PROJET RÉALISÉ PAR

L'ÉQUIPE ALCOLOCO DU GROUPE DE TD1

RAPPORT DE GROUPE DES UE BASES DE DONNÉES + SCIENCES DES DONNÉES 2

Bouteyre Maxime Boccaccio Mélissa Seveyrat Camille Petiot Mika

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Mai 2025

Soumis comme contribution partielle
pour le cours Science des données 2 et Bases de données

Déclaration de non plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation:

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature:  Date: 02/05/2025

Signature:  Date: 02/05/2025

Signature:  Date: 02/05/2025

Signature: _____ Date: _____

May 2, 2025

Remerciements

Nos plus sincères remerciements vont à nos enseignantes de Bases de données et de Sciences des données, pour leurs conseils tout au long du projet.

Leurs orientations méthodologiques et leurs retours nous ont permis de mieux structurer notre travail et d'avancer de manière plus efficace dans nos analyses.

Nous remercions également chaque membre du groupe pour sa participation, qui a rendu ce projet à la fois collaboratif et enrichissant.

May 2, 2025

Résumé

L'analyse des accidents de la route constitue un enjeu majeur pour orienter les politiques de sécurité et de prévention.

Ce projet porte sur les accidents survenus dans le département de l'Hérault en 2023, avec pour objectif d'identifier les facteurs associés à la gravité des blessures.

L'étude s'appuie sur l'exploitation de données ouvertes, nettoyées et croisées entre plusieurs sources (usagers, véhicules, lieux et caractéristiques liés accidents). Trois grandes familles de facteurs ont été examinées : humains (âge, sexe, équipement), environnementaux (météo, luminosité) et liés à l'infrastructure routière (type de route, vitesse maximale autorisée, type de collision).

Certains facteurs se sont révélés être très influents sur la gravité des accidents, notamment la plupart des facteurs liés à la route. Mais à l'inverse, d'autres caractéristiques semblent moins discriminantes sur le plan statistique.

Ces observations permettent de formuler des recommandations concrètes en matière d'aménagement, de réglementation et de sensibilisation des usagers.

Table des matières

Chapitre 1

Introduction

1.1 Introduction

L'étude des accidents de la route survenus au cours d'une année est utile pour prendre des mesures de sécurité, d'aménagement urbain ou encore de sensibilisation visant à réduire leur nombre. C'est pourquoi l'étude des caractéristiques des accidents ou encore des acteurs de ces accidents est primordiale pour obtenir des résultats réprésentatifs. Dans le cadre d'un travail universitaire nous nous sommes interrogés sur la problématique suivante :

"Quels sont principaux les facteurs contribuant aux accidents de la route dans l'Hérault en 2023, et comment adapter les mesures de prévention pour réduire leur nombre et leur gravité ?"

La question des accidents corporels de la route reste encore aujourd'hui au coeur des mesures de prévention publique. Dans un monde où les populations ne cessent de croître et où les gens ne circulent que davantage, se demander de quelle manière améliorer la sécurité des usagers est nécessaire. Elle peut d'ailleurs se présenter sous deux aspects, comment promouvoir de nouvelles manières de circuler mais aussi comment adapter les infrastrutures déjà existantes au vu des données que l'on dispose.

1.2 Responsabilités et composition de l'équipe

Principaux rôles de chacun :

- **Bouteyre Maxime (Étudiant n°22313124)** : Responsable de la collecte des données, Participation à la rédaction aux parties Introduction et BD du rapport, Réalisation de requêtes SQL, co-responsable de la coordination de la vidéo de présentation.
- **Seveyrat Camille (Étudiant n°2230344)** : Participation à la collecte des données, Responsable de la rédaction et de mise en page du rapport, Réalisation de requêtes SQL, Réalisation de l'ensemble des graphiques, des tests statistiques et conclusions.
- **Boccaccio Mélissa (Étudiant n°22400372)** : Responsable de la collecte de données, Responsable du nettoyage et filtrage des données, Réalisation de requêtes SQL, Co-responsable de la coordination de la vidéo de présentation, Réalisation des diapositives et montage de la vidéo de présentation.
- **Petiot Mika (Étudiant n°22313118)** : Responsable de la collecte de données, Réalisation de requêtes SQL.

Chapitre 2

Base de données

2.1 Provenance des données

Les données utilisées pour l'étude ont été extraites sur le site data.gouv et obtenables via le lien suivant : data.gouv.fr.

Elles portent sur les accidents de la route en 2023. Notre base de données est composée de 4 jeux de données différents contenant des informations sur les usagers, les lieux d'accidents, les véhicules impliqués ou encore les caractéristiques de l'accident en lui-même.

Le premier fichier CSV du jeu de données, intitulé `usagers_filtre.csv`, regroupe des informations concernant les usagers de la route. Dans le cadre de cette étude, nous avons décidé de retirer deux colonnes initialement présentes : `etatp` et `place`. La première indiquait si un piéton était accompagné, tandis que la seconde précisait la place occupée par l'usager dans le véhicule. Ces deux variables ont été jugées non pertinentes pour répondre à notre problématique.

Le second fichier CSV, intitulé `vehicule-2023.csv`, contient des informations relatives aux véhicules impliqués dans les accidents. Deux colonnes ont été retirées de ce fichier : `senc` et `occutc`. La première indiquait le sens de circulation du véhicule, tandis que la seconde renseignait le nombre de passagers à bord d'un transport en commun (uniquement si le véhicule concerné en était un). Ces variables ont été jugées peu pertinentes pour notre analyse.

Le troisième fichier CSV, intitulé `lieux-2023.csv`, porte sur les lieux des accidents. Plusieurs colonnes ont été supprimées : `V1`, `V2`, `vosp`, `pr`, `pr1`, `lartpc` et `larrout`. Ces variables contenaient soit des informations trop détaillées et difficiles à exploiter, soit des données manquantes n'apportant pas d'intérêt analytique.

Enfin, le quatrième fichier CSV, intitulé `caracteristiques-2023.csv`, regroupe les principales caractéristiques des accidents : date, heure, conditions de luminosité, conditions météorologiques, etc. Ce fichier constitue la base principale pour comprendre le contexte global des événements étudiés.

2.2 Orientation de l'étude

Afin de structurer notre analyse, nous avons choisi de regrouper les variables disponibles en trois grandes catégories de facteurs susceptibles d'influencer la survenue et la

gravité des accidents de la route : les facteurs humains, les facteurs environnementaux et les facteurs liés à l'infrastructure routière.

Cette organisation permet d'aborder l'étude de manière logique et complète :

- les facteurs humains englobent les caractéristiques des usagers (âge, sexe, comportement, équipement...) ;
- les facteurs environnementaux concernent les conditions dans lesquelles se produit l'accident (météo, luminosité, état de la chaussée...) ;
- les facteurs liés à la route incluent les aspects du réseau routier (catégorie de route, vitesse autorisée, présence d'intersections...).

Cette structuration reflète la manière dont les accidents sont généralement abordés dans les analyses de sécurité routière, en distinguant les causes liées aux individus, au contexte, et au système de circulation lui-même.

