|  |
| --- |
| DéSigéo 2021 - Statistiques |
| ANALYSE STATISTIQUE DES CRIMES EN FRANCE |
|  |

|  |
| --- |
| David Delord  Antoine Duruflé  Sébastien Quinot |
| 03/01/2021 |

Table des matières

[Introduction 1](#_heading=h.gjdgxs)

[I. Préparation des données 2](#_heading=h.30j0zll)

[II. Premières analyses 8](#_heading=h.1fob9te)

[III. Etude statistique des relations causes-conséquences 12](#_heading=h.3znysh7)

[Conclusion de l’étude 18](#_heading=h.2et92p0)

# Introduction

Le film *Minority report* (2002) de Steven Spielberg inspiré de l’ouvrage du même nom (1956) de Philip K. Dick présentent un futur dans lequel des êtres humains mutants peuvent prédire des crimes grâce à des dons de précognition. Ces œuvres posent la question de déterminisme. Le futur est-il écrit, et par extension, sommes-nous prédestinés dans les infractions que l’on commet ?

Sans aller aussi loin dans la science-fiction et la philosophie, nous nous sommes intéressés aux relations pouvant exister entre les infractions commises et un certain nombre de facteurs sociaux, démographiques, géographiques.

Quels facteurs influencent le nombre d’infractions **recensées** ?

L’influence sera décrite ici comme la relation statistique pouvant exister entre deux variables. Le nombre d’infractions est quant à lui celui des infractions ayant fait l’objet d’un procès-verbal.

Nous verrons dans un premier temps l’origine de nos données et la manière dont elles ont été « assemblées ». Dans un deuxième temps nous rechercherons aux seins des causes et des conséquences séparées de potentielles relations. Enfin, nous ferons une analyse fine, selon différents niveaux de granularité, des relations pouvant exister entre causes et conséquences.

# Préparation des données

Conformément au traitement des données vu en cours de *Business Intelligence* avec Mme Atigui (CNAM), nous avons suivi une phase ETL (*Extract, Transform, Load* : extraction, transformation, chargement) afin de préparer les données à leur futur traitement dans le cadre d’une analyse statistique.

En effet, les sources de données sont d’origines variées et parfois sous des formes différentes et nécessitent donc un traitement de mise en forme maîtrisé afin de conserver la qualité d’origine et favoriser alors une exploitation optimale. Dans ce sens, l’uniformisation des données (format identique) est également apparue primordiale pour permettre un usage aisé desdites données.

\*

1. Extraction des données (*Extract*) :

Le but que nous nous étions assignés était de rechercher d’éventuels liens ou corrélations entre les actes criminels et des données sur la population, telles que le revenu médian, le niveau d’étude ou le taux de pauvreté. Nous voulions donc croiser ces deux types de données suivant un découpage territorial et/ou de responsabilité administrative.

Tout d’abord, le choix des données a reposé sur deux facteurs principaux : leur disponibilité et leur véracité. En effet, bien que n’ayant pas d’objectif de publication ou de communication particulière, nous avons souhaité exploiter des données consolidées et validées. Par ailleurs, ces données devaient être accessibles sans difficulté et gratuitement. Dès lors, nous nous sommes intéressés aux données en libre accès fournies par des organismes d’autorité. Rapidement, nous nous sommes orientés vers le site data.gouv.fr . Ce site gouvernemental offre de nombreuses données, avec la garantie d’un contrôle réalisé par l'État, et ainsi porte un gage de qualité. Nous y avons trouvé les données concernant les actes criminels, par commissariats et gendarmeries ainsi que par tribunaux de grande instance. Par ailleurs, nous avons également consulté le site de l’INSEE : institut national de la statistique et des études économiques. Y sont disponibles les données récoltées grâce aux différents organismes de l’Etat, notamment concernant les ressources financières, le niveau de formation ou le taux de pauvreté.

Nous avons également opté pour l’utilisation des données de 2017 pour deux raisons principales : elles étaient disponibles pour cette année-là sur les deux sites visités et semblaient être les plus complètes.

Nous disposions alors d’une part de données concernant les actes criminels recensés pour les différents commissariats, brigades de gendarmerie et d’autre part pour les tribunaux de grande instance, à leurs échelles géographiques respectives (découpages administratifs particuliers) et suivant 107 libellés distincts de crimes et délits. Par ailleurs, nous avions obtenu des données pour chaque commune française concernant leur surface, leur population, leur revenu médian, le niveau scolaire et le taux de pauvreté (revenu médian des plus pauvres et leur intensité au seuil de 60%).

Fort de ces données, une première orientation est apparue concernant l’échelle de travail. Nous avons opté pour trois niveaux de travail sur les données obtenues : département, gendarmeries/commissariats et tribunaux de grande instance.

1. Traitement des données (Transform) :

Ces différents jeux de données n’étaient pas immédiatement compatibles et exploitables. Un travail de traitement s’est avéré nécessaire pour disposer de données sous un même format, aux échelles souhaitées et disposant de clés étrangères pour identifier les correspondances et permettre des agrégations et les opérations et analyses à venir.

Nous disposions initialement de fichiers de type CSV, EXCEL et shapefiles et la transformation de ces divers fichiers vers un format unique a dû être réalisée en tout premier lieu. Le choix du format CSV a été fait pour permettre une lecture et exploitation facilitées par les différents outils que nous pourrions avoir a employé : PostgreSQL (travail sur les bases de données), QGis (travail sur les bases de données et visualisation des données) et R (travail sur les données et étude statistique). Suivant les opérateurs et leur maîtrise des divers outils, cette première transformation a été réalisée avec PostgreSQL, R, voire EXCEL. En effet, la quantité de données limitée a permis l’emploi de tous les outils sans distinction de performance évidente.

Ensuite, nous avons cherché à lier les fichiers détenus : la base INSEE vers la base commissariat/brigade de gendarmerie et la base INSEE vers la base tribunal de grande instance. Nous avons donc cherché une clé permettant de lier ces bases. La première orientation fut d’employer le nom de chaque commune et nous avons rendu ceux-ci plus employables en retirant les caractères spéciaux (accents, cédilles, traits d’union, etc.). En effet, l’usage de ces caractères, communs en français, est une réelle difficulté lorsque nous souhaitons exploiter les données parce que les différents outils ne disposent pas nécessairement du même codage. Cet usage perturbe donc l’accès aux données et leur manipulation. Après avoir réalisé ces transformations, avec PostgreSQL et R, nous avons constaté que cette méthode n’était pas viable et ne satisfaisait pas nos attentes. En effet, certains noms devenaient doublons perdant ainsi leur unicité et donc leur capacité à être clé.

Nous avons ensuite employé une autre méthode en nous appuyant sur les clés naturelles de chaque base : le CODGEO (code INSEE) pour les données INSEE, ou en créant des clés pour les autres bases : id\_service pour les commissariats/brigade de gendarmeries et id pour les tribunaux de grande instance. Nous avons également dû effectuer une inversion des colonnes/lignes pour ces deux dernières afin de permettre les futures agrégations de ces bases (l’identification étant faite par colonne pour les commissariats et tribunaux et par lignes pour les données INSEE). Ensuite, nous avons créé deux tables de correspondance qui permettent de lier d’une part les données INSEE aux commissariats/gendarmeries et d’autre part les données INSEE aux tribunaux de grande instance. Pour ce faire, nous avons employé l’outil d’intersections (INTERSECTS ()) de PostGIS sur PostgreSQL qui permet d’identifier les entités qui intersectent une tierce entité, grâce à leurs géométries respectives. De là nous avons pu établir les deux tables de correspondance souhaitées qui nous ont permis de réaliser l’analyse des données.

