1.5em 0pt

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediene
Faculté d'Informatique
Département IA et SD

Rapport Pour La Partie 01 De Projet

Module : Data Mining

Exploitation des données et Extraction des règles d'associations

Travail présenté par :

- **BENKOUITEN Aymen——————191931046409**
- **NOUGHI Tarek——————191931041952**

2023/2024

Table des matières

I Données statistiques **2**

0.1 Objectifs : . . . 3
0.2 Manipulation de dataset-1 : . . . 3
0.2.1 Fourniture d'une description globale du dataset-1 : . . . 3
0.2.2 Fourniture d'une description détaillée de chaque attribut : . . . 3
0.3 Analyse des caractéristiques des attributs du dataset . . . 4
0.3.1 Calcul des mesures de tendance centrale pour déduire les symétries . . . 4
0.3.2 Construction des boîtes à moustache pour la détections des données aberrante . . . 5
0.3.3 Construction des histogrammes des données . . . 6

II Données temporelles **7**

0.4 Manipulation du dataset . . . 8
0.4.1 Importation et visualisation du contenu du dataset . . . 8
0.4.2 Description globale du dataset . . . 8
0.4.3 Description de chaque attribut du dataset . . . 9
0.4.4 Les graphes de distribution pour chaque colonne numérique du notre Dataset : . . . 10
0.4.5 Les Histogrammes pour chaque colonne numérique du notre Dataset : . . . 12
0.4.6 Les boîtes à moustaches pour chaque colonne numérique du notre Dataset : 14
0.4.7 Matrice de corrélation entre les attributs de notre dataset : . . . 16
0.5 Prétraitement . . . 17
0.5.1 Traitement des valeurs manquantes et Date Formats : . . . 17
0.5.2 Traitement des valeurs aberrantes . . . 18
0.6 Visualisation . . . 25
0.6.1 Distribution des Cas Confirmés et Tests Positifs par Zones . . . 25
0.6.2 Évolution Temporelle des Tests COVID-19, des Tests Positifs et du Nombre de Cas pour une Zone Spécifique : . . . 28
0.6.3 Distribution des Cas COVID-19 Positifs par Zone et par Année : . . . 30
0.6.4 Relation Graphique entre la Population et le Nombre de Tests Effectués : 31
0.6.5 Zones les Plus Fortement Impactées par le Coronavirus : . . . 32
0.6.6 Le rapport entre les cas confirmés, les tests effectués et les tests positifs au fil du temps pour chaque zone : . . . 32

III Extraction de motifs fréquents, règles d'associations et corrélations **34**
0.7 Introduction . . . 35
0.8 Analyse globale du dataset et prétraitement . . . 35
0.8.1 Description globale du dataset . . . 35
0.8.2 Description de chaque attribut . . . 35
0.8.3 Traitement des valeurs manquantes . . . 36
0.8.4 Traitement des valeurs aberrantes . . . 37
0.8.5 Réduction des données via la discrétisation . . . 37
0.9 Extraction des motifs fréquents en utilisant Apriori . . . 37
0.9.1 Préparation des données transactionnelles . . . 37
0.9.2 Création des tables Ck et Lk et génération des motifs fréquents . . . 38
0.9.3 Extraction des règles d'associations et corrélations . . . 39
0.9.4 Mesures de corrélation des règles d'association . . . 40
0.9.5 Sélection des Règles d'Association Fortes . . . 41
0.9.6 Expérimentation avec Min-Supp et Min-Conf . . . 41
0.10 Conclusion . . . 42

Table des figures

1 la visualisation des primières lignes de notre dataset 8
2 Description globale de notre dataset 8
3 Description de chaque attribut de notre dataset 9
4 Les tendances centrale des attributs de dataset 9
5 Description de chaque attribut de notre dataset 11
6 Histogramme de chaque attribut de notre dataset 13
7 Les boîtes à moustaches de chaque attribut de notre dataset 15
8 Matrice de corrélation 16
9 Notre dataset avant le clennage. 17
10 Notre dataset après le clennage. 18
11 Les boîtes à moustaches de chaque attribut de notre dataset 19
12 Les boîtes à moustaches (médiane) de chaque attribut de notre dataset 21
13 Les boîtes à moustaches (moyenne) de chaque attribut de notre dataset 23
14 La boîte à moustache de l'attribut "case count" 24
15 Le boîte à moustache (moyenne) de l'attribut "case count" 24
16 Nombre total de cas confirmés par zone 25
17 Nombre total de tests positifs par zone 26
18 Répartition des cas confirmés par zone (Tree Map) 27
19 Répartition de tests positifs par zone (Tree Map) 28
20 Évolution hebdomadaire des tests COVID-19, tests positifs et cas confirmés . . . 29
21 Évolution mensuelle des tests COVID-19, tests positifs et cas confirmés 29
22 Évolution annuelle des tests COVID-19, tests positifs et cas confirmés 30
23 Répartition des cas COVID-19 positifs par zone et par année 31
24 Rapport entre la population et le nombre de tests effectués 32
25 Zones les Plus Fortement Impactées par le Coronavirus 32
26 Le rapport entre les cas confirmés, les tests effectués et les tests positifs au fil du temps pour la zone "95128" 33
27 Description globale du dataset 35
28 Description globale du dataset 36
29 Distribution des valeurs manquantes 36
30 Distribution des valeurs aberrantes 37
31 "get$_{ki}temsets"function$ 38
32 "generate$_{r}ules"function$ 39
33 "correlation" function 40
34 "correlation" function 42

Liste des tableaux

1 Y– Description de dataset-1- 3
2 Description des collones avec Non-Null et Count et Dtype 4
3 Description statistique des attributs 5

Introduction Générale

Dans l'ère actuelle, le flux de données générées quotidiennement atteint des proportions monumentales, provenant d'une multitude de sources telles que les réseaux sociaux, les transactions commerciales et les capteurs IoT, parmi d'autres. Ces données représentent une richesse d'informations inestimable, mais leur potentiel reste souvent sous-exploité faute de techniques appropriées pour en extraire des connaissances exploitables.

Ce projet se concentre sur l'exploration des mécanismes d'exploitation des données, mettant en lumière l'extraction de règles d'associations. Il est articulé en deux phases distinctes : la première étape s'attache à l'analyse et au prétraitement des données. La seconde phase vise à extraire les motifs fréquents ainsi que les règles d'association.

Pour cette étude, trois ensembles de données sont mobilisés : deux d'entre eux sont dédiés à la première phase, couvrant respectivement les données statiques et temporelles. Le troisième ensemble de données est exclusivement réservé à la seconde phase de l'analyse.

Première partie

Données statistiques

0.1 Objectifs :

Cette section vise à établir clairement les objectifs fondamentaux de l'analyse des données statiques. L'analyse approfondie et le nettoyage du Dataset 1 sont des étapes cruciales pour préparer les données en vue de leur utilisation ultérieure dans la classification et le clustering. Les objectifs spécifiques comprennent :

0.2 Manipulation de dataset-1 :

N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	OM	Fertility
138	8.6	560	7.46	0.62	0.7	5.9	0.24	0.31	0.77	8.71	0.11	1.204	0
213	7.5	338	7.62	0.75	1.06	25.4	0.3	0.86	1.54	2.89	2.29	1.8232	0
163	9.6	718	7.59	0.51	1.11	14.3	0.3	0.86	1.57	2.7	2.03	1.9092	0
157	6.8	475	7.64	0.58	0.94	26	0.34	0.54	1.53	2.65	1.82	1.6168	0
270	9.9	444	7.63	0.4	0.86	11.8	0.25	0.76	1.69	2.43	2.26	1.4792	1
220	8.6	444	7.43	0.65	0.72	11.7	0.37	0.66	0.9	2.19	1.82	1.2384	0

TABLE 1 – Y– Description de dataset-1-

0.2.1 Fourniture d'une description globale du dataset-1 :

Le dataset-1 comprend des informations relatives aux caractéristiques du sol, représentées par différentes colonnes : N, P, K, pH, EC, OC, S, Zn, Fe, Cu, Mn, B, OM, et Fertility. Chaque ligne représente des mesures spécifiques de ces attributs pour des échantillons de sol différents.

Fourniture d'une description détaillée de chaque attribut :

0.2.2 Fourniture d'une description détaillée de chaque attribut :

N : Niveau de Nitrogen (N)

P : Niveau de Phosphorus (P)

K : Niveau de Potassium (K)

pH : Valeur du pH du sol

EC : Conductivité électrique du sol

OC : Carbone organique du sol

S : Soufre (S)

Zn : Zinc (Zn)

Fe : Fer (Fe)

Cu : Cuivre (Cu)

Mn : Manganèse (Mn)

B : Bore (B)

OM : Matière organique (OM)

Fertility : Indicateur de fertilité du sol

Chaque attribut représente une mesure spécifique des composants ou des propriétés du sol, fournissant une gamme complète d'informations sur la composition et les caractéristiques des échantillons de sol dans le dataset-1.

Column	Non-Null Count	Dtype
N	885	int64
P	885	object
K	885	int64
pH	885	float64
EC	885	float64
OC	884	float64
S	885	float64
Zn	885	float64
Fe	885	float64
Cu	884	float64
Mn	885	float64
B	885	float64
OM	885	float64
Fertility	885	int64

TABLE 2 – Description des collones avec Non-Null et Count et Dtype

0.3 Analyse des caractéristiques des attributs du dataset

0.3.1 Calcul des mesures de tendance centrale pour déduire les symétries

Attribut	Moyenne	Médiane	Mode	Max	Min	q0	q1	q2	q3	q4
N	247.0	257	207	383	6	6	201.0	257	307.0	307.0
P	14.56	7.5	8.3	125.0	2.9	2.9	12.4	7.5	10.6	10.7
K	501.34	475	444	1560	11	11	412.0	475	581.0	581.0
pH	7.51	7.5	7.5	11.15	0.9	0.9	7.35	7.5	7.63	7.63
EC	0.54	0.55	0.62	0.95	0.1	0.1	0.43	0.55	0.64	0.64
OC	0.62	0.68	0.88	24.0	0.1	0.1	0.39	0.68	1.07	1.07
S	7.55	6.64	5.13	31.0	0.64	0.64	4.7	6.64	8.75	8.75
Zn	0.47	0.36	0.28	42.0	0.07	0.07	0.28	0.36	0.47	0.47
Fe	4.13	3.56	6.32	44.0	0.21	0.21	2.035	3.56	6.31	6.32
Cu	0.95	0.93	1.25	3.02	0.09	0.09	0.63	0.93	1.25	1.25
Mn	8.65	8.34	7.54	31.0	0.11	0.11	6.21	8.34	11.45	11.48
B	0.59	0.41	0.34	2.82	0.06	0.06	0.27	0.41	0.61	0.61
OM	1.06	1.01	1.51	41.28	0.172	0.172	0.6536	1.0148	1.34	1.34
Fertility	0.59	1	1	2	0	0	0.0	1	1.0	1.0

TABLE 3 – Description statistique des attributs

Les symétriques sont les colonnes où la moyenne = la médiane = le mode, pour conclure ça visuellement nous exploitons les graphes de densité pour chaque attribut et on tire les symétriques on au aura donc 14 graphes.

Analyse :
Les graphes de densité révèlent une observation intéressante : les attributs K, EC, Mn et OC affichent des valeurs de moyenne, médiane et mode presque égales, soulignant ainsi une grande similitude dans leurs distributions respectives. En revanche, pour pH, Zn et OM, la similitude entre la moyenne, médiane et mode reflète des distributions très proches.

Conclusion :
Cette observation suggère une uniformité marquée dans les valeurs de K, EC, Mn et OC, suggérant une stabilité ou une homogénéité remarquable dans ces aspects du sol. Cependant, pour pH, Zn et OM, la symétrie des distributions indique une dispersion équilibrée des valeurs autour de la moyenne, dénotant une variabilité plus étendue dans ces paramètres du sol.

0.3.2 Construction des boîtes à moustache pour la détections des données aberrante

Analyse :
Après l'analyse des boîtes à moustaches pour chaque attribut, on remarque la présence de valeurs aberrantes dans l'ensemble des colonnes. Cependant, certaines colonnes se démarquent par un nombre significativement plus élevé de valeurs aberrantes, notamment N, K, pH, S, Mn et B. En revanche, les colonnes EC, Zn, Fe et OM affichent une fréquence moindre de valeurs aberrantes.

Conclusion :

En conclusion, une gestion et une interprétation appropriées des données devraient prendre en considération ces variations dans la détection des valeurs aberrantes, en accordant une attention particulière aux attributs présentant une instabilité plus prononcée. Cette observation pourrait guider les analyses futures et les mesures correctives visant à assurer une interprétation précise des caractéristiques du sol.
Il devient crucial de réfléchir à des solutions pour gérer ces valeurs aberrantes, une étape que nous explorerons dans un futur proche pour garantir la fiabilité des analyses et des conclusions tirées de ce jeu de données.

