IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

A journey to data scientist - UE coeur BC - TAF Data Sciences

Intitulé du projet :

“Analyse Prédictive : Impact des innovations éducatives sur les taux de criminalité à Chicago”

Groupe 12

Soumiya RAZZOUK
Chenjie QIAN
Yann LEGENDRE
Hicham CHEKIRI

I- Contexte & Problématique

Contexte :

La ville de Chicago, confrontée à des défis liés à la **sécurité** et à l'**éducation**, cherche à évaluer l'impact potentiel de la **construction** de **nouvelles** écoles sur le **taux de criminalité** dans des zones **spécifiques**.

Ce projet vise à développer un modèle de machine learning capable de **prédire** l'effet qu'aurait l'établissement d'une nouvelle école, dans un rayon de **1 kilomètre**, sur le **taux de criminalité** de la région.

Problématique métier :

L'étude examine l'impact de la scolarité à Chicago sur le taux de criminalité dans les alentours des écoles, en se concentrant sur des facteurs comme after school hours, school hours, nature de l'école et le dress code. Ces données pourront guider le développement de politiques éducatives et de prévention de la criminalité au niveau local.

Comment alors l'établissement de nouvelles écoles, à travers des caractéristiques spécifiques telles que les horaires après le cours, les heures de cours, la nature de l'école, le dress code et la nature des écoles à proximité peut il influencer de manière significative les taux de criminalité dans les zones délimitées par un rayon de 1km ?

Objectif :

Développer un modèle de machine learning capable d'anticiper et évaluer cet impact, permettant ainsi aux décideurs locaux de Chicago de prendre des décisions éclairées quant à la construction de nouvelles écoles et à la promotion d'un environnement éducatif plus sécurisé.

Problématique Data Science:

- **Impact des variables éducatives sur la criminalité :**
  Comment les différents aspects des politiques scolaires, y compris la nature de l'établissement (primaire, collège, lycée), les heures après l'école, la durée des journées scolaires, et les règles de tenue vestimentaire, influencent-ils les taux de criminalité dans les quartiers adjacents aux écoles de Chicago ?"

- **Analyse des Changements de Politiques Scolaires et Leur Effet sur la Criminalité Locale :**
  Quel est l'impact de modifications spécifiques des variables d'entrée sur les taux de criminalité dans les quartiers environnants ? Notre objectif est de mesurer l'augmentation ou la diminution du taux de criminalité consécutive à ces changements

Relation criminalité et éducation et choix des variables à étudier :

Le lien entre la criminalité et l'éducation, ainsi que la sélection des variables à étudier, se fondent sur des observations significatives. Selon un article du "Police Chief Magazine" (1), les programmes après l'école jouent un rôle crucial dans la réduction de la criminalité en offrant des activités encadrées pendant les heures où les jeunes sont plus susceptibles de commettre ou d'être victimes de crimes. Cet article souligne que l'engagement dans ces programmes est associé à de meilleures performances académiques et à un comportement plus positif chez les élèves.

Concernant les effets des codes vestimentaires, "Leicestershirevillages" (2) aborde la question des uniformes scolaires en soulignant qu'ils contribuent à instaurer une atmosphère de discipline et de cohésion, réduisant potentiellement les comportements violents et les infractions.

Ces constats fournissent une base rationnelle pour explorer l'impact potentiel de ces deux facteurs éducatifs sur les taux de criminalité au sein des quartiers de Chicago.

Notre recherche s'inspire aussi des travaux approfondis de Willits, Broidy et Denman (3), qui ont exploré les relations complexes entre les écoles et les taux de criminalité dans les quartiers. Leur étude suggère que les écoles, en particulier les établissements d'enseignement secondaire, peuvent influencer indépendamment les taux de criminalité des quartiers environnants, au-delà des dynamiques structurelles et sociales préexistantes.

En complément des enseignements de la recherche de Willits, notre étude s'appuie sur le travail de Gordon A. Crews (4), qui explore la relation complexe entre l'éducation et le crime. Crews souligne l'importance de l'éducation non seulement comme moyen de prévention du crime, mais aussi en tant que facteur influençant les comportements criminels.

