Université de sciences et de la technologie Houari Boumediene

Faculté d'Informatique
Département d'Intelligence Artificielle
Master 2 Systèmes Informatiques intelligents

De l'Analyse des Données à l'Extraction de Connaissances

Partie 1 : Exploitation des données et Extraction des règles d'associations

Réalisé Par :

FERKOUS Sarah
KHEMISSI Maroua

Professeur : M. Belkadi Widad

Année universitaire : 2023 / 2024

Table des matières

Liste des tableaux

Table des figures

Introduction

Dans l'ère numérique actuelle, la quantité de données générées quotidiennement atteint des proportions colossales. Ces données, provenant de diverses sources telles que les réseaux sociaux, les transactions commerciales, les capteurs IoT, et bien d'autres, représentent une mine d'informations précieuses.
Cependant, leur valeur intrinsèque reste souvent sous-exploitée tant que les techniques adéquates ne sont pas mises en œuvre pour en extraire des connaissances exploitables.
Au coeur de ce projet réside l'ambition d'explorer les mécanismes d'exploitation des données et d'approfondir la compréhension de l'extraction des règles d'associations, nous optons pour deux parties, la première étape l'analyse et prétraitement des données, Une deuxième étape consiste à extraire les motifs fréquents et les règles d'association.
Trois ensembles de données sont mobilisés dans cette étude : deux d'entre eux sont dédiés à la première phase, concernant respectivement les données statiques et les données temporelles, tandis que le troisième ensemble de données est réservé à la deuxième phase de l'analyse.

Chapitre 1

Données statistiques

1.1 Objectifs

L'objectif principal de cette section consiste à conduire une analyse approfondie et à mettre en œuvre un processus de nettoyage du dataset 1. Cette phase revêt une importance cruciale, car elle vise à préparer ces données en vue de leur utilisation ultérieure pour la classification et le clustering.
Le dataset 1 renferme des informations relatives aux caractéristiques du sol, et sa préparation adéquate.
Les étapes clés de cette analyse incluront l'extraction d'informations cruciales, le nettoyage des données pour assurer leur fiabilité, et la création de visualisations significatives

1.2 Manipulation de dataset

Visualisation du contenu du dataset

	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	OM	Fertility
0	138	8.6	560	7.46	0.62	0.70	5.90	0.24	0.31	0.77	8.71	0.11	1.2040	0
1	213	7.5	338	7.62	0.75	1.06	25.40	0.30	0.86	1.54	2.89	2.29	1.8232	0
2	163	9.6	718	7.59	0.51	1.11	14.30	0.30	0.86	1.57	2.70	2.03	1.9092	0
3	157	6.8	475	7.64	0.58	0.94	26.00	0.34	0.54	1.53	2.65	1.82	1.6168	0
4	270	9.9	444	7.63	0.40	0.86	11.80	0.25	0.76	1.69	2.43	2.26	1.4792	1
5	220	8.6	444	7.43	0.65	0.72	11.70	0.37	0.66	0.90	2.19	1.82	1.2384	0

Table 1.1 – Description de dataset-1-

Le dataset présenté dans le tableau 1.1 contient des informations liées à différents paramètres agronomiques et chimiques associés à des échantillons de sol. Les colonnes représentent les variables mesurées, telles que N (azote), P (phosphore), K (potassium), pH, EC (conductivité électrique), OC (carbone organique), S (soufre), Zn (zinc), Fe (fer), Cu (cuivre), Mn (manganèse), B (boron), OM (matière organique), et Fertility (fertilité).

Chaque ligne du dataset corresponds à une observation ou un échantillon distinct, avec des valeurs spécifiques pour chaque variable. Par exemple, la première ligne indique des valeurs spécifiques pour N, P, K, pH, etc.

L'objectif de ce dataset pourrait être d'analyser la relation entre ces variables et la fertilité du sol, étant donné la présence de la colonne "Fertility". Une analyse de visualisation pourrait être réalisée pour mieux comprendre les tendances, les corrélations ou les schémas au sein des donnée c'est ce qu'on va voire par la suite

Une description globale du dataset

	N	K	pH	EC	OC	S	Zn	Fe
Count	885.000	885.000	885.000	885.000	884.000	885.000	885.000	885.000
Mean	246.998	501.339	7.512	0.544	0.618	7.546	0.469	4.127
Std	77.359	129.105	0.465	0.141	0.841	4.418	1.889	3.108
Min	6.000	11.000	0.900	0.100	0.100	0.640	0.070	0.210
25%	201.000	412.000	7.350	0.430	0.380	4.700	0.280	2.050
50%	257.000	475.000	7.500	0.550	0.590	6.640	0.360	3.560
75%	307.000	581.000	7.630	0.640	0.780	8.750	0.470	6.320
Max	383.000	1560.000	11.150	0.950	24.000	31.000	42.000	44.000

Table 1.2 – Statistiques des données

	Cu	Mn	B	OM	Fertility
Count	884.000	885.000	885.000	885.000	885.000
Mean	0.952	8.654	0.593	1.064	0.592
Std	0.466	4.301	0.575	1.446	0.578
Min	0.090	0.110	0.060	0.172	0.000
25%	0.630	6.210	0.270	0.6536	0.000
50%	0.930	8.340	0.410	1.0148	1.000
75%	1.250	11.470	0.610	1.3416	1.000
Max	3.020	31.000	2.820	41.280	2.000

Table 1.3 – Suite

Description de chaque attribut

Il est à noter que la colonne OC a une valeur de Non-Null Count inférieure aux autres, indiquant qu'il y a une observation manquante dans cette colonne. Cela pourrait nécessiter une gestion spécifique, comme l'imputation des valeurs manquantes ou l'analyse de l'impact potentiel sur les résultats.
La colonne P semble contenir des objets, ce qui pourrait nécessiter une vérification pour s'assurer qu'elle est correctement interprétée.

Column	Non-Null-Count	Dtype	Unique-Value-Count
N	885 non-null	int64	61
P	885 non-null	object	93
K	885 non-null	int64	63
pH	885 non-null	float64	107
EC	885 non-null	float64	71
OC	884 non-null	float64	69
S	885 non-null	float64	153
Zn	885 non-null	float64	70
Fe	885 non-null	float64	387
Cu	884 non-null	float64	167
Mn	885 non-null	float64	429
B	885 non-null	float64	127
OM	885 non-null	float64	68
Fertility	885 non-null	int64	3

TABLE 1.4 – Description des colonnes du dataset

Après l'affichage des valeurs uniques de chaque attribut on constate qu'on doit convertir les valeurs erronés dans la colonne P par (nan)

```
df.P.unique()
```

```
array(['8.6', '7.5', '9.6', '6.8', '9.9', '7.2', '7', '14.9', '8.1',
       '5.3', '8.3', '7.7', '?', '6.1', '9.4', '5.5', '5', '5.7', '7.9',
       '10.7', '5.9', '4.8', '6.4', '6.6', '4.4', '4.6', '9', '59.2',
       '14.5', '9.2', '78.9', '12.3', '11.2', '12.9', '11.4', '14.3',
       '10.5', '10.3', '12.7', '11.8', '14.7', '13.4', '3.9', '11',
       '11.6', '18.4', '20.4', '19.3', '18.2', '19.5', '20.8', '21.5',
       '21.1', '10.1', '8.8', '2.9', '15.6', '12.5', '13.8', '76.8',
       '103.1', '63.6', '61.4', '70.2', '68', '65.8', '74.6', '81.1',
       '98.7', '85.5', '72.4', '13.2', '12.1', '111.8', '125', '15.1',
       '118.4', '14', '13.6', '15.4', '16', '17.3', '94.3', '92.1',
       '89.9', '96.5', '122.8', '114', '107.5', '83.3', '87.7', '105.3',
       '120.6'], dtype=object)
```

FIGURE 1.1 – Les valeurs uniques de l'attribut P

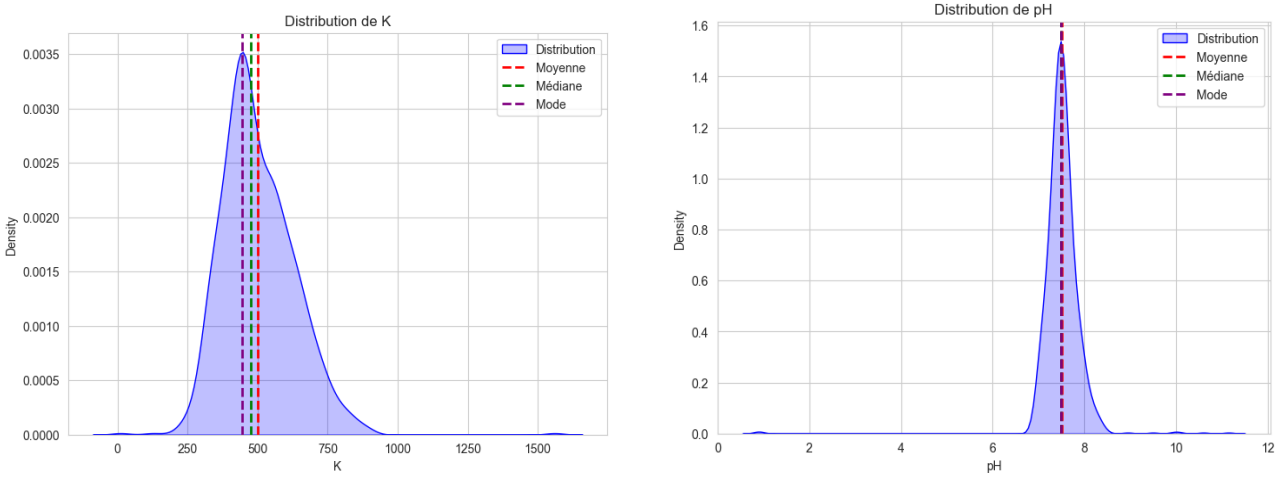
1.3 Analyse des caractéristiques des attributs du dataset

1.3.1 Calcul des mesures de tendance centrale pour déduire les symétries

	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

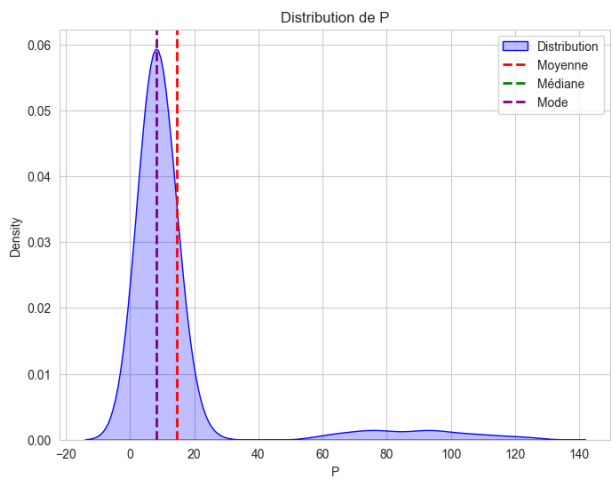
TABLE 1.5 – Description statistique des attributs.

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 aura donc 14 graphes.

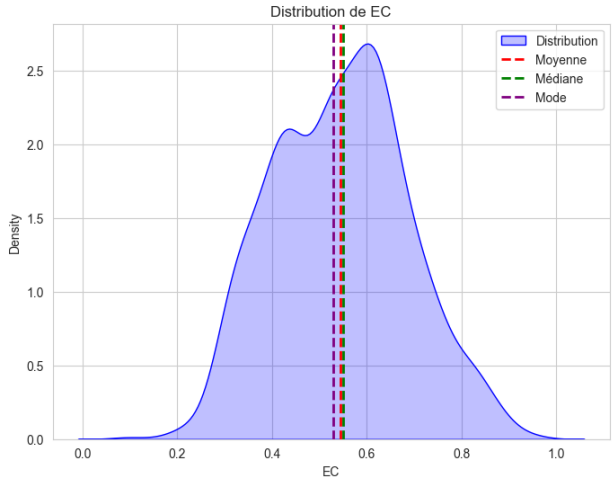


(a) K (b) pH

FIGURE 1.2 – Graphe Densité de K et pH

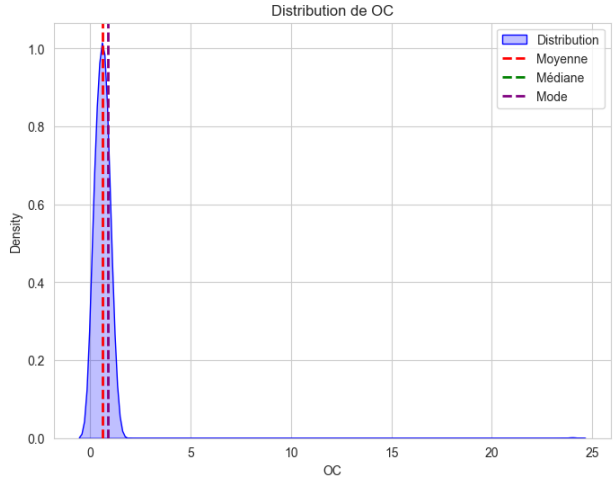


(a) P

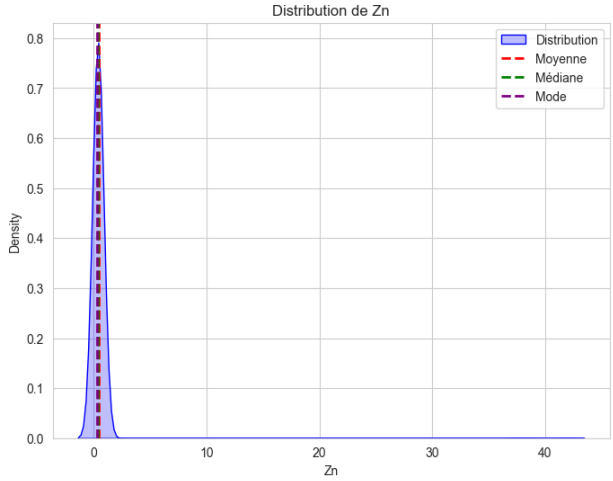


(b) EC

FIGURE 1.3 – Graphe Densité de P et EC

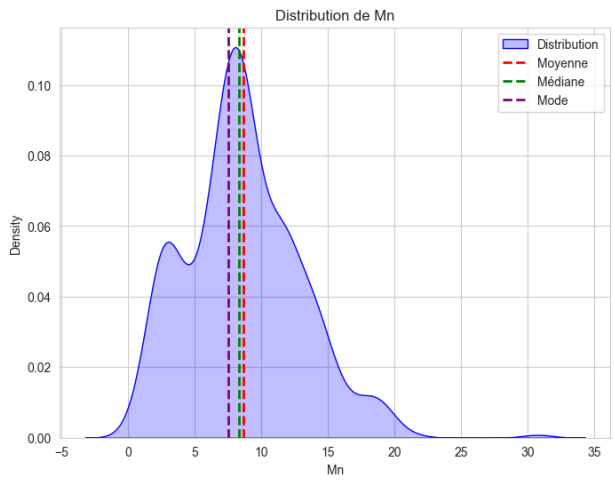


(a) OC

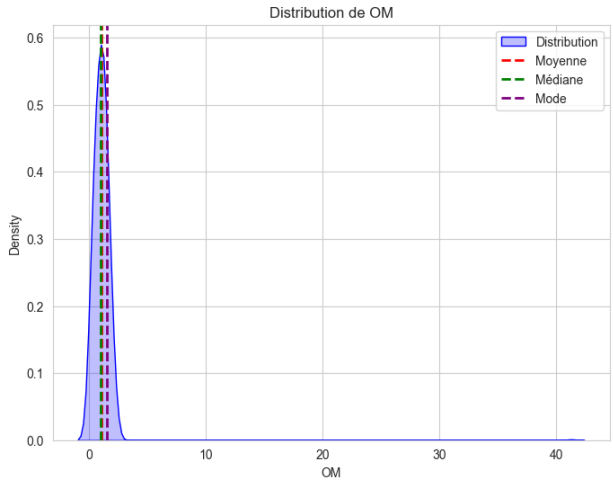


(b) Zn

FIGURE 1.4 – Graphe Densité de OC et Zn



(a) Mn



(b) OM

FIGURE 1.5 – Graphe Densité de Mn et OM

Analyse

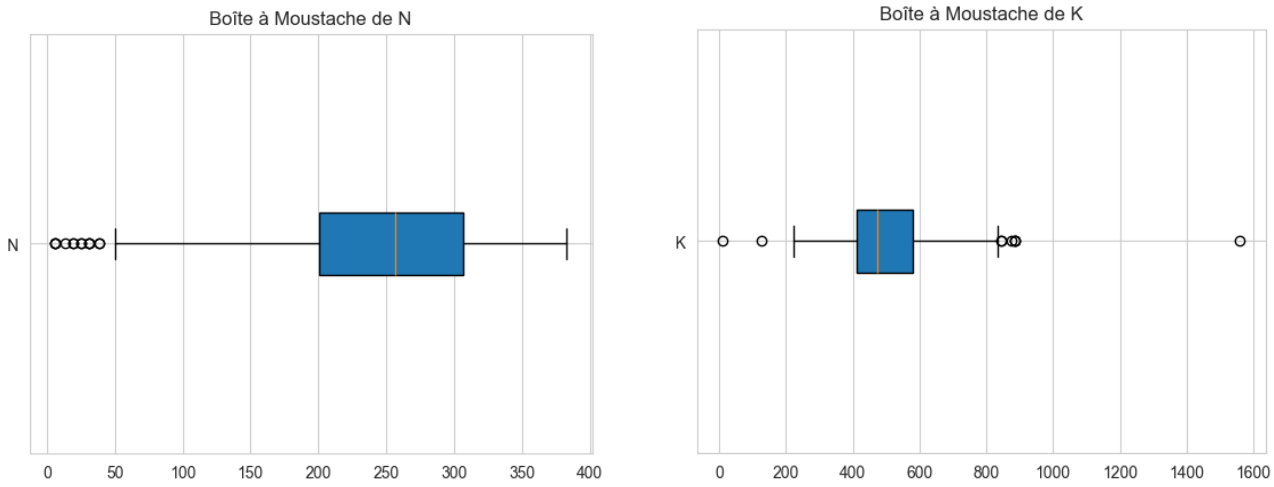
D'après les graphes de densité visualisés ci-dessus, on constate que pour K, EC, Mn et OC, les valeurs de moyenne, médiane et mode sont presque égales car les distributions sont presque identiques.

