**Préparez des données pour un organisme de santé publique**

**Enoncé :**

L'agence Santé publique France souhaite améliorer sa base de données Open Food Facts et fait appel aux services de votre entreprise.

Cette base de données open source est mise à la disposition de particuliers et d’organisations afin de leur permettre de connaître la qualité nutritionnelle de produits.

Aujourd’hui, pour ajouter un produit à la base de données d'Open Food Facts, il est nécessaire de remplir de nombreux champs textuels et numériques, ce qui peut conduire à des erreurs de saisie et à des valeurs manquantes dans la base.

L’agence Santé publique France confie à votre entreprise la création d’un système de suggestion ou d’auto-complétion pour aider les usagers à remplir plus efficacement la base de données. Dans un premier temps, vous devez vous concentrer sur la prise en main des données, et d’abord les NETTOYER et les EXPLORER.

Vous êtes missionné sur le projet de nettoyage et exploration des données en interne, afin de déterminer la faisabilité de cette idée d’application de Santé publique France. Voici le mail de briefing de la part d’une de vos collègues, Alma, qui était en contact direct avec le client :

*Bonjour,*

*Le jeu de données Open Food Facts est disponible sur le site officiel (ou disponible à ce lien en téléchargement) : // https://fr.openfoodfacts.org/data*

*Les variables sont définies à cette adresse : //https://world.openfoodfacts.org/data/data-fields.txt Les champs sont séparés en quatre sections :*

*> Les informations générales sur la fiche du produit : nom, date de modification, etc.*

*> Un ensemble de tags : catégorie du produit, localisation, origine, etc.*

*> Les ingrédients composant les produits et leurs additifs éventuels.*

*> Des informations nutritionnelles : quantité en grammes d’un nutriment pour 100 grammes du produit.*

*Afin de simplifier ton approche, je te propose de commencer par établir la faisabilité de suggérer les valeurs manquantes pour une variable dont plus de 50% des valeurs sont manquantes.*

*Voici les différentes étapes pour nettoyer et explorer les données :*

*1) Traiter le jeu de données.*

*> Repérer des variables pertinentes pour les traitements à venir, et nécessaires pour suggérer des valeurs manquantes,*

*> Nettoyer les données en :*

*- mettant en évidence les éventuelles valeurs manquantes parmi les variables pertinentes sélectionnées, avec au moins 3 méthodes de traitement adaptées aux variables concernées,*

*- identifiant et en traitant les éventuelles valeurs aberrantes de chaque variable.*

*> Automatiser ces traitements pour éviter de répéter ces opérations*

*Attention, le client souhaite que le programme fonctionne si la base de données est légèrement modifiée ! ((ajout d’entrées, par exemple)*

*2) Tout au long de l’analyse, produire des visualisations afin de mieux comprendre les données.*

*Effectuer une analyse univariée pour chaque variable intéressante, afin de synthétiser son comportement.*

*Et un mot à ce sujet : le client nous demande de réaliser une présentation qui permet d’expliquer les analyses faites à un public néophyte. Sois donc attentif à la lisibilité : taille des textes, choix des couleurs, netteté suffisante, et varie les graphiques (box-plots, histogrammes, diagrammes circulaires, nuages de points…) pour illustrer au mieux ton propos.*

*3) Sélectionner / créer des variables à l’aide d’une analyse multivariée.*

*Effectuer les tests statistiques appropriés pour vérifier la significativité des résultats.*

*4) Rédiger un rapport d’exploration et une conclusion pour expliquer la faisabilité de l’application demandée.*

*5) Même si les données n’incluent pas de données personnelles, on doit expliquer dans une présentation en quoi ce projet respecte les 5 grands principes du RGPD. Santé publique France aimerait publier quelque chose sur le site Open Food Facts pour couper court aux questions sur le respect des RGPD que nous recevons parfois.*

*Bon courage !*

*Alma*

**Livrables :**

1. Un notebook contenant l’ensemble de vos traitements des données ainsi que vos analyses (une analyse univariée, multivariée, une analyse en composante principale (ACP), ainsi que les différentes questions de recherches associées).

