# PAGE DE GARDE

# SOMMAIRE

# Introduction au projet

## Contexte

Chaque année depuis 2005, l'Observatoire National Interministériel de la Sécurité Routière (ONISR) met à disposition (via le site data.gouv) des bases de données relatives aux accidents corporels de la circulation routière.

Pour chaque accident corporel (accident survenu sur une voie ouverte à la circulation publique, impliquant au moins un véhicule et causant au moins une victime), des informations sont recueillies par les forces de l'ordre présentes sur les lieux. Ces informations sont consignées dans un document appelé bulletin d’analyse des accidents corporels. L'ensemble de ces bulletins constitue le fichier national des accidents corporels de la circulation, communément appelé « Fichier BAAC », géré par l'ONISR.

Les bases de données issues de ce fichier recensent tous les accidents corporels survenus au cours d'une année donnée en France métropolitaine, dans les départements et territoires d'Outre-mer avec une description simplifiée. Elles contiennent des informations sur la localisation de l'accident, les caractéristiques de l'accident et du lieu, ainsi que sur les véhicules impliqués et leurs victimes, qui sont rangées dans quatre bases de données : usagers, véhicules, lieux et caractéristiques.

Dans le cadre de notre projet « fil rouge » de notre formation de Data Analyst, nous sommes amenés à manipuler ces bases de données.

## Objectifs

* Quels sont les principaux objectifs à atteindre ? Décrivez en quelques lignes.
* Pour chacun des membres du groupe, préciser le niveau d’expertise autour de la problématique adressée ?
* Êtes-vous entré en contact avec des experts métiers pour affiner la problématique et les modèles sous-jacents ? Si oui, détaillez l’apport de ces intéractions.
* (Avez-vous connaissance d’un projet similaire au sein de votre entreprise, ou bien dans votre entourage ? Quel est son état d’avancement ? En quoi vous a-t-il aidé dans la réalisation de votre projet ? En quoi votre projet contribue-t-il à l’améliorer ?)

L’objectif de ce projet est d’essayer de prédire la gravité des accidents routiers en France. Les prédictions seront basées sur les données de l’ONISR.

# Compréhension et manipulation des données

* Quel(s) jeu(x) de donnée(s) avez vous utilisé pour atteindre les objectifs de votre projet ?
* Ces données sont-elles disponibles librement ? Dans le cas contraire, qui est le propriétaire de la donnée ?
* Décrivez la volumétrie de votre jeu de données ?
* Quelles variables vous semblent les plus pertinentes au regard de vos objectifs ?
* Quelle est la variable cible ?
* Quelles particularités de votre jeu de données pouvez-vous mettre en avant ?
* Etes-vous limités par certaines de vos données ?

L’objectif de cette partie est d’effectuer l’exploration, la visualisation et la préparation de ces bases de données.

# 1 – Jeux de données

Dans le cadre de notre projet sur les accidents routiers en France et leur gravité, nous avions accès à deux sources de données principales :

- https://www.data.gouv.fr

- https://www.kaggle.com

## 1.1 Jeux de données issus de Data.gouv.fr

Sur la plateforme Data.gouv.fr, nous disposions des données couvrant la période de 2005 à 2022, réparties en plusieurs fichiers :

- Usager\_annee.csv

- Véhicules\_annee.csv

- Lieux\_annee.csv

- Caractéristiques\_annee.csv

- Véhicule\_immatricules\_baac\_annee.csv

- Des fichiers descriptifs en format PDF, variables selon les années

- Des bases de données déjà agrégées

### a. Récupération des fichiers

Pour extraire ces fichiers, nous avons utilisé l'outil \*\*Octoparse\*\* afin d'obtenir un fichier CSV contenant l'ensemble des liens vers les différents fichiers de données.

Ensuite, à l'aide de \*\*Python\*\*, nous avons développé une fonction spécifique permettant de télécharger et d'extraire automatiquement toutes les données disponibles sur la plateforme (voir Annexe 1 pour le code).

### b. Analyse des données

Nous avons utilisé un modèle de template pour examiner en détail les colonnes de chaque fichier, en fournissant des informations telles que :

- Le nom de la colonne

- La description de la colonne

- Le type de données

- Le taux de valeurs manquantes (NaN)

- La distribution des valeurs

- Des observations et remarques pertinentes

Cette analyse a révélé que certaines colonnes ont évolué au fil du temps, n'apparaissant qu'à partir de certaines années, ou cessant d'être collectées à partir d'une certaine date.

