Travail 2

IFT3700

Introduction à la science des données

TRAVAIL PRÉSENTÉ À

Alain Tapp

22 décembre 2022

Par : Hugo Carrier 20197563

Alexandre Eyrolle 20134593

Dereck Piché 20177385

1. Introduction

L'objectif de ce devoir était de collecter une multitude de données sur tous les pays et territoires du monde pour ensuite traiter ces données. Les informations collectées allaient du PIB jusqu'au taux de publications scientifiques. Ce devoir était un exercice d'apprentissage multifacette pour le domaine des sciences des données. Nous avons dû être capable d'importer des données non-uniforme de différentes pages Wikipédia. Nous avons aussi dû e nettoyer ces données pour ensuite les traiter au travers de divers algorithmes de prédiction et de réduction de dimensionnalité. Ce rapport évoquera en détails tout ce processus avec les difficultés et les choix que nous avons fait tout du long de ce projet.

Le code complet est présent sur https://github.com/Sethorvo/TP-2-IFT-3700 .

2. Les données

2.1. Importation

Le mandat demandait de recueillir de l'information sur 40 tableaux différentes provenant tous de Wikipédia. Pour ci-faire, nous avons utilisé du code python ainsi que l'utilisation de la librairie pandas pour recueillir l'information de chacune des 40 pages. Cependant, étant donné que la disposition des tables sur ces pages change d'une à l'autre, nous avons dû manuellement sélection les bonnes tables et les bonnes colonnes pour recueillir l'information désirée. Quelquefois il arrivait même que pandas ne soit pas en mesure de détecter la table et que seulement certaines des colonnes désirées contenant des pourcentages étaient lus comme une longue liste de NaN. Plusieurs méthodes ont été testées pour importer ces colonnes donc la conversion forcer en string puis en valeur numérique mais rien n'a réussi. Nous avons dû nous résoudre à les importer manuelle en utilisant un fichier csv même. Heureusement, ce n'est arrivé que dans le cas que de deux colonnes.

Une fois toutes les données recueillies, nous avons dû corriger des erreurs d'importations. Les données recueillies sont sous forme de chaîne de charactères. Certaines d'entre elles contiennes des charactères en trop et d'autres était composé d'intervalles. Le résultat désiré était des tables de nombres. Nous avons donc commencé par enlever tous charactère en trop par essaies et erreurs. Nous terminons donc avec une plusieurs étapes qui traitent les

chaînes de charactère en python. Par la suite, nous avons implémenter une fonction qui lit les intervalles et calcule la médiane de celle-ci.

Trois autres cas importants restaient à être corrigés. Le premier étant de transformer tous les charactères utilisés pour indiquer que l'information n'est pas disponible par un NaN (not a number) de numpy.

Le deuxième cas est de transformer les pourcentages en chiffre à virgule. Parce que l'information avait déjà été corrigée de tout autres charactères, il suffisait de lire le dernier charactère de la chaîne. Si nous trouvons un signe de pourcentage, nous l'enlevons et divisons le chiffre par 100, sinon nous transformons simplement la chaîne en chiffre.

Le troisième cas important provient de la table contenant l'âge criminelle. Les États-Unis ont causé un problème à notre importation de données. Cette information change dépendamment des États. Nous avons donc fait le choix d'utilisé l'âge fédéral mentionné sur la page elle-même, soit un âge de 11 ans.

2.2. Traitement des données

Une fois l'importation et la correction de l'importation terminé, nous nous retrouvons avec un tableau de 40 colonnes contenant diverses informations chiffrées. La prochaine étape est de supprimer les lignes manquants trop d'information et puis de remplir tous les chiffres manquants de ce tableau.

Nous avons commencé avec 223 pays. Telle que demandé, nous avons supprimer toutes les lignes ayant 12 ou plus valeurs manques. 74 pays se retrouve dans cette catégorie. Après avoir supprimer les lignes, il reste 149 lignes d'informations

Une fois que ces données ont pu être filtrées, nous avons commencé a assemblé des données importantes pour la suite du projet en collectant tout un tas d'informations pour chacune des colonnes, des données telle que la médiane, le max et le min, la variance, la moyenne. Des informations que nous avons mis dans un tableau séparé afin d'y accéder pour le reste du projet.