2.3 Descriptif des tables

1. Les usagers

Nom colonne	**Type**	**Signification**	**Caractéristique**
id_usager	Int	Identifiant unique de l'usager (y compris piétons)	Unique, clef primaire
Num_acc	Int	Numéro d'identifiant de l'accident	Unique, clef étrangère
num_veh	Int	Identifiant du véhicule pour chaque usager	Code alphanumérique
catu	Int	Catégorie d'usager	
grav	Int	Gravité de blessure de l'usager	
sexe	Int	Sexe de l'usager	
an_nais	Int	Année de naissance de l'usager	
trajet	Int	Motif du déplacement au moment de l'accident	
secu1 / secu2 / secu3	Int	Présence et utilisation d'un équipement de sécurité	

Table 2.1: Description des variables de la table `usagers_filtre` (1548×15)

2. Les véhicules

Nom colonne	**Type**	**Signification**	**Caractéristique**
id-vehicule	Int	Identifiant unique du véhicule (avec usagers rattachés)	Unique, clef primaire
Num_acc	Int	Numéro d'identifiant de l'accident	Clef étrangère
num-veh	Int	Identifiant du véhicule pour chaque usager	Code alphanumérique
catv	Int	Catégorie du véhicule	
obs	Int	Obstacle fixe heurté	
obsm	Int	Obstacle mobile heurté	
choc	Int	Point de choc initial	
manv	Int	Manœuvre principale avant l'accident	
motor	Int	Type de motorisation du véhicule	

Table 2.2: Description des variables de la table `vehicule_filtre` (1154 × 10)

3. Les lieux

Nom colonne	**Type**	**Signification**	**Caractéristique**
id_lieux	Int	Identifiant du lieu	Unique, clef primaire
Num_acc	Int	Numéro identifiant de l'accident	Clef étrangère
catr	Int	Catégorie de la route	
voie	Varchar	Numéro ou nom de la route	
surf	Int	État de la surface de la route	
infra	Int	Aménagements - Infrastructures	
vma	Int	Vitesse maximale autorisée	

Table 2.3: Description des variables de la table `lieux_filtre` (660 × 12)

4. Les caractéristiques

Nom colonne	Type	Signification	Caractéristique
Num_acc	Int	Numéro identifiant de l'accident	Unique, clef primaire
id_lieux	Int	Identifiant du lieu de l'accident	Clef étrangère
jour	Int	Jour de l'accident	
mois	Int	Mois de l'accident	
hrmn	Int	Heure et minute de l'accident	
lum	Int	Conditions d'éclairage lors de l'accident	
dep	Int	Code INSEE du département	
com	Int	Code INSEE de la commune	
agg	Int	Localisation (agglomération ou hors)	
inte	Int	Présence d'une intersection	
atm	Int	Condition atmosphérique	
col	Int	Type de collision	
longi	Int	Longitude	
lat	Int	Latitude	

Table 2.4: Description des variables de la table `caract_filtre` (660×16)

2.4 Modèles MCD et MOD

- Modèles MCD et MOD réalisés sous Mocodo online :

```
LIEUX: id_lieux, Num_Acc, catr, voie, surf, infra, vma

CARACTERISTIQUE, 11 LIEUX , 1N VEHICULE, 1N USAGER: Num_Acc, id_lieux, jour, mois,
hrmn, lum, dep, com, agg, int, atm, col, long, lat

VEHICULE: id_vehicule, Num_Acc, num_veh, catv, obs, obsm, choc, manv, motor
CONDUIRE, 11 VEHICULE, 1N USAGER
USAGER: id_usager, Num_Acc, id_vehicule, num_veh, catu, grav, sexe, an_nais, trajet,
secu1, secu2

ETRE PASSAGER, 01 USAGER, 1N VEHICULE
```

Figure 2.1: Modèle Organisationnel de Données (MOD)

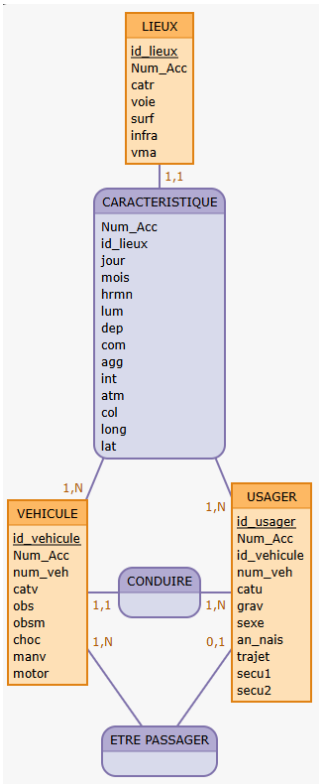


Figure 2.2: Modèle Conceptuel de Données (MCD)

2.5 Import des données ^

- Avant l'import des données dans phpMyAdmin (Mamp), nous avons procédé à un nettoyage puis un filtrage des données. Nous avons dans un premier temps retiré des tables certaines de leurs colonnes, qui contenaient des données peu utiles pour répondre à la problématique. Dans un souci de taille limite pour l'import des données, nous avons aussi du réduire le nombre maximal de nos individus dans nos tables. Pour cela nous avons décidé de concentrer notre étude sur le département de l'Hérault en conservant uniquement les données pour lesquelles la variable dep de la table caract_filtre valait 34. Cette étape de filtration a été réalisée à l'aide d'un programme Python.
- Pour pouvoir enfin importer les données dans SQL (Mamp), il a fallu créer les tables pour les contenir. Nous avons donc élaboré un script SQL permettant de les créer.

2.6 Requêtes réalisées

Nous avons réalisé plusieurs requêtes SQL directement sur la base de données afin d'obtenir un premier aperçu des relations possibles entre certaines variables.
Ces requêtes ont permis de dégager des tendances générales, d'identifier des sous-groupes intéressants, ou encore de vérifier la qualité des données.

Les requêtes SQL complètes ne sont pas toutes affichées sur le document PDF, afin de pouvoir alléger le rapport. Les résultats eux sont tous affichés, présentés sous forme de tableau, avec une analyse des résultats pour chacune des requêtes.

Facteurs humains

- Répartition de la mortalité de l'accident par tranche d'âge et par sexe:

```sql
SELECT u.sexe, CASE WHEN u.age < 18 THEN 'Moins de 18' WHEN u.age BETWEEN 18 AND 35
THEN '18-35' WHEN u.age BETWEEN 36 AND 60 THEN '36-60' ELSE '60+' END AS tranche_age,
COUNT(*) AS nb_graves FROM usagers_filtre u JOIN vehicules_filtre v
ON u.id_vehicule = v.id_vehicule JOIN caract_filtre c ON v.Num_acc = c.Num_acc
WHERE u.grav = 2 GROUP BY u.sexe, tranche_age ORDER BY u.sexe, tranche_age;
```

Figure 2.3: Exemple de code SQL - Première requête

sexe	tranche_age	nb_graves
1	18-35	16
1	36-60	25
1	60+	10
1	Moins de 18	3
2	18-35	5
2	36-60	2
2	60+	9
2	Moins de 18	1

Analyse des résultats

On observe une prédominance masculine dans l'implication dans les accidents mortels (environ 76% d'hommes) et que les tranche d'âge les plus touchées chez les deux sexes sont : pour les femmes celles âgées de plus de 60 ans et chez les hommes, ceux âgés de 35 à 60 ans.

- Influence du port de la ceinture sur la gravité des blessures des usagers (limité aux véhicules où une ceinture est normalement utilisée) :

grav	nb_accidents	nb_avec_ceinture	nb_sans_ceinture
Indemne	952	783	169
Tué	83	39	44
Blessé hospitalisé	464	228	236

Blessé léger	670	454	216

Analyse des résultats

Total sans ceinture = 665
Total avec ceinture = 1 504

Proportion tués avec ceinture = 2,593%
Proportion tués sans ceinture = 6,617%
Proportion blessés hospitalisés avec ceinture = 15,16%
Proportion blessés hospitalisés sans ceinture = 35,489%

En ne portant pas la ceinture de sécurité, les usagers s'exposent à un risque environ 2,5 fois plus élevé de mourir ou d'être blessé lors d'un accident.

