



# **Rapport technique 1 :** **Statistiques Descriptives**

Projet Dasci - UE B/C

**Groupe 11** - Marianne BELLERY, Nur BACKARI, Emma RAULET, Miguel GUZMAN

# Table des matières

<b>1) Problématique métier</b>	<b>3</b>
Contexte	3
Objectifs du projet	3
<b>2) Jeux de données utilisés</b>	<b>4</b>
Base de données des accidents corporels en 2021.	4
<b>3) Description statistique des données</b>	<b>4</b>
Description des données	5
Analyse statistique des données	6
Vérification de la qualité des données	6
<b>4) Préparation des données du dataset ‘caractéristiques’</b>	<b>6</b>
Nettoyage des données	6
Construction des nouvelles données	7
Intégration des données	7
Reformatage des données	8
<b>5) Exploration des données</b>	<b>8</b>
<b>6) Préconisations d’algorithmes de machine learning</b>	<b>10</b>
<b>7) Bibliographie</b>	<b>10</b>

# 1) Problématique métier

## Contexte

D'après l'Observatoire national interministériel de la sécurité routière, 2944 personnes ont perdu la vie dans un accident de la route en 2021. Malgré que ce chiffre soit en baisse par rapport à l'année de référence 2019, le nombre d'accidents en France s'élèvent à plus de 50 000 par an. Afin d'améliorer la sécurité des automobilistes, motards, cyclistes, piétons et tous les utilisateurs du réseau routier, on cherche à faire des préconisations d'aménagements routiers à mettre en place pour limiter le nombre d'accidents corporels sévères et mortels.

## Objectifs du projet et clients visés

L'objectif de ce projet est de développer un outil qui permette de proposer des recommandations aux acteurs responsables des routes des solutions à mettre en place pour réduire les accidents graves liés à des manquements d'aménagements routiers. Les objectifs secondaires sont d'abord d'analyser les causes et circonstances d'accidents corporels graves afin de comprendre les potentiels défauts liés au réseau routier. Ensuite, d'être capable de faire des recommandations automatiques d'aménagements à mettre en place en fonction des aménagements déjà mis en place et de l'analyse de la zone accidentogène.

Nos clients sont les administrations gouvernementales responsables de l'entretien des infrastructures routières françaises. En France, il existe plusieurs types de route (autoroutes, départementales,...) et donc plusieurs types de responsables qui s'occupent de la construction, la maintenance, la gestion, l'entretien et les travaux.

La prise en charge de l'entretien des routes se fait donc :

- pour les voies communales, par le conseil municipal de la commune
- pour les voies départementales, par le conseil départemental. (sauf Alsace)
- pour les voies nationales, par les Directions Interdépartementales des Routes (DIR).
- pour les autoroutes non concédées, par les DIR.
- pour les autoroutes concédées, par sociétés concessionnaires en contrat avec l'Etat.

La **valeur ajoutée** de notre produit réside dans le gain de temps et d'efficacité, permis par l'automatisation et la recherche de solution, apporté aux administrations en charge du réseau routier.

Cette problématique répond à l'objectif développement durable **11** "Villes et communautés durables", qui a pour but de faire en sorte que les villes et les établissements humains soient ouverts à tous, sûrs, résilients et durables.

## 2) Jeux de données utilisés

### Base de données des accidents corporels en 2021.

Les jeux de données que nous allons utiliser sont extraits du site **data.gouv**. (téléchargeables directement via le site en csv), plus précisément de la base de données annuelle des accidents corporels de la circulation routière allant de 2005 à 2021. Cette base de données mise à jour par le gouvernement est composée de 4 jeux de données par année de 2005 à 2021. On étudiera les jeux de données associés à l'année 2021. Les datasets utilisés proviennent uniquement de data.gouv, donc sont mis à disposition par l'état, il n'y a aucun problème juridique lié à leur exploitation.

Les 4 jeux de données sont les suivants :

- Caractéristiques - Lieux - Usagers - Véhicules

Afin de mener une première analyse des accidents corporels de la route, on s'intéressera au dataset '**caractéristiques**' (caractéristiques-2021.csv) que nous avons complété par les dataset 'lieux', 'usagers' et 'véhicules'. Le dataset '**caractéristiques**' donne des informations générales sur les circonstances de l'accident (jour/nuit, conditions météo, ...). Le dataset '**lieux**' donne des informations sur le lieu de l'accident et la typologie de la route. Le dataset '**usagers**' donne des informations sur le comportement et l'état des usagers et le dataset '**véhicules**' donne des informations sur les modèles véhicules prenant part à l'accident et les dommages causés.

## 3) Description statistique des données

Le jeu de données 'caractéristiques' contient 15 colonnes.

