Réalisez une étude de santé publique

100 heuresMis à jour le mardi 2 mars 2021

Prérequis

Pour ce projet, il vous faudra savoir manipuler au choix les bases du langage **R** ou **Python,** et savoir utiliser dans ces langages des objets de type *Dataframes* (objets natifs du langage R, et disponibles dans la librairie **Pandas** de Python. Ce projet vous fera travailler - via les dataframes - les concepts de l'algèbre relationnelle, qu'il est bon de maîtriser. Il vous faudra également maîtriser les bases du langage **SQL**, notamment les **requêtes** SQL de type **SELECT**. Si vous êtes déjà habitués à ce type de requêtes, alors vous connaissez sûrement sans le savoir les concepts d'algèbre relationnelle nécessaires à ce projet. Enfin, il vous faudra savoir installer et requêter (en SQL) un **système de gestions de bases de données** au choix (des indications sont données dans l'énoncé).

Mise en situation

Vous êtes intégré à une nouvelle équipe de chercheurs de la Food and Agriculture Organization of the United Nations (FAO), l'un des organes qui compose l'ONU et dont l'objectif est d' « aider à construire un monde libéré de la faim ».

Votre équipe est chargée de réaliser une étude de grande ampleur sur le thème de la sous-nutrition dans le monde.

Le problème de la faim est complexe et peut avoir de multiples causes, différentes selon les pays. L'étape préliminaire de cette étude sera donc d'établir un "état de l'art" des recherches déjà publiées, mais également de mener une étude statistique destinée à orienter les recherches vers des pays particuliers, et de mettre en lumière différentes causes de la faim. Ainsi, une poignée de data analysts (dont vous !) a été sélectionnée pour mener cette étape préliminaire. Lors de la première réunion, vous avez été désigné pour mettre une place la base de données que votre équipe pourra requéter (en SQL) à souhait pour réaliser cette étude statistique.

Les données

Les données sont disponibles sur ce lien et sont constituées de 5 fichiers:

- fr_animaux.csv : multiples indicateurs de production des produits animaux en 2013
- fr_population.csv: population mondiale par pays en 2013
- fr_vegetaux.csv: multiples indicateurs de production des produits végétaux en 2013
- fr_céréales.csv: quantité de céréales produites au niveau mondial en 2013
- fr_sousalimentation.csv: nombre de personnes sous alimentées dans le monde de 2013 à 2017.

Votre mission

Vous êtes prêt ? Alors c'est parti !

Si vous réalisez ce projet en Python, il est vivement conseillé d'utiliser un notebook Jupyter, et de savoir utiliser les librairies Python pour la data science. Ce cours vous y aidera. (Les chapitres sur Monty Hall et le Théorème central limite ne sont pas du niveau Data Analyst, mais les autres sont fait pour vous !)

Identifiez les grandes tendances

Une fois les fichiers CSV téléchargés, il vous faudra les charger dans R ou Python, afin d'identifier les grandes tendance et vous familiariser avec le jeu de données.

Pour chacune des tables téléchargées, identifiez les possibles clés primaires, et testez-les (pour plus d'informations, https://openclassrooms.com/courses/initiez-vous-a-lalgebre-relationnelle-avec-le-langage-sql/ne-perdez-pas-de-vue-vos-cles). Cette étape vous permettra de comprendre la "structure" de vos tables, et vous sera utile lorsque vous aurez à effectuer des jointures.

On commence !

Créez un dataframe contenant les informations de population de chaque pays. Calculez le nombre total d'humains sur la planète. Critiquez votre résultat. En cas d'anomalie, analysez et effectuer les corrections nécessaires.

Question 1 : donnez le résultat de votre calcul pour l'année 2013.

Parmi les documents sur les Bilans alimentaires que vous avez téléchargés, il y a des informations redondantes. En effet, pour un pays donné, certaines de ces informations peuvent se calculer à partir d'autres :

- Production
- Importations
- Exportations
- Variation de stock
- Disponibilité intérieure
- Semences
- Pertes
- Nourriture, aussi appelée Disponibilité alimentaire
- Aliments pour animaux
- Traitement
- Autres utilisations

Question 2 : Identifiez ces redondances, en donnant votre réponse sous forme de formule mathématique (pas besoin de coder ici). C'est une équation à 3 termes de type \(a_1 + a2 + [...] = b_1 + b_2 + [...] = c_1 + c_2 + [...]\)) faisant intervenir chacune des 11

quantités données ci dessus. Illustrez cette équation avec l'exemple du blé en France. Pour avoir un indice, cliquez sur "Définitions et Standards" sur la page de téléchargement des données.

