22/05/2024

Rapport

Analyse de données – TP3

Marianne Corbel

Rapport

Analyse de données – TP3

Exercice 1

Cet exercice vise à fournir un programme capable de nettoyer les données d'un fichier CSV contenant les informations suivantes sur les ressources de blé dans le monde, présentant les données suivantes :

- La région
- L'année
- La production de blé (en millions de tonnes métriques)

Les actions suivantes doivent être réalisées sur les données :

- Chargement du fichier CSV
- Vérification de la cohérence des données
- Remplacement des valeurs manquantes
- Correction des erreurs de saisie
- Normalisation des données

Les données à traiter doivent se présenter sous la forme suivante :

	Région	Année	Production
0	Asie de l'Est et Pacifique	1960	200.67
1	Asie de l'Est et Pacifique	1961	217.70
2	Asie de l'Est et Pacifique	1962	234.44
3	Asie de l'Est et Pacifique	1963	254.36
4	Asie de l'Est et Pacifique	1964	258.1194

Le fichier data/production.csv sera utilisé lors de cet exercice. Il reprend une partie des données du fichier complet et normalisé data/base_data/base_production.csv, mais avec des fautes de saisie et des valeurs erronées volontairement pour cet exercice.

Les données proviennent de l'étude « Production céréalière (tonnes métriques) » réalisée par l'Organisation des Nations Unies pour l'alimentation et la culture et ont été triées et traitées dans le cadre de l'exercice.

Vérification des colonnes

Le programme procède dans un premier temps à la vérification des colonnes du tableau. Les colonnes obligatoires sont **Région**, **Année** et **Production**.

Si l'une de ces colonnes est manquante, le programme ne poursuivra pas l'assainissement des données.

Si des colonnes supplémentaires non-obligatoires sont trouvées, elles seront supprimées du dataset.

```
 --- TRI DES COLONNES... ---
>>> Colonne 4 supprimée <<< (Colonne `Test` non-indispensable)
 --- Terminé ! ---
```

Formatage des données

1. Régions

Notre étude porte sur six régions du monde, prédéfinies à l'avance dans le fichier /data/regions.csv ayant la forme suivante :

	Région
0	Asie de l'Est et du Pacifique
1	Europe et Asie centrale
2	Amérique Latine et Caraïbes
3	Afrique du Nord et Moyen-Orient
4	Asie du Sud
5	Afrique subsaharienne

Si aucune correspondance totale n'est trouvée entre la région d'une ligne des données et les régions de références ci-dessus, le programme va tenter de déduire la région correcte en calculant la correspondance entre la région erronée et celles-ci.

Si l'un des ratios est suffisamment élevé (0.85 par défaut), la valeur erronée sera **remplacée par la région de référence** à la correspondance la plus élevée.

Région erronée	**Région correcte**	**R. de correspondance**
asoe du sud	Asie du Sud	0.91

Dans le cas où aucune correspondance significative est trouvée ou que la région est vide, la ligne entière est **retirée des données**.

```
 --- NETTOYAGE DES DONNEES... ---
>>> Ligne 6 supprimée <<< (Région `t Caraïbes` invalide)
>>> Ligne 7 modifiée <<< (Région `Amériqudsfe Latine et Caraïbes` remplacée par `Amérique Latine et Caraïbes`)
```

2. Colonnes numériques

Les actions suivantes sont réalisées sur les valeurs à traiter :

- Remplacement de toutes les virgules par des points ;
- Suppression de tous les caractères n'étant pas un chiffre, un tiret ou un point.

>>> Année

- Type : Entier
- Null_or_empty : False
- Min – Max : [1960 : 2021]
- Limit_to_extremes : False

Si la conversion est impossible, que la valeur est nulle ou que la valeur convertie dépasse de l'intervalle autorisé, la ligne est supprimée.