0.3.3 Construction des histogrammes des données

Les histogrammes de données offrent une analyse polyvalente, apportant divers éclairages :

- **Tendance et symétrie des données :** Les histogrammes permettent une évaluation rapide des tendances, du mode et de la symétrie des données. Par exemple, la confirmation de la symétrie pour pH, OM et Zn souligne une cohérence notable dans ces données.
- **Centrage et dispersion :** Ils fournissent une représentation visuelle claire du centrage (moyenne, médiane) ainsi que de la dispersion des données, offrant une vue immédiate de la variabilité des valeurs analysées.
- **Identification des anomalies :** Ces graphiques permettent de distinguer les valeurs rares, les plus fréquentes et les anomalies. Par exemple, ils mettent en lumière une présence très limitée de données à des valeurs spécifiques, comme 30 pour Fe et 3 pour Cu, soulignant des points singuliers dans le jeu de données.

Deuxième partie

Données temporelles

0.4 Manipulation du dataset

0.4.1 Importation et visualisation du contenu du dataset

	zcta	time_period	population	Start date	end date	case count	test count	positive tests	case rate	test rate	positivity rate
0	95129	32	39741	2020-10-11	2020-10-31	22.0	2543.0	23.0	2.6	304.7	0.9
1	95129	43	39741	2021-05-30	2021-06-19	NaN	3315.0	14.0	1.1	397.2	0.4
2	95129	40	39741	2021-03-28	2021-04-17	34.0	4816.0	37.0	4.1	577.1	0.8
3	95129	55	39741	2022-02-06	2022-02-26	110.0	10194.0	175.0	13.2	1221.5	1.7
4	95129	44	39741	2021-06-20	2021-07-10	14.0	3033.0	17.0	1.7	363.4	0.6
5	95129	54	39741	2022-01-16	2022-02-05	624.0	13479.0	817.0	74.8	1615.1	6.1
6	95129	25	39741	2020-05-17	2020-06-06	NaN	762.0	NaN	0.4	91.3	0.4
7	95129	30	39741	2020-08-30	2020-09-19	20.0	1773.0	20.0	2.4	212.4	1.1
8	95129	31	39741	2020-09-20	2020-10-10	12.0	2120.0	12.0	1.4	254.0	0.6
9	95129	66	39741	2022-09-25	2022-10-15	66.0	1571.0	78.0	7.9	188.2	5.0

FIGURE 1 – la visualisation des primières lignes de notre dataset

Les données ci-dessous représentent les statistiques des cas de COVID-19 pour plusieurs zones sur différentes périodes de temps. Chaque ligne correspond à une période spécifique et contient des informations telles que le nombre de cas, les tests effectués et le taux de positivité... et si l'on voulait commenter la première ligne de nos données, on pourrait dire :

- **Ligne 1 (10/11/2020 - 10/31/2020) :** Durant cette période, la zone ZCTA 95129 a enregistré 22 cas confirmés de COVID-19, sur un total de 2543 tests réalisés, entraînant un taux de positivité de 0.9%. Ces chiffres soulignent une prévalence relativement basse du virus, avec un nombre de cas proportionnellement faible par rapport aux tests effectués.

0.4.2 Description globale du dataset

	zcta	time_period	population	case count	test count	positive tests	case rate	test rate	positivity rate
count	337.00	337.00	337.00	311.00	325.00	310.00	337.00	337.00	337.00
mean	94663.60	43.69	50260.55	225.99	4938.12	380.20	19.39	454.84	5.83
std	506.65	15.22	17632.83	401.76	3672.16	2027.55	32.59	311.01	9.22
min	94085.00	18.00	23223.00	0.00	11.00	11.00	0.00	0.10	0.00
25%	94086.00	31.00	36975.00	39.50	2428.00	47.25	3.30	249.70	1.30
50%	95035.00	43.00	50477.00	91.00	4352.00	108.50	8.10	427.10	3.00
75%	95128.00	56.00	66256.00	235.00	6659.00	282.00	19.10	614.90	6.60
max	95129.00	155.00	79655.00	3627.00	20177.00	35000.00	260.70	1615.10	100.00

FIGURE 2 – Description globale de notre dataset

Ces statistiques offrent une vue d'ensemble des valeurs centrales et de la dispersion des données, permettant de mieux comprendre la variabilité des cas de COVID-19 dans les différentes zones étudiées.

0.4.3 Description de chaque attribut du dataset

```
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   zcta            337 non-null    int64
 1   time_period     337 non-null    int64
 2   population      337 non-null    int64
 3   Start date      191 non-null    datetime64[ns]
 4   end date        191 non-null    datetime64[ns]
 5   case count      311 non-null    float64
 6   test count      325 non-null    float64
 7   positive tests  310 non-null    float64
 8   case rate       337 non-null    float64
 9   test rate       337 non-null    float64
 10  positivity rate 337 non-null    float64
dtypes: datetime64[ns](2), float64(6), int64(3)
memory usage: 29.1 KB
```

```
zcta              0
time_period       0
population        0
Start date      146
end date        146
case count       26
test count       12
positive tests   27
case rate         0
test rate         0
positivity rate   0
dtype: int64
```

FIGURE 3 – Description de chaque attribut de notre dataset

Cette description présente les caractéristiques des colonnes du jeu de données sur les cas de COVID-19 pour plusieurs zones :

- **ZCTA, Time Period, Population, Case Rate, Test Rate, Positivity Rate :** Ces colonnes n'ont aucune valeur manquante, ce qui signifie qu'elles sont complètes pour toutes les entrées (0 valeurs manquantes).
- **Case Count, Test Count, Positive Tests :** Ces colonnes ont un nombre relativement plus faible de valeurs manquantes, avec 26, 12 et 27 valeurs manquantes respectivement. Ces données manquantes pourraient nécessiter une investigation plus approfondie pour comprendre pourquoi elles sont absentes et décider de la meilleure approche pour traiter ces lacunes dans l'analyse.

Voici les tendances centrale des attributs de notre dataset :

Attribut	Moyenne	Médiane	Mode	Max	Min	q1	q2	q3
zcta	94680.87	95035.0	95127	95129	94085	94086.0	95035.0	95128.0
time_period	45.75	46.0	26	155	21	34.0	46.0	57.0
population	51116.1	50477.0	66256	79655	23223	36975.0	50477.0	66256.0
case count	231.95	95.0	12.0	3627.0	12.0	42.0	95.0	243.5
test count	5208.56	4525.0	1295.0	20177.0	31.0	2675.5	4525.0	6871.0
positive tests	388.56	112.0	20.0	35000.0	12.0	51.0	112.0	284.5
case rate	21.5	9.3	2.9	260.7	0.7	4.75	9.3	22.9
test rate	495.34	458.8	265.5	1615.1	2.2	279.2	458.8	644.6
positivity rate	5.11	3.2	1.1	64.5	0.3	1.4	3.2	6.6

FIGURE 4 – Les tendances centrale des attributs de dataset

0.4.4 Les graphes de distribution pour chaque colonne numérique du notre Dataset :

Cet figure montre une série de graphiques de distribution pour chaque colonne numérique du notre Dataset, mettant en évidence la forme et la répartition des données, ainsi que les mesures centrales (moyenne, médiane et mode) pour chaque distribution. Cela permet d'avoir une vue d'ensemble des distributions des différentes variables numériques dans notre dataset.

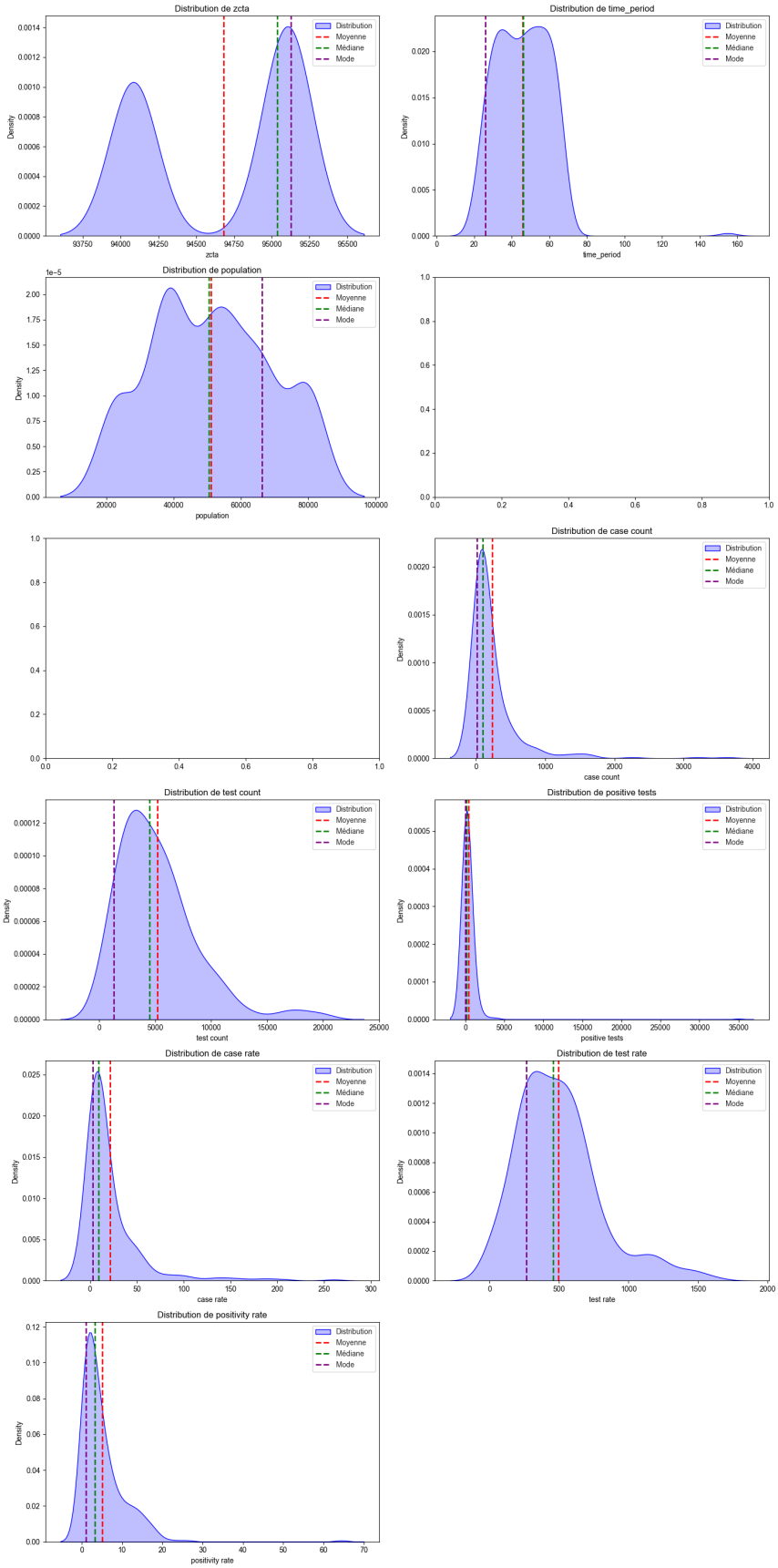

FIGURE 5 – Description de chaque attribut de notre dataset

0.4.5 Les Histogrammes pour chaque colonne numérique du notre Dataset :

Ces histogrammes donnent un aperçu visuel des distributions des différentes variables numériques dans notre dataset, aidant à comprendre la répartition des valeurs et la concentration autour de certaines plages.

