***Projet SARA***

******

Projet réalisé par :

Cécile Pilon

Fadimatou Abdoulaye

Stéphane Maillard

Christophe Levra

Abdoulaye Tall

Encadré par :

Christophe - *Datascientest*



**REMERCIEMENTS**

Toute l’équipe projet SARA remercient DataScientest pour la qualité de la formation en Data Science et la détermination des encadreurs pendant tout le cursus. Les Masterclass, Boost Hours et les sessions de gestion de carrière ont été fortement appréciés et bénéfiques pour nos apprentissages en Machine Learning. Nous aimerions également adresser nos remerciements à notre mentor, Christophe, qui nous a accompagnés tout au long de ce projet. Sa disponibilité et son soutien sont très appréciés.

**SOMMAIRE**

[**1. Introduction 5**](#_c685imp7ry6b)

[1.1 CONTEXTE 5](#_o1lwe4bai30l)

[1.1.1 Point de vue technique et scientifique 6](#_2acplfif4ux)

[1.1.2 Point de vue économique 6](#_51wcbzsqohld)

[1.1.3 Point de vue règlementaire 7](#_36ma4cr39um2)

[1.1.4 Point de vue facteurs environnementaux 7](#_t3p9psdtj56l)

[1.1.5 Point vue sociétal et de la santé publique 8](#_x9tt4rsg2zpo)

[1.2 OBJECTIFS ET ENJEUX 9](#_kjkwv2jgiv1p)

[**2. Compréhension et manipulation des données 10**](#_pgzvihk4rvus)

[2.1 PRÉSENTATION DU DATASET 10](#_lwn78dkakduv)

[2.1.1 Source des données 10](#_fbi2yvn3tdbi)

[2.1.2 Exploration distincte des quatre jeux de tests 10](#_v3j00y3gi7ku)

[2.1.3 Mapping Conceptuel de Données 14](#_f4n4s266tweb)

[2.1.4 Fusion des dataframes 15](#_wszdqqa525s1)

[2.2 NETTOYAGE DES DONNÉES 15](#_ujd9s38vdn5y)

[2.2.1 Gestion des Doublons 15](#_dhogsmn75vlg)

[2.2.2 Traitement des valeurs manquantes après fusion des jeux de données. 15](#_ehl3i761c1mq)

[2.3 VISUALISATIONS 19](#_uly6j3sy1r3b)

[2.3.1 Description des variables disponibles, en particulier la variable d'intérêt ('grav') 19](#_j9sjj5wze0b)

[2.3.2 Analyse temporelle 20](#_h7qg4eszgyxi)

[Annuelle 20](#_9ywmqutp1ix9)

[Mensuelle 21](#_adl34crn2t4)

[Hebdomadaire 22](#_48c14j4bwjgv)

[2.3.3 Analyse sociologique 23](#_wjuly8ew7n4e)

[Gravité des accidents selon le motif de déplacement 24](#_j8yy70ogfqo0)

[Gravité des accidents selon le genre des usagers 25](#_pxmw9ce1ev9)

[2.3.4 Analyse géographique 26](#_15l94f17mvww)

[2.3.5 Autres observations 27](#_ohpkaang7hcn)

[2.4 ANALYSES STATISTIQUES 28](#_jplxhxnjbed0)

[2.4.1 Analyse bivariée entre 'grav' et les autres variables catégorielles : 28](#_n5v1u1t8cqga)

[Chi² 28](#_j3vo9mtu0myn)

[2.4.2 ANOVA 29](#_t3r6ncd7a76v)

[**3. CONCLUSION 30**](#)

Liste des figures

| Figure 1. Distribution de la gravité des accidents | ... |
| --- | --- |
| Figure 2. Distribution de nombre d'accidents par an | ... |
| Figure 3. Distribution du nombre d’accidents par mois | ... |
| Figure 4. Distribution du nombre d’accidents par jour de la semaine et heure de la journée | ... |
| Figure 5. Evolution mensuelle du nombre d'accidents répartis sur les jours de la semaine | ... |
| Figure 6. Distribution de la gravité des accidents selon le motif de déplacement | ... |
| Figure 7. Répartition de la gravité des accidents selon le genre des usagers | ... |
| Figure 8. Répartition du nombre d’accidents par tranches d’âge | ... |
| Figure 9. Fréquence des accidents en métropole | ... |
| Figure 10. Répartition de la gravité des accidents selon les conditions atmosphériques | ... |
| Figure 11. Répartition de la gravité des accidents selon la localisation et l’action du piéton | ... |

Liste des tables

| Table 1. Récapitulatif des variables par type et par dataframe après exploration | ... |
| --- | --- |
| Table 2. Nombre de variables dans le dataset fusionné | ... |

# 1. Introduction

## 1.1 CONTEXTE

Les accidents de la route constituent une préoccupation majeure de santé publique pour les Pouvoirs publics, les Forces de l’ordre et l’ensemble de la société.

Chaque année, un grand nombre d'accidents se produisent, entraînant des blessures graves, voire la perte de vies humaines. Dans le but de réduire ces conséquences tragiques, il devient impératif de comprendre les facteurs qui influencent la gravité des accidents et de développer des outils prédictifs précis. Leur prévention nécessite la collaboration de nombreux acteurs (autorités gouvernementales, forces de l’ordre, experts en sécurité routière, chercheurs universitaires, compagnies d'assurances, associations de victimes d'accidents de la route, etc.) et une approche multidisciplinaire pour réduire leur fréquence et leur gravité.

Le présent projet, nommé par l’équipe projet "SARA (**Système d'Analyse et de Risque Automobile**) ", vise à utiliser des méthodes de Data Science, en se basant sur des données historiques, pour concourir à cette démarche collaborative de prédiction de la fréquence et de la gravité des accidents routiers en France.

L'exploitation de ces bases occulte néanmoins certaines données spécifiques relatives aux usagers et aux véhicules et à leur comportement dans la mesure où la divulgation de ces données porterait atteinte à la protection de la vie privée des personnes physiques aisément identifiables ou ferait apparaître le comportement de telles personnes alors que la divulgation de ce comportement pourrait leur porter préjudice.

L'importance de ce projet réside dans sa capacité à fournir des informations essentielles aux décideurs, aux autorités locales et nationales ainsi qu'aux acteurs de la sécurité routière. En anticipant la gravité potentielle d'un accident avant qu'il ne se produise, il devient possible de mettre en place des mesures préventives ciblées, d'ajuster les politiques de sécurité routière et de mobiliser les ressources adéquates pour réduire les risques et les conséquences néfastes.

Au cours de ce projet, nous mettrons en œuvre des techniques d'analyse de données avancées, telles que l'apprentissage automatique et l'exploration de données, afin de développer un modèle prédictif robuste et fiable.

Nous évaluerons également la performance de ce modèle en utilisant des métriques appropriées pour mesurer sa précision, sa sensibilité et sa spécificité.

Nous espérons que les résultats de cette étude pourront contribuer à améliorer la sécurité routière en France en fournissant des informations prédictives précieuses pour réduire le nombre d'accidents graves et les conséquences qui en découlent.

En conclusion, ce projet vise à combler une lacune importante dans la prévention des accidents routiers en France en développant un modèle de prédiction de la gravité des accidents. En exploitant les avantages de l'analyse de données, nous espérons contribuer à la réduction des risques et à l'amélioration de la sécurité routière, pour le bien-être de tous les usagers de la route.

Ce projet présente **un intérêt technique et scientifique ainsi que des enjeux multiples (économique, réglementaire, environnemental, sociétal).**

### 1.1.1 Point de vue technique et scientifique

Dans le domaine de la Data Science, ce projet portant sur la prédiction de la gravité des accidents routiers en France, présente un intérêt technique et scientifique.

Ce projet nécessite l'utilisation d'outils Python et de bibliothèques comme pandas, numpy, matplotlib, seaborn, missingno, scikit-learning.

Il implique l'utilisation d'algorithmes de Machine Learning et de techniques statistiques pour traiter un jeu de données volumineux, déséquilibré et diversifié.

Le projet nécessite l'application de méthodes de manipulation et de nettoyage des données, et d'extraction des variables caractéristiques pertinentes. Il nécessitera également l’utilisation de techniques de modélisation prédictive permettant un apprentissage automatique, de visualisation des résultats et enfin, d’évaluation des performances du modèle.

**Ces modèles de prédiction contribueront à la compréhension scientifique des causes et des conséquences des accidents de la route. Ainsi, ces informations pourront concourir à la prévention et à la réduction de leur impact.**

### 1.1.2 Point de vue économique

La réduction du nombre d'accidents routiers et de leur gravité est un enjeu économique important en France.