Question 3 : Calculez (pour chaque pays et chaque produit) la disponibilité alimentaire en kcal puis en kg de protéines.

Vous ferez cela à partir de ces informations :

- Population de chaque pays ;
- Disponibilité alimentaire donnée pour chaque produit et pour chaque pays en kcal/personne/jour.
- Disponibilité alimentaire en protéines donnée pour chaque produit et pour chaque pays en g/personne/jour.

Question 4 : A partir de ces dernières informations, et à partir du poids de la disponibilité alimentaire (pour chaque pays et chaque produit), calculez pour chaque produit le ratio "énergie/poids", que vous donnerez en kcal/kg. Vous pouvez vérifier la cohérence de votre calcul en comparant ce ratio aux données disponibles sur internet, par exemple en cherchant la valeur calorique d'un oeuf.

Indication : La disponibilité alimentaire en kcal/personne/jour est calculée par la FAO en multipliant la quantité Nourriture par le ratio énergie/poids (en kcal/kg), puis en le divisant par la population du pays puis par 365. Ici, on vous demande juste de retrouver le ratio énergie/poids que la FAO a utilisé dans son calcul.

En suivant la même méthodologie, calculez également le pourcentage de protéines de chaque produit (pour chaque pays). Ce pourcentage est obtenu en calculant le ratio "poids de protéines/poids total" (attention aux unités utilisées). Vous pouvez vérifier la cohérence de votre calcul en comparant ce ratio aux données disponibles sur internet, par exemple en cherchant la teneur en protéines de l'avoine.

Question 5 : Citez 5 aliments parmi les 20 aliments les plus caloriques, en utilisant le ratio énergie/poids.
Étonnamment, il arrive que ce ratio soit différent en fonction du pays. Il faudra donc réaliser pour chaque aliment une moyenne sur les différents pays. Vous créerez donc une nouvelle table grâce à une agrégation. **Attention à bien retirer les valeurs égales à 0 afin de ne pas fausser le calcul de la moyenne**.
Citez 5 aliments parmi les 20 aliments les plus riches en protéines.

Pour approfondir la réflexion, il est important de se documenter sur la qualité des protéines, notamment sur l'indice PDCAAS. Voici les articles Wikipedia correspondant : français, anglais. Cet aspect n'est pas demandé dans la soutenance, c'est juste pour la culture générale.

Question 6 : Calculez, pour les produits végétaux **uniquement**, la disponibilité intérieure mondiale exprimée en kcal.

Question 7 : Combien d'humains pourraient être nourris si toute la disponibilité intérieure mondiale de produits végétaux était utilisée pour de la nourriture ? Donnez les résultats en termes de calories, puis de protéines, et exprimez ensuite ces 2 résultats en pourcentage de la population mondiale.

Question 8 : Combien d'humains pourraient être nourris si toute la disponibilité alimentaire en produits végétaux la nourriture végétale destinée aux animaux et les pertes de produits végétaux étaient utilisés pour de la nourriture ? Donnez les résultats en termes de calories, puis de protéines, et exprimez ensuite ces 2 résultats en pourcentage de la population mondiale.

Question 9 : Combien d'humains pourraient être nourris avec la disponibilité alimentaire mondiale ? Donnez les résultats en termes de calories, puis de protéines, et exprimez ensuite ces 2 résultats en pourcentage de la population mondiale.

Question 10 : A partir des données téléchargées qui concernent la sous-nutrition, répondez à cette question : Quelle proportion de la population mondiale est considérée comme étant en sous-nutrition ?

Établissez la liste des produits (ainsi que leur code) considéré comme des céréales selon la FAO.

Repérez dans vos données les informations concernant les céréales (par exemple en créant une colonne de type booléen nommée "is_cereal").