- Port des équipement de sécurité (casque ou ceinture) selon le type de véhicule des usagers, triés par tranche d'âge :

type_vehicule	tranche_age	total_usagers	nb_avec_equipement_volontaire
Moto	18-35	91	88
Vélo	18-35	15	6
Voiture	18-35	354	310
Moto	36-60	95	94
Vélo	36-60	12	10
Voiture	36-60	365	314
Moto	60+	25	25
Vélo	60+	17	10
Voiture	60+	220	171
Moto	Moins de 18	7	7
Vélo	Moins de 18	4	1
Voiture	Moins de 18	67	52

Analyse des résultats

Voiture : les taux d'équipement sont élevés mais légèrement décroissants avec l'âge :

- Moins de 18 ans : 77,6 %
- 18–35 ans : 87,6 %
- 36–60 ans : 86,0 %
- 60 ans et plus : 77,7 %

Moto : les taux sont excellents dans toutes les tranches d'âge :

- Moins de 18 ans : 100 %
- 18–35 ans : 96,7 %
- 36–60 ans : 98,9 %

- 60 ans et plus : 100 %

Vélo : les taux sont beaucoup plus faibles, en particulier chez les plus jeunes :

- Moins de 18 ans : 25 %
- 18–35 ans : 40 %
- 36–60 ans : 83,3 %
- 60 ans et plus : 58,8 %

On voit donc que les usagers de moto sont les mieux protégés, avec des taux proches ou égaux à 100 %. Les moins bien protégés sont les cyclistes, surtout chez les jeunes.

Facteurs environnementaux

- Nombre d'accidents et gravité des blessures par rapport à l'état de la surface :

surf	nb_accidents	nb_morts	blesses_hosp	blesses_legers
1	1406	62	353	398
2	119	7	28	38
9	12	2	0	8
3	4	0	1	2
5	3	0	0	1
8	2	0	0	2
7	1	0	0	1

Code	**Signification**	**Code**	**Signification**
5	Enneigée	-1	Non renseigné
6	Boue	1	Normale
7	Verglacée	2	Mouillée
8	Corps gras – huile	3	Flaques
9	Autre	4	Inondée

Table 2.9: Codes et significations pour la variable `surf`, répartis en deux colonnes.

Analyse des résultats

Dans l'Hérault, le nombre de jours avec une météo favorable (route sèche) est élevé, ce qui explique la prédominance des accidents sur surface normale.

$$\begin{aligned}\text{Taux de tués sur surface mouillée} &= \frac{7}{119} \approx 5{,}9\% \\ \text{Taux de tués sur surface normale} &= \frac{62}{1406} \approx 4{,}4\% \\ \text{Taux de tués sur surface "Autre"} &= \frac{2}{12} \approx 16{,}7\%\end{aligned}$$

Les surfaces inhabituelles (neige, verglas, flaques, huile, etc.) sont moins fréquentes mais associées à une plus grande gravité des accidents.

- Impact de la luminosité sur le nombre d'accidents et influence des éclairages publics sur la gravité des blessures des usagers

a) Nombre d'accidents par niveau de luminosité :

condition_eclairage	nb_accidents
Plein jour	441
Nuit sans éclairage public	94
Nuit avec éclairage public allumé	71
Crépuscule ou aube	42
Nuit avec éclairage public non allumé	11

D'après les résultats, le nombre d'accidents est plus élevé en plein jour. Cependant, on ne peut pas tirer de conclusion intéressante avec ce seul résultat : la circulation est bien plus active le jour que la nuit, ce qui explique le nombre beaucoup plus élevé d'accidents le jour.

Mais on peut tout de même voir que la nuit, l'absence d'éclairages publics semble être un facteur augmentant le nombre d'accidents.

On peut donc se poser la question :

b) La présence d'éclairages publics la nuit a-t-elle une influence sur la gravité des blessures des usagers ?

condition_eclairage	gravite	nb_usagers
Éclairage absent ou non allumé	Blessé hospitalisé	61
Éclairage absent ou non allumé	Blessé léger	77
Éclairage absent ou non allumé	Indemne	102
Éclairage absent ou non allumé	Non renseigné	1
Éclairage absent ou non allumé	Tué	23
Éclairage allumé	Blessé hospitalisé	32
Éclairage allumé	Blessé léger	56
Éclairage allumé	Indemne	60
Éclairage allumé	Tué	4

Analyse des résultats

- **Sans éclairage** : environ $\frac{105}{176} \times 100 \approx 60\%$ des accidents
- **Avec éclairage** : environ $\frac{71}{176} \times 100 \approx 40\%$ des accidents

Condition d'éclairage	**Tués**	**Nombre d'accidents**	**Taux de tués par accident**
Sans éclairage	23	105	$\frac{23}{105} \approx 21{,}9\%$
Avec éclairage public allumé	4	71	$\frac{4}{71} \approx 5{,}6\%$

Table 2.12: Répartition des tués selon la condition d'éclairage

On peut donc dire qu'ici, les accidents de nuit **sans éclairage** sont **près de 4 fois plus mortels** (en proportion) que ceux de nuit **avec éclairage** !

Facteurs liés à la route

- Nombre d'accidents, nombre d'usagers impliqués et nombre de morts parmi ces usagers par catégorie de véhicule :

catv	nb_accidents	total_usagers	nb_morts
7	528	977	36
33	117	140	12
10	65	93	5
31	28	29	4
32	38	49	3
2	39	47	2
30	19	22	2
13	4	6	2
1	46	49	2
50	23	26	1
35	3	6	1
80	7	8	1

Analyse des résultats

Les voitures légères sont impliquées dans le plus grand nombre d'accidents, ce qui reflète leur forte présence sur les routes. En revanche, les deux-roues motorisés, notamment les grosses cylindrées, présentent un risque de mortalité nettement plus élevé. Certains véhicules comme les utilitaires légers, vélos électriques ou quads sont aussi associés à des accidents graves, malgré une fréquence plus faible.

- Répartition du nombre d'accidents selon le type d'infrastructure, en agglomération et hors agglomération :

infra	agg	nb_accidents
Aucun aménagement	2	298
Aucun aménagement	1	257
Carrefour aménagé	2	27
Carrefour aménagé	1	14
Autre	1	12
Autre	2	10
Pont / autopont	1	8

Pont / autopont	2	7
Bretelle d'échangeur	1	7
Chantier	1	7
Zone piétonne	2	4
Chantier	2	4
Bretelle d'échangeur	2	2
Voie ferrée	2	2

Analyse des résultats

La majorité des accidents surviennent sur des routes sans aménagement spécifique, que ce soit en agglomération ou hors agglomération. Les carrefours aménagés arrivent loin derrière. Les autres types d'infrastructures (ponts, chantiers, zones piétonnes...) sont très peu représentés. De manière générale, les accidents sont plus fréquents en agglomération.

Chapitre 3

Matériel et Méthodes

3.1 Logiciels

Toutes les analyses statistiques ont été réalisées à l'aide du logiciel R, via l'environnement RStudio. Les packages principaux utilisés sont `DBI`, `ggplot2`. Pour la gestion de projet, nous avons utilisé l'outil collaboratif Notion afin de partager les scripts R.

3.2 Modélisation statistique

Dans cette étude, nous avons utilisé différents outils statistiques pour analyser les relations entre variables.

- **Analyse descriptive préliminaire** :
  Graphiques univariés (histogrammes, diagrammes en barres) pour observer la répartition des variables.

  Avantages et limites

  - Avantage : donne une première compréhension intuitive des données.
  - Limite : descriptif uniquement, sans test d'hypothèse.

- **Graphiques croisés** :
  Graphiques bivariés (diagrammes en barres) pour visualiser les interactions entre variables catégorielles.

  Avantages et limites

  - Avantage : facilite l'interprétation visuelle d'une éventuelle association.
  - Limite : ne remplace pas un test statistique formel.

- **Test du χ^2 d'indépendance** :
  Utilisé pour évaluer l'existence d'une association entre deux variables qualitatives.