Les accidents routiers entraînent des pertes humaines, des blessures graves, des problèmes psychiques, des coûts médicaux et d’indemnisation élevés, des dommages matériels et des perturbations du trafic routier.

En développant un modèle de prédiction de la gravité des accidents, **ce projet vise à contribuer à la prévention des risques d’accident et à l'amélioration de la sécurité routière. Ce qui peut avoir un impact économique positif en réduisant les coûts associés aux accidents et en favorisant une utilisation plus efficace des ressources et investissements alloués à la sécurité routière.**

### 1.1.3 Point de vue règlementaire

Les accidents de la route sont régis par des réglementations et des lois visant à assurer la sécurité des usagers de la route. Les autorités publiques collectent et maintiennent des bases de données sur les accidents routiers survenus sur le territoire français.

Le projet utilisera ces données pour analyser les tendances, identifier les facteurs de risque.

Les résultats de ce projet pourraient donc avoir des implications réglementaires en matière de sécurité routière. En effet, des évolutions réglementaires pourraient, en partie, expliquer l’évolution de la fréquence et de la gravité des accidents routiers (ex : rehaussement de la vitesse autorisée, suppression de la perte de points pour les excès de vitesse inférieurs à 5 km/heure).

**Par ailleurs, les autorités compétentes pourraient utiliser les informations issues de ce projet pour renforcer les réglementations visant à réduire les accidents et à améliorer la sécurité sur les routes.**

### 1.1.4 Point de vue facteurs environnementaux

Les différents paramètres environnementaux, tels que les conditions météorologiques, les caractéristiques des routes et les divers autres facteurs environnementaux locaux, peuvent jouer un rôle important dans la survenance et la gravité des accidents de la route.

**En prenant en compte ces facteurs environnementaux, le projet cherchera à évaluer leur impact sur les accidents routiers. Ces résultats pourraient aider à développer des stratégies de prévention adaptées à chaque contexte environnemental.**

### 1.1.5 Point vue sociétal et de la santé publique

Les accidents de la route ont un impact significatif sur la société dans son ensemble et plus spécifiquement, dans le domaine de la santé publique.

Ils peuvent entraîner des pertes de vies humaines, des blessures, des traumatismes physiques et psychiques avec des conséquences émotionnelles, sociales et économiques pour les victimes, leurs proches et, voir, pour leur environnement professionnel.

**Ce projet pourra contribuer à une meilleure compréhension des facteurs de risque et des conséquences.**

**Les résultats du projet pourront donc être utilisés par les professionnels de la santé publique pour sensibiliser les citoyens français aux comportements sécuritaires (mise en place de mesures de prévention ciblées) et améliorer la prise en charge des victimes de la route et de leurs proches.**

Pour conclure sur la présentation du contexte, **en prenant en considération les aspects techniques, scientifiques, économiques, réglementaires, environnementaux, sociétaux précités,** ce projet peut fournir des informations utiles contribuant à améliorer la sécurité routière, à réduire les accidents graves et à faciliter la prise de décision en matière de prévention et de planification des mesures de sécurité routière.

Il pourra également être utile pour adapter la prise en charge des victimes et de leurs proches (ex : adaptation des ressources matérielles, des effectifs et des expertises des unités de soins en fonction de la zone géographique et de la saisonnalité).

**Ce projet peut ainsi bénéficier à plusieurs acteurs parties prenantes de la sécurité routière** telles que les autorités gouvernementales, les forces de l’ordre, les experts en sécurité routière, les chercheurs universitaires, les compagnies d'assurance et les associations de victimes d'accidents de la route.

**Ces acteurs pourront utiliser les enseignements résultant des livrables de ce projet pour définir leurs actions en matière de sécurité routière.**

## 1.2 OBJECTIFS ET ENJEUX

Selon les estimations de l’Observatoire national interministériel de la sécurité routière (ONISR), **3 260 personnes** ont perdu la vie sur les routes de France métropolitaine en 2022, contre 2 944 en 2021 (+10,7 %), et 3 244 en 2019 (+0,5 %), année de référence.

Ces estimations provenant des sources gouvernementales, confirment la nécessité de mettre en place des outils concrets de facilitation de la prise de décision en matière de sécurité routière. Le projet SARA (Système d'Analyse et de Risque Automobile) vise à prédire la gravité des accidents routiers en France en appliquant des techniques de Data Science et d’apprentissage automatique sur un ensemble de données historiques.

Plus spécifiquement, l’objectif est d'identifier les facteurs et les caractéristiques qui ont une influence significative sur la gravité des accidents, tels que les conditions météorologiques, l'emplacement géographique, les caractéristiques des véhicules, les caractéristiques des conducteurs, etc.

En utilisant des techniques de modélisation prédictive basées sur les variables pertinentes, l’équipe projet testera plusieurs modèles capables d'estimer la gravité des accidents avec une certaine précision. Enfin, elle évaluera les performances de chaque modèle testé pour comparer ses prédictions avec les données historiques.

L’usage de ces modèles contribuera à la compréhension scientifique des causes et des conséquences des accidents de la route, ainsi qu'à la prévention et à la réduction de leur impact.

En effet, comme développé dans la partie « Contexte », ce projet permettra notamment de fournir des informations utiles pour la prise de décision en matière de prévention des risques d’une part, et de planification et de déploiement des mesures de sécurité routière d’autre part. Il pourra également être utile pour adapter la prise en charge des victimes et de leurs proches.

# 2. Compréhension et manipulation des données

## 2.1 PRÉSENTATION DU DATASET

### 2.1.1 Source des données

Nous avons utilisé quatre jeux de données concernant **les lieux, les véhicules, les usagers et les caractéristiques** des accidents survenus entre 2005 et 2021 en France. Ces données brutes proviennent des “Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2021” mises à disposition, en accès public libre, sur le site gouvernemental https://www.data.gouv.fr/.

Nous avons regroupé ces quatre jeux de données à partir des observations communes (numéro d’accident, identifiant du véhicule, numéro du véhicule et l’année de l’accident) afin de produire à l’issue d’une première analyse exploratoire, un jeu de données unique et centralisé pour bien identifier et cerner les variables présentes dans ces quatre jeux de données.

### 2.1.2 Exploration distincte des quatre jeux de tests

L'exploration des fichiers CSV de données sur les usagers, les véhicules, les lieux et les caractéristiques routières a été menée en suivant une série d'actions pour mieux comprendre et analyser les données disponibles. Les principales étapes entreprises pour ce prétraitement sont résumées ci-dessous :

**Chargement des fichiers CSV :**

Les fichiers CSV contenant les données sur les usagers, les véhicules, les lieux et les caractéristiques routières ont été chargés individuellement dans un notebook spécifique. Cela a permis d'accéder aux données et de les manipuler à l'aide d'outils et de bibliothèques adaptés.

Au cours de notre exploration des quatre jeux de données sur la période 2005 à 2021 issus du site gouvernemental, nous avons découvert l’absence de données sur les années 2020 et 2021 pour le jeu de données concernant les lieux d’accidents, contrairement aux trois autres jeux de données. De façon similaire, nous avons constaté dans le cas des usagers que les numéros d’accidents étaient mal reportés sur l’année 2020. Dans les deux cas, nous avons récupéré les fichiers des années spécifiées sur le site gouvernemental, puis nous l’avons intégré pour compléter nos données sur les lieux et les usagers.

**Affichage des informations générales :**

Pour chaque fichier CSV, des informations essentielles ont été affichées, notamment le type de données, le nombre total de variables (colonnes) et le nombre de valeurs non nulles pour chacune d'entre elles. Cette étape a permis d'avoir une vue d'ensemble des données disponibles et de détecter d'éventuelles valeurs manquantes.

**Affichage des premières lignes :**

Pour se familiariser avec le contenu des fichiers CSV, les premières lignes ont été affichées. Cela a permis d'obtenir un aperçu des données et d'identifier les noms des variables.

**Description des données pour chaque type de variable :**

Pour chaque variable, une description statistique appropriée a été réalisée en fonction du type de données. Pour les variables numériques, cela inclut la moyenne, l'écart-type, les valeurs minimales et maximales, et les quartiles. Pour les variables catégorielles, des comptages des différentes catégories ont été effectués.

**Transformation des variables en date :**

Si des variables contenaient des informations temporelles, elles ont été converties au format de date pour mieux faciliter leur utilisation dans des analyses ultérieures. Il s’agit de *l'année de naissance de l’usager* ainsi que *l’année, le mois, le jour, l’heure de l’accident*.