- Fichier Caractéristiques: Ce fichier décrit les circonstances générales de chaque accident. Il contient 17 colonnes, dont deux sont des doublons (\*Accident\_ID\* et \*Num\_ACC\*). Nous avons constaté un grand nombre de valeurs manquantes dans les données géographiques (coordonnées GPS, adresses).

- Fichier Lieux : Ce fichier décrit le lieu principal de l'accident, y compris lorsque l'accident se produit à une intersection. Il contient 19 colonnes. De nombreuses valeurs sont manquantes, notamment pour des variables liées à la route (ex. indices de route comme 2bis, 3ter), ou des détails comme la distance par rapport à une borne, la largeur du terre-plein central, et la vitesse maximale autorisée.

- Fichier Véhicules : Ce fichier décrit les véhicules impliqués dans l'accident. Il contient 11 colonnes. De nombreuses valeurs sont manquantes, en particulier pour les variables liées au nombre d'occupants, à l'identifiant du véhicule et au type de moteur.

- Fichier Usager : Ce fichier décrit les personnes impliquées dans l'accident. Il contient 17 colonnes. Les valeurs manquantes concernent principalement les variables relatives à la sécurité (celles-ci ayant évolué dans le temps), l'identifiant du véhicule et celui de l'usager.

Nous avons décidé de ne pas inclure, pour le moment, le fichier Véhicule\_immatriculé\_baac dans notre analyse.

## 1.2 Jeux de données issus de Kaggle

Le jeu de données fourni par Kaggle est déjà agrégé et contient 55 colonnes. Toutefois, il ne couvre que la période jusqu'à 2016.

A FAIRE TB

## 1.3 Choix de la source de données et de la variable cible

Après cette première phase d'analyse, nous avons déterminé que la gravité de l'accident serait notre variable cible pour l'étude.

Après une discussion collective, nous avons fait le choix de conserver les données issues de Data.gouv.fr pour deux raisons principales :

1. Maîtrise des données : En sélectionnant cette source, nous avons un contrôle total sur la construction du jeu de données, incluant le choix des années à analyser et des jointures à réaliser.

2. Actualité des données : Les données sur Data.gouv.fr sont plus récentes. Nous estimons que l'évolution des infrastructures routières, les progrès en matière de sécurité des véhicules, ainsi que les mesures mises en place par l'État en faveur de la sécurité routière, sont des éléments importants à prendre en compte dans l'analyse de la gravité des accidents.

# 2 – Analyse du jeu de données

Une fois la source choisie, nous avons analyser avec Python plus en détails les composantes que nous avions à notre disposition en vue de créer un DF pour la data visualisation.

L'objectif de l'analyse a été de traiter et d'examiner des fichiers CSV provenant de différentes années et de divers types. Voici les principales étapes et résultats de l'analyse :

## 2-1. Contexte et Objectif

L'analyse portait sur plusieurs fichiers CSV situés dans un dossier spécifique, couvrant les années de 2005 à 2022. Les fichiers appartenaient à différents types, notamment des caractéristiques, des lieux, des usagers, des véhicules, et des immatriculations de véhicules.

## 2-2. Méthodologie

Les fichiers ont été traités en essayant plusieurs séparateurs (virgule, point-virgule, tabulation) pour assurer une lecture correcte des données. Pour chaque fichier, un DataFrame a été créé et les statistiques descriptives ont été calculées pour chaque colonne. Ces statistiques incluent le type de données, le taux de valeurs manquantes, le mode, les valeurs minimales et maximales, la moyenne, la médiane, et l'écart-type.

## 2-3. Résultats

- Résultats d'Analyse : Un résumé détaillé a été produit pour chaque fichier, incluant le nombre de colonnes et de lignes, les noms des colonnes, et les informations sur le succès ou l'échec de la lecture du fichier.

- Statistiques des Colonnes: Des statistiques détaillées ont été extraites pour chaque colonne des fichiers analysés, fournissant des informations clés sur la distribution des données.

## 2-4. Exportation des Résultats

Les résultats d'analyse et les statistiques des colonnes ont été préparés pour une éventuelle exportation en fichier Excel, facilitant la présentation et l'interprétation des données.

En conclusion, cette analyse a permis de compiler des informations précieuses sur la structure et la qualité des données contenues dans les fichiers CSV, offrant une base solide pour des analyses ultérieures.