Pour pH, Zn et OM, la moyenne = médiane = mode car leurs distributions sont identiques. Pour le reste des colonnes, elles ne représentent aucune symétrie.

Conclusion

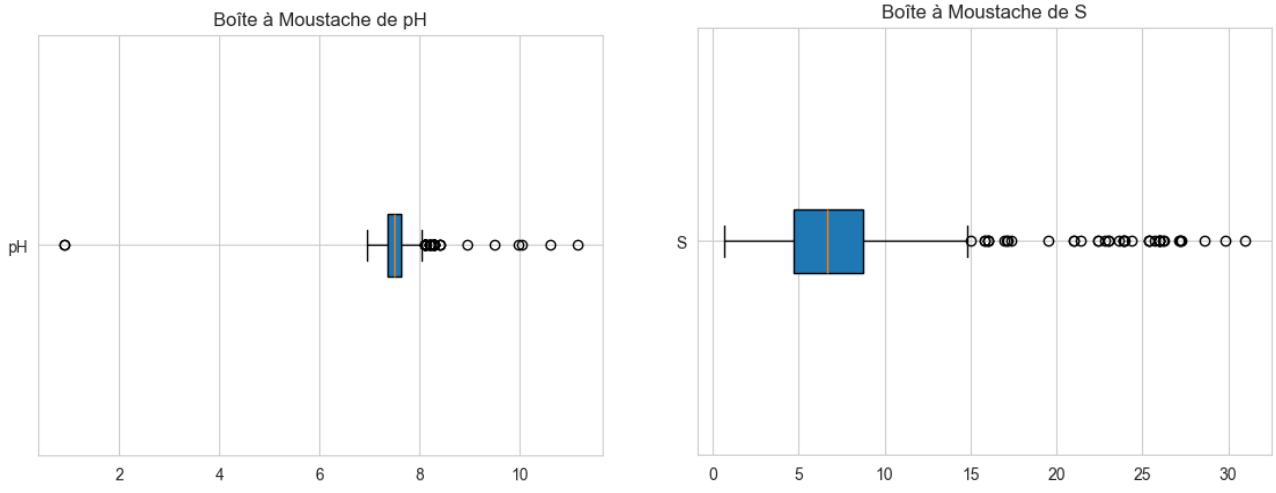
Cette constatation suggère une certaine homogénéité dans les valeurs de K, EC, Mn et OC, indiquant une stabilité ou une uniformité dans ces paramètres du sol. En revanche, pour pH, Zn et OM, la symétrie des distributions suggère une répartition équilibrée des valeurs autour de la moyenne. Pour les autres attributs, la symétrie n'est pas présente.

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



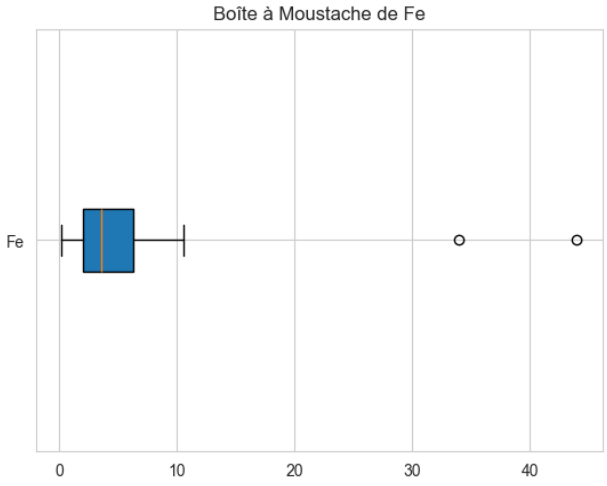
(a) N (b) K

FIGURE 1.6 – Boîte à moustaches de N et K



(a) pH (b) S

FIGURE 1.7 – Boîte à moustaches de pH et S



(a) Fe

FIGURE 1.8 – Boîte à moustaches de Fe

Analyse

Après la construction des boîtes à moustaches pour chaque attribut, on constate qu'il y a des valeurs aberrantes dans tous les attributs. Cependant, certaines colonnes présentent un nombre plus élevé de valeurs aberrantes par rapport aux autres, notamment N, K, pH, S, Mn et B. En revanche, les colonnes EC, Zn, Fe et OM montrent peu de valeurs aberrantes.

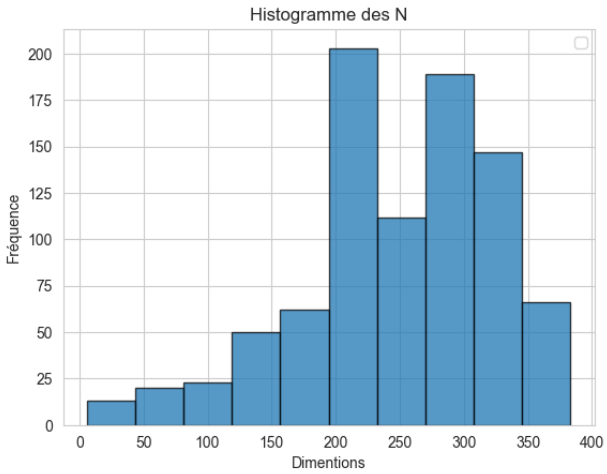
Il est clair qu'on peut déterminer les symétries à partir de ces graphes, si le trait orange (médiane) est au milieu de la boîte.

Conclusion

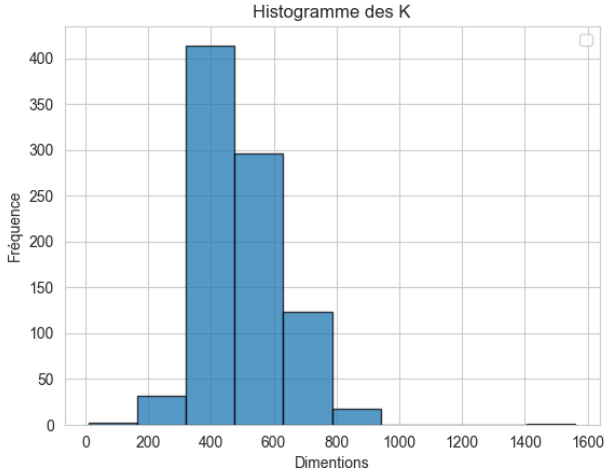
la gestion et l'interprétation des données devraient tenir compte de ces variations dans la détection des valeurs aberrantes, en mettant l'accent sur les attributs présentant une plus grande instabilité. Ceci pourrait orienter les analyses futures et les actions de correction pour assurer une interprétation précise des caractéristiques du sol.

Il faut penser à une solution pour gérer ces valeurs aberrantes, ce qu'on va voire dans un futur proch.

1.3.3 Construction des histogrammes des données

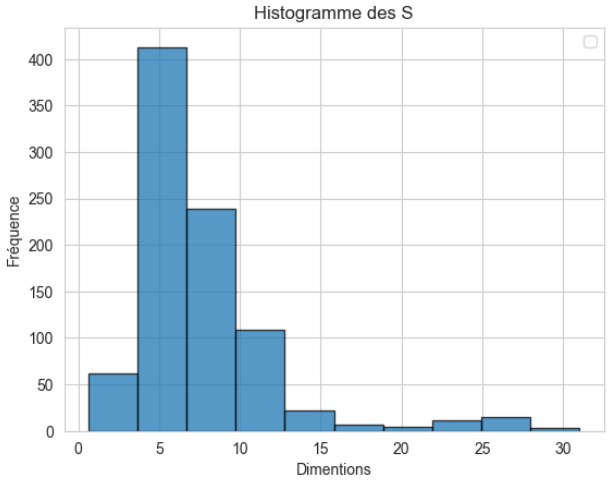


(a) N

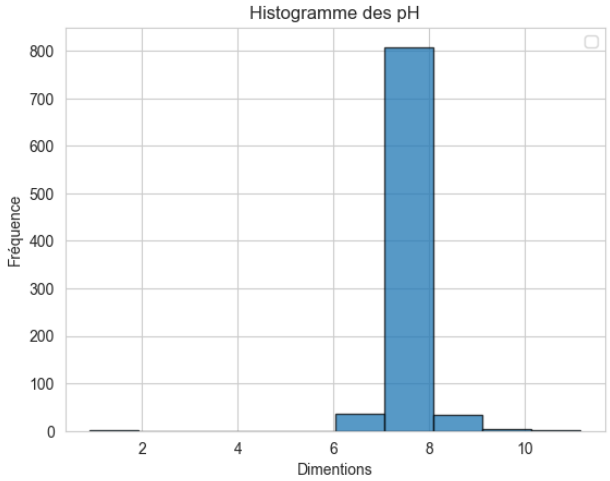


(b) K

FIGURE 1.9 – Histogramme de N et K



(a) S



(b) pH

FIGURE 1.10 – Histogramme de S et pH

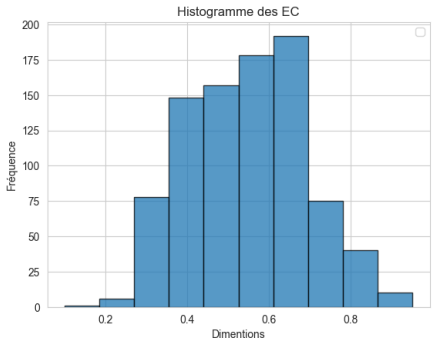


FIGURE 1.11 – Histogramme de EC

Discussion

L a visualisation des histogrammes de données présente plusieurs utilités :

1. **Distribution des données :**
   Avec ces histogrammes, on peut facilement déduire les tendances, et détecter la symétrie des données en utilisant une simple projection et une comparaison miroir. Par exemple elle confirme la symétries pour pH, OM et Zn car nous avons réussit a éfectuer le miroir entre les données.
2. **Centrage et dispersion :**
   Il fournissent une indication visuelle du centrage (moyenne, médiane, mode) et de la dispersion des données, permettant une compréhension rapide de la variabilité des valeurs.
   Par exemple, la mode de EC se situe entre 0.6 et 0.7, La médiane de P est de 250.
3. **Détection des anomalies :**
   On peut savoir quelles sont les valeurs rares, les plus fréquentes, et les anomalies.
   Par exemple, pour les variables Fe et Cu, on constate un nombre très limité de données à 30 et 40 pour Fe et à 3 pour Cu.

1.3.4 Construction des diagrammes de dispersion des données

Pour la construction des diagrammes de dispersion, nous avons calculé la matrice de corrélation suivante :

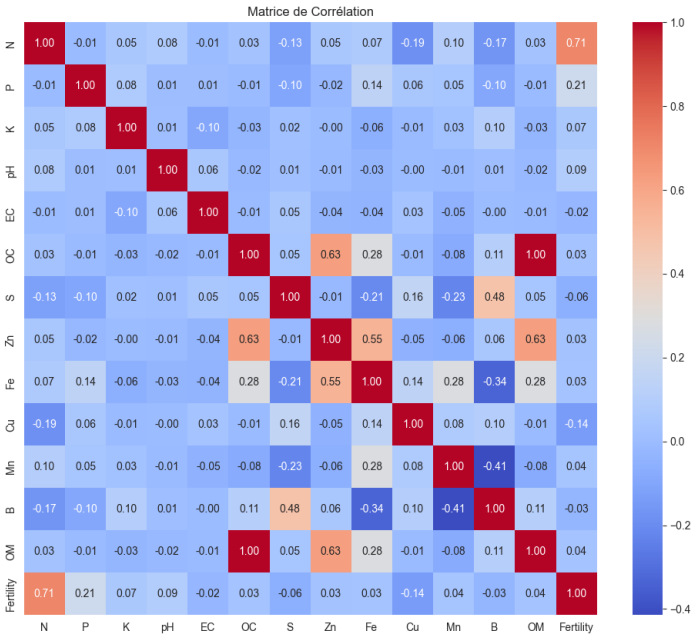


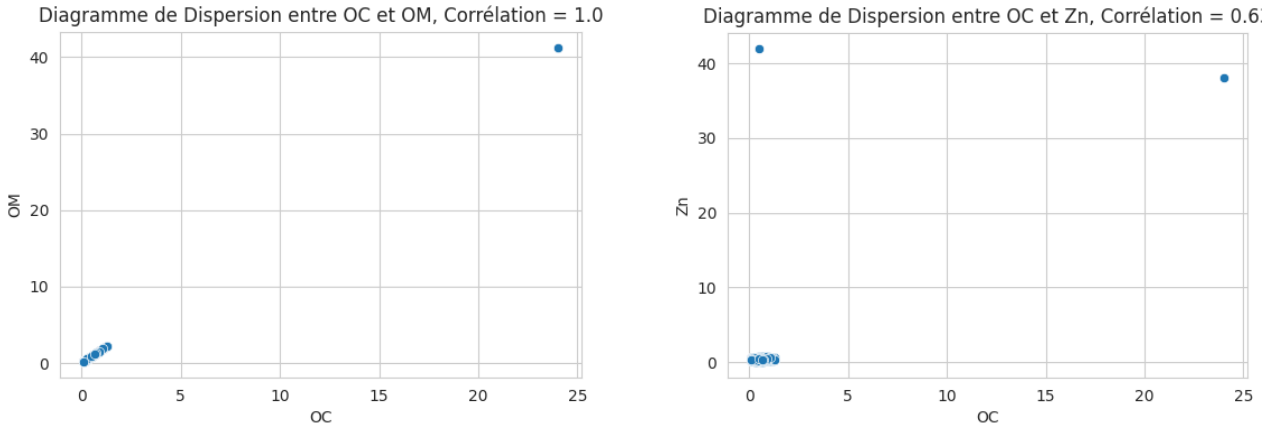
FIGURE 1.12 – La matrice de corrélation de dataset-1-

La matrice de corrélation est une table qui montre les coefficients de corrélation entre de multiples variables.
Le coefficient de corrélation mesure la force et la direction de la relation linéaire entre deux variables.

- r=1 indique une corrélation positive parfaite, ce qui signifie que les variables évoluent ensemble dans la même direction.

- r=-1 indique une corrélation négative parfaite, ce qui signifie que les variables évoluent ensemble dans des directions opposées.
- r=0 indique l'absence de corrélation linéaire entre les variables.

