standardisation des données**

Le jeu de données contient plus de 250 variables explicatives. Ce nombre étant très important, nous décidons d'appliquer une Analyse en Composantes Principales (PCA) pour réduire la dimensionnalité des données à 100 composantes. Après avoir appliqué la PCA, nous normalisons le jeu de données de dimension réduite.

Cette procédure vise à simplifier le jeu de données tout en préservant l'information essentielle pour la modélisation.

Un graphique nous a aidé à choisir le nombre de composantes en affichant la variance expliquée par composante :

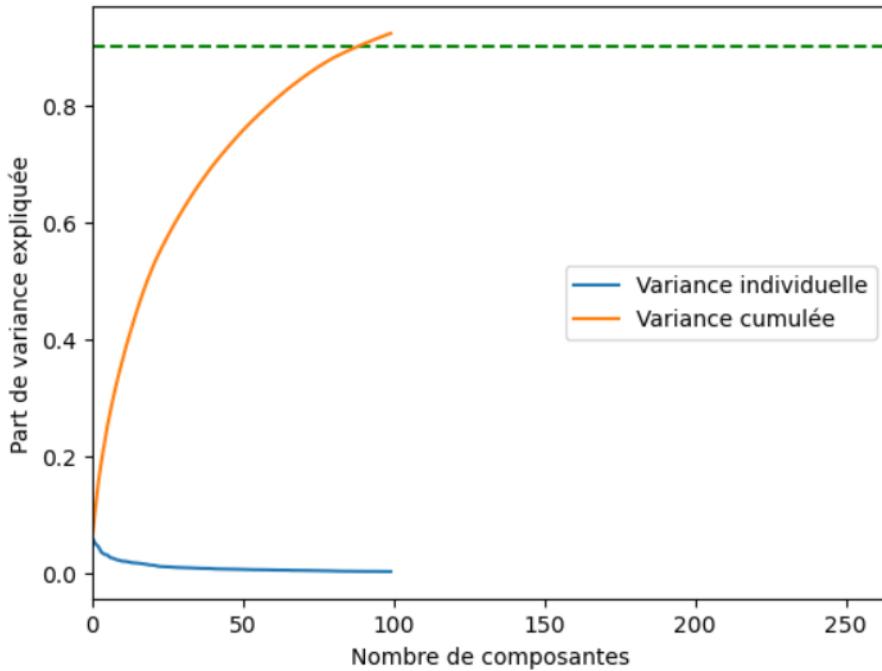


FIGURE 1 – Variance expliquée en fonction du nombre de composantes

Nous observons ici que 100 composantes principales suffisent à capturer plus de 90% de la variance totale des données originales. A notre avis, cela offre un bon compromis entre la réduction de dimensionnalité et la préservation de l'information.

- **Séparation du jeu de données**

Les données sont divisées en ensembles d'entraînement (80%) et de test (20%), et le random_state est fixé à 1234 pour assurer la reproductibilité de cette division.

- **Définition des modèles**

Le dictionnaire 'modeles' fixe la liste des modèles à implémenter sur la base de quelques éléments mentionnés ci-après :

- "prmmsgscv" : paramètres à optimiser via une grille de recherche (GridSearchCV).
- "prmfixes" : paramètres fixes du modèle.
- "perf" : destiné à stocker les performances du modèle.
- "nom" : nom court du modèle.
- "libelle" : intitulé long du modèle.
- "classe" : chemin d'importation de la classe du modèle dans scikit-learn.
- "instance" : instance du modèle avec des paramètres par défaut ou spécifiques.

Extrait du code Python

```

1 modeles = {
2
3     "SVC" : {
4         "prmmsgscv" : {"gamma" : ["scale", "auto"]}
5                 },
6         "prmfixes" : {},
7         "pred" : None,
8         "perf" : [],
9         "nom" : "SVC",
10        "libelle" : "Classification à support de vecteurs",
11        "classe" : "sklearn.svm.SVC",
12        "instance" : SVC(),

```

```

13     "grid"      : None
14 },
15
16 "LR" : {
17     "prmsgscv" : {'solver' : ['liblinear', 'lbfgs'],
18                  'C'       : [0.003, 0.005, 0.01, 0.02, 0.04]},
19     "prmfixes" : {},
20     "perf"     : {},
21     "nom"      : "LogisticRegression",
22     "libelle"   : "Régression logistique",
23     "classe"    : "sklearn.xxx",
24     "instance"  : LogisticRegression(max_iter = 10000),
25     "grid"      : None
26 },
27
28 (...)
```

Voici un extrait du texte produit :

Tailles	Total	Graves (True)		Non graves (False)	
		nombre	prop	nombre	prop
Jeu d'entraînement	142113	71012	49.97%	71101	50.03%
Jeu de test	35529	17809	50.13%	17720	49.87%
Total	177642	88821	50.00%	88821	50.00%

Nombre de variables explicatives (après PCA) : 100

- - - Classification à support de vecteurs [1/11] - - -

Paramètres essayés :

SVC()

Entraînement avec GridSearchCV : Recherche des meilleurs paramètres

Paramètres retenus :

Prédiction :

Durée	:	41 min 30 s
Score sur le jeu d'entraînement	:	0.8500%
Score sur le jeu de test	:	0.7813
Scores F1	:	0.7703 0.7913

	precision	recall	f1-score	support
False	0.81	0.74	0.77	17720
True	0.76	0.83	0.79	17809
accuracy			0.78	35529
macro avg	0.78	0.78	0.78	35529
weighted avg	0.78	0.78	0.78	35529

Pas de feature_importances_

Matrice de confusion :

Enregistrement du modèle entraîné : SVC.mdl, 65892763octets, 62Mo

(...)

Classe prédictive	False	True
Classe réelle		
False	13029	4691

True 3080 14729

```
[[ 36.67145149  13.20329871 ]
 [ 8.66897464  41.45627516 ]]
```

Ce code est, d'abord, destiné à la mise au point du notebook, puis à l'affichage de résultats, son affichage n'est pas optimisé pour la lisibilité mais pour le suivi des entraînements.

2.2 Évaluation des performances des modèles

- Initialisation de comptes-rendus d'analyse

Le code initialise des variables et structures de données pour l'analyse et l'évaluation des modèles de machine learning :

- Création d'un DataFrame vide (tbl_perf) pour stocker les performances des modèles, avec des colonnes pour différentes métriques (score, durée, précision, rappel, etc.).
- Initialisation de 'logres' pour stocker les messages de log.
- Calcul et stockage de statistiques sur les ensembles de données : nombre total d'observations, nombre d'observations pour l'entraînement et le test, nombre d'observations 'graves' et 'non graves' pour chaque ensemble, etc.
- Affichage d'un tableau récapitulatif des statistiques calculées, montrant la répartition des classes dans les ensembles d'entraînement et de test.
- Affichage du nombre de variables explicatives après PCA.
- Initialisation d'un compteur de temps (ttot1) et d'un index de modèle (idx_modele) pour le suivi chronométré et numéroté des modèles.

Ce code prépare l'environnement pour l'évaluation systématique des modèles, en organisant la structure pour stocker et afficher les résultats de manière cohérente.

Voici un extrait des résultats obtenus :

	Libellé	Entraînement	Test	f1 0	f1 1	AUC
0	Classification à support de vecteurs	0.849979	0.781277	0.770286	0.791265	0.781162
1	Régression logistique	0.755167	0.754567	0.750372	0.758623	0.754529
2	Hist Gradient Boosting Classifier	0.973542	0.773819	0.765496	0.781571	0.773735

	Libellé	Précision	Recall 0	Recall 1	f bêta=2
0	Classification à support de vecteurs	0.783624	0.735271	0.827054	0.812357
1	Régression logistique	0.754774	0.739616	0.769442	0.765078
2	Hist Gradient Boosting Classifier	0.775022	0.740181	0.807288	0.796801

	Libellé	Vrais nég.	%	Faux pos.	%	Faux nég.	%	Vrais pos.	%
0	Classification à support de vecteurs	13029	36.6715	4691	13.2033	3080	8.66897	14729	41.4563
1	Régression logistique	13106	36.8882	4614	12.9866	4106	11.5568	13703	38.5685
2	Hist Gradient Boosting Classifier	13116	36.9163	4604	12.9584	3432	9.65971	14377	40.4655

	Libellé	Durée entraînement	Id.	Taille modèle
0	Classification à support de vecteurs	62490.9	41 min 30 s	6.58928e+07
1	Régression logistique	0.439238	0 s	1647
2	Hist Gradient Boosting Classifier	66.7282	1 min 6 s	2.37262e+07

- **Optimisation des modèles et évaluation des performances**

Un code est créé pour effectuer une boucle d'entraînement et d'évaluation pour plusieurs modèles de classification figurant dans le dictionnaire préalablement défini. Voici les principales étapes :

- Pour chaque modèle dans le dictionnaire 'modeles' :
 - Affiche les informations sur le modèle et ses paramètres.
 - Utilise GridSearchCV pour optimiser les hyperparamètres du modèle.
 - Entraîne le modèle avec les meilleurs paramètres trouvés.
 - Fait des prédictions sur l'ensemble de test.
- Calcule diverses métriques de performance :
 - Scores d'entraînement et de test.
 - Scores F1, AUC, précision, rappel, et F-beta.
 - Matrice de confusion.
- Tente d'afficher l'importance des caractéristiques si disponible.
- Enregistre le modèle entraîné dans un fichier.
- Stocke toutes les métriques de performance dans le DataFrame 'tbl_perf'.
- Après avoir traité tous les modèles :
 - Affiche le tableau récapitulatif des performances (tbl_perf).
 - Calcule et affiche le temps total d'exécution.
- Génère un rapport complet dans un fichier texte "entraînement.txt".

Ce code permet une évaluation systématique et comparative de plusieurs modèles de classification, en enregistrant leurs performances et caractéristiques.

Ensuite, un autre code vient le compléter pour utiliser la fonction 'tabulate' en vue d'afficher de manière formatée et lisible les performances des différents modèles de classification stockées dans le DataFrame 'tbl_perf'. Il crée cinq tableaux distincts :

- Le premier tableau montre les scores d'entraînement, de test, les scores F1 pour les classes 0 et 1, et l'AUC.
- Le deuxième tableau présente la précision, le recall pour les classes 0 et 1, et le score F-beta avec beta=2.
- Le troisième tableau affiche les valeurs brutes de la matrice de confusion (vrais négatifs, faux positifs, faux négatifs, vrais positifs).
- Le quatrième tableau montre les mêmes informations que le précédent, mais en pourcentages.
- Le dernier tableau indique la durée d'entraînement (en secondes et en format lisible) ainsi que la taille du modèle enregistré.

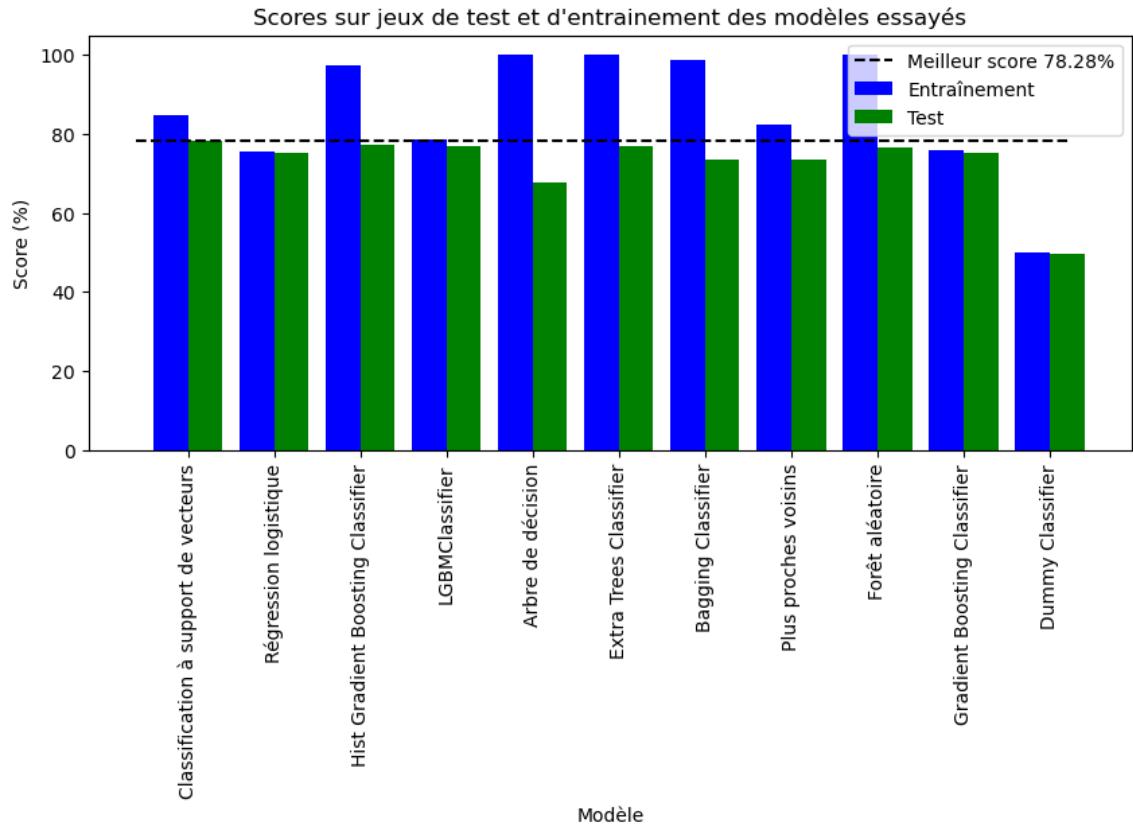
2.3 Analyse des résultats

- **Graphique de comparaison des scores d'entraînement et de test par modèle**

Un code est défini pour créer un graphique à barres comparant les scores d'entraînement (en bleu) et de test (en vert) pour les différents modèles de classification :

- Il crée une figure de taille 10x4.
- Définit la largeur des barres et calcule les positions sur l'axe x pour chaque modèle.
- Trace deux séries de barres :
 - Barres bleues pour les scores d'entraînement
 - Barres vertes pour les scores de test, légèrement décalées
- Ajoute une ligne horizontale en pointillés noirs indiquant le meilleur score de test obtenu.
- Configure le titre, les étiquettes des axes et la légende.
- Affiche la légende et le graphique.

Nous obtenons le résultat suivant :



Ce graphique permet de visualiser rapidement les performances de chaque modèle et de comparer leurs scores d'entraînement et de test, facilitant l'identification du meilleur modèle et la détection d'éventuels problèmes de surapprentissage.

Notamment, il en ressort les éléments suivants :

- **Performance générale :**

Le meilleur score de test de 78.28% (indiqué par la ligne pointillée horizontale) est obtenu par "Classification à support de vecteurs" (SVM), et la plupart des modèles atteignent des performances comprises entre 65% et 80% sur les données de test (sauf le "Dummy Classifier" qui a servi de modèle de référence).

- **Surapprentissage :**

Plusieurs modèles montrent des signes de surapprentissage, notamment les modèles "Arbre de décision", "Hist Gradient Boosting Classifier", "Extra Trees Classifier" et "Bagging Classifier" qui affichent un écart important entre le score d'entraînement et le score de test.

- **Modèles les plus équilibrés :**

Les modèles de "Régression logistique" et de "LGBMClassifier" montrent des performances quasi-similaires sur les données d'entraînement et de test, ce qui suggère leur capacité de bonne généralisation sur de nouvelles données.

Pour ces différentes raisons, nous choisissons le modèle "LGBMClassifier" comme base à notre analyse SHAP qui sera effectuée plus bas.

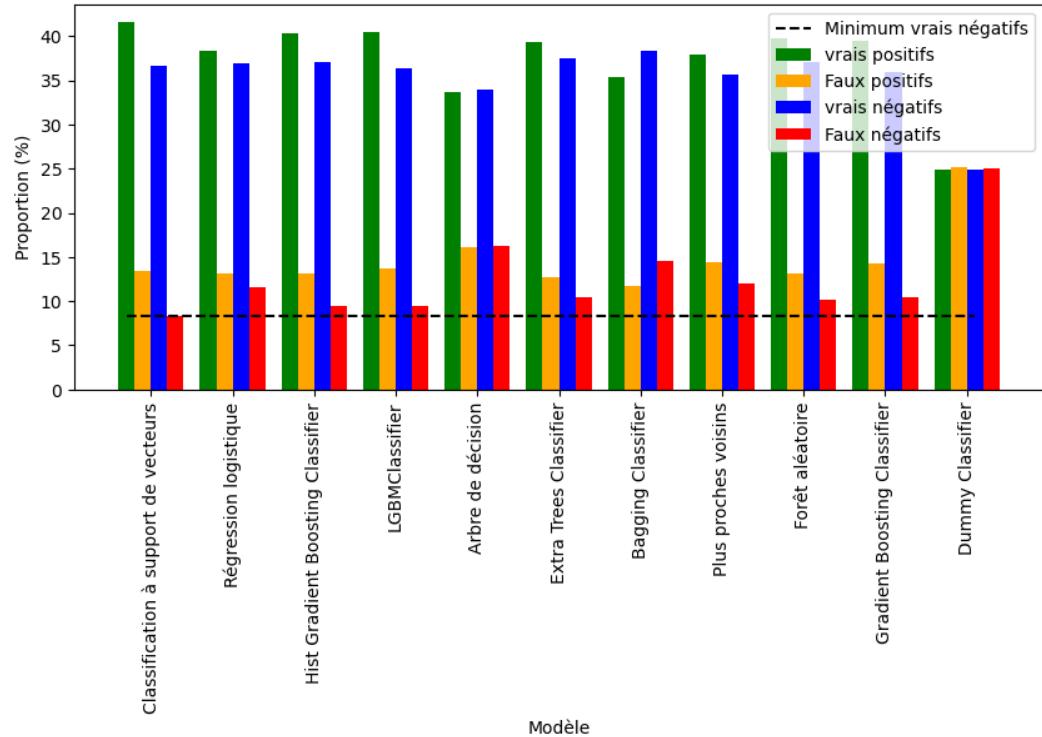
- **Graphique d'évaluation des performances par modèle**

Dans la continuité, un code est défini pour créer un graphique à barres empilées comparant les performances des différents modèles de classification :

- Il crée une figure de taille 10x4.
- Définit la largeur des barres et calcule quatre séries de positions sur l'axe x, une pour chaque catégorie de prédiction.
- Génère quatre séries de barres côté à côté pour chaque modèle représentant :

- en vert, les vrais positifs (TP)
- en orange, les faux positifs (FP)
- en bleu, les vrais négatifs (TN)
- en rouge, les faux négatifs (FN)
- Calcule et ajoute une ligne horizontale en pointillés noirs indiquant le minimum de faux négatifs parmi les modèles (excluant le "Dummy Classifier").
- Configure les étiquettes des axes et la légende.
- Affiche la légende et le graphique.

Nous obtenons le résultat suivant :



Ce graphique permet de visualiser la répartition des prédictions (TP, FP, TN, FN) pour chaque modèle, offrant une comparaison détaillée de leurs performances en termes de classification. Il aide à identifier les modèles les plus équilibrés et performants en montrant clairement leurs forces et faiblesses dans chaque catégorie de prédiction.

Il en ressort les points suivants :

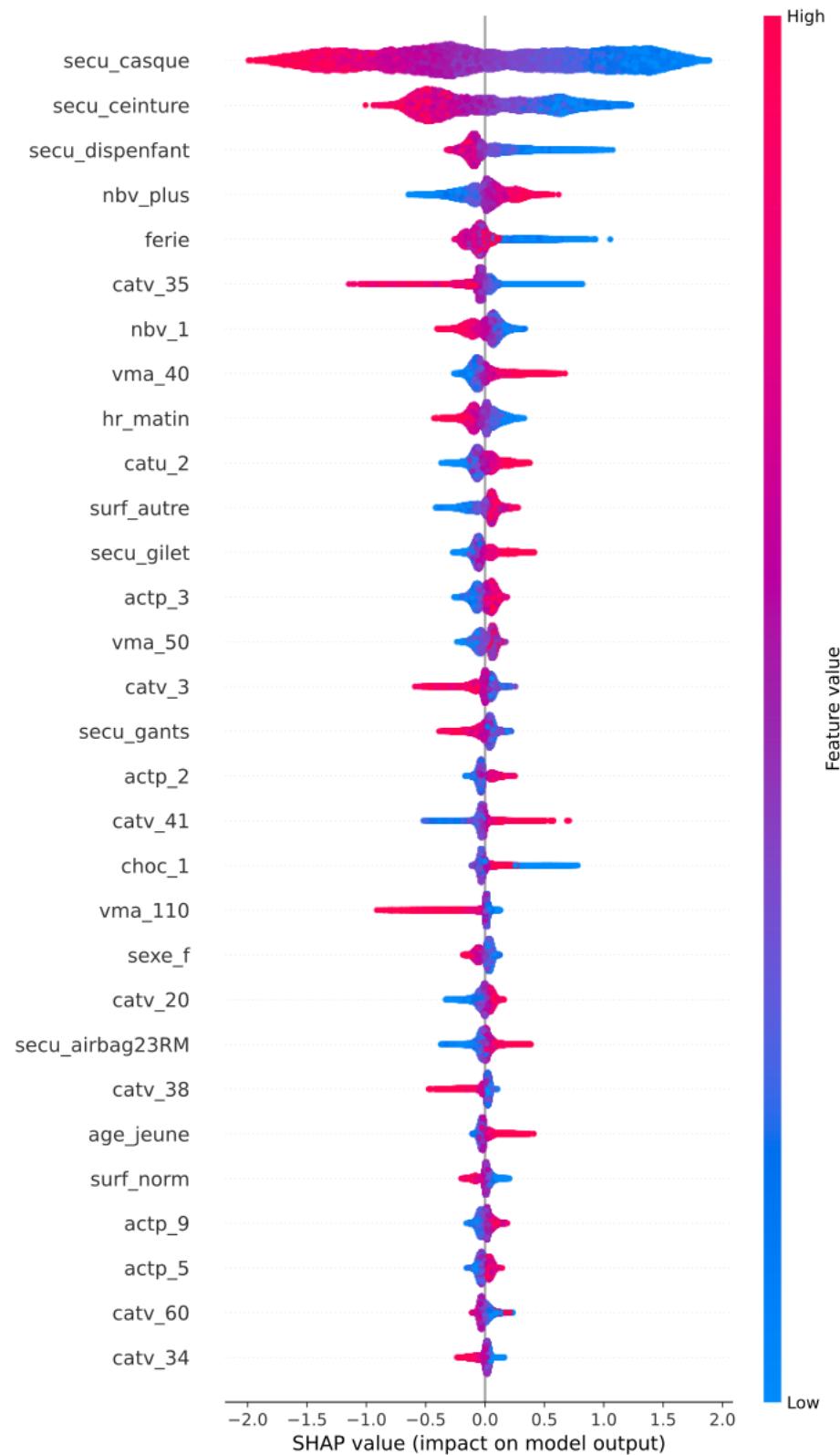
- Analyse générale de la distribution des modèles :
 - Forte proportion de Vrais Positifs (vert) environ 35-40%, et de Vrais Négatifs (bleu) environ 35%.
 - Faible proportion de Faux Positifs (orange) et de Faux Négatifs (rouge), environ 10-15%.
- Performance et équilibre des modèles :
 - La Classification à support de vecteurs (SVM) montre le plus haut taux de Vrais Positifs (environ 41%) et un bon taux de Vrais Négatifs.
 - Le modèle de Régression logistique montre un meilleur équilibre entre Vrais Positifs et Vrais Négatifs, avec des taux relativement bas de fausses prédictions.

• Analyse SHAP

Les dernières cellules contiennent le code d'une analyse SHAP pour interpréter les prédictions du modèle LGBM. LGBM est un des meilleurs modèles et il a une taille réduite sur disque. La première cellule du notebook importe la librairie SHAP. L'évaluation avec SHAP est réalisée avec deux cellules à la fin du notebook ; cette décomposition permet de réduire les temps de mise au point. La première cellule crée un TreeExplainer utilisant le classificateur LGBM entraîné et les données du jeu de test, puis lui fait faire une évaluation avec les données de test. La deuxième cellule fait l'affichage des variables explicatives les plus pertinentes et leur impact sur la cible.

Nous pouvons voir des effets que nous connaissons tous et qu'on pourrait qualifier d'"évidences" : le port du casque réduit nettement le risque de blessures graves et de décès, les cyclistes sont très exposés, la ceinture protège,...

Le graphique montre que notre modèle voit les impacts des circonstances des accidents. Il devra être utilisé avec une combinaison de circonstances.



3 Modélisation 3

3.1 Implémentation d'un modèle de deep learning

- Importation des librairies

Pour cette modélisation, plusieurs bibliothèques essentielles sont importées : pandas et numpy pour la manipulation des données, seaborn et matplotlib.pyplot pour la visualisation, scikit-learn pour la préparation et l'évaluation des données, ainsi que GridSearchCV pour l'optimisation des hyperparamètres. Pour les réseaux de neurones, le code utilise Keras via TensorFlow, incluant KerasClassifier, Sequential, Dense, Dropout, ainsi que les optimiseurs Adam et RMSprop, tout en intégrant des fonctions pour l'analyse des courbes ROC et le module warnings pour gérer les avertissements.

- Chargement et préparation des données

Le code créé définit d'abord le chemin d'accès au répertoire contenant les données prétraitées. Il procède ensuite au chargement de l'ensemble de données à partir d'un fichier CSV, puis prépare ces données. Il convertit toutes les colonnes en type entier, sépare donc les variables explicatives (features) de la variable cible 'grav_grave', divise les données en ensembles d'entraînement (80%) et de test(20%) avec une stratification pour maintenir la proportion de classes dans les ensembles, et normalise les features à l'aide de StandardScaler.

Cette préparation assure que les données sont dans un format approprié pour l'entraînement du modèle de réseau de neurones.

- Préparation du modèle

Le code créé définit une fonction 'create_model' qui construit un réseau de neurones séquentiel pour la classification binaire, avec :

- Trois couches denses (64, 32, et 16 neurones) utilisant l'activation ReLU.
- Des couches de dropout (taux de 0.4) entre chaque couche dense pour réduire le surapprentissage.
- Une couche de sortie avec un seul neurone et une activation sigmoid pour la classification binaire.

Le modèle est compilé avec la fonction de perte 'binary_crossentropy', adaptée à la classification binaire, et utilise l'accuracy comme métrique.

Un wrapper KerasClassifier est créé autour de ce modèle, permettant son utilisation avec des outils de scikit-learn comme GridSearchCV. Il est configuré avec des paramètres par défaut pour l'optimiseur, le nombre d'époques, la taille de batch, et l'initialisation des poids.

Cette approche permet une flexibilité dans l'optimisation des hyperparamètres et facilite l'intégration du modèle de deep learning dans un pipeline de machine learning.

- Configuration et exécution de la recherche d'hyperparamètres

Le code créé configure et exécute une recherche d'hyperparamètres pour optimiser le modèle de réseau de neurones :

- Il définit une grille de paramètres à tester, incluant différentes tailles de batch, nombres d'époques, optimiseurs, et initialisations de poids.
- GridSearchCV est configuré avec le modèle KerasClassifier et la grille de paramètres définie. Une validation croisée à 3 plis est spécifiée.
- La recherche d'hyperparamètres est lancée en ajustant le modèle sur les données d'entraînement normalisées.
- Enfin, le code affiche le meilleur score obtenu et les hyperparamètres correspondants.

Cette approche permet d'explorer systématiquement différentes configurations du modèle pour trouver celle qui offre les meilleures performances, en utilisant la validation croisée pour une évaluation robuste.

Nous obtenons le résultat suivant :

Best score : 0.7827, using Best parameters : 'batch_size' : 32, 'epochs' : 50, 'model_init' : 'he_normal', 'optimizer' : 'adam'

- Entraînement du meilleur modèle

Le code créé finalise la création et l'entraînement du modèle optimal :

- Il extrait les meilleurs hyperparamètres identifiés par la recherche GridSearchCV.
- Un nouveau modèle Keras est créé en utilisant ces hyperparamètres optimaux, notamment l'optimiseur et l'initialisation des poids.
- Ce modèle optimisé est ensuite entraîné sur les données d'entraînement normalisées.
- L'entraînement utilise une division de validation (20% des données d'entraînement), le nombre d'époques et la taille de batch optimaux.
- L'historique de l'entraînement est conservé, permettant de suivre l'évolution des métriques au fil des époques.

Cette étape permet d'obtenir un modèle final qui intègre les meilleures configurations identifiées lors de la recherche d'hyperparamètres, tout en fournissant des informations sur le processus d'entraînement.

3.2 Evaluation des performances

- **Evaluation et prédictions**

Le code créé évalue les performances du modèle optimisé sur l'ensemble de test :

- Il utilise la méthode evaluate() pour calculer la perte (loss) et la précision (accuracy) du modèle sur les données de test normalisées.
- Ces métriques de performance sont affichées avec une précision de 4 décimales.
- Ensuite, le modèle est utilisé pour faire des prédictions sur l'ensemble des données de test.
- Les prédictions, qui sont initialement des probabilités, sont converties en classes binaires (0 ou 1) en utilisant un seuil de 0.5 (point d'équilibre par défaut).

Cette évaluation permet de mesurer la performance du modèle, donnant une estimation plus réaliste de sa capacité de généralisation.

Nous obtenons le résultat suivant :

Test loss : 0.4792

Test accuracy : 0.7781

- **Métriques de performance**

Un code permet d'évaluer les performances du modèle en affichant un rapport de classification et une matrice de confusion. Également, il calcule et affiche le score AUC-ROC, qui correspond à la mesure de la capacité du modèle à distinguer entre les classes.

Le code créé effectue une évaluation des performances du modèle :

- Il génère et affiche un rapport de classification, qui inclut la précision, le rappel, le score F1 et le support pour chaque classe, ainsi que les moyennes.
- Une matrice de confusion est calculée et affichée, permettant de visualiser les vrais positifs, faux positifs, vrais négatifs et faux négatifs.
- Enfin, le score AUC-ROC est calculé et affiché. Ce score mesure la capacité du modèle à distinguer entre les classes.