Si la valeur convertie est différente de la valeur originale mais qu'elle est conforme, elle est modifiée dans les données.

```
 --- NETTOYAGE DES DONNEES... ---
>>> Ligne 6 modifiée <<< (Année `2011` remplacée par `2011`)
>>> Ligne 7 supprimée <<< (Année `aa` invalide)
```

>>> Production

- Type : Flottant
- Null_or_empty : True
- Min – Max : [0 :]
- Limit_to_extremes : True

Si la valeur est nulle, elle est temporairement conservée ; si elle est négative, elle est réduite au minimum autorisé « 0 ».

Si la valeur convertie est différente de la valeur originale mais qu'elle est conforme, elle est modifiée dans les données. Autrement, la ligne est supprimée.

```
 --- NETTOYAGE DES DONNEES... ---
>>> Ligne 9 modifiée <<< (Production `-73.1306` remplacée par `0`)
>>> Ligne 10 modifiée <<< (Production `aaa` remplacée par `nan`)
>>> Ligne 12 modifiée <<< (Production `116.542aa3` remplacée par `116.5423`)
```

Traitement des doublons

Si plusieurs lignes présentent les mêmes valeurs dans les colonnes Région et Année, seule la première est conservée et toutes les autres sont supprimées des données.

```
 --- TRAITEMENT DES DOUBLONS... ---
>>> Ligne 10 supprimée <<< (Duplication de la ligne 9 sur les colonnes `Région` et `Année`)
>>> Ligne 17 supprimée <<< (Duplication de la ligne 16 sur les colonnes `Région` et `Année`)
 --- Terminé ! ---
```

Formatage des valeurs vides

A ce stade, seule la colonne production peut présenter des valeurs nulles ou vides.

Le programme tentera d'y remédier en remplaçant la valeur Production manquante par la moyenne de celles des deux années les plus proches (supérieure et inférieure).

Si au moins une des deux valeurs les plus proches est nulle également, le programme ne cherchera pas à trouver la valeur suivante la plus proche et supprimera directement la ligne concernée, ceci afin d'éviter des valeurs trop floues ou erronées.

```
 --- TRAITEMENT DES VALEURS VIDES... ---
>>> Ligne 9 modifiée <<< (Production nulle remplacée par la moyenne des productions les plus proches)
>>> Ligne 11 supprimée <<< (Production nulle n'a pas pu être remplacée [manque de valeurs ou valeurs autour également nulles])
 --- Terminé ! ---
```

Sauvegarde des données assainies

Si les données assainies doivent être enregistrées dans un nouveau fichier externe, le programme les sauvegardera par défaut dans le fichier /data/saved_data/production_ex1.csv

```
--- ECRITURE DANS LE FICHIER... ---
Les données ont été écrites dans le fichier `data/production_cleared.csv`.
--- Terminé ! ---
```

Rapport

Analyse de données – TP3

Exercice 2

Cet exercice vise à mettre en œuvre plusieurs algorithmes de détection d'anomalies sur les données d'un fichier CSV contenant les informations suivantes sur les ressources de blé dans le monde, présentant les données suivantes :

- La région
- L'année
- La production de blé (en millions de tonnes métriques)

Les actions suivantes doivent être réalisées sur les données :

- Chargement du fichier CSV
- Calcul du z-score des données
- Application de l'algorithme DBSCAN
- Application de l'algorithme Isolation Forest

Les données à traiter viennent de la même étude que celles de l'exercice 1 et doivent se présenter sous le même format, à savoir :

	Région	Année	Production
0	Asie de l'Est et Pacifique	1960	200.67
1	Asie de l'Est et Pacifique	1961	217.70
2	Asie de l'Est et Pacifique	1962	234.44
3	Asie de l'Est et Pacifique	1963	254.36
4	Asie de l'Est et Pacifique	1964	258.1194

Le fichier data/production.csv sera également utilisé lors de cet exercice, et les valeurs sauvegardées dans data/saved_data/production_ex2.csv (sauf pour Isolation Forest).

Z-SCORE

Dans la mesure où l'Année et la Région sont des colonnes dont les valeurs ne peuvent réellement être aberrante, surtout après l'assainissement de leurs données, le z-score sera calculé sur la colonne Production.