Par la suite, avec la médian calculé pour chacune des colonnes, nous avons remplis les informations manquantes. Une fois que nous avons eu une table complète, il suffit appliquer une régression linaire sur chacune des colonnes. C'est-à-dire, une à la suite de l'autre, une colonne devenait notre valeur dépendant aux autres colonnes. Nous avons utilisé la méthode LinearRegression offert dans la librairie sklearn de python. Une fois le modèle appris, nous avons recalculer les valeurs qui avait été précédemment remplacées par les médianes. Nous avons refait cette procédure une deuxième fois et ce, pour chacune des colonnes.

Une fois toutes ces étapes terminées, nos données étaient prêtes pour les prochaines étapes.

3. Corrélations

La première étape pour calculer les corrélations est de déterminer laquelle il faut utiliser. La corrélation de Pearson est la plus connu et est la plus utilisée. Cependant, lorsque les données sont analysées, il est possible de remarquer que la majorité des colonnes ne suivent pas une distribution normale (histogramme dans annexe 6.1). De plus, certaines colonnes contiennent des valeurs loin des autres. Ainsi, pour ces raisons, nous avons décidé d'utiliser la corrélation de rang, aussi nommé la corrélation de Spearman.

La librairie de pandas du langage python nous permet rapidement de faire se calcul. Lorsque nous avons un objet de type « dataframe » il suffit d'utiliser la méthode corr en mentionnant la corrélation désirée.

Lorsque nous trions les corrélations moyennes par ordre descendant, nous obtenons qu'en moyenne, les colonnes qui ont une plus grande force de corrélation avec les autres sont indice de développement humain, la mortalité en dessous de l'âge de 5 ans et l'âge médian.

Si nous avions utilisé la corrélation de Pearson, par curiosité afin de les comparer, nous aurions obtenus des résultats légèrement différents. L'âge médian aurait été en deuxième position, et en troisième, nous aurions retrouvé espérance de vie.

4. Prédictions

4.1. Analyse des prédicteurs

i. Régression linéaire

La quantité de données est un gros problème de cette analyse. Comme mentionné plus tôt, une fois corrigé, nous avons seulement 149 lignes d'information.

Une régression linéaire perd un degré de liberté pour chaque paramètre et en perd un autre pour une constante. Ainsi, nous avons seulement 108 degrés liberté pour évaluation de notre modèle. Si le modèle est évalué avec tous ces données, le plus petit coefficient de détermination est de 0.975. Cela nous indiquerait que pour toutes les colonnes, il est possible de pouvoir utilisé les autres colonnes afin de régresser la valeur manquante. Dans les faits, nos modèles apparaissent avoir un meilleur fit qu'il en ait réellement. Avec peu de degré de liberté et beaucoup de paramètre, il y a un surapprentissage qui se crée. Afin de contrer cet effet, nous allons séparer les données pour l'apprentissage et pour les tests. De cette façon, nous aurons une meilleure image du pouvoir prédictifs réel du modèle, en d'autres mots, sont exactitude.

Nous avons exécuté cette dernière procédure en gardant 40 % des valeurs pour exécuter nos tests. Avec ces valeurs, nous avons calculer la prédiction, puis nous avons trouvé un nouveau coefficient de détermination qui se retrouve hors échantillon cette fois-ci. Certaines valeurs ont encore donné de bons résultats. D'autres ont leurs prédictions pires que prévu. Par exemple l'âge de responsabilité criminelle avait un coefficient de 0.996, mais une fois les données séparer, puis tester, le coefficient hors échantillon d'apprentissage est devenu 0.58. On remarque donc qu'il est particulièrement difficile d'évaluer le modèle correctement avec si peu de données. Le manque de données peut laisser croire que le modèle sera plus précis que prévu.

ii. Classifieur de Bayes Naïf

Pour le classifieur de bayes, nous avons utilisé la version Multinomial avec la discrétisation des données. Pour effectuer cette discrétisation, on a utilisé la médiane des colonnes. Si une donnée est plus grande que la médiane de sa colonne, on la met à 1, dans le cas contraire

on la met à zéro. Pour ce numéro, nous avons arrangé les scores dans une matrice n par n, ou n est le nombre de colonne. L'élément [i][j] de la matrice est alors le score de prédiction de du classifieur de bayes qui utilise les données de la colonne i pour prédire la colonne j. Pour le score, nous avons utilisé la séparation standard 0.8 / 0.2 pour la séparation entre l'ensemble d'entrainement et l'ensemble test. Puisque nous essayons de prédire une catégorie, nous avons pour fonction de coût simplement le nombre de prédictions adéquates sur le nombre de données à prédire.