Consécutivement, nous avons cherché à simplifier les données que nous avions sélectionnées. En effet, les tables recensant les crimes et délits représentaient 107 champs distincts soit un nombre trop important de critères à traiter. De plus, cette diversité créait une répartition trop importante des données et rendait l’analyse particulièrement incertaine et sans doute complexe. Nous avons alors opté pour une agrégation des données par grands groupes de crimes et délits et avons défini 8 « familles » : violence, destruction - dégradation, homicides, vols, viols - crimes sexuels, stupéfiants, falsification - escroquerie - fraude et autres. Plus loin, nous avons également traité les informations concernant le niveau de diplômes. Initialement distinguées suivant 7 catégories (sans diplôme, BEPC, CAP – BEP, BAC, BAC +2, BAC + 3/4, BAC + 5 et plus), nous avons agrégé ces données suivant trois nouvelles catégories : sans diplôme, BAC et inférieur, études supérieures. Cette opération nous a permis d’obtenir des données moins fines mais plus exploitables. Le cas échéant, nous serons en mesure de les désagréger pour en faire une exploitation plus précise en fonction des résultats obtenus avec ces premiers jeux de données.

Détenant dorénavant des tables compatibles *via* des tables de correspondance adaptées, permettant leur jointure, nous avons pu ensuite finaliser les données pour leur exploitation dans le cadre de l’étude statistique désirée.

1. Chargement des données (Load) :

Après avoir construit des tables exploitables concernant les données initialement acquises, nous avons constitué des tables regroupant les différentes données que nous voulions mettre en regard. Nous avons réalisé quelques opérations supplémentaires d’agrégation pour permettre un emploi simplifié de ces bases.

L’utilisation des tables de correspondance nous a donc permis de réaliser rapidement et efficacement les jointures souhaitées et adaptées aux travaux d’analyse que nous voulions réalisés. Nous avons donc constitué trois tables, à savoir :

* une table données INSEE et départements,
* une table données INSEE et commissariats/brigades de gendarmerie ,
* une table données INSEE et tribunal de grande instance.

Après cette première étape, nous avons pu agréger les données suivant les 3 niveaux souhaités : agrégation par département, agrégation par commissariat/brigade de gendarmerie, agrégation par tribunal de grande instance.

La préparation de ces données a été cruciale pour leur future exploitation. Elle a également permis la prise en main et la connaissance fine des informations contenues dans celles-ci.

Une première remarque importante a déjà pu être formulée quant à la qualité des données exploitées. Bien que semblant conformes aux attentes, nous avons constaté que de nombreuses données étaient lacunaires. En effet, il est apparu très clairement que des zones entières ne disposaient pas de données concernant les crimes et délits disponibles pour la période souhaitée. Toutefois, cette constatation n’a pas remis en cause notre projet et nous avons admis pour la suite de l’analyse que ces manques ne perturberaient pas l’exploitation prévue des données.

\*

Après cette première étape ETL essentielle visant la préparation des données, nous avons pu constituer les trois bases de données que nous souhaitions en vue des analyses à venir. Ce travail nous a notamment permis de constater les difficultés inhérentes à cette phase initiale et d’identifier les premiers écueils auxquels nous avons été confrontés : disponibilité des données, complétude des données, format des données ou encore codage de celles-ci.

Extrait 1 : base de données crimes et délits par gendarmerie

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Code index** | **Libellé index \ CGD** | **CGD BELLEY** | **CGD BOURG EN BRESSE** | **CGD GEX** | **CGD TREVOUX** | **CGD CHATEAU THIERRY NOGENTEL** | **Etc.** |
| **1** | Règlements de compte entre malfaiteurs | 0 | 0 | 0 | 0 | 0 | … |
| **2** | Homicides pour voler et à l'occasion de vols | 0 | 0 | 0 | 0 | 0 | … |
| **3** | Homicides pour d'autres motifs | 0 | 1 | 1 | 0 | 0 | … |
| **4** | Tentatives d'homicides pour voler et à l'occasion de vols | 0 | 0 | 0 | 0 | 0 | … |
| **5** | Tentatives homicides pour d'autres motifs | 2 | 1 | 0 | 4 | 0 | … |
| **6** | Coups et blessures volontaires suivis de mort | 0 | 0 | 1 | 0 | 0 | … |
| **7** | Autres coups et blessures volontaires criminels ou correctionnels | 259 | 301 | 312 | 391 | 140 | … |
| **8** | Prises d'otages à l'occasion de vols | 0 | 1 | 1 | 0 | 0 | … |
| **9** | Prises d'otages dans un autre but | 0 | 0 | 0 | 1 | 0 | … |
| **10** | Séquestrations | 4 | 2 | 10 | 6 | 2 | … |
| **Etc.** | … | … | … | … | … | … | … |

Extrait 2 : base de données INSEE

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **id** | **CODGEO** | **nom** | **surf\_ha** | **nb\_menage** | **nb\_pers** | **revenu\_med** | **nb\_dipl\_mi** | **Etc.** |
| **11084** | 85002 | L'Aiguillon-sur-Vie | 2333 | 893 | 2039 | 19920 | 354 | … |
| **24460** | 35331 | Talensac | 2167 | 989 | 2551 | 22300 | 321 | … |
| **13829** | 67169 | Gries | 1222 | 1155 | 2885 | 25700 | 484 | … |
| **14093** | 29137 | Logonna-Daoulas | 1242 | 948 | 2236 | 23990 | 150 | … |
| **27037** | 35014 | Bais | 3544 | 906 | 2286 | 19720 | 457 | … |
| **30872** | 76296 | Gainneville | 461 | 959 | 2570 | 25180 | 383 | … |
| **227** | 95504 | Presles | 1020 | 1497 | 3842 | 30360 | 448 | … |
| **12318** | 44024 | Brains | 1550 | 1088 | 2937 | 23640 | 214 | … |
| **11715** | 86178 | Nieuil-l'Espoir | 2113 | 1011 | 2609 | 21930 | 374 | … |
| **7081** | 29181 | Plouédern | 1962 | 1179 | 2978 | 22590 | 291 | … |
| **Etc.** | … | … | … | … | … | … | … | … |

 Extrait 3 : table de correspondance INSEE – commissariats/brigades de gendarmerie