En calculant le z-score de toutes les valeurs de Production indépendamment de l'Année et de la région, et en affichant toutes les lignes dont le z-score est inférieur ou supérieur à 3, on peut récupérer un tableau comme celui-ci :

	Région	Année	Production	z-score
54	Asie de l'Est et Pacifique	2014	918.0339	3.006875
56	Asie de l'Est et Pacifique	2016	935.1992	3.083971
57	Asie de l'Est et Pacifique	2017	921.7285	3.023468
60	Asie de l'Est et Pacifique	2020	945.2051	3.128911
61	Asie de l'Est et Pacifique	2021	954.1380	3.169033
205	Afrique du Nord et Moyen-Orient	1979	930.4532	3.062655

On pourrait être tenté de supprimer toutes ces valeurs puisqu'elles semblent particulièrement éloignées de la moyenne et qu'un z-score supérieur à 3 indique conventionnellement la présence de valeur aberrante.

Les valeurs pour la Région « Asie de l'Est et Pacifique » semblent cependant correctes et cohérentes les unes avec les autres, et la valeur pour la Région « Afrique du Nord et Moyen-Orient » semble proche des autres.

Pour en avoir le cœur net, on peut récupérer une partie des valeurs les plus extrêmes du dataset pour ces deux régions :

	Région	Année	Production
0	Asie de l'Est et Pacifique	1960	200.6704
1	Asie de l'Est et Pacifique	1961	217.7085
2	Asie de l'Est et Pacifique	1962	234.4499
3	Asie de l'Est et Pacifique	1963	254.3683
4	Asie de l'Est et Pacifique	1964	258.1194

	Région	Année	Production
61	Asie de l'Est et Pacifique	2021	954.1380
60	Asie de l'Est et Pacifique	2020	945.2051
56	Asie de l'Est et Pacifique	2016	935.1992
57	Asie de l'Est et Pacifique	2017	921.7285
54	Asie de l'Est et Pacifique	2014	918.0339

	Région	Année	Production
186	Afrique du Nord et Moyen-Orient	1960	17.2273
191	Afrique du Nord et Moyen-Orient	1965	21.3558
189	Afrique du Nord et Moyen-Orient	1963	22.2303
188	Afrique du Nord et Moyen-Orient	1962	23.2177
187	Afrique du Nord et Moyen-Orient	1961	23.9776

	Région	Année	Production
205	Afrique du Nord et Moyen-Orient	1979	930.4532
240	Afrique du Nord et Moyen-Orient	2014	300.0981
231	Afrique du Nord et Moyen-Orient	2005	74.6479
244	Afrique du Nord et Moyen-Orient	2018	73.1306
229	Afrique du Nord et Moyen-Orient	2003	70.7098

Deux choses sont à noter :

- L'évolution de la production en Asie semble effectivement cohérente et croissante donc les valeurs ne seraient à priori pas aberrantes du tout ;
- Les deux valeurs de production les plus élevées pour l'Afrique du Nord sont clairement bien trop élevées. La première avait été relevée par le z-score, mais la seconde est passée inaperçue.

Cela nous indique qu'utiliser le z-score sur la totalité du dataset n'est pas pertinent : des productions anormales pour des régions arides comme l'Afrique du Nord ne peuvent pas être les mêmes que pour des régions à climat tempéré et hautement dépendantes de la culture du blé comme l'Asie.

Ainsi, nous utiliserons les algorithmes de détection d'anomalies sur les valeurs de chacune des régions plutôt que sur le dataset entier :

```
 --- NETTOYAGE DES VALEURS ABERRANTES... ---
>>> Ligne 101 supprimée <<< (Le z-score `6.28` lié à la Production dans
        la Région `Europe et Asie centrale` rend la valeur 2000.0 aberrante)
>>> Ligne 166 supprimée <<< (Le z-score `4.42` lié à la Production dans
        la Région `Amérique Latine et Caraïbes` rend la valeur 500.75 aberrante)
>>> Ligne 205 supprimée <<< (Le z-score `7.42` lié à la Production dans
        la Région `Afrique du Nord et Moyen-Orient` rend la valeur 930.45 aberrante)
>>> Ligne 309 supprimée <<< (Le z-score `3.38` lié à la Production dans
        la Région `Asie du Sud` rend la valeur 700.03 aberrante)
 --- Terminé ! ---
```