Selon Crews, il existe une corrélation généralement acceptée entre le niveau d'éducation et la probabilité de s'engager dans un comportement criminel. Un niveau d'éducation plus élevé est souvent associé à une probabilité réduite de comportement criminel. Cette notion soutient l'idée que l'éducation peut jouer un rôle crucial en fournissant des compétences académiques et professionnelles, tout en cultivant une conscience sociale et morale qui dissuade les comportements criminels.

Notre étude vise à explorer et quantifier ces influences dans le but d'informer et d'optimiser les politiques publiques locales.

Solutions offerte et résultats attendus:

- Un modèle de régression qui permet de mesurer les corrélations entre les paramètres d'entrée (politiques liés au heures de cours, dress code …) et le taux de criminalité. Notre proposition de valeur réside dans la capacité d'anticiper et prédire l'impact que peut avoir l'établissement d'une nouvelle école dans une zone prédéfinie.

Client Potentiel :

Le principal client de cette solution serait le gouvernement de la ville de Chicago, plus précisément les départements de l'Éducation et de la Sécurité publique qui sont constamment à la recherche de méthodes basées sur les preuves pour formuler et ajuster les politiques publiques.

Les décideurs au sein de la mairie de Chicago, en collaboration avec le département de la police et les agences de planification urbaine, représentent les acteurs clés qui pourraient financer et mettre en œuvre les recommandations issues de cette étude pour favoriser une communauté plus sûre et plus éduquée.

Source de données :

Nos données ont été extraites du CHICAGO DATA PORTAL, un portail de données ouvert et publique géré par la ville Chicago

II - Data Preparation:

1- Importation des données :

Nous avons commencé par importer les données en utilisant la bibliothèque Pandas de Python.
Ceci inclut les données scolaires de Chicago pour les années académiques de **2016-2017** à **2022-2023**, ainsi que les données sur les crimes de 2001 à présent.
Chaque ensemble de données a été chargé séparément pour assurer une manipulation et une analyse précises.

Vue sur la dataset school :

School_ID	Legacy_Unit_ID	Finance_ID	Short_Name	Long_Name	Primary_Category	Is_High_School	Is_Middle_School	Is
400011	4730	66151	LOCKE A	Alain Locke Charter School	ES	False	True	
609958	3690	29121	GUNSAULUS	Frank W Gunsaulus Elementary Scholastic Academy	ES	False	True	
400049	5870	67071	LEGACY	Legacy Charter School	ES	False	True	
400134	9051	0	YCCS - ADDAMS	YCCS-Jane Addams Alternative HS	HS	True	False	

Vue sur la Dataset Crime :

ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward	Communi Ar
43	HN549294	08/25/2007 09:22:18 AM	074XX N ROGERS AVE	0560	ASSAULT	SIMPLE	OTHER	False	False	...	49.0	
89	HH109118	01/05/2002 09:24:00 PM	007XX E 103 ST	0820	THEFT	$500 AND UNDER	GAS STATION	True	False	...	NaN	Na
121	JG415333	09/06/2023 05:00:00 PM	002XX N Wells st	1320	CRIMINAL DAMAGE	TO VEHICLE	PARKING LOT / GARAGE (NON RESIDENTIAL)	False	False	...	42.0	32
88	JG423627	08/31/2023 12:00:00 PM	023XX W JACKSON BLVD	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	STREET	False	False	...	27.0	28

Après l'importation, nous avons procédé à un examen initial des données pour comprendre leur structure. Ceci a inclus l'observation des premières lignes de chaque jeu de données pour identifier les variables pertinentes, telles que l'ID de l'école, le type d'école (primaire, collège, lycée), les programmes périscolaires, le code vestimentaire pour les données scolaires, et l'ID de l'affaire, le type de crime, la localisation pour les données criminelles.