Un aperçu des actions qui ont été effectuées :

* La variable ‘hrmn’ a été transformée en variable ‘heure’, considérant que nous n’avions pas intérêt à conserver l’information des minutes.
* La variable ‘an\_nais’ correspondant à l’année de naissance des personnes impliquées a été transformée en ‘age’ pour connaître l'âge exact des usagers accidentés au moment de l'accident. Les âges ainsi obtenus ont été découpés en tranches d’âge (variable ‘tranches\_ages’) en fonction de la distribution des accidents par âge. L’objectif est de permettre l’étude des tendances et des différences de comportement en fonction de l'âge des personnes impliquées dans les accidents lors des accidents et d'identifier les groupes d'âge qui peuvent être plus vulnérables sur les routes.

**Transformation des variables catégorielles :**

Nous constatons que les données dont nous disposons renferment un nombre important de variables catégorielles. Nous avons choisi de convertir le type de ces variables en catégories pour permettre une meilleure compréhension des données.

En effet, la conversion d'une variable catégorielle comportant des valeurs numériques en catégories présente plusieurs avantages :

* Réduction de la taille mémoire : Lorsque nous convertissons une variable catégorielle en type catégorie, cela permet de stocker les données de manière plus compacte en mémoire. Les catégories sont représentées par des entiers, et les valeurs sont stockées dans une table de correspondance. Cela est bénéfique puisque nous travaillons avec un grand ensemble de données.
* Optimisation des opérations : En utilisant des catégories, certaines opérations peuvent être optimisées, ce qui peut accélérer le traitement des données. Par exemple, lors d'un tri ou d'un regroupement (groupby), les catégories sont plus rapides à manipuler que des valeurs numériques.
* Facilitation de l'analyse : En convertissant des valeurs numériques en catégories, nous attribuons des étiquettes significatives aux catégories, ce qui facilite l'interprétation des résultats et améliore la lisibilité des analyses. Par exemple, la variable “catégorie de véhicule (‘catv’)” possédait initialement 40 modalités différentes. Nous l’avons réduit à 21 en tenant compte du type de permis nécessaire pour chaque catégorie.
* Gestion des valeurs manquantes : Lors de la conversion en catégories, nous pouvons spécifier une catégorie spéciale pour les valeurs manquantes, facilitant ainsi leur gestion lors des analyses ou du traitement des données.
* Limitation des opérations mathématiques inappropriées : Si les valeurs numériques ont une signification catégorielle plutôt que continue, les opérations mathématiques (comme les moyennes ou les sommes) sur ces valeurs peuvent ne pas avoir de sens. La conversion en catégories permet d'éviter de telles opérations inappropriées.
* Préparation pour l’apprentissage automatique : la plupart des algorithmes de machine learning requièrent des données numériques pour fonctionner. Cela signifie que nous devons transformer nos variables catégorielles en une forme numérique adaptée avant de les utiliser dans certains algorithmes.

**Réalisation d'une statistique descriptive :**

Une analyse statistique descriptive a été réalisée pour explorer les tendances et les caractéristiques clés des données. Cela inclut des visualisations telles que des histogrammes, des diagrammes à barres et des diagrammes en boîte pour identifier les tendances et les schémas dans les données.

**Nouvelles modalités de variables et variables ajoutées :**

En 2019, la base de données des accidents a évolué, de nouvelles modalités de certaines variables et des nouvelles variables ont été ajoutées.

Pour donner un exemple de nouvelles modalités de variables ajoutées : l’indicateur « blessé hospitalisé » n’est plus labellisé par l’autorité de la statistique publique depuis 2019.

Les nouvelles variables ajoutées sont : “motor” (Type de motorisation du véhicule), id\_vehicule (identifiant du véhicule) et “secu3” (troisième équipement de sécurité utilisé).

**Étude des valeurs manquantes :**

Les valeurs manquantes ont été identifiées et étudiées pour évaluer leur impact sur les analyses ultérieures. Différentes approches ont été envisagées pour gérer les valeurs manquantes, telles que l'imputation ou la suppression des colonnes concernées.

**Suppression de certaines variables :**

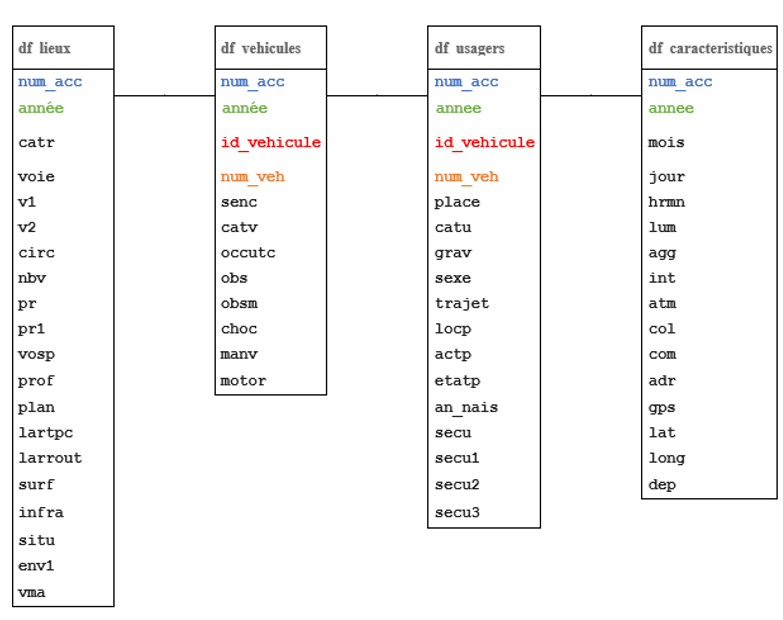
Après une évaluation approfondie des variables et de leur pertinence pour l'objectif de l'analyse, certaines variables jugées non pertinentes ou redondantes ont été supprimées pour simplifier les données et se concentrer sur les caractéristiques les plus importantes.

En somme, l'exploration des fichiers CSV de données sur les usagers, les véhicules, les lieux et les caractéristiques routières a été une étape cruciale pour comprendre la nature des données disponibles, identifier les tendances et les relations, et préparer les données pour des analyses ultérieures plus approfondies. Les actions entreprises ont permis de mieux cerner les caractéristiques des accidents routiers et d'orienter les prochaines étapes de traitement et de modélisation des données.

| **Colonne1** | **Usagers** | **Véhicules** | **Caractéristiques** | **Lieux** | **Total** |
| --- | --- | --- | --- | --- | --- |
| Catégorie | 12 | 7 | 6 | 8 | **33** |
| Chaîne | 3 | 2 | 6 | 0 | **11** |
| Entier | 1 | 1 | 2 | 1 | **5** |
| Décimal | 2 | 1 | 2 | 1 | **6** |
| Date | 3 | 1 | 2 | 1 | **7** |
| Total | **21** | **12** | **18** | **11** | **62** |

[Table 1](#table_varAvantFusion). Récapitulatif des variables par type et par dataframe après exploration

### 2.1.3 Mapping Conceptuel de Données



### 2.1.4 Fusion des dataframes

Pour construire le dataframe global et complet, nous avons suivi une stratégie de fusion étape par étape en utilisant les variables communes entre les dataframes.

Voici l'ordre dans lequel nous avons réalisé ces fusions :

* Nous avons commencé par fusionner les dataframes "Usagers" et "Véhicules", car ces deux dataframes contiennent les variables clés 'num\_acc', 'num\_veh', 'annee' et 'id\_vehicule' qui nous permettent d'associer les informations spécifiques sur les usagers et les véhicules impliqués dans les accidents.
* Parallèlement, nous avons fusionné les dataframes ‘caractéristiques’ avec le dataframe "Lieux". Nous avons utilisé les variables communes 'num\_acc' et “annee” pour effectuer cette fusion.
* Enfin, le dataframe issu de la deuxième fusion a été fusionné avec le premier. Encore une fois, nous avons utilisé les variables communes 'num\_acc' et “annee” pour réaliser cette dernière fusion.

En suivant cette approche de fusion progressive en utilisant des variables communes, nous avons réussi à créer un dataframe complet (voir [Table 2](#tab_varFusion).) qui rassemble toutes les informations importantes concernant les usagers, les véhicules, les lieux et les caractéristiques des accidents, nous permettant ainsi d'effectuer des analyses approfondies et de tirer des conclusions pertinentes.

## 2.2 NETTOYAGE DES DONNÉES

### 2.2.1 Gestion des Doublons

2709 observations en doublons ont été supprimées.