1. corrélation positive



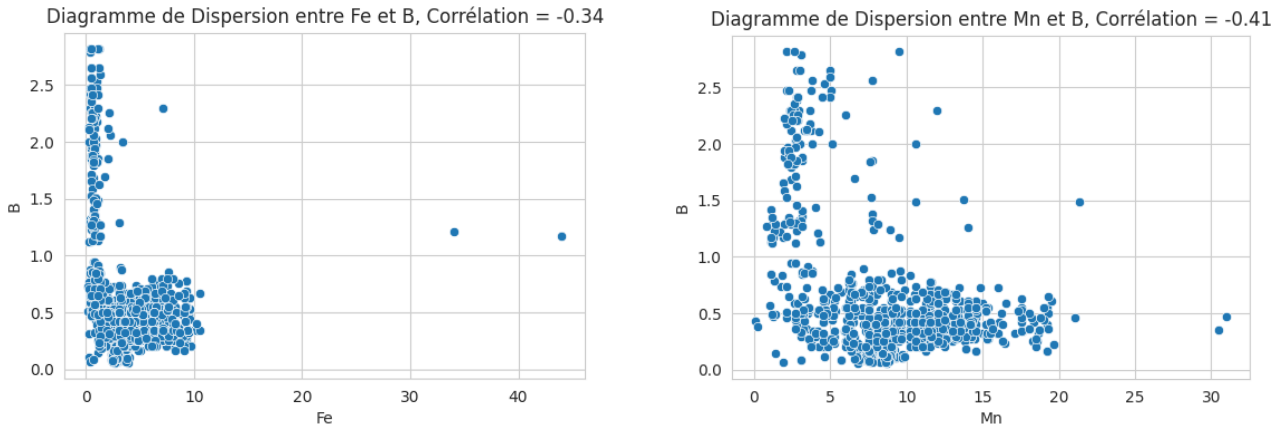
(a) OC-OM (b) OC-Zn

FIGURE 1.13 – Diagramme de dispersion entre OC-OM et OC-Zn

Analyse

Nous remarquons d'après le graphe (a) que chaque fois que la concentration en OC augmente, la concentration en OM augmente également, avec une corrélation très forte égale à 1.
La même observation s'applique à OC et Zn de graphe (b), à l'exception du degré de corrélation qui est égal à 0.63.

2. corrélation négative



(a) Fe-B (b) Mn-B

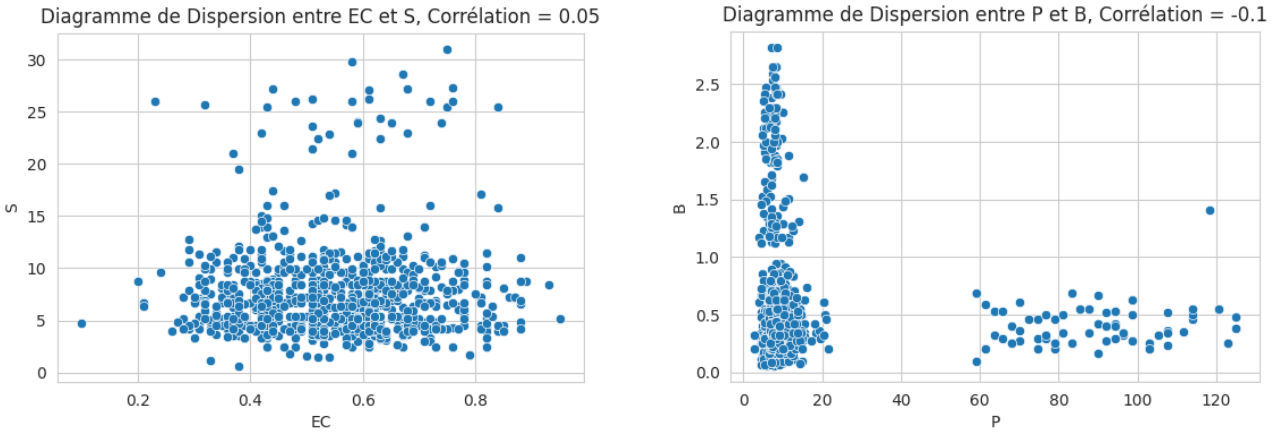
FIGURE 1.14 – Diagramme de dispersion entre Fe-B et Mn-B

Analyse

Nous observons une tendance particulière dans la variation de Fe et B dans le graphe (a), où chaque augmentation de Fe s'accompagne d'une augmentation de B, mais dans des directions opposées.

La même observation s'applique au graphique (b) pour la relation entre Mn et B. La différence réside dans le coefficient de corrélation, étant plus élevé pour OC-OM que pour Mn-B.
On note que |r|<0,5 c'est donc une faible corrélation.

3. Absence de corrélation



(a) EC-S (b) P-B-Zn

FIGURE 1.15 – Diagramme de dispersion entre EC-S et P-B

Analyse
On constate qu'il n'y a pas de relation linéaire systématique ni entre EC et S, ni entre P et B. Les données sont distribuées de manière aléatoire.

Conclusion
En conclusion, l'analyse des graphes (a) et (b) révèle des relations intéressantes entre certaines variables, dans le cas de la concentration en OC et OM, ainsi que pour OC et Zn. Et il y a aussi différentes valeurs de coefficients de corrélation, c'est-à-dire différentes corrélations entre les propriétés du sol.

1.4 Prétraitement

1.4.1 Traitement des valeurs manquantes et aberrantes

Choix de la méthode de remplacement des valeurs manquantes

Les valeurs manquantes dans les données peuvent avoir un impact significatif sur le processus de data mining. Réduction de la taille de l'échantillon, Perte d'informations, Robustesse des modèles... Voila l'affichage des valeurs manquantes dans notre dataset :

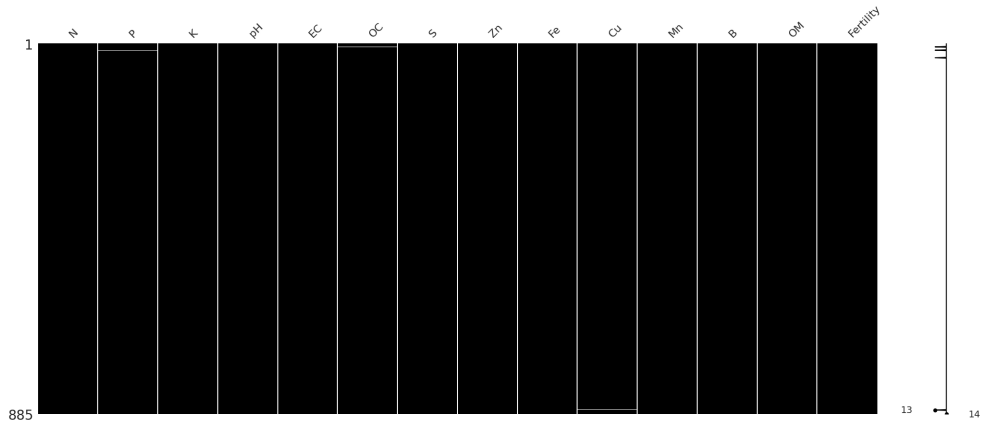


FIGURE 1.16 – Les valeurs nulles pour chaque attributs

Analyse

Après avoir observé les valeurs manquantes représentées dans le graphe de la figure 1.16 pour chaque attribut, il est noté que leur nombre n'est pas significativement élevé : 2 pour P, 1 pour OC et Cu. Afin de traiter ces valeurs manquantes, trois approches sont envisagées :

1. **Suppression des Valeurs Manquantes :**
   L'une des méthodes consiste à supprimer les lignes qui contiennent des valeurs manquantes. Dans ce cas, un total de 4 lignes seraient éliminé de l'ensemble de données.
2. **Remplacement par la Moyenne :**
   Une autre approche suggère de remplacer les valeurs manquantes par la moyenne de leur colonne respective. Ainsi, pour chaque valeur manquante, elle serait remplacée par la moyenne calculée à partir des valeurs disponibles dans la même colonne.
3. **Remplacement par la mediane**

Conclusion

Le choix entre ces trois méthodes dépend des objectifs spécifiques de l'analyse et de la nature des données. La suppression des lignes peut être envisagée si le nombre de valeurs manquantes est limité et n'affecte pas de manière significative la taille de l'ensemble de données. Cependant, si les valeurs manquantes sont réparties de manière homogène et leur suppression risque de créer un biais, le remplacement par la moyenne ou la mediane peut être une alternative plus appropriée pour préserver l'intégrité des données.

Choix de la méthode de traitement des valeurs aberrantes

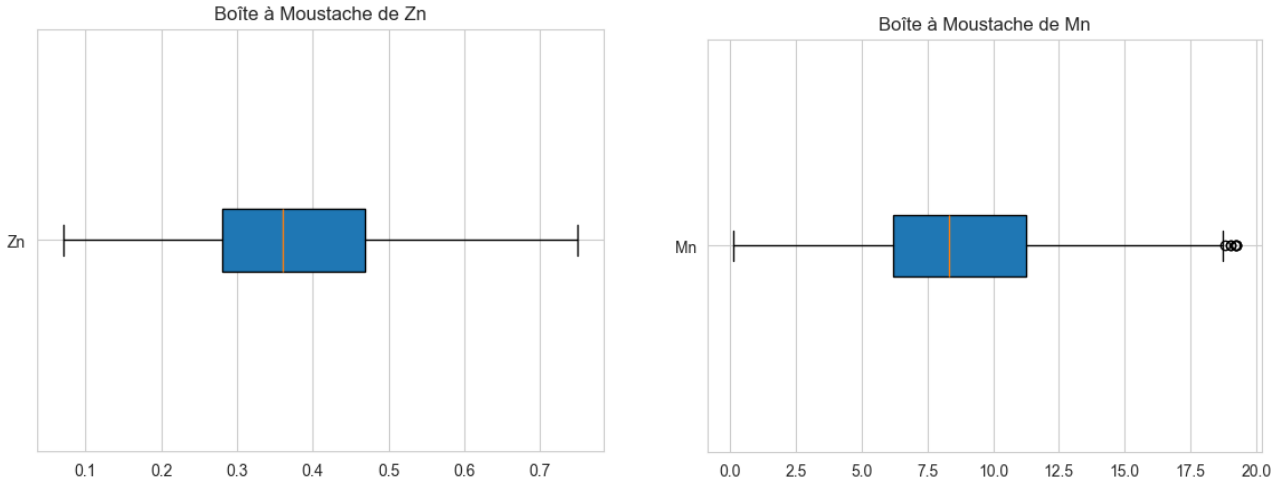
Les valeurs aberrantes, également appelées outliers, peuvent avoir plusieurs effets sur les analyses de données, Influence sur les mesures de tendance centrale, Impact sur la dispersion des données, Déformation des corrélations... Toutes les calculs effectués précédemment doivent être répétés après l'élimination des valeurs aberrantes, car dans ce contexte, les résultats seront plus précis.
L'affichage des ouliers est deja fait avec les boxplot dans la partie Construction des boîtes à moustache pour la détections des données aberrante.
Dans ce rapport nous proposons deux méthode pour les traiter, le calcule de moyenne et la médiane. Voici comment ces deux mesures peuvent être utilisées pour traiter les outliers :

I. Traitement des outliers par la moyenne :

a. Identification : Soit visuellement (boîte à moustaches) soit en calculant l'écart type

b. Remplacement : remplacer ces valeurs aberrantes par la moyenne des valeurs non aberrantes.

Cette méthode élimine bien les outliers de N, P, K, EC, OC, Zn, Fe, Cu, OM. Il reste cependant quelques-uns dans le pH et le B, mais le nombre n'est pas réduit pour S et Mn.



(a) Cas parfait (b) Cas non reduit

FIGURE 1.17 – Boite a moustache sans valeurs aberrantes

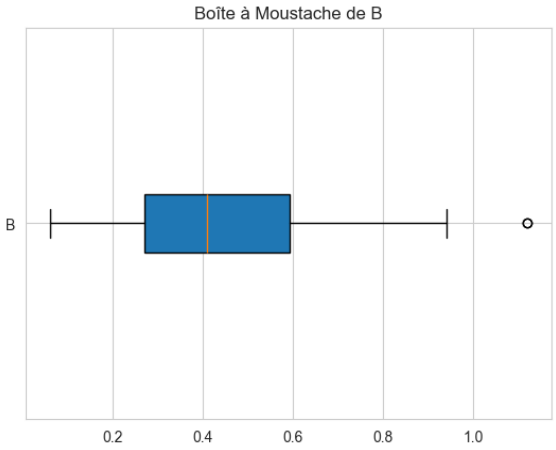


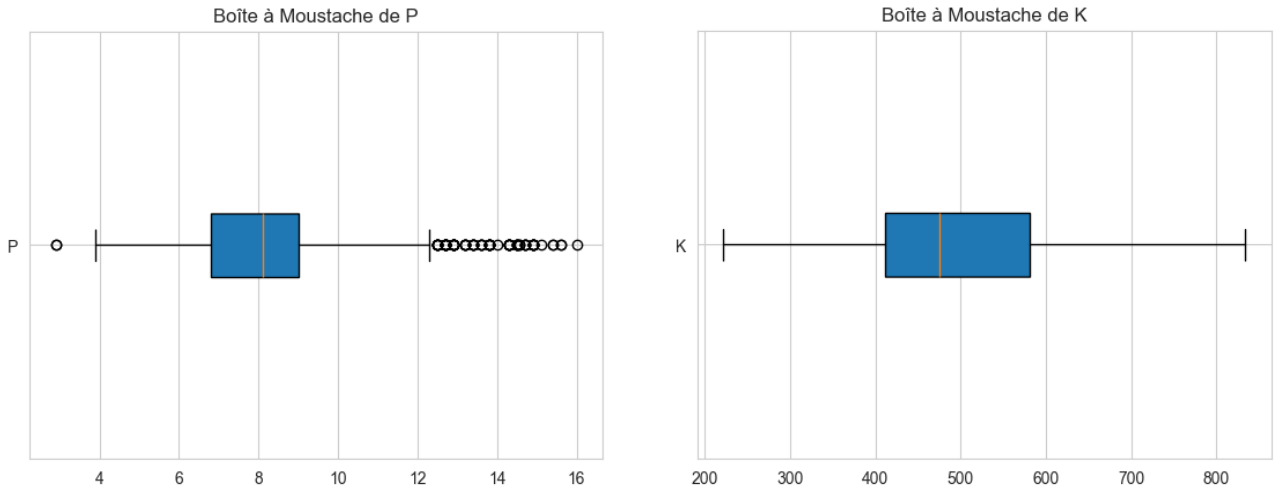
FIGURE 1.18 – Cas moyen

II. Traitement des outliers par la médiane :

a. Identification : Tout comme pour la moyenne

b. Remplacement : Une alternative consiste à remplacer les outliers par la médiane de l'ensemble des données.

Cette méthode élimine bien les outliers de N, K, Ec, OC, Fe, Cu, OM. Il reste cependant quelques-uns dans le pH et le Zn, mais le nombre n'est pas réduit pour P, S, Mn et B.



(a) Cas non reduit (b) Cas parfait

FIGURE 1.19 – Boite a moustache sans valeurs aberrantes

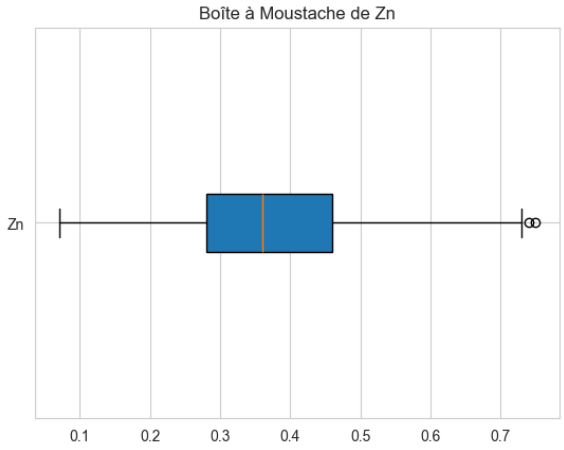


FIGURE 1.20 – Cas moyen

Après avoir appliqué les deux méthodes, on remarque bien que les deux peuvent éliminer les outliers, mais pas définitivement. Il en reste quelques-uns (la moyenne a traité les outliers mieux que la médiane). Nous proposons donc de choisir la première méthode, celle de la moyenne, et d'éliminer le reste des outliers par suppression tant que le nombre d'outliers n'est pas élevé.

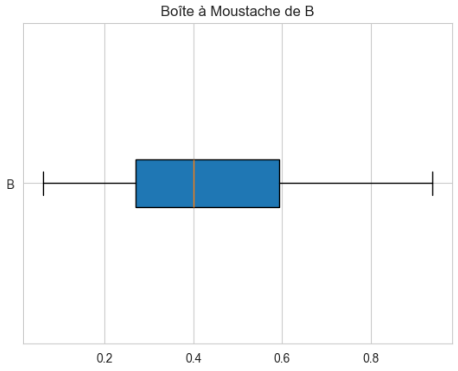


FIGURE 1.21 – Élimination des valeurs aberrantes de B

Conclusion

En fin de compte, le choix entre la moyenne et la médiane dépend de la nature spécifique de vos données et des objectifs de votre analyse. Il peut également être utile d'explorer les deux approches et de comparer les résultats pour prendre une décision éclairée, comme nous l'avons fait ici en choisissant la moyenne.

