|  |  |  |  |
| --- | --- | --- | --- |
|  | 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)](https://donnees.banquemondiale.org/indicateur/AG.PRD.CREL.MT) » 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.

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

## 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 :

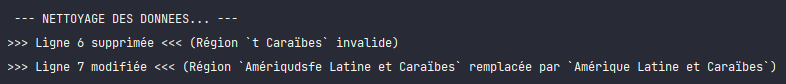
|  |  |
| --- | --- |
|  | Région |
| 0 | Asie de l’Est et du Pacifique |
| 1 | Europe et Asie centrale |
| 2 | Amérique Latine et Caraïbes |
| 3 | Afrique du Nord et Moyen-Orient |
| 4 | Asie du Sud |
| 5 | Afrique subsaharienne |

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**.



### 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.

Une image contenant texte, Police, capture d’écran

Description générée automatiquement

#### >>> 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.

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

## 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.

Une image contenant texte, Police, capture d’écran

Description générée automatiquement

## 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.

Une image contenant capture d’écran, texte, Police

Description générée automatiquement

## 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

Une image contenant capture d’écran, texte, Police

Description générée automatiquement

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 :

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

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 :

Une image contenant texte, capture d’écran, Police

Description générée automatiquement Une image contenant texte, capture d’écran, Police

Description générée automatiquement

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 :

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

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 :

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

## DBSCAN

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

Une image contenant texte, capture d’écran, Police, information

Description générée automatiquement

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 :

Une image contenant capture d’écran, texte, Tracé

Description générée automatiquement Une image contenant texte, capture d’écran

Description générée automatiquement Une image contenant texte, capture d’écran, Tracé

Description générée automatiquement Une image contenant texte, capture d’écran, Tracé

Description générée automatiquement

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 :

Une image contenant texte, capture d’écran, Police, information

Description générée automatiquement

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

Une image contenant texte, capture d’écran, ligne, Tracé

Description générée automatiquement Une image contenant texte, capture d’écran, ligne, Tracé

Description générée automatiquement Une image contenant texte, capture d’écran, ligne, Tracé

Description générée automatiquement Une image contenant texte, capture d’écran, ligne, Tracé

Description générée automatiquement

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.