Sujet – Projet de SAS

# Dataset

Le dataset présente l’intégralité des données relatives aux accidents corporels (mortels ou non) de la circulation routière relevé par les forces de l’ordre.

* La rubrique **CARACTERISTIQUES** qui décrit les circonstances générales de l’accident
* La rubrique **LIEUX** qui décrit le lieu principal de l’accident même si celui-ci s’est déroulé à une intersection
* La rubrique **VEHICULES** impliqués
* La rubrique **USAGERS** impliqués

Le n° d'identifiant de l’accident (Cf. "Num\_Acc") présent dans ces 4 rubriques permet d'établir un lien entre toutes les variables qui décrivent un accident. Quand un accident comporte plusieurs véhicules, il faut aussi pouvoir relier chaque véhicule à ses occupants. Ce lien est fait par la variable id\_vehicule.

La plupart des variables contenues dans les quatre fichiers précédemment énumérés peuvent contenir des cellules vides ou un zéro ou un point. Il s’agit, dans ces trois cas, d’une cellule non renseignée par les forces de l'ordre ou sans objet.

# Questions

1. Pré-traitement des données
   1. Détection des valeurs nulles, les remplacer ?
   2. Joindre les quatre fichiers en un seul
   3. Régler soucis d'encodage
2. Statistiques descriptives (qu'on pourrait faire en SAS et en SQL)
   1. TOP x / FLOP x
   2. Evolution entre les années, croisé avec d'autres variables catégorielles
   3. Corrélation entre plusieurs variables
3. Statistiques inférentiels
4. Classification