4.2. Quels sont la paire de colonne qui donne les résultats les plus précis

i. Régression linéaire

Afin de trouver la paire de colonne qui permet de prédire les résultats les plus précis avec une régression linaire, nous avons d'abords choisi un coefficient pour déterminer la précision. Dans une régression linéaire, la précision représente l'écart entre les points et la droite. En d'autres mots, elle dépend de la variance de l'erreur de la régression. Nous avons fait le choix d'utilisé le coefficient détermination pour évaluer cette erreur. Cette fois-ci, nous prenons la valeur trouver dans les variables d'apprentissage car nous cherchons la précision et non exactitude. L'exactitude est la capacité à prédire, en moyenne, la bonne valeur.

Le coefficient de détermination ne permet de comparer facilement chacune des régressions. Il vari toujours entre 0 et 1.

Afin de comparer chacune des paires, nous avons simplement fait la régression avec chaque pair. Heureusement pour nous, nous n'avons pas beaucoup de données. Cette procédure à comme inconvénient d'être plutôt longue étant donnée la grande quantité de régression qui doivent être calculer.

On remarque après avoir roulé le code que lorsqu'une colonne fait partie de la paire d'un autre, l'inverser sera souvent vrai.

ii. Classifieur de Bayes Naïf

Les résultats des classements du classifieur de bayes naïf nous ont grandement surpris. En effet, nous avons remarqué que la colonne 40, qui indique la température moyenne

annuelle, était de loin la plus performante selon notre mesure de précision et notre binarisation.

4.3. Quels sont les meilleurs prédicteurs de chacune des colonnes

i. Régression linéaire

Pour répondre à cette question à l'aide d'une régression linaire, nous avons premièrement standardisé nos données. Les données standardisées sont donc mises sur le même piédestal, c'est-à-dire que leur moyenne et leur variation n'influence pas la valeur des paramétrisation.

Une fois standardisé, il suffit de faire la régression. Une fois calculer, nous avons comparé les coefficient et sortie les deux plus grand coefficient en valeur absolue. Ces deux coefficients ont en moyenne le plus d'impact sur la colonne associé.

Le résultat est présenté dans le tableau à l'annexe 6.1. On remarque que le taux de fertilité a toujours le coefficient le plus grand (sauf pour lui-même car nous l'avons exclu).

En deuxième position, on retrouve le taux de croissance de population et l'indice de développement humain comme majorité.

ii. Classifieur de Bayes Naïf

La deuxième colonne la plus utile pour les prédictions était la colonne 38, qui représente la consommation moyenne de kilocalories par personne et par jour. La consommation de nourriture est également très utile pour distinguer les pays les plus riches des pays plus défavorisés. Cette différence de richesse est elle-même très utile pour déterminer beaucoup d'autres facteurs chez les pays.