2- DATA cleaning et traitement des valeurs nulles :

- Nous avons remarqué que la variable **after_school_hours** présente plusieurs valeurs nulles (*nan*) et nous avons supposé que ces écoles ne font pas ce genre de programme. Pour cela nous avons remplacé *nan* par 0 avec la méthode fillna.

```
scolaire1617['After_School_Hours']
✓ 0.0s
0                    NaN
1          3:00pm - 5:00pm
2                    NaN
3                    NaN
4                    NaN
              ...
656    3:00 pm - 5:30 pm
657                  NaN
658                  NaN
659                  NaN
660       3:00 pm-4:00 pm
Name: After_School_Hours, Length: 661, dtype: object
```

====>

After_School_Hours
0
0
0
0
0

3 - Fusion des datasets :

- **Fusion des données scolaires en une dataset :** Les données scolaires de 2016 à 2023 ont été consolidées en un unique Dataset, permettant une analyse des variables éducatives à travers le temps et qui aboutit à une base de données consolidée comprenant 4,593 lignes et 102 colonnes.

- Nous avons **fusionné** les données sur les **écoles** et les **crimes à Chicago** de 2016 à 2023, calculé le ratio de la criminalité autour des écoles d'une année à l'autre et exclut l'année 2023 :

	School_ID	Legacy_Unit_ID	Finance_ID	Short_Name	Long_Name	School_Type	Primary_Category	Is_High_School	I
0	610163	5770	30081	STOCK	Frederick Stock Elementary School	Neighborhood	ES	N	
1	610558	9598	46611	GOODE HS	Sarah E. Goode STEM Academy	Citywide-Option	HS	Y	
2	609750	1750	49051	SIMPSON HS	Simpson Academy HS for Young Women	Citywide-Option	HS	Y	
3	610571	9636	65015	OMBUDSMAN - WEST HS	Ombudsman Chicago-West	Citywide-Option	HS	Y	
4	610123	5370	24911	PENN	William Penn Elementary School	Neighborhood	ES	N	

Is_Elementary_School	...	SignificantlyModifiedMOD	transition	geometry	Crime_Count	Crime_Count_Next_Year
N	...	NaN	NaN	POINT (-5350956.717336828 11536273.469028268)	23	16.0
N	...	NaN	NaN	POINT (-5395265.04219293 11537655.527686711)	221	181.0
Y	...	NaN	NaN	POINT (-5379823.426302454 11527539.629460143)	167	165.0
N	...	NaN	NaN	POINT (-5377513.540669364 11529195.959269354)	518	435.0
Y	...	NaN	NaN	POINT (-5378677.600331739 11533981.78093037)	739	657.0