  Hypothèses et présupposés

  Pour utiliser le test du χ^2 d'indépendance, plusieurs hypothèses doivent être respectées :

- Les variables analysées doivent être **catégorielles**.
- Les effectifs théoriques dans le tableau croisé doivent être suffisamment élevés (généralement au moins 5 dans chaque cellule).
- Les observations doivent être indépendantes.

Avantages et limites

- Avantages :
  - Facile à utiliser pour étudier l'association entre deux variables qualitatives.
  - Résultats facilement interprétables via la p-value.
- Limites :
  - Nécessite des effectifs suffisants dans chaque modalité.
  - Ne donne pas d'information sur la force ou la direction de la relation.
  - Sensible aux tailles d'échantillon très grandes ou très petites.

Équations mathématiques associées
Le test du χ^2 repose sur la statistique suivante :

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

où :

- O_{ij} = effectif observé pour la cellule i, j,
- E_{ij} = effectif théorique attendu pour la cellule i, j.

L'hypothèse nulle H_0 est : **"Les deux variables sont indépendantes."**

Chapitre 4

Analyse exploratoire des données

4.1 Introduction

Avant de procéder à des analyses plus poussées, nous avons réalisé une **analyse exploratoire** des données.
L'objectif est de mieux comprendre la structure des variables, détecter d'éventuelles anomalies, et orienter les analyses statistiques futures.

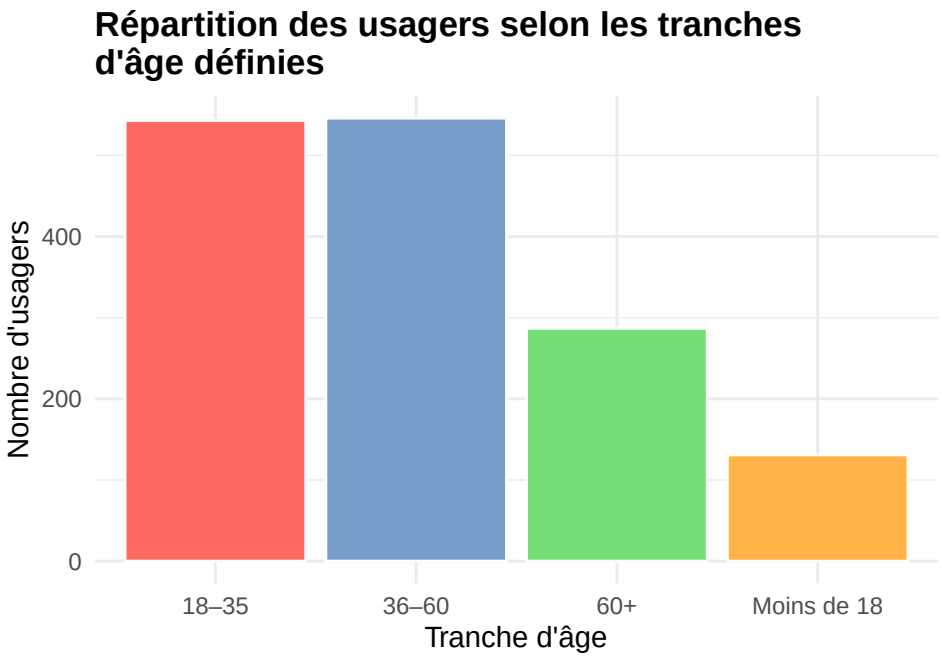
Nous avons effectué des analyses univariées et bivariées sous forme de diagramme ou de tableau. Chacun est accompagné d'un commentaire succinct.

4.2 Graphiques et Tableaux

Facteurs humains

Répartition des usagers par tranche d'âge

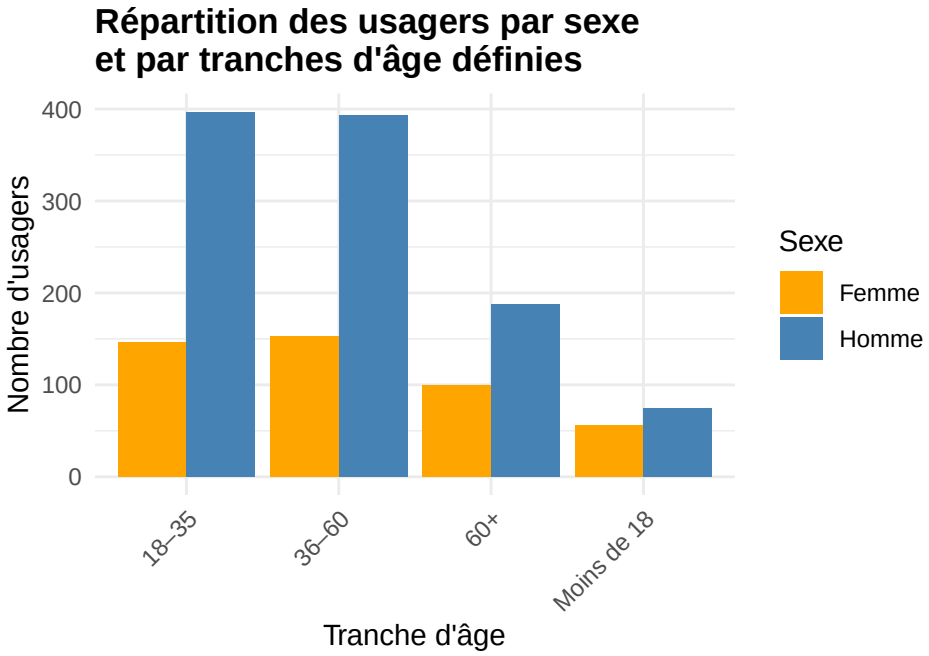
- Diagramme en barres - univarié



Commentaire : La majorité des usagers impliqués ont entre 18 et 60 ans, avec une concentration particulièrement marquée dans la tranche 18–35 ans. Les tranches $<$18 ans et 60+ sont présentes, mais en moindre proportion.

Répartition des usagers selon le sexe et la tranche d'âge

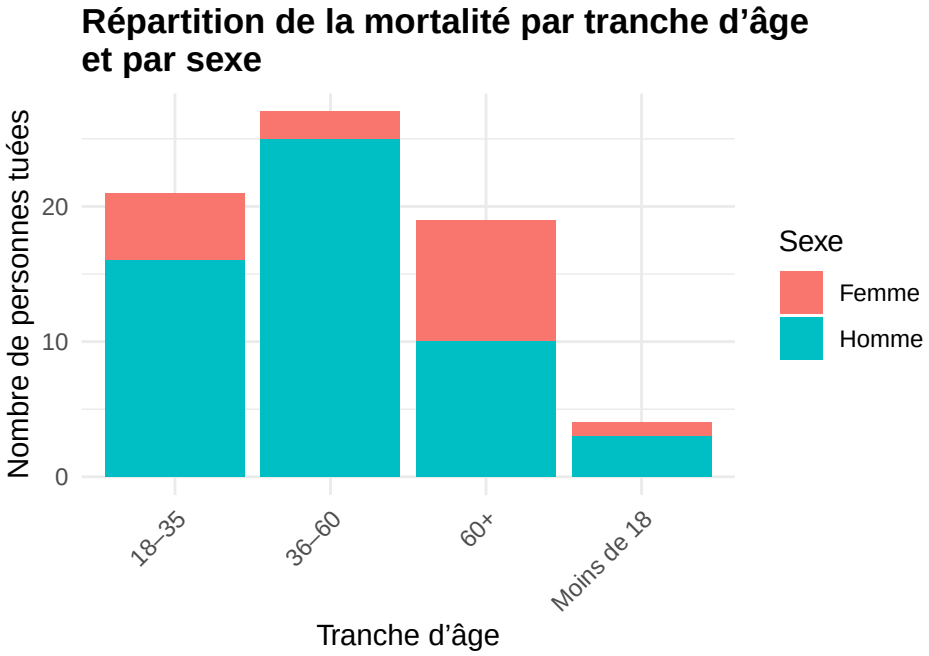
- Diagramme en barres groupées - bivarié



Commentaire : Les hommes sont largement majoritaires dans toutes les tranches d'âge, avec une surreprésentation marquée entre 18 et 60 ans. La répartition reste inégale dans les tranches les plus jeunes et les plus âgées.

