BUT SCIENCE DE DONNEES

MOISE OKITALOHATA

JOHANNE NZAOU

Alim'confiance

SAE 2-06 : Analyse de données, Reporting et Data-visualisation

Sujet :

Le but de l'étude est d'exploiter le dernier jeu de données brutes alimconfiance.csv pour analyser les données des contrôles sanitaires effectués tout au long de la chaîne alimentaire. Elle doit également étudier l'efficacité de ce dispositif sur différents secteurs et territoires en répondant à la question : La mise en place de ce dispositif a-t-elle bien contribué à améliorer le niveau sanitaire des établissements ?

Voici les objectifs de cette SAE :

Objectif 1 : Mener un travail préparatoire des données :

- Comprendre les données dans leur contexte et identifier les variables pertinentes.
- Vérifier les anomalies et mettre en place les traitement associés.
- Nettoyer et transformer les données initiales pour proposer des données de meilleure qualité : propres et structurées prêtes à être analysées.

Objectif 2 : Produire des fichiers de statistiques avec les indicateurs pertinents indicateurs, selon différents découpages du territoire.

Objectif 3 : Mettre en place une datavisualisation permettant de naviguer dans les données et les statistiques pour :

- Cartographier les établissements contrôlés et les résultats de leur contrôle, et ce pour toute la chaîne alimentaire (abattoirs, commerce de détails, restaurants, restauration collectives...) qui est contrôlé, fréquence, résultats).
- Montrer l'évolution du niveau sanitaire au fil du temps et ce dans tous les établissements de la chaîne alimentaire.

Objectif 4 : Faire une analyse unidimensionnelle et bidimensionnelle des données : calcul de paramètres, d'indicateurs et production des tableaux et graphiques statistiques

- Analyser l'évolution des contrôles et leurs résultats sur une période donnée.
- Valider les résultats obtenus en les confrontant à des résultats publiés officiellement. Il s'agit de comparer les indicateurs et les tendances avec ceux publiés par le gouvernement par exemple et de chercher les explications en cas de divergence.

Objectif 5 : Pousser cette analyse plus loin, intégrer d'autres données externes :

- Trouver un autre jeu de données en lien avec la sécurité alimentaire pour enrichir les données d'Alim-confiance. Exemples : les caractéristiques des établissements (type, notation, prix...), de la population, de la santé, des territoires ...
- Formuler une ou deux questions pertinentes et procéder à l'analyse en intégrant les données externes permettant d'apporter des éléments de réponses.
- Formuler les conclusions statistiques.tatistiques.entaire Formuler les conclusions statistiques.

Sommaire

Première Partie : Chargement des données.

Cette fonction permet d'importer des données à partir d'un fichier CSV. Elle ouvre le fichier en utilisant l'encodage UTF-8, lit les données en utilisant csv.DictReader avec ; comme délimiteur, et retourne une liste de dictionnaires, chaque dictionnaire représentant une ligne du fichier CSV. Utilisation: Pour utiliser cette fonction, passez le nom du fichier CSV comme argument. Par exemple, pour charger un fichier nommé "Region_clean.csv", utilisez:

In [34]:
```
import Johanne_Moise
donnees = Johanne_Moise.charger("Region_clean.csv")
```

In [35]:
```
fichier_csv = "Region_clean.csv"
donnees = Johanne_Moise.charger(fichier_csv)
```

In [35]:
```
Johanne_Moise.afficher(donnees)
```

Pour visualiser les données de notre fichier Excel, nous avons coder une fonction CHARGER qui nous a permis de charger notre fichier et une fonction AFFICHER qui nous a permis de bien visualiser nos données.

Deuxième partie : Visualisation des données.

Premièrement, nous avons commencé à développer une fonction nommée *Charger_selection* qui permet de charger uniquement une sélection spécifique de données à partir d'un fichier Excel. Cette fonction requiert que l'utilisateur spécifie un intervalle, avec une ligne de début et une ligne de fin, pour déterminer précisément quelle plage de données charger. Ci-dessous, nous avons pris un exemple de la ligne 100 à la ligne 105 :

Note: Si les indices spécifiés sont en dehors de la plage des données disponibles, une liste vide sera retournée ou la fonction vous demandera d'entrer une plage de donnée disponible.

In [26]:
```
donnees_selection = Johanne_Moise.charger_selection(fichier_csv, 100, 105)
Johanne_Moise.afficher(donnees_selection)
```