Ces métriques offrent une vue complète des performances du modèle pour pouvoir appréhender son efficacité dans différents aspects de la classification.

Nous obtenons les résultats suivants :

Classification Report :

	precision	recall	f1-score	support
0	0.80	0.75	0.77	17765
1	0.76	0.81	0.79	17764
accuracy			0.78	35529
macro avg	0.78	0.78	0.78	35529
weighted avg	0.78	0.78	0.78	35529

Confusion Matrix :

AUC-ROC Score : 0.8570

```
[[ 13240    4525 ]
 [ 3358   14406 ]]
```

3.3 Analyse des résultats

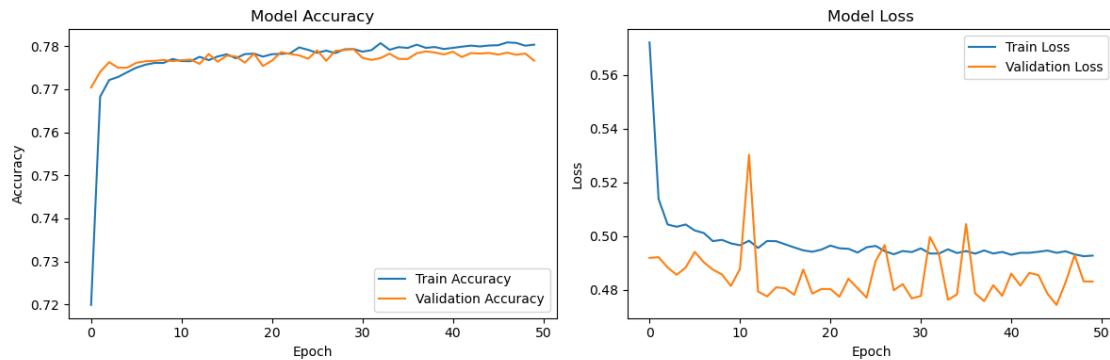
- **Tracé de l'historique d'entraînement**

Le code créé génère une visualisation de l'historique d'entraînement du modèle :

- Il génère une figure avec deux sous-graphiques côte à côté.
- Le premier graphique montre l'évolution de la précision (accuracy) au fil des époques, à la fois pour l'ensemble d'entraînement et l'ensemble de validation.
- Le second graphique illustre l'évolution de la perte (loss) au cours de l'entraînement, également pour les ensembles d'entraînement et de validation.

Cette visualisation permet d'observer rapidement comment le modèle a progressé durant l'entraînement, d'identifier d'éventuels problèmes de surapprentissage ou de sous-apprentissage, et de voir à quel moment les performances ont commencé à se stabiliser.

Nous obtenons les résultats suivants :



Interprétation globale :

- **Apprentissage efficace :**

Le modèle apprend efficacement, atteignant taux de précision de 0.76 sur les données d'entraînement et de validation.

- **Excellent généralisation :**

Le modèle généralise très bien aux données de validation, avec des performances presque identiques à celles sur l'ensemble d'entraînement.

- **Convergence rapide :**

Le modèle atteint un niveau de performance élevé dès les 5 premières époques, avec une stabilisation progressive par la suite.

- **Surapprentissage minimal :**

L'écart très faible entre les performances d'entraînement et de validation indique l'absence de surapprentissage significatif.

- **Stabilité de la perte (loss) :**

Après une diminution initiale rapide, la fonction de perte se stabilise autour de 0.49 pour l'entraînement et oscille légèrement pour la validation entre 0.48 et 0.52.

- **Optimisation réussie :**

Le modèle semble bien équilibré entre apprentissage et généralisation, suggérant que les techniques de régularisation (comme le dropout) ont été efficaces.

En conclusion, ces graphiques montrent un modèle très bien optimisé, avec une bonne capacité d'apprentissage et de généralisation. Le risque de surapprentissage est minimal, ce qui est idéal pour la tâche de prédiction de la gravité des accidents de la route. Les performances sont stables et cohérentes entre l'entraînement et la validation, indiquant que le modèle devrait bien se comporter sur de nouvelles données non vues.

• Fonctions d'analyse et de visualisation

Le code créé définit trois fonctions importantes pour l'analyse et la visualisation des résultats du modèle :

- La fonction 'get_feature_importance' calcule l'importance des caractéristiques du modèle de réseau de neurones en utilisant les poids de la première couche avec 'weights', puis en calculant la moyenne des valeurs absolues de ces poids avec 'importance'. Elle retourne un dictionnaire associant les noms des caractéristiques à leur importance.
- La fonction 'plot_feature_importance' trace un graphique à barres montrant les 20 caractéristiques les plus importantes du modèle (triées par ordre décroissant).
- La fonction 'plot_roc_curve' trace la courbe ROC pour évaluer la performance du classificateur binaire. Elle utilise 'roc_curve' pour calculer le taux de faux positifs (fpr) et le taux de vrais positifs (tpr) en fonction des vraies étiquettes (y_true) et des probabilités prédictives (y_pred_proba). L'AUC est calculée en utilisant auc(fpr, tpr).

Ces fonctions permettent une analyse visuelle approfondie des résultats du modèle, en mettant en évidence les caractéristiques les plus influentes et en évaluant la capacité du modèle à distinguer entre les classes à différents seuils de classification.

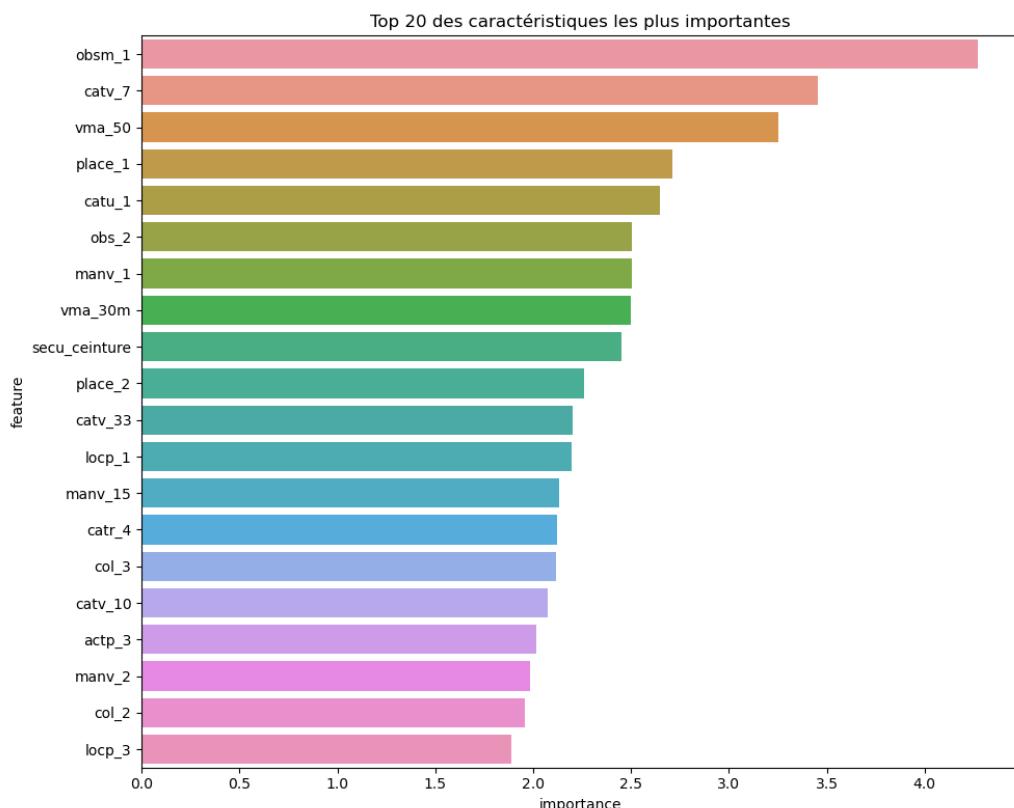
• Utilisation des fonctions d'analyse et de visualisation

Le code créé applique les fonctions d'analyse et de visualisation précédemment définies :

- Il calcule d'abord l'importance des caractéristiques du meilleur modèle en utilisant la fonction 'get_feature_importance'. Le résultat est stocké dans un dictionnaire.
- Ce dictionnaire est ensuite utilisé pour générer un graphique à barres des caractéristiques les plus importantes via la fonction 'plot_feature_importance'.
- Ensuite, le code génère des prédictions probabilistes sur l'ensemble de test normalisé.
- Ces prédictions sont utilisées avec les vraies étiquettes pour tracer la courbe ROC en appelant la fonction 'plot_roc_curve'.

Cette séquence permet de visualiser rapidement quelles caractéristiques ont le plus d'impact sur les prédictions du modèle et d'évaluer graphiquement sa performance en termes de compromis entre le taux de vrais positifs et le taux de faux positifs.

Nous obtenons un premier résultat :

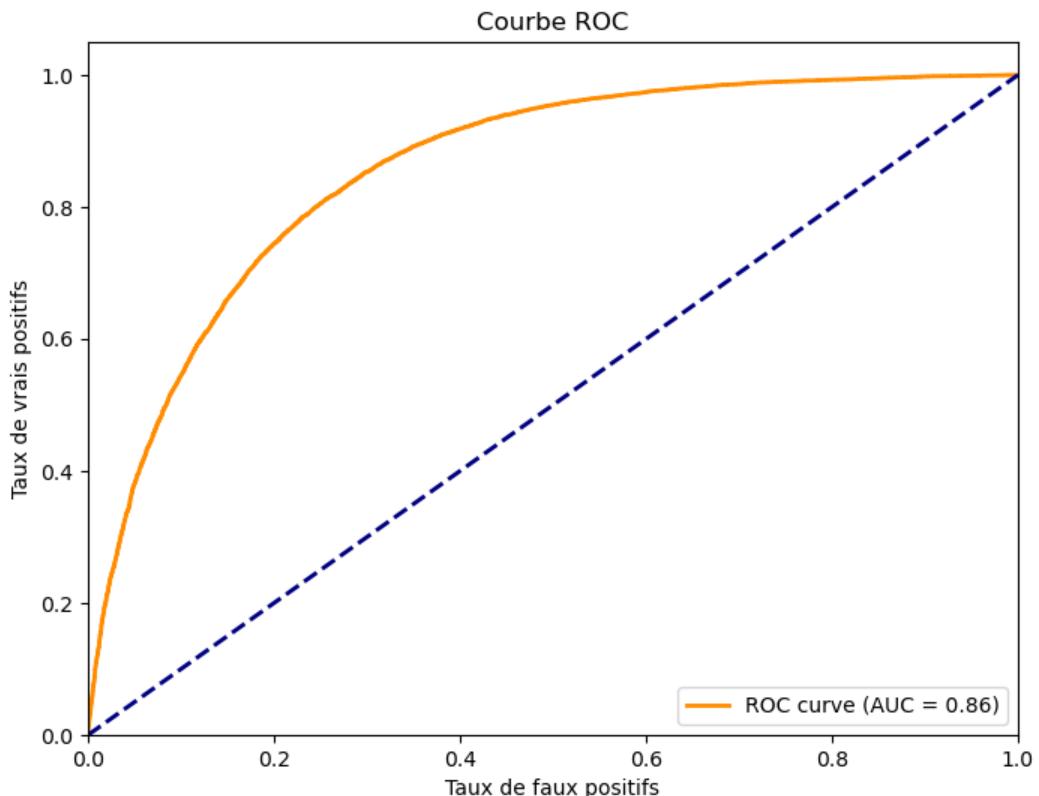


Interprétation du graphique :

- 'obsm_1' est la première caractéristique la plus importante, liée à un type d'obstacle mobile impliqué dans l'accident.
- Plusieurs catégories de véhicules (catv_7, catv_10, catv_33) sont présentes, indiquant que différents types de véhicules ont des impacts variés sur la gravité des accidents.
- 'vma_50' et 'vma_30m' indiquent une vitesse maximale autorisée, montrant que la limite de vitesse dans la zone de l'accident est un facteur important.
- Différentes valeurs de 'place' (place_1, place_2) suggèrent que la position des occupants dans le véhicule est un facteur important.
- 'catu_1' indique que le type d'usager peut influencer la gravité de l'accident.
- 'secu_ceinture' souligne l'importance du port de la ceinture de sécurité dans la détermination de la gravité des blessures.

Cette analyse met en évidence l'importance des facteurs liés au type de véhicule, aux conditions de circulation (comme la vitesse autorisée) dans la détermination de la gravité des accidents de la route et bien d'autres. Ces informations peuvent être cruciales pour cibler les efforts de prévention et améliorer la sécurité routière.

Nous obtenons ensuite un deuxième résultat :



Interprétation du graphique :

— Courbe ROC :

La courbe orange représente la performance du modèle à différents seuils de classification. Elle montre le compromis entre le taux de vrais positifs (sensibilité) et le taux de faux positifs (1 - spécificité) à mesure que le seuil de décision varie.

— Forme de la courbe :

La courbe s'élève rapidement vers le coin supérieur gauche, ce qui est souhaitable. Cela indique que le modèle atteint un bon taux de vrais positifs tout en maintenant un faible taux de faux positifs.

— **Comparaison avec la ligne de base :**

La courbe est nettement au-dessus de la ligne diagonale en pointillés (qui représente une classification aléatoire), ce qui confirme que le modèle est bien meilleur qu'une prédition basée sur le hasard.

— **Robustesse :**

La courbe lisse et régulière suggère que le modèle est robuste et performant sur une large gamme de seuils de décision.

Cette courbe ROC indique que le modèle a une excellente capacité à distinguer entre les accidents graves et non graves. Avec un AUC de 0.86, il démontre une forte performance de classification, ce qui est particulièrement important dans le contexte de la prédiction de la gravité des accidents de la route.

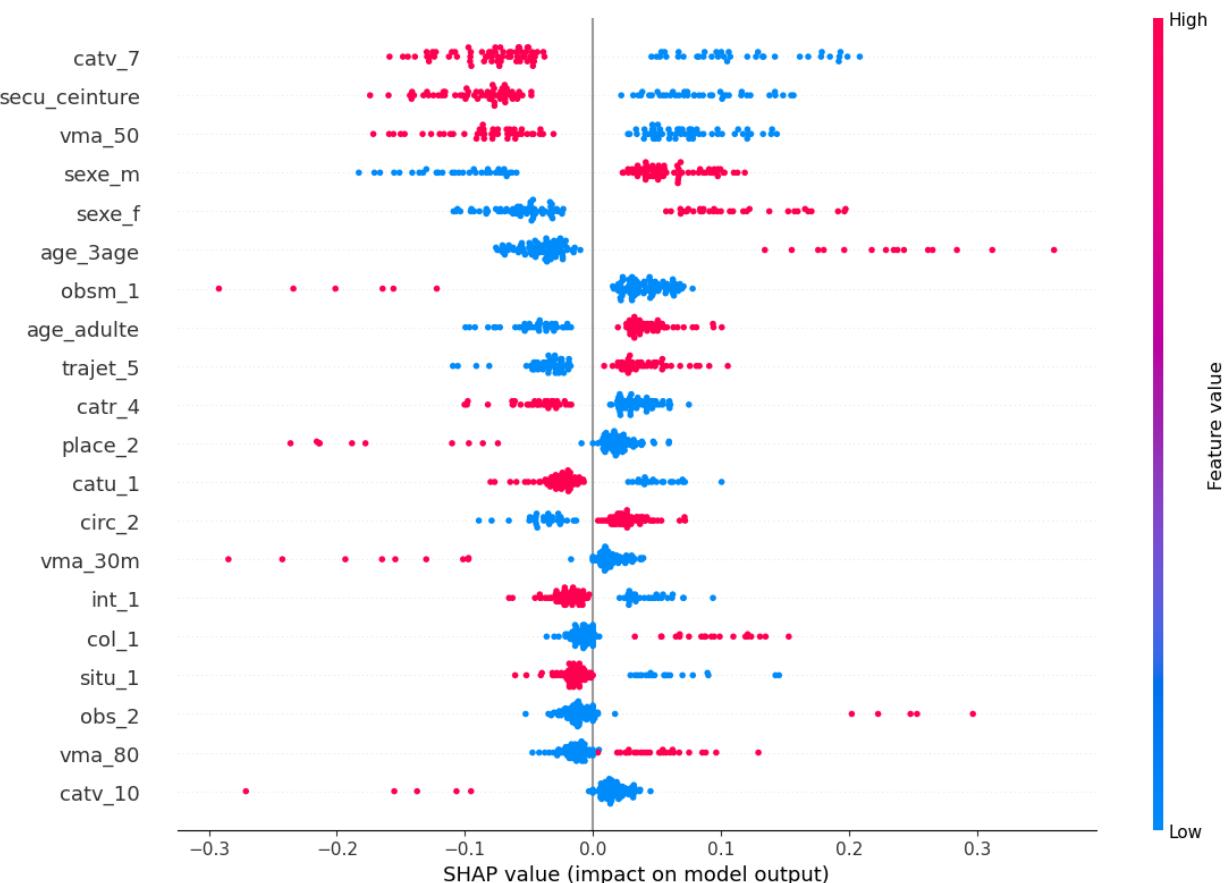
- **Analyse SHAP**

Le code créé met en œuvre une analyse SHAP pour interpréter les prédictions du modèle :

- Il commence par importer la bibliothèque SHAP.
- Une fonction de prédiction est définie pour être utilisée avec l'explainer SHAP.
- Un KernelExplainer SHAP est créé, utilisant un échantillon de 100 instances des données d'entraînement comme données de fond.
- Les valeurs SHAP sont calculées pour un sous-ensemble (max 100 échantillons) de l'ensemble de test.
- Les dimensions des valeurs SHAP et de l'échantillon de test sont vérifiées et affichées.
- Les valeurs SHAP sont ajustées si nécessaire pour s'assurer qu'elles sont dans le format approprié.
- Enfin, un graphique récapitulatif SHAP est généré, visualisant l'importance et l'impact des caractéristiques sur les prédictions du modèle.

Cette analyse SHAP fournit une interprétation détaillée de la façon dont chaque caractéristique influence les prédictions du modèle, offrant des insights sur l'importance et la direction de l'impact des variables.

Nous obtenons le résultat suivant :



Interprétation du graphique :

Ce graphique nous montre quels éléments sont les plus importants pour prédire si un accident de la route sera grave ou non. Chaque ligne représente un facteur différent, comme le port de la ceinture de sécurité ou le type de véhicule. Plus la ligne est haute dans le graphique, plus ce facteur est important. Les points sur chaque ligne nous disent deux choses :

- Leur position :
 - À gauche, ils réduisent le risque d'accident grave.
 - À droite, ils augmentent ce risque.
- Leur couleur :
 - Rouge signifie une valeur élevée pour ce facteur.
 - Bleu signifie une valeur basse.

Par exemple, pour la ceinture de sécurité (deuxième ligne) :

- Les points rouges à gauche signifient que le port de la ceinture (valeur élevée) tend à réduire la gravité de l'accident.
- Les points bleus à droite indiquent que l'absence de ceinture (valeur basse) tend à augmenter la gravité de l'accident.

En résumé, ce graphique SHAP offre une vue détaillée de comment chaque caractéristique influence les prédictions du modèle sur la gravité des accidents. Il met en évidence l'importance de facteurs tels que l'utilisation de la ceinture de sécurité, le type de véhicule, la vitesse autorisée, et les caractéristiques démographiques des personnes impliquées dans l'accident.

4 Conclusion

La comparaison approfondie des différents modèles de classification pour prédire la gravité des accidents de la route a révélé plusieurs enseignements importants :

1. Performance des modèles :

Le réseau de neurones et le SVC ont démontré les meilleures performances globales, avec des taux de rappel respectifs de 0.81 et 0.83. Cela signifie que ces modèles sont particulièrement efficaces pour correctement identifier les accidents graves (taux de recall > 0.80). Cependant, le Hist Gradient Boosting Classifier a également montré des résultats intéressants au regard de l'équilibre établi entre la performance et le temps de calcul.

2. Importance de l'interprétabilité :

L'utilisation de techniques comme SHAP a permis d'identifier les facteurs les plus influents dans la prédiction de la gravité des accidents, offrant des insights précieux au-delà de la simple performance prédictive.

3. Stabilité des prédictions :

Les différents modèles ont montré une certaine cohérence dans leurs prédictions, ce qui renforce la confiance dans les résultats obtenus et suggère que les caractéristiques identifiées comme importantes sont robustes à travers différentes approches de modélisation.

Les implications pratiques pour la prédiction de la gravité des accidents de la route :

1. Conception de politiques de sécurité routière ciblées :

L'identification des facteurs les plus influents dans la gravité des accidents (grâce à l'interprétabilité des modèles) peut guider l'élaboration de politiques de sécurité routière plus efficaces. Par exemple, si certaines conditions météorologiques ou caractéristiques routières sont fortement associées à des accidents graves, des mesures préventives spécifiques peuvent être mises en place.

2. Personnalisation des campagnes de sensibilisation :

Les insights tirés des modèles peuvent être utilisés pour créer des campagnes de sensibilisation à la sécurité routière plus ciblées et efficaces, en se concentrant sur les facteurs de risque les plus importants identifiés par les modèles.

3. Amélioration de la conception des véhicules :

Les constructeurs automobiles pourraient exploiter ces résultats pour orienter leurs efforts de recherche et développement vers des technologies de sécurité innovantes. En se concentrant sur les facteurs

identifiés comme les plus déterminants dans la gravité des accidents, ils seraient en mesure de concevoir et d'implémenter des dispositifs de sécurité plus efficaces et ciblés.

4. Amélioration des systèmes de triage et d'intervention :

La capacité à prédire avec précision la gravité des accidents peut permettre aux services d'urgence d'optimiser l'allocation des ressources. Par exemple, en envoyant immédiatement des équipes médicales plus spécialisées pour les circonstances d'accidents prédisposés comme graves.

En résumé, bien que chaque modèle présente ses propres forces et faiblesses, l'approche multi-modèles adoptée dans cette étude offre une compréhension riche et nuancée des facteurs influençant la gravité des accidents de la route. L'application pratique de ces connaissances a le potentiel d'améliorer la sécurité routière, en permettant des interventions plus ciblées et efficaces à plusieurs niveaux, de la prévention à la gestion des accidents.

CONCLUSION

Notre étude sur la prédiction de la gravité des accidents de la route en France a abouti à des résultats plutôt encourageants. Grâce à notre analyse des données d'accidents sur la période 2019-2022, et à l'utilisation de techniques avancées d'apprentissage automatique, nous avons pu atteindre notre objectif principal : identifier correctement plus de 80% des accidents graves.

Au cœur de nos découvertes, la classification par vecteurs de support (SVC) s'est particulièrement distinguée en parvenant à identifier correctement 83% des accidents graves (soit un recall de 0.83). Cette performance représente une avancée concrète dans notre capacité à anticiper les conséquences dramatiques des accidents. Le réseau de neurones et le Gradient Boosting, avec des recalls de 0.81, ont également démontré leur robustesse, ce qui confirme la fiabilité de nos approches prédictives.

Cette capacité à détecter 83% des accidents graves s'accompagne d'un taux de faux positifs d'environ 20%, ce qui signifie qu'un accident sur cinq est classé comme grave alors qu'il ne l'est pas. Dans le contexte de la sécurité routière, ce compromis est plutôt acceptable : une sur-mobilisation occasionnelle des ressources d'urgence est largement préférable à la non-détection d'accidents réellement graves qui pourrait avoir des conséquences fatales.

Un autre aspect précieux de notre étude réside dans l'identification des facteurs déterminants de la gravité des accidents. Grâce à l'analyse SHAP, nous avons pu mettre en lumière des éléments cruciaux tels que l'impact significatif du port de la ceinture de sécurité, l'influence du type de véhicule, et l'importance des conditions de circulation. Ces découvertes offrent des leviers d'action concrets pour les politiques de sécurité routière.

Malgré tout, notre travail présente des limites qu'il convient de reconnaître. La période d'étude relativement courte (limitée à quatre années) ne permet pas de capturer certaines tendances à long terme. L'absence de certaines données potentiellement pertinentes, comme l'expérience des conducteurs ou l'état technique détaillé des véhicules, laisse des zones d'ombre dans notre compréhension globale du phénomène. De plus, la limitation contextuelle au territoire de la France ne permet pas de généraliser nos résultats à d'autres pays.

Ces déconvenues, loin d'être des obstacles insurmontables, ouvrent la voie à de nouvelles perspectives de recherche. L'amélioration du taux de détection des accidents graves reste un objectif prioritaire qui pourrait être atteint par plusieurs approches complémentaires :

- Le développement de modèles plus sophistiqués prenant en compte les dépendances spatiales et temporelles.
- L'intégration de données en temps réel (conditions de circulation, météo, capteurs embarqués des véhicules modernes).
- L'extension de l'étude sur une période plus longue pour capturer les tendances à long terme.

Ceci étant, les résultats de notre travail présentent tout de même un intérêt potentiel à plusieurs niveaux. Pour les services d'urgence, nos modèles pourraient permettre une meilleure anticipation des ressources nécessaires lors des interventions. Pour les constructeurs automobiles, nos résultats peuvent guider au développement de dispositifs de sécurité plus efficaces. Pour les décideurs publics, nos découvertes offrent des bases minimales pour l'élaboration de politiques de prévention ciblées.

En définitive, bien que notre étude ne prétende pas résoudre tous les aspects de la sécurité routière, elle pose un premier jalon dans la quête d'une meilleure compréhension des facteurs de gravité des accidents de la route. Et même si le chemin à parcourir pour atteindre l'objectif d'une sécurité routière optimale soit encore long, nos modèles de prédiction constituent déjà une avancée concrète pour la préservation des vies humaines sur nos routes.

