



## MODELISATION STATISTIQUE

Comprendre le lien statistique entre Zones de Sécurité Prioritaire, effectifs des forces de l'ordre et nombre de plaintes déposées en France métropolitaine en 2021.

Le 1 septembre 2023,

A l'attention de Madame Marie DE ANTONIO,  
biostatisticienne au CHU de Clermont-Ferrand,  
et de Monsieur Vincent AUDIGIER, chercheur  
et professeur associé au Conservatoire National  
des Arts & Métiers.

**Guillaume CHEVRON**  
[chevron.guillaume@gmail.com](mailto:chevron.guillaume@gmail.com)  
+33 6 76 77 34 26

le cnam

## Table des matières

I. Introduction à l'étude statistiques du nombre de crimes et délits en France métropolitaine.....	2
A. Comprendre les enjeux de sécurité sur le territoire français .....	2
B. Décrypter la signification des infractions enregistrées par les Services .....	2
C. Appréhender l'évolution des infractions dans le temps .....	2
D. Identifier les contraintes de l'analyse statistique des infractions .....	3
E. Combiner analyse descriptive exploratoire et méthodes statistiques de classification et de régression .....	3
II. Analyse approfondie des données .....	4
A. Comprendre les données sources issues de l'État 4001.....	4
1. Portée de l'étude.....	4
2. Nature des infractions enregistrées.....	4
3. Localisation des infractions.....	4
4. Décalage temporel.....	4
5. Signification des données par catégories.....	4
6. Évolution des méthodologies d'enregistrement.....	4
7. Changements dans la liste des Services.....	5
8. Ruptures statistiques et corrections techniques.....	5
9. Interprétation des données des services de Police et de Gendarmerie.....	5
B. Intégrer des données sociodémographiques complémentaires à l'étude.....	5
1. Population par département en 2023 .....	5
2. Nombre de forces de l'ordre par Département en 2023 .....	5
3. Nombre de Zones de Sécurité Prioritaire par département en 2023.....	5
C. Préparer et structurer les données recueillies.....	5
1. Agrégation des données des établissements de Police et de Gendarmerie par département.....	5
2. Transformation et transposition des données départementales.....	5
3. Intégration de données démographiques et sécuritaires complémentaires.....	5
4. Pondération des données et relations linéaires entre les variables clés .....	6
5. Visualisation géographique des données .....	7
III. Analyse exploratoire approfondie des tendances et des motifs émergents.....	8
A. Analyse de la relation entre infractions et population.....	8
1. Tendance linéaire et points anomalies.....	8
2. Interprétation statistique et impact de la proximité .....	8
B. Analyse des effectifs de forces de l'ordre et corrélation avec les infractions .....	9
1. Relations complexes entre forces de l'ordre et infractions.....	9
2. Charge de travail des forces de l'ordre par région .....	9
C. Variabilité régionale et impact statistique .....	10
1. Zoom sur la distribution statistique régionale.....	10
2. Variations moyennes régionales et facteurs démographiques .....	10
3. Linéarité entre infractions et populations à l'échelle régionale .....	10
D. Impact géographique et analyse spatiale.....	10
1. Visualisation cartographique .....	11
2. Interprétation statistique et autocorrélation spatiale .....	11
3. Test de Durbin-Watson et validation des hypothèses .....	11
E. Analyse de la catégorisation des infractions .....	11
1. Enjeux, objectifs et limites de l'analyse .....	11
2. Analyse de la catégorisation des infractions .....	12
F. Le besoin de modéliser pour mieux prévenir .....	13
IV. Régression logistique pénalisée.....	14
A. Vers un modèle de classification robuste pour prédire les ZSP .....	14
1. Modéliser la présence de ZSP grâce à une méthode de classification.....	14
2. Explorer et comparer les méthodes de classification .....	14
3. Comprendre les forces d'un modèle de régression logistique .....	14
4. Identifier les contraintes mathématiques à prendre en compte .....	15
B. La mise en place d'une régression logistique .....	15
1. Réduction de la dimensionnalité et préparation des données.....	15
2. Scission en échantillons d'entraînement et échantillons de test .....	15
3. Configuration des paramètres du modèle .....	16
3. Paramétrage et exécution du modèle .....	17
4. Evaluation de la performance globale du modèle .....	18
V. Régression linéaire multiple pas-à-pas.....	21
A. Gestion et arbitrage des valeurs aberrantes .....	21
B. Choix du modèle de régression et contraintes à prendre en compte .....	21
C. Vérification des hypothèses .....	21
1. Hypothèse d'absence de multicolinéarité et calcul du VIF.....	21
2. Hypothèse de contribution globale des variables explicatives .....	22
3. Hypothèse d'homoscédasticité .....	22
4. Hypothèse de normalité des erreurs .....	22
D. Exécution d'une régression linéaire multiple pas-à-pas .....	23
1. Principes de la méthode pas-à-pas .....	23
2. Minimisation du RMSE .....	23
3. Sélection des variables retenues .....	23
4. Détermination et interprétation de la fonction de régression .....	24
VI. Conclusion de l'étude .....	24

## I. Introduction à l'étude statistiques du nombre de crimes et délits en France métropolitaine

### A. Comprendre les enjeux de sécurité sur le territoire français

Dans un contexte mondial en constante évolution, la France est confrontée à une série d'enjeux de sécurité cruciaux. Ces enjeux sont exacerbés par un calendrier exceptionnellement chargé en événements internationaux majeurs, dont la Coupe du Monde de Rugby en 2023 et les Jeux Olympiques de 2024. Ces compétitions sportives de renommée mondiale attirent des milliers de participants, de supporters et de touristes dans le pays, créant ainsi des défis complexes en matière de sécurité et suscitant des préoccupations quant à la sécurité des participants et du public. Les autorités françaises doivent être en mesure de garantir un environnement sûr pour tous les visiteurs et participants, en anticipant et en prévenant tout risque potentiel, qu'il s'agisse de menaces terroristes, de troubles civils ou d'autres incidents imprévus.

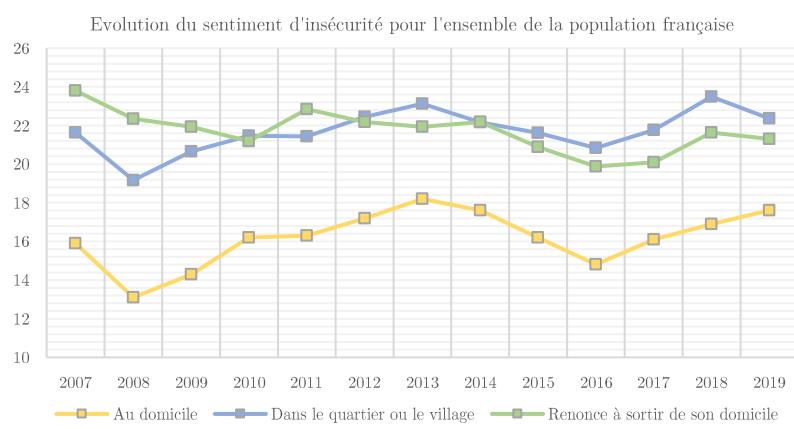
Par ailleurs, la France doit faire face à des enjeux de sécurité liés à la gestion des foules. Les rassemblements massifs lors de ces événements peuvent entraîner des défis en matière de contrôle de la circulation, de gestion des infrastructures et de maintien de l'ordre public. La coordination efficace des forces de l'ordre et des services de sécurité est essentielle pour éviter tout incident majeur.

Enfin, la France est confrontée à des défis persistants en matière de criminalité et de sécurité urbaine. Certaines régions du pays connaissent des taux de criminalité plus élevés que d'autres, ce qui nécessite une allocation stratégique des ressources policières pour prévenir et réagir aux infractions.

Il est important de noter que cette analyse ne prétend pas résoudre le problème complexe du sentiment d'insécurité, ce dernier étant influencé par de nombreux facteurs, y compris la perception individuelle et subjective de la sécurité, qui ne sont pas pris en compte dans cette étude.

Cette problématique est adressée par l'enquête CVS (Cadre de Vie et Sécurité) menée chaque année et qui permet de mesurer l'incidence de certaines catégories d'infractions (atteintes aux biens et aux personnes) et la fréquence des plaintes déposées auprès des services de police et de gendarmerie pour chaque type d'atteinte.

Il est d'ailleurs intéressant de noter qu'au cours des années 2007-2019, ce même sentiment d'insécurité en France a montré une relative stabilité, malgré un paysage médiatique parfois propice à susciter des inquiétudes.



Dans ce contexte, notre analyse s'efforce de fournir des informations statistiques aux autorités en charge de la sécurité et fera l'économie de l'étude de la gravité des crimes et dits enregistrées, que nous appellerons désormais « infractions », du fait de leur complexité juridique pour se concentrer sur la compréhension de leur fréquence. Elle vise ainsi à contribuer modestement à une planification stratégique à long terme pour l'allocation efficace des ressources humaines dans chaque département français métropolitain. En d'autres termes, son objectif est d'apporter un regard complémentaire aux autorités afin qu'elles puissent prendre des décisions éclairées et relever ces défis sécuritaires de manière proactive, en se basant sur une compréhension statistique des liens entre la présence de Zones de Sécurité Prioritaires (ZSP), la catégorisation et le nombre d'infractions commises ainsi que les ressources policières disponibles.

### B. Décrypter la signification des infractions enregistrées par les Services

L'analyse des infractions enregistrées par les services de police nationale et de gendarmerie nationale constitue un pilier essentiel de l'évaluation de la sécurité publique. Les infractions, qu'elles constituent des crimes graves ou des délits mineurs, fournissent un aperçu des tendances de la criminalité et de la délinquance. Cette information est cruciale pour comprendre comment les schémas criminels évoluent au fil du temps et comment les politiques de sécurité doivent être ajustées en conséquence.

Les données sur les infractions enregistrées ne sont pas simplement des statistiques froides ; elles racontent des histoires complexes sur les dynamiques sociales et géographiques. Elles nous renseignent sur les délits et crimes les plus fréquents, les tendances saisonnières, les variations régionales et les groupes de population les plus touchés. Ces informations guident les actions des forces de l'ordre, orientent les politiques de prévention et contribuent à façonner des stratégies de sécurité ciblées.

### C. Appréhender l'évolution des infractions dans le temps

L'évolution des infractions dans le temps est un indicateur clé pour évaluer l'efficacité des politiques de sécurité et les initiatives de prévention. Les données historiques permettent d'identifier les tendances à long terme et d'évaluer les impacts des interventions passées. Les fluctuations dans les taux d'infractions peuvent être liées à des facteurs variés tels que les changements économiques, les modifications de l'environnement urbain ou les évolutions sociales.

Les études et analyses menées par le passé ont montré à maintes reprises l'importance de suivre les évolutions des infractions dans le temps. Les rapports annuels sur la criminalité publiés par le ministère de l'Intérieur fournissent un aperçu des tendances nationales, régionales et départementales. Les recherches universitaires et les enquêtes gouvernementales ont contribué à dégager des schémas, à identifier des facteurs de risque et à éclairer les politiques de sécurité.

#### **D. Identifier les contraintes de l'analyse statistique des infractions**

L'analyse statistique des infractions ne vient pas sans défis. Dans notre étude, nous concentrons notre analyse au niveau départemental. Cette démarche permet d'obtenir une vue d'ensemble plus granulaire, mais elle signifie également que nous travaillons avec un petit échantillon de données. Les généralisations à l'ensemble de la population doivent être abordées avec prudence, car la diversité des départements peut influencer les résultats. Par ailleurs, une analyse au niveau communal aurait été plus précise mais aurait présenté un véritable défi de collecte des données nécessaires.

De plus, notre étude se concentre sur le nombre d'infractions enregistrées plutôt que sur leur gravité. Évaluer la gravité des infractions demande une compréhension approfondie du système juridique et une granularité encore plus fine des données. La gravité d'une infraction est multidimensionnelle, et tenter de la mesurer d'un point de vue statistique peut être complexe. Il existe des infractions mineures qui peuvent néanmoins entraîner des conséquences graves pour les victimes, tandis que certaines infractions plus graves peuvent parfois être moins fréquentes.

#### **E. Combiner analyse descriptive exploratoire et méthodes statistiques de classification et de régression**

En analysant les infractions de manière statistique, nous sommes incités à porter un regard critique sur les politiques publiques, notamment la notion de Zones de Sécurité Prioritaire (ZSP). Ces zones constituent une approche ciblée et proactive visant à améliorer la sécurité publique. Les ZSP ont été conçues pour réduire la criminalité et l'insécurité en concentrant les efforts sur des zones spécifiques. Elles visent à optimiser l'utilisation des ressources de maintien de l'ordre en ciblant les zones où la criminalité est la plus préoccupante. Crées dans le but de renforcer la présence policière et de répondre aux préoccupations locales, elles reposent sur la conviction que l'intensification des ressources dans les zones à risque peut entraîner une réduction significative de la criminalité.

Notre étude se fixera comme objectif d'une part de déterminer si l'analyse des infractions enregistrées dans les départements peut fournir des insights significatifs pour identifier la présence d'une ZSP, et d'autre part d'établir une estimation de l'effectif des forces de l'ordre nécessaires pour chaque département. Pour ce faire, nous adopterons une approche combinant une analyse descriptive exploratoire avec des méthodes statistiques de classification et de régression. Cette démarche visera donc à jeter un éclairage sur le poids des ZSP vis-à-vis les niveaux d'infractions enregistrées ainsi que leur impact sur la sécurité publique, et à offrir des informations complémentaires aux pouvoirs publics concernant l'allocation des ressources humaines pour renforcer la sécurité.

Nous reconnaissons les limites et les contraintes inhérents à cette étude, en particulier du fait de l'approche départementale adoptée. Nous sommes également conscients que le choix de mettre l'accent sur le nombre d'infractions plutôt que sur leur gravité représente un parti pris. Néanmoins, nous sommes optimistes quant à la possibilité que notre analyse offre des éléments de réponse pertinents en ce qui concerne la corrélation entre les infractions enregistrées et la présence de ZSP. En associant l'analyse statistique à une compréhension approfondie des enjeux de sécurité, nous visons à fournir des éclairages pour la prise de décisions politiques et pour renforcer la sécurité publique. Chaque avancée dans notre compréhension contribue à cette quête continue d'un environnement plus sûr et plus résilient, au bénéfice de l'ensemble des citoyens.

## **II. Analyse approfondie des données**

Pour éclairer notre démarche, nous avons entrepris une exploration approfondie des données disponibles. Cette étape cruciale nous permet de comprendre les tendances, les distributions et les relations au sein des données, ce qui fournit des bases solides pour les décisions ultérieures. Les données sources de cette étude comprennent plusieurs ensembles complémentaires.

### **A. Comprendre les données sources issues de l'État 4001**

#### 1. Portée de l'étude

L'État 4001, qui recense les infractions constatées par la police nationale et la gendarmerie nationale, représente un outil fondamental pour évaluer la criminalité en France. Instauré dans les années 1960, son rôle est de normaliser la collecte d'informations sur les infractions. Les données issues de cet état fournissent une vue détaillée sur les types d'infractions, leur fréquence et leur répartition géographique, constituant ainsi le socle même de notre étude.

Cependant, il est important de reconnaître que le périmètre de l'État 4001, limité aux crimes et délits excluant les contraventions, peut altérer la perception globale de l'étendue de la criminalité. Bien que les années correspondant aux limitations soient antérieures à l'année d'étude (2021), nous devons garder à l'esprit que l'absence des contraventions et la possible omission d'infractions à la définition juridique moins stricte pourraient influencer les conclusions, nécessitant une approche prudente dans l'interprétation des résultats.

En outre, il est important de reconnaître que l'utilisation de ces données peut engendrer d'autres limitations et biais d'analyse. En les identifiant et en les examinant attentivement, nous pourrons mieux contextualiser les résultats issus des méthodes statistiques que nous avons adoptées. Ces informations contribueront à éviter des interprétations erronées et à fournir des conclusions plus précises et nuancées. Ainsi, la validité de nos résultats sera renforcée, et ces derniers deviendront plus pertinents pour répondre aux questions cruciales que nous avons soulevées dans notre introduction.