1.4.2 Réduction des données (élimination des redondances)

L'objectif est généralement de diminuer la complexité des données tout en minimisant la perte d'informations importantes. L'élimination des redondances dans ce contexte se réfère à la suppression des informations en double ou inutiles dans l'ensemble de données. Les dimensions du jeu de données

```
NewData.shape
```

(883, 14)

verticales

```
def eliminer_redondance_colonnes(df):
    # Supprimer les colonnes redondantes (doublons)
    df_sans_redondance = df.T.drop_duplicates().T
    return df_sans_redondance
```

Élimination de redondances : Identifiez et supprimez les variables qui présentent une corrélation élevée entre elles, car elles peuvent apporter des informations similaires. Par exmple r=1 entre OC et OM donc on peut supprimer OM.

horizontales

```
def eliminer_redondance_lignes(df):
    # Supprimer les lignes redondantes (doublons)
    df_sans_redondance = df.drop_duplicates()
    return df_sans_redondance
```

```
NewData.shape
```

(880, 14)
On remarque que le nombre de lignes a diminué de 3 lignes.

1.4.3 Normalisation des données

La normalisation des données est un processus visant à ajuster l'échelle des variables dans un ensemble de données. L'objectif principal de la normalisation est de rendre les données comparables et de faciliter le processus d'analyse en garantissant que chaque variable contribue de manière équitable aux calculs, indépendamment de ses unités d'origine ou de son ordre de grandeur.

Méthode Min-Max

Cette méthode redimensionne les valeurs d'une variable pour qu'elles tombent dans une plage spécifique, généralement entre 0 et 1. La formule est la suivante :

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn	B	OM
0	0.264	0.435	0.552	0.459	0.56	0.513	0.371	0.25	0.00967	0.338	0.449	0.0568	0.513
1	0.489	0.351	0.19	0.606	0.733	0.82	0.488	0.338	0.0629	0.721	0.145	0.606	0.82
2	0.339	0.511	0.81	0.578	0.413	0.863	0.965	0.338	0.0629	0.736	0.135	0.606	0.863
3	0.321	0.298	0.413	0.624	0.507	0.718	0.488	0.397	0.0319	0.716	0.133	0.606	0.718
4	0.661	0.534	0.363	0.615	0.267	0.65	0.788	0.265	0.0532	0.796	0.121	0.606	0.65

Table 1.6 – Portion Résultat Normalisation Min-Max

Méthode z-score

Cette méthode transforme les données pour qu'elles aient une moyenne de 0 et un écart-type de 1.
La formule est la suivante :

$$Z = \frac{X - \mu}{\sigma}$$

	N	P	K	pH	EC	OC	S	Zn	Fe	Cu	Mn
0	-1.550	-0.101	0.522	-0.101	0.539	0.365	-0.348	-1.150	-1.440	-0.395	0.044
1	-0.512	-0.474	-1.354	0.606	1.462	1.580	0.294	-0.647	-1.229	1.281	-1.406
2	-1.204	0.238	1.857	0.473	-0.242	1.749	2.936	-0.647	-1.229	1.346	-1.454
3	-1.287	-0.712	-0.196	0.694	0.255	1.175	0.294	-0.312	-1.352	1.259	-1.466
4	0.276	0.339	-0.458	0.650	-1.023	0.905	1.958	-1.067	-1.268	1.607	-1.521

Table 1.7 – Portion Résultat normalisées Z-score

1.5 Conclusion

Le premier chapitre de notre étude nous a introduit aux différentes étapes nécessaires pour nettoyer une base de données et spécifiquement avec des données statistiques. Nous avons suivi plusieurs approches afin d'obtenir une version propre et efficace de notre jeu de données.
Au cours de cette phase, nous avons traité divers problèmes souvent rencontrés dans un ensemble de données, tels que les valeurs manquantes, les valeurs aberrantes, la réduction de données, ainsi que la normalisation.
Chaque étape de ce processus revêt une importance particulière, et chacune d'entre elles a un impact significatif sur les résultats que nous obtenons après avoir utilisé ce jeu de données nettoyé. Ces préparations sont cruciales pour garantir la fiabilité et la validité des analyses ultérieures.
Dans la deuxième partie de notre projet, consacrée à l'apprentissage supervisé et non supervisé, nous tirerons pleinement parti de ce jeu de données préalablement nettoyé.

Chapitre 2

Données temporelles

2.1 Objectifs

L'objectif principal de cette section est de mener une analyse approfondie et de mettre en œuvre un processus de nettoyage pour le dataset 2. Cette étape revêt une importance cruciale, car elle vise à préparer ces données en vue de leur utilisation ultérieure pour répondre à des questions spécifiques et découvrir des corrélations intéressantes dans le contexte des tests de COVID aux états unis au fil du temps.

2.2 Manipulation du dataset

2.2.1 Importation et visualisation le contenu du dataset

Voici le résultat de l'importation et la visualisation des premières lignes de notre dataset.

ZCTA	Time Period	Population	Start Date	End Date	Case Count	Test Count	Positive Tests
95129	32	39741	10/11/2020	10/31/2020	22.0	2543.0	23.0
95129	43	39741	5/30/2021	6/19/2021	NaN	3315.0	14.0
95129	40	39741	3/28/2021	4/17/2021	34.0	4816.0	37.0
95129	55	39741	2/6/2022	2/26/2022	110.0	10194.0	175.0
95129	44	39741	6/20/2021	7/10/2021	14.0	3033.0	17.0
94085	59	23223	1-May	21-May	165.0	2315.0	192.0
94085	63	23223	24-Jul	13-Aug	150.0	1348.0	190.0
94085	61	23223	12-Jun	2-Jul	219.0	1696.0	255.0
94085	27	23223	28-Jun	18-Jul	53.0	1379.0	61.0
94085	57	23223	20-Mar	9-Apr	30.0	1949.0	34.0

Table 2.1 – Dataset 2 Description-1

ZCTA	Start Date	End Date	Case Rate	Test Rate	Positivity Rate
95129	10/11/2020	10/31/2020	2.6	304.7	0.9
95129	5/30/2021	6/19/2021	1.1	397.2	0.4
95129	3/28/2021	4/17/2021	4.1	577.1	0.8
95129	2/6/2022	2/26/2022	13.2	1221.5	1.7
95129	6/20/2021	7/10/2021	1.7	363.4	0.6
94085	1-May	21-May	33.8	474.7	8.3
94085	24-Jul	13-Aug	30.8	276.4	14.1
94085	12-Jun	2-Jul	44.9	347.8	15.0
94085	28-Jun	18-Jul	10.9	282.8	4.4
94085	20-Mar	9-Apr	6.2	399.6	1.7

Table 2.2 – Dataset 2 Description-2

La dataset fournie semble être liée aux tests de COVID-19 effectués dans differents regions. Voici une explication des colonnes présentes dans la dataset :

zcta (ZIP Code) : Il s'agit du code postal associé à des cites aux états unis.

time period : C'est une valeur numérique représentant une période de temps spécifique associée aux tests de COVID. Par exemple, 32 pourrait correspondre à une semaine particulière.

population : Indique la population totale de la région associée au code postal.

Start date et end date : Ces colonnes indiquent le début et la fin de la période pour laquelle les données des tests de COVID sont enregistrées.

case count : Le nombre de cas de COVID-19 confirmés pendant la période spécifiée.

test count : Le nombre total de tests de COVID-19 effectués pendant la période.

positive tests : Le nombre de tests de COVID-19 qui ont été positifs pendant la période.

case rate : Taux de cas, représentant le nombre de cas pour 100 000 personnes dans la population.

test rate : Taux de tests, représentant le nombre de tests pour 100 000 personnes dans la population.

positivity rate : Taux de positivité, représentant le pourcentage de tests positifs par rapport au nombre total de tests effectués.

Cette dataset est conçue pour permettre l'analyse des données liées à la pandémie de COVID-19 dans une région spécifique, en mettant en évidence des aspects tels que le nombre de cas, le nombre de tests effectués, les taux de cas et de tests, ainsi que le taux de positivité au fil du temps. Cette analyse pourrait aider à comprendre la propagation du virus, l'efficacité des tests, et d'autres tendances importantes dans la région associée.

2.3 Analyse des caractéristiques des attributs du dataset

2.3.1 Description de chaque attribut du dataset

Voici la Description de chaque attribut de notre dataset. L'analyse des statistiques descriptives de

Nom	Valeur non null	Type	Nombre de valeur unique
zcta	337	int64	7
time_period	337	int64	51
population	337	int64	7
Start date	337	object	99
end date	337	object	99
case count	311	float64	205
test count	325	float64	321
positive tests	310	float64	210
case rate	337	float64	209
test rate	337	float64	323
positivity rate	337	float64	128

TABLE 2.3 – la Description de chaque attribut du dataset 2

chaque colonne de la dataset fournit des informations importantes sur la nature et la distribution des données. Voici quelques observations basées sur les statistiques fournies :

Il y a sept codes postaux différents dans la dataset avec 51 périodes temporelles distinctes, indiquant probablement des semaines ou des intervalles de temps spécifiques. pour les colonnes case count, test count, positive tests, case rate, test rate, positivity rate Ces colonnes représentent des mesures numériques. Elles ont différentes valeurs non nulles et un nombre variable de valeurs uniques. Les valeurs manquantes (NaN) sont présentes dans case count , test count et positive tests.

2.3.2 Calcul des mesures de tendance centrale pour déduire les symétries

Un indicateur de tendance centrale est une valeur résumant un ensemble de données pour une variable quantitative ou ordinale. Il s'agit de résumer une série statistique par une seule valeur qui est la plus représentative autour de laquelle se concentrent les données d'une distribution. Il existe trois mesures de tendances centrales : La moyenne arithmétique, la médiane et le mode. en plus des quartiles .

Voici les tendances centrale des attributs de notre dataset.

on peut deduire qu' on a une Distribution étalée à droite pour toutes les attributs de notre dataset. car on a : Moyenne > Mediane > Mode

Attribute	Mean	Median	Mode	Max	Min	Q0	Q1	Q2	Q3	Q4
case count	225.99	91.0	0.0	3627.0	0.0	0.0	39.5	91.0	235.0	3627.0
test count	4938.12	4352.0	1295.0	20177.0	11.0	11.0	2428.0	4352.0	6659.0	20177.0
positive tests	380.2	108.5	20.0	35000.0	11.0	11.0	47.25	108.5	282.0	35000.0
case rate	19.39	8.1	0.0	260.7	0.0	0.0	3.3	8.1	19.1	260.7
test rate	454.84	427.1	0.1	1615.1	0.1	0.1	249.7	427.1	614.9	1615.1
positivity rate	5.83	3.0	1.1	100.0	0.0	0.0	1.3	3.0	6.6	100.0

TABLE 2.4 – Les tendances centrale des attributs de dataset 2

2.4 Prétraitement

2.4.1 Date Formats

Après avoir trié notre ensemble de données, il a été observé que les formats des dates variaient considérablement, et que certaines années étaient manquantes. Étant conscient de l'importance cruciale de ces données, leur suppression aurait constitué un risque. Cependant, Nous avons effectué une corrélation entre l'attribut "time period" et les années de début et de fin.

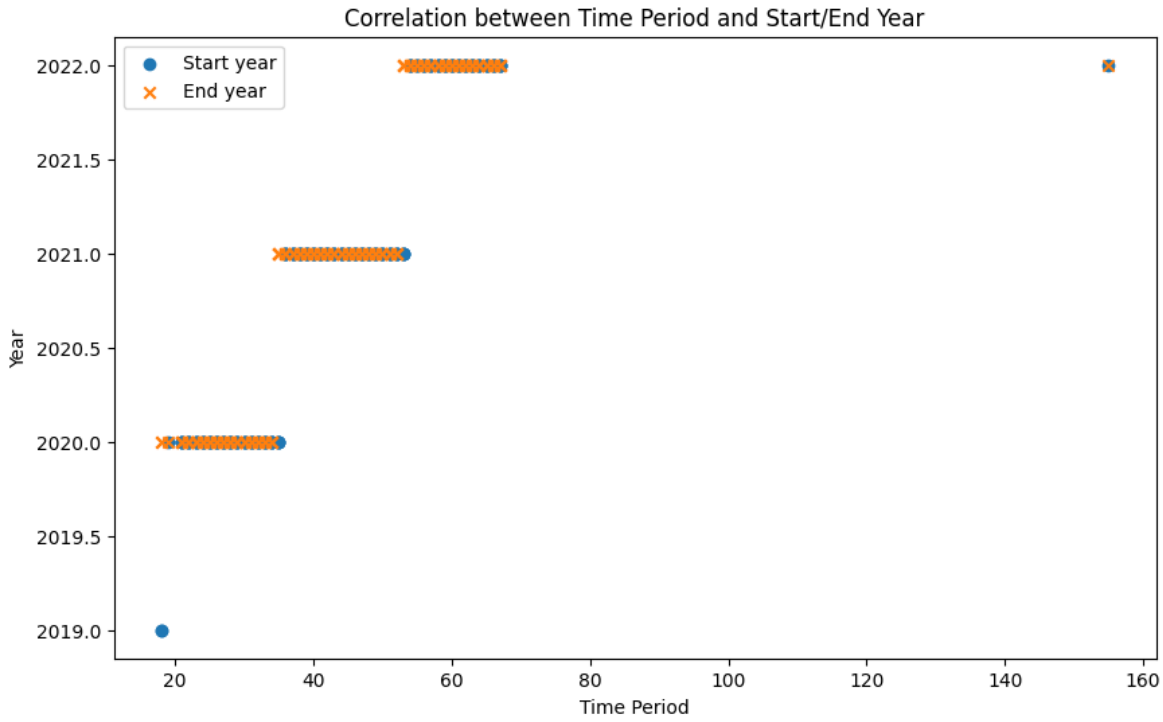


FIGURE 2.1 – Graphe de correlation entre Time period et Start and End Year

Cette corrélation s'est révélée positive, donc l'utilisation de l'attribut "time period" pour imposer des limites temporelles aux données. En conséquence, nous avons utilisé cette corrélation positive pour formater les dates de manière uniforme, tout en restreignant les années aux plages définies par l'attribut "time period". Cette approche garantit la cohérence des données tout en respectant les contraintes temporelles inhérentes à notre ensemble de données.

2.4.2 Traitement des valeurs manquantes et aberrantes

Valeurs manquantes :

Les valeurs manquantes et aberrantes sont des contraintes courantes rencontrées lors de la phase d'analyse des données. Grandement handicapante lors des différentes opérations effectuées nécessitant le dataset, il est impératif d'y remédié au plus tôt. C'est pour cela que dans cette sous-secton, nous nous attardons sur différentes méthodes permettant de les gérer. Comme on peut le constater, les attributs concernés sont : Case Count , Test Count et Positive Tests,

pour le choix de traitement des valeurs manquantes on a utilise les mesures de tendances (Mean , Mediane) .

Le choix entre remplacer les valeurs manquantes par des mesures centrales (comme la moyenne, la médiane ou le mode) ou supprimer les lignes contenant des valeurs manquantes , Voici quelques raisons courantes d'opter pour le remplacement plutôt que la suppression :

Préservation des données existantes : Supprimer des lignes contenant des valeurs manquantes peut entraîner une perte d'informations potentiellement utiles. Si les autres attributs de la ligne sont pertinents, les conserver peut contribuer à une analyse plus complète.

Maintien de la taille de l'échantillon : La suppression des lignes avec des valeurs manquantes peut réduire la taille de l'échantillon, ce qui peut affecter la précision statistique et la puissance de l'analyse.

Conservation de la distribution des données : Le remplacement par des mesures centrales peut permettre de conserver la distribution générale des données, ce qui peut être important dans certaines analyses.

Prévention de biais potentiel :la suppression peut entraîner un biais dans les résultats de l'analyse. Le remplacement par des mesures centrales peut réduire ce risque.