Question 11 : En ne prenant en compte que les céréales destinées à l'alimentation (humaine et animale), quelle proportion (en termes de poids) est destinée à l'alimentation animale ?

Sélectionnez parmi les données des bilans alimentaires les informations relatives aux pays dans lesquels la FAO recense des personnes en sous-nutrition.

Repérez les 15 produits les plus exportés par ce groupe de pays.

Parmi les données des bilans alimentaires au niveau mondial, sélectionnez les 200 plus grandes importations de ces produits (1 importation = une quantité d'un produit donné importée par un pays donné)

Groupez ces importations par produit, afin d'avoir une table contenant 1 ligne pour chacun des 15 produits. Ensuite, calculez pour chaque produit les 2 quantités suivantes :

- le ratio entre la quantité destinés aux "Autres utilisations" (Other uses) et la disponibilité intérieure.
- le ratio entre la quantité destinée à la nourriture animale et la quantité destinée à la nourriture (animale + humaine)

Question 12 : Donnez les 3 produits qui on t la plus grande valeur pour chacun des 2 ratios (vous aurez donc 6 produits à citer)

Question 13 : Combien de tonnes de céréales pourraient être libérées si les USA diminuaient leur production de produits animaux de 10% ?

Question 14 : En Thaïlande, quelle proportion de manioc est exportée ? Quelle est la proportion de personnes en sous-nutrition?

Entrez vos données dans une base de données relationnelle

Bon, finis les calculs mathématiques , passons à de la technique, de la vraie !

Transformez vos données en un format propice à l'utilisation souhaitée par les utilisateurs finaux, qui utiliserons votre base via le langage SQL. Une fois les données correctement formatées, vous les intégrerez dans une base de données.

Le choix de la technologie utilisée est libre. Vous en trouverez un petit tour d'horizon dans le chapitre Les bases de données du cours *Comprendre le web*.

Si vous n'avez jamais utilisé de système de gestion de base de données relationnelles (SGBR), nous vous conseillons d'utiliser *SQL-lite*. Pour savoir comment l'utiliser, la vidéo de ce chapitre peut vous être utile.

Si vous êtes plus aventurier , un cours sur la technologie *MySQL* est également disponible sur OpenClassrooms : les chapitres utiles seront les 10 premiers (de *Introduction* à *Suppression et modification de données*).

Pour exporter un dataframe de R ou Python vers une base de données, l'une des solutions est de passer via un fichier CSV. Sous Python, utilisez to_csv :

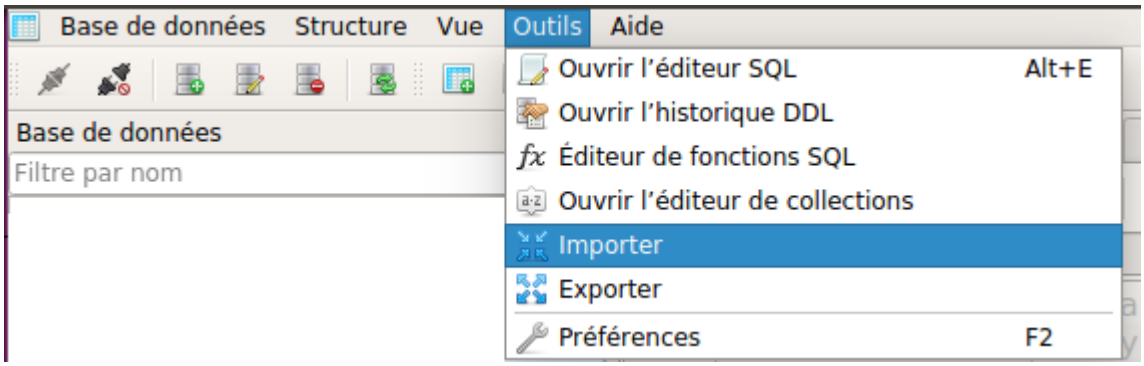
```
df.to_csv("export.csv", index = False)
```

Sous R, utilisez write.csv :

```
write.csv(df, file = "export.csv", row.names=FALSE)
```

Ensuite, en fonction de votre SGBDR, recherchez comment importer un fichier CSV dans une table. Pour ceci, votre moteur de recherche internet est votre meilleur ami (par exemple, recherchez : *"import a csv file into PostgreSQL"*).