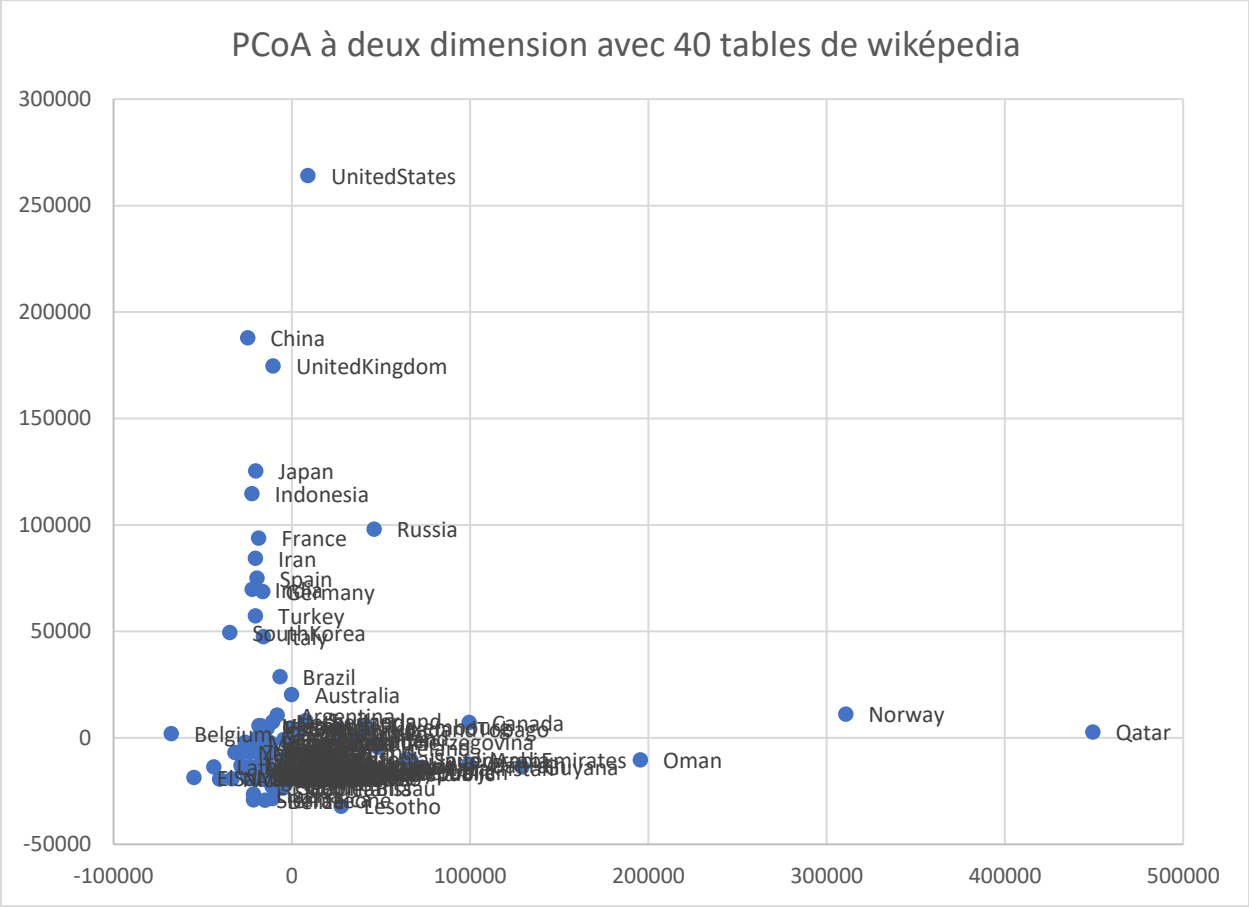
5. Réduction de dimensionnalité

5.1. Réduction à deux dimensions

Afin de visualiser nos données, nous avons opté pour une analyse en composantes principales (aussi connu sous l'acronyme PCoA, pour Principal component analysis en

anglais). Nous avons choisi cette méthode car nos données ont de nombreuses dimensions. De plus, dans le mandat qui nous a été confié, nous devons fournir une interprétation d'une réduction de dimension à deux et à cinq. Dans ce cas, prendre le processus de PCoA nous semblait plus facile.

Les images obtenues par la réduction de dimensions nous indiquent quelques informations importantes. En deuxième position, on retrouve le



D'abord, on remarque que certains de ces pays sont beaucoup plus différent que la moyenne des pays. Lorsqu'on analyse l'image, on remarque un noyau de pays concentré et d'autre pays éloigné. Les pays éloignés ont soit une des deux dimensions élevées mais pas les deux en même temps.

Selon la dimension mis en y, on remarque que les États-Unis ont la valeur la plus éloigné. Pour la dimension en y, on remarque que Qatar et le pays avec la variable la plus grande.

Une analyse en composantes principales réduit les dimensions en trouvant celles qui explique le mieux le problème. Une analogie possible pour bien comprendre le processus est celui d'un ressort. Si un scientifique enregistre quelques vidéos d'un ressort qui bouge selon des angles différents, il pensera probablement que le ressort bouge en 3 dimensions, or, lorsqu'une PCoA est appliquer, il ne trouvera qu'une dimension. Le ressort ne fait que s'étirer et se comprimer, ainsi il ne bouge que sur un axe.