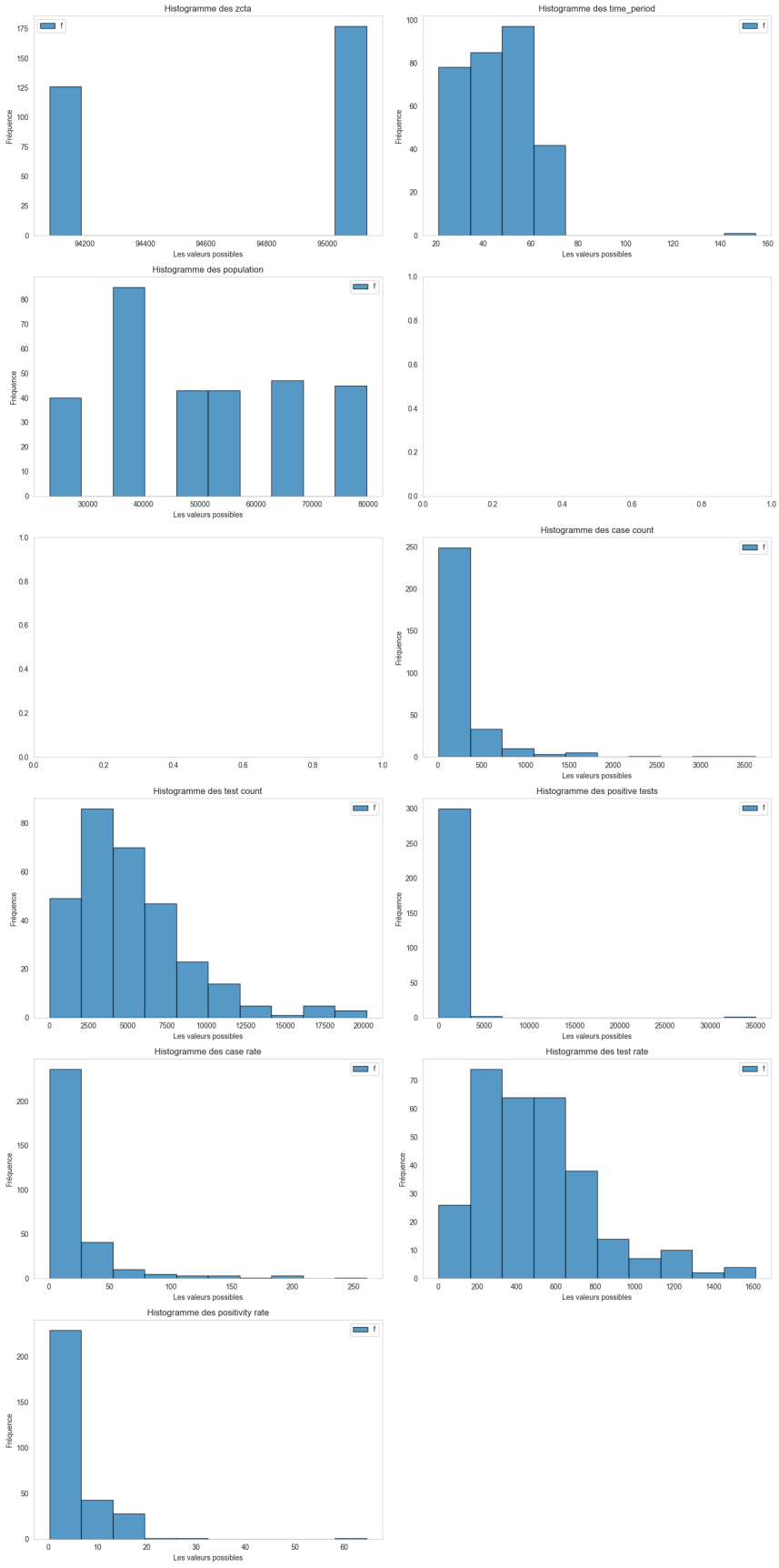

FIGURE 6 – Histogramme de chaque attribut de notre dataset

0.4.6 Les boîtes à moustaches pour chaque colonne numérique du notre Dataset :

Les boîtes à moustaches sont utiles pour visualiser la dispersion des données, la médiane et les quartiles, ainsi que pour identifier les valeurs aberrantes potentielles dans les distributions des différentes variables numériques.

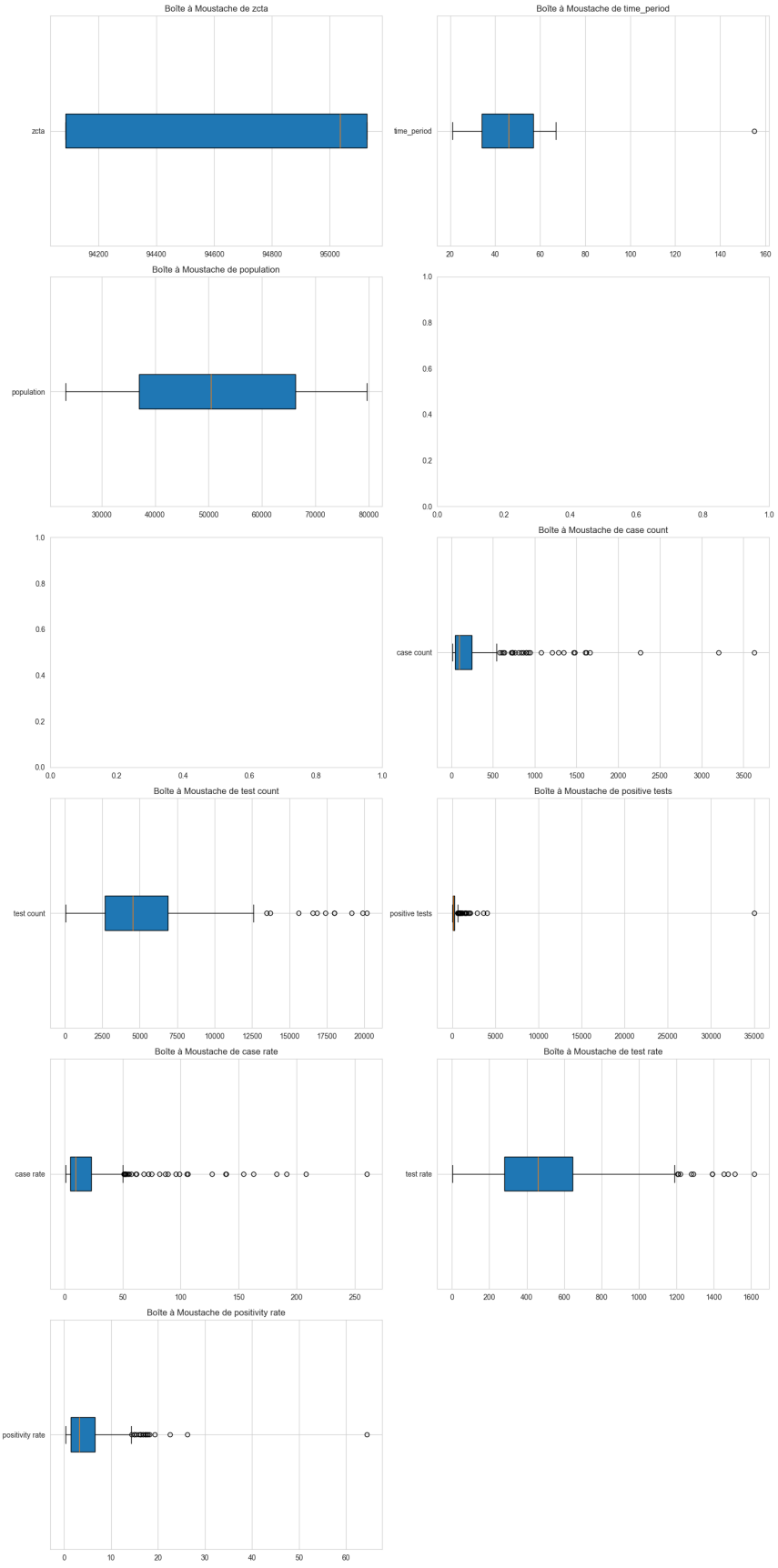

FIGURE 7 – Les boîtes à moustaches de chaque attribut de notre dataset

0.4.7 Matrice de corrélation entre les attributs de notre dataset :

Les valeurs plus proches de 1 ou -1 indiquent une corrélation forte, tandis que les valeurs proches de 0 indiquent une corrélation faible.

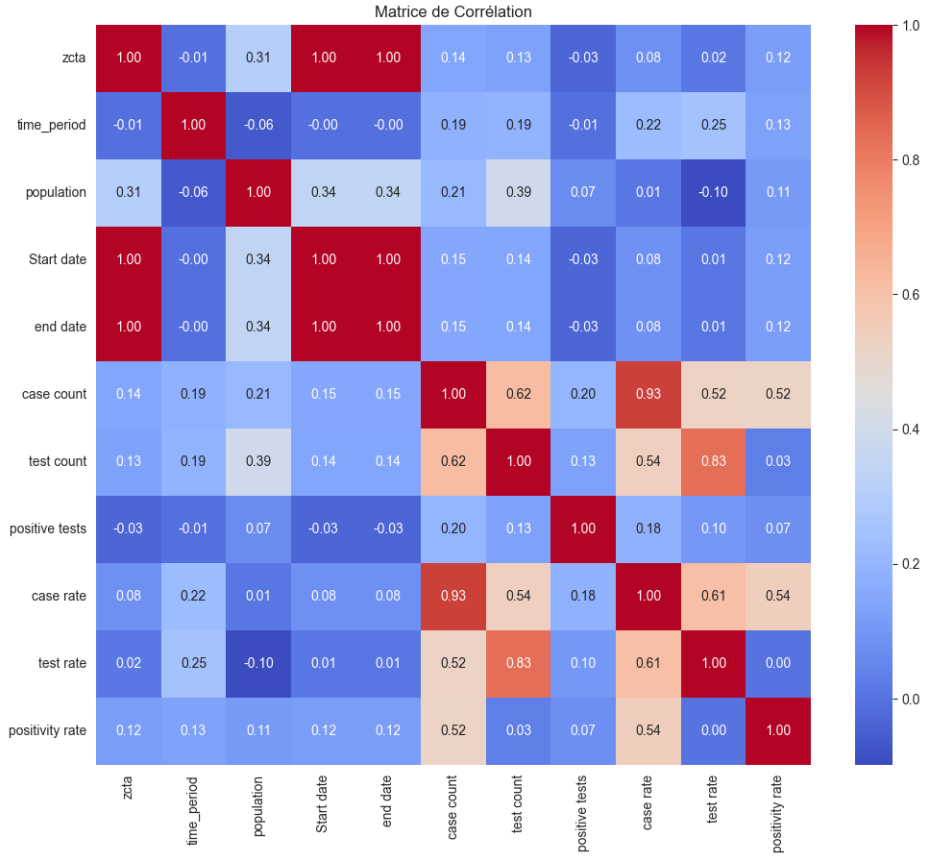

FIGURE 8 – Matrice de corrélation

La matrice de corrélation offre une vue synthétique des relations linéaires entre les variables. Cet outil simplifie l'identification des associations fortes ou faibles entre les données, facilitant ainsi la sélection ciblée des variables pour des analyses approfondies, contribuant ainsi à une prise de décision éclairée.

0.5 Prétraitement

0.5.1 Traitement des valeurs manquantes et Date Formats :

	zcta	time_period	population	Start date	end date	case count	test count	positive tests	case rate	test rate	positivity rate
0	95129	32	39741	10/11/2020	10/31/2020	22.0	2543.0	23.0	2.6	304.7	0.9
1	95129	43	39741	5/30/2021	6/19/2021	NaN	3315.0	14.0	1.1	397.2	0.4
2	95129	40	39741	3/28/2021	4/17/2021	34.0	4816.0	37.0	4.1	577.1	0.8
3	95129	55	39741	2/6/2022	2/26/2022	110.0	10194.0	175.0	13.2	1221.5	1.7
4	95129	44	39741	6/20/2021	7/10/2021	14.0	3033.0	17.0	1.7	363.4	0.6
...	...	...	...	...	...	...	...	...	...	...	...
332	94085	59	23223	1-May	21-May	165.0	2315.0	192.0	33.8	474.7	8.3
333	94085	63	23223	24-Jul	13-Aug	150.0	1348.0	190.0	30.8	276.4	14.1
334	94085	61	23223	12-Jun	2-Jul	219.0	1696.0	255.0	44.9	347.8	15.0
335	94085	27	23223	28-Jun	18-Jul	53.0	1379.0	61.0	10.9	282.8	4.4
336	94085	57	23223	20-Mar	9-Apr	30.0	1949.0	34.0	6.2	399.6	1.7

FIGURE 9 – Notre dataset avant le clennage.

L'intégration de données présentait une complexité liée à la variation des formats de date dans les colonnes **"Start date"** et **"end date"** de notre ensemble de données. Deux formats distincts étaient identifiés : **'%d-%b'** et **'%m/%d/%Y'**, mais des lacunes dans certaines données au format **'%d-%b'** privaient parfois ces entrées de l'information sur l'année.

Start date	min_period	max_period
2020-01-01	21	35
2021-01-01	36	53
2022-01-01	54	155

Ces données ont joué un rôle crucial dans la résolution de la diversité des formats de date. En se basant sur ces découvertes, les dates au format **'%d-%b'** ont été harmonisées en adoptant le format **['%m/%d/%Y']**, tout en comblant les lacunes avec les informations pertinentes extrapolées du scatter plot.

Cette approche méthodique a permis d'unifier et de rectifier les divergences de formats de date, assurant ainsi la cohérence et la précision des données pour une analyse future. Ce processus démontre l'efficacité d'une analyse minutieuse et de l'utilisation habile des insights obtenus pour résoudre des défis complexes de normalisation des données.

	zcta	time_period	population	Start date	end date	case count	test count	positive tests	case rate	test rate	positivity rate	year
0	95129	32	39741	2020-10-11	10/31/2020	22.0	2543.0	23.0	2.6	304.7	0.9	2020
2	95129	40	39741	2021-03-28	04/17/2021	34.0	4816.0	37.0	4.1	577.1	0.8	2021
3	95129	55	39741	2022-02-06	02/26/2022	110.0	10194.0	175.0	13.2	1221.5	1.7	2022
4	95129	44	39741	2021-06-20	07/10/2021	14.0	3033.0	17.0	1.7	363.4	0.6	2021
5	95129	54	39741	2022-01-16	02/05/2022	624.0	13479.0	817.0	74.8	1615.1	6.1	2022
...	...	...	...	...	...	...	...	...	...	...	...	...
332	94085	59	23223	2022-05-01	05/21/2022	165.0	2315.0	192.0	33.8	474.7	8.3	2022
333	94085	63	23223	2022-07-24	08/13/2022	150.0	1348.0	190.0	30.8	276.4	14.1	2022
334	94085	61	23223	2022-06-12	07/02/2022	219.0	1696.0	255.0	44.9	347.8	15.0	2022
335	94085	27	23223	2020-06-28	07/18/2020	53.0	1379.0	61.0	10.9	282.8	4.4	2020
336	94085	57	23223	2022-03-20	04/09/2022	30.0	1949.0	34.0	6.2	399.6	1.7	2022

FIGURE 10 – Notre dataset après le clennage.