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CODGEO** | **libelle\_commune** | **institution** | **id\_service** | **service** | **code\_postal** | **codes\_postaux** | **Etc.** |
| **1001** | L'Abergement-Clémenciat | GN | 1008574 | Gendarmerie - Brigade de Châtillon-sur-Chalaronne | 1400 | 1400 | … |
| **1002** | L'Abergement-de-Varey | GN | 1008601 | Gendarmerie - Brigade de Saint-Rambert-en-Bugey | 1640 | 1640 | … |
| **1004** | Ambérieu-en-Bugey | GN | 1008600 | Gendarmerie - Brigade d'Ambérieu-en-Bugey | 1500 | 1500 | … |
| **1005** | Ambérieux-en-Dombes | GN | 1008634 | Gendarmerie - Brigade de Villars-les-Dombes | 1330 | 1330 | … |
| **1006** | Ambléon | GN | 1008593 | Gendarmerie - Brigade de Belley | 1300 | 1300 | … |
| **1007** | Ambronay | GN | 1008600 | Gendarmerie - Brigade d'Ambérieu-en-Bugey | 1500 | 1500 | … |
| **1008** | Ambutrix | GN | 1008603 | Gendarmerie - Brigade de Lagnieu | 1500 | 1500 | … |
| **1009** | Andert-et-Condon | GN | 1008598 | Gendarmerie - Brigade de Virieu-le-Grand | 1300 | 1300 | … |
| **1010** | Anglefort | GN | 1008596 | Gendarmerie - Brigade de Culoz | 1350 | 1350 | … |
| **Etc.** | … | … | … | … | … | … | … |

 Extrait 4 : table commissariat/brigade de gendarmerie - crimes et délits - INSEE

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **id** | **service** | **id\_service** | **violence** | **destruction**  **dégradation** | **Etc.** | | | **pop** | **ss\_dipl** | **Etc.** | | **rev\_median\_pauv** |
| **326** | Brigade de proximité de Rodez | 271 | 910 | 570 |  |  |  | 15140 | 1851 |  |  | 9630 |
| **429** | Brigade territoriale autonome de Chambéry | 2571 | 2114 | 2170 |  |  |  | 9295 | 933 |  |  | 9550 |
| **287** | Brigade de proximité de Villefranche-de-Lauragais | 957 | 7476 | 5922 |  |  |  | 13938 | 1776 |  |  | 9500 |
| **472** | Brigade de proximité de Pornic | 3498 | 657 | 624 |  |  |  | 19489 | 2975 |  |  | 9395 |
| **43** | Commissariat de police de Pontarlier | 1016 | 242 | 202 |  |  |  | 20126 | 3898 |  |  | 9010 |
| **475** | Brigade de proximité de Castelsarrasin | 3224 | 1074 | 894 |  |  |  | 7647 | 1365 |  |  | 8950 |
| **49** | Brigade territoriale autonome de Chamonix-Mont-Blanc | 3105 | 645 | 411 |  |  |  | 12239 | 1320 |  |  | 8870 |
| **370** | Brigade de proximité de Dole | 1255 | 3234 | 2128 |  |  |  | 11162 | 1738 |  |  | 8790 |
| **566** | Brigade territoriale autonome de Montbrison | 1347 | 10089 | 5282 |  |  |  | 34118 | 5508 |  |  | 8760 |
| **400** | Brigade territoriale autonome de Chantilly | 2102 | 1644 | 1902 |  |  |  | 34609 | 4649 |  |  | 8727 |

# Premières analyses

Suite aux traitements de nos données, nous avons dans un premier temps étudié les jeux séparément de manière à en comprendre les relations et à les simplifier si nécessaire.

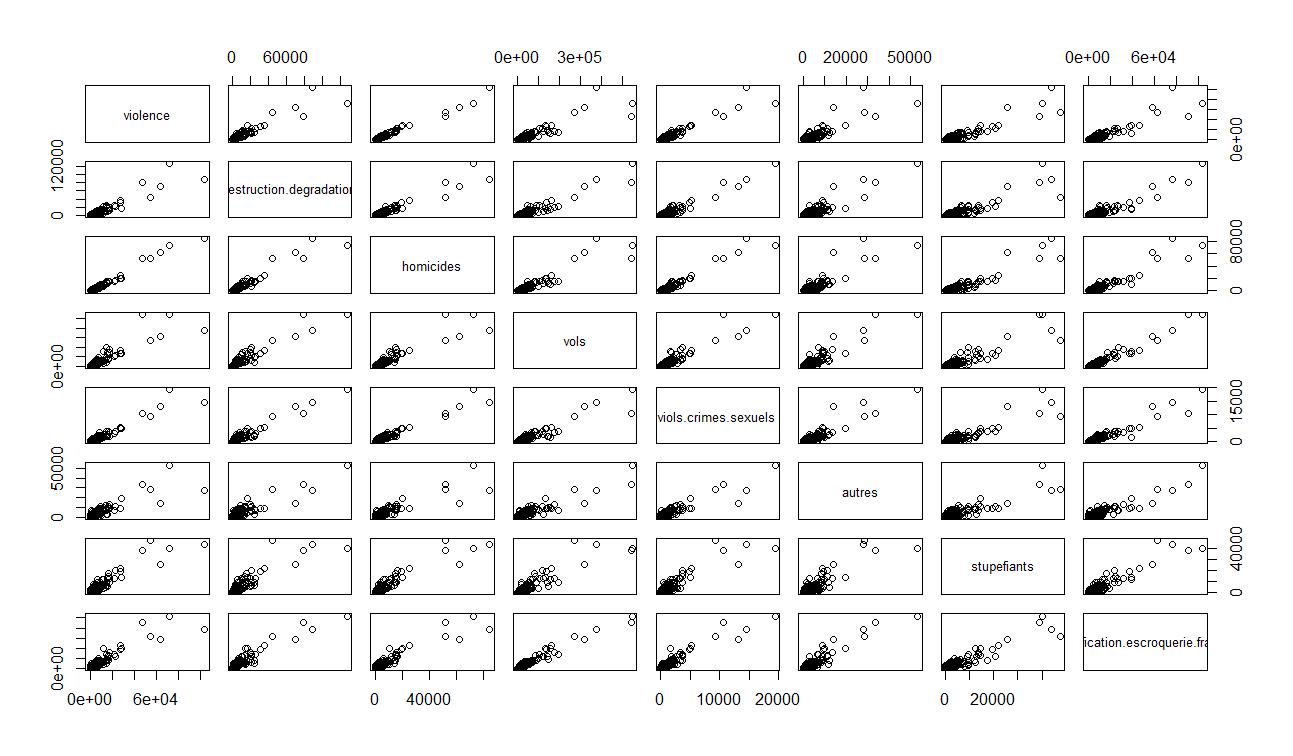
1. Infractions

Comme expliqué précédemment, à partir des données extraites de 107 types d’infractions différentes nous avons agrégé celles-ci en huit classes différentes :

violence, destruction - dégradation, homicides, vols, viols - crimes sexuels, stupéfiants, falsification - escroquerie - fraude et autres.

Ces classes ont été constituées en sommant les quantités d’infractions qui nous semblaient les plus proches.

