TP Pandas

Pour apprendre à manipuler pandas, nous partirons d'un jeu de données open data hébergé sur le site : data.gouv.fr

Il s'agit d'une liste de fréquentation au cours du temps de différents musées franciliens. Premières étapes :

- Déposer les fichiers envoyés « freq_musees.csv » et « communes.csv » dans votre dossier de travail Jupiter

- Charger le fichier freq_musees.csv dans un dataframe nommé « df »

- Afficher un extrait du dataframe

- Afficher les 13 premières lignes du dataframe

- Afficher seulement les colonnes « etablissements » et « localite » à l'écran

- Compter le nombre de musées disponibles dans ce dataframe

- Le premier row est pas terrible, enlevez-le

- Nous ne nous intéresserons pas à certaines des colonnes du dataframe, autant les enlever. Enlever les colonnes « grat_06 », « grat_07 », « grat_08 », « grat_09 », « grat_10 », « commentaires », « wgs84 » en créant un nouveau dataframe « df_clean »
- Créer un nouveau dataframe « df_test » d'après « df_clean » contenant uniquement les colonnes « num_ref », « etablissements » et « total_06 ».
- Vérifier le type de la colonne « total_06 » de « df_test ». Que constatez-vous ?
- Créer un data frame df_test2 à partir de df_test

- Caster la colonne « total_06 » en int. Que constatez-vous ?
- Le format du fichier csv de départ n'est pas très propre (par exemple, le nombre 288179 est écrit avec un espace dans le fichier : 288 179). Au moment du cast, Pandas considère ce nombre comme un string et le passe à null, ce que nous voulons éviter. Nous devons donc modifier la colonne pour enlever les espaces. Faites-le sur le data frame df_test.
- Enfin, passez le type de « total_06 » de « df_test » à Integer, le problème devrait être réglé.
- Maintenant que vous avez compris comment faire le cast proprement sur une colonne, créez un nouveau dataframe « df_clean_and_cast » en appliquant le cast sur toutes les colonnes le nécessitant à partir de « df_clean » (qui est complet).
  Liste des colonnes à caster : « total_06 », « total_07 », « total_08 », « total_09 », « total_10 », « evolution_07_06_en », « evolution_08_07_en », « evolution_09_08_en », « evolution_10_09_en »
- Trouver combien il y a de musées parisiens
- Afficher par ordre décroissant la liste des musées les plus fréquentés pendant l'année 2010
- Afficher par ordre décroissant la liste des musées parisiens les plus fréquentés pendant l'année 2010
- Afficher la liste des musées ayant une fréquentation supérieure à 1000000 millions de personnes en 2010
- Afficher la liste des musées ayant une fréquentation entre 500000 et 1000000 millions de visiteurs en 2010
- Créez un nouveau dataframe « df_enrich » à partir de « df_clean_and_cast » contenant une colonne supplémentaire : la somme de toutes les fréquentations de l'année 2006 à 2010 (somme_freq)
- Créez un nouveau dataframe « df_enrich2 » à partir de « df_enrich » écartant tous les musées dont la fréquentation n'a pas été remontée au moins une année (de sorte à écarter tous les « null »). Notez ensuite le nombre de musées suite au filtre appliqué.

- Calculez la moyenne de fréquentation sur les 5 années pour chacun des musées et ajouter une nouvelle colonne « moy_freq » au dataframe « df_enrich2 »
- Créez un nouveau dataframe « df_enrich3 » à partir de « df_enrich2 » ne gardant que les musées qui ont une fréquentation qui a augmenté d'année en année de 2006 à 2010. Affichez les ainsi que leur nombre.
- Renommer la colonne « localite » du « df_enrich3 » en « ville »
- A partir du « df_enrich2 », créer un nouveau dataframe « df_enrich4 » et calculez l'évolution de fréquentation des musées entre les années 2006 et 2010 dans une nouvelle colonne « evolution_10_06_en »
- Créez un nouveau dataframe « df_enrich5 » à partir de « df_enrich4 ». Créez trois nouvelles colonnes à partir de la colonne « num_ref ». Une colonne « departement_id » contenant les deux premiers chiffres, une colonne « code_insee » contenant les 5 premiers et une colonne « num_musee_insee » contenant les deux derniers chiffres.
- Le fichier communes.csv contient la liste des communes de France.
- Créez un dataframe « df_communes » à partir du fichier communes.csv et ne gardez que les champs « code_insee », « commune », « département », « region », « population »
- Renommer la colonne « Code INSEE » en « code_insee » pour avoir le même nom que sur df_enrich5
- Faire un join entre le dataframe « df_communes » et le « df_enrich5 » sur la colonne « code_insee » dans un nouveau dataframe « df_join ».
- Faire un dataframe « df_group_by_code_insee » contenant la somme des fréquentations des musées par code_insee
- Enregistrer le dataframe « df_group_by_code_insee » dans un fichier csv