Objets perdus

Cet exercice est un exercice à la maison noté et obligatoire comptant pour l'année 2022 du module Big Data 3. Vous pouvez faire cet exercice en binôme mais chaque soumission devra être individuelle — avec mention du binôme le cas échéant.

Nous allons dans cet exercice utiliser plusieurs jeu de données ouvertes — open data — de la SNCF.

Vous devrez soumettre votre réponse avant le 8 avril 2022 minuit en envoyant un mail à christophe.blefari@gmail.com détaillant votre réponse. Un notebook `jupyter` sera bien entendu accepté. Toutefois, vous êtes libres du format de soumission tant qu'il contient :

- Du code Python fonctionnel exécutable facilement
- Des explications mise en formes et lisibles

Les données

Nous allons utiliser les jeux de donnée des objets trouvés et déclarés perdus en gares.

Le fichier `objets-trouves-restitution.csv` fait ~91Mo et contient les données des objets qui ont été *trouvés* et parfois restitués. Il y a 7 colonnes dans ce fichier.

Le fichier `objets-trouves-gares.csv` fait ~156Mo et contient les données des objets qui ont été *déclarés perdus* par les usagers de la SNCF. Les usagers de la SNCF peuvent signaler des objets perdus par exemple via le site troov.com. Vous pouvez visiter le site pour comprendre comment le jeu de données et construit.

Questions

1. Expliquer et définir chacune des 7 colonnes du fichier des objets trouvés.

2. Charger le jeu de données en *pandas*. Remarquez-vous quelque chose au moment de chargement ?

3. Décrire le jeu de données : donner les types, étudier les valeurs nulles, la cardinalité des différentes catégories, etc.
4. Ce jeu de données contient deux colonnes de dates. Convertir ces colonnes en format de date plus utilisable pour la suite. C'est-à-dire un format où il est plus simple de récupérer l'année, le mois ou le jour par exemple.
5. Donner la date de début du jeu de données.
6. Dans cette question vous effectuerez l'analyse sur le jeu de données des objets trouvés mais aussi des objets déclarés perdus.
    a. Donner l'année, le mois et le jour de la semaine où il y a le plus d'objets trouvés et déclarés perdus. Créer des graphique en colonnes affichant le résultat.
    b. Tracer trois courbes affichant l'évolution du nombres d'objets déclarés perdus, trouvés et restitué depuis le début du jeu de donnée.
7. Trouver les 3 gares où il y a le plus d'objets restitués en pourcentage des objets trouvés.
8. L'objectif de cette question est d'afficher sur une carte de France un diagramme en bulle avec le volume d'objets déclarés perdus par gare
    a. Calculer le volume d'objets déclarés perdus par gare
    b. Associer à chaque gare une position GPS pour pouvoir l'afficher sur la carte
    c. Afficher la donnée sur la carte — si plusieurs bulles se superposent trop sur une ville vous pouvez aussi essayer de faire un affichage différent ou de fusionner les bulles
    d. **Bonus —** Ajouter un sélecteur de "type d'objets" pour avoir une carte par catégorie
9. Question libre — Dans cette question vous avez carte libre. Visualisation, calcul ou transformation. Creuser et trouver quelque chose d'intéressant dans la donnée.
10. Créer un modèle de machine learning qui pourra prédire le nombre d'objets perdus en 2022. Cette question a les contours volontairement flous, c'est à vous de les définir.
11. **Bonus —** Produire une image PNG au format 2000 x 2000 qui affiche les résultats à toutes les questions précédentes.