#### 2. Nature des infractions enregistrées

Les infractions enregistrées dans l'État 4001 excluent les contraventions, se concentrant sur les crimes et délits. Cette distinction pourrait potentiellement mener à une sous-évaluation des infractions mineures traitées comme des contraventions. De plus, la condition que seules les infractions ayant une assise juridique suffisante pour être portées devant un tribunal soient enregistrées pourrait entraîner l'omission d'infractions socialement significatives, mais ne satisfaisant pas entièrement aux critères légaux.

Dans le contexte de notre étude, cette sélection des infractions pourrait altérer la perception globale de l'étendue de la criminalité, mais étant donné que les années correspondant aux limitations (2015 et 2016) sont antérieures à l'année d'étude (2021), les ajustements ne seraient pas nécessaires pour interpréter les résultats de l'année 2021.

#### 3. Localisation des infractions

Les infractions sont enregistrées par l'administration qui les constate, ce qui peut créer des distorsions géographiques. Les infractions commises dans une région peuvent être enregistrées ailleurs en raison des compétences territoriales des services, ce qui pourrait brouiller la compréhension de la distribution géographique réelle des infractions. Ainsi, lors de l'analyse des tendances géographiques, il est crucial de prendre en compte ces écarts.

Cependant, étant donné que les années correspondant aux limitations sont antérieures à l'année d'étude, il est peu probable que ces distorsions aient un impact majeur sur l'interprétation des résultats de l'année 2021.

#### 4. Décalage temporel

Le décalage entre la commission d'une infraction et son enregistrement peut complexifier l'analyse des tendances annuelles, y compris pour 2021. Les infractions commises en fin d'année ou avant la période de référence, notamment celles liées aux cambriolages de logements, aux vols violents sans armes, aux vols simples, et aux coups et blessures volontaires, peuvent être enregistrées ultérieurement, créant des distorsions chronologiques et compliquant la détection des évolutions réelles. Cependant, il est important de noter que cette limitation est davantage liée aux années 2015 et 2016.

#### 5. Signification des données par catégories

L'interprétation des catégories d'infractions exige une approche attentive, y compris pour 2021. Les hausses inhabituellement élevées pour les violences sexuelles, les mauvais traitements et l'abandon d'enfants pour l'année 2015 reflètent potentiellement de problèmes de révélation et de traitement. Ces augmentations soudaines pourraient être dues à une amélioration des mécanismes de signalement ou à une plus grande sensibilisation à ces problématiques, plutôt qu'à une augmentation réelle des incidents. Cependant, l'année 2021 n'est pas impactée par ce phénomène.

#### 6. Évolution des méthodologies d'enregistrement

L'évolution des méthodes d'enregistrement peut influencer les tendances observées dans les données. Cependant, pour l'année 2021, il est important de noter que les variations ne correspondent pas nécessairement à des changements de comportement criminel, mais peuvent aussi refléter des évolutions dans les pratiques d'enregistrement.

Il est donc crucial de prendre en compte ces éventuelles modifications des protocoles de collecte des données dans le cas d'une étude temporelle de la criminalité, par le biais de séries chronologiques par exemple, car elles pourraient influencer la fréquence apparente des infractions sans refléter nécessairement une variation réelle dans les taux de criminalité. Toutefois, seule l'année 2021 sera analysée, ce qui nous permet de faire l'économie de cette problématique.

## 7. Changements dans la liste des Services

Les variations dans la liste des services de police et de gendarmerie, ainsi que les modifications territoriales de leurs compétences, peuvent compliquer la comparaison des données d'une année à l'autre. Cependant, les ajustements structurels s'étant principalement produits autour des années 2015 et 2016, ces variations organisationnelles pourraient avoir un impact moindre sur les données de l'année 2021.

## 8. Ruptures statistiques et corrections techniques

Il est essentiel de prendre en compte les ruptures statistiques et les corrections techniques lors de l'analyse comparative entre différentes périodes. En particulier, les évolutions méthodologiques, notamment la fragilisation de certains index spécifiques en 2015, entraînant des doubles comptabilisations ou des indexations incorrectes (comme pour les index 96, 97, 99 et 100 par exemple), doivent être examinées avec prudence dans l'interprétation des variations observées. Cette précaution est particulièrement significative pour garantir des conclusions précises concernant les évolutions réelles des taux de criminalité pour l'année 2021.

## 9. Interprétation des données des services de Police et de Gendarmerie

La compréhension des différences dans les organisations territoriales et les compétences des services de police et de gendarmerie est cruciale pour interpréter correctement les variations dans les données. Il est nécessaire de contextualiser les différences observées en 2021 entre les régions en tenant compte des nuances organisationnelles pour éviter des interprétations erronées des variations observées.

# **B. Intégrer des données sociodémographiques complémentaires à l'étude**

## 1. Population par département en 2023

Les données sur la population par département nous permettent de prendre en compte les variations démographiques lors de l'analyse des taux d'infractions, ce qui est essentiel pour contextualiser les chiffres absous d'infractions enregistrées et calculer des taux standardisés. Ces taux standardisés facilitent une comparaison équitable entre les départements et les régions, en tenant compte des différences démographiques. Cette information provient des données démographiques officielles de l'année 2023.

## 2. Nombre de forces de l'ordre par Département en 2023

L'indicateur du nombre de forces de l'ordre mobilisées par la Police Nationale et la Gendarmerie Nationale pour chaque département, rapporté pour mille habitants, joue un rôle de premier plan dans l'évaluation de la capacité de maintien de l'ordre dans chaque région. Comprendre cette variable est essentiel, car elle pourrait potentiellement influencer les niveaux d'infractions enregistrées. Par conséquent, son intégration à notre analyse devient impérative afin d'éviter des conclusions erronées. Les données pertinentes sont extraites des rapports officiels concernant les effectifs des forces de l'ordre pour l'année 2023.

## 3. Nombre de Zones de Sécurité Prioritaire par département en 2023

Les Zones de Sécurité Prioritaire (ZSP) sont le cœur même de notre étude, car elles identifient les points chauds de la criminalité. Étant donné leur importance, le nombre de ZSP par département revêt un intérêt particulier pour évaluer l'efficacité des mesures prises dans ces zones sensibles. Il reflète la concentration des problèmes de sécurité et peut aider à analyser les résultats de nos recherches. Ces données proviennent des rapports officiels sur les ZSP pour l'année 2023.

# **C. Préparer et structurer les données recueillies**

La qualité d'une analyse statistique repose en grande partie sur la rigueur avec laquelle les données sont préparées et structurées en amont. Dans cette section, nous plongerons dans le processus de préparation des données qui nous a permis d'aboutir à un ensemble de données prêt à être exploré et analysé. Cette étape revêt une importance capitale, car elle assure la fiabilité et la pertinence des résultats qui découlent de nos analyses ultérieures.

## 1. Agrégation des données des établissements de Police et de Gendarmerie par département

La première étape de notre préparation des données consiste à consolider les informations provenant d'une multitude d'établissements de police et de gendarmerie à travers le pays. Pour ce faire, nous avons regroupé ces données en fonction des départements français. Cette agrégation nous a permis d'obtenir un aperçu global des types d'infractions enregistrées, une liste qui compte près d'une centaine d'infractions différentes, pour chaque département de la France métropolitaine. Afin de maintenir la cohérence de notre étude, les départements et territoires d'outre-mer ont été exclus de l'analyse, se concentrant ainsi uniquement sur la France métropolitaine. De plus, nous avons entrepris des rectifications minutieuses au niveau des libellés de certains départements en corrigeant des erreurs telles que des apostrophes mal placées ou des problèmes de casse.

## 2. Transformation et transposition des données départementales

Une étape cruciale dans notre processus de préparation des données a été la transformation de la structure même de notre jeu de données. Initialement, les départements étaient considérés comme des dimensions, des variables ou des colonnes du jeu de données. Cependant, dans le but de mieux appréhender la répartition géographique des infractions, nous avons entrepris de transposer les données. Cette étape essentielle consiste à convertir les départements en individus, c'est-à-dire en lignes. Cette transformation a été motivée par notre désir d'analyser les relations entre les départements et les types d'infractions, en privilégiant une perspective plus globale et granulaire.

## 3. Intégration de données démographiques et sécuritaires complémentaires

Afin d'enrichir notre ensemble de données, nous avons procédé à l'intégration de données démographiques et sécuritaires pertinentes. Nous avons inclus des informations sur la population départementale, ce qui nous permettra d'ajuster le nombre d'infractions enregistrées en fonction de la taille de la population.

De plus, nous avons incorporé le nombre de Zones de Sécurité Prioritaire (ZSP) par département, ainsi que le nombre d'effectifs des forces de l'ordre. Il convient de noter que le nombre d'effectifs des forces de l'ordre a été pondéré pour mille habitants, afin de prendre en compte les variations démographiques entre les départements. L'intégration de ces données a ajouté des dimensions cruciales à notre analyse en nous permettant de mieux évaluer les liens entre les infractions, les paramètres de sécurité et la démographie.

#### 4. Pondération des données et relations linéaires entre les variables clés

Après avoir agrégé et enrichi les données, une étape cruciale a été d'appliquer une pondération au nombre d'infractions en fonction de la population départementale. Cette démarche vise à normaliser les infractions enregistrées, permettant ainsi le calcul des infractions pour mille habitants. Cette mesure présente l'avantage d'offrir une perspective relative et équitable de la prévalence des infractions entre les différents départements.

Une fois les données pondérées, nous nous sommes intéressés aux éventuelles relations linéaires entre les variables clés. Ces variables englobent la population départementale, le nombre pondéré d'infractions pour mille habitants, le nombre d'agents des forces de l'ordre pour mille habitants, le nombre d'infractions traitées par un agent des forces de l'ordre pour mille habitants, ainsi que le nombre de Zones de Sécurité Prioritaire (ZSP) par département. Pour établir ces relations, nous avons utilisé une matrice des corrélations, un outil mathématique puissant permettant de quantifier les liens entre ces différentes variables.

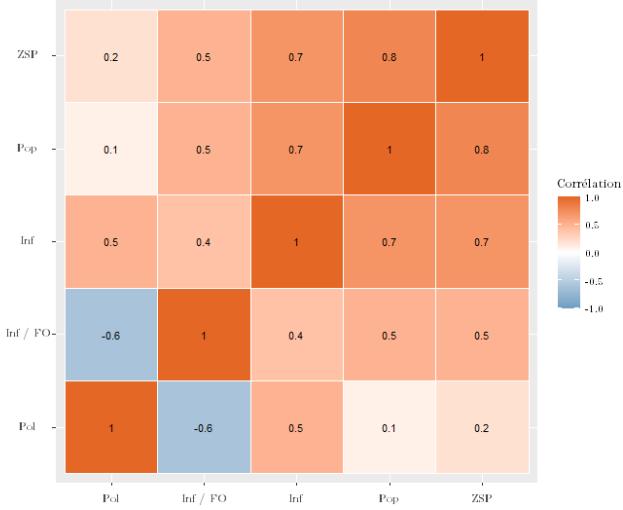
Nos observations ont révélé des résultats intrigants.

En premier lieu, une corrélation linéaire solide (0,8) a été mise en évidence entre la population départementale et le nombre de ZSP. Cette découverte suggère que les choix en matière de sécurité publique peuvent être en partie influencés par cette métrique. Par exemple, dans certaines régions, comme les départements densément peuplés de la région Île-de-France, on peut s'attendre à une plus grande concentration de ZSP.

Concernant le nombre d'agents des forces de l'ordre par département, nous avons constaté une corrélation pratiquement nulle avec la population départementale (environ 0,1). Cette observation peut sembler contre-intuitive au premier abord, notamment compte tenu des politiques publiques qui visent souvent à allouer davantage de ressources dans les régions à forte densité de population.

Exploration des relations linéaires entre les principales variables observées

Analyse de la matrice des corrélations linéaires



Cependant, d'autres facteurs, tels que la répartition géographique des infractions, la mobilité des agents et les priorités de sécurité locales, peuvent influencer cette relation complexe.

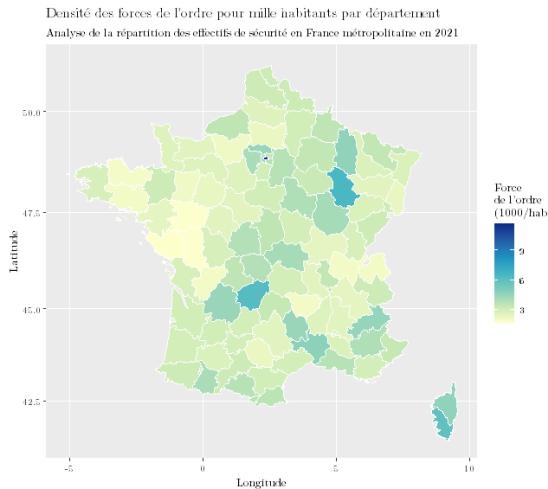
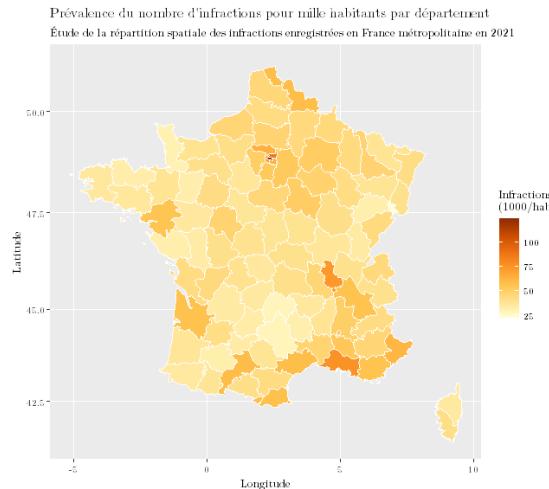
De plus, nous avons constaté une corrélation substantielle d'environ 0,7 entre le nombre pondéré d'infractions et deux autres variables clés : le nombre de ZSP et le nombre d'habitants. Plus spécifiquement, cela révèle une association significative entre le nombre d'infractions enregistrées et ces deux paramètres. Il est crucial de noter que le nombre de ZSP est également étroitement lié au nombre d'habitants, comme mentionné précédemment. Cette relation implique que les zones à forte densité de population et une concentration de ZSP présentent également une propension accrue à afficher un nombre plus élevé d'infractions enregistrées.

Il est essentiel de rappeler que la corrélation mesure la force et la direction de la relation linéaire entre deux variables. Cependant, elle ne permet pas de conclure une relation de cause à effet. Cette distinction est cruciale pour une analyse rigoureuse et souligne la nécessité d'une approche méthodologique approfondie, que nous poursuivrons en utilisant des méthodes de régression et de classification pour obtenir des éclairages plus précis et éclairés.

## 5. Visualisation géographique des données

Avant d'entamer l'analyse statistique exploratoire des données, il est opportun de commencer par une visualisation géographique. Nous avons ainsi créé plusieurs cartes de la France métropolitaine, chacune mettant en évidence une facette particulière des données collectées.

Ces cartes géographiques offrent un aperçu visuel immédiat des variations régionales en termes d'infractions enregistrées et des paramètres de sécurité. Elles nous permettent de saisir intuitivement la répartition spatiale des données, ainsi que d'identifier d'éventuelles tendances et corrélations émergentes. Cette étape préliminaire de visualisation sert de point de départ à notre analyse plus approfondie, en nous aidant à mieux appréhender la distribution géographique des phénomènes étudiés.