Nous avons un problème semblable, nous avons pris une grande quantité de données entre chaque pays et nous essayons de les démêler. On s'attend donc à ce que certains facteurs différencient mieux nos données. Afin d'essayer d'associé une explication à nos réductions de dimensions, nous avons recalculer la corrélation avec les données de bases. On s'attends à ce que la force de la corrélation soit presque de 1 avec certaines colonnes et c'est le cas.

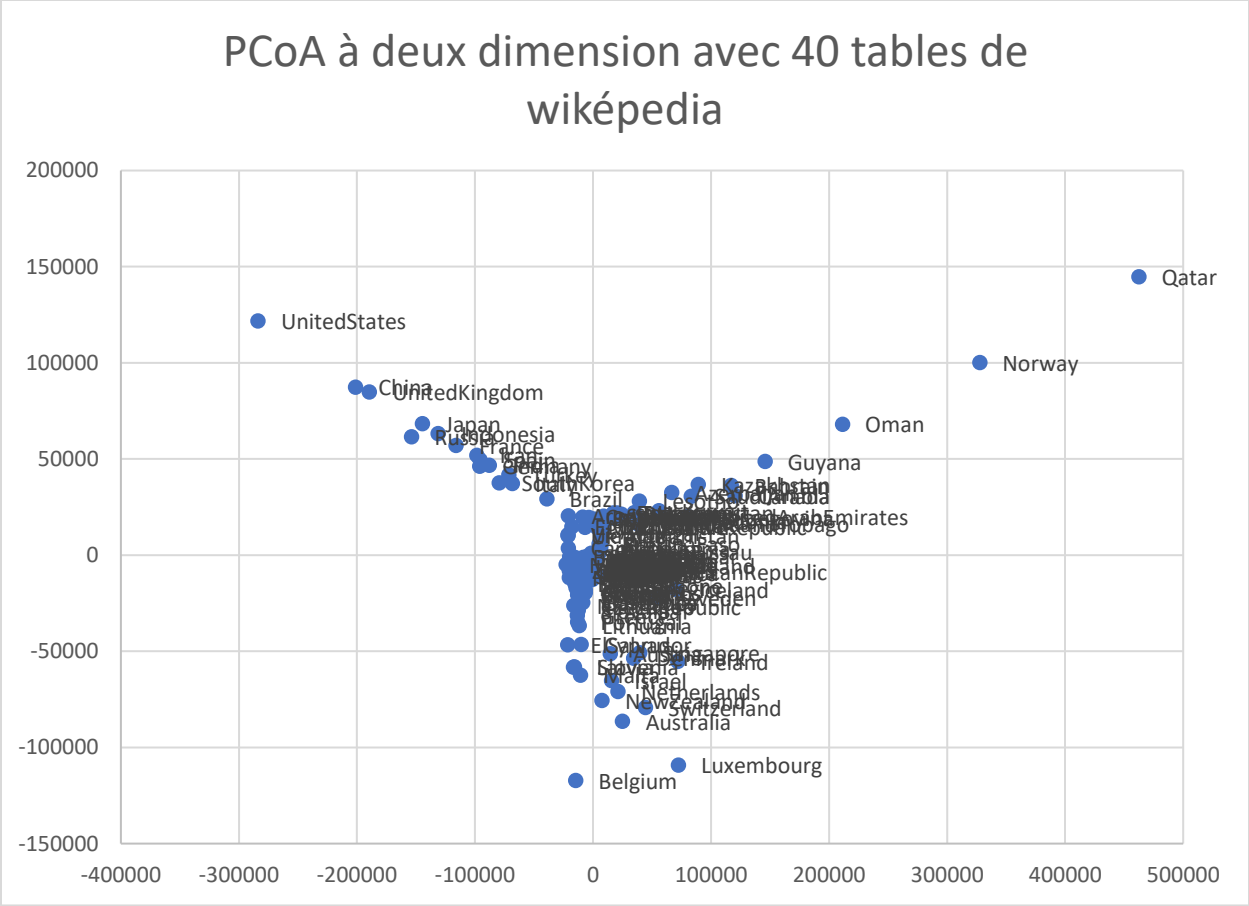
En dimension deux, nous retrouvons que notre dimension en x à une corrélation de 0.99877 avec la production de pétrole. Il est possible par la suite de voir que le Qatar et la Norvège sont deux grands producteurs de pétroles.