3- Choix et transformation des variables

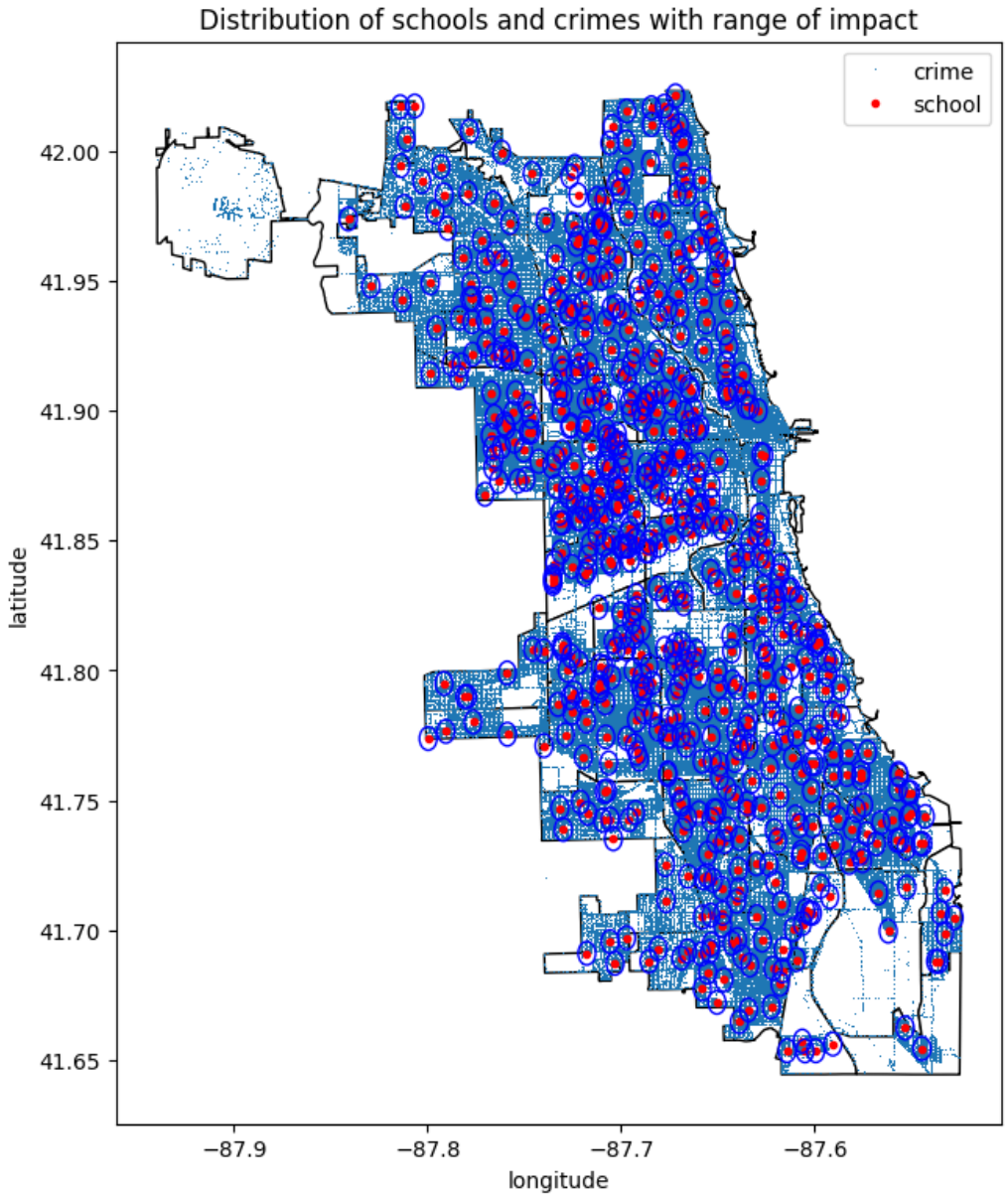
A : Sélection des variables :

Nous avons affiné notre ensemble de données en sélectionnant des variables que nous voulons étudier, ce qui a donné lieu à un tableau final de 3941 lignes et 14 colonnes:

['Is_High_School', 'Is_Middle_School', 'Is_Elementary_School', 'Is_Pre_School', 'After_School_Hours', 'School_Hours', 'Dress_Code', 'Student_Count_Total', 'Count_High_School_Near', 'Count_Middle_School_Near', 'Count_Elementary_School_Near', 'Count_Pre_School_Near','Crime_Count']
Tel que :

Is_High_School : booléen
Is_Middle_School : booléen
Is_Elementary_School : booléen
Is_Pre_School : booléen
After_School_Hours : converti en minutes
School_Hours : converti en minutes
Dress_Code : booléen
Student_Count_Total : entier : Nombre total d'élèves inscrits à l'école
Count_High_School_Near : entier : nombre de lycée situé dans une zone géographique de 500 m.
Count_Middle_School_Near : entier
Count_Elementary_School_Near : entier
Count_Pre_School_Near : entier
Crime_Count : entier : Nombre des crimes dans le voisinage de l'école.

Avec count high school near, count middle school near, count elementary school near, count pre school near et crime count sont des variables que nous avons créées à partir de notre database. Et pour faire, nous avons sélectionné un cercle de 500 m qui entoure chaque école et nous avons repéré dans cet entourage les écoles, le nombre de crimes, la nature des écoles existantes et le nombre des élèves. La figure ci-dessous illustre cette division que nous avons faite :



Les pixels (points) en bleu clair représentent les crimes, les points rouges représentent les écoles et le cercle bleu foncé délimite un périmètre de 500 m de chaque école. Cette figure nous donne une idée sur la répartition des crimes aux entourages des écoles.

