**Question mentor Projet 3**

**Globale**

Question sur le cleaning des données :

Valeurs négatives puis suppression des doublons (Exemple ligne 24).

Est-ce suffisant ?

**OK pour Nicolas**

**Vérifier sur Internet comment afficher les doublons avec la fonction Drop\_duplicates**

**Question par question**

Question 4 : Ratio max de 21900 kcal/kg quand les données internet donnent un max de 900 ?

**Faire un script pour contrôler rapidement les données à partir de ces composantes**

**Ou**

**Comparer avec les autres valeurs des autres pays pour voir si elle est cohérente**

**Remarques :**

* Améliorer la présentation avec tabulation Markdown et sommaire
* Lorsque je supprime une valeur incohérente ou un doublon, si je supprime la ligne en renseignant son index, l’information est statique et ce numéro peut changer si de nouvelles données sont intégrées dans la table. Il faut donc trouver une façon dynamique de renseigner la ligne à supprimer, de sorte à ce que les mêmes lignes soient supprimer si d’autres sont rajoutées entre temps.
* Présentation et avancement Ok
* Pour la question 2, même si ce n’est pas demandée, une méthode efficace serait de créer les colonnes pour chaque donnée et faire des soustractions et additions de sorte à obtenir 0 sur une dernière colonne pour ainsi constater l’équilibre entre chaque égalité.

**Précision si envoi**

Ligne 120 provisoire : affichage des 30 valeurs de ratio les plus élevées pour voir si la valeur max est bien supérieure à la seconde ou s’il y a un ensemble de valeurs trop élevées. Les valeurs de ratio sur Internet ne dépassant pas 10 000 kcal/kg (les valeurs sont données à chaque fois pour 100g donc on voit des valeurs inférieures à 1000 kcal/100g)

**Choses qui ne vont pas dans projet 3:**

- Question 12 : pourquoi la jointure pd.merge(AutresUtilisationsMondiale15Produits, DispIntMondiale15Produits, on='Produit') nécessite l’argument on=‘Produit’ alors que cette colonne est commune aux deux tableaux?

- Revoir question 7

Session du 7 mai :

Recherche Google : SQLiteStudio truncates first letter csv

Idéalement la recherche doit tenir en 5 mots max.

Bosser Atom

Origine du problème des premières lettres en moins dans SQLite : Il y a des caractères invisibles à chaque fin de ligne du fichier csv, CR (\n) et CF ( \r), qui permettent d’enclencher le retour à la ligne. **Il manquait CR**. Il faut le rajouter grâce à Atom.

Requête mentor :

SELECT pays,

substr(pays, 2, LENGTH(pays)) AS pays\_custom,

code\_pays,

annee,

population

FROM population\_crlf\_underscores;

Notes pour soutenance:

Pourquoi avoir supprimer « Chine » et garder « Chine, continentale » ?

* Parce que les données sur la Chine continentale est une donnée officielle dans le tableau population
* Parce que la Chine continentale apparaît dans tous les autres tableaux alors que la Chine n’apparaît pas dans le tableau cereales.