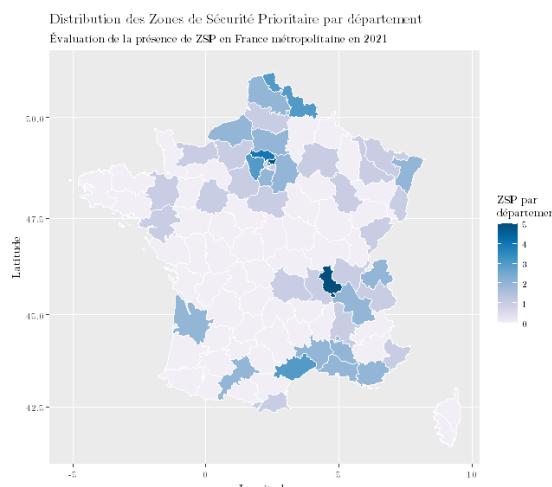
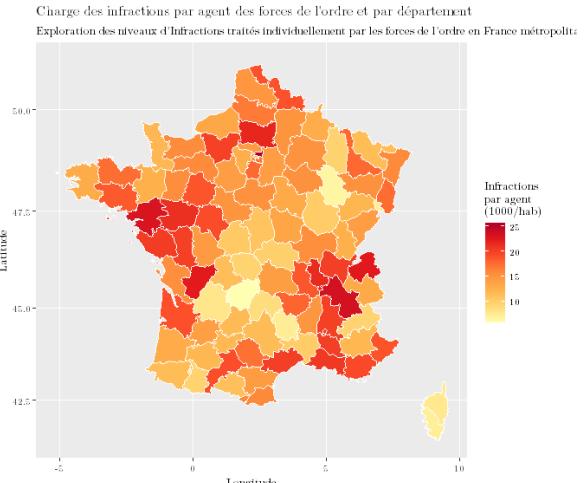


En examinant les deux cartes ci-dessus, nous pouvons observer que la distribution des régions densément peuplées ne présente pas nécessairement une corrélation directe avec le nombre d'agents des forces de l'ordre par département. Une observation remarquable est Paris, qui semble constituer une exception à cette tendance.

Pour illustrer davantage cette observation, prenons des exemples concrets tels que la Corse du Sud, la Corrèze et la Haute-Marne. Bien que ces départements présentent des densités de population allant du faible au modéré, ils exhibent des effectifs d'agents relativement élevés pour 1000 habitants.

Cependant, cette association n'est pas nettement évidente, ce qui suggère la nécessité de considérer le rapport entre ces deux variables, à savoir le nombre d'infractions par agent.

La carte de droite révèle une disparité substantielle dans la charge de travail des agents des forces de l'ordre. Des exemples significatifs incluent le département de la Loire-Atlantique, de l'Isère et de la Charente, où l'on peut observer des charges de travail plus conséquentes, dépassant les 25 infractions par agent. En revanche, la Corrèze affiche des valeurs inférieures à 10 infractions par agent.



Ces disparités mettent en évidence des dynamiques locales qui ne sont pas aisément expliquées par la densité de population seule.

Enfin, la carte de gauche montre une concentration notable dans les régions du Rhône et de Paris, où le nombre de ZSP dépasse cinq. De plus, des foyers significatifs apparaissent dans des régions telles que l'Hérault, la Seine-Saint-Denis, les Yvelines et le Nord, où le nombre de ZSP varie entre trois et quatre.

Cette distribution suggère la possibilité d'une corrélation linéaire entre la densité de population et la présence de ZSP. Cependant, elle met également en lumière que la relation entre le nombre d'agents des forces de l'ordre et la population n'est pas nécessairement linéaire, soulignant ainsi l'existence de facteurs multiples et complexes à l'œuvre dans la planification des ressources de sécurité.

### III. Analyse exploratoire approfondie des tendances et des motifs émergents

#### A. Analyse de la relation entre infractions et population

##### 1. Tendance linéaire et points anomalies

L'examen minutieux de la relation entre le nombre pondéré d'infractions (pour mille habitants) et la population départementale ouvre la voie à des découvertes fascinantes, nécessitant une exploration plus approfondie. Cette investigation vise à éclairer les liens potentiels entre le nombre d'habitants et le nombre pondéré d'infractions enregistrées, en prenant en compte l'influence des interactions sociales et de la proximité géographique sur la dynamique de la criminalité.

L'élaboration d'une représentation graphique sous forme de nuage de points, associant chaque département à son nombre pondéré d'infractions et à sa population, révèle une tendance linéaire qui se dessine entre ces deux variables. Cette observation suggère qu'en général, à mesure que la population d'un département croît, le nombre d'infractions pour mille habitants tend à augmenter.

Néanmoins, cette relation n'est pas uniforme, avec deux points qui émergent en tant qu'anomalies : les départements de Paris et du Nord. En effet, malgré sa population considérable, le département de Paris présente un nombre d'infractions bien supérieur à la moyenne projetée pour sa population. À l'opposé, le département du Nord, bien qu'affichant une population substantielle, enregistre un taux d'infractions inférieur à la moyenne attendue.

Cela suggère que des facteurs plus complexes que la simple densité de population influencent le lien entre population et criminalité. La compréhension statistique de cette observation met en évidence l'existence de multiples déterminants qui agissent en tandem pour influencer le nombre pondéré d'infractions. Alors que la proximité géographique et les interactions sociales jouent un rôle, il est essentiel de considérer d'autres facteurs tels que la prévalence des activités criminelles, les mesures de sécurité locales et les conditions socio-économiques.

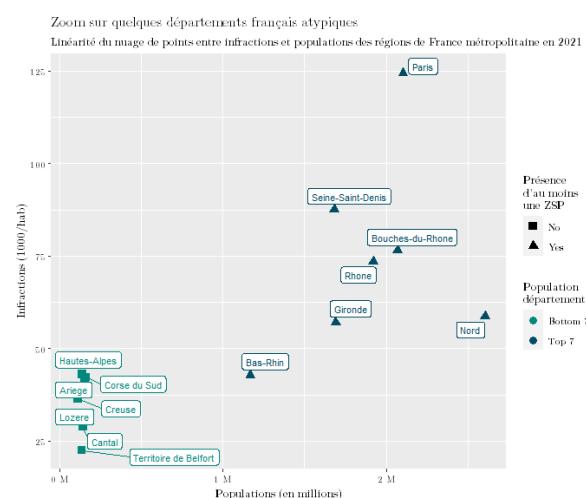
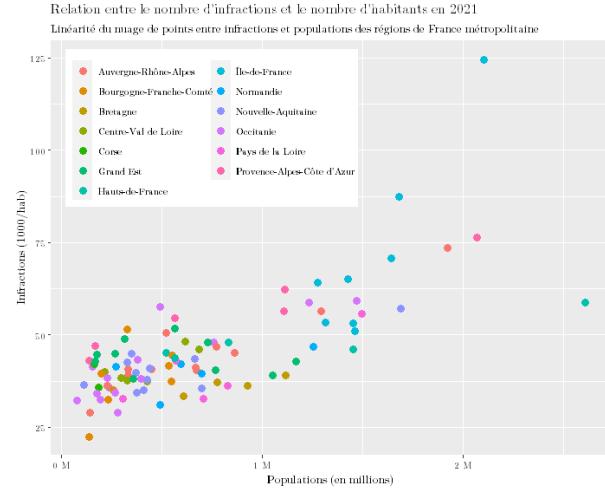
Comprendre comment ces facteurs interagissent peut s'avérer déterminant en vue de formuler des politiques publiques ciblées visant à réduire la criminalité et à améliorer la sécurité. Ces éléments peuvent expliquer pourquoi Paris, en tant que centre urbain densément peuplé, présente un nombre d'infractions élevé, tandis que le Nord, avec une population similaire, enregistre des taux plus bas.

##### 2. Interprétation statistique et impact de la proximité

Poursuivons l'étude des particularités de certains départements et comparons les sept départements plus peuplés aux sept départements les moins peuplés.

Les deux groupes sont très distinctement séparés de telle sorte que tous les départements les moins peuplés (Bottom 7) s'agencent en bas à gauche du nuage de points, tandis que les départements les plus peuplés (Top 7) se distribuent du milieu à en haut à droite du graphique. Cela met en évidence le rôle central de la population dans l'établissement des tendances de criminalité.

Toutefois, une observation intrigante émerge : bien que le Bottom 7 affiche une forte cohésion, avec des points regroupés étroitement, le Top 7 présente une dispersion bien plus élevée avec des distances euclidiennes importantes par rapport à leur centre de gravité.



## B. Analyse des effectifs de forces de l'ordre et corrélation avec les infractions

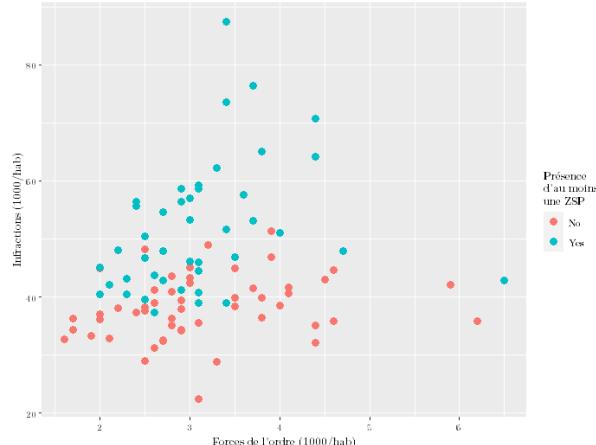
### 1. Relations complexes entre forces de l'ordre et infractions

Notre investigation s'approfondit en examinant la répartition des forces de l'ordre en France métropolitaine afin de tenter d'établir des liens avec la présence éventuelle de Zones de Sécurité Prioritaire (ZSP). Pour ce faire, nous créons un nuage de points qui relie le nombre d'infractions pour mille habitants au nombre de forces de l'ordre pour mille habitants. La présence ou l'absence d'au moins une ZSP est indiquée en légende.

Une première observation peut déjà être effectuée : Paris, se situant en haut à droite, se distingue nettement des autres départements. Cette valeur aberrante contraste avec le reste des points, qui se regroupent principalement dans le quart inférieur gauche du nuage.

Pour une analyse plus claire, nous excluons Paris de ce graphique, révélant une dispersion importante des valeurs.

Relation entre infractions et ressource policière (hors Paris) en France métropolitaine en 2021  
Analyse statistique de l'impact des forces de l'ordre et des Zones de Sécurité Prioritaire



### 2. Charge de travail des forces de l'ordre par région

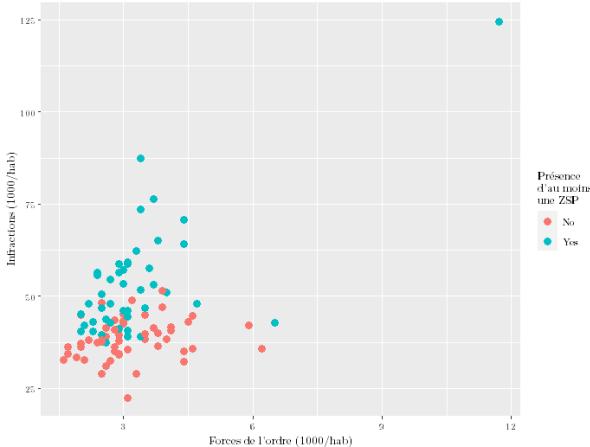
Notre analyse s'étend également au nombre d'infractions traitées par agent des forces de l'ordre. Pour ce faire, il est important de noter que l'on ne pourra pas parler de « charge de travail » dans la suite de l'analyse compte tenu du fait que le temps passé au traitement d'une infraction dépend de la nature de l'infraction elle-même. Aussi, les données étudiées ne permettent pas de conclure quant au niveau réel de sollicitation des agents des forces de l'ordre et à leur charge de travail associée. L'étude du nombre d'infractions traitées par agent des forces de l'ordre nous encourage à construire une boîte à moustaches, illustrant les variations régionales de cette mesure.

Plusieurs régions se distinguent par leurs médianes significativement différentes. Les Pays de la Loire, la région PACA et l'Auvergne Rhône Alpes affichent des médianes supérieures à 17 infractions par agent. À l'inverse, la Corse présente une médiane d'environ 7 infractions par agent. Sa faible étendue interquartile et sa médiane distinctive suggèrent un comportement différent par rapport aux autres régions.

Par ailleurs, une grande dispersion des valeurs est également observée à l'intérieur de chaque région, avec des écarts importants entre les valeurs maximales et minimales. Les régions Nouvelle Aquitaine, Auvergne Rhône Alpes, Occitanie, PACA et Grand Est en sont de bons exemples. En outre, certaines valeurs aberrantes, comme le maximum en Île-de-France (Paris) et en Normandie (Eure), ainsi que le minimum en Bourgogne Franche Comté (Territoire de Belfort), ajoutent à la complexité de la répartition des infractions et des ressources des forces de l'ordre.

En conclusion, cette étape de l'analyse souligne l'interaction complexe entre les infractions, les effectifs de forces de l'ordre et d'autres facteurs, comme la présence de ZSP. La répartition des ressources et la charge de travail des forces de l'ordre varient considérablement d'une région à l'autre, reflétant les défis uniques auxquels chaque région est confrontée en matière de sécurité nationale.

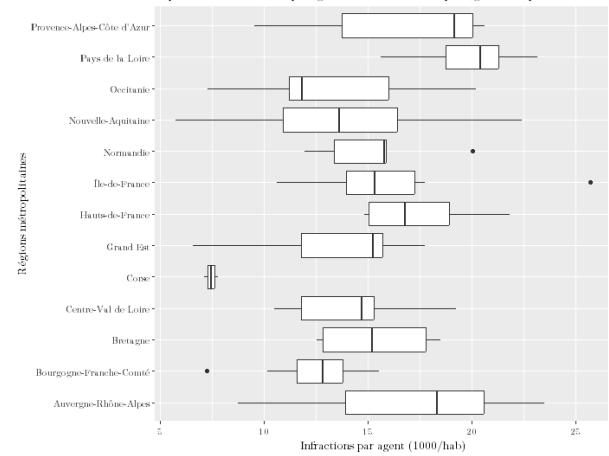
Relation entre infractions et ressource policière en France métropolitaine en 2021  
Analyse statistique de l'impact des forces de l'ordre et des Zones de Sécurité Prioritaire



Ce nouveau graphique ne permet pas de détecter une linéarité évidente entre le nombre de forces de l'ordre et les infractions pour mille habitants. Aussi, et après avoir fait précédemment l'étude des corrélations, on peut alors valider graphiquement que le nombre de forces de l'ordre ne semble pas être directement lié à la présence d'au moins une ZSP.

Cependant, une dichotomie horizontale autour des 40 infractions pour mille habitants semble émerger, séparant les départements ayant au moins une ZSP de ceux n'en possédant aucune. En d'autres termes, et comme précédemment évaluer avec la matrice des corrélations, la répartition des forces de l'ordre semble plus étroitement liée au nombre pondéré d'infractions. Cette relation complexe entre ces variables souligne la nécessité de considérer d'autres facteurs influençant les niveaux de criminalité.

Disparités régionales dans la charge de traitement par les forces de l'ordre  
Répartition des infractions par agent des forces de l'ordre par région métropolitaine en 2021



## C. Variabilité régionale et impact statistique

### 1. Zoom sur la distribution statistique régionale

La quête de compréhension des forces qui influent sur le nombre d'infractions se poursuit, mais cette fois à l'échelle régionale. En regroupant les départements en fonction de leurs régions respectives et en scrutant la distribution statistique des infractions pondérées par région, nous ouvrons de nouvelles perspectives, révélant des disparités significatives dans les tendances régionales.

Cette exploration approfondie enrichit notre vision de l'impact géographique sur la criminalité et nous permet de mieux appréhender les implications pour la sécurité nationale. Notre démarche débute par une analyse des distributions du nombre pondéré d'infractions dans chaque région, illustré à travers des boîtes à moustaches.

Comme prévu, l'Île-de-France émerge comme une région avec une étendue interquartile plus importante que celles des autres régions, soulignant la diversité considérable du nombre d'infractions au sein de ses départements. Dans cette optique, le département de Paris, avec ses caractéristiques spécifiques, se détache comme une valeur aberrante au sein de la région, établissant une dynamique différente par rapport aux autres régions.