Ainsi, beaucoup plus d'anomalies sont détectées. Cependant, la précédente valeur de 300 liée à un z-score de 2.02 du Moyen-Orient n'a pas été considérée comme aberrante, probablement à cause du 900 tirant la moyenne vers le haut.

On peut donc tenter de relancer le nettoyage plusieurs fois pour « lisser » les données :

```
 --- NETTOYAGE DES VALEURS ABERRANTES... ---
>>> Ligne 240 supprimée <<< (Le z-score `6.82` lié à la Production dans
        la Région `Afrique du Nord et Moyen-Orient` rend la valeur 300.1 aberrante)
 --- Terminé ! ---
```

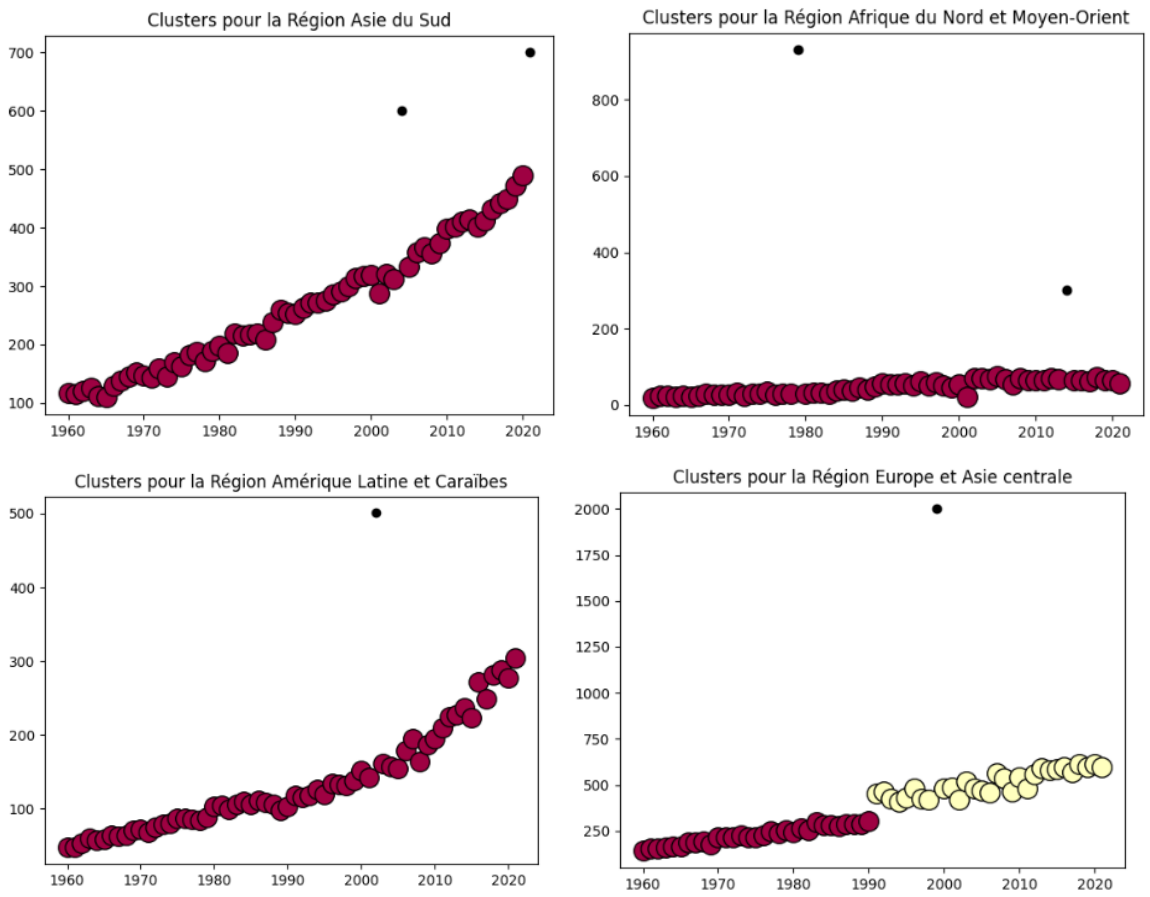
DBSCAN

Pour ne pas perdre de temps, l'algorithme sera également appliqué par région et non sur le dataset entier :

```
 --- NETTOYAGE DES VALEURS ABERRANTES (DBSCAN)... ---
>>> Ligne 101 supprimée <<< (la valeur de Production 2000.0 dans la région Europe et Asie centrale a été jugée aberrante par DBSCAN)
>>> Ligne 166 supprimée <<< (la valeur de Production 500.75 dans la région Amérique Latine et Caraïbes a été jugée aberrante par DBSCAN)