### 2.2.2 Traitement des valeurs manquantes après fusion des jeux de données.

**Suppression des variables avec un nombre important de valeurs manquantes**

Un certain nombre de variables contiennent un trop grand nombre de valeurs manquantes. Il s’agit des variables :

* ‘id\_vehicule’, c’est une variable qui a été rajoutée dans la base de données en 2019. On ne la retiendra pas dans l’étude.
* 'motor', c’est également une variable qui a été rajoutée en 2019. On ne la retiendra pas non plus.
* 'adr', il s’agit de l’adresse postale du lieu de l’accident. Elle est renseignée uniquement pour les accidents survenus en agglomération et le format de saisie des données est très hétérogène et par conséquent, inexploitable (y compris, après nettoyage des données).
* 'secuTrois', c’est une variable qui a également été rajoutée dans la base de données en 2019. Elle comprend 99% de valeurs manquantes et la seule valeur renseignée dans les 1% de données restantes est “Autre”. On ne retiendra donc pas cette variable dans l’étude.
* ‘gps’, ‘lat’ et ‘long’ : ces variables étaient initialement envisagées pour permettre l'identification des emplacements géographiques des accidents routiers en utilisant les relevés GPS. L'objectif était de dresser une cartographie des risques routiers en identifiant les endroits dangereux ou propices à ces événements.

Cependant, lors de l'exploration de ces variables, nous avons rencontré plusieurs difficultés. La variable 'gps' contenait plus de 50% de valeurs manquantes et ne présentait aucun doublon, ce qui la rendait peu exploitable pour notre objectif spécifique.

Quant aux variables 'lat' et 'long', elles présentaient plus de 80% de valeurs manquantes. De plus, elles avaient plusieurs valeurs mal renseignées et étaient exprimées dans deux formats différents, à savoir UTM et degré décimal. Nous avons entrepris des tentatives de conversion des valeurs au format degré décimal afin d'identifier les adresses correspondantes, mais nous avons constaté que certaines valeurs se trouvaient hors de la zone UTM entre 2005 et 2021, ce qui rendait cette conversion problématique.

Dans une deuxième approche, nous avons effectué une exploration plus poussée en prenant en compte la variable 'annee' pour tenter de réaliser les conversions à partir d'une date précise. Malheureusement, nous avons remarqué qu'un très faible nombre de points GPS étaient disponibles pour effectuer des correspondances aux adresses.

En somme, nous avons constaté que ces trois variables ('gps', 'lat' et 'long') étaient très peu renseignées entre 2005 et 2021, ce qui les rend inutilisables pour notre projet. Malgré nos efforts pour explorer ces variables et réaliser la cartographie des risques routiers, leur manque de données fiables et significatives ne nous permet pas d'atteindre cet objectif spécifique.

**Suppression de variables intermédiaires créés lors de l’exploration et la visualisation de chaque datasets**

Pour l'âge des usagers accidentés, compte tenu de la création de nouvelles variables lors de l’exploration et la visualisation de chaque datasets, après fusion des datasets, nous disposions de quatre variables en plus de la variable "tranches\_ages".

Afin de simplifier notre analyse, nous avons supprimé deux des trois variables liées à l'âge, à savoir "age\_acc" (exprimée en nombre de jours) et "age\_acc\_seconds".

Nous conserverons uniquement la variable "age\_acc\_an" qui donne l'âge de l'accident en années.

Par ailleurs, nous disposons également de la variable "an\_naiss" qui pourrait être utile pour d'éventuelles visualisations, nous la conserverons donc.

Ces ajustements nous permettront de travailler de manière plus concise et pertinente sur les données relatives à l'âge des personnes impliquées dans les accidents.

Comme nous disposons déjà de la date de l'accident, nous avons supprimé les variables relatives à l'année, au mois et au jour de l'accident pour éviter toute redondance.

**Gestion des valeurs manquantes pour la variable “tranches\_ages”**

Nous avons choisi de supprimer les observations pour lesquelles la tranche d'âge n'était pas renseignée, car cela pourrait induire des biais dans l'analyse, en supposant que l'année de naissance n'était pas non plus renseignée.

Ces ajustements contribuent à améliorer la qualité et la fiabilité des données relatives à la sécurité des usagers, ce qui est essentiel pour des analyses pertinentes et des conclusions éclairées.

**Gestion des valeurs manquantes dans les variables relatives à la sécurité des usagers**

Dans le cadre du traitement des variables relatives à la sécurité des usagers, nous avons effectué une analyse visuelle des données manquantes en utilisant la méthode matrix de la librairie missingno.

En examinant la matrice de visualisation des données manquantes, nous avons identifié trois variables spécifiques, à savoir 'place' (place occupée dans le véhicule par l'usager), 'secuUn' (1er équipement de sécurité utilisé), et 'secuDeux' (2nd équipement de sécurité utilisé), qui semblaient former un groupe cohérent concernant la sécurité des personnes impliquées dans l'accident.

En prenant en compte cette observation, nous avons procédé au remplacement des valeurs manquantes de ces variables par la méthode de remplacement par la valeur la plus fréquente (mode()). Cette approche nous a permis de préserver la cohérence et la validité des données relatives à la sécurité des usagers, en évitant toute distorsion dans l'analyse ultérieure.

Ces ajustements ont été effectués dans le but d'obtenir des données plus complètes et fiables, ce qui est essentiel pour réaliser des analyses précises et pertinentes sur les aspects de sécurité liés aux accidents.

**Gestion des valeurs manquantes dans les autres variables :**

Dans la continuité de notre traitement des données manquantes, nous avons décidé de remplacer les valeurs manquantes (NAN) par la mention 'Non renseigné' pour certaines variables.

Nous avons appliqué cette approche aux variables pour lesquelles le pourcentage de données manquantes est inférieur ou égal à 5%. Cela nous permet de conserver un niveau de précision acceptable tout en évitant de perdre des informations essentielles.

Voici les variables concernées avec le pourcentage de données manquantes associé :

* 'atm' (Conditions atmosphériques) : 0.01%
* 'int' (Intersection) : 0.01% (incluant les valeurs -1 qui représentent également un manque d'information)
* 'choc' (Type de choc) : 0.01%
* 'senc' (Sens du véhicule) : 0.01%
* 'manv' (Manœuvre du véhicule) : 0.02%
* 'trajet' (Motif du trajet) : 0.02%
* 'obsm' (Obstacle fixe heurté par le véhicule) : 0.04%
* 'obs' (Obstacle mobile heurté par le véhicule) : 0.05%
* 'locp' (Localisation du piéton lors de l'accident) : 2.57%
* 'etatp' (État du piéton lors de l'accident) : 2.57%
* 'actp' (Action du piéton lors de l'accident) : 2.57%

En utilisant la méthode `fillna` avec les valeurs spécifiées, nous avons complété les données manquantes avec 'Non renseigné' pour ces variables. Cela nous permet de garantir que ces données sont prises en compte lors de nos analyses futures, tout en étant transparents sur la nature des valeurs manquantes.

Après toutes ces manipulations, nous avons obtenu un jeu de données fusionné qui compte 41 variables et 2 291797 observations.

| **Colonne1** | **Dataset Fusionné** |
| --- | --- |
| Catégorie | 31 |
| Chaîne | 3 |
| Entier | 2 |
| Décimal | 3 |
| Date | 2 |
| **Total** | **41** |

[Table 2](#table_varFusion). Nombre de variables dans le dataset fusionné

## 2.3 VISUALISATIONS

### 

### 2.3.1 Description des variables disponibles, en particulier la variable d'intérêt ('grav')

Notre dataset résultant de la fusion des 4 rubriques possède 2192779 lignes et 64 colonnes. La majorité de ces variables peuvent être dans le contexte de l’étude considérer comme catégorielles. En effet, pour une très large majorité des variables, les modalités de celle- ci correspondent à des codes correspondant eux-mêmes aux observations des forces de l’ordre intervenues sur les lieux de l’accident.

Par exemple, notre variable cible ‘grav’ est encodé de la manière suivante :

* + 1 = les personnes indemnes : impliquées non décédées et dont l’état ne nécessite aucun soin médical du fait de l’accident,
  + 2 = les personnes tuées : personnes qui décèdent du fait de l’accident, sur le coup ou dans les trente jours qui suivent l’accident,
  + 3 = les blessés dits « hospitalisés » : victimes hospitalisées plus de 24 heures,
  + 4 = les blessés légers : victimes ayant fait l'objet de soins médicaux mais n'ayant pas été admises comme patients à l'hôpital plus de 24 heures.

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

[Figure 1](#figur_distrGravite). Distribution de la gravité des accidents

Afin de faciliter l'analyse des données, nous avons effectué une permutation dans l’ordre des modalités de la variable "grav" pour refléter un ordre de gravité croissant. Nous avons permuté les valeurs "Tué" (2) par "Blessé léger" (4) et les valeurs "Blessé léger" (4) par "Tué" (2).