Valeurs aberrantes :

Les valeurs aberrantes sont des valeurs distantes comparées à la majorité des valeurs d'un attribut. Ces valeurs peuvent s'avérer problématiques dans de nombreux cas et il est important de les prendre en charge afin d'éviter toutes erreurs ou anomalies dans de futur traitement. Dans le cas où elles sont peu nombreuses. On peut se permettre de juste les supprimer. Mais dans la majorité des cas, il est préférable de les remplacer par l'une des mesures de tendances. pour les attributes case count , test count, positive tests obtenons les boites à moustache suivantes :

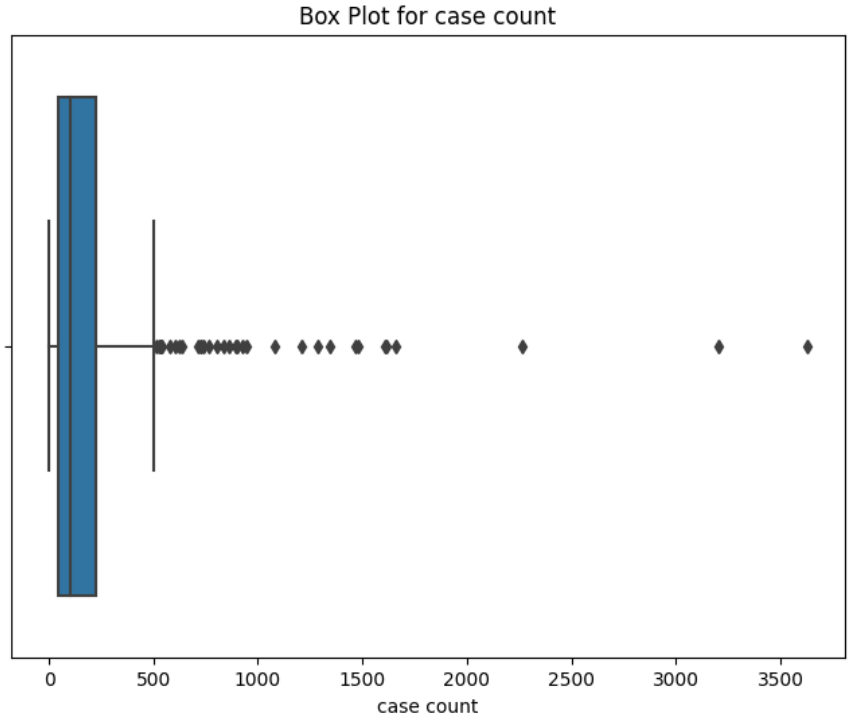


FIGURE 2.2 – Boite à moustache d'attribute case count avant le traitement

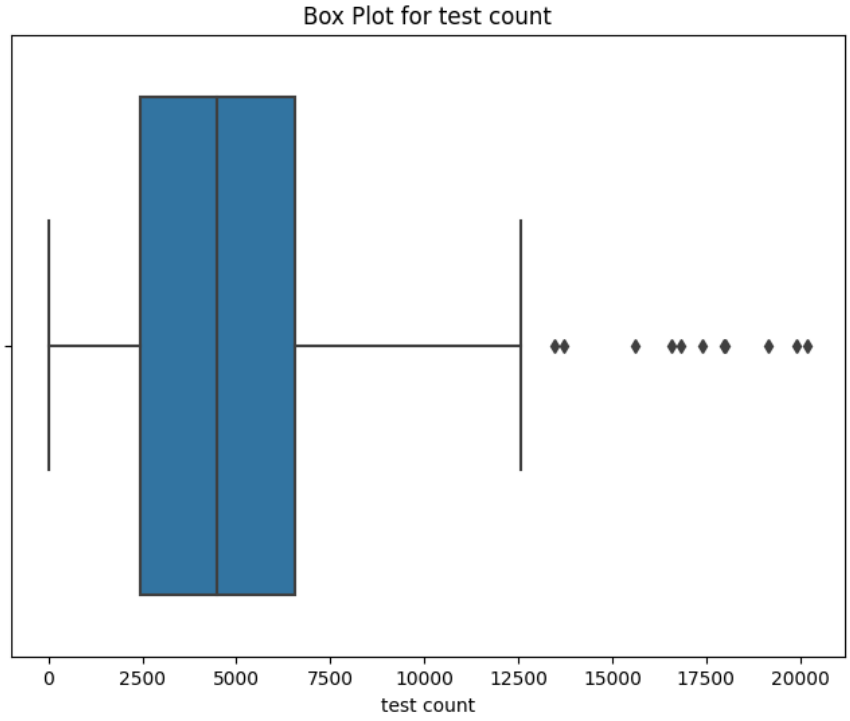


FIGURE 2.3 – Boite à moustache d'attribute test count avant le traitement

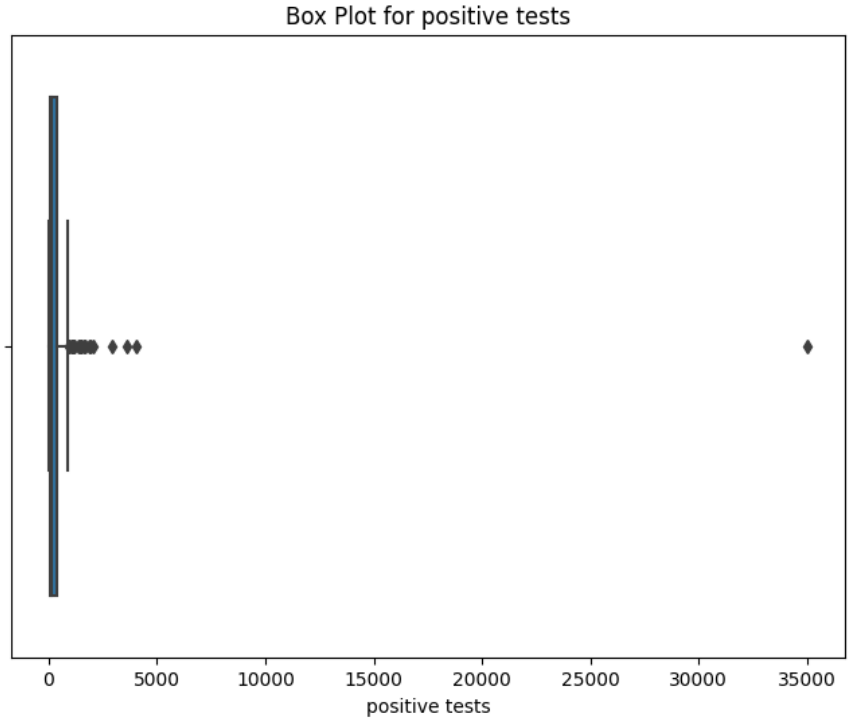


FIGURE 2.4 – Boite à moustache d'attribute positive tests avant le traitement

et apres avoir remplacer ses valeurs avec la mesure de tendance median on a obtenue :

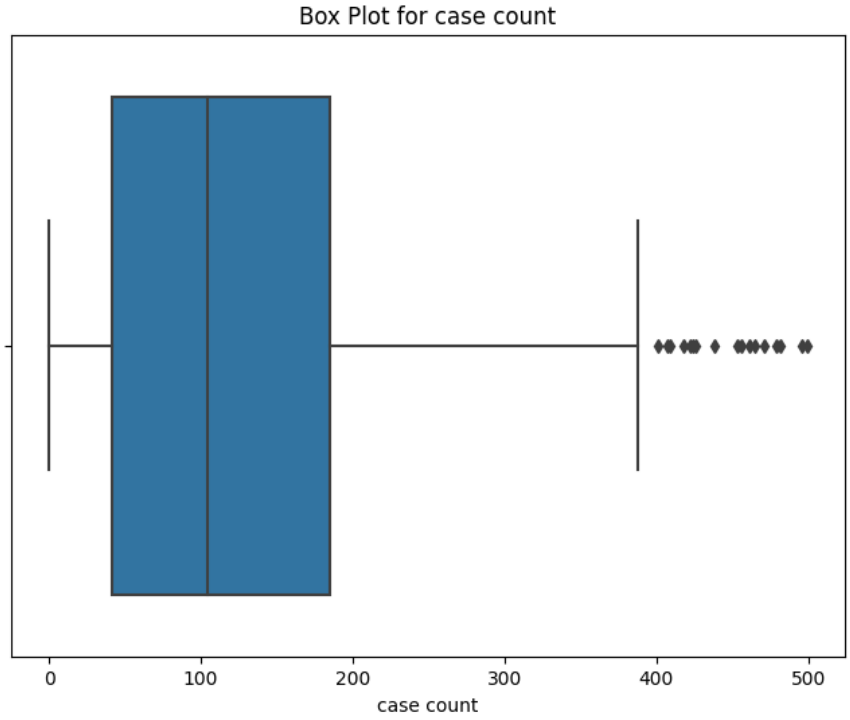


FIGURE 2.5 – Boite à moustache d'attribute case count apres le traitement

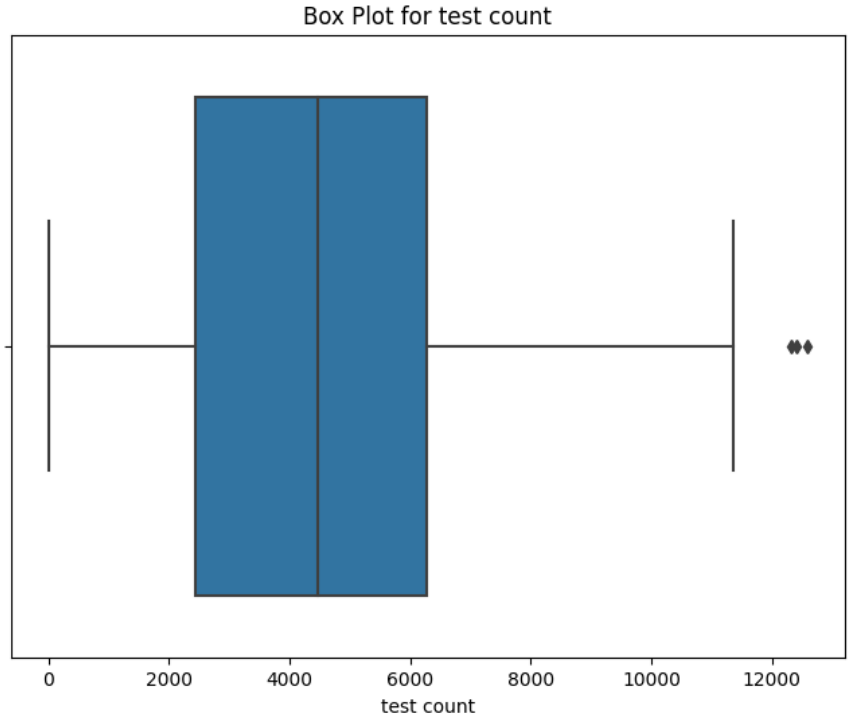


FIGURE 2.6 – Boite à moustache d'attribute test count apres le traitement

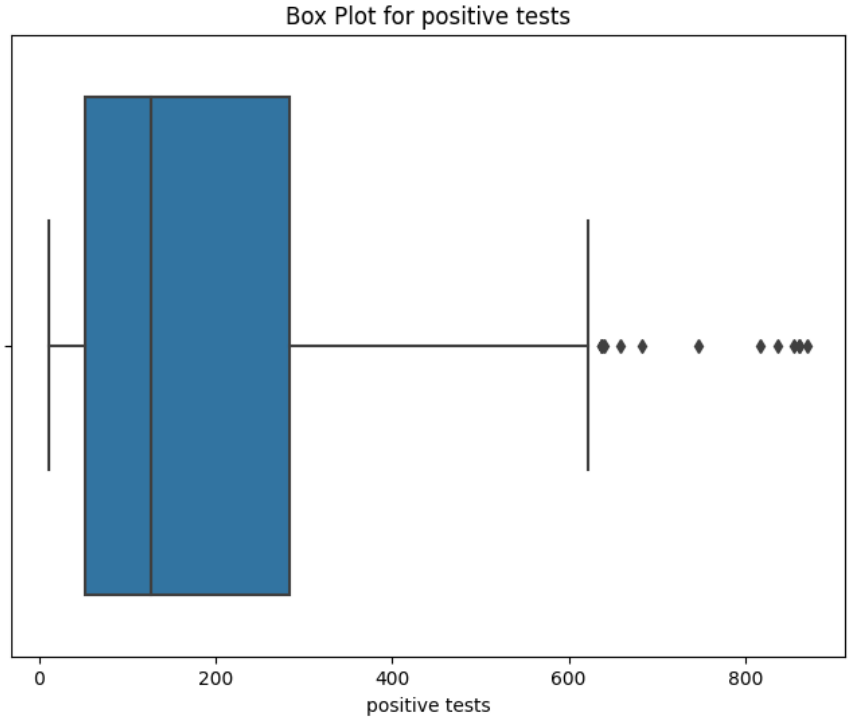


FIGURE 2.7 – Boite à moustache d'attribute positive tests apres le traitement

D'apres les figures on remarque que lorsque on remplace les valeurs aberrantes par la médiane, cela signifie que les valeurs qui étaient à l'origine considérées comme aberrantes (en dehors de la plage définie par 1,5 fois l'écart interquartile) sont désormais remplacées par une valeur médiane.

Cela a pour effet de rendre la distribution plus centrée autour de la médiane, mais cela ne garantit pas nécessairement que toutes les valeurs aberrantes ont été supprimées. Si la distribution des données

est telle que certaines valeurs sont extrêmes mais ne sont pas considérées comme aberrantes selon le critère de l'écart interquartile, elles ne seront pas affectées par ce processus.

Si notre objectif est de supprimer toutes les valeurs qui sont considérées comme des valeurs aberrantes, on peut envisager une approche plus stricte, par exemple on les supprimant complètement.

Mais vu que le domaine etudie est covid , on prefere de garder toutes les valeurs car ca peut indiques des informations importants si on ajoute des autres elements et facteurs au notre context.

2.5 Analyse et Visualisation du contenue du dataset

Après avoir appliqué les prétraitements nécessaires à notre ensemble de données, nous sommes désormais en mesure d'obtenir des conclusions pertinentes quant aux divers aspects de la pandémie de COVID-19 que nous avons examinés.

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. Ces représentations graphiques offrent une vue d'ensemble claire de la situation, permettant une analyse rapide des zones les plus touchées.

