L’objectif est de l’exploitation des données est de prédire la gravité corporelle des accidents en fonction des éléments relatifs aux données geotemporelles, équipements, véhicules, et caractéristiques des victimes.
Les jeu de données concernent 218404 accidents, sur les années 2019 à 2022.
Cela concerne 494182 victimes, et 373584 véhicules.

Nettoyage
Après jointure des jeux de données, nous avons donc 494182 lignes, et 51 variables (hors identifiants).
La variable d’intérêt est ‘grav’, classant la gravité des blessures des victimes d’indemne à tué .

Nous recodons les trois colonnes secu1, secu2, secu3 pour avoir une colonne binaire par équipement, et supprimons les trois colonnes d’origine

Nous éliminons 8 colonnes ont plus de 25 % de valeurs manquantes.

Nous éliminons également les colonnes identifiant les routes pour ne garder que leur caractérisation.
Enfin nous éliminons les lignes contenant une valeur manquante.
Nous conservons donc 442992 lignes soit environ 90 % du jeu de données initial.

Nous convertissons les longitudes et latitudes.

Nous calculons l’âge de l’usager dans l’année de l’accident.

Nous ne gardons que l’heure de la variable hrmn
Nous créons une variable date, et supprimons le jour et l’année.

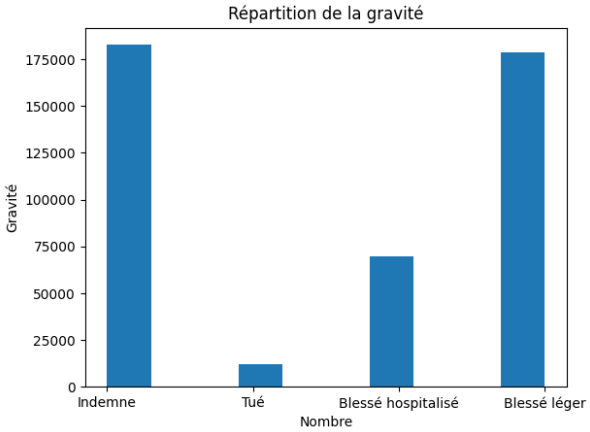
Nous supprimons les colonnes pr et pr1 trop difficiles à interpréter (la localisation sera plus précise avec les latitudes et longitudes) : il faudra peut-être regrouper ces coordonnées en cluster pour les ajouter dans les variables explicatives.

Nous supprimons dep et com pour les mêmes raisons.

Il reste 40 variables explicatives.
Nous convertissons toutes les variables en entier (sauf latitude et longitude).

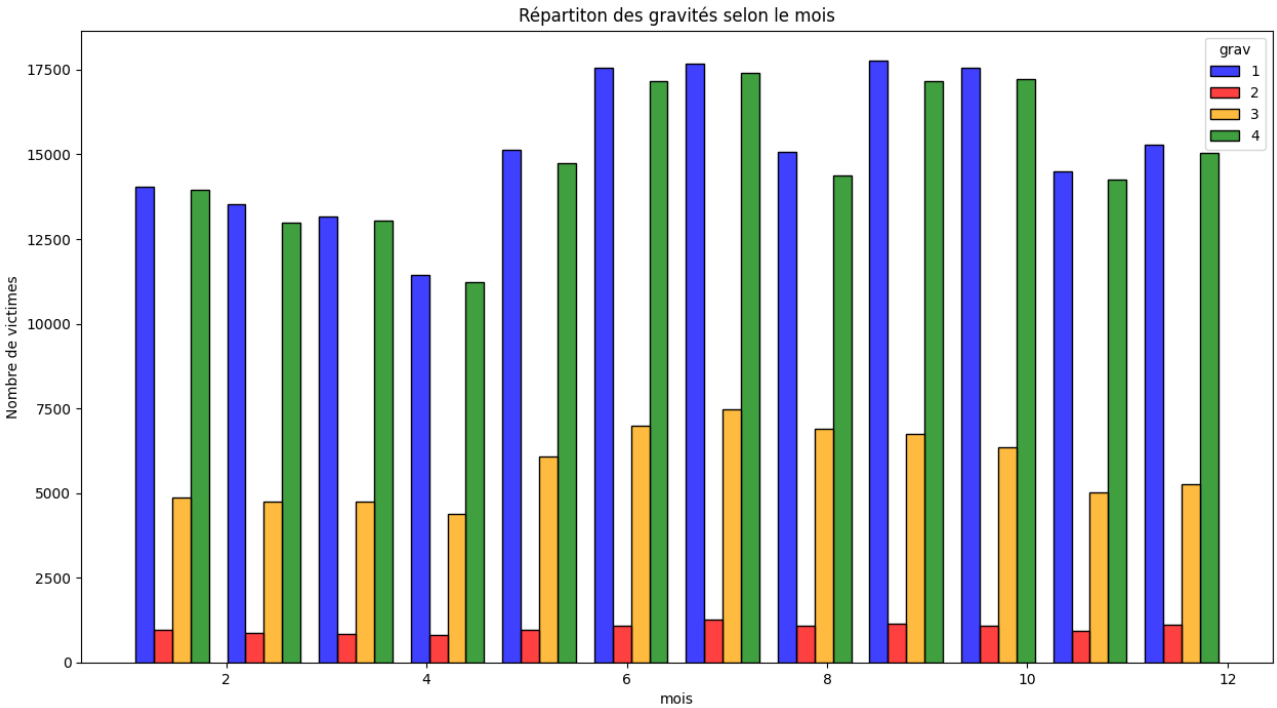
459 doublons ?

