Voyageons sans attestation

À travers cet exercice voyageons ensemble (si vous le voulez bien). À plus de 10km et sans attestation.

Pour cela vous devrez analyser un jeu de données sur les hébergements français classés. Vous trouverez ainsi des données sur les campings, hôtels, villages de vacances, etc. étoilés ou non.

Le fichier CSV de 21000 lignes a été récupéré sur DataGouv et contient 16 colonnes.

Cet exercice sera parcouru en classe commune et fera l'objet d'une note bonus pour l'examen final du cours *Python pour la datascience* pour 2021.

Consignes

Les analyses sont à privilégier en pandas, mais comme plusieurs solutions existent, potentiellement des questions peuvent être répondues en python simple.

Les questions sont dans une grande majorité indépendantes. Ce qui veut dire que vous pouvez faire une question sans avoir fait les précédentes. En revanche parfois au travers d'une questions vous découvrirez ou nettoierez les données de tel sorte que ça vous aide pour les questions suivantes.

Nous ferons la simplification de langage de dire que ce dataset contient toutes les données des hébergements français. **De plus à chaque fois qu'une question sera posée, seront attendus la réponse à la question mais aussi le code pour obtenir la réponse. La présentation des réponses devra être soignée.**

Questions

Q1. Pour la première question il est attendu que vous exploriez le jeu de données afin de bien comprendre ce qu'il contient. Dans cette question donnez une définition de ce que contient chaque colonne. Créez une DataFrame nommée `df` qui contiendra les données que nous allons analyser dans les questions suivantes.

> *Aide* — Pour vous aider à répondre à cette question vous pouvez calculer l'exhaustivité des colonnes suivantes par exemple : `TYPOLOGIE ÉTABLISSEMENT`, `CLASSEMENT`

Q2. Trouvez le top 3 des codes postaux qui accueillent le plus d'établissements. Donnez une explication aux résultats trouvés.

Q3. Calculez la capacité d'accueil moyenne des hôtels français pour chaque niveau d'étoiles (colonne `CLASSEMENT`).

Q4. Quel est le département le plus dense en établissements ?

Q5. Le but de cette question est de créer un choroplèthe de la France affichant en nuancier de couleur le nombre de camping par régions française.

> *Aide* — si vous êtes bloqués vous pouvez vous concentrer sur la France métropolitaine.
>
> Q5.1 Trouvez un dataset sur internet liant les codes postaux français aux 18 régions, téléchargez-le et chargez le dans une DataFrame.
>
> Q5.2 Créez une colonne `REGION` qui contient le numéro de la région qui correspond à la ligne et une colonne `REGION_NAME` avec le nom de la région.
>
> Q5.3 Pour créer le choroplèthe vous aurez besoin de récupérer les données GeoJSON des régions françaises qui se trouvent ici. Grâce à ces données affichez le choroplèthe. Pour cela vous êtes libres de la librairie Python à utiliser.

Q6. Dans cette question nous allons trouver tous les établissements à 4 ou 5 étoiles qui se trouvent à maximum 50 kms de Clermont-Ferrand.

> *Aide* — Nous allons utiliser la fonction `geocoder` de la librairie `geopy`.
>
> Q6.1 Créez deux nouvelles colonnes `LATITUDE` et `LONGITUDE` qui contiennent pour chaque ligne la latitude et la longitude de l'établissement grâce à `geopy`.
>
> Q6.2 Filtrer les lignes pour garder seulement les établissements pertinents pour notre questions
>
> Q6.3 Calculer une colonne `DISTANCE_CF` contenant la distance en kilomètre avec Clermont-Ferrand. Vous pouvez pour cette question calculer une distance à vol d'oiseau.
>
> Q6.4 Afficher sur une carte de la France métropolitaine tous les points GPS et un cercle de 50kms autour de Clermont pour visualiser nos résultats.

Q7. Question ouverte de Machine Learning : créer un modèle prédictif du nombre d'étoiles d'un établissement à partir de variables explicatives comme la `TYPOLOGIE`,

le `NOM COMMERCIAL`, la `CAPACITÉ` et les variables de `NOMBRE`.

Réponses

Vous répondrez en envoyant un email l'adresse christophe.blefari@gmail.com avec en pièce jointe un fichier `.ipynb` ou `.py` (à votre guise) ou un lien vers un Github public avant le lundi 26 avril 2021 12h.

La notation de cet exercice apportera des points bonus au module *Python pour la datascience* de manière significative.