Une première représentation graphique de ces données nous donne le résultat suivant :



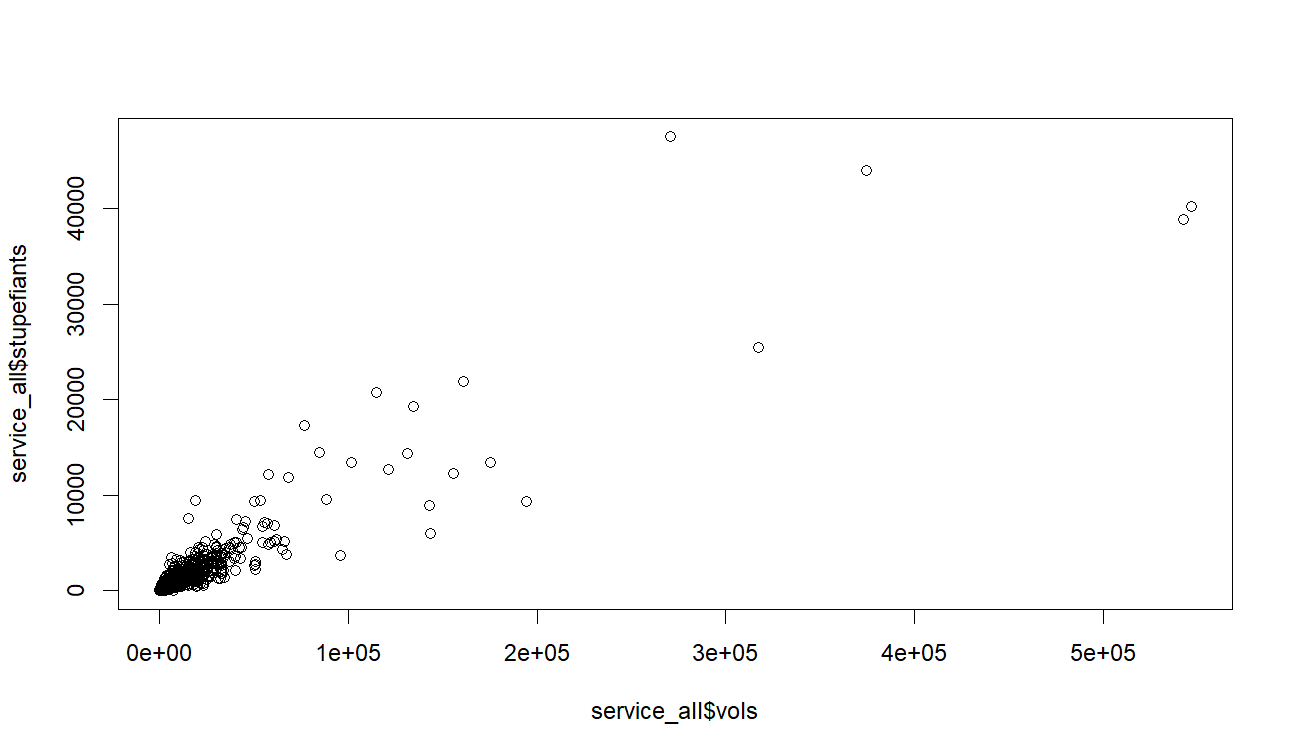
Ce « plot » met en évidence une dépendance linéaire des infractions les unes avec les autres.

La représentation de la matrice des variances/covariances confirme le plot initial.

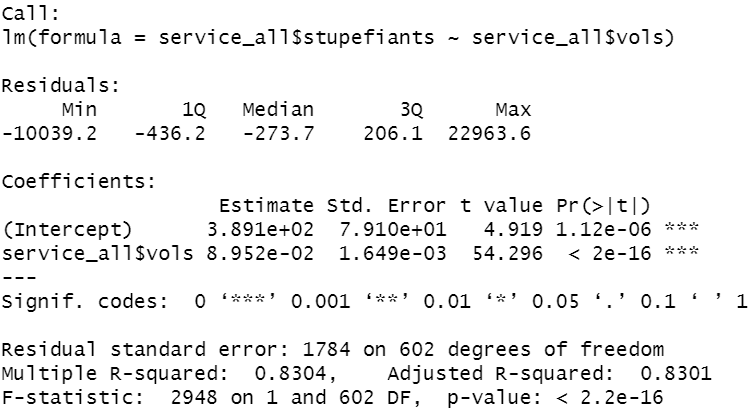
Une image contenant texte

Description générée automatiquement

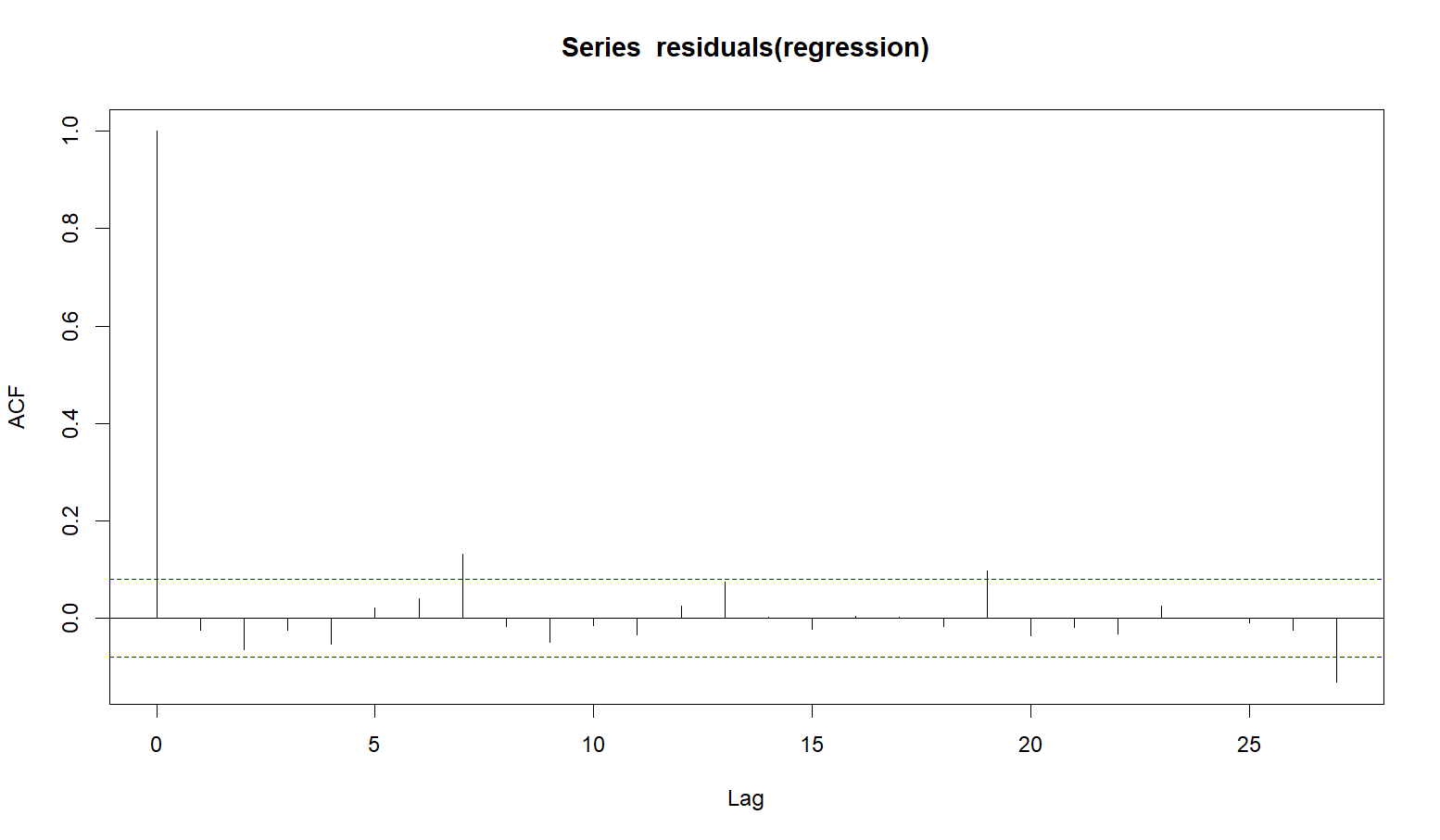
Nous pouvons donc affirmer que : si le nombre d’un des huit types d’infractions choisis augmente alors le nombre de tous les autres types d’infractions augmente.