Répartition de la mortalité par tranche d'âge et sexe

- Diagramme en barres empilées - bivarié

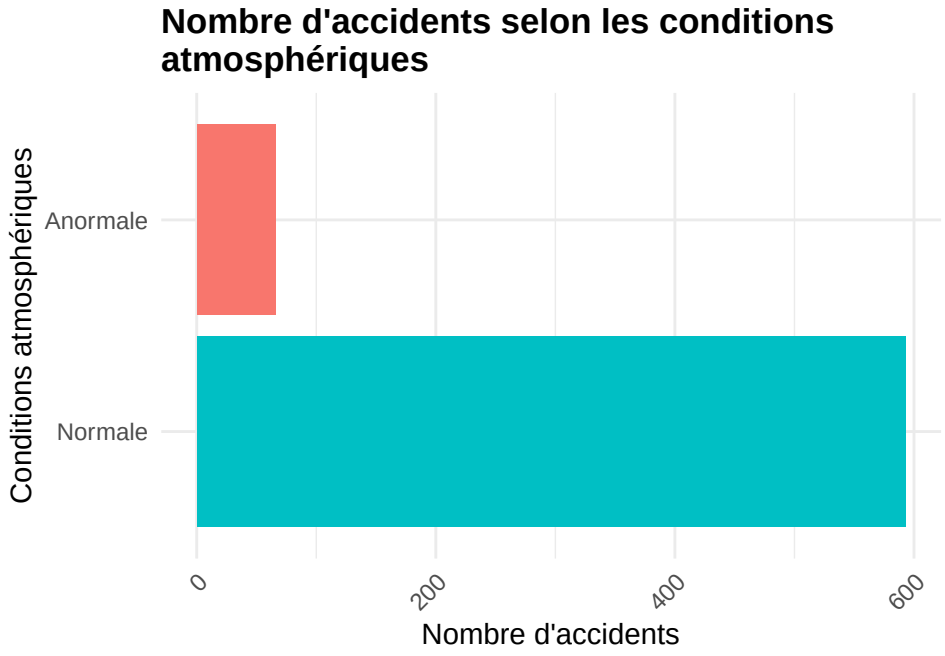


Commentaire : La mortalité est la plus élevée chez les usagers âgés de 36 à 60 ans, suivis des 18–35 ans et des 60 ans et plus. Les hommes restent majoritaires parmi les personnes tuées, quelle que soit la tranche d'âge.

Facteurs environnementaux

Nombre d'accidents selon les conditions atmosphériques

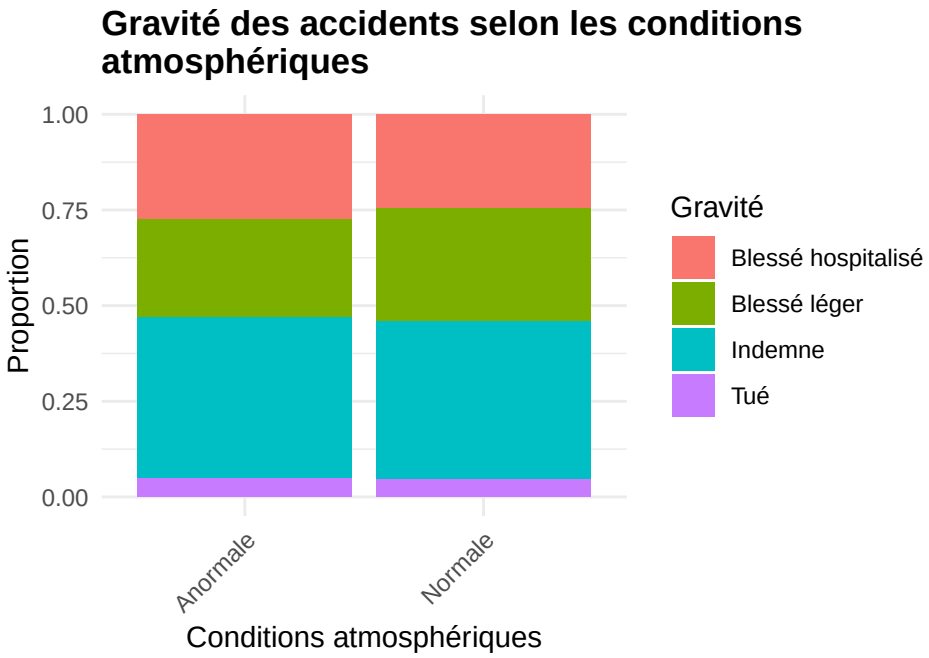
- Diagramme en barres - univarié



Commentaire : La modalité "Anormale" a été créée par un regroupement de 8 modalités de départ (neige, pluie, brouillard, etc.), car elles présentaient des effectifs trop faibles por une analyse statistique fiable.

Gravité des accidents selon les conditions atmosphériques

- Diagramme en barres empilées - bivarié

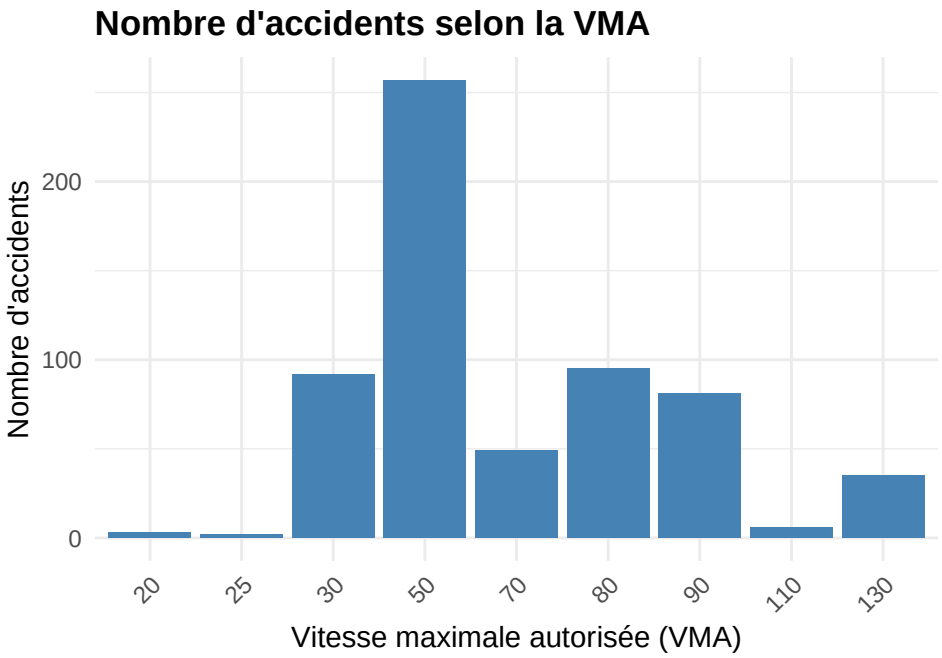


Commentaire : La répartition des niveaux de gravité ne varie que très peu selon les conditions atmosphériques. On observe seulement une petite différence de proportion dans les cas des blessés hospitalisés (plus importants en "Anormale"), et dans les cas des blessés légers (plus importants dans la catégorie "Normale")

Facteurs liés à la route

Nombre d'accidents selon la vitesse maximale autorisée

- Diagramme en barres - univarié



Commentaire : La VMA est principalement de 50 km/h (en régime urbain), avec une fréquence importante. Les autres vitesses (30, 70, 80...) sont moins fréquentes.