Partie Annexe

ANNEXE : ETUDE DES VARIABLES

1. Caractéristiques

Rows x columns Rows duplicated

Caracteristiques (1176873, 16) 0

a. Num_Acc

Description	Numéro d'identifiant de l'accident.																																											
Type	int64																																											
Etendue des valeurs	<table> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td>1176873</td><td>1176873</td><td>200500000001</td><td>1</td></tr> </tbody> </table>						count	unique	top	freq	Num_Acc	1176873	1176873	200500000001	1																													
	count	unique	top	freq																																								
Num_Acc	1176873	1176873	200500000001	1																																								
Valeurs nulles	<table> <thead> <tr> <th></th><th>Type</th><th>Val_notnull</th><th>Val_null</th><th>%_null</th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td>int64</td><td>1176873</td><td>0</td><td>0.0</td></tr> </tbody> </table>						Type	Val_notnull	Val_null	%_null	Num_Acc	int64	1176873	0	0.0																													
	Type	Val_notnull	Val_null	%_null																																								
Num_Acc	int64	1176873	0	0.0																																								
Outliers	<table> <thead> <tr> <th></th><th>outliers_count</th><th>outliers_unique</th><th>outliers_list</th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td>0</td><td>0</td><td>[]</td></tr> </tbody> </table>						outliers_count	outliers_unique	outliers_list	Num_Acc	0	0	[]																															
	outliers_count	outliers_unique	outliers_list																																									
Num_Acc	0	0	[]																																									
Répartition	<table> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td></td><td></td></tr> <tr> <td>200500000001</td><td>1</td><td>0.0</td></tr> <tr> <td>200500000002</td><td>1</td><td>0.0</td></tr> <tr> <td>200500000003</td><td>1</td><td>0.0</td></tr> <tr> <td>200500000004</td><td>1</td><td>0.0</td></tr> <tr> <td>200500000005</td><td>1</td><td>0.0</td></tr> <tr> <td>...</td><td>...</td><td>...</td></tr> <tr> <td>202200055298</td><td>1</td><td>0.0</td></tr> <tr> <td>202200055299</td><td>1</td><td>0.0</td></tr> <tr> <td>202200055300</td><td>1</td><td>0.0</td></tr> <tr> <td>202200055301</td><td>1</td><td>0.0</td></tr> <tr> <td>202200055302</td><td>1</td><td>0.0</td></tr> </tbody> </table>						Count	% valeurs	Num_Acc			200500000001	1	0.0	200500000002	1	0.0	200500000003	1	0.0	200500000004	1	0.0	200500000005	1	0.0	202200055298	1	0.0	202200055299	1	0.0	202200055300	1	0.0	202200055301	1	0.0	202200055302	1	0.0
	Count	% valeurs																																										
Num_Acc																																												
200500000001	1	0.0																																										
200500000002	1	0.0																																										
200500000003	1	0.0																																										
200500000004	1	0.0																																										
200500000005	1	0.0																																										
...																																										
202200055298	1	0.0																																										
202200055299	1	0.0																																										
202200055300	1	0.0																																										
202200055301	1	0.0																																										
202200055302	1	0.0																																										
	1176873 rows × 2 columns																																											

Remarque	En 2022, la variable disparait au profit de Accident_Id.
-----------------	--

b. Accident_Id

Description	Identifiant de l'accident qui remplace Num_Acc à compter de 2022 (cf. supra)
Remarque	Une modification du nom de colonne est préalablement nécessaire à l'étape de concaténation des dataframes des différentes années.

c. an

Description	Année de l'accident.			
Type	int64			
Etendue des valeurs	count unique top freq an 1176873 18 5 87026			
Valeurs nulles	Type Val_notnull Val_null %_null an int64 1176873 0 0.0			
Outliers	outliers_count outliers_unique outliers_list an 218404 4 [2019.0, 2020.0, 2021.0, 2022.0]			

Répartition	Count	% valeurs
an		
5	87026	7.0
6	82993	7.0
7	83850	7.0
8	76767	7.0
9	74409	6.0
10	69379	6.0
11	66974	6.0
12	62250	5.0
13	58397	5.0
14	59854	5.0
15	58654	5.0
16	59432	5.0
17	60701	5.0
18	57783	5.0
2019	58840	5.0
2020	47744	4.0
2021	56518	5.0
2022	55302	5.0
Remarque	Les années ne sont calibrées de la même manière (2 chiffres de 2005 à 2018, 4 chiffres de 2019 à 2022) : +2000 à rajouter aux années de moins de 4 chiffres.	

d. mois

Description	Mois de l'accident.			
Type	int64			
Etendue des valeurs	count unique top freq			
	mois 1176873 12 10 111728			
Valeurs nulles	Type	Val_notnull	Val_null	%_null
	mois	int64	1176873	0 0.0

Outliers		outliers_count outliers_unique outliers_list		
	mois	0	0	[]
Répartition				
Count % valeurs				
mois				
1	90313	8.0		
2	79959	7.0		
3	90842	8.0		
4	91380	8.0		
5	101060	9.0		
6	111000	9.0		
7	106237	9.0		
8	89006	8.0		
9	109167	9.0		
10	111728	9.0		
11	99941	8.0		
12	96240	8.0		
Evolution				
Remarque		On observe que pendant la période du covid (2020) les proportions changent drastiquement.		

e. jour

Description	Jour de l'accident.
-------------	---------------------

Type	int64										
Etendue des valeurs	<table> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>jour</td><td>1176873</td><td>31</td><td>6</td><td>40053</td></tr> </tbody> </table>		count	unique	top	freq	jour	1176873	31	6	40053
	count	unique	top	freq							
jour	1176873	31	6	40053							
Valeurs nulles	<table> <thead> <tr> <th></th><th>Type</th><th>Val_notnull</th><th>Val_null</th><th>%_null</th></tr> </thead> <tbody> <tr> <td>jour</td><td>int64</td><td>1176873</td><td>0</td><td>0.0</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	jour	int64	1176873	0	0.0
	Type	Val_notnull	Val_null	%_null							
jour	int64	1176873	0	0.0							
Outliers	<table> <thead> <tr> <th></th><th>outliers_count</th><th>outliers_unique</th><th>outliers_list</th></tr> </thead> <tbody> <tr> <td>jour</td><td>0</td><td>0</td><td>[]</td></tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	jour	0	0	[]		
	outliers_count	outliers_unique	outliers_list								
jour	0	0	[]								
Répartition	<p>jour</p>										

	Count	% valeurs
jour		
1	37018	3.0
2	38144	3.0
3	38623	3.0
4	39044	3.0
5	39125	3.0
6	40053	3.0
7	39942	3.0
8	39415	3.0
9	39646	3.0
10	39924	3.0
11	39104	3.0
12	39942	3.0
13	39006	3.0
14	39174	3.0
15	39247	3.0
16	39560	3.0
17	38773	3.0
18	38996	3.0
19	38995	3.0
20	38625	3.0
21	39083	3.0
22	38517	3.0
23	38320	3.0
24	37547	3.0
25	36901	3.0
26	36885	3.0
27	37233	3.0
28	37844	3.0
29	35535	3.0
30	35553	3.0
31	21099	2.0

f. hrmn

Description	Heures et minutes de l'accident.
Type	[2005-2018] : int64 [2019-2022] : object
Etendue des valeurs	count unique top freq <hr/> hrmn 1176873 2877 1800 14635
Valeurs nulles	Type Val_notnull Val_null %_null <hr/> hrmn object 1176873 0 0.0

Outliers	outliers_count outliers_unique			outliers_list
	hrmn	1123522	2873	[00:00, 00:01, 00:02, 00:03, 00:04, 00:05, 00...
Répartition	Count % valeurs			
	hrmn			
	00:00	373	0.0	
	00:01	71	0.0	
	00:02	20	0.0	
	00:03	14	0.0	
	00:04	10	0.0	
	
	955	1581	0.0	
	956	54	0.0	
	957	73	0.0	
	958	77	0.0	
	959	46	0.0	
	2877 rows × 2 columns			
Remarque	Le changement de format des heures et minutes intervenu depuis 2019 nécessitera une uniformisation des valeurs horaires (hh:mm / hhmm / hmm). Après correction, le regroupement par tranches d'une heure sera privilégié.			

g. lum

Description	Lumière : conditions d'éclairage dans lesquelles l'accident s'est produit :
Modalités	- 1 : Plein jour - 2 : Crénuscuile ou aube - 3 : Nuit sans éclairage public - 4 : Nuit avec éclairage public non allumé - 5 : Nuit avec éclairage public allumé
Type	int64
Etendue des valeurs	count unique top freq
	lum 1176873 6 1 803169
Valeurs nulles	Type Val_notnull Val_null %_null
	lum int64 1176873 0 0.0

Outliers	outliers_count outliers_unique outliers_list																							
	lum	0	0																					
Répartition																								
Count % valeurs	<table border="1"> <thead> <tr> <th>lum</th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>-1</td> <td>7</td> <td>0.0</td> </tr> <tr> <td>1</td> <td>803169</td> <td>68.0</td> </tr> <tr> <td>2</td> <td>71424</td> <td>6.0</td> </tr> <tr> <td>3</td> <td>98296</td> <td>8.0</td> </tr> <tr> <td>4</td> <td>9921</td> <td>1.0</td> </tr> <tr> <td>5</td> <td>194056</td> <td>16.0</td> </tr> </tbody> </table>			lum	Count	% valeurs	-1	7	0.0	1	803169	68.0	2	71424	6.0	3	98296	8.0	4	9921	1.0	5	194056	16.0
lum	Count	% valeurs																						
-1	7	0.0																						
1	803169	68.0																						
2	71424	6.0																						
3	98296	8.0																						
4	9921	1.0																						
5	194056	16.0																						
Evolution																								

h. agg

Description	Localisation agglomération.										
Modalités	<ul style="list-style-type: none"> - 1 : Hors agglomération - 2 : En agglomération 										
Type	int64										
Etendue des valeurs	<table border="1"> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>agg</td> <td>1176873</td> <td>2</td> <td>2</td> <td>791831</td> </tr> </tbody> </table>		count	unique	top	freq	agg	1176873	2	2	791831
	count	unique	top	freq							
agg	1176873	2	2	791831							
Valeurs nulles	<table border="1"> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>agg</td> <td>int64</td> <td>1176873</td> <td>0</td> <td>0.0</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	agg	int64	1176873	0	0.0
	Type	Val_notnull	Val_null	%_null							
agg	int64	1176873	0	0.0							

Outliers	outliers_count outliers_unique outliers_list						
	agg 0 0 []						
Répartition	Count % valeurs						
agg	<p>Count % valeurs</p> <table> <tr> <td>1</td> <td>385042</td> <td>33.0</td> </tr> <tr> <td>2</td> <td>791831</td> <td>67.0</td> </tr> </table>	1	385042	33.0	2	791831	67.0
1	385042	33.0					
2	791831	67.0					
Evolution							

i. int

Description	Intersection.										
Modalités	<ul style="list-style-type: none"> - 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 										
Type	int64										
Etendue des valeurs	<table> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>int</td> <td>1176873</td> <td>11</td> <td>1</td> <td>820757</td> </tr> </tbody> </table>		count	unique	top	freq	int	1176873	11	1	820757
	count	unique	top	freq							
int	1176873	11	1	820757							

Valeurs nulles	Type	Val_notnull	Val_null	%_null
	int	int64	1176873	0 0.0
Outliers		outliers_count	outliers_unique	outliers_list
	int	103582	7	[-1.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]
Répartition				
Count % valeurs				
int				
-1 8 0.0				
0 107 0.0				
1 820757 70.0				
2 143573 12.0				
3 108854 9.0				
4 18906 2.0				
5 11463 1.0				
6 35325 3.0				
7 10147 1.0				
8 1486 0.0				
9 26247 2.0				
Evolution				

int

Evolution de la distribution int

j. atm

Description	Conditions atmosphériques.
-------------	----------------------------

Modalités	- -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																																							
Type	[2005-2008 ; 2015-2016 ; 2019-2022] : int64 [2009-2014 ; 2017-2018] : float64																																							
Etendue des valeurs	<table> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>atm</td><td>1176800</td><td>10</td><td>1.0</td><td>949701</td></tr> </tbody> </table>		count	unique	top	freq	atm	1176800	10	1.0	949701																													
	count	unique	top	freq																																				
atm	1176800	10	1.0	949701																																				
Valeurs nulles	<table> <thead> <tr> <th></th><th>Type</th><th>Val_notnull</th><th>Val_null</th><th>%_null</th></tr> </thead> <tbody> <tr> <td>atm</td><td>float64</td><td>1176800</td><td>73</td><td>0.01</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	atm	float64	1176800	73	0.01																													
	Type	Val_notnull	Val_null	%_null																																				
atm	float64	1176800	73	0.01																																				
Outliers	<table> <thead> <tr> <th></th><th>outliers_count</th><th>outliers_unique</th><th>outliers_list</th></tr> </thead> <tbody> <tr> <td>atm</td><td>227099</td><td>9</td><td>[-1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]</td></tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	atm	227099	9	[-1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]																															
	outliers_count	outliers_unique	outliers_list																																					
atm	227099	9	[-1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]																																					
Répartition	<table> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>atm</td><td></td><td></td></tr> <tr> <td>-1.0</td><td>22</td><td>0.0</td></tr> <tr> <td>1.0</td><td>949701</td><td>81.0</td></tr> <tr> <td>2.0</td><td>123089</td><td>10.0</td></tr> <tr> <td>3.0</td><td>25270</td><td>2.0</td></tr> <tr> <td>4.0</td><td>6481</td><td>1.0</td></tr> <tr> <td>5.0</td><td>7927</td><td>1.0</td></tr> <tr> <td>6.0</td><td>2933</td><td>0.0</td></tr> <tr> <td>7.0</td><td>13816</td><td>1.0</td></tr> <tr> <td>8.0</td><td>39535</td><td>3.0</td></tr> <tr> <td>9.0</td><td>8026</td><td>1.0</td></tr> <tr> <td>Nan</td><td>73</td><td>0.0</td></tr> </tbody> </table> <p>The heatmap displays the distribution of weather conditions over time. The x-axis represents the weather condition (Modalites) from 1 to 11, and the y-axis represents the years from 2005 to 2022. The color intensity corresponds to the count of observations for each combination. The most frequent condition is 'Normale' (Modalite 1), which shows a significant increase in frequency from 2005 to 2022. Other conditions like 'Pluie forte' (Modalite 3) and 'Neige - grêle' (Modalite 4) have much lower counts and remain relatively stable.</p>		Count	% valeurs	atm			-1.0	22	0.0	1.0	949701	81.0	2.0	123089	10.0	3.0	25270	2.0	4.0	6481	1.0	5.0	7927	1.0	6.0	2933	0.0	7.0	13816	1.0	8.0	39535	3.0	9.0	8026	1.0	Nan	73	0.0
	Count	% valeurs																																						
atm																																								
-1.0	22	0.0																																						
1.0	949701	81.0																																						
2.0	123089	10.0																																						
3.0	25270	2.0																																						
4.0	6481	1.0																																						
5.0	7927	1.0																																						
6.0	2933	0.0																																						
7.0	13816	1.0																																						
8.0	39535	3.0																																						
9.0	8026	1.0																																						
Nan	73	0.0																																						

Evolution	<p>Evolution de la distribution atm</p> <p>Proportion (%)</p> <p>Années</p> <p>Modalités</p> <ul style="list-style-type: none"> 1 2 3 4 5 6 7 8 9 -1
Remarque	Les NaN peuvent être remplacés par la valeur -1 qui signifie « non renseigné ».

k. col

Description	Type de collision.										
Modalités	<ul style="list-style-type: none"> -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 										
Type	[2005-2009 ; 2012-2015 ; 2019-2022] : int64 [2010-2011 ; 2016-2018] : float64										
Etendue des valeurs	<table> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>col</td> <td>1176854</td> <td>8</td> <td>6.0</td> <td>381967</td> </tr> </tbody> </table>		count	unique	top	freq	col	1176854	8	6.0	381967
	count	unique	top	freq							
col	1176854	8	6.0	381967							
Valeurs nulles	<table> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>col</td> <td>float64</td> <td>1176854</td> <td>19</td> <td>0.0</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	col	float64	1176854	19	0.0
	Type	Val_notnull	Val_null	%_null							
col	float64	1176854	19	0.0							
Outliers	<table> <thead> <tr> <th></th> <th>outliers_count</th> <th>outliers_unique</th> <th>outliers_list</th> </tr> </thead> <tbody> <tr> <td>col</td> <td>0</td> <td>0</td> <td>[]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	col	0	0	[]		
	outliers_count	outliers_unique	outliers_list								
col	0	0	[]								

Répartition	<table border="1"> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>col</td><td></td><td></td></tr> <tr> <td>-1.0</td><td>1600</td><td>0.0</td></tr> <tr> <td>1.0</td><td>115616</td><td>10.0</td></tr> <tr> <td>2.0</td><td>140178</td><td>12.0</td></tr> <tr> <td>3.0</td><td>341612</td><td>29.0</td></tr> <tr> <td>4.0</td><td>37885</td><td>3.0</td></tr> <tr> <td>5.0</td><td>37244</td><td>3.0</td></tr> <tr> <td>6.0</td><td>381967</td><td>32.0</td></tr> <tr> <td>7.0</td><td>120752</td><td>10.0</td></tr> <tr> <td>NaN</td><td>19</td><td>0.0</td></tr> </tbody> </table> <p>Years: 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022</p>		Count	% valeurs	col			-1.0	1600	0.0	1.0	115616	10.0	2.0	140178	12.0	3.0	341612	29.0	4.0	37885	3.0	5.0	37244	3.0	6.0	381967	32.0	7.0	120752	10.0	NaN	19	0.0
	Count	% valeurs																																
col																																		
-1.0	1600	0.0																																
1.0	115616	10.0																																
2.0	140178	12.0																																
3.0	341612	29.0																																
4.0	37885	3.0																																
5.0	37244	3.0																																
6.0	381967	32.0																																
7.0	120752	10.0																																
NaN	19	0.0																																
Evolution	<p>Modalités: 1, 2, 3, 4, 5, 6, 7, -1</p>																																	
Remarque	<p>Les NaN peuvent être remplacés par la valeur -1 qui signifie « non renseigné ».</p> <p>A noter, une augmentation des modalités non-renseignées en 2020 (covid).</p>																																	

I. com

Description	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.										
Type	[2005] : float64 [2006-2018] : int64 [2019-2022] : object										
Etendue des valeurs	<table border="1"> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>com</td> <td>1176871</td> <td>23037</td> <td>55</td> <td>33591</td> </tr> </tbody> </table>		count	unique	top	freq	com	1176871	23037	55	33591
	count	unique	top	freq							
com	1176871	23037	55	33591							

Valeurs nulles	Type	Val_notnull	Val_null	%_null
	com	object	1176871	2 0.0
Outliers	outliers_count	outliers_unique		outliers_list
	com	1130160	23036	[0.0, 01001, 01004, 01005, 01007, 01008, 0101...
	<p style="text-align: center;">Boxplots pour: com</p>			
Répartition	Count	% valeurs		
	com			
	0.0	1	0.0	
	01001	4	0.0	
	01004	26	0.0	
	01005	3	0.0	
	01007	7	0.0	
	
	98833	3	0.0	
	99	1818	0.0	
	99.0	200	0.0	
	N/C	1	0.0	
	NaN	2	0.0	
	23038 rows × 2 columns			
Remarque	Les codes « communes » ne sont pas enregistrés sous le même format, et présence de valeurs non pertinentes telles que N/C, 0 ou NaN (voir si le code département permet sa reconstitution, puisqu'à partir de 2019, la variable change d'aspect pour répondre à celui de la description).			

m.adr

Description	Adresse postale : variable renseignée pour les accidents survenus en agglomération.																																								
Type	object																																								
Etendue des valeurs	count unique top freq adr 1032364 483819 AUTOROUTE A86 4268																																								
Valeurs nulles	Type Val_notnull Val_null %_null adr object 1032364 144509 12.28																																								
Outliers	outliers_count outliers_unique outliers_list adr 1032364 483819 [A64, (Bd FELIX mercarder, (CAMP MAJOR), ...																																								
Répartition	Count % valeurs <table> <thead> <tr> <th>adr</th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>A64</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>(Bd FELIX mercarder</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>(CAMP MAJOR)</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>(ROUTE DE DIEPPE)</td> <td>2</td> <td>0.0</td> </tr> <tr> <td>(nouvelle rocade)</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>...</td> <td>...</td> <td>...</td> </tr> <tr> <td>île HTR DU PALAIS DU MAROC</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>île hauteur de Vilormel</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>île proximité RD1075/50A</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>ôté droit dans le sens</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>NaN</td> <td>144509</td> <td>12.0</td> </tr> </tbody> </table> <p>483820 rows × 2 columns</p>					adr	Count	% valeurs	A64	1	0.0	(Bd FELIX mercarder	1	0.0	(CAMP MAJOR)	1	0.0	(ROUTE DE DIEPPE)	2	0.0	(nouvelle rocade)	1	0.0	île HTR DU PALAIS DU MAROC	1	0.0	île hauteur de Vilormel	1	0.0	île proximité RD1075/50A	1	0.0	ôté droit dans le sens	1	0.0	NaN	144509	12.0
adr	Count	% valeurs																																							
A64	1	0.0																																							
(Bd FELIX mercarder	1	0.0																																							
(CAMP MAJOR)	1	0.0																																							
(ROUTE DE DIEPPE)	2	0.0																																							
(nouvelle rocade)	1	0.0																																							
...																																							
île HTR DU PALAIS DU MAROC	1	0.0																																							
île hauteur de Vilormel	1	0.0																																							
île proximité RD1075/50A	1	0.0																																							
ôté droit dans le sens	1	0.0																																							
NaN	144509	12.0																																							
Remarque	Cette variable s'avère non pertinente en raison de sa dispersion.																																								

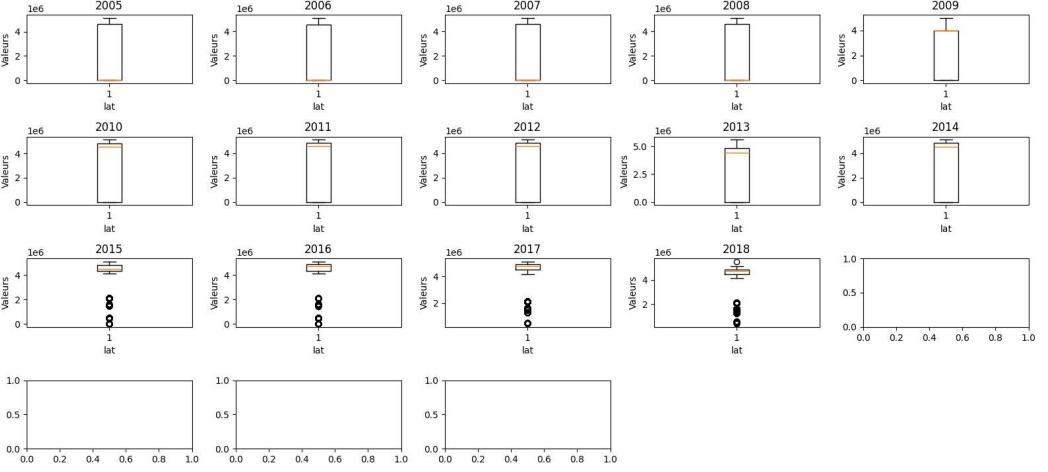
n.gps

Description	Codage GPS : 1 caractère indicateur de provenance : M = Métropole A = Antilles (Martinique ou Guadeloupe) G = Guyane R = Réunion Y = Mayotte
-------------	---

Type	object																																				
Etendue des valeurs	count unique top freq gps 480052 10 M 462639																																				
Valeurs nulles	Type Val_notnull Val_null %_null gps object 480052 696821 59.21																																				
Outliers	outliers_count outliers_unique outliers_list gps 17413 9 [0, A, C, G, P, R, S, T, Y]																																				
Répartition	<p>Count % valeurs</p> <table> <thead> <tr> <th>gps</th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>9</td> <td>0.0</td> </tr> <tr> <td>A</td> <td>7850</td> <td>1.0</td> </tr> <tr> <td>C</td> <td>2</td> <td>0.0</td> </tr> <tr> <td>G</td> <td>3387</td> <td>0.0</td> </tr> <tr> <td>M</td> <td>462639</td> <td>39.0</td> </tr> <tr> <td>P</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>R</td> <td>5316</td> <td>0.0</td> </tr> <tr> <td>S</td> <td>4</td> <td>0.0</td> </tr> <tr> <td>T</td> <td>2</td> <td>0.0</td> </tr> <tr> <td>Y</td> <td>842</td> <td>0.0</td> </tr> <tr> <td>Nan</td> <td>696821</td> <td>59.0</td> </tr> </tbody> </table> <p>Count</p> <p>Modalités</p> <p>Years</p> <ul style="list-style-type: none"> 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 	gps	Count	% valeurs	0	9	0.0	A	7850	1.0	C	2	0.0	G	3387	0.0	M	462639	39.0	P	1	0.0	R	5316	0.0	S	4	0.0	T	2	0.0	Y	842	0.0	Nan	696821	59.0
gps	Count	% valeurs																																			
0	9	0.0																																			
A	7850	1.0																																			
C	2	0.0																																			
G	3387	0.0																																			
M	462639	39.0																																			
P	1	0.0																																			
R	5316	0.0																																			
S	4	0.0																																			
T	2	0.0																																			
Y	842	0.0																																			
Nan	696821	59.0																																			
Evolution	<p>Evolution de la distribution gps</p> <p>Proportion (%)</p> <p>Années</p> <p>Modalités</p> <ul style="list-style-type: none"> S C G P A M O T R Y 																																				

Remarque	Cette variable présente un grand nombre de valeurs NaN, notamment depuis sa disparition en 2019, ce qui peut la rendre non pertinente.
-----------------	--

o. lat

Description	Latitude.				
Type	[2005-2018] : float64 [2019-2022] : object				
Etendue des valeurs	count unique top freq lat 689805 379834 0.0 117839				
Valeurs nulles	Type Val_notnull Val_null %_null lat object 689805 487068 41.39				
Outliers	outliers_count outliers_unique lat 549485 379832 [-12,6853290000, -12,6894530, -12,69219500... Boxplots pour: lat				
					

Répartition	Count % valeurs		
	lat		
	-12,6853290000	1	0.0
	-12,6894530	1	0.0
	-12,6921950000	1	0.0
	-12,7031220000	1	0.0
	-12,7044830	1	0.0

	942686.0	1	0.0
	944429.0	1	0.0
379835 rows × 2 columns			
Remarque	Le format des coordonnées ne semble pas uniforme avec un taux élevé de valeurs NaN.		

p. long

Description	Longitude			
Type	[2005-2008 ; 2010-2018] : float64 [2009 ; 2019-2022] : object			
Etendue des valeurs	count unique top freq long 689801 415250 0.0 107376			
Valeurs nulles	Type Val_notnull Val_null %_null long object 689801 487072 41.39			
Outliers	outliers_count outliers_unique outliers_list long 552144 415248 [-0,0003420000, -0,0004390, -0,0005150000,...]			

Répartition		Count	% valeurs
long			
	-0,0003420000	1	0.0
	-0,0004390	1	0.0
	-0,0005150000	1	0.0
	-0,0006440000	1	0.0
	-0,0012150	1	0.0

	9998.0	1	0.0
	99980.0	1	0.0
	99984.0	1	0.0
	99999.0	2	0.0
	NaN	487072	41.0
415251 rows × 2 columns			
Remarque	Le format des coordonnées ne semblent pas uniformes avec un taux élevé de valeurs NaN.		

q. dep

Description	Département : Code INSEE du département (2A Corse-du-Sud – 2B Haute-Corse).										
Type	[2005-2018] : int64 [2019-2022] : object										
Etendue des valeurs	<table> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>dep</td> <td>1176873</td> <td>204</td> <td>750</td> <td>99021</td> </tr> </tbody> </table>		count	unique	top	freq	dep	1176873	204	750	99021
	count	unique	top	freq							
dep	1176873	204	750	99021							
Valeurs nulles	<table> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>dep</td> <td>object</td> <td>1176873</td> <td>0</td> <td>0.0</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	dep	object	1176873	0	0.0
	Type	Val_notnull	Val_null	%_null							
dep	object	1176873	0	0.0							
Outliers	<table> <thead> <tr> <th></th> <th>outliers_count</th> <th>outliers_unique</th> <th>outliers_list</th> </tr> </thead> <tbody> <tr> <td>dep</td> <td>575077</td> <td>180</td> <td>[01, 02, 03, 04, 05, 06, 07, 08, 09, 1, 10, 1...</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	dep	575077	180	[01, 02, 03, 04, 05, 06, 07, 08, 09, 1, 10, 1...		
	outliers_count	outliers_unique	outliers_list								
dep	575077	180	[01, 02, 03, 04, 05, 06, 07, 08, 09, 1, 10, 1...								

Répartition	Count % valeurs	
	dep	
01	1291	0.0
02	625	0.0
03	694	0.0
04	605	0.0
05	700	0.0
...
977	58	0.0
978	125	0.0
986	46	0.0
987	539	0.0
988	1220	0.0

204 rows × 2 columns

2. Lieux

Rows x columns Rows duplicated

Lieux (1176873, 19) 0

a. Num_Acc

Description	Identifiant de l'accident identique à celui du fichier "rubrique CARACTERISTIQUES" repris dans l'accident.														
Type	int64														
Etendue des valeurs	<table> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td>1176873</td><td>1176873</td><td>200500000001</td><td>1</td></tr> </tbody> </table>						count	unique	top	freq	Num_Acc	1176873	1176873	200500000001	1
	count	unique	top	freq											
Num_Acc	1176873	1176873	200500000001	1											
Valeurs nulles	<table> <thead> <tr> <th></th><th>Type</th><th>Val_notnull</th><th>Val_null</th><th>%_null</th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td>int64</td><td>1176873</td><td>0</td><td>0.0</td></tr> </tbody> </table>						Type	Val_notnull	Val_null	%_null	Num_Acc	int64	1176873	0	0.0
	Type	Val_notnull	Val_null	%_null											
Num_Acc	int64	1176873	0	0.0											
Outliers	<table> <thead> <tr> <th></th><th>outliers_count</th><th>outliers_unique</th><th>outliers_list</th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td>0</td><td>0</td><td>[]</td></tr> </tbody> </table>						outliers_count	outliers_unique	outliers_list	Num_Acc	0	0	[]		
	outliers_count	outliers_unique	outliers_list												
Num_Acc	0	0	[]												

Répartition	Count % valeurs		
	Num_Acc		
200500000001	1	0.0	
200500000002	1	0.0	
200500000003	1	0.0	
200500000004	1	0.0	
200500000005	1	0.0	
...	
202200055298	1	0.0	
202200055299	1	0.0	
202200055300	1	0.0	
202200055301	1	0.0	
202200055302	1	0.0	
1176873 rows × 2 columns			

b. catr

Description	Catégorie de route.										
Modalités	<ul style="list-style-type: none"> - 1 : Autoroute - 2 : Route nationale - 3 : Route Départementale - 4 : Voie Communale - 5 : Hors réseau public - 6 : Parc de stationnement ouvert à la circulation publique - 7 : Routes de métropole urbaine - 9 : Autre 										
Type	[2005] : float64 [2006-2022] : int64										
Etendue des valeurs	<table> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>catr</td> <td>1176872</td> <td>8</td> <td>4.0</td> <td>571872</td> </tr> </tbody> </table>		count	unique	top	freq	catr	1176872	8	4.0	571872
	count	unique	top	freq							
catr	1176872	8	4.0	571872							
Valeurs nulles	<table> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>catr</td> <td>float64</td> <td>1176872</td> <td>1</td> <td>0.0</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	catr	float64	1176872	1	0.0
	Type	Val_notnull	Val_null	%_null							
catr	float64	1176872	1	0.0							
Outliers	<table> <thead> <tr> <th></th> <th>outliers_count</th> <th>outliers_unique</th> <th>outliers_list</th> </tr> </thead> <tbody> <tr> <td>catr</td> <td>127334</td> <td>4</td> <td>[1.0, 6.0, 7.0, 9.0]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	catr	127334	4	[1.0, 6.0, 7.0, 9.0]		
	outliers_count	outliers_unique	outliers_list								
catr	127334	4	[1.0, 6.0, 7.0, 9.0]								

Répartition	<table border="1"> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>catr</td><td></td><td></td></tr> <tr> <td>1.0</td><td>93089</td><td>8.0</td></tr> <tr> <td>2.0</td><td>89780</td><td>8.0</td></tr> <tr> <td>3.0</td><td>385809</td><td>33.0</td></tr> <tr> <td>4.0</td><td>571872</td><td>49.0</td></tr> <tr> <td>5.0</td><td>2077</td><td>0.0</td></tr> <tr> <td>6.0</td><td>7982</td><td>1.0</td></tr> <tr> <td>7.0</td><td>7275</td><td>1.0</td></tr> <tr> <td>9.0</td><td>18988</td><td>2.0</td></tr> <tr> <td>NaN</td><td>1</td><td>0.0</td></tr> </tbody> </table>		Count	% valeurs	catr			1.0	93089	8.0	2.0	89780	8.0	3.0	385809	33.0	4.0	571872	49.0	5.0	2077	0.0	6.0	7982	1.0	7.0	7275	1.0	9.0	18988	2.0	NaN	1	0.0
	Count	% valeurs																																
catr																																		
1.0	93089	8.0																																
2.0	89780	8.0																																
3.0	385809	33.0																																
4.0	571872	49.0																																
5.0	2077	0.0																																
6.0	7982	1.0																																
7.0	7275	1.0																																
9.0	18988	2.0																																
NaN	1	0.0																																
Evolution																																		
Remarque	La valeur NaN peut être supprimée.																																	