De cette manière, les catégories sont désormais classées par ordre de gravité croissante, allant des accidents les moins graves (Indemne et Blessé léger) aux accidents les plus graves (Blessé hospitalisé et Tué).

Cette analyse est essentielle pour comprendre les tendances de sécurité routière et identifier les situations où des mesures préventives doivent être prises pour réduire les accidents les plus graves. Par exemple, en se concentrant sur les catégories de gravité les plus élevées, les autorités peuvent cibler les interventions et les politiques visant à améliorer la sécurité sur nos routes.

Au vue du grand nombre de variables il ne nous semble pas judicieux de retranscrire l’analyse de toutes celles-ci dans ce rapport. Nous allons plutôt résumer les observations principales.

### 2.3.2 Analyse temporelle

#### Annuelle

L'analyse montre que le nombre d'accidents a légèrement diminué au fil des ans.

Une image contenant texte, capture d’écran, Police, ligne

Description générée automatiquement

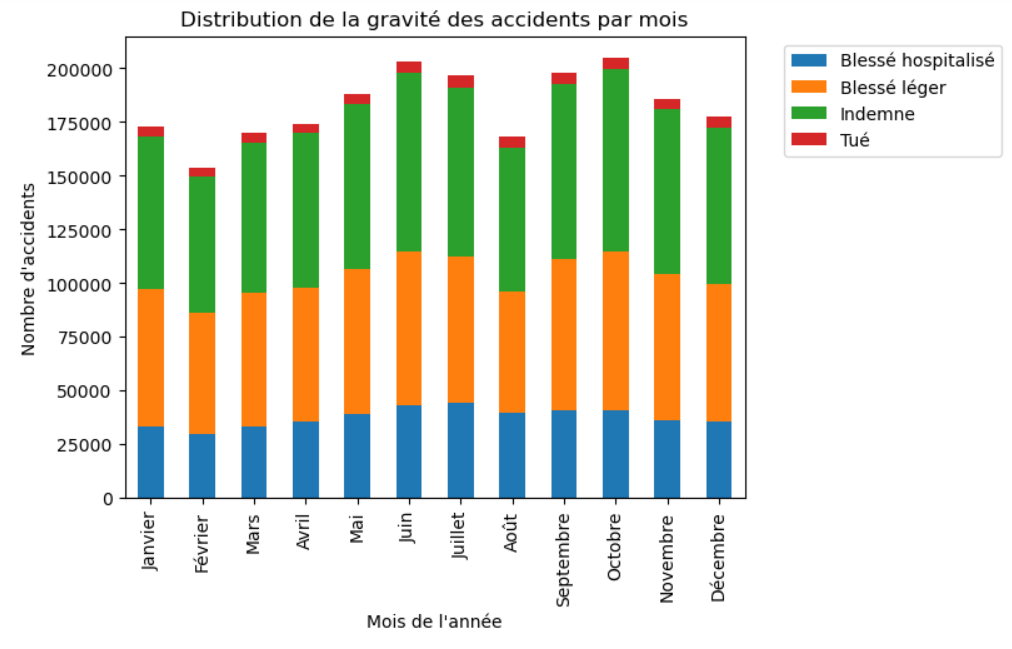
[Figure 2](#figur_NbreAccParAn). Distribution de nombre d'accidents par an

Le nombre d'accidents a diminué depuis 2005. On observe les effets de la politique de prévention routière, qui s'est fortement accentuée avec notamment l'apparition des radars fixes en 2003. Cependant, la distribution de la gravité des accidents n'a pas varié.

L’année 2020 est particulière et correspond aux confinements liés à l'épidémie de COVID-19

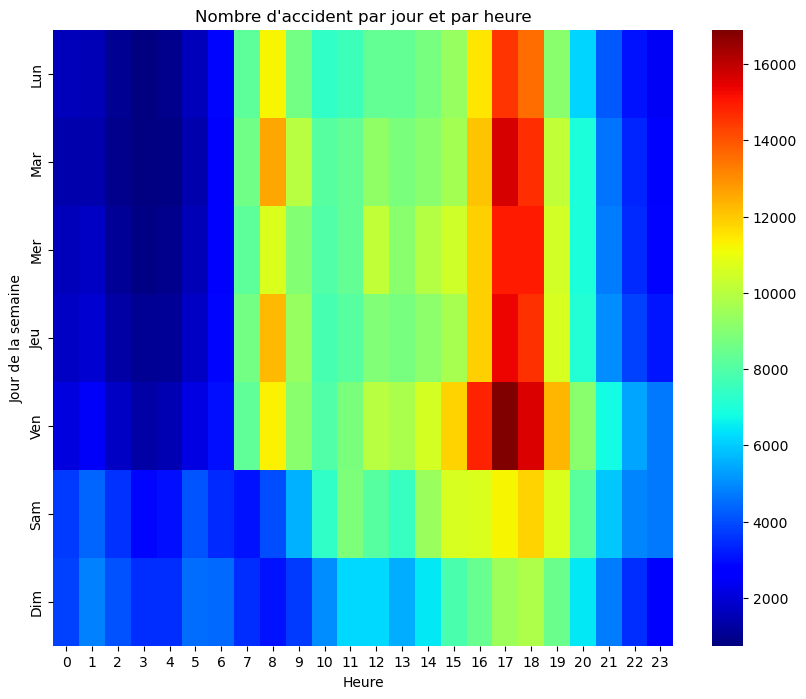
#### Mensuelle

Sur la période de l’étude, les mois de juin, juillet et octobre ont enregistré le plus grand nombre d'accidents. En revanche, le mois de février a enregistré le moins d'accidents.



[Figure 3](#figur_NbreAccMois). Distribution du nombre d’accidents par mois

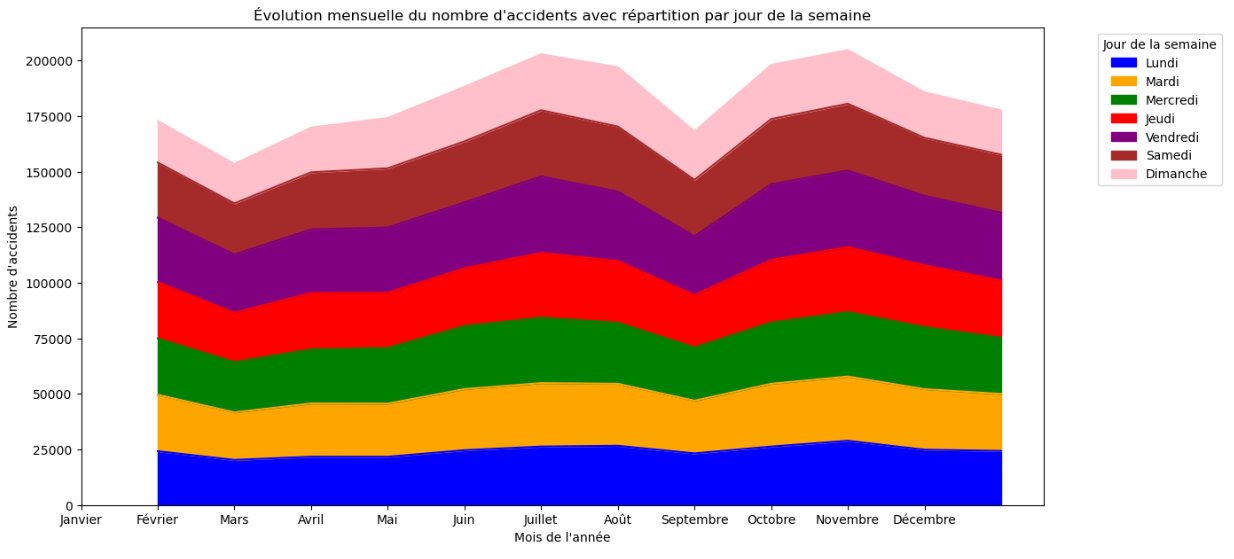
#### Hebdomadaire



[Figure 4](#figur_NbrAccSem). Distribution du nombre d’accidents par jour de la semaine et heure de la journée

On note la différence de la répartition des accidents entre la semaine et les week-ends :

* En semaine, les accidents se produisent pendant les heures de pointe correspondants aux horaires de travail.
* le week-end la répartition est plus uniforme même pendant la nuit.



[Figure 5](#figur_RpartNbrAccJJ). Evolution mensuelle du nombre d'accidents répartis sur les jours de la semaine

Selon les résultats affichés sur le graphique, nous pouvons observer que les accidents surviennent principalement le vendredi, la veille du week-end, et le samedi, peu importe le mois de l'année.

En outre, il est intéressant de noter que le nombre d'accidents est significativement élevé les dimanches du mois de juillet, tandis que les dimanches enregistrent le moins d'accidents pour les autres mois de l'année.