Nombre d'usagers impliqués selon la vitesse maximale autorisée

- Boxplots et tableau - bivarié

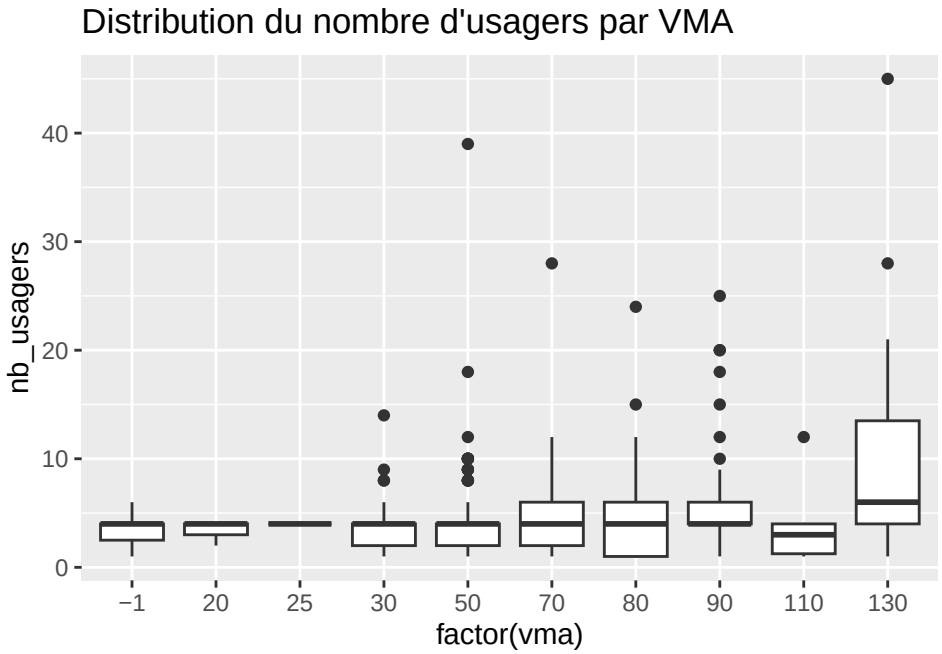


Table 4.1: Résumé du nombre d'usagers par VMA

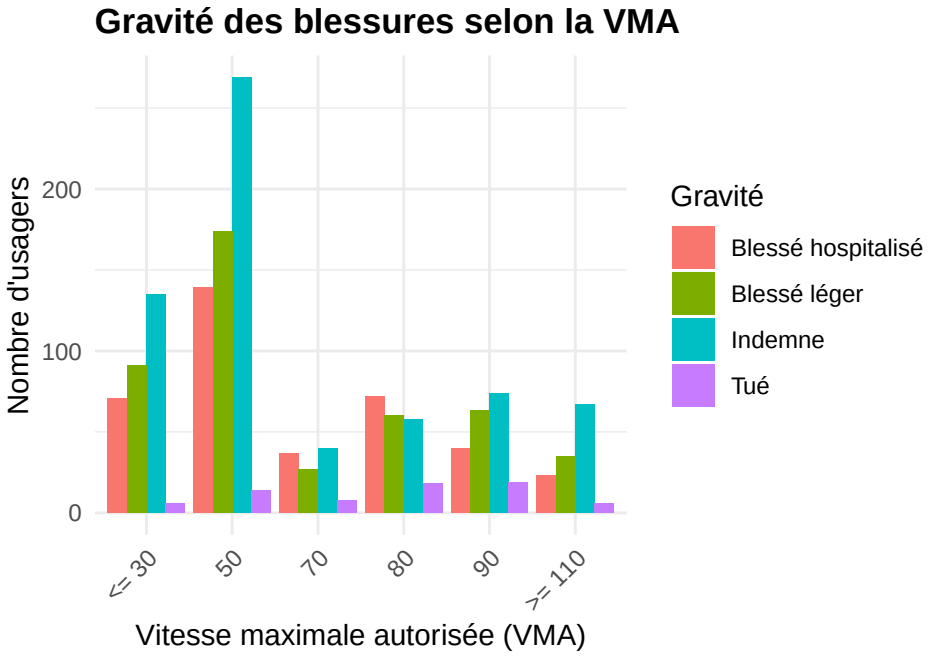
vma	nb_usager	moyenne	mediane	min	max	ecart_type
20	10	3.6	4	2	4	0.8
25	8	4.0	4	4	4	0.0
30	356	4.8	4	1	14	2.5
50	1045	6.4	4	1	39	7.0
70	226	8.5	6	1	28	7.8
80	391	7.2	6	1	24	5.5
90	437	9.2	6	1	25	6.5
110	24	7.6	8	1	12	4.6
130	310	18.0	15	1	45	13.1

Commentaire : Le graphique montre la distribution du nombre d'usagers impliqués selon la VMA, représentée sous forme de boxplot. On observe une tendance générale : plus la vitesse maximale autorisée est élevée, plus la médiane du nombre d'usagers impliqués par accident augmente. Cette évolution est confirmée par le tableau, qui présente les statistiques descriptives correspondantes. Par exemple, la médiane passe de 4 usagers à 50 km/h à 15 usagers à 130 km/h.

On observe également une forte augmentation de la dispersion pour les VMA élevées : l'écart-type atteint 13,1 à 130 km/h, contre seulement 2,5 à 30 km/h. Cela signifie que les accidents sur ces axes rapides impliquent non seulement davantage d'usagers en moyenne, mais sont aussi plus variables dans leur ampleur.

Gravité des blessures selon la vitesse maximale autorisée

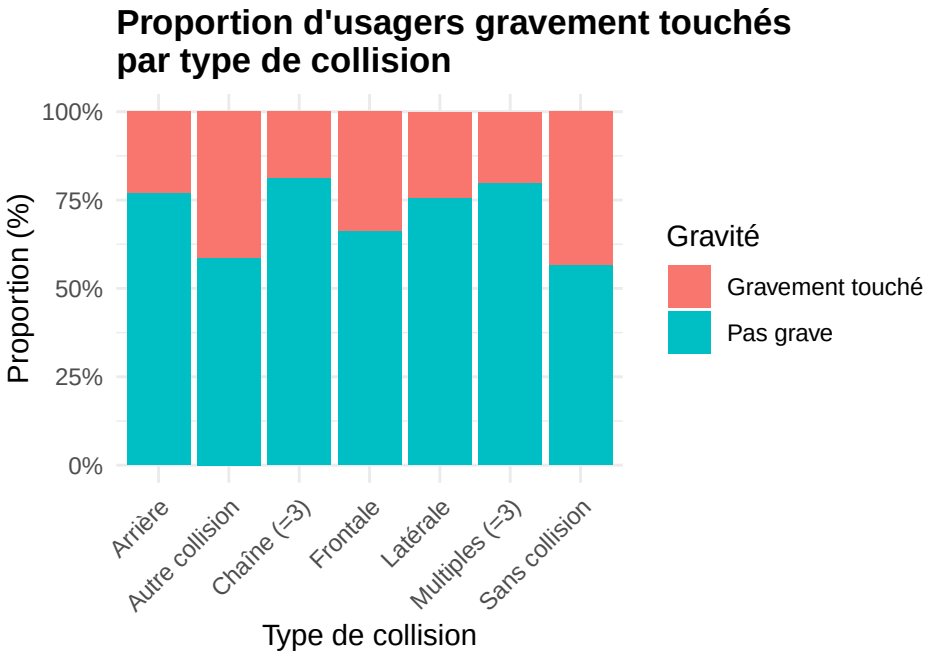
- Diagramme en barres - bivarié



Commentaire : On observe un nombre important d'usagers indemnes sur les routes limitées à 50 km/h, mais également des blessés et hospitalisés à toutes les vitesses. Les cas de décès, bien que rares, apparaissent à toutes les VMA, y compris les plus basses.