B : Transformations et codage des variables :

- Nous avons commencé par transformer les données textuelles des **heures après l'école "After School Hours"** et des **heures scolaires** "**School hours"** en **minutes**.
  3:00 - 5:00 PM => 120

- Après nous avons transformé les variables catégorielles en valeurs numériques :
  Y => 1
  N => 0
  True => 1

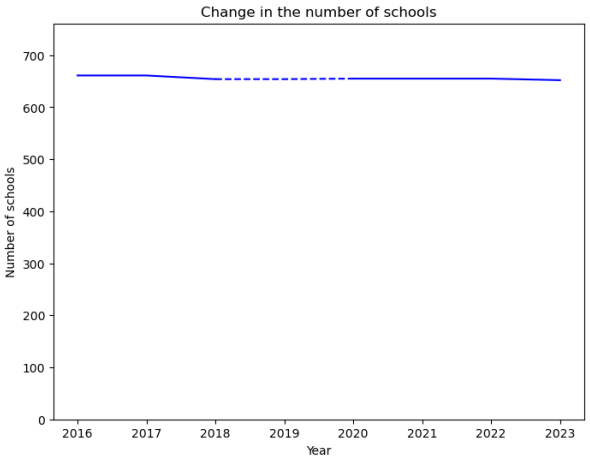
False => 0

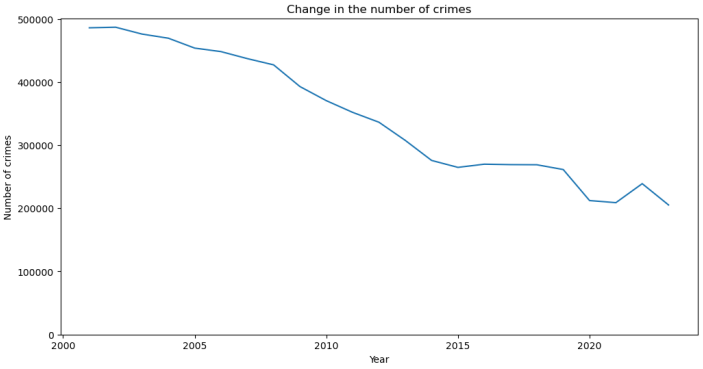
- Après nous avons converti le nombre total d'élèves en un format entier.

  Les nombres représentant le total des étudiants, initialement formatés avec des **virgules** pour les milliers, ont été nettoyés en supprimant les virgules et convertis en entiers pour une manipulation numérique précise

  “data['Student_Count_Total'] = data['Student_Count_Total'].str.replace(',', '').astype(int)”

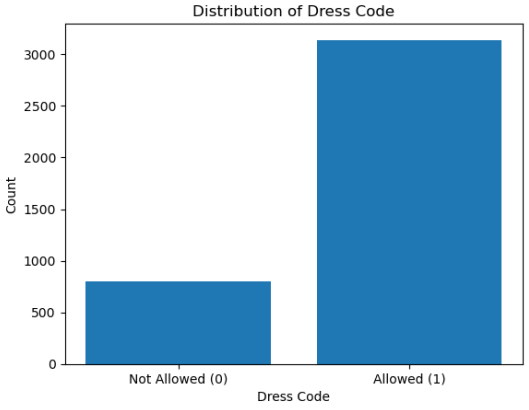
III - Description Statistiques des Données :

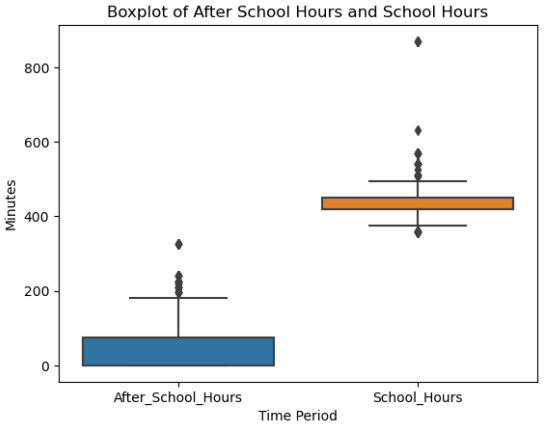




Nous avons commencé par visualiser les variations dans le nombre des écoles et le taux de criminalité. On remarque que le nombre d'écoles reste presque constant, cependant le nombre des crimes diminue significativement au fil des années (presque 300000).