Nous allons nous intéresser plus particulièrement à la relation entre les stupéfiants et les vols :

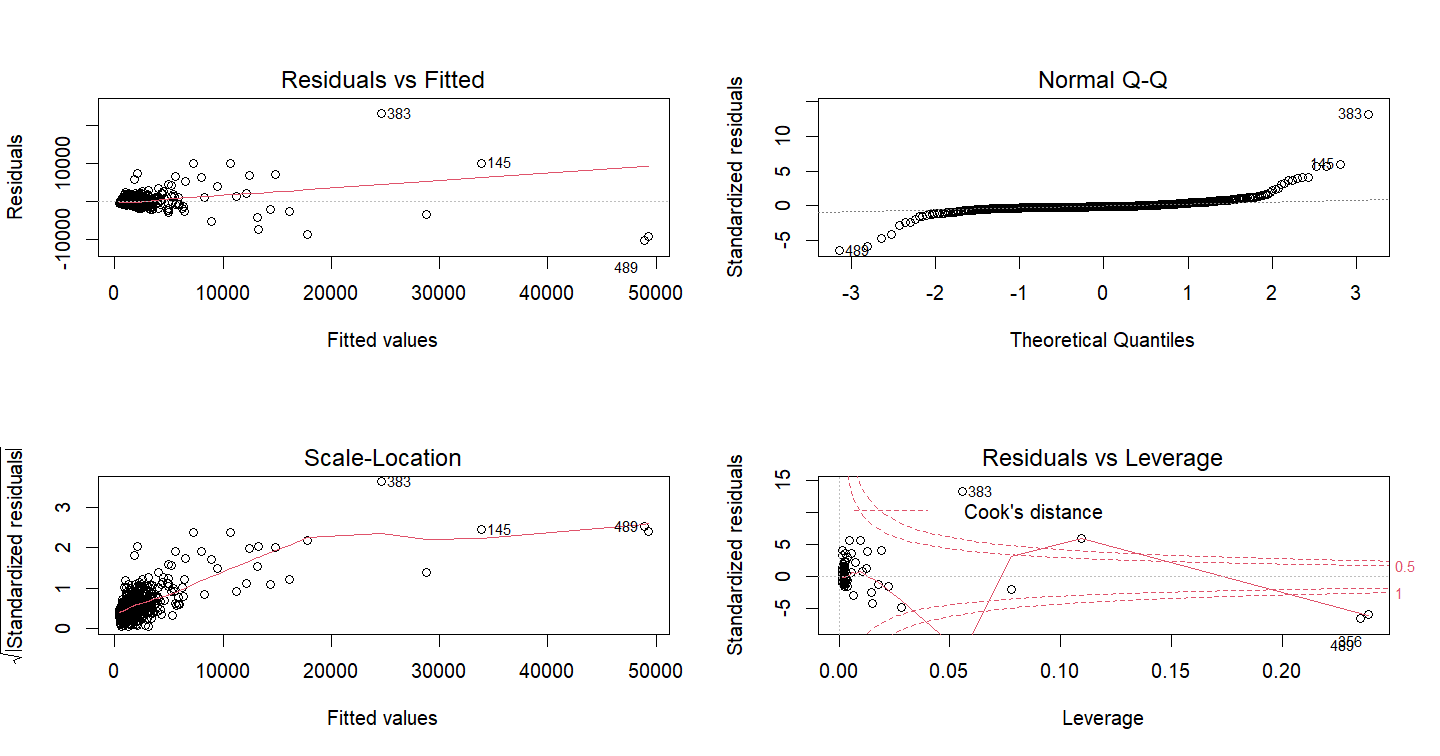
Nous proposons un calcul de régression linéaire pour éprouver notre affirmation :



La linéarité n’est donc pas si évidente. Bien que les p-values soient faibles, le R² n’est pas très élevé, le modèle ne prédit pas très bien les observations. La représentation des résidus nous le confirme.



De même que les relations residuals vs fitted et scale location qui semblent perturbées par les outliers.



L’exemple ci-dessus a été volontairement choisi parce que semblant un des moins linéaires, son R² est cependant de 0,83. Les valeurs des covariances nous ont néanmoins incités à considérer une forte dépendance entre ces variables et nous travaillerons sur la somme de toutes les infractions, quel que soit le niveau de granularité.

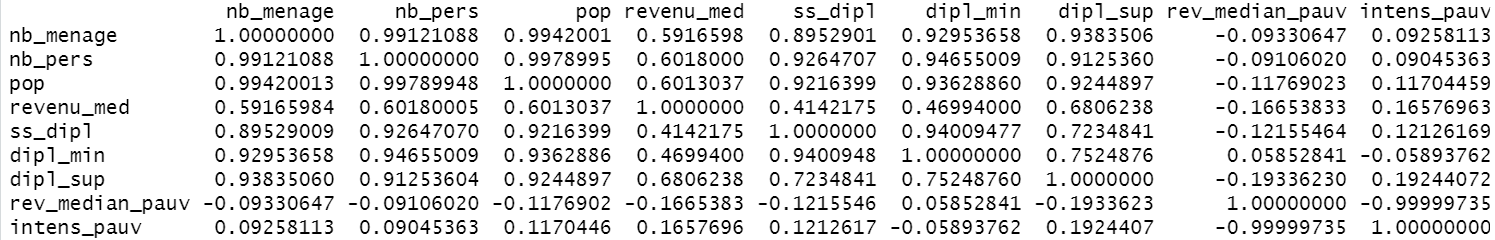
1. Données Insee

Concernant les données issues de l’Insee, une première représentation graphique met en évidence, sans surprise, que le nombre de personnes, le nombre de ménages et la population des départements sont fortement corrélés. De même pour le revenu médian des personnes dites « pauvres » et l’intensité de pauvreté. La linéarité est telle que nous pouvons supposer que l’intensité de pauvreté est calculée à partir de ce revenu médian.

Au sujet des diplômes, nous remarquons une dépendance linéaire au nombre d’habitants, quel que soit le niveau. Chaque personne n’étant comptée qu’une fois dans une des trois catégories, le résultat est ce que nous pouvions attendre, à savoir : « plus il y a de personnes dans un département, plus il y a de personnes avec ou sans diplôme ».