## Description des données

Nom	Description	Type	Utilisation
Num_Acc	Numéros d'identifiant de l'accident	Int	permet d'identifier les accidents et joindre les quatres datasets
jour mois	Mois de l'accident	Int	donne la date
an	Année de l'accident	Int	donne l'année
hrmn	Heure et minute de l'accident	ab:cd <i>a,b,c,d : int,int,int,int</i>	donne l'heure
lum	Lumière/éclairage	Int	variables explicatives donnant la qualité de l'éclairage lors de l'accident
dep	Département	Int	donne le département
com	Commune	Int	donne la commune
agg	Localisation	Int	donne la localisation
int	Type d'intersection	Int	variables explicatives donnant la typologie de la route
atm	Conditions atmosphériques	Int	variables explicatives donnant les conditions atmosphériques
col	Type de collision	Int	donne la localisation de
adr	Adresse postale	String	donne l'adresse précise
lat	Latitude	Float	permet de cartographier
long	Longitude	Float	permet de cartographier

## Analyse statistique des données

Dataset description:

Numerical Data

	Num_Acc	jour	mois	an	lum	\	Null Values per Feature
count	5.651800e+04	56518.000000	56518.000000	56518.0	56518.000000		Num_Acc 0
mean	2.021000e+11	15.764394	6.867087	2021.0	1.835398		jour 0
std	1.631549e+04	8.794004	3.295277	0.0	1.437602		mois 0
min	2.021000e+11	1.000000	1.000000	2021.0	1.000000		an 0
25%	2.021000e+11	8.000000	4.000000	2021.0	1.000000		hrmn 0
50%	2.021000e+11	16.000000	7.000000	2021.0	1.000000		lum 0
75%	2.021000e+11	23.000000	10.000000	2021.0	2.000000		dep 0
max	2.021001e+11	31.000000	12.000000	2021.0	5.000000		com 0
							agg 0
count	56518.000000	56518.000000	56518.000000	56518.000000			int 0
mean	1.642574	2.077374	1.623642	4.004600			atm 0
std	0.479246	2.020099	1.707091	1.962765			col 0
min	1.000000	1.000000	-1.000000	-1.000000			adr 573
25%	1.000000	1.000000	1.000000	3.000000			lat 0
50%	2.000000	1.000000	1.000000	3.000000			long 0
75%	2.000000	2.000000	1.000000	6.000000			
max	2.000000	9.000000	9.000000	7.000000			dtype: int64

Dataset description:

Categorical Data

	hrmn	dep	com	adr	lat	long
count	56518	56518	56518	55945	56518	56518
unique	1374	107	11150	29668	54618	54921
top	18:00	75	75116	AUTOROUTE A86	-17,5845220000	-149,5685780000
freq	769	5069	507	351	11	11

## Vérification de la qualité des données

Dans les datasets 'caractéristiques', on constate qu'il existe des données manquantes. Sinon, les données sont cohérentes et bien renseignées.

## 4) Préparation des données du dataset 'caractéristiques'

### Nettoyage des données

On remarque que 573 adresses ne sont pas renseignées. Afin que cela ne perturbe pas l'analyse des données, on décide de supprimer la colonne associée. On supprime donc la colonne '**adr**' qui recense les adresses des accidents. Cependant, cela ne nuit pas à la qualité des données puisqu'on utilise les attributs '**lat**' et '**long**' (latitude, longitude) pour placer les localisations des accidents sur la carte.

Notre analyse est localisée en France métropolitaine. On décide de trier les lignes et de supprimer les accidents ayant eu lieu dans les départements d'outre mer. On supprime alors les accidents associés à des numéros de départements supérieurs ou équivalents à 971.

### Construction des nouvelles données

Afin de faciliter l'analyse des conditions météorologiques et temporelles des accidents de voiture, nous avons décidé de construire un attribut dérivé des mois à propos de la saison (printemps, automne, été, hiver). Pour cela, nous avons utilisé la colonne 'mois' et fait des groupements de mois par saison.

De la même manière, on décide de construire un nouvel attribut dérivé de l'heure. On réalise des groupements d'heure traduisant des moments de la journée. Pour cela, on passe par un attribut intermédiaire. On distingue 6 catégories : 'tot le matin', 'matin', 'après-midi', 'nuit', 'tard la nuit'.