c. voie

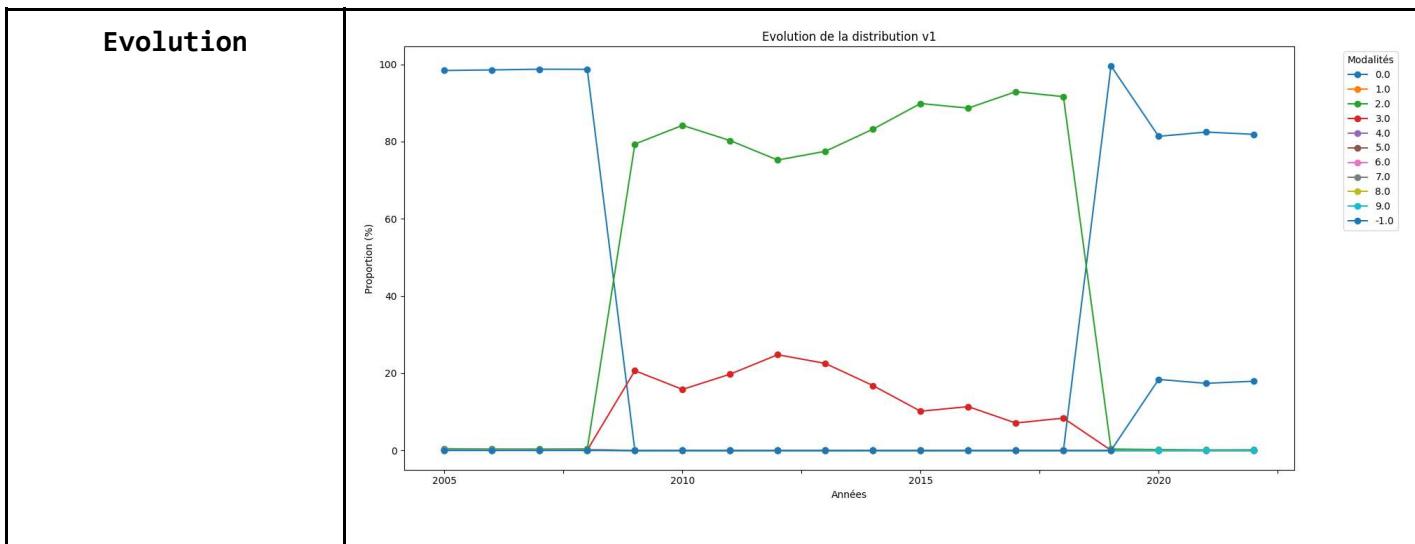
Description	Numéro de la route.										
Type	[2005-2015] : float64 [2016-2022] : object										
Etendue des valeurs	<table> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>voie</td> <td>1064888</td> <td>38865</td> <td>0.0</td> <td>429071</td> </tr> </tbody> </table>		count	unique	top	freq	voie	1064888	38865	0.0	429071
	count	unique	top	freq							
voie	1064888	38865	0.0	429071							
Valeurs nulles	<table> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>voie</td> <td>object</td> <td>1064888</td> <td>111985</td> <td>9.52</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	voie	object	1064888	111985	9.52
	Type	Val_notnull	Val_null	%_null							
voie	object	1064888	111985	9.52							

Outliers	<p>outliers_count outliers_unique outliers_list</p> <p>voie 563558 38862 [(R), ...</p> <p>Boxplots pour: voie</p>																																				
Répartition	<p>Count % valeurs</p> <table border="1"> <thead> <tr> <th>voie</th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>(R)</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>(AV)</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>(AV)</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>(AV)</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>(BD)</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>...</td> <td>...</td> <td>...</td> </tr> <tr> <td>x</td> <td>3</td> <td>0.0</td> </tr> <tr> <td>xxxx</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>ÉCHANGEUR DU RONDEAU</td> <td>2</td> <td>0.0</td> </tr> <tr> <td>Épalle</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>Nan</td> <td>111985</td> <td>10.0</td> </tr> </tbody> </table> <p>38866 rows × 2 columns</p>	voie	Count	% valeurs	(R)	1	0.0	(AV)	1	0.0	(AV)	1	0.0	(AV)	1	0.0	(BD)	1	0.0	x	3	0.0	xxxx	1	0.0	ÉCHANGEUR DU RONDEAU	2	0.0	Épalle	1	0.0	Nan	111985	10.0
voie	Count	% valeurs																																			
(R)	1	0.0																																			
(AV)	1	0.0																																			
(AV)	1	0.0																																			
(AV)	1	0.0																																			
(BD)	1	0.0																																			
...																																			
x	3	0.0																																			
xxxx	1	0.0																																			
ÉCHANGEUR DU RONDEAU	2	0.0																																			
Épalle	1	0.0																																			
Nan	111985	10.0																																			
Remarque	Les valeurs NaN sont élevées, ce qui peut rendre la variable non pertinente.																																				

d. v1

Description	Indice numérique du numéro de route (exemple : 2 bis, 3 ter etc.).
--------------------	--

Type	[2020-2022] : int64 [2005-2019] : float64																										
Etendue des valeurs	count unique top freq v1 541049 11 0.0 504288																										
Valeurs nulles	Type Val_notnull Val_null %_null v1 float64 541049 635824 54.03																										
Outliers	outliers_count outliers_unique outliers_list v1 36761 10 [-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]																										
Répartition	<table border="1"> <thead> <tr> <th>Count % valeurs</th> <th>v1</th> </tr> </thead> <tbody> <tr> <td>-1.0 28529 2.0</td> <td></td> </tr> <tr> <td>0.0 504288 43.0</td> <td></td> </tr> <tr> <td>1.0 1020 0.0</td> <td></td> </tr> <tr> <td>2.0 4302 0.0</td> <td></td> </tr> <tr> <td>3.0 887 0.0</td> <td></td> </tr> <tr> <td>4.0 396 0.0</td> <td></td> </tr> <tr> <td>5.0 232 0.0</td> <td></td> </tr> <tr> <td>6.0 461 0.0</td> <td></td> </tr> <tr> <td>7.0 392 0.0</td> <td></td> </tr> <tr> <td>8.0 247 0.0</td> <td></td> </tr> <tr> <td>9.0 295 0.0</td> <td></td> </tr> <tr> <td>Nan 635824 54.0</td> <td></td> </tr> </tbody> </table> <p>The chart displays the count of occurrences for each value of v1 across different years. The x-axis represents the value of v1, ranging from -1.0 to 10. The y-axis represents the count of occurrences, ranging from 0 to 500,000. The bars are stacked by year, with colors corresponding to the legend on the right. The legend lists years from 2005 to 2022, with each year having a unique color. The distribution is heavily skewed towards 0.0, with a significant portion of the data being null (NaN).</p>	Count % valeurs	v1	-1.0 28529 2.0		0.0 504288 43.0		1.0 1020 0.0		2.0 4302 0.0		3.0 887 0.0		4.0 396 0.0		5.0 232 0.0		6.0 461 0.0		7.0 392 0.0		8.0 247 0.0		9.0 295 0.0		Nan 635824 54.0	
Count % valeurs	v1																										
-1.0 28529 2.0																											
0.0 504288 43.0																											
1.0 1020 0.0																											
2.0 4302 0.0																											
3.0 887 0.0																											
4.0 396 0.0																											
5.0 232 0.0																											
6.0 461 0.0																											
7.0 392 0.0																											
8.0 247 0.0																											
9.0 295 0.0																											
Nan 635824 54.0																											



e. v2

Description	Lettre indice alphanumérique de la route.
Type	object
Etendue des valeurs	count unique top freq v2 56624 74 A 24722
Valeurs nulles	Type Val_notnull Val_null %_null v2 object 56624 1120249 95.19
Outliers	outliers_count outliers_unique outliers_list v2 18298 72 [□, -, D, ., 0, 1, 15, 1A, 2, 3, 34, 4, 5, ...

Répartition	Count % valeurs		
v2			
	□	537	0.0
	-	246	0.0
	D	1	0.0
	.	1	0.0
	0	1099	0.0

	v	2	0.0
	w	2	0.0
	y	3	0.0
	z	19	0.0
	NaN	1120249	95.0
75 rows × 2 columns			
Remarque	Les informations rassemblées sont parfois de type incohérent, d'autant plus qu'il y a un grand nombre de valeurs NaN.		

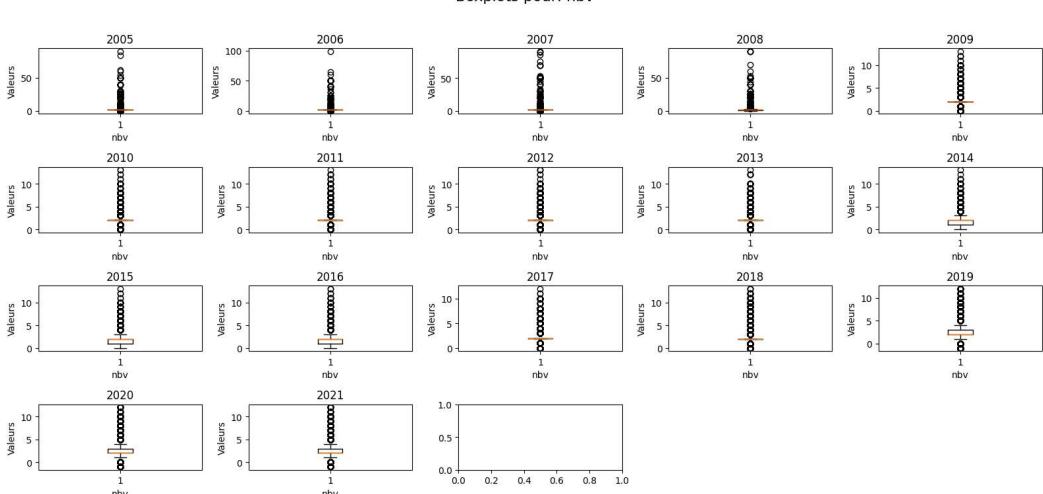
f. circ

Description	Régime de circulation										
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 1 : A sens unique - 2 : Bidirectionnelle - 3 : A chaussées séparées - 4 : Avec voies d'affectation variable 										
Type	int64										
Etendue des valeurs	<table> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>circ</td><td>1175299</td><td>6</td><td>2.0</td><td>741823</td></tr> </tbody> </table>		count	unique	top	freq	circ	1175299	6	2.0	741823
	count	unique	top	freq							
circ	1175299	6	2.0	741823							
Valeurs nulles	<table> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>circ</td><td>float64</td><td>1175299</td><td>1574</td><td>0.13</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	circ	float64	1175299	1574	0.13
	Type	Val_notnull	Val_null	%_null							
circ	float64	1175299	1574	0.13							
Outliers	<table> <thead> <tr> <th></th> <th>outliers_count</th> <th>outliers_unique</th> <th>outliers_list</th> </tr> </thead> <tbody> <tr> <td>circ</td><td>433476</td><td>5</td><td>[-1.0, 0.0, 1.0, 3.0, 4.0]</td></tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	circ	433476	5	[-1.0, 0.0, 1.0, 3.0, 4.0]		
	outliers_count	outliers_unique	outliers_list								
circ	433476	5	[-1.0, 0.0, 1.0, 3.0, 4.0]								

Répartition	<p>Count % valeurs</p> <table border="1"> <thead> <tr> <th>circ</th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>-1.0</td><td>12184</td><td>1.0</td></tr> <tr> <td>0.0</td><td>49966</td><td>4.0</td></tr> <tr> <td>1.0</td><td>208883</td><td>18.0</td></tr> <tr> <td>2.0</td><td>741823</td><td>63.0</td></tr> <tr> <td>3.0</td><td>155773</td><td>13.0</td></tr> <tr> <td>4.0</td><td>6670</td><td>1.0</td></tr> <tr> <td>NaN</td><td>1574</td><td>0.0</td></tr> </tbody> </table>	circ	Count	% valeurs	-1.0	12184	1.0	0.0	49966	4.0	1.0	208883	18.0	2.0	741823	63.0	3.0	155773	13.0	4.0	6670	1.0	NaN	1574	0.0
circ	Count	% valeurs																							
-1.0	12184	1.0																							
0.0	49966	4.0																							
1.0	208883	18.0																							
2.0	741823	63.0																							
3.0	155773	13.0																							
4.0	6670	1.0																							
NaN	1574	0.0																							
Evolution																									
Remarque	Les NaN peuvent être remplacés par la valeur -1 qui signifie « non renseigné ».																								

g. nbv

Description	Nombre total de voies de circulation.										
Type	[2005-2008 ; 2019-2021] : int64 [2009-2018] : float64 [2022] : object										
Etendue des valeurs	<table border="1"> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>nbv</td> <td>1174142</td> <td>70</td> <td>2.0</td> <td>351510</td> </tr> </tbody> </table>		count	unique	top	freq	nbv	1174142	70	2.0	351510
	count	unique	top	freq							
nbv	1174142	70	2.0	351510							
Valeurs nulles	<table border="1"> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>nbv</td> <td>object</td> <td>1174142</td> <td>2731</td> <td>0.23</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	nbv	object	1174142	2731	0.23
	Type	Val_notnull	Val_null	%_null							
nbv	object	1174142	2731	0.23							

Outliers	outliers_count outliers_unique outliers_list nbv 49288 61 [-1, #ERREUR, -1, 10, 10.0, 11, 11.0, 12, 12... 																																				
Répartition	Count % valeurs nbv <table border="1"> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>-1</td><td>561</td><td>0.0</td></tr> <tr> <td>#ERREUR</td><td>1</td><td>0.0</td></tr> <tr> <td>-1</td><td>1669</td><td>0.0</td></tr> <tr> <td>0</td><td>46729</td><td>4.0</td></tr> <tr> <td>0.0</td><td>67398</td><td>6.0</td></tr> <tr> <td>...</td><td>...</td><td>...</td></tr> <tr> <td>9.0</td><td>175</td><td>0.0</td></tr> <tr> <td>90</td><td>7</td><td>0.0</td></tr> <tr> <td>91</td><td>1</td><td>0.0</td></tr> <tr> <td>99</td><td>1</td><td>0.0</td></tr> <tr> <td>NaN</td><td>2731</td><td>0.0</td></tr> </tbody> </table> <p>71 rows × 2 columns</p>		Count	% valeurs	-1	561	0.0	#ERREUR	1	0.0	-1	1669	0.0	0	46729	4.0	0.0	67398	6.0	9.0	175	0.0	90	7	0.0	91	1	0.0	99	1	0.0	NaN	2731	0.0
	Count	% valeurs																																			
-1	561	0.0																																			
#ERREUR	1	0.0																																			
-1	1669	0.0																																			
0	46729	4.0																																			
0.0	67398	6.0																																			
...																																			
9.0	175	0.0																																			
90	7	0.0																																			
91	1	0.0																																			
99	1	0.0																																			
NaN	2731	0.0																																			
Remarque	Lorsqu'elles ne représentent pas des NaN, certaines valeurs paraissent aberrantes (on dépasse parfois les 50 voies de circulation).																																				

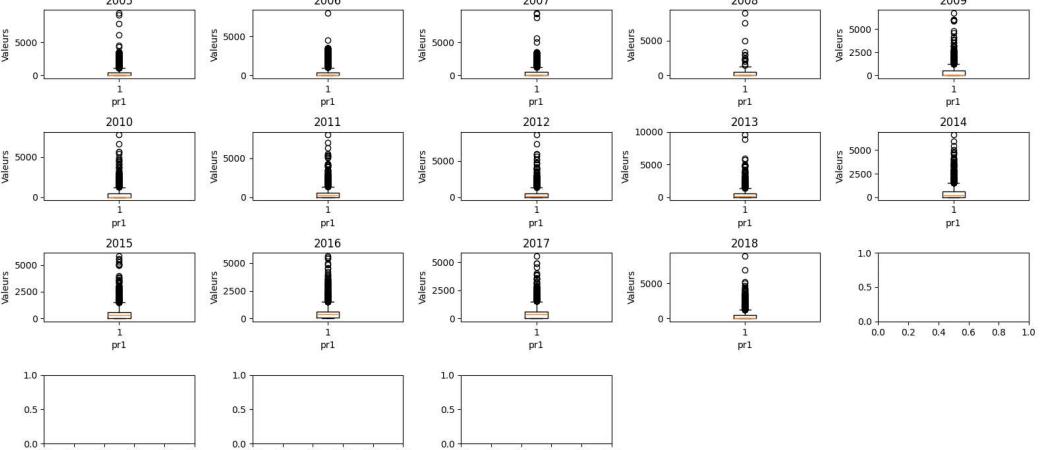
h. pr

Description	Numéro du PR de rattachement (numéro de la borne amont). La valeur -
--------------------	--

	1 signifie que le PR n'est pas renseigné.																																	
Type	[2005-2018] : float64 [2019-2022] : object																																	
Etendue des valeurs	count unique top freq <hr/> pr 701389 1413 0.0 150037																																	
Valeurs nulles	Type Val_notnull Val_null %_null <hr/> pr object 701389 475484 40.4																																	
Outliers	outliers_count outliers_unique outliers_list <hr/> pr 349171 1406 [0.01, 10, 10.0, 10.2, 10.5, 100, 100.0, 1000...																																	
Répartition	Count % valeurs <hr/> pr <table> <tbody> <tr><td>(1)</td><td>54073</td><td>5.0</td></tr> <tr><td>0</td><td>68855</td><td>6.0</td></tr> <tr><td>0.0</td><td>150037</td><td>13.0</td></tr> <tr><td>0.01</td><td>1</td><td>0.0</td></tr> <tr><td>1</td><td>13485</td><td>1.0</td></tr> <tr><td>...</td><td>...</td><td>...</td></tr> <tr><td>9900.0</td><td>1</td><td>0.0</td></tr> <tr><td>992</td><td>1</td><td>0.0</td></tr> <tr><td>9929.0</td><td>1</td><td>0.0</td></tr> <tr><td>999</td><td>19</td><td>0.0</td></tr> <tr><td>NaN</td><td>475484</td><td>40.0</td></tr> </tbody> </table> <p>1414 rows × 2 columns</p>	(1)	54073	5.0	0	68855	6.0	0.0	150037	13.0	0.01	1	0.0	1	13485	1.0	9900.0	1	0.0	992	1	0.0	9929.0	1	0.0	999	19	0.0	NaN	475484	40.0
(1)	54073	5.0																																
0	68855	6.0																																
0.0	150037	13.0																																
0.01	1	0.0																																
1	13485	1.0																																
...																																
9900.0	1	0.0																																
992	1	0.0																																
9929.0	1	0.0																																
999	19	0.0																																
NaN	475484	40.0																																
Remarque	Les valeurs semblent parfois incohérentes, et un grand nombre de NaN prédominent.																																	

i. pr1

Description	Distance en mètres au PR (par rapport à la borne amont). La valeur -1 signifie que le PR n'est pas renseigné.
Type	[2005-2018] : float64 [2019-2022] : object

Etendue des valeurs	count unique top freq pr1 699570 3708 0.0 198026																																				
Valeurs nulles	Type Val_notnull Val_null %_null pr1 object 699570 477303 40.56																																				
Outliers	outliers_count outliers_unique outliers_list pr1 223388 3696 [1, 1.0, 10, 10.0, 100, 1000, 1000.0, 1001, 1... Boxplots pour: pr1 																																				
Répartition	Count % valeurs pr1 <table border="1"> <thead> <tr> <th></th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>(1)</td> <td>54819</td> <td>5.0</td> </tr> <tr> <td>0</td> <td>75650</td> <td>6.0</td> </tr> <tr> <td>0.0</td> <td>198026</td> <td>17.0</td> </tr> <tr> <td>1</td> <td>6950</td> <td>1.0</td> </tr> <tr> <td>1.0</td> <td>9031</td> <td>1.0</td> </tr> <tr> <td>...</td> <td>...</td> <td>...</td> </tr> <tr> <td>998</td> <td>8</td> <td>0.0</td> </tr> <tr> <td>998.0</td> <td>41</td> <td>0.0</td> </tr> <tr> <td>999</td> <td>24</td> <td>0.0</td> </tr> <tr> <td>999.0</td> <td>258</td> <td>0.0</td> </tr> <tr> <td>NaN</td> <td>477303</td> <td>41.0</td> </tr> </tbody> </table> <p>3709 rows × 2 columns</p>		Count	% valeurs	(1)	54819	5.0	0	75650	6.0	0.0	198026	17.0	1	6950	1.0	1.0	9031	1.0	998	8	0.0	998.0	41	0.0	999	24	0.0	999.0	258	0.0	NaN	477303	41.0
	Count	% valeurs																																			
(1)	54819	5.0																																			
0	75650	6.0																																			
0.0	198026	17.0																																			
1	6950	1.0																																			
1.0	9031	1.0																																			
...																																			
998	8	0.0																																			
998.0	41	0.0																																			
999	24	0.0																																			
999.0	258	0.0																																			
NaN	477303	41.0																																			

Remarque	Les valeurs semblent parfois incohérentes, et un grand nombre de NaN prédominent.
-----------------	---

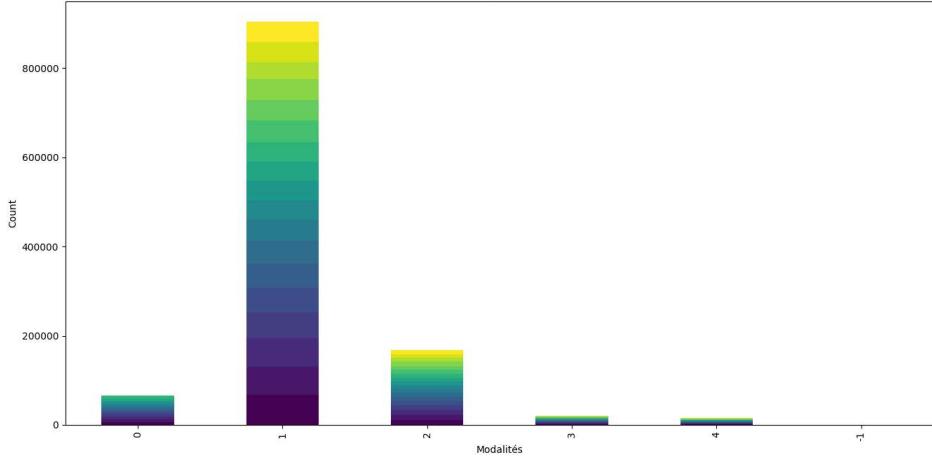
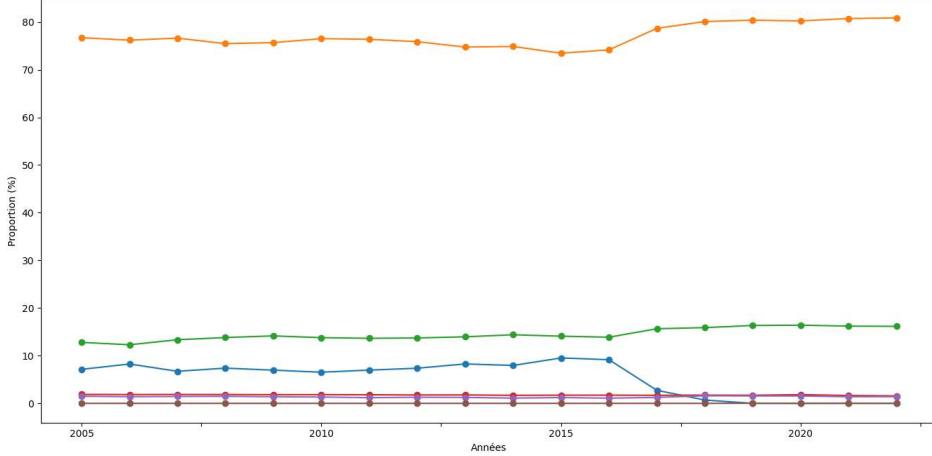
j. vosp

Description	Signale l'existence d'une voie réservée, indépendamment du fait que l'accident ait lieu ou non sur cette voie.																								
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 0 : Sans objet - 1 : Piste cyclable - 2 : Bande cyclable - 3 : Voie réservée 																								
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64																								
Etendue des valeurs	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">count</th> <th style="text-align: center;">unique</th> <th style="text-align: center;">top</th> <th style="text-align: center;">freq</th> </tr> </thead> <tbody> <tr> <td>vosp</td> <td style="text-align: center;">1174112</td> <td style="text-align: center;">5</td> <td style="text-align: center;">0.0</td> <td style="text-align: center;">1090745</td> </tr> </tbody> </table>		count	unique	top	freq	vosp	1174112	5	0.0	1090745														
	count	unique	top	freq																					
vosp	1174112	5	0.0	1090745																					
Valeurs nulles	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">Type</th> <th style="text-align: center;">Val_notnull</th> <th style="text-align: center;">Val_null</th> <th style="text-align: center;">%_null</th> </tr> </thead> <tbody> <tr> <td>vosp</td> <td style="text-align: center;">float64</td> <td style="text-align: center;">1174112</td> <td style="text-align: center;">2761</td> <td style="text-align: center;">0.23</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	vosp	float64	1174112	2761	0.23														
	Type	Val_notnull	Val_null	%_null																					
vosp	float64	1174112	2761	0.23																					
Outliers	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">outliers_count</th> <th style="text-align: center;">outliers_unique</th> <th style="text-align: center;">outliers_list</th> </tr> </thead> <tbody> <tr> <td>vosp</td> <td style="text-align: center;">83367</td> <td style="text-align: center;">4</td> <td style="text-align: center;">[-1.0, 1.0, 2.0, 3.0]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	vosp	83367	4	[-1.0, 1.0, 2.0, 3.0]																
	outliers_count	outliers_unique	outliers_list																						
vosp	83367	4	[-1.0, 1.0, 2.0, 3.0]																						
Répartition	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">Count</th> <th style="text-align: center;">% valeurs</th> </tr> </thead> <tbody> <tr> <td>vosp</td> <td></td> <td></td> </tr> <tr> <td>-1.0</td> <td style="text-align: center;">1322</td> <td style="text-align: center;">0.0</td> </tr> <tr> <td>0.0</td> <td style="text-align: center;">1090745</td> <td style="text-align: center;">93.0</td> </tr> <tr> <td>1.0</td> <td style="text-align: center;">29795</td> <td style="text-align: center;">3.0</td> </tr> <tr> <td>2.0</td> <td style="text-align: center;">17853</td> <td style="text-align: center;">2.0</td> </tr> <tr> <td>3.0</td> <td style="text-align: center;">34397</td> <td style="text-align: center;">3.0</td> </tr> <tr> <td>Nan</td> <td style="text-align: center;">2761</td> <td style="text-align: center;">0.0</td> </tr> </tbody> </table> <p>The heatmap displays the distribution of 'vosp' values across years and modalities. The y-axis represents the count of values, ranging from 0.0 to 1.0e6. The x-axis represents the modalities: 0, 1, 2, and 3. The color scale indicates the year, with 2005 being dark purple and 2022 being yellow. The highest counts are concentrated in modalité 0, with a significant peak in 2005. As the year progresses, the counts generally decrease across all modalities, with a notable presence of values in modalités 1, 2, and 3 during later years.</p>		Count	% valeurs	vosp			-1.0	1322	0.0	0.0	1090745	93.0	1.0	29795	3.0	2.0	17853	2.0	3.0	34397	3.0	Nan	2761	0.0
	Count	% valeurs																							
vosp																									
-1.0	1322	0.0																							
0.0	1090745	93.0																							
1.0	29795	3.0																							
2.0	17853	2.0																							
3.0	34397	3.0																							
Nan	2761	0.0																							

Evolution	
Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».

k. prof

Description	Profil en long décrit la déclivité de la route à l'endroit de l'accident.
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 1 : Plat - 2 : Pente - 3 : Sommet de côte - 4 : Bas de côte
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64
Etendue des valeurs	<pre>count unique top freq prof 1174924 6 1.0 904058</pre>
Valeurs nulles	<pre>Type Val_notnull Val_null %_null prof float64 1174924 1949 0.17</pre>
Outliers	<pre>outliers_count outliers_unique outliers_list prof 270866 [-1.0, 0.0, 2.0, 3.0, 4.0]</pre>

Répartition Count % valeurs prof <table border="1"> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>-1.0</td><td>38</td><td>0.0</td></tr> <tr> <td>0.0</td><td>65946</td><td>6.0</td></tr> <tr> <td>1.0</td><td>904058</td><td>77.0</td></tr> <tr> <td>2.0</td><td>168117</td><td>14.0</td></tr> <tr> <td>3.0</td><td>20807</td><td>2.0</td></tr> <tr> <td>4.0</td><td>15958</td><td>1.0</td></tr> <tr> <td>NaN</td><td>1949</td><td>0.0</td></tr> </tbody> </table>		Count	% valeurs	-1.0	38	0.0	0.0	65946	6.0	1.0	904058	77.0	2.0	168117	14.0	3.0	20807	2.0	4.0	15958	1.0	NaN	1949	0.0	
	Count	% valeurs																							
-1.0	38	0.0																							
0.0	65946	6.0																							
1.0	904058	77.0																							
2.0	168117	14.0																							
3.0	20807	2.0																							
4.0	15958	1.0																							
NaN	1949	0.0																							
Evolution																									
Remarque	0 n'apparaît pas dans la description. Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».																								