>>> Ligne 205 supprimée <<< (la valeur de Production 930.45 dans la région Afrique du Nord et Moyen-Orient a été jugée aberrante par DBSCAN)
>>> Ligne 240 supprimée <<< (la valeur de Production 300.1 dans la région Afrique du Nord et Moyen-Orient a été jugée aberrante par DBSCAN)
>>> Ligne 292 supprimée <<< (la valeur de Production 600.26 dans la région Asie du Sud a été jugée aberrante par DBSCAN)
>>> Ligne 309 supprimée <<< (la valeur de Production 700.03 dans la région Asie du Sud a été jugée aberrante par DBSCAN)
 --- Terminé ! ---
```

La totalité des valeurs anormales déjà détectées par le z-score l'ont également été pour le DBSCAN, et la ligne 292 a été détectée aberrante.

Voici les diagrammes concernant les régions présentant des aberrations :



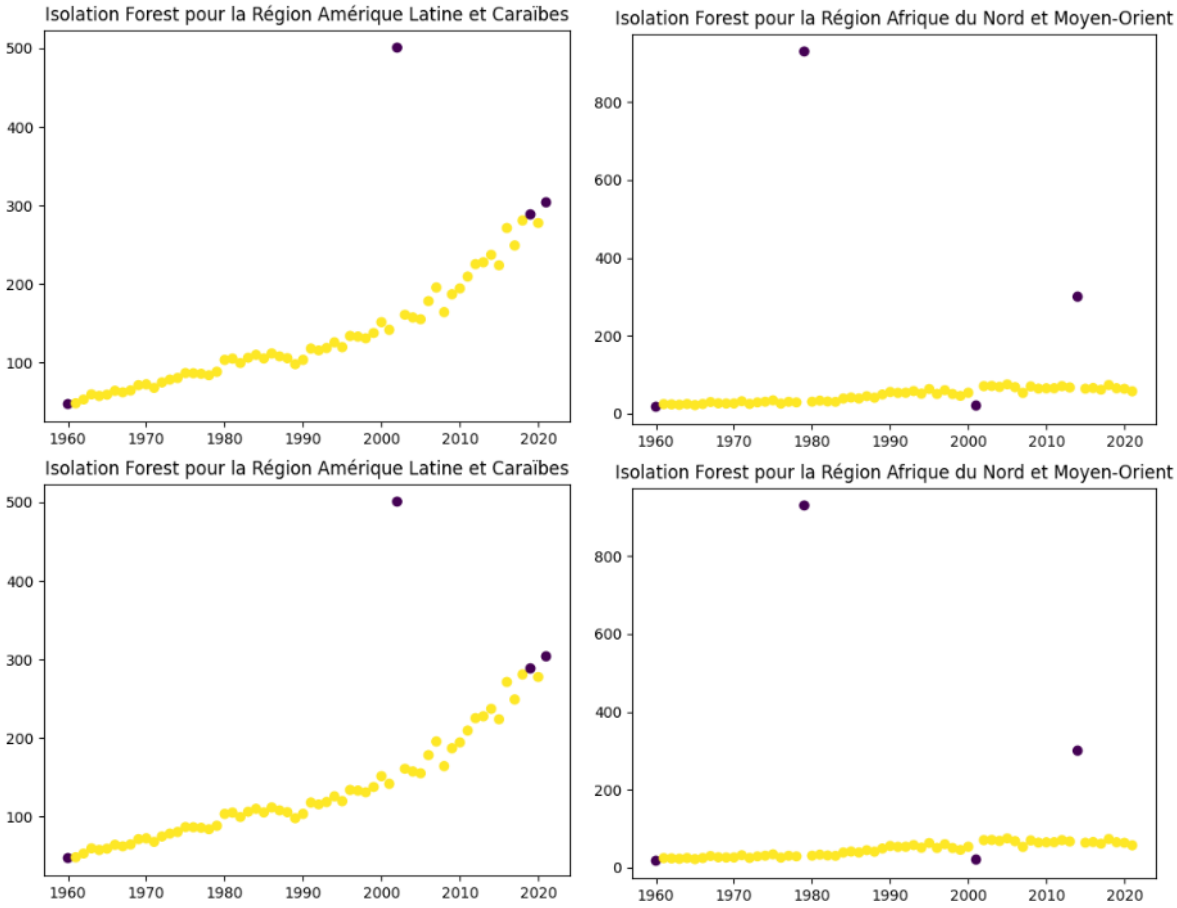
Ces diagrammes ont été obtenus avec un epsilon à 50, obtenus par essais successifs. Il est important de noter que des valeurs un peu plus éparpillées (moins nombreuses ou plus espacées dans le temps), nécessiteront un epsilon beaucoup plus élevé, sinon toutes les valeurs seront considérées aberrantes.