Cette analyse nous permet de percevoir la complexité des taux d'infractions dans la capitale et d'appréhender comment des particularités géographiques peuvent influencer la criminalité.

### 2. Variations moyennes régionales et facteurs démographiques

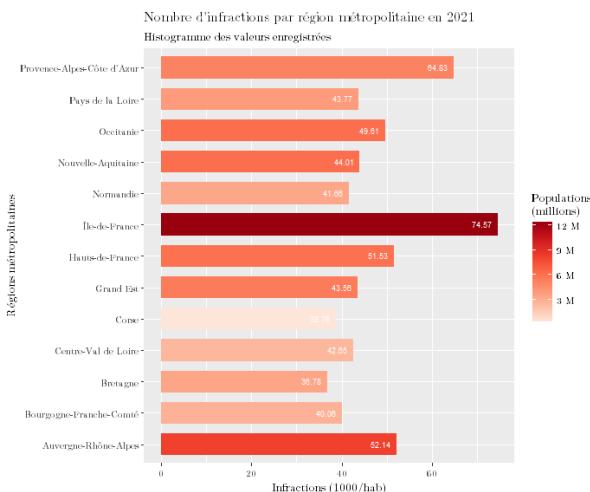
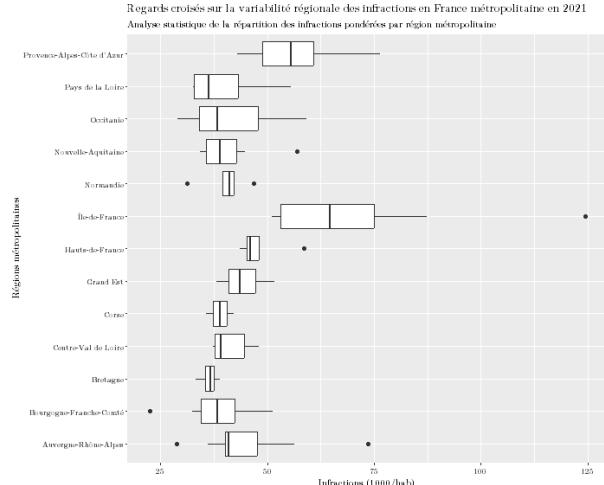
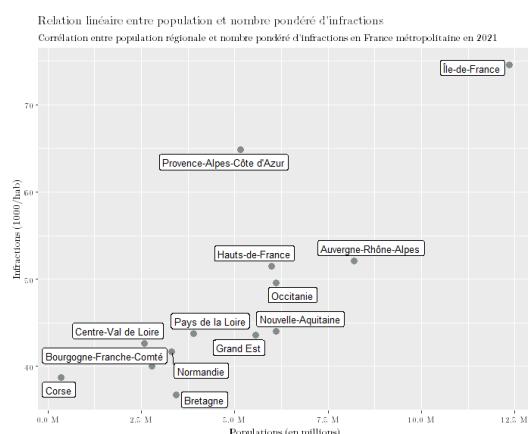
Une fois que nous avons examiné les distributions statistiques, notre attention se tourne vers les moyennes régionales du nombre d'infractions pour mille habitants.

Cette observation immédiate nous révèle des disparités saisissantes entre les régions. L'Île-de-France se distingue avec 74 infractions pour 1000 habitants, suivie de près par la région PACA avec 65 infractions. Par contraste, la région Auvergne Rhône-Alpes affiche 52 infractions. Cette observation met en évidence l'influence de la densité de population sur les taux d'infractions, avec une tendance à une augmentation des infractions dans les régions densément peuplées.

De plus, cette tendance semble s'atténuer lorsque les régions affichent des populations régionales moyennes, soulignant l'impact complexe des facteurs démographiques sur les taux de criminalité. Cette observation entraîne des répercussions significatives pour les politiques de sécurité nationale, qui doivent tenir compte de ces variations pour une allocation efficace des ressources et des efforts de prévention.

### 3. Linéarité entre infractions et populations à l'échelle régionale

Pour affiner notre compréhension, nous examinons graphiquement la relation entre la population régionale et le nombre d'infractions à travers un nuage de points régional.



Comme attendu, l'Île-de-France et la région PACA se distinguent nettement par leur comportement statistique, se démarquant du reste des régions. En excluant ces régions singulières, une tendance à la linéarité émerge entre la population régionale et les taux d'infractions pondérées. Cette linéarité sous-tend l'importance de la taille de la population dans la détermination des taux d'infractions.

Cependant, il est crucial de ne pas simplifier à l'excès cette relation, car d'autres variables complexes influencent également les variations régionales des taux de criminalité.

Cette analyse à l'échelle régionale confirme les constatations précédentes à l'échelle départementale, soulignant l'importance de tenir compte des spécificités géographiques et démographiques lors de la formulation de politiques de sécurité nationale et de stratégies de prévention de la criminalité.

## D. Impact géographique et analyse spatiale

Au terme de notre investigation, nous examinons l'impact de la géographie en étudiant la répartition géographique des infractions à travers la France métropolitaine.

Pour ce faire, nous avons recours à la cartographie, mettant en évidence les disparités régionales et sondant les liens potentiels entre la géographie et les taux pondérés d'infractions pour mille habitants par région.

### 1. Visualisation cartographique

En cartographiant le nombre pondéré d'infractions par région, une image instantanée des variations géographiques se dévoile sous nos yeux.

Cette représentation graphique offre une perspective visuelle sur les disparités régionales en matière de criminalité. Par exemple, l'Île-de-France et la Bourgogne Franche-Comté, bien que géographiquement proches, affichent des extrêmes opposés en termes d'infractions enregistrées. En effet, la Bourgogne Franche-Comté présente environ 40 infractions pour mille habitants, tandis que l'Île-de-France dépasse les 75 infractions pour mille habitants. Aussi, la criminalité ne semble pas être influencée par le voisinage.

### 2. Interprétation statistique et autocorrélation spatiale

Cette observation nous amène à conclure que la proximité géographique, qu'elle soit départementale ou régionale, semble ne pas jouer de rôle significatif dans la répartition du nombre d'infractions sur le territoire national. En d'autres termes, il n'y a sans doute pas (ou très peu) d'autocorrélation spatiale observable dans nos données. Cette constatation soulève des questions sur l'interconnexion entre la géographie et la criminalité et suggère que d'autres facteurs majeurs sont à l'œuvre.

Toutefois, il est important de noter que cette analyse se limite à l'échelle régionale que nous avons étudiée. L'absence d'autocorrélation spatiale ne doit pas être extrapolée à toutes les échelles géographiques. Pour approfondir cette conclusion, il est nécessaire de mener des analyses à différentes échelles et d'explorer les influences géographiques spécifiques à chaque contexte.

### 3. Test de Durbin-Watson et validation des hypothèses

Il convient de noter que, compte tenu des écarts significatifs entre les régions observées visuellement, il n'est pas nécessaire d'effectuer un test de Durbin-Watson pour confirmer l'hypothèse formulée. Le test de Durbin-Watson est une mesure statistique utilisée pour détecter la présence d'autocorrélation dans les résidus d'une régression. Il évalue si les erreurs dans le modèle statistique sont corrélées spatialement. En somme, ce test aurait pour objectif de valider ou d'invalider l'idée que les taux d'infractions sont influencés par des facteurs géographiques.

Cependant, vu les différences marquées entre les régions et l'absence d'autocorrélation spatiale évidente, il est plausible de conclure que la géographie, du moins à l'échelle départementale ou régionale, n'exerce qu'un impact limité sur le nombre pondéré d'infractions. Cette perspective complexe met en avant la nécessité de considérer un large éventail de facteurs lors de la formulation de politiques de sécurité nationale et de stratégies de prévention de la criminalité.

## **E. Analyse de la catégorisation des infractions**

### 1. Enjeux, objectifs et limites de l'analyse

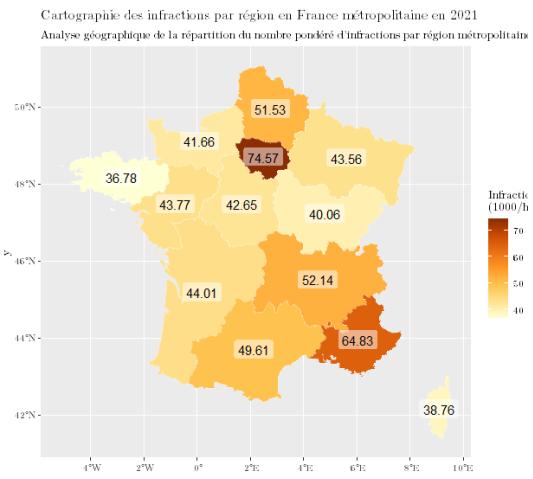
#### *a. Intégration des données de catégorisation des infractions*

Une fois avoir effectué l'analyse principale sur les départements, on peut également s'atteler à regarder en détail les types d'infractions afin d'apporter davantage de granularité quant aux crimes et délits commis. Pour ce faire, on transpose le jeu de données de telle sorte que les types d'infractions (non adressés jusqu'à présent) se trouvent en ligne. On va s'intéresser à la métropole dans son ensemble, en ne conservant pas de vision au département ou à la région mais pour l'ensemble du territoire. Aussi, nous allons prendre le total des infractions commis sur le territoire au détriment d'une vue à l'échelle départementale. Puis, on ajoute les catégories sous-catégories d'infractions à partir des données issues de l'Etat 4001.

Il est important de garder à l'esprit que certaines infractions peuvent être comptabilisées dans plusieurs catégories (ou sous-catégories) du fait de leur nature. Pour prendre un exemple concret, les vols à main armée avec arme à feu contre des entreprises de transports de fonds appartiennent à la fois à la catégorie des « atteintes aux biens », à la catégorie des « atteintes aux personnes » et ainsi qu'à la catégorie de la « Criminalité organisée et délinquance spécialisée ». Aussi, la somme des infractions de l'ensemble des catégories (ou sous-catégories) ne permet pas d'expliquer le phénomène dans son ensemble. Dit autrement, il faut revenir à la granularité de départ, à la maille de l'infraction, pour pouvoir mesurer le phénomène global. Toutefois, la catégorisation (et sous-catégorisation) des infractions amène un regard très intéressant vis-à-vis de leur répartition et leur influence.

#### *b. Limites dans l'interprétation des catégories d'infractions*

Il est important de mentionner le fait que certaines catégories d'infractions sont plus ou moins significatives de l'insécurité subie par les citoyens. En effet, Les enquêtes auprès de la population permettent d'apprécier la proportion d'infractions subies pour lesquelles une plainte est déposée et cette métrique peut beaucoup varier en fonction du type d'infractions commis. Par exemple, 90% pour les vols de voiture, de 70 à 80% pour les cambriolages de résidences principales, de 20 à 30% pour les violences physiques hors ménage et moins de 10% pour les violences sexuelles hors ménage et les violences physiques et sexuelles au sein du ménage). De même, pour les infractions pour lesquelles il n'existe pas de victimes physiques ou morales constituées (infractions à la législation sur les stupéfiants, sur le travail, sur le droit des étrangers, sur la protection de l'environnement ou le proxénétisme par exemple), le nombre d'infractions enregistrées retrace l'activité des forces de sécurité et témoigne de l'intensité de leurs efforts pour repérer les infractions et en confondre leurs auteurs présumés, et très peu l'évolution réelle de la délinquance.



### c. Relation entre sécurité et sentiment d'insécurité :

La variabilité des taux de plainte pour différentes catégories d'infractions est cruciale et l'enquête CVS<sup>(1)</sup> de 2020 illustre cette diversité. Par exemple, seulement 0,2% des personnes âgées de 14 ans ou plus ont signalé avoir subi un vol avec violences physiques, tandis que 1,1% ont déclaré être victimes d'un cambriolage ou d'une tentative de cambriolage de leur domicile. Ces chiffres révèlent des perceptions et des attitudes variées des citoyens envers la criminalité, influencées par des facteurs tels que la gravité perçue du crime, la confiance envers les forces de sécurité et les conséquences potentielles pour les victimes.

Bien que la corrélation entre les taux de plainte et le sentiment d'insécurité soit complexe, des tendances se dégagent. Parfois, des infractions peu signalées génèrent un sentiment d'insécurité élevé en raison de leur nature traumatisante ou médiatisée. D'autre part, des taux de signalement élevés ne garantissent pas nécessairement un sentiment d'insécurité élevé si la confiance envers les forces de sécurité est établie. Cependant, il est important de noter que l'étude ne prend pas en compte la variabilité des taux de plainte, bien que ce phénomène puisse être mis en lumière par des analyses complémentaires. Toutefois, cela ne diminue pas l'importance des implications pour la sécurité nationale. En effet, en considérant la diversité des perceptions, les gouvernements peuvent développer des stratégies ciblées pour répondre aux préoccupations des citoyens, améliorer la confiance dans les mesures de sécurité et optimiser l'allocation des ressources pour un environnement sécurisé et rassurant.

## 2. Analyse de la catégorisation des infractions

### a. Vue d'ensemble des catégories d'infractions

L'analyse fine des catégories et sous-catégories d'infractions permet d'obtenir un aperçu détaillé des tendances criminelles en France. L'histogramme ci-dessous présente le nombre total d'infractions enregistrées par an pour chaque catégorie.

Les "Atteintes aux biens" se distinguent avec près de 1,6 million d'infractions, suivies des "Atteintes aux personnes" (environ 650 000), des "Infractions économiques, financières et escroqueries" (environ 460 000), des "Infractions révélées par l'action des services" (environ 330 000), des "Autres infractions" (environ 320 000), et enfin de la "Criminalité organisée et délinquance spécialisée" (environ 30 000).

Cependant, il est important de noter que certaines infractions sont décomptées dans plusieurs catégories, ce qui peut créer des distorsions.

Compte tenu de l'ampleur des "Atteintes aux biens", responsables de près de 1,6 million d'infractions chaque année, il devient pertinent d'approfondir notre compréhension de cette catégorie. L'histogramme des infractions spécifiques liées aux "Atteintes aux biens" ci-dessous illustre les cinq types prédominants : "Autres vols simples contre des particuliers dans des locaux ou lieux publics" (260 000 cas), "Autres vols simples contre des particuliers dans des locaux privés" (160 000 cas), "Cambriolages de locaux d'habitation principale" (160 000 cas), "Vols à la roulotte" (210 000 cas), et "Vols à la tire" (120 000 cas).

Ces résultats peuvent être interprétés à la lumière de l'enquête CVS, révélant des implications significatives pour les taux de plainte et le sentiment d'insécurité.

Par exemple, les infractions telles que les "Autres vols simples contre des particuliers dans des locaux ou lieux publics" et les "Vols à la roulotte" pourraient générer un sentiment d'insécurité plus élevé, car elles touchent directement les biens des individus dans des environnements publics.

En revanche, les "Autres vols simples contre des particuliers dans des locaux privés" et les "Cambriolages de locaux d'habitation principale" pourraient avoir des taux de plainte plus élevés, car les victimes sont plus enclines à signaler des infractions qui se produisent dans leur domicile.