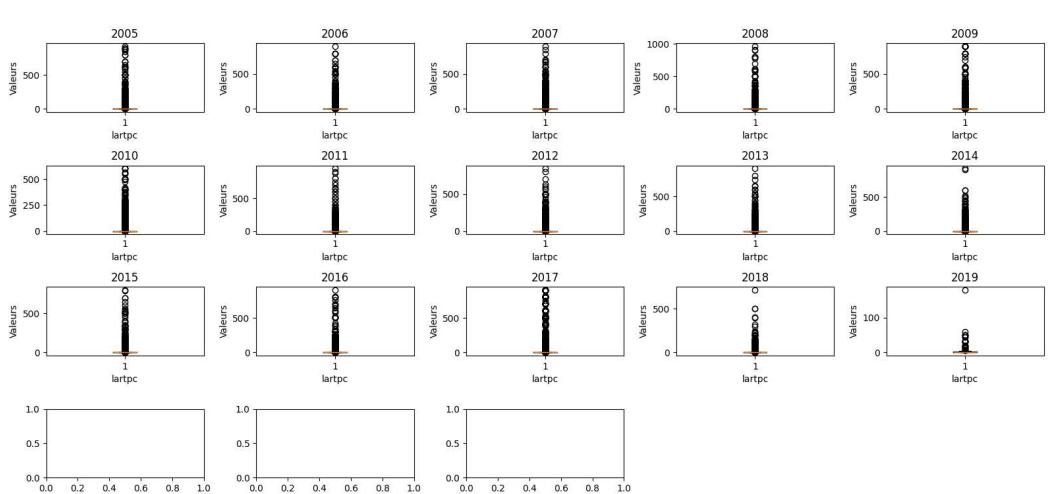
I. plan

Description	Tracé en plan.										
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 1 : Partie rectiligne - 2 : En courbe à gauche - 3 : En courbe à droite - 4 : En « S » 										
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64										
Etendue des valeurs	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">count</th> <th style="text-align: center;">unique</th> <th style="text-align: center;">top</th> <th style="text-align: center;">freq</th> </tr> </thead> <tbody> <tr> <td style="text-align: left;">plan</td> <td style="text-align: center;">1174592</td> <td style="text-align: center;">6</td> <td style="text-align: center;">1.0</td> <td style="text-align: center;">903356</td> </tr> </tbody> </table>		count	unique	top	freq	plan	1174592	6	1.0	903356
	count	unique	top	freq							
plan	1174592	6	1.0	903356							

Valeurs nulles	Type Val_notnull Val_null %_null																								
	plan float64 1174592 2281 0.19																								
Outliers	outliers_count outliers_unique outliers_list																								
	plan 271236 5 [-1.0, 0.0, 2.0, 3.0, 4.0]																								
Répartition	<p>Count % valeurs</p> <table border="1"> <thead> <tr> <th>plan</th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>-1.0</td> <td>34</td> <td>0.0</td> </tr> <tr> <td>0.0</td> <td>66417</td> <td>6.0</td> </tr> <tr> <td>1.0</td> <td>903356</td> <td>77.0</td> </tr> <tr> <td>2.0</td> <td>99412</td> <td>8.0</td> </tr> <tr> <td>3.0</td> <td>90257</td> <td>8.0</td> </tr> <tr> <td>4.0</td> <td>15116</td> <td>1.0</td> </tr> <tr> <td>NaN</td> <td>2281</td> <td>0.0</td> </tr> </tbody> </table> <p>Years: 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022</p>	plan	Count	% valeurs	-1.0	34	0.0	0.0	66417	6.0	1.0	903356	77.0	2.0	99412	8.0	3.0	90257	8.0	4.0	15116	1.0	NaN	2281	0.0
plan	Count	% valeurs																							
-1.0	34	0.0																							
0.0	66417	6.0																							
1.0	903356	77.0																							
2.0	99412	8.0																							
3.0	90257	8.0																							
4.0	15116	1.0																							
NaN	2281	0.0																							
Evolution	<p>Evolution de la distribution plan</p> <p>Modalités: 0, 1, 2, 3, 4, -1</p>																								
Remarque	<p>On observe une modalité 0, non répertoriée qui semble disparaître autour de 2018.</p> <p>Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».</p>																								

m.lartpc

Description	Largeur du terre-plein central (TPC) s'il existe (en m).
Type	[2005-2008] : int64 [2009-2019] : float64

	[2020-2022] : object
Etendue des valeurs	count unique top freq lartpc 902767 711 0.0 479834
Valeurs nulles	Type Val_notnull Val_null %_null lartpc object 902767 274106 23.29
Outliers	outliers_count outliers_unique outliers_list lartpc 108168 707 [0,4, 0,8, 1, 1,5, 1,6, 1,0, 1,5, 10, 10,2, 1... <p style="text-align: center;">Boxplots pour: lartpc</p> 

Répartition	Count	% valeurs
lartpc		
0	279196	24.0
0,4	1	0.0
0,8	1	0.0
0.0	479834	41.0
1	145	0.0
...
98	21	0.0
98.0	38	0.0
99	12	0.0
99.0	14	0.0
NaN	274106	23.0
712 rows × 2 columns		
Remarque	Les valeurs aberrantes sont très élevées.	

n. larrouut

Description	Largeur de la chaussée affectée à la circulation des véhicules ne sont pas compris les bandes d'arrêt d'urgence, les TPC et les places de stationnement (en m).
Type	[2009-2019] : float64 [2005-2008] : int64 [2020-2022] : object
Etendue des valeurs	count unique top freq larrouut 1064032 1138 0.0 211137
Valeurs nulles	Type Val_notnull Val_null %_null larrouut object 1064032 112841 9.59
Outliers	outliers_count outliers_unique outliers_list larrouut 417667 1126 [-81, 1, 1, 4, 1.0, 10, 10, 2, 10, 25, 10, 3, 10, ...]

Répartition	Count	% valeurs
larrout		
	-1	149480 13.0
	-81	1 0.0
	0	76818 7.0
	0.0	211137 18.0
	1	26 0.0

	990.0	4 0.0
	995	1 0.0
	999	2 0.0
	999.0	4 0.0
	NaN	112841 10.0
1139 rows × 2 columns		
Remarque	Les valeurs existantes semblent très éparpillées.	

o. surf

Description	Etat de la surface.										
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 1 : Normale - 2 : Mouillée - 3 : Flaque - 4 : Inondée - 5 : Enneigée - 6 : Boue - 7 : Verglacée - 8 : Corps gras – huile - 9 : Autre 										
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64										
Etendue des valeurs	<table> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>surf</td> <td>1174949</td> <td>11</td> <td>1.0</td> <td>921298</td> </tr> </tbody> </table>		count	unique	top	freq	surf	1174949	11	1.0	921298
	count	unique	top	freq							
surf	1174949	11	1.0	921298							
Valeurs nulles	<table> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>surf</td> <td>float64</td> <td>1174949</td> <td>1924</td> <td>0.16</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	surf	float64	1174949	1924	0.16
	Type	Val_notnull	Val_null	%_null							
surf	float64	1174949	1924	0.16							

Outliers	outliers_count outliers_unique			outliers_list																																																				
	surf	253651	10 [-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]																																																					
Répartition																																																								
Count % valeurs																																																								
surf																																																								
<table> <thead> <tr> <th></th> <th>Count</th> <th>% valeurs</th> <th></th> </tr> </thead> <tbody> <tr> <td>-1.0</td> <td>64</td> <td>0.0</td> <td></td> </tr> <tr> <td>0.0</td> <td>29139</td> <td>2.0</td> <td></td> </tr> <tr> <td>1.0</td> <td>921298</td> <td>78.0</td> <td></td> </tr> <tr> <td>2.0</td> <td>202151</td> <td>17.0</td> <td></td> </tr> <tr> <td>3.0</td> <td>1671</td> <td>0.0</td> <td></td> </tr> <tr> <td>4.0</td> <td>580</td> <td>0.0</td> <td></td> </tr> <tr> <td>5.0</td> <td>3285</td> <td>0.0</td> <td></td> </tr> <tr> <td>6.0</td> <td>701</td> <td>0.0</td> <td></td> </tr> <tr> <td>7.0</td> <td>6948</td> <td>1.0</td> <td></td> </tr> <tr> <td>8.0</td> <td>2735</td> <td>0.0</td> <td></td> </tr> <tr> <td>9.0</td> <td>6377</td> <td>1.0</td> <td></td> </tr> <tr> <td>Nan</td> <td>1924</td> <td>0.0</td> <td></td> </tr> </tbody> </table>					Count	% valeurs		-1.0	64	0.0		0.0	29139	2.0		1.0	921298	78.0		2.0	202151	17.0		3.0	1671	0.0		4.0	580	0.0		5.0	3285	0.0		6.0	701	0.0		7.0	6948	1.0		8.0	2735	0.0		9.0	6377	1.0		Nan	1924	0.0		
	Count	% valeurs																																																						
-1.0	64	0.0																																																						
0.0	29139	2.0																																																						
1.0	921298	78.0																																																						
2.0	202151	17.0																																																						
3.0	1671	0.0																																																						
4.0	580	0.0																																																						
5.0	3285	0.0																																																						
6.0	701	0.0																																																						
7.0	6948	1.0																																																						
8.0	2735	0.0																																																						
9.0	6377	1.0																																																						
Nan	1924	0.0																																																						
Evolution																																																								
Remarque	Les valeurs Nan peuvent être remplacées par -1 qui signifie « non renseigné ».																																																							

p. infra

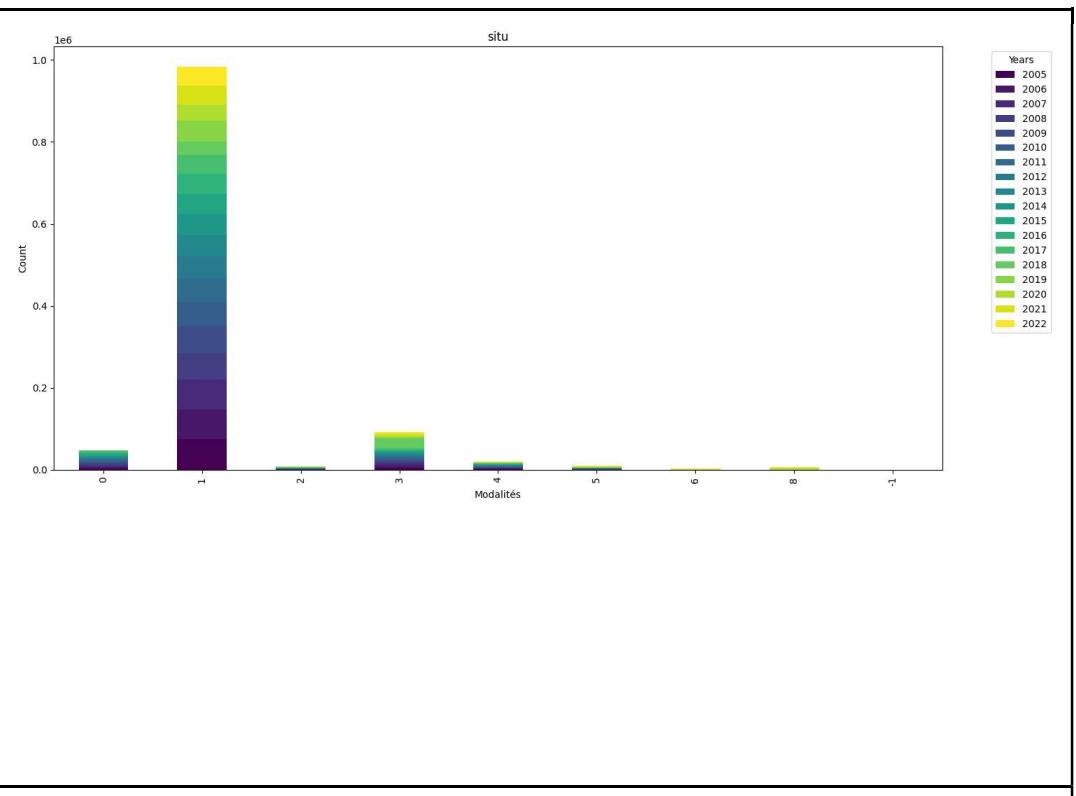
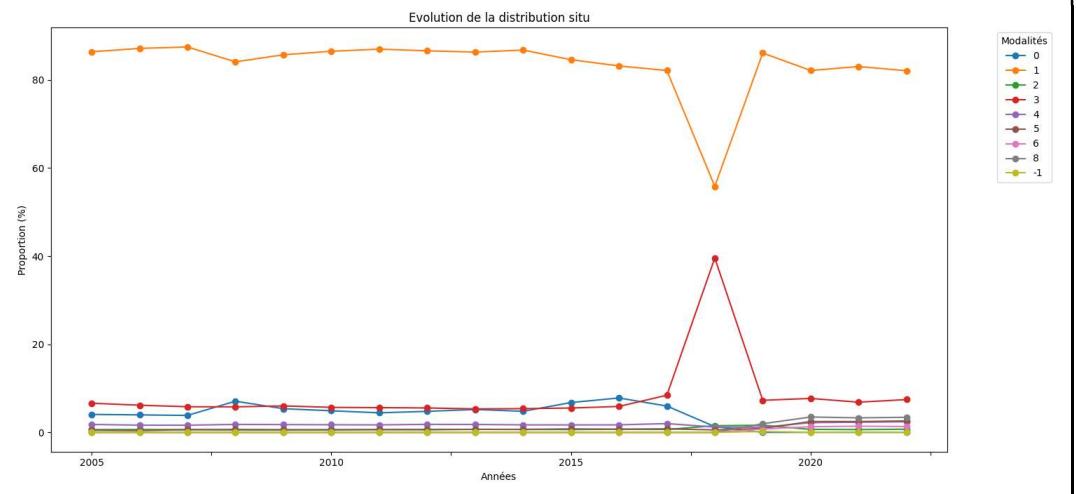
Description	Aménagement - Infrastructure.
Modalités	- -1 : Non renseigné

	<ul style="list-style-type: none"> - 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 																																										
Type	[2005-2008 ; 2019-2002] : int64 [2009-2018] : float64																																										
Etendue des valeurs	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">count</th><th style="text-align: center;">unique</th><th style="text-align: center;">top</th><th style="text-align: center;">freq</th></tr> </thead> <tbody> <tr> <td>infra</td><td style="text-align: center;">1171465</td><td style="text-align: center;">11</td><td style="text-align: center;">0.0</td><td style="text-align: center;">1032089</td></tr> </tbody> </table>		count	unique	top	freq	infra	1171465	11	0.0	1032089																																
	count	unique	top	freq																																							
infra	1171465	11	0.0	1032089																																							
Valeurs nulles	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">Type</th><th style="text-align: center;">Val_notnull</th><th style="text-align: center;">Val_null</th><th style="text-align: center;">%_null</th></tr> </thead> <tbody> <tr> <td>infra</td><td style="text-align: center;">float64</td><td style="text-align: center;">1171465</td><td style="text-align: center;">5408</td><td style="text-align: center;">0.46</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	infra	float64	1171465	5408	0.46																																
	Type	Val_notnull	Val_null	%_null																																							
infra	float64	1171465	5408	0.46																																							
Outliers	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">outliers_count</th><th style="text-align: center;">outliers_unique</th><th style="text-align: center;">outliers_list</th></tr> </thead> <tbody> <tr> <td>infra</td><td style="text-align: center;">139376</td><td style="text-align: center;">10</td><td style="text-align: center;">[-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]</td></tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	infra	139376	10	[-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]																																		
	outliers_count	outliers_unique	outliers_list																																								
infra	139376	10	[-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]																																								
Répartition <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">Count</th><th style="text-align: center;">% valeurs</th></tr> </thead> <tbody> <tr> <td>infra</td><td></td><td></td></tr> <tr> <td>-1.0</td><td style="text-align: center;">2236</td><td style="text-align: center;">0.0</td></tr> <tr> <td>0.0</td><td style="text-align: center;">1032089</td><td style="text-align: center;">88.0</td></tr> <tr> <td>1.0</td><td style="text-align: center;">10754</td><td style="text-align: center;">1.0</td></tr> <tr> <td>2.0</td><td style="text-align: center;">17283</td><td style="text-align: center;">1.0</td></tr> <tr> <td>3.0</td><td style="text-align: center;">17768</td><td style="text-align: center;">2.0</td></tr> <tr> <td>4.0</td><td style="text-align: center;">4113</td><td style="text-align: center;">0.0</td></tr> <tr> <td>5.0</td><td style="text-align: center;">68524</td><td style="text-align: center;">6.0</td></tr> <tr> <td>6.0</td><td style="text-align: center;">8329</td><td style="text-align: center;">1.0</td></tr> <tr> <td>7.0</td><td style="text-align: center;">699</td><td style="text-align: center;">0.0</td></tr> <tr> <td>8.0</td><td style="text-align: center;">1633</td><td style="text-align: center;">0.0</td></tr> <tr> <td>9.0</td><td style="text-align: center;">8037</td><td style="text-align: center;">1.0</td></tr> <tr> <td>Nan</td><td style="text-align: center;">5408</td><td style="text-align: center;">0.0</td></tr> </tbody> </table>		Count	% valeurs	infra			-1.0	2236	0.0	0.0	1032089	88.0	1.0	10754	1.0	2.0	17283	1.0	3.0	17768	2.0	4.0	4113	0.0	5.0	68524	6.0	6.0	8329	1.0	7.0	699	0.0	8.0	1633	0.0	9.0	8037	1.0	Nan	5408	0.0	<p style="text-align: center;">infra</p> <p style="text-align: right;">Years 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022</p>
	Count	% valeurs																																									
infra																																											
-1.0	2236	0.0																																									
0.0	1032089	88.0																																									
1.0	10754	1.0																																									
2.0	17283	1.0																																									
3.0	17768	2.0																																									
4.0	4113	0.0																																									
5.0	68524	6.0																																									
6.0	8329	1.0																																									
7.0	699	0.0																																									
8.0	1633	0.0																																									
9.0	8037	1.0																																									
Nan	5408	0.0																																									

Evolution	
Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».

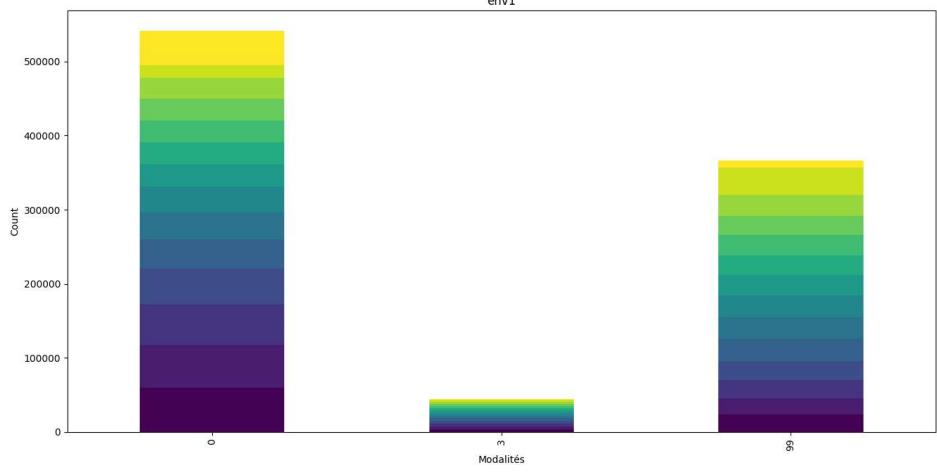
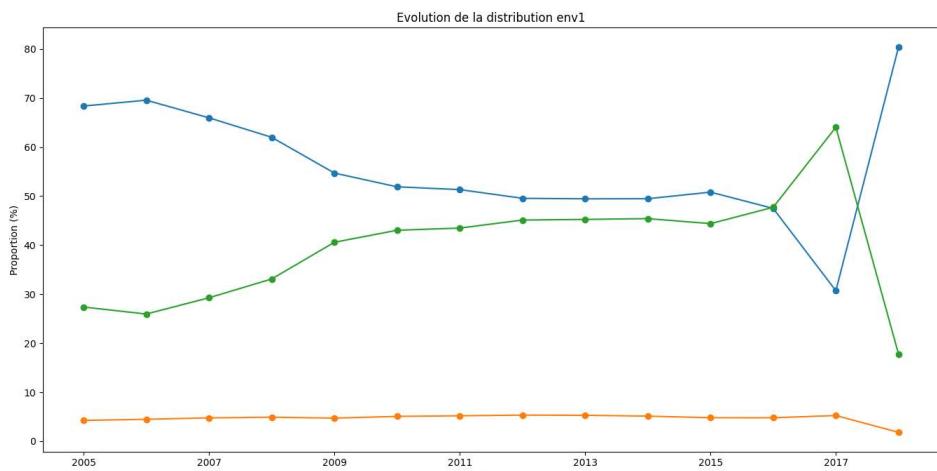
q. situ

Description	Situation de l'accident.										
Modalités	<ul style="list-style-type: none"> - -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 										
Type	[2009-2018] : float64 [2005-2008] : int64 [2019-2022] : object										
Etendue des valeurs	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">count</th> <th style="text-align: center;">unique</th> <th style="text-align: center;">top</th> <th style="text-align: center;">freq</th> </tr> </thead> <tbody> <tr> <td>situ</td> <td style="text-align: center;">1171903</td> <td style="text-align: center;">9</td> <td style="text-align: center;">1.0</td> <td style="text-align: center;">983318</td> </tr> </tbody> </table>		count	unique	top	freq	situ	1171903	9	1.0	983318
	count	unique	top	freq							
situ	1171903	9	1.0	983318							
Valeurs nulles	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">Type</th> <th style="text-align: center;">Val_notnull</th> <th style="text-align: center;">Val_null</th> <th style="text-align: center;">%_null</th> </tr> </thead> <tbody> <tr> <td>situ</td> <td style="text-align: center;">float64</td> <td style="text-align: center;">1171903</td> <td style="text-align: center;">4970</td> <td style="text-align: center;">0.42</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	situ	float64	1171903	4970	0.42
	Type	Val_notnull	Val_null	%_null							
situ	float64	1171903	4970	0.42							
Outliers	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">outliers_count</th> <th style="text-align: center;">outliers_unique</th> <th style="text-align: center;">outliers_list</th> </tr> </thead> <tbody> <tr> <td>situ</td> <td style="text-align: center;">188585</td> <td style="text-align: center;">8</td> <td style="text-align: center;">[-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	situ	188585	8	[-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0]		
	outliers_count	outliers_unique	outliers_list								
situ	188585	8	[-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0]								

<h3>Répartition</h3> <table border="1"> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>situ</td><td></td><td></td></tr> <tr> <td>-1.0</td><td>142</td><td>0.0</td></tr> <tr> <td>0.0</td><td>47458</td><td>4.0</td></tr> <tr> <td>1.0</td><td>983318</td><td>84.0</td></tr> <tr> <td>2.0</td><td>8456</td><td>1.0</td></tr> <tr> <td>3.0</td><td>92269</td><td>8.0</td></tr> <tr> <td>4.0</td><td>20686</td><td>2.0</td></tr> <tr> <td>5.0</td><td>10424</td><td>1.0</td></tr> <tr> <td>6.0</td><td>2580</td><td>0.0</td></tr> <tr> <td>8.0</td><td>6570</td><td>1.0</td></tr> <tr> <td>NaN</td><td>4970</td><td>0.0</td></tr> </tbody> </table>		Count	% valeurs	situ			-1.0	142	0.0	0.0	47458	4.0	1.0	983318	84.0	2.0	8456	1.0	3.0	92269	8.0	4.0	20686	2.0	5.0	10424	1.0	6.0	2580	0.0	8.0	6570	1.0	NaN	4970	0.0	 <p>histogramme de distribution situ par année</p> <p>Y-axis : Count (0.0 à 1.0e6)</p> <p>X-axis : Modalités (0 à 71)</p> <p>Legendre des années : 2005 (violet), 2006 (violet), 2007 (violet), 2008 (violet), 2009 (bleu), 2010 (bleu), 2011 (bleu), 2012 (bleu), 2013 (bleu), 2014 (bleu), 2015 (bleu), 2016 (vert), 2017 (vert), 2018 (vert), 2019 (vert), 2020 (vert), 2021 (jaune), 2022 (jaune)</p>
	Count	% valeurs																																			
situ																																					
-1.0	142	0.0																																			
0.0	47458	4.0																																			
1.0	983318	84.0																																			
2.0	8456	1.0																																			
3.0	92269	8.0																																			
4.0	20686	2.0																																			
5.0	10424	1.0																																			
6.0	2580	0.0																																			
8.0	6570	1.0																																			
NaN	4970	0.0																																			
<h3>Evolution</h3>	 <p>Evolution de la distribution situ</p> <p>Y-axis : Proportion (%) (0 à 80)</p> <p>X-axis : Années (2005 à 2022)</p> <p>Legendre des modalités : 0 (bleu), 1 (orange), 2 (vert), 3 (rouge), 4 (violet), 5 (marron), 6 (rose), 8 (gris), -1 (jaune)</p> <p>Detailed description: The graph shows the percentage distribution of road surface types from 2005 to 2022. Type 1 (orange) is the dominant category, peaking around 85% in 2006 and 2007, then fluctuating between 80% and 85%. Type 3 (red) shows a sharp peak of about 40% in 2019. Other categories like 0, 2, 4, 5, 6, 8, and -1 remain below 10% throughout the period.</p>																																				
<h3>Remarque</h3>	<p>On observe de brusques variations des modalités sur accotement et sur chaussée aux alentours de 2019. Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».</p>																																				

r. env1

Description	Point école : proximité d'une école
Modalités	<ul style="list-style-type: none"> - 0 - 3 - 99
Type	<ul style="list-style-type: none"> [2005-2008] : int64 [2009-2018] : float64

Etendue des valeurs	count unique top freq														
	env1 953029 3 0.0 541532														
Valeurs nulles	Type Val_notnull Val_null %_null														
	env1 float64 953029 223844 19.02														
Outliers	outliers_count outliers_unique outliers_list														
	env1 0 0 []														
Répartition															
	Count % valeurs env1 <table border="1"> <thead> <tr> <th></th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>0.0</td> <td>541532</td> <td>46.0</td> </tr> <tr> <td>3.0</td> <td>44824</td> <td>4.0</td> </tr> <tr> <td>99.0</td> <td>366673</td> <td>31.0</td> </tr> <tr> <td>NaN</td> <td>223844</td> <td>19.0</td> </tr> </tbody> </table> 		Count	% valeurs	0.0	541532	46.0	3.0	44824	4.0	99.0	366673	31.0	NaN	223844
	Count	% valeurs													
0.0	541532	46.0													
3.0	44824	4.0													
99.0	366673	31.0													
NaN	223844	19.0													
Evolution															
Remarque	La variable disparaît à partir de 2018.														

S. vma

Description	Vitesse maximale autorisée sur le lieu et au moment de l'accident.
Type	[2019-2022] : int64
Etendue des valeurs	count unique top freq
	vma 218404 47 50.0 115319

Valeurs nulles	Type	Val_notnull	Val_null	%_null	
	vma	float64	218404	958469 81.44	
Outliers		outliers_count	outliers_unique	outliers_list	
	vma	7701	21 [-1.0, 0.0, 1.0, 2.0, 3.0, 4.0, 130.0, 140.0,...		
		Boxplots pour: vma			
Répartition	Count	% valeurs			
vma					
-1.0	3393	0.29			
0.0	1	0.0			
1.0	67	0.01			
2.0	47	0.0			
3.0	9	0.0			
...			
770.0	1	0.0			
800.0	1	0.0			
900.0	4	0.0			
901.0	1	0.0			
Nan	958469	81.44			
48 rows × 2 columns					
Remarque	La variable apparaît en 2019. Lorsqu'elle n'évoque pas des situations aberrantes, elle correspond souvent à des NaN.				