Ces constatations mettent en évidence des tendances importantes en termes de jours de la semaine et de mois où la prudence sur les routes est particulièrement essentielle pour assurer la sécurité de tous les usagers.

**Journalier**

Nous pouvons croiser ces informations avec les résultats du graphique concernant les motifs de déplacement des personnes accidentées.

### 2.3.3 Analyse sociologique

Si les causes des accidents de la circulation peuvent être multifactorielles, une dimension sociale reste encore à explorer : l'impact du genre sur la fréquence et la gravité des accidents routiers.

La problématique sociologique que nous abordons dans ce projet se concentre sur la question de savoir si le genre des conducteurs influe sur le nombre d'accidents de la circulation en France. Les stéréotypes de genre peuvent jouer un rôle dans les comportements au volant, les choix de véhicules et les réactions face aux situations de conduite. Ainsi, il est essentiel d'analyser comment les différences de genre peuvent influencer les taux d'accidents routiers et la gravité des conséquences pour les conducteurs et les passagers.

Il est essentiel de se pencher sur la manière dont les variables de genre jouent un rôle dans cette dynamique. Les normes sociales, les comportements au volant, les choix de véhicules et les conditions de conduite peuvent varier selon le genre, et cela peut potentiellement se refléter dans les taux de blessures graves et de décès parmi les conducteurs.

#### Gravité des accidents selon le motif de déplacement

Une image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquement

[Figure 6](#figur_DistrGravAccMotifDepl). Distribution de la gravité des accidents selon le motif de déplacement

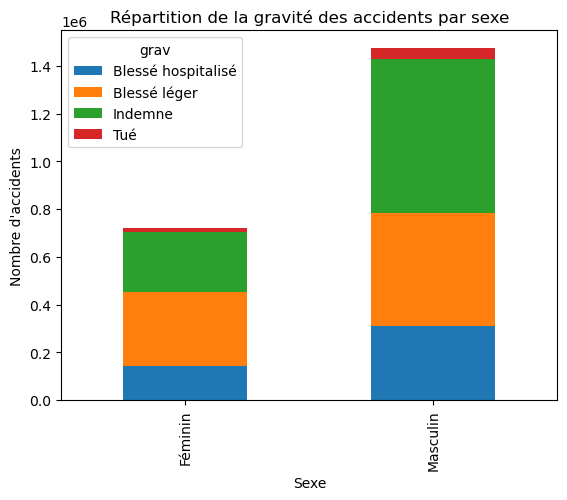
Il s'avère que plus de 37% des accidents sont survenus au cours d'un déplacement pour le motif 'promenade-loisirs'. Cette constatation pourrait expliquer les jours de la semaine et les mois présentant le plus grand nombre d'accidents.

En effet, nous avons indiqué ci-avant avoir observé que les accidents surviennent principalement le vendredi, la veille du week-end, et le samedi, peu importe le mois de l'année. Ces jours de la semaine sont généralement associés à des activités de loisirs et de détente, ce qui pourrait entraîner une augmentation du trafic routier pendant ces périodes.

En outre, il est intéressant de noter que le nombre d'accidents est significativement élevé les dimanches du mois de juillet, tandis que les dimanches enregistrent le moins d'accidents pour les autres mois de l'année. Le dimanche est souvent considéré comme un jour de repos et de loisirs, ce qui pourrait expliquer pourquoi les déplacements pour le motif 'promenade-loisirs' sont plus fréquents à cette période.

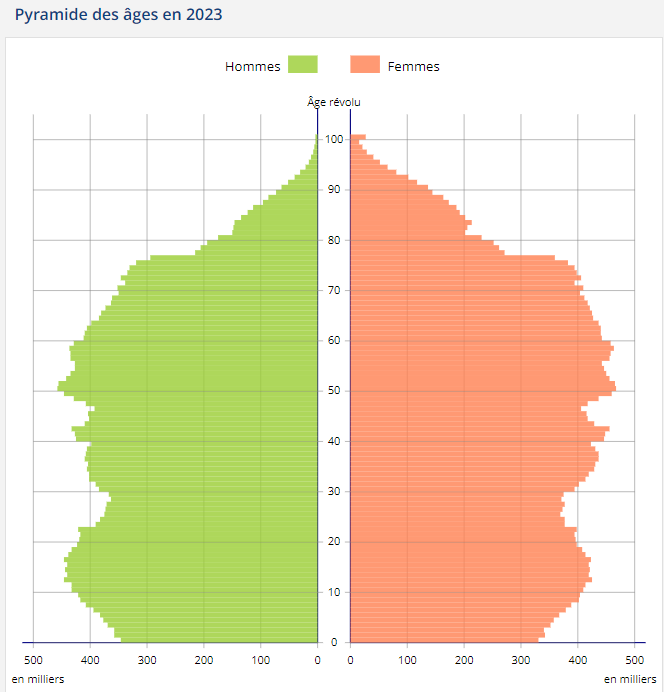
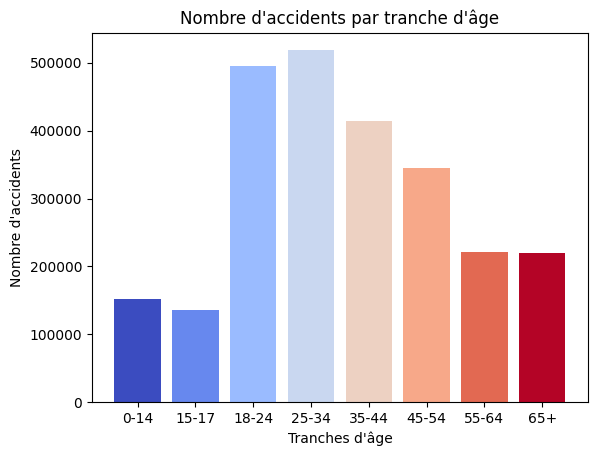
Ces résultats mettent en évidence des tendances importantes en termes de jours de la semaine et de mois où la prudence sur les routes est particulièrement essentielle pour assurer la sécurité de tous les usagers, en particulier lors des déplacements pour le motif “promenade-loisirs”.

#### Gravité des accidents selon le genre des usagers



[Figure 7](#figur_ReprtGravGenre). Répartition de la gravité des accidents selon le genre des usagers

Si les femmes semblent avoir moins d'accidents que les hommes ([Figure 7](#fig_ReprtGravGenre).), la proportion de chaque modalité de la gravité semble être la même.

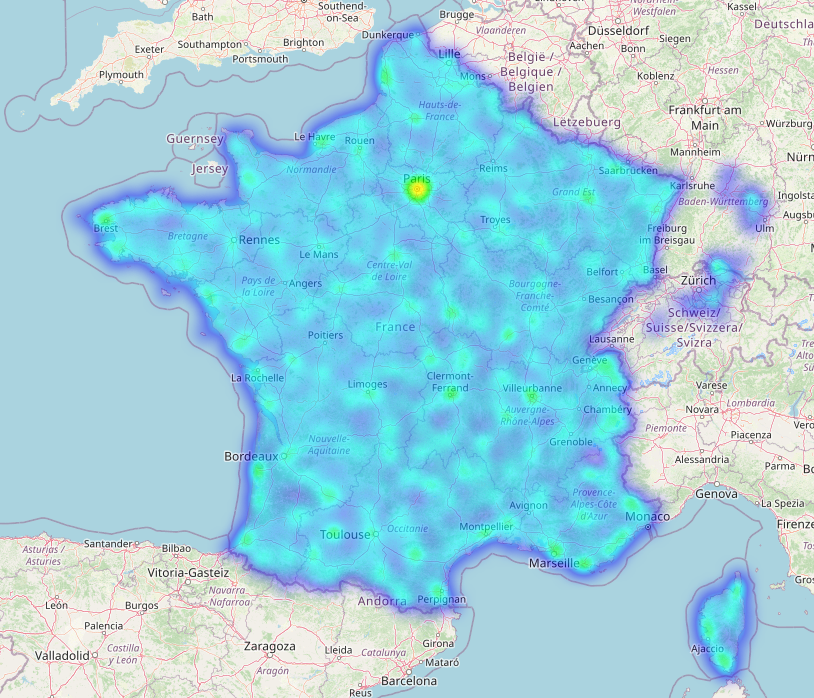


[Figure 8](#figur_TranchesAge). Répartition du nombre d’accidents par tranches d’âge