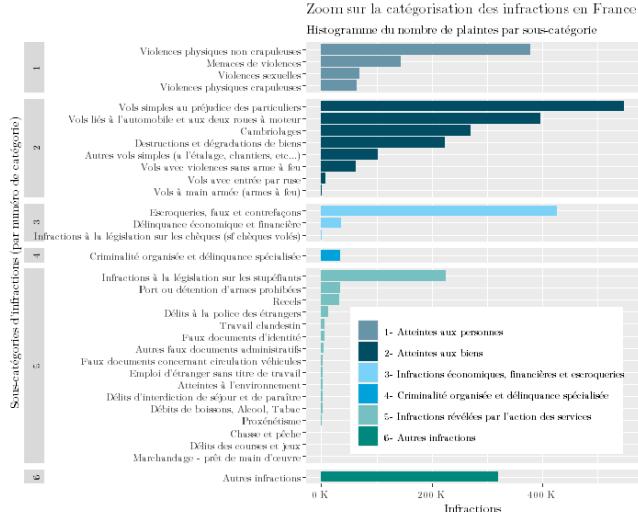
<sup>(1)</sup> L'enquête « Cadre de vie et sécurité » (CVS), dite de « victimisation », a été conduite chaque année de 2007 à 2021 (à l'exception de 2020 et avec un protocole de collecte particulier en 2021 compte tenu de la crise sanitaire). Elle vise à connaître les faits de délinquance dont les ménages et les individus ont pu être victimes dans les deux années précédant l'enquête, qu'ils aient, ou pas, donné lieu à une déclaration dans les services de police ou de gendarmerie mais également à recueillir, auprès de l'ensemble de la population (victimes et non victimes), leur opinion concernant leur cadre de vie et la sécurité, à analyser le sentiment d'insécurité ainsi que le niveau de satisfaction envers l'action de la justice et des forces de sécurité. Les informations issues de l'enquête CVS sont distinctes et complémentaires des données enregistrées par la police et la gendarmerie nationales car les victimes ne déposent pas toujours plainte. Combinées, elles offrent des outils précieux pour évaluer et analyser tant la délinquance que le sentiment d'insécurité.

Cette analyse souligne l'importance de prendre en compte à la fois les taux de plainte et les perceptions de l'insécurité pour une compréhension plus complète des infractions et de leurs effets.

#### b. Focus sur la sous-catégorisation des infractions

L'analyse des catégories nous permet d'avoir un premier aperçu de la distribution des types d'infractions. On peut ensuite descendre à la granularité des sous-catégories afin de comprendre davantage comment chacune des catégories est constituée.

L'examen de l'histogramme ci-dessous révèle des tendances significatives au sein des différentes catégories d'infractions, permettant d'approfondir notre compréhension des profils criminels en France. Chaque catégorie se distingue par des sous-catégories ayant des implications importantes pour les taux de plainte et le sentiment d'insécurité :



- Atteinte aux personnes : La sous-catégorie des "Violences physiques" ressort clairement avec près de 400 000 cas annuels. Cela reflète une préoccupation majeure en matière de sécurité, étant donné que ces crimes impliquent directement des atteintes à l'intégrité physique des individus. Les taux de plainte pour ces infractions pourraient être relativement élevés, car les victimes pourraient être plus enclines à signaler des agressions physiques graves.
- Atteinte aux biens : Les "Vols simples au préjudice des particuliers" constituent la sous-catégorie la plus dominante, totalisant près de 600 000 cas par an. Cette prévalence peut entraîner un sentiment d'insécurité, surtout si ces vols touchent directement les biens personnels des individus. Les "Vols liés à l'automobile et aux deux roues à moteurs" (environ 400 000 cas/an) ainsi que les "Cambriolages" (environ 250 000 cas/an) soulèvent également des préoccupations majeures en matière de sécurité résidentielle et de pertes matérielles. Les taux de plainte pour ces infractions pourraient varier en fonction de la perception individuelle de la gravité du préjudice.
- Escroqueries et infractions économiques et financières : Les "Escroqueries, faux et contrefaçons" prédominent dans cette catégorie avec environ 230 000 cas par an. Les escroqueries bancaires et économiques peuvent avoir un impact sur la confiance des citoyens dans les transactions financières et ébranler le sentiment de sécurité économique. Les taux de plainte pour ces infractions pourraient être influencés par la méfiance envers les transactions en ligne et la détection des fraudes.
- Criminalité organisée et délinquance spécialisée : Malgré son faible nombre absolu d'infractions (environ 30 000 cas), cette catégorie englobe des crimes graves. Ces infractions pourraient ne pas susciter un sentiment d'insécurité élevé en raison de leur nature spécifique, mais leur impact sur la société peut être considérable.
- Infractions révélées par l'action des services : Les "Infractions à la législation sur les stupéfiants" dominent cette catégorie, totalisant 80% des infractions. Les taux de plainte pourraient varier en fonction de la perception individuelle de la gravité des problèmes liés aux drogues, influençant ainsi le sentiment d'insécurité dans certaines communautés.

Autres infractions : Bien qu'une seule sous-catégorie soit représentée ici, elle enregistre plus de 300 000 cas par an. Elle englobe notamment les délits au sujet de la garde des mineurs, les atteintes à la dignité et à la personnalité, et les violations de domicile. Cette diversité d'infractions peut contribuer à une variété de taux de plainte et de sentiments d'insécurité en fonction de la nature de chaque infraction et il est difficile d'en tirer des conclusions.

#### F. Le besoin de modéliser pour mieux prévenir

Notre plongée minutieuse dans les tendances et les motifs émergents a permis de saisir la complexité de la situation sécuritaire en France. L'analyse détaillée des données, de leur distribution régionale aux corrélations entre infractions et populations, a livré des enseignements essentiels pour cerner les défis qui se présentent.

La transition vers la prochaine phase de notre étude est marquée par une approche plus concrète et méthodologique. Nous aspirons à résoudre deux questions fondamentales. La première consiste à déterminer la présence ou l'absence de Zones de Sécurité Prioritaires (ZSP) dans chaque département. Cette initiative vise à éclairer les stratégies de déploiement des ressources sécuritaires. La seconde question s'attaque à l'estimation précise des effectifs requis des forces de l'ordre pour chaque département, en tenant compte des données explorées jusqu'à présent.

Pour atteindre ces objectifs, nous mettrons en œuvre une méthode de classification ainsi qu'une méthode de régression. Cette approche pragmatique constitue le fil conducteur entre les tendances mises en lumière dans la partie précédente et les résultats concrets que nous cherchons à obtenir. Les enseignements tirés de l'analyse exploratoire des données, allant des tendances régionales à la variabilité statistique, fournissent la base nécessaire pour orienter nos modèles quantitatifs. Ainsi, la complémentarité de l'exploration des données et de la modélisation statistique permettra de passer de l'observation à l'action, en fournit des recommandations tangibles pour renforcer la sécurité de manière ciblée et efficace.

## **IV. Régression logistique pénalisée**

### **A. Vers un modèle de classification robuste pour prédire les ZSP**

#### 1. Modéliser la présence de ZSP grâce à une méthode de classification

Après une analyse minutieuse des tendances et des schémas qui se dessinent à travers nos données, nous franchissons une étape cruciale : la modélisation visant à évaluer la présence ou l'absence de Zones de Sécurité Prioritaires (ZSP) au sein des départements français. Cette transition reflète notre désir de traduire les enseignements de l'analyse descriptive en des approches pratiques, capables de prédire des scénarios concrets. Ainsi se pose la question : peut-on élaborer un modèle qui puisse classifier les départements en fonction de la présence ou de l'absence de ZSP ?

Notre objectif est double. Tout d'abord, il nous faut développer une méthode de classification qui puisse utiliser les informations inhérentes aux infractions enregistrées dans les départements pour prendre des décisions éclairées sur la nécessité de ZSP. Cela requiert une exploration rigoureuse des méthodes de classification principales, ainsi qu'une évaluation approfondie de leur applicabilité à notre contexte particulier. La sélection de la méthode la plus adaptée prendra en compte des considérations telles que la taille de notre jeu de données, la complexité des relations entre variables et la performance prédictive.

Il est essentiel de souligner que cette approche de classification se connecte directement à notre premier objectif pour cette étape. En optant pour cette méthodologie, notre but est d'utiliser les informations découvertes grâce à l'analyse descriptive pour identifier, à partir de données historiques, les départements où des ZSP pourraient être nécessaires. Ce modèle apportera une dimension prédictive à notre étude, transformant les observations passées en une base pour anticiper les développements futurs en matière de sécurité.

La variété des méthodes de classification que nous allons examiner dans cette phase démontre notre engagement à trouver la solution la plus pertinente pour notre objectif. En prenant en compte des facteurs tels que la taille de notre jeu de données, la complexité des relations entre variables et la capacité prédictive, nous explorerons diverses approches pour sélectionner la méthodologie optimale. Dans la continuité de notre démarche, nous franchissons une nouvelle étape exploratoire, avec pour objectif ultime de fournir des solutions pratiques et des recommandations en matière de sécurité publique.

#### 2. Explorer et comparer les méthodes de classification

Dans cette étape cruciale, nous plongeons dans l'univers des méthodes de classification, déterminés à identifier celle qui répond le mieux à notre objectif de prédire la présence ou l'absence de Zones de Sécurité Prioritaires (ZSP). Plusieurs approches se présentent à nous, chacune avec ses avantages et ses inconvénients, ajustables à notre contexte :

- Arbres de Décision : Cette méthode hiérarchique divise les données en sous-groupes en utilisant des seuils sur les variables. Pour chaque nœud, l'attribut qui maximise l'homogénéité des classes est choisi comme nœud de décision. Les calculs mathématiques sont liés à la recherche de l'attribut optimal et au calcul des critères d'homogénéité, tels que l'indice de Gini ou l'entropie.
- Forêts Aléatoires : C'est une extension des arbres de décision où plusieurs arbres sont construits, et chaque arbre vote pour la classe finale. Les avantages incluent la réduction du surajustement (dit « over-fitting ») par le vote majoritaire. Le processus mathématique implique la création de multiples arbres et l'agrégation des prédictions.
- Machine à Vecteurs de Support (SVM) : SVM cherche à maximiser la marge entre les classes en trouvant un hyperplan optimal. La formulation mathématique implique la minimisation de la norme du vecteur de poids tout en satisfaisant les contraintes d'assignation correcte des classes.
- K-Nearest Neighbors (K-NN) : Cette méthode attribue une classe en fonction de ses voisins les plus proches. Mathématiquement, il calcule les distances entre les points et attribue la classe majoritaire parmi les K plus proches voisins.
- Naïve Bayes : Basé sur le théorème de Bayes, il suppose que les attributs sont indépendants entre eux, d'où « naïve ». Les calculs impliquent l'estimation des probabilités conditionnelles pour chaque classe.
- Régression Logistique : Elle modélise la probabilité que la variable dépendante appartienne à une catégorie en fonction des variables indépendantes. Les mathématiques impliquent la fonction logistique pour transformer la somme pondérée des variables en une probabilité.

Parmi toutes ces méthodes, la régression logistique se présente comme un choix judicieux pour notre analyse en raison de ses nombreuses forces qui répondent à nos besoins spécifiques. Toutefois, il est important d'en comprendre ses forces, ses contraintes potentielles et les moyens de les surmonter.

#### 3. Comprendre les forces d'un modèle de régression logistique

La régression logistique est reconnue pour sa simplicité à être mise en place. Cette simplicité serait un atout pour cette tentative de résolution d'un problème complexe, original et multidimensionnel.

De même, cette méthode a comme avantage majeur sa capacité à fournir des informations sur l'importance de chaque caractéristique. Nous pourrions ainsi interpréter les coefficients du modèle comme des indicateurs de l'importance des différentes variables, ce qui renforcerait la pertinence de notre analyse.

Elle excelle également dans la gestion des données à faible dimensionnalité. Cela est un élément majeur à prendre en considération car l'ensemble de données initial comporte plus d'une centaine de variables qui représentent chacune un type d'infractions. Aussi, il nous faudrait sans doute travailler avec les catégories et sous-catégories d'infractions pour réduire la complexité.

Par ailleurs, elle est très efficace lorsque les données peuvent être linéairement séparées. Dans notre cas, l'analyse exploratoire nous a permis de conclure que le nombre total d'infractions était une variable liée linéairement à la présence d'une ZSP. La séparation était graphiquement visible sur les différents nuages de points tracés précédemment.

Enfin, et outre les résultats de classification, la régression logistique fournit des probabilités calibrées, ce qui pourrait être une force supplémentaire pour notre analyse en fournissant des informations sur la confiance des prédictions.

#### 4. Identifier les contraintes mathématiques à prendre en compte

Ces forces viennent également avec leur lot de contraintes. A ce titre, l'un des inconvénients majeurs de la régression logistique est qu'elle a tendance à sur-ajuster les données lorsque la dimensionnalité est élevée. Comme explicité précédemment, notre jeu de données initial comporte davantage de variables que d'individus. Ce problème est donc d'autant plus préoccupant.

Pour le surmonter, nous pourrions procéder à une Analyse en Composantes Principales pour réduire la dimensionnalité du jeu de données via la construction de composantes comme combinaisons linéaires des variables étudiées. Toutefois, cette méthode pourrait complexifier l'interprétation des résultats et l'explicabilité du modèle. L'utilisation des catégories et sous-catégories semble être une solution plus adaptée.

La multicolinéarité, c'est-à-dire la forte corrélation entre de nombreuses variables, est également une autre contrainte majeure. Nous l'avons identifiée lors de l'analyse exploratoire grâce à l'étude d'une matrice de corrélation qui nous avait indiqué une forte corrélation entre certaines variables. Pour contourner ce problème, nous pourrions envisager d'appliquer des techniques de régularisation telles que la régularisation Lasso (L1) ou Ridge (L2). Pour rappel, la régularisation L1 pénalise les coefficients inutiles en les ramenant à zéro, tandis que la régularisation L2 limite leurs valeurs.

La taille réduite de notre ensemble de données pourrait entraîner une prédiction parfaite de toutes les classes dépendantes. Pour éviter cela, nous pourrions tenter d'appliquer des méthodes de validation croisée pour évaluer la robustesse et la généralisation de notre modèle. Toutefois, et compte tenu de la centaine d'individus étudiés, nous risquons d'être confronté à ce problème malgré tout. Une analyse plus fine, au niveau communal par exemple, permettrait sans doute une plus grande robustesse.

Enfin, la régression logistique suppose une relation linéaire entre les variables indépendantes et la réponse. C'est une bonne nouvelle pour la suite car notre analyse exploratoire a précédemment montré que la plupart des variables semblaient avoir une liaison linéaire avec la présence ou non d'au moins une ZSP. Nous pourrions également effectuer un Test de Fisher de la contribution globale pour valider cette hypothèse et ainsi renforcer la pertinence de la régression logistique dans notre contexte.

A la lumière de tous ces éléments, la régression logistique se révèle être un choix logique et adapté à notre analyse malgré les contraintes qui en découlent. En adaptant notre approche, en réduisant la dimensionnalité de nos données et en envisageant des techniques de régularisation, nous serions en mesure de surmonter la plupart de ces contraintes et de tirer parti de la puissance de la régression logistique pour notre étude.

### **B. La mise en place d'une régression logistique**

La mise en place d'une régression logistique commence par la préparation des données et leur division en deux échantillons distincts : un échantillon d'entraînement et un échantillon de test. Cette étape fondamentale contribue de manière cruciale à la robustesse et à la validité de notre modèle.

#### 1. Réduction de la dimensionnalité et préparation des données

Comme mentionné précédemment, nous réduisons le jeu de données initial en agrégeant les informations au niveau des catégories d'infractions, au lieu de conserver le détail des types d'infractions individuels. Cette agrégation nous permet d'obtenir une vue plus globale tout en réduisant la complexité du modèle.

A cette agrégation, nous y incorporons également trois variables supplémentaires : le nombre pondéré de plaintes enregistrées par agent des forces de l'ordre, le nombre pondéré de forces de l'ordre et le nombre d'habitants par département de France métropolitaine. Ces nouvelles variables apportent une dimension supplémentaire à notre modèle, tenant compte de la réaction des forces de l'ordre et de la réponse du public face aux infractions.