Si vous utilisez le SGBDR SQLite, et que vous le manipulez grâce au logiciel SQLiteStudio, alors cliquez sur "Outils" puis sur "Importer" (à faire une fois que votre base de données est créée, comme indiqué sur cette vidéo) :



Une autre solution, plus élégante, consiste à utiliser une librairie Python ou R permettant de se connecter directement à votre SGBDR, comme par exemple les librairies *RMySQL*, *psycopg2*, etc.

Votre base devra contenir ces différentes tables :

> Une table appelée **population**, contenant l*a population de chaque pays pour 2013. Elle devra contenir 4* colonnes : **pays**, **code_pays**, **annee**, **population**.

Question 15 : Proposez une clé primaire pertinente pour cette table.

- Une table appelée **dispo_alim** contenant pour chaque pays et pour chaque produit en 2013, les informations suivantes:
  - la nature du produit (deux valeurs possibles : “animal” ou “végétal”)
  - disponibilité alimentaire en tonnes
  - disponibilité alimentaire en Kcal/personne/jour
  - disponibilité alimentaire de protéines en g/personne/jour
  - disponibilité alimentaire de matières grasses en g/personne/jour
- Elle devra contenir ces colonnes : pays, code_pays, année, produit, code_produit, origin, dispo_alim_tonnes, dispo_alim_kcal_p_j, dispo_prot, dispo_mat_gr .

Question 16 : Proposez une clé primaire pertinente pour cette table.

- Une table appelée **equilibre_prod** contenant pour chaque pays et pour chaque produit en 2013, les quantités suivantes :
  - disponibilité intérieure
  - aliments pour animaux
  - semences
  - pertes
  - transformés
  - nourriture
  - autres utilisations
- Elle devra contenir ces colonnes : pays, code_pays, année, produit, code_produit, dispo_int, alim_ani, semences, pertes, transfo, nourriture, autres_utilisations.

Question 17 : Proposez une clé primaire pertinente pour cette table.

Une table appelée **sous_nutrition**, contenant le nombre de personnes en sous-alimentation pour chaque pays en 2013. Elle devra contenir 4 colonnes : **pays**, **code_pays**, **année**, **nb_personnes**.

Question 18 : Vous vous en doutez... proposez encore une fois une clé primaire pertinente pour cette table !

Question 19 : Écrivez les requêtes SQL permettant de connaître...

- Les 10 pays ayant le plus haut ratio **disponibilité alimentaire/habitant** en termes de protéines (en kg) par habitant, **puis** en termes de kcal par habitant.
- Pour l'année 2013, les 10 pays ayant le plus faible ratio **disponibilité alimentaire/habitant** en termes de protéines (en kg) par habitant.
- La quantité totale (en kg) de produits perdus par pays en 2013.
- Les 10 pays pour lesquels la proportion de personnes sous-alimentées est la plus forte.
- Les 10 produits pour lesquels le ratio **Autres utilisations/Disponibilité intérieure** est le plus élevé.

Question 20 : pour quelques uns des produits identifiés dans cette dernière requête SQL, supposez quelles sont ces "autres utilisations" possibles (recherchez sur internet !)

Enrichissez votre analyse

Lors de votre présentation, vous présenterez les pistes étudiées jusqu'à maintenant grâce aux données de la FAO. En réalité, vous n'êtes pas le/la premier·e à étudier ce phénomène (bien heureusement!). Veillez donc à bien confirmer vos intuitions par des recherches sur internet : votre mentor vous fournira des sources.

Pour une bonne analyse, le data analyst doit comprendre le domaine qu'il étudie. On appelle cela les "connaissances métier". A partir des sources fournies par votre mentor, il vous sera donc également demandé de citer d'autres causes de la faim, et d'enrichir votre analyse de nouveaux chiffres. Si vous êtes motivés, vous pouvez même vérifier certains des chiffres citées dans les sources à partir des données de la FAO 😉

Notamment, cherchez les réponses à ces questions :

- Combien de personnes décèdent des causes de la faim ?
- Quelles sont les prévisions de population mondiale en 2050 ?