Une image contenant table

Description générée automatiquement



Bien que certaines valeurs soient très fortement liées, et dans la mesure où nous ne pouvons pas facilement les agréger, nous conserverons l’ensemble des variables dans la suite de notre étude

# Etude statistique des relations causes-conséquences

Niveaux de granularité :

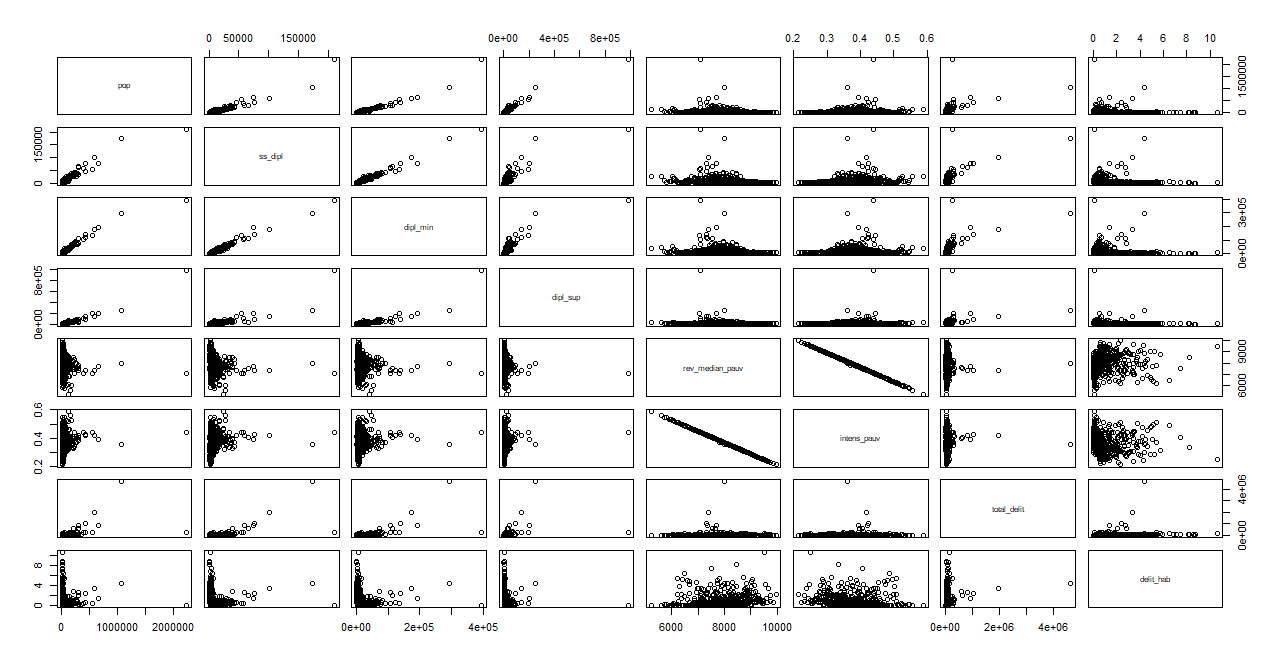
* Emprise gendarmerie / police : il y a un niveau de regroupement supérieur
* Emprise de tribunal de grande instance (ajout du nombre de magistrat)
* Département : ajout de quelques données démographiques, par exemple : chômage

Chacun de ces niveaux a un intérêt : le niveau de plus forte granularité (emprise gendarmerie police) conserve le plus de variance (non lissé par des moyennes et des valeurs non remplies), les 2 niveaux supérieurs apportent chacun des données supplémentaires (nombre de magistrat par emprise de tribunal judiciaire et taux de chômage par département) et permettent de lisser les données manquantes par les moyennes.

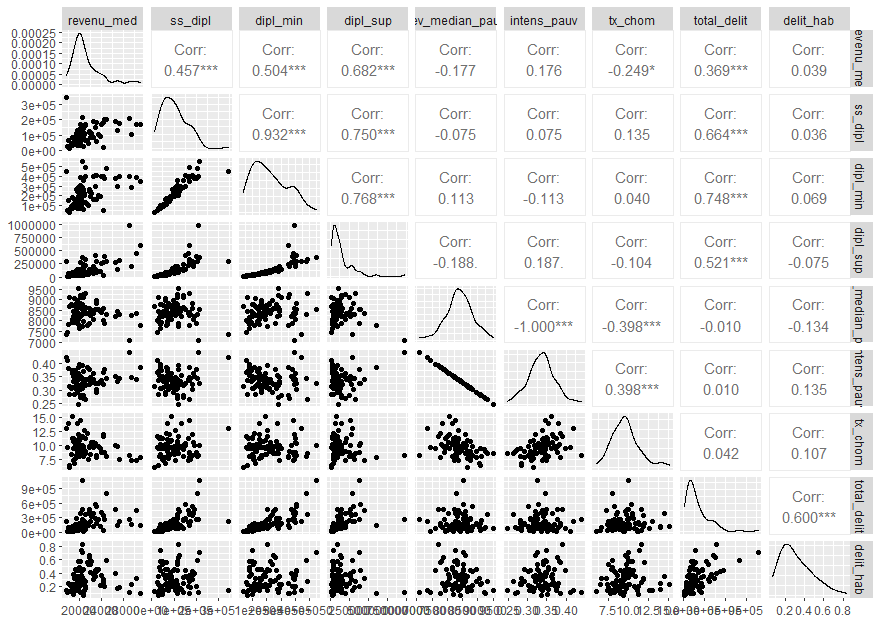
Nous avons effectué l’analyse statistique pour chacun de ces niveaux de granularité en utilisant un script R prenant en entrée le jeu de données correspondant. Nous allons ici décrire le processus en utilisant le niveau départemental, les données pour les deux autres jeux de données sont données en annexes au rapport.

Détail de l’analyse du niveau départemental (scriptR\_dept en annexe) :

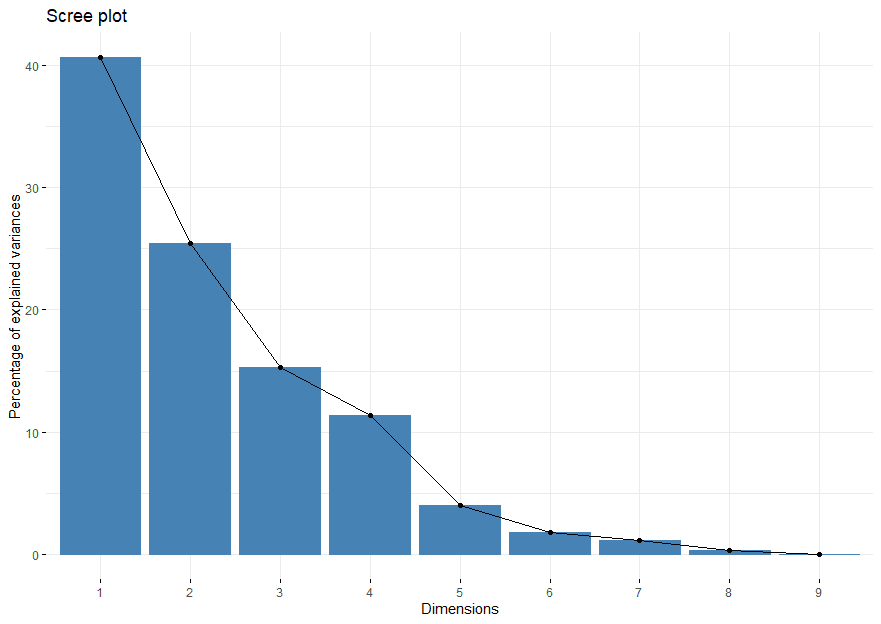
1. Chargement du jeu de données dans R
2. Enrichissement du jeu de données avec les données externes du chômage (récupérées sur le site de l’INSEE) grâce au code\_departement (on prend le soin de vérifier que la jointure donne bien le bon nombre de lignes, c'est-à-dire le nombre de départements !)
3. On enrichit le jeu de données avec une colonne delit\_total (somme de tous les délits par catégorie), et une colonne delit\_hab (nombre total de délits par habitant)
4. On va sélectionner quelques colonnes pour faire nos analyses (on a déjà vu que certaines données sont très corrélées, inutile donc de toutes les étudier)
5. On fait une visualisation initiale de notre jeu de données grâce à un plot , on a à ce point un jeu de données exploitable pour l’étude statistique en elle-même