La dimension en y à une haute corrélation (0.99432) avec le nombre de livres publiés dans une année. On retrouve donc les États-Unis dans le haut des livre publiés.

Il est important de mentionner que ce n'est pas une science exacte. Même si la corrélation est de proche de 1, nous ne savons pas si elle une égalité ou non. Par exemple, un autre facteur que nous n'avons pas dans nos données et qui serait aussi corrélé avec les données pourrait définir la dimension.

Finalement, nous avons essayé une réduction grâce Isomap afin de comparer les résultats.

Nous retrouvons encore une fois les mêmes pays aux extrémités. Cependant, la forme du graphique à changer. La forme en « L » que nous avions a changé pour une forme en Y.



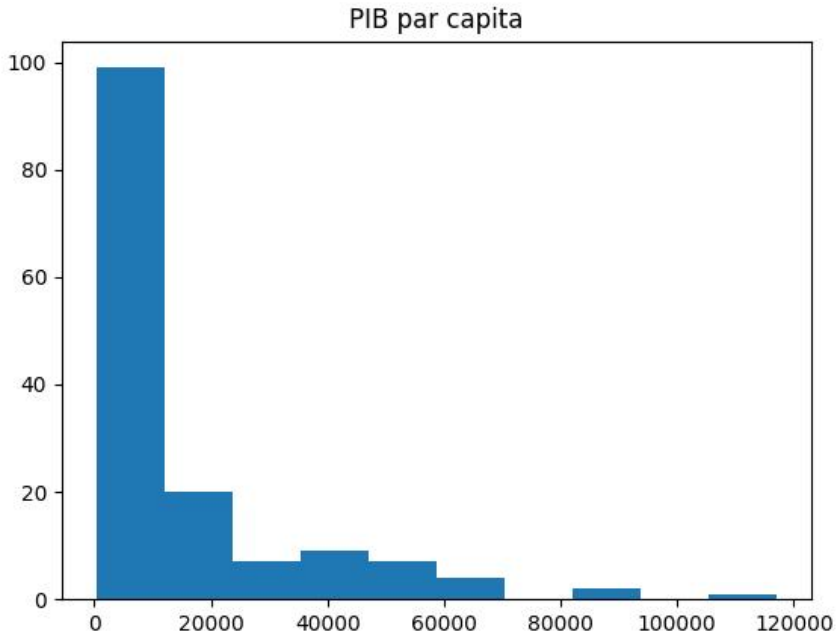
5.2. Réduction à 5 dimensions

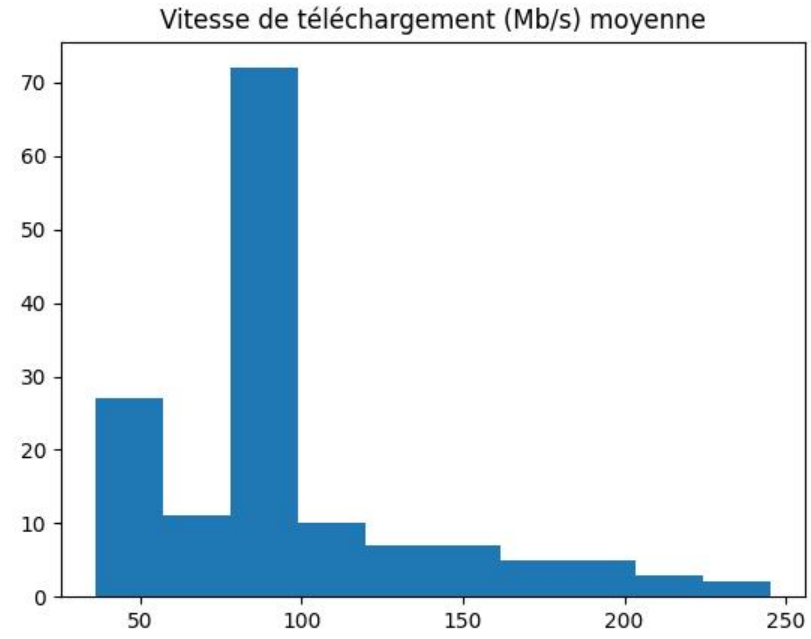
Lorsqu'on regarde une analyse en composantes principales à 5 dimensions, on retrouve les mêmes corrélations pour les deux premières dimensions. Par la suite, la force de corrélation maximale en ordre de dimension est de 0.88 avec le PIB par capita, de -0.55 avec le salaire minimum et de 0.84651 avec les dépenses en santé.