0.5.2 Traitement des valeurs aberrantes

Dans le processus de traitement des valeurs aberrantes, une approche basée sur les mesures de tendance telles que la moyenne (Mean) et la médiane a été adoptée pour rectifier les données. Initialement, j'ai remplacé les valeurs aberrantes par la médiane, puis par la moyenne, afin de comparer les résultats et de déterminer la méthode la plus appropriée pour le traitement des valeurs aberrantes.

Avant le remplacement par la moyenne, les boxplots illustraient la distribution des données avec les valeurs aberrantes :

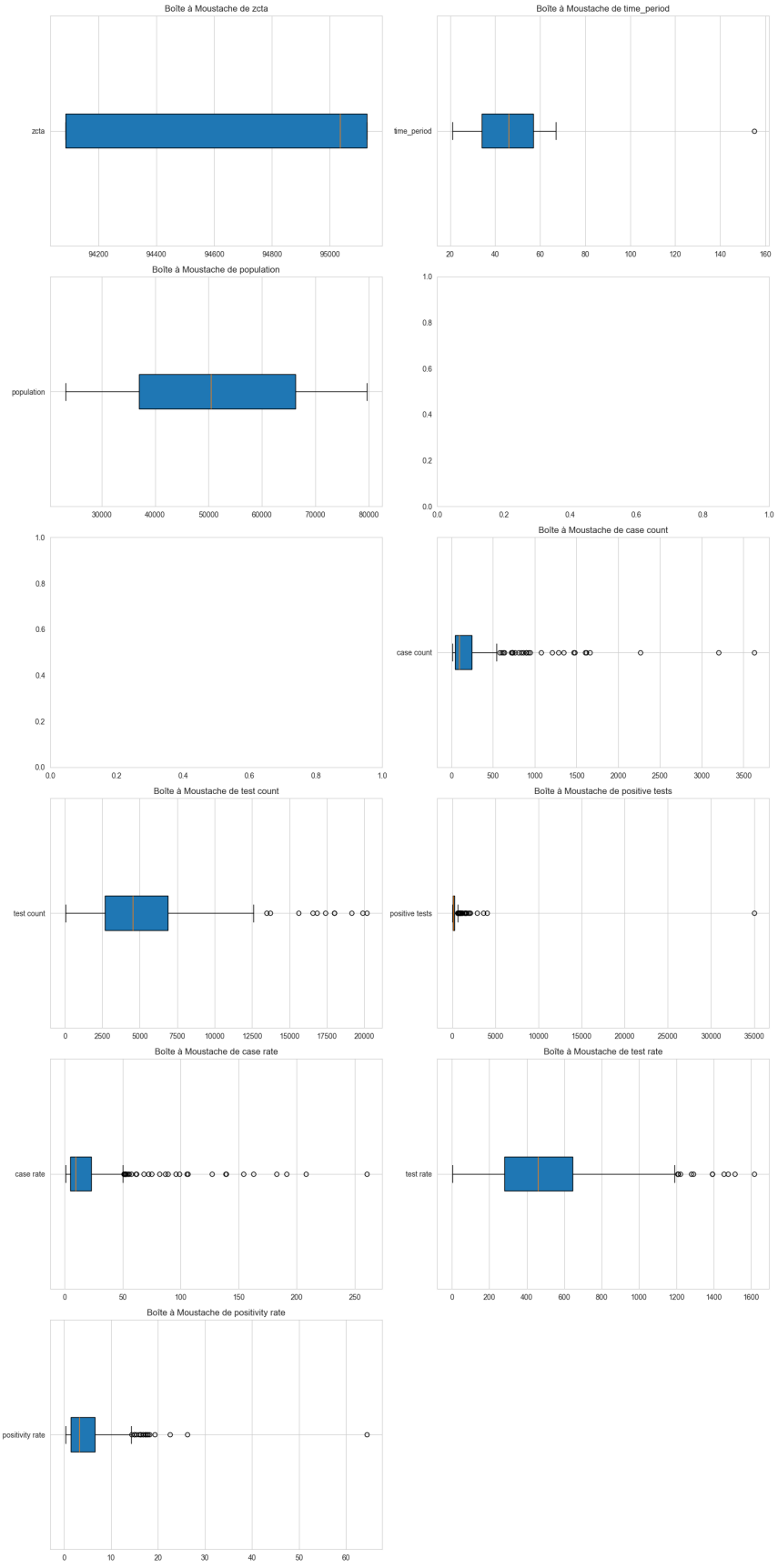

FIGURE 11 – Les boîtes à moustaches de chaque attribut de notre dataset

Cette représentation graphique montrait la dispersion des données, mettant en évidence les valeurs aberrantes avant tout traitement.

Après avoir remplacé ces valeurs aberrantes par la moyenne, les boxplots ont été actualisés, donnant lieu à une nouvelle visualisation :

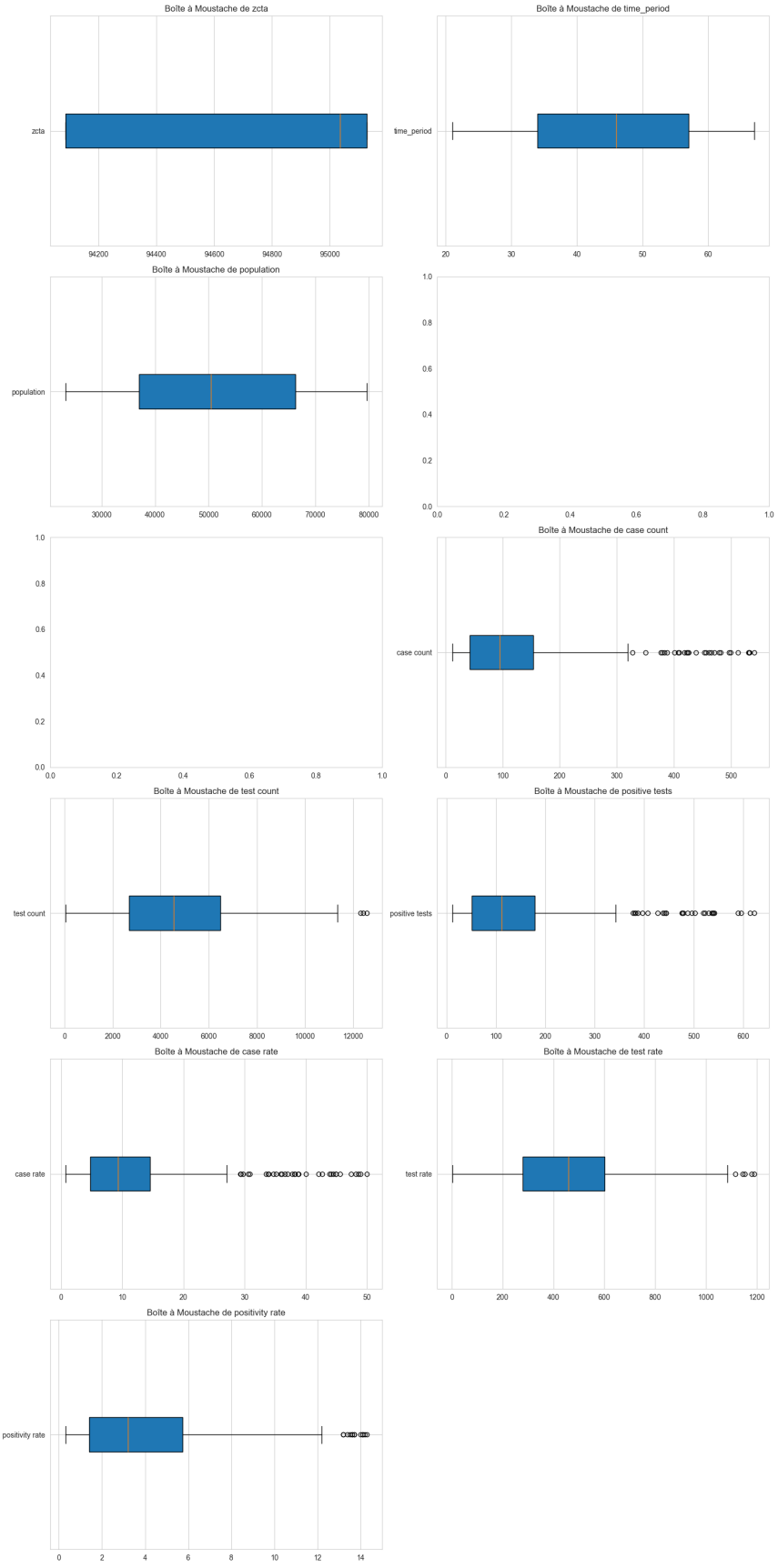

FIGURE 12 – Les boîtes à moustaches (médiane) de chaque attribut de notre dataset

Cette version modifiée des boxplots illustrait la distribution des données après le remplacement des valeurs aberrantes par le médiane.

Et enfin, après avoir remplacé ces valeurs aberrantes par la moyenne, les boxplots ont été actualisés, donnant lieu à une nouvelle visualisation :

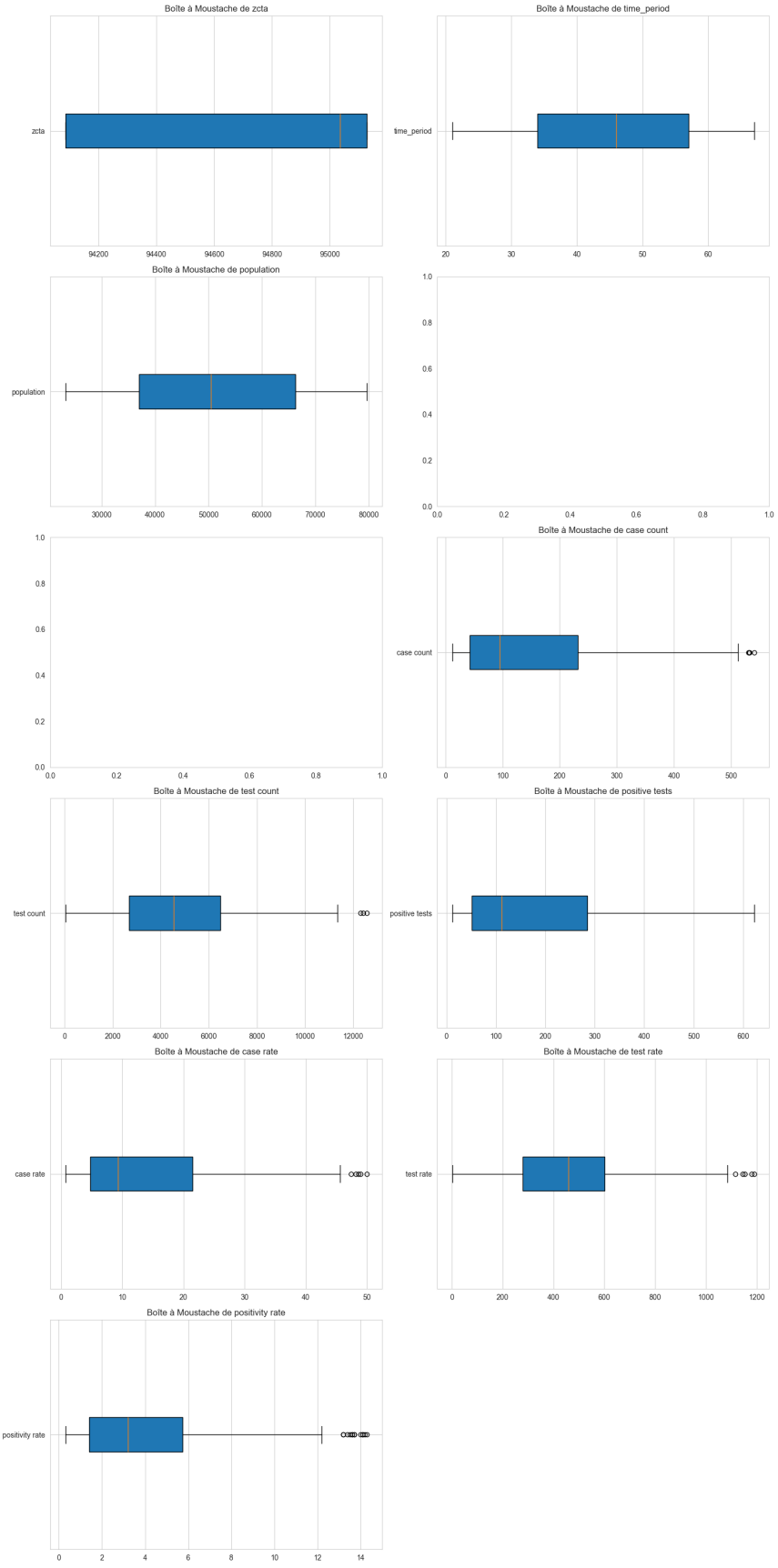

FIGURE 13 – Les boîtes à moustaches (moyenne) de chaque attribut de notre dataset

Cette version modifiée des boxplots illustrait la distribution des données après le remplacement des valeurs aberrantes par la moyenne.

