Transformation d'un tableau en base de données relationnelle

De nombreuses données sont gérées sous des feuilles de calcul, sous la forme de simples tableaux. C'est bien souvent le cas pour des applications de petite taille, ou lorsque les créateurs/utilisateurs n'ont pas les compétences ou connaissances nécessaires à la conception et exploitation d'une base de données normalisée. C'est au moment où les besoins évoluent, où les volumes augmentent, où les personnes se renouvèlent, où les bugs apparaissent, ... que la décision est bien souvent prise d'une migration vers un SGBD.

Dans cet exemple, nous considérons des données issues d'une campagne d'expériences menée en Neurosciences. Les données sont disponibles ici :
https://perso.liris.cnrs.fr/marc.plantevit/ENS/LIFBDW2/TP/Dataset_experiment_simplified.csv

Ce fichier contient le résultat d'une expérimentation où des sujets ont senti un ensemble d'odorants – une fois chacune – dans un ordre précis. L'expérience est divisée en plusieurs blocs (*block*) et le sujet bénéficie d'une pause entre chaque bloc. Pour chaque odeur, l'activité physiologique des individus est enregistrée : rythme cardiaque (FP), le rythme respiratoire (AR) et la température (ST). Pour chaque odeur, le sujet doit noter (sur une échelle entière de 1 à 9) une partie de son ressenti émotionnel lors d'une première session (hedonic1, intensity1, relaxation1, stress1, anxiety1) et l'autre partie (sur une échelle réelle de 0 à 1) au cours d'une autre session (intensity2, hedonic2, familiarity2, edibility2, disgust, surprise, neutral, pleasure, joy). Le CID est un numéro d'identification de l'odorant.

Travail demandé :

1) Étudiez les données. Dans un fichier texte, pour chaque colonne, décrivez la signification de l'attribut, son rôle dans les données, éventuellement la façon dont il est construit.
2) Importez ces données (disponibles en CSV) sous PostgreSQL dans une relation unique nommée « olfaction-csv », avec les types de données.
3) Faites l'inventaire des DF qui vous paraissent « plausibles » dans la sémantique de ces données ; vérifiez qu'elles soient bien satisfaites grâce à des requêtes SQL. Relevez des erreurs éventuelles et justifier votre choix lors de leur correction. On pourra éventuellement créer une relation répertoriant les tuples considérés comme erronés.
4) Faites également l'inventaire des DI, discutez leur pertinence.
5) Dressez un schéma Entité-Association modélisant les données. Traduisez le dans le modèle relationnel associé. Étudiez la forme normale de la modélisation proposée.
6) Implémentez la base de données correspondante avec les contraintes appropriées. On n'oubliera pas les contraintes pour limiter la valeur du domaine des attributs. Migrez les données par des requêtes SQL qui seront proprement conservées dans la documentation de la migration.
7) Écrire une requête qui retourne les sujets qui ont mis les notes les plus élevées pour stress1.
8) Écrire une requête pour identifier la configuration (sujet, odorant) où la fréquence cardiaque a le plus augmenté.
9) Écrire une fonction *main_statistics* qui étant donnée un odorant et un attribut physiologique, retourne les principales statistiques (valeurs min, max, moyenne, écart type).

10) De façon similaire, écrire une fonction une fonction *main_statistics_groupe* qui étant données un odorant, un attribut physiologique et un attribut relatif aux individus, retourne les statistiques pour chaque modalité de cet attribut. On testera par exemple, l'intensité (*intensity1*) de l'odeur *CID=12178* pour les sujets fumeurs et non-fumeurs.
11) A des fins d'analyse des résultats, on souhaite discrétiser les valeurs des attributs physiologiques de la première session en 3 sous-ensembles (peu, moyen, fort). Trois discrétisations des notes sont possibles :
    a) Échelles où {1,2,3} correspond à *peu*, {4,5,6} à *moyen* et {7,8,9} à *fort*.
    b) Percentile : 1er tiers pour peu, 2nd tiers pour moyen et 3ème tiers pour fort.
    c) Clustering : en trouvant automatiquement une partition en 3 clusters grâce à l'algorithme des k-moyennes : https://fr.wikipedia.org/wiki/K-moyennes.

    Notons que les discrétisations (b) et (c) se font pour chaque individu indépendamment des autres. Écrire les fonctions associées.
12) Résumez l'ensemble de votre import au sein d'un seul script ; une fois lancé, celui-ci crée ou récrée la base de données, la peuple à partir du fichier CSV. Ce fichier doit aussi contenir les requêtes et les fonctions et doit être suffisamment commenté.

Modalités d'évaluation du TP

Le TP est sur les séances restantes; chaque **binôme** rendra deux fichiers au plus tard le 18 décembre 2020, en utilisant TOMUSS (**un seul dépôt par binôme**) :

- Remplir dans la colonne « binôme », le nom du binôme (par convention, on prendra nom1_nom2 où nom1 précède nom2 dans l'ordre lexicographique).
- Un fichier « nom1_nom2_conception.pdf » dans lequel on trouvera les réponses aux questions 3, 4 et 5.
- Un fichier « nom1_nom2_script.sql » dans lequel on trouvera le script de la question 12 (si Tomuss n'accepte pas les fichiers avec extension .sql, alors utilisez une extension .txt).
- Chaque fichier devra contenir dans son entête, les noms, prénoms et numéros d'étudiant du binôme.