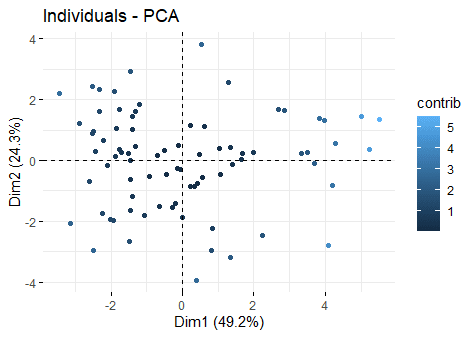
1. Vu les représentations du plot, on va créer un autre jeu de données dans lequel on va enlever les quelques outliers qui risqueraient de fausser notre analyse. On procède par colonnes, en :
   1. Étudiant les colonnes en triant les valeurs par ordre décroissant, et en regardant les valeurs extrêmes (dues soit à des anomalies soit à des conditions particulières)
   2. Remplissant au fur et à mesure une liste d’outliers, que l’on pourra ensuite enlever à notre jeu de données principal
   3. Dans le jeu de données départemental, on identifie comme outliers : NORD, RHONE, PAS DE CALAIS et SEINE MARITIME.
   4. On obtient sans ces outliers la représentation suivante :

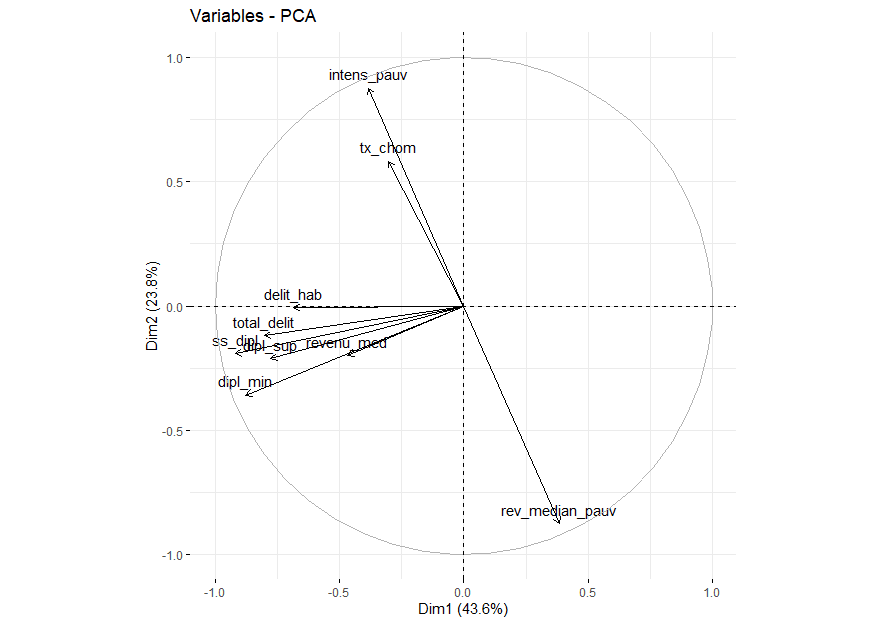


1. Après avoir identifié ces outliers, on va pouvoir réaliser 2 analyses en composantes principales (avec et sans outliers), en utilisant la fonction dudi.pca ainsi que les fonctions qui lui sont associées (get\_eigenvalue, fviz\_eig, var\_pca, fviz\_pca\_ind et fviz\_pca\_var)
   1. ACP sans outliers :

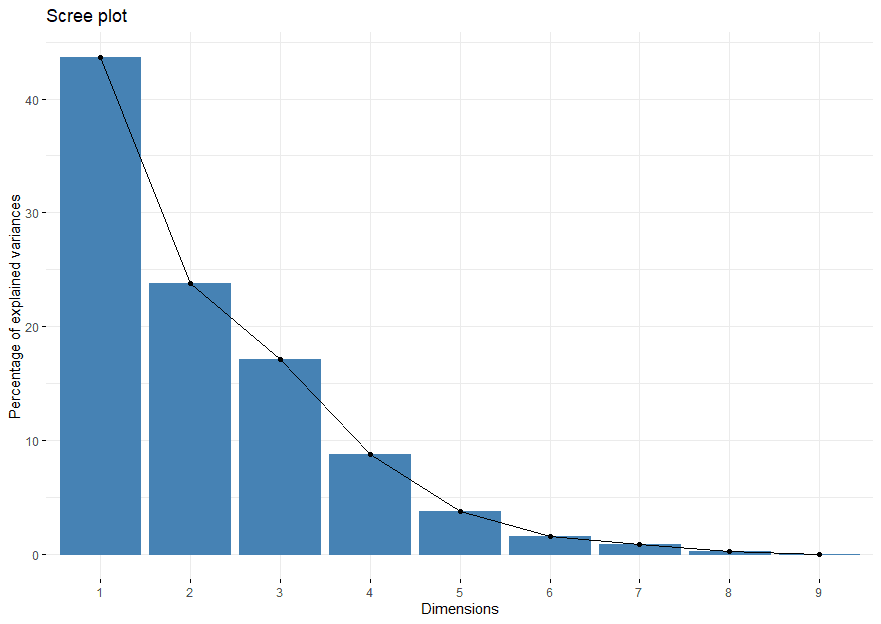


(On gardera 4 dimensions)