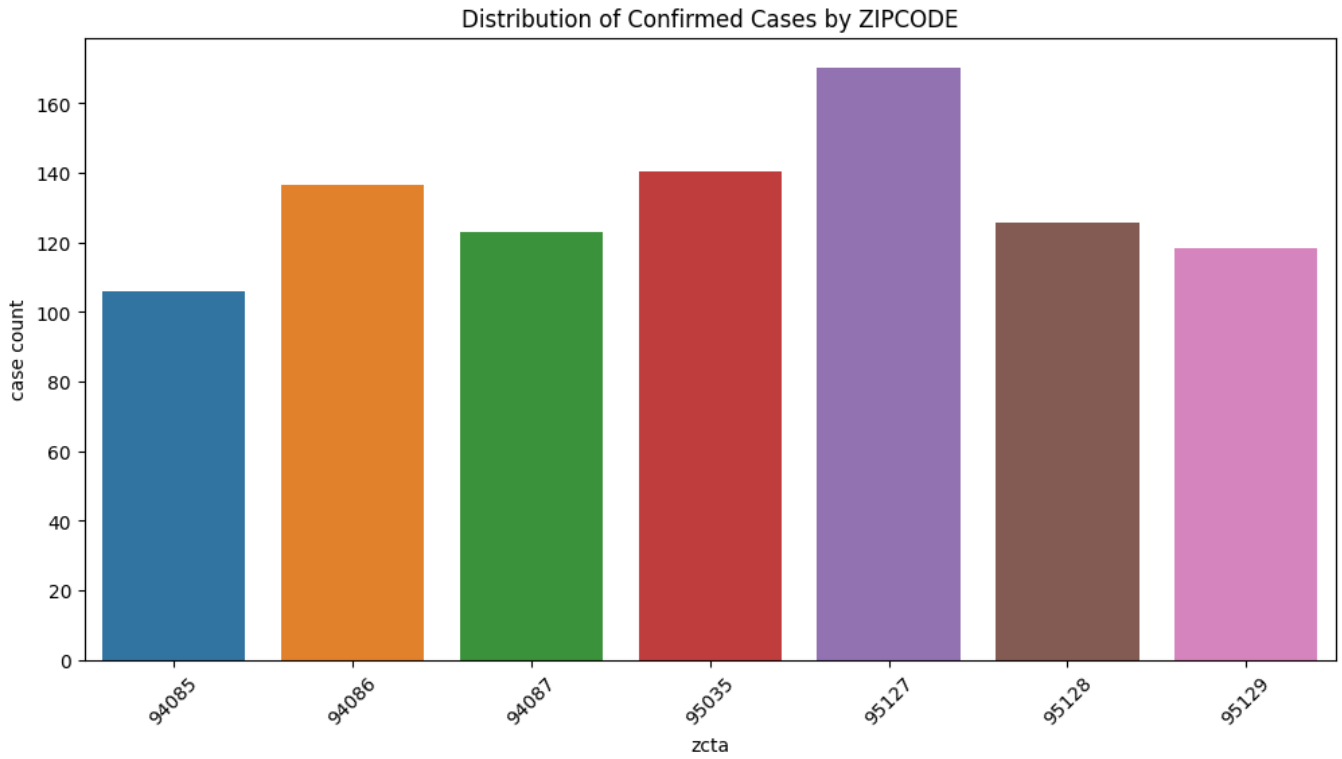


FIGURE 2.8 – Distribution des Cas Confirmés par Zones

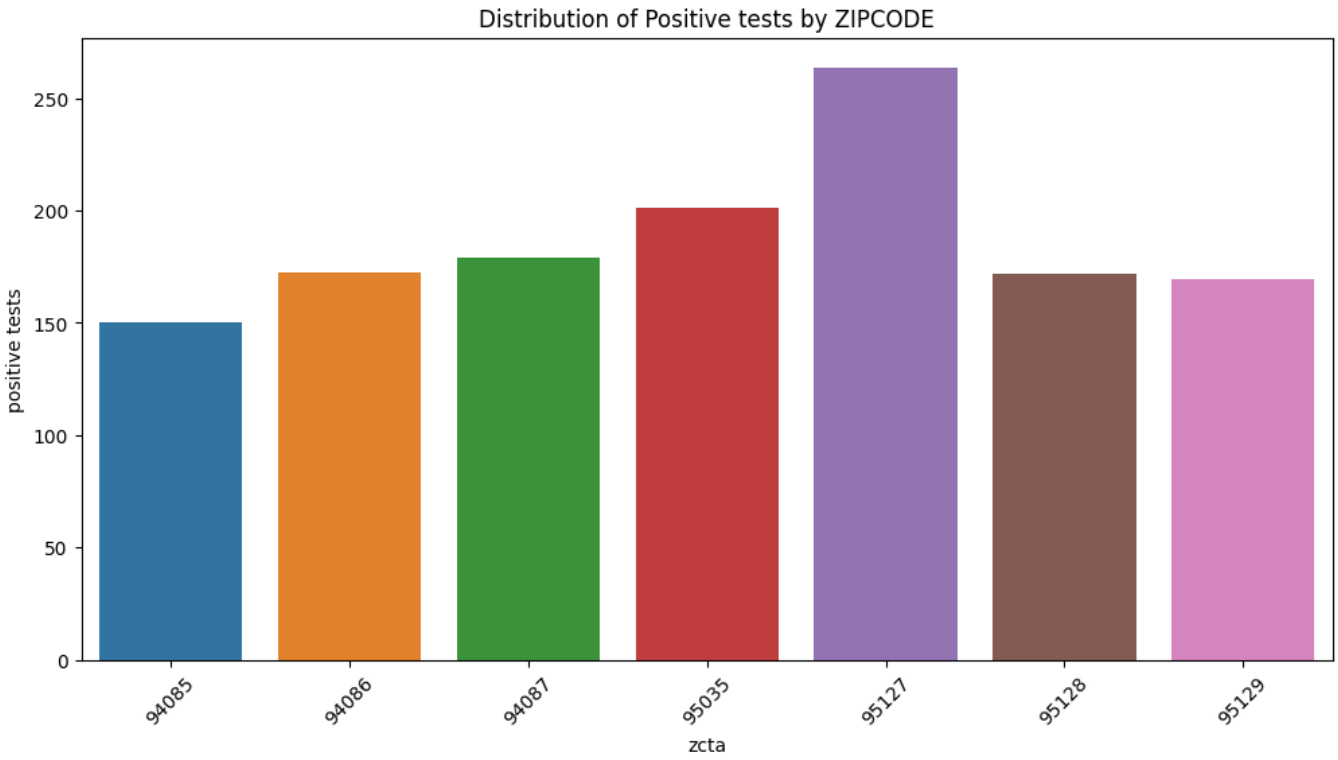


FIGURE 2.9 – Distribution des Tests Positifs par Zones

En observant le graphique, on peut noter que la zone ayant le code postal 95127 enregistre le plus grand nombre de cas confirmés et tests positives, ce qui est cohérent étant donné sa population importante, totalisant 66 256 et ce bar chart montre la distribution de population par zipcode.

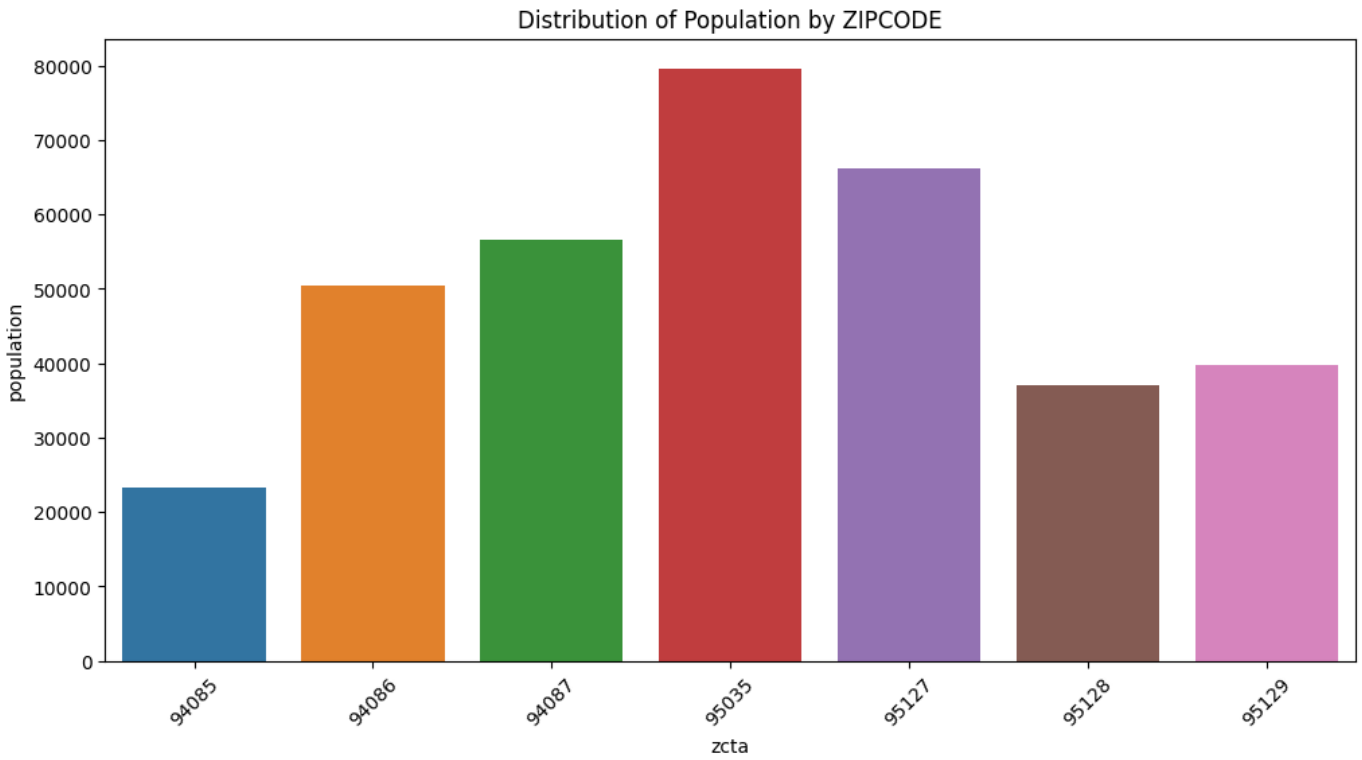


FIGURE 2.10 – Distribution de population par Zones

2. Évolution Temporelle des Tests COVID-19, des Tests Positifs et du Nombre de Cas pour

une Zone Spécifique

Nous avons examiné comment les tests COVID-19, les tests positifs et le nombre de cas évoluent au fil du temps, en adoptant une perspective hebdomadaire, mensuelle et annuelle. Les tendances ont été présentées à l'aide d'un graphique linéaire, offrant ainsi une compréhension approfondie des variations temporelles. la zone choisie est 95127 etant donnees que c'etait la zone avec plus d'informations (tests positifs , nombres des cas)

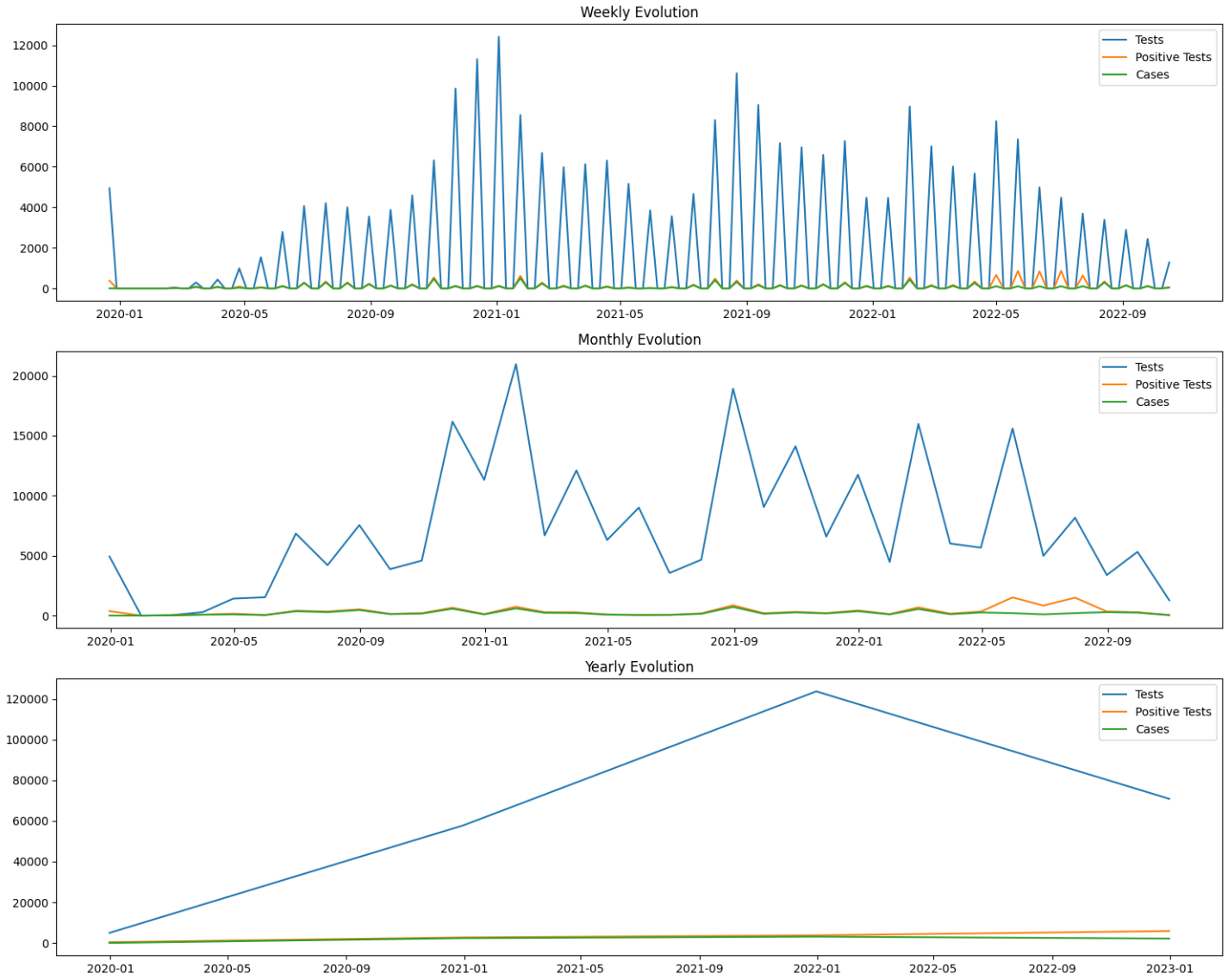


FIGURE 2.11 – Évolution temporelle des Tests COVID-19, des Tests Positifs et du Nombre de Cas pour la zone 95127

Analyse global :

D'après les graphiques, on remarque que le nombre de tests positifs et le nombre de cas déclarés sont les mêmes jusqu'à l'année 2022 (mois 5). Cependant, à partir de cette date, ils ont cessé de déclarer les cas, ce qui explique pourquoi le nombre de tests positifs est légèrement plus élevé que le nombre de cas déclarés. Au début (de janvier 2020 à mai 2020), le nombre de tests effectués, de cas confirmés et de tests positifs était le même, mais il a commencé à augmenter de manière exponentielle. À présent, le nombre de tests effectués est nettement supérieur au nombre de cas confirmés et de tests positifs.

Analyse hebdomadaire et mensuelle :

Chaque semaine, on constate un nombre élevé de tests, avec une moyenne de 5000. Cependant, il est remarquable que, de novembre 2020 à janvier 2021, ainsi qu'autour du mois de novembre 2021, le nombre de tests effectués a considérablement augmenté. Cette hausse pourrait être attribuée probablement à des préparations culturelles liées au Nouvel An ou à Halloween.

Analyse annuelle :

On observe une corrélation entre l'augmentation du nombre de tests effectués, de résultats positifs et de cas confirmés au cours du premier semestre de l'année 2020. Cependant, à partir de mai 2020, le nombre de tests effectués connaît une croissance exponentielle, atteignant son niveau le plus élevé en 2022, pour ensuite connaître une baisse. Pendant ce temps, les tests positifs et les cas confirmés restent stables de le début.

3. Distribution des Cas COVID-19 Positifs par Zone et par Année

Une analyse approfondie de la distribution des cas COVID-19 positifs a été réalisée en utilisant un graphique à barres empilées (Stacked Bar Chart). Cette représentation visuelle permet de visualiser comment les cas positifs sont répartis par zone au cours des différentes années, facilitant la détection de tendances spécifiques.

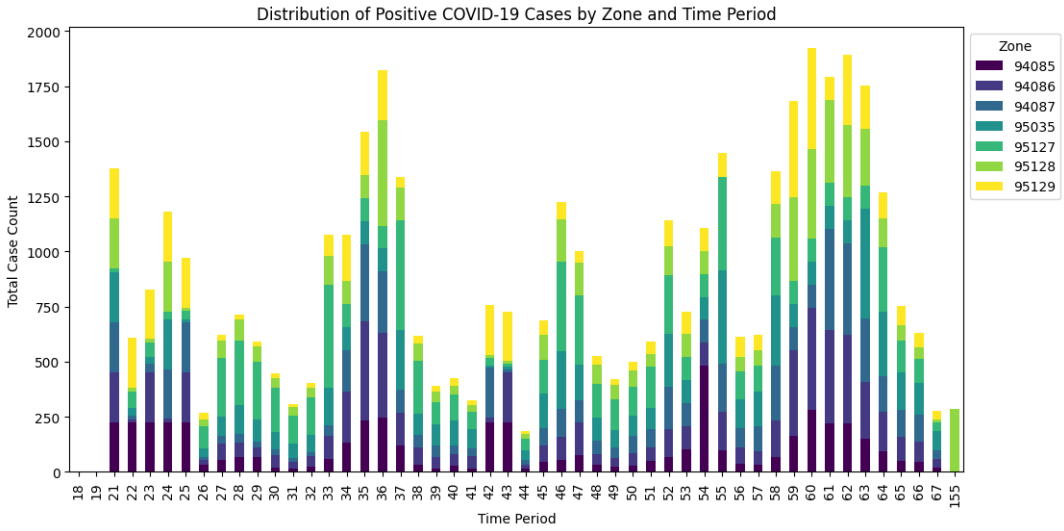


FIGURE 2.12 – Distribution des Cas COVID-19 Positifs par Zone et par Année

D'après le stacked bar, on observe que la zone avec le code postal 95129 présente le nombre le plus élevé de cas de Covid-19 positifs pour toutes les périodes temporelles, tandis que la zone 94085 affiche le nombre le plus bas de cas positifs. Pour la période temporelle 155, les données sont disponibles uniquement pour la zone 95128, mettant en évidence un taux élevé de cas positifs aux période 36 et entre les périodes 59 et 63 pour toutes les zones.

4. Relation Graphique entre la Population et le Nombre de Tests Effectués

Afin de représenter efficacement la relation entre la population et le nombre de tests effectués, nous avons élaboré un diagramme de dispersion (scatter plot) où l'axe des x représente la population et l'axe des y représente le nombre de tests effectués. Cela nous permettra de visualiser la relation entre ces deux variables.