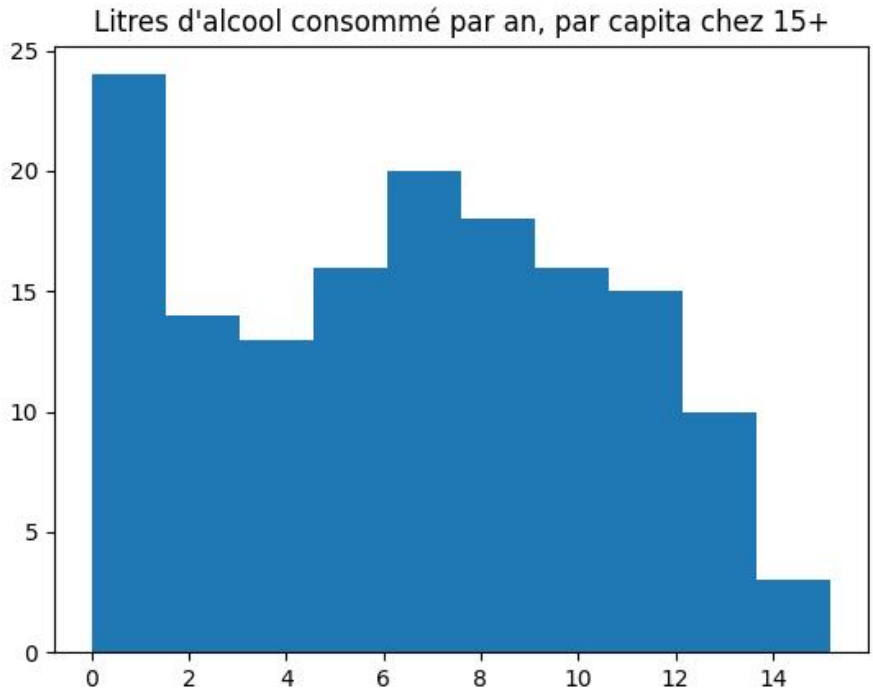
Information aussi très intéressante, lorsqu'on réutilise les forces de corrélation calculer plus tôt, on remarque que les colonnes mentionnées plus haut sont très peu corrélé avec les autres données. Par exemple, la production de pétrole avait une force de corrélation de 0.12542, la plaçant 5ième à partir de la fin.