Et si on prend un exemple plus précise, pour l'attribut **"case count"** , nous obtenons la boite à moustache suivante :

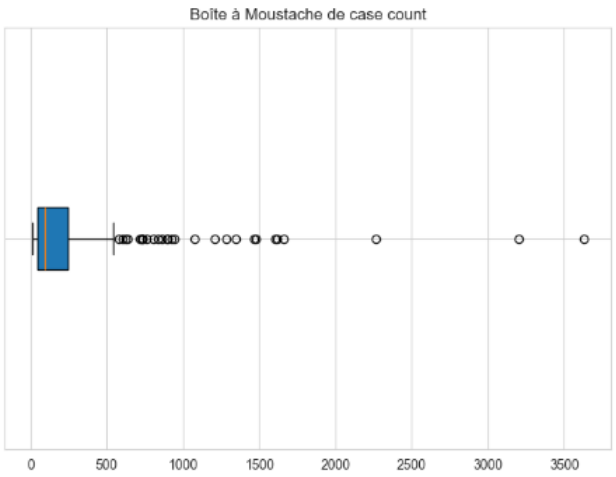

FIGURE 14 – La boîte à moustache de l'attribut "case count"

Et apres avoir remplacer ses valeurs avec la mesure de tendance mean on a obtenue :

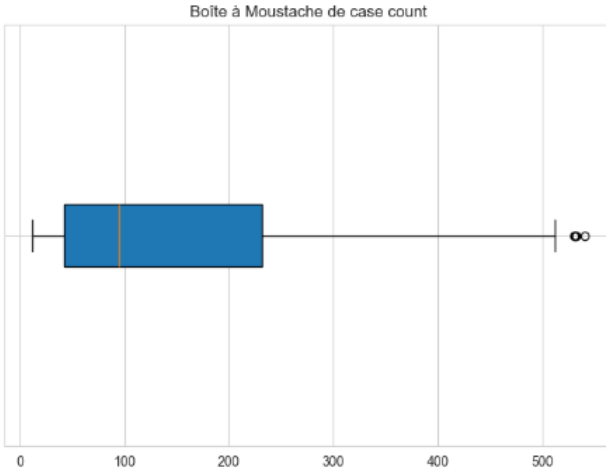

FIGURE 15 – Le boîte à moustache (moyenne) de l'attribut "case count"

L'analyse comparative entre les deux approches, utilisant la médiane et la moyenne pour traiter les valeurs aberrantes, a permis de déterminer celle qui était la plus adaptée aux spécificités des données (mean). Cette méthodologie a été cruciale pour garantir la qualité et la pertinence des données, ainsi que pour assurer une analyse précise et fiable par la suite.

0.6 Visualisation

Après avoir effectué les traitements requis sur nos données, nous sommes à présent aptes à déduire des conclusions significatives concernant les multiples facettes de la pandémie de COVID-19 que nous avons étudiées.

0.6.1 Distribution des Cas Confirmés et Tests Positifs par Zones

Pour visualiser la répartition du nombre total de cas confirmés et de tests positifs par zones, nous avons utilisé un diagramme en arbre (Tree Map) et un graphique à barres (Bar Chart). Ces représentations graphiques offrent une vue d'ensemble claire de la situation, permettant une analyse rapide des zones les plus touchées.

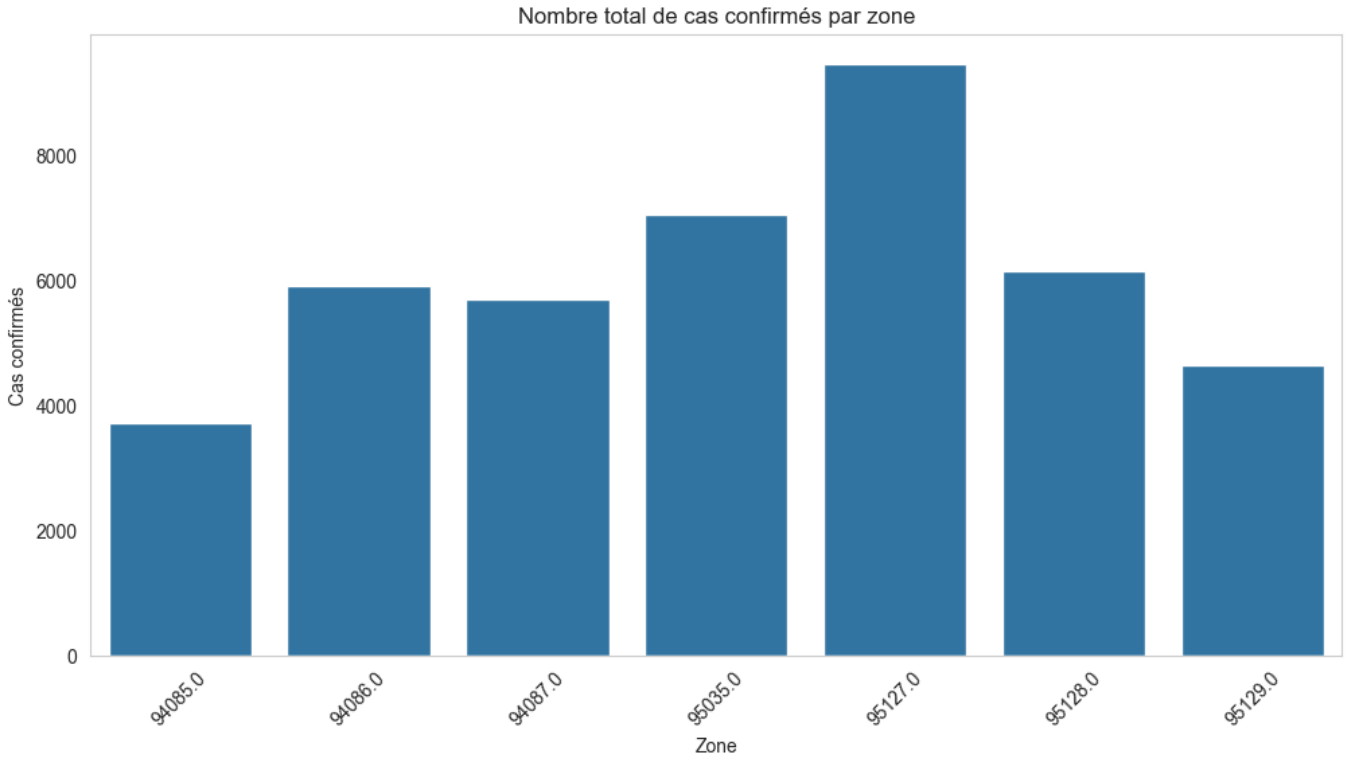

FIGURE 16 – Nombre total de cas confirmés par zone

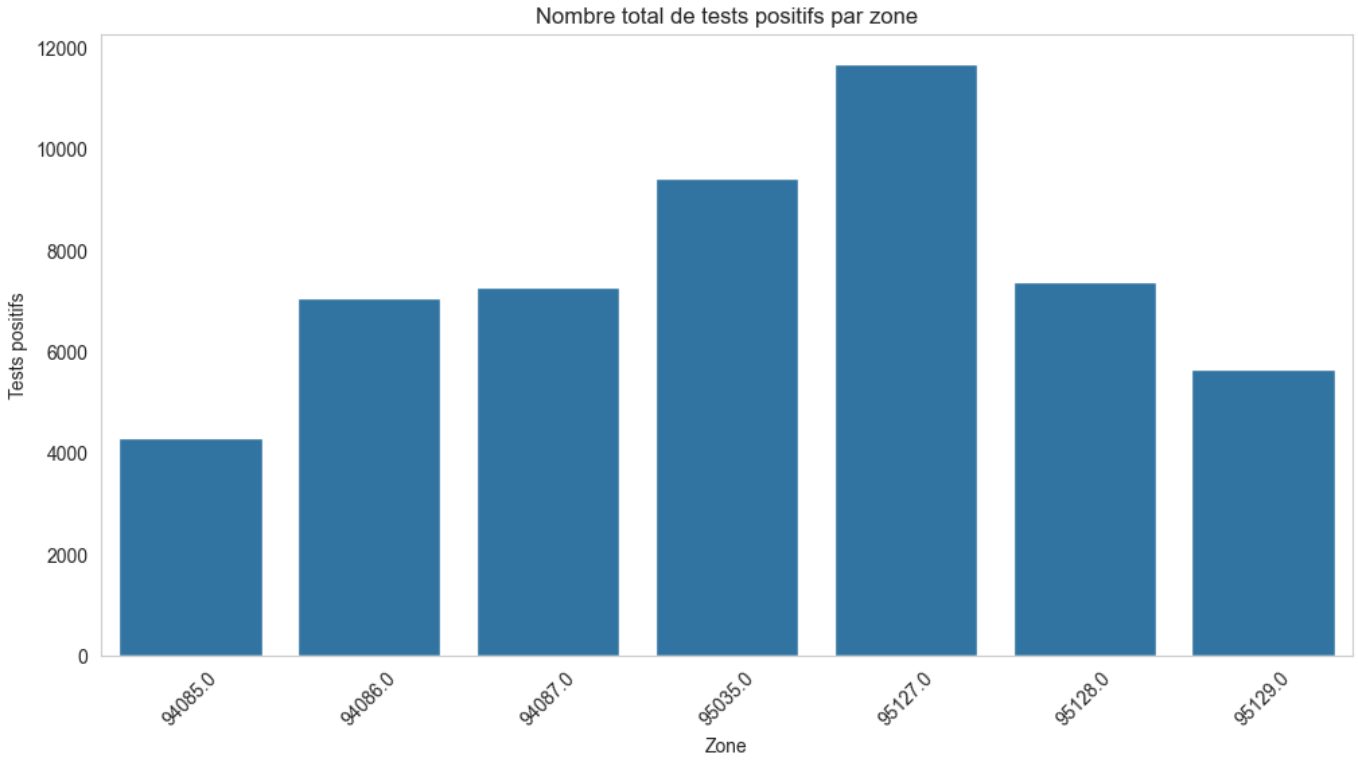

FIGURE 17 – Nombre total de tests positifs par zone

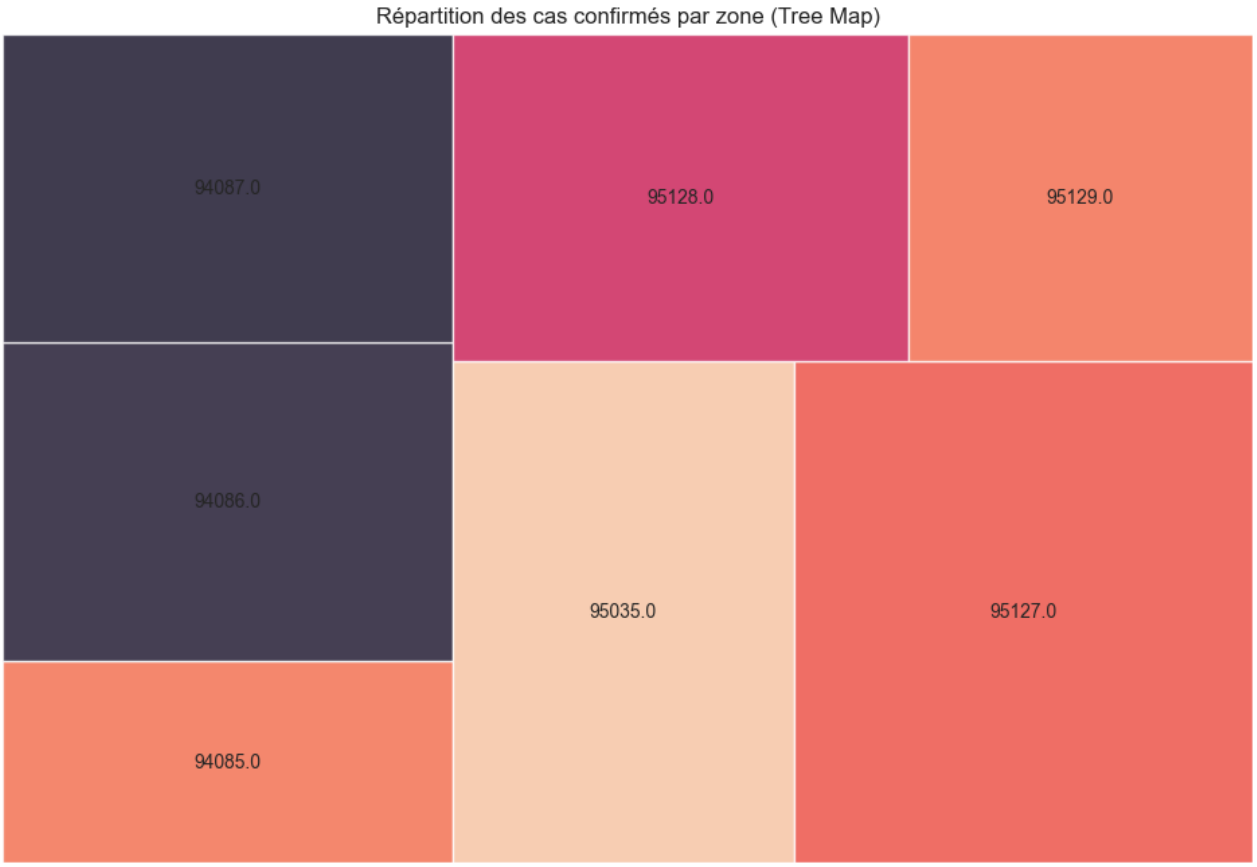

FIGURE 18 – Répartition des cas confirmés par zone (Tree Map)

