Mission - Préparez des données pour un organisme de santé publique

Comment allez-vous procéder ?

Cette mission suit un scénario de projet professionnel. Vous pouvez suivre les étapes pour vous aider à réaliser vos livrables.

Nous vous conseillons de diviser votre temps de la façon suivante :

- Prendre 30 min avant de démarrer :
    - lire toute la mission et ses documents liés
    - prendre des notes sur ce que vous avez compris
    - consulter les étapes pour vous guider
    - préparer une liste de questions pour la session Q/A avec votre professeur
    - Télécharger le jeu de données sur votre ordinateur

- 15 min : Q/A avec votre professeur.
- 60 min :Nettoyage de données (selection des données qui vous intéressent en lignes et colonnes/filtrage)
- 15 min : Premier nettoyage de notebook (ne laisser pas cette partie à la fin)
- 15 min : Pause à 16h
- 45 min : Faire la seconde partie d'exploration et d'analyse (cette fois avec les données propres et nettoyées)
- 30 min : Nettoyage de notebook
- 30 min : Ajout de commentaires/Interprétation/Conclusion/Suggestions pour SANTE PUBLIC FRANCE

Prêt à mener la mission ?

L'agence Santé publique France souhaite **améliorer et comprendre 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. Mais aussi, plusieurs données ne sont pas nécessaires pour dire si un produit est de qualité ou pas, pourtant nous les stockons quand mêle.

L'agence Santé publique France confie à votre entreprise le nettoyage, la préparation et la visualisation des données afin de mieux analyser les produits vendus en France et leur qualité et de potentiellement garder par la suite que les informations les plus importantes sur les produits.

Le jeu de données Open Food Facts est disponible sur le site officiel (ou disponible à ce lien en téléchargement). Les variables sont définies à cette adresse. 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.

Voici les différentes étapes pour **nettoyer et explorer les données** :
1) **Traiter le jeu de données.**

- **Repérer des variables pertinentes**
- **Nettoyer les données** en :
    - mettant en évidence les éventuelles **valeurs manquantes parmi les variables pertinentes sélectionnées**
    - traitant les valeurs manquantes
    - identifier et traiter les éventuelles **valeurs aberrantes** de chaque variable.

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 des analyses multivariées en montrant les interactions des variables entre elles.

Pour cette analyse nous devons donc être attentif à la lisibilité : taille des textes, choix des couleurs, netteté suffisante, et variez les graphiques (boxplots, histogrammes, diagrammes circulaires, nuages de points…) pour illustrer au mieux les liens entres les données.

L'intention de ce projet est de permettre à l'étudiant de **mettre en œuvre** des compétences de Data Analyst : cleaning, préparation des données et analyse exploratoire avec utilisation des **différentes librairies** pour générer les graphiques, **PLotly, Maplotlib**. De nombreuses réponses au projet sont possibles.

Voici un aperçu des étapes suggérées (Les apprenants ne sont pas tenus de réaliser les étapes du projet dans cet ordre). Cette section met également en évidence les difficultés potentielles. Veuillez noter :

Milestone 1 : Nettoyage – Filtrage des features et produits

20% progression

- **Livrable** :
  - Notebook de nettoyage, partie filtrage des données et produits
- **Problèmes et erreurs courants** :
  - Garder trop de variables, qui complexifierait le nettoyage et l'analyse
  - Se limiter à un nettoyage « technique » (taux de remplissage) sans se poser la question des besoins métiers
- **Recommandations** :
  - Dans ce projet nous traitons seulement les données concernant la france, trouver donc la variable et le filtre qui vont permettre cela
  - Lister l'ensemble des features du fichier, quantitatives (numériques) ou qualitatives (catégorielles)
  - Afficher le taux de remplissage des features, pour se rendre compte que la très grande majorité des features est très peu remplie
  - Combiner plusieurs approches :
    - **F**aire un choix fort de suppression des features pas suffisamment remplies. Un seuil de taux de remplissage autour de 40% est pertinent, à affiner éventuellement ensuite si nécessaire
    - Globalement ne pas hésiter à supprimer rapidement de nombreuses features, afin de faciliter les traitements et les analyses suivantes, quitte à rajouter après coup quelques features
    - L'étudiant devrait avoir gardé environ 8 à 10 features après cette étape
  - Supprimer les produits en double, selon un critère à définir par l'étudiant (code produit par exemple)

Milestone 2 : Nettoyage – valeurs aberrantes

40% progression

- **Livrable** :
    - Notebook de nettoyage, partie traitement des valeurs aberrantes
- **Recommandations** :
    - Dans ce projet il est important de privilégier une approche orientée métier pour traiter les valeurs aberrantes des features numériques :
        - Pour les valeurs qui doivent être entre 0 et 100g, une valeur négative ou >100g est aberrante
        - Pour l'énergie, vérifier sur Internet la valeur maximale possible pour 100g, pour l'utiliser comme seuil de valeur aberrante
    - Autres traitements possibles de détection de valeurs aberrantes :
        - La somme des ingrédients est > 100g
        - Attention satured_fat est inclus dans fat, idem pour sugar dans carbohydrates (Ne pas oublier ce contrôle)

Milestone 3 : Nettoyage – valeurs manquantes

60% progression

- **Livrable** :
    - Notebook de nettoyage, partie traitement des valeurs manquantes
- **Recommandations** :
    - Les traitements précédents ne doivent pas avoir pour conséquence de supprimer toutes les valeurs manquantes (NaN), sinon la compétence de cette étape ne pourra pas être mises en évidence
    - La stratégie de traitement des valeurs manquantes doit être adaptée à chaque feature en fonction du contexte métier du projet :
        - Remplissage par « 0 » si l'on estime que la saisie est bonne, mais que les produits ont dans cas en réalité une valeur nulle pour la feature : cela peut être le cas de la feature « fiber », car de nombreux produits ne contiennent pas de fibre
        - Remplissage par la médiane, à condition de le faire sur des produits homogènes, donc par catégorie de produit (pnns_group_1 par exemple ou autre variables catégorielles)

Milestone 4 : Analyse exploratoire – Analyse uni-variée et multi-variée

80% progression

- **Livrable** :

    - Notebook d’analyse exploratoire – partie analyse univariée et multi-variée
- **Recommandations** :
    - Une fois le fichier nettoyé, il est attendu de réaliser une analyse exploratoire complète, même si certaines analyses ont déjà été réalisées pour les besoins du nettoyage sur des données moins nettoyées
    - Analyse uni-variée :
        - Features numériques : il est attendu un describe(), des graphiques de type distribution et boxplot
        - Features catégorielles : il est attendu des graphiques de type barplot ou camembert
    - Analyse bi-variée :
        - Features numériques entre elles : par exemple scatterplot, pairplot, heatmap de corrélation
        - Feature numérique / feature catégorielle : par exemple feature « fat », avec boxplot selon le nutrigrade (A, B, C, D, E)

Milestone 5 : Analyse exploratoire – Interprétation

100% progression

- **Livrable** :
    - 1 ou 2 notebooks à vous de voir comme vous souhaitez vous organiser
    - Première possibilité :
        - Un notebook pour montrer le nettoyage de donnée
        - Un notebook pour montrer l’analyse et la réponse à la problématique
    - Seconde possibilité :
        - Un notebook contenant les 2
    - nommage nom1_nom2.ipynb