Livrables

Voici les livrables attendus, que vous transmettrez dans une archive .ZIP :

- Le **code** en R ou en Python qui vous a permis de calculer chacune des réponses aux questions 1 à 14.
- Le **code** en R, Python et/ou SQL vous ayant permis d'enregistrer les données dans la base de données
- Les **requêtes SQL** vous ayant permis de répondre aux questions 15 à 20.

Pour faciliter votre passage au jury, déposez sur la plateforme, dans un dossier nommé "*P3_nom_prenom*", tous les livrables du projet. Chaque livrable doit être nommé avec le numéro du projet et selon l'ordre dans lequel il apparaît, par exemple "*P3_01_coderéponses*", "*P3_02_codedonnées*", et ainsi de suite.

Soutenance

Pour la soutenance, mettez-vous dans la peau du data analyst qui présente les résultats de son étude préliminaire à son équipe (cf paragraphe "Mise en contexte" de cet énoncé).

Voici le déroulé de la soutenance. Même si les durées sont indicatives, merci de traiter toutes les étapes.

7 min	Mise en contexte auprès de votre équipe. Donnez les chiffres clés (nombre de personnes qui décèdent à cause de la faim, chiffres de la sous-nutrition mondiale, évolution possible au cours des années suivantes). Donnez ensuite différentes causes de la faim.
7 min	Approfondissez l'une de ces causes (ou plusieurs de ces causes) en présentant les résultats de vos calculs. Vous conclurez cette partie en répondant à ces 2 questions : • la faim dans le monde résulte t'elle d'un manque de production, ou de problèmes technologiques ? • quelles sont les prévisions de population en 2050 ? Aura t'on besoin d'augmenter drastiquement la production alimentaire ? Il n'est pas nécessaire de répondre scolairement à chacune des questions 1 à 14 durant la soutenance, mais vous devez présenter les chiffres les plus importants que vous avez trouvé en les incluant dans votre analyse.
3 min	Détaillez des données téléchargées : source, combien de fichiers CSV, à quoi correspondent-ils, à quoi correspond chaque ligne d'un fichier, etc. (il est possible de montrer danc votre présentation des premières lignes des fichiers en les détaillant). Détaillez de la même manière les principales tables (dataframes ou tables sql) utilisées lors de l'analyse.
3 min	Détaillez quelques opération d'algèbre relationnelle utilisées sur les dataframes, dont au moins : • une agrégation • une jointure (en justifiant le type de jointure : interne, externe gauche, etc.) • une restriction Pour chacune des opérations d'algèbre relationnelle détaillée, indiquez les clés primaires des tables avant et après opération. Il conseillé de donner des captures d'écran des tables pour expliquer les opérations utilisées

5 min	Montrez le résultat de chacune des requêtes de la question 19 (sous forme de capture d'écran). Détaillez le code SQL de 2 d'entre elles. Répondez à la question 20. Il n'est pas nécessaire de répondre aux questions 15 à 18 lors de la soutenance.
5 min	Questions-réponses de l'examinateur

Ressources complémentaires

En R

https://www.rstudio.com/resources/cheatsheets/

De nombreux pense-bêtes, très utile pour avoir l'essentiel des fonctions importantes des grands packages d'analyse de données en R.

https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html

Introduction complète de dplyr en anglais.

http://informatique-mia.inra.fr/r4ciam/ODBC.html#req

Exemples d'insertion de données dans une base SQL en R via le package RODBC.

En Python

https://openclassrooms.com/courses/decouvrez-les-librairies-python-pour-la-data-science

Cours OpenClassrooms "Découvrez les librairies Python pour la Data Science"

http://apprendre-python.com/page-database-data-base-donnees-query-sql-mysql-postgre-sqlite

Cours très bien fait sur l'interaction avec une base SQL en Python.

Compétences évaluées

- Utiliser les librairies spécialisées pour la Data Science
- Utiliser une documentation technique
- Récupérer des données à partir d'une source identifiée

- Appliquer l'algèbre relationnelle en R ou Python
- Maîtriser les bases de R ou Python
- Effectuer des requêtes complexes en SQL