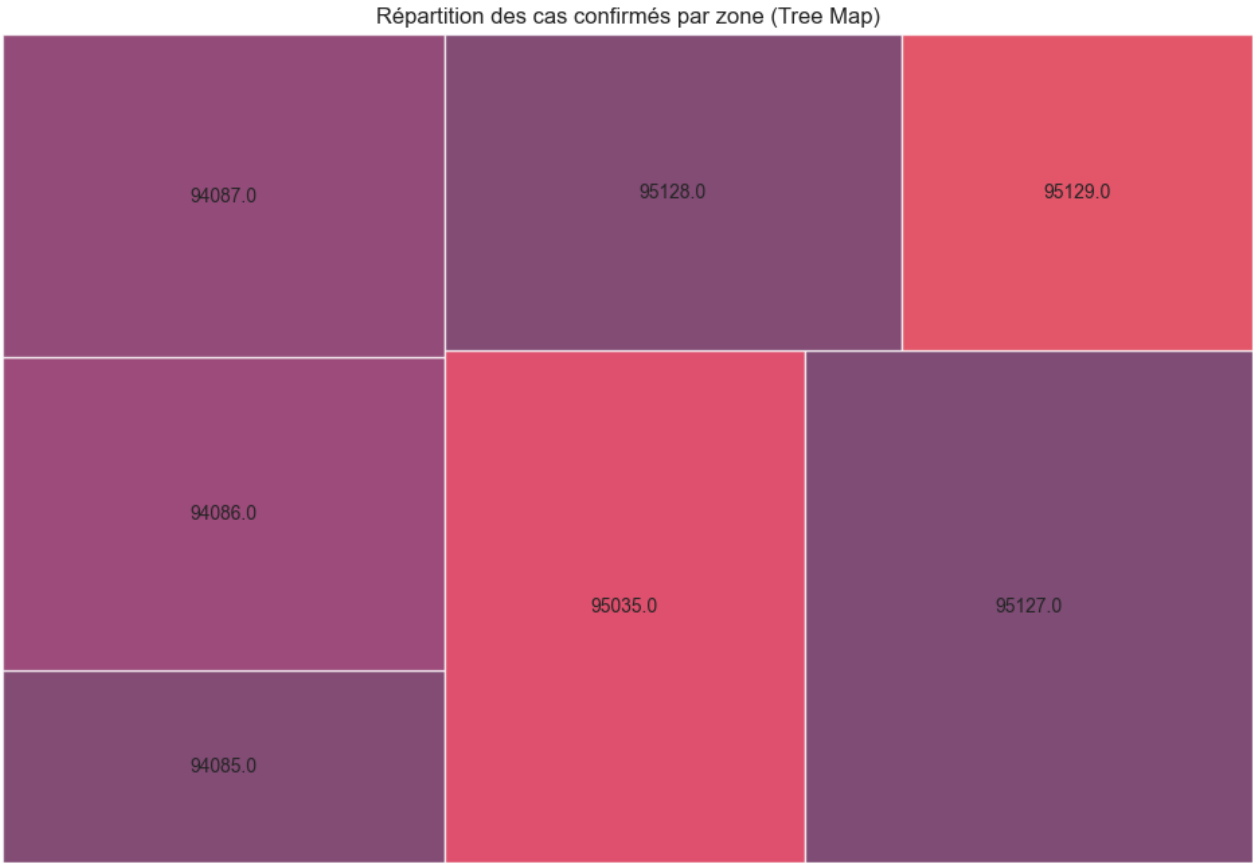

FIGURE 19 – Répartition de tests positifs par zone (Tree Map)

0.6.2 Évolution Temporelle des Tests COVID-19, des Tests Positifs et du Nombre de Cas pour une Zone Spécifique :

Nous avons analysé l'évolution dans le temps des tests COVID-19, des cas positifs, et du nombre de cas, en adoptant une approche à différentes échelles : hebdomadaire, mensuelle et annuelle. Ces tendances ont été représentées graphiquement par des courbes linéaires, offrant ainsi une vision détaillée des fluctuations temporelles.

Hebdomadaire :

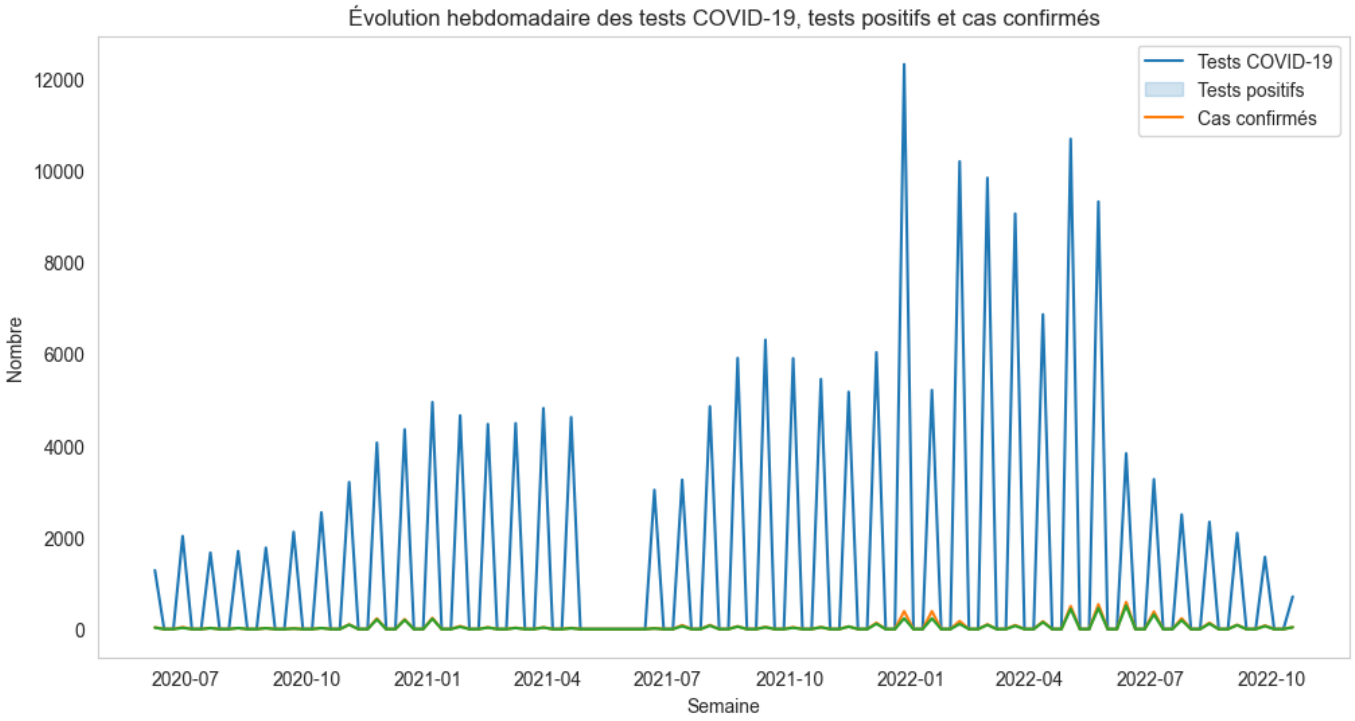

FIGURE 20 – Évolution hebdomadaire des tests COVID-19, tests positifs et cas confirmés

Mensuelle :

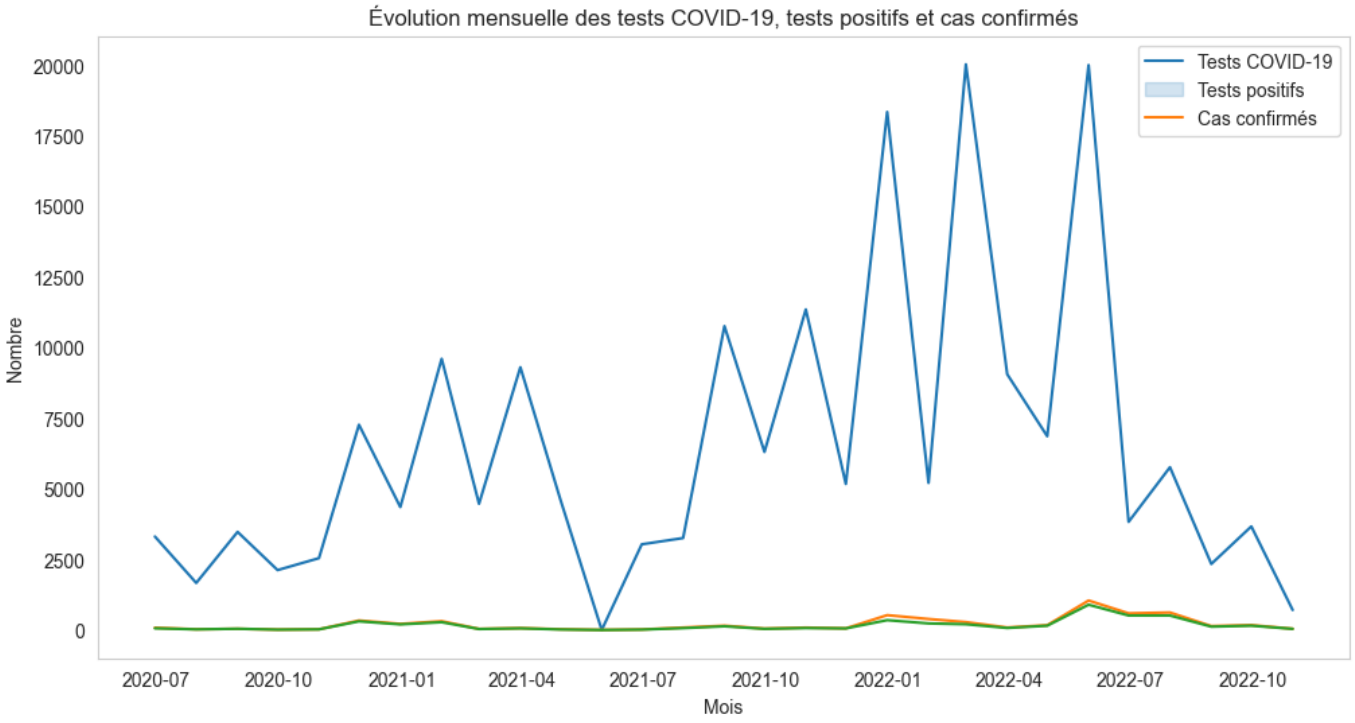

FIGURE 21 – Évolution mensuelle des tests COVID-19, tests positifs et cas confirmés

Annuelle :