3. Usagers

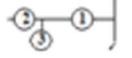
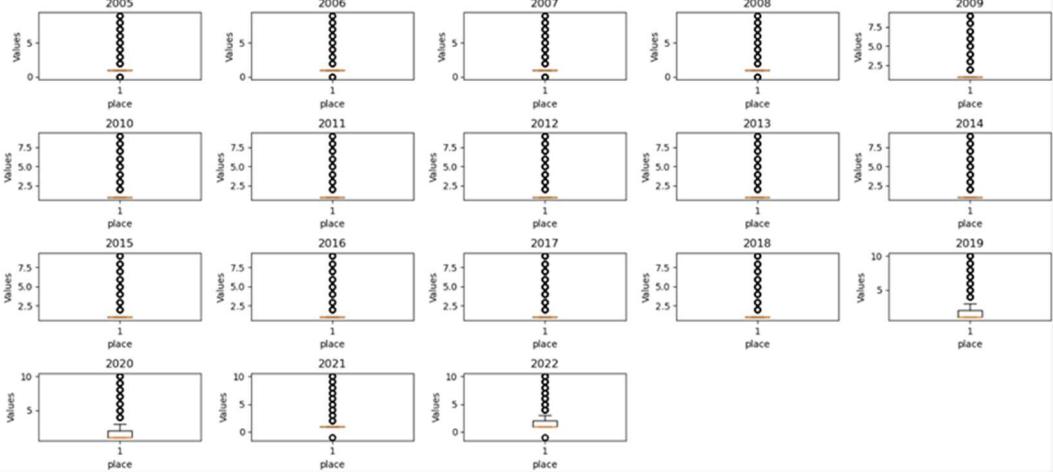
Rows x columns Rows duplicated

Usagers (2636377, 17) 2858

a. Num_Acc

Description	Identifiant de l'accident identique à celui du fichier "rubrique CARACTERISTIQUES" repris pour chacun des usagers décrits impliqués dans l'accident.																																																														
Type	int64																																																														
Etendue des valeurs	<table> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td>2636377</td><td>1176873</td><td>200600016834</td><td>86</td></tr> </tbody> </table>						count	unique	top	freq	Num_Acc	2636377	1176873	200600016834	86																																																
	count	unique	top	freq																																																											
Num_Acc	2636377	1176873	200600016834	86																																																											
Valeurs nulles	<table> <thead> <tr> <th></th><th>Type</th><th>Val_notnull</th><th>Val_null</th><th>%_null</th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td>int64</td><td>2636377</td><td>0</td><td>0.0</td></tr> </tbody> </table>						Type	Val_notnull	Val_null	%_null	Num_Acc	int64	2636377	0	0.0																																																
	Type	Val_notnull	Val_null	%_null																																																											
Num_Acc	int64	2636377	0	0.0																																																											
Outliers	<table> <thead> <tr> <th></th><th>outliers_count</th><th>outliers_unique</th><th>outliers_list</th><th></th></tr> </thead> <tbody> <tr> <td>Num_Acc</td><td>0</td><td>0</td><td>[]</td><td></td></tr> </tbody> </table>						outliers_count	outliers_unique	outliers_list		Num_Acc	0	0	[]																																																	
	outliers_count	outliers_unique	outliers_list																																																												
Num_Acc	0	0	[]																																																												
Répartition	<table> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th><th></th></tr> <tr> <th>Num_Acc</th><th></th><th></th><th></th></tr> </thead> <tbody> <tr> <td>200500000001</td><td>6</td><td>0.0</td><td></td></tr> <tr> <td>200500000002</td><td>2</td><td>0.0</td><td></td></tr> <tr> <td>200500000003</td><td>2</td><td>0.0</td><td></td></tr> <tr> <td>200500000004</td><td>4</td><td>0.0</td><td></td></tr> <tr> <td>200500000005</td><td>2</td><td>0.0</td><td></td></tr> <tr> <td>...</td><td>...</td><td>...</td><td></td></tr> <tr> <td>202200055298</td><td>2</td><td>0.0</td><td></td></tr> <tr> <td>202200055299</td><td>1</td><td>0.0</td><td></td></tr> <tr> <td>202200055300</td><td>1</td><td>0.0</td><td></td></tr> <tr> <td>202200055301</td><td>3</td><td>0.0</td><td></td></tr> <tr> <td>202200055302</td><td>2</td><td>0.0</td><td></td></tr> <tr> <td colspan="4">1176873 rows × 2 columns</td><td></td><td></td></tr> </tbody> </table>						Count	% valeurs		Num_Acc				200500000001	6	0.0		200500000002	2	0.0		200500000003	2	0.0		200500000004	4	0.0		200500000005	2	0.0			202200055298	2	0.0		202200055299	1	0.0		202200055300	1	0.0		202200055301	3	0.0		202200055302	2	0.0		1176873 rows × 2 columns					
	Count	% valeurs																																																													
Num_Acc																																																															
200500000001	6	0.0																																																													
200500000002	2	0.0																																																													
200500000003	2	0.0																																																													
200500000004	4	0.0																																																													
200500000005	2	0.0																																																													
...																																																													
202200055298	2	0.0																																																													
202200055299	1	0.0																																																													
202200055300	1	0.0																																																													
202200055301	3	0.0																																																													
202200055302	2	0.0																																																													
1176873 rows × 2 columns																																																															

b. place

Description	Permet de situer la place occupée dans le véhicule par l'usager au moment de l'accident.																																			
	Transport en commun																																			
	Voiture																																			
	Moto / Side-car																																			
	 <table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>4</td><td>7</td><td>7</td><td>7</td><td></td><td></td><td></td></tr> <tr><td>5</td><td>8</td><td>8</td><td>8</td><td></td><td>7</td><td>7</td></tr> <tr><td>5</td><td>8</td><td>8</td><td>8</td><td></td><td>8</td><td>8</td></tr> <tr><td>5</td><td>8</td><td>8</td><td>8</td><td></td><td>8</td><td>8</td></tr> <tr><td>3</td><td>9</td><td>9</td><td>9</td><td></td><td>9</td><td>9</td></tr> </table>	4	7	7	7				5	8	8	8		7	7	5	8	8	8		8	8	5	8	8	8		8	8	3	9	9	9		9	9
4	7	7	7																																	
5	8	8	8		7	7																														
5	8	8	8		8	8																														
5	8	8	8		8	8																														
3	9	9	9		9	9																														
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64																																			
Etendue des valeurs	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: left;">count</th> <th style="text-align: left;">unique</th> <th style="text-align: left;">top</th> <th style="text-align: left;">freq</th> </tr> </thead> <tbody> <tr> <td>place</td> <td>2513249</td> <td>12</td> <td>1.0</td> <td>1962529</td> </tr> </tbody> </table>		count	unique	top	freq	place	2513249	12	1.0	1962529																									
	count	unique	top	freq																																
place	2513249	12	1.0	1962529																																
Valeurs nulles	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: left;">Type</th> <th style="text-align: left;">Val_notnull</th> <th style="text-align: left;">Val_null</th> <th style="text-align: left;">%_null</th> </tr> </thead> <tbody> <tr> <td>place</td> <td>float64</td> <td>2513249</td> <td>123128</td> <td>4.67</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	place	float64	2513249	123128	4.67																									
	Type	Val_notnull	Val_null	%_null																																
place	float64	2513249	123128	4.67																																
Outliers	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: left;">outliers_count</th> <th style="text-align: left;">outliers_unique</th> <th style="text-align: right;">outliers_list</th> </tr> </thead> <tbody> <tr> <td>place</td> <td>550720</td> <td>11</td> <td style="text-align: right;">[-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]</td> </tr> </tbody> </table> <p style="text-align: center;">Boxplots for column: place</p> 		outliers_count	outliers_unique	outliers_list	place	550720	11	[-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]																											
	outliers_count	outliers_unique	outliers_list																																	
place	550720	11	[-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]																																	

Répartition	<table border="1"> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>Modalité</td><td></td><td></td></tr> <tr> <td>-1.0</td><td>25</td><td>0.0</td></tr> <tr> <td>0.0</td><td>60766</td><td>2.0</td></tr> <tr> <td>1.0</td><td>1962529</td><td>78.0</td></tr> <tr> <td>2.0</td><td>281895</td><td>11.0</td></tr> <tr> <td>3.0</td><td>60272</td><td>2.0</td></tr> <tr> <td>4.0</td><td>52156</td><td>2.0</td></tr> <tr> <td>5.0</td><td>27703</td><td>1.0</td></tr> <tr> <td>6.0</td><td>2647</td><td>0.0</td></tr> <tr> <td>7.0</td><td>9362</td><td>0.0</td></tr> <tr> <td>8.0</td><td>8118</td><td>0.0</td></tr> <tr> <td>9.0</td><td>9184</td><td>0.0</td></tr> <tr> <td>10.0</td><td>38592</td><td>2.0</td></tr> </tbody> </table>		Count	% valeurs	Modalité			-1.0	25	0.0	0.0	60766	2.0	1.0	1962529	78.0	2.0	281895	11.0	3.0	60272	2.0	4.0	52156	2.0	5.0	27703	1.0	6.0	2647	0.0	7.0	9362	0.0	8.0	8118	0.0	9.0	9184	0.0	10.0	38592	2.0
	Count	% valeurs																																									
Modalité																																											
-1.0	25	0.0																																									
0.0	60766	2.0																																									
1.0	1962529	78.0																																									
2.0	281895	11.0																																									
3.0	60272	2.0																																									
4.0	52156	2.0																																									
5.0	27703	1.0																																									
6.0	2647	0.0																																									
7.0	9362	0.0																																									
8.0	8118	0.0																																									
9.0	9184	0.0																																									
10.0	38592	2.0																																									
Evolution																																											
Remarque	Certaines modalités (-1 et 0) ne sont pas répertoriées dans la description.																																										

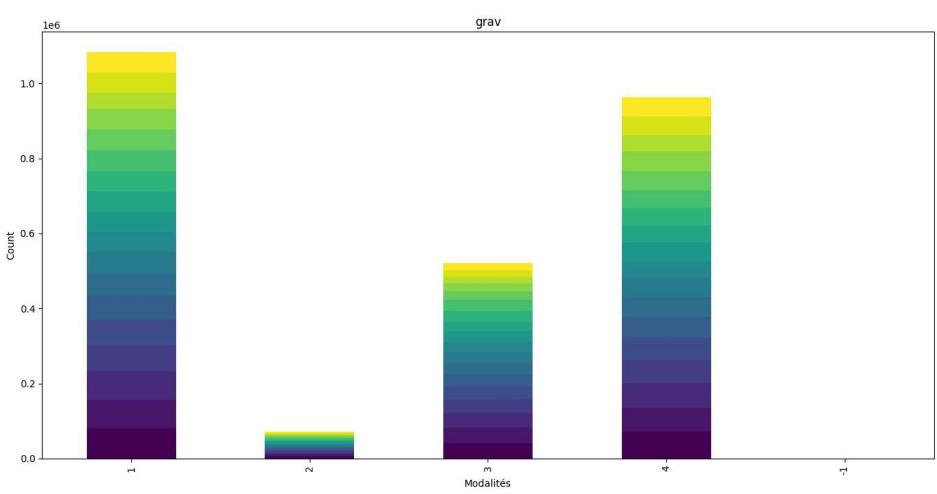
c. catu

Description	Catégorie d'usager.										
Modalités	<ul style="list-style-type: none"> - 1 : Conducteur - 2 : Passager - 3 : Piéton 										
Type	int64										
Etendue des valeurs	<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>catu</td> <td>2636377</td> <td>4</td> <td>1</td> <td>1961486</td> </tr> </tbody> </table>		count	unique	top	freq	catu	2636377	4	1	1961486
	count	unique	top	freq							
catu	2636377	4	1	1961486							

Valeurs nulles	Type	Val_notnull	Val_null	%_null		
	catu	int64	2636377	0 0.0		
Outliers	outliers_count	outliers_unique	outliers_list			
	catu	3560	1	[4.0]		
	Boxplots pour: catu					
Répartition	Count	% valeurs				
Modalité						
1	1961486	74.0				
2	454622	17.0				
3	216709	8.0				
4	3560	0.0				
Evolution	Evolution de la distribution catu					

Remarque	La modalité 4 non répertoriée dans la description, finit par disparaître à partir de 2018.
-----------------	--

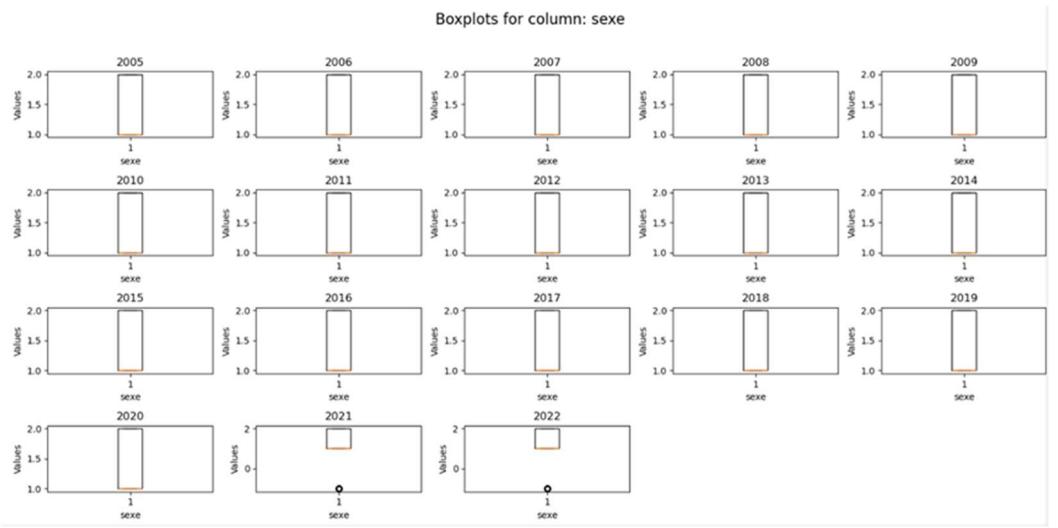
d. grav

Description	Gravité de blessure de l'usager.																					
Modalités	<ul style="list-style-type: none"> - 1 : Indemne - 2 : Tué - 3 : Blessé hospitalisé - 4 : Blessé léger 																					
Type	int64																					
Valeurs nulles	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>grav</td> <td>int64</td> <td>2636377</td> <td>0</td> <td>0.0</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	grav	int64	2636377	0	0.0											
	Type	Val_notnull	Val_null	%_null																		
grav	int64	2636377	0	0.0																		
Etendue des valeurs	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>grav</td> <td>2636377</td> <td>5</td> <td>1</td> <td>1082746</td> </tr> </tbody> </table>		count	unique	top	freq	grav	2636377	5	1	1082746											
	count	unique	top	freq																		
grav	2636377	5	1	1082746																		
Outliers	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th>outliers_count</th> <th>outliers_unique</th> <th>outliers_list</th> </tr> </thead> <tbody> <tr> <td>grav</td> <td>0</td> <td>0</td> <td>[]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	grav	0	0	[]													
	outliers_count	outliers_unique	outliers_list																			
grav	0	0	[]																			
Répartition	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>Modalité</td> <td></td> <td></td> </tr> <tr> <td>-1</td> <td>301</td> <td>0.0</td> </tr> <tr> <td>1</td> <td>1082746</td> <td>41.0</td> </tr> <tr> <td>2</td> <td>70628</td> <td>3.0</td> </tr> <tr> <td>3</td> <td>520817</td> <td>20.0</td> </tr> <tr> <td>4</td> <td>961885</td> <td>36.0</td> </tr> </tbody> </table> 		Count	% valeurs	Modalité			-1	301	0.0	1	1082746	41.0	2	70628	3.0	3	520817	20.0	4	961885	36.0
	Count	% valeurs																				
Modalité																						
-1	301	0.0																				
1	1082746	41.0																				
2	70628	3.0																				
3	520817	20.0																				
4	961885	36.0																				

Evolution	<p style="text-align: center;">Evolution de la distribution grav</p> <p style="text-align: center;">Années</p>
Remarque	Autour de 2018, un changement de saisie a lieu sur les modalités 4 et 3 (pour la comptabilisation des hospitalisations). La modalité -1 n'est pas répertoriée dans la description.

e. sexe

Description	Sexe de l'usager										
Modalités	- 1 : Masculin - 2 : Féminin										
Type	int64										
Etendue des valeurs	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">count</th> <th style="text-align: center;">unique</th> <th style="text-align: center;">top</th> <th style="text-align: center;">freq</th> </tr> </thead> <tbody> <tr> <td>sexé</td> <td style="text-align: center;">2636377</td> <td style="text-align: center;">3</td> <td style="text-align: center;">1</td> <td style="text-align: center;">1773190</td> </tr> </tbody> </table>		count	unique	top	freq	sexé	2636377	3	1	1773190
	count	unique	top	freq							
sexé	2636377	3	1	1773190							
Valeurs nulles	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">Type</th> <th style="text-align: center;">Val_notnull</th> <th style="text-align: center;">Val_null</th> <th style="text-align: center;">%_null</th> </tr> </thead> <tbody> <tr> <td>sexé</td> <td style="text-align: center;">int64</td> <td style="text-align: center;">2636377</td> <td style="text-align: center;">0</td> <td style="text-align: center;">0.0</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	sexé	int64	2636377	0	0.0
	Type	Val_notnull	Val_null	%_null							
sexé	int64	2636377	0	0.0							
Outliers	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">outliers_count</th> <th style="text-align: center;">outliers_unique</th> <th style="text-align: center;">outliers_list</th> </tr> </thead> <tbody> <tr> <td>sexé</td> <td style="text-align: center;">5806</td> <td style="text-align: center;">1</td> <td style="text-align: center;">[-1.0]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	sexé	5806	1	[-1.0]		
	outliers_count	outliers_unique	outliers_list								
sexé	5806	1	[-1.0]								

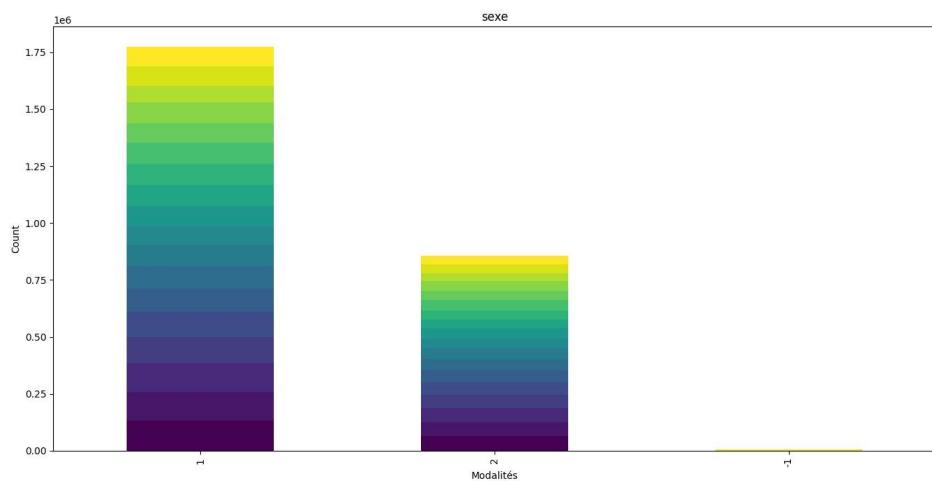


Répartition

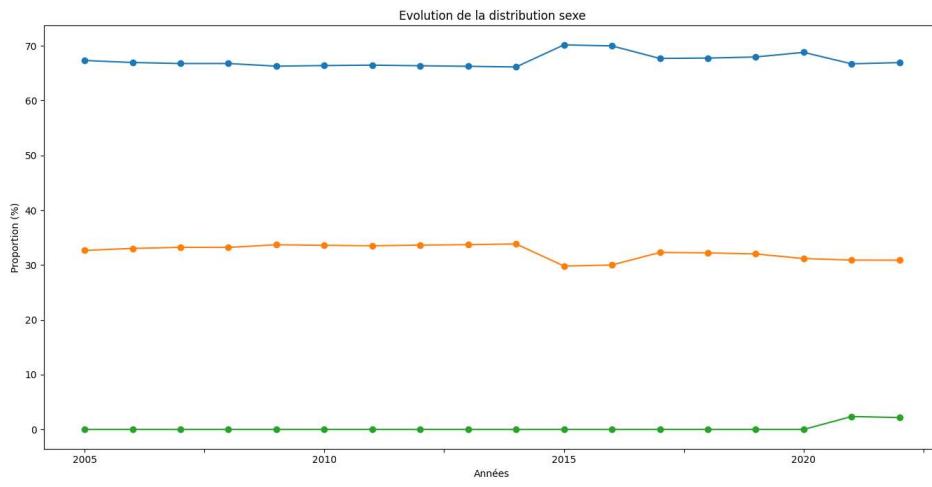
Count % valeurs

Modalité

-1	5806	0.0
1	1773190	67.0
2	857381	33.0



Evolution



Remarque

La valeur -1 n'est pas répertoriée dans la description et apparaît comme outlier.

f. trajet

Description

Motif du déplacement au moment de l'accident.

Modalités	- -1 : Non renseigné - 0 : Non renseigné - 1 : Domicile - travail - 2 : Domicile - école - 3 : Courses - achats - 4 : Utilisation professionnelle - 5 : Promenade - loisirs - 9 : Autre																														
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64																														
Etendue des valeurs	<table> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>trajet</td><td>2635883</td><td>8</td><td>5.0</td><td>978415</td></tr> </tbody> </table>		count	unique	top	freq	trajet	2635883	8	5.0	978415																				
	count	unique	top	freq																											
trajet	2635883	8	5.0	978415																											
Valeurs nulles	<table> <thead> <tr> <th></th><th>Type</th><th>Val_notnull</th><th>Val_null</th><th>%_null</th></tr> </thead> <tbody> <tr> <td>trajet</td><td>float64</td><td>2635883</td><td>494</td><td>0.02</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	trajet	float64	2635883	494	0.02																				
	Type	Val_notnull	Val_null	%_null																											
trajet	float64	2635883	494	0.02																											
Outliers	<table> <thead> <tr> <th></th><th>outliers_count</th><th>outliers_unique</th><th>outliers_list</th></tr> </thead> <tbody> <tr> <td>trajet</td><td>0</td><td>0</td><td>[]</td></tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	trajet	0	0	[]																						
	outliers_count	outliers_unique	outliers_list																												
trajet	0	0	[]																												
Répartition	<table> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>Modalité</td><td></td><td></td></tr> <tr> <td>-1.0</td><td>6900</td><td>0.0</td></tr> <tr> <td>0.0</td><td>734734</td><td>28.0</td></tr> <tr> <td>1.0</td><td>344904</td><td>13.0</td></tr> <tr> <td>2.0</td><td>54763</td><td>2.0</td></tr> <tr> <td>3.0</td><td>71040</td><td>3.0</td></tr> <tr> <td>4.0</td><td>255752</td><td>10.0</td></tr> <tr> <td>5.0</td><td>978415</td><td>37.0</td></tr> <tr> <td>9.0</td><td>189375</td><td>7.0</td></tr> </tbody> </table>		Count	% valeurs	Modalité			-1.0	6900	0.0	0.0	734734	28.0	1.0	344904	13.0	2.0	54763	2.0	3.0	71040	3.0	4.0	255752	10.0	5.0	978415	37.0	9.0	189375	7.0
	Count	% valeurs																													
Modalité																															
-1.0	6900	0.0																													
0.0	734734	28.0																													
1.0	344904	13.0																													
2.0	54763	2.0																													
3.0	71040	3.0																													
4.0	255752	10.0																													
5.0	978415	37.0																													
9.0	189375	7.0																													

Evolution	
Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».

g. secu

Description	<p>Enseigne variable sur 2 caractères :</p> <ul style="list-style-type: none"> - le premier concerne l'existence d'un Équipement de sécurité <ul style="list-style-type: none"> 1 - Ceinture 2 - Casque 3 - Dispositif enfants 4 - Equipement réfléchissant 9 - Autre - le second concerne l'utilisation de l'Équipement de sécurité <ul style="list-style-type: none"> 1 - Oui 2 - Non 3 - Non déterminable 										
Type	[2005 ;2007-2008] : int64 [2006 ; 2009-2018] : float64										
Etendue des valeurs	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: left;">count</th> <th style="text-align: left;">unique</th> <th style="text-align: left;">top</th> <th style="text-align: left;">freq</th> </tr> </thead> <tbody> <tr> <td>secu</td> <td>2085658</td> <td>24</td> <td>11.0</td> <td>1197467</td> </tr> </tbody> </table>		count	unique	top	freq	secu	2085658	24	11.0	1197467
	count	unique	top	freq							
secu	2085658	24	11.0	1197467							
Valeurs nulles	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: left;">Type</th> <th style="text-align: left;">Val_notnull</th> <th style="text-align: left;">Val_null</th> <th style="text-align: left;">%_null</th> </tr> </thead> <tbody> <tr> <td>secu</td> <td>float64</td> <td>2085658</td> <td>550719</td> <td>20.89</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	secu	float64	2085658	550719	20.89
	Type	Val_notnull	Val_null	%_null							
secu	float64	2085658	550719	20.89							
Outliers	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: left;">outliers_count</th> <th style="text-align: left;">outliers_unique</th> <th style="text-align: right;">outliers_list</th> </tr> </thead> <tbody> <tr> <td>secu</td> <td>126934</td> <td>8</td> <td style="text-align: right;">[40.0, 41.0, 42.0, 43.0, 90.0, 91.0, 92.0, 93.0]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	secu	126934	8	[40.0, 41.0, 42.0, 43.0, 90.0, 91.0, 92.0, 93.0]		
	outliers_count	outliers_unique	outliers_list								
secu	126934	8	[40.0, 41.0, 42.0, 43.0, 90.0, 91.0, 92.0, 93.0]								

	<p style="text-align: center;">Boxplots for column: secu</p>																																			
Répartition Count % valeurs <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>secu</th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr><td>0.0</td><td>68374</td><td>2.59</td></tr> <tr><td>1.0</td><td>3568</td><td>0.14</td></tr> <tr><td>2.0</td><td>2669</td><td>0.1</td></tr> <tr><td>3.0</td><td>7707</td><td>0.29</td></tr> <tr><td>10.0</td><td>5631</td><td>0.21</td></tr> <tr><td>...</td><td>...</td><td>...</td></tr> <tr><td>90.0</td><td>73</td><td>0.0</td></tr> <tr><td>91.0</td><td>7653</td><td>0.29</td></tr> <tr><td>92.0</td><td>7693</td><td>0.29</td></tr> <tr><td>93.0</td><td>105121</td><td>3.99</td></tr> <tr><td>NaN</td><td>550719</td><td>20.89</td></tr> </tbody> </table> <p>25 rows × 2 columns</p>	secu	Count	% valeurs	0.0	68374	2.59	1.0	3568	0.14	2.0	2669	0.1	3.0	7707	0.29	10.0	5631	0.21	90.0	73	0.0	91.0	7653	0.29	92.0	7693	0.29	93.0	105121	3.99	NaN	550719	20.89
secu	Count	% valeurs																																		
0.0	68374	2.59																																		
1.0	3568	0.14																																		
2.0	2669	0.1																																		
3.0	7707	0.29																																		
10.0	5631	0.21																																		
...																																		
90.0	73	0.0																																		
91.0	7653	0.29																																		
92.0	7693	0.29																																		
93.0	105121	3.99																																		
NaN	550719	20.89																																		
Remarque	La variable disparaît à partir de 2019 au profit de secu1/2/3. Forte proportion de valeurs NaN.																																			

h. locp

Description	Localisation du piéton.
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 0 : Sans objet - 1 : Sur chaussée : A + 50 m du passage piéton

	<ul style="list-style-type: none"> - 2 : Sur chaussée : A - 50 m du passage piéton - 3 : Sur passage piéton : Sans signalisation lumineuse - 4 : Sur passage piéton : Avec signalisation lumineuse - 5 : Sur trottoir - 6 : Sur accotement - 7 : Sur refuge ou BAU - 8 : Sur contre allée - 9 : Inconnue 										
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64										
Etendue des valeurs	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">count</th><th style="text-align: center;">unique</th><th style="text-align: center;">top</th><th style="text-align: center;">freq</th></tr> </thead> <tbody> <tr> <td style="color: #0070C0;">locp</td><td style="color: #0070C0;">2580016</td><td style="color: #0070C0;">11</td><td style="color: #0070C0;">0.0</td><td style="color: #0070C0;">2162665</td></tr> </tbody> </table>		count	unique	top	freq	locp	2580016	11	0.0	2162665
	count	unique	top	freq							
locp	2580016	11	0.0	2162665							
Valeurs nulles	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">Type</th><th style="text-align: center;">Val_notnull</th><th style="text-align: center;">Val_null</th><th style="text-align: center;">%_null</th></tr> </thead> <tbody> <tr> <td style="color: #0070C0;">locp</td><td style="color: #0070C0;">float64</td><td style="color: #0070C0;">2580016</td><td style="color: #0070C0;">56361</td><td style="color: #0070C0;">2.14</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	locp	float64	2580016	56361	2.14
	Type	Val_notnull	Val_null	%_null							
locp	float64	2580016	56361	2.14							
Outliers	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">outliers_count</th><th style="text-align: center;">outliers_unique</th><th style="text-align: right; vertical-align: bottom;">outliers_list</th></tr> </thead> <tbody> <tr> <td style="color: #0070C0;">locp</td><td style="color: #0070C0;">417351</td><td style="color: #0070C0;">10</td><td style="text-align: right; vertical-align: bottom;">[-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]</td></tr> </tbody> </table> <p style="text-align: center;">Boxplots for column: locp</p>		outliers_count	outliers_unique	outliers_list	locp	417351	10	[-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]		
	outliers_count	outliers_unique	outliers_list								
locp	417351	10	[-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]								

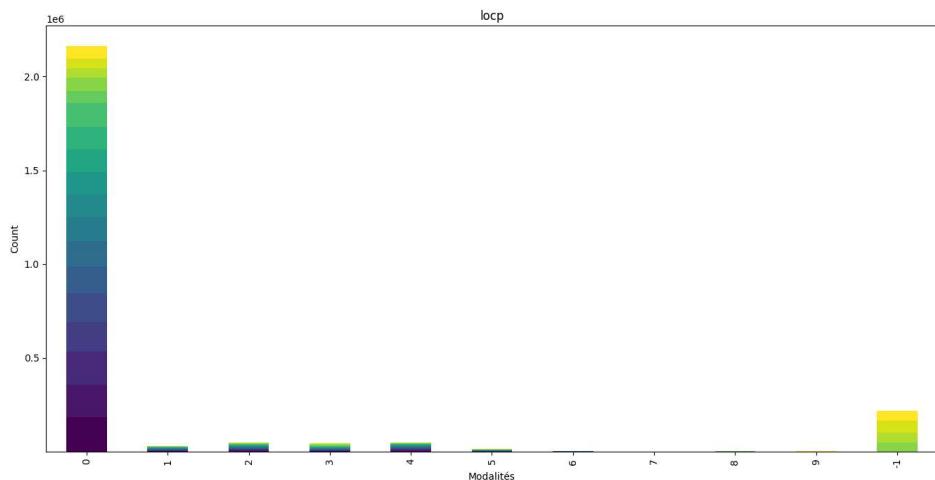
Répartition

Count % valeurs

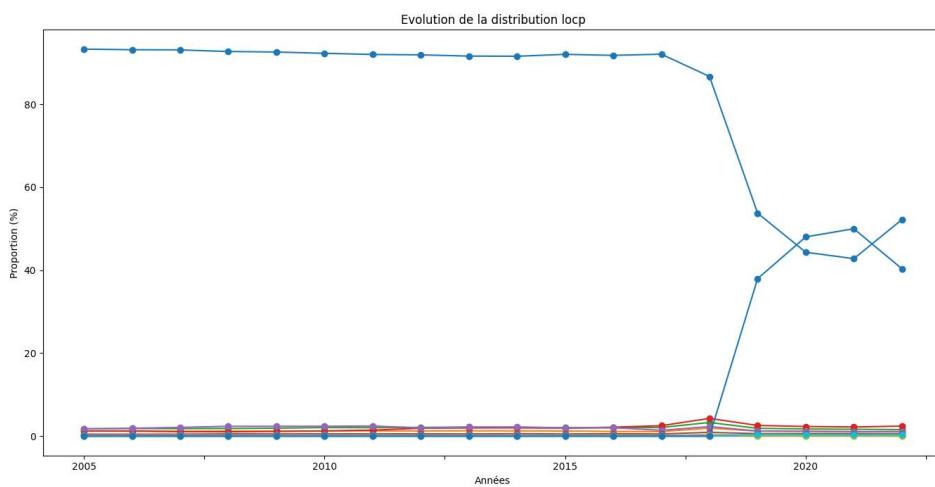
locp

	Count	% valeurs
-1.0	216753	8.22
0.0	2162665	82.03
1.0	31299	1.19
2.0	50467	1.91
3.0	46270	1.76
...
6.0	4703	0.18
7.0	247	0.01
8.0	2367	0.09
9.0	1844	0.07
NaN	56361	2.14

12 rows × 2 columns



Evolution



Remarque

Attention aux valeurs -1 et 0 pour lesquelles il pourrait y avoir un amalgame.
Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».

i. actp

Description

Action du piéton.