Proportion des usagers gravement touchés par type de collision

- Diagramme en barres empilées (en pourcentages) - bivarié



Commentaire : Le graphique présente la répartition des usagers selon la gravité des blessures, pour chaque type de collision. On observe des différences de proportions entre les types : par exemple, les "sans collision", "autres collisions" et "frontale" comptent une part plus importante d'usagers gravement touchés.

Chapitre 5

Analyse et Résultats

Dans cette partie, nous avons cherché à mettre en évidence d'éventuelles associations entre différentes variables. Étant donné que l'ensemble des variables étudiées sont qualitatives, nous avons utilisé uniquement des tests du χ^2 d'indépendance, qui permettent d'évaluer si deux variables catégorielles sont statistiquement liées.
Chaque analyse présente les hypothèses, les résultats du test et une interprétation.

Facteurs humains

Relation entre l'âge des usagers et le fait d'être tué lors d'un accident

```
##
##  Pearson's Chi-squared test
##
## data:  table_age_mort
## X-squared = 2.6508, df = 3, p-value = 0.4487
```

Hypothèse nulle : Le fait d'être tué est indépendant de la tranche d'âge.
Hypothèse alternative : Il existe une dépendance entre la tranche d'âge et le fait d'être tué.

Résultat du test :

- Valeur du χ^2 = 2,6508
- Degrés de liberté = 3
- p-value = 0,4487

Interprétation :

- La p-value est supérieure à 0.05, donc nous **ne rejetons pas l'hypothèse d'indépendance**.
- Il n'existe **pas de lien statistiquement significatif** entre la tranche d'âge des usagers et la mortalité dans notre jeu de données.
- L'âge ne semble donc **pas influencer significativement** la probabilité d'être tué lors d'un accident, du moins dans les effectifs observés.

Relation entre le sexe des usagers et le fait d'être tué lors d'un accident

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_sexe_mort
## X-squared = 1.0719, df = 1, p-value = 0.3005
```

Hypothèse nulle : Le fait d'être tué est indépendant du sexe de l'usager.
Hypothèse alternative : Il existe une dépendance entre le sexe et le fait d'être tué.

Résultat du test

- Valeur du $\chi^2 = 1{,}0719$
- Degrés de liberté = 1

- p-value = 0,3005

Interprétation :

- La p-value est supérieure à 0.05, donc nous **ne rejetons pas l'hypothèse d'indépendance**.
- Il n'existe **pas de lien statistiquement significatif** entre le sexe des usagers et la mortalité dans notre échantillon.
- Cela signifie que, dans notre base, le fait d'être un homme ou une femme **n'a pas d'influence significative** sur la probabilité d'être tué dans un accident.

Facteurs environnementaux

Relation entre les conditions météorologiques et la gravité des accidents

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_chi2_atm
## X-squared = 0.65078, df = 1, p-value = 0.4198
```

Hypothèse nulle : La gravité des accidents est indépendante des conditions météorologiques (normales ou anormales).
Hypothèse alternative : Il existe une dépendance entre les conditions météo et la gravité des accidents.

Résultat du test :

- Valeur du $\chi^2 = 0{,}651078$
- Degrés de liberté = 1

- p-value = 0,4198

Interprétation :

- La p-value est supérieure à 0.05, donc nous **ne rejetons pas l'hypothèse d'indépendance**.
- Il n'existe **pas de lien statistiquement significatif** entre les conditions météorologiques regroupées (normales vs anormales) et la gravité des accidents dans notre base de données.

Pour respecter les conditions d'application du test du χ^2, ce résultat repose sur un regroupement large des conditions météorologiques en une seule catégorie "Anormale". Cette simplification inclut des situations très différentes (neige, brouillard, etc.), souvent peu représentées dans notre base, ce qui limite la portée des conclusions spécifiques à chacune de ces conditions.

Facteurs liés à la route

Relation entre la vitesse maximale autorisée et la gravité des blessures

```
##
##  Pearson's Chi-squared test
##
## data:  table_gravite_vma
## X-squared = 14.143, df = 6, p-value = 0.02807
```

Hypothèse nulle : La gravité des blessures est indépendante de la VMA. **Hypothèse alternative** : Il existe une dépendance entre la VMA et la gravité des blessures.

Résultat du test :

- Valeur du $\chi^2 = 14.143$
- Degrés de liberté $= 6$
- p-value $= 0.02807$

Interprétation :

- La p-value est inférieure à 0.05, donc nous **rejetons l'hypothèse d'indépendance**.
- Il existe **une relation statistiquement significative** entre la vitesse limite et la gravité des blessures.
- Plus la vitesse autorisée est élevée, plus la gravité des blessures tend à augmenter.

Relation entre le type de collision et la gravité des blessures

```
##
##  Pearson's Chi-squared test
##
## data:  table_chi2_collision
## X-squared = 49.508, df = 6, p-value = 5.898e-09
```

Hypothèse nulle : Le type de collision est indépendant de la gravité des blessures.
Hypothèse alternative : Il existe une dépendance entre le type de collision et la gravité des blessures.

Résultat du test :

- Valeur du $\chi^2 = 49{,}508$
- Degrés de liberté = 6
- p-value = 5,898e-09

Interprétation :

- La p-value est très largement inférieure à 0.05, donc nous **rejetons l'hypothèse d'indépendance**.
- Il existe **une relation statistiquement significative** entre le type de collision et la gravité des blessures.
- Certains types de collision, comme les chocs frontaux ou multiples, sont associés à une proportion plus élevée de blessures graves.

Chapitre 6

Discussion

L'objectif de cette étude était d'identifier les facteurs influençant la survenue et la gravité des accidents de la route dans l'Hérault en 2023. Les analyses statistiques ont permis de clarifier certaines pistes tout en révélant les limites d'autres hypothèses.

Concernant les facteurs humains, les résultats sont nuancés. Ni le sexe ni la tranche d'âge ne montrent d'association statistiquement significative avec la probabilité d'être tué dans un accident. Il est important de noter que nos tests statistiques ici ne portaient que sur des facteurs personnels non modifiables pour l'usager, et non sur des éléments comportementaux comme le port d'un équipement de sécurité. L'absence de lien significatif peut aussi s'expliquer par des effectifs trop faibles dans certaines sous-catégories ou par une répartition trop dispersée des cas graves. Pourtant, les analyses descriptives avaient mis en évidence une surreprésentation des hommes parmi les victimes décédées, ainsi qu'une plus grande vulnérabilité chez les usagers âgés. Ce décalage souligne l'intérêt de croiser les approches descriptives et inférentielles : un test du χ^2 ou un graphique ne peuvent pas suffire à eux seuls à capturer la complexité de ces phénomènes.

Pour les facteurs environnementaux, aucun lien clair n'a été identifié entre la gravité des blessures et les conditions météorologiques. Toutefois, la différence d'effectif entre les deux catégories de conditions météorologiques, et le regroupement de certaines situations peu fréquentes dans l'Hérault (neige, brouillard, etc.) sous une catégorie large "Anormale" ont très sûrement masqué des effets réels.

En revanche, les facteurs liés à la route se révèlent plus déterminants. La vitesse maximale autorisée est significativement liée à la gravité des blessures : les accidents sur les routes à 110 ou 130 km/h présentent une part plus importante de blessés graves ou de décès, tandis que les zones limitées à 30 ou 50 km/h concentrent des blessures plus légères. Ce constat renforce l'intérêt des politiques de limitation de vitesse dans les zones sensibles. Le type de collision est également fortement associé à la gravité : les collisions frontales ou multiples entraînent proportionnellement plus de blessures graves. Cela confirme l'influence directe de la violence de l'impact et justifie des mesures spécifiques sur les zones à risque élevé (carrefours mal sécurisés, routes sans séparateur central...).