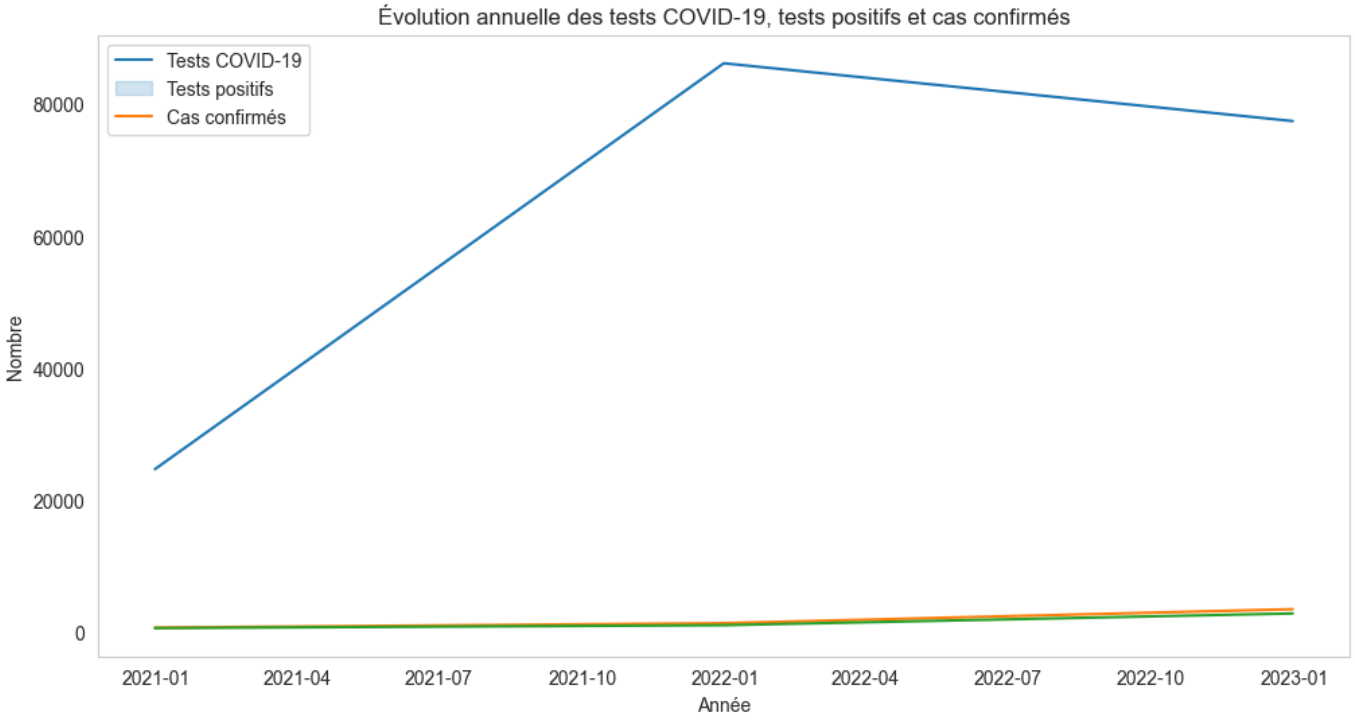

FIGURE 22 – Évolution annuelle des tests COVID-19, tests positifs et cas confirmés

0.6.3 Distribution des Cas COVID-19 Positifs par Zone et par Année :

Une étude détaillée de la répartition des cas positifs de COVID-19 a été menée à l'aide d'un graphique à barres empilées. Ce visuel met en lumière la répartition des cas positifs par zone au fil des années, simplifiant ainsi l'identification de tendances spécifiques.

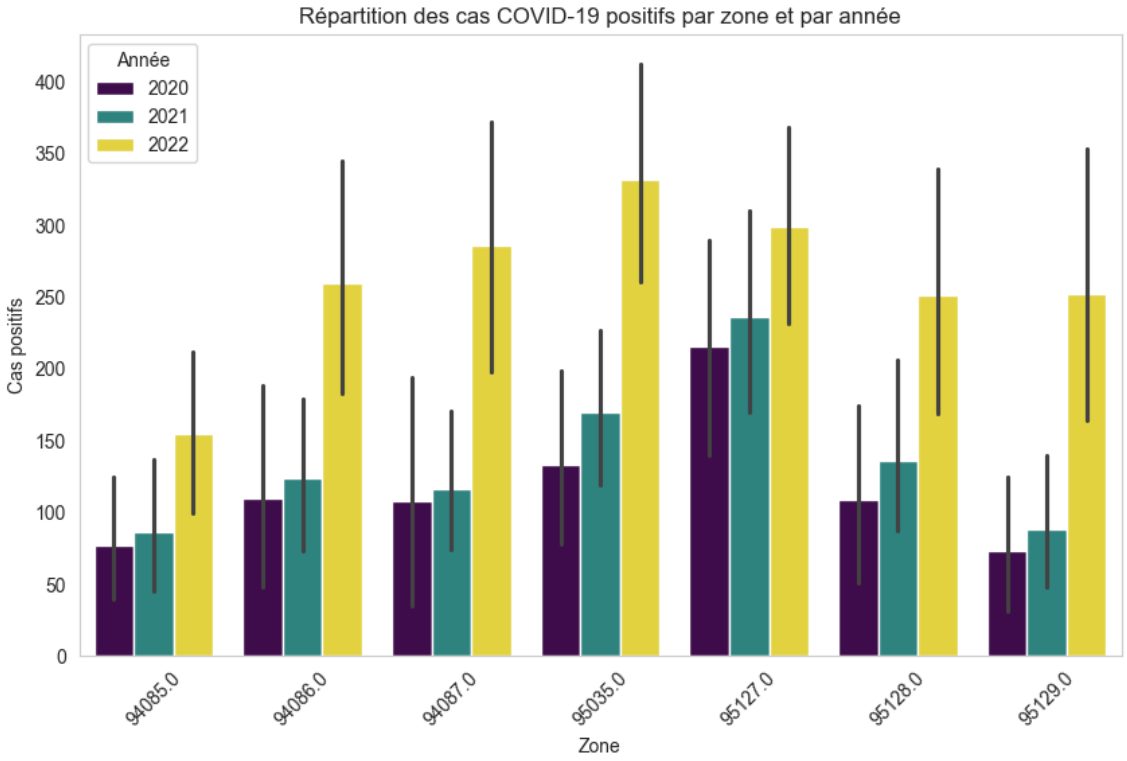

FIGURE 23 – Répartition des cas COVID-19 positifs par zone et par année

0.6.4 Relation Graphique entre la Population et le Nombre de Tests Effectués :

Le concept de ce graphe à bulles est de visualiser la relation entre la population et le nombre de tests effectués, tout en montrant les zones géographiques ou les différentes données (représentées par les bulles) en fonction des tailles et positions sur le graphique.

FIGURE 24 – Rapport entre la population et le nombre de tests effectués

0.6.5 Zones les Plus Fortement Impactées par le Coronavirus :

En répertoriant les cinq zones les plus touchées par le coronavirus - dans l'ordre **[95127, 95035, 95128, 94086, 94087]** - nous avons pu identifier les zones épidémiologiques majeurs. Ces données revêtent une importance cruciale pour orienter les interventions et allouer les ressources là où elles sont les plus urgentes, favorisant ainsi une gestion plus efficiente de la crise sanitaire.

Les 5 zones les plus fortement impactées par le coronavirus :

	zcta	case count
4	95127.0	9456.630363
3	95035.0	7063.584158
5	95128.0	6143.861386
1	94086.0	5921.907591
2	94087.0	5694.861386

FIGURE 25 – Zones les Plus Fortement Impactées par le Coronavirus

0.6.6 Le rapport entre les cas confirmés, les tests effectués et les tests positifs au fil du temps pour chaque zone :

Cette représentation graphique est un graphique en ligne (line plot) affichant l'évolution des cas confirmés, des tests effectués et des tests positifs pour toutes les zones. Cette visualisation, basée sur les données de la période sélectionnée (de '01/01/2021' a '31/12/2022') et la zone "95128".

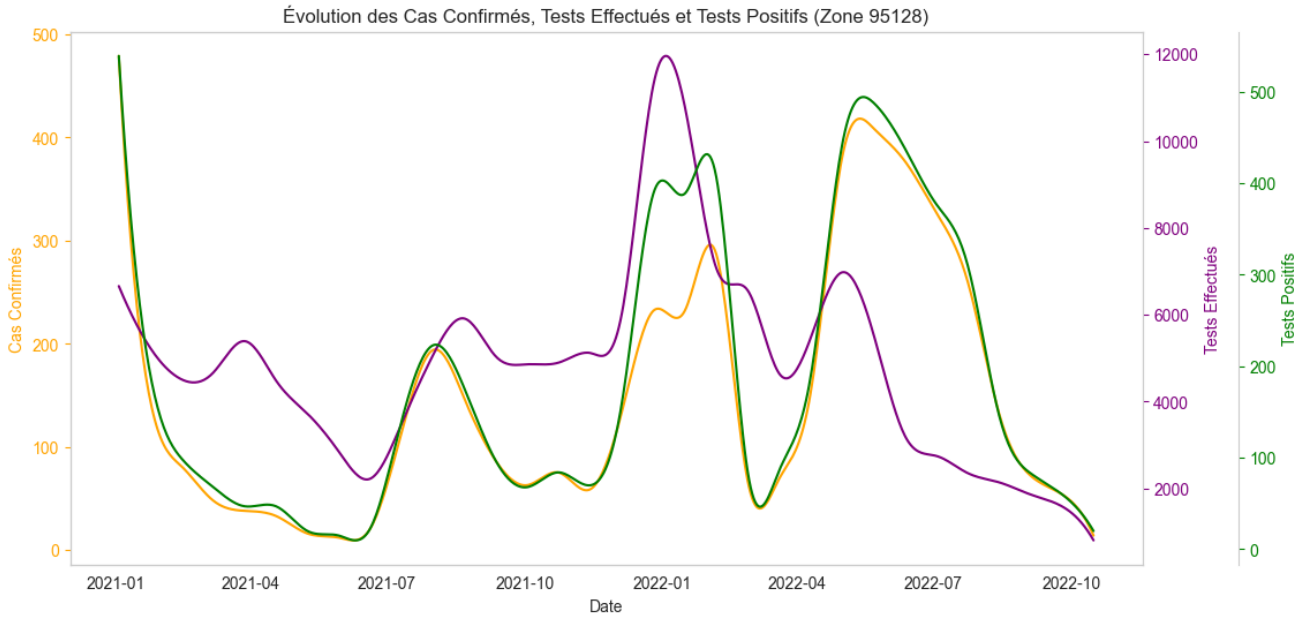

FIGURE 26 – Le rapport entre les cas confirmés, les tests effectués et les tests positifs au fil du temps pour la zone "95128"

Troisième partie

Extraction de motifs fréquents, règles d'associations et corrélations

0.7 Introduction

Cette partie vise à analyser un ensemble de données agricoles afin de déduire des informations pertinentes pour la gestion des cultures et des pratiques agricoles.

0.8 Analyse globale du dataset et prétraitement

0.8.1 Description globale du dataset

Le jeu de données "dataset3.csv" comprend des informations sur la température, l'humidité, les précipitations, le type de sol, la culture et le type d'engrais pour différentes observations.

```
   Temperature Humidity Rainfall        Soil     Crop Fertilizer
0        24.87    82.84   295.61      Clayey     rice        DAP
1        28.69    96.65   178.96    laterite  Coconut   Good_NPK
2        20.27    81.64   270.44  silty_clay     rice        MOP
3        25.07    95.02    192.9       sandy  Coconut       Urea
4        25.04     95.9    174.8     coastal  Coconut       Urea
5        20.82    84.13   230.22   clay_loam     rice       Urea
6        25.95    93.41   172.05    alluvial  Coconut       Urea
7        26.49    80.16   242.86      Clayey     rice        DAP
8        25.01    95.59   165.81     coastal  Coconut       Urea
9        21.87    80.19   224.56  silty_clay     rice       Urea
Nombre de lignes du dataset : 295
Nombre de colonnes du dataset :  6
Noms des colonnes :  ['Temperature', 'Humidity', 'Rainfall', 'Soil', 'Crop', 'Fertilizer']
```

FIGURE 27 – Description globale du dataset

0.8.2 Description de chaque attribut

- **Temperature :** Mesurée en degrés.
- **Humidity :** Mesurée en pourcentage.
- **Rainfall :** Mesurée en millimètres.
- **Soil :** Caractéristiques du sol.
- **Crop :** Type de culture associé aux données.
- **Fertilizer :** Type d'engrais utilisé.

	Nom	Type
0	Temperature	float64
1	Humidity	float64
2	Rainfall	float64
3	Soil	object
4	Crop	object
5	Fertilizer	object

FIGURE 28 – Description globale du dataset

Nous remarquons que les variables Temperature, Humidity and Rainfall sont des variables continues, alors que les variables Soil, Crop et Fertilizer sont des variables catégoriques de type object.

