


 PROJET À COMPLÉTER

Préparez des données pour un organisme de santé publique

MISSION COURS ÉVALUATION

 80 heures

Mis à jour le mercredi 15 janvier 2020

L'agence "Santé publique France" a lancé **un appel à projets pour rendre les données de santé plus accessibles**. L'agence souhaite faire explorer et visualiser des données, pour que ses agents puissent les exploiter.

Êtes-vous prêt à utiliser vos compétences en analyse de données pour améliorer la santé de vos concitoyens ? 😊



Voici un extrait de l'appel à projets :

L'agence Santé publique France souhaite rendre les données de santé publique plus accessibles, pour qu'elles soient utilisables par ses agents. Pour cela, nous faisons appel à

vous pour réaliser une première exploration et visualisation des données, afin que nos agents puissent ensuite s'appuyer sur vos résultats.

Votre analyse sera basée sur le jeu de données [Open Food](#) (ou disponible à [ce lien](#) en téléchargement). Les variables sont définies dans ce [document](#).

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.

L'analyse devra être compréhensible pour un large public. Seront notamment évalués des critères formels comme : la taille des textes, le choix des couleurs, la netteté, la diversité des graphiques.

Vous fournirez un prototype simplifié de votre exploration sous forme de page web. Vous utiliserez les fonctionnalités de graphiques interactifs pour rendre votre prototype aussi réaliste que possible.

Vous présenterez votre idée et le prototype de votre exploration au jury de l'appel à projets.

Après avoir lu l'appel à projets, voici les différentes étapes que vous avez identifiées :

1) **Traiter le jeu de données** afin de **repérer des variables pertinentes** pour les traitements à venir. **Automatiser ces traitements** pour éviter de répéter ces opérations.

Le programme doit fonctionner 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.

3) **Confirmer ou infirmer des hypothèses à l'aide d'une analyse multivariée descriptive et explicative. Effectuer les tests statistiques appropriés** (une méthode descriptive, et une méthode explicative) pour vérifier la significativité des résultats.

Dans ce projet, vous avez pour mission de réaliser une analyse exploratoire à l'aide de Python. Attention, cette phase d'exploration peut être longue ! C'est à vous de la limiter dans le temps. Référez-vous aux compétences évaluées dans le projet et à la durée suggérée pour vous donner un cadre de travail.

Livrables

Voici les livrables, extraits de l'appel à projets :

- Un **Jupyter Notebook** présentant votre nettoyage et votre analyse du jeu de données.
 - Ce support devra être autoporteur car il sera lu par le jury de l'appel à projets.
- Une **page web présentant votre analyse** à présenter lors de la soutenance orale. Vous pourrez utiliser le [package Voilà !](#) qui permet de transformer aisément un Jupyter Notebook en page web.
- Un **support de présentation** (type Power Point) pour votre oral qui comprendra :
 - une présentation de l'appel à projets
 - votre démarche de nettoyage et d'exploration des données

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_notebook*", "*P3_02_pageweb*", et ainsi de suite.

Soutenance

La soutenance se déroulera en visioconférence et durera 30 minutes.

Les 20 premières minutes sont consacrées à la présentation de votre analyse au jury de l'appel à projets (représenté par le mentor évaluateur) :

- Une présentation de l'appel à projets
- Votre démarche méthodologique de nettoyage et d'exploration de données
- Le prototype réalisé

Suivront ensuite 10 minutes de questions.

Ressources complémentaires

- Le [dossier Github](#) du package *Voilà !*, qui explique comment installer cet outil et vous montre des exemples d'utilisation dont vous pourrez vous inspirer.

montre des exemples d'utilisation dont vous pourrez vous inspirer.

- Un [article Medium](#) en anglais expliquant le principe de fonctionnement du package *Voilà!*

Référentiel d'évaluation

Compétences

Effectuer des opérations de nettoyage sur des données structurées

Critères d'évaluation

Le nettoyage des données est **complet** si :

- ☐ les éventuelles valeurs manquantes de chaque colonnes ont été identifiées, quantifiées et traitées
- ☐ les lignes dupliquées ont été identifiées, quantifiées et traitées
- ☐ au moins une fonction a été écrite, testée et utilisée pour nettoyer le jeu de données

Le nettoyage des données est **pertinent** si :

- ☐ une méthodologie de traitement des valeurs manquantes pour chaque colonne est justifiée et mise en oeuvre (ex : remplacer les valeurs manquantes d'une colonne par la valeur moyenne de la colonne)
- ☐ une méthodologie de traitement des lignes dupliquées est justifiée et mise en oeuvre (ex : les lignes doublons ont été supprimés)

Le nettoyage des données est **présentable** si :

- ☐ les fonctionnalités d'édition de cellule Markdown du Jupyter Notebook sont utilisées dans au moins trois cellules pour décrire les choix méthodologiques et rendre lisible le document (titres, mise en forme, alternance de cellule d'exécution de code Python et de cellule de texte explicatif)
- ☐ la démarche de nettoyage des données est visible dans la structure du document (découpage du document en partie avec des titres clairs et mis en évidence, des commentaires à l'intérieur

des parties pour expliciter la démarche, ...)

Effectuer une analyse statistique multivariée

L'analyse statistique multivariée est **complète** si :

- ☐ au moins une méthode d'analyse descriptive est appliquée sur le jeu de données (ex : ACP)
- ☐ au moins une méthode d'analyse explicative est appliquée sur le jeu de données (ex : ANOVA)
- ☐ au moins une fonction a été écrite, testée et utilisée pour effectuer une analyse statistique multivariée

L'analyse statistique multivariée est **pertinente** si :

- ☐ la méthode d'analyse descriptive appliquée sur le jeu de données est expliquée et justifiée
- ☐ la méthode d'analyse explicative appliquée sur le jeu de données est expliquée et justifiée

Communiquer ses résultats à l'aide de représentations graphiques lisibles et pertinentes

La communication des résultats à l'aide de représentations graphiques est **complète** si :

- ☐ au moins trois types différents de graphiques ont été utilisés (ex : histogramme, boîte à moustache, nuage de points)
- ☐ la justification des types de graphiques utilisés est explicitée dans le Jupyter Notebook.
- ☐ au moins une fonction a été écrite, testée et utilisée pour effectuer une représentation graphique

La communication des résultats à l'aide de représentations graphiques est **pertinente** si :

- ☐ les titres, valeurs des axes des abscisses et des ordonnées et légendes sont explicites
- ☐ au moins un graphique interactif est utilisé pour illustrer une analyse lors de la présentation.

La communication des résultats à l'aide de représentations graphiques est **présentable** si :

- ☐ les titres, valeurs des axes des abscisses et des ordonnées et légendes sont affichés de manière lisible

Compétences évaluées



Effectuer des opérations de nettoyage sur des données structurées



Communiquer ses résultats à l'aide de représentations graphiques lisibles et pertinentes



Effectuer une analyse statistique multivariée

OPENCLASSROOMS



ENTREPRISES



CONTACT



EN PLUS



Français



Télécharger dans
l'App Store