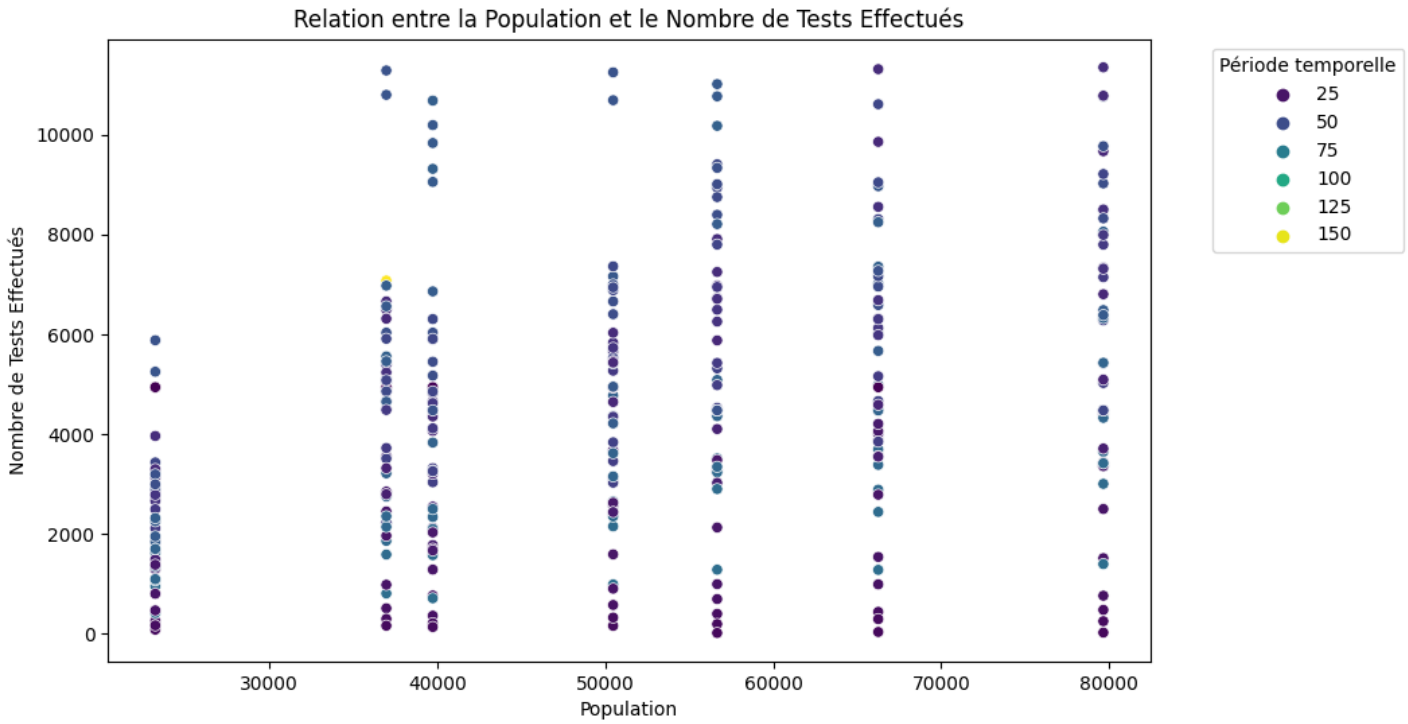


FIGURE 2.13 – Relation entre la Population et le Nombre de Tests Effectués

Nous remarquons d'après le graphe que chaque fois que le nombre de population s'augmente le nombre des testes s'augmente au rapport du temps avec une corrélation très forte égale à 1.

5. Zones les Plus Fortement Impactées par le Coronavirus

En identifiant les cinq zones les plus fortement impactées par le coronavirus par rapport au populations des zones on a élaboré un tree map ou on a fait le rapport entre les cas positifs et population pour chaque zone.

FIGURE 2.14 – Relation entre la Population et le nombre des cas

d'apres la figure on a les 5 zones plus fortement impactees sont en ordre : (94085,95128,95129,94086,95127) nous avons pu mettre en évidence les points chauds épidémiologiques. Ces informations sont essentielles pour cibler les interventions et les ressources là où elles sont le plus nécessaires, contribuant ainsi à une gestion plus efficace de la crise sanitaire.

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

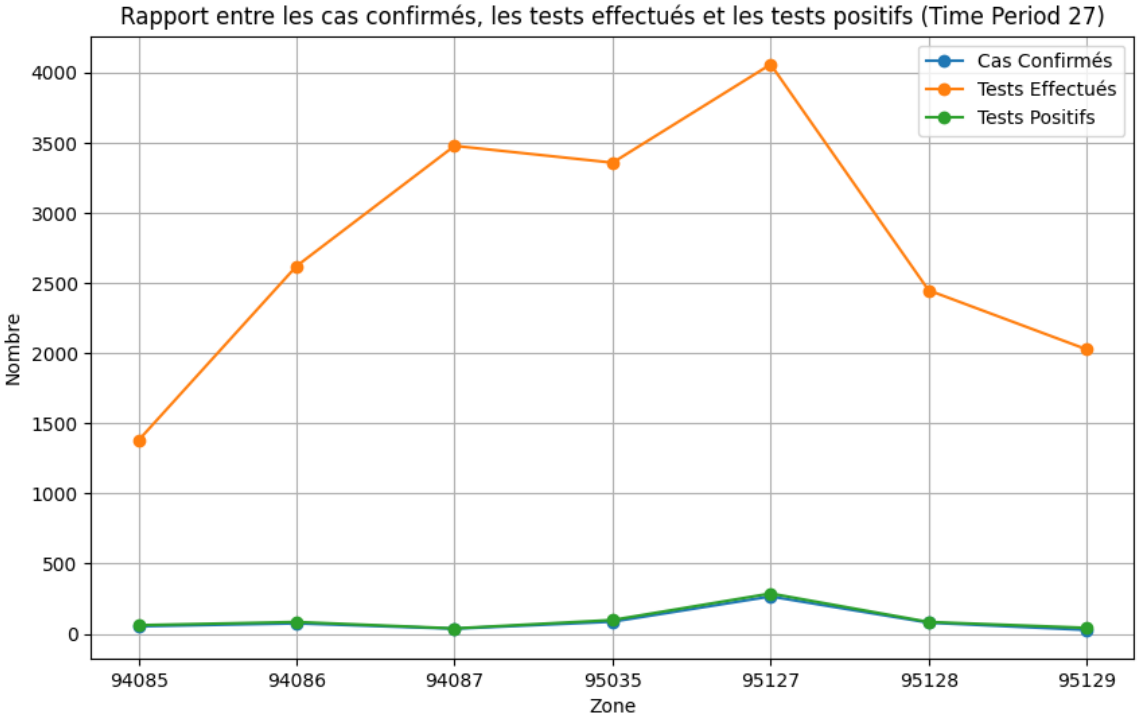


FIGURE 2.15 – Distribution des Cas Confirmés et Tests Positifs par Zones

Il est notable qu'une corrélation positive et significative existe entre les tests positifs et les cas confirmés pour toutes les zones. En revanche, il n'y a aucune corrélation entre les tests effectués et les cas confirmés, ainsi que les tests positifs. La seule explication réside dans le fait que le nombre de tests effectués est nettement plus élevé par rapport aux cas confirmés et aux tests positifs.

2.6 Conclusion

En conclusion, cette étude approfondie des données temporelles liées aux tests de COVID-19 dans différentes zones offre des insights précieux sur l'évolution de la pandémie. Voici quelques conclusions clés basées sur nos observations et analyses :

Évolution Temporelle des Tests et des Cas : Nous avons identifié une croissance exponentielle du nombre de tests effectués à partir de mai 2020, atteignant un pic en 2022, suivi d'une baisse. Pendant ce temps, les tests positifs et les cas confirmés sont restés relativement stables. Il est important de noter que le nombre de tests effectués a dépassé largement le nombre de cas confirmés.

Impacts Culturels sur les Tests : Des pics significatifs dans le nombre de tests effectués ont été observés autour des périodes culturelles, comme les préparations du Nouvel An et d'Halloween. Cela souligne l'influence des événements sociaux sur les comportements liés aux tests de COVID-19.

Distribution Géographique : Certaines zones, telles que celles avec les codes postaux 95127 et 95129, ont enregistré un nombre plus élevé de cas confirmés et de tests positifs. La distribution de la population par zone a également été prise en compte pour évaluer l'impact relatif.

Corrélation entre Tests Positifs et Cas Confirmés : Une corrélation positive et significative a été observée entre les tests positifs et les cas confirmés pour toutes les zones. Cependant, il n'y a pas de corrélation entre les tests effectués et les cas confirmés ou les tests positifs.

Visualisation Pertinente : Les graphiques, diagrammes en arbre et diagrammes de dispersion ont été efficacement utilisés pour représenter visuellement les tendances et les relations entre les variables, facilitant ainsi la compréhension des résultats.

En somme, cette étude offre une base solide pour une analyse plus approfondie de l'impact de la pandémie de COVID-19 dans différentes zones. Les conclusions fournies peuvent orienter les décideurs de la santé publique dans l'allocation efficace des ressources et la mise en œuvre de stratégies adaptées aux particularités de chaque région.

Chapitre 3

Extraction de motifs fréquents, règles d'associations et corrélations

3.1 Objectifs

L'objectif principal de cette section est d'analyser et d'extraire les motifs fréquents, les règles d'association et les corrélations à partir du dataset 3. Nous visons à mettre en lumière les relations existantes entre les attributs relatifs au climat (Température, Humidité, Précipitation), le sol, la végétation et l'utilisation d'engrais. Cette analyse nous permettra de dégager des informations essentielles pour prendre des décisions éclairées dans le contexte de la gestion des ressources environnementales et agricoles.

3.2 Importation et visualisation le contenu du dataset

Voici le résultat de l'importation et la visualisation des premières lignes de notre dataset.

	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	192.90	sandy	Coconut	Urea
4	25.04	95.90	174.80	coastal	Coconut	Urea
5	20.82	84.13	230.22	clay loam	rice	Urea
6	25.95	93.41	172.05	alluvial	Coconut	Urea

Table 3.1 – Description de dataset-3

Cette dataset est liée à l'agriculture, spécifiquement à la croissance des cultures, en fonction de divers paramètres environnementaux. Voici une explication des colonnes de la dataset :

Temperature (Température) : La température ambiante en degrés Celsius. Cela peut influencer la croissance des plantes, car différentes cultures ont des exigences de température spécifiques.

Humidity (Humidité) : Le pourcentage d'humidité relative dans l'air. L'humidité peut affecter la transpiration des plantes et leur capacité à absorber l'eau.

Rainfall (Précipitations) : La quantité de précipitations en millimètres. Les plantes ont des besoins en eau spécifiques, et les précipitations peuvent également affecter la qualité du sol.

Soil (Type de sol) : Le type de sol dans lequel la culture est plantée. Les types de sol, tels que l'argile, le limon, le sable, etc., ont des propriétés différentes qui peuvent influencer la croissance des plantes.

Crop (Culture) : Le type de culture plantée, par exemple, riz, noix de coco, etc. Différentes cultures ont des exigences spécifiques en termes de sol, de température et d'eau.

Fertilizer (Engrais) : Le type d'engrais utilisé pour la culture. Les engrais fournissent des éléments nutritifs essentiels aux plantes. Dans cet ensemble de données, des noms tels que "DAP", "Good NPK", "Urea", et "MOP" sont utilisés, représentant probablement différents types d'engrais.

3.3 Analyse des données

3.3.1 Les tendances centrales des attributs :

Voici les tendances centrale des attributs de notre dataset.

	Temperature	Humidity	Rainfall
Count	295.000000	295.000000	295.000000
Mean	25.522068	88.472271	205.330983
Std	2.495289	6.550542	43.276280
Min	20.050000	80.120000	131.090000
25%	23.810000	82.275000	172.480000
50%	25.760000	84.970000	202.940000
75%	27.170000	94.800000	231.560000
Max	29.870000	99.980000	298.560000

TABLE 3.2 – Statistiques descriptives pour Temperature, Humidity et Rainfall.

3.3.2 Matrice de corrélation

Pour analyser les relation entre les attributs, nous avons calculé la matrice de corrélation suivante :

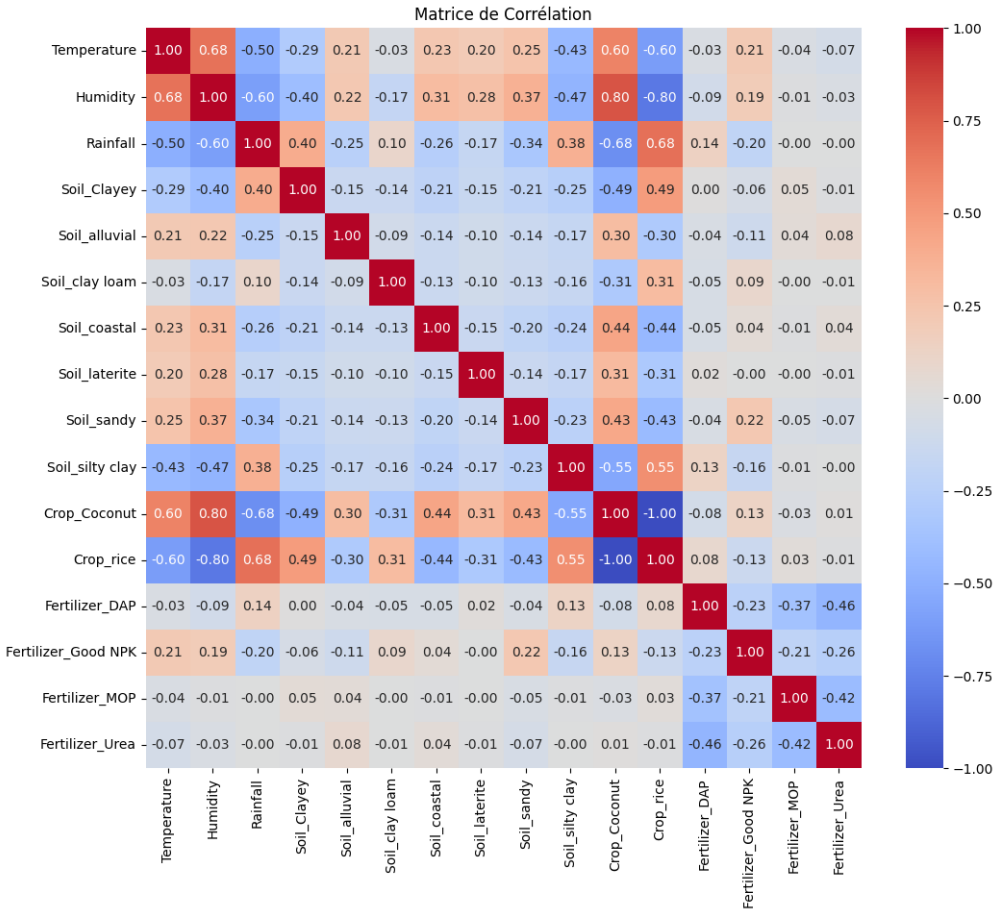


FIGURE 3.1 – la matrice de corrélation pour le dataset3

Conclusion

D'après la matrice de corrélation, on voit bien qu'il y a une corrélation > 0.5 entre température-humidité et température-précipitations. Ainsi, nous pouvons éliminer les deux attributs (humidité et précipitations) et ne conserver que la température pour au lieu les trois.

3.4 La discrétisation des données

La discrétisation des données est le processus de conversion de variables continues en catégories discrètes. Cela peut être nécessaire ou bénéfique dans le contexte de l'analyse de données pour la simplification du modèle vu que La discrétisation permet de simplifier les données en réduisant la complexité des valeurs continues, ce qui peut rendre l'analyse plus facile. On a choisie l'attribut Temperature pour la discrétisation

La discrétisation d'une variable comme la température en classes peut se faire de différentes manières, notamment avec les méthodes de largeur égale (equal width) et de fréquence égale (equal frequency).