2. Une présentation, à utiliser en support pour la soutenance, comprenant :

> Les opérations de nettoyage effectuées.

> La description et l'analyse univariée des différentes variables importantes avec les visualisations associées.

> L’analyse multivariée et les résultats statistiques associés, en lien avec l’idée d’application du client.

> 3 observations solidement étayées (graphes et/ou tests statistiques à l’appui au besoin) évaluant la pertinence et la faisabilité de l’idée d'application du client.

> La synthèse des différentes conclusions sur la faisabilité de votre projet.

**Référentiel d'évaluation :**

Compétences évaluées :

> Déterminer les objectifs du nettoyage des données et de la démarche de mise en œuvre afin de construire un jeu de données adapté à une problématique métier

> Effectuer des opérations de nettoyage sur des données structurées

> Effectuer des analyses statistiques univariées et multivariées

> Représenter des données grâce à des graphiques afin de justifier les analyses réalisées

Déterminer les objectifs du nettoyage des données et de la démarche de mise en œuvre afin de construire un jeu de données adapté à une problématique métier :

CE1 Vous avez défini les objectifs du nettoyage de données en fonction de la problématique métier

CE2 Vous avez défini la (bonne) démarche de préparation et nettoyage des données

Effectuer des opérations de nettoyage sur des données structurées, notamment l’identification d’outliers et le remplissage de données manquantes, dans le respect des normes en vigueur (RGPD) afin de proposer des données exploitables pour une problématique métier :

CE1 Vous avez éliminé les variables non pertinentes pour la problématique de l'application.

CE2 Vous avez proposé et justifié au minimum trois méthodes de traitement des valeurs manquantes, adaptées aux variables concernées (médiane, mise à 0, Iterative\_Imputer, KNN, suppression…).

CE3 Vous avez identifié, quantifié et traité les valeurs aberrantes de chaque variable, en tenant compte du contexte métier.

CE4 Vous avez traité les duplicats de variables et d'enregistrements.

CE5 Vous avez mis en œuvre une automatisation de certains traitements, à l'aide de fonctions et méthodes adaptées.

CE6 Vous avez veillé au respect des normes RGPD pendant les opérations de nettoyage. Dans ce projet cela signifie : d’expliquer les 5 grands principes RGPD et de montrer que votre travail n’a pas de lien avec les RGPD.

Effectuer des analyses statistiques univariées et multivariées, notamment des tests statistiques, à partir de données structurées et nettoyées afin de présenter leur distribution, valider leur cohérence et détecter des corrélations :

1) Evaluation de l’analyse statistique univariée :

CE1 Vous avez mis en évidence et analysé les éventuels outliers d’un point de vue statistique, afin de s’assurer s’il s’agit de valeurs aberrantes ou atypiques

CE2 Vous avez caractérisé correctement les distributions observées (uni, bi, multimodale)

CE3 Vous avez utilisé des métriques adaptées (moyenne ou médiane selon la dispersion)

CE4 Vous avez pu correctement définir le terme "quantiles".

2) Evaluation de l’analyse statistique bivariée et multivariée :

CE5 Vous avez présenté et analysé au minimum trois graphiques d’analyse bivariée (pair-plot ou graphique variables numérique / numérique, graphique variables numérique / catégorielle, heatmap de corrélation).

CE6 Vous avez expliqué, justifié et appliqué au minimum une méthode d’analyse multivariée descriptive sur le jeu de données (par exemple une ACP).

CE7 Vous avez expliqué, justifié et appliqué au minimum une méthode d’analyse multivariée explicative sur le jeu de données (par exemple une ANOVA, un test statistique).

Représenter des données grâce à des graphiques afin d’expliciter et justifier les analyses réalisées.

CE1 Vous avez identifié les cas où il était nécessaire de créer un graphique (en associant par exemple un graphe à la description de la distribution d'une variable).

CE2 Vous avez réalisé des graphiques lisibles (taille de texte suffisante, définition lisible).

CE3 Vous avez mis en œuvre au moins un de chacun des graphiques suivants : box-plot, bar-plot, pie chart, histogramme, scatter plot.