Modalités

- -1 : Non renseigné
- 0 : Non renseigné ou sans objet
- 1 : Se déplaçant dans le Sens véhicule heurtant

	<ul style="list-style-type: none"> - 2 : Se déplaçant dans le Sens inverse du véhicule - 3 : Traversant - 4 : Masqué - 5 : Jouant – courant - 6 : Avec animal - 9 : Autre - A : Monte/descend du véhicule - B : Inconnue 										
Type	[2005-2008] : int64 [2009-2018] : float64 [2019-2022] : object										
Etendue des valeurs	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">count</th><th style="text-align: center;">unique</th><th style="text-align: center;">top</th><th style="text-align: center;">freq</th></tr> </thead> <tbody> <tr> <td>actp</td><td style="text-align: center;">2579915</td><td style="text-align: center;">21</td><td style="text-align: center;">0.0</td><td style="text-align: center;">1224776</td></tr> </tbody> </table>		count	unique	top	freq	actp	2579915	21	0.0	1224776
	count	unique	top	freq							
actp	2579915	21	0.0	1224776							
Valeurs nulles	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">Type</th><th style="text-align: center;">Val_notnull</th><th style="text-align: center;">Val_null</th><th style="text-align: center;">%_null</th></tr> </thead> <tbody> <tr> <td>actp</td><td style="text-align: center;">object</td><td style="text-align: center;">2579915</td><td style="text-align: center;">56462</td><td style="text-align: center;">2.14</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	actp	object	2579915	56462	2.14
	Type	Val_notnull	Val_null	%_null							
actp	object	2579915	56462	2.14							
Outliers	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">outliers_count</th><th style="text-align: center;">outliers_unique</th><th style="text-align: center;">outliers_list</th></tr> </thead> <tbody> <tr> <td>actp</td><td style="text-align: center;">50975</td><td style="text-align: center;">16</td><td style="text-align: center;">[1, 1.0, 2, 2.0, 4, 4.0, 5, 5.0, 6, 6.0, 7, 8...</td></tr> </tbody> </table> <p style="text-align: center;">Boxplots for column: actp</p>		outliers_count	outliers_unique	outliers_list	actp	50975	16	[1, 1.0, 2, 2.0, 4, 4.0, 5, 5.0, 6, 6.0, 7, 8...		
	outliers_count	outliers_unique	outliers_list								
actp	50975	16	[1, 1.0, 2, 2.0, 4, 4.0, 5, 5.0, 6, 6.0, 7, 8...								

Répartition			
	Count	% valeurs	
Modalité			
-1	185666	7.0	
0	959099	37.0	
0.0	1224776	47.0	
1	5470	0.0	
1.0	6758	0.0	
2	2775	0.0	
2.0	3239	0.0	
3	70470	3.0	
3.0	88929	3.0	
4	1378	0.0	
4.0	2389	0.0	
5	4737	0.0	
5.0	7196	0.0	
6	239	0.0	
6.0	275	0.0	
7	60	0.0	
8	52	0.0	
9	6580	0.0	
9.0	8171	0.0	
A	422	0.0	
B	1234	0.0	
Remarque			Attention aux modalités -1, 0 pour lesquelles il pourrait y avoir un amalgame. Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».

j. etatp

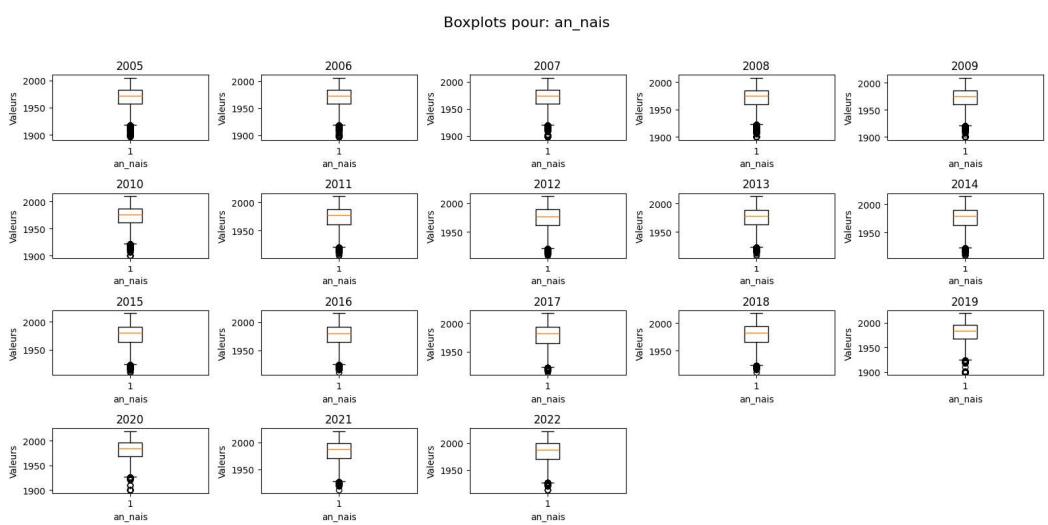
Description	Cette variable permet de préciser si le piéton accidenté était seul ou non.
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 1 : Seul - 2 : Accompagné - 3 : En groupe
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64

Etendue des valeurs		count	unique	top	freq	
		etatp	2579959	5	0.0 1914793	
Valeurs nulles		Type	Val_notnull	Val_null	%_null	
		etatp	float64	2579959	56418	2.14
Outliers		outliers_count	outliers_unique	outliers_list		
		etatp	665166	4	[-1.0, 1.0, 2.0, 3.0]	
<p>Boxplots for column: etatp</p>						
Répartition		Count	% valeurs			
Modalité		Count	% valeurs			
-1.0		456246	18.0			
0.0		1914793	74.0			
1.0		158481	6.0			
2.0		41509	2.0			
3.0		8930	0.0			

Evolution	<p>Evolution de la distribution etatp</p> <p>Proportion (%)</p> <p>Années</p> <p>Modalités</p> <ul style="list-style-type: none"> 0 1 2 3 -1
Remarque	À partir de 2019, la modalité -1 semble remplacer 0. Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».

k. an_nais

Description	Année de naissance de l'usager.										
Type	[2005-2018 ; 2021-2022] : float64 [2019-2020] : int64										
Etendue des valeurs	<table> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>an_nais</td> <td>2628018</td> <td>127</td> <td>1988.0</td> <td>66282</td> </tr> </tbody> </table>		count	unique	top	freq	an_nais	2628018	127	1988.0	66282
	count	unique	top	freq							
an_nais	2628018	127	1988.0	66282							
Valeurs nulles	<table> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>an_nais</td> <td>float64</td> <td>2628018</td> <td>8359</td> <td>0.32</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	an_nais	float64	2628018	8359	0.32
	Type	Val_notnull	Val_null	%_null							
an_nais	float64	2628018	8359	0.32							
Outliers	<table> <thead> <tr> <th></th> <th>outliers_count</th> <th>outliers_unique</th> <th>outliers_list</th> </tr> </thead> <tbody> <tr> <td>an_nais</td> <td>13290</td> <td>28</td> <td>[1896.0, 1897.0, 1898.0, 1899.0, 1900.0, 1901...]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	an_nais	13290	28	[1896.0, 1897.0, 1898.0, 1899.0, 1900.0, 1901...]		
	outliers_count	outliers_unique	outliers_list								
an_nais	13290	28	[1896.0, 1897.0, 1898.0, 1899.0, 1900.0, 1901...]								



Répartition

Count % valeurs

an_nais

1896.0	1	0.0
1897.0	3	0.0
1898.0	35	0.0
1899.0	2	0.0
1900.0	286	0.0
...
2019.0	1179	0.0
2020.0	758	0.0
2021.0	525	0.0
2022.0	198	0.0
NaN	8359	0.0

128 rows × 2 columns

Remarque

Attention aux outliers très bas, probablement le format de saisie (pas la date entière).

I. num_veh

Description	Identifiant du véhicule repris pour chacun des usagers occupant ce véhicule - code alphanumérique.
Type	object

Etendue des valeurs	count	unique	top	freq
	num_veh	2636377	181	A01 1601497
Valeurs nulles	Type	Val_notnull	Val_null	%_null
	num_veh	object	2636377	0 0.0
Outliers	outliers_count	outliers_unique		
	num_veh	63930	177	[A02, A03, A04, A05, A06, A07, A08, A09, A27,...
Répartition	Count	% valeurs		
	num_veh			
	A01	1601497	61.0	
	A02	589	0.0	
	A03	49	0.0	
	A04	7	0.0	
	A05	5	0.0	
	
	Z01	1	0.0	
	ZZ01	5	0.0	
181 rows × 2 columns				

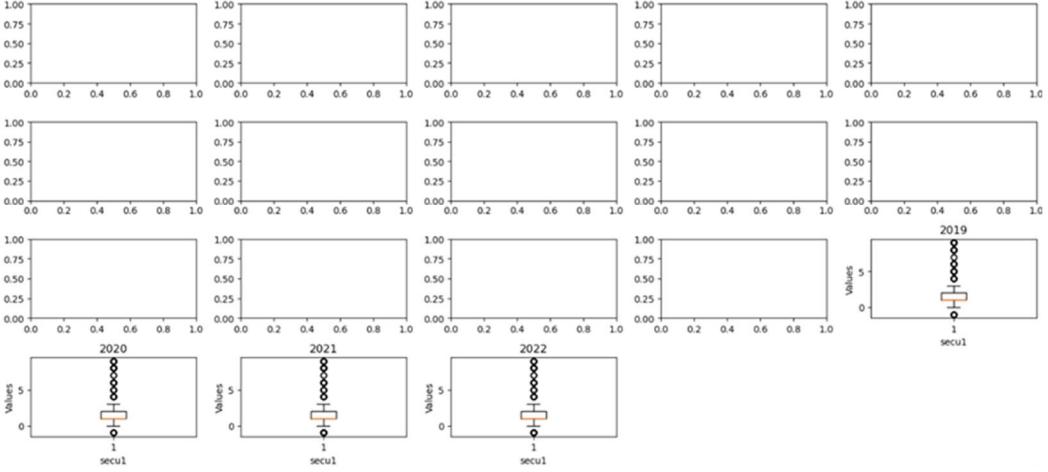
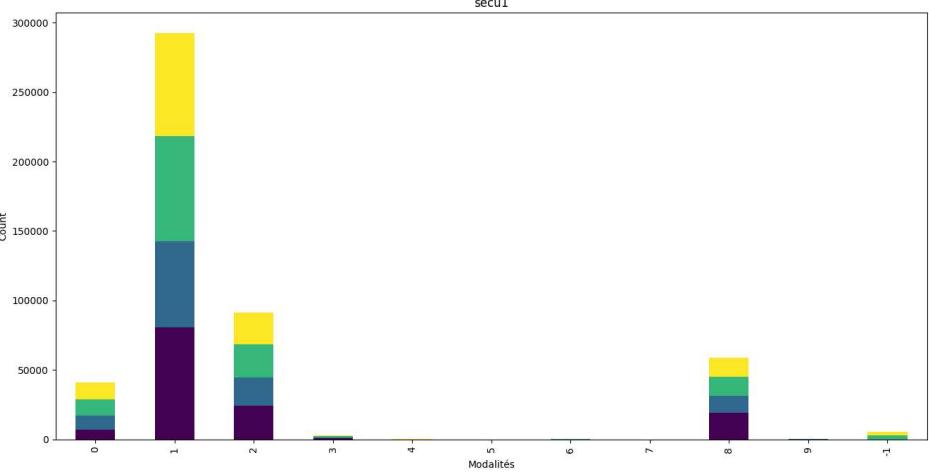
m.id_véhicule

Description	Identifiant unique du véhicule repris pour chacun des usagers occupant ce véhicule - code numérique.
Type	[2019-2022] : object
Etendue des valeurs	count
	unique
Valeurs nulles	top
	freq
Valeurs nulles	Type
	Val_notnull
Valeurs nulles	Val_null
	%_null
Valeurs nulles	id_véhicule
	object
Valeurs nulles	494182
	2142195
Valeurs nulles	81.26

Outliers	outliers_count outliers_unique			outliers_list
	id_vehicule	494182	369639	
Répartition	Count % valeurs			
	id_vehicule			
	100 882	1	0.0	
	100 883	1	0.0	
	100 884	1	0.0	
	100 885	1	0.0	
	100 886	1	0.0	
	
	813 950	1	0.0	
	813 951	1	0.0	
	813 952	1	0.0	
	813 953	1	0.0	
	Nan	2142195	81.0	
	369640 rows × 2 columns			

n. secu1

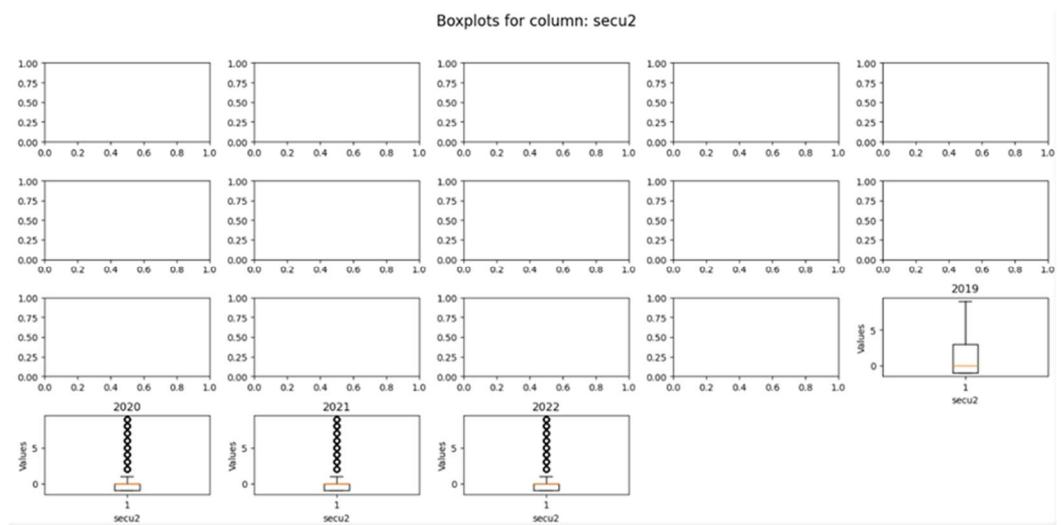
Description	Le renseignement du caractère indique la présence et l'utilisation de l'équipement de sécurité.										
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 0 : Aucun équipement - 1 : Ceinture - 2 : Casque - 3 : Dispositif enfants - 4 : Gilet réfléchissant - 5 : Airbag (2RM/3RM) - 6 : Gants (2RM/3RM) - 7 : Gants + Airbag (2RM/3RM) - 8 : Non déterminable - 9 : Autre 										
Type	[2019-2022] : int64										
Etendue des valeurs	<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>secu1</td><td>494182</td><td>11</td><td>1.0</td><td>292332</td></tr> </tbody> </table>		count	unique	top	freq	secu1	494182	11	1.0	292332
	count	unique	top	freq							
secu1	494182	11	1.0	292332							

Valeurs nulles	Type	Val_notnull	Val_null	%_null
	secu1	float64	494182	2142195 81.26
Outliers	outliers_count	outliers_unique	outliers_list	
	secu1	66208	7	[-1.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]
Boxplots for column: secu1				
				
Répartition	Count	% valeurs		
	Modalité			
-1.0	5591	1.0		
0.0	41182	8.0		
1.0	292332	59.0		
2.0	91302	18.0		
3.0	3158	1.0		
4.0	334	0.0		
5.0	219	0.0		
6.0	399	0.0		
7.0	15	0.0		
8.0	59115	12.0		
9.0	535	0.0		
				

Evolution	<p style="text-align: center;">Evolution de la distribution secu1</p>
Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».

o. secu2

Description	Le renseignement du caractère indique la présence et l'utilisation de l'équipement de sécurité.										
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 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 										
Type	int64										
Etendue des valeurs	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: left;">count</th> <th style="text-align: left;">unique</th> <th style="text-align: left;">top</th> <th style="text-align: left;">freq</th> </tr> </thead> <tbody> <tr> <td>secu2</td> <td>494182</td> <td>11</td> <td>-1.0</td> <td>193509</td> </tr> </tbody> </table>		count	unique	top	freq	secu2	494182	11	-1.0	193509
	count	unique	top	freq							
secu2	494182	11	-1.0	193509							
Valeurs nulles	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: left;">Type</th> <th style="text-align: left;">Val_notnull</th> <th style="text-align: left;">Val_null</th> <th style="text-align: left;">%_null</th> </tr> </thead> <tbody> <tr> <td>secu2</td> <td>float64</td> <td>494182</td> <td>2142195</td> <td>81.26</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	secu2	float64	494182	2142195	81.26
	Type	Val_notnull	Val_null	%_null							
secu2	float64	494182	2142195	81.26							
Outliers	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: left;">outliers_count</th> <th style="text-align: left;">outliers_unique</th> <th style="text-align: left;">outliers_list</th> </tr> </thead> <tbody> <tr> <td>secu2</td> <td>109988</td> <td>8</td> <td>[2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	secu2	109988	8	[2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]		
	outliers_count	outliers_unique	outliers_list								
secu2	109988	8	[2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]								

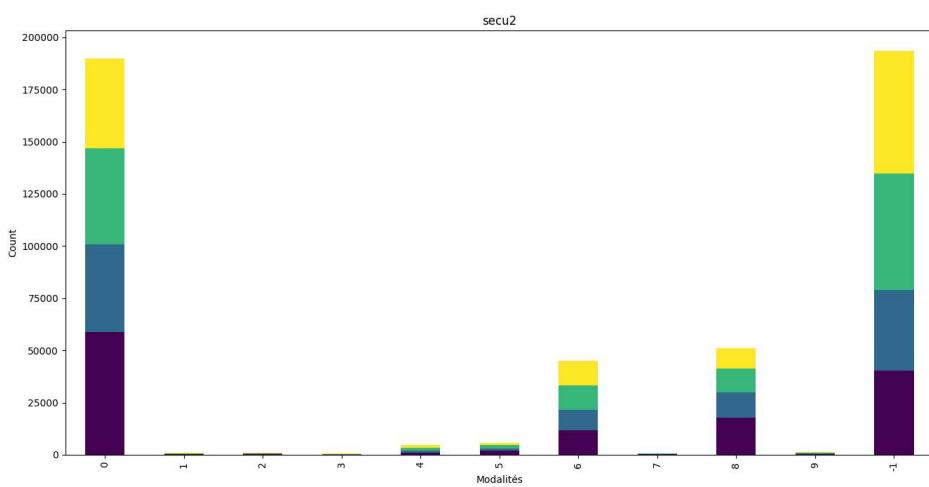


Répartition

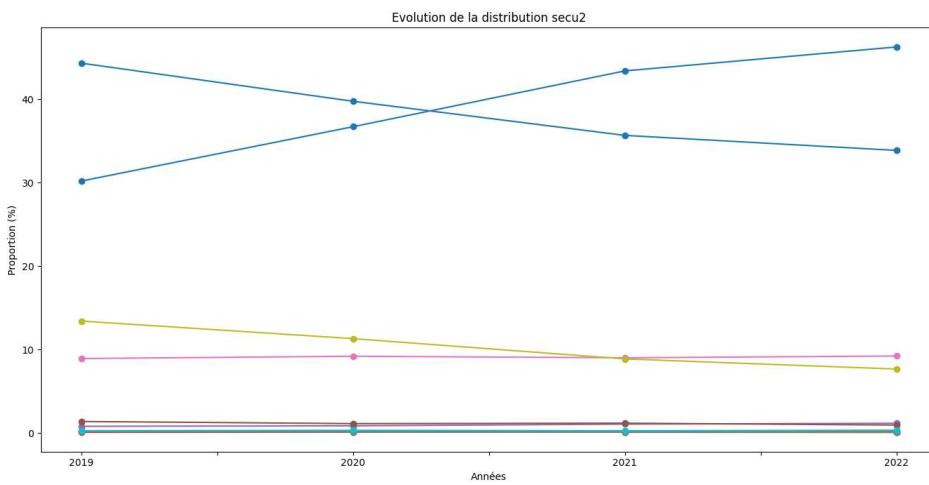
Count % valeurs

Modalité

-1.0	193509	39.0
0.0	189789	38.0
1.0	896	0.0
2.0	841	0.0
3.0	555	0.0
4.0	4845	1.0
5.0	5771	1.0
6.0	44898	9.0
7.0	656	0.0
8.0	50941	10.0
9.0	1481	0.0



Evolution



Remarque

Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».

p. secu3

Description	Le renseignement du caractère indique la présence et l'utilisation de l'équipement de sécurité :										
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 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 										
Type	int64										
Etendue des valeurs	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">count</th><th style="text-align: center;">unique</th><th style="text-align: center;">top</th><th style="text-align: center;">freq</th></tr> </thead> <tbody> <tr> <td>secu3</td><td style="text-align: center;">494182</td><td style="text-align: center;">11</td><td style="text-align: center;">-1.0</td><td style="text-align: center;">488588</td></tr> </tbody> </table>		count	unique	top	freq	secu3	494182	11	-1.0	488588
	count	unique	top	freq							
secu3	494182	11	-1.0	488588							
Valeurs nulles	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">Type</th><th style="text-align: center;">Val_notnull</th><th style="text-align: center;">Val_null</th><th style="text-align: center;">%_null</th></tr> </thead> <tbody> <tr> <td>secu3</td><td style="text-align: center;">float64</td><td style="text-align: center;">494182</td><td style="text-align: center;">2142195</td><td style="text-align: center;">81.26</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	secu3	float64	494182	2142195	81.26
	Type	Val_notnull	Val_null	%_null							
secu3	float64	494182	2142195	81.26							
Outliers	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">outliers_count</th><th style="text-align: center;">outliers_unique</th><th style="text-align: center;">outliers_list</th></tr> </thead> <tbody> <tr> <td>secu3</td><td style="text-align: center;">5594</td><td style="text-align: center;">10</td><td style="text-align: center;">[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0,...</td></tr> </tbody> </table> <p style="text-align: center;">Boxplots for column: secu3</p>		outliers_count	outliers_unique	outliers_list	secu3	5594	10	[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0,...		
	outliers_count	outliers_unique	outliers_list								
secu3	5594	10	[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0,...								

<h3>Répartition</h3> <table border="1"> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr><td>Modalité</td><td></td><td></td></tr> <tr><td>-1.0</td><td>488588</td><td>99.0</td></tr> <tr><td>0.0</td><td>1330</td><td>0.0</td></tr> <tr><td>1.0</td><td>87</td><td>0.0</td></tr> <tr><td>2.0</td><td>15</td><td>0.0</td></tr> <tr><td>3.0</td><td>7</td><td>0.0</td></tr> <tr><td>4.0</td><td>71</td><td>0.0</td></tr> <tr><td>5.0</td><td>38</td><td>0.0</td></tr> <tr><td>6.0</td><td>250</td><td>0.0</td></tr> <tr><td>7.0</td><td>12</td><td>0.0</td></tr> <tr><td>8.0</td><td>235</td><td>0.0</td></tr> <tr><td>9.0</td><td>3549</td><td>1.0</td></tr> </tbody> </table>		Count	% valeurs	Modalité			-1.0	488588	99.0	0.0	1330	0.0	1.0	87	0.0	2.0	15	0.0	3.0	7	0.0	4.0	71	0.0	5.0	38	0.0	6.0	250	0.0	7.0	12	0.0	8.0	235	0.0	9.0	3549	1.0	
	Count	% valeurs																																						
Modalité																																								
-1.0	488588	99.0																																						
0.0	1330	0.0																																						
1.0	87	0.0																																						
2.0	15	0.0																																						
3.0	7	0.0																																						
4.0	71	0.0																																						
5.0	38	0.0																																						
6.0	250	0.0																																						
7.0	12	0.0																																						
8.0	235	0.0																																						
9.0	3549	1.0																																						
<h3>Evolution</h3>																																								
Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».																																							

q. id_usager

Description	Identifiant unique de l'usager - code numérique.										
Type	object										
Etendue des valeurs	<table border="1"> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>id_usager</td> <td>255910</td> <td>255910</td> <td>133 818</td> <td>1</td> </tr> </tbody> </table>		count	unique	top	freq	id_usager	255910	255910	133 818	1
	count	unique	top	freq							
id_usager	255910	255910	133 818	1							
Valeurs nulles	<table border="1"> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>id_usager</td> <td>object</td> <td>255910</td> <td>2380467</td> <td>90.29</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	id_usager	object	255910	2380467	90.29
	Type	Val_notnull	Val_null	%_null							
id_usager	object	255910	2380467	90.29							

Outliers	outliers_count outliers_unique			outliers_list
	id_usager	255910	255910 [133 818, 133 819, 133 820, 133 821, 133 822,...	
Répartition	Count % valeurs			
	id_usager			
	133 818	1	0.0	
	133 819	1	0.0	
	133 820	1	0.0	
	133 821	1	0.0	
	133 822	1	0.0	
	
	999 994	1	0.0	
	999 996	1	0.0	
	999 998	1	0.0	
	999 999	1	0.0	
NaN 2380467 90.0				
255911 rows × 2 columns				

4. Véhicules

Rows x columns Rows duplicated

Véhicules (2009395, 11) 0

a. Num_Acc

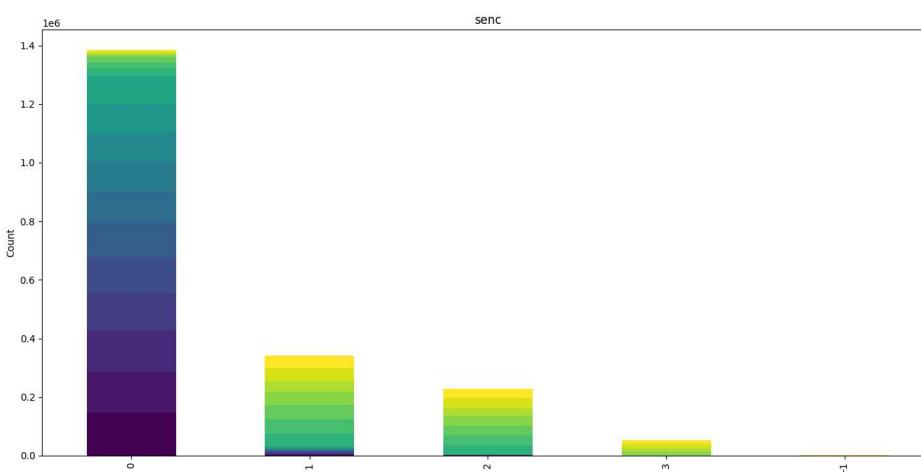
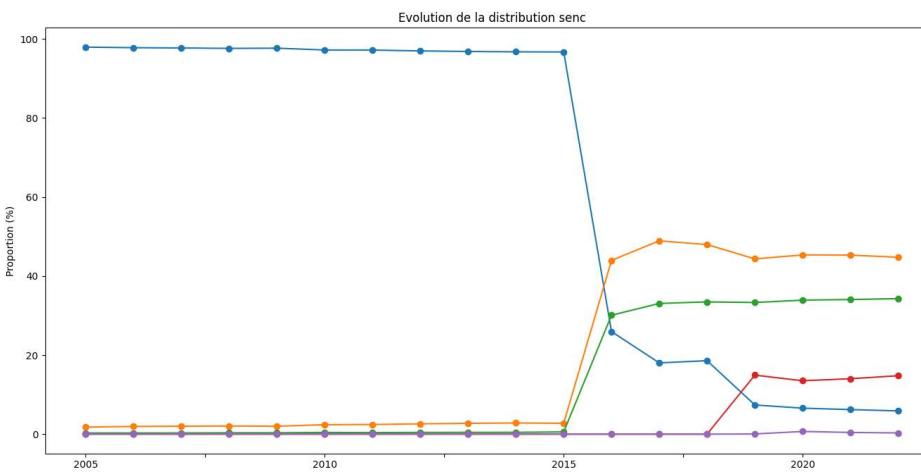
Description	Identifiant de l'accident identique à celui du fichier "rubrique CARACTERISTIQUES" repris pour chacun des véhicules décrits impliqués dans l'accident.
Type	int64
Etendue des valeurs	count unique top freq
	Num_Acc 2009395 1176873 200600074917 56
Valeurs nulles	Type Val_notnull Val_null %_null
	Num_Acc int64 2009395 0 0.0

Outliers	outliers_count outliers_unique outliers_list		
	Num_Acc	0	0
Répartition	Count % valeurs		
	Num_Acc		
	200500000001	2	0.0
	200500000002	2	0.0
	200500000003	2	0.0
	200500000004	3	0.0
	200500000005	1	0.0

	202200055298	1	0.0
	202200055299	1	0.0
	202200055300	1	0.0
	202200055301	2	0.0
	202200055302	2	0.0
	1176873 rows × 2 columns		

b. senc

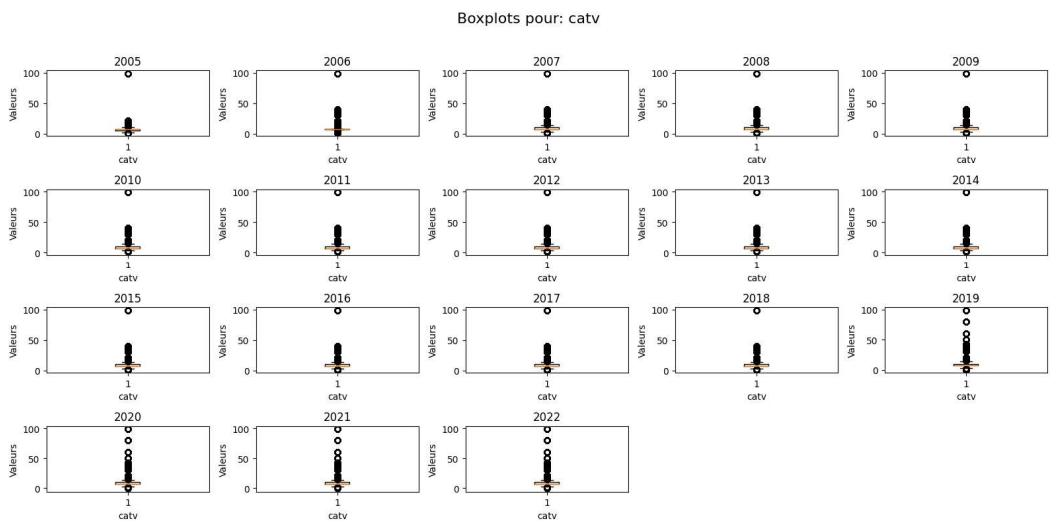
Description	Sens de circulation.
Modalités	- -1 : Non renseigné - 0 : Inconnu - 1 : PK ou PR ou numéro d'adresse postale croissant - 2 : PK ou PR ou numéro d'adresse postale décroissant - 3 : Absence de repère
Type	[2005-2015 ; 2019-2022] : int64 [2016-2018] : float64
Etendue des valeurs	count unique top freq
	senc 2009123 5 0.0 1384153
Valeurs nulles	Type Val_notnull Val_null %_null
	senc float64 2009123 272 0.01

