# Rapport sur l'Exploration des Données et Préprocessing

## Introduction

Ce rapport présente le processus d'exploration des données, le prétraitement et la préparation pour un modèle de machine learning visant à prédire la gravité des accidents sur un usager. Les données analysées proviennent de quatre catégories de dataframes : caractéristiques, lieux, véhicules, et usagers, couvrant les années 2005 à 2022. Toutefois, après une analyse approfondie, nous avons restreint notre étude à la période 2019-2022 en raison des évolutions dans les méthodes de saisie et structures des dataframes.

## Exploration Initiale des Données

Les données initiales comprenaient 72 dataframes répartis comme suit :

- **Caractéristiques** : Prend en compte les circonstances générales de l'accident.
- **Lieux** : Décrit l'endroit de l'accident.
- **Véhicules** : Décrit les véhicules impliqués dans l'accident.
- **Usagers** : Décrit les usagers impliqués dans l'accident.

Pour chaque catégorie, il y avait 18 dataframes correspondant aux années 2005 à 2022.

### Problématique des Méthodes de Saisie et Structures

Les méthodes de saisie et les structures des dataframes ont évolué au fil des ans, rendant certains dataframes incompatibles entre eux. Cela a conduit à la décision de concentrer notre analyse sur les données des années 2019 à 2022, période durant laquelle les méthodologies de saisie ont été stabilisées et standardisées.

## Analyse et Prétraitement des Données

### Description des Variables

Pour chaque variable dans les dataframes, nous avons cherché :

- Sa description.
- Son type (et l'évolution de celui-ci au cours des années).
- L'étendue des valeurs.
- Les valeurs nulles.
- Les outliers.
- La répartition des modalités au fil des années.

Voir documentation complète: Exploration des variables - rapport EM V240624.docx

## Prétraitement Initial

Nous avons effectué un premier prétraitement pour chaque colonne afin de gérer :

- Les valeurs manquantes.
- Les outliers.
- Les types de données.

## Harmonisation et Fusion des Dataframes

Pour assurer la compatibilité des dataframes, nous avons harmonisé leur structure de la manière suivante :

- Uniformisation des types et structures de chaque dataframe d'une même catégorie.
- Concatenation des dataframes homogénéisés pour chaque catégorie.
- Fusion des quatre catégories de dataframes en un unique dataframe, où chaque ligne représente un usager unique.

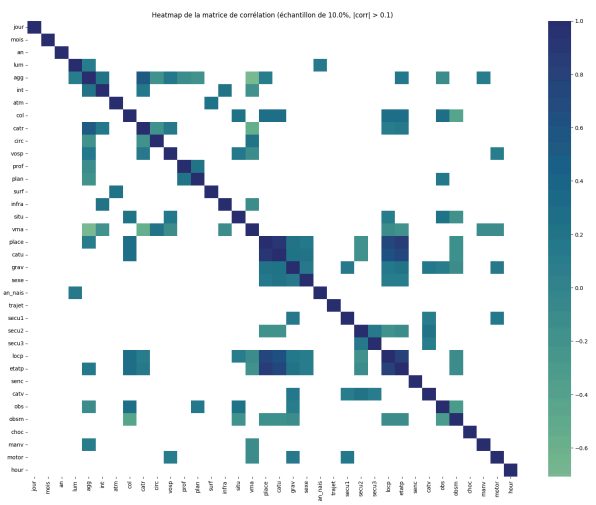
## Prétraitement Post-Fusion

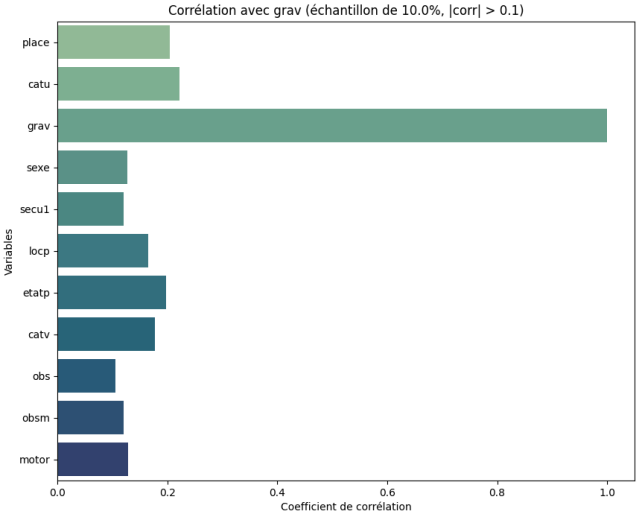
Après la fusion des dataframes, nous avons effectué un second traitement pour :

- Traiter les valeurs manquantes résultant de la fusion.
- Décider des colonnes à conserver pour l'analyse finale.
- Encoder les valeurs catégorielles.
- Standardiser les valeurs numériques.

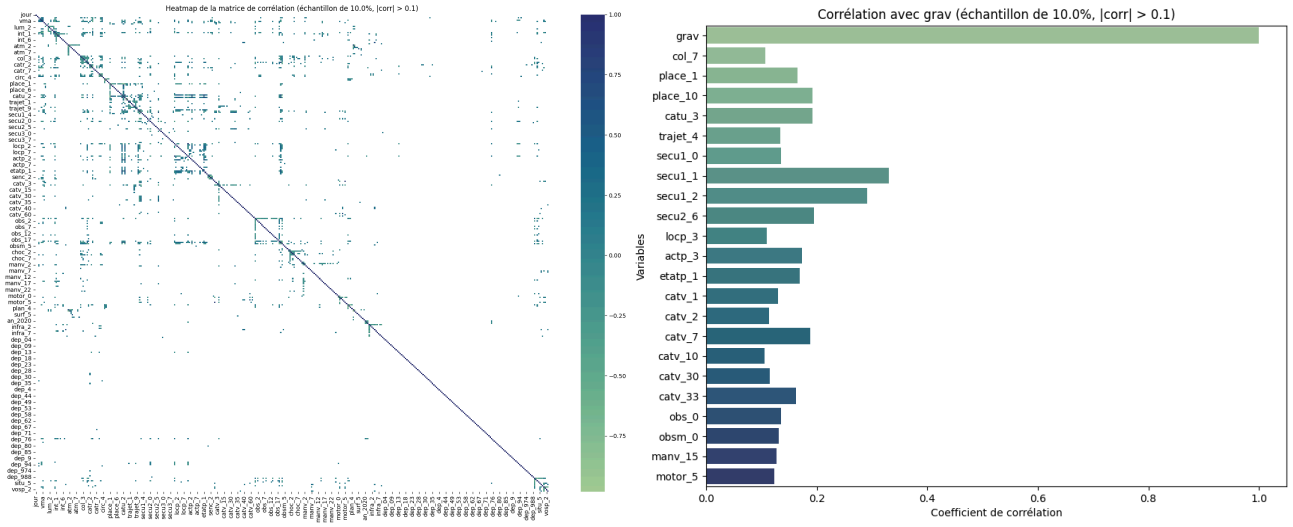
## Heatmaps

Avant encodage:



Après encodage:




# Conclusion

Le processus d'exploration et de prétraitement des données a permis de préparer un dataset consolidé et homogène pour le développement d'un modèle de machine learning. L'objectif final est de prédire la gravité d'un accident pour un usager en utilisant les données les plus récentes et les mieux structurées possibles. La prochaine étape consistera à développer, entraîner et évaluer des modèles de machine learning sur ce dataset préparé.