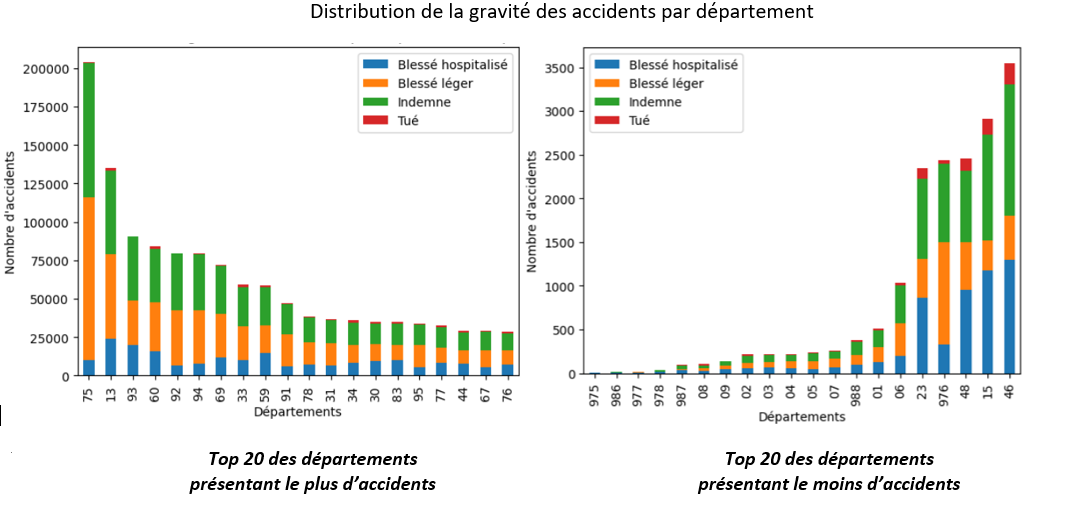
La répartition par tranche d'âge semble indiquer que les jeunes conducteurs ont plus d'accidents.

### 2.3.4 Analyse géographique

Ci-dessous est représentée une heatmap de la fréquence des accidents sur le territoire métropolitain. Sans surprise, les zones avec la plus forte densité d'accidents sont les zones à forte densité de population (les villes principalement) et donc de circulation.



[Figure 9](#figur_FreqAcc). Fréquence des accidents en métropole



### 2.3.5 Autres observations

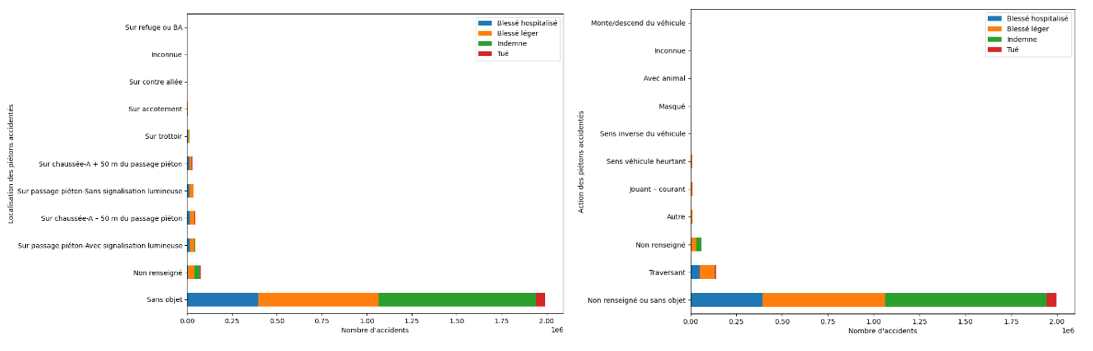
pour un grand nombre de variables, une modalité prédomine très largement les autres, par exemple pour les conditions atmosphériques :

Une image contenant texte, capture d’écran, logiciel, nombre

Description générée automatiquement

[Figure 10](#figur_ReprtAccCondAtmosph). Répartition de la gravité des accidents selon les conditions atmosphériques

Pour d’autres variables la modalité prédominante n’apporte aucune information par exemple concernant la localisation des piétons impliqués dans l’accident :



[Figure 11](#figur_RepartGravAccPietons). Répartition de la gravité des accidents selon la localisation et l’action du piéton

## 2.4 ANALYSES STATISTIQUES

Lors de cette étude, nous envisageons d’utiliser des analyses statistiques approfondies pour confirmer et étayer les informations présentées sur les graphiques. Ces analyses nous permettront d'approfondir notre compréhension des tendances et des relations entre les variables, en fournissant des mesures quantitatives pour étayer les observations visuelles. Nous allons utiliser des méthodes statistiques telles que la corrélation, l'analyse de variance (ANOVA), les tests de significativité et les régressions linéaires pour explorer les associations entre les différentes variables et déterminer leur impact sur les résultats des accidents routiers.

Grâce à ces analyses rigoureuses, nous allons pouvoir valider les informations clés présentées sur les graphiques, renforçant ainsi la fiabilité de nos conclusions et fournissant des preuves solides pour étayer nos recommandations en matière de sécurité routière.

### 2.4.1 Analyse bivariée entre 'grav' et les autres variables catégorielles :

### 

### Chi²

Comme indiqué précédemment, dans le cadre de cette étude, nous avons choisi d’utiliser le test du chi2 pour examiner les relations entre la variable dépendante catégorielle et chacune des variables catégorielles indépendantes.

C'est un test non paramétrique utilisé pour analyser les associations entre deux variables catégorielles. Il permet de déterminer si les fréquences observées dans les différentes catégories sont significativement différentes des fréquences attendues.

Pour analyser les relations de la variable 'grav', nous avons appliqué le test de Chi2 et calculé le coefficient de Cramer.

Il est à noter qu'aucune des variables ne semble présenter une forte relation avec notre variable cible. Les deux variables ayant le coefficient de Cramer le plus élevé sont la variable "Catégorie du véhicule (‘catv’)" et 'secuDeux', qui correspond au principal équipement de sécurité utilisé par l'usager impliqué dans l'accident.

Une image contenant texte, logiciel, Police, nombre

Description générée automatiquement

Une image contenant texte, diagramme, capture d’écran, Tracé

Description générée automatiquement

## 

### 2.4.2 ANOVA

L’Analyse de variance (ANOVA) est un test paramétrique utilisé pour comparer les moyennes de plus de deux groupes. L'ANOVA peut être utilisée lorsque la variable dépendante est catégorielle avec plus de deux niveaux et que la variable indépendante est catégorielle.

Après avoir utilisé l'Analyse de variance (ANOVA) pour comparer les moyennes de plus de deux groupes, nous avons présenté les résultats sous forme de statistiques F et de valeurs p (cf. annexe 2).

La statistique F mesure la variance entre les groupes par rapport à la variance au sein des groupes. Une valeur de F élevée indique des différences significatives entre les moyennes des groupes.

La valeur p est la probabilité d'observer ces différences si l'hypothèse nulle est vraie. Une valeur p faible (généralement inférieure à 0,05) indique que les différences sont statistiquement significatives.

En analysant les résultats, voici quelques conclusions possibles :

* Les variables "lum", "int", "atm", "col", "catr", "vosp", "prof", "plan", "surf", "infra", "situ", "obs", "obsm", "choc", "manv", "catv\_Label", "permis", "locp", "actp", "etatp", "an\_nais", "secuUn", "secuDeux", "an\_naiss" ont des valeurs de p très faibles (inférieures à 0,05), ce qui suggère qu'il y a des différences significatives dans la gravité des accidents entre les différentes catégories de ces variables. Ces variables jouent donc un rôle important dans l'explication des différences de gravité des accidents.
* En revanche, les variables "agg", "circ", "sexe" et "trajet" ont des valeurs de p élevées (supérieures à 0,05), ce qui indique que les différences dans la gravité des accidents entre les catégories de ces variables ne sont pas statistiquement significatives. Ces variables semblent donc moins influentes sur la gravité des accidents.

Ces conclusions concourent à la compréhension de l'impact des différentes variables sur la gravité des accidents routiers et de cibler celles qui sont les plus déterminantes.

L'ANOVA peut également être utilisée pour étudier la relation entre la gravité des accidents et les variables numériques.

L'utilisation de l'ANOVA pour les variables numériques pourrait nous permettre d'identifier les différences significatives dans la gravité des accidents en fonction des

valeurs numériques spécifiques de certaines variables (par exemple : l'âge

des personnes impliquées, le nombre d'occupants du véhicule, etc.). Cela vous

donnerait une compréhension plus approfondie de l'impact de ces variables

numériques sur la gravité des accidents routiers.

# 3. CONCLUSION

L'analyse du jeu de données révèle que peu de variables présentent une relation forte avec notre variable cible.

Cette conclusion est basée sur les résultats de deux tests statistiques, à savoir le test du Chi² et l'ANOVA, que nous avons réalisés pour évaluer les associations entre les variables explicatives et la gravité des accidents.