# 3 - Concaténation des fichiers par types et fusion d’un DF

A FAIRE TB

# 4 - Visualisation et statistiques

* Avez-vous identifié des relations entre différentes variables ? Entre variables explicatives ? et entre vos variables explicatives et la/les cible(s) ?
* Décrivez la distribution de ces données, répartition, outliers.. (pré/post traitement éventuellement)
* Présentez les analyses statistiques utilisées permettant de confirmer les informations présentes sur les graphiques.
* Tirer des conclusions des éléments relevés ci-dessus permettant de se projeter dans la partie de modélisation

Après avoir fusionné nos différents fichiers, nous avons procédé à la visualisation des données afin d’identifier des relations entre les différentes variables, observer la distribution des données, leur composition.

Afin de pouvoir réaliser la data visualisation des données de manière efficace, nous avons réalisé un premier pre-processing dont les étapes sont résumées ci-dessous :

* Remplacement des valeurs « -1 » (signifiant que la donnée n’a pas été renseignée par les forces de l’ordre) par un NaN ;
* Suppression des variables présentant un taux de valeurs manquantes trop important (gestion des NaN peu réalisable) ;
* Suppression des variables concernant les piétons. La plupart de ces variables sont inutilisables du fait du taux de NaN trop important ;
* Pour la localisation des accidents, du fait d'avoir les latitudes et longitudes, nous pouvons supprimer les autres variables. Nous gardons tout de même la variable dep (département) et com (commune).

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

Description générée automatiquement

Figure 1 : Proportion de valeurs manquantes par variables

Nous représentons ci-dessous le nombre d’accidents par jour durant la période d’étude considérée (2020 à 2022). De manière générale, le nombre d’accidents est stable sur ces années et oscille autour de 300 accidents par jour. A noter toutefois une forte diminution sur les mois de mars et avril 2020, correspondant à la période de confinement liée au Covid-19.

Une image contenant capture d’écran, texte, Tracé

Description générée automatiquement

Figure 2 : Répartition du nombre d'accidents par jour de 2020 à 2022

Le graphique suivant confirme également l’impact du confinement et des restrictions de circulation sur l’année 2020. Concernant les années 2021 et 2022, nous pouvons observer une répartition du nombre d’accidents par mois assez homogène.

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

Description générée automatiquement

Figure 3 : Répartition du nombre d'accidents par mois entre 2020 et 2022

Si nous nous intéressons à la variable gravité, nous pouvons remarquer que une répartition homogène entre les années. La variable gravité est décomposée en 4 classes : l’usager a été après l’accident soit indemne, soit tué, soit blessé hospitalisé ou soit blessé léger. Nous remarquons une hétérogénéité entre les classes avec seulement 2.5% des usagers tués lors d’un accident contre 44% indemne ou 38% étant blessé léger.

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

Description générée automatiquement

Figure 4 : Répartition de la gravité des accidents entre 2020 et 2022

Il est également possible de représenter la distribution de l’âge des usagers accidentés suivant la gravité de leur accident.

Une image contenant texte, diagramme, Tracé, ligne

Description générée automatiquement

Figure 5 : Distribution de l'âge des usagers lors de l'accident

On remarque un pic d'accidents chez les personnes nées aux alentours des 23 ans. Comme les données que nous utilisons proviennent des années 2020, 2021 et 2022, et que les jeunes conducteurs comptent parmi les catégories les plus à risque, ce résultat semble cohérent.

A noter toutefois d’une distribution clairement asymétrique et non normale. Un traitement sera à réaliser en vue du Machine Learning (la plupart des algorithmes de Machine Learning se basent sur l’hypothèse que les variables utilisées suivent une loi normale). Un exemple de traitement possible sera de discrétiser cette variable en créant des groupes.

# 5 -Pre-processing et feature engineering

* Avez-vous eu à nettoyer et à traiter les données ? Si oui, décrivez votre processus de traitement.
* Avez-vous dû procéder à des transformations de vos données de type normalisation/standardisation ? Si oui, pourquoi ?
* Envisagez-vous des techniques de réduction de dimension dans la partie de modélisation ? Si oui, pourquoi ?

La première étape du projet nous a permis de comprendre le jeu de données, de visualiser certaines tendances ainsi que d’avoir déterminé notre variable cible : la gravité des accidents.