Isolation Forest

Pour ne pas perdre de temps, l'algorithme sera également appliqué par région et non sur le dataset entier :

```
 --- NETTOYAGE DES VALEURS ABERRANTES (DBSCAN)... ---
>>> Ligne 0 supprimée <<< (la valeur de Production 200.67 dans la région Asie de l'Est et Pacifique a été jugée aberrante
>>> Ligne 1 supprimée <<< (la valeur de Production 217.71 dans la région Asie de l'Est et Pacifique a été jugée aberrante
>>> Ligne 60 supprimée <<< (la valeur de Production 945.21 dans la région Asie de l'Est et Pacifique a été jugée aberrante
>>> Ligne 61 supprimée <<< (la valeur de Production 954.14 dans la région Asie de l'Est et Pacifique a été jugée aberrante
>>> Ligne 62 supprimée <<< (la valeur de Production 141.13 dans la région Europe et Asie centrale a été jugée aberrante pa
>>> Ligne 101 supprimée <<< (la valeur de Production 2000.0 dans la région Europe et Asie centrale a été jugée aberrante p
>>> Ligne 122 supprimée <<< (la valeur de Production 613.03 dans la région Europe et Asie centrale a été jugée aberrante p
>>> Ligne 123 supprimée <<< (la valeur de Production 599.65 dans la région Europe et Asie centrale a été jugée aberrante p
>>> Ligne 124 supprimée <<< (la valeur de Production 47.4 dans la région Amérique Latine et Caraïbes a été jugée aberrante
>>> Ligne 166 supprimée <<< (la valeur de Production 500.75 dans la région Amérique Latine et Caraïbes a été jugée aberrant
```

Etrangement, une très grosse quantité de données aberrantes est détectée ici. Voici quelques diagrammes correspondants :



Beaucoup de valeurs aux extrêmités sont jugées aberrantes. Les tentatives d'ajuster la valeur de *contamination* se sont révelées infructueuses, dans la mesure où elles persistent à détecter des anomalies aux extrémités des courbes et ne considèrent parfois pas les réelles anomalies comme telles.

Cet algorithme devra donc être laissé de côté pour l'application de cet exercice.