#### 2. Scission en échantillons d'entraînement et échantillons de test

La scission du jeu de données en deux échantillons distincts revêt une importance cruciale dans le processus de modélisation. Cette étape est essentielle pour évaluer et valider les performances du modèle de régression logistique. Elle s'articule autour des points suivants :

- L'échantillon d'entraînement : cet échantillon est utilisé pour ajuster et entraîner le modèle de régression logistique. Il sert à apprendre les relations entre les variables explicatives (telles que les catégories d'infractions, le nombre de plaintes, et le nombre de forces de l'ordre) et la variable cible (la présence ou l'absence de Zones de Sécurité Prioritaires). Le modèle est ajusté sur cet échantillon pour comprendre comment les variables influencent la réponse.
- L'échantillon de test : cet échantillon est réservé à l'évaluation et à la validation du modèle une fois qu'il a été entraîné sur l'échantillon d'entraînement. Il s'agit d'une sorte de contrôle indépendant qui permet de mesurer la capacité du modèle à généraliser ses prédictions sur de nouvelles données. En d'autres termes, il nous permet de tester si le modèle peut effectuer des prédictions précises sur des données qu'il n'a pas encore vues.

Le choix du seuil de scission entre ces deux échantillons s'avère être un arbitrage délicat. Les bonnes pratiques empiriques encouragent à opter pour un découpage de 80% pour l'entraînement et 20% le test. Pour cette étude, nous arbitrerons pour un découpage 75%/25%.

En scindant nos données de cette manière, nous évitons le piège du surajustement (overfitting) du modèle, où ce dernier s'adapte trop précisément aux données d'entraînement et ne peut pas généraliser correctement. La séparation en échantillon d'entraînement et de test nous permet de quantifier la capacité prédictive réelle de notre modèle sur des données nouvelles et non biaisées. Cela garantit que notre modèle est fiable et peut être utilisé pour des prédictions futures en matière de sécurité publique.

### 3. Configuration des paramètres du modèle

Compte tenu des contraintes évoquées plus tôt, on souhaite effectuer une régression logistique avec pénalisation/régularisation. Rappeler ce qu'est la pénalisation et à quoi cela sert dans notre cas précis. Par exemple, les méthodes de régularisation (Ridge/L2, Lasso/L1) peuvent aider à réduire les effets de la multicolinéarité en pénalisant les coefficients des variables fortement corrélées. Pour ce faire, on décide d'utiliser la fonction `cv.glmnet()` du package 'glmnet'.

La démarche derrière l'utilisation de la fonction `cv.glmnet()` pour effectuer une régression logistique avec régularisation peut être décrite en trois étapes clés, chacune apportant une composante essentielle à la construction d'un modèle robuste et généralisable.

#### *a. Fonction de perte, matrice de caractéristiques et variable réponse*

Au cœur de notre démarche se trouve la fonction de perte, qui définit l'objectif de notre modèle. Dans ce contexte, nous utilisons la famille de distribution 'binomial', ce qui signifie que nous cherchons à résoudre un problème de classification binaire. La variable réponse, souvent notée "Y" dans les modèles linéaires généralisés, représente ici la présence ou l'absence de Zones de Sécurité Prioritaires (ZSP) dans les départements. Pour la rendre compatible avec notre modèle, nous la convertissons en une variable numérique binaire, prenant les valeurs 0 ou 1. L'objectif est de prédire la probabilité que Y soit égale à 1 en fonction des caractéristiques (X) que nous avons extraites de nos données brutes.

#### *b. Méthodes de régularisation*

La régularisation est une technique fondamentale dans la modélisation statistique qui nous permet de contrôler la complexité de notre modèle. Elle agit en ajoutant une pénalité aux coefficients associés à chaque caractéristique (variable) dans notre modèle de régression. La nature de cette pénalité est déterminée par le paramètre alpha, que nous avons spécifié. Les trois options courantes sont :

- Ridge (L2) : Cette régularisation ajoute une pénalité L2 aux coefficients. Elle favorise des coefficients plus petits, ce qui peut aider à prévenir le surajustement en réduisant la magnitude des coefficients.
- Lasso (L1) : La régularisation L1 ajoute une pénalité L1 aux coefficients. Elle encourage la parcimonie en poussant certains coefficients vers zéro, ce qui permet la sélection automatique des caractéristiques les plus importantes.
- Elastic Net : Cette régularisation combine à la fois les pénalités L1 et L2. Cela offre un compromis entre la parcimonie de Lasso et la réduction de magnitude de Ridge.

La régularisation est essentielle car elle empêche notre modèle de devenir trop complexe et surajusté aux données d'apprentissage, ce qui le rendrait peu généralisable à de nouvelles données. Dans notre cas de figure, on pourrait intuitivement pencher pour une régularisation L1 (Lasso), cette dernière étant particulièrement efficace dans la sélection automatique des caractéristiques du modèle. Elle ajoute une pénalité qui pousse certains coefficients vers zéro, ce qui équivaut à exclure ces caractéristiques du modèle final.

#### *c. Méthode de validation croisée avec règle du 1se*

La validation croisée (CV) est un pilier essentiel de l'évaluation de la performance de notre modèle. Dans notre contexte, nous avons opté pour une validation croisée k-fold, où k représente le nombre de plis dans lesquels nos données sont divisées. Cette méthode assure la robustesse de notre modèle et sa capacité à généraliser à de nouvelles données, ce qui est fondamental pour garantir la fiabilité de nos prédictions. Le processus de validation croisée s'exécute en suivant ces étapes :

1. Les données sont partitionnées en k plis, généralement de manière aléatoire.
2. Le modèle est formé k fois, chaque itération utilisant k-1 plis comme données d'apprentissage et le pli restant comme données de validation.
3. La performance du modèle est évaluée à chaque itération à l'aide de la fonction de perte associée à la régression logistique. Cela nous fournit une mesure de l'efficacité de notre modèle dans la prédiction de la variable cible, à savoir la présence ou l'absence de Zones de Sécurité Prioritaires (ZSP).
4. Les performances de toutes les itérations sont agrégées pour obtenir une évaluation globale de la performance du modèle.

La validation croisée permet d'éviter que le modèle ne soit biaisé par la manière dont les données sont divisées et donne la possibilité de sélectionner un modèle optimal en termes de paramètres de régularisation. Elle joue un rôle crucial dans la création d'un modèle fiable et généralisable pour la prédiction des ZSP dans les départements français.

Dans ce contexte, nous allons utiliser la règle du 1se, une approche couramment adoptée pour sélectionner la meilleure valeur de lambda lors de la validation croisée. Cette règle consiste à choisir la valeur de lambda qui produit un modèle dont la performance est à un écart-type (standard error) de la meilleure performance obtenue pendant la validation croisée. Son objectif est d'éviter la sélection d'un modèle excessivement complexe (surajusté) tout en maintenant un bon ajustement.

Le paramètre 'lambda.1se' correspond à la valeur de lambda associée à cette règle du 1-Standard Error et est sélectionné comme étant la meilleure valeur pour la pénalisation. Généralement, 'lambda.1se' est plus élevé que la valeur de lambda qui minimisera l'erreur de validation croisée, assurant ainsi que le modèle demeure parcimonieux tout en conservant sa performance prédictive. Cette approche garantit un équilibre optimal entre complexité et précision pour notre modèle de régression logistique régularisée.

### 3. Paramétrage et exécution du modèle

Après avoir correctement paramétré notre modèle de régression logistique avec régularisation, nous pouvons désormais l'exécuter pour évaluer les performances de chaque méthode de régularisation.

```
# Testing with the full model (without interactions)
model_1 = as.formula('zsp_status ~ police_workforce + infraction_per_personnel + population + category_1 + category_2 +
category_3 + category_4 + category_5 + category_6')

# Testing with the full model (with interactions between 'total' and 'population')
model_2 = as.formula('zsp_status ~ police_workforce + infraction_per_personnel + population + category_1 + category_2 +
category_3 + category_4 + category_5 + category_6' + total:population)
model_formula = model_1

# Building a function to perform logistic regression with regularization
logreg_model_with_regularization = function(regularization){
  cv.glmnet(x = as.matrix(model.matrix(model_formula, data=infraction_logreg_train)),
             y = as.numeric(infraction_logreg_train$zsp_status) - 1,
             alpha = regularization, # 0 = Ridge ; 0.5 = Elastic Net ; 1 = Lasso
             family = 'binomial')

# Performing logistic regression with different regularization techniques
set.seed(123456)
infraction_logreg_ridge_model = logreg_model_with_regularization(0)
infraction_logreg_elasnet_model = logreg_model_with_regularization(0.5)
infraction_logreg_lasso_model = logreg_model_with_regularization(1)
```

Pour évaluer la performance de ces modèles, plusieurs métriques sont disponibles, chacune apportant un éclairage différent sur la qualité de la modélisation. Dans le cadre de cette étude, nous allons nous pencher sur trois métriques essentielles :

- AIC (Critère d'Information d'Akaike) : L'AIC est un indicateur de la qualité d'ajustement du modèle aux données. Il tient compte de la capacité du modèle à expliquer la variation des données tout en pénalisant la complexité du modèle. Plus précisément, l'AIC estime la qualité de l'ajustement du modèle par rapport au nombre de paramètres qu'il utilise. Dans notre contexte de régression logistique, un AIC plus bas indique un meilleur ajustement du modèle aux données.
- BIC (Critère d'Information Bayésien) : Le BIC est un autre critère d'information qui, comme l'AIC, évalue la qualité de l'ajustement du modèle tout en pénalisant la complexité. Cependant, le BIC accorde une pénalité plus importante à la complexité du modèle que l'AIC. Par conséquent, le BIC favorise les modèles plus parcimonieux en pénalisant davantage les modèles avec un grand nombre de paramètre.
- R<sup>2</sup> Ajusté (Coefficient de Détermination Ajusté) : Le R<sup>2</sup> ajusté mesure la proportion de la variance totale dans la variable dépendante qui est expliquée par le modèle. Il est ajusté pour tenir compte du nombre de prédicteurs dans le modèle, ce qui en fait une métrique utile pour évaluer si l'ajout de variables au modèle améliore réellement son ajustement. Un R<sup>2</sup> ajusté plus élevé indique un modèle qui explique davantage la variance de la variable dépendante, tout en prenant en compte la complexité du modèle.

Nous avons construit une fonction personnalisée qui calcule ces trois métriques à partir des résultats de la fonction cv.glmnet(). En utilisant cette fonction, nous avons comparé les performances des différentes méthodes de régularisation. Voici les résultats obtenus avec une graine aléatoire (seed) fixée à 123456 :

Model	Alpha	AIC	BIC	Adjusted R-squared
Ridge	0	-38.01	-20.12	0.63
Elastic Net	0.5	-55.28	-38.79	0.82
Lasso	1	-62.94	-52.99	0.79

Nous pouvons constater que le modèle Lasso présente les meilleures performances, car il minimise à la fois le critère d'Akaike (AIC) et le critère d'Information Bayésien (BIC). Ces deux métriques visent à sélectionner des modèles qui offrent un bon équilibre entre l'ajustement aux données et la complexité du modèle. Le choix du modèle Lasso est donc justifié par son aptitude à expliquer efficacement la variation des données tout en utilisant un ensemble réduit de paramètres, ce qui favorise la généralisation à de nouvelles données.

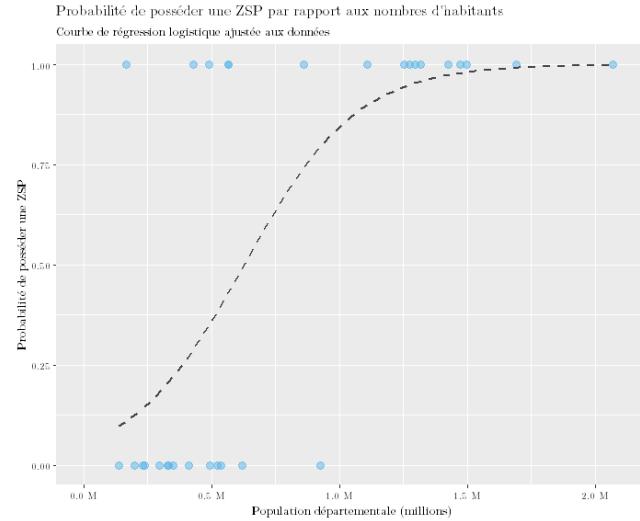
#### 4. Evaluation de la performance globale du modèle

##### *a. Courbe de régression logistique*

Pour évaluer la performance globale de notre modèle de régression logistique avec régularisation, nous avons construit deux courbes de régression logistique distinctes.

Une première courbe a été élaborée en utilisant la population totale des départements français en nombre d'habitants comme variable explicative, et une deuxième courbe en utilisant le nombre pondéré d'infractions dans chaque département. Ces courbes sont essentielles pour valider la capacité de notre modèle à prédire la présence ou l'absence de Zones de Sécurité Prioritaires (ZSP) dans ces départements.

Pour la première courbe, basée sur la population, tous nos points de données appartiennent à l'une des deux catégories : 0 (échec) ou 1 (réussite).



Dans ce contexte, ces valeurs signifient que notre modèle attribue à chaque département une prédiction binaire : il prédit soit que le département nécessite une ZSP (1), soit qu'il n'en nécessite pas (0). Cette première courbe est cruciale pour évaluer comment notre modèle se comporte en fonction de la taille de la population des départements. En analysant cette courbe, nous pouvons noter des caractéristiques spécifiques. Par exemple, pour un département ayant une population d'un million d'habitants, notre modèle prédit une probabilité d'environ 80% qu'il nécessite une ZSP. Cette prédiction est importante car elle nous indique comment notre modèle évalue le besoin potentiel de ZSP en fonction de la population.

La deuxième courbe, basée sur le nombre pondéré d'infractions par mille habitants, suit un schéma similaire à la première. Les valeurs de données sont également binaires : 0 (échec) ou 1 (réussite). Cette courbe évalue la performance de notre modèle en fonction du nombre pondéré d'infractions dans chaque département.

Pour un département ayant un nombre d'infractions pondéré de 45 pour mille habitants, notre modèle prédit une probabilité d'environ 50% qu'il nécessite une ZSP.

Cette prévision nous donne un aperçu de la manière dont notre modèle considère le besoin potentiel de ZSP en fonction du niveau d'infractions.

Ces courbes de régression logistique sont essentielles pour évaluer la performance globale de notre modèle. Elles nous permettent de comprendre comment notre modèle classe les départements en fonction des deux variables explicatives clés que sont la population et le nombre pondéré d'infractions.

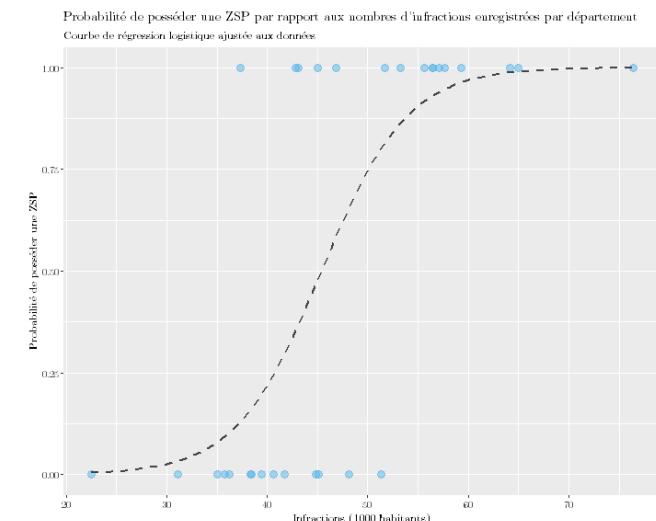
L'objectif est de vérifier si notre modèle peut discriminer efficacement entre les départements nécessitant des ZSP et ceux qui n'en ont pas besoin en se basant sur ces variables.

En interprétant ces courbes, nous pouvons également identifier les seuils de classification optimaux. Par exemple, nous pouvons déterminer à quel niveau de population ou de nombre pondéré d'infractions notre modèle prédit le mieux la nécessité de ZSP. Cela peut être particulièrement utile pour orienter les décisions politiques et les ressources vers les départements les plus susceptibles de bénéficier de ZSP.

##### *b. Calcul du ROC, du seuil optimal et de l'AUC*

Passons maintenant à l'évaluation de la performance de notre modèle, une étape importante dans notre analyse de la prédiction des Zones de Sécurité Prioritaires (ZSP) dans les départements français.