Outliers	outliers_count outliers_unique outliers_list																																	
	senc	53593	1 [3.0]																															
Répartition		 <p>A stacked bar chart titled "Répartition" showing the count of senc values from 0 to 3 across years from 2005 to 2022. The total count is 53593. The x-axis represents Modalités (0, 1, 2, 3) and the y-axis represents count (0.0 to 1.4e6). A color scale legend on the right indicates the years from 2005 to 2022, with darker shades representing earlier years.</p> <table border="1"> <thead> <tr> <th>Modalité</th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>-1.0</td> <td>1287</td> <td>0.0</td> </tr> <tr> <td>0.0</td> <td>1384153</td> <td>69.0</td> </tr> <tr> <td>1.0</td> <td>340696</td> <td>17.0</td> </tr> <tr> <td>2.0</td> <td>229394</td> <td>11.0</td> </tr> <tr> <td>3.0</td> <td>53593</td> <td>3.0</td> </tr> <tr> <td>NaN</td> <td>272</td> <td>0.0</td> </tr> </tbody> </table>			Modalité	Count	% valeurs	-1.0	1287	0.0	0.0	1384153	69.0	1.0	340696	17.0	2.0	229394	11.0	3.0	53593	3.0	NaN	272	0.0									
Modalité	Count	% valeurs																																
-1.0	1287	0.0																																
0.0	1384153	69.0																																
1.0	340696	17.0																																
2.0	229394	11.0																																
3.0	53593	3.0																																
NaN	272	0.0																																
Evolution		 <p>A line chart titled "Evolution de la distribution senc" showing the proportion (%) of senc values from 0 to 3 from 2005 to 2022. The proportion for modalité 0 remained at 100% until 2015, then dropped sharply to around 10%. Modalité 1 increased from 0% to about 45%. Modalité 2 increased from 0% to about 35%. Modalité 3 increased from 0% to about 15%.</p> <table border="1"> <thead> <tr> <th>Années</th> <th>0</th> <th>1</th> <th>2</th> <th>3</th> </tr> </thead> <tbody> <tr> <td>2005</td> <td>100</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>2010</td> <td>100</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>2015</td> <td>100</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>2020</td> <td>10</td> <td>45</td> <td>35</td> <td>15</td> </tr> <tr> <td>2022</td> <td>10</td> <td>45</td> <td>35</td> <td>15</td> </tr> </tbody> </table>			Années	0	1	2	3	2005	100	0	0	0	2010	100	0	0	0	2015	100	0	0	0	2020	10	45	35	15	2022	10	45	35	15
Années	0	1	2	3																														
2005	100	0	0	0																														
2010	100	0	0	0																														
2015	100	0	0	0																														
2020	10	45	35	15																														
2022	10	45	35	15																														
Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».																																	

c. catv

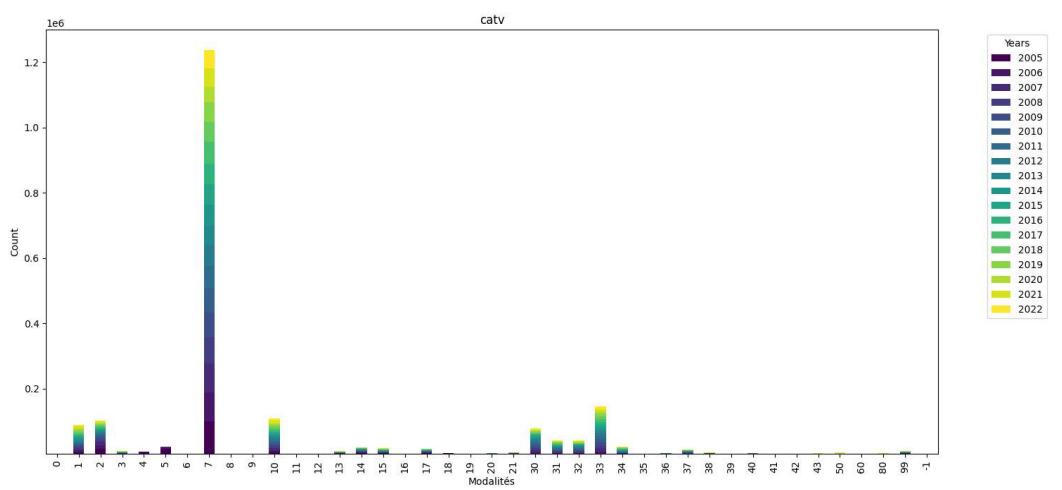
Description	Catégorie du véhicule.
Modalités	<ul style="list-style-type: none"> - 00 : Indéterminable - 01 : Bicyclette - 02 : Cyclomoteur <50cm3 - 03 : Voiturette (Quadricycle à moteur carrossé) - 04 : Référence inutilisée depuis 2006 (scooter immatriculé) - 05 : Référence inutilisée depuis 2006 (motocyclette) - 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 - 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 : VAE - 99 : Autre véhicule										
Type	int64										
Etendue des valeurs	<table> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>catv</td><td>2009395</td><td>41</td><td>7</td><td>1237634</td></tr> </tbody> </table>		count	unique	top	freq	catv	2009395	41	7	1237634
	count	unique	top	freq							
catv	2009395	41	7	1237634							
Valeurs nulles	<table> <thead> <tr> <th></th><th>Type</th><th>Val_notnull</th><th>Val_null</th><th>%_null</th></tr> </thead> <tbody> <tr> <td>catv</td><td>int64</td><td>2009395</td><td>0</td><td>0.0</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	catv	int64	2009395	0	0.0
	Type	Val_notnull	Val_null	%_null							
catv	int64	2009395	0	0.0							
Outliers	<table> <thead> <tr> <th></th><th>outliers_count</th><th>outliers_unique</th><th>outliers_list</th></tr> </thead> <tbody> <tr> <td>catv</td><td>599066</td><td>29</td><td>[-1.0, 0.0, 1.0, 2.0, 15.0, 16.0, 17.0, 18.0,...</td></tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	catv	599066	29	[-1.0, 0.0, 1.0, 2.0, 15.0, 16.0, 17.0, 18.0,...		
	outliers_count	outliers_unique	outliers_list								
catv	599066	29	[-1.0, 0.0, 1.0, 2.0, 15.0, 16.0, 17.0, 18.0,...								



Répartition

	Count	% valeurs
catv		
-1	13	0.0
0	1030	0.05
1	88885	4.42
2	101713	5.06
3	8072	0.4
...
43	1995	0.1
50	5116	0.25
60	754	0.04
80	1793	0.09
99	8283	0.41
41 rows × 2 columns		

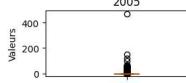
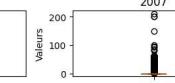
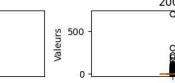
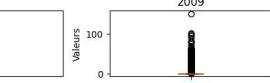
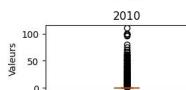
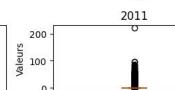
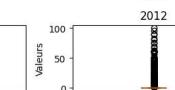
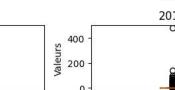
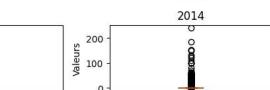
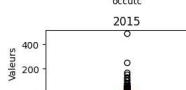
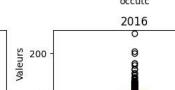
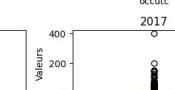
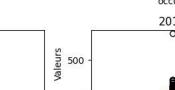
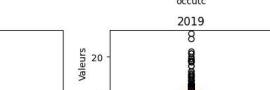
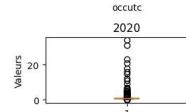
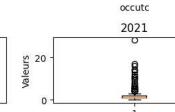
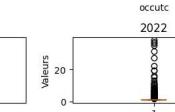


Remarque

On a une apparition de plus de modalités à partir de 2019.

d. occutc

Description	Nombre d'occupants dans le transport en commun.
Type	[2005-2018] : int64 [2019-2022] : float64

Etendue des valeurs		count	unique	top	freq
	occutc	1638885	124	0.0	1624683
Valeurs nulles		Type	Val_notnull	Val_null	%_null
	occutc	float64	1638885	370510	18.44
Outliers		outliers_count	outliers_unique		outliers_list
	occutc	14202	123	[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0,...	
	Boxplots pour: occutc				
					
					
					
					
Répartition		Count	% valeurs		
	occutc				
	0.0	1624683	81.0		
	1.0	7699	0.0		
	2.0	1127	0.0		
	3.0	581	0.0		
	4.0	301	0.0		
		
	480.0	1	0.0		
	490.0	1	0.0		
	700.0	1	0.0		
	900.0	1	0.0		
	NaN	370510	18.0		
	125 rows × 2 columns				

Remarque	Beaucoup de valeurs manquantes et certaines valeurs sont aberrantes (trop de passagers).
-----------------	--

e. obs

Description	Obstacle fixe heurté.										
Modalités	<ul style="list-style-type: none"> - -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 										
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64										
Etendue des valeurs	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">count</th> <th style="text-align: center;">unique</th> <th style="text-align: center;">top</th> <th style="text-align: center;">freq</th> </tr> </thead> <tbody> <tr> <td>obs</td> <td style="text-align: center;">2008389</td> <td style="text-align: center;">19</td> <td style="text-align: center;">0.0</td> <td style="text-align: center;">1740874</td> </tr> </tbody> </table>		count	unique	top	freq	obs	2008389	19	0.0	1740874
	count	unique	top	freq							
obs	2008389	19	0.0	1740874							
Valeurs nulles	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">Type</th> <th style="text-align: center;">Val_notnull</th> <th style="text-align: center;">Val_null</th> <th style="text-align: center;">%_null</th> </tr> </thead> <tbody> <tr> <td>obs</td> <td style="text-align: center;">float64</td> <td style="text-align: center;">2008389</td> <td style="text-align: center;">1006</td> <td style="text-align: center;">0.05</td> </tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	obs	float64	2008389	1006	0.05
	Type	Val_notnull	Val_null	%_null							
obs	float64	2008389	1006	0.05							
Outliers	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">outliers_count</th> <th style="text-align: center;">outliers_unique</th> <th style="text-align: center;">outliers_list</th> </tr> </thead> <tbody> <tr> <td>obs</td> <td style="text-align: center;">267515</td> <td style="text-align: center;">18</td> <td style="text-align: center;">[-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]</td> </tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	obs	267515	18	[-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]		
	outliers_count	outliers_unique	outliers_list								
obs	267515	18	[-1.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0...]								

Répartition		
	Count	% valeurs
obs		
-1.0	164	0.0
0.0	1740874	87.0
1.0	44192	2.0
2.0	28984	1.0
3.0	23176	1.0
4.0	23786	1.0
5.0	2925	0.0
6.0	22306	1.0
7.0	4818	0.0
8.0	21345	1.0
9.0	7049	0.0
10.0	2320	0.0
11.0	4577	0.0
12.0	12122	1.0
13.0	34044	2.0
14.0	14764	1.0
15.0	9904	0.0
16.0	10625	1.0
17.0	414	0.0
NaN	1006	0.0

Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».
----------	--

f. obsm

Description	Obstacle mobile heurté
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 0 : Aucun - 1 : Piéton - 2 : Véhicule - 4 : Véhicule sur rail - 5 : Animal domestique - 6 : Animal sauvage - 9 : Autre

Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64																														
Etendue des valeurs	count unique top freq obsm 2008617 8 2.0 1352641																														
Valeurs nulles	Type Val_notnull Val_null %_null obsm float64 2008617 778 0.04																														
Outliers	outliers_count outliers_unique outliers_list obsm 36611 5 [-1.0, 4.0, 5.0, 6.0, 9.0]																														
Répartition	<p>Count % valeurs</p> <table border="1"> <thead> <tr> <th></th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>-1.0</td> <td>199</td> <td>0.0</td> </tr> <tr> <td>0.0</td> <td>416695</td> <td>21.0</td> </tr> <tr> <td>1.0</td> <td>202670</td> <td>10.0</td> </tr> <tr> <td>2.0</td> <td>1352641</td> <td>67.0</td> </tr> <tr> <td>4.0</td> <td>1889</td> <td>0.0</td> </tr> <tr> <td>5.0</td> <td>1826</td> <td>0.0</td> </tr> <tr> <td>6.0</td> <td>3840</td> <td>0.0</td> </tr> <tr> <td>9.0</td> <td>28857</td> <td>1.0</td> </tr> <tr> <td>NaN</td> <td>778</td> <td>0.0</td> </tr> </tbody> </table>		Count	% valeurs	-1.0	199	0.0	0.0	416695	21.0	1.0	202670	10.0	2.0	1352641	67.0	4.0	1889	0.0	5.0	1826	0.0	6.0	3840	0.0	9.0	28857	1.0	NaN	778	0.0
	Count	% valeurs																													
-1.0	199	0.0																													
0.0	416695	21.0																													
1.0	202670	10.0																													
2.0	1352641	67.0																													
4.0	1889	0.0																													
5.0	1826	0.0																													
6.0	3840	0.0																													
9.0	28857	1.0																													
NaN	778	0.0																													
Evolution	<p>Evolution de la distribution obsm</p>																														
Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».																														

g. choc

Description	Point de choc initial.																																										
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 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) 																																										
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64																																										
Etendue des valeurs	<table> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>choc</td><td>2008998</td><td>11</td><td>1.0</td><td>738510</td></tr> </tbody> </table>		count	unique	top	freq	choc	2008998	11	1.0	738510																																
	count	unique	top	freq																																							
choc	2008998	11	1.0	738510																																							
Valeurs nulles	<table> <thead> <tr> <th></th><th>Type</th><th>Val_notnull</th><th>Val_null</th><th>%_null</th></tr> </thead> <tbody> <tr> <td>choc</td><td>float64</td><td>2008998</td><td>397</td><td>0.02</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	choc	float64	2008998	397	0.02																																
	Type	Val_notnull	Val_null	%_null																																							
choc	float64	2008998	397	0.02																																							
Outliers	<table> <thead> <tr> <th></th><th>outliers_count</th><th>outliers_unique</th><th>outliers_list</th></tr> </thead> <tbody> <tr> <td>choc</td><td>31156</td><td>1</td><td>[9.0]</td></tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	choc	31156	1	[9.0]																																		
	outliers_count	outliers_unique	outliers_list																																								
choc	31156	1	[9.0]																																								
Répartition	<table> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>choc</td><td></td><td></td></tr> <tr> <td>-1.0</td><td>197</td><td>0.0</td></tr> <tr> <td>0.0</td><td>133592</td><td>7.0</td></tr> <tr> <td>1.0</td><td>738510</td><td>37.0</td></tr> <tr> <td>2.0</td><td>235373</td><td>12.0</td></tr> <tr> <td>3.0</td><td>289599</td><td>14.0</td></tr> <tr> <td>4.0</td><td>190557</td><td>9.0</td></tr> <tr> <td>5.0</td><td>53591</td><td>3.0</td></tr> <tr> <td>6.0</td><td>67943</td><td>3.0</td></tr> <tr> <td>7.0</td><td>122529</td><td>6.0</td></tr> <tr> <td>8.0</td><td>145951</td><td>7.0</td></tr> <tr> <td>9.0</td><td>31156</td><td>2.0</td></tr> <tr> <td>NaN</td><td>397</td><td>0.0</td></tr> </tbody> </table> <p>The heatmap displays the distribution of 'choc' events across different modalities and years. The x-axis represents the modality (Modalités) from 0 to 11. The y-axis represents the year (Years) from 2005 to 2022. The color intensity indicates the count of events, with a legend on the right showing a gradient from dark purple (0) to bright yellow (over 700,000). The highest frequency is for Modality 1 (Avant) in 2019-2022, followed by Modality 3 (Avant gauche) and Modality 2 (Avant droit).</p>		Count	% valeurs	choc			-1.0	197	0.0	0.0	133592	7.0	1.0	738510	37.0	2.0	235373	12.0	3.0	289599	14.0	4.0	190557	9.0	5.0	53591	3.0	6.0	67943	3.0	7.0	122529	6.0	8.0	145951	7.0	9.0	31156	2.0	NaN	397	0.0
	Count	% valeurs																																									
choc																																											
-1.0	197	0.0																																									
0.0	133592	7.0																																									
1.0	738510	37.0																																									
2.0	235373	12.0																																									
3.0	289599	14.0																																									
4.0	190557	9.0																																									
5.0	53591	3.0																																									
6.0	67943	3.0																																									
7.0	122529	6.0																																									
8.0	145951	7.0																																									
9.0	31156	2.0																																									
NaN	397	0.0																																									

Evolution	<p style="text-align: center;">Evolution de la distribution choc</p> <p style="text-align: right;">Modalités</p> <ul style="list-style-type: none"> 0 1 2 3 4 5 6 7 8 9 -1
Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».

h. manv

Description	Manoeuvre principale avant l'accident.
Modalités	<ul style="list-style-type: none"> - -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 - 11 : Changeant de file A gauche - 12 : Changeant de file A droite - 13 : Déporté A gauche - 14 : Déporté A droite - 15 : Tournant A gauche - 16 : Tournant A droite - 17 : Dépassant A gauche - 18 : Dépassant A droite - 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
Type	[2005-2008 ; 2019-2022] : int64 [2009-2018] : float64

Etendue des valeurs	<table border="1"> <thead> <tr> <th></th><th>count</th><th>unique</th><th>top</th><th>freq</th></tr> </thead> <tbody> <tr> <td>manv</td><td>2008927</td><td>28</td><td>1.0</td><td>863725</td></tr> </tbody> </table>		count	unique	top	freq	manv	2008927	28	1.0	863725																													
	count	unique	top	freq																																				
manv	2008927	28	1.0	863725																																				
Valeurs nulles	<table border="1"> <thead> <tr> <th></th><th>Type</th><th>Val_notnull</th><th>Val_null</th><th>%_null</th></tr> </thead> <tbody> <tr> <td>manv</td><td>float64</td><td>2008927</td><td>468</td><td>0.02</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	manv	float64	2008927	468	0.02																													
	Type	Val_notnull	Val_null	%_null																																				
manv	float64	2008927	468	0.02																																				
Outliers	<table border="1"> <thead> <tr> <th></th><th>outliers_count</th><th>outliers_unique</th><th>outliers_list</th></tr> </thead> <tbody> <tr> <td>manv</td><td>0</td><td>0</td><td>[]</td></tr> </tbody> </table> <p style="text-align: center;">Boxplots pour: manv</p>		outliers_count	outliers_unique	outliers_list	manv	0	0	[]																															
	outliers_count	outliers_unique	outliers_list																																					
manv	0	0	[]																																					
Répartition <table border="1"> <thead> <tr> <th></th><th>Count</th><th>% valeurs</th></tr> </thead> <tbody> <tr> <td>manv</td><td></td><td></td></tr> <tr> <td>-1.0</td><td>142</td><td>0.01</td></tr> <tr> <td>0.0</td><td>156498</td><td>7.79</td></tr> <tr> <td>1.0</td><td>863725</td><td>42.98</td></tr> <tr> <td>2.0</td><td>233007</td><td>11.6</td></tr> <tr> <td>3.0</td><td>15728</td><td>0.78</td></tr> <tr> <td>...</td><td>...</td><td>...</td></tr> <tr> <td>23.0</td><td>52994</td><td>2.64</td></tr> <tr> <td>24.0</td><td>7274</td><td>0.36</td></tr> <tr> <td>25.0</td><td>1040</td><td>0.05</td></tr> <tr> <td>26.0</td><td>12678</td><td>0.63</td></tr> <tr> <td>Nan</td><td>468</td><td>0.02</td></tr> </tbody> </table> <p>29 rows x 2 columns</p>		Count	% valeurs	manv			-1.0	142	0.01	0.0	156498	7.79	1.0	863725	42.98	2.0	233007	11.6	3.0	15728	0.78	23.0	52994	2.64	24.0	7274	0.36	25.0	1040	0.05	26.0	12678	0.63	Nan	468	0.02	<p>Years</p> <ul style="list-style-type: none"> 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
	Count	% valeurs																																						
manv																																								
-1.0	142	0.01																																						
0.0	156498	7.79																																						
1.0	863725	42.98																																						
2.0	233007	11.6																																						
3.0	15728	0.78																																						
...																																						
23.0	52994	2.64																																						
24.0	7274	0.36																																						
25.0	1040	0.05																																						
26.0	12678	0.63																																						
Nan	468	0.02																																						

Remarque	Les valeurs NaN peuvent être remplacées par -1 qui signifie « non renseigné ».
-----------------	--

i. num_veh

Description	Identifiant du véhicule repris pour chacun des usagers occupant ce véhicule (y compris les piétons qui sont rattachés aux véhicules qui les ont heurtés) - Code alphanumérique.																																											
Type	object																																											
Etendue des valeurs	<table> <thead> <tr> <th></th> <th>count</th> <th>unique</th> <th>top</th> <th>freq</th> </tr> </thead> <tbody> <tr> <td>num_veh</td> <td>2009395</td> <td>189</td> <td>A01</td> <td>1160074</td> </tr> </tbody> </table>						count	unique	top	freq	num_veh	2009395	189	A01	1160074																													
	count	unique	top	freq																																								
num_veh	2009395	189	A01	1160074																																								
Valeurs nulles	<table> <thead> <tr> <th></th> <th>Type</th> <th>Val_notnull</th> <th>Val_null</th> <th>%_null</th> </tr> </thead> <tbody> <tr> <td>num_veh</td> <td>object</td> <td>2009395</td> <td>0</td> <td>0.0</td> </tr> </tbody> </table>						Type	Val_notnull	Val_null	%_null	num_veh	object	2009395	0	0.0																													
	Type	Val_notnull	Val_null	%_null																																								
num_veh	object	2009395	0	0.0																																								
Outliers	<table> <thead> <tr> <th></th> <th>outliers_count</th> <th>outliers_unique</th> <th colspan="2">outliers_list</th> </tr> </thead> <tbody> <tr> <td>num_veh</td> <td>40101</td> <td>184</td> <td colspan="2" rowspan="2">[A02, A03, A04, A05, A06, A07, A08, A09, A27,...</td> </tr> </tbody> </table>						outliers_count	outliers_unique	outliers_list		num_veh	40101	184	[A02, A03, A04, A05, A06, A07, A08, A09, A27,...																														
	outliers_count	outliers_unique	outliers_list																																									
num_veh	40101	184	[A02, A03, A04, A05, A06, A07, A08, A09, A27,...																																									
Répartition	<table> <thead> <tr> <th></th> <th>Count</th> <th>% valeurs</th> </tr> </thead> <tbody> <tr> <td>num_veh</td> <td></td> <td></td> </tr> <tr> <td>A01</td> <td>1160074</td> <td>58.0</td> </tr> <tr> <td>A02</td> <td>527</td> <td>0.0</td> </tr> <tr> <td>A03</td> <td>33</td> <td>0.0</td> </tr> <tr> <td>A04</td> <td>7</td> <td>0.0</td> </tr> <tr> <td>A05</td> <td>6</td> <td>0.0</td> </tr> <tr> <td>...</td> <td>...</td> <td>...</td> </tr> <tr> <td>ZB01</td> <td>1</td> <td>0.0</td> </tr> <tr> <td>ZZ01</td> <td>8</td> <td>0.0</td> </tr> <tr> <td>[01</td> <td>20</td> <td>0.0</td> </tr> <tr> <td>\01</td> <td>3</td> <td>0.0</td> </tr> <tr> <td>]01</td> <td>1</td> <td>0.0</td> </tr> </tbody> </table> <p>189 rows × 2 columns</p>						Count	% valeurs	num_veh			A01	1160074	58.0	A02	527	0.0	A03	33	0.0	A04	7	0.0	A05	6	0.0	ZB01	1	0.0	ZZ01	8	0.0	[01	20	0.0	\01	3	0.0]01	1	0.0
	Count	% valeurs																																										
num_veh																																												
A01	1160074	58.0																																										
A02	527	0.0																																										
A03	33	0.0																																										
A04	7	0.0																																										
A05	6	0.0																																										
...																																										
ZB01	1	0.0																																										
ZZ01	8	0.0																																										
[01	20	0.0																																										
\01	3	0.0																																										
]01	1	0.0																																										

j. id_vehicule

Description	Identifiant unique du véhicule repris pour chacun des usagers occupant ce véhicule - code numérique.
Type	object

Etendue des valeurs		count	unique	top	freq
	id_vehicule	373584	373584	100 882	1
Valeurs nulles	Type	Val_notnull	Val_null	%_null	
	num_veh	object	2009395	0	0.0
Valeurs uniques	Type	Val_notnull	Val_null	%_null	
	id_vehicule	object	373584	1635811	81.41
Outliers	outliers_count	outliers_unique			outliers_list
	373584	373584			[100 882, 100 883, 100 884, 100 885, 100 886,...
Répartition	Count	% valeurs			
	id_vehicule				
	100 882	1	0.0		
	100 883	1	0.0		
	100 884	1	0.0		
	100 885	1	0.0		
	100 886	1	0.0		
		
	813 950	1	0.0		
	813 951	1	0.0		
	813 952	1	0.0		
	813 953	1	0.0		
	Nan	1635811	81.0		
373585 rows × 2 columns					
Remarque	Apparaît à partir de 2019.				

k. motor

Description	Type de motorisation du véhicule.
Modalités	<ul style="list-style-type: none"> - -1 : Non renseigné - 0 : Inconnue - 1 : Hydrocarbures - 2 : Hybride électrique - 3 : Electrique

	<ul style="list-style-type: none"> - 4 : Hydrogène - 5 : Humaine - 6 : Autre 																																	
Type	[2019-2022] : int64																																	
Etendue des valeurs	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">count</th><th style="text-align: center;">unique</th><th style="text-align: center;">top</th><th style="text-align: center;">freq</th></tr> </thead> <tbody> <tr> <td>motor</td><td style="text-align: center;">373584</td><td style="text-align: center;">8</td><td style="text-align: center;">1.0</td><td style="text-align: center;">304898</td></tr> </tbody> </table>		count	unique	top	freq	motor	373584	8	1.0	304898																							
	count	unique	top	freq																														
motor	373584	8	1.0	304898																														
Valeurs nulles	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">Type</th><th style="text-align: center;">Val_notnull</th><th style="text-align: center;">Val_null</th><th style="text-align: center;">%_null</th></tr> </thead> <tbody> <tr> <td>motor</td><td style="text-align: center;">float64</td><td style="text-align: center;">373584</td><td style="text-align: center;">1635811</td><td style="text-align: center;">81.41</td></tr> </tbody> </table>		Type	Val_notnull	Val_null	%_null	motor	float64	373584	1635811	81.41																							
	Type	Val_notnull	Val_null	%_null																														
motor	float64	373584	1635811	81.41																														
Outliers	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">outliers_count</th><th style="text-align: center;">outliers_unique</th><th style="text-align: right; vertical-align: bottom;">outliers_list</th></tr> </thead> <tbody> <tr> <td>motor</td><td style="text-align: center;">68686</td><td style="text-align: center;">7</td><td style="text-align: right; vertical-align: bottom;">[-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0]</td></tr> </tbody> </table>		outliers_count	outliers_unique	outliers_list	motor	68686	7	[-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0]																									
	outliers_count	outliers_unique	outliers_list																															
motor	68686	7	[-1.0, 0.0, 2.0, 3.0, 4.0, 5.0, 6.0]																															
Répartition	<p>Count % valeurs</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th><th style="text-align: center;">Count</th><th style="text-align: center;">% valeurs</th></tr> </thead> <tbody> <tr> <td>motor</td><td></td><td></td></tr> <tr> <td>-1.0</td><td style="text-align: center;">865</td><td style="text-align: center;">0.0</td></tr> <tr> <td>0.0</td><td style="text-align: center;">28200</td><td style="text-align: center;">1.0</td></tr> <tr> <td>1.0</td><td style="text-align: center;">304898</td><td style="text-align: center;">15.0</td></tr> <tr> <td>2.0</td><td style="text-align: center;">5558</td><td style="text-align: center;">0.0</td></tr> <tr> <td>3.0</td><td style="text-align: center;">10997</td><td style="text-align: center;">1.0</td></tr> <tr> <td>4.0</td><td style="text-align: center;">191</td><td style="text-align: center;">0.0</td></tr> <tr> <td>5.0</td><td style="text-align: center;">19685</td><td style="text-align: center;">1.0</td></tr> <tr> <td>6.0</td><td style="text-align: center;">3190</td><td style="text-align: center;">0.0</td></tr> <tr> <td>Nan</td><td style="text-align: center;">1635811</td><td style="text-align: center;">81.0</td></tr> </tbody> </table>		Count	% valeurs	motor			-1.0	865	0.0	0.0	28200	1.0	1.0	304898	15.0	2.0	5558	0.0	3.0	10997	1.0	4.0	191	0.0	5.0	19685	1.0	6.0	3190	0.0	Nan	1635811	81.0
	Count	% valeurs																																
motor																																		
-1.0	865	0.0																																
0.0	28200	1.0																																
1.0	304898	15.0																																
2.0	5558	0.0																																
3.0	10997	1.0																																
4.0	191	0.0																																
5.0	19685	1.0																																
6.0	3190	0.0																																
Nan	1635811	81.0																																
Evolution	<p>Evolution de la distribution motor</p>																																	
Remarque	Les valeurs Nan peuvent être remplacées par -1 qui signifie « non »																																	

renseigné ».