Projet Netflix - Partie 3

Pour cette partie chargez à nouveau les données avec pd.read_csv() mais cette fois-ci n'utilisez pas l'option index_col.

19. Avec Workbench créez la base Netflix et les tables destiner à accueillir les données structurées

Veuillez à choisir le bon types et la bonne taille pour chacun des champs.

- Le champ date_added au format datetime
- Le champ duration au format int

Le nom des champs des tables et le noms des colonnes des dataframes devront être identique pour faciliter l'intégration des données.

20. Créez un dataframe nommé « show_listed_in » ayant pour seul colonne « show_id » et « listed_in ».

Action : Chaque cellule de la colonne « listed_in » ne devra contenir qu'une seul et unique catégorie de film. Si un film appartient à plusieurs catégories alors son ID apparaitra dans la colonne show_id autant de fois qu'il possède de catégorie.

Indices :

- Vous devez travailler sur un dataframe ou il n'y a pas de valeur manquante.
- Pour le premier film du dataframe créez une liste de ces catégories, puis créez une liste de même longueur contenant seulement l'ID du film qui se répète pour chaque élément de cette liste.
- Utilisez ces deux liste pour créer le dataframe « show_listed_in »
  df = pd.DataFrame({'show_id': liste1, 'listed_in': liste2}, columns = ['show_id', 'listed_in'])
- Répétez l'opération pour tous les films et enrichissez le dataframe au fur et à mesure avec la méthode append().

21. Créez un dataframe nommé « listed_in » ayant pour colonne « listed_in_id » et « listed_in ».

Action :

- La colonne listed_in devra contenir le noms des catégories une seule et unique fois.
- Créer ce dataframe en utilisant drop_duplicates() sur show_listed_in (conserver show_listed_in)
- Utiliser reset_index() et rename() pour créer la colonne listed_in_id
- Joindre listed_in_id au dataframe show_listed_in en utilisant merge() et supprimer la colonne listed_in du dataframe show_listed_in.

22. Faire de même avec les colonnes director et cast en créant des dataframe qui leur sont dédier.

- Une fois fait, supprimez les colonnes listed_in, director et cast du dataframe initial.

- Modifier le type de duration et date_added

21. Intégrez les dataframes dans la base de données structuré.

Action :

- Avec la méthode create_engine() de sqlalchemy établir une connexion à la base Netflix nouvellement.
- Utiliser la méthode to_sql() de pandas pour inscrire les dataframes dans les tables préalablement créez. Mettre les argument suivant :
    - con = engine (pour spécifier quel connexion utiliser)
    - if_exists = 'append' (Si la table existe déjà dans MySQL alors on vient ajouté les données à la suite)
    - index = False (pour que l'index ne soit pas pris en compte l'or de l'écriture)

Exemple :

```
donnees.to_sql('nom_de_la_table', con=engine, index = False,
               if_exists = 'append')
```