Afin d'évaluer notre modèle de régression logistique régularisée, on construit le ROC (Receiver Operating Characteristic) qui permet d'illustrer la manière dont notre modèle distingue les départements avec ZSP (1) ou sans ZSP (0). Plus précisément, il montre la relation entre le taux de faux positifs (FPR) et le taux de vrais positifs (TPR). Dans le cadre de cette étude, le FPR mesure le nombre de départements sans ZSP incorrectement classés, tandis que le TPR mesure combien de départements avec ZSP sont correctement classés. Un modèle idéal aurait un TPR de 1 et un FPR de 0.



L'évaluation du modèle continue avec le calcul de l'AUC (Area Under the Curve), un autre indicateur complémentaire d'estimation de la performance globale du modèle, qui quantifie l'aire sous la courbe ROC. Une AUC proche de 1 signifie que le modèle est excellent pour différencier les départements avec ou sans ZSP.

```
# Selecting the best model to fit
infraction_logreg_best_fit = predict(infraction_logreg_lasso_model,
                                      newx=as.matrix(model.matrix(model_formula, data=infraction_logreg_test)),
                                      s='lambda.1se', type='response')

# Building the ROC object and recording it into a list
roc_curve = roc(response=infraction_logreg_test$zsp_status, predictor=as.numeric(infraction_logreg_best_fit))

# Calculating the optimized threshold for the model selected
best_model_threshold = coords(roc_curve, 'best')$threshold
```

Dans notre cas, l'AUC s'élève à 0.95 tandis que l'intervalle de confiance de l'AUC se trouve entre 0.88 et 1.00. L'intervalle étant plutôt étroit, ceci suggère une estimation précise et une confirmation de la fiabilité de notre estimation de l'AUC.

Enfin, on peut également calculer le seuil optimal. Ce dernier détermine le point où le modèle équilibre judicieusement la classification des départements. Dans le cadre de cette étude, obtient un seul optimal d'environ 0.56. Cela signifie que lorsqu'une probabilité prédictive dépasse les 56%, le département est considéré comme nécessitant une ZSP.

Ces trois métriques permettent de tracer la courbe ROC. Cette courbe représente la capacité du modèle à distinguer les départements avec ZSP (1) ou sans ZSP (0). Elle est construite en traçant le taux de faux positifs (FPR) en fonction du taux de vrais positifs (TPR) à différents seuils de classification.

On observe graphiquement que la courbe se rapproche du coin supérieur gauche, où le TPR est élevé et le FPR est faible, indiquant une très bonne capacité de discrimination du modèle.

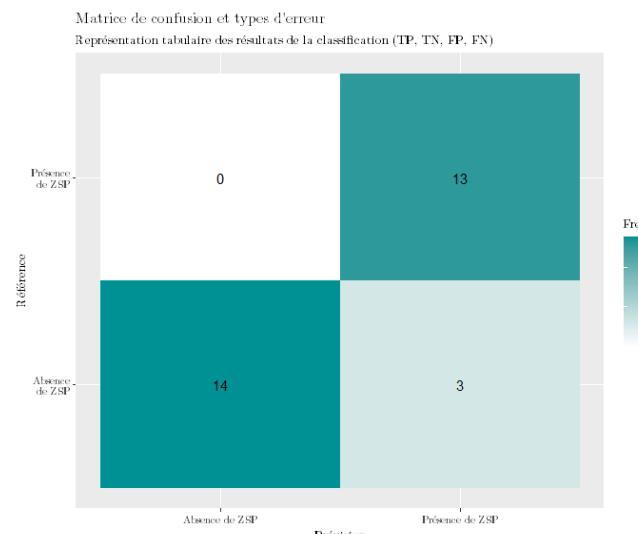
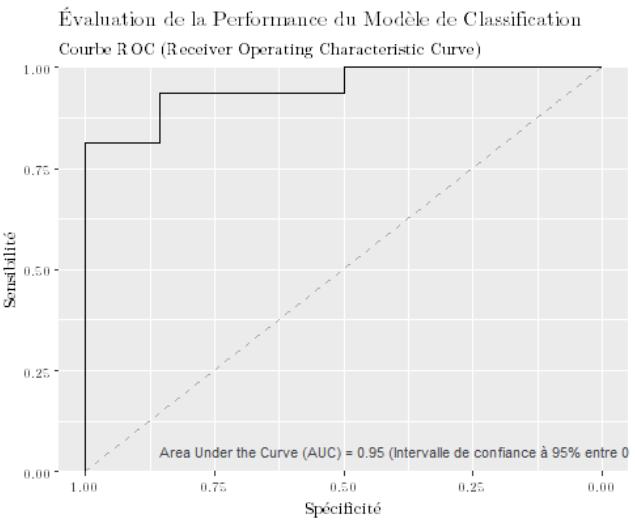
#### c. Matrice de confusion des résultats

La matrice de confusion est une représentation tabulaire des résultats de la classification effectuée par le modèle. Elle est constituée de quatre cellules qui résument la manière dont les observations ont été classées. Ces cellules sont les suivantes :

- Vrais positifs (VP) : Il s'agit du nombre d'observations positives réelles (départements avec au moins une ZSP) correctement classées comme positives par le modèle.
- Vrais négatifs (VN) : Il représente le nombre d'observations négatives réelles (départements sans ZSP) correctement classées comme négatives par le modèle.
- Faux positifs (FP) : Cela correspond au nombre d'observations négatives réelles classées à tort comme positives par le modèle. C'est une erreur de type I.
- Faux négatifs (FN) : Il s'agit du nombre d'observations positives réelles classées à tort comme négatives par le modèle. C'est une erreur de type II.

Sur la figure ci-contre, la matrice de confusion montre que sur les 30 départements de l'échantillon test, 27 ont été correctement classés (VP et VN), tandis que 3 ont été faussement identifiés comme ayant au moins une ZSP (FP). Le modèle a donc tendance à commettre des erreurs de type I (FP) plutôt que des erreurs de type II (FN). Il est donc peut-être préférable de privilégier une approche plus conservatrice lors de la prise de décision basée sur les prédictions du modèle.

Pour rappel, une erreur de type I se produit lorsque le modèle prédit à tort la présence d'une ZSP dans un département qui n'en a pas. Cela pourrait entraîner des ressources gaspillées, car des mesures de sécurité pourraient être prises inutilement dans ces départements. A contrario, une erreur de type II se produit lorsque le modèle ne parvient pas à détecter une ZSP dans un département qui en a effectivement une. Cela pourrait entraîner des conséquences graves en termes de sécurité publique, car des mesures nécessaires pourraient ne pas être prises.



Ainsi, la matrice de confusion nous permet de comprendre comment notre modèle se comporte en termes d'erreurs de classification, ce qui est essentiel pour une prise de décision éclairée dans le domaine de la sécurité publique.

Maintenant, pour répondre à la deuxième problématique soulevée, nous devrons choisir un modèle de régression approprié pour estimer l'effectif des forces de l'ordre nécessaires pour chaque département. Cette étape nécessitera une analyse distincte, car elle implique des variables et des paramètres différents de ceux de la première problématique.

*d. Précision, sensibilité et spécificité*

La précision, la sensibilité et la spécificité sont des métriques complémentaires à la bonne évaluation de la performance du modèle de régression logistique choisi. Voici les valeurs obtenues à partir du paramétrage décrit plus tôt :

La précision mesure la capacité du modèle à prédire correctement la classe des observations. Dans notre cas, un score de précision de 90% indique que le modèle a correctement classé 90% des observations de l'ensemble de test.

Accuracy	Sensitivity	Specificity
0.9000	1.0000	0.8125

Cela signifie que dans 90% des cas, le modèle a fait des prédictions exactes par rapport à la présence ou à l'absence de Zones de Sécurité Prioritaires (ZSP) dans les départements français.

La sensibilité, également appelée rappel, évalue la capacité du modèle à identifier les vrais positifs parmi les observations positives réelles. Avec un score de sensibilité de 100%, notre modèle a identifié toutes les observations positives réelles, ce qui signifie qu'il n'a pas manqué une seule ZSP parmi celles qui étaient réellement présentes. Cela montre que le modèle est très performant pour détecter les départements avec des ZSP.

La spécificité mesure la capacité du modèle à identifier correctement les vrais négatifs parmi les observations négatives réelles. Un score de spécificité de 81.25% indique que le modèle a correctement identifié 81.25% des départements sans ZSP parmi ceux qui n'en avaient effectivement pas. Cette valeur montre que le modèle a également une bonne capacité à exclure les départements qui ne sont pas des ZSP.

## V. Régression linéaire multiple pas-à-pas

Après avoir réalisé une analyse exploratoire approfondie pour comprendre les données observées et effectué une classification supervisée via une régression logistique pour déterminer la présence ou l'absence d'au moins une Zone de Sécurité Prioritaire (ZSP) dans chaque département français à partir des données d'infractions, cette section se penche sur l'estimation des effectifs nécessaires des forces de l'ordre pour chaque département. Compte tenu de la nature du problème et de la variable à expliquer, l'utilisation d'un modèle de régression semble être intuitivement la solution la plus appropriée.

### A. Gestion et arbitrage des valeurs aberrantes

L'identification et la gestion de possibles valeurs aberrantes, pouvant potentiellement influencer de manière significative les résultats d'une régression, représente la première étape de préparation du jeu de données. Grâce aux enseignements de l'analyse exploratoire, il semble que Paris ait un comportement atypique par rapport aux autres départements. De ce fait, la valeur est retirée du jeu de données car elle pourrait biaiser les résultats du modèle. De même, et à la lumière de leur influence respective, d'autres valeurs pourraient être considérées comme aberrantes (Bouches-du-Rhône, Haute-Marne). Toutefois, nous faisons le choix de les conserver dans les données étudiées afin de ne pas impacter la taille des échantillons.

Par ailleurs, et comme pour la classification supervisée, une scission du jeu de données en deux échantillons distincts est nécessaire avant de débuter les premiers travaux de modélisation. Un premier échantillon est dédié à l'entraînement du modèle, et l'autre à sa validation. Nous décidons d'opter pour un découpage classique de 80% pour l'entraînement et 20% pour la validation, conformément aux meilleures pratiques en modélisation. Cette division est cruciale pour garantir la fiabilité de notre modèle de régression.

### B. Choix du modèle de régression et contraintes à prendre en compte

Nous avons opté pour l'utilisation d'un modèle de régression linéaire multiple dans le cadre de notre analyse afin d'estimer les effectifs des forces de l'ordre nécessaires pour chaque département. Ce choix s'explique par la nécessité d'explorer les relations complexes entre les données d'infractions, de population, et les effectifs des forces de l'ordre.

Ce modèle présente des avantages significatifs. Tout d'abord, il permet une analyse multivariée en incluant plusieurs variables indépendantes dans le modèle, ce qui est essentiel pour comprendre l'impact simultané de plusieurs prédicteurs sur la variable dépendante. De plus, en utilisant un modèle de régression linéaire multiple, nous pouvons obtenir des estimations plus précises des coefficients de régression, améliorant ainsi la qualité de nos prévisions. Enfin, ce modèle nous permet d'explorer en profondeur les relations complexes entre les variables, aidant ainsi à comprendre les dynamiques sous-jacentes aux données.

Cependant, l'utilisation de la régression linéaire multiple comporte également des contraintes. Premièrement, elle suppose que la relation entre la variable dépendante (effectifs des forces de l'ordre) et les variables indépendantes est linéaire. Si cette hypothèse n'est pas satisfaite, les résultats peuvent être biaisés. Deuxièmement, cette méthode nécessite que les erreurs de prévision soient indépendantes les unes des autres pour assurer la validité des tests statistiques et des intervalles de confiance. Troisièmement, l'homoscédasticité est une hypothèse importante, ce qui signifie que la variance des erreurs doit être constante pour toutes les valeurs des variables indépendantes. Si cette condition n'est pas respectée, les estimations des coefficients peuvent être biaisées. De plus, les résidus du modèle doivent suivre une distribution normale pour que les tests statistiques soient valables. Enfin, il est crucial que les variables indépendantes ne présentent pas de forte corrélation entre elles, car cela peut rendre l'interprétation des coefficients difficile et provoquer des instabilités dans les estimations.

En conclusion, la régression linéaire multiple est un outil puissant pour explorer les relations complexes entre les variables, mais elle exige la vérification et le respect de certaines hypothèses essentielles pour garantir la validité des résultats.

## C. Vérification des hypothèses

### 1. Hypothèse d'absence de multicolinéarité et calcul du VIF

L'hypothèse d'absence de multicolinéarité est cruciale en régression linéaire multiple, car elle garantit que les variables indépendantes du modèle ne sont pas fortement corrélées entre elles, ce qui pourrait entraîner des problèmes d'instabilité dans les estimations des coefficients de régression.

Pour vérifier l'hypothèse d'absence de multicolinéarité, on peut calculer le Variance Inflation Factor (VIF), outil statistique utilisé pour évaluer la présence de multicolinéarité entre les variables indépendantes du modèle. Il mesure à quel point la variance d'un coefficient de régression est augmentée en raison de la corrélation entre cette variable indépendante et les autres. Plus précisément, le VIF calcule la proportion de la variance de l'estimateur du coefficient qui est due à la multicolinéarité. Un VIF élevé (généralement supérieur à 5 ou 10) suggère une forte corrélation entre la variable en question et les autres variables indépendantes, ce qui peut entraîner une instabilité dans les estimations des coefficients de régression.

Dans notre étude, les résultats du calcul du VIF pour chaque variable indépendante sont les suivants (avec un seed de 20230831) :

Population	Statut ZSP	Catégorie 1	Catégorie 2	Catégorie 3	Catégorie 4	Catégorie 5	Catégorie 6
3.452621	2.219459	5.885812	4.076989	2.785807	4.992963	3.968803	2.617469

Un VIF de 3.45 pour la variable "population" indique que la variance de l'estimateur du coefficient de cette variable est augmentée de 3.45 fois en raison de la multicolinéarité. De manière générale, des VIF inférieurs à 5 sont considérés comme acceptables, ce qui signifie que la multicolinéarité n'affecte pas de manière significative les estimations des coefficients.

En conclusion, l'absence de multicolinéarité est essentielle pour garantir la stabilité et la validité des estimations dans un modèle de régression linéaire multiple. Le calcul du VIF permet d'identifier les variables qui pourraient poser problème en raison de leur

corrélation avec d'autres, ce qui permet de prendre des mesures correctives si nécessaire. Dans notre étude, les VIF obtenus sont globalement acceptables, ce qui suggère que la multicolinéarité n'est pas un problème majeur pour notre modèle.

## 2. Hypothèse de contribution globale des variables explicatives

L'hypothèse de contribution globale des variables explicatives est fondamentale en régression linéaire multiple, car elle nous permet de déterminer si l'ensemble des variables indépendantes du modèle explique de manière significative la variation de la variable dépendante. En d'autres termes, elle vise à évaluer si le modèle linéaire multiple est globalement pertinent pour notre analyse.

Un Test de Fisher peut être utilisé pour cette évaluation. Il repose sur l'idée que si les coefficients de régression de toutes les variables explicatives sont nuls, alors le modèle n'est pas pertinent et la régression n'explique pas la variation de la variable dépendante. A l'inverse, si au moins un de ces coefficients est différent de zéro, alors le modèle est pertinent. En d'autres termes, en comparant un modèle constant avec un modèle complet et on cherche à rejeter l'hypothèse nulle  $H_0$  au profit de l'hypothèse de contribution  $H_1$  :

- Hypothèse nulle  $H_0$  à partir du modèle constant :  $Y_i = \alpha + \epsilon_i$  où  $\epsilon_i \sim N(0, \sigma^2)$
- Hypothèse non-nulle  $H_1$  à partir du modèle complet :  $Y_i = \beta + \sum_{j=1}^p \alpha_j x_{i,j} + \epsilon_i$  où  $\epsilon_i \sim N(0, \sigma^2)$

La p-value calculée à partir de ce test bidirectionnel (0.0004998) se trouve être bien en dessous d'un seuil de signification courant des 5%. Cela signifie que nous avons suffisamment de preuves pour rejeter l'hypothèse nulle selon laquelle tous les coefficients de régression sont nuls. En d'autres termes, le modèle dans son ensemble est statistiquement significatif et explique de manière globale la variation de la variable dépendante. Il est donc justifié de poursuivre notre analyse avec ce modèle.