Au final, la répartition des gravités est la suivante :



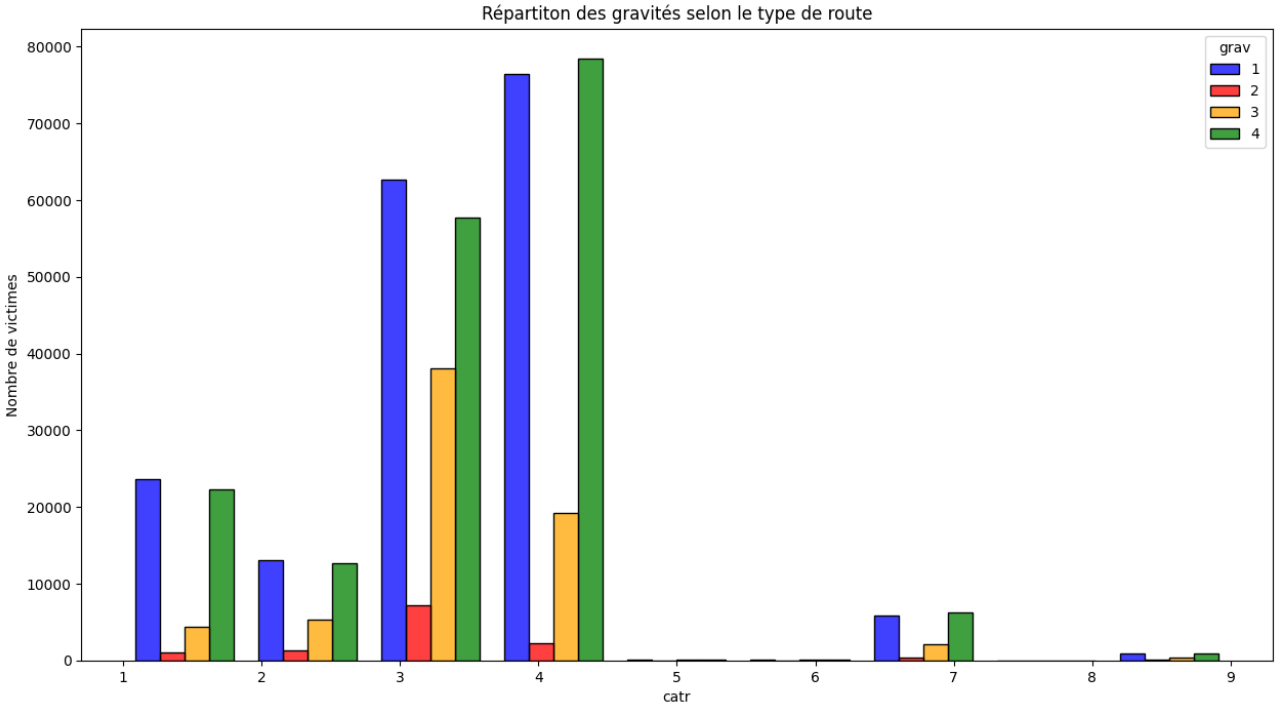
On constate un déséquilibre de la répartition de la gravité, dont il faudra sans doute tenir compte.

La répartition des gravités par mois suggère une saisonnalité.



Un test de chi² de contingence confirme qu'il y a un lien : p-value = 3.7209049978151205e-90

Le type de route semble aussi influencer la gravité : les routes départementales semblent notamment les plus meurtrières…



Là encore, le test de Chi² confirme ce lien.