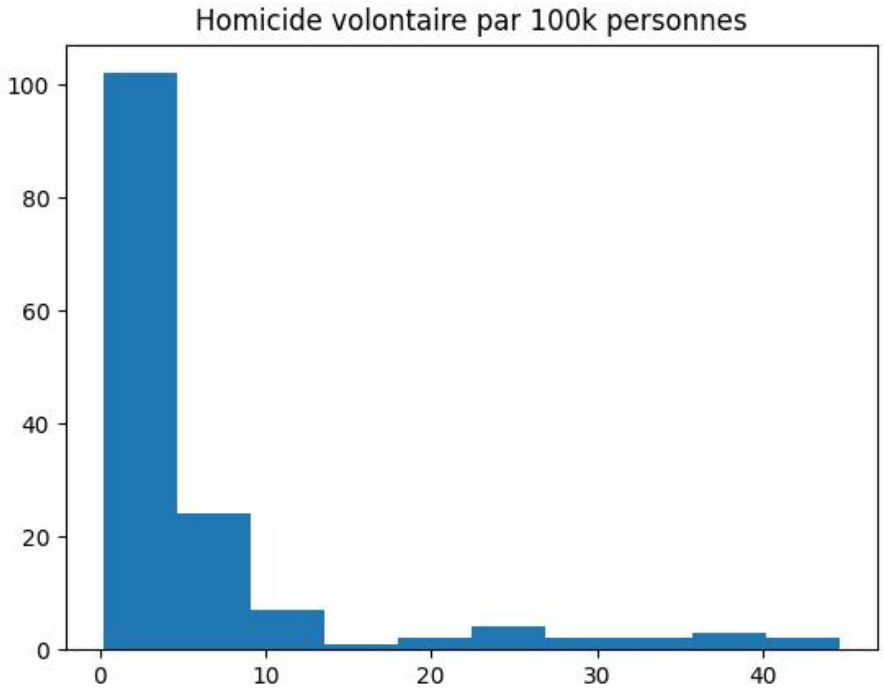
6. Annexe

6.1. Exemple d'histogramme de nos données :






Litres d'alcool consommé par an, par capita chez 15+
25
20
15
10
5
0
0
2
4
6
8
10
12
14


Homicide volontaire par 100k personnes
100
80
60
40
20
0
0
10
20
30
40

6.2. Résultat des plus grand coefficient pour chacune des colonnes

Colonne	Max 1	Max 2
PIB par capita	Taux de fertilité	Indice de développement humain
Vitesse de téléchargement (Mb/s) moyenne	Taux de fertilité	Taux de croissance de population
Litres d'alcool consommé par an, par capita chez 15+	Taux de fertilité	Indice de développement humain
Homicide volontaire par 100k personnes	Taux de fertilité	Taux de croissance de population
%PIB dépensé dans le militaire	Taux de fertilité	Indice de développement humain
Indice de développement humain	Taux de fertilité	Taux de croissance de population
Indice de démocratie	Taux de fertilité	Indice de développement humain
% d'éducation tertiare de 2 ans atteint	Taux de fertilité	Indice de développement humain
% d'importance de la religion	Taux de fertilité	Taux de croissance de population
% de chrétiens	Taux de fertilité	Indice de développement humain
% de musulmans	Taux de fertilité	Indice de développement humain
% de bouddhistes	Taux de fertilité	Indice de développement humain
% de juifs	Taux de fertilité	Indice de développement humain
Mortalité en dessous de 5 ans (mort/1k naissance)	Taux de fertilité	Indice de développement humain
Age de responsabilité criminelle	Taux de fertilité	Taux de croissance de population
Salaire minimum annuel	Taux de fertilité	Taux de croissance de population
Dette externe en % de PIB	Taux de fertilité	Taux de croissance de population
Indice de Gini en %	Taux de fertilité	Taux de croissance de population
Dépense en santé par capita	Taux de fertilité	Taux de croissance de population
Taux de suicide	Taux de fertilité	Indice de développement humain

Taux de fertilité	Indice de développement humain	Taux de croissance de population
Consommateur de tabac chez 15+	Taux de fertilité	Taux de croissance de population
Taux d'obésité	Taux de fertilité	Indice de développement humain
Taux d'utilisateur d'Internet	Taux de fertilité	Taux de croissance de population
Age médian	Taux de fertilité	Indice de développement humain
Score de liberté économique	Taux de fertilité	Indice de développement humain
Production de pétrole	Taux de fertilité	Taux de croissance de population
Taux de croissance de population	Taux de fertilité	Indice de développement humain
Espérance de vie	Taux de fertilité	Taux de croissance de population
Consommation de viande en kg par capita	Taux de fertilité	Indice de développement humain
Taux d'incarcération par 100k	Taux de fertilité	Indice de développement humain
Taux d'alphabétisation	Taux de fertilité	Taux de croissance de population
Age de premier marriage	Taux de fertilité	Indice de développement humain
Dépense en éducation, % du PIB	Taux de fertilité	Taux de croissance de population
Taux d'itinérance par 100k	Taux de fertilité	Indice de développement humain
Consommation de lait par capita	Taux de fertilité	Taux de croissance de population
Nombre d'articles scientifique par capita	Taux de fertilité	Taux de croissance de population
Livres publiés par année	Taux de fertilité	Indice de développement humain
Consommation de nourriture en kilocalories	Taux de fertilité	Indice de développement humain
Température annuelle moyenne	Taux de fertilité	Taux de croissance de population