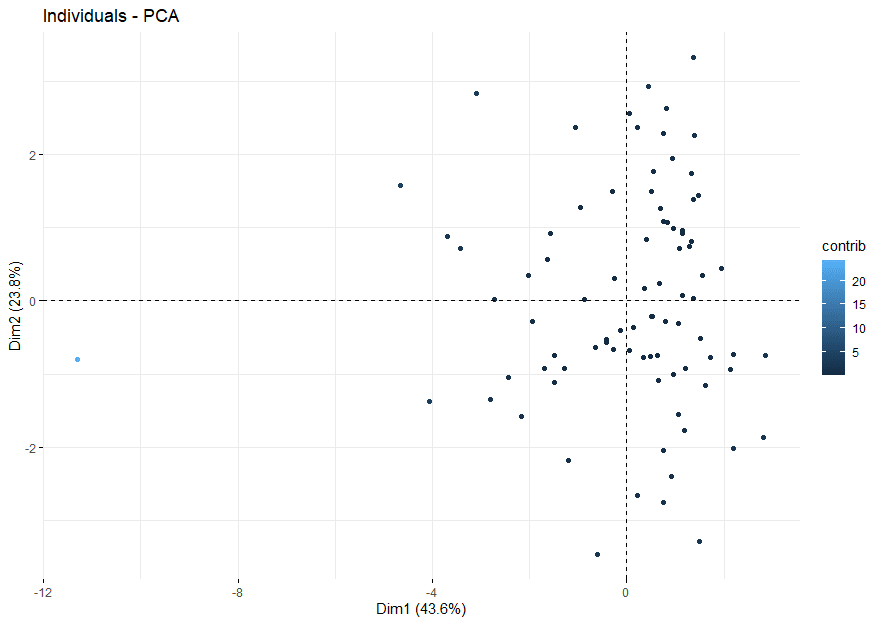


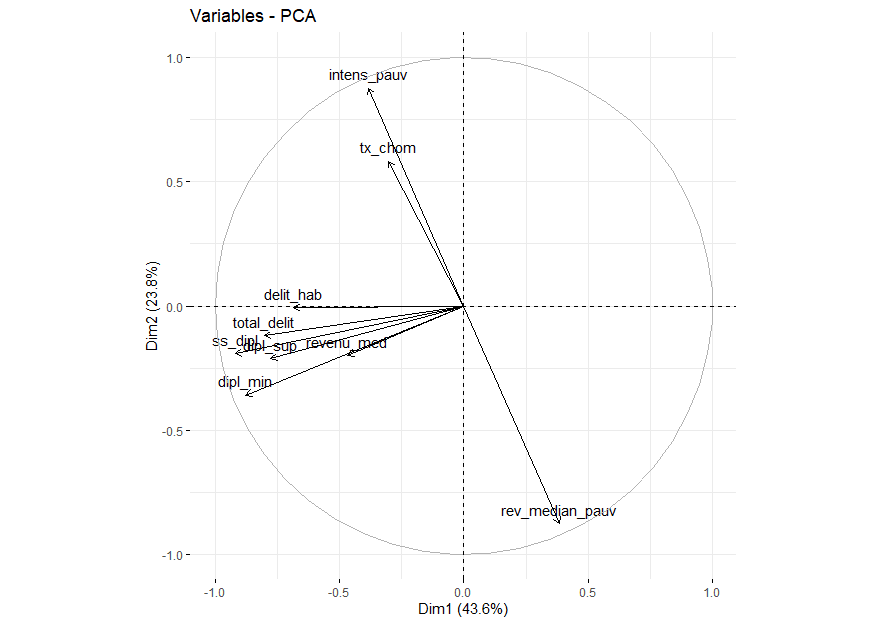


1. ACP avec outliers :



On ne conserve que 3 dimensions





On remarque que les résultats sont très proches avec ou sans les outliers, on choisira alors de n’étudier que les résultats qui conservent les outliers.

On a donc pu réduire le nombre de dimensions à 3, en conservant près de 90% de la variance globale. La plupart des variables sont bien représentées par les 2 premières dimensions de l’ACP.

**On remarque que les dimensions peuvent être classées en 2 catégories : très corrélées à l’intérieur d’une catégorie, et très décorrélées entre catégories :**

* **Catégorie 1 : intens\_pauvreté, tx\_chomage, rev\_med\_pauvr**
* **Catégorie 2 : delit\_hab, total\_delit, ss\_dipl, dipl\_min, dipl\_sup, rev\_med**

Contrairement à l’idée initiale de corrélation entre le taux de chômage et la criminalité, il semble que ces deux valeurs soient en fait décorrélées. La criminalité semble en revanche relativement corrélée au niveau d’éducation du territoire, que l’on mesure avec les indicateurs sur le nombre de diplômes dans le territoire.

Pour les deux autres niveaux de granularité, on trouve des résultats sensiblement différents, avec des corrélations différentes, et une sensibilité aux outliers plus importantes (les outliers sont moins lissés dans une agrégation plus importante, on en trouve d’ailleurs plus).

# 

# Conclusion de l’étude

Cette étude cherchait donc à mettre en évidence les liens statistiques existants entre certains facteurs sociaux et les chiffres de la criminalité dans le territoire français (métropolitain). Toutes ces données étaient issues de sites open sources gouvernementaux (INSEE, ministère de l’intérieur).

La première étape a été de construire un jeu de données le plus exploitable possible, avec des données initialement non liées entre elles. Cela nous a amené à rapidement envisager 3 niveaux géographiques d’études : le niveau des entités gendarmerie / police, le niveau tribunal d’instance et le niveau départemental, qui présentent chacun des avantages et des inconvénients (précision/complétude des données vs généralisation et données supplémentaires).

Une fois cette étape longue et complexe pour des raisons techniques, nous avons pu réaliser les études statistiques sur les données elles même (liens entre les données INSEE et de criminalité) et sur les 3 niveaux d’agrégation.

Les études sur les données d’entrée (INSEE) ont montré l’existence d’une corrélation importante entre les différents paramètres choisis (niveau d’éducation, salaires médians, intensité de la pauvreté). Il a été de même pour les données de sortie (criminalité) qui sont très liées linéairement (la répartition dans les différentes catégories choisies est pratiquement proportionnelle au nombre global de crimes commis).

L’étude finale de liens entre toutes ces données a été faite grâce à des analyses en composantes principales (ACP) sur chacun des niveaux d’agrégation (emprises polices/gendarmerie, cours d’appel, département) que nous avons enrichis de données (nombres de magistrats par cours d’appel et taux de chômage par département). Chacune de ces ACP a été réalisée deux fois, la première avec le jeu de données complet, et la seconde avec un jeu de données épuré des outliers qui peuvent fausser les résultats. Au final, ces derniers ne faussent que peu les résultats.

De ces ACP, nous avons déduit qu’il existe une relation de corrélation entre le nombre de crimes et le niveau d’éducation dans les différents niveaux, mais pas forcément avec le taux de chômage, contrairement à l’idée initiale que nous avions. Le nombre de magistrat par cours d’appel est lui aussi très corrélé au taux de criminalité (heureusement !), ce qui semble être la réponse logique du ministère de la justice à la répartition de la criminalité sur le territoire (adaptation du nombre de magistrat là où il y a le plus de crimes et délits).

**Au final, si vous avez à faire à la justice pour une raison qui n’est pas un crime ou un délit (affaires civiles : problème de voisinage, tutelle, affaires familiales etc…) : mieux vaut être dans une cours d’appel où il y a beaucoup de criminalité que dans une cours d’appel très peuplée. (Le nombre de magistrat est lié au nombre de crimes plutôt qu’à la taille de la population).**

**Autrement dit : si malheureusement vous êtes amené à divorcer, le divorce sera prononcé plus vite à Bobigny ou à Lille qu’à Quimper ou au Mans.**

Ce que nous retiendrons de cette étude :

* Plus de 50% du temps consacré à la création et à la fiabilisation du jeu de données
* Puissance de R et de l’ACP
* L’utilisation des données spatiales permet de contourner des problèmes techniques complexes
* Importance des formats d’encodage
* Passer d’un programme à l’autre nécessite des précautions importantes (R – QGIS – PostGreSQL/PostGIS)
* Puissance de l’open source (tant pour les données que pour les programmes)