3.4.1 Equal Width (Largeur égale) :

Dans la méthode de largeur égale, l'intervalle de valeurs est divisé en plusieurs classes de largeur égale. pour le nombre d'intervalles K on a utilisee la formule de Huntsberger vu que les valeurs de

l'attribut temperature sont tres proches .

$$k = 1 + \frac{10}{3} \cdot \log_{10}(n)$$

La valeur de n : 295

La valeur de k : 9

on obtenu des classes de Largeur : 1.0911111111111111.

et interales : [20.05, 21.141111111111112, 22.232222222222223, 23.323333333333334, 24.414444444444445, 25.505555555555556, 26.596666666666668, 27.68777777777778, 28.77888888888889]
Voila le resultat :

Temperature	Humidity	Rainfall	Soil	Crop	Fertilizer	Temperature_DEw
24.957179	82.84	295.61	Clayey	rice	DAP	4
28.285714	96.65	178.96	laterite	Coconut	Good NPK	7
25.895577	81.64	270.44	silty clay	rice	MOP	0
24.957179	95.02	192.90	sandy	Coconut	Urea	4
24.957179	95.90	174.80	coastal	Coconut	Urea	4

TABLE 3.3 – Descretisation equal width de l'attribut Temperateur

3.4.2 Equal Frequency (Fréquence égale) :

Dans la méthode de fréquence égale, les données sont divisées en classes de sorte que chaque classe contient le même nombre d'observations. Cela signifie que si vous avez 100 observations et que vous voulez 5 classes, chaque classe aurait 20 observations. Cette méthode peut conduire à des classes de largeurs différentes, car elle vise à maintenir le même nombre d'observations dans chaque classe plutôt que d'avoir des intervalles de largeur égale.
Le nombre d'intervalle est calculer avec la formule suivante : $\sqrt{N}$
tel que N est la taille de dataset 3, donc on aura 17 intervalles : [5, 14, 0, 6, 8, 10, 1, 15, 9, 13, 4, 7, 2, 12, 16, 3, 11]

Voila le resultat :

Temperature	Humidity	Rainfall	Soil	Crop	Fertilizer	Temperature_Def
24.713684	82.84	295.61	Clayey	rice	DAP	5
28.365000	96.65	178.96	laterite	Coconut	Good NPK	14
20.522632	81.64	270.44	silty clay	rice	MOP	0
25.138824	95.02	192.90	sandy	Coconut	Urea	6
25.138824	95.90	174.80	coastal	Coconut	Urea	6

TABLE 3.4 – Descretisation equal frequency de l'attribut Temperateur

3.5 Extraction des règles d'associations et corrélations

3.5.1 Algorithme Apriori

L'algorithme Apriori est une technique utilisée dans le domaine de l'exploration de données pour extraire des règles d'association à partir d'un ensemble de données. Les règles d'association mettent en évidence des relations fréquentes entre les éléments d'un ensemble.

L'idée fondamentale derrière l'algorithme Apriori est de rechercher des ensembles d'articles fréquents, c'est-à-dire des groupes d'articles qui apparaissent ensemble dans un ensemble de transactions avec une fréquence supérieure à un seuil prédéfini. En identifiant ces ensembles fréquents, l'algorithme peut générer des règles d'association qui décrivent les relations entre les articles.

L'extraction des règles d'association et des corrélations est une technique utilisée dans le domaine de l'apprentissage automatique pour découvrir des relations intéressantes entre différentes variables dans un ensemble de données.

L'extraction des motifs fréquents est la première étape de l'extraction des règles d'association. Elle consiste à extraire le contexte de l'ensemble d'attributs binaires I. Le problème de recherche des motifs fréquents associés aux données , consiste à déterminer le sous-ensemble Xk $\subset X$

des motifs fréquents ainsi que le support de chaque motif fréquent. Les algorithmes de recherche de ces motifs doivent parcourir la totalité de la base de données chaque fois qu'ils ont déterminé le support de motifs candidats. Dans la plupart de cas, l'espace de recherche est exponentiel, de l'ordre de 2|I| itemsets candidats

L'algorithme Apriori est particulièrement bien adapté pour limiter l'espace de recherche lors de l'extraction des règles d'association dans le contexte de l'analyse de données transactionnelles. car elle se repose sur le principe apriori, qui stipule que si un ensemble d'articles est fréquent, alors tous ses sous-ensembles doivent également l'être. Cela permet de réduire l'espace de recherche en éliminant les ensembles qui ne satisfont pas à cette condition.

Algorithm 1 Algorithme Apriori

Entrée : Ensemble de transactions D, seuil de support minimum $minSup$
Sortie : Ensemble de règles d'association
Initialisation : L_1 = ensembles d'articles fréquents de taille 1
$k \leftarrow 2$
while $L_{k-1} \neq \emptyset$ **do**
    $C_k \leftarrow$ Génération de candidats à partir de L_{k-1}
$L_k \leftarrow$ Filtrage des candidats avec un support minimum dans D
$k \leftarrow k + 1$

Génération des règles d'association : for *chaque ensemble* S *dans* L **do**
    chaque sous-ensemble non vide A de S
$B \leftarrow S - A$
Générer la règle d'association $A \Rightarrow B$

3.5.2 Les Transactions

Attribut	Valeurs uniques
Temperature_DEw	[4, 7, 0, 5, 1, 6, 3, 2]
Fertilizer	['DAP', 'Good NPK', 'MOP', 'Urea']
Crop	['rice', 'Coconut']
Soil	['Clayey', 'laterite', 'silty clay', 'sandy', 'coastal', 'clay loam', 'alluvial']

TABLE 3.5 – Les valeurs uniques des attributs

Crop	Soil
Coconut	['laterite', 'sandy', 'coastal', 'alluvial']
rice	['Clayey', 'silty clay', 'clay loam']

TABLE 3.6 – Relation entre l'attribut Crop et Soil

D'après ces résultats, on peut décider des attributs de nos transactions. Chaque transaction est constituée par les valeurs des attributs Température, Crop et Fertilizer. Le choix de ces attributs est lié à notre dataset et à nos besoins. Nous avons prouvé précédemment que la température, l'humidité et les précipitations sont fortement corrélées. C'est pourquoi nous choisissons un seul attribut parmi les trois, qui est la température. Nous ne prenons pas en compte le sol, car selon le tableau, si Soil est ['laterite', 'sandy', 'coastal', 'alluvial'], le Crop sera Coconut, et si Soil est ['Clayey', 'silty clay', 'clay loam'], le Crop sera Rice.
En fin de compte, la transaction se compose de Température, Crop et Fertilizer afin que nous puissions connaître les valeurs de température adéquates pour le rice et pour coconut, de même que pour le fertilisant.

	Transactions
1	4_rice_DAP
2	7_Coconut_Good NPK
3	0_rice_MOP
4	4_Coconut_Urea
5	4_Coconut_Urea
6	5_Coconut_MOP
7	5_rice_MOP
8	4_rice_MOP
9	5_rice_MOP
10	3_rice_MOP

TABLE 3.7 – Exemple de transactions

3.5.3 Les Regles d'association

Antecedent	Consequent	Confidence
'4'	'rice'	0.564103
'rice'	'4'	0.153846
'4'	'DAP'	0.307692
'DAP'	'4'	0.139535
'rice'	'DAP'	0.328671
'MOP', 'rice'	'0'	0.184211
'MOP', '0'	'rice'	0.538462
'rice', '0'	'MOP'	0.318182
'4'	'Coconut', 'Urea'	0.179487
'Coconut'	'4', 'Urea'	0.0460526
'Urea'	'4', 'Coconut'	0.0693069

TABLE 3.8 – Exemple de règles d'association avec confiance

3.5.4 Effectuer des expérimentations en variant les valeurs de MinSupp et MinConf

Min Support	Min Confidence	Association Rules
1	0.05	2*('4', 'rice'), ('rice', '4'), ('4', 'DAP'), ('DAP', '4'), ('rice', 'DAP')
1	0.1	
1	0.5	('4', 'rice'), ('DAP', 'rice'), ('7', 'Coconut'), ('Good NPK', 'Coconut'), ('MOP', 'rice')
1	0.05	('2', 'rice'), ('Good NPK', '7', 'Coconut'), ('3', 'DAP', 'rice'), ('Urea', '7', 'Coconut')
2	0.1	2*('4', 'rice'), ('rice', '4'), ('4', 'DAP'), ('DAP', '4'), ('rice', 'DAP')
2	0.5	
2	0.05	('4', 'rice'), ('DAP', 'rice'), ('7', 'Coconut'), ('Good NPK', 'Coconut'), ('MOP', 'rice')
2	0.1	('2', 'rice'), ('Good NPK', '7', 'Coconut'), ('3', 'DAP', 'rice'), ('Urea', '7', 'Coconut')
3	0.5	2*('4', 'rice'), ('rice', '4'), ('4', 'DAP'), ('DAP', '4'), ('rice', 'DAP')
3	0.05	
3	0.1	('4', 'rice'), ('DAP', 'rice'), ('7', 'Coconut'), ('Good NPK', 'Coconut')
3	0.5	('2', 'rice'), ('Good NPK', '7', 'Coconut'), ('3', 'DAP', 'rice'), ('Urea', '7', 'Coconut')

TABLE 3.9 – Tableau des expérimentations de MinSupp et MinConf

3.6 Extraction des fortes règles d'associations

La formule de confiance pour une règle d'association $A \rightarrow B$ est donnée par :

$$\text{Confiance}(A \rightarrow B) = \frac{\text{Nombre d'occurrences de } A \rightarrow B}{\text{Nombre d'occurrences de } A}$$

La formule de similarité cosinus pour une règle d'association $A \rightarrow B$ est donnée par :

$$\text{Cosinus}(A \rightarrow B) = \frac{|A \cap B|}{\sqrt{|A| \cdot |B|}}$$

La formule du coefficient de lift pour une règle d'association $A \rightarrow B$ est donnée par :

$$\text{Lift}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A) \times \text{Support}(B)}$$

Antecedent	Consequent	Lift	Cosine	Confidence
{'2'}	{'rice'}	2.06294	5.98315	1
{'3', 'DAP'}	{'rice'}	2.06294	7.80107	1
{'2', 'DAP'}	{'rice'}	2.06294	9.32406	1
{'2', 'Urea'}	{'rice'}	2.06294	11.0324	1
{'MOP', '2'}	{'rice'}	2.06294	11.0324	1
{'Good NPK', '7'}	{'Coconut'}	1.94079	11.9638	1
{'Urea', '7'}	{'Coconut'}	1.94079	7.97589	1
{'1'}	{'rice'}	1.86647	5.38325	0.904762
{'1', 'Urea'}	{'rice'}	1.83372	8.22305	0.888889
{'1', 'DAP'}	{'rice'}	1.80507	8.72186	0.875
{'7'}	{'Coconut'}	1.80216	4.5219	0.928571
{'3'}	{'rice'}	1.77839	4.58094	0.862069
{'6', 'Good NPK'}	{'Coconut'}	1.69819	8.45971	0.875
{'MOP', '7'}	{'Coconut'}	1.69819	8.45971	0.875
{'MOP', '6'}	{'Coconut'}	1.69819	8.45971	0.875
{'3', 'Urea'}	{'rice'}	1.68786	7.43803	0.818182
{'DAP', '7'}	{'Coconut'}	1.66353	9.04381	0.857143
{'6'}	{'Coconut'}	1.62485	3.64893	0.837209
{'6', 'DAP'}	{'Coconut'}	1.61732	6.90732	0.833333
{'6', 'Urea'}	{'Coconut'}	1.55263	6.1781	0.8
{'MOP', '3'}	{'rice'}	1.5472	8.72186	0.75

Table 3.10 – Fortes Regle d'association

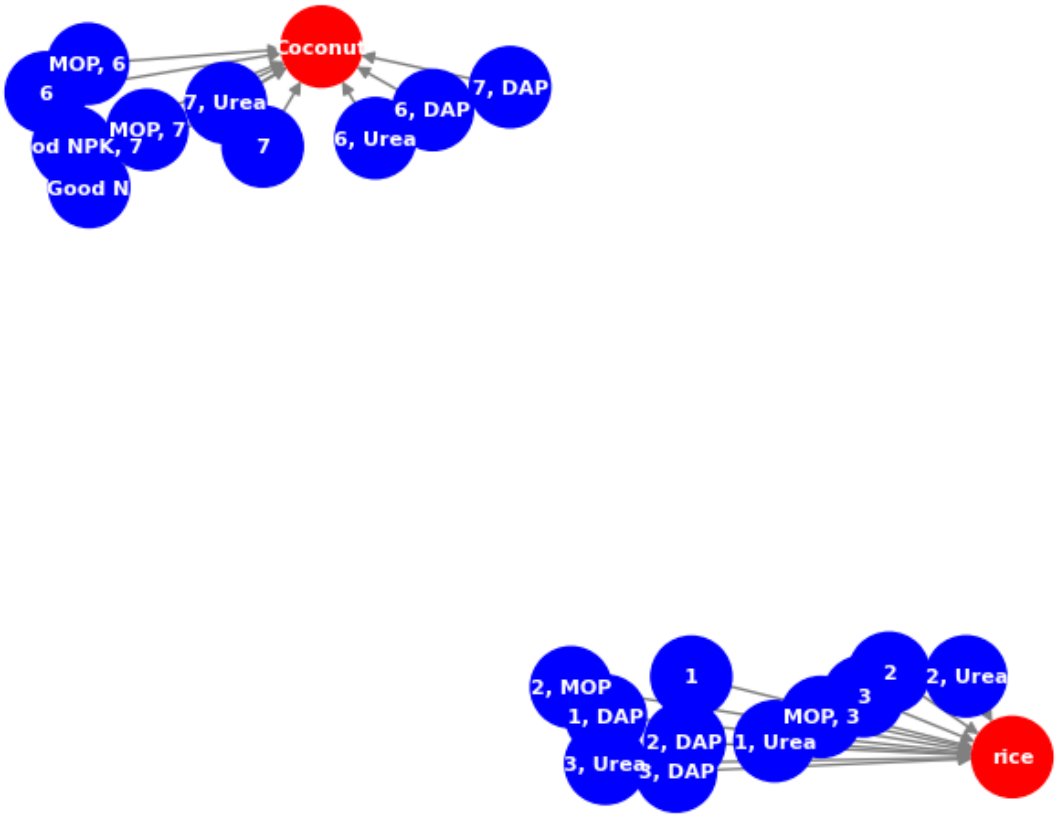


FIGURE 3.2 – Resultat des Fortes Regles D'association

Analyse

D'après la figure 3.4, il est observé que les valeurs de température optimales pour la culture du cocotier se situent entre 6 et 7. En ce qui concerne les types d'engrais recommandés, ils comprennent le MOP, l'Uréa, le DAP et le Good N. Pour la culture du riz, les températures idéales sont de 1 et 2, tandis que les engrais recommandés sont l'Uréa, le MOP et le DAP.

Conclusion

En conclusion, l'analyse effectuée à partir de la figure 3.4 met en évidence des recommandations spécifiques pour la culture du cocotier et du riz en fonction des conditions de température et des types d'engrais.

Ces informations fournissent des directives précieuses pour une gestion efficace des conditions environnementales et des intrants agricoles, contribuant ainsi à maximiser le rendement des cultures respectives.

Températures (Ew)	Températures Réelles
[1 - 2]	[21.87, 21.72, 21.97, 22.73, 21.77, 22.68, 21.33, 22.78, 22.23, 21.32, 22.09, 21.67, 22.3 , 23.06, 22.18, 23.24, 22.71, 21.45, 21.95, 21.84, 21.53, 21.59, 22.7 , 23.22, 23. , 21.41]
[6 - 7]	[28.69, 27.51, 28.11, 28.36, 28.28, 27.13, 27.08, 27.31, 27.59, 28.48, 28.39, 27.54, 27.57, 27.06, 28.13, 27.64, 28.74, 26.61, 26.92, 28.06, 27.19, 28.3 , 26.93, 27.8 , 27.01, 27.1 , 27.75, 28.27, 26.73, 28.44, 27.15, 27.56, 26.76, 27.46, 26.8 , 28.29, 26.87, 28.03, 26.88, 28.57, 27.02]

TABLE 3.11 – Récupération des Températures Réelles

3.7 Conclusion

Ce chapitre nous a offert l'opportunité d'explorer le monde réel et d'extraire des informations cruciales afin de formuler des recommandations dans le domaine de l'agriculture. L'accent a été mis sur la manière de comprendre pleinement notre ensemble de données et de le manipuler de manière à obtenir des résultats exploitables a l'aide des regles d'associations.

Conclusion Générale

En conclusion ce projet démontré l'importance cruciale de l'analyse approfondie et du prétraitement des données dans l'exploitation efficace de la richesse informationnelle contenue dans les vastes ensembles de données numériques. Les compétences acquises dans la manipulation, l'analyse, et l'extraction de connaissances à partir de données diverses représentent des atouts essentiels dans un contexte où la prise de décision éclairée repose de plus en plus sur la maîtrise de l'information numérique. Ce projet constitue ainsi une étape importante dans notre parcours vers une exploitation plus intelligente et efficace du monde des données.

Références

[1] Cours Data Mining - Prof DERIAS.H .

[2] TPs Data Mining - DR Belkadi Widad Hassina ,DR Khelfa Celia.

[3] Parfait Bemarisika. Extraction de règles d'association selon le couple support-MGK : Graphes implicatifs et Applications en didactique des mathématiques. Informatique [cs]. Université d'Antananarivo, 2016. Français. ffNNT : ff.