APP_Libelle_etablissement	SIRET	Adresse_2_UA	Code_postal	Libelle_commune	Numero_inspection	Date_inspection
APP_Libelle_activite_etablissement	Synthese_eval_sanit	geores	filtre	Synthese_eval		
THUY LONG	8,45E+13	111 RUE DE VAUGIRARD	75006	Paris 6e Arrondissement	23-051715-1	2023-07-04T02:00:00+02:00
Restaurant	Très satisfaisant	48.845215,2.320783	Restaurant	3		
KERONE	9,05E+13	21 RUE DU VIEUX COLOMBIER	75006	Paris 6e Arrondissement	23-083017-1	2023-10-11T02:00:00+02:00
Restaurant	Très satisfaisant	48.851775,2.330275	Restaurant	3		
CHEZ ISAAC	8,85E+13	6 RUE DES ABBESSES	75018	Paris 18e Arrondissement	23-033673-1	2023-04-25T02:00:00+02:00
Boulangerie-Pâtisserie	Très satisfaisant	48.884024,2.339107	Boulangerie-Pâtisserie	3		
VOYAGE	9,00E+13	21 RUE DE LA MONNAIE	75001	Paris 1er Arrondissement	23-042302-1	2023-06-06T02:00:00+02:00
Restaurant	Très satisfaisant	48.859612,2.343028	Restaurant	3		
SUR LE POUCE	4,84E+13	14 PASSAGE GEFFROY DIDELOT	75017	Paris 17e Arrondissement	23-096431-1	2023-11-22T01:00:00+01:00
Restaurant	Très satisfaisant	48.882458,2.317203	Restaurant	3		

Ensuite, nous avons également créé une fonction appelée *Charger_variables* qui charge uniquement certaines variables spécifiques d'un fichier Excel. Cette fonction est particulièrement utile lorsque nous souhaitons travailler uniquement avec certaines variables, évitant ainsi de charger et d'afficher l'ensemble des données présentes dans le fichier. Pour utiliser cette fonction, il faut passer le nom du fichier CSV et une liste des noms de colonnes comme arguments. Par exemple, pour charger les colonnes 'Code_postal', 'Synthese_eval_sanit' et 'Libelle_commune' :

In [36]:
```
variables_a_charger = ['Code_postal', 'APP_Libelle_activite_etablissement', 'Synthese_eval_sanit']
donnees_variables = Johanne_Moise.charger_variables(fichier_csv, variables_a_charger)
Johanne_Moise.afficher(donnees_variables, 100,105)
```

Code_postal	APP_Libelle_activite_etablissement	Synthese_eval_sanit
75006	Restaurant	Très satisfaisant
75006	Restaurant	Très satisfaisant
75018	Boulangerie-Pâtisserie	Très satisfaisant
75001	Restaurant	Très satisfaisant
75017	Restaurant	Très satisfaisant

Enfin, nous avons développé une fonction nommée *Charger_par_filtre* qui permet de charger des données selon un critère spécifique. Par exemple, dans le cas illustré, cette fonction est utilisée pour charger uniquement les données correspondant au code postal "75001".

In [28]:
```
donnees_filtrees = Johanne_Moise.charger_par_filtre(fichier_csv, Code_postal='75001')
Johanne_Moise.afficher(donnees_filtrees, 100,105)
```

APP_Libelle_etablissement	SIRET	Adresse_2_UA	Code_postal	Libelle_commune	Numero_inspection	Date_inspection
APP_Libelle_activite_etablissement	Synthese_eval_sanit	geores	filtre	Synthese_eval		
HOTEL MEURICE	3,45E+13	228, RUE DE RIVOLI	75001	Paris 1er Arrondissement	23-035994-1	2023-05-12T02:00:00+02:00
Restaurant	Très satisfaisant	48.865194,2.328101	Restaurant	3		
AU PAVILLON	8,51E+13	65 BD DE SEBASTOPOL	75001	Paris 1er Arrondissement	23-110504-1	2024-02-19T01:00:00+01:00
Restaurant	Satisfaisant	48.863292,2.350653	Restaurant	2		
ZAPI	9,23E+13	7 RUE DU VINGT NEUF JUILLET	75001	Paris 1er Arrondissement	23-105547-1	2023-12-14T01:00:00+01:00
Restaurant	Satisfaisant	48.865137,2.330504	Restaurant	2		
CHEVAL BLANC PARIS	7,98E+13	8 QUAI DU LOUVRE	75001	Paris 1er Arrondissement	23-107740-1	2023-12-20T01:00:00+01:00
Restaurant	Très satisfaisant	48.858791,2.342062	Restaurant	3		
MARYLIN'S CAFE	5,00E+13	41 BOULEVARD SEBASTOPOL	75001	Paris 1er Arrondissement	22-107969-1	2023-03-27T02:00:00+02:00
Restaurant	Satisfaisant	48.861388,2.349588	Restaurant	2		

Troisième partie : Traitement des anomalies.

Premièrement, nous avons créé une fonction appelée *detecter_erreurs_format* qui est conçue pour identifier les erreurs de format, notamment les caractères spéciaux ou les entrées mal rédigées qui n'ont pas de sens logique. Ensuite, nous avons mis en place une fonction nommée *detecter_valeurs_aberrantes* qui détecte les anomalies dans les données. Par exemple, pour les codes postaux qui devraient se situer entre 75000 et 75020, toute valeur en dehors de cette plage est considérée comme une erreur. Et enfin la fonction *supprimer_ligne* viendra supprimer les lignes ou il y'a ce genre d'erreurs.