Ces tests nous ont permis de quantifier la relation entre chaque variable explicative et la variable cible (gravité des accidents).

Malheureusement, les résultats indiquent qu'il n'existe que peu de variables ayant une influence significative sur la gravité des accidents, ce qui rend notre objectif de construire un modèle prédictif plus complexe.

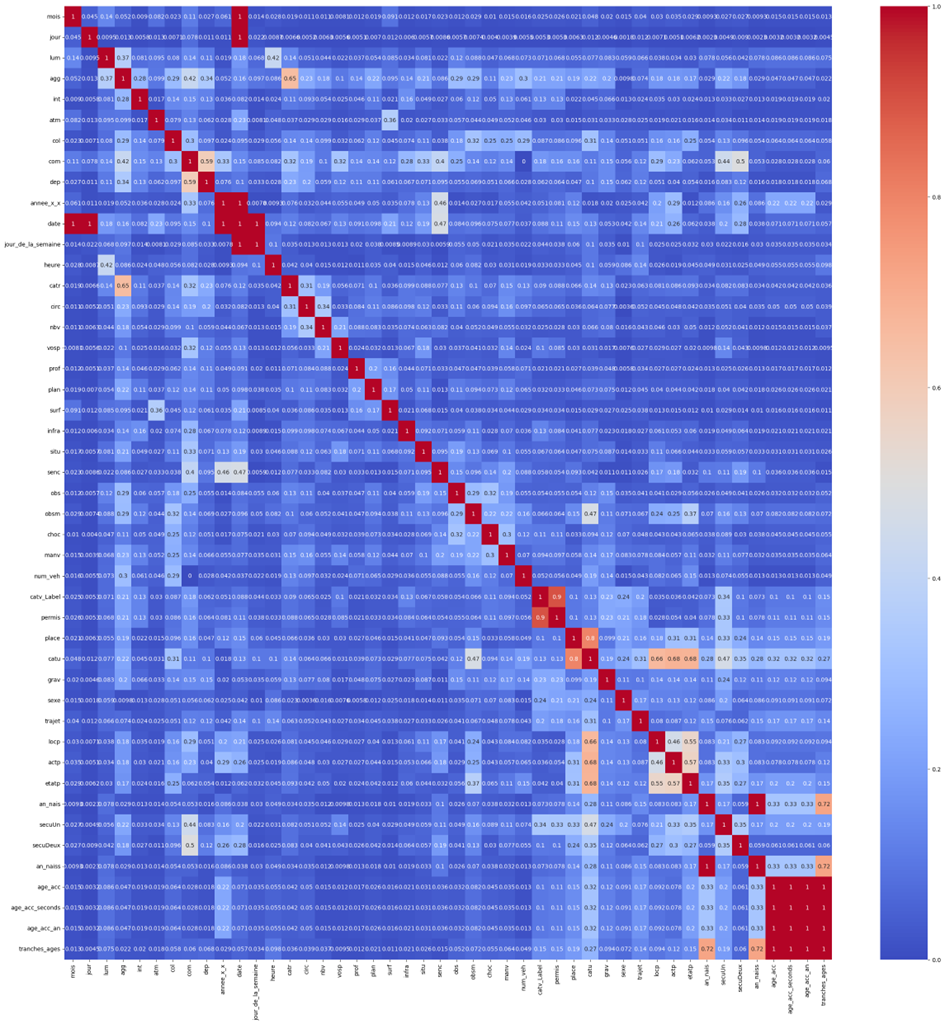
Face à cette situation, il se peut qu'il soit nécessaire d'enrichir notre dataset en ajoutant des données supplémentaires, ce qui pourrait nous permettre de capturer davantage de caractéristiques et de relations potentielles avec la gravité des accidents.

En attendant, nous avons décidé de procéder à la construction de nos premiers modèles prédictifs en utilisant les données disponibles. Nous avons ainsi entraîné deux modèles, à savoir le modèle RandomForest et l'arbre de décision, afin de voir comment ils se comportent malgré les limitations actuelles du jeu de données.

Ces modèles nous permettent de commencer à explorer les performances de prédiction et d'identifier les caractéristiques importantes dans l'estimation de la gravité des accidents. Cependant, nous sommes conscients que l'enrichissement du dataset en ajoutant des données supplémentaires pourrait améliorer considérablement les résultats de nos modèles.

En parallèle, nous continuons à peaufiner le preprocessing de nos données pour garantir que nous exploitons au mieux les informations actuellement disponibles. Une fois que nous aurons obtenu un dataset enrichi, nous pourrons alors entraîner nos modèles sur des données plus complètes et affiner leur performance de prédiction pour atteindre notre objectif de construire un modèle prédictif robuste et précis pour estimer la gravité d'un accident.

**Annexe 1 : Test Chi²**

****

**Annexe 2 : Test ANOVA**

<ipython-input-131-7fefc11b1c14>:16: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.

Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation:<https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy>

selected\_data[var] = selected\_data[var].astype(str)

ANOVA result for the categorical variable 'lum':

F-statistic = 5.915649939400401

p-value = 0.004607122463436838

ANOVA result for the categorical variable 'agg':

F-statistic = 1.2890383353519088

p-value = 0.2995343615910299

ANOVA result for the categorical variable 'int':

F-statistic = 8.258408179813243

p-value = 1.3376089918331384e-05

ANOVA result for the categorical variable 'atm':

F-statistic = 7.83520118460214

p-value = 7.694476591819332e-06

ANOVA result for the categorical variable 'col':

F-statistic = 3.4738537871413717

p-value = 0.010321773795376358

ANOVA result for the categorical variable 'catr':

F-statistic = 5.228051323939453

p-value = 0.0010084100383082441

ANOVA result for the categorical variable 'circ':

F-statistic = 6.397805370627458

p-value = 0.003277040335518708

ANOVA result for the categorical variable 'vosp':

F-statistic = 8.272959178115414

p-value = 0.0009880549157575429

ANOVA result for the categorical variable 'prof':

F-statistic = 6.560039027236189

p-value = 0.0029316643813264015

ANOVA result for the categorical variable 'plan':

F-statistic = 6.399622930307336

p-value = 0.0032729251309046655

ANOVA result for the categorical variable 'surf':

F-statistic = 7.833291160533194

p-value = 7.712457730765182e-06

ANOVA result for the categorical variable 'infra':

F-statistic = 8.24016940699886

p-value = 1.6449694214668471e-06

ANOVA result for the categorical variable 'situ':

F-statistic = 7.3795911247344845

p-value = 3.524204829563489e-05

ANOVA result for the categorical variable 'obs':

F-statistic = 6.852162737114381

p-value = 1.035158799298044e-08

ANOVA result for the categorical variable 'obsm':

F-statistic = 6.009504221329476

p-value = 0.0004012775440742489

ANOVA result for the categorical variable 'choc':

F-statistic = 5.125531531485514

p-value = 0.0001649193669987099

ANOVA result for the categorical variable 'manv':

F-statistic = 7.626589384387409

p-value = 2.83576585360372e-13

ANOVA result for the categorical variable 'catv\_Label':

F-statistic = 5.620969841011969

p-value = 6.517130846716092e-08

ANOVA result for the categorical variable 'permis':

F-statistic = 5.547610244823238

p-value = 6.384057997751008e-07

ANOVA result for the categorical variable 'catu':

F-statistic = 4.479521582167622

p-value = 0.0446495327313231

ANOVA result for the categorical variable 'sexe':

F-statistic = 1.7338882098705672

p-value = 0.23597204745697542

ANOVA result for the categorical variable 'trajet':

F-statistic = 4.347322552572357

p-value = 0.005272908635831232

ANOVA result for the categorical variable 'locp':

F-statistic = 7.5130272042557555

p-value = 4.356110657740322e-06

ANOVA result for the categorical variable 'actp':

F-statistic = 7.070246259693944

p-value = 1.0130205295853895e-09

ANOVA result for the categorical variable 'etatp':

F-statistic = 7.098193831092538

p-value = 0.00534126850862662

ANOVA result for the categorical variable 'an\_nais':

F-statistic = 3.230673931086638

p-value = 5.871791515094458e-16

ANOVA result for the categorical variable 'secuUn':

F-statistic = 4.457491631945983

p-value = 0.0009001217075645784

ANOVA result for the categorical variable 'secuDeux':

F-statistic = 7.118338549924037

p-value = 7.5669665551075305e-06

ANOVA result for the categorical variable 'an\_naiss':

F-statistic = 3.230673931086638

p-value = 5.871791515094458e-16

ANOVA result for the categorical variable 'tranches\_ages':

F-statistic = 1.3764810688337263

p-value = 0.2793523133323504