Visualisation des variables dress code, school hours et after school hours après codage :

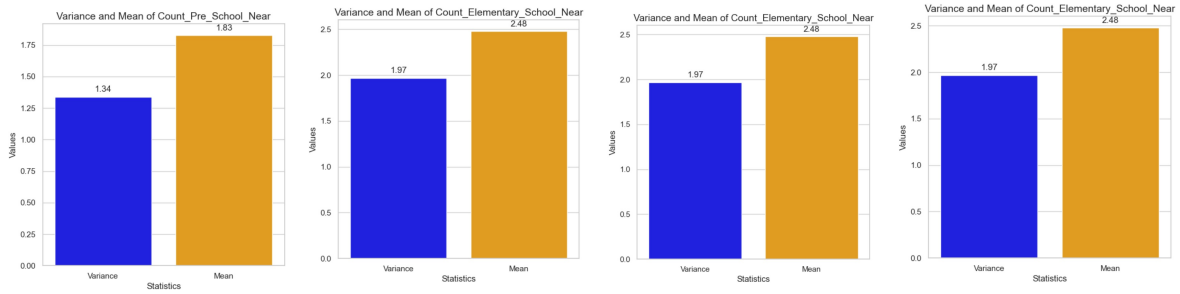




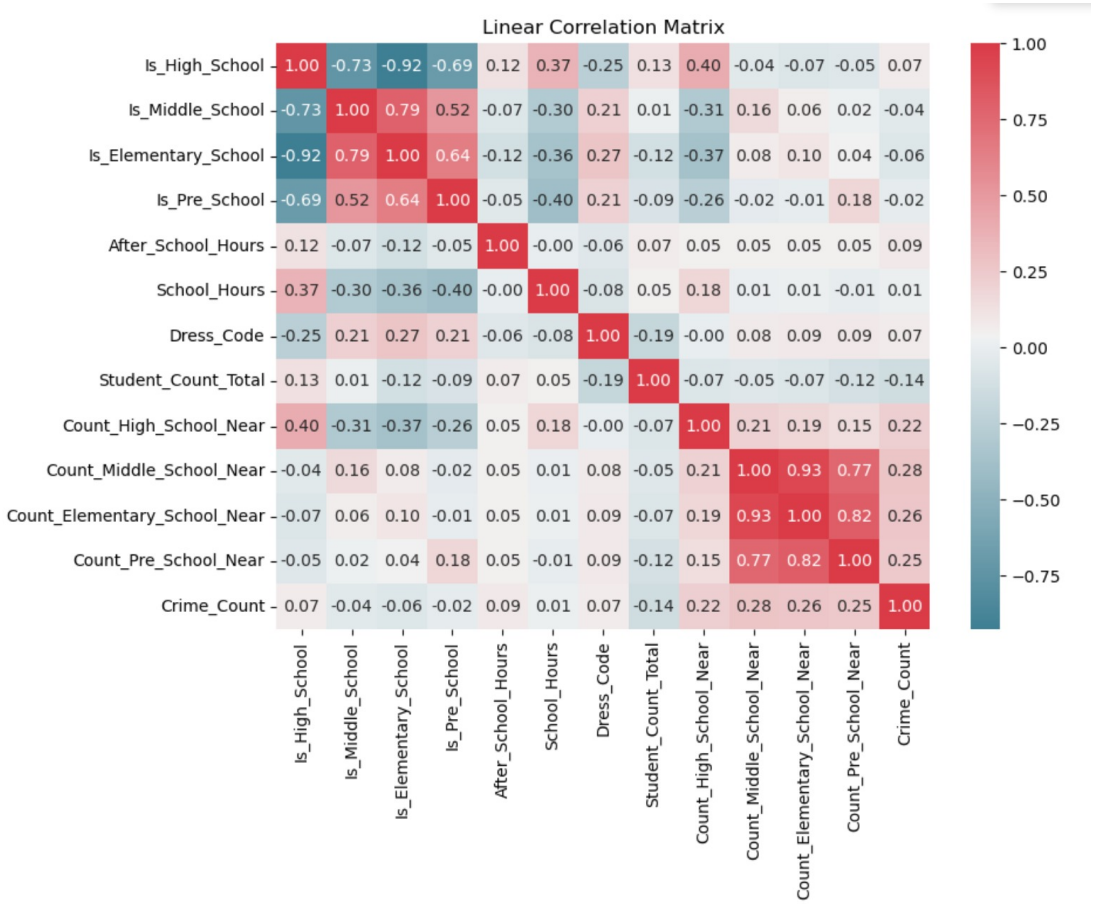
Description statistiques des variables numériques :