In [29]:
```
erreurs_format = Johanne_Moise.detecter_erreurs_format(donnees)
valeurs_aberrantes = Johanne_Moise.detecter_valeurs_aberrantes(donnees)
donnees_fil=(donnees_filtrees + Johanne_Moise.supprimer_lignes(donnees, erreurs_format + valeurs_aberrantes))
Johanne_Moise.afficher(donnees_fil, 100, 105)
```

APP_Libelle_etablissement	SIRET	Adresse_2_UA	Code_postal	Libelle_commune	Numero_inspection	Date_inspection
APP_Libelle_activite_etablissement	Synthese_eval_sanit	geores	filtre	Synthese_eval		
HOTEL MEURICE	3,45E+13	228, RUE DE RIVOLI	75001	Paris 1er Arrondissement	23-035994-1	2023-05-12T02:00:00+02:00
Restaurant	Très satisfaisant	48.865194,2.328101	Restaurant	3		
AU PAVILLON	8,51E+13	65 BD DE SEBASTOPOL	75001	Paris 1er Arrondissement	23-110504-1	2024-02-19T01:00:00+01:00
Restaurant	Satisfaisant	48.863292,2.350653	Restaurant	2		
ZAPI	9,23E+13	7 RUE DU VINGT NEUF JUILLET	75001	Paris 1er Arrondissement	23-105547-1	2023-12-14T01:00:00+01:00
Restaurant	Satisfaisant	48.865137,2.330504	Restaurant	2		
CHEVAL BLANC PARIS	7,98E+13	8 QUAI DU LOUVRE	75001	Paris 1er Arrondissement	23-107740-1	2023-12-20T01:00:00+01:00
Restaurant	Très satisfaisant	48.858791,2.342062	Restaurant	3		
MARYLIN'S CAFE	5,00E+13	41 BOULEVARD SEBASTOPOL	75001	Paris 1er Arrondissement	22-107969-1	2023-03-27T02:00:00+02:00
Restaurant	Satisfaisant	48.861388,2.349588	Restaurant	2		

Quatrième partie : Calcul d'indicateurs statistiques.

Cette fonction permet d'obtenir une description des données. Elle calcule et affiche le nombre de lignes, le nombre de colonnes et le nombre total de mots dans le fichier CSV. Utilisation: Pour utiliser cette fonction, passez le nom du fichier CSV comme argument. Par exemple, pour obtenir une description des données dans le fichier "Region_csv", utilisez:

In [31]:
```
Johanne_Moise.describe_donnees(fichier_csv)
```

```
Nombre de lignes dans le fichier : 3454
Nombre d'indicateurs (colonnes) : 12
Nombre total de mots dans toutes les valeurs : 64712
```

Cinquième partie : Exportation.

Cette fonction permet d'exporter les données modifiées dans un nouveau fichier CSV. Elle écrit les données dans un fichier en utilisant l'encodage ISO-8859-1 et le délimiteur ;. Utilisation: Pour utiliser cette fonction, passez les données à exporter et le nom du fichier de destination comme arguments. Par exemple, pour exporter les données nettoyées dans un fichier nommé "Region_Parisienne_clean.csv", utilisez:

In [ ]:
```
Johanne_Moise.export_data(donnees_fil, "Region_Parisienne_clean.csv")
```

Conclusion

Ce document décrit les fonctionnalités et l'architecture d'un programme conçu pour traiter les données d'un fichier CSV de manière efficace. Le programme offre des solutions complètes pour charger, filtrer, traiter et exporter des données tout en garantissant la qualité et l'intégrité des informations.

Les points clés du programme incluent :

1. **Chargement des Données** : Le programme permet d'importer des données complètes ou partielles, en sélectionnant des colonnes spécifiques ou en appliquant des filtres pour ne charger que certaines lignes.
2. **Traitement des Anomalies** : Il inclut des fonctions pour détecter les erreurs de format et les valeurs aberrantes, ainsi que des méthodes pour les supprimer, assurant ainsi que les données sont nettoyées et prêtes à l'analyse.
3. **Affichage des Données** : Les données peuvent être affichées de manière structurée, avec des contrôles intégrés pour éviter les erreurs d'indices lors de la sélection des plages de lignes à afficher.
4. **Description des Données** : Une fonction dédiée permet de fournir une vue d'ensemble des données, incluant le nombre de lignes et de colonnes ainsi que le volume total de texte.
5. **Exportation des Données** : Les données traitées peuvent être facilement exportées dans un nouveau fichier CSV, assurant une utilisation future sans perte de qualité.

Synthèse

Le programme présente une série de fonctions modulaires et interconnectées, chacune ayant un rôle précis dans le processus de manipulation des données. Du chargement initial à l'exportation finale, chaque étape est conçue pour maximiser la précision et l'efficacité du traitement des données.

En conclusion, ce programme offre une solution robuste pour gérer des fichiers CSV, en mettant l'accent sur la simplicité d'utilisation et la robustesse du traitement des données. Il permet aux utilisateurs de transformer des données brutes en informations précieuses, prêtes pour l'analyse ou la prise de décision.

In [ ]:
```
```