### Intégration des données

Afin d'accumuler les données sur les accidents de la route, nous avons décidé de combiner plusieurs datasets. Nous avons combiné le dataset 'caractéristiques' avec les datasets lieux, véhicules et usagers en effectuant la jointure sur l'attribut "**num\_acc**", l'identifiant de l'accident. La jonction permet de compléter le dataset caractéristiques avec les attributs utiles suivants :

Nom	Description	Type	Utilisation
grav	Gravité des blessures de l'un des usagers	Int	variable d'intérêt, indique la gravité des blessures subies par le blessé
catu	Catégorie d'usager du véhicule	Int	variable explicative sur l'usager
place	Place de l'usager dans le véhicule	Int	variable explicative sur l'usager
trajet	Motif de déplacement	Int	détail sur les habitudes de l'usager
secu1	Dispositif de sécurité utilisé	Int	circonstances de l'accident

secu2	Dispositif de sécurité utilisé	Int	circonstances de l'accident
secu3	Dispositif de sécurité utilisé	Int	circonstances de l'accident
catr	Caractéristique de la voie	Int	permet de lier la recommandation au responsable de la route
circ	Régime de circulation	Int	variables explicatives complémentaires sur le type de voie
surf	Etat de la surface au moment de l'accident	Int	détail des circonstances de l'accident
catv	Caractéristique du véhicule	Int	variables explicatives sur le véhicule
obs	Obstacle heurté	Int	détail dommage de l'accident
obsn	Obstacle mobile heurté	Int	détail dommage de l'accident

### Reformatage des données

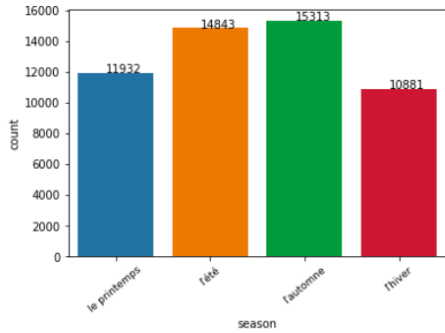
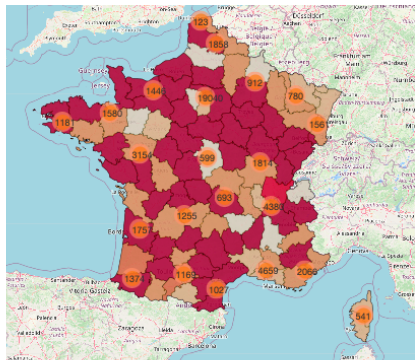
Lorsque nous avons exploré l'ensemble de données, nous avons vu que le type de colonne **"lat"** et **"long"** est un objet. Afin de ne pas avoir de problème lorsque nous traçons les colonnes **"lat"** et **"long"** sur la carte de France, on changeo le type de ces deux colonnes. Pour réaliser cette opération, on transforme les virgules en points. Ensuite, on transforme 'lat' et 'long' en float et les 'mois' en chaînes de caractères. Nous voulons également utiliser des noms de mois (janvier, février) au lieu de chiffres, nous devons donc également changer le type de colonne **mois**.

## 5) Exploration des données

Nom	Histogramme	Description	Commentaire
-----	-------------	-------------	-------------



Etat des victimes		Répartition des victimes impliquées dans les accidents en fonction de la gravité de leur état	La grande majorité des victimes sont “indemnes” ou “blessés légers”
Collisions		Répartition des accidents en fonction du type de collision	-un pic significatif se distingue : “Deux véhicules-par le côté” (le pic 6 étant “autre collision”) -les accidents “sans collision” (pic 7) sont largement minoritaires
Luminosité		Répartition des accidents en fonction des conditions d'éclairage	La majorité des accidents ont lieu en plein jour
Distribution mensuelle		Répartition mensuelle des accidents sur l'année 2021	La période de estivale (juin-juillet) et de rentrée (sept-oct) sont les plus accidentogènes

Distribution saisonnière	 <table><thead><tr><th>season</th><th>count</th></tr></thead><tbody><tr><td>le printemps</td><td>11932</td></tr><tr><td>l'été</td><td>14843</td></tr><tr><td>l'automne</td><td>15313</td></tr><tr><td>l'hiver</td><td>10881</td></tr></tbody></table>	season	count	le printemps	11932	l'été	14843	l'automne	15313	l'hiver	10881	Répartition saisonnière des accidents sur l'année 2021	La période de vacances (été) et de rentrée (automne) sont les plus accidentogènes
season	count												
le printemps	11932												
l'été	14843												
l'automne	15313												
l'hiver	10881												
Distribution spatiale		Répartition géographique des accidents sur l'année 2021	La région île-de-France représente la zone avec le plus d'accidents										

## 6) Préconisations d'algorithmes de machine learning

Afin de faire des recommandations pertinentes, on préconise l'utilisation d'algorithmes de clustering pour classifier les conditions et circonstances 'optimales' d'accidents corporels de la route. On s'intéressera en particulier aux dommages provoqués par des aménagements routiers. En vue de prédire la gravité des accidents corporels de la route, on utilisera plusieurs algorithmes de machines learning en particulier RandomForest.

## 7) Bibliographie

Types de routes en France et qui les gèrent :

<https://www.dir.est.developpement-durable.gouv.fr/la-route-qui-gere-quoi-r114.html>

Dataset listant les accidents français de la route et leurs caractéristiques :

<https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2021/>

Liste des aménagements routiers :

<https://tel.archives-ouvertes.fr/tel-02390950/document>