	After_School_Hours	School_Hours	Dress_Code	Student_Count_Total	Count_High_School_Near	Count_Middle_School_Near	Count_Elementary_School_Near	Count_Pre_School_Near	Crime_Count
count	3941.000000	3941.000000	3939.000000	3941.000000	3941.000000	3941.000000	3941.000000	3941.000000	3941.000000
mean	36.107079	437.644253	0.796903	521.949505	3.232175	7.483380	7.926161	5.915250	3738.566354
std	61.808392	31.113617	0.402355	402.726547	2.087342	3.283869	3.341764	2.626577	2122.109663
min	0.000000	357.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	135.000000
25%	0.000000	420.000000	1.000000	275.000000	2.000000	5.000000	6.000000	4.000000	2313.000000
50%	0.000000	420.000000	1.000000	429.000000	3.000000	7.000000	8.000000	6.000000	3398.000000
75%	75.000000	450.000000	1.000000	649.000000	4.000000	9.000000	10.000000	8.000000	4946.000000
max	325.000000	870.000000	1.000000	4514.000000	11.000000	19.000000	21.000000	18.000000	17334.000000

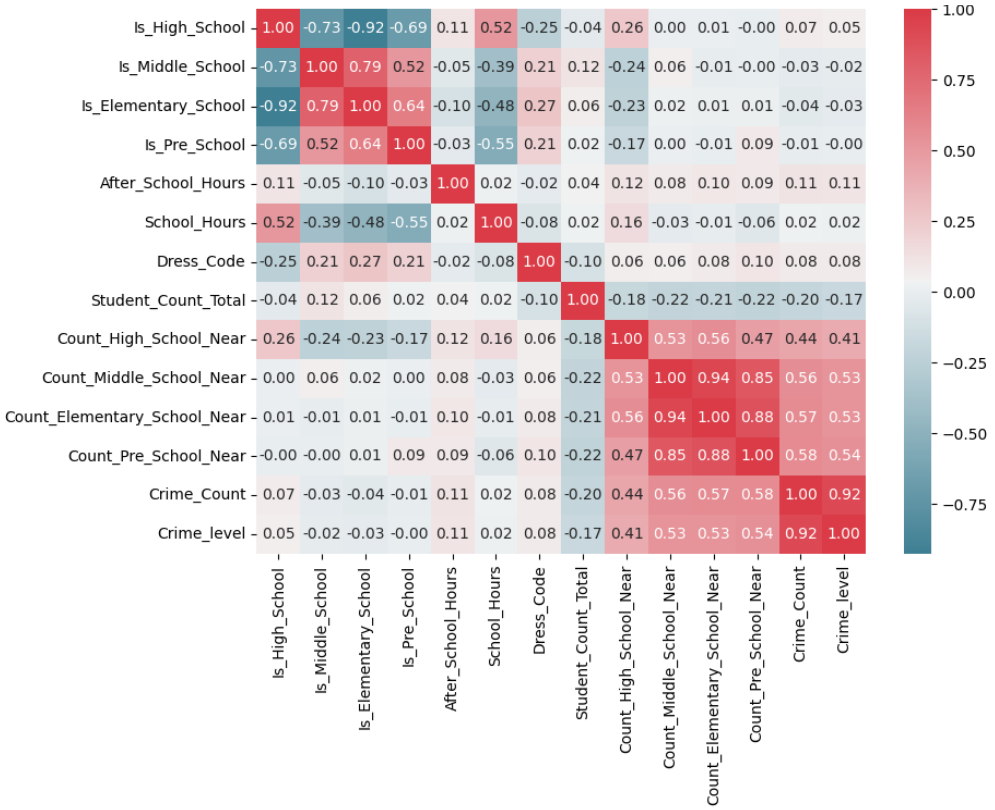
Variance et moyenne des variables significatives :