## 3. Hypothèse d'homoscédasticité

L'hypothèse d'homoscédasticité est un concept important en régression linéaire multiple, car elle concerne la variabilité constante des résidus (les erreurs de prédiction) sur l'ensemble de la plage des valeurs prédictes (fitted values) de la variable dépendante. En d'autres termes, l'homoscédasticité suppose que la dispersion des résidus reste constante et ne change pas en fonction des niveaux de la variable indépendante.

Le nuage de points des résidus en fonctions des valeurs prédictes est une méthode graphique couramment utilisée pour vérifier l'hypothèse d'homoscédasticité. Sur ce graphique, les résidus (c'est-à-dire les erreurs de prédiction) sont représentés en fonction des valeurs prédictes de la variable dépendante.

L'objectif est d'observer si la dispersion des résidus est constante le long de la plage des valeurs prédictes.

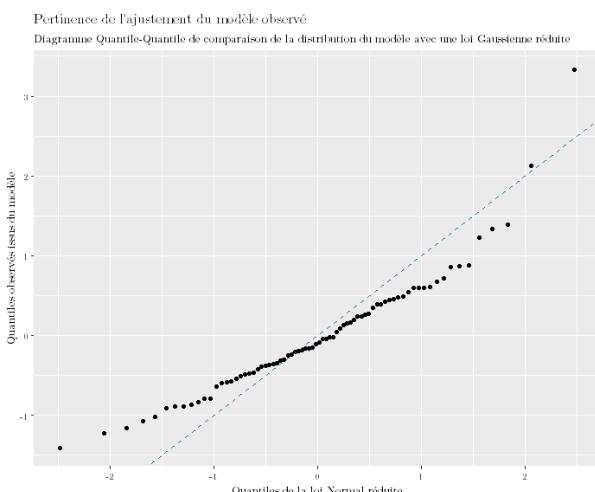
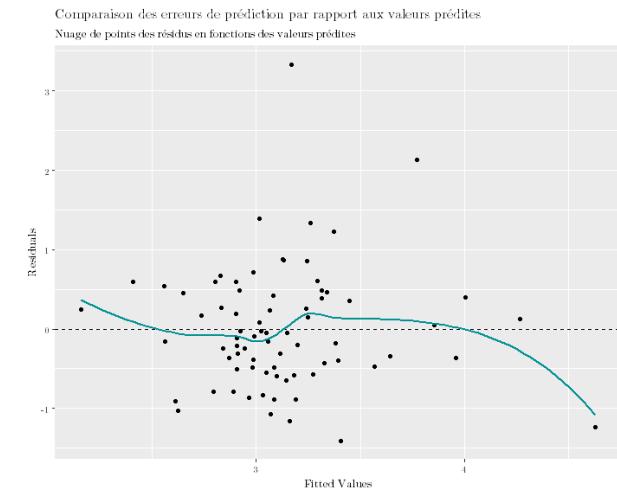
Dans le cas idéal, les points sur le graphe seraient répartis de manière aléatoire autour de la ligne horizontale zéro, sans tendance particulière à s'écartez de cette ligne. Cela indiquerait une homoscédasticité parfaite, ce qui signifie que la variance des résidus est constante à tous les niveaux de prédiction.

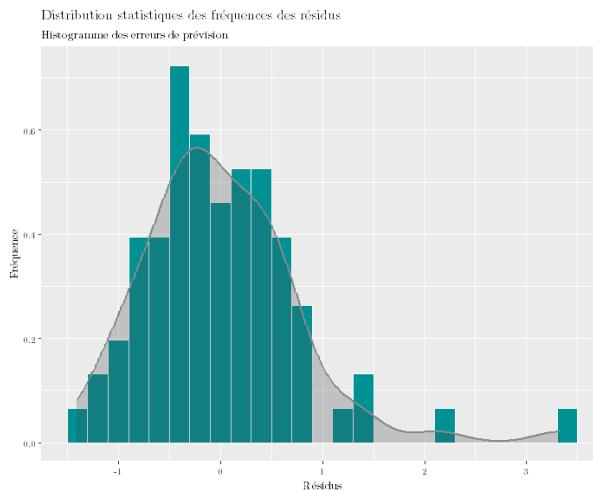
Cependant, dans la pratique, de légères déviations par rapport à la droite de régression peuvent être observées. C'est d'ailleurs le cas pour notre étude où l'on retrouve un léger décrochage aux extrémités de la droite. Ce phénomène s'explique par la présence de quelques valeurs inédites (Bouches-du-Rhône, Haute-Marne) qui n'ont pas été retirées du jeu de données par souci d'arbitrage entre précision du modèle et gestion des outliers. Cela influence très probablement un peu la dispersion des résidus. Toutefois, on considère que cette situation est acceptable tant que les déviations restent relativement faibles et que la tendance générale des résidus à rester autour de la ligne horizontale zéro est respectée. On peut donc valider l'hypothèse pour laquelle la variance des résidus est constante.

## 4. Hypothèse de normalité des erreurs

Enfin, la dernière hypothèse à vérifier dans l'optique de réaliser une régression linéaire est celle de normalité des erreurs. Elle stipule que les résidus du modèle de régression doivent suivre une distribution normale, c'est-à-dire une distribution en forme de cloche symétrique. Cette hypothèse est importante car de nombreux tests et intervalles de confiance en régression linéaire sont basés sur cette supposition de normalité.

Pour la vérifier, on peut utiliser plusieurs outils graphiques, dont le diagramme Quantile-Quantile de comparaison de la distribution du modèle avec une loi Gaussienne réduite. Ce graphique compare les quantiles des résidus aux quantiles d'une distribution normale théorique. Dans un scénario idéal, les points sur le graphe suivraient une ligne droite, signifiant que les résidus suivent une distribution normale.





En pratique, il est cependant assez courant d'observer de légères déviations de cette ligne droite aux extrémités. Ces déviations peuvent être causées par la présence de valeurs à forte influence dans les données.

Dans le cas de cette étude, et comme expliqué précédemment, nous sommes conscients que certains départements présentent des caractéristiques atypiques en termes de forces de l'ordre par rapport à leur population, ce qui peut influencer les résidus.

Un second outil de validation graphique, l'histogramme des fréquences d'apparition des résidus, peut également permettre de vérifier l'hypothèse de normalité.

En parcourant cet histogramme, on s'attend à voir une forme de cloche symétrique. Bien que l'histogramme puisse présenter une légère asymétrie à gauche, la forme globale se rapproche significativement d'une distribution gaussienne, vérifiant ainsi l'hypothèse de normalité.

## D. Exécution d'une régression linéaire multiple pas-à-pas

### 1. Principes de la méthode pas-à-pas

La régression linéaire pas-à-pas est une méthode qui vise à sélectionner de manière itérative les variables explicatives les plus pertinentes pour un modèle de régression linéaire multiple. À chaque itération, cette méthode sélectionne la variable qui contribue le plus à améliorer les performances du modèle en termes de minimisation de l'erreur (RMSE) ou de maximisation du coefficient de détermination (R-squared). Cette sélection itérative continue jusqu'à ce qu'aucune variable supplémentaire ne puisse améliorer davantage le modèle. On exécute cette régression à partir du code suivant :

```
# Setting up repeated k-fold cross-validation
train_control = trainControl(method='cv', number=10, seed=list(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13))
# Performing stepwise linear regression on multiple continuous variables
linreg_stepwise_model = train(police_workforce ~ ., data=infraction_linreg_train, # Training the model
                               method='leapSeq', # Using stepwise method
                               trControl=train_control, # Adding cross-validation parameter
                               trace=FALSE)
```

Par ailleurs, l'utilisation de la validation croisée, effectuée ici avec la fonction trainControl(), permet de s'assurer que le modèle est généralisable à de nouvelles données et évite le surajustement (overfitting). Elle consiste à diviser les données en plusieurs sous-ensembles (ou plis) pour évaluer la performance du modèle de régression sur des données non utilisées lors de son entraînement.

### 2. Minimisation du RMSE

On peut ensuite afficher les résultats de la régression linéaire pas-à-pas pour les différentes itérations réalisées, en fonction du nombre de variables explicatives incluses, avec l'objectif de minimiser le RMSE<sup>(2)</sup>.

nvmax	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
2	0.8368751	0.1732904	0.6524009	0.2917468	0.1353521	0.1667989
3	0.8693129	0.1487930	0.6714584	0.3031154	0.1387439	0.1494091
4	0.8128608	0.1150403	0.6468927	0.2447522	0.1005600	0.1153833

Dans ce tableau, la colonne « nvmax » fait référence au nombre de variables explicatives incluses dans le modèle à chaque itération. D'autres mesures de performance telles que le R-squared (coefficients de détermination) et le MAE (Mean Absolute Error) sont également affichées.

### 3. Sélection des variables retenues

On constate que la troisième itération, et l'entrée de la quatrième variable dans le modèle, permet de minimiser le RMSE. Jetons maintenant un œil aux variables retenues :

Iteration	Population	Statut ZSP	Catégorie 1	Catégorie 2	Catégorie 3	Catégorie 4	Catégorie 5	Catégorie 6
1	**	**	**	**	***	**	**	**
2	***	**	**	**	**	***	**	**
3	***	**	***	**	***	**	**	**
4	***	***	***	***	**	**	**	**

Les variables finalement retenues par le modèle sont les suivantes : le nombre d'habitants (Population), la présence ou non d'au moins une ZSP (Statut ZSP), le nombre de plaintes pour atteintes aux biens (Catégorie 1) ainsi que le nombre de plaintes pour atteintes aux personnes (Catégorie 2). Leurs coefficients associés sont les suivants :

(Intercept)	Population	Statut ZSP	Catégorie 1	Catégorie 2
1.816580	-5.147255e-07	-9.596441e-02	1.685883e-01	6.682653e-03

(2) Pour rappel, le RMSE (Root Mean Square Error) est une mesure de la précision du modèle de régression. Il quantifie l'écart moyen entre les valeurs prédites par le modèle et les valeurs réelles de la variable cible. Un RMSE plus faible indique un modèle de régression plus précis.

#### 4. Détermination et interprétation de la fonction de régression

Ces coefficients permettent d'interpréter le lien entre les variables explicatives (population, présence de ZSP et catégories d'infractions) et la variable de réponse (nombre d'agents des forces de l'ordre). Ils sont tous relativement petits car les variables associées sont très grandes. On peut alors décider de modifier les échelles afin d'améliorer l'interprétation et la compréhension de la fonction de régression.

Variables	Valeur unitaire initiale	Transformation	Valeur unitaire transformée
(Intercept)	1,816580	-	1,816580
Population	0,0000005147255	Pour un million d'habitant	0,5147255
Statut ZSP	0,09596441	-	0,09596441
Infractions de catégorie 1 (pour 1000 habitants)	0,1685883	Pour un million d'habitant	168,5883
Infractions de catégorie 2 (pour 1000 habitants)	0,006682653	Pour un million d'habitant	6,682653

On peut alors écrire la fonction de régression du modèle choisi de manière suivante :

$$\{Nombre\ d'agents\ des\ forces\ de\ l'ordre\ pour\ mille\ habitants\} = 1,81658 - 0,5147255 * \{Population\ en\ millions\ d'habitants\} - 0,09596441 * \{Statut\ ZSP\} + 168,5883 * \{Infractions\ de\ catégorie\ 1\ pour\ un\ million\ d'habitants\} + 6,682653 * \{Infractions\ de\ catégorie\ 2\ pour\ un\ million\ d'habitants\}$$

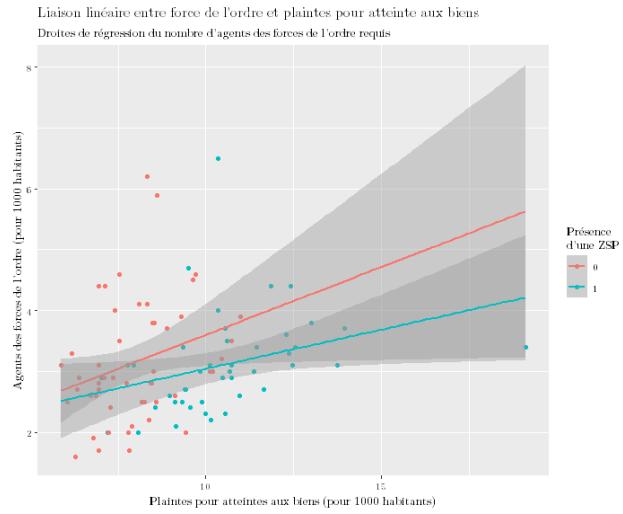
Ce modèle nous permet de quantifier l'impact de chacune de ces variables sur le nombre d'agents des forces de l'ordre pour mille habitants.

Paradoxalement, on observe que la population et la présence d'au moins une ZSP ont toutes deux une influence négative sur le nombre de force de l'ordre. A contrario, le nombre de plaintes déposées pour des infractions de catégorie 1 ou 2 exerce une influence positive importante sur le nombre de forces de l'ordre.

Par exemple, une augmentation d'une unité du nombre de plaintes pour atteintes aux biens pour un million d'habitant est associée à une augmentation de 168,5883 unités du nombre d'agents des forces de l'ordre, toutes choses étant égales par ailleurs. A ce titre, c'est cette variable qui influence le plus la réponse.

Pour mieux s'en rendre compte, visualisons le lien entre nombre de plaintes pour atteintes aux biens et nombre d'agents des forces de l'ordre.

Ce graphique permet d'observer visuellement la relation linéaire entre ces deux variables, tout en tenant compte de la présence ou non d'au moins une ZSP dans les départements français métropolitains. On y voit également l'influence de la présence de ZSP sur la réponse.



## VI. Conclusion de l'étude

La France fait face à un calendrier sécuritaire chargé en termes d'événements internationaux, tels que la Coupe du Monde de Rugby à la rentrée 2023 ou encore les Jeux Olympiques de 2024. C'est dans ce contexte particulièrement inédit que cette analyse fournit des informations cruciales pour les autorités chargées de la sécurité lors de ces événements, leur permettant d'ajuster leurs plans en conséquence et de garantir la sécurité des participants et du public.

L'analyse exploratoire, la classification par régression logistique puis la régression linéaire multiple pas-à-pas ont eu pour objectif de déterminer la présence ou l'absence de Zones de Sécurité Prioritaires (ZSP) pour chaque département et d'établir une estimation de l'effectif des forces de l'ordre nécessaires pour chaque département français métropolitain. Ces résultats offrent des informations précieuses pour les pouvoirs publics concernant l'allocation des ressources humaines pour renforcer la sécurité.

La présence d'au moins une Zone de Sécurité Prioritaire (ZSP) a été identifiée comme un facteur significatif influençant le nombre d'agents des forces de l'ordre nécessaires. Les départements avec au moins une ZSP ont tendance à nécessiter un nombre plus élevé d'agents pour maintenir la sécurité publique. Cela peut être dû à une concentration plus élevée de crimes ou délits dans ces zones, nécessitant une réponse policière plus importante.

De même, les types d'infractions ont également montré une influence significative. Par exemple, les atteintes aux biens et aux personnes ont été identifiées comme des facteurs importants dans la détermination du nombre d'agents nécessaires. Les politiques publiques pourraient donc envisager d'adapter le recrutement des forces de l'ordre en fonction de la nature des infractions prédominantes dans chaque département.

Cette analyse offre donc une perspective complémentaire aux autorités chargées de la sécurité en se fondant sur une compréhension statistique des liens entre ZSP, niveaux d'infractions et ressources policières.