0.8.3 Traitement des valeurs manquantes

Le dataset ne contient aucune valeur manquante comme nous pouvons l'apercevoir dans la figure suivante :

```
dataset.isna().sum()
✓ 0.0s

Temperature    0
Humidity       0
Rainfall       0
Soil           0
Crop           0
Fertilizer     0
dtype: int64
```

FIGURE 29 – Distribution des valeurs manquantes

0.8.4 Traitement des valeurs aberrantes

Dans notre dataset, nous ne retrouvons aucune valeur aberrante dans les divers attributs. Nous pouvons le remarquer ci-dessous :

```
check_outliers (dataset, ['Temperature', 'Rainfall', 'Humidity'] )
✓ 0.0s

Aucun outlier dans l'attribut Temperature
Aucun outlier dans l'attribut Rainfall
Aucun outlier dans l'attribut Humidity
```

FIGURE 30 – Distribution des valeurs aberrantes

0.8.5 Réduction des données via la discrétisation

La discrétisation est une technique utilisée pour transformer des données continues en données discrètes, facilitant ainsi leur analyse. Dans le contexte de ce projet, les attributs numériques tels que la température, l'humidité et les précipitations (Rainfall) ont été discrétisés à l'aide de deux méthodes principales :

1. **Discrétisation par intervalles égaux (Equal Width) :** Cette méthode divise la plage de valeurs d'un attribut en intervalles de largeur égale. Le code fourni calcule les intervalles pour chaque attribut numérique en spécifiant le nombre de bins (intervalles) souhaité. Par exemple, pour la température, l'humidité et les précipitations (Rainfall), ces valeurs sont divisées en bins selon des intervalles égaux.

2. **Discrétisation par fréquence égale (Equal Frequency) :** Cette méthode regroupe les valeurs en fonction de leur fréquence d'apparition. Le code utilise la méthode `equal_freq` pour attribuer des labels basés sur la fréquence des valeurs pour chaque attribut numérique, créant ainsi des intervalles de fréquence égale.

Ces approches permettent de transformer les données continues en catégories ou en intervalles, simplifiant ainsi leur analyse tout en conservant une représentation significative des informations. En ajustant le nombre de bins ou en utilisant des techniques alternatives, il est possible d'explorer différentes manières de discrétiser les données selon les besoins spécifiques de l'analyse.

0.9 Extraction des motifs fréquents en utilisant Apriori

0.9.1 Préparation des données transactionnelles

La préparation des données pour l'algorithme Apriori implique plusieurs étapes :

- **Chargement des données** : Les données sont chargées depuis le fichier CSV "dataset3.csv" en utilisant des bibliothèques Python telles que Pandas pour les manipuler.

- **Conversion des données** : Les données brutes peuvent nécessiter une conversion pour être utilisées avec l'algorithme Apriori. Dans le code fourni, les valeurs numériques sont converties en un format approprié pour l'analyse en remplaçant les virgules par des points et en convertissant les chaînes de caractères en nombres.

- **Sélection des attributs pertinents** : Dans le contexte de l'algorithme Apriori, il est essentiel de sélectionner les attributs pertinents pour l'extraction des motifs fréquents. Dans le code, seuls certains attributs tels que la température, l'humidité, les précipitations, le type de sol, la culture et le type d'engrais sont conservés pour l'analyse.

- **Création de transactions** : L'algorithme Apriori fonctionne avec des données transactionnelles où chaque transaction contient un ensemble d'éléments. Dans le contexte de l'agriculture, une transaction peut être définie comme un ensemble d'attributs associés à une observation spécifique. Le code doit organiser les données de manière à créer ces transactions pour chaque entrée du dataset, probablement sous forme de listes ou d'objets transactionnels.

En résumé, la préparation des données transactionnelles pour Apriori consiste à charger, nettoyer, sélectionner et organiser les données pour qu'elles soient compatibles avec l'algorithme d'extraction de motifs fréquents. Ces étapes garantissent que les données sont prêtes à être utilisées dans l'analyse pour découvrir les associations et les motifs significatifs entre les différents attributs agricoles.

0.9.2 Création des tables Ck et Lk et génération des motifs fréquents

L'algorithme Apriori est utilisé pour découvrir des motifs fréquents dans les données transactionnelles. Cette méthode itérative génère des ensembles candidats (Ck) et identifie les ensembles fréquents (Lk) de manière incrémentielle, en respectant un seuil de support minimum.

```
def get_k_itemsets(data, k):
    itemsets = []
    for row in data:
        itemsets.extend(combinations(row, k))
    return itemsets
```

FIGURE 31 – "get$_{ki}temsets$" *function*

Cette fonction `get_k_itemsets` utilise la méthode `combinations` pour générer tous les ensembles possibles de taille k à partir des données transactionnelles. Ces ensembles candidats seront évalués pour leur fréquence dans les transactions afin d'identifier les ensembles fréquents.

L'algorithme procède alors en évaluant la fréquence de chaque ensemble candidat via une méthode similaire à `calculate_support`. Les ensembles qui dépassent le seuil de support minimum défini sont considérés comme des ensembles fréquents.

L'itération se poursuit, générant des ensembles candidats de taille $k+1$ en combinant les ensembles fréquents existants de taille $k-1$. Cette approche incrémentielle permet de découvrir progressivement des motifs fréquents plus complexes dans les données transactionnelles.

0.9.3 Extraction des règles d'associations et corrélations

Une fois les ensembles fréquents (Lk) identifiés, l'étape suivante consiste à déduire les règles d'association à partir de ces ensembles. Ces règles d'association mettent en évidence les relations significatives entre différents attributs des données.

```python
def generate_rules(frequent_itemsets, min_confidence):
    rules = []

    for itemset, support in frequent_itemsets:
        if len(itemset) > 1:
            for i in range(1, len(itemset)):
                antecedent = itemset[:i]
                consequent = itemset[i:]

                support_antecedent = calculate_support(df.values.tolist(), antecedent)
                support_consequent = calculate_support(df.values.tolist(), consequent)
                confidence = support / support_antecedent
                rule_support = support
                if confidence >= min_confidence:
                    rules.append((antecedent, support_antecedent, consequent, support_consequent, rule_support, confidence))

    return rules
```

FIGURE 32 – "generate$_r ules$" $function$

Et voici le pseudo code :

Algorithm 1 Algorithme d'Extraction des Règles d'Association

Entrée : Ensemble de motifs fréquents F, seuil de confiance minimal $minConf$
Sortie : Ensemble de règles d'association R
    forall *motif fréquent* f *dans* F **do**
    end
    sous-ensemble non vide A de f
$B \leftarrow f \setminus A$
Calculer la confiance$(A \Rightarrow B)$ **if** $confiance \geq minConf$ **then**
    end
    Ajouter la règle $A \Rightarrow B$ à R
retourner R

Cette fonction parcourt les ensembles fréquents et génère des règles d'association pour chaque ensemble, en évaluant la confiance de chaque règle basée sur le support des ensembles impliqués. Les règles dont la confiance dépasse le seuil de confiance minimum spécifié sont considérées comme significatives et sont conservées pour une analyse plus approfondie.

De plus, une fois que les règles d'association sont obtenues, elles peuvent être évaluées en termes de corrélation pour mesurer la force et la fiabilité de ces règles. Pour ce faire, des métriques telles que le Lift, la Confiance et le Cosine sont calculées. La fonction `correlation(rules)` calcule ces métriques pour chaque règle d'association déduite.

0.9.4 Mesures de corrélation des règles d'association

Une fois les règles d'association extraites, les mesures de corrélation sont utilisées pour évaluer la force et la fiabilité de ces règles. Trois métriques couramment utilisées pour évaluer la corrélation sont : le Lift, la Confiance et le Cosine.

La fonction `correlation(rules)` est employée pour calculer ces métriques de corrélation pour chaque règle d'association déduite. Voici un extrait du code montrant la structure de cette fonction :

```python
def correlation(rules):
    rules_with_correlation = []
    for antecedent, support_antecedent, consequent, support_consequent, rule_support, confidence in rules:
        lift = rule_support/(support_antecedent*support_consequent)
        confidence = rule_support/max(support_antecedent, support_consequent)
        cosine = rule_support/math.sqrt(support_antecedent*support_consequent)

        rules_with_correlation.append((antecedent, consequent, lift, confidence, cosine))
    return rules_with_correlation
```

FIGURE 33 – "correlation" function

Et voici le pseudo code :

Algorithm 2 Algorithme d'Extraction des Règles d'Association

Entrée : Ensemble de motifs fréquents F, seuil de confiance minimal $minConf$
Sortie : Ensemble de règles d'association R
    forall *motif fréquent* f *dans* F **do**
    end
    sous-ensemble non vide A de f
$B \leftarrow f \setminus A$
Calculer la confiance$(A \Rightarrow B)$ **if** *confiance* $\geq minConf$ **then**

    end
    Ajouter la règle $A \Rightarrow B$ à R

retourner R

0.9.5 Sélection des Règles d'Association Fortes

La sélection des règles d'association fortes est une étape importante pour filtrer les résultats et ne conserver que les règles les plus significatives. Nous discutons des critères de sélection utilisés et présentons les règles d'association finales.

Algorithm 3 Sélection des Règles d'Association Fortes

1: **Entrée :** Ensemble de règles d'association R, seuil de support minimal $minSup$
2: **Sortie :** Ensemble de règles d'association fortes R_{fortes}
forall *règle d'association* r *dans* R **do**
3:
end
Calculer le support pour l'antécédent et le conséquent de la règle **if** *support de l'antécédent et du conséquent* $\geq minSup$ **then**
4:
end
Ajouter la règle r à R_{fortes}
5:
6:
7: **retourner** R_{fortes}

Cette fonction calcule les mesures de corrélation pour chaque règle d'association en utilisant les valeurs de support et de confiance précédemment calculées. Le Lift mesure la corrélation entre l'antécédent et le conséquent, tandis que la Confiance compare la fréquence d'occurrence de l'ensemble à celle de ses parties individuelles. Le Cosine calcule la corrélation basée sur les vecteurs représentant les ensembles.

En évaluant ces mesures de corrélation, les règles d'association peuvent être classées et évaluées en fonction de leur fiabilité et de leur importance dans les données. Ces mesures aident à sélectionner les règles les plus significatives pour des applications spécifiques, comme l'optimisation des stratégies de recommandation ou la prise de décision dans divers domaines.

0.9.6 Expérimentation avec Min-Supp et Min-Conf

Les seuils de support minimum (Min-Supp) et de confiance minimum (Min-Conf) jouent un rôle crucial dans le processus d'extraction des règles d'association. Ils permettent de contrôler la qualité et la quantité des règles extraites. L'ajustement de ces seuils peut influencer significativement le nombre et la nature des règles découvertes.[2]

Dans le cadre de l'expérimentation avec ces seuils, différents réglages sont testés pour évaluer l'impact sur les règles d'association extraites. Cette expérimentation vise à déterminer les valeurs optimales pour ces seuils afin d'obtenir des règles pertinentes et significatives.[1]

Dans notre code on a utilisé des valeurs spécifiées pour `min_confidence` et `min_support`, qui sont respectivement les seuils de confiance minimum et de support minimum. Voici un extrait du code démontrant leur utilisation :

```
min_confidence = 0.8
min_support = 0.2
frequent_itemsets = find_frequent_itemsets(df.values.tolist(), min_support=min_support)
association_rules = generate_rules(frequent_itemsets, min_confidence)
association_rules = sorted(association_rules, key=lambda x: x[3], reverse=True)
```

FIGURE 34 – "correlation" function

En ajustant ces valeurs de seuils, l'expérimentation permet d'observer comment la variation de ces seuils affecte le nombre et la qualité des règles d'association extraites. Par exemple, augmenter le seuil de support minimum peut réduire le nombre de règles extraites, mais celles-ci seront plus fiables. À l'inverse, une diminution du seuil de confiance minimum peut générer plus de règles, mais potentiellement moins pertinentes.

Cette expérimentation permet de comprendre la sensibilité des résultats à ces seuils et aide à choisir des valeurs appropriées en fonction des objectifs spécifiques de l'analyse. Il est recommandé de réaliser plusieurs essais avec différentes combinaisons de seuils pour déterminer les paramètres optimaux en fonction du contexte et des besoins du problème.

0.10 Conclusion

L'analyse des données agricoles a permis d'extraire des règles d'association significatives entre différents paramètres, offrant ainsi des perspectives potentielles pour améliorer les pratiques agricoles.

Conclusion Générale

La première étape de notre projet de Data Mining nous a plongés dans les processus essentiels de prétraitement et d'exploration des données. En analysant en détail les caractéristiques des ensembles de données statiques et temporels, nous avons compris à quel point il est crucial de saisir la nature, la distribution et les comportements des données avant toute analyse approfondie.

Le nettoyage des données a joué un rôle essentiel pour assurer la fiabilité et la pertinence des futures analyses. Nous avons utilisé diverses techniques, comme le traitement des valeurs manquantes, la gestion des données aberrantes et la normalisation, afin de supprimer les redondances et de préparer les données pour des analyses plus précises.

L'utilisation de l'algorithme Apriori pour extraire des règles d'association a été un élément clé de cette première phase du projet. En dérivant des règles à partir de motifs fréquents, nous avons pu mettre en évidence des relations potentiellement précieuses entre les attributs liés au climat, au sol, à la végétation et à l'utilisation d'engrais.

Cette phase initiale établit les bases de notre projet, nous préparant ainsi à explorer davantage les données pour des tâches de clustering et de classification dans la seconde partie du projet.

Bibliographie

[1] An Overview of Particle Swarm Optimization Variants. MuhammadImranRathiahAuthors. PublishedbyElsevierLtd.

[2] Genes-chrom. https://towardsdatascience.com/.