En résumé, cette étude met en évidence que si les caractéristiques individuelles non modifiables (âge et sexe) des usagers ne permettent pas à elles seules de prédire la gravité d'un accident, les facteurs structurels liés à la route et au type de choc en sont de puissants déterminants. Ces résultats orientent les recommandations vers une combinaison de mesures : aménagements d'infrastructure, régulation de la vitesse, et campagnes de prévention ciblées.

Chapitre 7

Conclusion et perspectives

Conclusions principales

Notre analyse des accidents de la route survenus dans l'Hérault en 2023 met en évidence plusieurs facteurs significativement associés à la gravité des accidents. Les résultats montrent notamment que :

- Les jeunes adultes, particulièrement les hommes âgés de 36 à 60 ans, sont visuellement surreprésentés parmi les usagers blessés gravement ou tués. Cette tranche d'âge semble plus exposée aux comportements à risque, ou moins protégée par les dispositifs de sécurité.Cependant, nous n'avons pas pu trouver de lien statistiquement significatif pour appuyer ce propos.
- Le port de la ceinture de sécurité est un facteur de réduction majeure de la gravité des blessures. Les usagers non équipés présentent une probabilité nettement plus élevée d'hospitalisation ou de décès.
- Les conditions environnementales telles que la faible luminosité (nuit sans éclairage), et l'état de la surface de la route (surface mouillée, verglacée) aggravent les conséquences des accidents.
- Les accidents graves sont plus fréquents sur les routes où la vitesse maximale autorisée est élevée (routes départementales ou nationales), en particulier en dehors des agglomérations. Moins d'accidents ont lieu sur ces routes-ci qu'en agglomération, mais leur gravité est beaucoup plus élevée.

Recommandations pour le commanditaire

À la lumière de ces constats, nous proposons les recommandations suivantes :

- Intensifier les campagnes de prévention ciblées vers les jeunes conducteurs, avec un accent particulier sur le port du casque à deux-roues, surtout à vélo, et de la ceinture à bord des véhicules, qu'on soit passager ou conducteur.
- Mieux signaler les zones à fort risque d'accident, notamment en périphérie ou sur les routes limitées à plus de 70 km/h, en renforçant la visibilité ou la signalisation.
- Renforcer les contrôles de sécurité routière en soirée et la nuit, moments où la gravité des accidents est accrue. Ajouter davantage d'éclairage public pourrait également contribuer à améliorer la sécurité dans ces créneaux horaires.

Perspectives à court terme (amélioration de l'analyse)

- Nettoyer plus finement les données manquantes ou peu fiables, par exemple en excluant les enregistrements où les informations essentielles ne sont pas renseignées, afin d'éviter les biais dans les résultats.

- Intégrer plus de modèles, ainsi que des plus robustes (régression logistique, arbres de décision) pour mieux modéliser les relations entre les variables et la gravité des accidents.
- Mettre en place une procédure automatique de suppression des doublons pour éviter les biais d'interprétation.
- Tester d'autres croisements de variables pertinents et plus ciblés sur une population.

Perspectives à long terme (domaine métier et science des données)

- Croiser les données d'accidents avec d'autres bases externes, comme les données hospitalières, de géolocalisation, ou encore celles liées à l'alcoolémie et à la consommation de stupéfiants, afin d'enrichir l'analyse de contexte.
- Mettre en oeuvre des modèles prédictifs de gravité d'accident en fonction des caractéristiques en temps réel (heure, météo, trafic, luminosité, etc.) pour fournir des informations utiles aux acteurs publics pour agir concrètement.
- Proposer une application, une plateforme permettant d'identifier les zones à fort risque et d'ajuster les politiques de sécurité.
- Étendre l'analyse à plusieurs années (analyse temporelle) pour observer des tendances et mesurer l'effet d'éventuelles campagnes de prévention.

Difficultés rencontrées

- De nombreuses données non renseignées ont complexifié certaines de nos analyses. – La codification des variables qualitatives sous forme de nombres entiers rend la lecture moins intuitive et a nécessité un travail préalable de documentation pour bien les interpréter.
- Les relations "plusieurs usagers et plusieurs véhicules pour un même accident" ont complexifié les jointures entre tables : un accident pouvait apparaître plusieurs fois si les regroupements n'étaient pas bien maîtrisés. Il a donc fallu vérifier rigoureusement les agrégations pour éviter des erreurs de comptage.
- La répartition très inégale de certaines observations a limité la puissance de certaines analyses.
- Le manque d'information contextuelle (ex. : niveau d'alcoolémie, état de fatigue) restreint la portée explicative de certaines conclusions.

Annexes

Codes

- Distribution du nombre d'accidents selon le motif de déplacement du véhicule associé à la gravité des blessures entrainées :

trajet	grav	nb_accident
Non renseigné	NA	504
Promenade - loisirs	1	225
Promenade - loisirs	4	186
Promenade - loisirs	3	183
Utilisation professionnelle	1	86
Domicile - travail	1	68
Domicile - travail	4	44
Domicile - travail	3	35
Autre	1	28
Promenade - loisirs	2	28
Autre	4	23
Autre	3	22
Courses - achats	1	19
Utilisation professionnelle	4	17
Domicile - école	1	14
Domicile - travail	2	13
Utilisation professionnelle	3	13
Domicile - école	4	11
Courses - achats	4	8
Courses - achats	3	7
Domicile - école	3	5
Utilisation professionnelle	2	3
Autre	2	2
Courses - achats	2	2
Domicile - école	2	1

Analyse des résultats

La distribution du nombre d'accidents en fonction du motif de déplacement et de la gravité des blessures présente une forte proportion de valeurs manquantes, avec

504 cas sur un total de 1119 non renseignés (soit environ 45 % des observations). Ce taux limite fortement l'interprétation globale.

Parmi les données disponibles, les motifs les plus fréquents sont :

- Promenade / loisirs (622 cas), avec une majorité de blessés légers ou indemnes.
- Domicile – travail (160 cas) et utilisation professionnelle (123 cas) apparaissent également comme des trajets à risques modérés.

Les accidents graves (tués) sont rares dans toutes les catégories (entre 1 et 3 cas par motif), ce qui empêche toute comparaison statistique solide.

Conclusion : en raison du taux élevé de données manquantes et de la faible fréquence des cas graves par catégorie, aucun lien clair ne peut être établi entre le motif de déplacement et la gravité des accidents.

- Répartition de la gravité de l'accident selon la catégorie de la route :

catr	gravite	nb_accidents
1	1	83
1	2	6
1	3	33
1	4	41
2	1	15
2	3	7
2	4	10
3	1	212
3	2	46
3	3	164
3	4	177
4	-1	1
4	1	273
4	2	12
4	3	139
4	4	181
5	3	1
5	4	1
6	1	7
6	3	4
6	4	1
7	1	50
7	2	7
7	3	34
7	4	34
9	1	3

9	4	5

Analyse des résultats

Cette analyse croise la gravité des blessures avec la catégorie de route. Bien que certaines tendances semblent émerger (ex. : davantage de tués sur routes départementales), les effectifs sont trop hétérogènes selon les types de routes pour tirer des conclusions solides. Certaines catégories comme les routes nationales ou les voies hors réseau public sont peu représentées, ce qui rend les comparaisons peu fiables. Aucune donnée sur le trafic ou l'exposition (ex. : nombre de véhicules circulant par jour) n'est disponible, ce qui empêche de calculer des taux standardisés.