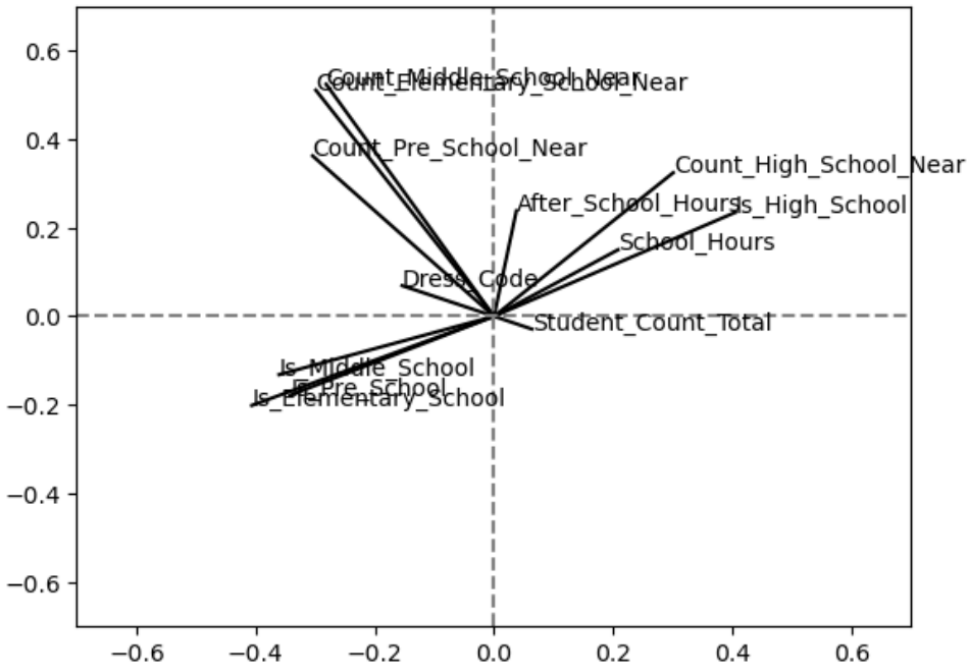
Etude des corrélations entre les variables :



La matrice (method spearman) ci-dessus illustre les corrélations entre les variables d'entrée et de sortie (crime_count) dans le cas d'un cercle de 500 m.
Les coefficients de corrélation obtenus ne montrent pas une grande corrélation surtout entre schools hours, nature de l'école et le taux de crime.
Pour cela, nous avons décidé d'élargir notre cercle et étudier le cercle de rayon 1 Km autour de l'école. Et voici la matrice de corrélation correspondante :



En modifiant le périmètre du rayon à 1 Km autour de l'école, on remarque que les coefficients de corrélation ont légèrement augmenté. Cette extension pourrait fournir des insights supplémentaires pour mieux comprendre les facteurs liés au taux de criminalité dans les environs de l'école.


Count_Middle_School_Near
Count_Elementary_School_Near
Count_Pre_School_Near
Count_High_School_Near
After_School_Hours
Is_High_School
School_Hours
Dress_Code
Student_Count_Total
Is_Middle_School
Is_Pre_School
Is_Elementary_School
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

Références :

1 : Susan Manheimer and Joshua Spaulding, “After School: The Prime Time for Juvenile Crime—Partnering with After-School Programs to Reduce Crime, Victimization, and Risky Behaviors Among Youth,” Police Chief Online, August 5, 2020.

2 : Laura , “The Effect Of School Uniforms On Crime” Leicestershirevillages, November 29, 2021

3 : Dale Willits ,“Schools, Neighborhood Risk Factors, and Crime”, SagePub ,2013

4 : Crews, G. (2009). Education and crime. In J. M. Miller 21st Century criminology: A reference handbook (pp. 59-66). SAGE Publications, Inc., https://www.doi.org/10.4135/9781412971997.n8