Afin de pouvoir utiliser le jeu de données et ainsi travailler avec les différents algorithmes de Machine Learning, nous avons dû transformer nos jeux de données. Les raisons principales sont les suivantes :

* Les algorithmes de Machine Learning nécessitent d’avoir en entrée des formats spécifiques, comme par exemple une classification binaire des variables, des formats numériques.
* Discrétiser certaines variables en sous-ensembles afin d’améliorer les résultats des algorithmes.
* Sélectionner les variables réellement utiles pour les modèles.

Parmi les différentes transformations réalisées, nous avons pu :

* Discrétiser les éléments de sécurité (variable ‘secu1’, ‘secu2’ et ‘secu3’). Ces variables contenaient une liste d’équipements de sécurité qui étaient présents lors de l’accident. Cela pouvait correspondre au port du casque, de la ceinture, l’utilisation du airbag. Les forces de l’ordre ont donc la possibilité de faire remonter la présence de jusqu’à 3 équipements. Afin de corriger une part de valeurs manquantes et pour avoir des variables utilisables, nous avons créé une variable pour chaque équipement, indiquant si celui-ci était présent lors de l’accident.
* Regrouper des modalités ensemble de certaines variables afin de recréer des sous-ensembles plus simple et ainsi améliorer les performances des modèles. Par exemple, la variable ‘place’ indique la position de l’usager dans le véhicule lors de l’accident. Les modalités possibles vont de 1 à 9, avec pour chaque numéro une position dans une voiture. Ces différentes modalités ont été regroupées en 3 sous-ensemble ;chaque sous-ensemble étant une nouvelle variable créée : ‘conducteur’, ‘passager\_avant’, ‘passager-arrière’.

### Recherche de corrélation

Afin d’avoir un premier aperçu sur les liens entre les variables et avec la variable cible, nous avons réalisé une matrice de corrélation afin de faire apparaitre l’importance de certaines variables vis-à-vis de la gravité de l’accident. Cette matrice de corrélation est présentée ci-dessous via une heatmap.

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

Description générée automatiquement

Figure 6 : Matrice de corrélation suivant les variables du jeu de données

Concernant la variable ‘gravité’, nous pouvons voir une corrélation plus marquée avec les variables suivantes :

* Port de la ceinture ;
* Si le véhicule impliqué est une moto ;
* Le type de motorisation ;
* L’obstacle heurté ;
* L’année de naissance de l’usager ;
* …

Il est également possible de voir que certaines variables (autre que la variable cible) sont corrélées entre elles. C’est le cas notamment du port du casque et si l’usager était à moto.

Nous avons également étudié la présence d’un lien statistique entre les variables explicatives et la variable cible. Nos variables sont des variables catégorielles. Nous avons donc utilisé le test du khi 2. Les hypothèses sont décrites ci-dessous :

* H0 : La variable est indépendante de la variable cible (gravité)
* H1 : La variable n'est pas indépendante de la variable cible (gravité)

Nous fixons la p-value à 0.05 : si la valeur de p est inférieure à 0.05, cela signifie que nous pouvons rejeter l’hypothèse H0 et affirmer que le résultat soutient l’hypothèse H1.

Après réalisation du test sur les variables, il en ressort que les p-values sont toutes nulles. Il existe donc bien une relation statistique entre les variables qualitatives et la variables cible. Cela signifie également qu’on ne peut pas utiliser simplement cette méthode pour sélectionner les variables pertinentes pour les modèles. Il sera donc nécessaire de regarder quelles sont les features les plus utiles à chaque modèle pour permettre de répondre à cette question.

Chronologie :

Fréquence de point

Partage des tâches du mode projet

24-06 : Attribution du projet et présentation de Christophe

01-07 : Cadrage initial et découverte de l’équipe

11-07 : Mise en place du drive, analyse des deux sources de données par binôme.

17-07 : Mise en commun exploration des données

18-07 : Choix de la source, mise en place d’un github, data viz et partage du travail

23-07 : Point sur avancement, sélection des variables, méthode de fusion du DF

01-08 : Point sur data viz et choix des visuels à présenter

06-08 : Mise au propre du github, et sélection finale des présentation

08-08 : Point avec Christophe, rajout de visuel type carte, début du préprocessing, amélioration de visuel existant

14-08: A compléter je n’étais pas là

09-09 : Point sur le préprocessing, début du rapport

13-09 : Point avec Christophe pour préprocessing et rapport